Foundation Model Attributions Reveal Shared Inflammatory Program Across Diseases

Maxwell P. Gold[1], Miguel Reyes[2], Nathaniel Diamant[1], Tony Kuo[3], Ehsan Hajiramezanali[1], Jane W. Newburger[4,5], Mary Beth F. Son[6], Pui Y. Lee[6], Gabriele Scalia[1], Aicha BenTaieb[1], Sharookh B. Kapadia[2], Anupriya Tripathi[1], Héctor Corrada Bravo[1], Graham Heimberg[1*], Tommaso Biancalani[1*]

1) gRED Computational Sciences, Genentech, South San Francisco, California

2) gRED Research Biology, Genentech, South San Francisco, California

3) Roche Informatics, F. Hoffmann-La Roche, Mississauga, Ontario, Canada

4) Department of Cardiology, Boston Children's Hospital, Boston, Massachusetts

5) Department of Pediatrics, Harvard Medical School, Boston, Massachusetts

6) Division of Immunology, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts

* to whom correspondence should be sent

Abstract

Determining a gene's functional significance within a specific cellular context has long been a challenge. We introduce a framework for quantifying gene importance by leveraging attributions learned by foundation models (FMs) trained on large corpora of single-cell RNA-sequencing (scRNA-seq) datasets. Attribution scores robustly quantify gene importance across datasets, emphasizing key genes in relevant cell populations, while minimizing technical artifacts. Therefore, we developed SIGnature, a tool that enables rapid search of gene signatures across multiple scRNA-seq atlases. We demonstrated its utility by querying the MS1 gene program across 400 diverse studies, identifying activation of this signature in myeloid cells from multiple hyperinflammatory conditions, including Kawasaki disease. In-house experimental validation confirmed that serum from Kawasaki disease patients induces the MS1 phenotype in monocytes. These findings highlight that SIGnature can help uncover shared mechanisms across conditions, demonstrating its power for large-scale signature scoring and cross-disease analysis.

Main

A gene's expression level is not a reliable indicator of its functional significance within a single cell. For instance, key regulators of cell fate, such as transcription factors, are typically detected at low levels, while highly abundant genes, like those involved in mitochondrial functions, may have limited influence on specific cellular capabilities. To gain a deeper understanding of gene function, researchers often rely on approaches that compare expression levels in one group relative to another. One widely used method is differential gene expression analysis, where algorithms like DESeq2 [1] and Gene Set Enrichment Analysis (GSEA) [2] can reveal genes or gene sets with significant changes in expression between experimental conditions or cell types. Alternatively, scoring methods, such as Gene Set Variation Analysis (GSVA) [3] and those implemented in *Seurat* [4] and *Scanpy* [5], quantify activation within an individual sample or cell.

While these approaches have yielded valuable biological insights, the resulting differential genes and sample-level scores have limited generalizability across experiments, as each study's unique design complicates extrapolating results outside of the original context. For instance, the cell type composition of an experiment impacts the background distribution used for many scoring methods [4–7] and constrains the possible comparisons for differential analysis. Interpretation of scores and differential genes across studies is further complicated by technical variations such as sequencing depth, library preparation, and batch effects, which can introduce substantial noise. To address these issues, various computational methods for normalization and dataset integration have been developed [8–11], but these approaches cannot be effectively applied to analyze the thousands of publicly available single cell transcriptomics experiments.

Here, we show that these limitations can be overcome by quantifying gene importance through *attributions* derived from single-cell RNA-sequencing (scRNA-seq) foundation models (FMs). Attribution, a concept from explainable artificial intelligence, quantifies the contribution of each feature in a sample to the predictions made by a model. For example, in image classification, the pixels with the

highest attribution scores are most responsible for identifying an object, such as a dog in a backyard (**Figure 1a**). Similarly, this concept can be applied to a foundation model (FM) trained in a self-supervised manner on a large corpus of single-cell RNA sequencing (scRNA-seq) profiles (**Figure 1b)**. In this context, genes with high attribution scores are those that significantly influence a cell's position within the latent space of the model **(Figure 1c)**. If the FM encodes biological function, as demonstrated in SCimilarity[12] (which will be used throughout the paper) and other models [13–22], then genes with high attribution scores reflect their functional importance in the cell. Through this approach, we can transform the expression vector of a cell into an attribution vector which can be reliably compared across studies. Indeed, we will demonstrate that attributions recover key markers of the cell's identity or regulators of its specialized functions, and are more resistant to technical artifacts than normalized counts, hence enabling cross-dataset analyses (**Figure 1d**).

