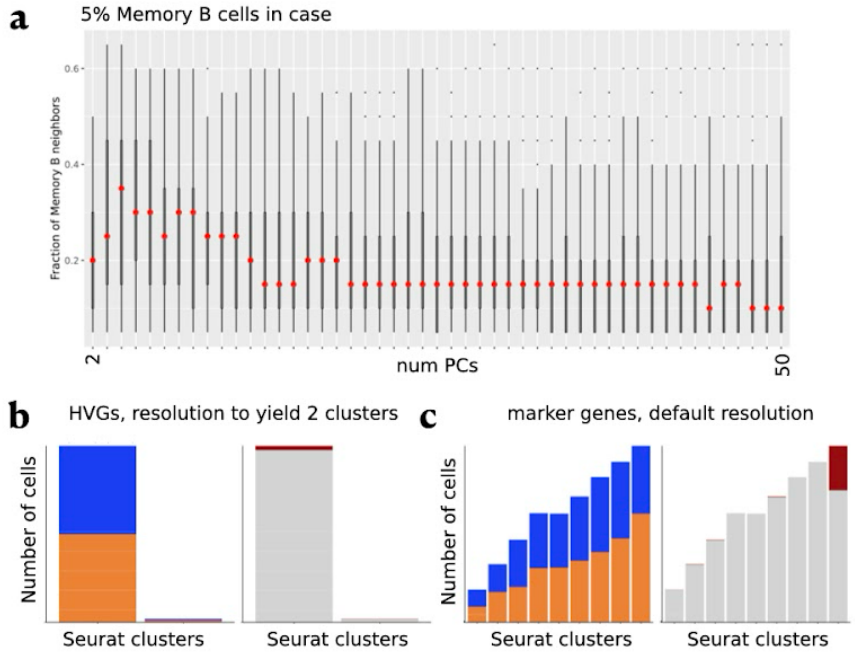
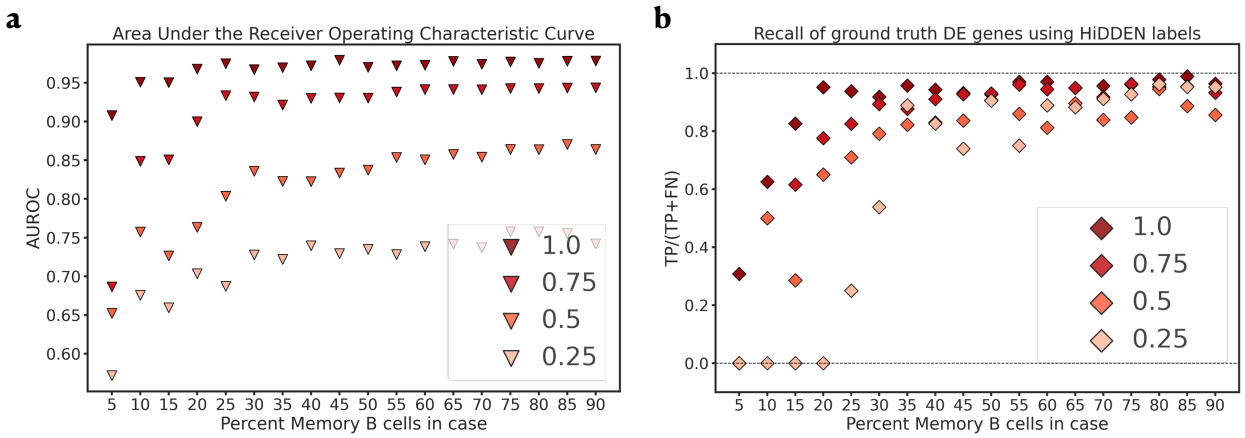
HiDDEN: A machine learning method for detection of disease-relevant populations in case-control single-cell transcriptomics data

