nature methods

Article

https://doi.org/10.1038/s41592-024-02175-z

SLIDE: Significant Latent Factor Interaction Discovery and Exploration across biological domains

Received: 23 November 2022

Accepted: 9 January 2024

Published online: 19 February 2024

Check for updates

Javad Rahimikollu[1,2,8], Hanxi Xiao[1,2,8], AnnaElaine Rosengart[1], Aaron B. I. Rosen[1,2], Tracy Tabib[3], Paul M. Zdinak[1], Kun He[4], Xin Bing[5], Florentina Bunea[6], Marten Wegkamp[6,7], Amanda C. Poholek[4] ✉, Alok V. Joglekar[1] ✉, Robert A. Lafyatis[3] ✉ & Jishnu Das[1] ✉

Modern multiomic technologies can generate deep multiscale profiles. However, differences in data modalities, multicollinearity of the data, and large numbers of irrelevant features make analyses and integration of high-dimensional omic datasets challenging. Here we present Significant Latent Factor Interaction Discovery and Exploration (SLIDE), a first-in-class interpretable machine learning technique for identifying significant interacting latent factors underlying outcomes of interest from high-dimensional omic datasets. SLIDE makes no assumptions regarding data-generating mechanisms, comes with theoretical guarantees regarding identifiability of the latent factors/corresponding inference, and has rigorous false discovery rate control. Using SLIDE on single-cell and spatial omic datasets, we uncovered significant interacting latent factors underlying a range of molecular, cellular and organismal phenotypes. SLIDE outperforms/performs at least as well as a wide range of state-of-the-art approaches, including other latent factor approaches. More importantly, it provides biological inference beyond prediction that other methods do not afford. Thus, SLIDE is a versatile engine for biological discovery from modern multiomic datasets.

Modern multiomic technologies can generate deep multiscale profiles. However, differences in data modalities, multicollinearity of the data, and large numbers of irrelevant features make the analyses and integration of high-dimensional omic datasets challenging. For example, multicollinearity can increase the variance of regression coefficients and lead to deflation of corresponding *P* values[1]. This is a major barrier to meaningful inference in a regression setting for high-dimensional multicollinear data. Further, human biological systems are complex, multifactorial and organized hierarchically, with complex interaction rules at each hierarchy. A linear model is often inadequate at capturing relevant higher-order relationships in such a system. Finally, while recent methods developed by us[2–7] and others[8–10] have harnessed these high-dimensional multiscale multimodal datasets to accurately predict different outcomes/groups of interest, they do not directly provide

[1]Center for Systems Immunology, Departments of Immunology and Computational & Systems Biology, University of Pittsburgh, Pittsburgh, PA, USA. [2]Joint CMU-Pitt PhD Program in Computational Biology, Pittsburgh, PA, USA. [3]Division of Rheumatology and Clinical Immunology, Department of Medicine, University of Pittsburgh, Pittsburgh, PA, USA. [4]Department of Pediatrics, University of Pittsburgh, Pittsburgh, PA, USA. [5]Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada. [6]Department of Statistics and Data Science, Cornell University, Ithaca, NY, USA. [7]Department of Mathematics, Cornell University, Ithaca, NY, USA. [8]These authors contributed equally: Javad Rahimikollu, Hanxi Xiao. ✉e-mail: poholeka@pitt.edu; joglekar@pitt.edu; lafyatis@pitt.edu; jishnu@pitt.edu

meaningful inference beyond prediction. In fact, approaches that do provide insights into the underlying mechanistic bases of outcome are tailored primarily for low-dimensional datasets, and often trade predictive power for inference[11].

In this Article, to address these, we present SLIDE, a novel data-distribution-free approach to analyze high-dimensional multiomic datasets and uncover latent factors that drive the outcome of interest (Fig. 1a). SLIDE makes no assumptions regarding the distribution of the underlying data as it significantly builds on a unique latent-factor regression framework developed by us[12,13]. It takes into account an extremely large search space of relationships to converge on a very small subset of biologically relevant and actionable latent factors. Critically, SLIDE incorporates both linear and nonlinear relationships, including complex hierarchical structures. It uncovers significant interacting latent factors in diverse contexts that span scales of organization from cellular/molecular phenotypes (for example, extent of clonal expansion of CD4 T cells) to organismal phenotypes (for example, disease severity of patients with diffuse systemic sclerosis). The discovery of these relationships is also coupled to rigorous false discovery rate (FDR) control via our unique analytical framework that creatively adapts ultramodern methods for FDR control[14]. SLIDE comes with provable statistical guarantees regarding identifiability of the latent factors and corresponding inference of significant interacting latent factors. This is fundamentally different from recent methods that rely on clever heuristics but do not have formal statistical guarantees or work only when strong biological priors are available. SLIDE has rigorous statistical guarantees, recapitulates known biological mechanisms and helps uncover novel biological mechanisms.

We tested the predictive performance of SLIDE on a range of datasets, and it outperformed/performed as well as several state-of-the-art approaches. Further, it provided novel inference not afforded by any existing approaches, thus being one of the only methods that simultaneously provides meaningful inference for high-dimensional data without compromising on predictive power. When analyzing datasets from patients with systemic sclerosis (SSc) to elucidate the basis of SSc pathogenesis, SLIDE recovered altered transcriptomic states in myeloid cells and fibroblasts, a well-studied basis of SSc disease severity[15–20]. But it also identified an unexplored keratinocyte-centric signature (validated by protein staining), and a novel mechanism involving an interaction between the altered transcriptomic states in myeloid cells and fibroblasts with human leukocyte antigen (HLA) signaling in macrophages. SLIDE also worked extremely well across a range of modern spatial modalities, including 10X Visium, Slide-seq, MERFISH and CODEX, in recapitulating immune and neuronal cell partitioning by 3D location. In the characterization of latent factors underlying clonal expansion of CD4 T cells, SLIDE recapitulated well-known inhibitory receptors and markers of activation/exhaustion, but also identified several novel markers that standard differential expression analyses would have missed. Overall, SLIDE is an engine for biological discovery from modern multiomic datasets.

Results

The SLIDE framework

SLIDE is an interpretable latent factor regression-based machine learning approach (Fig. 1b). It identifies significant latent factors capturing linear and nonlinear relationships (up to pairwise interactions) between observed data (X, typically high-dimensional, multicollinear) and the response of interest (Y) (Fig. 1b). SLIDE consists of three steps starting with the unsupervised identification of latent factors (Z) from the data (equation (1)),

$$X' = AZ + E \tag{1}$$

$X_{n\times p}$ represent the data matrix with n samples and p features. Using our previously described LOVE approach[13], X decomposes into two factors: $A_{p\times K}$ and $Z_{K\times n}$, with an error term E. A is the allocation matrix and represents the membership of each feature to a latent factor. Z is the latent factor matrix and represents a lower-dimensional representation ($K < p$) of the input data in latent space. Critically, this decomposition, unlike other factor analysis/non-negative matrix factorization (NMF) approaches, comes with theoretical guarantees regarding unique identifiability of the latent factors without assumptions regarding data-generating mechanisms. It permits overlapping latent factors, and there are no restrictive assumptions regarding orthogonality. The only assumption is to anchor each of the latent factors using two pure variables (that is, variables associated with only one latent factor).

The next step in SLIDE (Fig. 1c, equations (2)–(4)) is the identification of significant standalone latent factors using a regression model that utilizes the LOVE latent factors.

$$\mathrm{LP} = \sum_{j\in S1} \beta_j Z_j + \epsilon_1 \qquad S1 \text{ determined by knockoffs} \tag{2}$$

Here LP is the linear part of the SLIDE model. Without identifiability (for example, in a NMF setting), it would be meaningless to perform regression using the latent factors as they are stochastic and/or unstable. The identifiability guarantees allow us to meaningfully use these factors in a regression setting with corresponding guarantees on inference and FDR control in the regression model (Methods and Supplementary Note 1).

The identification of significant latent factors uses a multistage adaptation of an ultramodern framework for FDR-controlled variable selection—knockoffs[14]. This approach is based on differences or lack thereof between true and fake (knockoff) variables. These knockoff variables are approximately orthogonal (with a deviation magnitude of 1–s) to the response variable, preserving the covariance structure (Σ) as illustrated in equations (3)–(5). This means that the correlation between the original variable Z_i and the knockoff variable $\tilde{Z}_i$ is 1–s, with $s \approx 1$.

$$Z^T Z = \Sigma \tag{3}$$

$$\tilde{Z}^T \tilde{Z} = \Sigma \tag{4}$$

$$Z^T \tilde{Z} = \Sigma - \mathrm{diag(s)} \tag{5}$$

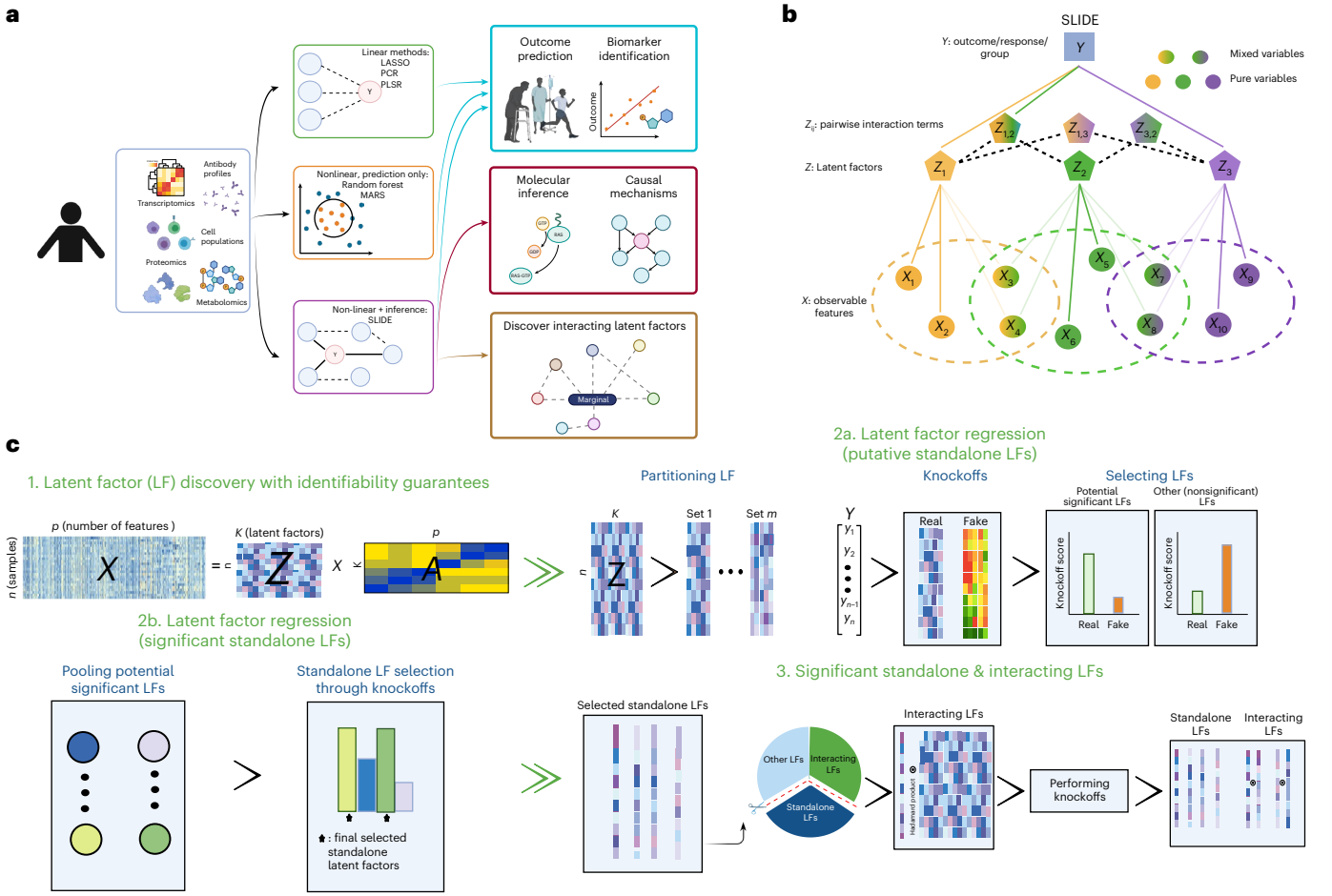
While the knockoff strategy has typically been used on observed variables, we adapted it for use on the latent factors. Here, the variable Z_j is statistically significant if it considerably outperforms its knockoff $\tilde{Z}_j$ based on W_j, the test statistic of interest, as defined by equations (6) and (7):