Building on attributions, we introduce the SIGnature (Scoring the Importance of Genes) python package that enables scalable quantification of gene importance on a large number of datasets. The main application of SIGnature is to leverage attributions to search a gene signature (*i.e.* a set of biologically related genes) across a large corpus of experiments (**Figure 1e)**. To demonstrate its utility, we focus on MS1 monocytes, a subpopulation identified in critically ill patients with conditions such as sepsis and severe COVID-19 [23,24]. Defined by a characteristic gene set, or signature, of approximately 100 upregulated genes, known as the MS1 signature, these cells are hypothesized to have immunosuppressive properties, although their causal role in disease mechanisms remains unclear. Using SIGnature to analyze over 400 diverse studies, we discovered that activation of the MS1 signature is associated with three additional inflammatory conditions: hemophagocytic lymphohistiocytosis (HLH), severe fever thrombocytopenia syndrome (SFTS), and Kawasaki disease (KD). These findings, which we validated through analysis of the original datasets, suggest that MS1-like cells may be implicated in the mechanisms underlying these hyperinflammatory diseases.

Focusing on Kawasaki disease (KD), we observed that the MS1 gene signature decreases following treatment with intravenous immunoglobulin (IVIG), the standard therapy for KD. While the mechanism of IVIG action remains poorly understood, we conjectured that MS1-like cells may be the therapeutic target. To test this hypothesis, we directly exposed MS1 cells to an IVIG factor but observed no significant response. This led us to propose an alternative hypothesis: that KD induces the MS1 phenotype. To test this, we exposed healthy monocytes to serum extracted from KD patients and found a significant increase in the MS1 phenotype compared to treating healthy monocytes with serum from febrile controls. These findings underscore SIGnature's ability to reveal shared disease mechanisms through large-scale gene set scoring.

The following section presents our results in detail. In the first part, we show that attribution scores capture key biological properties, such as marker genes/ programs, that remain consistent even across many integrated datasets. In the second part, we introduce the SIGnature package and its use of attributions for gene set search. By enabling testable hypotheses, such as the involvement of MS1-like cells in Kawasaki disease, SIGnature provides new insights into disease mechanisms and potential therapeutic targets.

Results

Attribution scores reflect gene importance

To begin, we need to choose a foundation model (FM) trained on a vast collection of gene expression profiles from human single cells. The FM's latent representation forms the "biological universe": each point in the latent representation needs to encode gene expression for a single cell, and the distances between points need to be biologically meaningful. FMs trained on scRNA-seq profiles are ideal for this purpose, as they leverage abundant data and allow investigation of a large number of genes. In our study, we tested several models, including SCimilarity [12], scGPT [15], and an SCVI-based FM [11] (**Extended Data Fig**. **1**), but ultimately focused on SCimilarity due to its loss function, which explicitly incorporates cell type annotations to create an embedding that groups functionally similar cells together.

After selecting a trained FM, we compute the attribution scores for each gene in every cell using our method, based on the Integrated Gradients algorithm but adapted to handle multi-dimensional embeddings [25] (**Figure 2a, Methods**). This results in the attribution matrix: a cell-by-gene matrix where the matrix entries are the attribution scores. These scores quantify how important each gene is in determining a cell's position within the latent representation. Simply put, for genes with low attribution scores, variation in their expression would minimally affect the cell's location in the latent space. On the other hand, altering the expression of genes with high attribution scores would lead to a substantial shift in the cell's placement (**Figure 1c**). The SIGnature package includes a pre-trained foundation model (SCimilarity) and the corresponding attribution matrix computed from 22M single cell profiles, 7M of which were the training set of SCimilarity. When applying SIGnature to external datasets, users only need to compute the attribution matrix for those datasets—there is no need for model training or fine-tuning.

Attribution scores computed on external datasets (i.e.not included in the FM's training) reveal genes that are critical to a cell's identity and functional role. We analyzed a lung scRNA-seq dataset from Adams *et al.* [26], comprising approximately 250,000 cells across 16 cell types. First, we computed the attribution matrix of the dataset. We then focused on a specific subpopulation, B cells, and calculated the mean attribution for each gene, comparing it to its mean log-normalized expression (**Figure 2b**). We found that genes with the highest attributions, such as *BANK1* [27], *CD79A* [28] and *MS4A1* [29], are strongly associated with the cell type's unique expression profile and specialized functions. In contrast, genes with the highest expression counts (*e.g. MALAT1*, *RPS27*, and *MT-CO1*) contain mitochondrial and ribosomal genes, as expected. A similar trend was observed by repeating the analysis in multiple subpopulations (**Figure 2c, Extended Data Fig. 2**). In fact, when ranking detected genes in each cell by their expression level and attribution score, cell type markers rank significantly higher in attribution scores (Wilcoxon test $p < 0.01$), whereas mitochondrial and ribosomal genes rank significantly higher in expression scores (Wilcoxon test $p < 0.01$) (**Figure 2d-e)**. This confirms that attribution scores computed on external datasets using SIGnature identify biologically relevant genes.