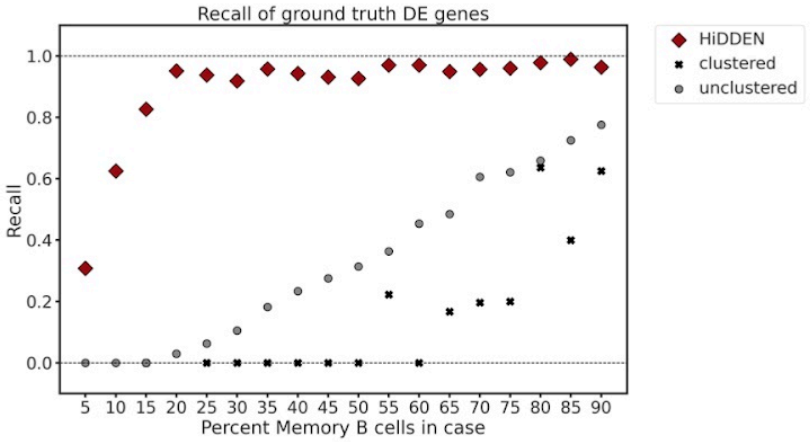
Supplementary Figures



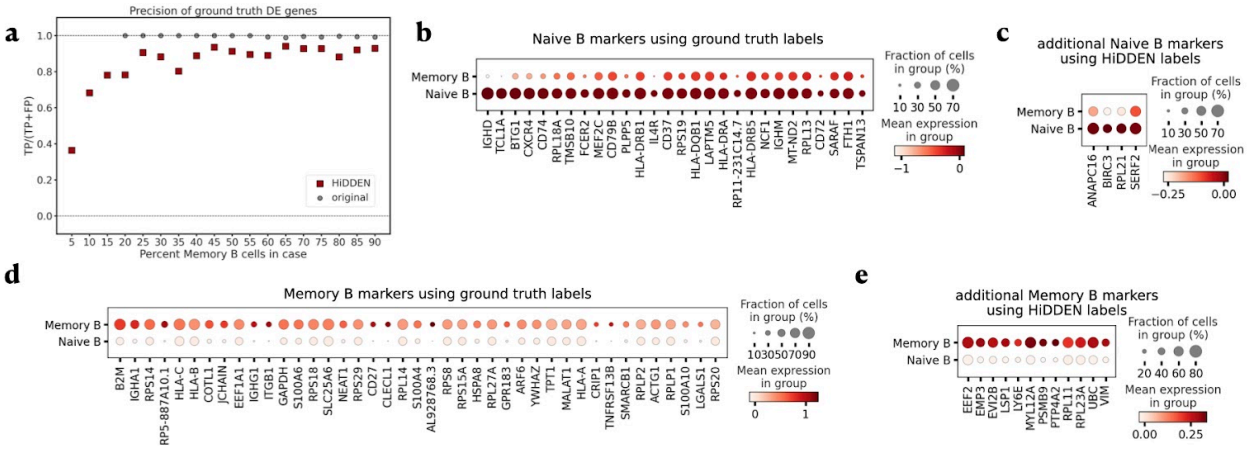
Supplementary Figure 1: **Separability of Naive B and Memory B cells in the latent space and sensitivity of standard dimensionality reduction and clustering workflow to influential parameters for the dataset with 5% Memory B cells in case (total n=1949 cells). A** Box plots of the distribution of fraction of Memory B cells among the 20-nearest neighbors of each Memory B cell (y-axis). The fraction of Memory B neighbors in the latent space is defined by a variable number of top PCS from 2 to 50 (x-axis). Distribution of case-control (left) and Memory B-Naive B (right) cell identities across Seurat clusters. Colors defined as follows: case-orange, control-blue, Memory B-red, Naive B-gray. Dimensionality reduction and resolution parameters chosen as follows: **B** Using highly variable genes for feature selection (default); resolution parameter chosen to yield two clusters. **C** Using Naive B and Memory B markers for feature selection; default resolution. Source data are provided as a Source Data file.



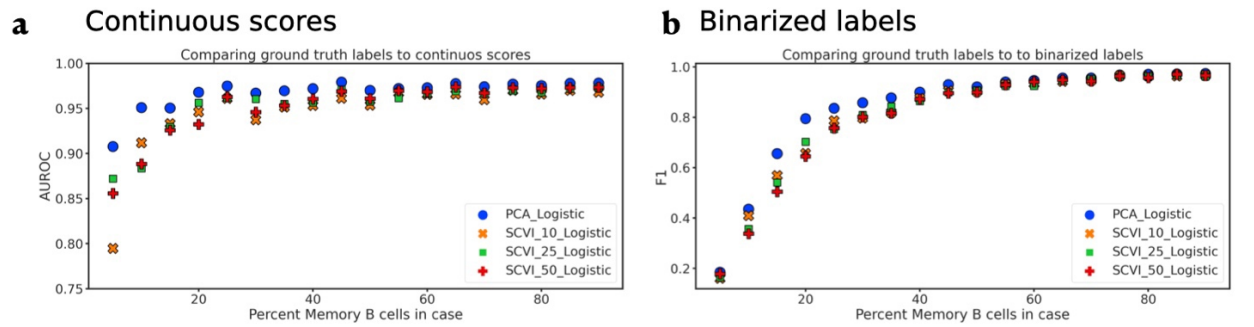
Supplementary Figure 2: **Effect of perturbation strength (as indicated by legend labels) and fraction perturbed cells in case (x-axis) on problem difficulty.** y-axis defined as follows: **A** Area under the Receiver Operating Characteristic Curve (AUROC) for classification of ground truth cell labels; **B** Recall of ground truth DE genes from DE testing on HiDDEN-refined labels. Source data are provided as a Source Data file.



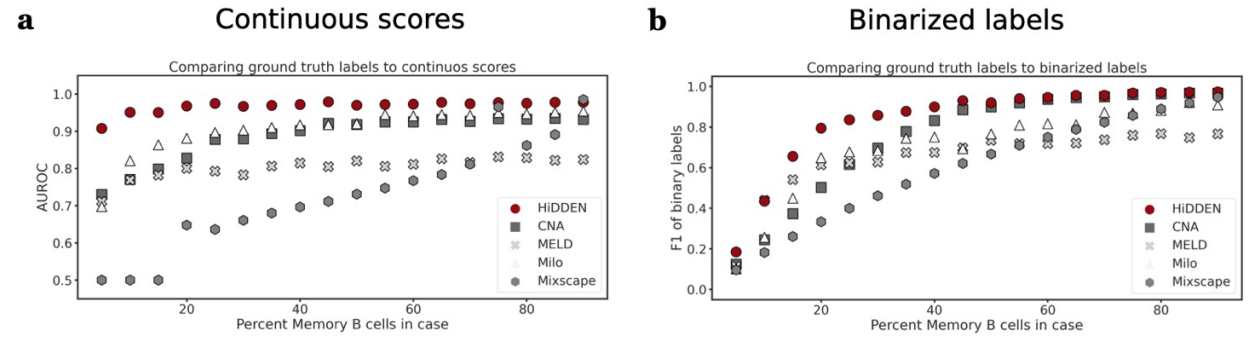
Supplementary Figure 3: **Recall of ground truth DE genes as a function of percent Memory B cells in case sample.** Legend labels define the DE testing approach as follows: HiDDEN - using the HiDDEN-refined binary labels on the unclustered data; unclustered - using the case-control labels on the unclustered data; clustered - using the case-control labels to perform DE testing per Seurat cluster and taking the union of DE genes across clusters. Source data are provided as a Source Data file.



Supplementary Figure 4: **DE genes found from DE testing using HiDDEN-refined binary but not found using Memory B-Naive B labels. A** Precision of ground truth DE genes from DE testing on HiDDEN-refined and original case-control labels as a function of percent Memory B cells in case sample. For the dataset with 20% Memory B cells in case: Dotplot of mean expression of Naive B (**B**) and Memory B (**D**) marker genes ordered by p-value. Dotplot of mean expression of HiDDEN-refined label 0 (**C**) and label 1 (**E**) DE genes ordered alphabetically. Source data are provided as a Source Data file.

