CAPTURING CLINICAL HETEROGENEITY IN RHEUMATOID ARTHRITIS

EVALUATING THE LIVI LATENT SPACE USING GENE EXPRESSION DATA

Author: Patrick Lo

Responsible Professor: Marcel Reinders

Supervisors: Inez den Hond & Kirti Biharie

Background

- Rheumatoid arthritis (RA) is a **heterogeneous** autoimmune disease: patients with the same diagnosis respond differently to the same therapy.
- Zhang et al.[1] stratified synovial RA cells from 70 RA donors into six **cell-type abundance phenotypes (CTAPs)**. This was done through clustering of relative abundances of six major cell-types (T, B/plasma, NK, myeloid, endothelial and fibroblast), produced from **counting pre-annotated cells.**

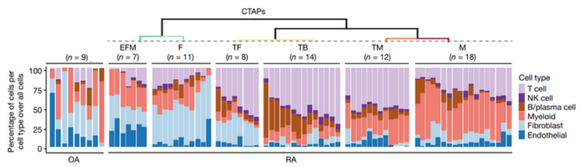

Figure 1: Hierarchical clustering of donors by their six relative cell type abundances producing CTAPs. Figure copied from Zhang et al[1].

- **Latent factor models** learn without relying on pre-annotated counts
- We evaluate the **Latent Interaction Variational Inference (LIVI)** model (Vagiaki et al.)[2]
- Authors left **two questions** open: 1. What is the optimal size of the latent space for a given dataset? 2. Can it capture disease status on diagnosed donors?

Research Question

- 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?
- Does the **decoder link** CTAP-related **factors to biological pathways**?

Methods

- Built on a **Variational Auto Encoder**
- Separates **donor-specific variation (D), cell-state variation (C) and global donor effects (V)** in separate latent embeddings
- Each of the latent factors can be decoded using a **linear decoder** mapping factors directly to genes. Only the **donor-effect decoder (W_{DxC})** is **sparse**, an **L1 penalty** restricts each donor factor to a few genes.
- LIVI is trained **without cell-type** or **diagnostic labels**

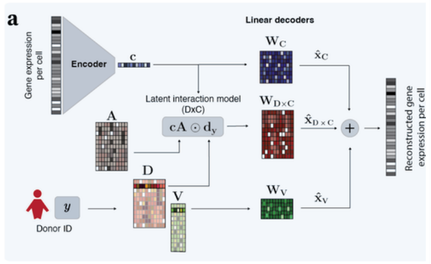

Figure 2: Architectural overview of the LIVI model. A (assignment matrix) assigns donor factors (D) to specific cell-states (C). Figure copied from Vagiaki et al.[2]

1. **Model fitting, dataset and pre-processing:**
   - We fit **15 cell-state factors (C)** and **5 global donor factors (V)** on **82 samples,** refitting the **donor space (D)** at **700, 100, 60** and **50 factors (D)**
   - The dataset is **314,011 cells** from **70 RA** and **9 OA** donors, each cell annotated by **six major cell-types** (Zhang et al.). All analysis uses the **70 RA donor**s (3 repeat biopsies removed), labelled by **CTAP**.
   - We compute **raw cell-type abundances** using Zhang et al. annotations
   - We applied **centred log-ratio (CLR) transform** to **relative cell-type abundances**, which maps proportions (0 to 1) to unbounded real numbers.
2. **Exploratory projections:**
   - **Principal Component Analysis (PCA)** and **Uniform Manifold Approximation Projection (UMAP)**
3. **Continuous association testing**
   - **Spearman correlations** (corrected using Benjamini Hochberg (BH)) on 70 RA donors for their six CLR cell-type abundances and their donor scores (6x700 matrix for 700D)
   - **Donor space co-correlation matrix:** correlate association profiles (all donor-factor correlations) of each cell-type with each other (6x6 matrix)
   - **Raw co-correlation matrix:** correlate CLR cell-type abundances of each cell-type with each other (6x6 matrix)
4. **Discrete separation testing:**
   - Run **Kruskal-Wallis** test (BH-corrected) comparing donor factor scores across the six CTAPs
5. **Cross-dimensionality testing:**
   - Compute spearman correlations between higher (700D) donor factors and lower (60D, 50D) donor factors to measure signal across dimensions.
6. **Gene extraction and pathway analysis:**
   - Each D factor maps one-to-one to a D×C factor, whose gene loadings were read from the donor-effect decoder ($W_{D\times C}$). Loadings were thresholded using the **Kneedle** algorithm (adaptive elbow) and over-representation analysis (ORA) was run on the Biological Processes 2025 gene set, both with and without filtering out ribosomal genes.

Results

Donor space does not cluster CTAP groups

- PCA and UMAP on donor space (700D) revealed that CTAP annotated donors **overlap greatly** (PC1 = 3.6%)
- Silhouette score −1 (highly mixed) to +1 (perfectly separated) on CTAP labels: **- 0.024 → No clustering**
- So we do not rely on discrete CTAP labels but analyze **relative cell-type abundances** (which define them)

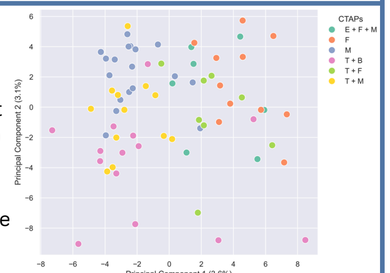

Figure 3: PCA of the 700 donor space

Donor space recovers cell-type structure underlying CTAPs

- Donor space co-correlation matrix (left) reveals how **similarly cell-types relate** to the donor scores
- Clustering the matrix reveals a **two block split: lymphoid** (B cell, NK, T cell) vs. **non-lymphoid** cells (endothelial, myeloid, fibroblast). Donor scores similarly track these cell-type abundances within each block (e.g., T-NK = 0.66)
- When compared to the raw co-correlation matrix (right) produced directly from CLR abundances we see the **same split**
- The cell-type abundance relationships which underlie CTAP are **captured by the donor scores**, running the analysis on fittings for **100, 60** and **50D produced the same result**

