LEXA: LEARNING A LEXICON OF GENE-PROGRAMME STATES FROM SINGLE-CELL TRANSCRIPTOMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-cell transcriptomic modeling faces a persistent trade-off between preserving complex cellular variation and providing biologically interpretable representations. Continuous latent models capture rich cellular information but are difficult to interpret at the level of individual dimensions, whereas factorization- and topic-based gene-programme methods provide interpretable components but retain less fine-grained variation. We propose LEXA, which couples programme-factorized vector quantization with scaffold-guided gene responsibility and gene-resolved additive decoding to organize cellular variation into programme-specific axes, where each cell selects one of multiple discrete states along each axis, forming a compositional lexicon with explicit gene-resolved molecular effects. We pretrained LEXA on CELLxGENE and evaluated on multiple held-out datasets against representative continuous, discrete, and gene-programme baselines, and found that it preserves substantially more cellular information than conventional gene-programme methods while yielding state-level molecular interpretations that are biologically coherent and reproducible. In an external COVID-19 cohort, both the atlas-pretrained and cohort-specific LEXA settings revealed severity-associated programme variation across immune lineages beyond conventional cell-subtype descriptions. These results demonstrate LEXA as a framework for reconciling information-rich representation with explicit molecular interpretation, turning cellular variation into enumerable and biologically testable molecular states.

1 INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) profiles cellular identity and state through high-dimensional gene-expression patterns shaped by multiple, concurrently active biological programmes (Wagner et al., 2016). Modeling such data involves a persistent trade-off between preserving complex cellular variation and providing biologically interpretable representations. Continuous latent models, such as variational autoencoders (VAEs) and more recent large-scale pretrained models, can retain rich information useful for downstream analysis (Lopez et al., 2018; Theodoris et al., 2023; Cui et al., 2024; Hao et al., 2024), but individual latent dimensions generally lack direct biological meaning. In contrast, factorization- and topic-based gene-programme methods provide interpretable, gene-grounded components, but their constrained programme decompositions can discard substantial fine-grained cellular information (Blei et al., 2003; Kotliar et al., 2019; Kunes et al., 2024). Moreover, each programme is typically represented by a fixed gene-loading pattern with a continuous score, capturing variation in programme strength but not explicitly representing alternative molecular configurations within the same programme. We therefore seek a representation that combines the information capacity of flexible latent models with the interpretability of gene-programme decomposition by representing each programme through distinct molecular states.

To bridge this gap, we propose LEXA (Lexicon of EXpression-state Axes). Building on VQ-VAE-style vector quantization, LEXA factorizes cellular variation into programme-specific axes, each equipped with its own discrete codebook, and uses a training-derived scaffold to encourage different axes to capture distinct gene programmes. A gene-resolved additive decoder reconstructs each transcriptome as the sum of axis-specific contributions and assigns every axis–state pair an exact signed gene-effect vector. Each cell is therefore represented by a combination of states selected across distinct programme-specific axes, forming a compositional lexicon that links discrete programme states

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

directly to gene-expression changes. The same framework can be pretrained on heterogeneous atlases to learn transferable programme structure, or fitted *de novo* within a target cohort to resolve more context-specific molecular states.

We evaluated LEXA on the CELLxGENE atlas and multiple held-out datasets. We first assessed representational capacity by examining the trade-off between representation size and downstream prediction performance, showing that LEXA retains substantial cellular information despite its discrete and structured representation. We then evaluated whether the learned state-level molecular effects are biologically meaningful by testing their agreement with phenotype-associated expression changes and coherent biological pathways; LEXA outperformed established interpretable gene-programme methods in these analyses. We also tested reproducibility across independently initialized training runs and found that the LEXA-learned programme effects recurred across fits. Finally, in a case study on an external COVID-19 cohort, we applied both atlas-pretrained and cohort-specific LEXA representations and showed that they reveal severity-associated programme variation across immune lineages beyond conventional cell-subtype descriptions.

Together, these results establish LEXA as a framework that reconciles information-rich cellular representations with explicit molecular interpretation, enabling biologically meaningful variation to be resolved and compared beyond conventional cell labels or opaque latent representations.

2 Related work

Continuous representations in single-cell transcriptomics. A large class of single-cell methods represents cells in continuous latent spaces. scVI learns probabilistic latent representations while accounting for technical variation (Lopez et al., 2018), scArches extends reference models to new datasets (Lotfollahi et al., 2022), and Harmony integrates cells by adjusting continuous coordinates (Korsunsky et al., 2019). More recently, single-cell foundation models such as Geneformer, scGPT, and scFoundation learn transferable cell representations from large single-cell collections (Theodoris et al., 2023; Cui et al., 2024; Hao et al., 2024). These models can preserve rich cellular variation and support a broad range of downstream tasks, but the biological meaning of individual latent dimensions is generally not explicit.

Gene-programme decomposition in single-cell data. Gene-programme methods provide a complementary route to interpretability by decomposing transcriptomes into reusable gene-level components. Topic models represent cells as mixtures of latent topics (Blei et al., 2003), while cNMF uses consensus non-negative matrix factorization to recover recurrent gene-expression programmes and their cell-level usages (Kotliar et al., 2019). Spectra further incorporates prior biological knowledge to guide programme discovery (Kunes et al., 2024). These approaches are naturally compositional and interpretable, but their constrained programme decompositions can discard fine-grained cellular information. Moreover, each programme is typically represented by a fixed gene-loading pattern with a continuous cell-level score, capturing variation in programme strength but not explicitly representing alternative molecular configurations within the same programme.

Discrete representations for single-cell transcriptomes. Recent work has also explored discrete representations of single-cell data. CellVQ maps cells to a learned codebook (Wang et al., 2026), while CellTok represents transcriptomes through discrete token sequences (Xiao et al., 2026). These approaches build on broader ideas from vector quantization and discrete representation learning (van den Oord et al., 2017; Jégou et al., 2011), but the codes themselves are not explicitly endowed with biological meaning or anchored across training runs, and their combinations are not guaranteed to correspond to coherent or interpretable biological states.

3 The LEXA framework

LEXA combines the information capacity of continuous latent models, the gene-level interpretability of programme decomposition, and the enumerable states of discrete representations. As shown in Figure 1, it represents each cell as a composition of states selected across multiple programme-specific axes. Each axis contains a finite codebook of alternative states, allowing the same programme to adopt distinct expression configurations across cells. A training-derived scaffold encourages different axes to capture distinct gene programmes, while a shared additive decoder maps each

selected state to a gene-resolved contribution. The transcriptome is reconstructed as the sum of these axis-specific contributions, linking information preservation directly to the programme-state representation. Consequently, each axis–state pair is associated with an explicit molecular effect in gene-expression space.

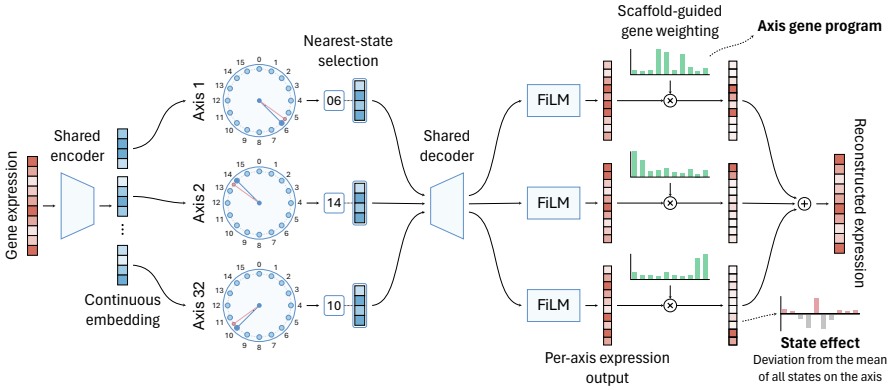

Figure 1: Overview of LEXA. Each cell is encoded along multiple axes and assigned one spherical codebook state per axis, producing a compositional discrete profile grounded in gene expression.

3.1 Discrete encoding

Let $\mathbf{x}_i \in \mathbb{R}^G$ denote the normalized log-expression vector of cell i. A shared encoder maps $\mathbf{x}_i$ to a structured latent representation $\mathbf{z}_i \in \mathbb{R}^{A\times D}$, whose A slices define the axis-specific latent vectors $\mathbf{z}_{ia} \in \mathbb{R}^D$, for $a \in \{1,\dots,A\}$. Each programme axis a has a codebook consisting of K learnable state vectors $\mathbf{e}_{am}$, where $m \in \{1,\dots,K\}$. Before quantization, encoder outputs and codebook vectors are normalized to the same radius, removing magnitude effects so that nearest-neighbor state assignment depends only on angular similarity:

$$c_{ia} = \arg\min_m \|\mathbf{z}_{ia} - \mathbf{e}_{am}\|_2^2 . \tag{1}$$

Each cell is thus represented by one discrete state per axis,

$$\mathbf{c}_i = (c_{i1},\dots,c_{iA}),$$

yielding a compositional profile across programme-specific axes. The resulting $A \times K$ axis–state entries constitute the finite lexicon, while their combinations support up to K^A possible cell profiles. The nominal coding rate is

$$R = A\log_2 K \quad \text{bits per cell.} \tag{2}$$

This quantity reflects the capacity of the discrete representation rather than its empirical entropy.

3.2 Scaffold-guided gene weighting

To encourage different programme axes to capture distinct gene-expression patterns, LEXA assigns each gene a learnable distribution of responsibility across axes. This distribution is initialized from a scaffold S derived exclusively from the training data. The scaffold summarizes two complementary sources of variation: gene-expression differences across major cell types and donor-level variation after accounting for dataset-by-cell-type effects. Estimates from repeated training-cell samples are aligned and aggregated to obtain a stable prior over which axes should emphasize each gene. Importantly, the scaffold does not fix gene-to-axis assignments; it only provides the initialization and regularization target for trainable gene-responsibility parameters. Full details are provided in Appendix A.2.

Specifically, trainable logits L_{ag} are initialized from the scaffold prior and converted into gene-responsibility weights through a softmax across axes:

$$M_{ag} = \frac{\exp(L_{ag}/\tau_M)}{\sum_{b=1}^{A}\exp(L_{bg}/\tau_M)}, \qquad \sum_a M_{ag} = 1. \tag{3}$$

For each gene g, M_{ag} specifies the relative responsibility of axis a for that gene. Because the weights are normalized across axes but remain soft, a gene may contribute to multiple programmes. The scaffold therefore encourages programme specialization while allowing gene responsibilities to adapt during model training.

3.3 Additive decoding

LEXA reconstructs gene expression from the selected latent state on each programme axis using a shared nonlinear decoder with axis-specific FiLM modulation (Perez et al., 2018). For an input latent vector $q \in \mathbb{R}^D$, the decoder is defined as

$$h(q) = \mathrm{GELU}(W_1 q + b_1), \tag{4}$$

$$d_a(q) = W_2\,\mathrm{GELU}\left(\gamma_a \odot h(q) + \beta_a\right) + b_2, \tag{5}$$

where W_1 and W_2 are shared across all programme axes, while γ_a and β_a are learned axis-specific modulation parameters. Thus, the decoder shares its main parameters across axes while allowing each axis to map its selected latent state to a distinct gene-expression pattern.

For cell i, the selected codebook vector on axis a is $\mathbf{e}_{a,c_{ia}}$. Its decoder output is weighted gene-wise by the learned responsibility M_{ag}, and the weighted outputs are summed across axes to reconstruct the transcriptome:

$$\widehat{x}_{ig} = \sum_{a=1}^{A} M_{ag}\, d_a(\mathbf{e}_{a,c_{ia}})_g\,. \tag{6}$$

Thus, LEXA reconstructs each cell by additively combining one axis-conditioned expression pattern from each programme axis, with M_{ag} controlling the relative responsibility of each axis for gene g. Further implementation details are provided in Appendix A.3.

3.4 Learning Objective

LEXA jointly optimizes the encoder, codebooks, gene-responsibility weights, and decoder using a reconstruction objective together with several regularization terms:

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{VQ}}\mathcal{L}_{\mathrm{VQ}} + \lambda_{\mathrm{use}}\mathcal{L}_{\mathrm{use}} + \lambda_S\mathcal{L}_S + \lambda_{\mathrm{bal}}\mathcal{L}_{\mathrm{bal}} + \lambda_{\mathrm{div}}\mathcal{L}_{\mathrm{div}}. \tag{7}$$

The vector-quantization term combines codebook updating and encoder commitment:

$$\mathcal{L}_{\mathrm{VQ}} = \mathcal{L}_{\mathrm{code}} + \beta_{\mathrm{com}}\mathcal{L}_{\mathrm{com}}, \qquad \beta_{\mathrm{com}} = 1. \tag{8}$$

Here, $\mathcal{L}_{\mathrm{rec}}$ measures expression reconstruction error, $\mathcal{L}_{\mathrm{VQ}}$ optimizes vector quantization and encoder commitment, $\mathcal{L}_{\mathrm{use}}$ encourages broad codebook usage, $\mathcal{L}_S$ retains the scaffold-guided gene-responsibility structure, $\mathcal{L}_{\mathrm{bal}}$ prevents responsibility from concentrating on a small number of axes, and $\mathcal{L}_{\mathrm{div}}$ discourages redundant states within the same codebook.

