TRANSFORMER-BASED INTEGRATIVE PATIENT REPRESENTATIONS FROM SINGLE-CELL RNA DATA

Benedikt von Querfurth[1], **Johannes Lohmöller**[1], **Jan Pennekamp**[1,2], **Tore Bleckwehl**[2], **Rafael Kramann**[2], **Klaus Wehrle**[1], **Sikander Hayat**[2]
[1] RWTH Aachen University
[2] University Hospital RWTH Aachen
{querfurth,lohmoeller,pennekamp,wehrle}@comsys.rwth-aachen.de
{tbleckwehl,rkramann,shayat}@ukaachen.de

ABSTRACT

Single-cell RNA sequencing (scRNA-Seq) is a powerful tool to explore cellular heterogeneity in healthy and diseased states, yet its translation into clinical insights has been limited. To bridge the gap between detailed cellular analysis and broader patient-level representations usable for phenotyping, we introduce a novel transformer-based architecture capable of embedding single-cell data into meaningful patient-level embeddings. This approach utilizes a self-supervised learning phase to construct integrative patient representations, which are then refined using contrastive learning techniques. On a dataset covering 7 million cells across 1223 individuals with diverse disease states, we show that learned embeddings are meaningful representations for a variety of downstream analytical tasks. Here, our approach proves robust against unbalanced datasets and shows indications of learning similarities between related diseases, such as COVID-19 and flu.

1 INTRODUCTION

Recent improvements in single-cell technologies have significantly improved our understanding of cellular diversity and its implications on human health. This progress is largely attributable to the ability to profile heterogeneous cell populations across various modalities at single-cell resolution. Driven by advances in deep learning and mass sequencing, downstream analytical approaches now facilitate detailed examination of gene expression (Soneson & Robinson, 2018), cellular pathways (Luecken & Theis, 2019b), composition (Buettner et al., 2021), and inter-cellular communication (Dimitrov et al., 2022), fostering not only an enriched understanding of cellular states (Dann et al., 2023) and disease mechanisms (Kuppe et al., 2022; Hoeft et al., 2023; Schreibing et al., 2022; Pekayvaz et al., 2024), but also holding promise for developing therapeutic interventions (Dann et al., 2024; Bartfai et al., 2012; Van de Sande et al., 2023; Amrute et al., 2022).

Furthermore, the application of single-cell methodologies extends beyond cellular and cluster analysis (Keener, 2019). Indeed, there is a pressing need to develop techniques to draw patient-level conclusions from such analyses, thereby enhancing the clinical utility of single-cell data for patient stratification (Leader et al., 2021; Khaliq et al., 2022), drug-target discovery (Van de Sande et al., 2023) and, overall, precision medicine. Although single-cell datasets are inherently high-dimensional, recent advances in representation learning (Vaswani et al., 2017) have succeeded in distilling these complexities into meaningful representations that preserve biological integrity on the cell level, such as scGPT (Cui et al., 2024) and Geneformer (Theodoris et al., 2023). Yet, the potential of such embeddings to also facilitate patient-level comparisons remains largely underexplored as there are only few concrete methods to learn and analyze such higher layer abstractions (Joodaki et al., 2024; Chen et al., 2020; Liu et al., 2024).

Here, we propose a novel method designed to generate patient-level embeddings from single-cell transcriptomics data. Sourcing data from multiple cells and the BERT architecture (Devlin et al., 2019), we refer to this method as multi-cell BERT (mcBERT). Our approach addresses the critical need for a cohesive, patient-centric representation. By employing a novel training pipeline that integrates a transformer encoder and multiple data sources, our method processes individual cell gene

counts to produce a compact, disease-capturing patient vector that condenses relevant information learned from single-cell gene expressions. Specifically, the training of our model involves an initial patient-level pretraining phase using a self-supervised data2vec methodology (Baevski et al., 2022), which does not require additional metadata beyond a donor identifier. This phase prepares the model to accurately map single-cell data to patient identities, setting the stage for downstream disease-specific comparative analyses. We fine-tune our pipeline using principles from contrastive learning to show that the learned representations are meaningful, specifically, to extract patient-level clinically relevant phenotypical information from individual tissues.

We validated mcBERT across multiple datasets—encompassing over 7 million cells from diverse tissues and pathological conditions (see Table 1)—demonstrate its efficacy in integrating data, mitigating batch effects, and delineating diseases within a disease-oriented latent space. This method to derive patient-level representations from raw cellular data marks a significant contribution toward the practical application of single-cell technologies in disease diagnosis and treatment stratification.

2 METHOD

To learn meaningful patient-level representations, mcBERT leverages the BERT framework (Devlin et al., 2019), well-known in Natural Language Processing (NLP). The architecture and training methods of mcBERT are designed for calculating a patient-level representation expressing the donors' phenotype based on its single-cell RNAseq expressions of a tissue. More formally, given a patient represented by its normalized single-cell RNAseq readouts $M_{sc} \in \mathbb{R}^{m_{genes} \times n_{cells}}$ consisting of n randomly selected single cells expressed by their m most Highly Variable Genes (HVG), the objective is to project the data to a patient-level vector $e \in \mathbb{R}^{d_{embd}}$ whose embedding space defines the donor-specific phenotype. That is, two patients A and B with similar phenotypes are projected to embeddings e_A and e_B exhibiting a high similarity score $S(x, y)$ with $S(e_A, e_B) = 1$, while a third dissimilar donor C should be projected to e_C with $S(e_C, e_A) = 0$ and $S(e_C, e_B) = 0$ (see Fig. 1).

2.1 ARCHITECTURE

After determining the organ-specific m HVGs and randomly selecting n cells, the first processing step of mcBERT is embedding the cells individually. Here, a cell embedding layer features a linear transformation with an input size of $2 + m$ and an output size of d with $d < m$, simplifying computation in subsequent attention layers. Here, the input dimensionality of $2 + m$ is determined by one one-hot encoding for the classification token and the masking token which is needed for the self-supervised learning stage and the m HVGs. Unlike traditional BERT (Devlin et al., 2019), we omit positional embeddings, making the model positionally invariant. Related methods like PILOT (Joodaki et al., 2024) and Harmony (Korsunsky et al., 2019) utilize Principal Component Analysis (PCA) for dimensionality reduction (Heumos et al., 2023). PCA exclusively relies on the statistical properties of the input, the different genes of numerous cells, and represents them through linear combinations. However, our approach replaces PCA with a fully connected layer without activation function, which serves as a learnable dimensionality reduction tool, similar to scBERT (Yang et al., 2022).