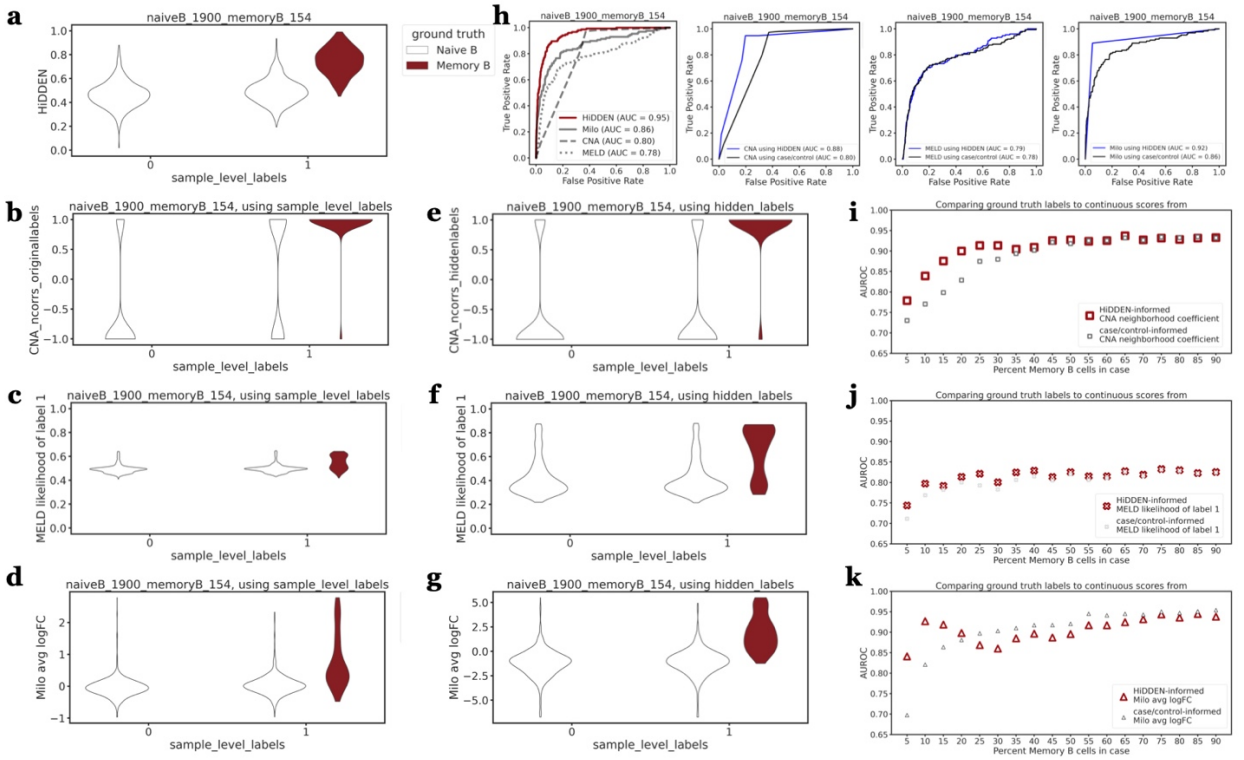


Supplementary Figure 5: **A direct comparison between a linear and a deep-learning version of HiDDEN.** Four variations of the dimensionality reduction component are considered: PCA (linear) and an scVI autoencoder with a latent dimension of size 10, 25, and 50. **A** AUROC score for using continuous perturbation scores for classification of ground truth cell labels as a function of percent Memory B cells in case. **B** F1-score measuring the accuracy of binarized labels against ground truth Naive B / Memory B labels as a function of percent Memory B cells in case sample. Source data are provided as a Source Data file.

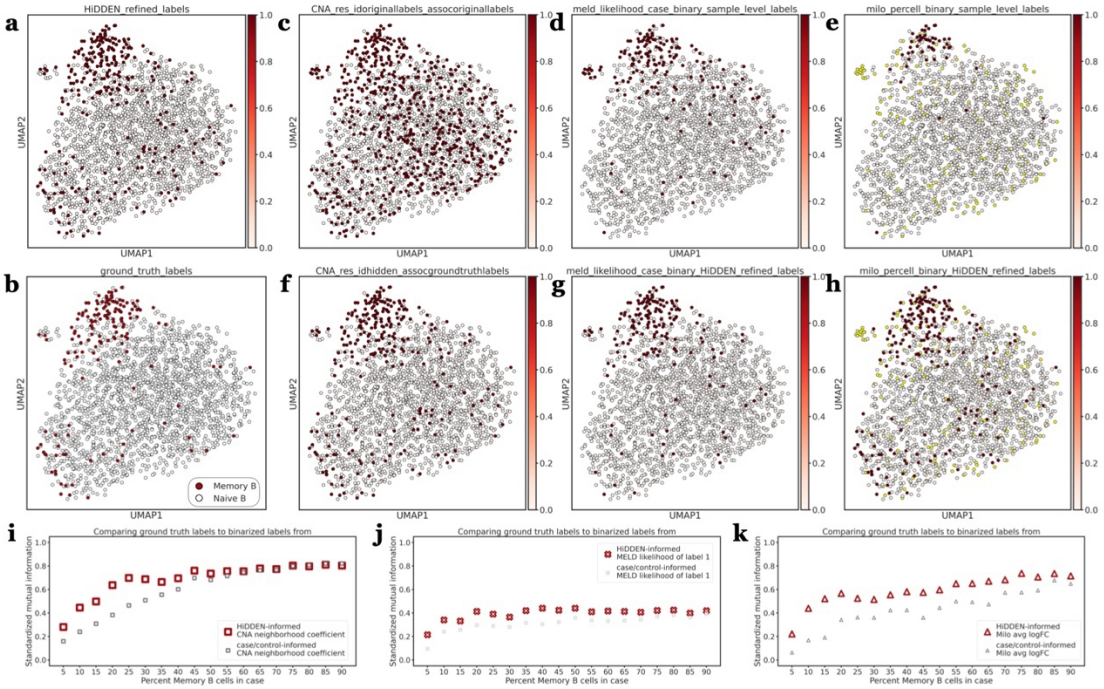


Supplementary Figure 6: **A direct comparison between HiDDEN and CNA, MELD, Milo, and Mixscape. A** AUROC score for comparing ground truth labels to continuous perturbation scores as a function of percent Memory B cells in case sample for each of the five methods. **B** F1 score measuring

