#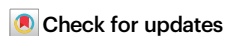 HiDDEN: a machine learning method for detection of disease-relevant populations in case-control single-cell transcriptomics data

Aleksandrina Goeva [1] ✉, Michael-John Dolan[1], Judy Luu[1], Eric Garcia[1], Rebecca Boiarsky[1,2], Rajat M. Gupta [1,3] & Evan Macosko [1,4] ✉

In case-control single-cell RNA-seq studies, sample-level labels are transferred onto individual cells, labeling all case cells as affected, when in reality only a small fraction of them may actually be perturbed. Here, using simulations, we demonstrate that the standard approach to single cell analysis fails to isolate the subset of affected case cells and their markers when either the affected subset is small, or when the strength of the perturbation is mild. To address this fundamental limitation, we introduce HiDDEN, a computational method that refines the case-control labels to accurately reflect the perturbation status of each cell. We show HiDDEN's superior ability to recover biological signals missed by the standard analysis workflow in simulated ground truth datasets of cell type mixtures. When applied to a dataset of human multiple myeloma precursor conditions, HiDDEN recapitulates the expert manual annotation and discovers malignancy in early stage samples missed in the original analysis. When applied to a mouse model of demyelination, HiDDEN identifies an endothelial subpopulation playing a role in early stage blood-brain barrier dysfunction. We anticipate that HiDDEN should find wide usage in contexts that require the detection of subtle transcriptional changes in cell types across conditions.

High-dimensional transcriptional profiling of cells has enabled the comprehensive characterization of cellular changes in response to perturbations, such as disease[1–4], treatment with a drug[5], or gene knockouts[6–8]. Existing computational strategies address different aspects of this general question, each accompanied by a set of assumptions[9]. Differential expression and differential abundance approaches aim to identify changes in gene expression and cell type proportion between perturbation conditions with the caveat that their power to infer the biological alterations is compromised when the condition labels do not correctly represent the presence or absence of an effect in individual cells. For example, many perturbations only affect a subset of the cells in a given cell type while the rest of the cells are largely unaffected[10]. Condition-agnostic approaches aim to identify

perturbation-affected groups of neighboring cells within the latent space, which may be clouded by the presence of several additional axes of biological or technical variation[11–13] making it challenging to tease out the perturbation-relevant signal.

Detecting cell-level transcriptional changes across experimental conditions is one of the big promises of high-resolution single-cell expression data[9]. In recent years, several methods have been proposed to characterize perturbation effects in single-cell data. The standard analysis workflow performs label-agnostic dimensionality reduction and clustering, followed by comparisons of cell attributes across condition labels within clusters. CNA[12] provides a cluster-free approach to identifying regions in the latent space of uneven mixing of condition labels. MELD[14] produces a continuous measure of the perturbation

[1]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA, USA. [2]CSAIL and IMES, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Divisions of Cardiovascular Medicine and Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [4]Massachusetts General Hospital, Department of Psychiatry, Boston, MA, USA. ✉e-mail: aleksandrina.goeva@utoronto.ca; emacosko@broadinstitute.org

effect by distributing the condition labels among neighbors in the cell state manifold. Milo[15] performs differential abundance testing among experimental conditions in the presence of continuous trajectories. Mixscape[8] removes known confounding sources of variation and dissects successfully from unsuccessfully perturbed cells in gene knockout screens where it is expected that a high proportion of cells in the case sample will be perturbed. These approaches rely on at least one of the following assumptions: 1) the condition labels correctly represent the presence or absence of an effect in individual cells; 2) the perturbation effect is a dominant signal in the latent space; or 3) that any confounding sources of variation are known and can be removed. However, these assumptions might not always be met−often, perturbation effects are small relative to the biological heterogeneity and technical noise, or the proportion of affected cells is small and therefore the condition labels are mostly incorrect.

To address these challenges, we developed a statistical framework called HiDDEN, which refines the labels of individual cells within perturbation conditions to accurately reflect their status as affected or unaffected. We systematically generate ground truth

datasets of cell type mixtures and demonstrate that HiDDEN can accurately identify marker genes from affected subpopulations of cells that are undetected by standard approaches to single-cell analysis. We used HiDDEN to recapitulate manual annotation of neoplastic cells in human multiple myeloma precursor conditions and discover malignancy in previously considered healthy early-stage samples, as well as to identify an endothelial cell subpopulation that regulates blood-brain barrier function during the early stages of demyelination in a mouse model.

Results

Overview of problem and method

In many case-control experiments, only a subset of the cells in case samples are affected by the perturbation (Fig. 1a). The standard analysis workflow of jointly clustering gene expression profiles of case and control cells can fail to distinguish affected from unaffected cells, resulting in mixed clusters (Fig. 1b) due to multiple sources of variance competing with the perturbation signal. Differential expression using the sample-level labels within a mixed cluster can fail to recover the
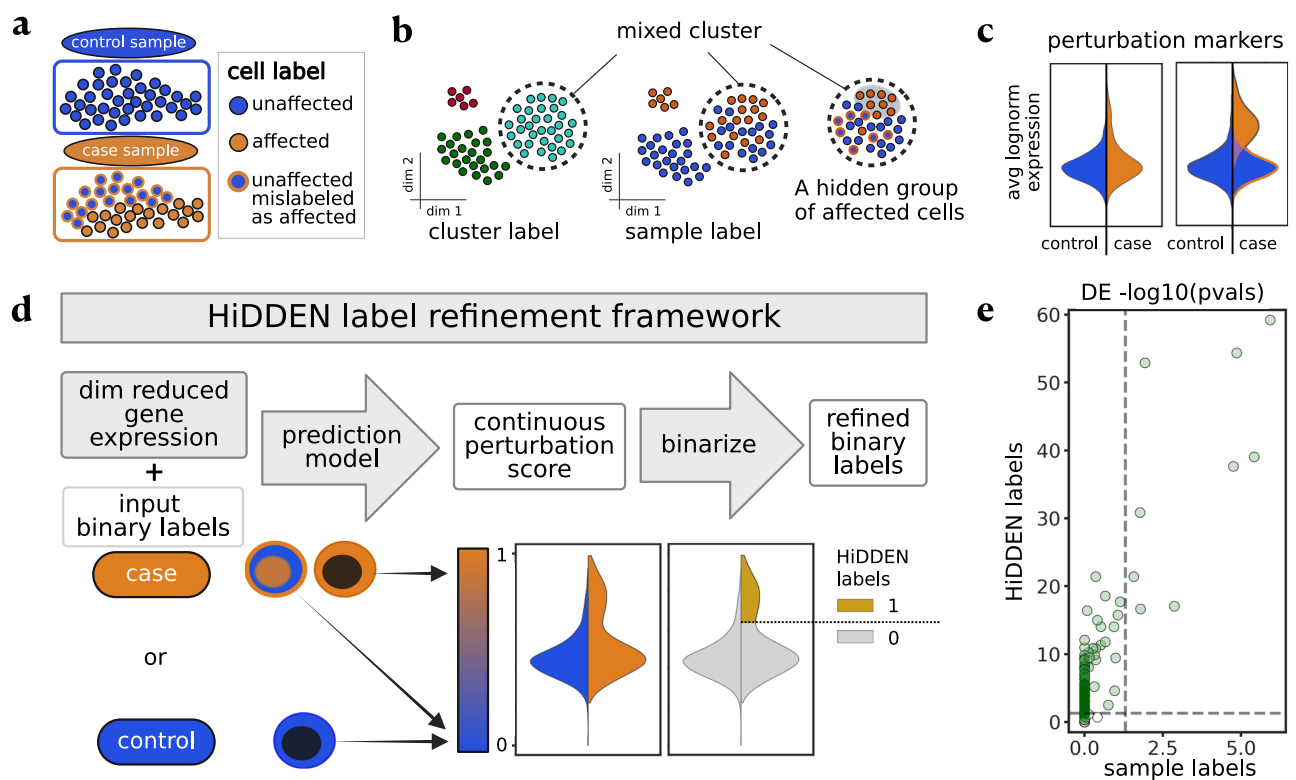

Fig. 1 | Overview of problem and HiDDEN label refinement framework. a Setup of a case-control single-cell experiment, in which cells of a given cell type in control samples are labeled as unaffected, while cells in case samples can be either affected or unaffected by the perturbation. **b** Standard clustering can produce clusters containing cells with mixed case-control sample-level labels while the subset of truly affected cells can be hidden. Colors as defined in A. **c** Representative violin plots of the average log normalized expression of perturbation markers split by sample-level labels (left) and highlighting the difference in the distributions of affected and unaffected cells within the case sample (right). Colors as defined in A. Area not scaled to count. **d** Overview of the HiDDEN label refinement framework. First, gene expression profiles are summarized through a dimensionality reduction method. Then, a prediction model takes the reduced expression profiles and the sample-level binary labels and transforms them into per cell continuous

perturbation scores. Finally, the continuous scores of cells originating from the case samples can be binarized through a classification method into HiDDEN-refined binary labels (Methods). **e** Representative scatterplot of -log10 adjusted *p* values per gene computed using differential expression (DE) on case-control sample labels (x-axis) and HiDDEN-refined binary labels (y-axis). *P* values are calculated using a one-sided Wilcoxon rank sum test with Benjamini-Hochberg correction. Horizontal and vertical dashed lines drawn at -log10(0.05) significance threshold. Ground truth DE genes colored in green. Standard DE analysis on case-control labels captures only a small number of ground truth markers, while HiDDEN successfully recovers many of them. Figure 1 panel **a** Created in BioRender. Lab, M. (2024) BioRender.com/f66p361. Figure 1 panel B created in BioRender. Lab, M. (2024) BioRender.com/j24d711. Figure 1 panel **d** created in BioRender. Lab, M. (2024) BioRender.com/z12o210.
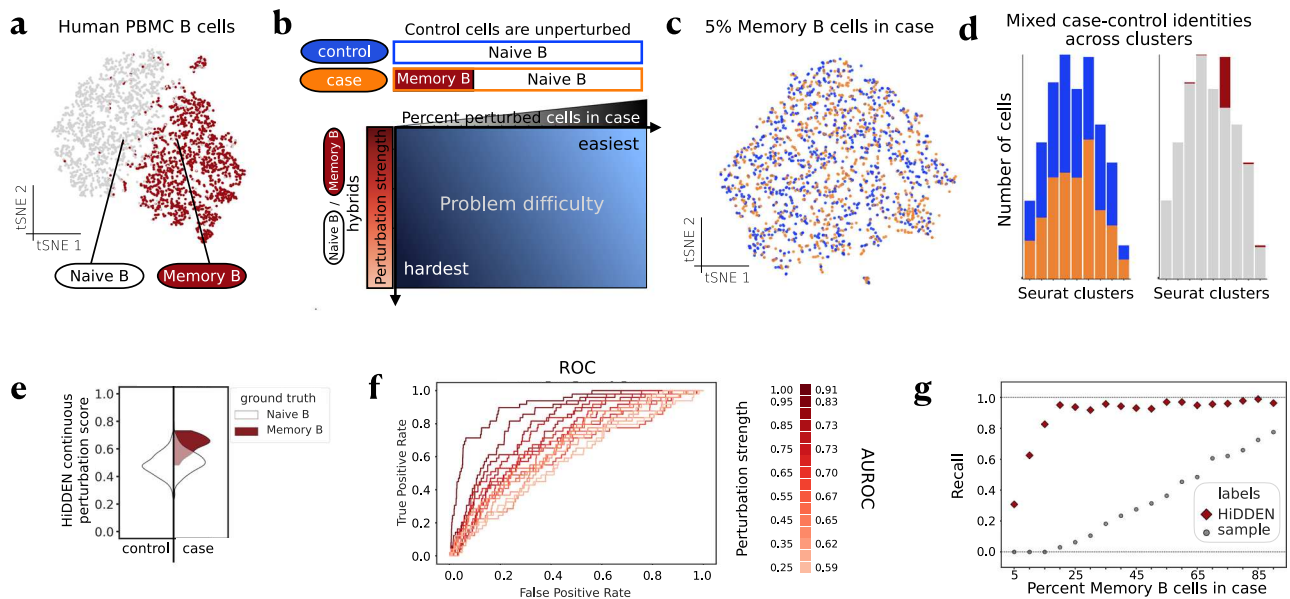
Fig. 2 | HiDDEN detects biological signal missed by the standard analysis workflow in simulated ground truth mixtures of two cell types. a tSNE embeddings of gene expression of Naive B and Memory B cells. **b** Schematic of problem difficulty and definition of synthetic datasets along two axes: percent perturbed cells in case sample (x-axis) and strength of the perturbation (y-axis). Detecting the perturbation is most challenging when there are few affected cells and the difference between affected and unaffected cells is small. **c** tSNE embeddings of a representative simulated dataset containing 5% Memory B cells in the case sample. Cells colored by case-control labels. Colors as defined in B. **d** Distribution of case-control (left) and Memory B-Naive B (right) cell identities across Seurat clusters. Colors as defined in A and B. **e** Violin plot of the distributions of the continuous perturbation score of Naive B and Memory B cells split over control and case and colored by ground truth labels, for the dataset containing 5% Memory B cells in the case sample. **f** Area under the Receiver Operating Characteristic (AUROC) curves for classification of ground truth cell labels as a function of perturbation strength for the dataset containing 5% perturbed cells in the case sample with the AUROC indicated in the legend for a sampling of the curves. **g** Recall of ground truth DE genes by DE testing on indicated labels as a function of percent Memory B cells in case sample. Source data of (**a, c, d, e, f, g**) are provided as a Source Data file.