Next, we process the condensed cell representations through a transformer encoder. Specifically, we feed all of the sampled cells of one donor through multiple subsequent self-attention layers, capable of extracting important correlations across the cells and genes via multiple layers of abstraction.

Post-encoding, we obtain n cell embeddings. To consolidate these embeddings into a single patient-level vector, we apply global average pooling across all embeddings, a method also employed in vision transformers (Dosovitskiy et al., 2021) and offering comparable efficacy to using the $[CLS]$ token. The final embedding vector represents the aggregated gene expression profile, prepared for downstream tasks focused on disease-specific embeddings.

2.2 TRAINING

mcBERT undergoes a two-stage training process on this data. Initially, we employ an unsupervised learning strategy similar to data2vec (Baevski et al., 2022) to capture patient-level correlations across multiple single-cell sequences. Subsequently, the model is fine-tuned using a semi-supervised

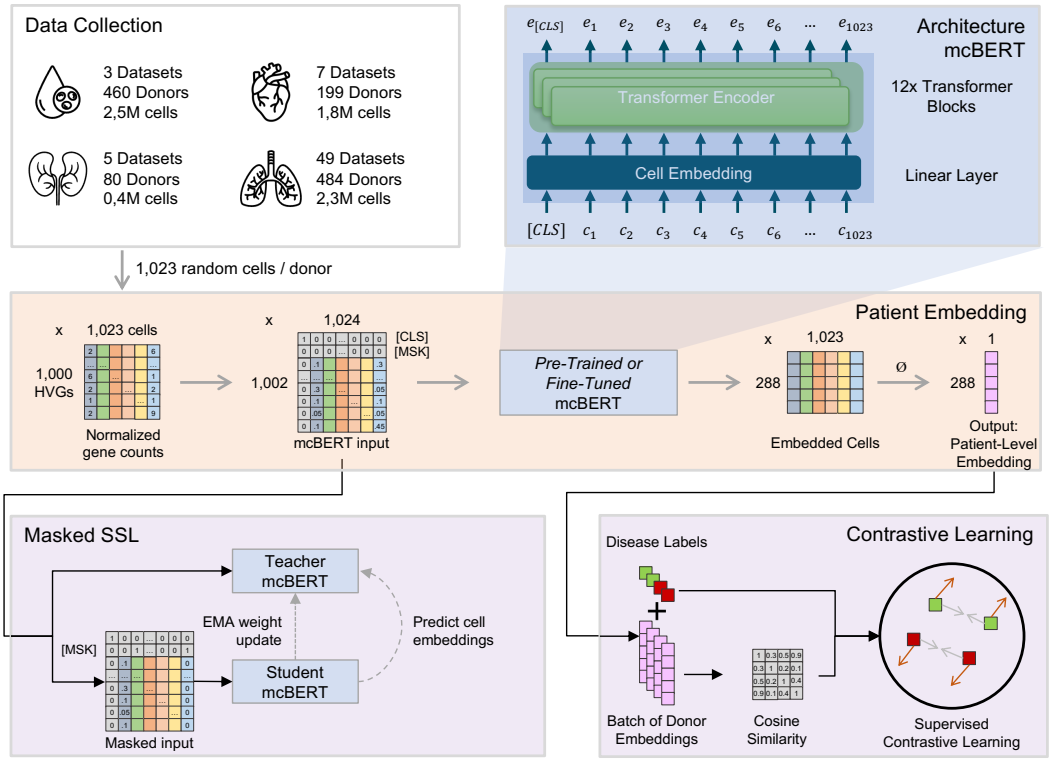

Figure 1: Overview of the patient-level training and embedding process using mcBERT using $n = 1023$ cells, the top $m = 1000$ HVGs, and an embedding dimensionality of $d = 288$.

contrastive learning approach, aiming to cluster donors with similar diseases closer in the embedding space, as indicated by higher cosine similarity.

Self-Supervised Masked Learning. Self-supervised learning (SSL) generally helps to discover underlying structures and correlations in datasets without a specified training target and is widely used for pretraining single-cell neural networks. Unlike cell-based transformer models that commonly employ masking strategies for training on gene correlations in single cells (Yang et al., 2022; Cui et al., 2024), our approach raises this masking strategy to the patient level, focusing on the training of cell interrelationships in a self-supervised manner. We adopt a masking strategy akin to data2vec (Baevski et al., 2022), aiming for an initial contextual understanding at the patient level involving over a thousand single cells.

Drawing parallels with Natural Language Understanding (NLU), the training of the student model utilizes a smooth L1 loss combined with the Adam optimizer and a learning rate of 1×10^{-5}. We randomly mask 15 % of input cells, following the suggestions of data2vec for text processing (Baevski et al., 2022). The loss function compares the student's embeddings of masked cells against the teacher's embeddings of the same, unmasked cells:

$$\mathcal{L}_{\text{data2vec}}(y_c, f_c^s(x)) = \begin{cases} \frac{1}{2}(y_c - f^s_{c(x)})^2/\beta & |y_c - f_c^s(x)| \leq \beta \\ |y_c - f_c^s(x)| - \frac{1}{2}\beta & \text{otherwise} \end{cases} \tag{1}$$

This self-supervised method, proven effective in other pretraining contexts (Assran et al., 2023), enables the model to predict masked cell representations accurately using the contextual information from remaining unmasked cells, fostering an understanding of patient-specific cell interrelations. This general understanding allows for further training on more specific tasks.

Supervised Contrastive Learning. Using the resulting patient-level embeddings e_A and e_B, we conduct patient comparisons by calculating the cosine similarity and distance:

$$S_C(e_A, e_B) = \frac{e_A * e_B}{||e_A|| * ||e_B||} \tag{2}$$

$$d_C(e_A, e_B) = 1 - S_C(e_A, e_B) \tag{3}$$

We then apply a contrastive cosine embedding loss function:

$$\mathcal{L}_{\text{cos\_sim}}(e_A, e_B) = \begin{cases} (1 - S_C(e_A, e_B))^2 & \text{if same disease} \\ max(0, (S_C(e_A, e_B)))^2 & \text{otherwise} \end{cases} \tag{4}$$

Here, for each patient in the dataset, we randomly select a second patient, such that it has a 50 % probability of belonging to the same disease class. As for unsupervised training, we select n random cells stratified by cell type. These cells are separately processed by the model, before comparison via the contrastive loss function.