$$M_j = \max(\lambda) \text{ where } |b_{j(\lambda)}| > 0 \tag{6}$$

$$W_j = \max(M_j, \tilde{M}_j) \times \mathrm{sgn}(M_j - \tilde{M}_j) \tag{7}$$

In this approach, we identify important latent factors by maximizing the L1 regularization hyperparameter λ, such that for the original and knockoff variables, the absolute value of their corresponding coefficients $|b_{j(\lambda)}|$ remains positive as defined in equation (6). If a latent factor Z_j is strongly associated with the response variable y, increasing λ will result in a high value of M_j and $b_{j(\lambda)}$ will remain positive. If its corresponding knockoff is unimportant, $\tilde{M}_j$ will be small (that is, this latent factor is truly important). Conversely, if the value of M_j is small and the corresponding value for the knockoff $\tilde{M}_j$, is similar/higher, the latent factor is unimportant. As a result, a higher score for the test statistic W_j selects important latent factors with corresponding unimportant knockoffs. Further, our adaptation of the knockoff approach is a

https://doi.org/10.1038/s41592-024-02175-z



d

Method	Identifiability?	No assumptions regarding data-generating mechanisms?	Not restricted to only single-cell data modalities?	Can handle correlated latent factors?	Can handle nonlinear relationships?	Mechanism to control FDR?	Inference beyond prediction?	Can identify factors related to a specific outcome of interest?
SLIDE	✓	✓	✓	✓	✓	✓	✓	✓
LOVE	✓	✓	✓	✓	✗	✗	✗	✗
MOFA+	✗	✗	✓	✗	✗	✗	✓	✗
scVI	✗	✗	✗	✓	✓	✗	✗	✗
Linear scVI	✗	✗	✗	✓	✗	✗	✓	✗
VAE	✗	✓	✓	✓	✓	✗	✗	✗
PLSR	✗	✓	✓	✓	✗	✗	✓	✓
PCR	✗	✓	✓	✗	✗	✗	✗	✓
LASSO	✗	✓	✓	✗	✗	✗	✓	✓
PHATE	✗	✓	✓	✓	✗	✗	✗	✗

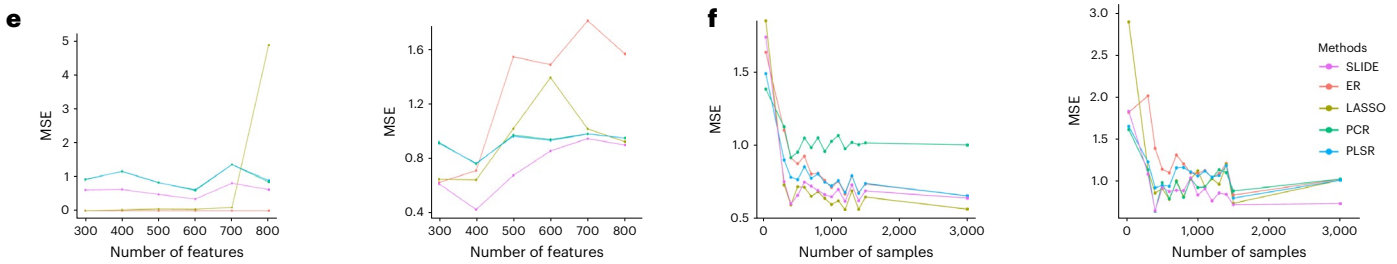


Fig. 1 | SLIDE—a novel interpretable machine learning method for Significant Latent Factor Interaction Discovery and Exploration. **a**, Schematic illustrating the vast array of datasets on which SLIDE can be applied and the key advances over existing analytical frameworks for the analyses of these datasets. **b**, Conceptual overview of the SLIDE algorithm. **c**, Schematic summarizing the implementation and different steps in SLIDE. **d**, Key conceptual innovations of SLIDE. **e**, Comparison of the predictive performance of ER, LASSO, PCR, PLSR and SLIDE on simulated datasets across a range of number of features without (left) and with (right) interaction terms. MSE, mean squared error. **f**, Comparison of the predictive performance of ER, LASSO, PCR, PLSR and SLIDE on simulated datasets across a range of sample sizes without (left) and with (right) interaction terms.

multistage stage procedure (Fig. 1c). Initially, latent factors are divided into sets, to which we apply knockoffs and identify putative significant latent factors. In stage 2, these latent factors undergo another round of selection via knockoffs to converge on a set of standalone significant latent factors. We repeat these two stages to identify stable (corresponding stability parameter, where 'spec' is the frequency of selection

from repeated application of knockoffs) standalone significant latent factors (Methods and Supplementary Note 1).

The final step in SLIDE incorporates nonlinear relationships—we identify significant interactors of the standalone significant latent factors (S1) that is, each interaction term has at least one standalone significant latent factor (equations (8) and (9)).

$$\mathrm{NP}_j = \beta_j Z_j + \sum_i C_{ij} Z_i \odot Z_j + \epsilon_2 \text{ where } i \neq j, j \in S1, \quad i \in \{1 \ldots K\} \text{ and } Z_i \odot Z_j \in S2 \tag{8}$$

$$y' = \sum_j \theta_j \mathrm{NP}_j + \epsilon_3 \text{ where } NP_j \in S3 \text{ from knockoffs} \tag{9}$$

Here $\beta \in R^{1xK}$ and C_{ij} is the effect size of the interaction term between the two latent variables Z_i and Z_j. S2 is the set of putative interactions involving the standalone significant latent factors. Knockoffs are applied again to extract the significant interaction terms (Fig. 1c). However, these knockoffs are on the pairwise interaction terms to identify significant interacting latent factors. If C_{ij} for variable Z_j is zero for $i \in \{1 \ldots K\}$, the latent factor is standalone significant without any interactors. S3 is the final set of significant latent factors (standalone and interacting). Overall, the combination of eight unique properties: (1) identifiability of the latent factors, (2) lack of assumptions regarding data generating mechanisms, (3) applicability to any data modality (single cell, spatial, bulk and so on), (4) the ability to handle correlated factors, (5) ability to handle nonlinear relationships, (6) FDR control, (7) the ability to provide inference beyond prediction and (8) specificity in identifying significant latent factors related to specific outcomes of interest enable SLIDE to outperform existing approaches (Fig. 1d).

Using simulations (Methods), we compared the performance of SLIDE to other state-of-the-art methods including essential regression (ER)[12], least absolute shrinkage and selection operator (LASSO)[21], partial least squares regression (PLSR)[22] and principal components regression (PCR)[23] with and without interaction terms (Fig. 1e,f). SLIDE performs as well as state-of-the-art approaches when there are no interaction terms present (Fig. 1e,f). In the presence of interaction terms, it consistently outperforms these methods (Fig. 1e,f). Importantly, all approaches other than SLIDE and LASSO use the full model (all features/clusters) for prediction. However, SLIDE only uses a small number of prioritized latent factors for prediction. Next, as simulations use only synthetic datasets, we sought to test the performance of SLIDE across a diverse range of biological contexts.

SLIDE uncovers novel interacting latent factors that explain SSc pathogenesis

Using SLIDE, we first sought to discover interacting latent factors underlying SSc disease severity. We analyzed single-cell RNA sequencing (scRNA-seq) data from 24 subjects with SSc[15,24] across the severity spectrum (Fig. 2a), where disease severity was quantified using the Modified Rodnan Skin Score (MRSS). We identified 35 unique clusters and retained clusters with at least 20 cells for each of the 24 subjects for downstream analyses (Fig. 2b). Next, we applied SLIDE on these cell-type-specific transcript abundances to predict SSc severity and infer corresponding significant interacting latent factors of outcome (Methods and Supplementary Fig. 1a,b). We benchmarked SLIDE against a wide range of state-of-the-art approaches—ER[12], LASSO[21], a variational autoencoder (VAE), MOFA+–regression (linear regression coupled to MOFA+ (ref. 25)), PHATE–regression (linear regression coupled to PHATE[26]), PLSR[22] and PCR[23]. Although MOFA+ and PHATE are unsupervised approaches, for a fair comparison across the methods, we used the clusters/latent factors uncovered by these methods (MOFA+ and PHATE) in a model to regress to MRSS. SLIDE was able to accurately predict SSc severity and outperformed five of our seven benchmarks—PLS, PCR, PHATE–regression, MOFA+–regression and a VAE in terms of prediction accuracy (Fig. 2c and Supplementary Fig. 1c). Interestingly, the two other latent factor-based approaches—MOFA+ and VAE both underperformed SLIDE in terms of prediction performance. LASSO and ER (developed by us) were the only methods with comparable prediction performance (Fig. 2c and Supplementary Fig. 1c). However, LASSO only identified a small set of predictive biomarkers that were uninformative of the actual molecular basis underlying SSc pathogenesis. On the other hand, SLIDE identified nine significant latent factors that could be used to infer the mechanistic basis of SSc pathogenesis. Further, while the performance of SLIDE and ER were comparable, ER used the entire set of latent factors to predict outcome, while SLIDE used only nine. Thus, SLIDE provides the same predictive power as ER but has stronger inference with fewer latent factors (Supplementary Fig. 1d).

The nine latent factors uncovered by SLIDE spanned a range of cell-intrinsic and cell-extrinsic circuits (Fig. 2d), encompassing altered transcriptomic states that have been characterized and recognized to be critical in SSc pathogenesis. These states include modulated inflammatory states/signaling in myeloid cells and fibroblasts, including SFRP2 fibroblasts, which are well-known bases of SSc pathogenesis (Fig. 2d)[15–20]. Other canonical mechanisms recapitulated include cross-talk between interferon signaling and myeloid inflammatory signaling (Fig. 2d)[15–20]. Key genes that contribute to these altered transcriptomic states include cytokines and chemokines (for example, CCL19), signaling molecules (for example, WIF1), interferon signaling genes (for example, IGFBP5), components of mechano-transduction (for example, THBS1) and alarmins/damage sensing molecules (for example, S100A9). These agree well with previous studies by us and others[15–20]. In addition to recovering well-known mechanisms, we converged on several novel mechanisms. The first involves a previously unelucidated role of keratinocytes in SSc pathogenesis (Fig. 2d). We have recently validated this keratinocyte functional signature by protein staining[27]. We also converged on another novel mechanism involving interactions between altered myeloid/endothelial cell inflammation and keratinocyte–fibroblast–endothelial cell crosstalk. This interaction hinges on altered HLA signaling (Fig. 2d). While our work is the first to study this at the transcriptomic level, there is evidence for this mechanism in recent genetic studies[28].

We compared the predictive power of standalone significant latent factors to size-matched random ones, and our actual model outperformed the random set at different stability parameters for the selection of significant latent factors via the repeated application of knockoffs (Fig. 2e and Supplementary Fig. 1e). We also assessed the quality of the interacting latent factors by fixing the standalone factors and swapping the interactors with a size-matched randomly chosen set. As expected, the model's performance decreased, highlighting the importance of having the right interacting latent factors for predicting SSc pathogenesis mechanisms.

Canonical markers of SSc severity (including those captured by LASSO) such as CCL19, IGFBP5, WIF1, SAA1 and THBS1 (refs. 15–20) had significant high linear correlations with MRSS, but almost no nonlinear relationships (Fig. 2f–h). Further, genes such as APOE, S100A9 have both significant linear and nonlinear relationships with MRSS (Fig. 2f–h). Some of these are entirely novel, and others have begun to be characterized in SSc[15–20]. Finally, several have only nonlinear relationships with MRSS (Fig. 2f–h). Most of these have been missed by previous approaches. Evaluating these nine latent factors with MRSS revealed strong relationships with most of them, showing that SLIDE accurately captures context-specific biological group structures with valuable information about SSc pathogenesis (Fig. 2i).

