COMPUTATIONAL BIOLOGY

# Annotation-free discovery of disease-relevant cells in single-cell datasets

**Erin Craig[1]†, Timothy J. Keyes[1,2]†, Jolanda Sarno[2,3,4], Jeremy P. D'Silva[1], Pablo Domizi[2], Maxim Zaslavsky[5], Albert Tsai[6], David Glass[7], Garry P. Nolan[6], Trevor Hastie[8,1], Robert Tibshirani[1,8], Kara L. Davis[2,9]***

Copyright © 2025 The Authors, some rights reserved; exclusive licensee American Association for the Advancement of Science. No claim to original U.S. Government Works. Distributed under a Creative Commons Attribution NonCommercial License 4.0 (CC BY-NC).

**In single-cell datasets, patient labels indicating disease status (e.g., "sick" or "not sick") are typically available, but individual cell labels indicating which of a patient's cells are associated with their disease state are generally unknown. To address this, we introduce mixture modeling for multiple-instance learning (MMIL), an expectation-maximization approach that trains cell-level binary classifiers using only patient-level labels. Applied to primary samples from patients with acute leukemia, MMIL accurately separates leukemia from nonleukemia baseline cells, including rare minimal residual disease (MRD) cells; generalizes across tissues and treatment time points; and identifies biologically relevant features with accuracy approaching that of a hematopathologist. MMIL can also incorporate cell labels when they are available, creating a robust framework for leveraging both labeled and unlabeled cells. MMIL provides a flexible modeling framework for cell classification, especially in scenarios with unknown gold-standard cell labels.**

## INTRODUCTION

A common goal in the study of single-cell biology is to identify rare cell populations associated with human disease (*1*). This is because, when analyzing clinical samples collected from patients, we often know whether a person has a disease, but we do not necessarily know which cells in their body (which contains a heterogeneous mixture of disease-associated and baseline cells) play a role in their illness. Because single-cell heterogeneity is a hallmark of human tissues in major diseases such as cancer (*2*), autoimmunity (*3*), and infection (*4*), the ability to accurately identify disease-associated cells has far-reaching implications in systems biology, diagnostics, and the development of targeted therapeutics.

However, classifying individual cells as disease-associated or not disease-associated can be difficult. In clinical scenarios, highly trained pathologists spend hours examining tissue samples using numerous assays to determine whether disease-associated cells are present, to identify the features that make such cells abnormal, and to decide whether their observations are of diagnostic or prognostic relevance. Furthermore, the increasing availability of high-parameter single-cell technologies in research laboratories has enabled deep cellular profiling of patient samples at an unprecedented scale (*5*). In this context, determining gold-standard cell labels—that is, manually assigning cells to the categories "disease associated" or "not disease associated"—can be prohibitively challenging or expensive, and it is not obvious how to predict such cell labels in a data-driven manner without labeled training data. This is particularly true in the context of complex diseases, for which true disease-associated cell populations remain unidentified, making gold-standard cell annotation impossible.

Recent efforts have proposed solutions to these challenges through reference mapping. For example, Seurat (*6*) uses a weighted nearest-neighbors approach to map cells to annotated reference datasets, and SCimilarity (*7*) leverages a large, pretrained foundation model to find similar cells in a labeled atlas. While these methods are powerful, they rely on the availability and completeness of labeled reference datasets. In many settings, such a dataset is unavailable, such as when disease-associated populations are poorly characterized or completely unknown.

Here, we present mixture modeling for multiple-instance learning (MMIL; pronounced as "Emile" - /əˈmiːl/), a method that, armed only with patient disease status, uses expectation-maximization (EM) (*8*) to train a classifier to label individual cells as disease-associated or baseline. Multiple-instance learning (MIL) is designed for datasets consisting of multiple instances (in our case, cells) from each unit (patient), and the typical goal is to predict unit labels (patient disease status). Unlike most other MIL algorithms, MMIL's primary goal is to predict cell labels using a model trained only with patient disease status (Fig. 1A), and, unlike many other single-cell methods, MMIL does not require a labeled reference dataset.

MMIL assumes that all cells from healthy donors are baseline cells, and cells from patients may be disease-associated or baseline. At a high level, MMIL uses similarity of a cell from a patient to cells from healthy donors to estimate the probability that the cell is disease associated. To do this, MMIL alternately estimates cell labels and trains a classifier; at each iteration, the classifier and cell label estimates improve until the algorithm converges on a solution. MMIL is versatile and can be easily implemented as a wrapper around any classifier trained by optimizing the binomial log likelihood, including lasso logistic regression models, gradient-boosted trees and neural networks. When available, MMIL naturally incorporates any gold-standard cell labels; they guide model fitting when used alongside a dataset of unlabeled cells. MMIL can be used to improve the model calibration of any cell-level model, making predictions more interpretable by aligning estimated probabilities of disease-association with their prevalence.

[1]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [2]Division of Hematology, Oncology, Stem Cell Transplant and Regenerative Medicine, Department of Pediatrics, Stanford University, Stanford, CA, USA. [3]Tettamanti Center, Fondazione IRCCS San Gerardo dei Tintori, Monza, Italy. [4]School of Medicine and Surgery, University of Milano-Bicocca, Milan, Italy. [5]Department of Genetics, Stanford University, Stanford, CA, USA. [6]Department of Pathology, Stanford University, Stanford, CA, USA. [7]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, Seattle, WA, USA. [8]Department of Statistics, Stanford University, Stanford, CA, USA. [9]Center for Cancer Cell Therapy, Stanford University, Stanford, CA, USA.
*Corresponding author. Email: kardavis@stanford.edu
†These authors contributed equally to this work.

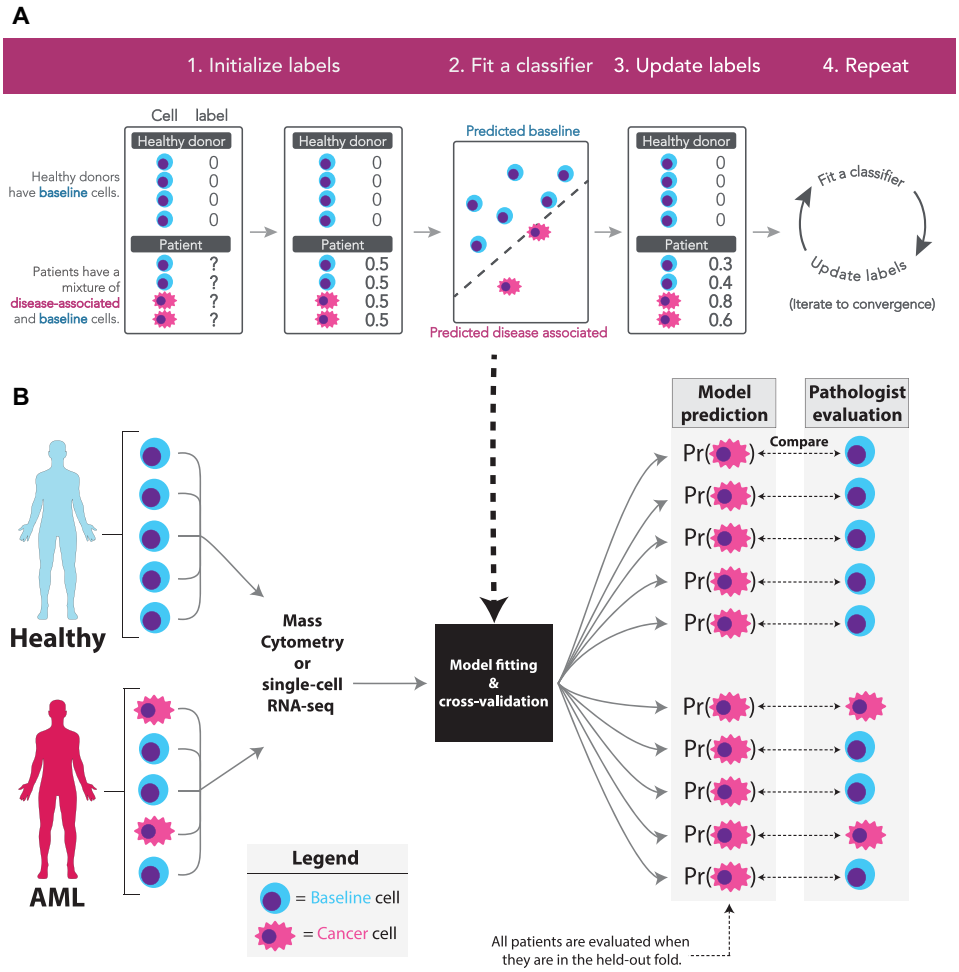


**Fig. 1. Mixture modeling for multiple-instance learning.** (**A**) Process to train a mixture model for multiple instance learning (MMIL). We initialize the patient's cell labels as 0.5: In this example, we assume that the prevalence of disease-associated cells in patients (ρ) is 50%, so each cell is initially given a 50/50 chance of being disease associated. After training the first classifier, we improve our estimates of the patient's cell labels. This process is repeated until convergence. (**B**) Schematic of model training and evaluation on the AML cohorts from (*13*) and (*20*).

Herein, we demonstrate MMIL's application to the detection of leukemia cells both at diagnosis and after initiation of treatment. At diagnosis, determination of acute leukemia is typically straightforward given the malignant proliferation of immature white blood cells but, in some cases, can be challenging due to phenotypic characteristics resembling the baseline cells from which they originate (*9*). We use MMIL to train lasso logistic regression models with single-cell data collected from patients with either acute myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL) as well as healthy controls. In both diseases, we demonstrate MMIL's ability to (i) identify cancer cells accurately; (ii) generalize to cells sampled from different patients, different time points after initiation of treatment, and different tissues than those on which the models were originally trained; and (iii) select single-cell features known to be associated with leukemia. We demonstrate MMIL's strong performance when provided only with unlabeled data and its ability to incorporate partially labeled data as well.

We further demonstrate MMIL's performance in disease monitoring by its accurate identification of minimal residual disease (MRD) cells in ALL. Identification of treatment-resistant MRD cells in clinical contexts is challenging and labor-intensive, even among highly trained clinicians using state-of-the-art blood profiling technologies (*10*, *11*), yet critical, as their presence increases relapse risk and guides treatment stratification (*12*). Last, we demonstrate how MMIL can be used for discovery, as we show its prediction of MRD-associated cells from ALL diagnosis samples and how these cells are enriched at relapse from the same patients.

The applications in leukemia diagnosis and treatment monitoring demonstrate that MMIL has broad implications in the field of single-cell analysis. Because MMIL is a direct, flexible method for identifying disease-associated cells using only patient labels, it has the potential to reveal biological insights that inform patient monitoring, treatment, and diagnosis.

## RESULTS

### An algorithm to classify cells without complete cell labels

Given a classification task to distinguish healthy from disease-associated cells, typically, we would (i) gather a dataset of cells where we know which cells are disease associated and which are baseline

and (ii) train a model to predict cell disease status using tools such as logistic regression, random forests, or neural networks. This workflow relies on knowing which cells are disease associated and which are not. However, in clinical and research settings, we frequently do not know which cells are disease associated; we only know that cells sampled from healthy people are baseline. MMIL is an algorithm that trains classifiers in this setting.

MMIL is an iterative process that relies on the following intuition. If we had a trained classifier, then we could use it to predict which cells are baseline and which are not; and if we had a prediction of which cells were baseline and which were not, then we could use these predicted labels to train a classifier. MMIL fits a classifier by alternating between these two steps and is described in Fig. 1A and Algorithm 1 (and in more detail in Materials and Methods, Algorithm 2, and Supplementary Methods). MMIL hinges on the assumption that cells from healthy people are not disease associated; this allows the model to characterize baseline cells. Disease-associated cells, then, are cells that are distinct from the baseline class. MMIL uses two parameters, $\rho$ and $\zeta$, that are not identifiable; see Supplementary Methods for details. For the parameter $\rho$, the proportion of baseline cells in patients, we performed a sensitivity analysis and found that using its correct value improves the test area under the receiver operating characteristic curve (AUROC), but the performance difference from using an incorrect value was relatively minor (fig. S1A). Further, the test observed log likelihood tends to favor the true $\rho$ (fig. S1B), which suggests that the log likelihood may be useful in selecting $\rho$. The second parameter $\zeta$, the fraction of patient-derived cells in the prediction population, is often directly estimable from data. Note that Algorithm 1 assumes that a classifier can be trained using soft (probabilistic) labels; see Supplementary Methods for a modification for learning algorithms that require hard labels (0 or 1).

**Input**: Cells with labels indicating whether they originated from a patient or healthy donor.

**Assumptions**: The proportion of disease-associated cells in patients is $1 - \rho$.
Cells from healthy donors are not disease-associated.
The proportion of cells from patients in the prediction population is $\zeta$.

1. Create a dataset with:
   - Cells from healthy donors, assigned the label $y = 0$ (baseline).
   - Cells from patients, labeled $1 - \rho$ (prior probability of being disease-associated).

   *The labels reflect uncertainty. We are confident that healthy donors' cells are not disease-associated. But we do not know whether patients' cells are disease-associated – the best we can do is assume that each cell is disease-associated with probability* $1 - \rho$.

2. Alternate:
   a. Train a classifier using the augmented dataset.
      *The classifier characterizes the baseline class using the data from healthy donors; the disease-associated class consists of cells that are different from the baseline class.*

   b. Use the trained classifier to predict the probability that each patient's cell is disease-associated. Update their labels using these predictions, $\rho$ and $\zeta$. (See Methods for details.)
      *After training, we use the classifier to improve our guess for whether each patient's cell is disease-associated. Then, the new guesses are used to train a new classifier.*

   Stop alternating when the model predictions stop changing.

Output: A fitted model to predict labels for new cells.

**Algorithm 1. Mixture modeling for multiple-instance learning (MMIL) is a method to learn cell labels using patient labels.** Our dataset consists of cells with labels indicating whether they were sampled from patients or healthy donors. We assume that cells from healthy people are not disease associated: This assumption allows us to train a model that uses these cells to characterize the "baseline" class. The model can then tease out the "disease-associated" class by finding cells from patients that are distinct from the baseline class. This approach is an application of the expectation-maximization (EM) algorithm.

**Algorithm 2 MMIL: Mixture Models for Multiple Instance Learning**
**Train a mixture model for multiple instance learning data**

**Input**: Covariates $X \in \mathbb{R}^{n \times p}$.
Binary person labels $z \in \mathbb{R}^n$:
$n_1$ are positive *(patients)*
$(n - n_1)$ are negative *(donors)*.