perturbation markers due to the incorrect labels decreasing detection power (Fig. 1c).

The standard analysis of single-cell data is not tailored to identifying perturbation-associated signals. However, combining gene expression profiles and sample-level labels in a novel way allows us to leverage that at least some of the labels are correct and empowers HiDDEN to utilize the shared variability in features corresponding to correctly labeled cells. HiDDEN transforms the sample-level labels into cell-specific continuous perturbation-effect scores and assigns new binary cell labels, revealing their status as affected or unaffected (Fig. 1d, Methods). The resulting binary labels can accurately capture the perturbation signature and boost power to detect genes whose expression is affected by the perturbation (Fig. 1e).

HiDDEN detects biological signal missed by the standard analysis workflow in simulated ground truth datasets of cell-type mixtures

To simulate the biological change in cell function induced by a perturbation, we conducted simulations using the single-cell RNA-seq profiles of Naive B and Memory B cells from a dataset of peripheral blood mononuclear cells (PBMC)[16] (Fig. 2a, Methods). Naive B and Memory B cells have relatively similar expression profiles but with biologically relevant differences[17,18], making them suitable for modeling perturbation-induced changes. To mimic the outcome of a perturbation experiment, we constructed a control sample consisting of Naive B (representing unperturbed) cells and a case sample consisting of both Naive B and Memory B (representing perturbed) cells (Fig. 2b, Methods). We observed that, as Memory B and Naive B cells became increasingly imbalanced, the ability of a commonly used single-cell analysis pipeline (Methods) to identify the Memory B cluster became impaired. For example, having 5% Memory B cells in the case condition results in a highly heterogeneous latent space produced by the

standard dimensionality reduction workflow, making it impossible to detect a locus of perturbed cells (Fig. 2c). Indeed, using a standard clustering workflow with default parameter values (Methods) fails to recover a cluster that purely represents the Memory B labeled cells (Fig. 2d). Exposing the ground truth labels reveals that even the cluster with the highest enrichment of case-labeled cells contains a majority of Naive B cells (Fig. 2d). The recovery of the Memory B cluster could not be improved by varying the number of principal components (PCs) used to construct the latent space, adjusting the resolution parameter of the clustering algorithm, or varying the gene selection, including by utilizing Naive B and Memory B markers in lieu of highly variable genes (Supplementary Fig. 1). By contrast, HiDDEN was far better able to identify the Memory B signature within this artificial mixture and distinguish Memory B from Naive B cells (Fig. 2e).

This example reveals a more general feature of how cell types are detected in single-cell data. To comprehensively characterize the problem difficulty and assess the power of our method to detect the perturbation signal, we constructed a collection of ground truth case-control datasets by varying two key aspects (Fig. 2b, Methods). First, to study the effect of perturbation strength, we defined perturbed cells as hybrids of Naive B and Memory B cells of variable relative weight (Methods). Decreasing the strength of the transcriptional difference between perturbed and unperturbed cells increased the difficulty of identifying the perturbed cell cluster (Fig. 2f, Supplementary Fig. 2, Methods). However, even when only 5% of the case cells are even slightly perturbed, HiDDEN continuous perturbation scores identified the biological differences between perturbed and unperturbed cells with high accuracy (Supplementary Fig. 2a). Second, to explore the influence of class imbalance, we varied the percent of perturbed cells in the case sample (Supplementary Table 1). Strikingly, in datasets with fewer than 20% Memory B cells, using the standard analysis pipeline with the sample-level labels completely failed to retrieve any of the Naive B and

Memory B marker genes, and overall retrieved only a fraction even in datasets with high sample label accuracy (Fig. 2g, Methods). The marker gene recovery was not improved even when we considered the union of case-control label-derived Differentially Expressed (DE) genes per cluster since the reduction in number of cells dramatically hinders the power of DE testing to recapitulate the markers (Supplementary Fig. 3, Methods). By contrast, the HiDDEN-refined binary labels had superior power to detect the ground truth markers. Furthermore, HiDDEN-refined labels appeared to provide accuracy beyond the ground truth labels for this dataset. Specifically, the genes identified by DE testing on HiDDEN-refined labels, but not by DE testing on ground truth labels, identified additional genes that are consistent with markers of Naive B and Memory B cells (Supplementary Fig. 4), suggesting that HiDDEN labels possess corrective power for a slight amount of misclassification that might have occurred in the original annotation.

Dimensionality reduction is a key component of the HiDDEN analysis framework that defines the input features of the label-prediction model (Fig. 1d). We provide a collection of dimensionality reduction approaches that the user can select from, or a pre-computed embedding can be plugged in. To examine the performance of different dimensionality reduction strategies in our Naive B / Memory B ground truth datasets, we compared approaches ranging from linear methods to deep-learning alternatives and found that a simple dimensionality reduction method performs as well as or better than an auto-encoder (Supplementary Fig. 5, Methods).

Several computational methods have recently been proposed to characterize perturbation effects in single-cell data, each designed to tackle a particular aspect of this general problem. Specifically, CNA[12] provides cluster-free detection of perturbation-affected areas of the latent space, while MELD[14] offers the identification of a perturbation gradient. A third method, Milo[15], performs differential abundance testing over continuous trajectories. Mixscape[8] identifies cells that have escaped a gene knockout perturbation in pooled CRISPR screens. When applied to our target task of refining the sample-level case status into perturbed and unperturbed cell labels, HiDDEN continuous perturbation scores and binary-refined labels outperformed the corresponding continuous and binarized scores from CNA, MELD, Milo, and Mixscape across ground truth Naive B / Memory B mixtures (Supplementary Fig. 6, Methods). Each of these methods relies on veritable cell-level labels as input, such that HiDDEN-refined labels could augment their respective performances. Indeed, when HiDDEN is applied first, there was an improvement in the performance of CNA, MELD, and Milo-derived continuous scores (Supplementary Fig. 7, Methods) and binarized labels (Supplementary Fig. 8, Methods) to recover the ground truth perturbation labels.

The HiDDEN method has a single model parameter: the number of features in the predictive model. To explore how this parameter affects the stability and accuracy of HiDDEN results, we generated two heuristics for automatically choosing it and demonstrated that either strategy works well, indicating that the parameter is not especially influential for model performance (Supplementary Fig. 9, Methods).

HiDDEN recapitulates manual annotation of neoplastic cells in human multiple myeloma precursor conditions and discovers malignancy in previously considered healthy early-stage samples

To test the ability of our method to capture perturbation signal in a real dataset, we applied HiDDEN to single-cell RNA-seq profiles of human bone marrow plasma cells from patients with multiple myeloma (MM), its precursor conditions smoldering multiple myeloma (SMM) and monoclonal gammopathy of undetermined significance

(MGUS), and healthy donors with normal bone marrow (NBM)[4] (Methods). Precursor samples can contain a mixture of neoplastic and normal cells (Fig. 3a) and the authors of the original study defined two orthogonal strategies for describing the malignancy status of precursor samples and their cells. The first strategy is a per-sample computational analysis excluding immunoglobulin light chain genes followed by manual annotation resulting in binary labels defining cells as healthy or malignant. The second strategy is a tumor-purity estimate of the proportion of malignant cells in each precursor sample from a model based on the distribution of immunoglobulin gene expression.

According to the manual annotation, three MGUS and five SMM samples contain a mixture of malignant and healthy cells. Application of HiDDEN continuous perturbation scores to distinguish cells in these mixed samples showed remarkable agreement with this manual annotation (Fig. 3b, Supplementary Fig. 10, Methods). Furthermore, sample purity estimates derived from HiDDEN binary labels agreed with their corresponding point estimates of malignant cell proportions and outperformed manual annotation-based estimates in the majority of mixed precursor samples (Fig. 3c, Methods, Supplementary Table 2).

We next turned our attention to the three MGUS samples with the lowest tumor purity. In those samples, the manual annotation strategy failed to identify any neoplastic cells. By contrast, HiDDEN was able to discover malignant cells in these early-stage patients that were missed by the manual annotation (Fig. 3e, Methods, Supplementary Table 2). To computationally validate that these were indeed malignant cells, we assessed whether the distinguishing genes of these cells matched with known signatures of healthy plasma and malignancy. Indeed, previously described gene signatures distinguishing normal from malignant cells[4] were heavily differentially enriched between the HiDDEN-defined normal and malignant cells in these samples (Fig. 3f, Supplementary Fig. 11, Methods).

Of note, HiDDEN successfully recapitulates both types of malignancy estimates in the presence of pronounced patient-specific batch effects in this dataset (Supplementary Fig. 12a). Mirroring the analysis in the original study, we deployed a batch-sensitive strategy to fitting the HiDDEN model, namely training it on all NBM, all MM, and one precursor sample at a time. Additionally, we also developed a batch-agnostic strategy, where we fit all samples together (Methods). HiDDEN outputs under both strategies were closely aligned and almost indistinguishable (Supplementary Fig. 12B-F). We provide a heuristic to automatically choose the optimal number of features used in the training of the prediction model and demonstrate that HiDDEN outputs were closely aligned and almost indistinguishable across a wide range of values for the tunable model parameter (Supplementary Fig. 13, Methods).

We leveraged our refined definition of healthy and neoplastic cells in precursor states to derive markers of early disease. We find a total of 8208 differentially expressed genes, 2400 of which significantly overlap (hypergeometric test, p-value = 3.066e-31) with basic malignancy markers derived from a comparison of healthy and multiple myeloma patients, and 5808 of which are uniquely found using the HiDDEN-refined labels in precursor samples (Fig. 3d, Methods).

