Capturing Clinical Heterogeneity in Rheumatoid Arthritis

Evaluating the LIVI Latent Space using Gene Expression Data

Patrick Alexander Lo[1]

Supervisors: Prof.dr.ir Marcel Reinders[1], Inez den Hond[1], Kirti Biharie[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering

June 21, 2026

Name of the student: Patrick Alexander Lo
Final project course: CSE3000 Research Project
Thesis committee: Prof.dr.ir. Marcel Reinders, Inez den Hond, Kirti Biharie, Dr. Christoph Lofi

Abstract

Rheumatoid arthritis (RA) is a heterogeneous autoimmune disease: patients who share the same diagnosis respond differently to the same therapy. Zhang et al. stratified the RA synovium into six cell-type abundance phenotypes (CTAPs) by clustering counted, pre-annotated cell-type abundances. The LIVI model, built on a variational autoencoder developed to map trans-eQTLs in non-diagnosed donors, instead learns donor structure directly from gene expression, separating donor-level variation from cell-state variation. The developers of the model left two questions open: whether LIVI can also capture disease status in a diagnosed cohort, and what the optimal number of donor-level embeddings is for a given dataset. We address these by applying LIVI to a CITE-seq dataset of 314,011 cells from 70 RA and 9 OA donors across four different numbers of donor embeddings. In this research, we show that although LIVI is given no cell-type or diagnostic labels, its donor space recovers the underlying cell-type relationships between the six cell types defining CTAPs: a lymphoid (T, B, NK) versus non-lymphoid (myeloid, endothelial, fibroblast) block, which is consistent across all four dimensions. The CTAPs themselves do not form discretely separable groups in the donor space, but at lower dimensionalities, individual donor factors begin to distinguish them along the same axis. Reading the genes behind these factors was limited. We hypothesize this is due to LIVI's sparsity penalty, which was tuned for detecting trans-eQTLs on a much larger cohort, leaving ribosomal pathways dominating the loadings. Therefore, LIVI's donor space captures disease state information, but on a broader scale compared to Zhang et al.'s discretely defined CTAPs. For this particular dataset, the signal becomes stronger at lower dimensionalities, but the interpretability of the signal is limited.

1 Introduction

Rheumatoid arthritis (RA) is a chronic, systemic autoimmune disease that causes synovial inflammation and joint destruction [1]. A central obstacle to treatment is heterogeneity: patients with the same diagnosis respond differently to the same therapy. Resolving this requires deconstructing the disease into its underlying cellular mechanisms. Using CITE-seq data, which measures both RNA expression and surface protein levels in each cell [2], Zhang et al.[3] hierarchically clustered RA synovial samples based on the relative abundances of six major cell types (T, B/plasma, NK, myeloid, endothelial, and stromal/fibroblast). This stratified the disease into six cell-type abundance phenotypes (CTAPs), each named after their enriched populations: endothelial-fibroblast-myeloid (EFM), fibroblast (F), T cell-fibroblast (TF), T cell-B cell (TB), T cell-myeloid (TM), and myeloid (M) (Figure 1). The resulting grouping is robust, as alternative cell-type groupings reproduce the same CTAPs. Furthermore, they observed that CTAPs are dynamic and predict treatment response, giving identification of a patient's CTAP direct clinical utility.

However, this classification relies on predefined, counted cell-type abundances. An alternative is to use unsupervised latent factor models, which learn structure directly from gene expression by creating latent embeddings without counting pre-annotated cell types. Such models have already been applied to RA. Palla and Ferrero [4] benchmarked four of them (scVI [5] with a linear decoder [6], scCoGAPS [7], LDA [8], and scHPF [9]) across different sizes of the embedding space, based on the stability of their latent factors, classification of diseased versus control cells, and the number of unique gene sets their latent factors map to. Using scHPF, the model with the highest mean gene-set coverage, they identified the OSMR signalling pathway in a subset of RA synovial fibroblasts. However, in their analysis,

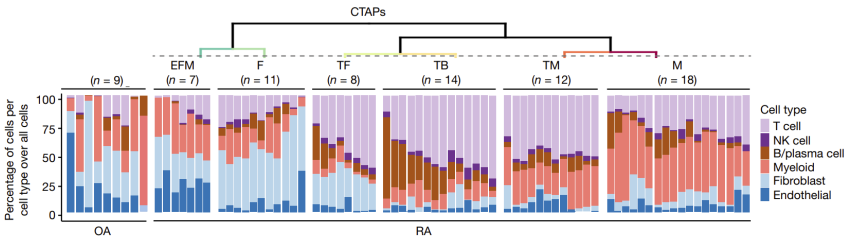


Figure 1: Distribution of cell types across rheumatoid arthritis subtypes. Samples were grouped into six Cell-Type Abundance Phenotypes (CTAPs) using hierarchical clustering. Each bar represents one patient sample, with colors showing the proportion of the six major cell types. Nine osteoarthritis (OA) samples are included for comparison. Figure is directly copied from Zhang et al.[3].

the latent space is represented only over cells and genes, through a latent-by-gene loading matrix and a cell-by-latent factor matrix. Since there are no separate donor-level embeddings, variation between donors and variation between individual cell states (which also differs between a donor's own cells) are captured together by the same factors. This limits the extraction of donor-specific information.

The recently published Latent Interaction Variational Inference (LIVI) model [10] addresses this limitation. LIVI splits the variation into separate spaces: a donor-specific latent space (D), a cell-state latent space (C), and a global donor latent space (V), which captures broad changes in gene expression across the donors.

LIVI was previously applied to the OneK1K cohort (over one million PBMCs from 982 non-diagnosed donors) [11], where it mapped trans-eQTLs at single-cell resolution and found the cell types and genes linked to polygenic risk for autoimmune diseases, including RA. However, the authors left two questions open: whether LIVI is able to capture disease status when applied to a diagnosed cohort, and what the optimal number of donor (D) factors is for a given dataset, since they tuned this parameter specifically for OneK1K. We address both by testing how well LIVI's unsupervised donor space (D) captures Zhang et al.'s CTAPs directly from gene expression, and how this recovery changes with the number of donor factors. Specifically, we ask: does the donor space encode the cell-type structure behind the CTAPs? Can individual factors distinguish CTAP groups discretely? How does dimensionality affect each of these? And does the decoder link CTAP-related factors to biological pathways?

We show that LIVI's donor space, separated from cell-state variation and learned without cell-type labels or diagnostics, recovers the cell-type structure the CTAPs are built on: a split between lymphoid (T, B, NK) and non-lymphoid (myeloid, fibroblast, endothelial) cells. This structure holds at every number of donor factors we tested, so it does not depend on the size of the donor space. The CTAPs themselves, however, do not form separate donor groups in the space as a whole. Instead, LIVI captures the continuous cell-type composition underlying them. Only at lower numbers of donor factors do some individual factors begin to separate the CTAP groups, along that same lymphoid versus non-lymphoid axis. Finally, the genes behind these factors were difficult to interpret, as ribosomal programmes dominate the loadings, which suggests that LIVI's interpretability does not fully transfer to our dataset.

Background

LIVI [10] is a model for single-cell RNA expression measured across a group of donors. It separates the variation in gene expression into cell-state variation, which is shared by donors across canonical cell states, and donor variation, which differs between individual donors. In this way, donor effects can be studied while maintaining single-cell resolution. The model is a variational autoencoder (VAE) [12] with a linear decoder, allowing each latent factor's gene loadings to be read directly as a list of contributing genes, unlike a standard VAE.

LIVI is trained on a single-cell gene expression matrix (cells × genes) together with the donor identity of each cell. During training, it is given no cell-type annotations, donor diagnostic labels or donor genotypes. From this, it learns three individual latent spaces. In our configuration, it is fit to 314,011 cells from 82 synovial samples (from 79 unique donors, see Section 4.1), for 17,049 genes, with 15 cell-state, 700 donor, and 5 global factors:

- **Cell-state factors (C):** canonical cell-state variation shared across donors, 15 factors per cell ($314,011 \times 15$).
- **Donor factors (D):** donor variation separated from cell state, 700 factors per sample (82×700). **Cell-state-specific donor factors ($D \times C$):** how each sample's donor factors vary within specific cell states, formed by combining the cell-state and donor factors via an assignment matrix A (700×15) that links each donor factor to the cell states in which it is active.
- **Global donor factors (V):** donor variation independent of cell state, 5 factors per sample (82×5).