**Assumptions**: The proportion of baseline cells in patients is $P(y=0 \mid z=1) = \rho$.
Cells from healthy donors are not disease-associated: $P(y=1 \mid z=0) = 0$.
The proportion of cells from patients in the prediction population is $P(z=1) = \zeta$.

1. Create a label vector $y^{(0)}$.
   Set $y^{(0)}$ to 0 for cells from donors and to $1 - \rho$ for cells from patients.

2. Iterate to convergence. At the $i^{th}$ step:

   ***a. Maximization step:*** Fit a model $\eta^{*(i)}(x)$ using X and $y^{(i-1)}$.
   Account for biased sampling: make the case-control intercept adjustment to obtain:

$$\eta^{(i)}(x) = \eta^{*(i)}(x) + \log\left(\frac{(1-\rho)n_1}{n-(1-\rho)n_1}\right) - \log\left(\frac{(1-\rho)\zeta}{1-(1-\rho)\zeta}\right).$$

   ***b. Expectation step:*** Use $\eta^{*(i)}(x)$ to define $y^{(i)}$.
   Cells from donors have known label $y^{(i)} = 0$.
   Cell j from a patient has the label $y_j^{(i)}$ such that:

$$\operatorname{logit} y_j^{(i)} = \eta^{(i)}(x_j) - \log\left(\frac{\rho\zeta}{1-(1-\rho)\zeta}\right).$$

Output: A fitted model $\hat{\eta}(x)$ to predict labels for new cells.

**Algorithm 2. Train a mixture model for multiple instance learning data.**

## MMIL identifies cancer cells in AML using only unlabeled cells

We applied MMIL to two publicly available single-cell datasets collected from patients with AML (Fig. 1B). First, we consider a mass cytometry (CyTOF) dataset of 800,000 cells collected from 13 patients with AML at the diagnosis time point and three healthy bone marrow controls (*13*). Cells were analyzed for the presence of 32 surface proteins and 11 intracellular proteins (*14*, *15*). Each AML sample is a mixture of cancer cells and baseline cells, and this cohort contains patients with a wide range of leukemic cells in the bone marrow (ranging from 89.2 to 13.9%). Typically, a hematopathologist must analyze a blood or bone marrow specimen to diagnose and track the progression of disease over time. Hematopathologists do this by manually enumerating the proportion of leukemic blasts in the sample based on each cell's size, shape, and expression of leukemia-associated markers measured by microscopy or flow cytometry. Particularly in the context of myeloid leukemias, blast enumeration can be a nuanced and challenging process, with inherent variability even among experienced pathologists (*13*, *14*, *16*).

Using this dataset, we trained three models: (i) the "oracle" model, a lasso logistic regression model trained using pathologist-provided, gold-standard annotations; (ii) the "naive" model, a lasso model trained using only patient labels; and (iii) a lasso model trained with MMIL. All three models were implemented as lasso-regularized logistic regression models to facilitate feature selection, with the goal of identifying specific subsets of features associated with acute leukemia. When training the MMIL model, we assumed that 75% of cells in patients with cancer are baseline on the basis of the clinical guideline that patients with at least 25% of blasts in their bone marrow receive a leukemia diagnosis (*9*). We estimated model performance using leave-one-patient-out cross-validation (LOOCV). In each fold of our cross-validation procedure, models were trained using all three healthy control marrows and 12 of the 13 AML marrows; then, the performance of each model was evaluated by comparing predictions on the held-out patient's cells to the pathologist's gold-standard labels (Fig. 1, A and B). To assess MMIL's robustness to the choice of $\rho$, we performed a sensitivity analysis (fig. S2) and found that values above $\rho = 0.375$ all had similar AUROC across held-out patients. In

addition, we compared the performance of the MMIL model to several additional baselines: (i) two unsupervised clustering algorithms [PhenoGraph (*17*) and FlowSOM (*18*)] commonly used to analyze mass cytometry data; (ii) a weighted *K*-nearest neighbor (KNN) classifier, commonly used in single-cell data analysis; and (iii) an alternative MIL model implemented using the milr R (MILR) package (*19*), which applies logistic regression to identify relevant features in weakly-labeled data.

At the cell level, MMIL achieved a mean AUROC of 0.751 across all held-out patients during the cross-validation. By comparison, the naive model achieved worse performance with an AUROC of 0.658, and the oracle model achieved superior performance with an AUROC of 0.945, as expected given its access to cell labels during training (Fig. 2A). MMIL outperformed both PhenoGraph and FlowSOM, which achieved AUROC values of 0.602 and 0.616, respectively (fig. S3). It also outperformed the KNN classifier, which achieved an AUROC of 0.652, 0.681, and 0.711 with *K* values of 10, 25, and 100, respectively (fig. S4) and the MILR model, which achieved an AUROC of 0.490 (fig. S5). At the patient level, when a probability threshold of 0.5 was used to classify cells as leukemic blasts (see Materials and Methods), the blast percentages that MMIL assigned to each patient in the dataset correlated strongly with pathologist-assigned blast percentages (Pearson $\rho = 0.88$ compared to the oracle model's $\rho = 0.95$). The naive model's blast percentage assignments

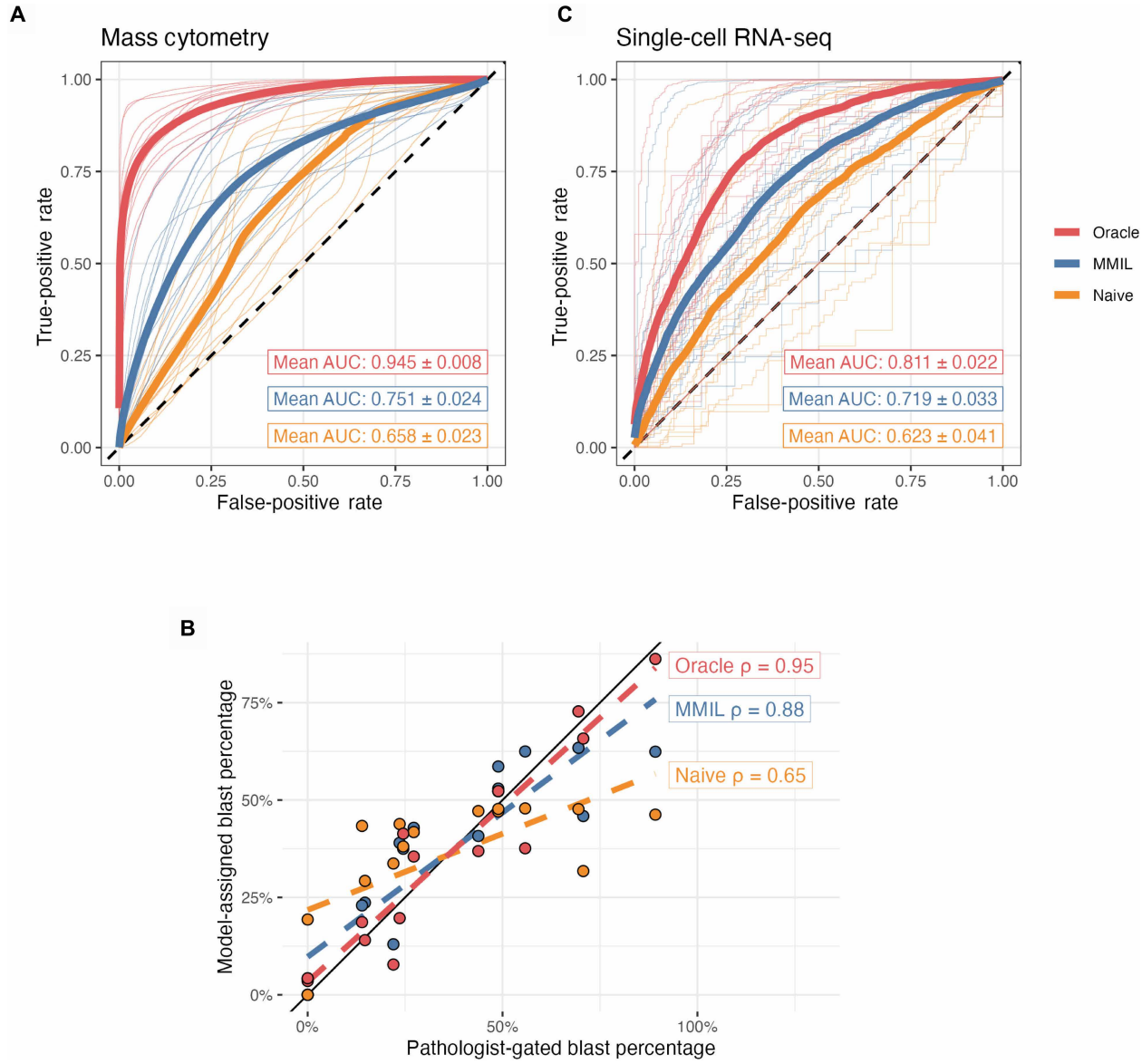


**Fig. 2. MMIL detects cancer cells in AML using patient labels only.** (**A**) Receiver operating characteristic (ROC) curves demonstrating individual (thin) and average (thick) performance of oracle, MMIL, and naive models trained to detect leukemic blasts in AML CyTOF dataset. Insets indicate mean area under the ROC curve (AUC) across all patients. (**B**) Scatterplots representing the relationship between the gold-standard, pathologist-enumerated blast percentage for each patient (*x* axis) and the model-assigned blast percentage for each patient for the oracle (red), MMIL (blue), and naive (yellow) lasso models. Inset text represents the Pearson correlation coefficients between the values on the *x* and *y* axes. (**C**) ROC curves demonstrating individual (thin) and average (thick) performance of oracle, MMIL, and naive models trained to detect leukemic blasts in AML scRNA-seq dataset. Insets indicate mean AUC across all patients.

correlated less with the pathologist gold-standard (Pearson $\rho = 0.65$) (Fig. 2B).

MMIL is designed to be compatible with high-dimensional settings, but, as with many methods, performance may degrade as dimensionality increases. In simulations (Supplementary Methods), we find that applying principal components analysis (PCA) before model fitting substantially improves MMIL's performance (fig. S6, A and B), especially when the principal components (PCs) are computed using patient data only. To determine MMIL's performance with a true high-dimensional dataset, used the single-cell RNA-sequencing (scRNA-seq) dataset from van Galen *et al.* (*20*) using a similar experimental design as applied to the CyTOF dataset above. After filtering, the dataset consisted of 22,600 cells from 21 individuals, five healthy donors, and 16 patients with AML. We used PCA to reduce dimensionality and used the first 15 PCs as input features for MMIL (fig. S6C). MMIL again outperformed the naive model in LOOCV (Fig. 2C), indicating MMIL's applicability across common single-cell technologies.

In their original study, Tsai *et al.* (*13*) identified two sets of markers that could be used to distinguish between baseline myeloid and AML cell populations. Among these were the cytoskeletal proteins β actin, lamin A/C, and lamin B1; the granule-associated proteins serpinB1, VAMP-7, ribosomal RNA (rRNA), lactoferrin, and lysozyme; CD45; and the cell size marker wheat germ agglutinin (WGA) lectin. Two of the most important markers for the pathologists' AML blast enumeration were rRNA and lactoferrin. Figure 3 (A and B) shows that lactoferrin, lamin A/C, CD45, and rRNA were selected by both the MMIL and oracle models, whereas serpinB1 and WGA were uniquely selected by the MMIL model. However, lamin B1 and β actin were uniquely selected by the oracle lasso model, and only the oracle model selected rRNA with a positive coefficient. Thus, MMIL closely follows but does not entirely reproduce the pathologist's decision-making process, suggesting an ability to augment (not just replicate) existing clinical decision processes.

To visualize MMIL's predictions, we constructed several uniform manifold approximation and projection (UMAP) (*21*) plots using all cells and markers in the dataset (Fig. 3, C to F). Cells proximal to each other in UMAP space have similar MMIL probabilities (Fig. 3C) and have similar pathologist labels (Fig. 3D), indicating that MMIL identifies phenotypically coherent cell populations. Regions of UMAP space with high MMIL probabilities correspond to regions containing large numbers of pathologist-identified AML blasts. However, MMIL does not simply assign high probabilities to all cells collected from patients with AML (Fig. 3E and fig. S7); rather, it assigns higher probabilities to cells in areas occupied primarily by cells from patients with cancer, but not from healthy donors (Fig. 3F).

Protein-specific UMAP plots further indicate that MMIL assigns high disease-association probabilities to cells with unique marker expression patterns, generally, those low in lactoferrin and high in serpinB1, CD16, and/or CD56. Furthermore, while rRNA was not explicitly selected as a disease-associated feature in the MMIL model, cells with high probabilities also tended to express high levels of rRNA (figs. S7 and S8).

### MMIL identifies cancer cells in AML using both labeled and unlabeled cells

Next, we sought to evaluate MMIL's performance in the setting where both labeled and unlabeled cells are present, a scenario in which most patients with leukemia in a dataset do not have cell labels, but a small number of them do have cell labels. We hypothesized that a model capable of leveraging both the expert labels of a clinician and a larger corpus of unlabeled data would outperform models trained using only labeled or only unlabeled data (*22*).

However, using expert labels to train a model on a difficult classification task can be risky: Clinical misidentification of leukemic blasts is a known challenge in hematopathology, with a high degree of documented interobserver variability and methodological disagreement (fig. S9) (*16*). This is particularly concerning for training machine learning models, as the degradation of model performance due to label error is a widely characterized problem (*23*, *24*). Thus, we also hypothesized that models trained on both labeled and unlabeled data would be more robust to label error than models trained on labeled data alone, as they would be less susceptible to overfitting on the imperfect labels.

To test these hypotheses, we developed a semisupervised version of MMIL (termed "1-shot MMIL") using the training procedure illustrated in Fig. 4A. Briefly, 1-shot MMIL is trained identically to MMIL with one change: During model training, the gold-standard cell labels for a single patient with AML (termed the "1-shot patient") are used for supervision instead of probabilistic labels. Like the cells from healthy donors, the 1-shot patient's cell labels remain fixed throughout all iterations of the EM. Otherwise, model training and LOOCV are carried out as before.

To benchmark 1-shot MMIL's performance, we compare it to 0-shot MMIL (MMIL as presented in Fig. 1A), the naive model (a naive model trained as described before), the 1-shot naive model (a model in which the 1-shot patient's gold-standard cell labels are used during model training, but all cells from other patients use their inherited patient labels), and the 1-shot oracle model (a model trained on the healthy donors' cells and the 1-shot patients' cells using their gold-standard cell labels). Furthermore, to benchmark 1-shot MMIL's robustness to imperfect labeling, we again fit each 1-shot model after randomly permuting 25% of the 1-shot patient's cell labels, a proportion chosen to match the variability observed between pathologists when enumerating leukemic blasts in AML (fig. S9 and Materials and Methods). As before, all models were lasso-regularized logistic regression classifiers.