HiDDEN identifies an endothelial subpopulation affected in the early stages of demyelination

To explore HiDDEN's ability to identify rare, subtle perturbations, we applied the method to single-nucleus RNA-seq (snRNA-seq) profiles from a time-resolved dataset of a mouse model of demyelination[19,20] (Fig. 4a, Methods). In this experiment, case animals received a corpus callosum injection containing lysophosphatidylcholine (LPC), a compound toxic to oligodendrocytes, while control animals are injected with saline (PBS) (Methods). LPC induces white matter loss, demyelination, which is rapidly repaired in a stereotyped manner over three weeks. Several cell types showed dramatic changes in response to this
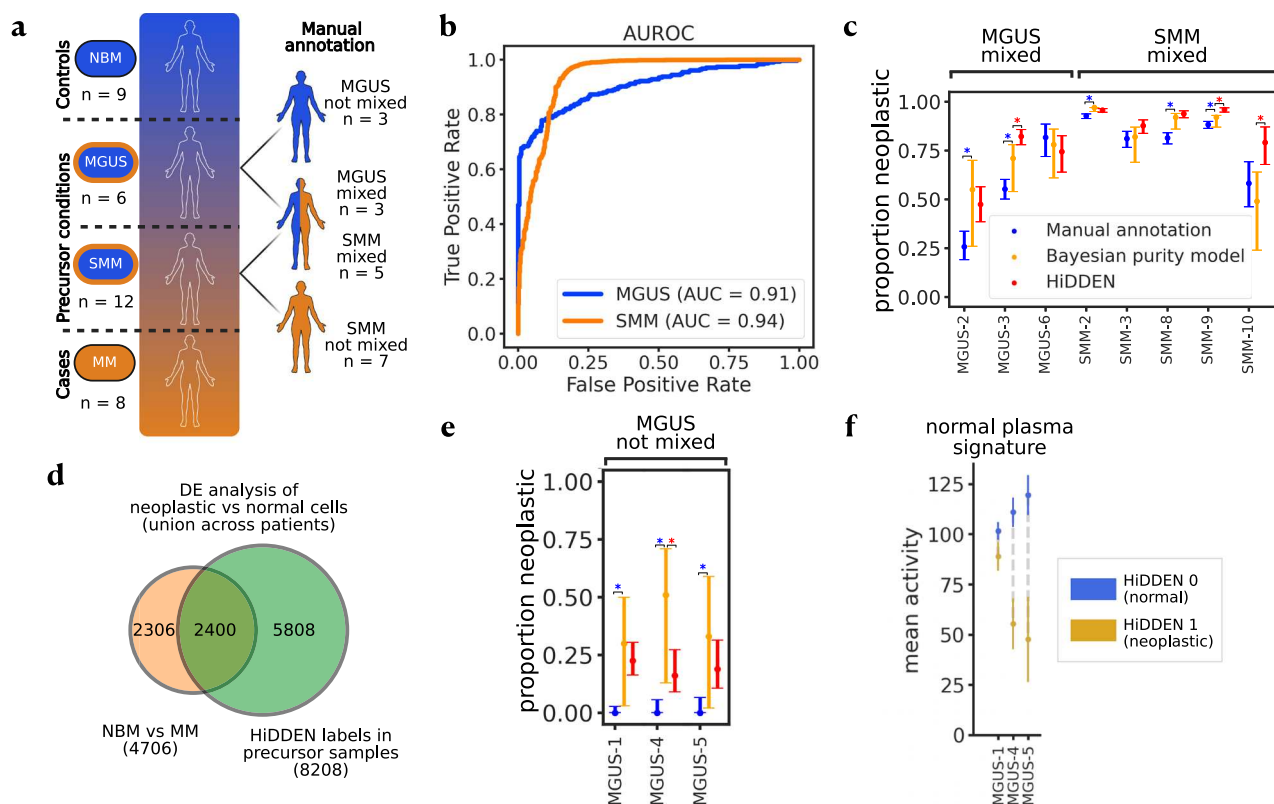
Fig. 3 | Application of HiDDEN to a human bone marrow dataset with previously published annotations. a Schematic of the dataset[4] which includes human plasma cells from healthy donors, multiple myeloma patients, and two precursor states. Precursor samples possibly contain a mixture of healthy and malignant cells. Manual annotation of healthy and malignant cells per precursor patient were reported previously[4]. **b** AUROC for predicting per-cell malignancy status in mixed samples averaged for each precursor state. **c** Comparison of manual annotation, Bayesian purity model, and HiDDEN predictions for estimating the neoplastic proportion (y-axis) in mixed MGUS and SMM samples (x-axis). Data are presented as an estimated proportion of neoplastic cells with 95% confidence intervals computed over $n = 116$ cells for MGUS-2, $n = 321$ for MGUS-3, $n = 82$ for MGUS-6, $n = 1857$ for SMM-2, $n = 349$ for SMM-3, $n = 711$ for SMM-8, $n = 1253$ for SMM-9, and $n = 67$ for SMM-10. Significance for testing the difference between manual annotation and HiDDEN-based estimate with the Bayesian ground truth point estimate is calculated using a two-sided Beta-Binomial test (Methods), indicated with an asterisk for Bonferroni-adjusted p values < 0.01. Exact p-values are reported in the Source Data file and Supplementary Table 2. **d** Venn diagram of DE genes comparing neoplastic

with normal cells based on NBM/MM samples and HiDDEN refined labels in precursor samples identified 2400 significantly overlapping genes (one-sided hypergeometric test, p-value = 3.066e-31) and 5808 genes uniquely found using HiDDEN. **e** Comparison of manual annotation, Bayesian purity model, and HiDDEN predictions for estimating the neoplastic proportion (y-axis) in non-mixed MGUS samples (x-axis). Colors as defined in **c**. Data are presented as estimated proportion of neoplastic cells of each sample with 95% confidence intervals computed over $n = 133$ cells for MGUS-1, $n = 62$ for MGUS-4, and $n = 53$ for MGUS-5. Significance is established using the same approach as in **c**. Exact p values are reported in the Source Data file and Supplementary Table 2. **f** Computational validation of cells predicted to be malignant by HiDDEN in low purity MGUS samples. Mean activity ± SEM (y-axis) of genes assigned to a normal plasma signature for the normal and abnormal populations within each sample (x-axis). SEM is computed over $n = 30$ normal and $n = 103$ neoplastic cells for MGUS-1, $n = 10$ normal and $n = 52$ neoplastic cells for MGUS-4, and $n = 10$ normal and $n = 43$ neoplastic cells for MGUS-5. Figure 3 panel **a** created in BioRender. Lab, M. (2024) BioRender.com/x13c819. Source data of (**b, c, d, e, f**) are provided as a Source Data file.

injury (Supplementary Fig. 14a) as expected[21–23], but the effects on endothelial cells (ECs) appeared modest (Fig. 4b,c). As vascular cells of the brain, ECs play critical roles in homeostasis, myelin formation and tissue repair, but the altered genes and pathways underlying these functions in demyelination are poorly understood[24,25].

We first examined ECs during remyelination using an existing analytic pipeline[26]. The standard dimensionality reduction workflow produced a homogeneous distribution of sample-level labels in the latent space (Fig. 4c, Methods), and clustering failed to identify a perturbation-enriched subpopulation (Fig. 4d, e, Methods). The case-control identities were similarly mixed across time points (Supplementary Fig. 14c). By contrast, fitting HiDDEN to the ECs across all time points (Methods) generated a bimodal distribution of continuous perturbation scores for case cells at the earliest time point, suggesting an underlying mixture of affected and unaffected cells (Fig. 4f, Supplementary Fig. 14d). We split the bimodal distribution of continuous scores and used the resulting binary cell-labels to define demyelination-affected and unaffected EC subpopulations, denoted

LPC1 and LPC0, respectively (Fig. 4g, Methods). Together this demonstrates that HiDDEN can reveal demyelination-specific effects on endothelial cells not apparent using conventional analysis approaches.

We next analyzed the differential response of LPC1 and LPC0 ECs to demyelination. The LPC1 subpopulation was characterized by 28 unique markers (Fig. 5a, Supplementary Table 3, Methods), a subset of which we experimentally validated to be lesion-specific with in-situ hybridization at the 3dpi timepoint (Fig. 5b). To understand the biological functions of these changes, we applied Gene Set Enrichment Analysis. This revealed the observed changes in gene expression were consistent with alterations that occur in the context of inflammation and demyelination, such as increased angiogenesis[27], blood-brain barrier breakdown[28,29], and increased production of extracellular matrix components (Fig. 5c, d, Methods). Together, this suggests that the LPC1 EC subset, revealed by the HiDDEN method, have altered core endothelial functions specifically during the early stages of white matter damage.
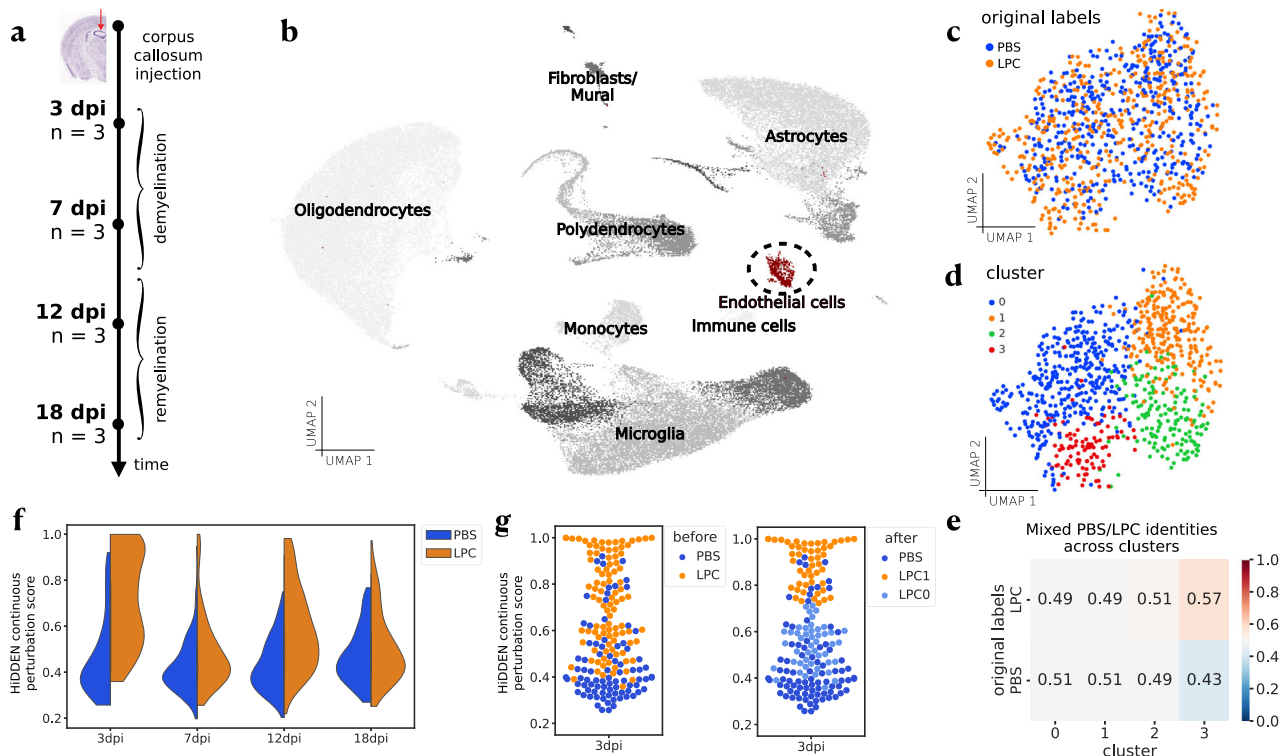
Fig. 4 | Application of HiDDEN to ECs from a mouse demyelination time-course experiment. a Overview of experimental design. Corpus callosum injection with saline (PBS) and a compound toxic to oligodendrocytes (LPC) used to induce demyelination with *n* = 3 mice per condition per time point across four time points. **b** UMAP embeddings of non-neuronal cells from PBS (control) and LPC conditions across all time points colored by annotation of major cell type. ECs highlighted in red. UMAP embeddings of ECs across all time points colored by PBS/LPC sample-level labels (**c**), and Seurat cluster labels (**d**). **e** Relative abundance of case-control cell identities across Seurat clusters. **f** Violin plots of HiDDEN continuous perturbation scores split over PBS and LPC labels and grouped by time point. **g** Swarmplot of HiDDEN continuous perturbation scores for the 3dpi cells colored by original PBS/LPC labels (left) and with color indicating the refinement of LPC cells into affected (LPC1) and unaffected (LPC0) (right). Source data of (**b, c, d, e, f, g**) are provided as a Source Data file.