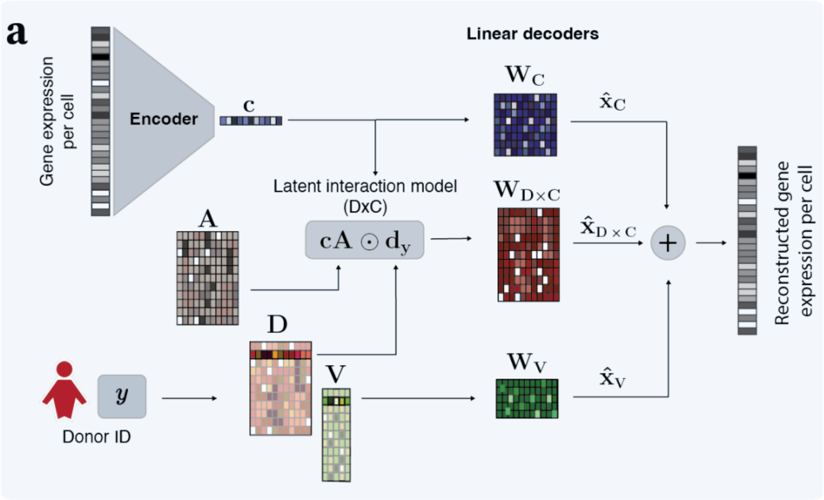


Figure 2: The LIVI architecture, which separates single-cell gene expression into cell-state embeddings ($\mathbf{c}$), global donor embeddings ($\mathbf{V}$), and cell-state-specific donor embeddings ($\mathbf{D} \times \mathbf{C}$) using the cell-state matrix (C), assignment matrix ($\mathbf{A}$), and donor embeddings ($\mathbf{D}$). Figure is directly copied from Vagiaki et al.[10].

To ensure stability and optimize the separation of donor-specific variation, LIVI uses a stepwise training scheme. First, the cell-state space (C) is learned and frozen. Subsequently, the donor spaces (D and V) are learned on top of it. Because the donor space is learned unsupervised and separately from cell state, it is well-suited to test whether it can independently recover external phenotypes such as CTAPs. After training, the linear decoders (W_C, $W_{D\times C}$, W_V) map their factors back to genes, allowing the driving genes to be read off directly. Out of these, only $W_{D\times C}$ is made sparse through a row-wise L_1 penalty, which ensures each donor factor only maps to a small set of genes, enhancing interpretability. W_C and W_V, on the other hand, are dense. The assignment matrix A is pushed towards 0/1 values by an $A(1-A)^2$ penalty and links each donor factor to the specific cell states it is active in, rather than all of them.

2 Results

We applied LIVI to an RA single-cell RNA-seq dataset (Table 1, [3]). The dataset also contains 9 OA donors, which served as a control. The exploratory analysis of the donor space and all association testing only made use of the 70 RA donors as described in Section 4.3. The model was fitted and trained as described in Section 4.2, which produced 15 cell-state factors (C) and 5 global factors (V). To study how the donor space (D) depends on its size, we refitted the model for four different sizes of D (700, 100, 60, 50). Unless stated otherwise, the analyses below use the 700-factor donor space. The effect of dimensionality is evaluated separately in Section 2.5.

Data Type	Details
Cells	314,011
Genes	17,049
Surface proteins	58
Tissue samples	82
Donors	79 (70 RA, 9 OA control)
CTAP groups (RA)	EFM: 7, TF: 8, F: 11, TM: 12, TB: 14, M: 18
Treatment response (RA)	28 treatment-naive, 27 MTX-inadequate, 15 anti-TNF-inadequate

Table 1: Overview of the Zhang et al. synovial CITE-seq cohort. All donors had moderate-to-high disease activity (CDAI $\geq$ 10) [13].

2.1 LIVI's cell-state space recovers the canonical cell types

Before evaluating the donor space (D), we confirmed that the model trained as expected by verifying the separately trained cell-state space (C). We projected all 314,011 cells by their 15 cell-state factors from the 700-factor run using a Uniform Manifold Approximation and Projection (UMAP) and coloured them by their annotated cell type (Figure 3). This projection shows that the embeddings separate the six annotated cell types into distinct clusters. Now we can ask whether the donor space (D), which is trained separately afterward without cell-type or diagnostic labels, can capture donor-specific variation across the CTAPs.

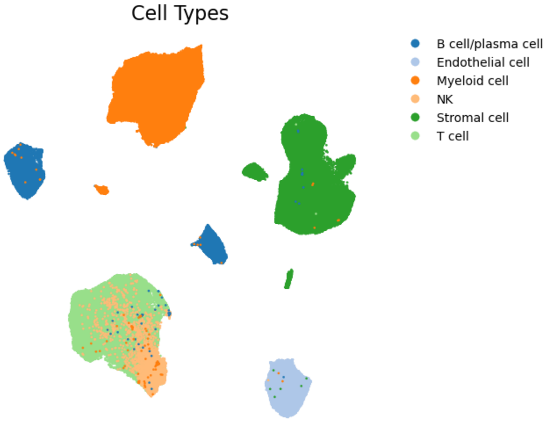


Figure 3: LIVI's cell-state space recovers the canonical cell types. UMAP of the 15 cell-state (C) embeddings from the 700-factor run, over all 314,011 cells, coloured by their annotated cell type.

2.2 The donor space does not separate CTAPs into discrete clusters

To explore the donor space, we projected the 700 donor scores with both a two-dimensional Principal Component Analysis (PCA) and a UMAP (Figure 4). This revealed two things. First, there is no clustering of donors, as the leading components explain little of the variance (PC1 = 3.6%, PC2 = 3.1%). The donor scores form a uniform, highly overlapping space. Second, when donors are annotated by CTAP, the groups overlap without forming distinct regions in both projections. To quantify this visual finding, we computed the silhouette score, a metric ranging from -1 (highly mixed) to $+1$ (perfectly separated) that measures how well data points fit within their assigned clusters [14]. The CTAP labels achieved a score of -0.024 in the 700D donor space. This result confirms that donors do not group by CTAP at this dimensionality.

2.3 Donor space recovers cell-type structure underlying CTAPs

Since the exploratory analysis showed no discrete grouping, we turned from using the donors' discrete CTAP labels to calculating their continuous cell-type abundances, which are counted from Zhang et al.'s cell-type annotations and not produced by LIVI (Section 4.3). Subsequently, we asked whether the donor factors can be associated with any of these abundances using Spearman correlation testing. However, since CTAPs are defined on relative cell-type abundances, these absolute counts had to be converted into compositional percentages. To prevent artifacts from emerging, we applied the centered log-ratio (CLR) transform (Section 4.5), producing a 6-part vector for each of the 70 donors which consists of the relative abundance of each of the six major cell types (T cell, NK, B cell, Myeloid, Endothelial, Fibroblast). We then correlated the abundance of each cell type with the 700 donor scores from LIVI's donor embedding (D, 70 donors $\times$ 700 factors) across all 70 donors, producing a 700×6 matrix of factor-cell-type associations.

Out of these associations, only 18 factors correlated significantly (BH-corrected, $p_{adj} < 0.05$) with the abundance of at least one cell type. Therefore, we decided to analyze the entire

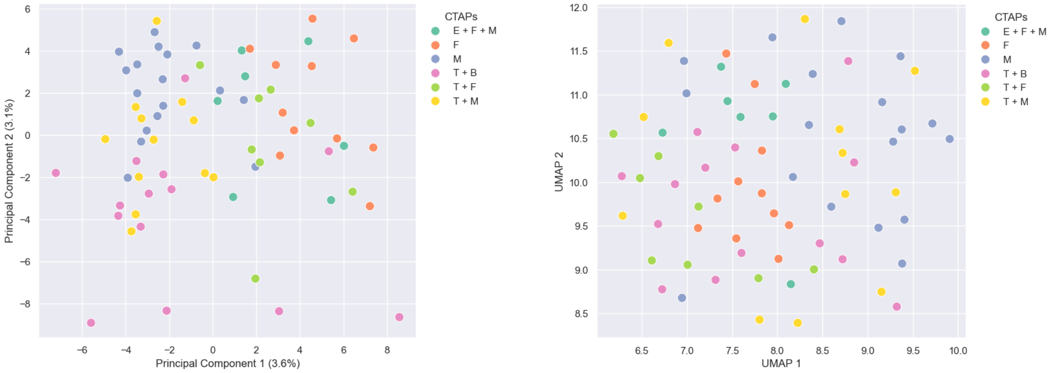
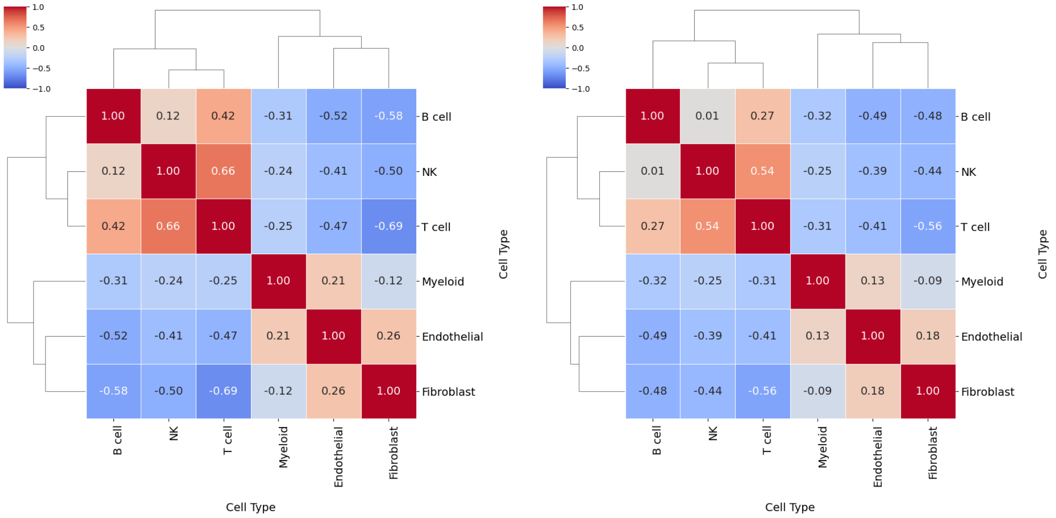