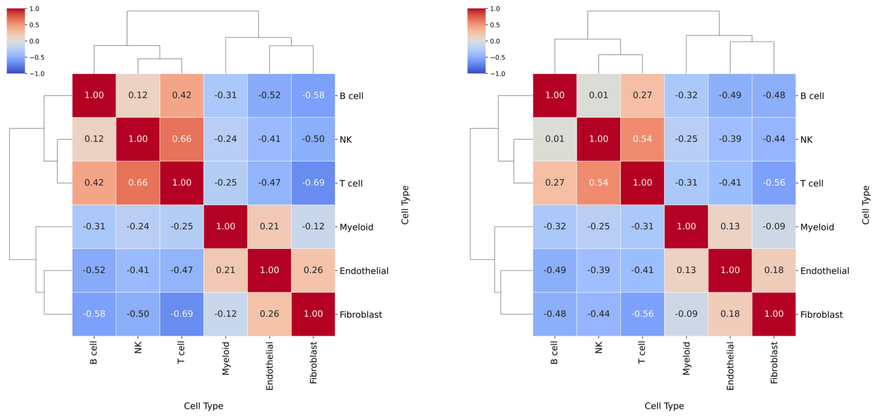

Figure 4: Donor space co-correlation matrix (left) and raw abundance co-correlation matrix (right)

Discrete separability concentrates at lower dimensionality

- **No significant** CTAP discriminating factors were found at **700D** and **100D.**
- At 60D, 5 factors emerged and, at 50D, 2 factors, which split CTAPs along the lymphoid / non-lymphoid axis identified by continuous test
- Proportion of factors **significantly correlated** with at least one cell-type abundance **rose when dimensionality decreases**
- Cross-dimensionality testing (700D x 60D and 700D x 50D) revealed **no factors strongly correlated** (max |r| = 0.50)

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 1: CTAP signal versus donor-space dimensionality.

Ribosomal programmes dominate factor loadings

- To extract genes, all donor factors were thresholded on both their positive and negative tail using 100 x IQR: **0 genes survived,** loadings are dense.
- Thresholding gene weights of **7 discriminating factors** (Kneedle) resulted in **ribosomal and translational pathways** dominating
- When these are filtered out beforehand, pathways with **weaker significance** emerge, associated with RA: **MHC-I antigens** (HLA-A, HLA-B, HLA-C and B2M)[3]

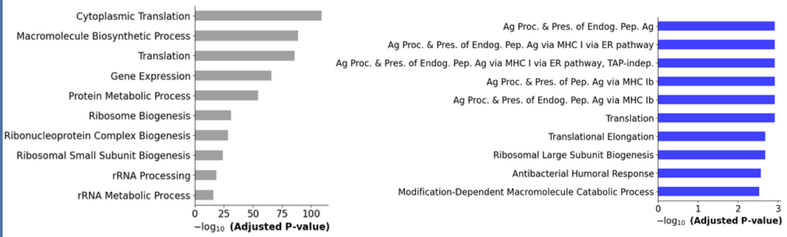

Figure 5 Over-representation analysis (Enrichr, GO Biological Process 2025) for discriminating Factor 13 (negative tail)

Conclusions

- The donor space **captures underlying cell-type relationships** which CTAPs are built on, **without** having seen any **pre-annotated** cell-types or diagnostic **labels**. (**lymphoid** vs **non-lymphoid** block)
- These broad relationships are captured **across all dimensions** (700, 100, 60, 50)
- At **lower dimensionalities discriminating factors emerge,** separating CTAP groups along the same lymphoid versus non-lymphoid axis.
- However, the loadings of these factors are dominated by **ribosomal genes** and are much **denser**. We suspect the **sparsity penalty** is the bottleneck.
- So the LIVI donor space is able to capture **disease-relevant structure** on a diagnosed cohort but on a **broad scale** (lymphoid, non-lymphoid axis)
- The optimal number of dimensions for our given dataset **depends on the goal**. If you want to capture discriminating factors, **lower dimensions** provide **higher resolution**

Limitations

Our cohort is **small (70 donors)** with uneven CTAP groups (7-18 each, Figure 1), so discrete tests have limited power. Furthermore, CLR values **still sum to zero** which inevitably influences the two-block structure we observe

References

1. Zhang, F., Jonsson, A. H., Nathan, A., Millard, N., Curtis, M., Xiao, Q., Gutierrez-Arcelus, M., et al. (2023). Deconstruction of rheumatoid arthritis synovium defines inflammatory subtypes. Nature, 623(7987), 616–624. https://doi.org/10.1038/s41586-023-06708-y
2. Vagiaki, D., Heinen, T., Saraswat, M., Clarke, B., & Stegle, O. (2026). Mapping trans-eQTLs at single-cell resolution using Latent Interaction Variational Inference. bioRxiv. https://doi.org/10.64898/2026.02.04.703363
3. Xiao, X., Hao, J., Wen, Y., Wang, W., Guo, X., & Zhang, F. (2016). Genome-wide association studies and gene expression profiles of rheumatoid arthritis. Bone & Joint Research, 5(7), 314–319. https://doi.org/10.1302/2046-3758.57.2000502

Contact

Email: P.A.Lo@student.tudelft.nl