The blood-brain barrier is an active hub for cell-cell interactions between the ECs comprising blood vessel walls and the surrounding cell types. In particular, endothelial-endothelial, endothelial-fibroblast, and endothelial-astrocyte interactions are crucial in tightly regulating blood-brain barrier permeability[30] and have been implicated in neurodegenerative disease pathogenesis[29]. To investigate if these intercellular pathways could be dysregulated during a demyelinating event, we examined differences in cellular communication of the affected LPC1 and unaffected LPC0 endothelial subpopulations with neighboring cell types, we used a computational method for targeted hypothesis testing of ligand-receptor expression (Fig. 5e, Supplementary Fig. 15, Methods). The changes in communication were consistent with increased angiogenesis, blood-brain barrier breakdown, and increased extracellular matrix. The anti-angiogenic interactions of *Flt1* with *Vegfa*/*Vegfb* and of *Sema3a* with *Npr2* were decreased in LPC1 endothelial cells, supporting increased angiogenesis. In addition, interactions between collagen and integrin components were increased in LPC1, pointing to the remodeling of the extracellular matrix. Interactions supporting the tight junctions between endothelial cells, such as *Jam2* and *Jam3* with integrins, were also decreased in LPC1, suggesting compromised barrier function. Furthermore, we found that LPC1 endothelial cells have an increased expression of *Vcam1*, which acts as a ligand for recruiting immune cells from the bloodstream to cross the blood-brain barrier (Fig. 5f). In summary, changes in EC function during white matter damage are poorly understood. While established approaches failed to identify changes in ECs during de- and remyelination, HiDDEN revealed a temporally-specific alteration of a subset of ECs during demyelination which likely drives blood-brain barrier breakdown and immune cell influx. As few biological contexts or perturbations are truly uniform, this illustrates the power and broad utility of HiDDEN to isolate *bonafide* effects from complex biological systems in vivo.

Discussion

With the increased amount of annotated single-cell atlases there is an increased opportunity to automate the labeling of existing cell types in novel datasets. However, when we are seeking to identify the perturbation effect in a single-cell case-control study of a novel disease or treatment, we would not have any existing annotated data to draw from. Towards this end, in this work we developed HiDDEN, a computational method for the identification of subtle perturbation effects in single-cell data. HiDDEN accurately refines the condition labels of case cells into affected and unaffected for more sensitive detection of perturbation signals. We leveraged the HiDDEN output to find hard-to-detect disease-affected subpopulations of cells and characterized their marker genes using differential expression testing. We provide a computationally efficient Python implementation of HiDDEN at https://github.com/tudaga/LabelCorrection[31], making it scalable to large datasets. At the same time, in the application to endothelial cells, we found that HiDDEN can detect subtle perturbation changes involving tens of genes even in small datasets on the order of a hundred cells. In the application to human bone marrow plasma cells, we showed that HiDDEN can be successfully applied to samples from heterogeneous conditions with pronounced batch effects without the necessity of sophisticated preprocessing or alignment. Therefore, HiDDEN has the potential to be applicable to single-cell atlases with batch effects and a high variability in retrieved numbers of cells across cell types.

The identification of phenotype-associated cells has important uses in both genomic and translational studies. In biological contexts
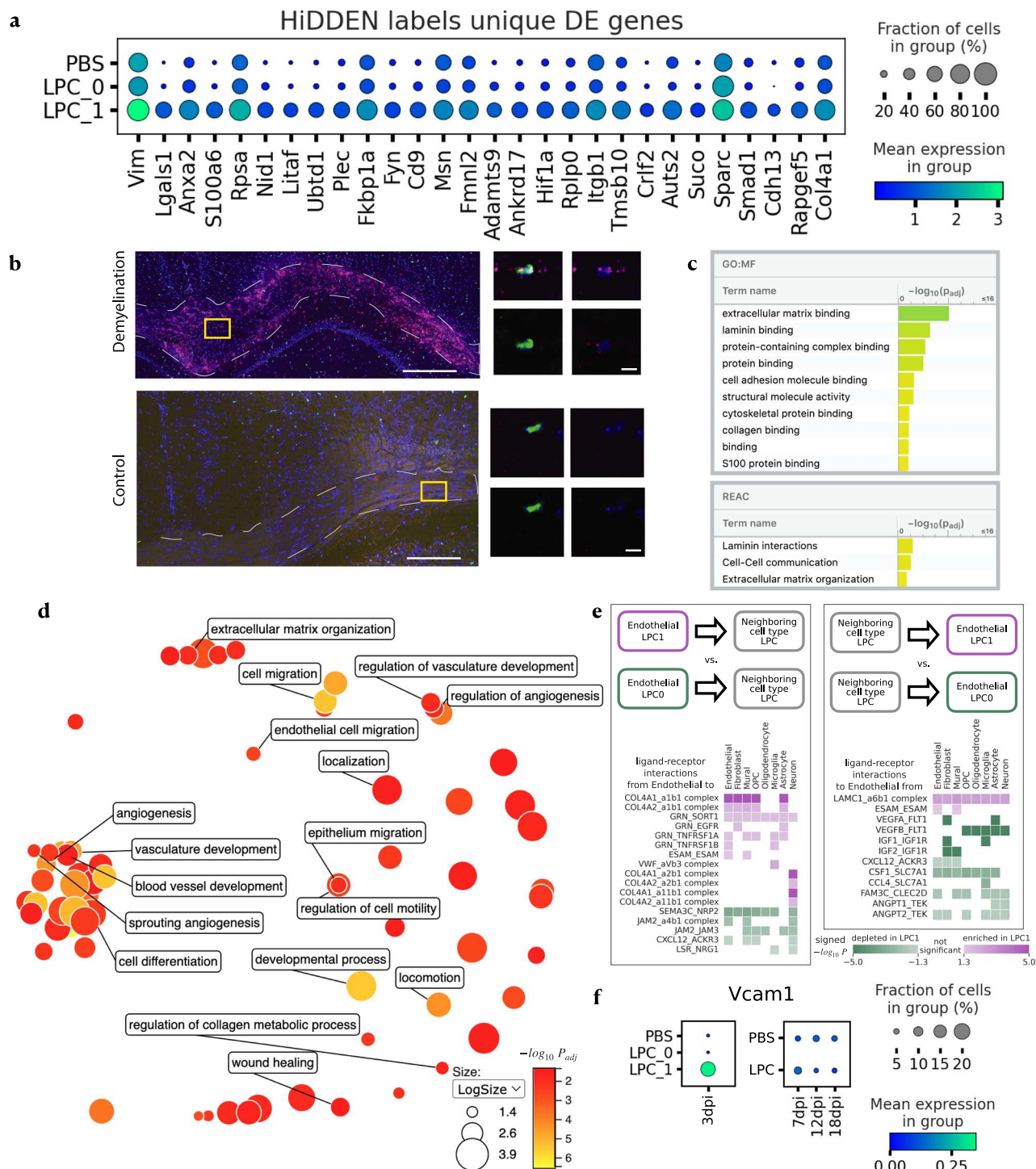
Fig. 5 | Characterization of the demyelination-affected endothelial sub-population (LPC1) identified by HiDDEN. a Dotplot of mean expression at 3 dpi of LPC1 marker genes ordered by *p*-value. *P*-values are calculated using a one-sided Wilcoxon rank sum test with Benjamini-Hochberg correction. **b** Validation of gene expression with fluorescent in-situ hybridization (FISH) and confocal microscopy showing presence of endothelial cells (*Flt1*-positive cells, green) coexpressing *Lgals1* (red) and *S100a6* (magenta) specifically present in demyelinating lesion (top) and not control (bottom) brains, 3 days after injection. Left: Overview of a demyelinating or control white matter lesion. Corpus callosum outlined in gray dashed line. Right: High-resolution confocal images of single endothelial cells. All images are representative of *n* = 2–3. Scale bars are left top = 200um, left bottom =100um, right = 10um. **c** Significantly enriched GO molecular function terms (top) and Reactome

pathways (bottom) based on LPC1 marker genes ordered by significance. The reported *p*-values are computed using the over-representation statistical test in g:Profiler and are Bonferroni-adjusted. **d** ReviGO plot summarizing the significantly enriched GO biological processes based on LPC1 marker genes colored by significance with selected labels. Significance values are computed in the same was as described in **c**. **e** Significantly enriched (purple) and depleted (green) ligand-receptor interactions between LPC1 (relative to LPC0) endothelial cells and neighboring cell types split by interaction direction: from endothelial to neighboring cell type (left), and from neighboring cell type to endothelial (right). *P* values are calculated using a one-sided permutation test. **f** Dot plots of mean expression of Vcam1 across time and condition labels. Source data of (**a, c, d, e, f**) are provided as a Source Data file.

where perturbation effects are small relative to the biological heterogeneity and technical noise, or the proportion of affected cells is small, the condition labels are mostly incorrect and a refinement of the perturbation label at the single-cell level is needed. We show that HiDDEN outperforms existing methods in producing accurate perturbation labels. Furthermore, we demonstrate that HiDDEN-refined binary labels can be used to boost the performance of existing approaches relying on cell-level labels accurately representing the presence of a perturbation effect, including CNA and MELD, as well as methods for differential abundance across conditions, such as Milo.

In this paper, we focused on applications of HiDDEN to detect the presence or absence of a disease effect at single-cell resolution. However, as a future direction, the HiDDEN framework could be applied to other challenges, in additional contexts, in which the aim is to focus the latent space on a particular distinction, for example, to explore subtle genotype effects (i.e. eQTLs) and sexual dimorphism. The HiDDEN framework is amenable to extensions to spatial and multi-omics data, as well as applications beyond a binary output, such as multi-stage disease progressions or time-course experiments, with appropriate modification to the dimensionality reduction and prediction modules of the framework.

HiDDEN has several limitations. First, given that the perturbation effect would likely differ across cell types, the method needs to be applied one cell type at a time. Second, HiDDEN can single out an affected subpopulation, but that does not imply that the perturbation effect is homogeneous amongst the affected cells. Additional downstream analyses need to be carried out to disentangle the potential presence of multiple perturbation responses within the same cell type. Third, currently we do not provide a statistical test to distinguish whether the perturbation effect is binary or if the strength of the effect forms a continuum. As a result, it is up to the researcher to interpret the distribution of the continuous perturbation scores produced by HiDDEN in case and control cells and decide whether to proceed with clustering either set.