Canonical biomarker approaches including LASSO focused on a handful of individual genes and do not capture any information regarding functional groups. Pathway-centric approaches do have group information, but these groups are predefined and not tailored to the specific context being analyzed. Only a handful of

https://doi.org/10.1038/s41592-024-02175-z

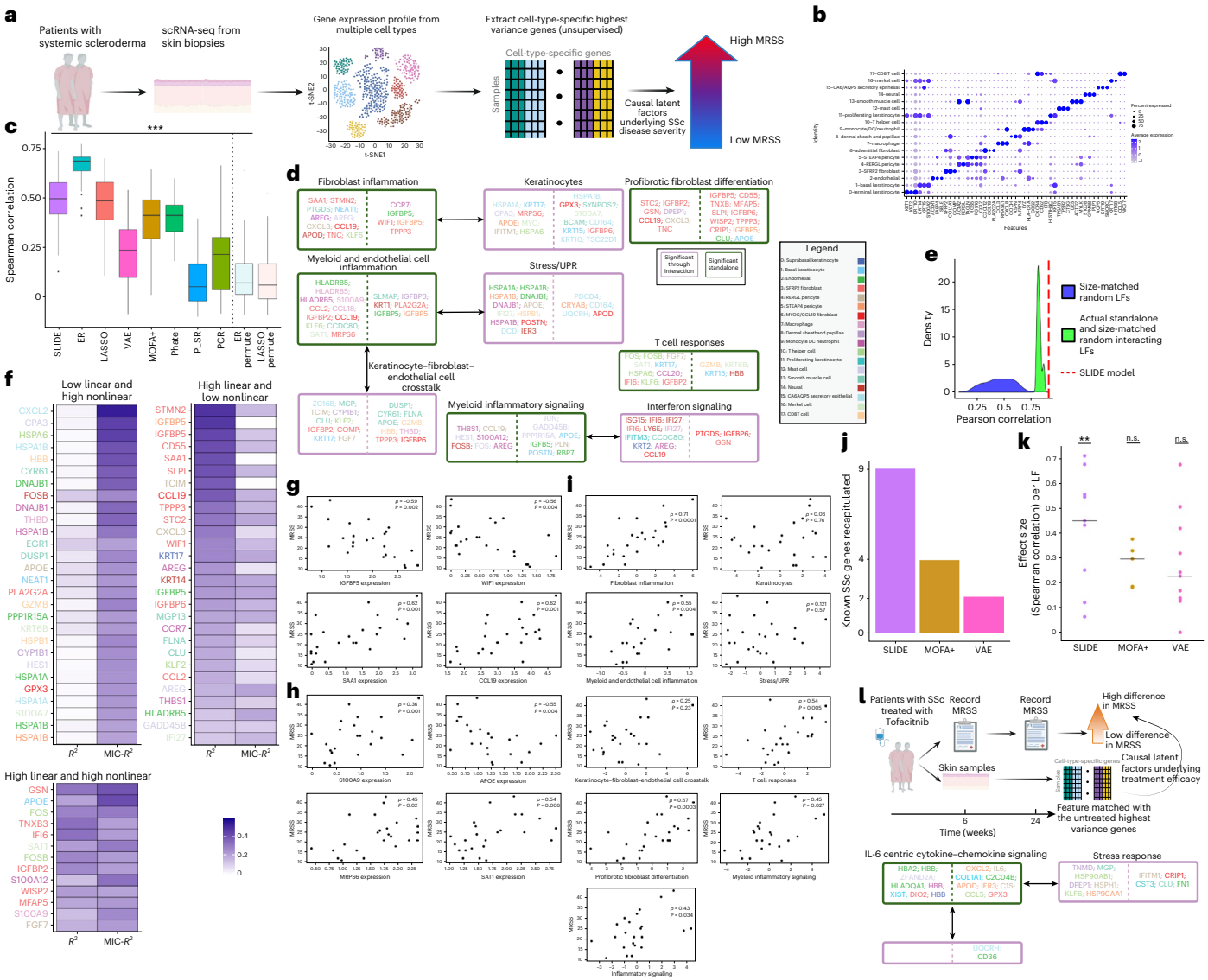


Fig. 2 | SLIDE uncovers novel interacting latent factors that explain SSc pathogenesis. a, Schematic summarizing the overall setup. t-SNE, t-distributed stochastic neighbor embedding. **b**, Cellular cluster identities defined by top cell-type-specific differentially expressed genes (DEGs). **c**, Spearman correlations between true MRSS and MRSS predicted using different methods—SLIDE (spec = 0.1), ER, LASSO, VAE, MOFA+—regression, PHATE—regression, PLSR and PCR. Model performance plotted across 50 replicates of *k*-fold cross-validation with permutation testing. ***exact *P* from a permutation test <0.01. **d**, Significant interacting latent factors identified by SLIDE. Green boxes denote significant standalone latent factors, and purple boxes denote significant interacting latent factors. Color corresponds to the cell type. Genes on the left and right of the dashed line have negative and positive correlations with MRSS, respectively. **e**, Performance of the real model (spec = 0.1) relative to (1) the distribution of the performance of models built using size-matched random latent factors (blue) and (2) the distribution of the performance of models built using the actual significant standalone latent factors and size-matched random interacting latent factors (green). **f**, Linear (Spearman correlations) and nonlinear (MIC) relationships between key components of the latent factors and MRSS. **g**, MRSS and expression of genes with a significant linear relationship with MRSS. **h**, MRSS and expression of genes with a significant nonlinear relationship with MRSS. UPR, unfolded protein response. **i**, Scatter plot between each significant latent factor from SLIDE and MRSS. **j**, The number of known drivers, identified from previously published bulk RNA-seq studies recovered by the SLIDE, VAE and MOFA+ models. **k**, Effect sizes of the SLIDE, MOFA+ and VAE latent factors in stratifying patients by their MRSS. *P* calculated by a Mann–Whitney *U* test. The null distribution is built with random size-matched non-significant SLIDE latent factors. **$P < 0.05$. n.s., not significant. **l**, Significant standalone and interacting latent factors underlying changes in MRSS on treatment with tofacitinib. For box plots, the box spans from the first to the third quartile, and the whiskers extend from the first quartile −1.5 interquartile range (IQR) to the third quartile +1.5 IQR.

recent approaches (for example, MOFA+ or VAE-based methods) try to identify context-specific groups. To better evaluate how SLIDE performs relative to these approaches, we compared the quality of the inferred latent factors across the relevant approaches (that is, approaches that use latent factors or equivalent entities in the model). We first benchmarked these approaches by comparing their recovery of known drivers (from prior bulk RNA-seq studies) of SSc pathogenesis[16–20,29,30]. While MOFA+ and VAE captured only four and two known genes, respectively, underlying the severity of SSc, SLIDE captured nine (Fig. 2j and Supplementary Fig. 1f,g). This demonstrates the superior performance of SLIDE in recapitulating known markers of SSc severity. Next, we moved beyond individual genes to context-specific groups. SLIDE latent factors, compared to both MOFA+ and VAE latent factors were significantly more correlated to MRSS (Fig. 2k and Supplementary Fig. 1f,g) demonstrating that SLIDE also better captures context-specific groups that can stratify by disease severity. SLIDE also outperforms other unsupervised clustering approaches, confirming that it hones in on

meaningful significant latent factors underlying outcomes of interest (Supplementary Fig. 1h–j).

To more rigorously test the biological significance of these latent factors, we moved to a 'human perturbation experiment' where 10 of these 24 subjects had recently been in a clinical trial with tofacitinib (tofa) (Fig. 2l)[31]. We used SLIDE to identify significant interacting latent factors that underlie the reduction in disease severity (change in MRSS from pre to post). Remarkably, SLIDE very accurately honed in on the known IL6/JAK/STAT-centric molecular mechanism[31] underlying tofa treatment. This demonstrates SLIDE's power in meaningful inference of complex high-dimensional datasets (Fig. 2l). And it was able to do so even at an early time point (we used 6 week scRNA-seq data to predict outcomes at week 24, Fig. 2l), demonstrating the sensitivity of SLIDE in capturing subtle changes over the course of treatment.

SLIDE uncovers latent factors underlying immune cell partitioning by 3D localization

We applied SLIDE to spatial transcriptomic datasets to uncover latent factors underlying the 3D spatial partitioning of immune cells in different contexts. First, 10X Visium was performed in a murine allergy model[32,33] where animals were treated intranasally with house dust mite (HDM) for five consecutive days and mediastinal lymph nodes (mLNs) were isolated from these animals after the third (D3) and fifth (D5) day followed by spatial RNA-seq (Fig. 3a and Methods). Clustering results of the spatial regions were overlayed with fluorescence microscopy images, designating spatial labels of border, central and intermediate zones (Fig. 3b). Border regions showed B cell enrichment, while central areas showed CD4 T cell and dendritic cell enrichment (Fig. 3b). These labels denoted only spatial locations, not cell types. This allows us to test the biological significance of the factors uncovered by the different methods—if they are indeed meaningful, they should reflect this immune cell partitioning.

SLIDE was able to accurately predict spatial labels for the D3 samples, and outperformed PLS, PCR and PHATE–regression in terms of prediction accuracy (Fig. 3c and Supplementary Fig. 2a–c). Further, SLIDE provides the same predictive power as ER but stronger inference with fewer latent factors (Supplementary Fig. 2d). The SLIDE latent factors, relative to the other methods that have similar prediction performance (LASSO, MOFA+ and VAE), also provided more meaningful inference of factors underlying immune cell partitioning by 3D location.

Interestingly, although SLIDE was only given spatial labels, the identified latent factors consisted of genes that mark B cells, CD4 T cells and dendritic cells (DCs), aligning with fluorescence microscopy images (Fig. 3d). The seven latent factors represent multiple immune cell canonical functions including broad adaptive immune responses, antigen processing and presentation and specific humoral responses (Fig. 3d). When compared to a size-matched set of random latent factors, the actual latent factors performed significantly better at different stability parameters for the selection of significant latent factors via the repeated application of knockoffs (Fig. 3e and Supplementary Fig. 2e). When keeping the actual standalone latent factors fixed but shuffling the interactors, the performance of this model (at different stability parameter settings) was significantly lower compared to the actual model (Fig. 3e and Supplementary Fig. 2e). While some genes in the significant latent factors had significant linear relationships with spatial labels (Fig. 3f and Supplementary Fig. 2f), several others only had nonlinear relationships (Fig. 3f and Supplementary Fig. 2f).

Next, we found significant relationships between individual latent factors and the spatial region labels (Fig. 3g). SLIDE captures true context-specific biological group structure where each individual context-specific group (latent factor) has meaningful information regarding the spatial region label of interest. These inferences provided by SLIDE surpass those provided by other methods that had comparable prediction performance—LASSO and MOFA+. LASSO inherently (because of L1 regularization) provides only individual biomarkers. The SLIDE latent factors had significantly higher effect sizes than the MOFA+ latent factors in identifying immune cell partitioning by spatial location (Fig. 3h and Supplementary Fig. 2g–i).

We also sought to evaluate whether SLIDE could recapitulate spatial partitioning at D5 (Fig. 3i). While there is noticeable cell migration from D3 to D5 post HDM treatment, the overall orientation of cells remains the same[33]. SLIDE outperformed all the benchmarks in terms of prediction (Fig. 3j and Supplementary Fig. 3a,b). Among the six latent factors (Fig. 3k) selected by SLIDE, we indeed observe recapitulation of both individual genes (for example, Trbc2, Cd3d and Ms4a4b) and broader signatures from the D3 analyses. Moreover, while there are some differences in membership in the latent factors, the overall processes represented by the latent factors remain similar across the D3 and D5 analyses (Fig. 3k). However, the MOFA+ and VAE latent factors are unstable and fail to recapitulate this trend (Supplementary Fig. 3c,d). As earlier, SLIDE outperforms both a size-matched set of random latent factors and a size-matched set of latent factors where the standalone factors are 'real', but the interactors are shuffled (Fig. 3l). We also evaluated SLIDE's stability and interpretability on another replicate of this experiment. As expected, SLIDE captured similar latent factors (Fig. 3m and Supplementary Fig. 4a–h). SLIDE accurately and stably, across time points and replicates, captures immune cell partitioning in an allergy model of asthma.