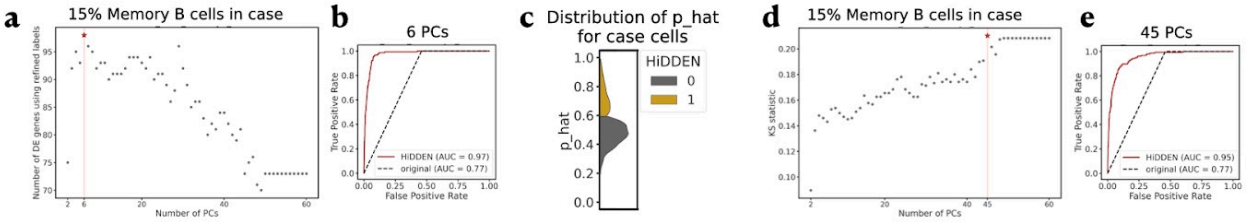
the accuracy of binarized labels against ground truth Naive B / Memory B labels as a function of percent Memory B cells in case sample for each of the four methods. Source data are provided as a Source Data file.



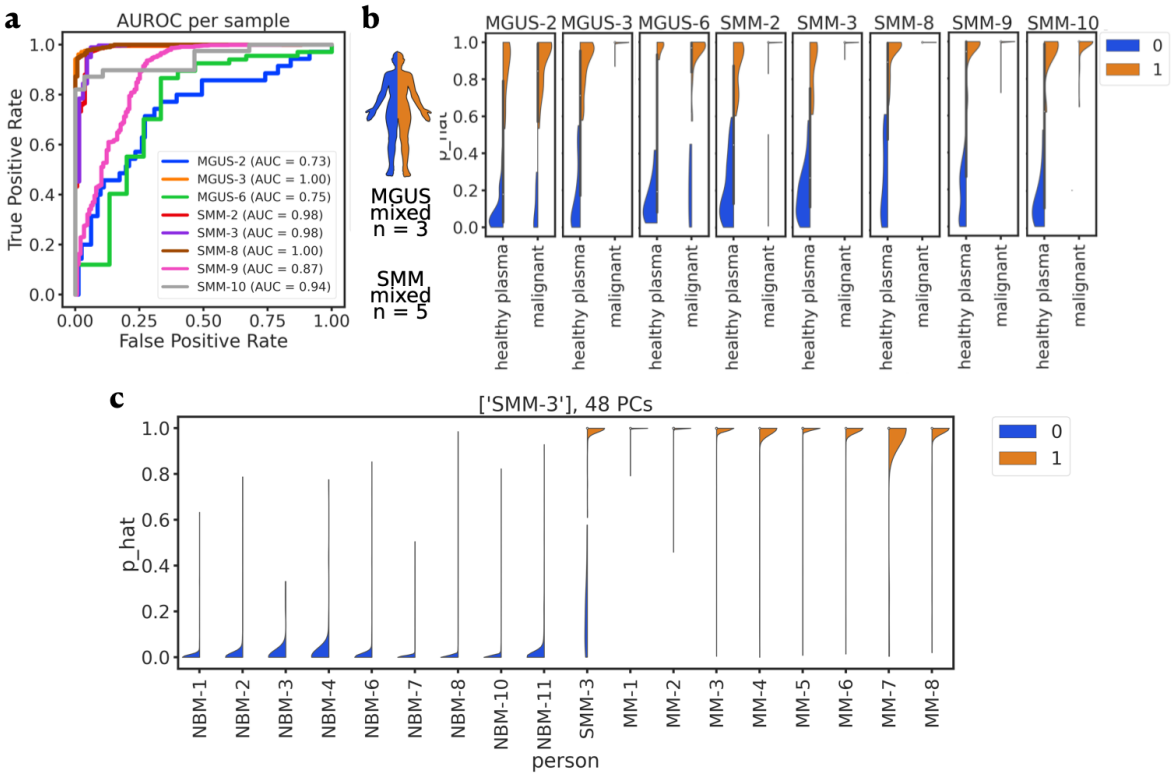
Supplementary Figure 7: **A comparison of HiDDEN-informed vs. case/control-informed performance of CNA, MELD, and Milo continuous scores.** Violin plots (area not scaled to count) of the distribution of continuous scores (method and continuous score details detailed on y-axis) of Naive B and Memory B cells split over control and case samples and colored by ground truth labels, for the dataset containing 15% Memory B cells in the case sample. Results in **A, B, C, D** use case/control sample-level labels as input. Results in **E, F, G** use HiDDEN-refined binary labels. **H** Area under the Receiver Operating Characteristic (ROC) curves for classification of ground truth cell labels as a function of percent Memory B cells in case sample with the Area Under the ROC (AUROC) and method indicated in the legend, for the dataset containing 15% perturbed cells in the case sample. **I, J, K** AUROC score for using continuous perturbation scores for classification of ground truth cell labels as a function of perturbed cells in case sample, with the method name reflected in the plot title. Source data are provided as a Source Data file.