While this standard contrastive learning setting is commonly applied for, e.g., embeddings of sentences as done in Sentence-BERT (Reimers & Gurevych, 2019), it often exhibits unstable training behavior and less globally meaningful embeddings, as only two random data samples are compared based on which the network is tuned. When not only the bilateral relationship of the training instances but all general labels (like diseases) are known, one can use all of the labels of one training batch and train on pushing apart patients with different diseases while pulling together the same instances, as proposed with the Supervised Contrastive (SupCon) loss (Khosla et al., 2021). Given the instability of traditional contrastive learning and the availability of disease labels, we employ the SupCon loss function. Hyperparameter testing indicates that choosing an AdamW optimizer (Loshchilov & Hutter, 2019), together with a learning rate of 1×10^{-5} and a batch size of 48, yields good performance and the most stable training behavior. Furthermore, with respect to the selected scRNA-seq datasets, a number of $n = 1023$ cells is a viable choice, as the selected datasets contain at least a median of 1023 cells per donor. However, if a donor is represented by less than 1023 cells, oversampling is used to compensate. For the number of top HVGs, $m = 1000$ has been experimentally shown to be a good value for cellular representativeness.

3 Results

mcBERT is designed to embed complex transcriptomics data from hundreds of sequenced single cells from an individual patient into a low-dimensional vector that encapsulates the donor's phenotype. To evaluate our method, we conduct analyses using diverse single-cell datasets derived from multiple tissues (see Table 1), highlighting the broad, universal applicability and flexibility of our approach.

Primary tasks include evaluating disease similarity directly through the mean cosine similarity across patient embeddings in which, optimally, a cosine similarity of 0 is achieved for patients with dissimilar diseases and a score of 1 with the same disease. Next, precise disease classification based on the local cosine neighborhood of a patient is evaluated via a k-nearest neighbor classifier using accuracy as a score. Finally, the overall clustering quality is analyzed using the Silhouette score (Rousseeuw, 1987) to estimate the global clustering qualities ranged from -1 (worst) to 1 (best) based on the separation of the clusters and the Adjusted Random Index (ARI) (Rand, 1971). Given the number of different diseases of the embedded patients, a hierarchical clustering with average linkage based on the cosine distance first separates the patients into clusters which are then compared with the actual disease labels. The ARI is delimited by 0 for worst clustering results and 1 for best clustering results. These metrics are designed to assess the model's capability not only to distinguish between diseases but also to generalize across different datasets and diverse biological characteristics of the analyzed tissues.

To validate the resulting model, we begin with experiments on single datasets to establish the baseline effectiveness of patient-level embeddings. Progressively, we increase the complexity of our evaluations by incorporating multiple datasets from varied laboratories to ascertain the robustness and generalizability of mcBERT. This strategy includes testing the model on previously unseen datasets and diseases, thereby verifying its ability to capture and generalize meaningful biological signatures rather than merely memorizing dataset-specific anomalies. Given the utilization of diverse datasets covering multiple tissues, these systematic experiments demonstrate the utility of the learned embeddings for summarizing findings from single-cell data to a patient level.

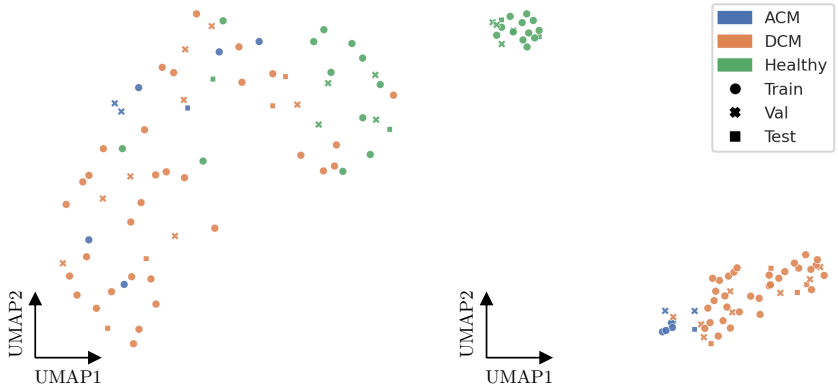

Figure 2: Baseline embeddings (left) versus mcBERT embeddings (right) of the first experiment only using one heart tissue dataset (Reichart et al. (2022), 566k cells from 70 donors) visualized using a cosine similarity UMAP. The healthy and diseased patients are well separated, with ACM and DCM forming two sub-clusters among the disease cluster, reflecting the feasibility of patient-centered embeddings using mcBERT.

3.1 Single Dataset

We evaluate the capability of raw gene counts from single cells to stratify patients based on disease status using a cardiac tissue dataset comprising 70 donors categorized as Healthy, Dilated Cardiomyopathy (DCM), and Arrhythmogenic Cardiomyopathy (ACM). The dataset is split into training, validation, and testing subsets (70 %, 10 %, and 20 %, respectively) in a stratified way to maintain proportional representation across disease categories. We conduct evaluations under 5-fold cross-validation. After training, we select the best non-overfitted model based on the validation set, which is subsequently used for embedding the donors of the test dataset. For this data, Fig. 2 shows the cosine similarity topology of one fold of the embeddings via UMAP dimensionality reduction, indicating good separability of healthy and diseased patients, as well as between different diseases.

Using the same cells per patient, we compare the performance of mcBERT against a baseline that uses the average cosine similarities of the raw gene counts, providing a naïve patient-level representation. The baseline approach exhibits poor clustering performance, with an ARI score of only 0.02 and a Silhouette score of 0.03, indicating a largely random similarity topology among patients. This is further evidenced by the high cosine similarity among patients with different diseases (0.717) that is close to the cosine similarity of patients of the same diseases (0.774), undermining the potential of raw inputs to effectively distinguish between disease states. Conversely, mcBERT demonstrates superior clustering capabilities, achieving an ARI of 0.766 and significantly improved mean cosine similarities within the same disease (0.889) and significantly reduced similarities across different diseases (0.555). The Silhouette score of 0.635 regarding the whole embedding space underlines these capabilities. These metrics illustrate the enhancement in the model's ability to embed disease characteristics meaningfully and stay consistent when expanding the experiment to leave-one-dataset-out cross-validation (compare Table 3).