Discrete state selection is trained with a straight-through estimator, so the forward pass uses the selected codebook vector while reconstruction gradients propagate to the encoder. Full definitions of all loss terms and optimization settings are provided in Appendix A.4.

3.5 Interpreting Gene Programmes in Axes and States

LEXA provides two complementary levels of molecular interpretation (see Appendix B.1 for details). Each axis is characterized by the genes for which that axis has high responsibility, whereas each state within an axis is characterized by its signed gene-expression effect relative to the average state on the same axis. These two levels distinguish the overall biological domain represented by an axis from the alternative molecular configurations represented by its individual states.

For each axis a, we use the learned gene-responsibility weights M_{ag} to identify the genes most strongly associated with that axis. Genes with larger M_{ag} contribute more strongly to defining the molecular scope of the axis. We therefore rank genes by M_{ag} and use the highest-responsibility genes to summarize the biological programme represented by the axis.

Within a given axis, different discrete states represent alternative expression configurations of the same broader programme. For state m on axis a, we define a gene-resolved state effect

$$\phi_{amg} = u_{amg} - \frac{1}{K}\sum_{j=1}^{K} u_{ajg},$$

where u_{amg} denotes the decoder-derived contribution of state m to gene g. Positive and negative values of ϕ_{amg} indicate genes expressed above or below the within-axis state average, respectively. We therefore interpret each state using genes with the strongest positive and negative effects, which describe the molecular configuration that distinguishes that state from alternative states on the same axis.

Functional annotations are obtained by gene set enrichment of these genes, as described in Appendix B.2. This hierarchical interpretation allows an axis to describe a broad gene programme while its states resolve distinct molecular configurations within that programme.

3.6 Training setup

In this study, we used 32 programme axes with 16 states per axis and 64-dimensional codebook vectors, yielding 512 axis–state entries and a nominal representation capacity of 128 bits per cell. Additional architectural and optimization details are provided in Appendix A.5.

LEXA was trained on a curated CELLxGENE atlas comprising 13.8 million cells from 210 datasets with a shared vocabulary of 10,310 protein-coding genes (CZI Cell Science Program et al., 2025). Datasets were selected based on data quality, donor and cell-type diversity, and balanced tissue coverage to reduce overrepresentation of highly abundant tissue categories. Full construction and preprocessing details are provided in Appendix A.6.

4 LEXA learns compact and informative representations

LEXA deliberately constrains cellular representations to a finite set of programme states, raising the question of how much information is lost relative to less structured continuous representations. We therefore first assessed representational capacity by examining the trade-off between representation size and downstream prediction performance. Rather than asking whether a discrete representation can match unconstrained continuous embeddings in absolute predictive accuracy, this analysis quantifies the cost of imposing LEXA's discrete and programme-factorized structure.

Comparison with alternative representations. We compared LEXA with several classes of representations, including conventional dimensionality reduction and integration methods PCA and Harmony (Korsunsky et al., 2019), probabilistic latent-variable models scVI (Lopez et al., 2018) and scArches (Lotfollahi et al., 2022); continuous gene-programme models cNMF (Kotliar et al., 2019) and Spectra (Kunes et al., 2024); and discrete representations Product-VQ (Jégou et al., 2011; van den Oord et al., 2017) and CellVQ (Wang et al., 2026). We evaluated the learned representations on four donor-disjoint datasets. The OneK1K and SLE datasets (Yazar et al., 2022; Perez et al., 2022) contained held-out donors from datasets represented during atlas training, whereas COVID-A and COVID-B were entirely unseen cohorts (Stephenson et al., 2021; Yoshida et al., 2022). We assessed major cell-type information using k-nearest-neighbor (kNN) and MLP classification (see Appendix C for details).

Across the four datasets and both kNN and MLP evaluations, LEXA generally outperformed programme-based methods and VQ-based discrete representations, while performing slightly below continuous representations such as PCA, Harmony, scVI, and scArches (Figure 2). This suggests that LEXA preserves substantial cell-identity information despite the additional constraints imposed by its compact discrete representation.

Dissecting the design principles of LEXA. To separate the effects of programme factorization and discretization, we compared LEXA with architecture-matched continuous controls. In the two independent COVID cohorts, the modular continuous autoencoder retained or even improved downstream performance relative to the unconstrained continuous autoencoder, indicating

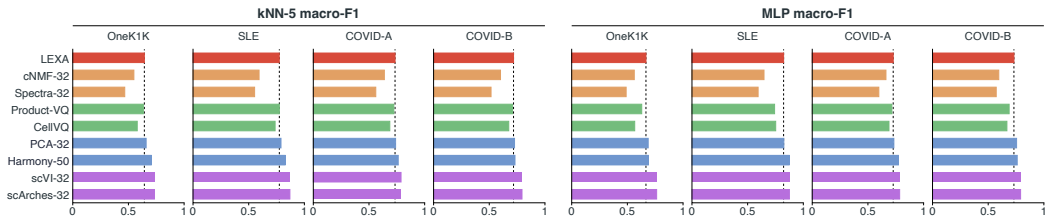

Figure 2: Macro-F1 scores of kNN and MLP classifiers on four datasets. Colors denote different method categories. Higher values indicate better major cell-type classification performance.

that programme-wise factorization itself does not substantially limit representational capacity (Figure 3A). The additional reduction observed with LEXA therefore primarily reflects the finite-state bottleneck introduced by discretization.

We next examined how assigning genes to programme axes affects representation quality. Replacing the scaffold-guided responsibilities with freely learned weights produced performance close to that of full LEXA across both COVID cohorts, whereas assigning all genes uniformly led to a small but more consistent reduction in macro-F1 (Figure 3B). Thus, the scaffold can impose an axis-specific biological organization without materially reducing predictive information.

We next examined whether radius normalization improves the assignment of cells to discrete programme states. Compared with Euclidean quantization, the default radius-normalized LEXA representation achieved higher kNN macro-F1 in both COVID cohorts and substantially improved MLP performance in COVID-B (Figure 3C). Performance remained stable across intermediate radius values, whereas the largest tested radius reduced local predictive accuracy. The selected spherical configuration performed more consistently than standard Euclidean assignment across datasets. We therefore use a radius of four as the default configuration.

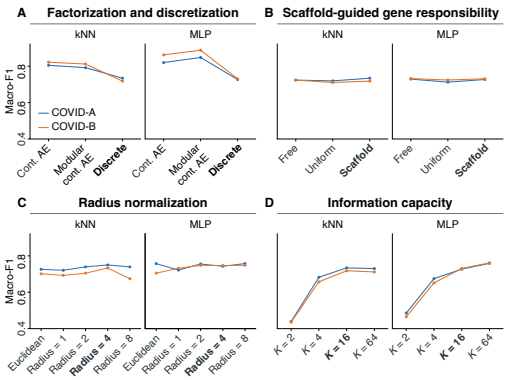

Figure 3: LEXA design ablations across independent COVID-19 cohorts. Macro-F1 scores from kNN and MLP evaluation are shown for COVID-A and COVID-B while varying representation factorization and discretization (A), scaffold-guided gene responsibility (B), radius normalization (C), and information capacity (D). Bold labels indicate the default LEXA configuration.

Finally, we varied the number of states per programme axis to examine the trade-off between coding capacity and information retention. Performance improved substantially from two to four states and continued to increase up to 16 states, whereas expanding the codebook to 64 states provided only limited additional benefit (Figure 3D). We therefore selected 16 states per axis as the default configuration, balancing compactness with downstream performance.

Together, these results show that LEXA retains substantial cell-identity information despite its compact discrete representation. The results are consistent with the finite-state bottleneck accounting for a substantial part of the performance gap relative to continuous embeddings, supporting a favorable trade-off between representational capacity and biological structure.

5 Biological coherence and reproducibility of gene-programme states learned by LEXA

LEXA recovers biological programmes with reduced fragmentation. We next asked whether the gene programmes learned by LEXA correspond to coherent biological processes. We compared

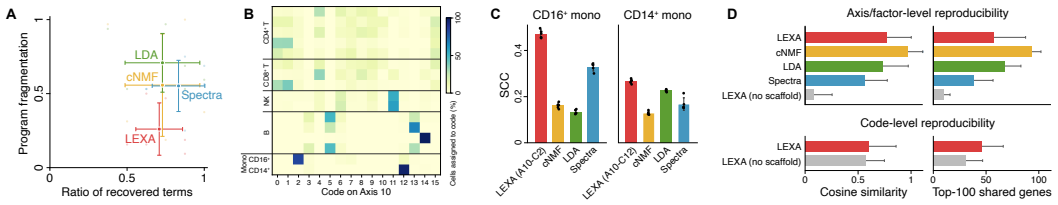

Figure 4: LEXA organizes biological programmes into compact axes. (A) Coverage and fragmentation of six predefined biological programme classes across different methods. (B) Distribution of the 16 codes within LEXA Axis 10 across immune-cell subtypes. Only subtypes containing more than 1,000 cells are shown. (C) Concordance between subtype-associated learned gene programmes and monocyte subtype expression contrasts. (D) Cross-seed reproducibility of learned gene programmes. For (A), (C), and (D), mean $\pm$ SD is shown.

LEXA with cNMF, Topic Modeling based on Latent Dirichlet Allocation (LDA) (Blei et al., 2003), and Spectra, all fitted independently to the OneK1K dataset. To provide a common reference, we curated a GO-BP panel covering both immune-cell-specific processes and broadly shared cellular functions, with eight representative terms per programme class (Appendix D.1). For cNMF, LDA, and Spectra, programme enrichment was evaluated using the top 500 genes of each factor or topic. For LEXA, enrichment was first evaluated for the top 500 genes of each state, and the strongest state-level match within each axis was used as the axis-level statistic, followed by statistical correction (Appendix D.2).

We assessed programme organization using two complementary measures: term coverage, which quantifies how many reference GO terms are recovered, and fragmentation, which measures how broadly each programme is distributed across parent axes or factors. LEXA achieved programme coverage comparable to cNMF and LDA and moderately below Spectra, while showing substantially lower fragmentation (Figure 4A). Thus, LEXA recovered similar biological content using fewer parent axes, indicating greater programme-level specificity and less functional redundancy.

Discrete codes resolve cell-type-specific states within a shared LEXA axis. Programme-level concentration alone does not establish whether the axis–code hierarchy provides additional biological resolution. We therefore examined the 16 codes within LEXA Axis 10, an immune-associated axis linked to leukocyte-mediated cytotoxicity and TNF/NF-κB inflammatory processes in the reference-panel analysis. The distribution of Axis 10 codes differed markedly across immune-cell subtypes (Figure 4B). For instance, C2 accounted for 81.7% of CD14-low CD16$^+$ monocytes, whereas C12 accounted for 95.9% of CD14$^+$ monocytes. Thus, two codes within the same parent axis were strongly associated with the two major monocyte subtypes.

We next asked whether the gene effects learned for these codes reflected the transcriptional differences between the two monocyte states. We compared the C2 and C12 gene-effect profiles with the corresponding CD16$^+$–CD14$^+$ expression contrasts, and performed analogous comparisons for the most subtype-associated factors from cNMF, LDA, and Spectra, using Spearman correlation coefficient (SCC) to quantify concordance (Appendix D.3). LEXA showed the strongest concordance with held-out monocyte expression differences for both subtype-associated programmes when each method was evaluated over its own top-500 genes (Figure 4C). C2 captured the CD16$^+$-associated transcriptional contrast most accurately, while C12 similarly showed the highest concordance for the reciprocal CD14$^+$-associated contrast.

These results show that the subtype specificity of Axis 10 codes is reflected not only in their cellular distributions but also in their learned gene effects. Importantly, LEXA represents these distinct subtype-associated programmes as different discrete states within a shared parent axis, whereas the comparison methods capture them through separate latent factors. Axis 10 therefore illustrates the hierarchical organization of LEXA, in which a shared biological dimension is resolved into finer-grained, functionally distinct cellular states.

LEXA learns reproducible programme axes and states. We next assessed the reproducibility of LEXA across three random seeds by aligning learned components with Hungarian matching

(Appendix D.4). At the axis level, matched LEXA axes had a mean cosine similarity of 0.765 and shared approximately 58 of their top 100 genes across runs (Figure 4D). LEXA and LDA showed comparable axis-level reproducibility, both exceeding Spectra, whereas cNMF showed the highest reproducibility, consistent with its use of consensus across multiple NMF fits. Matched LEXA codes also showed reproducible gene effects, with a mean cosine similarity of 0.598 and an average top-100 gene overlap of 46.

Scaffold-guided gene weighting was particularly important for stabilizing the parent axes. Without the scaffold, both axis-level cosine similarity and top-100 gene overlap decreased. At the code level, removing the scaffold had a smaller effect on cosine similarity but reduced top-100 gene overlap from 46 to 31. These results indicate that the scaffold stabilizes the programme decomposition across independent runs, particularly the gene composition of the learned axes and codes.