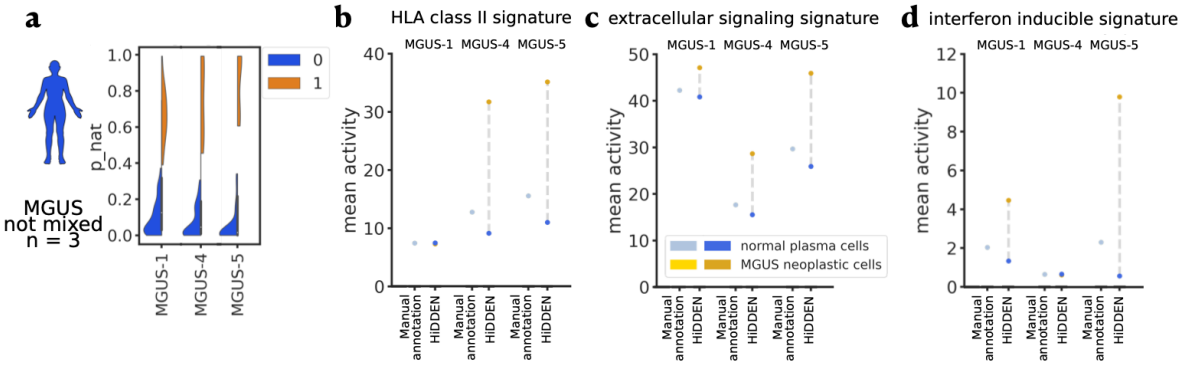
Supplementary Figure 8: **A comparison of HiDDEN-informed vs. case/control-informed performance of CNA, MELD, and Milo binarized labels.** UMAP embeddings of Naive B and Memory B cell gene expression for the dataset containing 15% Memory B cells in the case sample colored by: **A** HiDDEN-refined binary labels, **B** ground truth Memory B / Naive B labels, **C, F** CNA binarized neighborhood coefficient, **D, G** MELD binarized likelihood of label Memory B, **E, H** Milo binarized average log fold-change (yellow dots mark NA log fold-change values corresponding to cells excluded by Milo from the analysis); (using case/control or HiDDEN-refined labels, respectively) **I, J, K** Standardized mutual information between ground truth Naive B / Memory B labels and binarized labels as a function of percent Memory B cells in case sample, with the method name reflected in the plot title. Source data are provided as a Source Data file.



Supplementary Figure 9: **Two heuristics for choosing the number of features (top PCs) for the prediction model part of HiDDEN.** For the dataset with 15% Memory B cells in case: **A** Number of DE genes using HiDDEN-refined binary labels (y-axis) as a function of number of PCs (x-axis) with the optimal (6 PCs) marked by a vertical dashed red line and a star. **B** Receiver Operating Characteristic Curve (ROC) for classification of ground truth cell labels of the model using 6 PCs with AUROC and line style and color as defined in the legend. The second heuristic computes the two-sample Kolmogorov-Smirnov statistic (**Methods**) between the distribution of HiDDEN continuous perturbation scores (p_hat) for HiDDEN-refined binary label 0 and label 1. **C** Density plots of the two distributions of p_hat scores when using 6 PCs with colors as indicated in legend. **D** KS statistic as defined in C (y-axis) as a function of number of PCs (x-axis) with the optimal (45 PCs) marked by a vertical dashed red line and a star. **E** Receiver Operating Characteristic Curve (ROC) for classification of ground truth cell labels of the model using 45 PCs with AUROC and line style and color as defined in the legend. Source data are provided as a Source Data file.

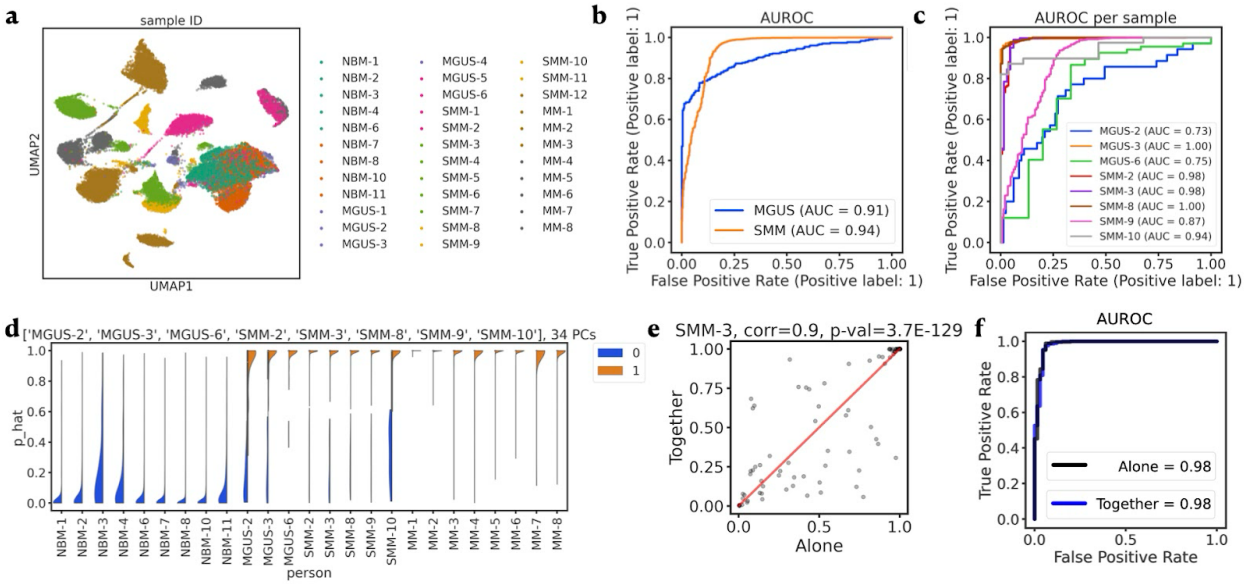


Supplementary Figure 10: **HiDDEN predictions across mixed precursor samples and positive (MM) and negative (NBM) controls. A** Area under the Receiver Operating Characteristic Curves (AUROC) for each mixed precursor sample plotted individually. **B** Violin plot of the distribution of the continuous perturbation score (y-axis) across cells manually annotated as healthy plasma or malignant (x-axis) split over and colored by HiDDEN-refined binary labels 0 and 1; one panel per mixed precursor sample; area scaled by count. **C** Density plots of the distribution of the continuous perturbation score (y-axis) from the batch-sensitive fitting strategy for the mixed precursor SMM-3 across all samples included in the model (x-axis). Colors correspond to HiDDEN-refined labels; area scaled by count.
Part of Supplementary Figure 11/panel B, created in BioRender. Lab, M. (2024) BioRender.com/v73y738. Source data are provided as a Source Data file.