This first experiment showed that raw gene counts of 1023 single-cell data potentially exhibit sufficient information to separate healthy from diseased donors, and the proposed mcBERT architecture with the training method is suitable to extract the relevant correlations inside the gene counts. Therefore, in the following, more complex experiments are conducted to explore the limits and potentials of both mcBERT and the training procedure.

3.2 Integrating Multiple Datasets

Mixing multiple datasets for training single-cell models requires addressing inherent challenges such as inconsistent cell-type annotations and technical batch effects (Korsunsky et al., 2019). To assess mcBERT's data integration capabilities, we use the union of all our heart tissue datasets and only standardize cell-type annotations but otherwise, perform no harmonization of gene counts or further

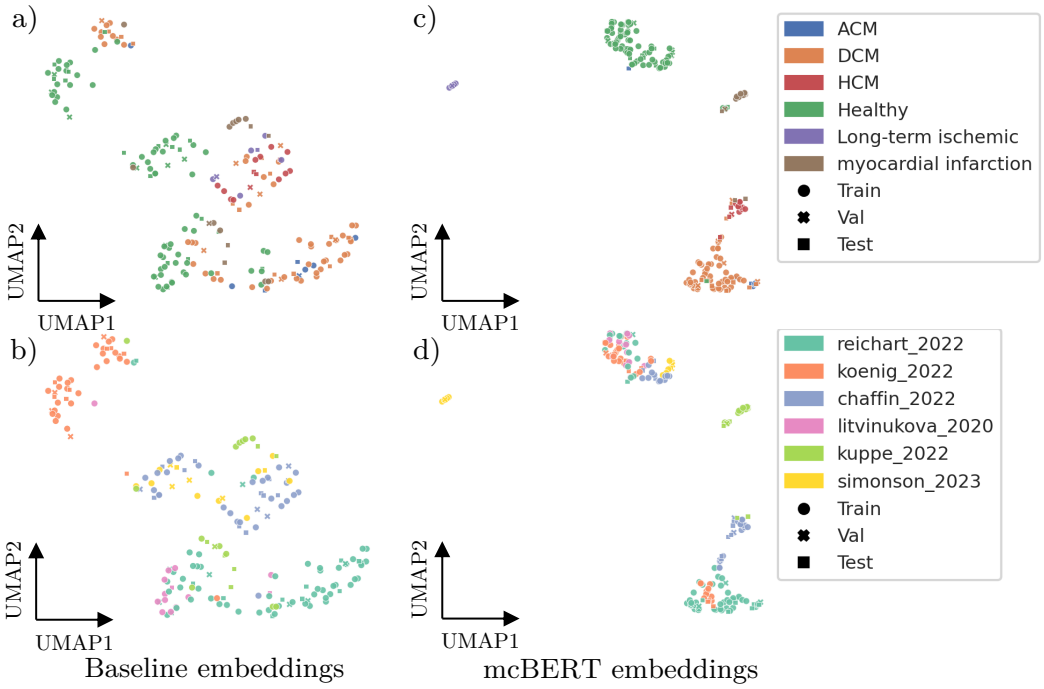

Figure 3: Baseline (a,b) versus mcBERT (c,d) embeddings of multiple heart tissue datasets colored by disease (a,c) and dataset origins (b,d). While local dataset-specific clusters are observed in the baseline, mcBERT correctly embeds the diseases across multiple datasets (see Table 3).

batch correction. We resort to the same stratified-splitting configuration and 5-fold cross-validation as before.

Fig. 3) visualizes the embeddings colored by disease and dataset, respectively. Comparing mcBERT with our baseline highlights the baseline's inability to contextualize disease embeddings beyond local dataset characteristics, evidenced by a ARI score of 0.05 which takes into account the global embedding in contrast to a seemingly well-performing model indicated by a k-NN accuracy of 0.72. This inability underscores the need for a more sophisticated approach, such as that offered by mcBERT.

Comparatively, the fine-tuned embeddings demonstrate improved clustering across datasets, forming well-clustered groups by disease state, with subclusters for related conditions, such as Hypertrophic Cardiomyopathy (HCM), DCM, and ACM for the heart datasets, or COVID-19 together with Influenza for the PBMC datasets. As these clusters consist of samples from heterogeneous datasets for those disease conditions covered in multiple datasets, this clustering indicates that mcBERT effectively captures and transfers cell-biological knowledge rather than technical batch artifacts or further dataset-specific features.

Applying mcBERT's pretraining and fine-tuning methodology to samples from kidney, lung, and Peripheral Blood Mononuclear Cell (PBMC) datasets (see Fig. 4) yields consistent results with those from heart tissues, validating our method's capability to generalize across different biological contexts. Despite the raw input baseline showing relatively high k-NN classification accuracy, significant improvements can be observed in all other metrics. For example, on average, the fine-tuned mcBERT increases the margin of the mean cosine similarity between the same and differently diseased patients from 0.063 to 0.455, demonstrating the model's effectiveness in overcoming batch effects and integrating multi-tissue datasets. However, this effect cannot be observed when using the model directly after the pretraining stage, showing both the necessity of the supervised learning stage and the inability to directly derive disease-related information in a self-supervised setting with data2vec.

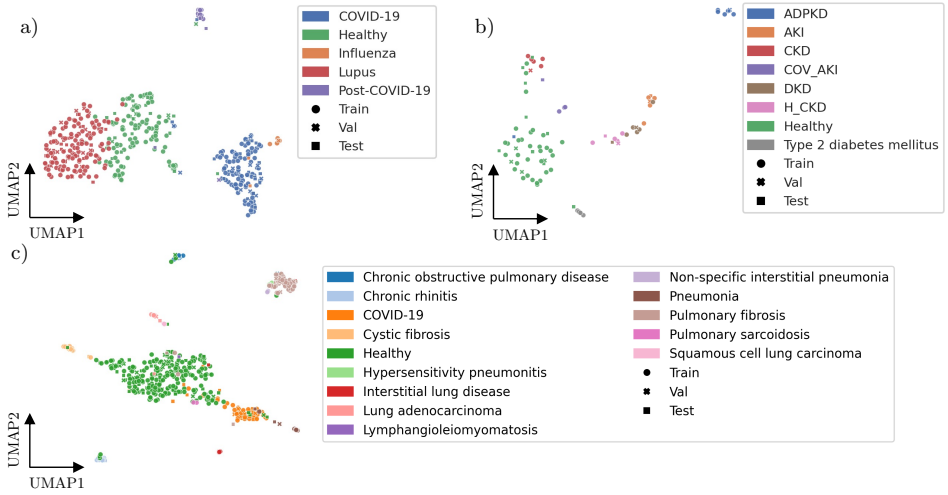

Figure 4: Patient-level mcBERT embeddings of different tissue datasets. a) PBMC, b) Kidney, and c) Lung colored with respect to the patient phenotype. Overall, we observe that mcBERT embeddings can be used to cluster phenotypically similar patients (see Table 3).