We performed 13 1-shot experiments: One for each patient with AML in the dataset to serve as the 1-shot patient. The results of these experiments are summarized in Fig. 4 (B to E). In the 0-shot case, we once again observed that MMIL outperformed the naive model (Fig. 4B, left). In the 1-shot case, we found that MMIL's performance improved substantially compared to 0-shot MMIL (average AUROC increase of 0.058) while maintaining its superior performance to the 1-shot naive model (Fig. 4B, middle). We also found that training an oracle model using only the 1-shot patient (and removing any unlabeled data) outperforms 1-shot MMIL; however, after training labels were permuted, 1-shot MMIL outperforms both the naive model and the oracle model, whose performance degrades substantially after training on the imperfect labels (Fig. 4B, right). Further, using simulated data, we tested the impact of adding labeled patients on model performance and found that, as expected, performance improved with each additional labeled patient but plateaued after about seven to eight labeled patients (fig. S10). Together, these results suggest that MMIL can improve its performance using even a small number of gold-standard labels while remaining more robust to noisy labeling than alternative models.

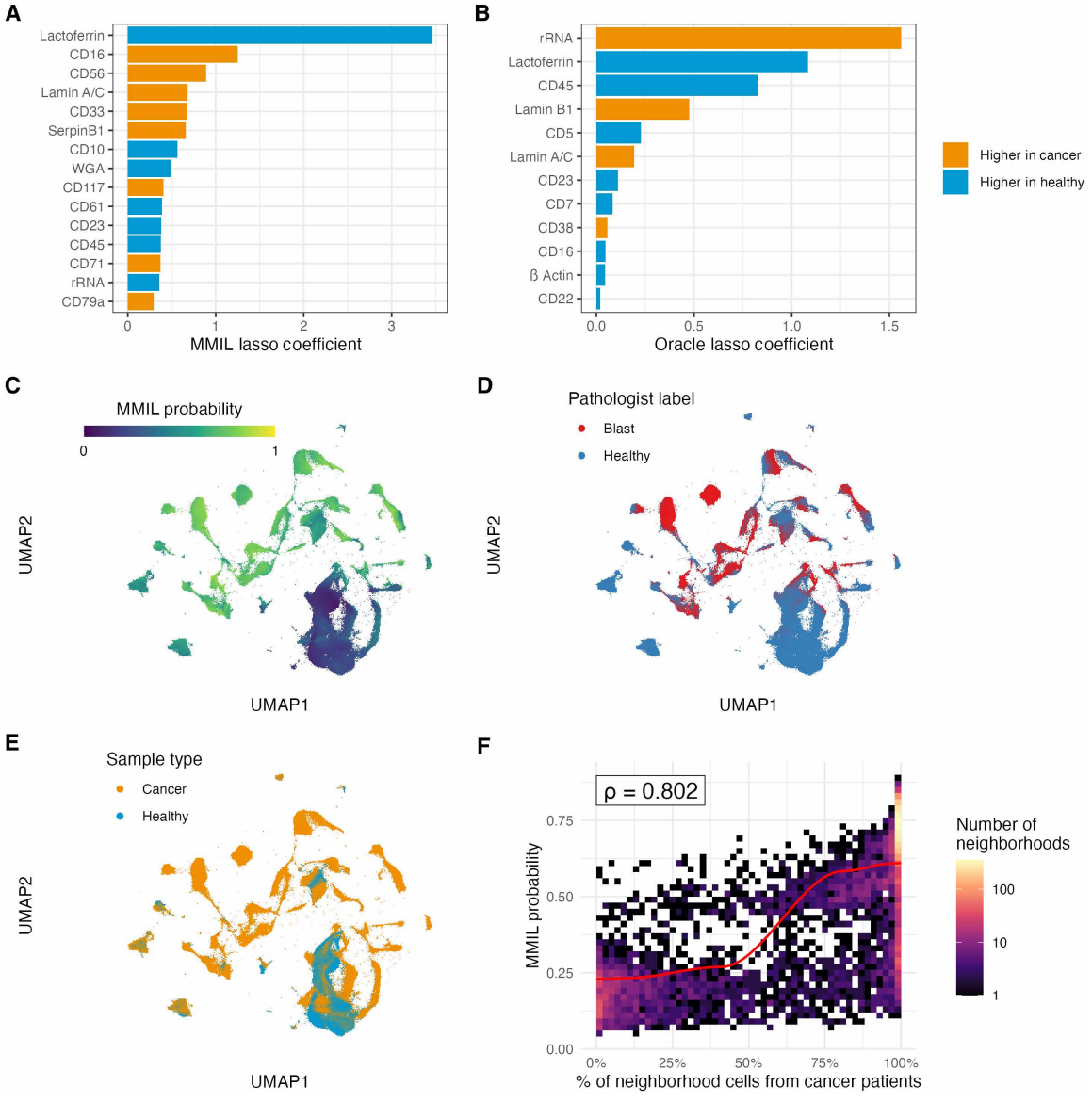


**Fig. 3. MMIL identifies regions of high-dimensional phenotype space containing cells from patients with AML, but not healthy controls.** (**A**) Nonzero coefficients for the MMIL model trained to detect leukemic blasts in AML. (**B**) Nonzero coefficients for the "oracle" lasso model trained to detect leukemic blasts in AML. (**C**) A scatterplot of uniform manifold approximation and projection (UMAP) coordinates colored by MMIL probabilities. Cells with scores of 0 have a small chance of being AML-associated (i.e., leukemic blasts), whereas cells with probability scores near 1 have a high chance of being AML-associated. (**D**) UMAP plot as in (C), but with cells labeled as leukemic blasts in red and cells annotated as baseline cells by a pathologist in blue. Note the general agreement of probabilities in (C) to red regions in (D). (**E**) UMAP plot as in (C), but with cells from patients with cancer shown in orange and cells from healthy controls in blue. Note that regions with overlapping orange and blue cells are assigned low MMIL probabilities in (C). (**F**) Count heatmap of two-dimensional bins demonstrating the correlation between the average MMIL probability in a phenotypic neighborhood (*y* axis) and the proportion of cells from patients with cancer it contains (*x* axis). Bins are colored by the density of neighborhoods in that region, and the red line represents the locally weighted moving average across the *x* axis. Inset text indicates the Spearman correlation between the values on the *x* and *y* axes. In (A) to (E), UMAP coordinates were calculated using all protein markers.

On the basis of these results, we wondered how the inclusion of the 1-shot patient during model training affected the feature sets selected by MMIL. To interrogate this, we examined the 1-shot experiment with the largest average AUROC improvement between the 0-shot and 1-shot models across all patients (patient AML-5An was the 1-shot patient). In this experiment, rRNA was selected as the most cancer-associated feature by the oracle model (Fig. 4C). rRNA was not identified as a cancer-associated marker by the 0-shot MMIL model but was identified by the 1-shot model (Fig. 4D). In addition, several markers (CD64, CD3, CD8, and CD19) that were included in the 0-shot MMIL model but that were not used in manual gating of leukemic blasts of Tsai *et al.* (*13*) were removed in the 1-shot model. This held as a general trend across all 1-shot experiments, across which rRNA was never selected as a cancer-associated marker by 0-shot MMIL but was selected as a cancer-associated marker in 10 of the 13 corresponding 1-shot MMIL models. At the

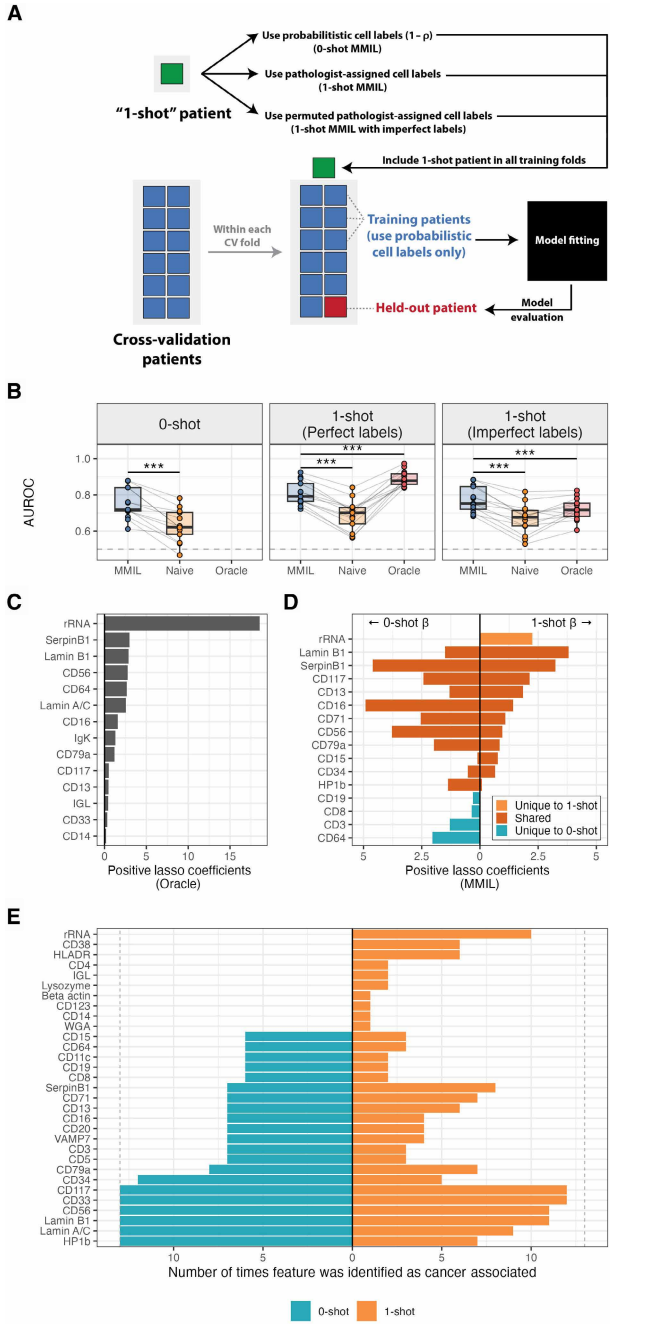


**Fig. 4. MMIL can train on labeled and unlabeled data simultaneously to incorporate expert knowledge while remaining robust to imperfect labeling.** (**A**) Schematic of semisupervised 0-shot and 1-shot MMIL experiments (also see Materials and Methods). (**B**) Boxplots indicating average AUROC for MMIL (blue), naive (orange), and oracle (red) models across 0-shot (left), 1-shot (with perfect labels; middle), and 1-shot (with imperfect labels; right) training procedures. ***$P < 0.0001$ using a paired Student's *t* test with Benjamini-Hochberg correction for multiple comparisons. (**C**) Positive lasso coefficients for an oracle model fit on a single patient (AML-5An). (**D**) Positive MMIL coefficients after 0-shot (left) and 1-shot (right) learning. Note that rRNA, the feature with the largest oracle lasso coefficient in (C) and Fig. 3B, was selected with a positive coefficient only after 1-shot learning. (**E**) Two-sided bar plot indicating how many times a feature was included in the MMIL model with a positive coefficient after 0-shot (left, blue) and 1-shot (right, orange) training. Dashed gray lines indicate the maximum number of times a feature could have been included (13, the total number of 1-shot experiments).

same time, markers CD3, CD5, CD15, and CD11c were often selected by 0-shot MMIL but are not canonically considered markers of AML blasts, and these were selected less frequently by 1-shot MMIL. This change in feature selection suggests that providing MMIL with expert knowledge allows it to prioritize features of known diagnostic relevance (such as rRNA) while downweighting markers that do not align with an expert's input. Together, these results demonstrate that MMIL can adapt its feature selection process via semisupervision, leveraging even a small number of expert labels to prioritize clinically relevant markers over noninformative ones.

## MMIL identifies and tracks cancer cells throughout treatment progression in ALL

We next applied MMIL to a dataset of 1.1 million blood and bone marrow cells obtained from three patients with ALL and three healthy controls. This dataset contains samples collected from multiple tissues (blood and bone marrow) and multiple time points (diagnosis, day 8 posttreatment initiation, and day 15 posttreatment initiation). Using this dataset, we evaluated the ability of an MMIL model trained with pretreatment cells to accurately identify and track residual leukemic cells during treatment.

The identification of residual leukemic cells throughout treatment is crucial for assessing a patient's risk of relapse and, therefore, for clinical decision-making. Thus, MMIL's ability to maintain strong performance despite distributional shift, i.e., a shift in the distribution of leukemic cell phenotypes over time due to treatment effects or changes in disease state, would be paramount for its application to disease monitoring in cancer.

To evaluate MMIL's ability to detect residual leukemic cells over the course of treatment, we analyzed all cells in the ALL cohort for the presence of 29 proteins using CyTOF. Gold-standard labels indicating which cells were leukemic blasts were provided by expert gating. Oracle, naive, and MMIL models were trained using only bone marrow cells from the diagnostic time point, and model performance for cell label prediction was once again evaluated using LOOCV. Next, we applied our diagnostic models to paired blood samples from the diagnostic time point to evaluate each model's ability to generalize to a different tissue context that is more easily monitored clinically. Then, we applied each model to paired blood samples collected at days 8 and 15 posttreatment initiation to evaluate its ability to track a patient's disease burden over time.

Trained without access to gold-standard labels, MMIL demonstrates robust performance and generalizability, with an AUROC of 0.941 at diagnosis and 0.815 at day 15 posttreatment initiation. The naive model begins with AUROC of 0.923 at diagnosis, but this degrades across time to 0.662 at day 15 (Fig. 5). This decline in performance aligns with clinical challenges encountered over the course of a patient's treatment, including the diminishing proportion of leukemic blasts over time (often less than 0.1% after treatment) and phenotypic plasticity in resistant cells posttherapy (*25*). Notably, the naive model's brittle performance in response to these shifts highlights its susceptibility to false positives. MMIL, however, retains strong performance even as the percentage of leukemic cells becomes low, suggesting its potential for robust disease monitoring and MRD detection in leukemia.