SLIDE enables discovery of significant latent factors underlying spatial partitioning for a wide range of spatial data modalities

Next, we evaluated SLIDE on a wide range of other spatial data modalities and technologies—Slide-seq[34], MERFISH[35] and CODEX[36]. We used SLIDE to again examine immune cell partitioning by spatial localization within a lymph node in a murine model of asthma. However, we now used spatial data generated using Slide-seq instead of the 10X Visium platform (Fig. 4a,b and Methods). Immunofluorescence confirmed that border regions were enriched for B cells (blue) and the central regions for CD4 T cells (green) and DCs (pink, Fig. 4b). However, as earlier, the actual immune cell partitioning was not used in the labels at all—the labels only corresponded to spatial location. SLIDE was able to accurately predict spatial labels, outperforming PLS, PCR and PHATE–regression (Fig. 4c and Supplementary Fig. 5a,b). The SLIDE latent factors, relative to the other methods that have similar prediction performance (LASSO, MOFA+ and VAE) also provided more meaningful inference of factors underlying partitioning by 3D location (Supplementary Fig. 5c,d).

Using only spatial labels, SLIDE identified six latent factors consisting of genes that mark B cells, CD4 T cells and DCs, in agreement with the true spatial partitioning (Fig. 4d). As earlier, the latent factors uncovered by SLIDE included processes related to innate and adaptive immune responses (Fig. 4d). More interestingly, SLIDE uncovered two additional processes: antibody-mediated complement activation, recapitulating a well-known but complex role of the complement system in allergic asthma[37] and PPAR signaling, hinting at a relatively novel mechanism of pathogenic type II immune responses in lung inflammation as asthma mediated by PPARγ expressed by DCs and T cells[38]. We also analyzed the predictive power and SLIDE's actual latent factors performed significantly better than a random size-matched latent factors and those with shuffled interactors (Fig. 4e). Further, SLIDE captured genes which only had nonlinear relationships that would have been missed by traditional regression methods (Fig. 4f). The inferences provided by SLIDE surpass those provided by other methods that had comparable prediction performance—in particular, MOFA+ (Fig. 4g).

Next, we employed SLIDE to dissect differences in the spatial localization of five different subclasses of glutamatergic neurons, including five extra telencephalic projecting (L5 ET), layer 5/6 near-projecting (L5/6 NP), layer 6 CT (L6 CT), layer 6b (L6b) and intratelencephalic (IT) neurons, in the murine primary motor cortex[35] (Fig. 4h). SLIDE accurately predicted

https://doi.org/10.1038/s41592-024-02175-z

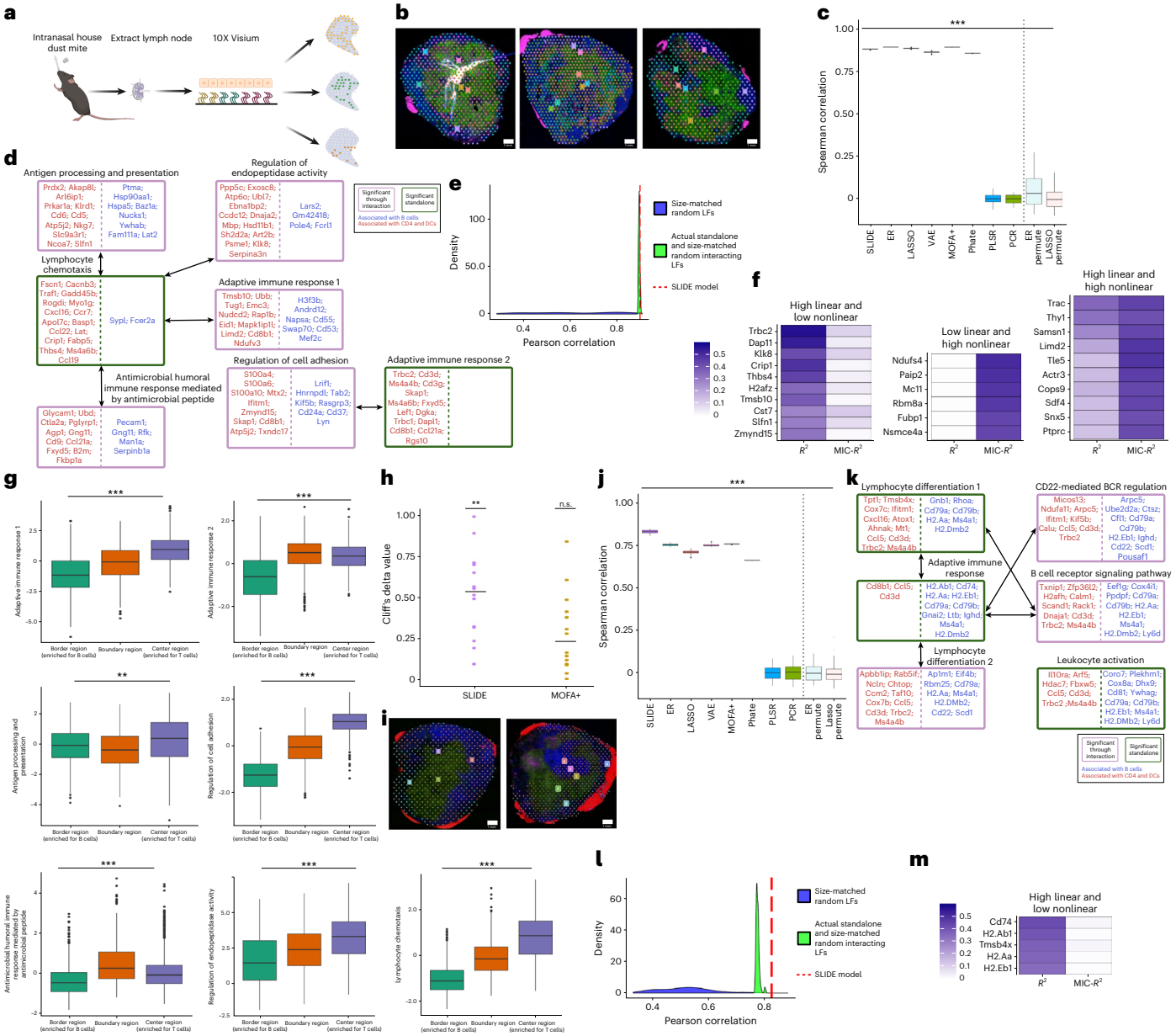


Fig. 3 | SLIDE uncovers latent factors underlying immune cell partitioning by 3D localization in a murine model of asthma. a, Schematic of the 10X Visium experiment. **b**, K-nearest neighbors (KNN) clustering of the spatial regions overlayed with microscopic images of three D3 technical replicates (blue, B cells; green, CD4 T cells; pink, dendritic cells). **c**, Spearman correlations between true and predicted spatial region for D3 lymph nodes using different methods—SLIDE (spec = 0.1), ER, LASSO, VAE, MOFA+–regression, PHATE–regression, PLSR and PCR. Model performance is plotted across 50 replicates of fivefold cross-validation framework with permutation testing. ***exact *P* from a permutation test <0.01. **d**, Significant interacting latent factors for D3 samples. Green, significant standalone latent factors; purple, significant interacting latent factors. **e**, Performance of the real model (spec = 0.1) for D3 samples relative to null models as described in Fig. 2e. **f**, Linear (Spearman correlations) and nonlinear (MIC) relationships between key components of the D3 latent factors and spatial region. **g**, Box plots illustrating the distributions (across cells) of each SLIDE latent factor across spatial regions. *P* values are calculated using Kruskal–Wallis test. ***$P < 0.01$, **$P < 0.05$. **h**, Effect sizes of the SLIDE latent factors from **g** and top size-matched MOFA+ latent factors (each dot corresponds to a latent factor) in discriminating by spatial localization. *P* from a two-sided Mann–Whitney *U* test. The null distribution is built with random size-matched nonsignificant SLIDE latent factors. **$P = 0.028$. n.s., not significant. **i**, KNN clustering of the spatial regions overlayed with microscopic images of two D5 technical replicates (blue, B cells; green, CD4 T cells; pink, dendritic cells). **j**, Spearman correlations between true spatial region and spatial region predicted for D5 lymph nodes using different methods—SLIDE (spec = 0.1), ER, LASSO, VAE, MOFA+, PHATE–regression, PLSR and PCR. Model performance is plotted across 50 replicates of *k*-fold cross-validation with permutation testing. ***$P < 0.01$. **k**, Significant interacting latent factors for D5 samples identified by SLIDE. Other conventions correspond to **d**. BCR, B cell receptor. **l**, Performance of the real model for D5 samples relative to models as described in Fig. 2e. **m**, Linear Spearman correlations and nonlinear relationships (quantified using MIC) between key components of the D5 latent factors and spatial region. For box plots, the box spans from the first to the third quartile, and the whiskers extend from the first quartile −1.5 interquartile range (IQR) to the third quartile +1.5 IQR.

the neuron localization and captured highly interpretable latent factors that represent and capture multiple well-known aspects of neuronal differentiation and axonal development (Fig. 4i–k). Furthermore, SLIDE identified several genes that only had nonlinear relationships (Fig. 4l).

SLIDE was also applied to spatial proteomic data from healthy (BALBc) and lupus (MRL/lpr) mice[36]. Here, instead of focusing on immune cell partitioning by spatial location (a cellular phenotype), we focus on spatial differences in protein abundance between healthy and