Despite these limitations, HiDDEN is a sensitive approach to identifying perturbation effects that would otherwise be missed by existing approaches, especially when a small fraction of cells is affected or when the perturbation effect is subtle relative to naturally observed variation in single-cell RNA-seq data. As our quest for better understanding human disease at single-cell resolution continues, computational methods that can pull out hard-to-detect transcriptional changes across conditions will become central in realizing the promise of high-resolution single-cell expression data.

Methods

HiDDEN: A computational method for revealing subtle transcriptional heterogeneity and perturbation markers in case-control studies

Intuition. In a case-control experiment, typically all cells in control samples will be unaffected, and possibly only a subset of the cells in case samples will be affected by the perturbation (Fig. 1a). Using the gene expression profiles alone can fail to separate out the affected from unaffected cells (Fig. 1b). Using the sample-level labels alone can fail to recover the perturbation markers (Fig. 1c). However, combining the two allows us to leverage that at least some of the labels are correct and allows a prediction model to utilize the shared variability in features corresponding to correctly labeled cells. As a result, we transform the sample-level labels into cell-specific perturbation effect scores and can assign binary cell labels representing their status as affected or unaffected (Fig. 1d). We use this information to find hard-to-detect affected subpopulations of cells, characterize their marker genes, and contrast their cellular communication patterns with those of unaffected cells.

Notation. Let $X \epsilon R^{N \times M}$ denote the matrix containing the gene expression profiles of N cells across M genes. Let $Z \epsilon R^{N \times K}$ denote the reduced representation of the N cells in a K-dimensional latent space of features. Let $Y \epsilon \{0,1\}^N$ denote the binary vector encoding the sample-level label of each cell, where 0 stands for control and 1 stands for case. We train a predictive model denoted by $h(.)$ on the reduced representation Z and the binary sample-level labels Y. Due to the binary nature of the case-control labels Y, the predictive model is a binary classifier modeling the probability of label 1 given the input features, i.e., $P(Y=1, |, Z) = h(Z)$. We train the parameters of the classifier on a dataset of interest and denote the fitted value of $P(Y=1, |, Z)$ with $\hat{p}$. Finally, we cluster the continuous scores $\hat{p}$ to derive refined binary labels, denoted by $\widetilde{Y}$, reflecting the status of each cell as affected or unaffected by the perturbation regardless of which sample it originated from.

Construction of the latent space. We transform the $N \times M$ gene expression matrix X into an $N \times K$ matrix Z containing an information-rich reduced representation of the gene expression profile of each cell in the dataset. Throughout the applications to real data in this work we used principal component analysis (PCA) for this task, which is a commonly used dimensionality reduction technique for single-cell RNA-seq data. In principle, any other information-preserving dimensionality reduction method can be used to construct the latent features in lieu of PCA, such as non-negative matrix factorization (NMF), or the latent representations from a state-of-the-art single-cell expression autoencoder[11]. We did not find convincing evidence that using more sophisticated dimensionality reduction techniques improved model performance (Supplementary Fig. 5) and thus adopted PCA as the most computationally efficient option.

Estimation of the continuous perturbation score. For each cell, we derive a continuous score, $\hat{p} \epsilon [0,1]$, reflective of the strength of the perturbation effect on that cell relative to the rest of the cells in the dataset. Throughout this work we use logistic regression for this task. That is, the predicted probability of label 1 given the input features is given by

$$\hat{p} = P(Y=1, |, Z) = h(Z\hat{\beta}) = g^{-1}(Z\hat{\beta}) \tag{1}$$

where h is the logit link function, i.e., g is the logistic function, and $\hat{\beta}$ is the K-dimensional vector of fitted regression parameters. In principle, we can use any classifier $h: R^{N \times K} \rightarrow [0,1]^N$, including large-parameter non-linear models such as neural networks. In practice, we opted for logistic regression as a simple yet powerful model with a canonical parameter optimization routine that does not introduce additional hyperparameters, training heuristics, and increased computational resources and time demands.

Derivation of the refined binary label. Each cell in a dataset from a case-control experiment possesses a binary sample-level label reflecting whether it originated from a case or control sample. As these coarse labels do not reflect the individual cell identity of being affected or unaffected by the perturbation, we derive a new refined binary label that captures the presence or absence of a perturbation effect in each cell. When we want to distinguish between affected and unaffected cells in the case sample, we cluster the continuous perturbation scores $\hat{p}$ for all the cells with initial label 1 into two groups. The cells in the group with lower $\hat{p}$ scores receive new label $\widetilde{Y}=0$ and the cells in the group with higher $\hat{p}$ scores get a new label $\widetilde{Y}=1$ matching their old label. To do this, we use k-means clustering with $k=2$. In principle, any other clustering algorithm, such as Gaussian mixture models with two components for example, can be utilized in lieu of k-means. We compared these two clustering methods across our ground truth Naive B / Memory B mixtures and concluded that the two clustering strategies

tend to perform similarly, especially for the datasets with 15% to 75% Memory B cells in the case condition. For datasets with less than 15% Memory B cells, the Gaussian mixture model approach had overall more power to detect ground truth marker genes. However, for datasets with more than 75% Memory B cells in the case condition, k-means performed better, especially due to a lower false discovery rate.

Choosing the number of latent dimensions. When selecting K, the number of latent dimensions, our guiding principle is that K should be chosen in a data-dependent manner aiming to retain informative transcriptional heterogeneity while avoiding overfitting the not-entirely-correct sample-level labels. For example, note that using $K > \text{rank}(X)$ will yield predictions $\hat{p}$ synonymous with the sample-level labels. In practice, this implies we should choose $K \ll \min(N, M)$. Therefore, we develop two novel data-driven heuristics to quantify the amount of informative heterogeneity retained in the latent space by measuring the perturbation signal downstream of redefining the binary labels.

The first heuristic is to use the number of differentially expressed genes defined by the refined labels $\widetilde{Y}$. A large number of DE genes indicates a meaningful signal in $\widetilde{Y}$, and in turn in $\hat{p}$ and the latent space Z. Intuitively, when K is too small, the continuous scores $\hat{p}$ do not contain enough heterogeneity to yield labels $\widetilde{Y}$ that distinguish DE genes. Conversely, when K is too large, we are overfitting to the sample-level labels resulting in low power to detect the perturbation markers. To achieve an appropriate balance, we scan a range of values for K that traverses the concave relationship between K and the number of DE genes and choose the number of latent dimensions maximizing it.

The second heuristic is to use the strength of the difference between the values of $\hat{p}$ for cells in the case sample with new label $\widetilde{Y} = 1$ and new label $\widetilde{Y} = 0$. We quantify the probability that these two sets of values are drawn from the same distribution using the two-sample Kolmogorov-Smirnov (KS) test[32]. The larger the value of the KS test statistic, the more different the sample distributions of the perturbation score are between the cells predicted to be affected and unaffected. This value is generally an increasing function of K, therefore we pick the smallest value of K that maximizes the KS test statistic.

Note that we do not need access to ground truth labels of the perturbation effect for neither heuristic. When ground truth data is available, we find that the ability of the refined binary labels $\widetilde{Y}$ to represent the true perturbation effect per cell is similarly high for a wide range of values of K and that either heuristic yields a choice in that range (Supplementary Fig. 9).

Assessing performance on semi-simulated ground truth data
To demonstrate and quantify the problem difficulty and to assess the power of our method, we conducted simulations using a real single-cell dataset. We used the RNA profiles of $n = 1900$ Naive B and $n = 1630$ Memory B cells from a dataset of peripheral blood mononuclear cells (PBMC) freely available from 10x Genomics[16]. We first describe our design of simulated case-control datasets by combining the two B cell subpopulations. We then describe the challenge of separating the two subpopulations using the standard single cell clustering analysis workflow. Finally, we describe how we train our method and the metrics we use to assess its power to detect the biological signal and to compare it against related methods.

Generation of ground truth datasets. The RNA profiles of all $n = 30672$ cells in the human PBMC data from 10x Genomics were clustered and annotated independently of the ATAC-seq profiles following standard approaches detailed in the Seurat Weighted Nearest Neighbor Analysis vignette[33]. This resulted in 27 annotated cell types of

which we subset all Naive B and Memory B cells for the subsequent generation of ground truth datasets.

For the tSNE representation of all Naive B and all Memory B cells in Fig. 2a, we normalized the gene counts, performed variable gene selection, scaled the normalized counts, performed dimensionality reduction using PCA, built the nearest-neighbor graph, and ran tSNE, all with default hyperparameter values using the standard functions in Seurat v 3.2.3[26].

To comprehensively describe the problem difficulty and test the performance of our method, we constructed a collection of ground truth case-control datasets by varying two aspects (Fig. 2b). In each dataset, the control sample consists entirely of Naive B cells, which we refer to as unperturbed, whereas the case dataset consists of both unperturbed and perturbed cells, which are either Memory B or s of Naive B and Memory B cells. Each dataset is indexed by (1) the percent perturbed cells in the case sample and (2) the strength of the perturbation.

To explore the effect of the percent perturbed cells in the case sample, we randomly drew $(100 - p)\%$ of the cells in the case from the Naive B cells and $p\%$ from the Memory B cells. The remaining Naive B cells were all allocated to the control sample. We explored 18 values of the percent perturbed cells $p \in \{5, 10, 15, \ldots, 85, 90\}\%$. The resulting number of Naive B and Memory B cells across case and control per dataset is provided in Supplementary Table 1.

To explore the effect of the strength of the perturbation, we varied the extent to which perturbed and unperturbed cells in the case sample differ from each other. Let $W^{(m)}$ and $W^{(n)}$ denote the weight of Memory B and Naive B contribution, respectively, for each hybrid cell, where $W^{(m)} + W^{(n)} = 1$, and $W^{(m)}, W^{(m)} \geq 0$, i.e. $W^{(m)}$ denotes the strength of the perturbation. Let i index the perturbed (Memory B) cells in the case sample and let N_i denote the total number of UMIs in cell i. Each Memory B / Naive B hybrid cell has the same number of UMIs, N_i, as the Memory B cell it originates from, $N_i^{(m)}$ of which are subsampled from the originating Memory B cell and $N_i^{(n)}$ of which are drawn from the Naive B centroid profile, as described below.

Let $\vec{x}_i$ denote the gene expression profile of the original Memory B cell i. Let $\vec{p}_i$ denote the normalized counts, i.e., the relative proportion of counts across genes. We drew a Memory B / Naive B hybrid gene expression profile $\vec{h}_i$ by first subsampling $N_i^{(m)} = \text{ceiling}(W^{(m)} * N_i)$ counts from the original Memory B expression profile $\vec{x}_i$:

$$\vec{h_i^{(m)}} \sim \text{Multinomial}\left(N_i^{(m)}, \vec{p}_i\right),$$

where the ceiling function for any real number x, integer z, and the set of integers $\mathbb{Z}$ is defined as $\text{ceiling}(x) = \min\{z \in \mathbb{Z} | z \geq x\}$, i.e. the ceiling is the smallest integer greater than or equal to x; and Multinomial denotes the Multinomial probability distribution.

We then drew $N_i^{(n)} := N_i - N_i^{(m)}$ counts from the Naive B centroid, $\vec{p^{(n)}}$, defined as the average normalized counts $\overrightarrow{\text{tor}}$ across all Naive B cells in the dataset:

$$\vec{h_i^{(n)}} \sim \text{Multinomial}\left(N_i^{(n)}, \vec{p^{(n)}}\right),$$

and finally, summed the two count $\overrightarrow{\text{tors}}$ $\vec{h_i} = \vec{h_i^{(m)}} + \vec{h_i^{(n)}}$ to compose the hybrid profile.