3.3 Cell Embeddings

Typical single-cell analysis frameworks (Wolf et al., 2018; Luecken & Theis, 2019b; Stuart et al., 2019; Svensson et al., 2020) use unsupervised clustering techniques to identify groups of cells with potentially similar properties such as cell lineage, cell state (Xu et al., 2023), and pathway activity (Aibar et al., 2017). These properties reflect the baseline and dysregulated disease states of groups of cells and are thus vital aspects for disease-centric patient-level neural networks. At the single-cell level, these properties are expressed through, for example, slight changes in gene expressions upon which clustering analyses can be based (Luecken & Theis, 2019a; Zhang et al., 2023). To enable mcBERT to learn biologically relevant cell representations, it processes hundreds of cells simultaneously and contextualizes them in the transformer Encoder. Furthermore, the variation in cell-type distribution across donors, both natural and introduced through training randomization, ensures that the embeddings generated by mcBERT are not merely reflections of statistical cell distributions.

Besides the transformer consuming multiple cells at once, the training is tailored to the concept of clustering cell types based on the similarity of gene-expression values. It involves a pretraining phase in which, through cell masking, mcBERT learns typical gene distributions across multiple cells for different phenotypes. This data2vec-like approach requires that the transformer discerns relevant information from surrounding cells to accurately predict the masked cells. This method's impact is evident when analyzing the integration of the gene-wise dataset; for instance, three independent PBMC datasets initially integrate poorly, showing a scaled iLISI score of 0.076 (larger is better, see Appendix A.3 for an overview of metrics). However, post pretraining, this score improves to 0.325 without any dataset-origin information of the donors while maintaining a scaled cLISI score of 0.987 (see Fig. 5). The benefit of this pretrained cell embedding is underlined by the little to no changes of it after fine-tuning (compare Fig. 5 mid versus right). To effectively predict the phenotypes of the patients, the fine-tuned mcBERT does not separate the cells again to e.g., pick up dataset-specific batch effects, but relies on using the integrated cell embeddings, emphasizing the distilled biological knowledge.

Thereby, mcBERT demonstrates its capability to transfer knowledge from samples within one dataset to others at the cell level by embedding similar cell types across datasets in a comparable manner. We attribute this capability primarily to the masking strategy employed during pretraining, which focuses on learning from the cellular context within a donor's sample to predict the embedding of masked cells, highlighting the importance of this self-supervised training phase.

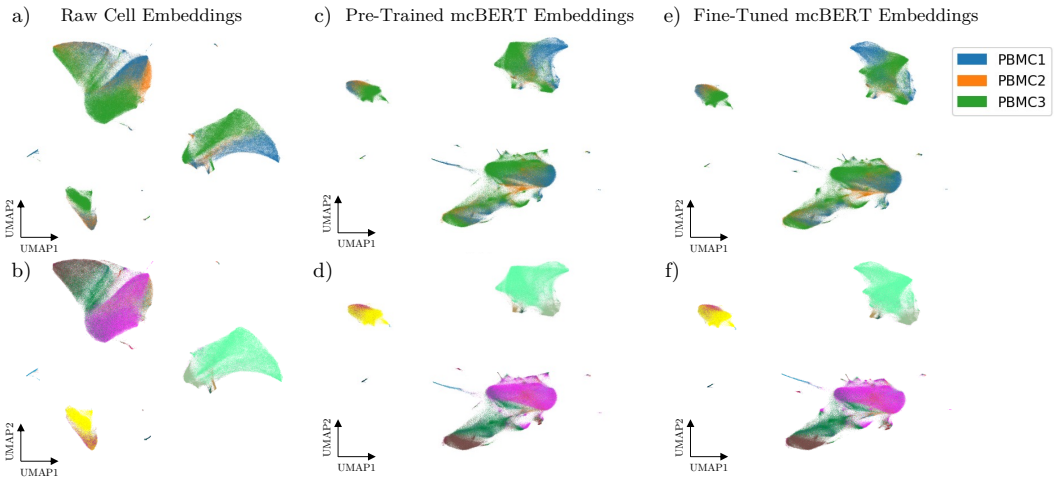

Figure 5: PBMC cell embedding of the first linear layer of mcBERT before training (a, b), after pretraining (c, d), and with subsequent supervised fine-tuning (e, f). Cells are colored by the dataset (a, c, e) and the cell types (b, d, f, legend omitted for clarity). The cell embeddings show clear dataset-wise disparities in the raw gene counts, which are corrected by the pretraining step and stay consistently corrected after fine-tuning.

Additionally, this step further helps to prevent early overfitting during fine-tuning and contributes to a meaningful patient-level embedding. By setting a low learning rate during fine-tuning, the model avoids re-learning dataset-specific technical drifts, ensuring that disease characteristics not present in the training dataset do not skew the embeddings. This approach results in consistent iLISI scores and UMAP visualizations between the pretrained and fine-tuned phases, as shown in Fig. 5).

4 Conclusion

Traditional scRNA-seq methods typically focus on cell-type level analyses, which leads to challenges in identifying sample-level features and comprehensive sample-vs-sample comparisons. To bypass these limitations, we introduce mcBERT, a novel method that mitigates this situation by abstracting from cell-level to patient-level information via embeddings that represent single-cell data per sample. These embeddings enable patient-level comparisons that can be used to differentiate samples based on their disease phenotype.