As before, we analyzed MMIL's coefficients to examine which protein markers best distinguish between baseline and leukemic

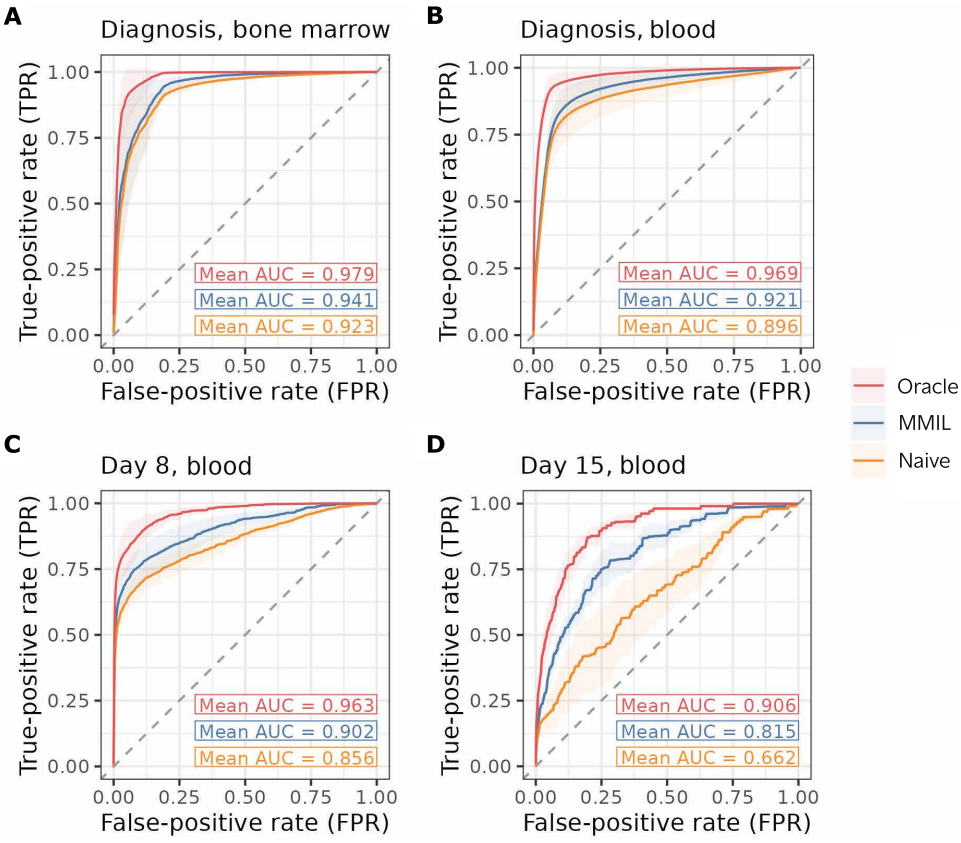


**Fig. 5. MMIL identifies leukemia cells across distinct tissues and treatment time points in pediatric ALL without training on known cell labels.** We compare the performance of MMIL to the oracle model (trained using gold-standard labels) and the naive model (trained using patient labels in place of cell labels). After training on diagnostic bone marrow specimens (**A**), MMIL generalizes to different tissues and treatment time points better than the naive model, evidenced by its performance on blood samples collected at (**B**) diagnosis, (**C**) day 8 posttreatment initiation, and (**D**) day 15 posttreatment initiation. Note that the oracle model is provided as a reference for the highest achievable classifier performance, as it is trained on gold-standard labels that are typically unknown.

cells. Among the markers with positive (cancer-associated) coefficients identified by MMIL (figs. S11 and S12) were CD10, CD19, paired box protein 5 (PAX5), and CD34, each of which has been previously described as aberrantly overexpressed in ALL (*14*, *25*, *26*). MMIL also identified CD58, a well-known marker of clinical MRD (*26*). Conversely, all the proteins with negative coefficients in the MMIL model are phenotypic markers of mature B cells, such as surface immunoglobulin Ms (IgMs), intracellular immunoglobulin M (IgMi), CD20, and the lambda chain of the B cell receptor (figs. S11 and S12). Each of these proteins are expressed primarily during stages of B cell development that are not reached by leukemic cells, whose differentiation is blocked at earlier stages of B cell development (*14*). Thus, we demonstrate that MMIL identifies residual leukemia cells of a phenotype consistent with known features of resistant cells, demonstrating its robustness against shifting distributions and cell frequencies.

## MMIL identifies MRD-associated cells without expert, gold-standard labels

Last, we applied MMIL to a critical challenge: Identifying disease-associated cell populations when obtaining gold-standard labels is impossible, even for highly trained experts. The presence of treatment-resistant MRD cells can only be determined after a patient begins treatment and is monitored for the presence of residual leukemia cells. It is not now possible to predict if someone has resistant cells before they start therapy. We tested MMIL's ability to perform this task.

To achieve this, we trained MMIL on a dataset of 3.1 million cells collected from the diagnostic time points of 51 patients with ALL (*14*). Among these, 12 patients were clinically evaluated as MRD-negative ($MRD^-$) and 39 were evaluated as MRD-positive ($MRD^+$) following induction chemotherapy. Our objective was to separate the $MRD^-$ and $MRD^+$ groups using only the diagnostic time point data.

At diagnosis, MMIL demonstrated superior performance in distinguishing MRD-positive and MRD-negative patients in the held-out cross-validation fold compared to existing methods such as PhenoGraph, FlowSOM, and pseudobulk analyses. MMIL performed particularly well across several metrics—mean, median, quantile, and threshold-based decision boundaries—outperforming the alternative approaches with higher patient-level AUROCs in the held-out fold (Fig. 6, A and B).

Upon investigating the cell-level features driving this separation, we identified several markers that were pivotal in distinguishing MRD-positive patients from MRD-negative patients at diagnosis. Notably, activated signaling proteins associated with MRD-positive

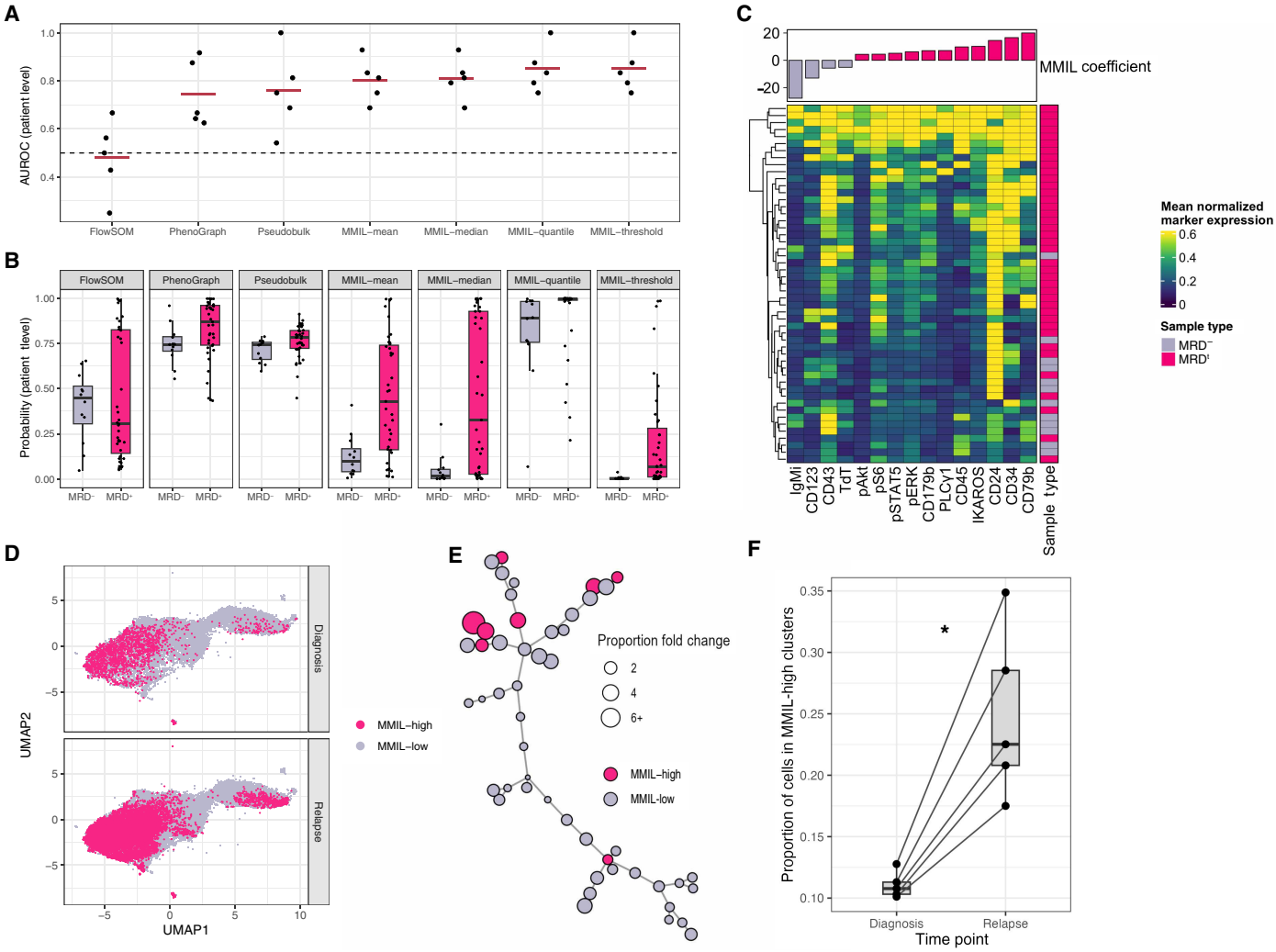


**Fig. 6. MMIL prospectively identifies cells that predict MRD at diagnosis and that expand in paired relapse samples.** (**A**) Comparison of patient-level AUROCs across methods: FlowSOM, PhenoGraph, pseudobulk, and MMIL (mean, median, quantile, and threshold; see Materials and Methods) using cross-validation to separate MRD-positive and MRD-negative patients at the diagnostic time point. Each point is the AUROC for a fold of fivefold cross-validation. MMIL consistently outperforms other methods in classification performance. (**B**) Comparison of MMIL-assigned patient-level probabilities across MRD-positive and MRD-negative patients in the diagnostic cohort, showing MMIL's ability to separate groups. Probabilities were calculated in the held-out fold. (**C**) Heatmap showing mean expression of the 15 features with the largest lasso coefficients in the MMIL model across diagnostic samples. Mean expression values were calculated from cells with MMIL probabilities at or above the ρth quantile (99th percentile) in each patient's diagnostic samples, where ρ is the parameter used to train the MMIL model. Dendrograms (sample-wise hierarchical clustering) are shown at left, and sample annotations are shown at right. Features are ordered by lasso coefficient (bar plot). TdT, terminal deoxynucleotidyl transferase. (**D**) UMAP projection of paired relapse samples from a single patient with ALL (UPN10), showing expansion of high-probability MMIL phenotypes at relapse. Also see fig. S13. (**E**) Minimum-spanning tree (MST) visualization of FlowSOM clustering results for diagnostic and relapse samples from patient UPN10 demonstrating expansion of high MMIL-probability clusters. Clusters are annotated as MMIL-high and MMIL-low based on MMIL probability scores of the cells they contain (Materials and Methods). Also see fig. S13. (**F**) Quantification of the relative abundance of high MMIL-probability cells in relapse samples versus diagnostic samples across all five patients with diagnosis-relapse sample pairs, demonstrating significant increase at relapse. *$P < 0.05$ for a paired $t$ test between the relapse and diagnostic time points ($t_4 = 5.27$; $P = 0.006$).

cells included those known to correlate with treatment resistance (Fig. 6C) (*14*).

Next, we applied the diagnostic MMIL model to previously unseen paired relapse samples from a subset of five patients in the initial cohort. We found that cell phenotypes with high MMIL probabilities for MRD at diagnosis expanded significantly at relapse. This expansion was visualized using UMAP plots, where regions of high MMIL probabilities at diagnosis showed an increased presence in relapse samples (Fig. 6D and fig. S13). Last, FlowSOM clustering compared the expansion of MMIL probability-high and MMIL probability-low clusters between diagnosis and relapse (Fig. 6E and fig. S13). We quantified the extent of this expansion, demonstrating a significant increase in the relative abundance of cells with high MMIL probabilities in relapse samples compared to diagnostic samples (Fig. 6F). Together, these findings highlight MMIL's unique ability to detect and track disease-associated cell populations across time points, even in the absence of gold-standard labels. In addition, they demonstrate MMIL's ability to identify single-cell features

associated with poor treatment response, which is not now possible by available methods.

## DISCUSSION

Biomedical studies of single-cell data often involve analyzing many unlabeled cells from individual patients to identify disease-associated cells. Finding such cells has broad-reaching applications in biomedical science, such as aiding our comprehension of disease-contributing cell populations and improving disease diagnosis and monitoring. MMIL is a method to train a cell classifier using patient labels only. Our approach has several attractive features. First, it adapts to different choices of classification algorithms: It is straightforward to implement as a wrapper that repeatedly calls established classification software that optimizes the binomial log likelihood. This makes it easy to implement our method within a preexisting machine learning pipeline. Second, in the past, it has been unclear how to calibrate models without labeled data; because our approach can estimate cell labels with logistic regression, we can use it to perform Platt scaling, a standard calibration method (Materials and Methods). This is useful because interpretable probability estimates empower straightforward downstream clinical decision-making. Last, MMIL also supports partially labeled datasets: If a small number of cells have known positive labels, they can be included during model training to guide the classifier. This means that MMIL can learn from both labeled and unlabeled data simultaneously, allowing it to leverage existing knowledge while remaining robust to noisy or imperfect gold-standard labels.

Applied to a dataset of cells from patients with and without leukemia, a setting where distinguishing cancer cells from baseline cells is critical but challenging, MMIL performed and generalized better than the naive approach of training a classifier using patient labels in place of cell labels. As designed, MMIL was compatible with high-dimensional settings, such as scRNA-seq datasets. In this setting, we recommend incorporating dimensionality reduction and using regularized classifiers when applying MMIL in extremely high-dimensional settings. Furthermore, we observed that including a small amount of labeled data, a single pathologist-annotated AML sample, not only improved MMIL's performance but also led it to incorporate expert knowledge into its own decision-making for identifying clinically and/or biologically relevant markers. We also showed that MMIL is more robust to noisy labeling than traditional supervised learning methods, allowing for robust performance in the classically difficult task of leukemic blast identification.

