DiSCERN - Deep Single Cell Expression ReconstructioN for improved cell clustering and cell subtype and state detection.

Fabian Hausmann[a,b,1], Can Ergen-Behr[a,1], Robin Khatri[a,b], Mohamed Marouf[a], Sonja Hänzelmann[a,b], Nicola Gagliani[c,d,e,f], Samuel Huber[c,d], Pierre Machart[a,b,*], Stefan Bonn[a,b,*]

[a] *Institute of Medical Systems Biology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany.*
[b] *Center for Biomedical AI, University Medical Center Hamburg-Eppendorf, Hamburg, Germany.*
[c] *Section of Molecular Immunology and Gastroenterology, I. Department of Medicine, University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany*
[d] *Hamburg Center for Translational Immunology (HCTI), University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany*
[e] *Department of General, Visceral and Thoracic Surgery, University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany*
[f] *Immunology and Allergy Unit, Department of Medicine Solna, Karolinska Institute, 17176 Stockholm, Sweden*

Abstract

Single cell sequencing provides detailed insights into biological processes including cell differentiation and identity. While providing deep cell-specific information, the method suffers from technical constraints, most notably a limited number of expressed genes per cell, which leads to suboptimal clustering and cell type identification. Here we present DISCERN, a novel deep generative network that reconstructs missing single cell gene expression using a reference dataset. DISCERN outperforms competing algorithms in expression inference resulting in greatly improved cell clustering, cell type and activity detection, and insights into the cellular regulation of disease. We used DISCERN to detect two novel COVID-19-associated T cell types, cytotoxic $CD4^+$ and $CD8^+$ Tc2 T helper cells, with a potential role in adverse disease outcome. We utilized T cell fraction information of patient blood to classify mild or severe COVID-19 with an AUROC of 81 % that can serve as a biomarker of disease stage. DISCERN can be easily integrated into existing single cell sequencing workflows and readily adapted to enhance various other biomedical data types.

Keywords: Single cell RNA-seq, RNA sequencing, imputation, cell clustering,

*Corresponding authors
Email addresses: `pierre.machart@neclab.eu` (Pierre Machart), `sbonn@uke.de` (Stefan Bonn)
[1] Authors contributed equally

cell type identification, expression reconstruction, Deep Learning, Machine Learning, auto encoder, batch effect correction, transfer learning, probabilistic modeling, reference atlas mapping, COVID-19, T helper cell, transcription factor analysis, single nuclear RNA-seq
2010 MSC: 00-01, 99-00

1. Introduction

Single-cell RNA sequencing (scRNA-seq) technologies allow the dissection of gene expression at single-cell resolution, which improves the detection of known and novel cell types and the understanding of cell-specific molecular processes [1, 2]. The extension of the basic scRNA-seq technology with epitope sequencing of cell-surface protein levels (CITE-seq), allows for the simultaneous surveillance of the gene and protein surface expression of a cell [3]. Another recent technological innovation was TCR-seq, which enables the simultaneous sequencing of essential immune cell features and the variable segments of T cell antigen receptors (TCRs) that confer antigen specificity ([4, 5]).

While several commercial platforms have enabled researchers to use single cell sequencing methods with relative ease and at reasonable cost, the analysis of the high-dimensional scRNA-seq data still remains challenging [6, 7]. The main technical downside of single cell sequencing that impedes downstream analysis is the sparsity of gene expression information and high technical noise. Depending on the platform used, single cell sequencing detects around three thousand genes per cell, giving almost an order of magnitude less genes detected than bulk RNA-sequencing [8]. The term 'dropout' refers to genes that are expressed by a cell but cannot be observed in the corresponding scRNA-seq data, a technical artifact that afflicts predominantly lowly to medium expressed genes, as their transcript number is insufficient to reliably capture and amplify them. This missing expression information limits the resolution of downstream analyses, such as cell clustering, differential expression, marker gene and cell type identification [9].

To improve the lack and stochasticity of gene expression information in single cell experiments, several in silico gene imputation methods have been designed based on different principles. Gene imputation infers gene expression in a given cell type or state, based on the information from other biologically similar cells of the same dataset. Several methods utilizing this principle have been developed [10], amongst them DCA, MAGIC, and scImpute [11, 12, 13]. DCA is an autoencoder-based method for denoising and imputation of scRNAseq data using a zero-inflated negative binomial model of the gene expression. MAGIC uses a nearest neighbor diffusion graph to impute gene expression and scImpute estimates gene expression and drop-out probabilities using linear regression. All of these algorithms use information from similar cells with measured expression of the same dataset for imputation.

While current imputation methods provide improved gene expression information, they still rely on the comparison of similar cells with largely absent gene

expression information. Genes that are not expressed in neighboring cells cannot be imputed, limiting the value of classical imputation. In an ideal case, it would be possible to obtain information of the expected true gene expression per cell, or at least expression information with less technical noise, to reconstruct the true expression at single cell level.

Recent work has shown the effectiveness of deep generative models (e.g. Autoencoders and Generative Adversarial Networks) to infer realistic scRNA-seq data and augment scarce cell populations using Generative Adversarial Networks [14] or the prediction of perturbation response using Autoencoders [15]. We hypothesized that a deep generative model could allow for the reconstruction of missing single cell gene expression information (low quality - lq) by using related data with more genes expressed (high-quality - hq) as a reference, a completely novel approach to gene expression inference (Figure 1A). In other words, lq data with many missing gene expression values and bad clustering could be transformed into data with few missing genes and improved clustering if the "style" of a related hq dataset could be transferred to it. In the best case, it would be possible to infer gene expression information for single cell data (lq) by using purified bulk RNA-seq data (hq), obtaining over ten thousand genes expressed per cell. We envision that this novel approach, when properly calibrated, is transformative for the analysis of single cell data, gaining deep mechanistic insights into data beyond what is currently measurable. It is important to note that the concept of using hq data to reconstruct gene expression in lq data is fundamentally different from classical imputation algorithms that infer gene expression based on nearby cells from the same dataset, as outlined above.

Based on the above considerations, we developed DISCERN, a novel deep generative neural network for directed single cell expression reconstruction. DISCERN allows for the realistic reconstruction of gene expression information by transferring the style of hq data onto lq data, in latent and gene space. Our experiments on real and simulated data show that DISCERN outperforms several existing algorithms in gene expression inference across a wide array of single cell datasets and technologies, improving cell clustering, cell type and activity detection, and pathway and gene regulation identification. To obtain deep insights into the cellular changes underlying COVID-19, we reconstructed single cell expression data of patient blood and lung immune data. While in our initial analysis [16] of blood data we detected few immune cell types, expression reconstruction with DISCERN resulted in the detection of 28 cell types and states in blood, including two novel disease-associated T cell types, cytotoxic $CD4^+$ and $CD8^+$ Tc2 T helper cells. Reconstructing a second COVID-19 blood dataset with disease severity information, we were able to classify mild and severe COVID-19 with an AUROC of 81 %, obtaining a potential biomarker of disease stage. DISCERN can be easily integrated into existing workflows, as an additional step after count mapping. Given that DISCERN is not limited by a predefined distribution of data, we believe that it can be readily adapted to enhance various other biomedical data types, especially other omics data such as proteomics and spatial transcriptomics.

2. Results

2.1. The DISCERN algorithm for directed expression reconstruction

We aim to realistically reconstruct gene expression in scRNA-seq data by using a related hq dataset. Ideally, this expression reconstruction algorithm should meet several requirements [7]. First, it needs to be **precise** and model gene expression values realistically. It shouldn't remove information of cellular identity to form 'average cells' or collapse different cell types or states into one. Second, the network should be **robust** to the presence of different cell types in hq and lq data, or an imbalance in their relative ratios. It shouldn't, for instance, 'hallucinate' hq-specific cells into the lq data. Lastly, the network should be directional, as the user should be able to choose the target (reference) dataset.

With these prerequisites in mind, we designed a deep neural network for directed single cell expression reconstruction (DISCERN) (Figure S1B) that is based on a modified Wasserstein Autoencoder [17]. A unique feature of DISCERN is that it transfers the "style" of hq onto lq data to reconstruct missing gene expression, which sets it apart from other batch correction methods such as [18], which operate in a lower dimensional representation of the data (e.g. PCA, CCA). To allow DISCERN to accurately reconstruct single cell RNA-seq expression based on reference data, the structure of the network had to be adapted in several ways. First, we implemented Conditional Layer Normalization (CLN) [19, 20, 14] to allow for directed expression reconstruction of lq data based on reference hq data (Figure S1B & S2). Second, we added two decoder heads to the network to enable it to model dataset-specific dropout rates and gene expression separately. Lastly, we extended DISCERN's loss function with a binary cross-entropy term for learning the probability of dropouts to increase general inference fidelity. Further algorithmic details of DISCERN can be found in the methods and figs. S1 and S2.