We then focused specifically on transcription factors (TFs), which are crucial to cell functionality but challenging to detect with scRNA-seq due to their low expression. From the Adams *et al.* lung dataset [30], we only considered cell type markers annotated as TFs. When we compared genes ranked by their attribution versus expression for a given cell, we observed that TF marker genes have higher attribution ranks across all 16 cell types (**Figure 2d**). Additionally, we considered a new dataset focused on CD4+ T cells [31] and analyzed three transcription factors—*GATA3*, *RORC*, and *FOXP3*—recognized as markers for T cell subsets: T helper 2 cells (Th2), T helper 17 cells (Th17), and regulatory T cells (Tregs), respectively. In this study [31], CD4+ T cells were treated with cytokines to induce these specific molecular phenotypes and we found that in each case, when the marker TF was detected, its within-cell attribution ranks were significantly higher than its expression ranks (Wilcoxon test, $p < 0.01$) (**Figure 2f**). These results indicate that the biological information encoded in

the foundation model used for attribution calculations effectively captures the importance of these TFs, despite their low baseline expression.

Finally, we show that attributions are robust to technical artifacts, such as total mRNA counts and dropout noise. Indeed, we note that attribution scores of marker genes are less correlated with total counts per cell compared to expression (**Figure 2g**). For all 16 cell types, attributions showed weaker correlation with cDNA library complexity measures (such as the library size and number of unique genes detected), compared to expression values (**Figure 2h**, **Extended Data Fig. 3**). We also demonstrate that attributions are robust to dropout noise by randomly downsampling counts from B cells from the previously analyzed lung dataset: when removing 50% of the counts from each cell, we observed that 93 of the top 100 genes ranked by mean attribution are consistent with the original, unmanipulated B cell (**Extended Data Fig. 4**).

Attributions facilitate identification of gene programs that are robust across studies

We demonstrated that attributions are biologically meaningful and robust within individual studies. Building on this, we investigated whether attribution values could enable cross-dataset analyses, particularly by aggregating multiple datasets to enhance the identification of biologically relevant gene programs. Typically, dimensionality reduction techniques like non-negative matrix factorization (NMF) are used in scRNA-seq data to identify gene programs [32]. However, these techniques are rarely applied to multi-experiment data due to the difficulty of separating biological signals from technical noise. Here, we show that performing NMF on the attribution matrix, rather than directly on gene expression data, yields insightful results.

We jointly analyzed T cells from three different experiments [26,31,33]. Using NMF, we extracted gene programs from the aggregated attributions, without any feature selection or batch correction (Methods). Our analysis revealed interpretable gene programs, such as CD8+ T cell-specific factors (*CD8A*, *CD8B*, *CD3D*) and cytokine-induced factors (*ISG15*, *IFI6*) which we found enriched in the corresponding cell

types (Figure 2i, Extended Data Fig. 5). Notably, one factor highlights regulatory T cells (Tregs) across all three studies and has high loading scores for key markers *FOXP3* [34] and *IL2RA* [35] (Figure 2j). To assess the robustness of this gene program, we used the learned weights to calculate usage scores for 3200 T-cells from 16 tissues [12] and observed that factor usage is significantly elevated in Tregs compared to other CD4+ T cells (Figure 2k, paired t-test, $p < 0.01$).

We confirmed that these results are robust across various NMF settings, including different random seeds, feature sets, and numbers of dimensions. By sampling random parameter configurations (Methods), 90% of configurations produced a Treg-associated factor, and 98% of those considered FOXP3 as a top 10 gene. In contrast, performing the same analysis with normalized expression yielded less consistent results: factors were often difficult to interpret with only 58% of configurations yielding a Treg-associated factor, and 4% of those cases identifying FOXP3 as a top 10 gene (Methods; Extended Data Fig. 5). These findings highlight the utility of attributions for investigating gene importances across datasets, which we shall leverage in the next section for signature scoring.

Attributions enable signature scoring across scRNA-seq datasets

Signature (or gene set) scoring measures the collective activity of a group of genes within biological samples [4–7,36,37]. Traditional methods assess signature activation by comparing expression levels to background genes, but results are often dataset-dependent and hard to compare across studies. (**Extended Data Fig**. **6**). Alternatives like UCell [36] and JASMINE [37] use within-cell gene rankings to improve comparability, but they face challenges due to the sparsity of scRNA-seq data. To address these issues, we explored averaging attribution scores across genes as a more robust approach for signature scoring.

We begin by demonstrating that the mean attribution scores of marker genes highlight the corresponding cell types. We focused on a single study [38] of peripheral blood mononuclear cells (PBMCs) and assigned a

gene signature for each cell type using the marker genes from CellMarker2.0 [39] and CIBERSORT [40] (**Methods**). In each case, the average attribution score was highest in the expected cell type **(Figure 3a**). Furthermore, cell types with similar functional roles, such as the cytotoxically capable NK cells, CD8+ T cells, and δ-γ T cells, showed partially overlapping activation patterns.