6 Case Study: revealing COVID-19 severity-associated gene-programme states by LEXA

We next used an external peripheral-blood COVID-19 cohort (Stephenson et al., 2021), excluded from the atlas training data, to examine whether LEXA could resolve immune programme states associated with clinical severity (Appendix E.1). The original study primarily characterized severity-associated changes in conventional immune-cell subtypes. We asked whether LEXA could reveal additional disease-associated variation within these annotated populations by resolving cells according to their programme states.

Atlas transfer and cohort-specific refinement. We first applied the atlas-trained LEXA model directly to the cohort and then tested whether donor-level state occupancy changed systematically with disease severity within each major immune-cell class via regression analysis (Appendix E.2). The transferred lexicon revealed widespread severity-associated shifts in programme-state occupancy across multiple immune lineages, in contrast to conventional cell-subtype proportions, most of which were not significantly associated with disease severity (Figure 5A). For example, Atlas Axis 30, which captured a broad immune-activation and cytoskeletal programme (Appendix E.3), contained distinct severity-associated states in unconventional T cells and the platelet/megakaryocyte compartment (Figure 5A), illustrating how a shared atlas-level programme can manifest through different states across immune lineages.

Because the atlas-level lexicon captures broad programmes shared across heterogeneous datasets, its axes remained relatively coarse-grained in this disease-specific cohort. We therefore fitted LEXA *de novo* within the COVID-19 cohort to resolve these broad programmes into more context-specific components. For example, Atlas Axis 30 was separated into two cohort-specific axes with distinct molecular emphases: one retained a broad immune-response and cytoskeletal programme, whereas the other emphasized lymphocyte activation and proximal signalling (Appendix E.2). These cohort-refined states showed stronger and more lineage-resolved severity associations than the corresponding atlas states (Figure 5A).

Severity-associated variation beyond conventional CD4 T-cell subtypes. We next focused on $CD4^+$ T cells, where conventional subtype abundance provides a well-characterized severity signal. Consistent with the original study, IL-22^+ $CD4^+$ T cells decreased with increasing disease severity. However, this was the only CD4 subtype showing an association in our donor-level analysis, and even this signal was modest, reaching only the 10% FDR threshold (Figure 5B). In contrast, the cohort-fitted LEXA representation revealed multiple severity-associated programme states within $CD4^+$ T cells, several with substantially stronger statistical support than the subtype-level association. Their associations with disease severity remained significant after adjusting for subtype composition (Figure 5B), suggesting that the LEXA states capture severity-associated molecular variation beyond conventional subtype composition, encompassing both within-subtype heterogeneity and programmes shared across annotated subtypes.

Biological interpretation of severity-associated programme states. The severity-associated states mapped to several biologically coherent programme groups (details in Appendix E.3). Cohort Axis 20 separated an endoplasmic reticulum (ER)/protein-processing state that increased with sever-

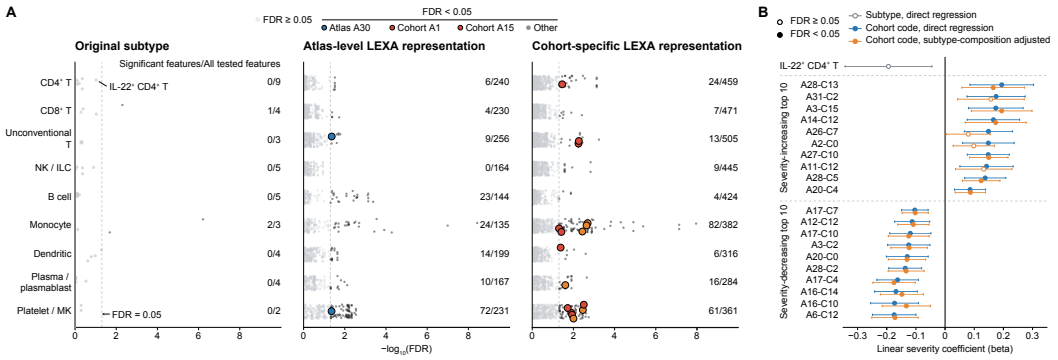

Figure 5: LEXA identifies COVID-19 severity-associated programme states beyond conventional cell-subtype composition. (A) Severity-association significance for conventional subtype proportions, atlas-level states, and cohort-level states across major immune-cell classes. Only codes with a prevalence of at least 1% and presence in at least 10 donors were included. (B) Linear severity coefficients and 95% confidence intervals for IL-22⁺ CD4⁺ T cells and cohort codes with the ten strongest positive and negative associations.

ity from a chromosome-segregation/cell-cycle state that decreased. This opposing redistribution is consistent with SARS-CoV-2-induced ER-stress responses and with the impaired proliferative capacity reported for circulating T cells in critically ill patients (Wang et al., 2023; Bidar et al., 2022). Cohort Axis 16 captured mitochondrial respiratory-chain and biosynthetic programmes, with multiple states decreasing with severity, in line with the mitochondrial dysfunction, impaired oxidative phosphorylation, and apoptosis-associated metabolic states reported in T cells from patients with severe COVID-19 (Thompson et al., 2021). Cohort Axis 17 represented ubiquitin–proteasome and autophagy-related programmes, with several states declining in a manner consistent with SARS-CoV-2-associated remodeling of ubiquitination, proteostasis, and autophagic flux (Stukalov et al., 2021; Gassen et al., 2021).

Together, these results show that LEXA provides a hierarchical view of disease-associated cellular variation: atlas-level states capture transferable programme structure, whereas cohort-specific states resolve context-dependent molecular configurations that are obscured by conventional subtype labels. This enables clinically associated changes to be traced from donor-level state redistribution to explicit gene-level programmes.

7 Discussion

A central challenge in single-cell representation learning is to reconcile expressive cellular embeddings with biological interpretability. LEXA addresses this by representing each cell as a composition of discrete programme states with explicit gene effects. Rather than using discretization only for compression, LEXA turns it into a finite vocabulary of recurring cellular programmes, suggesting a view of cell state not as a single point in latent space, but as a composition of reusable biological states.

Our results show that this structured bottleneck retains substantial cell-identity information while organizing biological variation more coherently. LEXA concentrates related processes into fewer programme axes, resolves distinct states within shared biological dimensions, and learns structures that remain reproducible across model re-initialization. Such a programme-state lexicon provides a natural interface between representation learning and biological reasoning, allowing cellular differences and population changes to be described in terms of interpretable state combinations.

LEXA nevertheless has important limitations. Its additive decoder does not explicitly model interactions between axes, the scaffold incorporates major cell-type annotations, and the fixed numbers of axes and states constrain lexicon granularity. State effects are model-derived rather than causal, and cross-fit correspondence is not universal. Future work could extend this framework toward hierarchical, interaction-aware, and perturbation-informed representations.

REPRODUCIBILITY STATEMENT

Implementation, training, hyperparameter, and preprocessing details are provided in Appendix A. Axis- and state-level gene-effect extraction and functional annotation procedures are described in Appendix B. Representation evaluation procedures are documented in Appendix C. Biological-coherence analyses and repeated-fit reproducibility are described in Appendix D. The COVID-19 case-study protocol is detailed in Appendix E. An anonymized implementation of LEXA is available at `https://anonymous.4open.science/r/LEXA-5913`.

AI USE STATEMENT

Generative AI tools were used to assist with code development and language refinement. All analyses were executed on the underlying data and numerical outputs, and AI-generated suggestions were reviewed against the implemented code. The authors determined the research ideas, model and experimental designs, interpretation of results, and take responsibility for the complete content of the manuscript.

REFERENCES

Frank Bidar, Sarah Hamada, Morgane Gossez, Remy Coudereau, Jonathan Lopez, Marie-Angelique Cazalis, Claire Tardiveau, Karen Brengel-Pesce, Marine Mommert, Marielle Buisson, Filippo Conti, Thomas Rimmelé, Anne-Claire Lukaszewicz, Laurent Argaud, Martin Cour, Guillaume Monneret, Fabienne Venet, and RICO Study Group. Recombinant human interleukin-7 reverses t cell exhaustion ex vivo in critically ill COVID-19 patients. *Annals of Intensive Care*, 12(1):21, 2022. doi: 10.1186/s13613-022-00982-1.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. URL `https://www.jmlr.org/papers/v3/blei03a.html`.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M. Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, J. Michael Cherry, Tiffany Chi, Jennifer Chien, Leah Dorman, Pablo Garcia-Nieto, Nayib Gloria, Mim Hastie, Daniel Hegeman, Jason Hilton, Timmy Huang, Amanda Infeld, Ana-Maria Istrate, Ivana Jelic, Kuni Katsuya, Yang Joon Kim, Karen Liang, Mike Lin, Maximilian Lombardo, Bailey Marshall, Bruce Martin, Fran McDade, Colin Megill, Nikhil Patel, Alexander Predeus, Brian Raymor, Behnam Robatmili, Dave Rogers, Erica Rutherford, Dana Sadgat, Andrew Shin, Corinn Small, Trent Smith, Prathap Sridharan, Alexander Tarashansky, Norbert Tavares, Harley Thomas, Andrew Tolopko, Meghan Urisko, Joyce Yan, Garabet Yeretssian, Jennifer Zamanian, Arathi Mani, Jonah Cool, and Ambrose Carr. Cz cellxgene discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, 53(D1):D886–D900, 2025. doi: 10.1093/nar/gkae1142.

Nils C. Gassen, Jan Papies, Thomas Bajaj, Jackson Emanuel, Frederik Dethloff, Robert Lorenz Chua, Jakob Trimpert, Nicolas Heinemann, Christine Niemeyer, Friderike Weege, Katja Hönzke, Tom Aschman, Daniel E. Heinz, Katja Weckmann, Tim Ebert, Andreas Zellner, Martina Lennarz, Emanuel Wyler, Simon Schroeder, Anja Richter, Daniela Niemeyer, Karen Hoffmann, Thomas F. Meyer, Frank L. Heppner, Victor M. Corman, Markus Landthaler, Andreas C. Hocke, Markus Morkel, Nikolaus Osterrieder, Christian Conrad, Roland Eils, Helena Radbruch, Patrick Giavalisco, Christian Drosten, and Marcel A. Müller. SARS-CoV-2-mediated dysregulation of metabolism and autophagy uncovers host-targeting antivirals. *Nature Communications*, 12(1): 3818, 2021. doi: 10.1038/s41467-021-24007-w.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21:1481–1491, 2024. doi: 10.1038/s41592-024-02305-7.

Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:117–128, 2011. doi: 10.1109/TPAMI.2010.57.

Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Priit Adler, Jaak Vilo, and Hedi Peterson. g:profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Research*, 51(W1):W207–W212, 07 2023. ISSN 0305-1048. doi: 10.1093/nar/gkad347.

Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po-Ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, 16:1289–1296, 2019. doi: 10.1038/s41592-019-0619-0.

Dylan Kotliar, Adrian Veres, M. Aurel Nagy, Shervin Tabrizi, Eran Hodis, Douglas A. Melton, and Pardis C. Sabeti. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-seq. *eLife*, 8:e43803, 2019. doi: 10.7554/eLife.43803.

Russell Z. Kunes, Thomas Walle, M. Land, Tal Nawy, and Dana Pe'er. Supervised discovery of interpretable gene programs from single-cell data. *Nature Biotechnology*, 42:1084–1095, 2024. doi: 10.1038/s41587-023-01940-3.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15:1053–1058, 2018. doi: 10.1038/s41592-018-0229-2.

Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D. Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso, Nir Yosef, Marta Interlandi, Sergei Rybakov, Alexander V. Misharin, and Fabian J. Theis. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, 40:121–130, 2022. doi: 10.1038/s41587-021-01001-7.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3942–3951, 2018. doi: 10.1609/aaai.v32i1.11671.

Richard K. Perez, M. Grace Gordon, Meena Subramaniam, Min Cheol Kim, George C. Hartoularos, Sasha Targ, Yang Sun, Anton Ogorodnikov, Raymund Bueno, Andrew Lu, Mike Thompson, Nadav Rappoport, Andrew Dahl, Cristina M. Lanata, Mehrdad Matloubian, Lenka Maliskova, Serena S. Kwek, Tony Li, Michal Slyper, Julia Waldman, Danielle Dionne, Orit Rozenblatt-Rosen, Lawrence Fong, Maria Dall'Era, Brunilda Balliu, Aviv Regev, Jinoos Yazdany, Lindsey A. Criswell, Noah Zaitlen, and Chun Jimmie Ye. Single-cell rna-seq reveals cell type–specific molecular and genetic associations to lupus. *Science*, 376(6589):eabf1970, 2022. doi: 10.1126/science.abf1970.