We first demonstrate DISCERN's capabilities to faithfully reconstruct gene expression using five pancreas single cell expression datasets of varying quality (Tables S1 and S2). The pancreas data is widely used for benchmarking and it is ideal to evaluate expression reconstruction for many cell types and sequencing technologies. We consider a dataset as hq when the average number of genes detected per cell (GDC) (e.g. smartseq2, GDC 6214) is much higher than in a comparable lq dataset (Table S2). Conversely, a dataset is lq when the average cell has lower counts and fewer genes expressed than a comparable hq dataset (e.g. indrop, GDC 1887). Throughout this text, we will name sequencing technologies with capital (e.g. Smart-Seq2, InDrop) and datasets with lower case first letters (smartseq2, indrop). We trained DISCERN on these five pancreatic single cell datasets and assessed the integration of data in gene space and the average expression reconstruction per cell type. While uncorrected data cluster by batch and not by cell type, DISCERN-integrated data show good batch mixing and clustering of cells by cell type across all five datasets (fig. 1B & fig. S2). To get a clearer picture of DISCERN's expression reconstruction capabilities we next calculated correlation coefficients of measured expression

between the lowest quality inDrop and highest quality Smart-Seq2 data, before and after expression reconstruction using DISCERN. The mean expression reconstruction of indrop-lq to smartseq2-hq and smartseq2-hq to indrop-lq data is very accurate, showing a Pearson correlation of $r = 0.95$ ($p < 0.001$), while mean expression correlation between uncorrected indrop-lq and smartseq2-hq data is only $r = 0.77$ due to strong batch effects (Figure 1C & D, Figures S3 and S4). The improved quality of indrop-lq data reconstructed to smartseq2-hq level is validated by the strong increase of genes expressed per cell, ranging from ≈2000 genes per cell in the uncorrected indrop-lq data to ≈6000 genes in the indrop-lq data after reconstruction (Figure S5).

We next investigated the effect of reconstruction of three cell type-specific genes, before and after correction across the five pancreas datasets (Figure S6). Insulin expression in the pancreas should be largely restricted to beta cells [21], which can be observed in the uncorrected smartseq2-hq and celseq2 datasets, while the indrop-lq batch shows a diffuse pattern of insulin expression across cell types (Figure S6A left panel). This diffuse insulin expression is corrected by reconstructing the smartseq2-hq expression pattern from the indrop-lq data (Figure S6A middle panel). In general, the expected specificity of insulin expression in beta cells can be recovered for all datasets when using DISCERN's reconstruction using the smartseq2-hq reference. Conversely, the reconstruction from hq to the indrop-lq reference results in diffuse insulin expression across all reconstructed datasets (Figure S6A right panel). We obtained similar results for the pancreatic acinar cell-specific gene REG1A and the delta cell-specific gene SST, both of which show diffuse expression across cell types in the uncorrected inDrop data and cell-specific expression after reconstruction using smartseq2-hq reference (Figure S6B & C). Interestingly, DISCERN can not only recover biological expression information, but it is also able to apply sequencing method-specific effects after reconstruction. The smartseq2-hq dataset, for instance, displays nearly no ribosomal protein coding gene expression after sequencing as previously reported by [8], while data sequenced using InDrop, Cel-Seq, or Cel-Seq shows prominent ribosomal protein coding gene expression (Figure S6D, left panel). When reconstructing smartseq2-hq data to indrop-lq data, ribosomal protein coding gene expression is re-instantiated (Figure S6D, right panel).

We further corroborated DISCERN's capability to integrate and reconstruct gene expression in the more complex difftec dataset (Tables S1 and S2), consisting of 14 single cell peripheral blood mononuclear cell (PBMC) datasets across a wide range of technologies. Similar to pancreas, the difftec dataset is widely used for benchmarking and it is ideal to evaluate expression reconstruction for even more cell types and sequencing technologies. The different single cell technologies show large variation in quality, with an GDC ranging from 422 in Seq-Well to 2795 in Smart-seq2. We trained DISCERN on these 14 PBMC single cell datasets and observed very good integration in gene space (Figure S7). We then reconstructed chromium-v2-lq (GDC 795) using a chromium-v3-hq reference (GDC 1514) and observed high mean gene expression correlation between the reconstructed and reference datasets (Figures S8 and S9). These results across 19 single cell datasets provide first evidence for the high-quality data in-

tegration and expression reconstruction that can be obtained with DISCERN.

2.2. Specific and robust gene expression inference

We next investigated the precision and robustness of DISCERN's expression reconstruction in more detail and compared DISCERN's performance to several state-of-the-art algorithms for expression imputation and data integration.

Since expression reconstruction can be seen as a generalization of expression imputation, we compared DISCERN to DCA, MAGIC, and scImpute, three state-of-the-art imputation algorithms [11, 12, 13]. Expression reconstruction can also be viewed as a batch correction task in gene space, which is why we additionally compared DISCERN to scGEN and Seurat [15, 18]. It is important to note, however, that neither Seurat nor scGEN were designed for the expression reconstruction task. Seurat and scGEN use a lower dimensional representation in which a linear transformation aligns different batches. Seurat uses canonical correlation analysis and scGEN uses the bottleneck layer representation of an autoencoder to calculate and apply linear transformations.

To investigate the precision of gene expression reconstruction, we created an artificial dataset by dividing the smartseq2-hq pancreas data into two batches, smartseq-lq and smartseq2-hq. In the smartseq-lq batch, the top one KEGG pathways per cell type were removed by setting the expression of genes contained in these pathways to zero, while the smartseq2-hq remained unaltered. Therefore, a reconstruction of smartseq-lq data using smartseq2-hq reference (reconstructed-hq) should ideally recover the smartseq-lq expression to its original state, prior to the removal of the genes. DISCERN is able to reconstruct the mean expression for all cell types, achieving a correlation $r = 0.99$ (Figure 2A). DCA ($r = 0.66$), MAGIC ($r = 0.34$), scImpute ($r = 0.80$), and Seurat ($r = 0.76$) have significantly lower correlation between the smartseq2-hq and reconstructed-hq gene expression (Figure 2A). scGen shows only slightly reduced performance ($r = 0.98$) compared to DISCERN, especially in the reconstruction of highly expressed genes (Figure 2A) and low abundant cell types (Figure S10, Megakaryocytes). We obtained similar results on the difftec dataset, with DISCERN ($r = 0.98$) outperforming DCA ($r = 0.47$), Magic ($r = 0.21$), scImpute ($r = 0.04$), Seurat ($r = 0.92$), and scGEN ($r = 0.94$) (Figure S10). To further investigate gene expression reconstruction specificity, we compared the correlation of reconstructed-hq to smartseq2-hq data after performing differential gene expression (DEG) for each cell type against all other cell types (Figure 2B, upper panel). DISCERN is able to recover the correct DEG t-statistics with a median correlation of 0.92, improving over state-of-the-art tools by more than 15 percentage points. In the corresponding experiment using the difftec dataset, DISCERN achieves a median correlation of 0.85, which is a 25 percentage point improvement over competing methods (Figure S11).

Since the genes were initially selected using KEGG gene set enrichment analysis, the reconstruction of the corresponding pathways was investigated by performing KEGG gene set enrichment analysis on the DEG results. DISCERN is able to recover the pathway expression enrichment scores with a median correlation of 0.93, exceeding the performance of Seurat and scGEN by more than

11 percentage points on median (Figure 2B, lower panel). In the corresponding experiment using the difftec dataset, DISCERN achieves a median correlation of 0.77, outperforming Seurat and scGen by more than 16 percentage points (Figure S12).

While DISCERN outperforms competing algorithms in expression and pathway reconstruction correlation, it achieves the second-best correlation for the DEG fold-change (FC) of reconstructed-hq to smartseq2-hq data for the pancreas (Figure S13) and reconstructed-hq to chromium-v3-hq difftec datasets (Figure S14). In both cases Seurat achieves slightly better correlation, which is due to the fact that DISCERN slightly underestimates FC in favor of superior DEG variance estimation.

Next, we show DISCERN's expression reconstruction robustness with respect to varying sizes of lq to hq data. It is conceivable to assume that a large amount of hq data would benefit the expression reconstruction of the lq data, which makes it important to understand at what ratio good results can be expected. Interestingly, DISCERN seems to be very robust across a wide range of smartseq2-lq to smartseq2-hq ratios, with correlations of 0.98 (ratio of lq/hq 0.14) to 0.93 (ratio of lq/hq 18.4), while the second-best performing algorithm scGen showed a 11 percentage point decrease in performance (0.82 for ratio of lq/hq 18.4) (Figure 2C, Figure S15). We observed similar results for the correlation of t-statistics, showing a slight dependence of DISCERN's performance on the lq/hq ratio (Figure S16). In general, all methods show better performance with a small ratio of lq/hq data, while DISCERN shows least dependence and outperforms other algorithms in the correlation of expression and t-statistics, especially in the case of high lq/hq ratio.

Another aspect of expression reconstruction robustness is the dependence of the algorithm on the cell type or cell state similarity of the lq and hq datasets. In the optimal case, DISCERN would not require that the lq and hq datasets have overlapping cell types to perform an accurate expression reconstruction, which is theoretically possible if the network learns the general gene-regulatory expression logic of the hq data (see discussion). To understand the dependence on dataset similarity, we removed a complete cell type, pancreas alpha cells, from the smartseq2-hq data and left the alpha cells in the smartseq2-lq data. We then additionally varied the number of common cells in the lq and hq data, starting with no overlapping cells (only alpha cells in the lq and all cells except alpha in the hq data) and ending with almost complete overlap (all cells overlap between the smartseq2-hq and -lq data, except for the alpha cells only present in lq data) (Figure 2D). When evaluating DEG correlation, DISCERN was the only method consistently achieving better performance than uncorrected data, outperforming Seurat and scGen by more than 15 percentage points (Figure 2D). Similarly, DISCERN was the only method consistently achieving better performance than uncorrected data in the FC correlation task (Figure S17).