We then demonstrated that attributions enable robust signature scoring across datasets by aggregating two studies: Deng *et al.*[33], containing both CD8+ and CD4+ T cells, and Cano-Gamez *et al.* [31], which includes only CD4+ T cells. We focused on a signature consisting of two known markers for the CD8+ population: CD8 and CD3. For each cell in the two datasets, we calculated the mean attribution score over these markers. Plotting the distribution of attributions for the annotated subpopulations confirmed that this signature distinguishes CD8+ *vs*. CD4+ populations (**Figure 3b**). In contrast, when using the classical *Scanpy* scoring method that is based on expression, CD4+ T cells from the Cano-Gamez dataset showed higher scores than CD8+ T cells from Deng *et al.* (**Figure 3c**). This discrepancy highlights the challenges of comparing dataset-specific scores across experiments and emphasizes the advantage of using attribution-based scoring.

Building on the previous example, we expanded the analysis to a larger scale, analyzing 1.2 million annotated cells from 15 studies comprising 84 cell types (**Supplementary Table 1**). We examined two common immune cell populations, B cells and mast cells, and observed consistent enrichment of the signature in the target cell type (**Figures 3d**). We confirmed that this pattern also holds for rarer cellular subpopulations, such as alveolar epithelial cells in the lung (**Figure 3e**) and specific kidney subpopulations (**Figure 3f**). Since SCimilarity uses cell type annotations during training, we also evaluated whether attributions could highlight cell populations and states not known to the model. From this analysis, we found that attributions can distinguish cell cycle states (**Extended Data Fig. 8**), as well as the cDC1 and cDC2 subtypes of dendritic cells that were not explicitly used during SCimilarity's training (**Extended Data Fig. 9**).

Having established the effectiveness of mean attribution scores in identifying relevant cell populations, we explored their potential for accurate cell type prediction. Since signature scoring of marker genes is commonly employed for cell typing, we compared the mean attribution scores to several established methods: ANS [6] , JASMINE [37], Scanpy [5,7], and UCell [36]. To evaluate cell typing performance, we employed a supervised and an unsupervised approach (**Methods**). For supervised learning, we stratified our dataset into training and testing sets based on cell type and trained a binary logistic regression classifier using gene set scores to predict a specific cell type. In the unsupervised case, we applied Otsu thresholding [41] to the mean gene scores to distinguish cells of the target cell type from others. Average attributions are more predictive than the average expression for all 16 cell types in the supervised task and unsupervised test (**Extended Data Fig. 10**). When comparing against all tested gene set scoring methods, average attributions has the highest F-score in the majority of cases (23/32), with the next best method (ANS) winning 8/32 tests (**Figures 3g-i**, **Extended Data Fig. 10**).

These results conclude the first part of the Results section, demonstrating that attribution scores are effective for detecting signature activation. By using precomputed attributions, SIGnature (Scoring Importance of Genes tool) delivers rapid scoring, covering 22 million cells in minutes, compared to hours or days with conventional methods [42]. In the following sections, we show how SIGnature uncovers a shared inflammatory cell state across conditions, enabling the formulation of testable hypotheses.

Querying attributions across 22M cells reveals MS1-like state in multiple inflammatory conditions

Previous sections established that attribution scores are biologically significant and robust to technical artifacts. Now, we turn to a core application of the SIGnature method: identifying cells with activation of specific gene signatures across large-scale scRNA-seq studies. The key advantage of using attributions is to provide a scoring measure against the entire biological context (atlases), in contrast to traditional gene

scoring/enrichment analysis which focus on comparison within specific groups. Combined with its robustness to technical artifacts, this enables SIGnature to perform gene set scoring across multiple studies.

To demonstrate the potential of our approach, we focus on a poorly understood subpopulation of monocytes, characterized by activation of the MS1 gene signature. These cells share characteristics with myeloid-derived suppressor cells [23,43] and are linked to adverse outcomes in sepsis and COVID-19 [23,24]. This MS1 population has a known gene signature of approximately 100 upregulated genes [23,24]. We aimed to query this signature across 412 studies of various human diseases containing 22 million cells (**Methods**), thereby exploring whether the MS1-like state is present in conditions beyond those previously recognized. Note that 56 of these studies (~7M cells) have annotated cell types and constitute the training set of SCimilarity. For the remaining studies, we predict cell type using the SCimilarity model.