Figure 4: The 700D donor space does not separate CTAPs into discrete clusters. (left) Two-dimensional PCA of the 700 donor factor scores (70 donors). Each point is a donor, coloured by their CTAP. PC1 and PC2 explain only 3.6% and 3.1% of the variance (scree plot, Figure S1). (right) UMAP of the same scores, as a non-linear check. Both projections show extensive overlap between CTAP groups, with no group forming a distinct region.

association profile of each cell type, which consists of all of its 700 Spearman correlations, one per donor factor. Correlating the profiles of each cell type with each other resulted in a 6 × 6 correlation matrix (Figure 5a). This matrix gives us insight into how similar two cell types relate to the donor space (D) with respect to their abundance. For example, T cell-NK ($r = 0.66$) indicates that factors whose scores track T cell abundance tend to track NK abundance as well.

To observe whether these co-correlations form meaningful structures, we applied hierarchical clustering as a visualization. This shows that the matrix clearly splits the co-correlations into two opposing blocks: a lymphoid block (B, NK, and T cells) and a non-lymphoid block (myeloid, endothelial, and fibroblast cells). The strongest signals are the anti-correlations between these two groups (e.g., T cell-fibroblast $= -0.69$ and B cell-endothelial $= -0.52$). Within the lymphoid block, myeloid cells are the most loosely attached, sharing weaker links with the lymphoid cells (weakest is M-NK $= -0.24$) compared to endothelial (weakest is endothelial-NK $= -0.41$) and fibroblast (weakest is fibroblast-NK $= -0.50$) cells. Furthermore, they also show weaker links within the non-lymphoid block itself (myeloid-fibroblast $= -0.12$). This fits the multiple roles synovial myeloid cells play across the CTAPs: they dominate alone in CTAP M, appear with lymphoid T cells in CTAP TM, and with non-lymphoid cells in CTAP EFM.

To verify that this block structure reflects genuine cell-type relationships, we built the same 6 × 6 matrix directly from the CLR-transformed cell-type abundances of the 70 RA donors, using none of the donor factors (Figure 5b). This raw abundance matrix shows the same lymphoid versus non-lymphoid split. Even though both matrices are derived from the same cell-type abundances, the donor factors that reproduce this structure were learned unsupervised, without predefined cell-type annotations and separately from the cell-state space (C). This match shows that the unsupervised donor space recovers the natural cell-type groupings on which CTAPs are built, rather than this being a circular result.



(a) Donor space co-correlation matrix

(b) Raw abundance co-correlation matrix

Figure 5: Cell types group into a lymphoid versus non-lymphoid block in both the donor factor space and the raw abundances. Each entry is a Spearman correlation (r-value) between two cell types (700D). The matrix is clustered using average linkage on Euclidean distance for visualization purposes. (a) The donor space matrix correlates the donor score association profiles of each cell type (vector of length 700) with each other. A positive value means that the cell types relate to the factor scores similarly. (b) The raw abundance matrix correlates the cell types' CLR abundances (vector of length 6) across donors directly without donor factors involved.

2.4 No single factor separates CTAPs at 700D

After establishing that the 700D space is able to capture the cell-type relationships that underlie CTAPs, we tested whether this information concentrates enough in any single donor factor to separate the discrete CTAP groups. Finding such a factor would be highly valuable, as it could potentially serve as a decodable marker for a specific CTAP, which has been shown to predict treatment response [3]. To test this, we ran a discrete non-parametric Kruskal-Wallis test (BH-corrected, Section 4.7) for each donor factor.

However, at both 700D and 100D, zero factors reached significance. This indicates that at the high dimensionalities LIVI was tuned for, the broad cell-type relationships are present (Figure 5a), but no single factor separates the discrete CTAP groups. This suggests that the relevant signal is spread across many factors rather than concentrated in any one.

2.5 Discrete separability concentrates at lower dimensionality

To test the hypothesis we established in Section 2.4, we wanted to see whether reducing the number of donor factors concentrates the CTAP signal enough for individual factors to emerge, which can separate CTAPs discretely. We repeated both analyses, the continuous Spearman co-correlation tests and the discrete Kruskal-Wallis test, at 100D, 60D, and 50D.

Two patterns emerged. First, the continuous lymphoid versus non-lymphoid structure produced by the co-correlations of cell-type abundances and their donor association profiles (Figure 5) persists across all dimensionalities (Figure C). However, the proportion of continuous factors that are significantly correlated with at least one of the cell-type abundances increases as dimensionality shrinks (Table 2). Furthermore, discriminating factors seem to emerge at lower dimensionalities as well, with 5 factors separating CTAPs at 60D and 2 at 50D. This implies that a stronger signal is present at a lower number of donor factors.

D factors	Continuous (Spearman) (n, %)	Discrete (KW)
700	18 (2.6%)	0
100	3 (3%)	0
60	9 (15%)	5
50	12 (24%)	2

Table 2: CTAP signal versus donor-space dimensionality. For each of the four latent dimensionalities, we show the number (and percentage) of donor factors significantly correlated with at least one cell-type abundance (Spearman, BH-corrected) and the number of factors that can discretely separate the CTAP groups (Kruskal-Wallis, BH-corrected).

However, this does not mean the donors themselves form CTAP clusters in the donor space (D). Even at 50D, a PCA and UMAP of the donor scores show significant overlap between CTAP groups (Figure S2), with a silhouette score of -0.016 (compared to -0.024 at 700D), indicating no clustering by CTAP. The separation we found therefore comes from a few individual factors that distinguish the groups on their own, rather than the donor space as a whole separating into CTAP groups.

To see which CTAPs each factor separates, we ran a post-hoc Dunn's test (Table F). For most of the discriminating factors, the significant pairs place the non-lymphoid CTAPs (M, F, EFM) on one side and the T-cell-rich CTAPs (TB, TF, TM) on the other. This shows the same lymphoid versus non-lymphoid division as seen in the continuous analysis

(Section 2.3). The exception is Factor 8, which separates the fibroblast CTAP (F) from the rest (Table F), even though it shows no significant correlation with any individual cell-type abundance (Table 3).

Factor	Dim.	η^2_H	Cell-type abundance correlations (CLR Spearman r)
Factor 13	50D	0.34	M +0.44 / T −0.45, B −0.42
Factor 60	60D	0.32	E +0.41, M +0.38 / T −0.41, B −0.38
Factor 51	60D	0.31	T +0.55
Factor 8	50D	0.30	None significant
Factor 48	60D	0.26	E +0.45 / T −0.49
Factor 39	60D	0.22	B +0.42 / F −0.38
Factor 32	60D	0.20	M +0.48

Table 3: Discriminative factors and their cell-type associations. The seven factors that discretely separate the CTAPs (Kruskal-Wallis, BH-corrected), with their dimensionality, effect size (η^2_H), and significant associations with cell-type abundance (Spearman, BH-corrected): endothelial (E), fibroblast (F), myeloid (M), T cell (T), NK cell (NK), B cell (B).

Finally, we asked whether the seven discriminating factors emerging at lower dimensionalities are simply re-expressions of factors that already exist at 700D (and did not survive BH-correction) or are actual combinations of factors at higher dimensionalities. Therefore, we correlated every 60D factor against all 700D factors across the 70 shared donors, taking the absolute value of each correlation since we only care about how closely two factors track each other, not the direction (Section 4.8). However, after clustering the resulting cross-correlation matrices (Figure D), no clear pattern emerged, showing no low-dimensional factor matching any single 700D factor closely (best $|r| = 0.50$). This suggests that the signal is indeed being spread across many factors rather than carried by one, although the lack of clear structure means we cannot confirm this.