https://doi.org/10.1038/s41592-024-02175-z

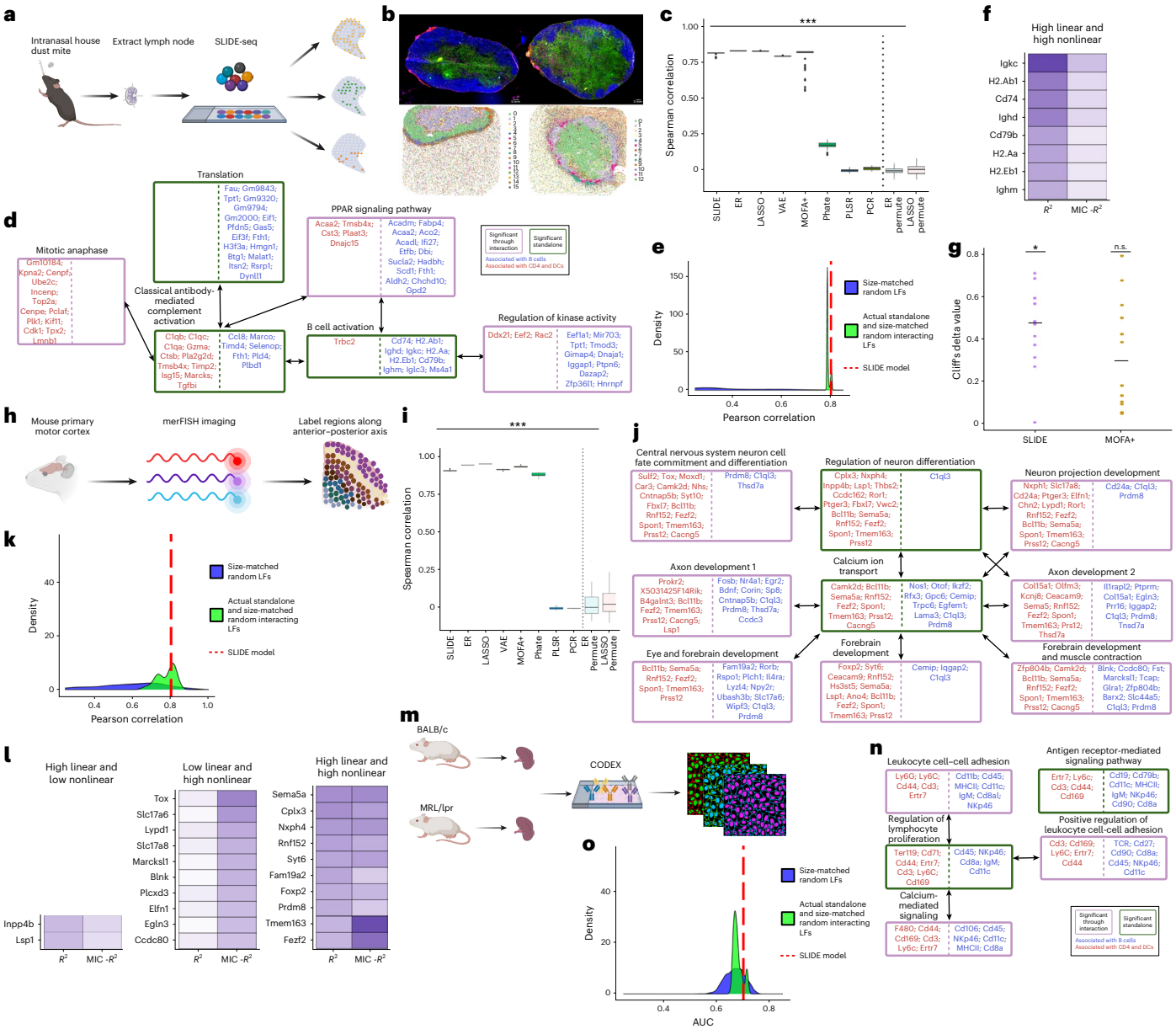


Fig. 4 | SLIDE uncovers latent factors underlying spatial localizations and phenotypes from different spatial transcriptomic and proteomic modalities. **a**, Schematic summarizing the Slide-seq experiment. **b**, KNN clustering of the spatial regions and microscopic images of two technical replicates of mLNs. (blue, B cells; green, CD4 T cells; pink, dendritic cells). **c**, Spearman correlations between true and predicted spatial region for D3 lymph nodes using different methods—SLIDE (spec = 0.1), ER, LASSO, VAE, MOFA+—regression, PHATE—regression, PLSR and PCR. Model performance is plotted across 50 replicates of fivefold cross-validation with permutation testing. ***$P < 0.01$. **d**, Significant interacting latent factors identified by SLIDE. Green, significant standalone latent factors; purple, significant interacting latent factors. **e**, Performance of the real model (spec = 0.1) relative to null models as described in Fig. 2e. **f**, Linear (Spearman correlations) and nonlinear (MIC) relationships between key components of the latent factors and spatial region. **g**, Effect sizes of the SLIDE latent factors from **d** and top size-matched MOFA+ latent factors (each dot corresponds to a latent factor) in discriminating by spatial localization. *P* from a Mann–Whitney *U* test. The null distribution is built with random size-matched nonsignificant SLIDE latent factors. **$P < 0.05$. n.s., not significant. **h**, Schematic summarizing MERFISH data from different subsets of glutamatergic neurons spatially distributed across the murine motor cortex. **i**, Spearman correlations between true spatial region and spatial region predicted for day 3 treated lymph nodes using different methods—SLIDE (spec = 0.1), ER, LASSO, VAE, MOFA+—regression, PHATE—regression, PLSR and PCR. Model performance plotted across 50 replicates of fivefold cross-validation framework with permutation testing. ***$P < 0.01$. **j**, Significant interacting latent factors identified by SLIDE. Green, significant standalone latent factors; purple, significant interacting latent factors. **k**, Performance of the real model for D5 samples relative to null models as described in Fig. 2e. **l**, Linear Spearman correlations and nonlinear relationships (quantified using MIC) between key components of latent factors and spatial region. **m**, Schematic summarizing CODEX data from BALBc and MRL/lpr murine spleens. **n**, Significant interacting latent factors identified by SLIDE. Green, significant standalone latent factors; purple, significant interacting latent factors. **o**, Performance of the real model for D5 samples relative to null models as described in Fig. 2e. AUC, area under the receiver operating characteristic curve. For box plots, the box spans from the first to the third quartile, and the whiskers extend from the first quartile −1.5 interquartile range (IQR) to the third quartile +1.5 IQR.

SLE individuals (organismal phenotypes) (Fig. 4m). SLIDE uncovered interesting latent factors reflective of well-known rewired cellular programs in SLE, including altered antigen processing and presentation, cell proliferation and adhesion as well as Ca^{2+} signaling (Fig. 4n). The actual latent factors performed significantly better than a random size-matched set of latent factors, as well as a size-matched set of actual

https://doi.org/10.1038/s41592-024-02175-z

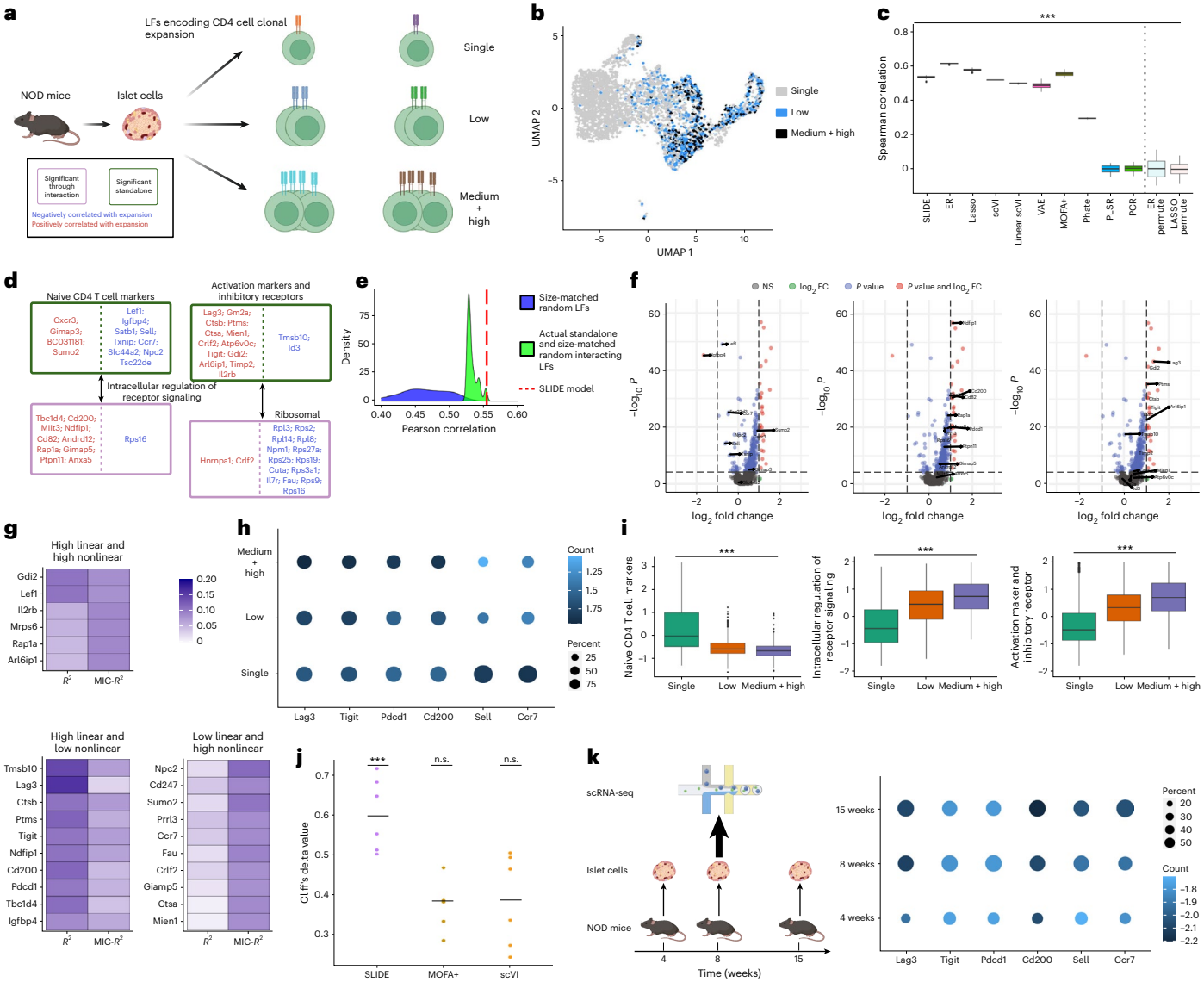


Fig. 5 | SLIDE elucidates novel interacting latent factors underlying the clonal expansion of CD4 T cells in T1D. a, Schematic summarizing scRNA-seq and TCR-seq data from NOD mice used to infer mechanisms underlying clonal expansion of CD4 T cells. **b**, UMAP visualization of the three stages of clonal expansion. **c**, Spearman correlations between true stage of clonal expansion and stage of clonal expansion predicted using different methods—SLIDE (spec = 0.1), ER, LASSO, PLS, PCR and PHATE–regression. Model performance is measured across 50 replicates of fivefold cross-validation with permutation testing. $^{**}P < 0.01$. **d**, Significant interacting latent factors (LFs) identified by SLIDE. Green, significant standalone latent factors; purple, significant interacting latent factors. **e**, Performance of the real model (spec = 0.1) relative to null models as described in Fig. 2e. **f**, Volcano plots illustrating genes in the significant latent factors. Highlighted genes indicate members in latent factors identified by the SLIDE model. *P* values from a Wald test. **g**, Linear Spearman correlations and nonlinear relationships (quantified using MIC) between key components of the latent factors and extent of clonal expansion. FC, fold change. **h**, Dot plots illustrating frequency (circle size) and median expression (color intensity) of well-known markers of T cell activation, exhaustion and inhibitory receptors at the three stages of clonal expansion. Frequency/expression calculated using data from our study. **i**, Box plots illustrating the distributions of each SLIDE latent factor (across) cells at the three different stages of clonal expansion. Kruskal–Wallis test is performed to calculate *P* values. $^{***}P < 0.01$. **j**, Effect sizes of the SLIDE latent factors from **d** (excluding ribosomal) and top-sized matched MOFA+ and scVI latent factors in stratifying CD4 T cells by extent of clonal expansion. *P* value is calculated using a Mann–Whitney *U* test. The null distribution is built with random size-matched nonsignificant SLIDE latent factors. $^{***}P < 0.01$. n.s., not significant. **k**, Dot plots illustrating frequency (circle size) and median expression (color intensity) of well-known markers of T cell activation, exhaustion and inhibitory receptors at the three stages of clonal expansion. Frequency/expression calculated using data from Unanue and colleagues. For box plots, the box spans from the first to the third quartile, and the whiskers extend from the first quartile −1.5 interquartile range (IQR) to the third quartile +1.5 IQR.

latent factors with shuffled interactors (Fig. 4o). Overall, SLIDE works very well across a wide variety of spatial datasets.

SLIDE elucidates novel interacting latent factors underlying clonal expansion in T1D

Finally, we sought to analyze paired multiomic datasets using SLIDE and uncover interacting latent factors underlying clonal expansion in type-1 diabetes (T1D). Using paired scRNA-seq and T-cell receptor sequencing (TCR-seq) data on islet-derived CD4 T cells in a nonobese diabetic (NOD) mouse model, we labeled cells (Fig. 5a) on the basis of their clonal expansion levels—single (1 clone), low (2–10 clones) or medium/high (>10 clones). Here we use paired multiomic data drawn from different distributions and examine the ability of SLIDE to identify factors underlying a cellular phenotype at single-cell resolution.

While transcriptomic profiles showed differences between cells at different stages of clonal expansion (Fig. 5b), there is significant

intragroup heterogeneity in the profiles of individual cells (Fig. 5b). SLIDE transcends this heterogeneity and accurately predicts and infers extent of clonal expansion, outperforming several benchmarks including PLS, PCR and PHATE–regression in terms of prediction accuracy (Fig. 5c and Supplementary Fig. 6a–c). LASSO, ER, scVI[39], VAE and MOFA+–regression had comparable prediction performance (Fig. 5c and Supplementary Fig. 6c); however, SLIDE provides the same predictive power as ER but stronger inference with fewer latent factors (Supplementary Fig. 6d).