We next took a closer look at the integration and expression reconstruction performance when no cell types overlap between the lq (alpha cells only) and hq (all other cells) data. Notably, Seurat seems to over-integrate cell types, mixing smartseq2-hq beta and gamma cells with reconstructed-hq alpha cells from

other batches (Figure S18), while scGEN and DISCERN keep the smartseq2-hq and reconstructed-hq exclusive cell types separate (Figure 2E & Figure S18). This over-integration seems to be causal for Seurat's poor DEG correlation performance ($r = 0.28$), while DISCERN ($r = 0.55$) is the only method achieving better performance than uncorrected cells ($r = 0.47$) (Figure 2F). Thus, DISCERN is able to keep existing expression correlations and improves the detection of cell type specific genes by reconstruction using an hq batch as reference. In conclusion, DISCERN is both a precise and robust method for expression reconstruction that outperforms existing methods by a significant margin.

2.3. Improving cell cluster, type, and trajectory identification

The comparison to competing methods provided evidence for DISCERN's superior expression reconstruction. Now, we will delineate how DISCERN's expression reconstruction improves downstream cell clustering, cell type and activity state identification, marker gene determination, and gene regulatory network and cell trajectory analysis.

To understand if cell-determining gene expression and pathways could be recovered with expression reconstruction, we used a single nuclear sequencing (sn-lq) and scRNA-seq (sc-hq) data pair that was prepared from the same liver metastasis biopsy [22]. We reconstructed sn-lq data using the sc-hq reference, obtaining reconstructed-hq data. While single nuclear sequencing provides reduced expression information in the average counts per cell as compared to scRNA-seq (Table S2) [22], it is still the method of choice to obtain cell-specific expression information when intact single cells cannot be recovered from a tissue (e.g. after tissue fixation or freezing). It is important to note that nuclear transcripts reflect current gene activity, which in part might not correlate with transcripts that have lifetimes of up to days. Before integration, the sn-lq and sc-hq datasets cluster by batch and not by cell type, while after expression reconstruction with DISCERN cells cluster by type and not by batch (Figure S19). This is reflected in an expression correlation of 0.49 (sc-hq vs. sn-lq) before and 0.93 after reconstruction (sc-hq vs. reconstructed-hq) (Figure S20). Reconstruction resulted in the expression of T cell receptor signaling genes in reconstructed T cells (Figure S21) and antigen presentation genes in macrophages (Figure S22), providing evidence that DISCERN faithfully recreates cell-determining genes and pathways based on the hq data.

It is intriguing to observe that many genes of the antigen presentation pathway in macrophages are not expressed in the sc-hq reference, most probably due to dropout (Figure S22). We rationalized that bulk RNA sequencing (RNA-seq) data of purified cell types (e.g. FACS sorted immune cells) is a suitable hq proxy for the expected gene expression per cell. RNA-seq data of purified cells is readily available from public repositories, making it possible to obtain thousands of purified immune cell RNA-seq samples (see methods). We therefore set out to increase cluster, cell type, gene regulatory network, and trajectory identification of scRNA-seq data by reconstructing gene expression using a related RNA-seq reference (Figure S23). For the scRNA-seq data we chose a cord blood mononuclear citeseq dataset (cite-lq) that was labeled with 15 antibodies (Table S3) to

allow for surface protein-based cell type discovery [23]. The CITE-seq information allowed us to confirm expression reconstruction by DISCERN in cases where gene expression is absent but protein expression and cell identity are validated via antibody labeling. For the RNA-seq data, we selected 9.852 purified immune samples (bulk-hq) and proceeded to reconstruct cite-lq (GDC 798) using a bulk-hq (GDC 13.104) reference to obtain reconstructed-hq data with DISCERN. We first investigated the correspondence of gene expression prior (cite-lq) and post reconstruction (bulk-hq) with antibody-based surface protein labeling of *CD3D*, *CD4*, *CD8A*, *CD2*, *B3GAT1*, *FCGR3A*, *CD14*, *ITGAX* and *CD19* (Figure 3A, Figure S24). For several proteins (CD8A, B3GAT1, CD4), the corresponding cite-lq gene expression was absent and cell type-specifically re-instantiated in the reconstructed-hq expression data with DISCERN (Figure 3A, Figure S24). In cases where cell type-specific gene and protein expression matched cite-lq data (*CD3D*, *CD14*) the expression in reconstructed-hq data was left unaltered (Figure S24). In some instances, we observed low cell type-specific expression in the cite-lq data (*CD8A*, *CD2*, *FCGR3A*, *CD19*) that matched protein expression (Figure S24). In these cases, gene expression was increased in the correct cell types in the reconstructed-hq data. In general, we observed increased agreement between cell type-specific surface protein and gene expression after reconstruction, showing that DISCERN doesn't invent or 'hallucinate' cell types but reconstructs the expected expression specific for each cell type. We further corroborated these results by selecting eight known cell type-specific cytosolic proteins and investigated their expression before and after expression reconstruction. *MS4A1* (B cells), *IL7R* ($CD4^+$ T cells), *MS4A7* (Monocytes), *GNLY* and *NKG7* (NK cells) showed consistent expression before and after reconstruction (Figure S25). The chemokine receptors *CCR2* (Monocytes, activated T cells), *CXCR1* (NK cells), and *CXCR6* ($CD8^+$ T cells) showed the correct cell type-specific expression only after expression reconstruction (Figure S25) [24]. It is notoriously hard to obtain cell subtype-specific information from blood mononuclear scRNA-seq data, especially for $CD4^+$ T helper cells due to their limited activation status in healthy individuals. This doesn't mean that polarized $CD4^+$ T helper cells do not exist in healthy blood, as they are commonly detected after stimulation using FACS (Table S3) [25]. This lack of resolution in scRNA-seq impedes clustering, marker gene, and trajectory analyses, a drawback that could be overcome using DISCERN's expression reconstruction. We therefore compared $CD4^+$ T cell (gene expression of *CD4* > 1 and *CD3E* > 2.5) clustering and subtype identification using cite-lq and reconstructed-hq data. While clustering with the leiden algorithm [26] using highly variable genes of cite-lq data resulted in an unstructured distribution of $CD4^+$ T cell subtypes (Figure 3B), clustering of reconstructed-hq data yields detailed insights into T helper cell subtypes of blood mononuclear data (Figure 3C). After reconstruction, we were able to characterize TH17, TH2, TH1, HLA-DR expressing TREG (Active_TREG), naive $CD4^+$ T cells (CD4_naive), effector-memory $CD4^+$ T cells (CD4_EM), central-memory $CD4^+$ T cells (CD4_CM), and effector cells expressing IFN-regulated genes (IFN_regulated) (Figure 3C). We selected published cell-determining marker genes and observed that many of

them were dropped out in the uncorrected data but present after reconstruction (Figure S26). The absence of marker genes in uncorrected data results in poor clustering and cell type identification, while single positive cells are detectable in the respective neighborhood identified by reconstructed counts (Figure S26). Importantly, we observed that in all cases the DISCERN-estimated proportions of T helper subsets fall within the range of expected proportions as assessed by previous FACS studies (Table S3, Figure S27). These findings are important, as they prove once more that DISCERN discovers the correct cell subtypes and cell proportions, in this case substantially outperforming the available CITE-seq information in cell subtype resolution.

To further verify the cell type annotations, we extracted the top cluster-determining genes from the reconstructed-hq data. Members of the TNF-receptor superfamily are known to be expressed in T helper cell subtypes [27], which can be observed after reconstruction in TH17 cells and partially in TH1, TH2, Active_TREG and IFN_regulated cells (Figure S28). Similarly, reconstructed TH1 cells show the expected high expression of granzymes *GZMK* and *GZMA* [28], while *MIAT* and *HLA* expression are found in activated TREG cells after reconstruction (Active_TREG cluster, Figure S28) [29, 30]. *NOG* expression is detected in reconstructed CD4_naive cells, as previously described [31]. In addition, reconstructed CD4_naive, CD4_EM and CD4_CM show low expression of the genes important for the T helper subtypes TH1, TH2, TH17, Active_TREG and IFN_regulated. We further corroborated our cell type annotation of reconstructed-hq data by observing the expected expression of several established T cell subtype markers (Figure S29).

Similar to improved clustering and cell subtype detection, DISCERN reconstructed-hq data resulted in improved gene regulatory network inference with SCENIC [32]. SCENIC infers transcription factor-regulated gene expression modules of single cell data. While cite-lq data resulted in a scattered distribution of transcription factor networks across several T helper cell subtypes, SCENIC with reconstructed-hq data showed transcription factor regulation in the correct subtypes (Figure 3D). After expression reconstruction the IKZF2 regulon is detected in activated TREG cells [33] and the MAF regulon is found in differentiated $CD4^+$ T cells but not in naive $CD4^+$ T cells [34]. A weak signal of the MAF regulon is already detectable in the cite-lq data, yet strongly increased in reconstructed-lq, while maintaining differentiated T helper cell specificity (Figure 3D). Furthermore, after reconstruction with DISCERN we could identify the TH17 associated master transcriptional regulators RORC(+) and RORA(+) [35], which were scattered over all TH17 cells before reconstruction (Figure S30).