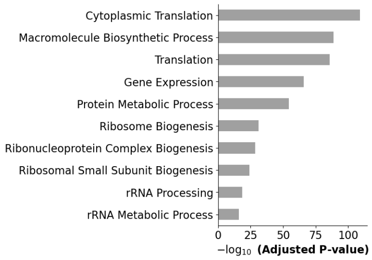
2.6 LIVI's sparsity penalty does not transfer to Zhang et al. dataset

To investigate whether the discriminative factors are associated with biological pathways, we read their gene loadings directly from the sparse decoder. Each D factor corresponds one-to-one to a $D \times C$ factor, whose loadings are read from $W_{D \times C}$ (Section 4.9). However, we found that LIVI's gene-selection threshold did not carry over to our dataset. The 100 × IQR threshold that Vagiaki et al. proposed to isolate trans-eQTL genes returned a median of zero genes per factor at every dimensionality, against a median of two on OneK1K. This indicates that the donor-effect loadings produced on our dataset (generated by Zhang et al. [3]) are far denser than those produced by LIVI on OneK1K. We accordingly selected genes per factor by thresholding using an adaptive elbow (Kneedle) [15] on both the positive and the negative loading tail of the gene weights (Section 4.9).

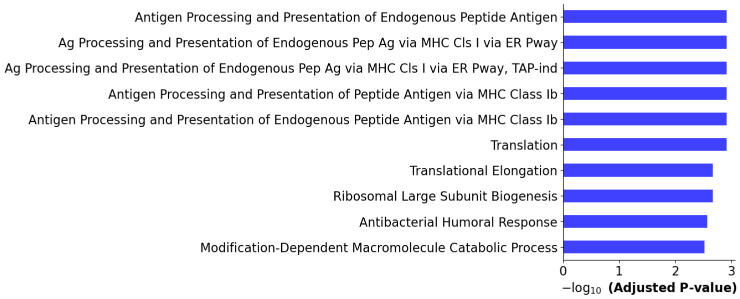
2.7 Ribosomal programmes dominate factor loadings

After thresholding, we ran over-representation analysis (ORA) using Enrichr [16] on the GO Biological Process 2025 gene set. We tested each factor's positive and negative gene tails (Section 4.9) and kept all significant pathways ($FDR < 0.05$). Across all seven discriminative factors, we notice similar results, as loadings are dominated by ribosomal and translation-related programmes, such as cytoplasmic translation (GO:0002181) and macromolecule biosynthetic process (GO:0009059). Figure 6a illustrates this for Factor 13 (negative tail), showing large significance for these pathways ($-\log_{10} p_{adj} = 109.01$ and $-\log_{10} p_{adj} = 88.85$).

When these ribosomal genes are filtered out beforehand, we suddenly find immune-related pathways (antigen presentation, cytokine signalling), some of which are related to RA (Figure 6b). For Factor 13, the surviving pathways ($FDR < 0.05$) relate to antigen processing and presentation of endogenous peptide antigen via MHC class I (GO:0002483 and GO:0002476), which are driven by HLA-A, HLA-B, HLA-C, and B2M. MHC class I antigens have been associated with RA specifically, where an integrative GWAS-expression analysis found HLA-A and HLA-C to be among the top-ranked synovial gene modules [17]. However, since the enrichment after filtering is far weaker ($-\log_{10} p_{adj} = 2.92$) compared to the ribosomal signal and rests on only these three HLA genes, we only report it as an observation and do not interpret it further. The before- and after-filtering enrichment for all discriminative factors is given in Table E.



(a) Before Filtering



(b) After Filtering

Figure 6: Over-representation analysis (Enrichr, GO Biological Process 2025) for Factor 13 (negative tail) at $S = 1.0$, before and after removing ribosomal genes. The bars show the top terms by $-\log_{10}$ adjusted P. (a) Before filtering. (b) After filtering. Note the difference in x-axis scale between the two panels. The results for all factors are given in Figure E.

3 Discussion

Previous latent factor models evaluated on RA single-cell RNA expression data have successfully managed to find biological information associated with RA without relying on pre-annotated cell-type abundances [4]. However, the latent factors of these models capture donor-level and cell-state variation together. LIVI separates these two, which is what led to our question: can its donor space, learned without any labels, recover the CTAPs, and does this change with the number of donor factors?

We have seen that the donor space encodes the cell-type structure underlying the CTAPs, even though it never sees cell-type or diagnostic labels. The factors recover a lymphoid (T/B/NK) versus non-lymphoid (endothelial/fibroblast/myeloid) block that matches the co-correlations of the raw abundances, capturing the same cell-type relationships the CTAPs are built on. Although both matrices are derived from the same cell-type abundances, the factors that reproduce this structure were learned without cell-type labels, so the match shows that the unsupervised donor space recovers the cell-type structure rather than being a circular artifact. Furthermore, this continuous structure is invariant to dimensionality. What changes with dimensionality is whether the signal concentrates enough to emerge in single factors: at 700D, no single factor separates the CTAP groups, but at lower dimensionalities, individual factors appear that distinguish them along the same lymphoid versus non-lymphoid axis.

These results address the two questions Vagiaki et al. [10] left open. First, LIVI's donor space does capture disease-relevant structure in a diagnosed cohort, though it mainly captures broad distinctions in cell-type abundance (the lymphoid versus non-lymphoid axis), while some individual factors can discretely separate RA subtypes across patients. Second, on the question of how many donor factors to use for our given datset, our results show that the appropriate number depends on the goal. Capturing broad general cell-type abundance structure is largely insensitive to dimensionality, whereas detecting individual factors that discretely separate CTAPs requires a dimensionality far below the default of 700, which was tuned for OneK1K.

Sparsity and ribosomal noise. During the extraction of gene loadings for enrichment analysis, we noticed that the weights produced by the $W_{D \times C}$ decoder are far less sparse on our cohort than on OneK1K. The enrichment was dominated by ribosomal and translation programmes, and the original $100 \times$ IQR threshold selected a median of zero genes per factor. We suspect that the L_1 penalty of 0.001 is the cause, since it was calibrated for a dataset with three times the number of cells and more than 10 times the number of donors. However, we cannot establish that definitively here. Alternatively, it could be that the processes driving cell-type abundance are larger gene programmes compared to the single-gene trans-eQTLs LIVI was built to detect.

Limitations. Our research also poses some limitations. First of all, our dataset is small, consisting of 70 donors and 314,011 cells, compared to the 982 donors and over a million cells LIVI was evaluated on. In addition, the CTAP groups have unequal sizes (7-18 donors, Section 4.1). This limits the power of the discrete tests and means some of our findings could differ when run on a larger cohort. Furthermore, even though we use the CLR transform to reduce the effects of compositional data (on which CTAPs are defined), the six transformed values for each donor still sum to zero. This zero-sum constraint inevitably affects the formation of two opposing blocks in the cross-correlation matrices. However, it does not dictate that these blocks must specifically form a lymphoid versus non-lymphoid split. For this reason, the exact correlation values should be carefully interpreted.

Future Work. To further investigate our findings, the most valuable extension would be applying LIVI to a larger RA cohort. With more donors per CTAP, the discrete tests would gain power, which would validate the discriminative factors we found at low dimensionality. Ultimately, this could show whether discrete separation appears even at higher dimensionalities when enough donors are available. Second, to evaluate the sparsity hypothesis we proposed, a penalty-weight ablation study should be run. By varying the L_1 penalty, we could test whether the ribosomal-dominated loadings arise from the penalty being miscalibrated for a cohort of this size, or from other limitations of the smaller dataset.

4 Methods

4.1 Dataset

The dataset we use was generated by Zhang et al.in their research introducing the CTAP stratification [3], using cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) [2]. It consists of 314,011 cells, labeled for 58 surface proteins, across 17,049 genes, from 82 synovial tissue samples: 73 from rheumatoid arthritis (RA) patients and 9 from osteoarthritis (OA) controls. These 82 samples come from 79 unique donors: 70 with RA (3 repeat biopsies) and 9 with OA. The 70 RA patients spanned over a range of clinical treatment responses, with 28 treatment-naive, 27 methotrexate-inadequate, and 15 anti-TNF-inadequate responders, all of which had moderate-to-high disease activity (clinical disease activity index, CDAI ≥ 10) [13]. Each RA patient was assigned a CTAP by Zhang et al.through hierarchical clustering of their relative cell-type abundances (T, B/plasma, NK, myeloid, stromal, endothelial). This produced six unevenly sized groups (EFM: 7, TF: 8, F: 11, TM: 12, TB: 14, M: 18, Figure 1).

4.2 Fitting the model

LIVI [10] was fitted following the configuration of Vagiaki et al., with 15 cell-state (C) and 5 global donor (V) factors, and both sparsity penalty weights ($\lambda_{D\times C}$ and λ_A) set to 10^{-3}. The cell-state factors and their decoder were pre-trained for 90 warmup epochs and frozen, after which the donor components (D and V) were trained to convergence with a learning rate of 8×10^{-4} and a batch size of 1024. To study how the donor space depends on its size, the model was refitted for four donor dimensions (700, 100, 60, and 50 D factors), keeping the number of cell-state and global factors fixed. All runs were produced on the Delft AI Cluster [18].