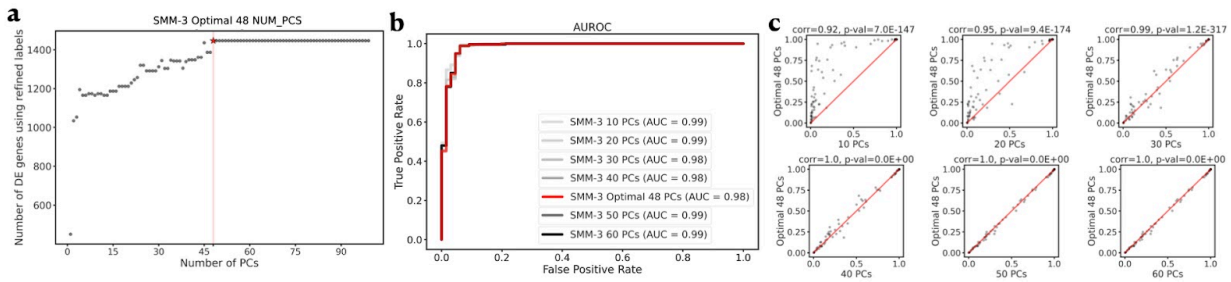


Supplementary Figure 11: **Predictions and computational validation for the low tumor purity MGUS samples considered non-mixed according to the manual annotation**.

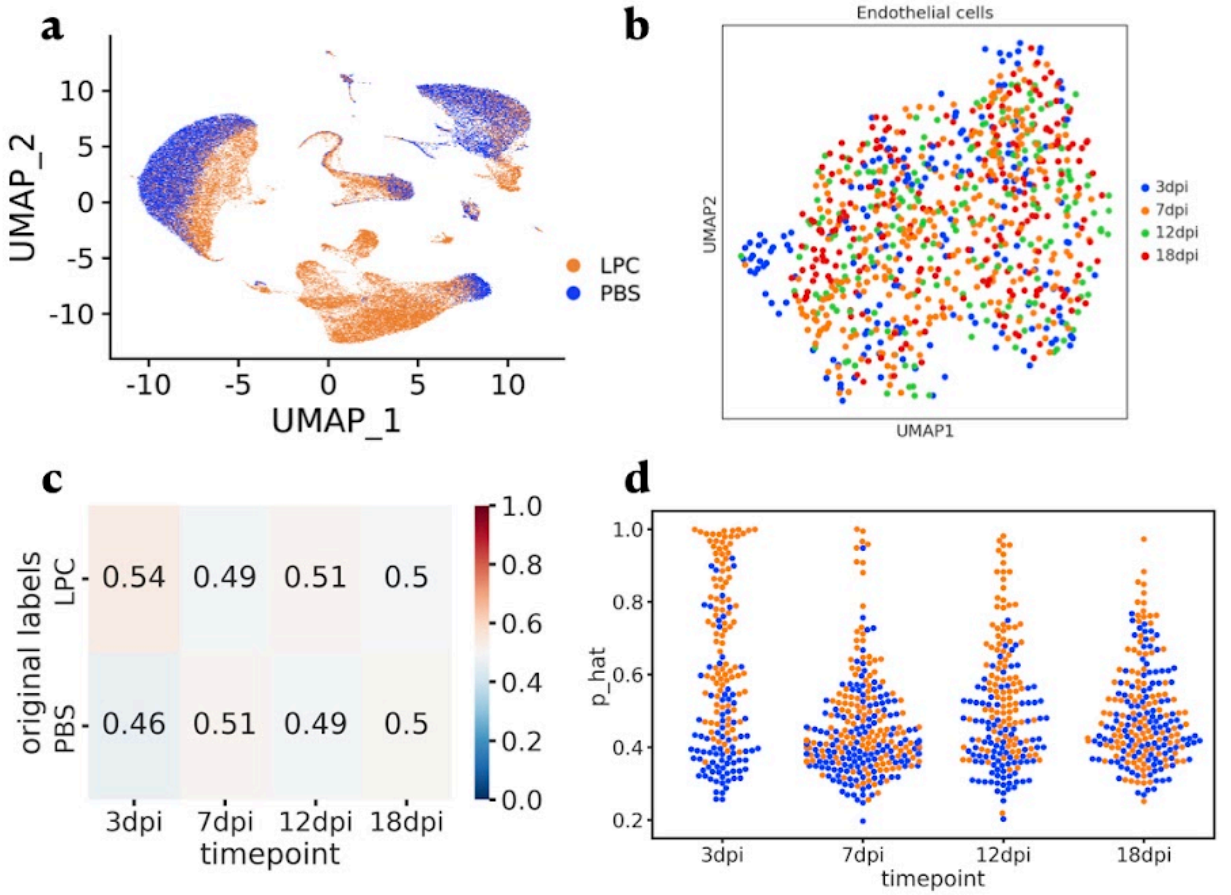
A Violin plots of the distribution of the continuous perturbation score (y-axis) across the three low tumor purity MGUS precursors (x-axis) split over and colored by HiDDEN-refined binary labels 0 and 1; area scaled by count. **B, C, D** Computational validation of cells predicted to be malignant by HiDDEN in low tumor purity MGUS samples. Mean activity (y-axis) of genes assigned to three biologically interpretable signatures from the original study separately computed per HiDDEN label across the three low tumor purity MGUS samples (x-axis).
Part of Supplementary Figure 11/panel A, created in BioRender. Lab, M. (2024) BioRender.com/x94m813. Source data are provided as a Source Data file.