Finally, we wanted to investigate if DISCERN could also enhance cell trajectory analyses with Slingshot of the citeseq data [36]. We focused on the differentiation of effector and other T helper cell subtypes and found five lineages that either pass through or terminate in the effector cell cluster in reconstructed-hq data (Figure 3C). Two trajectories were of special interest to us: Lineage1 from CD4_naive to TH1 cells (Figure S31) and Lineage2 from CD4_naive to TH17 cells (Figure S32). While the expression change along the trajectory in uncor-

rected data (Figure S31A, Figure S32A) is hardly visible, cell type-specific clusters can be easily observed after DISCERN reconstruction (for lineage details see Figure S31B, Figure S32B). The detailed insights into cell differentiation that we obtained with reconstructed data are in stark contrast to the Slingshot results obtained with cite-lq data. While terminal effector molecules can be detected with cite-lq data, intermediate stages remain hidden, which prohibits the detection of trajectories and results in a shuffling of marker gene expression (Figure S31). Taken together these results highlight how expression reconstruction using DISCERN improves downstream analyses and yields deeper biological insights into cell type and state identification, gene regulation, and developmental trajectories of cells.

2.4. Discovering COVID-19 disease-relevant cells in lung and blood

The previous sections have demonstrated DISCERN's utility to reconstruct single cell expression data based on an hq reference, vastly improving the detection of cell (sub-) types and their signaling. Given these advantages, we wondered if DISCERN's expression reconstruction could deepen our understanding of cell type-composition and signaling changes of immune cells in COVID-19 disease (Figure S33), using two published datasets [37, 16]. To obtain best reconstruction results, we again resorted to using bulk-hq immune reference data (Table S1) [38], as outlined in the previous section.

First, we used a COVID-19 blood dataset (covid-blood-lq) with limited cell type resolution, which was originally analyzed by our group using Seurat (Table S1) [16]. While $CD4^+$, $CD8^+$, and NK cells formed separate clusters we were unable to visibly distinguish subpopulations of these cells in covid-blood-lq data [16]. Reconstruction of gene expression using bulk-hq data led to the identification of 24 subtypes of $CD4^+$ and $CD8^+$ T cells in covid-blood-hq data (Figure S34). Several cell clusters identified in covid-blood-hq data showed the correct cell type-specific marker gene expression in covid-blood-lq data, albeit in fewer cells, reduced in magnitude, and in some cases less specific (Figures S35 and S36). Reconstruction also led to the identification of $CD4^+$ TH17 helper cells that express *RORC* Figure 4A & B, Figure S37). Based on the molecular footprint of these TH17 cells they were further subdivided into TH17_cluster1 that exhibits a memory T cell phenotype with elevated *IL7R* expression and TH17_cluster2 that exhibits an activated T cell phenotype with elevated *MHC-II*, *CCR4* and *RBPJ* expression (Figure 4B, Figure S37). The expression of *RBPJ* is of particular interest, as it is linked to TH17 cell pathogenicity, suggesting a role of pathogenic TH17 cells in COVID-19 [39]. It is common practice to stimulate memory T cells in vitro to trigger IL-17A production and a shift towards a TH17 phenotype was previously described in COVID-19 [40]. With DISCERN we are able to distinguish these cells in COVID-19 patient blood without stimulation, identifying cytokine producing memory cells with a TH17-like phenotype (Figure S37).

To further validate the existence of activated TH17 cells in COVID-19 patient blood, we next analyzed the corresponding lung data (covid-lung) of the

patients for shared T cell receptor clones (Figure S38). The underlying assumption is that cells with the same T cell receptor in lung and blood originate from the same progenitor and therefore have a high probability of belonging to the same cell type. For this comparison we used the cell type annotation and representation of our original analysis of the covid-lung data, in which memory T and TH17 cells were readily observed without reconstruction [16]. TH17_cluster1 cells showed strong clonal overlap with covid-lung $CD4^+$ memory T cells (Figure S38) and expressed comparable levels of *RORC* to covid-lung effector memory TH17 cells (Figure S39), indicating that these $CD4^+$ central memory T cells could be TH17 (-like) cells. TH17_cluster2 in blood exhibited strong clonal overlap with effector memory and resident memory TH17 cells in covid-lung data (Figure S38) that express *RORC* and *IL-17A* (Figure S39). Using the clonotype information of resident memory cells producing *IL-17A* in inflamed lung (TRM17), we further corroborated the existence of the newly identified population of IL-17A-producing TH17 cells in reconstructed COVID-19 blood data (Figure S38). In general, the T cell receptor clonal information in blood and lung therefore corroborated our cell type annotation in covid-blood-hq data.

To understand the role of T cell subtypes in COVID-19 disease progression we analyzed a second blood single cell dataset (covid-blood-severity-lq) containing disease-severity information for 130 COVID-19 patients [37]. To obtain optimal cell type resolution, we combined the covid-blood-severity-lq T cell data[37] with CD3+ covid-blood-lq cells [16] and reconstructed gene expression for the combined dataset using bulk T cell sequencing reference data[38], resulting in covid-blood-severity-hq data. Many of the 15 $CD4^+$ T cell clusters identified in covid-blood-severity-hq data (Figure S40) were also present in the covid-blood-hq data, further validating the consistency of our cell type identification. This is also corroborated by the available surface protein data for covid-blood-severity data, substantiating that naive cells are CD45RA, memory cells are CD45RO, and effector cell types are CD45RO positive (further details in Figure S41). We compared the clusters that we identified in the covid-blood-hq with clusters identified in the covid-blood-severity-hq data and found confined and overlapping regions of TFH, TH17_cluster1, and TH17_cluster2 cells (Figure S42). We also compared the identified clusters to clusters defined in the original publication (Figure S43). Cells identified as TFH in the original publication show significant overlap with naive $CD4^+$ T cells (defined on transcriptome and protein level) and $CD4^+$ $IL22^+$ cells (CD4.IL22) show marked overlap with TREG cells. These results confirm once more the precise and robust cell type identification that can be achieved with DISCERN.

Interestingly, we also identified two rather unexpected cell types after reconstruction. One cluster is positive for *CD4* and negative for *CD8A* while otherwise expressing a signature of $CD8^+$ effector memory cells with high expression of *GZMB*, *GZMH* and *PRF1* (Figure 4D & 4E). This signature points to a $CD4^+$ cytotoxic phenotype and indeed virus-reactive $CD4^+$ cytotoxic cells were described to be increased in blood during COVID-19 [41]. The other cell type expresses *CD8*, *IL6R*, and *GATA3*, while being negative for *SLAMF7* (Fig-

ure 4D & 4E). These cells were described in the literature to be CD8$^+$ T helper cells [42], exert T helper function, and have been shown to lack cytotoxicity. They lack expression of a significant number of cytokines and key transcription factors pointing to a TH17 or TH22 phenotype. On a protein level these cells express *CCR4*, while being negative for CCR6, making them cytolytic CD8$^+$ T helper type 2 cells (Tc2) cells. Part of this cluster overlaps with CD4 single-positive cells and might explain why T helper type 2 cells are missing in the CD4 cell clustering.

Overall, the highly specific and sensitive cell type identification in covid-blood-severity-hq data enabled us to correlate the five COVID-19 disease severity categories to shifts in cell type and activity information. We first validated the decrease in TFH cells with increasing disease severity, as described in the original work (Figure S44) [37]. TH17 cells have been extensively studied using flow cytometry and in accordance with our results MHC-II positive as well as *CCR4* positive cells were described in COVID-19 patients (Figure 4B) [40]. We observed a strong decrease in naive T helper cells in severe disease, most pronounced for naive TREGs, while the fraction of TH17 cells showed little correlation with disease severity (Figure S44). Of the two mixed cell types we detected in COVID-19 data, cytotoxic CD4$^+$ cells were increased in moderate and severe disease (Figure S45). A similar increase is visible in patients with severe respiratory disease without COVID-19 (Figure S46) and these cells might therefore be a general marker of severe respiratory illness. Cytolytic CD8$^+$ Tc2 cells are increased in patients with severe symptoms and in those who died from COVID-19 (Figure S45) and are described to be reduced after recovery from COVID-19 [43]. This positive correlation and the known role of Tc2 cells in fibroblast proliferation induction and tissue remodeling could pinpoint a mechanistic role of these cells in lung fibrosis as witnessed in severe COVID-19 patients. The possibility to observe these cells in reconstructed single cell data may pave the way to study the functional role of these cells in adverse COVID-19 outcome.

The relatively strong correlation of some cell types with COVID-19 outcome suggests that blood cell fraction information might be used for patient severity prediction. We trained a Gradient Boosting Machine (GBM) using leave-one-out-cross-validation (LOOCV) on the fractions of all T cell types and performed a forward feature elimination, to obtain a sparse, optimal model for patient blood-based severity prediction. We first classified patients into three groups, mild (union of asymptomatic and mild, $n = 26$), moderate ($n = 26$), and severe (union of severe and critical, $n = 19$), reaching an AUROC of 0.63 (Table S4). We noticed that the mild and moderate groups were indistinguishable for the classifier (Figure S47). Training a GBM classifier on mild and severe cases substantially increased classification performance, reaching an AUROC of 0.81 and accuracy, and F1 score of 0.82 (Table S4, Figure 4F &G). Compared to the original T cell types and fractions reported (accuracy 0.61) [37], DISCERN reconstructed T cell fractions are 33 % more accurate in the prediction of COVID-19 disease severity (Figure 4G, Table S4). This classification improvement is remarkable, given that DISCERN has no notion of disease severity when it reconstructs gene expression. These results further demonstrate DISCERN's

precise and robust expression reconstruction that enabled the discovery of a potential new blood-based biomarker for COVID-19 severity prediction.