4.3 Data pre-processing

The model was trained on all 82 samples, producing an 82×700 donor embedding (D). We removed the three samples that were flagged as repeat biopsies, leaving 79 unique donors. From these, all exploratory projections and association testing used the donor factor scores from D of the 70 RA donors, which were annotated for their CTAP by Zhang et al. Furthermore, we took the 314,011 cells, annotated by their donor and cell type, and computed the six cell-type abundances by cross-tabulating the counts per donor.

4.4 Exploratory projections

Cell-state visualization. To verify that the cell-state space had learned canonical cell types before analyzing the donor space, we took the 15 cell-state embeddings (C) from all 314,011 cells and projected them using UMAP, annotating each cell by the cell-type annotation provided by Zhang et al.

Donor-space PCA and UMAP. To explore the donor embeddings, we projected the donor factor scores (D) to two dimensions using both PCA and UMAP (as a non-linear comparison). We did this for every dimensionality (700, 100, 60, 50). To quantify whether there is a clear separation of the CTAP groups, we computed a silhouette score of the CTAP labels for the donor embeddings, which ranges from -1, indicating mixed clusters, to $+1$, indicating separated clusters [14].

4.5 Compositional transform

The cell-type counts from Section 4.3 were converted to relative abundances, which CTAPs are defined on, for our association tests. These relative abundances are compositional: because the six proportions sum to one, a rise in one cell type forces the others down. This dependence, together with the values being bounded between 0 and 1, distorts correlations computed on the raw proportions. We therefore applied a centered log-ratio (CLR) transform. For each donor, we added a pseudocount of 1 to each of the six counts, so that cell types with zero cells do not lead to the logarithm of 0. Then we divided each count by the geometric mean of the donor's six counts and took the natural logarithm. This maps the values onto real numbers instead of proportions bounded between 0 and 1, which removes the bound that distorts correlations. The transformed values still sum to zero per donor, so the transform reduces the compositional effect rather than removing it entirely.

4.6 Continuous association testing

To find associations between the transformed relative cell-type abundances and the donor embeddings, we performed three types of correlation tests. These tests were performed across all four dimensionalities (700, 100, 60, 50). For illustrative purposes, we will refer to the donor embeddings at 700D in the text below.

Donor factor-abundance associations. For each of the donor factors and each of the six cell types, we computed the Spearman correlation between the factor's donor scores (D) and that cell type's CLR abundance, across the 70 donors. This gives a matrix of correlations of size 700×6, whose p-values were corrected using Benjamini-Hochberg across all 4,200 factor × cell-type tests ($p_{adj} < 0.05$).

Donor factors co-correlation. To describe the structure of these associations, we took each cell type's column in the 700×6 factor-cell-type correlation matrix as its association profile across the factors. Subsequently, we Spearman-correlated these six cell-type profiles against each other, producing a 6×6 cell-type co-correlation matrix. This matrix reflects whether two cell types relate to the donor factors in similar ways. We built this matrix from all factor-cell-type correlations, without significance filtering, both to avoid selecting on the structure we are testing and to keep it comparable across dimensionalities.

Raw abundance co-correlation. As a reference, we computed the same 6×6 matrix directly from the CLR abundances, Spearman-correlating each pair of cell types across donors. Because this construction uses no donor factors, it reflects the cell-type relationships with respect to their abundance. Since the CTAPs were derived from the same abundance data, a matching donor-factor matrix, recovers the same cell-type structure underlying the CTAPs as donor factors are learned without cell-type annotations.

Each of the co-correlation matrices was clustered with average linkage on Euclidean distance for visualization purposes only.

4.7 Discrete separation testing

For each factor, a non-parametric Kruskal-Wallis test was run, which compares factor scores across the six CTAP groups ($p < 0.05$). A factor was kept if it was significant after BH-correction. Kruskal-Wallis was considered appropriate since at these small, unequal group sizes ($n = 7 - 18$), methods validating for normality and homoscedasticity (e.g., Shapiro-Wilk, Levene) are very unreliable. Kruskal-Wallis does not rely on these two assumptions. Effect sizes are measured using η^2 (eta-squared, computed from the H statistic) using the equation:

$$\eta_H^2 = \frac{H - k + 1}{n - k} \tag{1}$$

Significant CTAP pairs were identified with Dunn's post-hoc test (BH-corrected).

4.8 Cross-dimensionality testing

To check whether the low-dimensional factors are simply re-expressions of factors already present at 700D, we correlated every factor from a lower dimensionality against every 700D factor (Spearman, across the 70 shared donors). We took the absolute value of each correlation, since a strong match can be either positive or negative and we only care about how closely two factors track each other, not the direction. We did this for the 60D and 50D factors against all 700D factors, producing two cross-correlation matrices (Figure D). Each matrix was clustered with average linkage on Euclidean distance for visualization purposes only.

4.9 Gene extraction and pathway analysis

To extract the genes associated with each discriminative factor, we read them directly from the sparse $D \times C$ decoder $W_{D \times C}$, which maps each factor to its gene loadings (each D factor corresponds to a $D \times C$ factor decoded by $W_{D \times C}$). Because LIVI's default $100 \times$ IQR threshold returned a median of zero genes per factor on our dataset (far less than on OneK1K), we instead extracted genes using the Kneedle algorithm [15], part of the Python `kneed` library. Kneedle adaptively identifies the point of maximum curvature (the knee) to separate the high-weight genes from the near-zero genes in the flat tail. We thresholded the gene weights at the knee point for both the positive and negative loadings. We applied Kneedle at three sensitivity settings ($S = 1.0, 0.5, 0.01$, Table G), which selected a median of 98 to 342 genes per factor across 50D and 60D. The enrichment results were consistent across these settings, so we report $S = 1.0$, which is the `kneed` default.

Since ribosomal genes dominated the loadings, we ran the enrichment both on all genes and after removing ribosomal genes by removing genes starting with RPL and RPS. Afterwards, we compared the two.

For pathway enrichment, Enrichr [16] over-representation analysis (ORA) was run via the `gseapy` library [19] against the GO Biological Process 2025 database. Both tails were tested for every factor, and we report the tail carrying the significant enrichment.

All analyses were performed in Python 3.10.20 using gseapy (1.2.1), scanpy (1.11.5), pandas (2.3.3), numpy (2.2.6), scipy (1.15.2), scikit-learn (1.7.2), statsmodels (0.14.6), and kneed (0.8.6).

5 Responsible Research

In this section, we reflect on the ethical aspects of our research and discuss the reproducibility of the methods we use.

Data access and privacy. Throughout our analysis, we used de-identified, individual-level human omics data (CITE-seq and metadata) from the AMP RA/SLE program published by Zhang et al. Due to the sensitive nature of this data, it is classified as Controlled Access within the ARK Portal. Therefore, we secured the data under the approval of a controlled data use certificate. To this end, we made sure the acquired data had not been shared or redistributed.

Statistical limitations. Since the cohort we used in this research is small and the CTAP groups are unbalanced, the statistical power and validity of our results are limited. To mitigate this problem as much as possible, we emphasize this limitation in our report and apply best statistical practices by using BH correction, reporting effect sizes, and filtering discrete factors on significance.

Reproducibility. The code to produce the results in this report is fully public and hosted at https://github.com/PatrickLo2004/TUDelft-LIVI-Downstream. However, since the clinical dataset cannot be legally shared, anyone who wants to recreate the analysis must apply for Controlled Access through the Synapse ARK Portal. Throughout this report, we documented all configurations and fittings of the model (15/700/5 plus 100/60/50), the specific donors used (70 RA), software versions (and other parameters, e.g., $S = 1.0$ for Kneedle). For training, we used a fixed seed, making the fitting of the model reproducible.

Usage of Generative AI. During this research, generative AI was used to help refine the writing of the report, plotting figures and debugging analysis code. For the biological interpretations of our results, it was used to form a general idea of the concepts and the terminology. Example prompts include: "Please provide a layout with the same structure as this figure from Vagiaki et al.", "Why does this code return this error when I run it?", and "Which of these pathways are associated with T cells?". Furthermore, no individual-level human omics data, patient metadata, or any other sensitive clinical information from the AMP RA/SLE cohort was ever input into these AI tools. All design decisions, experiments, and interpretations of this research were made independently by the authors. Nonetheless, all responses, code, and text were critically tested, evaluated, and verified against the original sources, and we take full responsibility for the content of this report.

References

[1] E. M. Gravallese and G. S. Firestein, "Rheumatoid arthritis - common origins, divergent mechanisms," *New England Journal of Medicine*, vol. 388, no. 6, pp. 529–542, 2 2023.