Emily Stephenson, Gary Reynolds, Rachel A. Botting, Fernando J. Calero-Nieto, Michael D. Morgan, Zewen Kelvin Tuong, Karsten Bach, Waradon Sungnak, Kaylee B. Worlock, Masahiro Yoshida, Natsuhiko Kumasaka, Katarzyna Kania, Justin Engelbert, Bayanne Olabi, Jarmila Stremenova Spegarova, Nicola K. Wilson, Nicole Mende, Laura Jardine, Louis C. S. Gardner, Issac Goh, Dave Horsfall, Jim McGrath, Simone Webb, Michael W. Mather, Rik G. H. Lindeboom, Emma Dann, Ni Huang, Krzysztof Polanski, Elena Prigmore, Florian Gothe, Jonathan Scott, Rebecca P. Payne, Kenneth F. Baker, Aidan T. Hanrath, Ina C. D. Schim van der Loeff, Andrew S. Barr, Amada Sanchez-Gonzalez, Laura Bergamaschi, Federica Mescia, Josephine L. Barnes, Eliz Kilich, Angus de Wilton, Anita Saigal, Aarash Saleh, Sam M. Janes, Claire M. Smith, Nusayhah Gopee, Caroline Wilson, Paul Coupland, Jonathan M. Coxhead, Vladimir Yu Kiselev, Stijn van Dongen, Jaume Bacardit, Hamish W. King, Cambridge Institute of Therapeutic Immunology and Infectious Disease-National Institute of Health Research (CITIID-NIHR) COVID-19 BioResource Collaboration, Anthony J. Rostron, A. John Simpson, Sophie Hambleton, Elisa Laurenti, Paul A. Lyons, Kerstin B. Meyer, Marko Z. Nikolić, Christopher J. A. Duncan, Kenneth G. C. Smith, Sarah A. Teichmann, Menna R. Clatworthy, John C. Marioni, Berthold Göttgens, and Muzlifah Haniffa. Single-cell multi-omics analysis of the immune response in COVID-19. *Nature Medicine*, 27:904–916, 2021. doi: 10.1038/s41591-021-01329-2.

Alexey Stukalov, Virginie Girault, Vincent Grass, Ozge Karayel, Valter Bergant, Christian Urban, Darya A. Haas, Yiqi Huang, Lila Oubraham, Anqi Wang, M. Sabri Hamad, Antonio Piras, Fynn M. Hansen, Maria C. Tanzer, Igor Paron, Luca Zinzula, Thomas Engleitner, Maria Reinecke, Teresa M. Lavacca, Rosina Ehmann, Roman Wölfel, Jörg Jores, Bernhard Kuster, Ulrike Protzer, Roland Rad, John Ziebuhr, Volker Thiel, Pietro Scaturro, Matthias Mann, and Andreas Pichlmair. Multilevel proteomics reveals host perturbations by SARS-CoV-2 and SARS-CoV. *Nature*, 594 (7862):246–252, 2021. doi: 10.1038/s41586-021-03493-4.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618:616–624, 2023. doi: 10.1038/s41586-023-06139-9.

Elizabeth A. Thompson, Katherine Cascino, Alvaro A. Ordonez, Weiqiang Zhou, Ajay Vaghasia, Anne Hamacher-Brady, Nathan R. Brady, Im-Hong Sun, Rulin Wang, Avi Z. Rosenberg, Michael Delannoy, Richard Rothman, Katherine Fenstermacher, Lauren Sauer, Kathyrn Shaw-Saliba, Evan M. Bloch, Andrew D. Redd, Aaron A. R. Tobian, Maureen Horton, Kellie Smith, Andrew Pekosz, Franco R. D'Alessio, Srinivasan Yegnasubramanian, Hongkai Ji, Andrea L. Cox, and Jonathan D. Powell. Metabolic programs define dysfunctional immune responses in severe COVID-19 patients. *Cell Reports*, 34(11):108863, 2021. doi: 10.1016/j.celrep.2021.108863.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1711.00937.

Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nature Biotechnology*, 34(11):1145–1160, 2016. doi: 10.1038/nbt.3711.

Jue Wang, Cheng Tan, Zhangyang Gao, Sida Shao, Shiping Liu, and Stan Z. Li. Illuminating cell states by a comprehensive and interpretable single cell foundation model. *Nature Communications*, 17:4037, 2026. doi: 10.1038/s41467-026-70071-5.

Xuefeng Wang, Wenqi Wang, Tiecheng Wang, Jianzhong Wang, Yuhang Jiang, Xue Wang, Ze Qiu, Na Feng, Weiyang Sun, Chang Li, Songtao Yang, Xianzhu Xia, Hongbin He, and Yuwei Gao. SARS-CoV-2 orf8 protein induces endoplasmic reticulum stress-like responses and facilitates virus replication by triggering calnexin: an unbiased study. *Journal of Virology*, 97(3):e00011–23, 2023. doi: 10.1128/jvi.00011-23.

Chuxi Xiao, Yuang Ding, Haiyang Bian, Yixin Chen, Lei Wei, and Xuegong Zhang. Tokenizing single-cell transcriptomes as a native language for large language models. *bioRxiv*, 2026. doi: 10.1101/2025.10.22.684047. URL https://www.biorxiv.org/content/10.1101/2025.10.22.684047v2. Version 2, revised 11 July 2026.

Seyhan Yazar, Jose Alquicira-Hernandez, Kristof Wing, Anne Senabouth, M. Grace Gordon, Stacey Andersen, Qinyi Lu, Antonia Rowson, Thomas R. P. Taylor, Linda Clarke, Katia Maccora, Christine Chen, Anthony L. Cook, Chun Jimmie Ye, Kirsten A. Fairfax, Alex W. Hewitt, and Joseph E. Powell. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. *Science*, 376:eabf3041, 2022. doi: 10.1126/science.abf3041.

Masahiro Yoshida, Kaylee B. Worlock, Ni Huang, Rik G. H. Lindeboom, Colin R. Butler, Natsuhiko Kumasaka, Cecilia Dominguez Conde, Lira Mamanova, Liam Bolt, Laura Richardson, Krzysztof Polanski, Elo Madissoon, Josephine L. Barnes, Jessica Allen-Hyttinen, Eliz Kilich, Brendan C. Jones, Angus de Wilton, Anna Wilbrey-Clark, Waradon Sungnak, J. Patrick Pett, Juliane Weller, Elena Prigmore, Henry Yung, Puja Mehta, Aarash Saleh, Anita Saigal, Vivian Chu, Jonathan M. Cohen, Clare Cane, Aikaterini Iordanidou, Soichi Shibuya, Ann-Kathrin Reuschl, Iván T. Herczeg, A. Christine Argento, Richard G. Wunderink, Sean B. Smith, Taylor A. Poor, Catherine A. Gao, Jane E. Dematte, NU SCRIPT Study Investigators, Gary Reynolds, Muzlifah Haniffa, Georgina S. Bowyer, Matthew Coates, Menna R. Clatworthy, Fernando J. Calero-Nieto, Berthold Göttgens, Christopher O'Callaghan, Neil J. Sebire, Clare Jolly, Paolo De Coppi, Claire M. Smith, Alexander V. Misharin, Sam M. Janes, Sarah A. Teichmann, Marko Z. Nikolić, and Kerstin B. Meyer. Local and systemic responses to SARS-CoV-2 infection in children and adults. *Nature*, 602(7896):321–327, 2022. doi: 10.1038/s41586-021-04345-x.

A ARCHITECTURE AND IMPLEMENTATION OF LEXA

A.1 INPUT AND ENCODER

Gene-expression profiles are mapped to a common vocabulary of 10,310 protein-coding genes shared across the atlas datasets (see Appendix A.6 for details). Raw count-like expression is normalized within each cell to a total count of 10,000 and transformed as $\log(1 + x)$. Genes not measured in a source dataset are excluded from the reconstruction objective using an observation mask rather than treated as biological zeros. Cells failing dataset-specific quality-control criteria are removed before model training.

The encoder is a multilayer perceptron with dimensions $10{,}310 \to 2{,}048 \to 2{,}048$. The first linear layer is followed by GELU activation and dropout with rate 0.1, and the final output is normalized by LayerNorm and reshaped into 32 axis-specific vectors of dimension 64. Each axis vector is projected onto a sphere of radius $r = 4$ before vector quantization. The codebook contains 16 freely trainable 64-dimensional vectors for each axis, which are likewise projected onto the same sphere. State assignment is performed by nearest-code lookup within each axis. Thus, each cell is represented by one selected state on each of 32 axes. The common gene vocabulary and input transformation define a shared measurement space; cohort-specific fitting uses the same input space but does not inherit learned encoder or codebook parameters from the atlas model.

A.2 SCAFFOLD CONSTRUCTION

LEXA uses a data-derived scaffold to provide an initial soft assignment of genes to programme axes. The scaffold can be viewed as an axis-by-gene responsibility matrix, where larger values indicate that a gene is more strongly associated with a given axis. It is used to initialize and regularize the trainable gene-responsibility weights rather than to impose a fixed gene-to-axis assignment.

The scaffold is constructed exclusively from training cells and combines two sources of biological variation. First, we compute expression contrasts between major cell types, with datasets weighted equally, to capture broad lineage- and identity-associated programmes. Second, we quantify donor-level residual variation after subtracting dataset-by-cell-type means, allowing the scaffold to capture variation that remains within annotated cell populations, such as activation or other context-dependent programmes. For each atlas scaffold repeat, 300,000 training cells are sampled and eligible donor-by-cell-type groups are summarized.

To reduce sensitivity to sampling and initialization, scaffold construction is repeated three times using different random seeds. The resulting axis-responsibility matrices are aligned across repeats and averaged to obtain a consensus scaffold. Informative genes retain memberships in up to three axes, whereas genes without sufficient scaffold signal are assigned uniformly across axes. Under the majority-of-repeats rule used for the atlas model, 1,548 of 10,310 genes remain uniformly assigned in the final consensus scaffold. Among genes with non-uniform assignments, the median effective number of contributing axes is 2.730.

Let S_{ag} denote the consensus scaffold weight for gene g on axis a, normalized across axes for each gene. We define a smoothed scaffold prior as

$$\bar{S}_{ag} = \frac{S_{ag} + \epsilon}{\sum_b (S_{bg} + \epsilon)}, \qquad \epsilon = 0.001.$$

The trainable gene-responsibility logits are initialized from this prior as

$$L_{ag} = \tau_M \log \bar{S}_{ag},$$

where τ_M is the mask temperature. No additional logit noise is used in the selected full-atlas model. Applying the softmax in Equation 3 yields the trainable gene-responsibility weights M_{ag}. Thus, the scaffold provides a biologically informed starting point, but gene responsibilities remain learnable during model training and are not fixed by the scaffold.

The scaffold is therefore data-derived: major cell-type labels contribute to the expression-contrast component used to initialize the gene-axis structure, whereas the neural training objective itself contains no supervised cell-type classification loss. For cohort-specific models, the scaffold is recomputed from the corresponding target-cohort training cells and aligned across independent scaffold-construction repeats.

A.3 DECODER IMPLEMENTATION

For a latent vector $q \in \mathbb{R}^{64}$, the axis-conditioned decoder is

$$h(q) = \mathrm{GELU}(W_1 q + b_1), \tag{9}$$

$$d_a(q) = W_2 \,\mathrm{Dropout}_{0.1}\left[\mathrm{GELU}\left(\gamma_a \odot h(q) + \beta_a\right)\right] + b_2. \tag{10}$$

Here, W_1 and W_2 are shared across axes, whereas γ_a and β_a are axis-specific FiLM parameters. Both the encoder hidden layer and the shared decoder hidden layer contain 2,048 units. The decoder dropout probability is 0.1 and is applied after the FiLM-modulated GELU activation. Decoder outputs are unconstrained in sign as they are additive contributions in normalized log-expression space.

For cell i, the selected codebook state on axis a is $\mathbf{e}_{a,c_{ia}}$. Reconstruction is

$$\widehat{x}_{ig} = \sum_a M_{ag}\, d_a(\mathbf{e}_{a,c_{ia}})_g.$$

A.4 TRAINING OBJECTIVE AND REGULARIZATION

LEXA is trained by jointly optimizing expression reconstruction, discrete vector quantization, codebook usage, scaffold consistency, axis balance, and within-codebook diversity.

Reconstruction loss. Let O_{ig} indicate whether gene g is observed for cell i. Reconstruction is optimized only over observed entries:

$$\mathcal{L}_{\mathrm{rec}} = \frac{\sum_{i,g} O_{ig}\,(x_{ig} - \widehat{x}_{ig})^2}{\sum_{i,g} O_{ig}}. \tag{11}$$

This prevents genes absent from a source dataset from being treated as biological zeros.