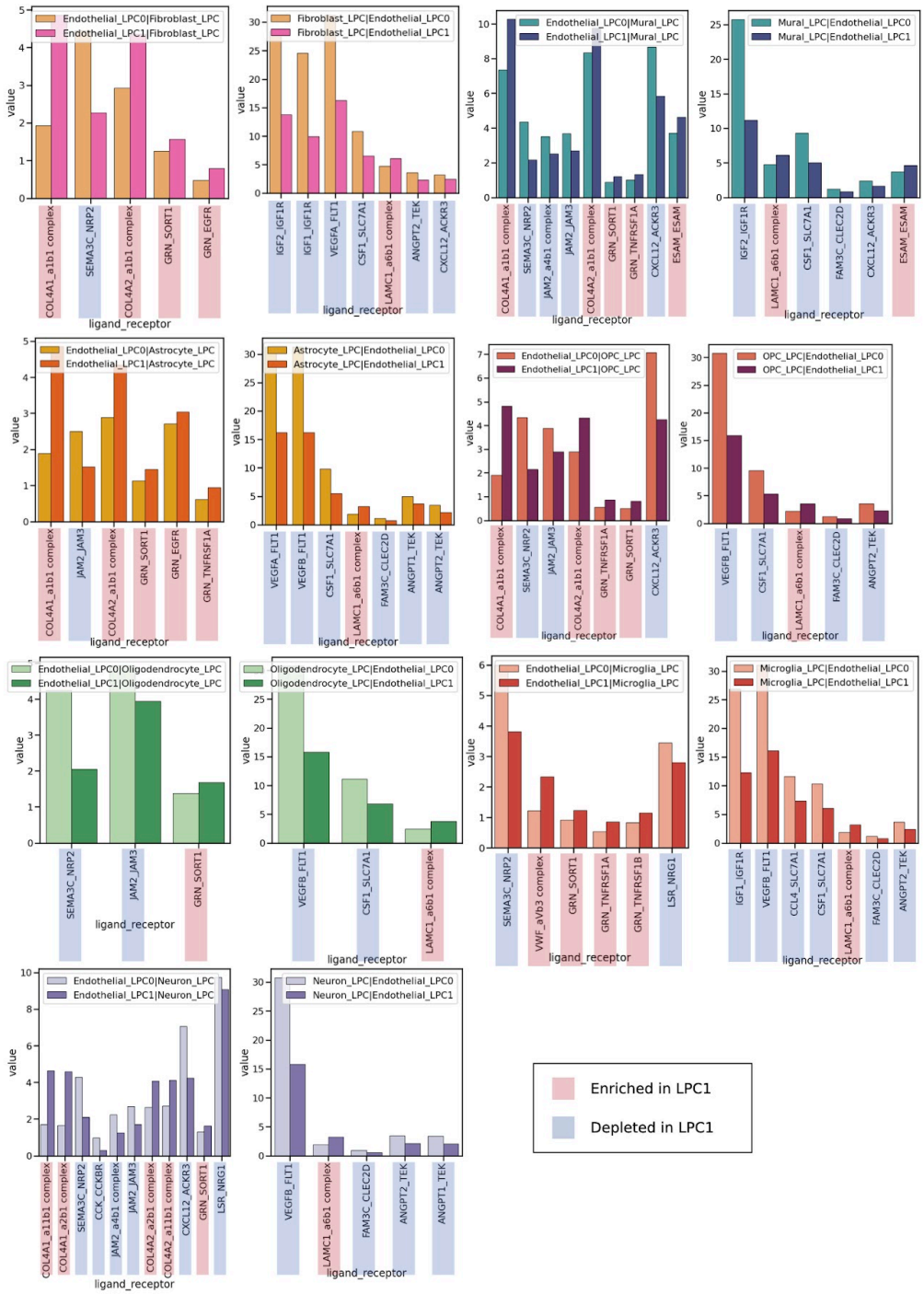
Supplementary Figure 12: **Comparison of batch-sensitive and batch-agnostic fitting strategies for mixed precursor sample SMM-3. A** UMAP embeddings of human bone marrow cells from all NBM, MGUS, SMM and MM patients from [4], colored by patient. **B** AUROC for predicting the per-cell manually annotated malignancy status in mixed samples averaged for each precursor state using the batch-agnostic strategy. Note: Figure 3B shows the results using the batch-sensitive strategy. **C** AUROC for each mixed precursor sample plotted individually using the batch-agnostic strategy. Note: Supplementary Figure 6A shows the results using the batch-sensitive strategy. **D** Density plots of the distribution of the continuous perturbation score (y-axis) from the batch-agnostic fitting strategy for all mixed precursors along with all NBM and all MM across (x-axis). Colors correspond to HiDDEN-refined labels; area scaled by count. Note: Supplementary Figure 10C shows the results using the batch-sensitive strategy for SMM-3. **E** Scatterplot of the per-cell continuous perturbation scores from the batch-sensitive (x-axis) and batch-agnostic (y-axis) strategies for SMM-3 with 45-degree red line, corr=0.9 (p-val=3.7e-129). **F** Comparison of AUROC for SMM-3 with colors corresponding to the two fitting strategies: Alone - batch-sensitive and Together - batch-agnostic. Source data are provided as a Source Data file.



Supplementary Figure 13: **Heuristic for choosing the optimum and sensitivity of the number of features (top PCs) for the bone marrow data illustrated under the batch-sensitive strategy for SMM-3. A** Number of DE genes using HiDDEN-refined binary labels (y-axis) as a function of number of PCs (x-axis) with the optimal (48 PCs) marked by a vertical dashed red line and a star. **B** ROC curves for classification of manually annotated malignancy cell labels in SMM-3 using a different number of PCs (as indicated by the legend labels). **C** A collection of scatterplots of the per-cell continuous perturbation scores using the optimal number of 48 PCs (y-axis) and each of the number of PCs explored in B (x-axis) with 45-degree red line; correlation and corresponding p-value indicated in the title of each plot. Source data are provided as a Source Data file.