[2] M. Stoeckius, C. Hafemeister, W. Stephenson, B. Houck-Loomis, P. K. Chattopadhyay, H. Swerdlow, R. Satija, and P. Smibert, "Simultaneous epitope and transcriptome measurement in single cells," *Nature Methods*, vol. 14, no. 9, pp. 865–868, 9 2017. [Online]. Available: https://doi.org/10.1038/nmeth.4380

[3] F. Zhang, A. H. Jonsson, A. Nathan, N. Millard, M. Curtis, Q. Xiao, M. Gutierrez-Arcelus *et al.*, "Deconstruction of rheumatoid arthritis synovium defines inflammatory subtypes," *Nature*, vol. 623, no. 7987, pp. 616–624, 2023. [Online]. Available: https://doi.org/10.1038/s41586-023-06708-y

[4] G. Palla and E. Ferrero, "Latent factor modelling of scRNA-seq data uncovers novel pathways dysregulated in cell subsets of autoimmune disease patients," *bioRxiv*, 2019. [Online]. Available: https://www.biorxiv.org/content/early/2019/11/25/853903

[5] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef, "Deep generative modeling for single-cell transcriptomics," *Nature Methods*, vol. 15, no. 12, pp. 1053–1058, 12 2018. [Online]. Available: https://www.nature.com/articles/s41592-018-0229-2

[6] V. Svensson, A. Gayoso, N. Yosef, and L. Pachter, "Interpretable factor models of single-cell RNA-seq via variational autoencoders," *Bioinformatics*, vol. 36, no. 11, pp. 3418–3421, 6 2020. [Online]. Available: https://doi.org/10.1093/bioinformatics/btaa169

[7] G. L. Stein-OâBrien, B. S. Clark, T. Sherman, C. Zibetti, Q. Hu, R. Sealfon, S. Liu, J. Qian, C. Colantuoni, S. Blackshaw, L. A. Goff, and E. J. Fertig, "Decomposing cell identity for transfer learning across cellular measurements, platforms, tissues, and species," *Cell Systems*, vol. 8, no. 5, pp. 395–411.e8, 5 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2405471219301462

[8] K. K. Dey, C. J. Hsiao, and M. Stephens, "Visualizing the structure of RNA-seq expression data using grade of membership models," *PLOS Genetics*, vol. 13, no. 3, p. e1006599, 3 2017. [Online]. Available: https://dx.plos.org/10.1371/journal.pgen.1006599

[9] H. M. Levitin, J. Yuan, Y. L. Cheng, F. J. Ruiz, E. C. Bush, J. N. Bruce, P. Canoll, A. Iavarone, A. Lasorella, D. M. Blei, and P. A. Sims, "De novo gene signature identification from single-cell RNA-seq with hierarchical poisson factorization," *Molecular Systems Biology*, vol. 15, no. 2, p. e8557, 2 2019. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC6386217/

[10] D. Vagiaki, T. Heinen, M. Saraswat, B. Clarke, and O. Stegle, "Mapping trans-eqtls at single-cell resolution using latent interaction variational inference," *bioRxiv*, 2026. [Online]. Available: https://www.biorxiv.org/content/early/2026/02/06/2026.02.04.703363

[11] S. Yazar, J. Alquicira-Hernandez, K. Wing, A. Senabouth, M. G. Gordon, S. Andersen, Q. Lu, A. Rowson, T. R. P. Taylor, L. Clarke, K. Maccora, C. Chen, A. L. Cook, C. J. Ye, K. A. Fairfax, A. W. Hewitt, and J. E. Powell, "Single-cell eqtl mapping identifies cell typeâspecific genetic control of autoimmune disease," *Science*, vol. 376, no. 6589, p. eabf3041, 2022. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.abf3041

[12] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," 2013. [Online]. Available: https://arxiv.org/abs/1312.6114

[13] D. Aletaha, V. P. Nell, T. Stamm, M. Uffmann, S. Pflugbeil, K. Machold, and J. S. Smolen, "Acute phase reactants add little to composite disease activity indices for rheumatoid arthritis: validation of a clinical activity score," *Arthritis*

Research & Therapy, vol. 7, no. 4, pp. R796–R806, 2005. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC1175030/

[14] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0377042787901257

[15] V. Satopa, J. Albrecht, D. Irwin, and B. Raghavan, "Finding a "kneedle" in a haystack: Detecting knee points in system behavior," in *31st International Conference on Distributed Computing Systems Workshops*, 2011, pp. 166–171.

[16] M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, and A. Ma'ayan, "Enrichr: a comprehensive gene set enrichment analysis web server 2016 update," *Nucleic Acids Research*, vol. 44, pp. W90–W97, 7 2016. [Online]. Available: https://doi.org/10.1093/nar/gkw377

[17] X. Xiao, J. Hao, Y. Wen, W. Wang, X. Guo, and F. Zhang, "Genome-wide association studies and gene expression profiles of rheumatoid arthritis," *Bone & Joint Research*, vol. 5, no. 7, pp. 314–319, 8 2016. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC5005471/

[18] Delft AI Cluster (DAIC), "The delft ai cluster (daic), rrid:scr_025091," 2024. [Online]. Available: https://daic.tudelft.nl/

[19] Z. Fang, X. Liu, and G. Peltz, "GSEApy: a comprehensive package for performing gene set enrichment analysis in Python," *Bioinformatics*, p. btac757, 2022. [Online]. Available: https://doi.org/10.1093/bioinformatics/btac757

Appendix A Scree plot of the 700D donor space

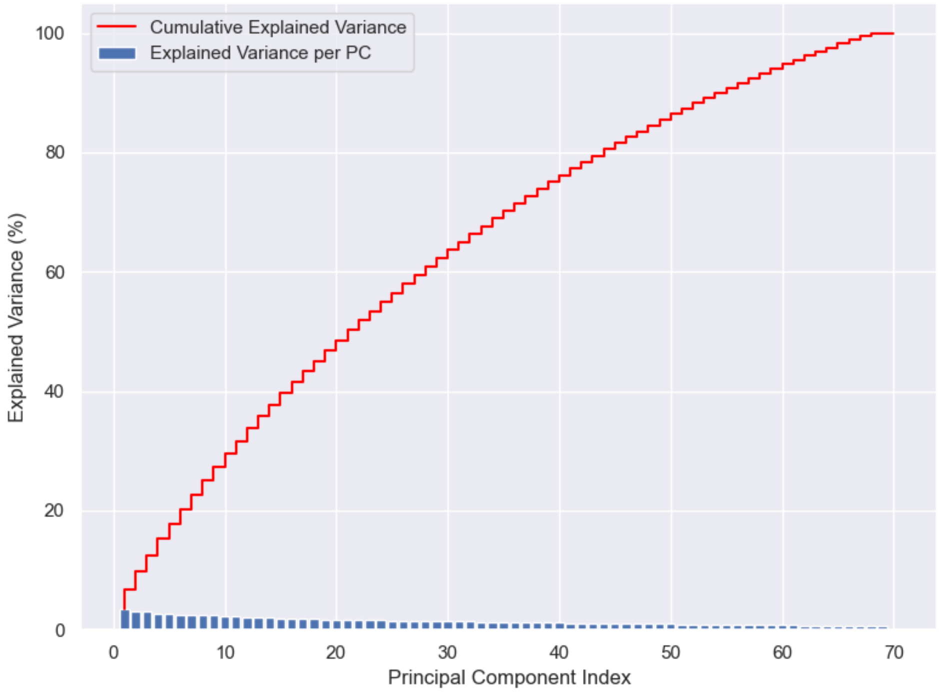
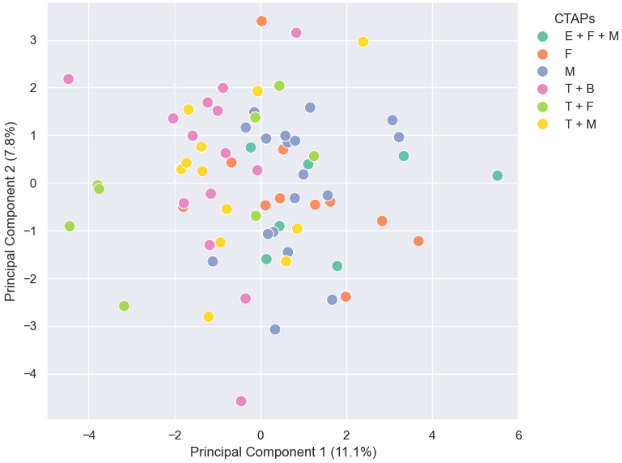
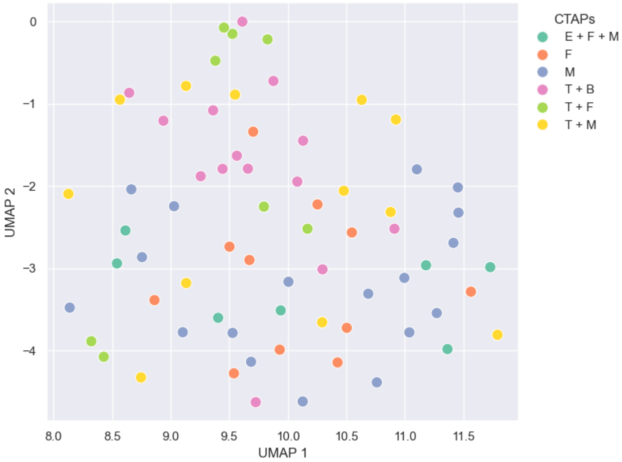


Figure S1: Scree plot of the $700D$ donor space, showing the percentage of variance explained by each successive principal component of the donor factor scores.