Specifically, mcBERT transforms normalized raw, high-dimensional single-cell gene expression data into manageable, low-dimensional representations per patient while preserving biological expressiveness. Thereby, and in combination with advancements from NLU and self-supervised learning, mcBERT enables meaningful phenotypical interpretations from single-cell datasets with direct clinical implications.

mcBERT demonstrates robust versatility across tissue types, extracting features from diverse datasets without prior data integration. Notably, the model clusters diseases with similar pathologies closely together, e.g., COVID-19 and flu, and abstracts from batch effects inherent in different datasets without requiring prior integration. We find this methodology to generalize well to novel data and disease classes, offering a significant advancement in the representation of complex, high-dimensional single-cell gene expression data. To our knowledge, mcBERT is one of the first methods to use transformers for systematically deriving patient-level insights from cell-level data. It shows potential for disease and phenotype detection and is readily available (see code). Although the high costs of scRNA-seq limit its routine clinical use, future integration of explainable AI techniques with mcBERT might contribute to an in-depth analysis of the model's decision-making processes. This analysis helps identify expressive cell types, states, and genes that can then be further utilized as marker genes for accurate and more cost-effective disease testing and drug discovery.

MEANINGFULNESS STATEMENT

We consider a meaningful representation to capture concise and useful abstractions from biological concepts, increasing their availability to (data-driven) analytical processes and beyond. Specifically, we show that our learned representations not only effectively describe the donor's disease state but, more importantly, that the embedded vector space is coherent in a way that similar disease states are correctly projected onto a similar representation, which directly increases meaningfulness for clinical use cases. This coherence is expressed by representations of similar diseases being similarly embedded by the model even without explicit training on such similarities.

ACKNOWLEDGMENTS

This work has been funded by the Excellence Strategy of the German federal and state governments.

REFERENCES

David J. Ahern, Zhichao Ai, and Mark Ainsworth et al. A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. *Cell*, 185(5):916–938.e58, 2022. ISSN 00928674. doi: 10.1016/j.cell.2022.01.012.

Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman et al. Scenic: single-cell regulatory network inference and clustering. *Nature methods*, 14(11):1083–1086, 2017.

Junedh M. Amrute, Xin Luo, Vinay Penna et al. Targeting the immune-fibrosis axis in myocardial infarction and heart failure. *BioRxiv*, pp. 2022–10, 2022.

Mahmoud Assran, Quentin Duval, Ishan Misra et al. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023. doi: 10.1109/CVPR52729.2023.01499.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu et al. data2vec: A general framework for self-supervised learning in speech, vision and language. arXiv, 2022.

Tamas Bartfai, Peter T. Buckley, and James Eberwine. Drug targets: single-cell transcriptomics hastens unbiased discovery. *Trends in pharmacological sciences*, 33(1):9–16, 2012.

Maren Buettner, Johannes Ostner, Christian L. Mueller et al. scCODA is a bayesian model for compositional single-cell data analysis. *Nature communications*, 12(1):6876, 2021.

Mark Chaffin, Irinna Papangeli, and Bridget Simonson et al. Single-nucleus profiling of human dilated and hypertrophic cardiomyopathy. *Nature*, 608(7921):174–180, 2022. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-022-04817-8.

William S. Chen, Nevena Zivanovic, David Van Dijk et al. Uncovering axes of variation among single-cell cancer specimens. *Nature methods*, 17(3):302–310, 2020.

Haotian Cui, Chloe Wang, Hassaan Maan et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 2024. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-024-02201-0.

Emma Dann, Ana-Maria Cujba, Amanda J. Oliver et al. Precise identification of cell states altered in disease using healthy single-cell references. *Nature Genetics*, 55(11):1998–2008, 2023.

Emma Dann, Erin Teeple, Rasa Elmentaite et al. Single-cell RNA sequencing of human tissue supports successful drug targets. *MedRxiv*, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Daniel Dimitrov, Dénes Türei, Martin Garrido-Rodriguez et al. Comparison of methods and resources for cell-cell communication inference from single-cell RNA-seq data. *Nature communications*, 13 (1):3224, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv, 2021.

Lukas Heumos, Anna C. Schaar, and Christopher Lance et al. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8):550–572, 2023. ISSN 1471-0064. doi: 10.1038/s41576-023-00586-w.

Konrad Hoeft, Gideon J.L. Schaefer, Hyojin Kim et al. Platelet-instructed SPP1+ macrophages drive myofibroblast activation in fibrosis in a CXCL4-dependent manner. *Cell Reports*, 42(2), 2023.

Mehdi Joodaki, Mina Shaigan, Victor Parra et al. Detection of PatIent-level distances from single cell genomics and pathomics data with optimal transport (PILOT). *Molecular systems biology*, 20 (2):57–74, 2024.

Amanda B Keener. Single-cell sequencing edges into clinical trials. *Nat. Med*, 25(9):1322–1326, 2019.

Ateeq M. Khaliq, Cihat Erdogan, and Zeyneb et al. Kurt. Refining colorectal cancer classification and clinical stratification through a single-cell atlas. *Genome biology*, 23(1):113, 2022.

Prannay Khosla, Piotr Teterwak, Chen Wang et al. Supervised contrastive learning. arXiv, 2021.

Andrew L. Koenig, Irina Shchukina, and Junedh Amrute et al. Single-cell transcriptomics reveals cell-type-specific diversification in human heart failure. *Nature Cardiovascular Research*, 1(3): 263–280, 2022. ISSN 2731-0590. doi: 10.1038/s44161-022-00028-6.

Ilya Korsunsky, Nghia Millard, Jean Fan et al. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods*, 16(12):1289–1296, 2019. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-019-0619-0.

Christoph Kuppe, Mahmoud M. Ibrahim, and Jennifer Kranz et al. Decoding myofibroblast origins in human kidney fibrosis. *Nature*, 589(7841):281–286, 2021.

Christoph Kuppe, Ricardo O. Ramirez Flores, Zhijian Li et al. Spatial multi-omic map of human myocardial infarction. *Nature*, 608(7924):766–777, 2022.

Blue B. Lake, Rajasree Menon, and Seth Winfree et al. An atlas of healthy and injured cell states and niches in the human kidney. *Nature*, 619(7970):585–594, 2023.

Andrew M. Leader, John A. Grout, and Barbara B. et al. Maier. Single-cell analysis of human non-small cell lung cancer lesions refines tumor classification and patient stratification. *Cancer cell*, 39(12):1594–1609, 2021.

Monika Litviňuková, Carlos Talavera-López, and Henrike et al. Maatz. Cells of the adult human heart. *Nature*, 588(7838):466–472, 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-2797-4.