In the context of ALL, we additionally found that MMIL offered modest improvement over the naive model when applied to patient samples from the same time point (diagnosis) as the samples on which the models were trained (Fig. 5, A and B). However, only MMIL maintained good performance for tracking cancer cells over the course of treatment. The naive model's performance degraded substantially when applied to samples collected at later time points, likely due to phenotypic shifts occurring in leukemic blasts throughout chemotherapy (*27*). By contrast, MMIL's performance was more robust at later time points, particularly at day 15 posttreatment initiation, with a mean area under the ROC curve (AUC) of 0.815 compared to the naive model's AUC of 0.662 (Fig. 5, C and D). MMIL's robust performance across multiple treatment time points highlights its ability to reliably detect biomarkers for disease at the single-cell level. The method was also able to identify cells in pretreatment samples that predict future MRD and are enriched in relapse samples.

Two parameters must be set to train MMIL, aside from any hyperparameters for the chosen classifier: We assume that $\zeta$, the proportion of cells from patients in the prediction population, and $\rho$, the proportion of baseline cells in patients, are known. Previous work (*28*, *29*) has shown that these parameters are not, in general, statistically identifiable. Where possible, these parameters should be estimated from other datasets and/or clinical knowledge. If necessary, a sensitivity analysis could be performed to evaluate how results are affected (Supplementary Methods).

We make several assumptions regarding healthy donors and patients. First, we assume that healthy donors have no disease-associated cells. In the case that healthy donors may also have a mixture of disease-associated and baseline cells, we instead recommend a clustering method such as MDA2 (*30*). Second, MMIL assumes binary patient-level labels. While this assumption is often useful, it may not reflect the graded or uncertain nature of disease labels (e.g., symptom severity or staging). An interesting future direction is to extend MMIL to fit an ordinal or multinomial model for settings where disease labels are not binary.

Of course, if technical batch effects exist in the datasets causing systematic differences between patients and healthy donors, then it is impossible for MMIL to distinguish batch effects from disease-associated differences; thus, appropriate upstream data quality should be addressed. Last, MMIL is an iterative process that repeatedly fits models. For large datasets where this is not feasible, we recommend using stochastic gradient descent to optimize the observed log likelihood directly (Supplementary Methods).

Thus, MMIL is an easy to implement method to train a cell-level classifier using patient labels and can be used with a variety of classification algorithms and data modalities. It has potential as a useful tool for biological hypothesis generation, diagnostics, disease monitoring, and biomarker discovery, especially for discovering important cell populations when gold-standard labels are unavailable or incomplete.

## MATERIALS AND METHODS

### Algorithm details

In our setting, we have data consisting of cells $x_i \in \mathbb{R}^p$, sampling labels $z_i$ indicating whether each cell was sampled from a patient or healthy donor, and unobserved cell labels $y_i$ indicating whether each cell is disease associated. We wish to train a classifier to predict $P(y_i = 1 \mid x_i)$, and we will do so using its logit, $\eta(x_i)$.

In the case that $\eta$ has hyperparameters (e.g., the regularization strength in lasso logistic regression) that need to be selected, we recommend using cross-validation: For each desired choice of hyperparameter, train a model using cells from a subset of patients, and record the value of the observed data log likelihood (in Supplementary Methods) on the cells from held-out patients. After repeating this across all desired hyperparameter values and all folds, choose the hyperparameters that maximize the cross validated log likelihood.

MMIL uses the assumption that $\rho$ and $\zeta$, the proportion of baseline cells in patients and the probability that a sampled cell came from a patient, are known. We make these assumptions because $\rho$ and $\zeta$ are not, in general, identifiable. Ward *et al.* (*28*) show that they are identifiable only if we make strong assumptions about the form of $\eta(x)$; even when they are identifiable, their estimates have high variance. This argument is reaffirmed and expanded upon by Hastie and Fithian (*29*).

For large datasets, the EM procedure may be slow and it may be more appropriate to optimize the observed log likelihood (Supplementary Methods) through stochastic gradient descent. For smaller datasets, using EM allows us to easily try many different model choices for η and to take advantage of off-the-shelf model fitting software in each loop of the algorithm.

As we have described it here, our algorithm relies on the use of classification software that can train a model using soft labels (probabilities between 0 and 1). In Supplementary Methods, we share a modification for classifiers that require hard labels (restricted to be 0 or 1). We additionally share additional details and proofs to support this method.

## Related work

In MIL, data consist of labeled bags, each with many unlabeled instances. Negative bags have only negative instances, and positive bags have at least one positive instance. In our case, the "bags" are people, each with many unlabeled cells; healthy people have only baseline cells (negative instances), and patients have at least one disease-associated cell (positive instance). MIL methods can be broadly categorized into three paradigms (*31*): instance space, bag space, and embedded space. Bag and embedded space methods aim to label bags; the instance space paradigm, like ours, aims to label instances, and so is our primary focus. For example, multiple-instance support vector machine (mi-SVM) performs SVM with the goal of assigning all negative instances to one halfspace and at least one instance from every positive bag to the other. This method trains an SVM iteratively. First, instances are assigned labels: All cells from healthy donors are labeled "baseline," and one cell from each patient is randomly labeled disease associated. Then, an SVM is fit. Using the fitted SVM, instances are relabeled: For each patient, the single cell with the highest margin is labeled disease-associated and all others are labeled baseline, and a model is trained. This refitting and relabeling process is repeated until the labels stabilize. One can modify mi-SVM to allow multiple positives per bag; this requires setting another hyperparameter to perform thresholding. Another related approach is multiple-instance logistic regression with a lasso penalty (MILR) (*19*), which treats the instance labels as missing data, and uses EM to optimize a lasso-penalized binomial log likelihood. This is the same general approach as ours. However, in the expectation step, MILR uses the joint probability of all instances in the same bag; as a result, MILR tends to be biased when the number of instances in each bag becomes large. Our approach, MMIL, instead treats instances as the primary objects of study and does not use their bag-level grouping, thereby avoiding issues related to computing joint probabilities of many instances. This is particularly important in medical settings where we may sample tens or hundreds of thousands of cells per person. Last, there are many neural network architectures designed for bag space MIL (*31*) that can be used to label cells. One common method is to apply attention (*23*) to simultaneously label patients and identify the cells responsible for the predicted bag label.

Our setting is also analogous to the presence-only or positive-unlabeled settings, where some data points have known positive labels, and the remaining labels are unknown. We may instead refer to our data as negative-only data: Cells from healthy people are baseline and, therefore, have known negative labels, and cells from patients are unlabeled and may have positive or negative labels. In presence-only data, unlabeled instances are randomly sampled from the population, and

$$P\left(y_i = 1 | x_i, i \text{ is unlabeled}\right) = P\left(y_i = 1 | x_i\right) \quad (1)$$

In our setting, though, unlabeled data have a higher-than-baseline probability of belonging to the positive class. The knowledge that a cell is from a patient raises the probability that it is disease associated

$$P\left(y_i = 1 | x_i, i \text{ is unlabeled}\right) = P\left(y_i = 1 | x_i, i \text{ is from a patient}\right) > P\left(y_i = 1 | x_i\right) \quad (2)$$

Ward *et al.* (*28*) present a method to train a classifier with presence-only data using EM. MMIL is a modification of their approach to accommodate this structural difference in our data.

Mixture discriminant analysis 2 (MDA2) (*29*, *30*), a clustering method, is related to the problem at hand. MDA2 considers data from $K$ classes. Each class is a mixture of Gaussian distributions, the centers of which are shared across all $K$ classes. Here, we have $K = 2$ classes (healthy donors and patients). The healthy class is composed of just one distribution: the distribution of baseline cells. Because patients have both disease-associated and baseline cells, the patient class is a mixture of two distributions: the baseline and disease-associated cell distributions. While MDA2 is a natural fit for our setting, it is a generative method that models the joint likelihood $P(X, y)$, and it assumes that the features $X$ (including, e.g., log-transformed expression or intensity values) follow a multivariate normal distribution. This assumption may not always hold in practice. In such cases, we may prefer to use discriminative methods that model the conditional likelihood $P(y \mid X)$ and that may also provide feature importance measures (e.g., lasso and ridge regression) or capture relationships between features (e.g., gradient-boosted trees and neural networks).

Alternately, a simple but problematic approach is to train cell-level classifiers using patient labels in lieu of cell labels (*4*, *31*, *32*). That is, all cells from patients are labeled disease associated, and all cells from healthy donors are labeled baseline, and a model is trained using these labels. In our comparison, we term this the naive approach. This may work when most cells sampled from patients are disease associated, as the inherited label is correct most of the time.

When patients have a smaller fraction of disease-associated cells, classifier performance will suffer because of high label noise. Ward *et al.* (*28*) further show that using this approach with logistic regression causes coefficients to be biased toward 0, thereby hindering discovery of biologically important features. To directly address label noise, approaches from the denoising literature (*24*, *33*) may be useful. Still, our method leverages the knowledge that cells from healthy people have known labels, which is not assumed by more general methods.

## Simulation reveals that calibration is possible without ground truth labels

We simulate a dataset with size $n = 1000$ and $P = 100$ such that $\rho = \zeta = 0.5$, and the relationship between $x_i$ and $y_i$ is given by $\text{logit}\, P\left(y_i \mid x_i\right) = \beta^T x_i$. In each simulation, the first 10 coefficients of $\beta$ are drawn from a normal distribution with mean and SD of 1, and the remainder are 0. Then, $x_i \in \mathbb{R}^{100}$ is sampled from a standard multivariate normal distribution with identity covariance; $y_i$ is from a

binomial distribution defined by $P(y_i \mid x_i)$. Instances with $y_i = 1$ (disease-associated cells) necessarily have $z_i = 1$ (sampled from patients), and instances with $y_i = 0$ (baseline cells) have $z_i = 1$ with probability $\frac{\rho\zeta}{\rho\zeta + (1-\zeta)}$. We repeat this process until we have 1000 instances (500 train and 500 test) satisfying the requirement $\rho = \zeta = 0.5$.

Then, as before, we trained the oracle model, the naive model, and MMIL. In addition, we also trained an mi-SVM (described above). We are interested in both prediction and inference, so we compare the models in terms of AUROC, area under the precision-recall curve (AUPRC), expected calibration error, and the sum of absolute differences between $\widehat{\beta}$and $\beta$.

For the oracle model, calibration is done with Platt scaling using the true labels *y*; this represents the best performance that we could reasonably expect. The naive model is calibrated using the incorrect labels *z* to stay consistent with the fact that it was trained using the incorrect labels. The MMIL and mi-SVM models are calibrated in two steps: (i) obtain held-out predictions for each point in the training set (or use a separate held-out dataset) and (ii) fit a MMIL model using only the held-out probabilities to predict the cell labels *y*. This results in a logistic regression model that adjusts the predicted probabilities to be better calibrated. This simulation is repeated 1000 times. For all objectives, the oracle model (trained using the true labels) performs best. When the true labels are unavailable, MMIL outperforms the naive approach (Table 1), and performing Platt scaling using MMIL greatly improves calibration.

## AML CyTOF dataset analysis

### *Data acquisition*

Normalized, singlet-gated AML data were downloaded from FlowRepository (ID: FR-FCMZ2E7), and gold-standard pathologist annotations were obtained via correspondence with the authorship team of Tsai *et al.* (*13*). Due to the ambiguity of their diagnosis, the one patient in the cohort with myelodysplastic syndrome was excluded from analysis. For compatibility with the blast enumeration in (*13*) and standard pathology laboratory procedure, only $CD45^+$ events (hematopoietic lineage cells) were analyzed.

### *Data cleaning and preprocessing*

Standard preprocessing steps for mass cytometry data analysis were performed as described previously (*13*). Specifically, ion counts were transformed using the hyperbolic arcsine function with a cofactor of 5, and all markers were scaled to their 99.9th percentile for comparability between markers. Additionally, all cells expressing a marker value over the 99.9th percentile were excluded from analysis to remove technical artifacts and outliers. All analyses were performed using the R package tidytof (*34*, *35*).

## Model fitting

### *Models*

*Model specification*. MMIL was applied to the AML cohort via a lasso logistic regression model trained using algorithm 1 described in the Supplementary Materials. We used a value of ($\rho = 0.75$) for all MMIL models, relying on the clinical knowledge that patients with AML with more than 25% blasts in their bone marrow receive the clinical diagnosis of leukemia (*16*). However, we also conducted an analysis of the performance of our MMIL models over different values of ρ, which suggested that our results were not particularly sensitive to this hyperparameter (see fig. S2). ζ was estimated using the training set of each fold (see below).

In addition to MMIL, two other models were fit and evaluated: the oracle model (a lasso logistic regression model trained using the true cell labels, as annotated by a pathologist) and the naive model (a lasso model trained using inherited patient labels instead of cell labels). For clarity, the cell labels used by MMIL are referred to as the "probabilistic labels" of each cell; the cell labels used by the oracle model are referred to as the "gold-standard" or "pathologist-annotated" labels of each cell; and the labels used by the naive model are referred to as the "inherited patient labels" of each cell. MMIL, the oracle model, and the naive model are referred to as the three "model classes" that we evaluated.

*Hyperparameter tuning*. Lasso models have a single hyperparameter: λ, the penalty term determining the amount of regularization applied to the model's coefficients. For each model class, we tuned over 10 values of lambda, equally spaced on a logarithmic scale between the lowest value ($10^{-5}$) and the highest value (1). The optimal hyperparameter for each model class was determined using cross-validation (see below). The model predictions of the model fit with the optimal penalty parameter were used for reporting in Fig. 2's receiver operating characteristic (ROC) curves and AUROC values.

*Cross-validation and model selection*. Model performance was estimated using LOOCV, a schema in which all cells from a single patient with AML are held out as a separate test set in each fold of the cross-validation. Specifically, we break the dataset into 13 folds such that each fold includes 12 patients with AML and all three healthy controls for model training and one held-out AML patient for model evaluation. For each fold and model class, we fit the model on the training set and evaluate its cell-level performance on the held-out AML patient according to gold-standard, pathologist-annotated cell

**Table 1. Comparison of three lasso logistic regression modeling approaches and one SVM in 1000 simulations.** Oracle was trained with the true labels y (unavailable in realistic settings) and naive with the observed labels *z*. The first two columns report predictive performance, and the third reports inferential performance: the sum of the absolute values of the differences between the fitted coefficients $\widehat{\beta}$ and the true coefficients β. The final two columns show the expected calibration error (ECE), before and after calibration; a lower ECE is preferred. Platt scaling with MMIL (logistic regression) improves calibration for MMIL (lasso logistic regression) and makes calibration possible for mi-SVM. The mean and one SD of each distribution are shown.

| | AUROC | AUPRC | Coef. L1 error | Uncalibrated ECE | Calibrated ECE |
|---|---|---|---|---|---|
| **Oracle** | 0.94 ± 0.02 | 0.86 ± 0.05 | 6.71 ± 1.58 | 0.37 ± 0.02 | 0.04 ± 0.01 |
| **MMIL** | 0.90 ± 0.03 | 0.80 ± 0.07 | 9.64 ± 2.24 | 0.11 ± 0.02 | 0.06 ± 0.02 |
| **Naive** | 0.88 ± 0.04 | 0.74 ± 0.07 | 11.08 ± 2.43 | 0.38 ± 0.02 | 0.25 ± 0.03 |
| **mi-SVM** | 0.71 ± 0.05 | 0.47 ± 0.07 | | | 0.17 ± 0.05 |

labels. The optimal lasso regularization penalty for each model class was chosen by selecting the penalty that optimized the average log-likelihood (for MMIL) or binomial deviance (for both other models) on the held-out sample across all folds. Last, for model interpretation, we refit each model with the optimal penalty on all 13 AML samples and all three healthy bone marrow samples.