Vector quantization and commitment. For cell i and axis a, let $\boldsymbol{z}_{ia}$ denote the encoder output and $\boldsymbol{e}_{a,c_{ia}}$ the selected codebook vector. The selected state is

$$c_{ia} = \arg\min_m \left\|\boldsymbol{z}_{ia} - \boldsymbol{e}_{am}\right\|^2. \tag{12}$$

The quantized representation uses the selected codebook vector in the forward pass, with gradients propagated to the encoder through a straight-through estimator. Because both the normalized encoder output and the codebook vectors lie on a sphere of radius r, the codebook and commitment losses are normalized by r^2:

$$\mathcal{L}_{\mathrm{code}} = \frac{1}{NA}\sum_{i,a} \frac{\left\|\mathrm{sg}[\boldsymbol{z}_{ia}] - \boldsymbol{e}_{a,c_{ia}}\right\|^2}{r^2}, \tag{13}$$

$$\mathcal{L}_{\mathrm{com}} = \frac{1}{NA}\sum_{i,a} \frac{\left\|\boldsymbol{z}_{ia} - \mathrm{sg}[\boldsymbol{e}_{a,c_{ia}}]\right\|^2}{r^2}, \tag{14}$$

where $\mathrm{sg}[\cdot]$ denotes the stop-gradient operator. The combined vector-quantization term is

$$\mathcal{L}_{\mathrm{VQ}} = \mathcal{L}_{\mathrm{code}} + \beta_{\mathrm{com}}\mathcal{L}_{\mathrm{com}}, \qquad \beta_{\mathrm{com}} = 1. \tag{15}$$

Codebook usage. To discourage collapse to a small subset of states, we compute a differentiable soft assignment

$$q_{iam} = \frac{\exp\left(-\|\boldsymbol{z}_{ia} - \boldsymbol{e}_{am}\|^2/\tau_q\right)}{\sum_j \exp\left(-\|\boldsymbol{z}_{ia} - \boldsymbol{e}_{aj}\|^2/\tau_q\right)}, \qquad \bar{q}_{am} = \frac{1}{N}\sum_i q_{iam}, \qquad \tau_q = 0.1. \tag{16}$$

The usage penalty is

$$\mathcal{L}_{\mathrm{use}} = 1 - \frac{1}{A\log K}\sum_a H(\bar{q}_a), \qquad H(p) = -\sum_m p_m \log p_m. \tag{17}$$

Minimizing this term encourages each axis to use multiple states rather than collapsing to a single code. Hard-assignment entropy is monitored as a diagnostic but is not used as a training objective.

Scaffold consistency and axis balance. Let M_{ag} denote the trainable responsibility of axis a for gene g, and $\bar{S}_{ag}$ the smoothed scaffold prior. Scaffold consistency is encouraged by

$$\mathcal{L}_S = \frac{1}{G} \sum_g \sum_a \left(M_{ag} - \bar{S}_{ag} \right)^2 . \tag{18}$$

This term encourages the learned gene responsibilities to remain close to the data-derived scaffold while allowing them to adapt during training.

To prevent the model from concentrating most genes on a small number of axes, we additionally use

$$\mathcal{L}_{\text{bal}} = \frac{1}{A} \sum_a \left(\frac{1}{G} \sum_g M_{ag} - \frac{1}{A} \right)^2 . \tag{19}$$

Within-codebook diversity. To encourage distinct states within the same axis, we penalize similarity between different codebook vectors:

$$\mathcal{L}_{\text{div}} = \frac{1}{AK(K-1)} \sum_a \sum_{m \neq n} \left(\frac{\boldsymbol{e}_{am}^{\mathsf{T}} \boldsymbol{e}_{an}}{r^2} \right)^2 . \tag{20}$$

Because all codebook vectors lie on a sphere of radius r, this term corresponds to the squared cosine similarity between distinct states and encourages angular separation within each axis.

Overall objective. The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda_{\text{VQ}} \mathcal{L}_{\text{VQ}} + \lambda_{\text{use}} \mathcal{L}_{\text{use}} + \lambda_S \mathcal{L}_S + \lambda_{\text{bal}} \mathcal{L}_{\text{bal}} + \lambda_{\text{div}} \mathcal{L}_{\text{div}}. \tag{21}$$

The coefficients used in the selected model are listed in Table 1.

A.5 Optimization and hyperparameters

Models were optimized using AdamW with a learning rate of 2×10^{-4} and weight decay of 10^{-4}. Gradients were clipped to a maximum norm of 1. The learning rate remained fixed throughout training. The configured coefficients of selected auxiliary losses were linearly scheduled over the first three epochs. Full-atlas training used a batch size of 2,048, whereas cohort-specific models used a batch size of 1,024. The selected full-atlas model was trained for at most 20 epochs or 10 million cell updates, with checkpoint selection based on validation reconstruction MSE. Validation evaluation was capped at 200,000 cells, and training used an early-stopping patience of three evaluations.

Additional optimization and training hyperparameters are provided in Table 1.

A.6 Atlas construction and preprocessing

We constructed the LEXA training atlas from 1,113 datasets downloaded from CELLxGENE (CZI Cell Science Program et al., 2025). Eligible scRNA-seq datasets were required to contain human adult samples, at least 1,000 cells and 15,000 measured features, primary-data cells, cell-type annotations, and donor identifiers. We further required a raw-count-like expression matrix with non-negative, approximately integer-valued entries. Eligible datasets were ranked according to cell number, donor diversity, and cell-type diversity, and were subsequently selected using tissue-wise round-robin sampling to reduce overrepresentation of highly abundant tissue categories. The resulting atlas comprised 210 datasets.

Protein-coding genes were identified, and we retained the strict intersection of protein-coding genes measured across all 210 datasets. This yielded a common vocabulary of 10,310 genes and prevented genes absent from a particular dataset from being interpreted as observed biological zeros. We applied minimal cell-level quality control, retaining cells with at least 20 detected genes and nonzero total counts and excluding cells explicitly annotated as non-primary observations. Cells were then sampled within each dataset according to predefined dataset-specific caps. The final atlas contained 13,826,509 cells. Expression counts were normalized to 10,000 counts per cell and transformed using $\log(1 + x)$.

Table 1: Optimization settings for the selected full-atlas LEXA model.

Parameter	Value
VQ coefficient, λ_{VQ}	0.003125
Commitment multiplier within VQ, β_{com}	1
Usage coefficient	0.1
Scaffold-prior coefficient	0.1
Axis-balance coefficient	1
Within-codebook diversity coefficient	0.005
AdamW learning rate	2×10^{-4}
Weight decay	10^{-4}
Gradient clipping	1
Auxiliary-loss coefficient schedule	Linear ramp over the first 3 epochs
Full-atlas batch size	2,048
Cohort-fitted batch size	1,024
Maximum epochs	20
Full-atlas cell updates	10,000,000
Validation cap	200,000
Early-stopping patience	3
Checkpoint criterion	Validation reconstruction MSE

To prevent information leakage across data partitions, training, validation, and test splits were defined at the donor or sample-group level. Entire groups were assigned to training, validation, or test partitions using a deterministic seeded hash, ensuring that no donor or sample group occurred in more than one partition. Datasets reserved for the downstream in-depth evaluations described below were excluded in their entirety from model training and validation. The resulting atlas contained 10,041,846 training cells, 1,884,246 validation cells, and 1,900,417 test cells.

B Axis and state interpretation

B.1 Axis- and state-level gene programmes.

Axis-level gene programmes. The learned responsibility matrix $M \in \mathbb{R}^{A \times G}$ assigns each gene a soft distribution of responsibility across axes. For axis a, genes are ranked by M_{ag}, and the highest-ranking genes are used to characterize the overall molecular scope of that axis.

State-level gene effects. For state m on axis a, define its responsibility-weighted decoder output for gene g as

$$u_{amg} = M_{ag}\, d_a(\mathbf{e}_{am})_g.$$

To isolate the molecular configuration specific to state m, we center this quantity relative to the uniform mean across K states on the same axis:

$$\phi_{amg} = u_{amg} - \frac{1}{K} \sum_{j=1}^{K} u_{ajg}.$$

By construction,

$$\sum_{m} \phi_{amg} = 0.$$

Positive values of ϕ_{amg} indicate genes relatively enhanced in state m, whereas negative values indicate genes relatively depleted relative to the within-axis state average.

State-effect identities. Because reconstruction is additive across axes, replacing state m with state n on axis a changes reconstructed expression by

$$\phi_{ang} - \phi_{amg},$$

independently of the states selected on the remaining axes.

Defining

$$b_g = \sum_a \frac{1}{K} \sum_m u_{amg},$$

the reconstruction can equivalently be written as

$$\widehat{x}_{ig} = b_g + \sum_a \phi_{a,c_{ia},g}.$$

These identities are algebraic consequences of the additive decoder and do not imply causal gene regulation.

B.2 Functional annotation of learned axes and codes

Axis- and code-level functions were annotated using different gene selection strategies. Axis annotations were derived directly from the learned gene masks. For each axis, genes were ranked by their mask weights, and the top 300 genes were submitted to Gene Ontology (GO) enrichment analysis. Thus, the axis annotation describes the overall molecular scope selected by the corresponding mask and does not depend on the severity regression or on any individual code.

Code annotations were derived from the code-specific gene effects within each axis. We first restricted the analysis to the 500 genes with the highest mask weights for the corresponding axis. These genes were then ranked according to their signed, centred effect for the code of interest. Up to 200 genes with the strongest positive effects and up to 200 genes with the strongest negative effects were analysed separately. The positive gene set represents genes relatively enhanced in that programme state, whereas the negative gene set represents genes relatively depleted in the same state. This procedure distinguishes the general biological domain encoded by an axis from the specific molecular configuration represented by an individual code.

GO enrichment was performed with g:Profiler (Kolberg et al., 2023) using a custom background comprising all 10,310 genes considered by the model. Biological process (BP), cellular component (CC), and molecular function (MF) ontologies were tested separately. Terms with a false-discovery-rate adjusted $p < 0.05$ were considered significant. Terms containing fewer than 5 or more than 1,500 annotated genes were excluded, together with highly generic terms such as "biological process", "cellular process", "metabolic process", "regulation of biological process", "response to stimulus", "molecular function", "binding", "protein binding", "cellular component", "cellular anatomical entity", and "intracellular anatomical structure". Within each ontology, retained terms were ranked by adjusted enrichment p-value, with smaller gene sets prioritized in the event of a tie and duplicated term names removed.

For concise presentation, the functional descriptions below report representative leading enrichment terms rather than the complete set of significant GO results. For each axis or signed code direction, we retained up to three BP, two CC, and two MF terms. When no term passed the enrichment and filtering criteria, no GO-based functional label was assigned. The results of all atlas-level axes are provided in Table 2.

Table 2: Functional annotations of the 32 atlas-level LEXA axes. NA denotes no significant term.

Axis	Functional summary	Representative GO terms
A1	An immune-signalling programme centred on small-GTPase regulation and vesicle-associated signalling.	**BP:** intracellular signaling cassette; regulation of small GTPase mediated signal transduction; immune system process. **CC:** vesicle membrane; cytoplasmic vesicle membrane. **MF:** GTPase regulator activity; nucleoside-triphosphatase regulator activity.
A2	NA.	**BP:** NA. **CC:** NA. **MF:** NA.
A3	An interferon-associated antiviral programme linking recognition of foreign RNA to immune defence.	**BP:** immune response; response to other organism; response to external biotic stimulus. **CC:** external side of plasma membrane; fibrillar center. **MF:** double-stranded RNA binding; ubiquitin-protein transferase activity.

Continued on next page

Table 2 continued

Axis	Functional summary	Representative GO terms
A4	An intracellular transport and macromolecule-catabolism programme involving endosomal and pericentriolar compartments.	**BP:** intracellular transport; protein transport; macromolecule catabolic process. **CC:** pericentriolar material; early endosome. **MF:** NA.
A5	NA.	**BP:** NA. **CC:** NA. **MF:** NA.
A6	A mitochondrial oxidative-phosphorylation programme centred on respiratory-chain electron transport and ATP synthesis.	**BP:** mitochondrial ATP synthesis coupled electron transport; ATP synthesis coupled electron transport; oxidative phosphorylation. **CC:** respiratory chain complex; mitochondrial inner membrane. **MF:** oxidoreduction-driven active transmembrane transporter activity; electron transfer activity.
A7	A developmental transcriptional programme involving tissue morphogenesis and sequence-specific regulatory activity.	**BP:** central nervous system development; anatomical structure morphogenesis; brain development. **CC:** NA. **MF:** beta-catenin binding; RNA polymerase II cis-regulatory region sequence-specific DNA binding.
A8	A chromosome-organization and DNA-replication programme coupled to nucleic-acid metabolism and nucleolar complexes.	**BP:** chromosome organization; DNA metabolic process; DNA replication. **CC:** nuclear protein-containing complex; nucleolus. **MF:** RNA binding; catalytic activity, acting on a nucleic acid.
A9	A proteostasis and membrane-lipid trafficking programme involving stress granules and ER–mitochondrial contact sites.	**BP:** aromatic amino acid transport; regulation of proteasomal ubiquitin-dependent protein catabolic process; lipid localization. **CC:** mitochondria-associated endoplasmic reticulum membrane contact site; cytoplasmic stress granule. **MF:** nuclear localization sequence binding; aromatic amino acid transmembrane transporter activity.
A10	An extracellular-matrix organization and remodelling programme involving matrix structural components and heparin binding.	**BP:** external encapsulating structure organization; extracellular matrix organization; extracellular structure organization. **CC:** extracellular matrix; external encapsulating structure. **MF:** extracellular matrix structural constituent; heparin binding.
A11	An oxidative small-molecule metabolic programme localized predominantly to mitochondria and peroxisomes.	**BP:** small molecule metabolic process; oxoacid metabolic process; organic acid metabolic process. **CC:** mitochondrion; peroxisomal matrix. **MF:** oxidoreductase activity; vitamin binding.
A12	A ribonucleoprotein-complex and ribosome-biogenesis programme centred on RNA processing and the nucleolus.	**BP:** ribonucleoprotein complex biogenesis; RNA processing; ribosome biogenesis. **CC:** ribonucleoprotein complex; nucleolus. **MF:** RNA binding; ribonucleoprotein complex binding.
A13	A myelination and axon-ensheathment programme linked to GTPase-regulatory activity.	**BP:** myelination; ensheathment of neurons; axon ensheathment. **CC:** myelin sheath; axon. **MF:** GTPase regulator activity; nucleoside-triphosphatase regulator activity.
A14	A mitochondrial energy and organelle-acidification programme combining respiratory-chain activity with lysosomal functions.	**BP:** lysosomal lumen acidification; mitochondrial ATP synthesis coupled electron transport; ATP synthesis coupled electron transport. **CC:** mitochondrial envelope; mitochondrial membrane. **MF:** NA.
A15	An RNA-processing programme involving mRNA metabolism and nuclear ribonucleoprotein granules.	**BP:** RNA processing; mRNA processing; mRNA metabolic process. **CC:** ribonucleoprotein granule; nuclear ribonucleoprotein granule. **MF:** RNA binding; mRNA binding.