First, we noticed that cells with high attributions for the MS1 signature are those predicted as monocytes and macrophages (**Extended Data Fig. 11**), consistent with literature [23,24], so we focused on those cell types for downstream analyses. We calculated the mean attribution score for the MS1 signature in these cells, and considered the top 10% as MS1-like cells. Subsequently, we investigated the prevalence of MS1-like cells across various diseases for each study and visualized the findings (**Figure 4a**). Our analysis recovered known biology, showing that the MS1 gene signature is most strongly associated with septic shock and a subset of COVID-19 patients. Additionally, when analyzing patients from Stephenson *et al.* [44], those with severe COVID-19 have a greater percentage of MS1-like cells than those with milder symptoms (**Extended Data Fig. 12**). Interestingly, we also identified MS1-like cells in severe hyperinflammatory conditions not previously associated with this cell population (**Figure 4a**): Kawasaki disease (KD), Severe Fever Thrombocytopenia Syndrome (SFTS), and hemophagocytic lymphohistiocytosis (HLH). Both KD and HLH are acute febrile conditions that often occur in children [45,46], whereas SFTS is a specific immune reaction to *Dabie bandavirus* (also known as SFTS virus) [47].

These findings are robust against various MS1 + hit thresholds (**Extended Data Fig. 13**), suggesting that the MS1 signature is a disease biomarker for various inflammatory conditions.

To validate these findings, we re-analyzed the original datasets for these conditions, revealing new insights that were absent from the original publications. In the SFTS study [48], peripheral blood mononuclear cells (PBMCs) were collected from four cohorts: control (healthy), active SFTS infection, SFTS recovery, and fatal SFTS infection. We found that the MS1 attribution scores are lowest in the control and recovery groups, elevated in the infected samples, and highest among the most severe phenotypes of fatal SFTS (**Figure 4b**, **Extended Data Fig. 14**). Similarly, HLH data [49] revealed that the prevalence of MS1-like monocytes decreases following treatment with Nivolumab, a PD-1 inhibitor: on day 0, the hit rate is 47%, which drops to 18% by day 7, 13% by day 21 and 8% by day 67 (**Figure 4c**). Importantly, analysis of the KD dataset [50] revealed a potential therapeutic intervention. Wang *et al.* performed scRNA-seq on PBMCs from six KD patients before and after treatment with intravenous immunoglobulins (IVIG), which is the the standard of care for treating KD, although its mechanism of action is not fully understood [45,50,51]. Upon re-examining this dataset, we observed that MS1-like cells were highly abundant in KD patients *vs.* healthy, but significantly less prevalent after administration of IVIG (**Figure 4d**). Our finding suggests that the IVIG could be beneficial in KD treatment by acting on MS1 monocytes.

Given the novel finding that MS1 cells are present in KD and decrease after treatment with IVIG, we sought further evidence to support that KD cells with high MS1 mean attribution scores are indeed similar to the original MS1 cells found in sepsis and COVID-19. To assess this, we performed consensus NMF [32] on KD monocytes to identify relevant gene programs (**Methods**). We then clustered the KD-specific gene coefficients with the previously calculated coefficients for COVID and sepsis datasets [24] and found that KD program 4 clusters strongly with the MS1 gene programs from other diseases (**Figure 4f**). This program has high coefficients for many core MS1 genes, like S100A9 and RETN and is the only factor whose usage significantly decreases after IVIG treatment in all patients (Mann-Whitney U test, $p < 0.01$)

(**Figure 4g**). Collectively, these observations suggest that the MS1-like cells from KD patients are quite similar to the ones from sepsis and COVID-19 and that IVIG treatment may impact this cell population, which we set out to test.

KD patient serum induces MS1 phenotype, but KD treatment, IVIG, does not directly affect MS1 cell induction

Our analysis indicated that PBMCs from Kawasaki Disease (KD) patients exhibit elevated levels of MS1-like monocytes. To assess whether factors in KD patient serum contribute to increased MS1 cell abundance, we conducted an induction experiment based on the approach utilized by Reyes *et al.* [24]. In their study, serum from bacterial sepsis or severe COVID-19 patients induced emergency myelopoiesis in hematopoietic stem and progenitor cells (HSPCs), resulting in high monocyte production and a significant increase in the proportion of MS1-like cells ($CD45^{+}$, HLA-DRA^{low}).

Implementing a similar methodology, we compared serum from KD patients with that from febrile infant controls to assess its effect on monocyte differentiation. Our results confirmed the cells treated with KD serum contain a significant increase in monocyte production and the proportion of HLA-DRA^{low} monocytes (**Figure 4g**). These findings suggest that circulating factors in KD patient serum may promote the expansion of this disease-associated monocyte population.

Additionally, by analyzing the original data [50], we observed a decrease in the percentage of MS1-like cells in KD patients following IVIG treatment. However, since the scRNA-seq analysis used to establish this association does not provide information on whether IVIG directly targets this cell population, we sought to investigate its direct impact. To do so, we evaluated the effect of IgG (the primary component of IVIG) on the rate of MS1 cell induction. Our findings showed no significant change in proportion of MS1

monocytes (**Figure 4h**), suggesting that IgG does not directly influence the generation of this disease-associated cell type. These findings highlight the need for further investigation into the mechanisms by which IVIG modulates immune responses in KD.