***0-Shot and 1-shot models***

In this section, we refer to models trained on a dataset for whom no cell labels are known "0-shot" models (the 0 denoting the number of patients who have received expert annotation before model training). Similarly, we refer to models trained on a dataset for whom a single patient's cell labels are known "1-shot" models (the 1 again denoting the number of patients who have received expert annotation before model training). We borrow this language from the few-shot machine learning literature, a paradigm that has been scarcely applied to MIL problems but that is highly applicable here.

0-Shot models are simply the same models described in the "Models" section above. In the 0-shot case, MMIL models are trained using probabilistic cell labels as described in Fig. 1A, naive models are trained using inherited patient labels, and oracle models are trained using gold-standard cell labels. Notably, it is not possible to fit a 0-shot oracle model, as oracle models require direct cell labels. This is why the left panel of Fig. 4B has a blank space for the oracle model.

By contrast, 1-shot models are trained identically to 0-shot models except for a single change: During model training, one AML patient (termed the 1-shot patient) is chosen such that their gold-standard cell labels are used for supervision instead of their probabilistic labels (for MMIL) or inherited patient labels (for naive models). Similarly to the cells from healthy patients, the 1-shot patient's cell labels remain fixed throughout all iterations of the EM. Otherwise, model training, cross-validation, and model selection are carried out as before.

Note that a single "1-shot experiment" in the Results refers to the following training procedure:

1) Choose one of the 13 patients with AML to designate as the 1-shot patient. Designate all other 12 patients with AML as "cross-validation patients."

2) Fit a 0-shot MMIL model and a 0-shot naive model using 12-fold cross-validation. In each fold of the cross-validation, 11 cross-validation patients, the 1-shot patient, and all three healthy controls are included in the training set, and one cross-validation patient is used as the held-out test set. For MMIL, the model with the penalty parameter that optimizes the average MMIL log likelihood in the test set across all folds is chosen as the best model and is used for error reporting. For the naive model, the model with the penalty parameter that optimizes the average binomial deviance in the test set across all folds is chosen as the best model and used for error reporting.

3) Fit a 1-shot MMIL model and a 1-shot naive model using 12-fold cross-validation. In each fold of the cross-validation, 11 cross-validation patients, the 1-shot patient, and all three healthy controls are included in the training set, and one cross-validation patient is used as the held-out test set. Additionally, fit an oracle model using only the 1-shot patient. For MMIL, the model with the penalty parameter that optimizes the average MMIL log likelihood in the test set across all folds is chosen as the best model and is used for error reporting. For the naive model, the model with the penalty parameter that optimizes the average binomial deviance in the test set across all folds is chosen as the best model and used for error reporting. For the oracle model, only a single patient is available for model training, so the model with the penalty parameter that optimizes the average binomial deviance for that patient is chosen as the best model and used for error reporting.

4) Repeat step 3 after randomly permuting 25% of the 1-shot patient's gold-standard labels.

Thus, to give each AML patient a turn being the 1-shot patient, steps 1 to 4 were repeated 13 times. Last, the AUROC for each left-out patient was averaged across all 13 1-shot experiments, giving an expected AUROC across all possible choices of 1-shot patient. These AUROC values are the reported values for each patient in Fig. 4B.

***Comparisons to unsupervised clustering algorithms***

To compare MMIL's ability to identify cancer cells relative to field-standard clustering algorithms commonly used to analyze high-dimensional cytometry data, we adapted the unsupervised clustering algorithms FlowSOM (*18*) and PhenoGraph (*17*). Because these algorithms are generally not used in the context of a train/test paradigm, we modified their application to allow for evaluation on held-out samples. Typically, these algorithms are applied to the entire dataset at once, with cytometrists manually annotating clusters based on their domain expertise. In this study, however, we applied FlowSOM and PhenoGraph to the same LOOCV folds as those described above to train and evaluate the MMIL, oracle, and naive models.

Specifically, we fit PhenoGraph- and FlowSOM-based models for classifying cells in a held-out patient using the following procedure. First, we clustered cells from all 12 patients in the training set of each fold using FlowSOM or PhenoGraph (using the tidytof R package's default parameters). We then annotated clusters as either "healthy" or "cancer" on the basis of the proportion of baseline cells within each cluster, greedily assigning clusters to the healthy class in descending order of their proportion of baseline cells until a predefined threshold (70, 80, or 90%) of all cells collected from healthy patients being assigned to the healthy class was met. Last, for each cell from the held-out patient, we calculated its 100-nearest neighbors in the training set using cosine similarity, and we assigned a probability of each test cell belonging to the cancer class by calculating the proportion of its nearest neighbors belonging to the cancer class. This procedure was repeated for all folds of the cross-validation such that each held-out patient's cell probabilities could be used to compute the ROC curves in fig. S3.

***Comparisons to KNN classifier***

To evaluate MMIL's performance relative to a baseline classifier commonly used in single-cell data analysis, we compared its performance to a KNN classifier. KNN is widely used in single-cell data analysis due to its simplicity and ability to assign cell labels based on local similarity. To ensure fair comparison, we applied the KNN classifier to the same LOOCV folds as those described for the MMIL, oracle, and naive models as well as the PhenoGraph and FlowSOM baselines described above.

Specifically, we fit the KNN classifier for classifying cells in a held-out patient using the following procedure. First, we used all cells from all 12 patients in the training set of each fold to calculate the 10-, 25-, or 100-nearest neighbors for each cell in the held-out patient using cosine similarity. For each test cell, we calculated the weighted vote of its nearest neighbors based on the sample type (cancer or healthy) from which the neighbor was collected. Each training cell's vote was inversely weighted by the number of cells in its sample type as well as the number of cells collected from the patient from which it originated (to ensure that each training patient contributed equally to the prediction regardless of sample size). The cancer

probability of each test cell was estimated as the weighted mean of votes from its nearest neighbors, where votes from cells from cancer samples were assigned a value of 1 and votes from healthy samples were assigned a value of 0. This procedure was repeated for all folds of the cross-validation such that each held-out patient's cell probabilities could be used to compute the ROC curves in fig. S4.

#### *Comparisons to MILR classifier*

*Model specification*. To further evaluate MMIL's performance relative to a baseline classifier commonly used in MIL, we compared its performance to a logistic regression model trained using the milr R package. MILR is an established method for handling weakly labeled data by modeling relationships between instances (cells) and their corresponding bags (patients). Each patient was treated as a separate bag, and cells within each bag were classified as disease associated or baseline based on the sample type annotation. The MILR model was trained using the binomial deviance loss function, consistent with the oracle and naive models described above.

*Hyperparameter tuning*. MILR (which uses lasso-regularized logistic regression) models have a single hyperparameter: λ, the penalty term for the model's coefficients. As for the MMIL models, we tuned over 10 values of lambda, equally spaced on a logarithmic scale between $10^{-5}$ and 1. The optimal hyperparameter was determined using LOOCV, selecting the lambda value that minimized the binomial deviance on the held-out sample. The model predictions of the MILR model fit with the optimal penalty were used for reporting in fig. S5's ROC curves and AUROC values.

*Cross-validation and model selection*. To ensure fair comparison, we applied the MILR model to the same LOOCV folds as those described for the MMIL, oracle, and naive models as well as the PhenoGraph, FlowSOM, and KNN baselines described above.

### Model evaluation and interpretation

#### *Single-cell model evaluation*

To evaluate each model class's performance at the single-cell level, we calculated the ROC curve and corresponding AUROC for each sample using gold-standard cell labels. ROC curves for each patient were calculated using the model from the cross-validation fold in which that patient was held out. In fig. S3, the average ROC curve across all patients was calculated by interpolating their individual ROC curves at 500 equally spaced points along the *x* axis [false-positive rate (FPR)] and averaging values on the *y* axis [true-positive rate (TPR)] across patients at each of these points. Ribbons around each averaged ROC curve represent the SEM around each interpolated point along the *x* axis.

#### *Blast percentages*

MMIL, naive, and oracle probabilities were assigned to each cell using the models refit on all patients with AML with the best penalty parameters identified by cross-validation. These probabilities were used to classify each cell in the dataset as either a leukemic blasts or not a leukemic blast with each of the three model classes. In the case of the oracle model, cell labels are known, so the probability cutoff was chosen to give the highest cell label classification accuracy. In the case of the naive model, cell labels are not known, so the probability cutoff was chosen to give the highest inherited patient label classification accuracy. Last, for MMIL, the cell labels are not known, but because of MMIL's strong calibration due to our model fitting procedure, a simple probability cutoff of 0.5 was chosen

Using these binary classifications for each cell, blast percentages were calculated for each patient. To recalibrate the blast percentages assigned by each model with the pathologist-enumerated blast percentages, a linear regression was fit for each model class as a post-processing step. Specifically, a linear regression of the model-assigned blast percentage onto the pathologist-enumerated blast percentage was used to derive the values plotted on the *y* axis of Fig. 2B.

#### *UMAP plots*

To analyze how cells assigned high MMIL probabilities arrange in high-dimensional space, we performed dimensionality reduction using UMAP using all cells and all markers in the dataset (*21*). The plotted MMIL probabilities were taken from the MMIL model refit on all patients with AML using the optimal penalty parameter identified by cross-validation, and UMAP was run using the default parameters of the tidytof R package (*34*, *35*).

In Fig. 2D, local neighborhoods were constructed using a two-step process. First, density-dependent downsampling (*34*, *36*) was used to select index cells from the full dataset such that all regions of phenotypic space are represented equally, with each index cell representing the center of a local neighborhood in high-dimensional space. All markers were used to calculate the local densities surrounding each cell in the dataset during density-dependent downsampling. This downsampling process selected 6849 index cells that were approximately evenly dispersed throughout the high-dimensional point cloud. After index cells were selected, the percentage of cells in each local neighborhood collected from AML patient samples was calculated by finding the 100-nearest neighbors of each index cell in the original dataset and computing the proportion of those neighbors from cancer samples.

#### *0-Shot and 1-shot Lasso coefficient analysis*

For model interpretation in Fig. 4, we examined the nonnegative coefficients of the MMIL model refit on all data using the optimal penalty parameter identified by cross-validation. We focused our analysis on positive (i.e., cancer-associated) lasso coefficients because, generally speaking, disease-associated features are more useful as positive biomarkers that can be used to diagnose or monitor disease. In Fig. 4E, features were counted as disease associated if they had a lasso coefficient of at least 0.01 in their corresponding MMIL model.

### AML scRNA-seq dataset analysis

#### *Data acquisition and preprocessing*

scRNA-seq data were downloaded from the Gene Expression Omnibus under accession number GSE116256 (*20*). For each sample, raw count matrices and accompanying metadata were read and merged into a single Seurat object. To ensure comparability across samples and genes, count data were normalized using Seurat's NormalizeData() function, and the top variable features were selected using FindVariableFeatures() with default parameters. The resulting gene expression matrix was scaled and subjected to PCA. On the basis of visual inspection of the ElbowPlot, the top 15 PCs were retained for downstream modeling, capturing ~12% of the total variance in the data.

#### *Metadata curation and filtering*

We extracted patient- and cell-level metadata from the Seurat object and manually annotated each sample with its disease status (healthy or cancer), time point, and patient identifier. Cell-level annotations were applied to each cell based on the refined random-forest model score provided for each cell in which was shown to have >95% sensitivity and >99% specificity for detecting malignant cells. Cells assigned to the "normal" class were assigned a cell-level label of 0,

and all other cells were assigned a cell-level label of 1 for evaluation (and for training oracle models).

To ensure consistency across samples, we restricted our analysis to cells collected at the diagnostic time point (D0), excluding samples from other days due to their sparse and inconsistently timed collection. Additionally, to focus exclusively on primary patient samples, we removed cell lines (MUTZ3 and OCI-AML3) from the analysis. After filtering, the final dataset included 22,600 cells from 21 individuals (5 healthy donors and 16 patients with AML).

## Model fitting

### *Model specification*

We trained MMIL, naive, and oracle models as described above. MMIL and naive models were trained using only patient-level labels (disease associated versus baseline) and the first 15 PCs as input features based on manual inspection of the elbow plot (fig. S6C). We fixed $\rho = 0.75$ on the basis of clinical knowledge that patients with ≥25% blasts in their bone marrow are diagnosed with leukemia. For each fold of cross-validation, we estimated $\zeta$, the proportion of disease-labeled cells, empirically using the training set. In addition, oracle models were trained using the cell-level annotations provided by van Galen *et al.*

### *Cross-validation and model selection*

Model performance was assessed using LOOCV. In each fold, one patient with cancer was held out for testing, and the remaining patients with cancer and all healthy donors were used for training. We trained MMIL models across a range of 10 lasso penalty values ($\lambda$), logarithmically spaced between $10^{-5}$ and 1. The optimal $\lambda$ for MMIL was selected as the value that maximized test set log-likelihood across folds. For the naive and oracle models, the optimal $\lambda$ was selected to minimize binomial deviance.

For each model class, the best-performing model from each fold was extracted and used for downstream evaluation. We calculated AUROC values on held-out patients and plotted ROC curves using gold-standard, pathologist-annotated labels for evaluation. Final model comparisons were based on mean AUROC across folds, and SEs of the AUROC were used to assess uncertainty.