3. Discussion

The sparsity of gene expression information and high technical noise in single cell sequencing technologies limits the resolution of cell clustering, cell type identification, and many other analyses. Several algorithms such as scImpute, MAGIC, and DCA have addressed this problem by imputing missing gene expression in single cell data by borrowing expression information from similar cells within the same dataset. While gene imputation clearly improves gene expression by inferring values for dropped out genes, this imputation relies on the comparison of similar cells with largely absent gene expression information in the same dataset. With DISCERN we take a completely novel approach to gene expression inference of single cell data, by realistic reconstruction of missing gene expression in scRNA-seq data using a related dataset with more complete gene expression information. We thus propose to call this procedure 'expression reconstruction' to highlight the fundamental difference to classical imputation and refer to the dataset with missing gene expression information as low quality (lq) and the reference dataset as high-quality (hq).

We provide compelling evidence that our reference-based reconstruction outperforms classical expression imputation algorithms as well as batch correction algorithms such as Seurat and scGen, when they are repurposed for expression reconstruction. To obtain an objective and thorough performance evaluation for expression inference, we used seven performance metrics on 19 datasets, including 12 single cell sequencing technologies. We focused our performance evaluation on three scenarios with available ground-truth information, i) the in silico creation of defined gene and pathway drop out events in scRNA-seq data, ii) published hq and lq data pairs from the same tissue (pancreas, difftec, sn/scRNA-seq datasets), and iii) CITE-seq protein expression as ground-truth for cell types (citeseq dataset). In total, DISCERN achieved best performance in 13 out of 15 experiments and obtained second rank in the remaining 2 comparisons. While DISCERN yields first place to Seurat in two FC expression correlation comparisons, it always obtains best results across all datasets in gene expression, gene regulatory network analysis, pathway reconstruction, and cell type and activity identification and is the most stable algorithm for different lq to hq size ratios and cell type overlaps.

It is important to note that DISCERN is a **precise** network that models gene expression values realistically while retaining prior and vital biological information of the lq dataset after reconstruction. The network is also **robust** to the presence of different cell types in hq and lq data, or an imbalance in their relative ratios, and is robust to 'hallucinating' hq-specific cells into the lq data. Several algorithmic choices are the foundation of DISCERN's precision and robustness. The network was designed to model the sequencing-technology-specific and the underlying biological signals in separate components of its architecture. Disentanglement of those two components is necessary to accurately

reconstruct expression information in the case where lq and hq datasets have different content, i.e. cell type compositions. If the component designed to model the effect of sequencing technology also captures the difference in the biological signal, the reconstruction will lead to a lack of integration across the two datasets where some cell types are still clustered by dataset (similar to scGen in fig. S19). On the contrary, if the component modeling the biological signal captures sequencing-technology-specific features, the reconstruction will lead to an over-integration of the datasets where cells of different types are mixed together (similar to Seurat in fig. S19). The demonstrated ability of DISCERN to avoid those shortcomings, even in scenarios where there is very little to no overlap between cell types across datasets, lies in the carefully crafted balance between the expressivity of its components. The representational capabilities of DISCERN, achieved via batch normalization, five loss terms, and a dual head decoder, would reduce DISCERN's usability, if they would require frequent dataset-specific tuning. The stability and usability was therefore a central concern in the design and evaluation phase of DISCERN, which resulted in an algorithm that gave very good results with a single set of default (hyper-) parameters. All comparisons to other algorithms, for instance, were performed with default settings. Only the expression reconstruction of the exceptionally large COVID-19 datasets required the fine-tuning of the learning rate, cross entropy term, sigma, and the MMD penalty term. Another important technical feature of DISCERN is that it can easily be integrated into existing workflows. It takes a normalized count matrix, as created by nearly all existing single cell analysis workflows, as input and produces a reconstructed expression matrix. This can be used for most downstream applications (i.e. cell clustering, cell type identification, cell trajectory analysis, and differential gene expression). DISCERN can be trained on standard processors (CPU) for small and medium-sized datasets and requires graphical processing units (GPU) for the expression reconstruction of large datasets. Altogether, the usability and robustness of DISCERN should enable even non-expert users to perform gene expression reconstruction.

A unique feature of DISCERN is the use of an hq reference to infer biologically meaningful gene expression. While we consider this a main strength of DISCERN, the dependence on a suitable reference dataset might also limit its application. We took great care in this manuscript to mitigate this concern by showing how DISCERN is able to reconstruct gene expression for many different types of lq and hq pairs, ranging from indrop - smartseq2 to single nucleus - single cell data pairs. Remarkable in this context is DISCERN's robustness to differences between the cell type compositions of lq and hq data pairs, with DISCERN being the only algorithm obtaining robust expression reconstruction when few cell types overlap. We have also shown that purified bulk RNA-seq samples can be used as hq reference, as successfully applied to PBMC and COVID-19 datasets in this study. We used 9852 FACS purified immune cell bulk sequencing samples [38], comprising 27 cell types, to successfully reconstruct single cell expression data. This implies that most single cell studies involving immune cells (with or without other cell types present) can be reconstructed with DISCERN using a single published bulk RNA-seq dataset.

Furthermore, public RNA-seq repositories such as NCBI GEO contain tens of thousands of samples of immune and non-immune cells that could serve as reference for most expression reconstruction experiments. In conclusion, we provide strong evidence that DISCERN is widely and easily applicable to many single cell experiments.

While DISCERN gave good reconstruction results using default parameters for most datasets we analyzed, we would like to highlight that the immense representational power of generative neural networks can remove or hallucinate biological information if not properly handled [6]. This is true for data integration [44] as well as for expression reconstruction algorithms and we would highlight two guiding principles for optimal results. For non-expert users, we would recommend the use of default settings and a careful selection of a related hq dataset. When datasets are large and complex, with many cell types in the lq and several non-overlapping cell types in the hq data, one should always ensure that training does not merge or mix non-overlapping cell types with other cells, by investigating that these cells keep their cell type-specific marker gene expression. Keeping these 'checks and balances' will usually result in good reconstruction results even for complex datasets such as covid-blood-severity.

To obtain novel insights into COVD-19 disease mechanisms and a new blood-based biomarker for disease severity we reconstructed two published datasets with DISCERN, Hamburg COVID-19 patients (covid-lung, -blood) and the COVID-19 cell atlas (covid-blood-severity). The application of DISCERN to the covid-blood dataset (COVID-19 patient blood) enabled us to detect 24 different immune cell types and activity states, which is quite remarkable given that we find these cells in blood. Two TH17 subtypes caught our attention, as they share the TCR clonality with the lung data from the same patients (covid-lung), suggesting bloodstream re-entry of lung TH17 cells. We linked these two subclusters to their functional role by separating them into a memory-like and activated-like phenotype. The clonal overlap of activated TH17 cells in blood with previously discovered lung-resident cells suggests that activated TH17 cells in blood are resident T cells from the lung reentering circulation. These cells might in part explain the multi-organ pathology observed in COVID-19, as activated T cells might travel via the blood to secondary organs and cause inflammation and tissue damage. Future work might demonstrate the effect of these activated T cells on tissue inflammation.

Given the detailed cell type and activity information we reached with gene expression reconstruction, we wondered if changes in blood immune cell populations might be useful as a biomarker for disease severity prediction. We used DISCERN to reconstruct the covid-blood and the covid-blood-severity datasets and again identified a plethora of different T cell subtypes in the blood of patients with COVID-19. Using these cell proportions, we were able to classify mild and severe disease using a GBM machine learning algorithm with 82 % accuracy, outperforming classification with the originally published T cell types by 21 percent points. This improvement is absolutely striking, as DISCERN has no notion of the classification groups. It simply reconstructs gene expression and thereby improves cell type detection. These results are a convincing

implicit proof not only of the usefulness of DISCERN but more importantly of its precision and robustness. While the use of this scRNA-seq-based biomarker would be too expensive and time-consuming for clinical care, it strongly suggests that FACS-based T cell fraction or count information from blood could be used to trace and predict the severity state and potentially the disease trajectory of COVID-19 patients.

Interestingly, we also discovered two atypical T cell types in reconstructed COVID-19 patient blood single cell data. While cytotoxic $CD4^+$ T cells have been observed in COVID-19, we can show that this increase is not COVID-19 specific and is also observed in other types of pneumonia. Interestingly, we also detected cytolytic $CD8^+$ Tc2 cells that express *CD8A*, *GATA3*, *IL6R* and are negative for *SLAMF6*. This cell type is linked to tissue fibrosis and steroid refractory disease in asthma [45]. The increase in $CD8^+$ Tc2 cells that we observe specifically in COVID-related death could be associated with COVID-19 patients that do not respond to steroids. Demonstration of increase of this cell type in patients dying of COVID-19 points to a potential therapeutic intervention with the drug Fevipiprant, which blocks $CD8^+$ Tc2 cell activation and its pro-fibrotic effects by inhibiting prostaglandin D2 signaling [46]. Functional analysis of these cells has to demonstrate whether these cells are an early marker of later death or whether it is a marker of already escalated treatment.

The basic concept of utilizing a high-quality reference to improve lower quality data might be applied to many other research areas where technological limitations restrict biological insights. The usage of deep generative networks and other artificial intelligence methodology to infer information beyond what is technically measurable could be transformative in future biomedical research.