Discussion

The central contribution of our manuscript is the introduction of gene-level attributions from a self-supervised foundation model (FM) as a metric of gene importance. Instead of relying on comparative abundances of gene expression, we assess the impact of each gene on the cell's representation from a scRNA-seq FM. By analyzing 1.2 million cells from 15 datasets, we found that attributions are elevated in genes associated with a cell's identity and specialized functions. Additionally, attributions are more robust than normalized counts to scRNA-seq QC metrics and do not require dataset-specific normalization, making them useful for cross-study analyses like NMF and gene set scoring. We demonstrated the utility of attributions for drug discovery by querying the MS1 gene set across millions of cells, which resulted in the discovery of novel disease associations and testable hypotheses for drug repurposing.

While explainable AI techniques have quantified gene importance in supervised transcriptomic classifiers [52–54] and methods like Cell2Sentence [55] enhanced scRNA-seq interpretability using foundation models, our key innovation is using attribution scores to harness the vast information embedded in scRNA-seq FMs like SCimilarity to create a standardized metric of gene importance. The biological insights captured by the FM boost signal for crucial genes with low baseline expression, like TFs, while technical information learned from training across diverse datasets enhances the generalizability of the output scores. Although count data can be used to derive gene importance in individual studies through feature selection and differential expression analysis, it is not straightforward how to apply these procedures to massive scRNA-seq atlases. Attributions, however, provide a unified framework for measuring gene importance applicable to any cell. These values are meaningful and robust without the need for dataset-specific normalization, allowing for precomputation across huge atlases and rapid downstream processing. Collectively, these features of attributions enabled the development of SIGnature, the first publicly available tool for cell-level gene set scoring across scRNA-seq atlases.

While attributions provide many advantages when analyzing gene importance, this approach also has limitations. First, attributions are less interpretable than gene expression; comparing abundances has a clearly defined meaning, whereas understanding why a gene has a specific attribution score is less straightforward. Second, the output values depend on the underlying FM and explainability method. Our work focused on applying IG to SCimilarity [12], but SCimilarity was trained on 10X scRNA-seq data without any tumor cells, suggesting that attributions collected against cancer data or non-10X platforms may be less reliable. Additionally, quantifying importance by measuring how changes in a gene affect a cell's representation is more suitable for FMs trained to generate geometrically-meaningful, cell-focused embeddings using reconstruction or contrastive loss, like SCimilarity and SCVI, than for transformer-based FMs, like scGPT, that build embeddings with masked prediction loss. Future research could explore new approaches to leverage transformer-based FMs or new multi-modal FMs [56,57], utilize attributions for guided deconvolution [58,59], or build new FMs with cell-focused embeddings tailored to specific biological functionalities. Collectively, this manuscript demonstrates the potential of FM attributions for measuring gene importance, paving the way for further research to harness the knowledge embedded in scRNA-seq FMs for diverse and specialized biological applications.

Methods

Calculating attributions using integrated gradients and foundation models

Several methods exist for calculating feature attributions [60–64], with Integrated Gradients [61] being among the most widely used. Originally developed for supervised models, these techniques highlight key input features that influence a model's prediction, such as individual pixels important for image classification. However, scRNA-seq foundation models are trained in an unsupervised (or semi-supervised) setting to generate biologically meaningful high-dimensional embedding spaces. This makes it challenging to determine how variations in gene expression influence the multiple latent variables within these representations. Moreover, computing attribution scores for each latent dimension is computationally intensive, limiting the feasibility of these methods in such contexts.

Here, building on our previous work [25], an efficient variation of IG for multi-dimensional embeddings was implemented using *Captum* [65]. This approach adds a final layer to the original scRNA-seq FM, which calculates a weighted sum of each embedding dimension to yield a single output. Formally, let $f\colon X(R^{n}) \rightarrow H(R^{d_h})$ represent the scRNA-seq FM. The attribution score for the ith dimension of an input X is defined as follows:

$$IG\left(\sum_{k=1}^{d_{hy}} [f_k(X) \ * f_k(X_i \bullet)],\ X\right)$$

The IG operation begins by calculating gradients for expression profiles interpolated between a zero-vector baseline and the observed cell state (**Figure 2a**). The zero baseline, a standard in IG applications for image classification, allows consistent use across cells without dataset-specific adjustments. These gradients are summed across interpolations, multiplied by the original gene

expression vector, and re-normalized to a standard scale factor (i.e. 1000), to yield the final attribution score for each gene, quantifying its influence on the cell's representation in the embedding. Leveraging the SCimilarity model [12], this method processes approximately 200 cells per second on a single A100 GPU.