We explored 17 values of the perturbation strength parameter $W^{(m)} \in \{0.25, 0.3, 0.35, \ldots, 0.95, 1\}$. Overall, spanning both the percent perturbed cells in case and the perturbation strength axes, we generated 72 datasets to characterize the problem difficulty and assess the performance of our method, as described below.

Clustering analysis of ground truth datasets. For the tSNE representation in Fig. 2c, we focused on the simulated dataset containing 5%

Memory B cells in the case. We used the standard Seurat functions and normalized the gene counts, performed variable gene selection with default parameter values, scaled the normalized counts, performed dimensionality reduction using PCA with default parameter values, built the nearest-neighbor graph with the default number of nearest neighbors, used the Leiden algorithm with the default value of the resolution parameter to find clusters, and ran tSNE. For the bar plots in Fig. 2d and Supplementary Fig. 1, we computed the abundance of case and control-labeled cells and the abundance of Naive B and Memory B cells in each cluster.

Several hyperparameters influence the clustering results, and we varied each one to study their effects on the problem difficulty. The challenge of capturing the biological signal and separating Memory B from Naive B cells using the standard pipeline lies in (1) the construction of the latent space; and (2) the resolution parameter of the clustering algorithm. To quantify the problem difficulty, we investigated the degree of separability of Naive B and Memory B cells in the latent space via the distribution of the number of Memory B nearest neighbors across Memory B cells. For a given simulated dataset, varying the number of principal components (PCs) used to build the nearest-neighbor graph impacts the separability of the latent space with including more PCs resulting in a more mixed latent space (Supplementary Fig. 1a). The choice of a feature selection strategy along with the choice of the resolution parameter of the Leiden clustering algorithm have a significant impact on the results. We observed that using highly variable gene selection with the resolution parameter chosen to yield two clusters (since we are aiming to separate two cell types) fails to isolate the Memory B cells (Supplementary Fig. 1b). In this simulated ground truth setting, we can compute the differentially expressed (DE) genes, i.e., marker genes, between the two classes and use them as the selected features. However, that choice alone is also not sufficient to yield improved clustering when using the default value of the resolution parameter (Supplementary Fig. 1c).

HiDDEN model training. The HiDDEN model training was done in python and consists of three steps: (1) we preprocess the raw gene expression counts, (2) we train a logistic regression model, and (3) we binarize the predictions for cells in the case sample. First, we followed the standard preprocessing routine for single-cell RNA-seq data in scanpy[26] which consists of filtering out cells with <20 genes and filtering out genes not expressed in any cells, followed by log-normalization, and then, we used the standard PCA dimensionality reduction routine in scanpy on the scaled gene features. For the comparison of the choice of dimensionality reduction technique in Supplementary Fig. 5, we trained a state-of-the-art gene expression autoencoder, using the scVI framework[11], varying the size of the latent space $K \in \{10, 25, 50\}$. Second, we used the LogisticRegression function from the sklearn.linear_model python library to train a logistic regression on the binary sample-level labels and the first K features. Finally, we used the KMeans function from the sklearn.cluster python library with n_clusters = 2 on the continuous perturbation scores output by the logistic regression for cells in the case sample.

Note that HiDDEN does not require parameter tuning. Since we use all genes when computing the PC embedding of the data, the only parameter in the model is K, the number of PCs used in the training of the logistic regression. As described earlier in this section, we provide two data-driven heuristics for automatically choosing an appropriate value for K. For each dataset and for each heuristic, we scanned all integer values for K in the range [2, 60]. The first heuristic is to choose K that maximizes the number of DE genes defined by the HiDDEN refined binary labels. To compute the number of DE genes downstream of a given value of $K \in [2, 60]$, we used the Wilcoxon rank-sum differential expression test in scanpy with adjusted p-value threshold <0.05 (Supplementary Fig. 9a). The second heuristic is to choose the smallest value of K that maximizes the value of the two-sample

Kolmogorov-Smirnov (KS) test statistic comparing the sampling distributions of $\hat{p}$ for cells in the case sample with new refined label $\widetilde{Y} = 1$ and $\widetilde{Y} = 0$ (Supplementary Fig. 9b). To compute the value of the KS test statistic, we used the ks_2samp function from the scipy.stats python library.

Assessing agreement between HiDDEN continuous perturbation scores and ground truth labels. To quantify the ability of the continuous perturbation scores output by HiDDEN to capture the biological difference between Memory B and Naive B cells in each simulated dataset, we computed the Area Under the Receiver Operating Characteristic Curve (AUROC) using the roc_auc_score function from the sklearn.metrics python library (Fig. 2f, Supplementary Figs. 2a, 5a, 6a, 7I-K). The AUROC score can take values in the range [0, 1], with higher values indicating better agreement between the continuous prediction scores and the ground truth Naive B / Memory B binary labels.

Assessing agreement between HiDDEN-refined binary labels and ground truth labels. To evaluate the agreement between ground truth Naive B / Memory B labels and the binary labels refined by HiDDEN in each simulated dataset, we measured the accuracy of retrieved ground truth markers. Here we considered two scenarios – (1) working with the (unclustered) dataset as a whole and (2) looking for perturbation markers across clusters produced by the standard Seurat workflow.

The unclustered case: To define the set of ground truth markers for Naive B and Memory B cells in each simulated dataset, we computed the DE genes using the Naive B / Memory B ground truth labels using the Wilcoxon rank-sum differential expression test in scanpy with adjusted p-value threshold <0.05. Analogously, we defined the set of HiDDEN-derived marker genes as well as a baseline set of marker genes using the refined binary labels and the sample-level case-control labels, respectively, under the same testing procedure. We then computed the number of True Positives (TP), False Negatives (FN), False Positives (FP), and associated metrics of Recall, Precision, and F1-score (Fig. 2g, Supplementary Figs. 2b, 3, 4a, 5b, 6b). Recall can take values in the range [0, 1], with higher values indicating a higher fraction of correctly retrieved ground truth markers. Precision can take on values in the range [0, 1], with lower values indicating a higher fraction of falsely discovered marker genes. The F1-score is calculated as the harmonic mean of the precision and recall. F1-score can take values in the range [0, 1], with higher values indicating better agreement between the HiDDEN refined binary labels and the ground truth Naive B / Memory B binary labels.

The clustered case: We proceeded analogously to the unclustered case with the difference that DE testing was performed per cluster and the union of all DE genes across clusters defined the final gene set (Supplementary Fig. 3).

Comparing HiDDEN to CNA, MELD, Milo, and Mixscape. While these methods are developed with different objectives within the larger question of characterizing the effect of a perturbation in single-cell data, all of them utilize expression profiles (or a neighborhood graph derived from them) and cell labels reflecting the condition of the sample a cell comes from (as well as other metadata, optionally) as input. All methods output a continuous score measuring the effect of the perturbation in each cell, which can further be binarized whenever appropriate. Therefore, we can compare the performance of these five methods along both continuous scores and binary labels. Towards that end, we use the ground truth datasets of Naive B and Memory B cell mixtures.

Training of CNA was performed in python following the jupyter notebook tutorial provided by the authors of the method[34]. The original CNA implementation has a hard-coded assumption that the dataset to be analyzed is composed of at least five samples. We relaxed this assumption to accommodate our B cell mixtures and obtained

continuous perturbation scores as the per-cell neighborhood coefficient using the CNA association function with case/control status as the sample-level attribute of interest and case/control status as sample id. The resulting CNA continuous score is a correlation value ranging from −1 to 1.

Training of MELD was done in python following the code used by the authors of Milo in their comparison section[35]. We computed the k-nearest neighbors graph based on the expression matrix subsetted to the top 2000 highly variable genes. Then we used the meld_op.transform function to compute the density of the case/control sample labels and transformed the density to likelihood per condition using the meld.utils.normalize_densities function. Continuous perturbation scores, which take on values between 0 and 1, were obtained as the likelihood of label 1, which denotes the case condition.

Training of Milo was performed in python following the Differential abundance analysis in python with milopy jupyter notebook tutorial provided by the authors of the method[36]. The underlying implementation of Milo is done in R and calls on the glmFit routine from the edgeR R package. This routine cannot estimate the negative binomial dispersion parameter if it is not given at least three samples. To overcome this hard-coded assumption, we randomly created three samples per condition. We created the partially overlapping cellular neighborhoods using the milo.make_nhoods function. This results in a collection of neighborhoods, each identified by an index cell. A fraction of the cells in the dataset are deliberately excluded from the Milo analysis as outliers. Cells included in the analysis can also belong to one or more neighborhoods at the same time. Then we used the case/control sample level labels to count the number of cells from each sample in each neighborhood using the milo.count_nhoods function. We used the milo.DA_nhoods function to perform differential abundance testing, which outputs a log fold-change test statistic per neighborhood. Since we aim to make comparisons at the level of individual cells, when a cell belonged to more than one neighborhood - we reported the average log fold-change across neighborhoods. For cells excluded from the analysis, the average log fold-change is NA. Continuous perturbation scores were obtained as the average log fold-change and can take on values between minus to plus infinity.

Training of Mixscape was performed in python following the code provided by the Theis lab as part of pertpy tools on github at https://github.com/theislab/pertpy/blob/development/pertpy/tools/_mixscape.py. We calculated perturbation signatures by subtracting the averaged expression profile of the 20 control neighbors from the expression profile of each cell using the perturbation_signature function. We then identified perturbed and non-perturbed cells within the case condition using the mixscape function with default hyperparameter values.

Comparison of the continuous perturbation scores from HiDDEN, CNA, MELD, Milo, and Mixscape (Supplementary Figs. 6a, 7) was performed using the AUROC described earlier in the Methods section as a metric for assessing agreement between continuous perturbation scores and ground truth labels.

Converting the continuous perturbation scores into binarized labels was done in an identical manner for HiDDEN, CNA, MELD, and Milo, as described earlier in the Methods section. Mixscape has a built-in function for computing binary perturbed/non-perturbed labels. Comparison of the binary labels across all five methods (Supplementary Figs. 6b, 8) was performed using the F1-score described earlier in the Methods section as a metric for assessing the agreement between HiDDEN-refined binary labels and ground truth labels.

Using HiDDEN-refined binary labels as input to CNA, MELD, and Milo. HiDDEN-refined binary labels demonstrate better agreement with ground truth labels than the sample-level case/control labels. Therefore, we explored the performance of CNA, MELD, and Milo when given HiDDEN binary labels in lieu of case/control labels as input

(Supplementary Figs. 7, 8). CNA continuous perturbation scores were obtained as the per-cell neighborhood coefficient using the CNA association function with Memory B / Naive B ground truth labels as the sample-level attribute of interest and HiDDEN-refined binary labels as sample id. MELD continuous perturbation scores were obtained as the density of the HiDDEN-refined binary labels transformed to likelihood of label 1. Milo continuous perturbation scores were obtained as the average log fold-change downstream of differential abundance testing per neighborhood using the HiDDEN-refined binary labels to count the number of cells per condition. All continuous scores were binarized in the same manner as above.

Assessing performance on human multiple myeloma and precursor states data
The first real dataset we analyzed consists of single-cell RNA-seq profiles of human bone marrow plasma cells from patients with multiple myeloma (MM) ($n = 8$ patients, $N = 10790$ cells), its precursor conditions smoldering multiple myeloma (SMM) ($n = 12$ patients, $N = 8431$ cells) and monoclonal gammopathy of undetermined significance (MGUS) ($n = 6$ patients, $N = 817$ cells), and healthy donors with normal bone marrow (NBM) ($n = 9$ patients, $N = 9329$ cells)[4].