The SLIDE latent factors, relative to the other methods that have similar prediction performance, also provided more meaningful inference of states underlying the extent of clonal expansion. The four latent factors uncovered by SLIDE included (1) markers of naive CD4 T cells, (2) activation markers and inhibitory receptors, (3) a latent factor that captured intracellular regulation of receptor signaling and (4) ribosomal proteins (Fig. 5d). As earlier, we tested if the interactions of (1) with (3) and (2) with (4) provided better prediction and additional inference that (1) and (2) alone would not provide. As expected, the actual latent factors significantly outperformed the random size-matched set at different stability parameters for the selection of significant latent factors via the repeated application of knockoffs (Fig. 5e and Supplementary Fig. 6e). When keeping the two standalone latent factors fixed but shuffling the interactors, the performance of this model significantly dropped (Fig. 5e and Supplementary Fig. 6e), highlighting the importance of correct interacting latent factors for prediction and corresponding inference of mechanisms underlying clonal expansion. These four significant latent factors also capture both linear and nonlinear relationships. (Supplementary Fig. 6f).

Importantly, the four significant latent factors included well-known inhibitory receptors and markers of clonal expansion/exhaustion including Lag3, Pdcd1 (Pd1) and Tigit[40] that standard DE analyses would have picked up (Fig. 5f). As expected, SLIDE grouped these inhibitory receptors in one latent factor. Interestingly, the intracellular signaling regulation latent factor also contained Ndfip1 (ref. 41), which was shown to induce apoptosis in self-reactive T cells. The association of other potential mediators of apoptosis such as Anxa5 (ref. 42), suggests a different pathway of action than the inhibitory receptor latent factor. Compared to DE analysis, SLIDE's grouping of genes with convergent functions led to the identification of inhibitory receptors and intracellular restriction on proliferation as two parallel mechanisms in clonally expanded T cells.

SLIDE also identified novel markers that standard DE analyses would have missed (Fig. 5g). Of particular interest is Ccr7, which is elevated in naive T cells, as well as memory T cells. Co-expression of Ccr7 with Sell and Lef1 are hallmarks of naive T cells[43], confirming that unexpanded $CD4^+$ T cells are naive in their phenotype. Overall, the significant interacting latent factors encapsulate additive effects and are far better at capturing the molecular basis of clonal expansion than individual canonical markers, which show weak univariate trends (Fig. 5h,i). SLIDE latent factors were also significantly better than the MOFA+ and sCVI latent factors in stratifying by extent of clonal expansion (Fig. 5j and Supplementary Fig. 6g–k). Thus, corresponding biological inference, both at the level of individual gene and latent factors, is superior to existing approaches.

To further validate and contextualize discoveries made by SLIDE, we analyzed scRNA-seq data from an independent recent study[44] on T1D disease progression markers. Several key markers aligned with our results (Fig. 5k) confirming the robustness of our findings. Importantly, some differences arose since our markers reflect the extent of clonal expansion while the study identified markers of T1D disease progression. Overall, SLIDE can identify highly context-specific markers of clonal expansion of CD4 T cells in a NOD model of T1D.

Discussion

With a surge in technologies for deep profiling, there is a deluge of high-dimensional datasets quantifying multiscale multimodal responses. Most current methods, such as black-box deep learning approaches or classification/regression techniques, focus primarily on prediction. Thus, they are useful in predictive contexts but do not offer insights into actual mechanisms of complex molecular, cellular or organismal phenotypes.

To address these key challenges, we present SLIDE, an interpretable latent factor regression-based machine learning approach for ubiquitous biological discovery from high-dimensional multiomic datasets. SLIDE incorporates nonlinear relationships and comes with rigorous guarantees regarding identifiability of the latent factors and corresponding inference. These give SLIDE a significant edge over other modern techniques (for example, VAEs) that incorporate nonlinear relationships, but are sensitive to parameter initialization and often get stuck with local minima[1]. Further, current state-of-the-art methods (for example, MOFA+) provide discovery of latent factors with no FDR control, while SLIDE creatively adapts knockoffs for rigorous FDR control. SLIDE is also compatible with different preprocessing and batch-effect correction methods and/or technological-platform-specific analysis tools, because it makes no assumptions regarding data-generating mechanisms (the input data has to be continuous, but there are no other distributional assumptions). SLIDE provides inference in addition to, and not at the cost of, predictive performance. SLIDE is currently limited to inference solely from the input data, but future approaches may further improve inference via the incorporation of prior knowledge.

Critically, SLIDE infers context-specific groups (latent factors). Canonical biomarker approaches do not have group information at all, while pathway-centric approaches have group information that is context independent and irrelevant in specific scenarios. SLIDE's context-specific group inference provides accurate guidance for downstream analyses and experimentation. Thus, SLIDE is a first-in-class interpretable machine learning framework for biological discovery.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-024-02175-z.

References

1. Altman, N. & Krzywinski, M. Regression diagnostics. *Nat. Methods* **13**, 385–386 (2016).
2. Peddireddy, S. P. et al. Antibodies targeting conserved non-canonical antigens and endemic coronaviruses associate with favorable outcomes in severe COVID-19. *Cell Rep.* **39**, 111020 (2022).
3. Das, J. et al. Delayed fractional dosing with RTS,S/AS01 improves humoral immunity to malaria via a balance of polyfunctional NANP6- and Pf16-specific antibodies. *Medicine* **2**, 1269–1286 e1269 (2021).
4. Suscovich, T. J. et al. Mapping functional humoral correlates of protection against malaria challenge following RTS,S/AS01 vaccination. *Sci. Transl. Med.* **12**, eab4757 (2020).
5. Lu, L. L. et al. Antibody Fc glycosylation discriminates between latent and active tuberculosis. *J. Infect. Dis.* **13**, 2093–2102 (2020).
6. Ackerman, M. E. et al. Route of immunization defines multiple mechanisms of vaccine-mediated protection against SIV. *Nat. Med.* **24**, 1590–1598 (2018).
7. Das, J. et al. Mining for humoral correlates of HIV control and latent reservoir size. *PLoS Pathog.* **16**, e1008868 (2020).
8. Li, S. et al. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nat. Immunol.* **15**, 195–204 (2014).
9. Vafaee, F. et al. A data-driven, knowledge-based approach to biomarker discovery: application to circulating microRNA markers of colorectal cancer prognosis. *NPJ Syst. Biol. Appl* **4**, 20 (2018).

10. Nakaya, H. I. et al. Systems biology of vaccination for seasonal influenza in humans. *Nat. Immunol.* **12**, 786–795 (2011).
11. Bzdok, D., Altman, N. & Krzywinski, M. Statistics versus machine learning. *Nat. Methods* **15**, 233–234 (2018).
12. Bing, X. et al. Essential regression: a generalizable framework for inferring causal latent factors from multi-omic datasets. *Patterns* **3**, 100473 (2022).
13. Bing, X., Bunea, F., Royer, M. & Das, J. Latent model-based clustering for biological discovery. *iScience* **14**, 125–135 (2019).
14. Barber, R. F. & Candés, E. J. Controlling the false discovery rate via knockoffs. *Ann. Stat.* **43**, 2055–2085 (2015).
15. Tabib, T. et al. Myofibroblast transcriptome indicates SFRP2^{hi} fibroblast progenitors in systemic sclerosis skin. *Nat. Commun.* **12**, 4384 (2021).
16. Stifano, G. et al. Skin gene expression is prognostic for the trajectory of skin disease in patients with diffuse cutaneous systemic sclerosis. *Arthritis Rheumatol.* **70**, 912–919 (2018).
17. Nazari, B. et al. Altered dermal fibroblasts in systemic sclerosis display podoplanin and CD90. *Am. J. Pathol.* **186**, 2650–2664 (2016).
18. Bhattacharyya, S. et al. Tenascin-C drives persistence of organ fibrosis. *Nat. Commun.* **7**, 11703 (2016).
19. Rice, L. M. et al. A longitudinal biomarker for the extent of skin disease in patients with diffuse cutaneous systemic sclerosis. *Arthritis Rheumatol.* **67**, 3004–3015 (2015).
20. Farina, G., Lafyatis, D., Lemaire, R. & Lafyatis, R. A four-gene biomarker predicts skin disease in patients with diffuse cutaneous systemic sclerosis. *Arthritis Rheum.* **62**, 580–588 (2010).
21. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **58**, 267–288 (1996).
22. Boulesteix, A. L. & Strimmer, K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Brief. Bioinform.* **8**, 32–44 (2007).
23. Bair, E., Hastie, T., Paul, D. & Tibshirani, R. Prediction by supervised principal components. *J. Am. Stat. Assoc.* **101**, 119–137 (2006).
24. Xue, D. et al. Expansion of fcγ receptor IIIa-positive macrophages, Ficolin 1-positive monocyte-derived dendritic cells, and plasmacytoid dendritic cells associated with severe skin disease in systemic sclerosis. *Arthritis Rheumatol.* **74**, 329–341 (2022).
25. Argelaguet, R. et al. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* **21**, 111 (2020).
26. Moon, K. R. et al. Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* **37**, 1482–1492 (2019).
27. Berkowitz, J. S. et al. Cell type-specific biomarkers of systemic sclerosis disease severity capture cell-intrinsic and cell-extrinsic circuits. *Arthritis Rheumatol.* **75**, 1819–1830 (2023).
28. Gourh, P. et al. HLA and autoantibodies define scleroderma subtypes and risk in African and European Americans and suggest a role for molecular mimicry. *Proc. Natl Acad. Sci. USA* **117**, 552–562 (2020).
29. Apostolidis, S. A. et al. Single cell RNA sequencing identifies HSPG2 and APLNR as markers of endothelial cell injury in systemic sclerosis skin. *Front. Immunol.* **9**, 2191 (2018).
30. Wu, M. et al. Identification of cadherin 11 as a mediator of dermal fibrosis and possible role in systemic sclerosis. *Arthritis Rheumatol.* **66**, 1010–1021 (2014).
31. Khanna, D. et al. Tofacitinib blocks IFN-regulated biomarker genes in skin fibroblasts and keratinocytes in a systemic sclerosis trial. *JCI Insight* **7**, e159566 (2022).
32. Gregory, L. G. & Lloyd, C. M. Orchestrating house dust mite-associated allergy in the lung. *Trends Immunol.* **32**, 402–411 (2011).
33. He, K. et al. Blimp-1 is essential for allergen-induced asthma and Th2 cell development in the lung. *J. Exp. Med.* **217**, e20190742 (2020).
34. Rodriques, S. G. et al. SLIDE-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).
35. Zhang, M. et al. Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature* **598**, 137–143 (2021).
36. Goltsev, Y. et al. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell* **174**, 968–981 e915 (2018).
37. Zhang, X. & Kohl, J. A complex role for complement in allergic asthma. *Expert Rev. Clin. Immunol.* **6**, 269–277 (2010).
38. Nobs, S. P. et al. PPARγ in dendritic cells and T cells drives pathogenic type-2 effector responses in lung inflammation. *J. Exp. Med.* **214**, 3015–3035 (2017).
39. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
40. Blank, C. U. et al. Defining 'T cell exhaustion'. *Nat. Rev. Immunol.* **19**, 665–674 (2019).
41. Altin, J. A. et al. Ndfip1 mediates peripheral tolerance to self and exogenous antigen by inducing cell cycle exit in responding CD4^{+} T cells. *Proc. Natl Acad. Sci. USA* **111**, 2067–2074 (2014).
42. Hu, Z. et al. Annexin A5 is essential for PKCθ translocation during T-cell activation. *J. Biol. Chem.* **295**, 14214–14221 (2020).
43. Szabo, P. A. et al. Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nat. Commun.* **10**, 4706 (2019).
44. Zakharov, P. N., Hu, H., Wan, X. & Unanue, E. R. Single-cell RNA sequencing of murine islets shows high cellular complexity at all stages of autoimmune diabetes. *J. Exp. Med.* **217**, e20192362 (2020).

Methods

Research reported in the manuscript complies with all ethical regulations.