Attribution scoring across scRNA-seq foundation models

Gradient-based measures of gene importance were implemented for three foundation models: SCimilarity [12] (v. 2023_01_rep0) , scGPT (whole-human model) [15], and an SCVI-based model trained on cellxgene data (2024-02-12-scvi-homo-sapiens) [11]. Scores were calculated on the Adams *et al.* [26] lung dataset using the 4014 genes considered by all three models. These cells were all re-processed for the reduced set of genes using log-normalization with a scale factor of 10,000.

Attribution scores were calculated for the SCimilarity embedding (128 dimensions) and the mean encoding layer of the SCVI model (200 dimensions) using the modified integrated gradients approach described above. While SCimilarity and SCVI use normalized expression as input, scGPT ranks genes within a cell, bins those rankings, and uses the bins as input. This ranking scheme is incompatible with integrated gradients, as interpolated values between 0 and the true expression vector always result in the same ranking and binning. However, gradients were calculated for each input feature against the CLS token of 512-dimensional embedding and multiplied by the true expression vector to get attribution scores.

To compare each foundation model with normalized expression, the analysis from **Figure 2d** was repeated for attributions collected using each model (**Extended Data Fig. 1**). In this approach, all the genes in a cell with non-zero values are rank-normalized from 0 to 100 using a given attribution score and then these ranks are averaged across four gene families for each cell type: cell type marker genes,

ribosomal genes, cell-type associated transcription factors, and mitochondrial genes. **Extended Data Fig. 1** shows that attributions collected against SCimilarity's embedding yield higher scores for markers and cell type-associated TFs and generally lower scores for ribosomal and mitochondrial genes.

Datasets and gene sets

The SCimilarity validation dataset was used for many of the analyses in this manuscript because it contains 1.2 million cells with known cell type annotations. **Supplementary Table 1** summarizes the 15 studies included in this dataset and lists the analyses where each was utilized. Additionally, many tests in this manuscript used established gene sets associated with particular cell types or molecular phenotypes. **Supplementary Table 2** lists the genes included in each gene set and the cell type annotates with which they were associated.

Plotting and statistics

Plotting was performed in python using *seaborn* (v. 0.13.2) [66] and *matplotlib* (v. 3.9.0) [67]. Statistical tests were implemented using *scipy* (v. 1.14.0) [68] and *statsmodels* (v. 0.14.4) [69] .

Top attributions per cell type

This analysis was first performed on a subset of the Adams *et al.* lung dataset [26], including the 16 cell types that have a corresponding marker gene set within MSigDB [2,70]. For each of these 16 cell types, the average log-normalized gene expression and the average attribution values were calculated for every gene. **Figure 2b** shows a scatterplot of gene values in B cells, where the x-axis indicates the standardized average expression values and the y-axis shows the standardized average attribution values. **Figure 2c** shows the top 5 ranking genes by average attribution and expression for each cell type.[71]

The average rank analysis (**Figure 2d-e**) was performed on a per cell basis, with non-zero gene values rank-normalized from 0 to 100 in ascending order using the "rankdata" method from *scipy* [68] and taking the average rank for tied values. For each cell type, the average rank percentile was calculated for each gene in every cell it was detected. **Figure 2d** shows the summary for four gene families: lung cell type marker genes from MSigDB [2,70], ribosomal genes annotated from KEGG [72,73], mitochondrial genes (starting with MT-), and marker genes from MSigDB that are also transcription factors [30]. The values for each individual cell type and gene family are shown in **Extended Data Fig. 2.**

Robustness analysis

Each B cell from Adams *et al.* [26] was downsampled by randomly removing X% of the counts in each cell for the percentages: 0%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%. These downsampled datasets were then log-normalized and used to calculate attribution scores. For each downsampling percentage, the average attribution score for each gene was calculated across all B cells and ranked in descending order. The top scoring genes were robust to downsampling, with 93 of the top 100 genes remaining in the top 100, even after 50% of the counts had been removed (**Extended Data Fig. 4**).

Correlation analysis

For each of the 16 cell types analyzed from Adams et al. [26], the average log-normalized expression and average attribution scores were collected for marker genes in each cell. The Spearman correlation was then calculated between these values and three scRNA-seq QC metrics: number of UMIs, number of unique genes, and percent of UMIs from mitochondrial genes. **Figure 2g** shows an example comparing the total UMI in each cell with the average counts and attributions for non-classical monocytes. The results across all cell types are summarized in **Figure 2h** and **Extended Data Fig. 3**.

Cross-study non-negative matrix factorization

Non-negative matrix factorization (NMF) was implemented using *scikit-learn* (v. 1.5.0) [74] on author-annotated T cells from three studies: Adams *et al.* [26], Cano-Gamez *et al*. [31], and Deng *et al.* [33]. Analysis included 14,056 genes detected across all three studies. NMF was performed 200 times for both expression and attributions, varying three parameters: number of factors (5, 10, 15, 20, 25), random state (114-133), and feature set (all 14,056 genes or 2,591 highly variable genes).