Acknowledgements

We thank Immo Prinz, Manuel Friese, Johannes Soeding, Robert Zinzen, Yu Zhao and Stefan Kurtz for their helpful comments and suggestions. FH, RK, MM, PM, & SB were supported by the LFF-FV 78, EU ERare-3 Maxomod, SFB 1286 Z2, FOR 296, and FOR 5068 research grants. Further support was obtained from the UKE R3 reduction of animal testing grant. CE was supported by DFG ER 981/1-1 and the clinician scientist programme of university Hamburg, NG was supported by ERC StG-715271, and SHu was supported by ERC CoG-865466 and has an endowed Heisenberg-Professorship awarded by the Deutsche Forschungsgemeinschaft. SHa was funded by the BMBF STOP-FSGS-01GM1518C and SFB 1192 B08 research grants.

Competing interests

The authors declare no competing interests.

Author contributions

SB initiated and SB, PM, FH, and CE conceptualized the study with help from MM. FH and CE implemented DISCERN, MM refactored the code, and PM reviewed the DISCERN implementation. FH, CE, and RK performed the analyses. SB, PM, NG, and SHu supervised the study. SB, FH, and CE wrote the manuscript. SHu, PM, NG, RK and SHa provided ideas, contributed to manuscript text and critically reviewed the manuscript. All authors read and approved the final manuscript.

Main figures

References

[1] N. Editorial, Method of the year 2013, Nat. Methods 11 (1) (2014) 1.

[2] Y. Zhao, U. Panzer, S. Bonn, C. F. Krebs, Single-cell biology to decode the immune cellular composition of kidney inflammation, Cell and tissue research 385 (2) (2021) 435–443.

[3] M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, P. Smibert, Simultaneous epitope and transcriptome measurement in single cells, Nature methods 14 (9) (2017) 865–868.

[4] A. A. Tu, T. M. Gierahn, B. Monian, D. M. Morgan, N. K. Mehta, B. Ruiter, W. G. Shreffler, A. K. Shalek, J. C. Love, Tcr sequencing paired with massively parallel 3' rna-seq reveals clonotypic t cell signatures, Nature immunology 20 (12) (2019) 1692–1699.

[5] J. A. Pai, A. T. Satpathy, High-throughput and single-cell t cell receptor sequencing technologies, Nature Methods 18 (8) (2021) 881–892.

[6] S. Oller-Moreno, K. Kloiber, P. Machart, S. Bonn, Algorithmic advances in machine learning for single-cell expression analysis, Current Opinion in Systems Biology 25 (2021) 27–33.

[7] D. Lähnemann, J. Köster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, A. Mahfouz, et al., Eleven grand challenges in single-cell data science, Genome biology 21 (1) (2020) 1–35.

[8] X. Wang, Y. He, Q. Zhang, X. Ren, Z. Zhang, Direct comparative analyses of 10x genomics chromium and smart-seq2, Genomics, proteomics & bioinformatics 19 (2) (2021) 253–266.

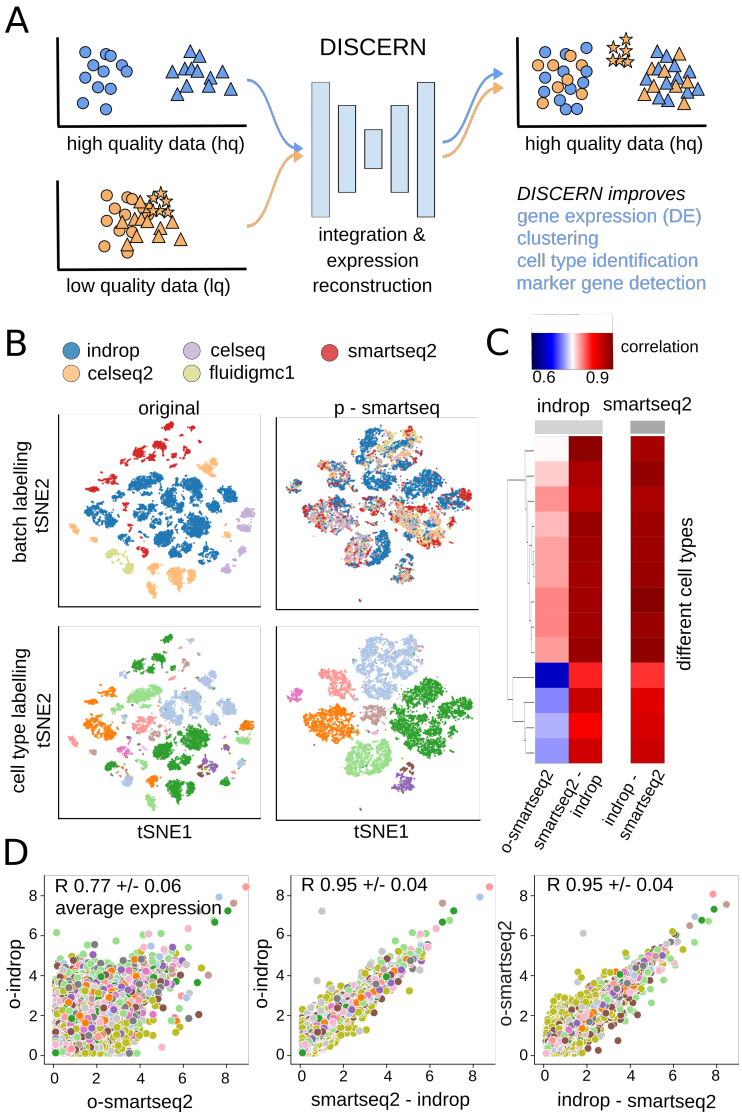

Figure 1: *Integration and expression reconstruction of single cell sequencing data.* **A**: DISCERN transfers the style of a high-quality (hq) dataset to a related low quality (lq) dataset, enabling gene expression reconstruction that results in improved clustering, cell type identification, marker gene detection, and mechanistic insights into cell function. The hq and lq datasets have to be related but not identical, containing for example several overlapping cell types but also exclusive cell types of cell activity states for one or the other dataset. **B**: t-SNE visualization of the pancreas dataset before reconstruction (original) and after transferring the style of the smartseq2 dataset using DISCERN (p-smartseq2). The upper row shows the dataset of origin before and after projection colored by batch and the lower row colored by cell type annotation (details of 13 cell types in supplements). **C** and **D**: Average gene expression (over all the cells of a given type) of the pancreas indrop and smartseq2 datasets before (first column and panel) and after smartseq2 to indrop (second column and panel), and after indrop to smartseq2 projection (third column and panel). **C**: Gene correlation by cell type shown in colored heatmap. **D**: Each colored point represents a single gene colored by the cell type, 'o' refers to original data. The mean Pearson correlation with one standard deviation over all cell types is shown in the figure title.

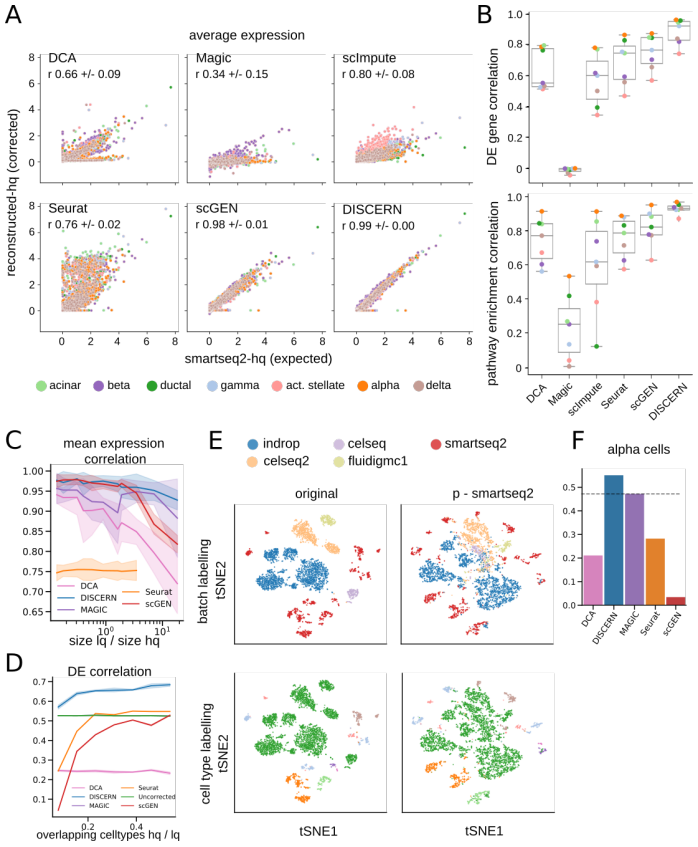

Figure 2: *Expression reconstruction benchmark of DISCERN and five state-of-the-art batch correction and imputation algorithms.* **A**: Comparison of the expression reconstruction performance of Seurat, scGEN, Magic, scImpute, DCA, and DISCERN using smartseq2 data. The smartseq2 data was split into a smartseq2-lq and a smartseq2-hq batch. The smartseq2-lq batch was modified such that the expression of all genes of a cell type determining pathway (top ranked by GSEA) was set to zero. The expression of the in silico altered pathway genes was then compared between reconstructed-hq data and the unaltered smartseq2-hq data. **B**: Differential gene expression and pathway enrichment correlation of the reconstructed-hq to the expected values before removal. The smartseq2-lq data was the same as in **A**. The DEG analysis was restricted to genes which were removed in the smartseq2-lq batch. Correlation of the DEG analysis was based on the t-statistic and for the pathway enrichment analysis on the normalized enrichment scores. **C**: Mean expression correlation of reconstructed-hq with the expected expression in smartseq-hq data for different ratios of lq to hq data. The standard deviation indicates the deviation in correlation of the cell types. The datasets were created as described in **A**. **D**: Alpha cells were removed from the smartseq-hq batch and left in the smartseq-lq batch. The number of other overlapping cell types between the hq and lq data was then altered by removing cell types from the lq data before expression reconstruction (x-axis). The y-axis shows the correlation of the t-statistics of alpha cells from lq-batches vs other cells from the smartseq2 batch with ground truth alpha cells from the smartseq2 batch vs other cells from the uncorrected smartseq2 batch. **E**: t-SNE visualization of the cell type removal experiment where alpha cells are removed from the smartseq2 batch and all non-alpha cells are removed from the lq-batches, such that there is no overlap between lq and hq. **F**: Pearson correlation of the t-statistics of alpha cells from lq-batches vs other cells from the smartseq2 batch with ground truth alpha cells from the smartseq2 batch vs other cells from the uncorrected smartseq2 batch. The dataset was the same as in **E** (no cell type overlap between hq and lq data). The dotted line indicates the correlation achieved without reconstruction.