Continued on next page

Table 2 continued

Axis	Functional summary	Representative GO terms
A16	An innate immune-response programme linking pattern-recognition receptors to secretory-granule and vesicle functions.	**BP:** immune response; immune system process; defense response. **CC:** secretory granule; secretory vesicle. **MF:** pattern recognition receptor activity; lipopolysaccharide immune receptor activity.
A17	NA.	**BP:** NA. **CC:** NA. **MF:** NA.
A18	A contractile and muscle-structural programme involving actin binding, myofibrils, and contractile fibres.	**BP:** muscle system process; muscle contraction; muscle cell development. **CC:** contractile muscle fiber; myofibril. **MF:** structural constituent of muscle; actin binding.
A19	A ciliary-motility and cell-adhesion programme involving cell–cell junctions and adhesion mediators.	**BP:** cilium movement; cell-cell adhesion; cell adhesion. **CC:** cornified envelope; cell-cell junction. **MF:** cell-cell adhesion mediator activity; cell adhesion mediator activity.
A20	A neuronal synaptic-signalling and transport programme involving cytoskeletal, kinesin, and neuron-projection functions.	**BP:** chemical synaptic transmission; anterograde trans-synaptic signaling; trans-synaptic signaling. **CC:** synapse; neuron projection. **MF:** cytoskeletal protein binding; kinesin binding.
A21	A photoreceptor and sensory-cilium programme associated with visual perception and light detection.	**BP:** visual perception; sensory perception of light stimulus; detection of light stimulus. **CC:** non-motile cilium; photoreceptor cell cilium. **MF:** NA.
A22	A vascular-development and endothelial-signalling programme involving receptor interactions and cell-surface junctions.	**BP:** circulatory system development; vasculature development; blood vessel development. **CC:** anchoring junction; cell surface. **MF:** signaling receptor binding; transmembrane signaling receptor activity.
A23	An epithelial adhesion and metabolic programme associated with extracellular vesicles and oxidoreductase activity.	**BP:** small molecule metabolic process; aldehyde metabolic process; tissue development. **CC:** extracellular exosome; extracellular vesicle. **MF:** cell-cell adhesion mediator activity; oxidoreductase activity.
A24	A kinase-regulatory programme involving histone H3S57 and DNA-dependent protein kinase activities.	**BP:** NA. **CC:** NA. **MF:** histone H3S57 kinase activity; DNA-dependent protein kinase activity.
A25	A DNA-replication-initiation programme involving origin-recognition complexes and zinc-ion transport.	**BP:** NA. **CC:** nuclear origin of replication recognition complex; origin recognition complex. **MF:** DNA replication origin binding; zinc ion transmembrane transporter activity.
A26	An ER-stress and protein-processing programme linking misfolded-protein handling to ER–Golgi transport.	**BP:** response to endoplasmic reticulum stress; endoplasmic reticulum to Golgi vesicle-mediated transport; response to topologically incorrect protein. **CC:** endoplasmic reticulum; organelle subcompartment. **MF:** misfolded protein binding; oligosaccharyl transferase activity.
A27	An immune and alpha-beta T-cell activation programme coupled to transcriptional regulation and chromosomal complexes.	**BP:** immune system process; alpha-beta T cell activation. **CC:** chromosome; transcription regulator complex. **MF:** enzyme binding; transcription regulator inhibitor activity.

Continued on next page

Table 2 continued

Axis	Functional summary	Representative GO terms
A28	A biogenic-amine biosynthetic programme with platelet- and erythroid-associated mask genes.	**BP:** biogenic amine biosynthetic process; amine biosynthetic process. **CC:** NA. **MF:** NA.
A29	A cell-projection organization programme involving microtubule-based movement and projection-associated compartments.	**BP:** plasma membrane bounded cell projection organization; cell projection organization; microtubule-based movement. **CC:** neuron projection; cell projection. **MF:** NA.
A30	A broad immune-activation programme linking cell-surface signalling with actin-cytoskeletal organization.	**BP:** immune response; immune system process; cell activation. **CC:** side of membrane; cell surface. **MF:** actin binding; signaling receptor activity.
A31	A transcriptional and chromatin-regulatory programme involving histone modification and ubiquitin-ligase complexes.	**BP:** positive regulation of RNA metabolic process; positive regulation of DNA-templated transcription; positive regulation of RNA biosynthetic process. **CC:** transferase complex; ubiquitin ligase complex. **MF:** chromatin binding; histone modifying activity.
A32	A transcription-coupled chromatin-remodelling and nuclear RNA-export programme.	**BP:** transcription elongation-coupled chromatin remodeling; mRNA export from nucleus; RNA export from nucleus. **CC:** RNA polymerase III transcription regulator complex; phagophore assembly site membrane. **MF:** NA.

C Details of representation ability analysis

C.1 Datasets

We evaluated cell-level representation quality on four scRNA-seq datasets spanning healthy population variation, autoimmune disease, and acute viral infection. OneK1K is a large population-scale peripheral blood mononuclear cell (PBMC) cohort profiling inter-individual variation across approximately one thousand donors (Yazar et al., 2022). SLE is a PBMC cohort from patients with systemic lupus erythematosus (SLE) and matched controls, providing a disease setting with substantial immune-state heterogeneity (Perez et al., 2022). COVID-A and COVID-B are two independent PBMC cohorts collected from individuals with COVID-19 across a range of disease severities and were used as external datasets to assess transfer to unseen cohorts (Stephenson et al., 2021; Yoshida et al., 2022).

OneK1K and SLE were represented among the datasets used to construct the atlas, but evaluation was restricted to donors excluded from atlas model training, thereby testing generalization to unseen individuals within previously observed cohorts. In contrast, COVID-A and COVID-B were entirely absent from atlas training and were used to assess transfer to independently collected cohorts. For all four datasets, discovery and held-out test partitions were defined at the donor or sample level, with no donor shared between partitions; for OneK1K and SLE, discovery cells were drawn from the atlas training partition and held-out cells from the atlas test partition, whereas COVID-A and COVID-B used independently constructed donor-disjoint splits. All datasets were represented using the common atlas vocabulary of 10,310 protein-coding genes, with expression counts normalized to 10,000 counts per cell and log1p-transformed when required by the corresponding representation method.

For evaluation, cell annotations were mapped to a common major-cell-class taxonomy comprising $CD4^+$ T cells, $CD8^+$ T cells, monocytes, dendritic cells (DCs), NK/innate lymphoid cells (ILCs), B cells, unconventional T cells, plasma cells/plasmablasts, and platelets/megakaryocytes (MKs). Ambiguous broad lymphocyte and progenitor annotations were not assigned to more specific classes. The same cells, partitions, and labels were used to evaluate all representations.

C.2 METHODS FOR COMPARISON

We compared LEXA with representative continuous, discrete, integration-based, and gene-programme representations. Whenever applicable, models were trained on a common benchmark core sampled from the donor-disjoint atlas partitions, comprising 500,000 training cells and 100,000 validation cells.

LEXA. LEXA represents each cell by one categorical state on each of 32 gene-programme axes, with 16 possible states per axis. For downstream representation evaluation, each categorical state was converted to a 16-dimensional one-hot vector and the 32 vectors were concatenated, yielding a 512-dimensional binary vector. Although represented as a 512-dimensional one-hot vector for computational convenience, the underlying categorical representation contains $32 \log_2 16 = 128$ bits per cell.

Product-VQ. Product-VQ was used as a capacity-matched discrete baseline. It used 32 independent latent positions with 16 states at each position and therefore had the same categorical capacity of 128 bits per cell as LEXA.

CellVQ. CellVQ representations were obtained using the official pretrained checkpoint and inference implementation. We used the model's single categorical cell token, which takes one of 4,096 possible values. The resulting token was represented as a 4,096-dimensional one-hot vector for downstream evaluation.

PCA. Principal component analysis (PCA) was fitted on the common 500,000-cell training core using 2,000 highly variable genes selected from the training data. Discovery and held-out cells were projected using the frozen training mean and loading matrix, and the first 32 principal components were retained as the cell representation.

Harmony. Harmony was applied to a 50-dimensional PCA representation. The common training core together with the unlabeled discovery and held-out cells from each target dataset were jointly supplied to Harmony, using dataset identity and assay as integration covariates. Harmony was run for a maximum of ten iterations. Because Harmony had access to the expression distribution of the held-out cells, but not their cell-type labels, it represents a transductive comparison in this benchmark.

scVI. A reference scVI model was trained on the 500,000-cell benchmark training core using 2,000 highly variable genes and a 32-dimensional latent space. For atlas-included datasets, existing dataset identifiers were used as reference batch categories. Discovery and held-out cells were encoded using the frozen reference model without further parameter updates.

scArches. For scArches, the trained scVI reference model was used to initialize query mapping. The model was adapted to each target dataset using all unlabeled discovery and held-out cells for ten epochs with a batch size of 1,024. Thus, scArches was evaluated as a query-adapted method and, unlike frozen LEXA, PCA, Product-VQ, and scVI representations, had access to the unlabeled expression distribution of the target cohort during adaptation.

cNMF. Consensus non-negative matrix factorization (cNMF) was fitted on the common training core to obtain 32 gene-expression programmes. These programmes were subsequently held fixed, and programme usages for discovery and held-out cells were estimated by non-negative matrix factorization. The resulting 32-dimensional usage vectors were normalized to sum to one and used as cell representations.

Spectra. Spectra was fitted with 32 factors using 100,000 cells sampled from the common training core over the same set of 2,000 highly variable genes. A subset was used because fitting Spectra on the complete 500,000-cell core was computationally prohibitive. Learned factors were held fixed during evaluation, and non-negative factor usages were inferred for discovery and held-out cells. Usage vectors were normalized to sum to one.

Continuous autoencoder. To quantify the cost of replacing an unconstrained continuous representation with LEXA's structured discrete representation, we trained an architecture-matched continuous autoencoder. The model encoded each cell into a 128-dimensional continuous latent vector. Representations were extracted from the frozen trained model and directly used for downstream classification.

Modular continuous autoencoder. We additionally constructed a modular continuous autoencoder to separate the effect of programme-wise factorization from that of discretization. The model retained the 32-axis organization used by LEXA but represented each axis by a 64-dimensional continuous vector rather than a categorical state. The 32 axis-specific vectors were concatenated, yielding a 2,048-dimensional continuous representation per cell. This comparison tests whether loss of predictive information arises primarily from decomposition into programme-specific components or from the subsequent finite-state bottleneck.

Gene-responsibility ablations. To assess the contribution of scaffold-guided gene responsibility, we compared the default LEXA model with two control variants. In the free-responsibility control, the scaffold prior was removed, while the gene-responsibility logits remained trainable and were optimized jointly with the rest of the model. This control therefore preserves the same parameterization and flexibility of M_{ag} but removes the data-derived initialization and scaffold-consistency constraint. In the uniform-responsibility control, gene responsibilities were fixed uniformly across all axes $M_{ag} = 1/A$, for every axis a and gene g, and were not updated during training.

C.3 Evaluation metrics

We evaluated the extent to which each representation retained major cell-type information using two complementary classification procedures: k-nearest-neighbour (kNN) classification and a nonlinear multilayer-perceptron (MLP) probe. Both classifiers were trained or constructed exclusively using the discovery partition and evaluated once on the donor-disjoint held-out partition.

kNN classification. kNN classification was performed directly in each method's native representation space using cosine distance. An approximate nearest-neighbour index was constructed from the discovery cells using NNDescent, retaining 30 candidate neighbours. For each held-out cell, the five nearest discovery cells were selected.