Supplementary Figure 14: **Sample-level and HiDDEN predictions across time points in the mouse demyelination data**. **A** UMAP embeddings of non-neuronal cells from both conditions across all time points colored by PBS/LPC sample-level labels. **B** UMAP embeddings of endothelial cells across both conditions colored by time point. **C** Relative abundance of case-control cell identities across time points. **D** Swarmplots of HiDDEN continuous perturbation scores (y-axis) across time points (x-axis); colors as defined in A. Source data are provided as a Source Data file.



Supplementary Figure 15: **Test statistic for contracting endothelial LPC0 and LPC1 ligand-receptor interactions.** A collection of panels each containing a set of plots with two bars. The bars are colored as indicated in each panel's legend and represent an interaction either from or to one of the endothelial LPC subset and a neighboring cell type. X-axis enlisting the names of significant ligand-receptor pairs, each pair colored to indicate whether it is enriched or depleted. Y-axis is the average of the mean ligand expression and mean receptor expression in their respective cell types. Source data are provided as a Source Data file.

Supplementary Tables

% Memory B in case sample	# Memory B in case sample	# Naive B in case sample	# Naive B in control sample
5	49	925	975
10	100	900	1000
15	154	873	1027
20	211	844	1056
25	271	814	1086
30	335	782	1118
35	403	748	1152
40	475	712	1188
45	552	674	1226
50	633	633	1267
55	721	589	1311
60	814	543	1357
65	915	492	1408
70	1023	438	1462
75	1140	380	1520
80	1267	316	1584
85	1404	248	1652
90	1555	172	1728

Supplementary Table 1: **Number of Naive B and Memory B cells in synthetic datasets varying the percent perturbed cells in the case sample.** Rows indicate the percent of Memory B cells in case. Columns are number of Memory B cells in case; number of Naive B cells in case; number of Naive B cells in control.

	Manual annotation				HiDDEN labels				Bayesian purity model
	malignant	healthy	frac malignant	p-value	malignant	healthy	frac malignant	p-value	frac malignant
MGUS-1	0	133	**0**	5.99E-02	30	103	**0.23**	1.63E-12	**0.3**
MGUS-2	35	81	**0.3**	9.63E-02	55	61	**0.47**	7.58E-12	**0.55**
MGUS-3	205	166	**0.55**	4.80E-06	305	66	**0.82**	3.57E-10	**0.71**
MGUS-4	0	62	**0**	7.51E-11	10	52	**0.16**	0.00E+00	**0.51**
MGUS-5	0	53	**0**	2.25E-02	10	43	**0.19**	2.41E-07	**0.33**
MGUS-6	67	15	**0.82**	4.26E-01	61	21	**0.74**	4.13E-01	**0.78**
SMM-2	1721	136	**0.93**	1.73E-03	1776	81	**0.96**	9.28E-16	**0.97**
SMM-3	283	66	**0.81**	7.06E-03	306	43	**0.88**	6.58E-01	**0.82**
SMM-8	579	132	**0.81**	7.81E-02	667	44	**0.94**	1.22E-15	**0.92**
SMM-9	1106	147	**0.88**	3.34E-06	1202	51	**0.96**	8.6E-06	**0.92**
SMM-10	39	28	**0.58**	3.26E-08	53	14	**0.79**	1.25E-01	**0.49**

Supplementary Table 2: **Comparison of manual annotation and HiDDEN outputs across mixed precursor samples.** Rows indicate precursor samples labeled as mixed according to the manual annotation. There are three meta-columns: the first reports statistics using the manual annotation, the second reports statistics using the HiDDEN-refined binary labels, and the third reports the ground truth sample purity point estimate according to the Bayesian purity model. Under the first two meta-columns, there are four columns reporting the number of cells annotated as malignant; number of cells annotated as healthy; the fraction of cells annotated as malignant; and the p-value from comparing the corresponding purity estimate to the ground truth which is calculated using a Beta-Binomial test and is Bonferroni-adjusted (**Methods**).

DE genes	unaffected	affected
HiDDEN_0 vs HiDDEN_1, 3dpi Endothelial cells	'**Atp10a**' '**Igf1r**' 'Pltp' '**Flt1**' 'Ndrg1' '**Sgms1**' '**Plcb1**' 'Ly6c1' '**Tmtc2**' '**Itm2a**' '**Slco1a4**' '**Ccny**' '**Abcc4**' '**Nostrin**' 'Cxcl12' '**Paqr5**' '**Ccdc141**' '**Tsc22d1**' 'Nfkbia' '**Tbc1d4**' '**Spock2**'	'**Vim**' 'Spp1' '**Lgals1**' '**Anxa2**' '**S100a6**' 'Ctsd' '**Rpsa**' 'Tmem252' '**Nid1**' '**Litaf**' '**Ubtd1**' '**Plec**' '**Fkbp1a**' 'Fabp5' '**Fyn**' '**Cd9**' '**Msn**' '**Fmnl2**' '**Adamts9**' '**Ankrd17**' 'Rock2' '**Hif1a**' 'Fabp7' '**Rplp0**' '**Itgb1**' '**Tmsb10**' 'Lyz2' '**Crlf2**' '**Auts2**' '**Suco**' 'Lgals3' '**Sparc**' '**Smad1**' '**Cdh13**' '**Rapgef5**' '**Col4a1**' 'Igfbp7'
PBS vs LPC, 3dpi Endothelial cells	'Ly6c1' 'Cxcl12' 'Pltp' 'Ndrg1' 'Nfkbia'	'Spp1' 'Lyz2' 'Ctsd' 'Fabp5' 'Rock2' 'Tmem252' 'Lgals3' 'Igfbp7' 'Swap70' 'Ctss' 'Fbxo11' 'Fabp7' 'Fryl' 'Magi1'

Supplementary Table 3: **Comparison of DE genes for 3dpi Endothelial cells found using sample-level and HiDDEN labels.** Rows indicate labels used in the DE testing (Methods). Columns indicate DE genes for unaffected (associated with HiDDEN_0 or PBS label) and affected cells (associated with HiDDEN_1 or LPC label). Gene names in **bold** are uniquely found using the HiDDEN-refined labels.