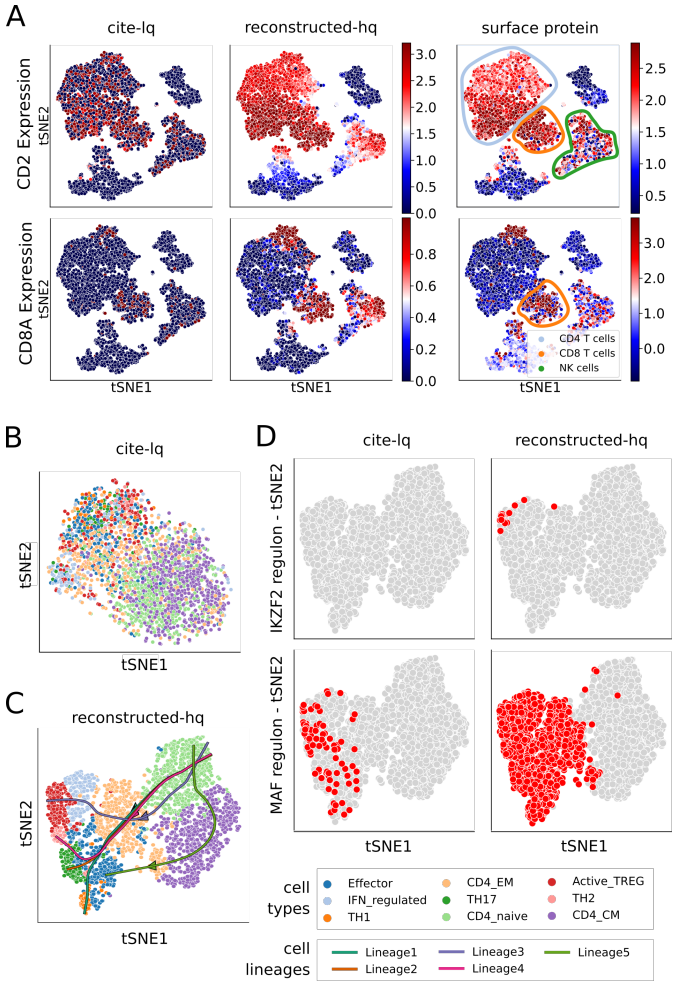

Figure 3: *Expression reconstruction improves downstream analyses including cell identification, gene regulation, and trajectory inference.* The cite-lq dataset was reconstructed using bulk-hq data and compared to ground truth CITE-seq (surface protein) information. The CITE-seq information was not used during training of DISCERN. **A**: t-SNE visualization of *CD2* (first row) and *CD8A* (second row) gene (first two columns) and protein (last column))expression. The first column depicts gene expression for uncorrected cite-lq, the second for reconstructed-hq, and the third protein surface expression ground truth information. Cell types commonly known to express these genes are highlighted with colored circles in the last column. **B**: t-SNE visualization of CD4$^+$ T cells in the cite-lq dataset. Cell types were assigned using louvain clustering on the reconstructed-hq data (see C) and show no clear clustering. **C**: t-SNE and trajectory information of CD4$^+$ T cell subtypes found by Slingshot analysis on reconstructed-hq data. While uncorrected data shows no clear cell type clustering (see **B**), reconstructed data shows a clear grouping of cell types. Trajectories were calculated using CD4_naive as starting point and TH2, TH17, TH1, Active_TREG, CD4_CM as endpoints. Lineage1 indicates TH1, Lineage2 TH17, Lineage3 Active_TREG, Lineage4 TH2, and Lineage5 Effector cell differentiation. **D**: Detection of regulons that are specific for CD4$^+$ T cell subtypes using pySCENIC. The first column shows regulons found in the uncorrected cite-lq and the second column in reconstructed-hq data.

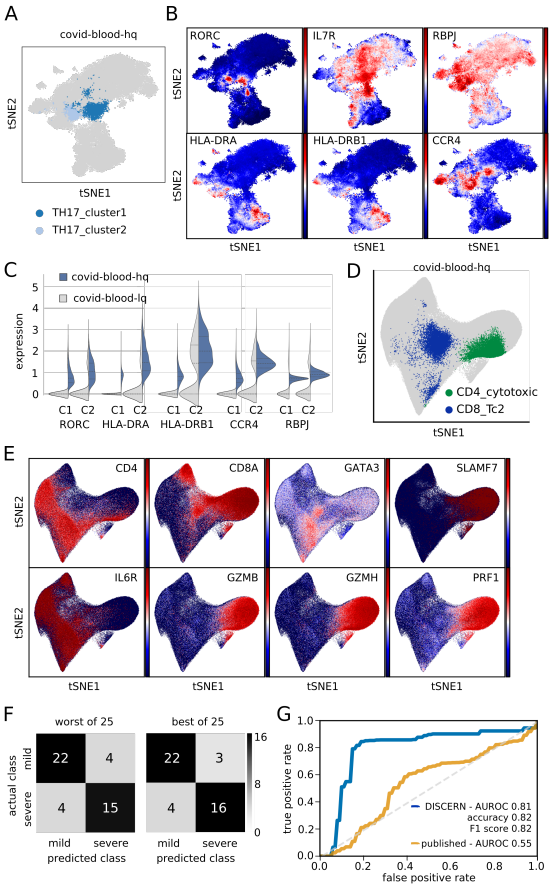

Figure 4: *Expression reconstruction improves COVID-19 cell type identification and allows for efficient disease severity prediction.* Two COVID-19 blood datasets were reconstructed and analyzed. Hamburg covid-blood-lq and covid-lung-lq data was reconstructed using bulk-hq data, resulting in the respective -hq datasets. Similarly, Cambridge covid-blood-severity-lq data, which contains disease severity information, was reconstructed using bulk-hq data. **A**: t-SNE representation of TH17 subclusters using reconstructed covid-blood-hq data. Clusters were defined using the leiden clustering algorithm on $CD4^+$ T cells. **B**: t-SNE representation colored by expression of reconstructed genes distinguishing TH17_cluster1 and TH17_cluster2 cells. TH17_cluster1 displays a central memory and TH17_cluster2 a more activated phenotype. **C**: Violin plots of expression levels for genes distinguishing TH17_cluster1 (C1) and TH17_cluster2 (C2) cells before (covid-blood-lq) and after (covid-blood-hq) reconstruction with DISCERN. **D**: Rare and unexpected cell types found in the reconstructed covid-blood-hq data with covid-blood-severity and bulk data. Cytotoxic $CD4^+$ T cells (CD4_cytotoxic) are displayed in green, $CD8^+$ Tc2 helper cells (CD8_Tc2) in blue, and all other cells in gray color. **E**: t-SNE representation of key marker genes in covid-blood-hq data for CD4_cytotoxic and CD8_Tc2 cells displayed in **D**. **F**: Best and worst confusion matrix for disease severity prediction using GBM classifiers trained on fractions of five T cell types (CD4_CM, CD4_cytotoxic, CD4_naive, CD8_EM, CD8_effector) using reconstructed covid-blood-severity-hq data. Category "critical" was combined with "severe" and "asymptomatic" with "mild". **G**: ROC curve of the GBM predictions outlined in **F** using reconstructed (blue color) covid-blood-severity-hq (CD4_CM, CD4_cytotoxic, CD4_naive, CD8_EM, CD8_effector) and published T cell information from uncorrected (yellow color) data (CD4.CM, CD4.Tfh, CD8.EM, NKT, Treg). Confidence intervals (color shades) indicate one standard deviation.

[9] A. K. Shalek, R. Satija, J. Shuga, J. J. Trombetta, D. Gennert, D. Lu, P. Chen, R. S. Gertner, J. T. Gaublomme, N. Yosef, et al., Single-cell rna-seq reveals dynamic paracrine control of cellular variation, Nature 510 (7505) (2014) 363–369.

[10] W. Hou, Z. Ji, H. Ji, S. C. Hicks, A systematic evaluation of single-cell rna-sequencing imputation methods, Genome biology 21 (1) (2020) 1–30.

[11] G. Eraslan, L. M. Simon, M. Mircea, N. S. Mueller, F. J. Theis, Single-cell rna-seq denoising using a deep count autoencoder, Nature communications 10 (1) (2019) 1–14.

[12] W. V. Li, J. J. Li, An accurate and robust imputation method scimpute for single-cell rna-seq data, Nature communications 9 (1) (2018) 1–9.