Initial LFD framework

The same latent factor discovery (LFD) framework is employed for all analyses. We first performed FDR thresholding on the covariance matrix, followed by the optimization of two key hyperparameters (delta and lambda) using *k*-fold cross-validation. The delta parameter controls the number of latent factors, while the lambda parameter has an impact on the sparsity of the allocation matrix of latent factors. Considering the large and continuous search space of delta and lambda, we perform the search in multiple steps and ranges. The first step of the framework performs a coarse grid search of delta in four different numerical ranges: 0–0.001, 0.001–0.01, 0.01–0.1 and 0.1–1. The subsequent models utilize the most optimal delta within each range and identify the final delta using cross-validation and permutation testing. Once an optimal delta has been identified, lambda is tuned using a coarse grid search coupled to cross-validation and permutation testing.

Simulations

The simulations aimed to evaluate model performance on different values of feature size (p) and sample size (n) with and without interaction terms. The numbers of latent factors (K), significant standalone and significant interacting latent factors were set 100, 2 and 5, respectively. For the models with no interactions, the number of interacting latent factors was set to 0. For varying features, we fixed the number of observations at 300. For simulations that have a varying n, we fixed the number of features to 1,000. To generate the simulated data, R package mvnorm was utilized to randomly generate datapoints K times, where K represents the number of latent factors.

For the model with interaction terms, the dependent variable is generated using equation (2) with coefficients generated randomly from standard normal distributions.

Analyses of transcriptomic profiles in patients with SSc

We analyzed scRNA-seq data from 24 previously described patients with SSc[15,24,27] where 10 of these patients were treated with tofacitnib for 24 weeks. Standard 10x Genomics sequencing pipeline was used including cellranger. The aligned samples were then normalized and clustered with Seurat. scRNA-seq data from 24 patients with SSc all using the same V2 chemistry was processed using a standard analytic pipeline. This consisted of alignment via cellranger and dimensionality reduction and clustering using Seurat (Methods). We used R (version 4.00) for the data analysis. Seurat (R package version 3.2.2) was adopted for cell population identification and visualization. To transform the 24 scRNA-seq samples into pseudo-bulk format, we utilized the 'Average Expression Function' to reduce the cell dimension by calculating the average expression of each gene across all cells in each cell-type-specific cluster per patient. Clusters with fewer than 20 cells on average across all patients were excluded. Eighteen clusters passed the filtration process, and the top 50 highest variance genes across patients for each cluster were chosen in an unsupervised manner as features for the subsequent analysis. For each of the retained clusters, we tried a range of feature engineering approaches and converged on a cell-type-specific pseudo-bulk average of the most variable genes, where variance was calculated in an unsupervised fashion without using the disease severity labels in any form. We utilized cell-type-specific pseudo-bulk average of the most variable genes as input features while keeping all preprocessing steps unsupervised and do not use the MRSS scores.

Post preprocessing, for the 24 untreated samples, the input data for the analysis are a sample by gene matrix with dimensions of 24 by 804. Using the LFD framework with tenfold cross-validation and 20 replicates, optimal delta and lambda values were identified (Supplementary Fig. 1a,b). With optimal parameter values set at 0.6 and 1 for delta and lambda, respectively, the final model produced 120 latent factors. We then used SLIDE to identify the significant standalone latent factors using the iterative knockoff procedure as described above. Corresponding parameters of spec (a frequency-based parameter to quantify the stability of stages 1 and 2 of the multistage knockoff approach), FDR and F (feature split size) are set to 0.3, 0.1 and 100, respectively. The analysis resulted in five significant standalone factors and four significant interacting latent factors. Fivefold cross-validation with 50 replicates was used to compare the predictive power SLIDE with ER, LASSO, PHATE, PLSR and PCR. Since PHATE is an unsupervised approach, we coupled it to a standard regression model (that is, we ran regression on PHATE1 and PHATE2). Glmnet and PhateR packages were used to build the LASSO and PHATE models, respectively. The PLS package was employed for PCR and PLSR model construction. Additionally, to implement MOFA+ and VAE, we utilized the Python packages MOFA2, and Keras, respectively, throughout the study. All these methods were implemented using the R interpreter package reticulate. For VAEs in Keras, we used a sigmoid activation function in the encoder and rectified linear unit (RELU) activation functions for the decoder. The number of latent factors in VAE were adjusted to match SLIDE. For MOFA+, we used the default settings, and a similar size-matched number of latent factors.

We used the Spearman correlation and maximal information criterion (MIC) to evaluate the linear and nonlinear relationships of important genes with MRSS, respectively. These important genes have high loadings that correspond to the significant latent factors obtained from the SLIDE model. The effect sizes of each significant latent factor chosen by SLIDE are quantified by Spearman correlation coefficient between the latent factors and MRSS (Fig. 2k). P value is calculated by constructing a null distribution as follows. Random nonsignificant latent factors, size-matched with the SLIDE output, were chosen 30 times followed by Spearman correlation calculation. A Mann–Whitney U test is performed to calculate the P value between the real correlation coefficients and the null correlation coefficients.

For the tofa-treated SSc analysis, post quality control, normalization and clustering using Seurat, we matched the cluster identities to the untreated samples. To transform the scRNA-seq dataset into pseudo-bulk format, we used the same averaging calculation mentioned above and feature matched with the untreated analysis resulted in the input sample by gene matrix with dimensions of 10 and 728. Using the LFD framework, and leave-one-out cross-validation (LOOCV), a delta of 0.009 (175 latent factors) and a lambda of 1 were identified as optimal hyperparameters. SLIDE was then applied to identify significant standalone and interacting latent factors. The SLIDE parameters of spec, FDR and F were set to 0.3, 0.1 and 100, respectively. The final SLIDE model produced one significant standalone latent factor and three significant interacting latent factors.

Analyses of spatial 10X Visium (RNA-seq) data in lymph nodes

Twenty-five micrograms of LPS-low HDM (Stallergenes-Greer Pharmaceuticals) in 25 µl of sterile 1× phosphate-buffered saline was delivered intranasally under anesthesia to C57BL/6 mice (Jax Laboratory, male mice 6–8 weeks) daily for 3 days. mLNs were isolated on day 4, snap frozen and embedded in chilled optimal cutting temperature compound (Tissue-Tek) on dry ice, and stored at −80 °C. mLN samples were cryosectioned (10 µm) at −20 °C on a cryostat (Leica) and mounted directly onto the 6.5 × 6.5 mm capture areas of a single Visium Spatial Gene Expression slide (10x Genomics). The slides were sealed in individual 50 ml Falcon tubes at −80 °C until further processing according to the manufacturer's protocol (10x Genomics). The Visium Spatial Tissue Optimization Slide & Reagent kit (10x Genomics) was used to determine optimal permeabilization timing of 18 min. Immunofluorescence staining was done using 'Methanol Fixation, Immunofluorescence Staining & Imaging for Visium Spatial Protocols (CG000312)'.

Slides were stained with anti-B220 eFluor 450 (RA3-6B2, eBioscience), anti-CD4 Alexa Fluor 488 (G.K.1.5, BioLegend) and anti-CD11c Biotin (N418, BioLegend) followed by secondary detection with streptavidin Alexa Fluor-647. Images were acquired using EVOS M7000 Imaging System (AMF7000) under the Visium assay mode. Tissue sections were then permeabilized, and messenger RNA molecules within cells captured by poly (dT) sequence on the slide surface, followed by on slide reverse transcription to generate complementary DNA. cDNA was amplified and further processed into sequencing libraries according to the manufacturer's protocol (10x Genomics). Libraries were sequenced on an Illumina Nextseq2000 at 50,000 read pairs per spot covered by tissue. Sequencing results were initially processed by spaceranger (10x Genomics) to align sequencing data with the image. Here we filtered the genes more than 900 zeros threshold, which resulted in matrix of 1,932 genes by 3,779 regions.

Seurat was employed for the quality control, normalization and clustering analyses. By overlaying the uniform manifold approximation and projection (UMAP) clusters and the fluorescent microscopy plot (Fig. 3b,i), cell types predominant in each cluster were identified. Genes that are not expressed in at least 900 regions were filtered to control the sparsity. Using the LFD framework with tenfold cross-validation and 20 replications the optimal value for delta parameter was obtained as 0.049 with 21 latent variables (Supplementary Fig. 2a,b). SLIDE was applied to identify factors underlying immune cell partitioning by spatial localization. We set an FDR threshold at 0.1 for each knockoff replicate and $F = 21$ for this dataset. Two significant standalone latent factors and five significant interacting latent factors were identified.

The effect sizes of the significant latent factors identified by SLIDE in partitioning by the spatial localization of immune cells are quantified by Cliff's delta values (Fig. 3h). Cliff's delta calculation is performed utilizing the R library, effsize. *P* value is calculated by constructing a null distribution as follows. Random nonsignificant latent factors, size-matched with the SLIDE output, were chosen 30 times followed by Cliff's delta calculation. Mann–Whitney–Wilcoxon test is performed to calculate the *P* value between the real Cliff's deltas and the null Cliff's deltas.

Analyses of spatial Slide-seq data in lymph nodes

Curio Seeker tile was removed and placed in a 1.5-ml Eppendorf LoBind tube with Hybridization Reaction Mix. Reverse transcription was performed to generate cDNA followed by tissue clearing and Curio Seeker Bead resuspension. Second strand synthesis was done followed by cDNA amplification, which after purifying was subjected to tagmentation (Nextera XT) and library generation. Libraries were sequenced on an Illumina Nextseq2000 at 200 M reads per tile. Demultiplexed FastQ files were initially processed by the Curio Seeker bioinformatics pipeline.

Seurat was employed for the quality control, normalization and clustering analyses. By overlaying the UMAP clusters and the fluorescent microscopy plot, predominant cell types in each cluster were identified. Genes that are not expressed in at least 600 regions were filtered to control the sparsity. After applying the filtering, the final matrix consisted of 11,421 regions and 3,851 genes. Using the LFD framework with tenfold cross-validation and 20 replications the optimal value for delta parameter was obtained as 0.6755 and lambda value of 1 with 47 latent variables (Supplementary Fig. 5a,b). SLIDE was applied to identify factors underline immune cell partitioning by spatial localization. For this dataset, we set an FDR threshold of 0.1 for each knockoff replicate and performed SLIDE with a spec of 0.5 and $F = 47$. As a result, we identified two significant standalone latent factors and three significant interacting latent factors. For VAEs in Keras, we used a sigmoid activation function in the encoder and RELU activation functions for the decoder. The number of latent factors is size matched with that of SLIDE latent factors. For the MOFA+, we used the default setting of the software. The libraries and the packages that were used for this analysis are the same as previous analysis. The effect sizes of the significant latent factors from SLIDE in stratifying by spatial localization are quantified by Cliff's delta values as described above.

Analysis of spatial transcriptomics data generated from MERFISH imaging

We first used the LFD framework to identify latent factors. The input data consisted of a cell-by-gene matrix, comprising 16,200 cells and 241 genes. Parameter tuning was performed via tenfold cross-validation with 20 replicates. The LFD framework, utilizing a delta value of 0.109 and lambda value of 1, ultimately identified 31 latent factors.

Subsequently, SLIDE was applied to discern the significant interacting latent factors that underlie the differences in subclasses of the glutamatergic neurons in the mouse primary motor cortex. In this analysis, we set the SLIDE parameters as follows: spec at 0.2, FDR at 0.1 and feature partition size at 31. This configuration led to the identification of two significant standalone latent factors and seven significant interacting latent factors. For VAEs in Keras, we used a sigmoid activation function in the encoder and RELU activation functions for the decoder. The number of latent factors in VAE is adjusted to match that of SLIDE. For MOFA+, we initially used default parameters followed by tuning of the number of latent factors. To perform benchmarking across methods, we utilized the same packages as in the previous analysis, ensuring consistency and comparability across the evaluations.

Analysis of spatial proteomics data generated by CODEX

We first applied the LFD framework to uncover latent factors. The dataset consisted of the cell-by-protein matrix comprising 10,000 cells and 30 proteins. The parameter tuning LFD framework was conducted via tenfold cross-validation with 20 replicates. Ultimately, utilizing a delta value of 0.081 and a lambda value of 0.5, the LFD framework identified a total of nine latent factors as the final model. For VAEs in Keras, we used a sigmoid activation function in the encoder and RELU activation functions for the decoder. The number of latent factors in VAE is adjusted to match that of SLIDE. For MOFA+, we initially used default parameters followed by tuning of the number of latent factors.