Tianyu Liu, Edward De Brouwer, and Tony Kuo et al. Learning multi-cellular representations of single-cell transcriptomics data enables characterization of patient-level disease states. In *NeurIPS 2024 Workshop on AI for New Drug Modalities*, 2024. URL https://openreview.net/forum?id=UJhqLXsb5y.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv, 2019.

Malte D. Luecken and Fabian J. Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019a.

Malte D. Luecken and Fabian J. Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019b.

Yoshiharu Muto, Parker C. Wilson, Nicolas Ledru et al. Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. *Nature Communications*, 12(1):2190, 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-22368-w.

Yoshiharu Muto, Eryn E. Dixon, and Yasuhiro Yoshimura et al. Defining cellular complexity in human autosomal dominant polycystic kidney disease by multimodal single cell analysis. *Nature Communications*, 13(1):6497, 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34255-z.

Kami Pekayvaz, Corinna Losert, Viktoria Knottenberg et al. Multiomic analyses uncover immunological signatures in acute and chronic coronary syndromes. *Nature Medicine*, pp. 1–15, 2024.

Richard K. Perez, M. Grace Gordon, and Meena Subramaniam et al. Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus. *Science*, 376(6589):eabf1970, 2022. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abf1970.

William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971. ISSN 0162-1459, 1537-274X. doi: 10.1080/01621459.1971.10482356.

Daniel Reichart, Eric L. Lindberg, and Henrike Maatz et al. Pathogenic variants damage cell composition and single cell transcription in cardiomyopathies. *Science*, 377(6606):eabo1984, 2022. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.abo1984.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. arXiv, 2019.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 03770427. doi: 10.1016/0377-0427(87)90125-7.

Felix Schreibing, Monica T. Hannani, Hyojin Kim et al. Dissecting CD8+ t cell pathology of severe SARS-CoV-2 infection by single-cell immunoprofiling. *Frontiers in immunology*, 13:1066176, 2022.

Lisa Sikkema, Ciro Ramírez-Suástegui, and Daniel C. Strobl et al. An integrated cell atlas of the lung in health and disease. *Nature Medicine*, 29(6):1563–1577, 2023. ISSN 1078-8956, 1546-170X. doi: 10.1038/s41591-023-02327-2.

Bridget Simonson, Mark Chaffin, and Matthew C. et al. Hill. Single-nucleus RNA sequencing in ischemic cardiomyopathy reveals common transcriptional profile underlying end-stage heart failure. *Cell Reports*, 42(2):112086, 2023. doi: 10.1016/j.celrep.2023.112086.

Charlotte Soneson and Mark D. Robinson. Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, 15(4):255–261, April 2018.

Tim Stuart, Andrew Butler, Paul Hoffman et al. Comprehensive integration of single-cell data. *cell*, 177(7):1888–1902, 2019.

Valentine Svensson, Eduardo da Veiga Beltrame, and Lior Pachter. A curated database reveals trends in single-cell transcriptomics. *Database*, 2020:baaa073, 2020.

Christina V. Theodoris, Ling Xiao, and Anant Chopra et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Bram Van de Sande, Joon Sang Lee, Euphemia Mutasa-Gottgens et al. Applications of single-cell RNA sequencing in drug discovery and development. *Nature Reviews Drug Discovery*, 22(6): 496–520, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar et al. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Parker C. Wilson, Yoshiharu Muto, Haojia Wu et al. Multimodal single cell sequencing implicates chromatin accessibility and genetic background in diabetic kidney disease progression. *Nature Communications*, 13(1):5253, 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-32972-z.

F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.

Yang Xu, Rafael Kramann, Rachel Patton McCord, and Sikander Hayat. Masi enables fast model-free standardization and integration of single-cell transcriptomics data. *Communications Biology*, 6(1): 465, 2023.

Fan Yang, Wenchuan Wang, Fang Wang et al. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence*, 4(10): 852–866, 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00534-z.

Masahiro Yoshida, Kaylee B. Worlock, and Ni Huang et al. Local and systemic responses to SARS-CoV-2 infection in children and adults. *Nature*, 602(7896):321–327, 2022. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-021-04345-x.

Shixiong Zhang, Xiangtao Li, Jiecong Lin et al. Review of single-cell rna-seq data clustering for cell-type identification and characterization. *Rna*, 29(5):517–530, 2023.

A Appendix

A.1 Dataset Details

See Table 1.

A.2 Architecture Details

The exact architectural details of mcBERT are shown in Table 2. Before processing the cells by the model, the precomputed 1000 HVGs are selected per tissue datasets. After the relatively lightweight dimensionality reduction step consisting of a single linear layer, a normalization layer, and a dropout layer with a dropout rate of 10 %, most of the computation falls to the transformer encoder. Here, each of the 12 consecutive transformer blocks contains 12 attention heads and has a hidden dimensionality of 288.

The final global average pooling layer is used to reduce the contextualized cell matrix to a single patient-level vector. The learning of a donor's cell correlations during the self-supervised training phase is not aimed at a single patient-level vector, that is why the pooling layer is omitted during the pretraining phase.

A.3 Evaluation Metrics

We evaluate the trained models on three different application tasks: (1) the direct patient similarity of the same and different disease class; (2) the disease classification using k-NN in relation to the training dataset; (3) the clustering using a hierarchical clustering approach.

The patient comparison task is the most natural task the model can be evaluated on as the contrastive loss used during fine-tuning trains the model specifically on this task. For numeric evaluation, we separately evaluate the mean cosine similarity of the test against the training samples for donors from the same class and from different classes, respectively. That is, we evaluate how close the trained model maps the donors' embeddings next to other donors belonging to the same disease on the one hand and how far they are mapped away from different diseases on the other hand. The formal definitions of the two metrics are:

$$\overline{\overset{\sim}{S_C}} = \frac{1}{N * M} \sum_{i}^{N} \sum_{j \atop \text{disease}(i) \sim \text{disease}(j)}^{M} S_C(e_i^{train}, e_j^{test}) \tag{5}$$

$$\overline{\overset{\nsim}{S_C}} = \frac{1}{N * M} \sum_{i}^{N} \sum_{j \atop \text{disease}(i) \nsim \text{disease}(j)}^{M} S_C(e_i^{train}, e_j^{test}) \tag{6}$$

where S_C is the cosine similarity and e denotes the patient-level embedding. Ideally, a mean cosine similarity of 1 is achieved for patients with the same disease and 0 for different diseases.