## Longitudinal ALL dataset

### *Data acquisition*

Deidentified bone marrow and peripheral blood primary samples from three patients with ALL were obtained under informed consent from Lucile Packard Children's Hospital and Stanford Hospital from the Pediatric Clinic of Istituto di Ricovero e Cura a Carattere Scientifico (IRCCS) San Gerardo dei Tintori Hospital (Fondazione Tettamanti, Monza, Italy). The use of these samples was approved by the Institutional Review Boards at Lucile Packard Children's Hospital, Stanford Hospital at Stanford University, and from the Pediatric Clinic of IRCCS San Gerardo dei Tintori Hospital. Cryopreserved primary bone marrow and peripheral blood samples were thawed rapidly in thawing media (RPMI 1640 supplemented with 10% fetal bovine serum, 1% penicillin-streptomycin, and glutamine, sodium heparin (20 U/ml), and Benzonase (0.025 U/ml). Cells were rested for 30 min at 37°C and cisplatin viability stained (*37*).

Each sample was analyzed for the expression of 29 proteins using mass cytometry as previously described (*14*). Briefly, cells were fixed with paraformaldehyde, washed in cell staining medium (CSM) twice, followed by one wash in phosphate-buffered saline (PBS) and one wash in PBS and 0.02% saponin. Blocking was performed with Human TruStain FcX receptor blocking solution (BioLegend, 422302). Cells underwent surface staining with the following surface markers: CD3, CD10, CD19, CD20, CD22, CD24, CD34, CD38, CD43, CD45, CD58, CD61, CD123, CD127, CD179a, CD179b, human leukocyte antigen–DR, IgM, and thymic stromal lymphopoietin receptor.

After surface staining, cells were washed, permeabilized, and intracellular stained with the following markers: IgMi, marker of proliferation Ki-67, lambda chain of the B cell receptor, PAX5, recombination activating gene 1, and terminal deoxynucleotidyl transferase. Once intracellularly stained, samples were washed in CSM, iridium intercalated, and washed in CSM, followed by two washes in ultrapure double-distilled water. To prepare for acquisition, cells were resuspended with normalization beads. Mass cytometry data were then acquired on a Helios (a third-generation CyTOF).

### *Data cleaning and preprocessing*

Standard preprocessing steps for mass cytometry data analysis were performed as described previously (*14*). Specifically, ion counts were transformed using the hyperbolic arcsine function with a cofactor of 5, and all markers were scaled to their 99.9th percentile for comparability between markers. Additionally, all cells expressing a marker value over the 99.9th percentile were excluded from analysis to remove technical artifacts and outliers. All analyses were performed using the R package tidytof (*34*, *35*).

## Model fitting

### *Model specification*

MMIL was applied to the ALL cohort via a lasso logistic regression model trained using algorithm 1 described in the Supplementary Materials. We used a value of ($\rho = 0.75$) for all MMIL models, relying on the clinical knowledge that patients with ALL with more than 25% blasts in their bone marrow receive the clinical diagnosis of leukemia (*16*). $\zeta$ was estimated using the training set of each fold (see below).

In addition to MMIL, two other models were fit and evaluated: the oracle model (a lasso logistic regression model trained using the true cell labels, as annotated by a pathologist) and the naive model (a lasso model trained using inherited patient labels instead of cell labels). For clarity, the cell labels used by MMIL are referred to as the probabilistic labels of each cell; the cell labels used by the oracle model are referred to as the gold-standard or pathologist-annotated labels of each cell; and the labels used by the naive model are referred to as the inherited patient labels of each cell. MMIL, the oracle model, and the naive model are referred to as the three model classes that we evaluated.

For the ALL dataset, only diagnostic bone marrow samples were used to train any models to mimic the clinical scenario in which a model is built at the time of diagnosis and used to track a patient's leukemia burden over time.

### *Hyperparameter tuning*

Lasso models have a single hyperparameter: $\lambda$, the penalty term determining the amount of regularization applied to the model's coefficients. For each model class, we tuned over 10 values of lambda, equally spaced on a logarithmic scale between the lowest value ($10^{-5}$) and the highest value (1). The optimal hyperparameter for each model class was determined using cross-validation (see below). The model predictions of the model fit with the optimal penalty parameter were used for reporting in Fig. 5's ROC curves and AUROC values.

### *Cross-validation and model selection*

Model performance was estimated using LOOCV, a schema in which all cells from a single patient with ALL are held out as a separate test

set in each fold of the cross-validation. Specifically, we break the dataset into three folds such that each fold includes two patients with ALL and three healthy patients for model training and one held-out patient with ALL for model evaluation. For each fold and model class, we fit the model on the training set and evaluate its cell-level performance on the held-out patient with ALL according to gold-standard cell labels. The optimal lasso regularization penalty for each model class was chosen by selecting the penalty that optimized the average log-likelihood (for MMIL) or binomial deviance (for both other models, as is ordinarily done) on the held-out sample across all folds. Last, for model interpretation, we refit each model with the optimal penalty on all three diagnostic ALL and all three healthy bone marrow samples.

#### *Model evaluation and interpretation*

*Single-cell model evaluation*. To evaluate each model class's performance at the single-cell level, we calculated the ROC curve and corresponding AUROC for each sample using gold-standard cell labels. ROC curves for each patient were calculated using the model from the cross-validation fold in which that patient was held out. The average ROC curve across all patients was calculated by interpolating their individual ROC curves at 500 equally spaced points along the $x$ axis (FPR) and averaging $y$ axis (TPR) values across patients at each of these points. Ribbons around each averaged ROC curve represent the SEM around each interpolated point along the $x$ axis.

Although models were only trained using cells from the diagnostic time point, the models were evaluated for all available samples at all time points. Accordingly, ROC curves and AUROCs from different tissues and time points were calculated separately.

*Model interpretation*. For model interpretation in fig. S11, we examined the nonzero coefficients of the MMIL model refit on all data using the optimal penalty parameter identified by cross-validation.

### Discovery ALL dataset

#### *Data acquisition*

Normalized, singlet-gated data from Good *et al.* (*14*) were downloaded from GitHub at the following link: https://github.com/karadavis-lab/DDPR. Only patient samples annotated with a clinical MRD risk group were retained for analysis ($n$ = 51; 12 MRD negative).

#### *Data cleaning and preprocessing*

Standard preprocessing steps for mass cytometry data analysis were performed as described previously (*14*). Specifically, ion counts were transformed using the hyperbolic arcsine function with a cofactor of 5, and all markers were scaled to their 99.9th percentile for comparability between markers. Additionally, all cells expressing a marker value over the 99.9th percentile were excluded from analysis to remove technical artifacts and outliers. Only cells in the "basal" stimulation condition were included for analysis, and diagnostic samples were downsampled to a maximum of 50,000 cells to prevent any single sample with a large number of cells from dominating model training. All analyses were performed using the R package tidytof (*34*, *35*).

### Model fitting

#### *MMIL model specification*

MMIL was applied to the Discovery ALL dataset via a lasso logistic regression model trained using algorithm 1 described in the Supplementary Materials. We used a value of $\rho = 0.99$ for all MMIL models because of our expectation that MRD-associated cells are likely to be rare at the diagnostic time point. $\zeta$ was estimated using the training set of each fold.

#### *Hyperparameter tuning*

Lasso logistic regression models trained with MMIL were tuned over five values of $\lambda$, equally spaced on a logarithmic scale between the lowest value ($10^{-5}$) and the highest value ($10^{-1}$). The optimal hyperparameter for each model class was determined using cross-validation (see below). The predictions of the model fit with the optimal penalty parameter were used for reporting in Fig. 6A's AUROC boxplots and Fig. 6B's probability boxplots.

#### *Cross-validation and model selection*

Model performance was estimated using fivefold cross-validation. Specifically, we break the dataset into five folds such that each patient is randomly assigned to a single fold to be included in the held-out set. For each fold, we fit an MMIL model on the training set and evaluated its cell-level performance on the held-out set using the MMIL binomial log-likelihood (as gold-standard cell labels are unknown). The optimal lasso regularization penalty for the MMIL model was chosen by selecting the penalty that optimized the average MMIL log-likelihood on the held-out data across all folds. Last, for model interpretation, we refit a final model using the optimal penalty hyperparameter $\lambda$ on all diagnostic samples.

### Model evaluation and interpretation

#### *Patient-level model evaluation*

Because gold-standard cell labels in the discovery scenario are unknown, it was not possible to strictly evaluate MMIL's performance at the single-cell level. Instead, we evaluated the MMIL model at the patient-level by aggregating the MMIL probabilities of all cells from each patient sample when they were in the held-out set during cross-validation. This yielded a single score between 0 and 1 for each patient, representing the predicted probability that the patient would be MRD positive at the end of induction chemotherapy, that could be used to calculate the patient-level AUROC for discriminating MRD-positive and MRD-negative patients within each fold (Fig. 6A) or to be plotted directly (Fig. 6B). We tested several aggregation functions, each of which is reported in Fig. 6:

1) MMIL-mean: Derive patient-level probability scores by computing the mean of all cell-level MMIL probabilities for each patient.

2) MMIL-median: Derive patient-level probability scores by computing the median of all cell-level MMIL probabilities for each patient.

3) MMIL-quantile: Derive patient-level probability scores by computing the 99th percentile of all cell-level MMIL probabilities for each patient (chosen to match the value of $\rho$ used to train the models).

4) MMIL-threshold: Derive patient-level probability scores by computing the proportion of cells in each patient with a cell-level MMIL probability larger than 0.99 (chosen to match the value of $\rho$ used to train the models).

We further compared MMIL's patient-level performance (using the aggregated patient-level probabilities as described above) to several simple baselines described below:

1) Pseudobulk: We conducted a pseudobulk analysis by aggregating protein features from all cells within each patient to create a single, average vector of protein expression for each patient. This approach effectively reduces the single-cell data into patient-level summaries, treating each patient as a single "bulk" sample. For each fold of cross-validation, we then fit a lasso logistic regression model

on the aggregated training data to predict the probability of a patient being MRD positive after induction chemotherapy. Pseudobulk models were trained using identical cross-validation folds as those used to train the MMIL models, and they were tuned over the same hyperparameter values for λ. We then applied the trained model to the held-out patient in each fold and computed the AUROC to evaluate its performance in distinguishing MRD-positive from MRD-negative patients within each fold. This method allows for comparison between MMIL's ability to leverage single-cell level probabilities and the performance of traditional bulk-like analyses.

2) PhenoGraph: We adapted PhenoGraph, an unsupervised clustering algorithm commonly used to analyze high-dimensional cytometry data to predict cell-level MRD-association probabilities using a similar procedure to that described above for fig. S3. Specifically, we used PhenoGraph to cluster cells from all cells in the training set of each fold of the cross-validation using the tidytof R package's default parameters. Clusters were then annotated as either "MRD negative" or "MRD positive" based on the proportion of MRD-negative cells within each cluster—clusters were assigned to the "MRD-negative" class in descending order of their proportion of MRD-negative cells until a predefined threshold (75%) of all cells collected from MRD-negative patients were assigned to the MRD-negative class. All other clusters were assigned to the "MRD-positive" class. For the held-out patients, we then computed the probability that each cell belonged to the MRD-positive class using a nearest neighbor classifier based on the 100-nearest neighbors in the training set. Last, we aggregated these cell-level probabilities using mean aggregation to derive a patient-level probability score. PhenoGraph models were trained using identical cross-validation folds as those used to train the MMIL models, and we computed the AUROC to evaluate performance in distinguishing MRD-positive from MRD-negative patients.

3) FlowSOM: We followed the same procedure as described above for PhenoGraph, but we used tidytof's implementation of FlowSOM (with default parameters) for clustering instead of PhenoGraph.

***Single-cell interpretation***

To evaluate MMIL's usefulness for identifying specific cell populations associated with MRD, we refit an MMIL model using all diagnostic data using the optimal λ identified by cross-validation and performed several analyses using the resulting cell-level probabilities:

1) MMIL feature selection heatmap: For each diagnostic patient sample, we selected the cells with MMIL probabilities at or above the 99th percentile (in accordance with our choice for ρ during model training) and plotted their average expression of the top 15 protein features with the largest lasso coefficients in the final model using the ComplexHeatmap (*38*) R package (Fig. 6C).

2) UMAP plots: Single-cell visualization of high-dimensional phenotype space was performed using the tidytof R package's implementation of UMAP under default parameters (Fig. 6D and fig. S13). As in the heatmap for Fig. 6C, UMAP embeddings were computed using the 15 protein features with the largest lasso coefficients in the MMIL model. In the UMAP plots, cells are labeled as "MMIL-high" if their probabilities are at or above the 90th percentile of MMIL probabilities for cells at the diagnostic time point. All cells below the 90th MMIL probability percentile at the diagnostic time point are labeled "MMIL-low."

3) Minimum-spanning tree (MST) cluster visualization: To visualize the expansion of clusters with high MMIL probabilities between the diagnostic and relapse time points, we constructed MSTs for each patient using FlowSOM clusters as nodes and the 15 protein features with the largest lasso coefficients in the MMIL model as input for computing edges. Fifty FlowSOM clusters were computed per patient using the tidytof R package using cosine distance as the distance metric for the self-organizing map. For each patient, the "tof_plot_clusters_mst" tidytof function was used to compute the MST over a weighted graph of cluster centroids using Euclidean distances between cluster centroids as edge weights. MSTs were visualized using tidytof's default force-directed layout for fully connected graphs with fewer than 1000 nodes (*39*). Clusters were categorized as MMIL-high or MMIL-low by sorting clusters in descending order of mean MMIL probability and then greedily assigning clusters to the MMIL-high category until 10% of a patient's total cells at the diagnostic time point were assigned to MMIL-high clusters. All remaining clusters were assigned to the MMIL-low category. The MST for each patient at diagnosis was compared to the MST at relapse by mapping the expansion or contraction of each cluster over time. To quantify this expansion/contraction, we calculated the fold change in the proportion of cells within each cluster between diagnosis and relapse. Visualizations display clusters color-coded by their MMIL-high status and scaled by the magnitude of the expansion (Fig. 6E and fig. S13). Last, the expansion of cells in MMIL-high clusters across all patients was quantified by calculating the number of cells in MMIL-high clusters divided by the total number of cells in the sample at diagnosis and relapse (separately) for each patient. These proportions were compared between the diagnostic and relapse time points using a paired Student's *t* test via the "t.test" function in R and plotted in Fig. 6F.

### Ethical approval declarations

This research complies with all relevant ethical regulations. The use of the primary samples was approved by the Institutional Review Board at Lucile Packard Children's Hospital at Stanford University. Healthy human bone marrow samples ($n = 3$) were purchased through AllCells. Deidentified bone marrow samples from pediatric patients with B cell precursor (BCP)-ALL were obtained, under informed consent, from the Pediatric Clinic of Maria Letizia Verga Center, San Gerardo Hospital (Fondazione Tettamanti, Monza, Italy; $n = 3$). The use of these primary samples was approved by the Institutional Review Boards at both institutions. Written informed consent was obtained from the parents of the patients or their legal representatives, who agreed to the use of biological material for research and clinical studies. To protect patients' privacy, samples have been deidentified.