Cell-type prediction was performed by inverse-distance-weighted voting over these five neighbours. For a neighbour at cosine distance d, the voting weight was defined as

$$w = \frac{1}{\max(d, 10^{-6})}.$$

The predicted cell type was the class with the largest sum of neighbour weights. This evaluation requires no additional learned mapping from representation to cell type and therefore primarily assesses whether cells with the same biological identity are locally proximal in the representation space.

MLP classification. We additionally used a nonlinear probe to test whether cell-type information remained recoverable even when it was not organized purely through local neighbourhood structure. Before classifier training, each representation dimension was standardized using a StandardScaler fitted exclusively on discovery cells; the same transformation was then applied to held-out cells.

The classifier was a MLP with two hidden layers containing 256 and 64 units, respectively. Each hidden layer used a ReLU activation and a dropout probability of 0.1. The network was trained using class-weighted cross-entropy loss and the AdamW optimizer with a learning rate of 10^{-3}, weight decay of 10^{-4}, and batch size of 512, for a maximum of 20 epochs.

Approximately 10% of discovery cells from each class were reserved as an internal validation set. Model selection was based on validation macro-F1, with early stopping after three consecutive epochs without an improvement greater than 10^{-4}. The selected model was subsequently evaluated once on the fixed held-out donor partition.

D Details of Biological coherence and reproducibility

D.1 GO-BP reference programme panel

We constructed a reference panel of Gene Ontology Biological Process (GO-BP) terms to evaluate whether the latent representations learned by different methods corresponded to recognizable bi-

ological programmes. The panel contained 48 terms organized into 6 programme classes, with 8 terms per class (Table 3). The classes were selected to cover both immune-cell-associated functions and biological processes shared more broadly across cell types.

Table 3: GO-BP reference panel used to compare biological programme coverage and fragmentation.

Programme class	Functional summary	GO-BP terms
Leukocyte-mediated cytotoxicity	Cytotoxic immune-effector functions shared across T cells, natural killer cells, and other cytotoxic leukocyte populations.	leukocyte-mediated cytotoxicity; T-cell-mediated cytotoxicity; regulation of T-cell-mediated cytotoxicity; positive regulation of natural-killer-cell-mediated cytotoxicity; CD8-positive alpha-beta T-cell activation; natural-killer-cell-mediated immunity; regulation of leukocyte-mediated cytotoxicity; positive regulation of immune effector process.
Antigen presentation	Processing and presentation of endogenous and exogenous antigens through MHC class I and class II pathways.	antigen processing and presentation of exogenous peptide antigen via MHC class II; antigen processing and presentation of exogenous peptide antigen; antigen processing and presentation of peptide antigen; antigen processing and presentation; antigen processing and presentation of peptide or polysaccharide antigen via MHC class II; antigen processing and presentation of exogenous antigen; antigen processing and presentation of peptide antigen via MHC class I; regulation of antigen processing and presentation.
Interferon response	Cellular responses to type I and type II interferons and regulation of interferon-mediated signalling.	cellular response to interferon-gamma; positive regulation of type I interferon-mediated signalling pathway; regulation of type I interferon-mediated signalling pathway; cellular response to interferon-beta; response to interferon-beta; response to interferon-alpha; response to type I interferon; response to interferon-gamma.
TNF/NF-κB inflammatory response	Inflammatory signalling involving tumour necrosis factor, NF-κB, interleukin signalling, and inflammatory cytokine production.	response to tumour necrosis factor; regulation of tumour-necrosis-factor-mediated signalling pathway; NIK/NF-κB signalling; positive regulation of NF-κB transcription-factor activity; interleukin-1-mediated signalling pathway; interleukin-6-mediated signalling pathway; tumour-necrosis-factor-mediated signalling pathway; cytokine production involved in inflammatory response.
Metabolism and mitochondrial function	Core energy metabolism involving mitochondrial respiration, fatty-acid oxidation, glycolysis, and the tricarboxylic-acid cycle.	regulation of fatty-acid beta-oxidation; oxidative phosphorylation; tricarboxylic-acid cycle; respiratory electron-transport chain; regulation of glycolytic process; ATP synthesis coupled electron transport; mitochondrial gene expression; fatty-acid beta-oxidation.
RNA and protein homeostasis	RNA processing and turnover, translation, protein folding, and endoplasmic-reticulum-associated protein quality control.	ubiquitin-dependent ERAD pathway; nuclear-transcribed mRNA catabolic process, nonsense-mediated decay; cleavage involved in rRNA processing; endoplasmic-reticulum unfolded-protein response; protein folding; regulation of mRNA processing; cytoplasmic translation; translational elongation.

D.2 PERMUTATION-BASED COMPARISON OF PROGRAMME COVERAGE AND FRAGMENTATION

We compared LEXA with cNMF, LDA, and Spectra using an identical set of 32 parent axes, factors, or topics per method. For each cNMF factor, LDA topic, and Spectra factor, the 500 genes with the largest gene weights were used to define its representative gene set. For LEXA, the top 500 genes were identified separately for each of the 16 codes within each axis.

For a GO-BP term g, the enrichment statistic for a factor or topic from a comparison method was the number of genes shared between the GO-BP gene set and the factor's top-500 genes. Because a LEXA axis contains 16 discrete codes, its axis-level statistic was defined as the maximum overlap across its codes:

$$T_{ag}^{\mathrm{LEXA}} = \max_{c\in\{1,\dots,16\}} |G_g \cap S_{ac}|\,, \tag{22}$$

where G_g denotes the measured genes in GO-BP term g, and S_{ac} denotes the top-500 genes of code c in axis a.

To account for the opportunity to search across multiple codes within each LEXA axis, the same maximization was applied to the null statistic. For each GO-BP term, 20,000 random gene sets containing the same number of measured genes were sampled without replacement from the OneK1K gene universe. For each random gene set, its overlap with the top-500 genes of all 16 codes was calculated, and the maximum overlap was retained. Thus, the empirical null distribution explicitly incorporated the 16-code search performed for LEXA. For cNMF, LDA, and Spectra, the corresponding null statistic was the overlap with the top-500 genes of the individual factor or topic.

Empirical one-sided p-values were calculated as

$$p_{ag} = \frac{1 + \sum_{b=1}^{B} \mathbb{I}\left(T_{ag}^{(b)} \geq T_{ag}^{\mathrm{obs}}\right)}{B+1}, \qquad B = 20{,}000. \tag{23}$$

Within each method, Benjamini–Hochberg correction was applied jointly to all $32 \times 48 = 1{,}536$ axis–term tests used in the primary benchmark. An axis–term association was considered significant at $q < 0.05$. A GO-BP term was considered recovered by a method if it was significantly associated with at least one of the method's 32 components.

For method m and programme class p, term coverage was defined as

$$\mathrm{Coverage}_{mp} = \frac{1}{8} \sum_{g\in p} \mathbb{I}\left(\min_{a} q_{mag} < 0.05\right). \tag{24}$$

Programme fragmentation was defined as the fraction of parent components associated with at least one term from that programme:

$$\mathrm{Fragmentation}_{mp} = \frac{1}{32} \sum_{a=1}^{32} \mathbb{I}\left(\min_{g\in p} q_{mag} < 0.05\right). \tag{25}$$

Coverage therefore quantifies how much of a predefined biological programme was recovered, whereas fragmentation quantifies how widely that programme was distributed across the latent representation. A representation with high coverage and low fragmentation recovers multiple subprocesses while concentrating them within a small number of parent components.

D.3 MONOCYTE-SUBTYPE GENE-PROGRAMME CONCORDANCE

We restricted the analysis to CD14$^+$ monocytes and CD14-low CD16$^+$ monocytes from the OneK1K dataset. Within each held-out fold, the observed monocyte-subtype expression contrast for gene g was calculated as

$$\Delta_g^{(k)} = \overline{X}_{g,\mathrm{CD16}}^{(k)} - \overline{X}_{g,\mathrm{CD14}}^{(k)},$$

where the means were calculated across all cells of the corresponding subtype from donors in held-out fold k.

For cNMF, LDA, and Spectra, cell-level factor activities were calculated from expression using the corresponding model-provided gene weights. Within each training fold, we calculated the standardized difference in factor activity between CD16$^+$ and CD14$^+$ monocytes. The factor with the largest

positive standardized difference was selected as the CD16-associated profile, and the factor with the largest negative standardized difference was selected as the CD14-associated profile. Factor selection was stable across all folds.

Gene-programme concordance was evaluated over the genes with the largest parent-axis or factor loadings. For both LEXA codes, we used the 500 genes with the highest loadings in the Axis 10 module mask. Thus, C2 and C12 were evaluated over the same Axis 10-defined gene set, while retaining their distinct signed code-specific gene-effect vectors. No additional normalization was applied to the supplied LEXA gene effects. For cNMF, LDA, and Spectra, each selected factor was evaluated over its own 500 highest-loading genes. Because the gene sets were defined separately by each learned programme, this analysis measures the internal transcriptional coherence of each method's high-loading genes rather than concordance over an identical externally specified gene set.

We used five-fold donor-blocked cross-validation, ensuring that all cells from a donor were assigned to the same fold. For a CD16-associated profile with learned gene-effect or loading vector w_g, concordance in held-out fold k was defined as

$$\rho_{\mathrm{CD16}}^{(k)} = \mathrm{SCC}\left(w_g, \Delta_g^{(k)}\right), \qquad g \in G_{500},$$

where G_{500} denotes the 500 highest-loading genes of the corresponding parent axis or factor.

For a CD14-associated profile, the expression contrast was reversed:

$$\rho_{\mathrm{CD14}}^{(k)} = \mathrm{SCC}\left(w_g, -\Delta_g^{(k)}\right), \qquad g \in G_{500}.$$

A positive correlation therefore indicates that genes receiving larger effects or loadings in the selected profile tend to be more highly expressed in the corresponding monocyte subtype.

D.4 Reproducibility of programme axes and codes

We evaluated reproducibility across three independently initialized training runs, with only the random seed varied. Each run produced 32 axes or factors, yielding three pairwise run comparisons.

Axis- and factor-level reproducibility. For LEXA, each axis was represented by its gene-responsibility profile; for cNMF, LDA, and Spectra, each factor was represented by its native gene-loading profile. To emphasize component-specific genes, each profile was residualized by subtracting the mean profile across the 32 components within the same run. For each pair of runs, we constructed a 32×32 cosine-similarity matrix and applied Hungarian matching to obtain a one-to-one alignment of axes or factors. Reproducibility was quantified by the cosine similarity between matched profiles and the overlap between their 100 genes with the largest absolute residualized values.

Code-level reproducibility. For each LEXA code, the gene-effect profile was defined as its decoder contribution multiplied by the corresponding axis-specific gene responsibilities and centered across the 16 codes within that axis. After aligning the parent axes, we constructed a 16×16 cosine-similarity matrix within each matched axis pair and applied a second Hungarian matching to align codes one-to-one. Code-level reproducibility was evaluated using gene-effect cosine similarity and overlap among the 100 genes with the largest absolute effects.

E Details of the COVID-19 case study

E.1 COVID-19 cohort and data processing

We used the COVID-A dataset generated by Stephenson et al. (2021), which profiled PBMCs from individuals spanning healthy, asymptomatic, mild, moderate, severe, and critical COVID-19. The original study jointly measured single-cell transcriptomes, cell-surface proteins, and lymphocyte antigen-receptor repertoires. The present analysis used only the single-cell RNA expression profiles and the accompanying donor- and sample-level metadata.

After harmonization, the processed dataset included COVID-19 and several control groups, including healthy, Lipopolysaccharide-challenged, and non-COVID controls. All data were retained for

training LEXA, whereas regression analyses of clinical progression were restricted to donors with COVID-19.

Several donors were sampled at more than one time point. To ensure that each donor contributed only one statistically independent observation, we retained the earliest available sample for each donor. The five COVID-19 severity groups were ordered as asymptomatic, mild, moderate, severe, critical. For regression analyses, these groups were encoded as an ordinal variable

$$s_d \in \{0, 1, 2, 3, 4\}.$$

The original cell annotations were grouped into nine major immune-cell classes for the donor-level analyses as described in Appendix C.1.

Each retained cell was encoded using two LEXA representations. The atlas-level representation was obtained by directly applying the CELLxGENE-trained model to the COVID-19 cells without updating its parameters. The cohort-level representation was obtained by fitting a separate LEXA model *de novo* to the COVID-A cohort. Both representations assigned each cell one of 16 states on each of 32 axes. Atlas and cohort axis indices are specific to their respective models and were therefore related using their gene-responsibility masks rather than by assuming index-wise correspondence.

E.2 Donor-level severity regression

All associations were evaluated at the donor level rather than the cell level. Within each major cell class, we counted the number of cells assigned to each LEXA state separately for every donor. Let Y_{dmac} denote the number of cells from donor d, major cell class m, axis a, and code c, and let N_{dm} denote the total number of cells from that donor in the corresponding major cell class. Code occupancy was therefore defined relative to the number of available cells in the same major class rather than relative to the donor's total peripheral-blood cell count.