Table 1: All datasets used for the evaluation studies of mcBERT.

Tissue	Name	#Donors	#Cells	Median #Cells/ Donor	Disease Labels
Heart	Litviňuková et al. (2020)	14	199,312	13,129	Healthy
	Simonson et al. (2023)	15	89,529	6,076	Long-term Ischemic, Healthy
	Kuppe et al. (2022)	20	189,349	8,494	Healthy, Myocardial Infarction
	Koenig et al. (2022)	38	216,972	5,298	Healthy, DCM
	Chaffin et al. (2022)	42	560,696	12,349	Healthy, HCM, DCM
	Reichart et al. (2022)	70	566,809	8,007	Healthy, DCM, ACM
Kidney	Muto et al. (2021)	5	19,985	3,804	Normal
	Wilson et al. (2022)	11	39,176	2,996	Normal, Type 2 Diabetes Mellitus
	adpkd1: Muto et al. (2022)	13	125,034	9,892	Control, ADPKD
	Kuppe et al. (2021)	15	51,849	1,493	Healthy, CKD
	kpmp: Lake et al. (2023)	36	200,338	4,995	DKD, H_CKD, AKI, Ref, COV_AKI
PBMC	PBMC1: Ahern et al. (2022)	124	836,148	6,312	COVID-19, Normal, Influenza
	PBMC2: Yoshida et al. (2022)	75	422,220	5,232	Normal, COVID-19, Post-COVID-19 Disorder
	PBMC3: Perez et al. (2022)	261	1,263,676	4,075	Normal, Systemic Lupus Erythematosus
Lung	Human Lung Cell Atlas (Full) Sikkema et al. (2023)	484	2,282,447	3,249	Normal, Pulmonary Fibrosis, Squamous Cell Lung Carcinoma, COVID-19, Lung Adenocarcinoma, Chronic Obstructive Pulmonary Disease, Pulmonary Sarcoidosis, Pneumonia, Lymphangioleiomyomatosis, Interstitial Lung Disease, Cystic Fibrosis, Chronic Rhinitis, Pleomorphic Carcinoma, Lung Large Cell Carcinoma, Hypersensitivity Pneumonitis, Non-Specific Interstitial Pneumonia
Total		1223	7,063,540		

Table 2: The mcBERT pipeline is subdivided into the different embedding steps.

	Output size	Number of parameters	Details of Step
Pre-Selected Genes	(1024, 1002)	-	-
Cell Embedding	(1024, 288)	288,864	1x Linear Layer
	(1024, 288)	576	1x LayerNorm
	(1024, 288)	-	1x Dropout, p=0.1
Transformer Encoder	(1024, 288)	25,282,944	12x Attention Blocks (12x Heads each)
(Pooling)	(1, 288)	-	1x Global Average Pooling
$\sum$	-	25,572,384	-

For the second application task, the disease classification, the k-NN algorithm is used with k=5 and the cosine distance as the distance metric to determine the closest related donor. As k-NN relies on known data points to classify new samples, the samples from the training sets were taken and act as a database to classify the new donors. After classification by the k-NN algorithm, the accuracy can be determined. However, if the disease is not known in the training database, the accuracy is always 0 % and would distort the mcBERT metrics and is therefore not taken into account.

Lastly, to evaluate the global topology of our patient-level embedding, we employ a hierarchical clustering approach using agglomerative clustering with an average linkage criterion and cosine similarity as the distance metric and the number of unique diseases as the number of clusters. Similar to the patient-level distance evaluation of PILOT (Joodaki et al., 2024), the adjusted random index (ARI) (Rand, 1971) of the clustering is calculated to determine the clustering result. In combination to the hierarchical clustering, the silhouette coefficient (Rousseeuw, 1987) of the embedding is calculated using the cosine distance.

Calculating the integration scores of the raw and embedded cells, respectively, is based on the Local Inverse Simpson's Index (LISI) as defined in (Korsunsky et al., 2019) using the *scib* Python library. Here, for better comparability, the cell-type LISI (cLISI) and integration LISI (iLISI) are scaled between 0 and 1. The given cLISI values are inverted so that 0 indicates a high variability of cell types and 1 indicates a good separation of cell types, instead of the original definition of cLISI where a higher value reflects a high degree of mixing of cell types.

A.4 Detailed Metrics of Experiments

Table 3 contains the cross-validated metrics for the conducted experiments. For the first experiment shown in the table, only the dataset by (Reichart et al., 2022) was used. The leave-one-dataset-out cross-validation (LOOCV) averages over all runs where a different testing set was used and the model trained for the remaining datasets. The remaining experiments incorporate all selected datasets shown in Table 1 for the specified tissue.

Table 3: Averaged results of the cross-validation metrics of all experiments. D2V denotes the pre-training stage using Data2Vec, FT the fine-tuning and LOOCV the leave-one-dataset-out cross validation.

Methodology		Silhouette Score	ARI	$\frac{\sim}{S_C}$	$\frac{\nsim}{S_C}$	k-NN Accuracy
Heart single dataset	Baseline	0.034	0.018	0.774	0.717	0.757
	D2V + FT	0.635	0.766	0.889	0.555	0.814
Heart LOOCV	Baseline	−0.063	0.065	0.700	0.666	0.483
	D2V + FT	0.608	0.824	0.858	0.319	0.586
Heart All	Baseline	−0.063	0.055	0.729	0.683	0.769
	D2V	−0.076	0.207	0.857	0.806	0.679
	FT	−0.126	0.000	1.000	1.000	0.718
	D2V + FT	0.561	0.789	0.868	0.312	0.819
Kidney All	Baseline	−0.188	0.103	0.554	0.504	0.682
	D2V	−0.101	0.132	0.524	0.487	0.592
	FT	−0.336	0.077	0.925	0.875	0.556
	D2V + FT	0.566	0.845	0.773	0.205	0.761
Lung All	Baseline	−0.272	0.110	0.284	0.223	0.695
	D2V	−0.204	−0.018	0.716	0.724	0.604
	FT	0.262	0.632	0.811	0.387	0.747
	D2V + FT	0.188	0.504	0.666	0.396	0.759
PBMC All	Baseline	0.078	0.388	0.900	0.808	0.785
	D2V	0.017	0.004	0.791	0.765	0.750
	FT	−0.032	0.423	0.941	0.647	0.723
	D2V + FT	0.412	0.538	0.855	0.428	0.828