Precursor samples can contain a mixture of neoplastic and normal cells (Fig. 3a) and the authors of the original study define two sources of ground truth describing the malignancy status of precursor samples and their cells. The first source is a manual annotation of binary labels reflecting whether a cell is healthy or malignant. The second source is a tumor-purity estimate of the proportion of malignant cells in each precursor sample.

HiDDEN model training. This dataset has pronounced patient-specific batch effects (Supplementary Fig. 12a). Therefore, echoing the analysis in the original study, we first deployed a batch-sensitive strategy to refine the malignancy status of cells in precursor samples using HiDDEN. Additionally, we developed a batch-agnostic strategy as well, to explore the ability of our method to perform well in the presence of strong batch effects. Mirroring the within-patient annotation approach in the original study, our batch-sensitive fitting approach considers each precursor sample one at a time. We trained the model on all NBM, one precursor sample, and all MM samples, where we give all NBM cells label 0, or healthy, and all the rest label 1, reflecting that they do not originate from healthy donors. The batch-agnostic fitting approach consists of fitting the model to all NBM samples, all precursor samples manually annotated to be mixed, and all MM samples together. Besides this, all other aspects of model training were carried out the same way between the two strategies. The results from the batch-specific strategy are featured in Fig. 3, and the results of the batch-agnostic approach, along with a comparison of the two, are included in Supplementary Fig. 12.

The HiDDEN model training was done in python. First, we followed the standard preprocessing routine for single-cell RNA-seq data in scanpy and log-normalized each sample separately. We then used the standard PCA dimensionality reduction routine in scanpy on the scaled gene features. Next, we used the LogisticRegression function from the sklearn.linear_model python library to train a logistic regression on the binary NBM / non-NBM labels and the first K PCs. Finally, we used the KMeans function from the sklearn.cluster python library with n_clusters = 2 on the continuous perturbation scores output by the logistic regression for cells in each precursor sample separately (Supplementary Table 2). We used all genes to compute the PC dimensionality reduction and automatically chose K, the number of PCs used in the logistic regression, using the heuristic for maximizing the number of DE genes downstream of the HiDDEN refined binary labels (Supplementary Fig. 13). The specific strategy for defining the DE genes in this dataset characterized by strong batch effects is described in detail below.

Under both fitting strategies, we used the same downstream metrics to evaluate the performance of our method at recovering the manually annotated cell-level labels and the purity sample-level estimates, as described below.

Assessing agreement between HiDDEN continuous perturbation scores and ground truth manually annotated labels. To quantify the ability of the continuous perturbation scores produced by HiDDEN to capture the manually annotated healthy-malignant binary labels in all mixed precursor samples, we computed the AUROC using the roc_auc_score function from the sklearn.metrics python library. The average AUROC across samples per precursor state from the batch-sensitive training strategy is depicted in Fig. 3b, and the per-sample curves and distributions of perturbation scores are included in Supplementary Fig. 12.

Assessing agreement between HiDDEN refined binary labels and tumor-purity sample estimates. The sample-level source of ground truth provided in the original study is an estimate of the proportion of malignant cells from a Bayesian hierarchical model based only on the expression of immunoglobulin light chain genes. Additionally, we estimated the per-sample tumor-purity using the manual labels and the HiDDEN-refined binary labels. Confidence bounds in all three cases were derived following the same approach as in the original study (Fig. 3c, e).

There are three quantities relevant to the ability of HiDDEN and manual annotation labels to recapitulate the ground truth point estimates of sample purity, i.e. the proportion of neoplastic cells in a sample:

1. sample purity estimates derived from HiDDEN-refined binary labels, denoted as p_1,
2. manual annotation-based estimates of sample purity reported in the paper the data originated from[4] denoted as p_2,
3. ground truth point estimates of proportion of malignant cells[4], denoted as p_3.

Deriving significance values that quantify the ability of HiDDEN-derived labels and manual annotation labels to match the ground truth point estimates of malignant cell proportions. we treat p_1 and p_2 as random variables, and the point estimate p_3 as a scalar. The probability distributions of p_1 and p_2 follow from the derivation of confidence bounds in the original study[4] and are described in more detail below.

For a given sample, let p be the proportion of neoplastic cells (denoting the population parameter and not just the estimate based on the sequenced cells from that sample). We model the observed data as

$$x \sim \text{Binomial}(n, p),$$

where n denotes the total number of sequenced cells in the sample, x of which are labeled as neoplastic. We put a noninformative uniform prior on p:

$$p \sim \text{Beta}(1, 1).$$

Due to the Beta-Binomial conjugacy, the posterior distribution of p is closed form:

$$p|n, x \sim \text{Beta}(x + 1, n - x + 1),$$

with p taking on values in $[0, 1]$.

For each of the 11 precursor samples we consider in Fig. 3c, e, we test a pair of hypotheses:

$$H_0 : p_1 = p_3$$

and

$$H_A : p_1 \neq p_3$$

$$H_0 : p_2 = p_3$$

$$H_A : p_2 \neq p_3.$$

Since we are testing analogous hypotheses for p_1 and p_2, below we describe the hypothesis test with respect to p_1.

The distribution of the test statistic (the difference between p_1 and p_3) under the null hypothesis is:

$$p_3|n, H_0 \sim \text{Beta}(x_3 + 1, n - x_3 + 1),$$

with support $[-p_3, 1 - p_3]$, where x_3 is the expected number of neoplastic cells under the null, calculated as $p_3 * n$ rounded to the nearest integer.

Therefore, the p-value is calculated as the probability to observe a test statistic equal to or more extreme (in either direction away from 0) according to the null distribution:

$$\mathbb{P}\left(\text{Beta}(x_3 + 1, n - x_3 + 1) \leq -|p_1 - p_3|\right) + \mathbb{P}\left(\text{Beta}(x_3 + 1, n - x_3 + 1) \geq |p_1 - p_3|\right). \tag{2}$$

The resulting p-values for testing $p_1 - p_3$ and $p_2 - p_3$ across all 11 mixed precursor samples are reported in Supplementary Table 2. The significance depicted in Fig. 3c, e is with respect to a Bonferroni-adjusted threshold of alpha=0.01/22 = 4.55E-04. A smaller p-value indicates a larger discrepancy with the ground truth. Whenever the HiDDEN-based approach produced a larger p-value compared to the manual annotation, we concluded that HiDDEN agreed better with the ground truth compared to the manual annotation approach.

Differential expression analysis using manually annotated and HiDDEN binary labels. To demonstrate the ability of the HiDDEN-refined binary labels to discover additional malignancy markers from precursor samples, we computed the DE genes using the manual annotation in NBM and MM samples and contrasted it against the DE genes found using the HiDDEN refined labels in precursor samples (Fig. 3d). Due to the presence of strong batch effects in the data, mirroring the DE testing strategy in the original paper, we find the DE genes per patient and take the union across patients. For a gene to be considered DE, it had to have an adjusted p-value <0.05 from the t-test differential expression testing routine in scanpy and a maximum absolute log-foldchange >1.5.

To assess the significance of the overlap between the two sets of DE genes, we ran a hypergeometric test using the dhyper function from the Stats R package. The background number of genes was calculated based on genes expressed in at least one cell from all precursor samples (18, 770 genes).

Validation of HiDDEN binary labels in non-mixed MGUS samples. The authors of the original study computed a Bayesian non-negative matrix factorization (NMF) to highlight gene signatures that are active in this patient cohort and validated them in external cohorts. Several signatures were annotated with a biological interpretation. There were three MGUS samples considered to consist of only healthy cells, according to the manual annotation. They are also the three precursor samples with lowest, although not zero, estimated sample purity according to the Bayesian purity model from the original study. The HiDDEN refined binary labels for these patients annotate some of their cells as malignant. To validate this annotation, we plotted the mean activity of the genes identified by the original study for each signature in the cells labeled as healthy and as malignant for each sample (Fig. 3f,

Supplementary Fig. 11). The confidence bounds (SEM) in Fig. 3f were derived following the same approach as in the original study.

Analysis of mouse endothelial cells from a time-course demyelination experiment

Generation of demyelination and control tissue. The second real dataset analyzed consists of single-nucleus RNA-seq profiles of mouse endothelial cells from a demyelination model with matched controls. Case and control animals received 500 nl of 1% lysophosphatidylcholine (LPC, Cat# 440154, Millipore Sigma, US) or saline vehicle (PBS) injection, respectively. This was delivered intracranially, using standard approaches, with a Nanoject III (Drummond, US) into the corpus callosum at the following stereotaxic coordinates: Anterior-Posterior: −1.2, Medio-lateral: 0.−5 relative to bregma and a depth of 1.4 mm normalized to the surface of the skull.

All mice were housed on a 12-h light/dark cycle between 68 °F and 79 °F and 30-70% humidity. All animal work was approved by the Broad's Institutional Animal Care and Use Committee (IACUC). The only mouse strain used was C57BL/6 J which was purchased directly from The Jackson Labs (cat# 000664). Mice were sacrificed at four time points: 3, 7, 12, and 18 days post injection (dpi) with $n = 3$ animals per time point per condition, totaling $n = 24$ animals (Fig. 4a). At the appropriate time point, mice were perfused with ice-cold pH 7.4 HEPES buffer (containing 110 mM NaCl, 10 mM HEPES, 25 mM glucose, 75 mM sucrose, 7.5 mM MgCl2, and 2.5 mM KCl) to remove blood from the brain. Brains were fresh frozen for 3 min in liquid nitrogen vapor and all tissue was stored at −80 °C for long-term storage. The full dataset will be described in a forthcoming paper (Dolan et al., in preparation).

Generation of single-nucleus RNA profiles. Frozen mouse brains were mounted onto cryostat chucks with OCT embedding compound within a cryostat. Brains were sectioned until reaching the injection site location, which was confirmed by the presence of hypercellularity using a Nissl stain (Histogene Staining Solution, KIT0415, Thermofisher). For saline controls, anatomical landmarks were used to determine the injection site. Lesions or control white matter was microdissected using a 1 mm biopsy punch (Integra Miltex, US), whose circular punch was bent into a rectangle shape with a sterile hemostat. Lesion punches were 300 μm deep.

Each excised tissue punch was placed into a pre-cooled 0.25 ml PCR tube using pre-cooled forceps and stored at −80 °C for a maximum of 24 hours. Nuclei were extracted from this frozen tissue using gentle, detergent-based dissociation, according to a protocol available at protocols.io (https://doi.org/10.17504/protocols.io.bck6iuze) with minor changes to maximize nuclei extraction, which will be described in a forthcoming paper (Dolan et al. in preparation). Nuclei were loaded into the 10x Chromium V3 system. Reverse transcription and library generation were performed according to the manufacturer's protocol (10x Genomics). Sequencing reads from mouse cerebellum experiments were demultiplexed and aligned to a mouse (mm10) premrna reference using CellRanger v3.0.2 with default settings. Digital gene expression matrices were generated with the CellRanger count function. Initial analysis and generation of overall UMAP and clustering (Fig. 4b) was performed with Seurat v3[26].

Clustering analysis of endothelial cells. For the UMAP representations in Fig. 4c, d, and Supplementary Fig. 14 of the profiles of $N = 891$ endothelial cells from $n = 6$ animals spanning both case and control conditions and all four timepoints, we used the standard scanpy functions to log-normalize and scale the gene counts, ran PCA, computed the nearest-neighbor graph with 10 neighbors in the latent space defined by the first 50 PCs, and ran the UMAP algorithm.

To cluster the endothelial cells (Fig. 4d), we used the standard preprocessing and clustering workflow in Seurat. We normalized the gene counts, performed variable gene selection with default parameter values, scaled the normalized counts, performed dimensionality reduction using PCA with default parameter values, built the nearest-neighbor graph with the default number of nearest neighbors, and used the Leiden algorithm with resolution parameter $= 0.5$ to find clusters.