Appendix B Donor space projections at 50D



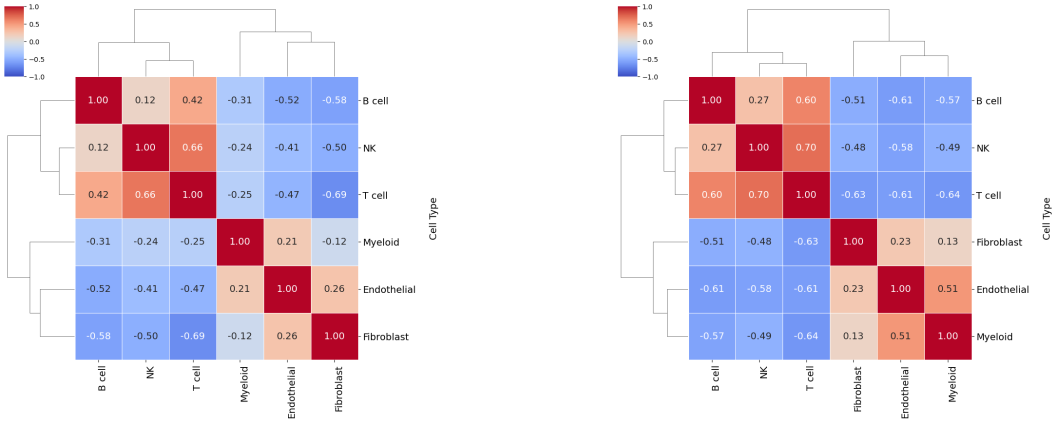
(a) PCA



(b) UMAP

Figure S2: Two-dimensional projections of the $50D$ donor factor scores: (a) PCA and (b) UMAP. Each point is a donor, coloured by CTAP.

Appendix C Cell-type co-correlation matrices across dimensionalities

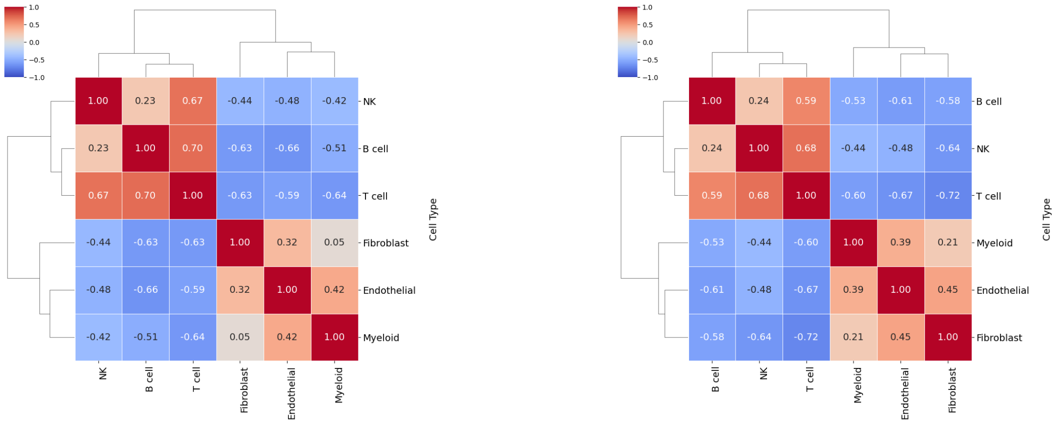


	B cell	NK	T cell	Myeloid	Endothelial	Fibroblast
B cell	1.00	0.12	0.42	-0.31	-0.52	-0.58
NK	0.12	1.00	0.66	-0.24	-0.41	-0.50
T cell	0.42	0.66	1.00	-0.25	-0.47	-0.69
Myeloid	-0.31	-0.24	-0.25	1.00	0.21	-0.12
Endothelial	-0.52	-0.41	-0.47	0.21	1.00	0.26
Fibroblast	-0.58	-0.50	-0.69	-0.12	0.26	1.00

	B cell	NK	T cell	Fibroblast	Endothelial	Myeloid
B cell	1.00	0.27	0.60	-0.51	-0.61	-0.57
NK	0.27	1.00	0.70	-0.48	-0.58	-0.49
T cell	0.60	0.70	1.00	-0.63	-0.61	-0.64
Fibroblast	-0.51	-0.48	-0.63	1.00	0.23	0.13
Endothelial	-0.61	-0.58	-0.61	0.23	1.00	0.51
Myeloid	-0.57	-0.49	-0.64	0.13	0.51	1.00

(a) $700D$

(b) $100D$



	NK	B cell	T cell	Fibroblast	Endothelial	Myeloid
NK	1.00	0.23	0.67	-0.44	-0.48	-0.42
B cell	0.23	1.00	0.70	-0.63	-0.66	-0.51
T cell	0.67	0.70	1.00	-0.63	-0.59	-0.64
Fibroblast	-0.44	-0.63	-0.63	1.00	0.32	0.05
Endothelial	-0.48	-0.66	-0.59	0.32	1.00	0.42
Myeloid	-0.42	-0.51	-0.64	0.05	0.42	1.00

	B cell	NK	T cell	Myeloid	Endothelial	Fibroblast
B cell	1.00	0.24	0.59	-0.53	-0.61	-0.58
NK	0.24	1.00	0.68	-0.44	-0.48	-0.64
T cell	0.59	0.68	1.00	-0.60	-0.67	-0.72
Myeloid	-0.53	-0.44	-0.60	1.00	0.39	0.21
Endothelial	-0.61	-0.48	-0.67	0.39	1.00	0.45
Fibroblast	-0.58	-0.64	-0.72	0.21	0.45	1.00

(c) $60D$

(d) $50D$

Figure S3: Cell-type co-correlation in the donor factor space at four dimensionalities: **(a)** $700D$, **(b)** $100D$, **(c)** $60D$, **(d)** $50D$. Each panel is a 6×6 matrix where each entry is the Spearman correlation between the association profiles of two cell types across the donor factors, a positive value means the two cell types relate to the factor scores similarly. Each matrix is clustered using average linkage on Euclidean distance for visualization purposes.

Appendix D Cross-dimensionality factor correlations

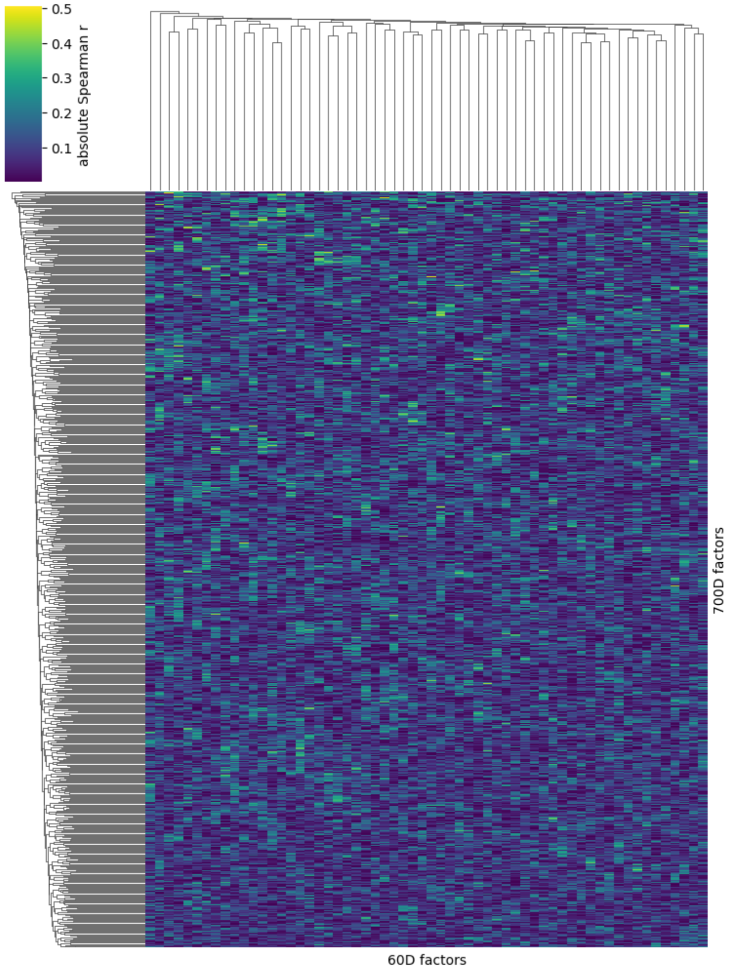


Figure S4: Cross-correlation between every 60D factor and every 700D factor. Each entry is the absolute Spearman correlation of the two factors' donor scores across the 70 RA donors. Rows and columns are clustered using average linkage on Euclidean distance for visualization purposes.