[13] D. Van Dijk, R. Sharma, J. Nainys, K. Yim, P. Kathail, A. J. Carr, C. Burdziak, K. R. Moon, C. L. Chaffer, D. Pattabiraman, et al., Recovering gene interactions from single-cell data using data diffusion, Cell 174 (3) (2018) 716–729.

[14] M. Marouf, P. Machart, V. Bansal, C. Kilian, D. S. Magruder, C. F. Krebs, S. Bonn, Realistic in silico generation and augmentation of single-cell rna-seq data using generative adversarial networks, Nature communications 11 (1) (2020) 1–12.

[15] M. Lotfollahi, F. A. Wolf, F. J. Theis, scgen predicts single-cell perturbation responses, Nature methods 16 (8) (2019) 715–721.

[16] Y. Zhao, C. Kilian, J.-E. Turner, L. Bosurgi, K. Roedl, P. Bartsch, A.-C. Gnirck, F. Cortesi, C. Schultheiß, M. Hellmig, et al., Clonal expansion and activation of tissue-resident memory-like th17 cells expressing gm-csf in the lungs of patients with severe covid-19, Science Immunology 6 (56) (2021) eabf6692.

[17] I. Tolstikhin, O. Bousquet, S. Gelly, B. Schoelkopf, Wasserstein auto-encoders, arXiv preprint arXiv:1711.01558.

[18] T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck III, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive integration of single-cell data, Cell 177 (7) (2019) 1888–1902.

[19] V. Dumoulin, J. Shlens, M. Kudlur, A learned representation for artistic style, arXiv preprint arXiv:1610.07629.

[20] J. L. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, arXiv preprint arXiv:1607.06450.

[21] Z. Fu, E. R Gilbert, D. Liu, Regulation of insulin synthesis and secretion and pancreatic beta-cell dysfunction in diabetes, Current diabetes reviews 9 (1) (2013) 25–53.

[22] M. Slyper, C. Porter, O. Ashenberg, J. Waldman, E. Drokhlyansky, I. Wakiro, C. Smillie, G. Smith-Rosario, J. Wu, D. Dionne, et al., A single-cell and single-nucleus rna-seq toolbox for fresh and frozen human tumors, Nature medicine 26 (5) (2020) 792–802.

[23] G. C. Linderman, J. Zhao, M. Roulis, P. Bielecki, R. A. Flavell, B. Nadler, Y. Kluger, Zero-preserving imputation of single-cell rna-seq data, Nature Communications 13 (1) (2022) 1–11.

[24] E. Bakos, C. A. Thaiss, M. P. Kramer, S. Cohen, L. Radomir, I. Orr, N. Kaushansky, A. Ben-Nun, S. Becker-Herman, I. Shachar, Ccr2 regulates the immune response by modulating the interconversion and function of effector and regulatory t cells, The Journal of Immunology 198 (12) (2017) 4659–4671.

[25] G. Monaco, B. Lee, W. Xu, S. Mustafah, Y. Y. Hwang, C. Carré, N. Burdin, L. Visan, M. Ceccarelli, M. Poidinger, et al., Rna-seq signatures normalized by mrna abundance allow absolute deconvolution of human immune cell types, Cell reports 26 (6) (2019) 1627–1640.

[26] V. A. Traag, L. Waltman, N. J. Van Eck, From louvain to leiden: guaranteeing well-connected communities, Scientific reports 9 (1) (2019) 1–12.

[27] M. Croft, Control of immunity by the tnfr-related molecule ox40 (cd134), Annual review of immunology 28 (2009) 57–78.

[28] T. Riaz, L. M. Sollid, I. Olsen, G. A. de Souza, Quantitative proteomics of gut-derived th1 and th1/th17 clones reveal the presence of cd28+ nkg2d-th1 cytotoxic cd4+ t cells, Molecular & Cellular Proteomics 15 (3) (2016) 1007–1016.

[29] L. Peng, Y. Chen, Q. Ou, X. Wang, N. Tang, Lncrna miat correlates with immune infiltrates and drug reactions in hepatocellular carcinoma, International immunopharmacology 89 (2020) 107071.

[30] D. P. Saraiva, A. Jacinto, P. Borralho, S. Braga, M. G. Cabral, Hla-dr in cytotoxic t lymphocytes predicts breast cancer patients' response to neoadjuvant chemotherapy, Frontiers in immunology (2018) 2605.

[31] M. S. Lee, K. Hanspers, C. S. Barker, A. P. Korn, J. M. McCune, Gene expression profiles during human cd4+ t cell differentiation, International immunology 16 (8) (2004) 1109–1124.

[32] S. Aibar, C. B. González-Blas, T. Moerman, V. A. Huynh-Thu, H. Imrichova, G. Hulselmans, F. Rambow, J.-C. Marine, P. Geurts, J. Aerts, et al., Scenic: single-cell regulatory network inference and clustering, Nature methods 14 (11) (2017) 1083–1086.

[33] A. M. Thornton, J. Lu, P. E. Korty, Y. C. Kim, C. Martens, P. D. Sun, E. M. Shevach, Helios+ and helios- treg subpopulations are phenotypically and functionally distinct and express dissimilar tcr repertoires, European journal of immunology 49 (3) (2019) 398–412.

[34] C. Imbratta, H. Hussein, F. Andris, G. Verdeil, c-maf, a swiss army knife for tolerance in lymphocytes, Frontiers in immunology 11 (2020) 206.

[35] X. O. Yang, B. P. Pappu, R. Nurieva, A. Akimzhanov, H. S. Kang, Y. Chung, L. Ma, B. Shah, A. D. Panopoulos, K. S. Schluns, et al., T helper 17 lineage differentiation is programmed by orphan nuclear receptors rorα and rorγ, Immunity 28 (1) (2008) 29–39.

[36] K. Street, D. Risso, R. B. Fletcher, D. Das, J. Ngai, N. Yosef, E. Purdom, S. Dudoit, Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics, BMC genomics 19 (1) (2018) 1–16.

[37] E. Stephenson, G. Reynolds, R. A. Botting, F. J. Calero-Nieto, M. D. Morgan, Z. K. Tuong, K. Bach, W. Sungnak, K. B. Worlock, M. Yoshida, et al., Single-cell multi-omics analysis of the immune response in covid-19, Nature medicine 27 (5) (2021) 904–916.

[38] M. Ota, Y. Nagafuchi, H. Hatano, K. Ishigaki, C. Terao, Y. Takeshima, H. Yanaoka, S. Kobayashi, M. Okubo, H. Shirai, et al., Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases, Cell 184 (11) (2021) 3006–3021.

[39] G. Meyer Zu Horste, C. Wu, C. Wang, L. Cong, M. Pawlak, Y. Lee, W. Elyaman, S. Xiao, A. Regev, V. Kuchroo, Rbpj controls development of pathogenic th17 cells by regulating il-23 receptor expression. cell rep 16 (2): 392–404 (2016).

[40] S. De Biasi, M. Meschiari, L. Gibellini, C. Bellinazzi, R. Borella, L. Fidanza, L. Gozzi, A. Iannone, D. Lo Tartaro, M. Mattioli, et al., Marked t cell activation, senescence, exhaustion and skewing towards th17 in patients with covid-19 pneumonia, Nature communications 11 (1) (2020) 1–17.

[41] B. J. Meckiff, C. Ramírez-Suástegui, V. Fajardo, S. J. Chee, A. Kusnadi, H. Simon, S. Eschweiler, A. Grifoni, E. Pelosi, D. Weiskopf, et al., Imbalance of regulatory and cytotoxic sars-cov-2-reactive cd4+ t cells in covid-19, Cell 183 (5) (2020) 1340–1353.

[42] L. Loyal, S. Warth, K. Jürchott, F. Mölder, C. Nikolaou, N. Babel, M. Nienen, S. Durlanik, R. Stark, B. Kruse, et al., Slamf7 and il-6r define distinct cytotoxic versus helper memory cd8+ t cells, Nature communications 11 (1) (2020) 1–12.

[43] J. Yang, M. Zhong, E. Zhang, K. Hong, Q. Yang, D. Zhou, J. Xia, Y.-Q. Chen, M. Sun, B. Zhao, et al., Broad phenotypic alterations and potential

dysfunction of lymphocytes in individuals clinically recovered from covid-19, Journal of Molecular Cell Biology 13 (3) (2021) 197–209.

[44] M. Lotfollahi, M. Naghipourfar, M. D. Luecken, M. Khajavi, M. Büttner, M. Wagenstetter, Ž. Avsec, A. Gayoso, N. Yosef, M. Interlandi, et al., Mapping single-cell data to reference atlases by transfer learning, Nature Biotechnology 40 (1) (2022) 121–130.

[45] C. Wagner, M. Griesel, A. Mikolajewska, A. Mueller, M. Nothacker, K. Kley, M.-I. Metzendorf, A.-L. Fischer, M. Kopp, M. Stegemann, et al., Systemic corticosteroids for the treatment of covid-19, Cochrane Database of Systematic Reviews (8).

[46] W. Chen, J. Luo, Y. Ye, R. Hoyle, W. Liu, R. Borst, S. Kazani, E. A. Shikatani, V. J. Erpenbeck, I. D. Pavord, et al., The roles of type 2 cytotoxic t cells in inflammation, tissue remodeling, and prostaglandin (pg) d2 production are attenuated by pgd2 receptor 2 antagonism, The Journal of Immunology 206 (11) (2021) 2714–2724.