COVID-19 severity was encoded as an ordinal variable s_d. For each eligible axis–code state, we fitted a negative-binomial generalized linear model,

$$Y_{dmac} \sim \mathrm{NB}(\mu_{dmac}, \alpha),$$

with

$$\log \mu_{dmac} = \log N_{dm} + \beta_{0,mac} + \beta_{s,mac} s_d + \boldsymbol{\gamma}_{mac}^{\mathsf{T}} \mathbf{x}_d.$$

Here, $\log N_{dm}$ is an offset and $\mathbf{x}_d$ contains donor age, sex, collection site, and days from symptom onset. Age and days from onset were mean-centred before model fitting. Sex and collection site were represented by indicator variables. Missing continuous covariates were replaced by the median among the included donors, and missing categorical covariates were assigned an explicit unknown level. We used a fixed negative-binomial dispersion parameter $\alpha = 1$ and heteroskedasticity-consistent HC0 covariance estimates.

The coefficient $\beta_{s,mac}$ quantifies the change in the logarithm of relative code occupancy associated with a one-level increase in disease severity. Its exponentiated value,

$$\exp(\beta_{s,mac}),$$

is the estimated multiplicative change in the expected occupancy of the code per severity level, conditional on the total number of cells in the corresponding major class and the included covariates. Consequently, these models test redistribution among programme states within a major cell class rather than changes in the absolute abundance of that class.

Axis–code states were tested only when they accounted for at least 1% of the cells in the corresponding major cell class and occurred in at least ten donors. Two-sided Wald tests were used to test

$$H_0 : \beta_{s,mac} = 0.$$

The resulting p-values were adjusted using the Benjamini–Hochberg procedure separately for each major cell class and LEXA representation. Unless stated otherwise, an axis–code state was considered severity-associated at FDR < 0.05.

For comparison with the LEXA states, the original cell-subtype compositions were analysed using the same donor-level framework. For subtype j nested within major cell class m, let Y_{dmj} denote the number of cells assigned to that subtype. We fitted

$$\log E[Y_{dmj}] = \log N_{dm} + \theta_{0,mj} + \theta_{s,mj} s_d + \boldsymbol{\eta}_{mj}^{\mathsf{T}} \mathbf{x}_d.$$

Thus, the subtype and LEXA-state analyses used the same donor-level sampling unit, severity variable, offset, and clinical covariates. Subtype-level p-values were adjusted within each major cell class using the Benjamini–Hochberg procedure.

To determine whether severity-associated LEXA states merely reflected changes in the relative abundance of conventional CD4$^+$ T subtypes, we performed an additional analysis restricted to the cohort-fitted CD4$^+$ LEXA representation. For each donor, we calculated the proportions of the ten CD4$^+$ subtypes. A pseudocount of 0.5 cells was added to every donor–subtype count before converting the counts to proportions.

Because subtype proportions are compositional and sum to one, entering all raw proportions into the regression would produce exact collinearity. We therefore transformed the ten-part composition into nine orthogonal isometric log-ratio coordinates. Let

$$\mathbf{p}_d = (p_{d1}, \ldots, p_{d,10})$$

denote the pseudocount-adjusted subtype composition of donor d. The isometric log-ratio representation was calculated as

$$\mathbf{z}_d = \mathbf{H} \log \mathbf{p}_d,$$

where $\mathbf{H}$ is a 9×10 orthonormal Helmert contrast matrix.

The subtype-adjusted model was

$$\log \mu_{dac} = \log N_{d,\mathrm{CD4}} + \beta^*_{0,ac} + \beta^*_{s,ac} s_d + \boldsymbol{\gamma}^{*\mathsf{T}}_{ac} \mathbf{x}_d + \boldsymbol{\delta}^{\mathsf{T}}_{ac} \mathbf{z}_d.$$

This model tests whether an axis–code association remains after conditioning on the complete conventional CD4 subtype composition. The adjusted p-values were corrected across all eligible cohort-level CD4 codes using the Benjamini–Hochberg procedure.

E.3 Analysis of representative axes and states

Matching atlas- and cohort-level axes. To relate the atlas-trained and cohort-fitted lexicons, we compared their gene masks. For each axis, genes were ranked by their mask weights, and the 500 highest-ranking genes were retained. For every pair of atlas and cohort axes, mask similarity was quantified using the Jaccard index,

$$J(A, B) = \frac{|G_A \cap G_B|}{|G_A \cup G_B|},$$

where G_A and G_B denote the respective top-500 mask-gene sets.

Using this analysis, Atlas Axis 30 was linked to Cohort Axes 1 and 15. Among the top 500 mask genes, Atlas Axis 30 and Cohort Axis 1 shared 179 genes, corresponding to a Jaccard similarity of 0.218 and a 7.38-fold enrichment over the overlap expected from the model gene universe. Atlas Axis 30 and Cohort Axis 15 shared 90 genes, corresponding to a Jaccard similarity of 0.099 and a 3.71-fold enrichment (Figure 6). We therefore interpreted Cohort Axes 1 and 15 as two related but more specific components of the broader programme represented by Atlas Axis 30.

Functions of axes and codes were analyzed as described in Appendix B.2. The results of representative axes and codes mentioned in the main text are provided in Tables 4 and 5.

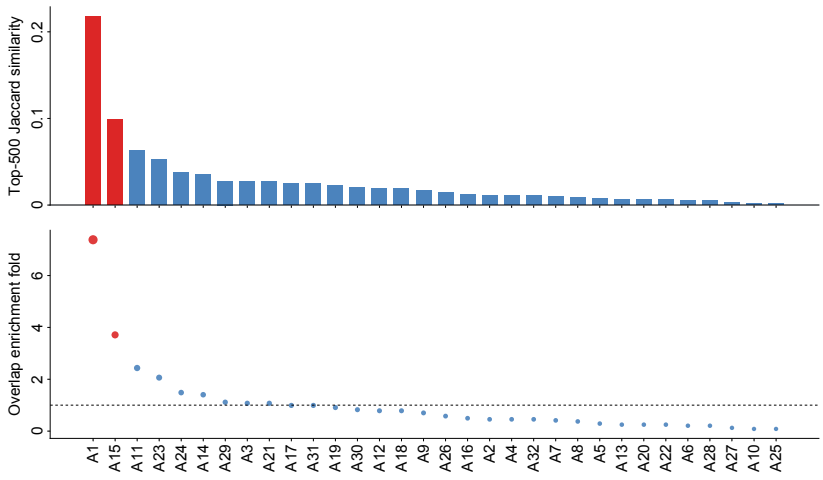

Figure 6: Atlas-to-cohort axis correspondence. Cohort axes were ranked by top-500 mask-gene overlap with Atlas Axis 30.

Table 4: Functional annotations of Atlas Axis 30 and its cohort-level refinements.

Lexicon	Axis	Functional summary	Representative GO terms
Atlas	A30	A broad immune-activation programme linking cell-surface signalling with actin-cytoskeletal organization.	**BP:** immune response; immune system process; cell activation. **CC:** side of membrane; cell surface. **MF:** actin binding; signalling receptor activity.
Cohort	A1	An immune-regulatory and cytoskeletal-adhesion programme connecting immune responses with actin-associated cellular organization.	**BP:** immune system process; immune response; regulation of immune system process. **CC:** focal adhesion; cell–substrate junction. **MF:** actin binding; cytoskeletal protein binding.
Cohort	A15	A more lymphocyte-focused activation programme involving proximal lymphocyte signalling and activation-associated regulation.	**BP:** lymphocyte activation; leukocyte activation; cell activation. **CC:** axonal growth cone. **MF:** no significant representative term.

Table 5: Functional annotations of representative severity-associated cohort-level axes and codes in $CD4^+$ T cells. For each code, annotations of genes with positive and negative effects are shown in red and blue, respectively.

Axis/code	Functional summary	Representative GO terms
A20	An axis contrasting ER/protein-processing activity with chromosome-segregation and cell-cycle activity.	**BP:** chromosome segregation; mitotic nuclear division; mitotic sister-chromatid segregation. **CC:** endoplasmic reticulum membrane; endoplasmic reticulum subcompartment. **MF:** protein disulfide isomerase activity; intramolecular oxidoreductase activity, transposing S–S bonds.
A20–C4	A severity-increasing ER/protein-processing state opposed to chromosome-segregation and cell-cycle activity.	**BP:** protein targeting to ER; establishment of protein localization to endoplasmic reticulum; response to endoplasmic reticulum stress. **CC:** endoplasmic reticulum membrane; endoplasmic reticulum subcompartment. **MF:** ribonucleoprotein complex binding; ATP hydrolysis activity. **BP:** chromosome segregation; cell-cycle process; mitotic sister-chromatid segregation. **CC:** chromosome, centromeric region; condensed chromosome. **MF:** no significant representative term.
A20–C0	A severity-decreasing chromosome-segregation/cell-cycle state opposed to ER-stress and ER protein-processing activity.	**BP:** chromosome segregation; mitotic sister-chromatid segregation; nuclear division. **CC:** chromosome, centromeric region; condensed chromosome. **MF:** no significant representative term. **BP:** response to endoplasmic reticulum stress; protein targeting to ER; establishment of protein localization to endoplasmic reticulum. **CC:** endoplasmic reticulum; endoplasmic reticulum membrane. **MF:** ribonucleoprotein complex binding; dolichyl-diphosphooligosaccharide–protein glycotransferase activity.
A16	An axis integrating mitochondrial respiratory-chain activity with RNA processing, translation, and protein biosynthesis.	**BP:** ATP synthesis coupled electron transport; mitochondrial ATP synthesis coupled electron transport; respiratory electron transport chain. **CC:** ribonucleoprotein complex; respiratory-chain complex. **MF:** RNA binding; oxidoreduction-driven active transmembrane transporter activity.

Continued on next page

Table 5 continued

Axis/code	Functional summary	Representative GO terms
A16–C10	A severity-decreasing state within the mitochondrial and biosynthetic axis, contrasting RNA-splicing/translation functions with mitochondrial respiratory-chain activity.	**BP:** RNA splicing via transesterification reactions with bulged adenosine as nucleophile; mRNA splicing via spliceosome; RNA splicing via transesterification reactions. **CC:** ribonucleoprotein complex; spliceosomal complex. **MF:** RNA binding; translation-factor activity. **BP:** mitochondrial ATP synthesis coupled electron transport; ATP synthesis coupled electron transport; respiratory electron transport chain. **CC:** respiratory-chain complex; proteasome complex. **MF:** electron-transfer activity; oxidoreduction-driven active transmembrane transporter activity.
A16–C14	A severity-decreasing state within the mitochondrial and biosynthetic axis, contrasting translation/protein biosynthesis with mitochondrial ATP synthesis and respiratory-chain activity.	**BP:** protein biosynthetic process; translation; formation of cytoplasmic translation-initiation complex. **CC:** ribonucleoprotein complex; eukaryotic translation initiation factor 3 complex. **MF:** RNA binding; translation-factor activity. **BP:** ATP synthesis coupled electron transport; mitochondrial ATP synthesis coupled electron transport; respiratory electron transport chain. **CC:** catalytic complex; respiratory-chain complex. **MF:** RNA binding; oxidoreduction-driven active transmembrane transporter activity.
A17	A protein-quality-control axis involving ubiquitin–proteasome, protein-catabolic, and autophagy-related processes.	**BP:** protein catabolic process; proteasomal protein catabolic process; macromolecule catabolic process. **CC:** ubiquitin-ligase complex; basal plasma membrane. **MF:** 2 iron, 2 sulfur cluster binding; proteasome binding.
A17–C4	A severity-decreasing protein-quality-control state contrasting macromolecule catabolism with autophagy-related functions.	**BP:** macromolecule catabolic process. **CC:** cytosolic ribosome; ribosome. **MF:** no significant representative term. **BP:** autophagy; process utilizing autophagic mechanism; protein catabolic process. **CC:** no significant representative term. **MF:** 2 iron, 2 sulfur cluster binding.
A17–C7	A severity-decreasing protein-quality-control state contrasting proteasomal degradation and ubiquitin-ligase functions with autophagy-related processes.	**BP:** proteasomal protein catabolic process; protein catabolic process; macromolecule catabolic process. **CC:** TORC2 complex; ubiquitin-ligase complex. **MF:** enzyme binding. **BP:** autophagy; process utilizing autophagic mechanism; cellular response to nitrogen levels. **CC:** autophagosome membrane; mitochondrion. **MF:** 2 iron, 2 sulfur cluster binding; proteasome binding.
A17–C10	A severity-decreasing protein-quality-control state contrasting ubiquitin-ligase-complex functions with autophagy and autophagosome-associated processes.	**BP:** no significant representative term. **CC:** ubiquitin-ligase complex; transferase complex. **MF:** no significant representative term. **BP:** process utilizing autophagic mechanism; autophagy; cellular response to nitrogen levels. **CC:** autophagosome membrane; autophagosome. **MF:** no significant representative term.