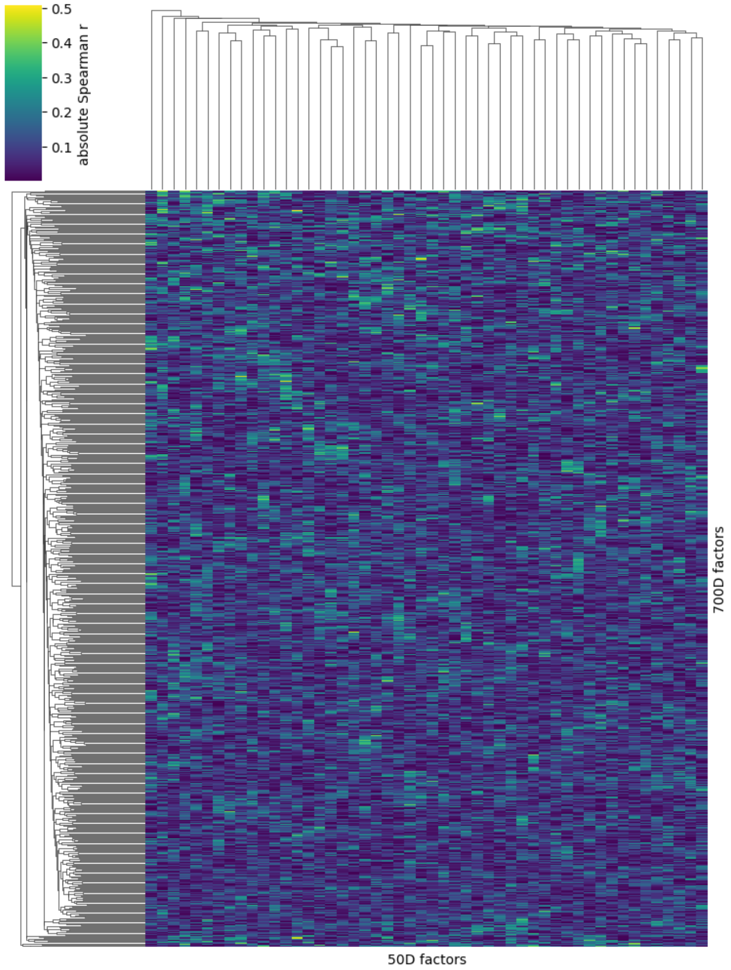


Figure S5: Cross-correlation between every $50D$ factor and every $700D$ factor. Each entry is the absolute Spearman correlation of the two factors' donor scores across the 70 RA donors. Rows and columns are clustered using average linkage on Euclidean distance for visualization purposes.

Appendix E Pathway enrichment for all discriminative factors, before and after filtering ribosomal genes

Factor	**Tail**	**Top three enriched terms before filtering**	$-\log_{10} p_{\mathrm{adj}}$
Factor 8 (50D)	negative	Cytoplasmic Translation	28.32
		Macromolecule Biosynthetic Process	22.79
		Translation	20.73
Factor 13 (50D)	negative	Cytoplasmic Translation	109.01
		Macromolecule Biosynthetic Proces	88.85
		Translation	85.83
Factor 32 (60D)	negative	Cytoplasmic Translation	55.11
		Macromolecule Biosynthetic Process	44.48
		Translation	41.43
Factor 39 (60D)	positive	Cytoplasmic Translation	53.05
		Macromolecule Biosynthetic Process	41.82
		Translation	41.03
Factor 48 (60D)	positive	Cytoplasmic Translation	7.99
		Translation	6.26
		Macromolecule Biosynthetic Process	6.02
Factor 51 (60D)	negative	Cellular Response to Zinc Ion	3.66
		Response to Cadmium Ion	3.66
		Cellular Response to Copper Ion	3.66
Factor 60 (60D)	negative	Cytoplasmic Translation	98.79
		Macromolecule Biosynthetic Process	79.81
		Translation	75.81

Table 4: Top three enriched terms before ribosomal filtering, for each discriminative factor. The table shows the three most significant GO Biological Process (2025) terms per factor tail found using Enrichr over-representation analysis at $S = 1.0$ (FDR < 0.05). Each factor is listed with the dimensionality at which it was discriminative and the tail tested.

Factor	**Tail**	**Top three enriched terms after filtering**	$-\log_{10} p_{\text{adj}}$
Factor 8 (50D)	negative	Antimicrobial humoral response	5.47
		Antimicrobial humoral immune response	4.05
		Antigen processing & presentation (MHC-I)	3.19
Factor 13 (50D)	negative	antigen processing and presentation of endogenous peptide antigen (MHC-I)	2.92
		antigen processing and presentation of endogenous peptide antigen via MHC class I via ER pathway	2.92
		Aantigen processing and presentation of endogenous peptide antigen via MHC class I via ER pathway, TAP-independent	2.92
Factor 32 (60D)	negative	Antimicrobial humoral response	2.46
		Antimicrobial humoral immune response	2.28
		ATP biosynthetic process	1.83
Factor 39 (60D)	positive	None	–
Factor 48 (60D)	negative	antigen processing and presentation of endogenous peptide antigen (MHC-I)	4.25
		antigen processing and presentation of endogenous peptide antigen via MHC class I via ER pathway	3.90
		antigen processing and presentation of endogenous peptide antigen via MHC class Ib	3.90
Factor 51 (60D)	negative	Cellular response to zinc ion	3.74
		Response to cadmium ion	3.74
		Cellular Response to Copper Ion	3.74
Factor 60 (60D)	negative	Cytokine-mediated signalling pathaway	3.34
		Antimicrobial humoral immune response	3.18
		Antimicrobial humoral response	2.82

Table 5: Top three enriched terms after ribosomal filtering, for each discriminative factor. The table shows the three most significant GO Biological Process (2025) terms per factor tail found using Enrichr over-representation analysis at $S = 1.0$ (FDR < 0.05). Each factor is listed with the dimensionality at which it was discriminative and the tail tested.

Appendix F CTAP pairs separated by each discriminative factor (Dunns)

Factor	**Separated CTAP pair**	p_{adj}
Factor 8 (50D)	F vs. M	2.4×10^{-3}
	F vs. EFM	3.2×10^{-3}
	F vs. TF	7.8×10^{-3}
	F vs. TM	4.1×10^{-2}
Factor 13 (50D)	M vs. TB	2.3×10^{-3}
	EFM vs. TB	2.3×10^{-3}
	EFM vs. TM	9.1×10^{-3}
	M vs. TM	1.5×10^{-2}
	EFM vs. TF	3.2×10^{-2}
Factor 32 (60D)	F vs. M	1.3×10^{-2}
	M vs. TB	1.3×10^{-2}
Factor 39 (60D)	M vs. TB	1.1×10^{-2}
	M vs. TM	1.1×10^{-2}
Factor 48 (60D)	EFM vs. TF	5.0×10^{-3}
	EFM vs. TB	1.1×10^{-2}
	EFM vs. TM	1.1×10^{-2}
	M vs. TF	2.0×10^{-2}
	F vs. TF	2.1×10^{-2}
Factor 51 (60D)	F vs. TB	5.2×10^{-3}
	F vs. TF	8.9×10^{-3}
	M vs. TB	1.3×10^{-2}
	F vs. TM	1.3×10^{-2}
	EFM vs. TB	1.7×10^{-2}
	M vs. TF	2.0×10^{-2}
	EFM vs. TF	2.0×10^{-2}
	M vs. TM	4.5×10^{-2}
	EFM vs. TM	4.5×10^{-2}
Factor 60 (60D)	EFM vs. TB	7.2×10^{-4}
	M vs. TB	9.8×10^{-4}
	EFM vs. TF	1.3×10^{-2}
	EFM vs. TM	2.6×10^{-2}
	F vs. TB	3.3×10^{-2}
	M vs. TF	3.6×10^{-2}

Table 6: CTAP pairs separated by each discriminative factor, from Dunn's post-hoc test (BH-corrected). The table shows all pairwise differences significant at $p_{adj} < 0.05$. Each factor is listed with the dimensionality at which it was discriminative, and pairs are ordered by p_{adj}. CTAPs: endothelial-fibroblast-myeloid (EFM), fibroblast (F), T-fibroblast (TF), T-B (TB), T-myeloid (TM), myeloid (M).

Appendix G Median genes detected for each Kneedle sensitivity parameter (S)

S	**50**D	**60**D
1.0	341.5	250.0
0.5	118.0	151.0
0.01	98.5	117.0

Table 7: Median number of genes selected per factor by the Kneedle algorithm, for each sensitivity setting S and donor-space dimensionality.