Next, SLIDE was employed to discern significant interacting latent factors. With the SLIDE parameter configuration set at spec 0.5, FDR 0.1 and a feature partition size of 9, the analysis identified two significant standalone latent factors and three significant interacting latent factors.

scRNA-seq and TCR-seq of islet infiltrating CD4 T cells

NOD mice (6-, 8- or 10-week-old female NOD/ShiLtJ) were euthanized by CO_2 asphyxiation and immediately dissected for pancreas perfusion. Pancreas perfusion was performed under a dissecting Zeiss microscope. Pancreatic duct was clamped using surgical clamps and 3 ml of 600 U ml^{-1} collagenase dissolved in Hank's Balanced Salt Solution (HBSS) was injected using a 30 G needle. Perfused pancreata were collected and incubated at 37 °C for 30 min. After the incubation, HBSS with R10 was added to quench collagenase. After washing twice with HBSS + R10, the tissue was plated on a 10-cm plate, individual islets were picked using a micropipettor. Islets were then incubated in dissociation buffer, centrifuged and resuspended in the staining mix (1:500 dilution of anti-Thy1.2-BV605 + 1:500 dilution of Live/Dead-APC-Cy7, and 1:100 dilution of cell hashing anti-mCD45 TotalSeq-C antibodies (BioLegend)). After staining, the cells were resuspended in phosphate-buffered saline + 0.04% bovine serum albumin and sorted on BD FACS Aria III sorter. After sorting the cells, they were counted and processed for scRNA-seq. Cells were processing using 10 × 5′ single cell gene expression kit v3 in a Chromium controller according to the manufacturer's protocols. V(D)J enrichment was done using the single-cell 5′ VDJ enrichment kit according to the manufacturer's protocols. Libraries were sequenced on HiSeq4000 (Novogene) with a 70:20:10 mix for gene expression:VDJ:hashing libraries.

Sequence data were downloaded and aligned to the mouse genome (Mm10) using cellranger (10x Genomics). TCR annotation was performed using cellranger vdj using mouse GRCm38 assembly. All three time points were sequenced and processed separately. Cellranger and cellranger vdj output files were used as inputs in Seurat for normalization, scaling, and dimensionality reduction. The packaged scRepertoire was used for TCR clonotype calling and analyses. The data were normalized using NormalizeData and scaled using ScaleData functions in Seurat. The scRepertoire functions combineTCR and combineExpression were used to add TCR clonotypes to each cell. HTODemux function in Seurat was used to demultiplex cell hashes and assign the correct mouse identity to each cell. At this point, all three time points were merged in Seurat using the merge function. After merging, integration was done using FindIntegrationAnchors and IntegrateData functions. Principle component analysis was performed using RunPCA. Top 20 principal components were used for UMAP, followed by cluster identification using FindNeighbors and FindClusters. $CD4^+$ T cells were subsetted using FeatureScatter and CellSelector functions, and reclustered. Cluster markers were defined by FindAllMarkers function. Clonotype data were sorted according to expansion and exported as a csv file. UMAP representations with clonotypes were generated using highlightClonotypes function in scRepertoire. Differentially expressed genes were identified using FindMarkers function using DESeq2 statistics and represented using EnhancedVolcano function. After obtaining scRNA-seq and TCR-seq data on islet-derived cells in a NOD mouse model, analysis was done through the standard 10x Genomics pipeline. We labeled cells based on their clonal expansion stages followed by the postprocessing of the scRNA-seq data in R (4.1.0) using Seurat (R package, version 3.2.2). The columns representing genes and rows representing cells are filtered on the basis of 1,200 threshold, meaning that if the sparsity exceeds 1,200, the cell row or gene column will be removed. The SLIDE input matrix was finalized with 1,776 genes and 2,482 cells.

The LFD framework is first utilized to discover latent factors. The input data is a cell by gene matrix, consisting of 2,484 cells and 1,776 genes. As described previously, tenfold cross-validation with 20 replications was performed for optimal parameter tuning (Supplementary Fig. 6a,b). The final model constructed by the LFD framework using delta as 0.0912 and lambda as 1 discovered 40 latent factors. We then performed SLIDE to identify significant interacting latent factors underlying differences in clonal expansion in CD4 T cells. We set the SLIDE parameter spec at 0.2, FDR at 0.1 and feature partition size at 40, resulting in the identification of two significant standalone and two significant interacting latent factors. For VAEs in Keras, we used a sigmoid activation function in the encoder and RELU activation functions for the decoder. The number of latent factors in VAE is adjusted to match that of SLIDE. As for MOFA+, we used the default setting, and the number of latent factors was fine-tuned within the software. In addition to MOFA+ and VAE, we also performed linear and nonlinear scVI analyses using the Python package scvi-tools and the R interpreter reticulate. We used scvi.model.LinearSCVI and scvi.model.SCVI functions for linear and nonlinear models, respectively. The effect sizes of the significant latent factors identified by SLIDE in stratifying by the extent of clonal expansions are quantified by Cliff's delta values as described above.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data including the SSc scRNA-seq, 10X Visium, Slide-seq, CD4 T cell scRNA-seq and TCR-seq datasets and associated documentation are available at https://github.com/jishnu-lab/SLIDE and at https://github.com/jishnu-lab/SLIDEpre. Corresponding stable releases are available at https://doi.org/10.5281/zenodo.10159961 and https://doi.org/10.5281/zenodo.10159957, respectively. The relevant datasets have also been deposited at the Gene Expression Omnibus (accession IDs: GSE245112 and GSE247410 for the spatial and T1D datasets, respectively). Source data are provided with this paper.

Code availability

All code and documentation is available at https://github.com/jishnu-lab/SLIDE and at https://github.com/jishnu-lab/SLIDEpre. Corresponding stable releases are available at https://doi.org/10.5281/zenodo.10159961 and https://doi.org/10.5281/zenodo.10159957, respectively.

Acknowledgements

J.D. was supported in part by NIAID DP2AI164325, NIAID R01AI170108 and NHGRI U01HG012041. The authors acknowledge support from the University of Pittsburgh Center for Research Computing through the high-performance computing resources provided. The authors acknowledge all members of the Das lab for helpful discussions.

Author contributions

J.D. conceived of the project and supervised all aspects. X.B., F.B. and M.W. developed the theoretical foundations of the method. J.R., H.X., A.R. and A.B.I.R. implemented SLIDE. R.A.L. designed and assembled the SSc cohort; T.T. carried out the corresponding scRNA-seq experiments. A.V.J. designed the T1D scRNA-seq/TCR-seq experiments, which were executed by P.M.Z. A.C.P. designed the 10X Visium and Slide-seq experiments, which were carried out by K.H. J.D. designed all computational analyses which were carried out by J.R. and H.X. J.R., H.X. and J.D. interpreted results with inputs from R.A.L., A.V.J. and A.C.P. J.D., J.R. and H.X. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-024-02175-z.

Correspondence and requests for materials should be addressed to Amanda C. Poholek, Alok V. Joglekar, Robert A. Lafyatis or Jishnu Das.

Peer review information *Nature Methods* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling editor: Madhura Mukhopadhyay, in collaboration with the *Nature Methods* team. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Jishnu Das, Robert Lafyatis, Alok Joglekar and Amanda Poholek

Last updated by author(s): 11/19/2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Please do not complete any field with "not applicable" or n/a. Refer to the help text for what text to use if an item is not relevant to your study.
For final submission: please carefully check your responses for accuracy; you will not be able to make changes later.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Not applicable.
Data analysis	All code and documentation is available at https://github.com/jishnu-lab/SLIDEpre and at https://github.com/jishnu-lab/SLIDE. Corresponding stable releases are available at https://zenodo.org/doi/10.5281/zenodo.10159957 and https://zenodo.org/doi/10.5281/zenodo.10159961 respectively.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All data including the SSc scRNA-seq, 10X Visium, Slide-seq, CD4 T cell scRNA-seq and TCR-seq datasets and associated documentation are available at https://github.com/jishnu-lab/SLIDEpre and at https://github.com/jishnu-lab/SLIDE.
Corresponding stable releases are available at https://zenodo.org/doi/10.5281/zenodo.10159957 and https://zenodo.org/doi/10.5281/zenodo.10159961 respectively.
The relevant datasets have also been deposited at the Gene Expression Omnibus (accession ids: GSE245112, GSE247410 for the spatial and T1D datasets respectively).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Not applicable
Reporting on race, ethnicity, or other socially relevant groupings	Not applicable
Population characteristics	Not applicable
Recruitment	Not applicable
Ethics oversight	Not applicable

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Samples sizes were not determined using a-priori power calculations.
Data exclusions	No data were excluded.
Replication	Most computational analyses were performed across 50 replicates of 5-fold cross-validation. Other experiments were at least in triplicate. Additional details in the manuscript.
Randomization	Splits for 5-fold cross-validation were randomized.
Blinding	All experiments were conducted blinded to experimental group/label.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	
Research sample	
Sampling strategy	
Data collection	
Timing	
Data exclusions	
Non-participation	
Randomization	

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	
Research sample	
Sampling strategy	
Data collection	
Timing and spatial scale	
Data exclusions	
Reproducibility	
Randomization	
Blinding	

Did the study involve field work? ☐ Yes ☐ No

Field work, collection and transport

Field conditions	
Location	
Access & import/export	
Disturbance	

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☐	☒ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	anti-B220 eFluor 450 (RA3-6B2, eBioscience), anti-CD4 Alexa Fluor 488 (N418, Biolegend), anti-CD11c Biotin (N418, Biolegend), streptavidin Alexa Fluor-647
Validation	All antibodies have been validated by the manufacturer as described on their website.

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	
Authentication	
Mycoplasma contamination	
Commonly misidentified lines (See ICLAC register)	

Palaeontology and Archaeology

Specimen provenance	
Specimen deposition	
Dating methods	

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	For spatial experiments, male C57/BL6J mice (6-8 weeks) were used. For the clonal expansion analyses, 6, 8, or 10 week old female NOD/ShiLtJ mice were used.
Wild animals	No wild animals were used in this study.
Reporting on sex	Both sexes were considered in method design and there are no relevant differences based on sex.
Field-collected samples	No field-collected samples were used in this study.
Ethics oversight	All animal studies were done in accordance with the Institutional Animal Care and Use Committee Protocol Number 21028806 and 23022626.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	Not applicable. We used scRNA-seq data from previously described samples as noted in the manuscript.
Study protocol	Not applicable. We used scRNA-seq data from previously described samples as noted in the manuscript.
Data collection	Not applicable. We used scRNA-seq data from previously described samples as noted in the manuscript.
Outcomes	Not applicable. We used scRNA-seq data from previously described samples as noted in the manuscript.

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☐	☐	Public health
☐	☐	National security
☐	☐	Crops and/or livestock
☐	☐	Ecosystems
☐	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☐	☐	Demonstrate how to render a vaccine ineffective
☐	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☐	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☐	☐	Increase transmissibility of a pathogen
☐	☐	Alter the host range of a pathogen
☐	☐	Enable evasion of diagnostic/detection modalities
☐	☐	Enable the weaponization of a biological agent or toxin
☐	☐	Any other potentially harmful combination of experiments and agents

Plants

Seed stocks

Novel plant genotypes

Authentication

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

Files in database submission

Genome browser session
(e.g. UCSC)

Methodology

Replicates

Sequencing depth

Antibodies

Peak calling parameters

Data quality

Software

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Instrument

Software

Cell population abundance

Gating strategy

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type

Design specifications

Behavioral performance measures

Imaging type(s)

Field strength

Sequence & imaging parameters

Area of acquisition

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software

Normalization

Normalization template

Noise and artifact removal

Volume censoring

Statistical modeling & inference

Model type and settings

Effect(s) tested

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference

(See Eklund et al. 2016)

Correction

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Graph analysis

Multivariate modeling and predictive analysis