For the heatmaps in Fig. 4e and Supplementary Fig. 14c, we computed the abundance of case (LPC) and control (PBS) labels in each cluster and across time points, respectively.

HiDDEN model training. Model training was performed in python on all $n = 891$ endothelial cells together. We followed the standard preprocessing routine for single-cell RNA-seq data in scanpy and log-normalized the gene counts, followed by PCA dimensionality reduction of the scaled gene features. For the continuous perturbation score in Fig. 4f and Supplementary Fig. 14d, we used the LogisticRegression function from the sklearn.linear_model python library to train a logistic regression on the binary PBS / LPC labels and the first K PCs. We used all genes to compute the PC embedding and automatically chose $K = 5$, the number of PCs used in the logistic regression, using the heuristic for maximizing the number of DE genes downstream of the HiDDEN refined binary labels. The strategy for defining the DE genes in this dataset is described in detail below.

For the HiDDEN refined binary labels in Fig. 4g, we used the KMeans function from the sklearn.cluster python library with n_clusters $= 2$ to split the continuous perturbation scores of all LPC endothelial cells at 3 dpi into two groups. We denote the group with lower perturbation scores LPC0, corresponding to endothelial cells unaffected by the LPC injection and similar to endothelial cells in the PBS control condition, and the group with higher perturbation scores LPC1, as the subset of endothelial cells affected by the LPC injection.

Differential expression analysis to define endothelial LPC1 markers. To define the set of endothelial LPC1 markers depicted in the dotplot in Fig. 5a, we took the unique perturbation-enriched genes found in DE analysis using the HiDDEN refined labels and not in DE analysis using the original PBS / LPC labels. We performed both DE analyses using the Wilcoxon rank-sum test in scanpy with a threshold for the adjusted p-value <0.05. The comprehensive output from both tests can be found in Supplementary Table 3, with the unique genes found using the HiDDEN refined binary labels highlighted in bold font.

Validation of endothelial LPC1 markers using RNAscope. Fresh-frozen, 14 μm sections of 3 days post injection (dpi) demyelinating or saline control tissue were mounted on cold Superfrost plus slides (Fisher Scientific, US). These slides were stored at −80 °C. We performed RNAscope Multiplex Fluorescent v2 (Advanced Cell Diagnostics, US) using probes targeting *S100a6* (412981), *Lgals1* (897151-C2) and *Flt1* (415541-C3), where *Flt1* is a general marker for endothelial cells. RNAscope was performed following the manufacturer's protocol for fresh frozen tissue and the following dyes were used at a concentration of 1/1500 to label specific mRNAs (TSA Plus fluorescein, TSA Plus Cyanine 3, TSA Plus Cyanine 5 from PerkinElmer, USA). Imaging was performed on an Andor CSU-X spinning disk confocal system coupled to a Nikon Eclipse Ti microscope equipped with an Andor iKon-M camera. Images were acquired using 20x air and 60x oil immersion objectives (Nikon). All images shown in Fig. 5b are representative images taken from at least 2 independent experiments.

Interpretation of endothelial LPC1 markers using gene ontology analysis. For the identification of gene ontology (GO) categories summarizing the list of unique endothelial LPC1 markers (Fig. 5c, d), we performed GO enrichment analysis in g:Profiler[37] with default settings and ReviGo[38] to summarize and visualize the results.

Ligand-receptor analysis to identify cell-cell communication changes between endothelial LPC1 and LPC0. To contrast the ligand-receptor communication of the two endothelial LPC subtypes with neighboring cell types in the tissue (Fig. 5e), we used a modification (Goeva et al., in preparation) to CellphoneDB[39]. We separated the output for each interaction in three bins based on the sign of the test statistic (Supplementary Fig. 15) and the magnitude of the p-value, reflected in the figure legend: significantly depleted in LPC1 with respect to LPC0, not significant (p-value ≥ 0.05), and significantly enriched in LPC1 with respect to LPC0.

Statistics and reproducibility. HiDDEN was evaluated on data originating from two publicly available datasets and one novel dataset, using as many samples as possible in these datasets (no statistical method was used to predetermine the sample size and no data were excluded from the analyses). Preprocessing steps were performed according to standard practice and reported for each dataset independently. The experiments involving running computational methods on previously published publicly available datasets did not require randomization. The investigators were not blinded to allocation during experiments and assessment of outcome. Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

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

Data availability
The endothelial single-nucleus RNA-seq data used in this study are available in the GEO database under accession code GSE276570. The rest of the datasets used in this study were already publicly available. The processes PBMC data used in this study are freely available from 10x Genomics and can be downloaded by following the link: https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k. The scRNA-seq human bone marrow plasma cell data from patients with multiple myeloma, precursor states, and healthy donors is available in the GEO database under accession code GSE193531. All relevant data supporting the key findings of this study are available within the article and its Supplementary Information files. Source data are provided with this paper.

Code availability
Code and scripts to reproduce analyses presented here are available on Github at https://github.com/tudaga/LabelCorrection[31].

References
1. Grubman, A. et al. A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.* **22**, 2087–2097 (2019).
2. Wilk, A. J. et al. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nat. Med.* **26**, 1070–1076 (2020).
3. Kamath, T. et al. A molecular census of midbrain dopaminergic neurons in Parkinson's disease. *bioRxiv* https://doi.org/10.1101/2021.06.16.448661 (2021).
4. Boiarsky, R. et al. Single cell characterization of myeloma and its precursor conditions reveals transcriptional signatures of early tumorigenesis. *Nat. Commun.* **13**, 7040 (2022).
5. Aissa, A. F. et al. Single-cell transcriptional changes associated with drug tolerance and response to combination therapies in cancer. *Nat. Commun.* **12**, 1628 (2021).
6. Dixit, A. et al. Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell* **167**, 1853–1866.e17 (2016).
7. Replogle, J. M. et al. Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq. *Cell* **185**, 2559–2575.e28 (2022).
8. Papalexi, E. et al. Characterizing the molecular regulation of inhibitory immune checkpoints with multimodal single-cell screens. *Nat. Genet.* **53**, 322–331 (2021).
9. Petukhov, V. et al. Case-control analysis of single-cell RNA-seq studies. *bioRxiv* (2022) https://doi.org/10.1101/2022.03.15.484475 (2022).
10. Keren-Shaul, H. et al. A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. *Cell* **169**, 1276–1290.e17 (2017).
11. Gayoso, A. et al. A Python library for probabilistic analysis of single-cell omics data. *Nat. Biotechnol.* **40**, 163–166 (2022).
12. Reshef, Y. A. et al. Co-varying neighborhood analysis identifies cell populations associated with phenotypes of interest from single-cell transcriptomics. *Nat. Biotechnol.* **40**, 355–363 (2022).
13. Kotliar, D. et al. Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *Elife* **8**, (2019).
14. Burkhardt, D. B. et al. Quantifying the effect of experimental perturbations at single-cell resolution. *Nat. Biotechnol.* **39**, 619–629 (2021).
15. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat. Biotechnol.* **40**, 245–253 (2022).
16. Datasets -Single Cell Multiome ATAC + Gene Exp. -Official 10x Genomics Support. https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k (2022).
17. Akkaya, M., Kwak, K. & Pierce, S. K. B cell memory: building two walls of protection against pathogens. *Nat. Rev. Immunol.* **20**, 229–238 (2020).
18. Bhattacharya, D. et al. Transcriptional profiling of antigen-dependent murine B cell differentiation and memory formation. *J. Immunol.* **179**, 6808–6819 (2007).
19. Cantuti-Castelvetri, L. et al. Defective cholesterol clearance limits remyelination in the aged central nervous system. *Science* **359**, 684–688 (2018).
20. Miller, R. H., Fyffe-Maricich, S. & Caprariello, A. C. Chapter 37 - Animal Models for the Study of Multiple Sclerosis. in *Animal Models for the Study of Human Disease (Second Edition)* (ed. Conn, P. M.) 967–988 (Academic Press, 2017).
21. Lloyd, A. F. & Miron, V. E. The pro-remyelination properties of microglia in the central nervous system. *Nat. Rev. Neurol.* **15**, 447–458 (2019).
22. Molina-Gonzalez, I. & Miron, V. E. Astrocytes in myelination and remyelination. *Neurosci. Lett.* **713**, 134532 (2019).
23. Shen, K. et al. Multiple sclerosis risk gene Mertk is required for microglial activation and subsequent remyelination. *Cell Rep.* **34**, 108835 (2021).
24. Yuen, T. J. et al. Oligodendrocyte-encoded HIF function couples postnatal myelination and white matter angiogenesis. *Cell* **158**, 383–396 (2014).
25. Zhou, T. et al. Microvascular endothelial cells engulf myelin debris and promote macrophage recruitment and fibrosis after neural injury. *Nat. Neurosci.* **22**, 421–435 (2019).
26. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21 (2019).
27. Girolamo, F., Coppola, C., Ribatti, D. & Trojano, M. Angiogenesis in multiple sclerosis and experimental autoimmune encephalomyelitis. *Acta Neuropathol. Commun.* **2**, 84 (2014).
28. Berghoff, S. A. et al. Blood-brain barrier hyperpermeability precedes demyelination in the cuprizone model. *Acta Neuropathol. Commun.* **5**, 94 (2017).

29. Nguyen, B., Bix, G. & Yao, Y. Basal lamina changes in neurodegenerative disorders. *Mol. Neurodegener.* **16**, 81 (2021).
30. Abbott, N. J., Rönnbäck, L. & Hansson, E. Astrocyte-endothelial interactions at the blood-brain barrier. *Nat. Rev. Neurosci.* **7**, 41–53 (2006).
31. Goeva, A. *HiDDEN: A Machine Learning Method for Detection of Disease-Relevant Populations in Case-Control Single-Cell Transcriptomics Data*, https://github.com/tudaga/LabelCorrection. (Zenodo, 2024). https://doi.org/10.5281/ZENODO.13823942.
32. Massey, F. J. Jr. The kolmogorov-smirnov test for goodness of fit. *J. Am. Stat. Assoc.* **46**, 68–78 (1951).
33. Weighted nearest neighbor analysis. https://satijalab.org/seurat/articles/weighted_nearest_neighbor_analysis.html (2022).
34. Notebook on nbviewer. https://nbviewer.org/github/yakirr/cna/blob/master/demo/demo.ipynb (2023).
35. *Run_meld.Py at Main · MarioniLab/Milo_analysis_2020.* (Github, 2023).
36. Notebook on nbviewer. https://nbviewer.org/github/emdann/milopy/blob/master/notebooks/milopy_example.ipynb (2023).
37. Raudvere, U. et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).
38. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One* **6**, e21800 (2011).
39. Efremova, M., Vento-Tormo, M., Teichmann, S. A. & Vento-Tormo, R. CellPhoneDB: inferring cell–cell communication from combined expression of multi-subunit ligand–receptor complexes. *Nat. Protoc.* **15**, 1484–1506 (2020).

Author contributions
A.G. developed the algorithm and performed all analyses. M.J.-D. acquired the demyelination data and performed the imaging validation experiments, with help from J.L. and E.G. RB and RMG assisted with biological interpretation of the analyses. A.G. and EM conceived the study and wrote the paper, with contributions 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/s41467-024-53666-8.

Correspondence and requests for materials should be addressed to Aleksandrina Goeva or Evan Macosko.

Peer review information *Nature Communications* thanks Rahul Dhodapkar, Qin Ma and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.