## Supplementary Materials

**This PDF file includes:**

Figs. S1 to S13

Supplementary Methods

## REFERENCES AND NOTES

1. G. S. Gulati, J. P. D'Silva, Y. Liu, L. Wang, A. M. Newman, Profiling cell identity and tissue architecture with single-cell and spatial transcriptomics. *Nat. Rev. Mol. Cell Biol.* **26**, 11–31 (2025).
2. D. Hanahan, Hallmarks of cancer: New dimensions. *Cancer Discov.* **12**, 31–46 (2022).
3. D. S. Pisetsky, Pathogenesis of autoimmune disease. *Nat. Rev. Nephrol.* **19**, 509–524 (2023).
4. M. H. Spitzer, P. F. Gherardini, G. K. Fragiadakis, N. Bhattacharya, R. T. Yuan, A. N. Hotson, R. Finck, Y. Carmi, E. R. Zunder, W. J. Fantl, S. C. Bendall, E. G. Engleman, G. P. Nolan, IMMUNOLOGY., An interactive reference framework for modeling a dynamic immune system. *Science* **349**, 1259425 (2015).

5. Y. C. Lo, Y. Liu, M. Kammersgaard, A. Koladiya, T. J. Keyes, K. L. Davis, Single-cell technologies uncover intra-tumor heterogeneity in childhood cancers. *Semin. Immunopathol.* **45**, 61–69 (2023).
6. Y. Hao, T. Stuart, M. H. Kowalski, S. Choudhary, P. Hoffman, A. Hartman, A. Srivastava, G. Molla, S. Madad, C. Fernandez-Granda, R. Satija, Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2024).
7. G. Heimberg, T. Kuo, D. J. DePianto, O. Salem, T. Heigl, N. Diamant, G. Scalia, T. Biancalani, S. J. Turley, J. R. Rock, H. Corrada Bravo, J. Kaminker, J. A. Vander Heiden, A. Regev, A cell atlas foundation model for scalable search of similar human cells. *Nature* **638**, 1085–1094 (2025).
8. A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the *EM* algorithm. *J. R. Stat. Soc. Ser. B* **39**, 1–22 (1977).
9. M. N. Dworzak, B. Buldini, G. Gaipa, R. Ratei, O. Hrusak, D. Luria, E. Rosenthal, J. P. Bourquin, M. Sartor, A. Schumich, L. Karawajew, E. Mejstrikova, O. Maglia, G. Mann, W. D. Ludwig, A. Biondi, M. Schrappe, G. Basso, International-BFM-FLOW-network, AIEOP-BFM consensus guidelines 2016 for flow cytometric immunophenotyping of Pediatric acute lymphoblastic leukemia. *Cytometry B Clin. Cytom.* **94**, 82–93 (2018).
10. A. Kruse, N. Abdel-Azim, H. N. Kim, Y. Ruan, V. Phan, H. Ogana, W. Wang, R. Lee, E. J. Gang, S. Khazal, Y. M. Kim, Minimal residual disease detection in acute lymphoblastic leukemia. *Int. J. Mol. Sci.* **21**, 1054 (2020).
11. X. Chen, B. L. Wood, Monitoring minimal residual disease in acute leukemia: Technical challenges and interpretive complexities. *Blood Rev.* **31**, 63–75 (2017).
12. D. A. Berry, S. Zhou, H. Higley, L. Mukundan, S. Fu, G. H. Reaman, B. L. Wood, G. J. Kelloff, J. M. Jessup, J. P. Radich, Association of minimal residual disease with clinical outcome in pediatric and adult acute lymphoblastic leukemia: A Meta-analysis. *JAMA Oncol.* **3**, e170580 (2017).
13. A. G. Tsai, D. R. Glass, M. Juntilla, F. J. Hartmann, J. S. Oak, S. Fernandez-Pol, R. S. Ohgami, S. C. Bendall, Multiplexed single-cell morphometry for hematopathology diagnostics. *Nat. Med.* **26**, 408–417 (2020).
14. Z. Good, J. Sarno, A. Jager, N. Samusik, N. Aghaeepour, E. F. Simonds, L. White, N. J. Lacayo, W. J. Fantl, G. Fazio, G. Gaipa, A. Biondi, R. Tibshirani, S. C. Bendall, G. P. Nolan, K. L. Davis, Single-cell developmental classification of B cell precursor acute lymphoblastic leukemia at diagnosis reveals predictors of relapse. *Nat. Med.* **24**, 474–483 (2018).
15. A. Jager, J. Sarno, K. L. Davis, Mass cytometry of hematopoietic cells. *Methods Mol. Biol.* **2185**, 65–76 (2021).
16. A. Hodes, K. R. Calvo, A. Dulau, I. Maric, J. Sun, R. Braylan, The challenging task of enumerating blasts in the bone marrow. *Semin. Hematol.* **56**, 58–64 (2019).
17. J. H. Levine, E. F. Simonds, S. C. Bendall, K. L. Davis, E.-a. D. Amir, M. D. Tadmor, O. Litvin, H. G. Fienberg, A. Jager, E. R. Zunder, R. Finck, A. L. Gedman, I. Radtke, J. R. Downing, D. Pe'er, G. P. Nolan, Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell* **162**, 184–197 (2015).
18. K. Quintelier, A. Couckuyt, A. Emmaneel, J. Aerts, Y. Saeys, S. Van Gassen, Analyzing high-dimensional cytometry data using FlowSOM. *Nat. Protoc.* **16**, 3775–3801 (2021).
19. P. Y. Chen, C. C. Chen, C. H. Yang, S. M. Chang, K. J. Lee, milr: Multiple-Instance Logistic Regression with lasso penalty. *R. J.* **9**, 446–457 (2017).
20. P. van Galen, V. Hovestadt, M. H. Wadsworth Ii, T. K. Hughes, G. K. Griffin, S. Battaglia, J. A. Verga, J. Stephansky, T. J. Pastika, J. Lombardi Story, G. S. Pinkus, O. Pozdnyakova, I. Galinsky, R. M. Stone, T. A. Graubert, A. K. Shalek, J. C. Aster, A. A. Lane, B. E. Bernstein, Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity. *Cell* **176**, 1265–1281.e24 (2019).
21. E. Becht, L. McInnes, J. Healy, C. A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, E. W. Newell, Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–44 (2019).
22. A. Yakimovich, A. Beaugnon, Y. Huang, E. Ozkirimli, Labels in a haystack: Approaches beyond supervised learning in biomedical applications. *Patterns* **2**, 100383 (2021).
23. L. D. Cunha, M. Yang, R. Carter, C. Guy, L. Harris, J. C. Crawford, G. Quarato, E. Boada-Romero, H. Kalkavan, M. D. L. Johnson, S. Natarajan, M. E. Turnis, D. Finkelstein, J. T. Opferman, C. Gawad, D. R. Green, LC3-associated phagocytosis in myeloid cells promotes tumor immune tolerance. *Cell* **175**, 429–441.e16 (2018).
24. H. Song, M. Kim, D. Park, Y. Shin, J. G. Lee, Learning from noisy labels with deep neural networks: A survey. *IEEE Trans. Neural. Netw. Learn. Syst.* **34**, 8135–8153 (2023).
25. J. Sarno, P. Domizi, Y. Liu, M. Merchant, C. B. Pedersen, D. Jedoui, A. Jager, G. P. Nolan, G. Gaipa, S. C. Bendall, F. A. Bava, K. L. Davis, Dasatinib overcomes glucocorticoid resistance in B-cell acute lymphoblastic leukemia. *Nat. Commun.* **14**, 2935 (2023).
26. M. Veltroni, L. De Zen, M. C. Sanzari, O. Maglia, M. N. Dworzak, R. Ratei, A. Biondi, G. Basso, G. Gaipa, I-BFM-ALL-FCM-MRD-Study Group, Expression of CD58 in normal, regenerating and leukemic bone marrow B cells: Implications for the detection of minimal residual disease in acute lymphocytic leukemia. *Haematologica* **88**, 1245–1252 (2003).
27. M. J. Borowitz, M. Devidas, S. P. Hunger, W. P. Bowman, A. J. Carroll, W. L. Carroll, S. Linda, P. L. Martin, D. J. Pullen, D. Viswanatha, C. L. Willman, N. Winick, B. M. Camitta, Children's Oncology Group, Clinical significance of minimal residual disease in childhood acute lymphoblastic leukemia and its relationship to other prognostic factors: A Children's Oncology Group study. *Blood* **111**, 5477–5485 (2008).
28. G. Ward, T. Hastie, S. Barry, J. Elith, J. R. Leathwick, Presence-only data and the em algorithm. *Biometrics* **65**, 554–563 (2009).
29. T. Hastie, W. Fithian, Inference from presence-only data; the ongoing controversy. *Ecography* **36**, 864–867 (2013).
30. T. Hastie, J. H. Friedman, R. Tibshirani, *The Elements of statistical learning: Data Mining, Inference, and Prediction* (Springer Series in Statistics, Springer, 2009).
31. J. Amores, Multiple instance classification: Review, taxonomy and comparative study. *Artif. Intell.* **201**, 81–105 (2013).
32. S. Andrews, I. Tsochantaridis, T. Hofmann, "Support vector machines for multiple-instance learning," in *Proceedings of the 16th International Conference on Neural Information Processing Systems (NIPS'02)* (MIT Press, 2002), pp. 577–584.
33. B. Nagarajan, R. Marques, E. Aguilar, P. Radeva, Bayesian DivideMix++ for enhanced learning with noisy labels. *Neural. Netw.* **172**, 106122 (2024).
34. T. J. Keyes, A. Koladiya, Y. C. Lo, G. P. Nolan, K. L. Davis, tidytof: A user-friendly framework for scalable and reproducible high-dimensional cytometry data analysis. *Bioinform. Adv.* **3**, vbad071 (2023).
35. W. J. Hutchison, T. J. Keyes, tidyomics Consortium, H. L. Crowell, J. Serizay, C. Soneson, E. S. Davis, N. Sato, L. Moses, B. Tarlinton, A. A. Nahid, M. Kosmac, Q. Clayssen, V. Yuan, W. Mu, J. E. Park, I. Mamede, M. H. Ryu, P. P. Axisa, P. Paiz, C. L. Poon, M. Tang, R. Gottardo, M. Morgan, S. Lee, M. Lawrence, S. C. Hicks, G. P. Nolan, K. L. Davis, A. T. Papenfuss, M. I. Love, S. Mangiola, The tidyomics ecosystem: Enhancing omic data analyses. *Nat. Methods* **21**, 1166–1170 (2024).
36. P. Qiu, E. F. Simonds, S. C. Bendall, K. D. Gibbs, R. V. Bruggner, M. D. Linderman, K. Sachs, G. P. Nolan, S. K. Plevritis, Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.* **29**, 886–891 (2011).
37. H. G. Fienberg, E. F. Simonds, W. J. Fantl, G. P. Nolan, B. Bodenmiller, A platinum-based covalent viability reagent for single-cell mass cytometry. *Cytometry A* **81**, 467–475 (2012).
38. Z. Gu, R. Eils, M. Schlesner, Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847–2849 (2016).
39. T. M. J. Fruchterman, E. M. Reingold, Graph drawing by force-directed placement. *Softw. Pract. Exper.* **21**, 1129–1164 (1991).


**Acknowledgments:** We thank A. Koladiya, N. D'Silva, and the Stanford University Department of Biomedical Data Science for helpful discussions. **Funding:** This work was supported by the Stanford Data Science Scholars Program (E.C.); Stanford Graduate Fellowship (E.C.); National Science Foundation, R01GM134483 (T.H.) and 19DMS1208164 (R.T.); National Institutes of Health, 5R01EB001988-16 (T.H.), 5R01EB001988-16 (R.T.), 1F31CA239365-01 (T.J.K.), U54CA209971 (G.P.N.), U19AI100627 (G.P.N.), U54CA209971 (G.P.N.), and R01 CA251858-01A1 (K.L.D.); Point Foundation (T.J.K.); Mark Foundation for Cancer Research (T.J.K. and K.L.D.); Andrew McDonough B+ (Be Positive) Foundation (T.J.K. and K.L.D.); Point Foundation Graduate Student Scholarship (T.J.K.); Associazione Italiana per la Ricerca sul Cancro AIRC, Start-UP grant no. 27325 (J.S.); Oxnard Foundation (K.L.D.); and Stanford Maternal and Child Health Research Institute (K.L.D.). **Author contributions:** Conceptualization: E.C., T.J.K., T.H., R.T., G.P.N., and K.L.D. Methodology: E.C., T.J.K., M.Z., T.H., R.T., G.P.N., P.D., and K.L.D. Investigation: E.C., T.J.K., J.S., A.T., D.G., T.H., R.T., G.P.N., and K.L.D. Visualization: E.C., T.J.K., J.P.D., R.T., and K.L.D. Funding acquisition: E.C., T.J.K., R.T., and K.L.D. Project administration: E.C., T.J.K., G.P.N., T.H., R.T., and K.L.D. Supervision: G.P.N., T.H., R.T., and K.L.D. Writing—original draft: E.C., T.J.K., J.S., M.Z., J.P.D., T.H., R.T., G.P.N., and K.L.D. Writing—review and editing: E.C., T.J.K., J.S., M.Z., J.P.D., D.G., T.H., R.T., G.P.N., P.D., and K.L.D. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. AML mass cytometry data are available without annotations on FlowRespository, with Accession ID FR-FCM-Z2E7, or with annotations on Dryad, with DOI 10.5061/dryad.jq2bvq8kj. The Dryad dataset can be accessed with the following link: https://doi.org/10.5061/dryad.jq2bvq8kj. Longitudinal pediatric ALL mass cytometry data are available on Dryad, with DOI 10.5061/dryad.8gtht76vw. The Dryad dataset can be accessed at this link: https://doi.org/10.5061/dryad.8gtht76vw. Discovery pediatric ALL mass cytometry data are available on Dryad, with DOI 10.5061/dryad.pvmcvdnxc. The Dryad dataset can be accessed at this link: https://doi.org/10.5061/dryad.pvmcvdnxc. The "mmil" R package to train MMIL models can be found at https://zenodo.org/records/15271400.

Submitted 21 December 2024
Accepted 25 July 2025
Published 27 August 2025
10.1126/sciadv.adv5019