For each run, a "Treg-associated factor" was identified by calculating t-statistics between Tregs and non-Tregs across all nine possible within-study and cross-study comparisons (*scipy* “ttest_ind”, alternative= “greater”). The factor with the highest minimum t-statistic across these nine comparisons was selected as the Treg-associated factor. The Treg-associated factor was considered to robustly highlight Tregs if that the worst performing Treg vs. non-Treg comparison was significant, using a Bonferroni-adjusted p-value of 2.7E-6 (alpha=0.01 divided by 3600 tests: 400 trials, each with 9 comparisons). **Figures 2i-k** show results from a specific parameter set (10 factors, all genes, random_state=114), though similar trends were observed across most parameter combinations.

Generalization analysis was performed using 3,200 cells from the SCimilarity database [12] that were not used for model training or validation. Sixteen tissues containing at least one hundred confidently predicted Tregs and CD4+ T cells (nn_prediction_dist < 0.02) were included, with one hundred cells randomly sampled from each population in each tissue. The NMF models trained on the original T cell cohort were applied to this independent dataset to assess generalization (**Extended Data Fig. 5**).

Dataset specificity and scanpy scoring

The “score_genes” function from *Scanpy* was implemented on two datasets: Deng *et al.* [33] and Cano-Gomez *et al.* [31]. The gene set queried included all isoforms of CD3 (*CD3D*, *CD3E*, *CD3G*, and

CD247) and of CD8 (*CD8A* and *CD8B*). **Figure 3b** shows that CD8+ and CD4+ T cells are separated in the Deng *et al.* [33] dataset, but CD4+ T cells from Cano-Gomez *et al.* [31] score higher for the CD8+ T cell marker genes than the CD8+ T cells from the Deng *et al.* [33] dataset. **Figure 2a** shows the average attributions for those 6 genes are clearly elevated in CD8+ T cells compared to CD4+ T cells. Density plots were created using the "kde_plot" function in *seaborn* [66] with standard parameters.

Additionally, **Extended Data Fig. 6** demonstrates that *Scanpy*'s "score_genes" function produces dataset-dependent results. When applied to cDC2 marker genes [75] across the same cells in different contexts (full Adams *et al.* [26] dataset versus a subset containing only dendritic cells and mast cells), the scores shift systematically, highlighting the method's sensitivity to background cell composition.

Implementing established gene set scoring methods

Gene set activation was evaluated by comparing mean attributions against five existing methods: Adjusted Neighborhood Scoring (ANS) [6], JASMINE [37], Mean Expression, Scanpy [5,7] , and UCell [36]. For all methods, data were log-normalized in Scanpy (v. 1.9.2) with a target sum of 10,000. Scores were calculated separately for each dataset to accommodate methods using background distributions.

Mean Expression averages log-normalized expression for genes of interest. Scanpy scoring uses the "score_genes" function with default parameters, which is based on the Tirosh *et al.* [4,7] approach, comparing query gene expression against background genes with similar expression levels. ANS [6] also uses control genes but identifies backgrounds through gene expression neighbor graphs rather than expression buckets. UCell [36] and JASMINE[37] are both rank-based gene set scoring approaches; UCell calculates Mann-Whitney U statistics from expression ranks [36], while JASMINE computes likelihood or odds ratios from within-cell gene ranks [37]. ANS, UCell, and JASMINE were implemented using the

ANS_signature_scoring package [6] , with ANS using a control_size of 50 and JASMINE using the "likelihood" model.

Gene set scoring across scimilarity validation dataset

The gene sets used in this study are summarized in **Supplementary Table 2.** Immune gene sets from CellMarker 2.0 [39] and CIBERSORT [40] were applied to PBMCs from Van Der Wijst et al.[38]. Seven cell types with matching gene sets from both sources were analyzed. For each cell, consensus scores were created by averaging marker gene scores from both sources for the same cell type. These consensus scores were then averaged by annotated cell type identity and displayed in **Figure 3c**.

For analysis across all 15 SCimilarity datasets, B cells and mast cells were considered as these populations appeared in the most unique studies (7 and 6 respectively). CIBERSORT [40] signatures were used to query these immune cell populations across datasets (**Figure 3d**). Lung and kidney epithelial cell signatures were collected from CellMarker 2.0 [39] (**Figure 3e,f)** and dendritic cells were assessed using both CellMarker 2.0 and from Table 1 of Collin *et al.* [75] (**Extended Data Figure 9**).

B cell and mast cell signatures were analyzed across 15 SCimilarity validation datasets, as these populations appeared in the most unique studies (7 and 6 respectively). **Extended Data Fig. 8** uses cell cycle gene sets for S and G2/M phases from Tirosh *et al.* [48] to show that attributions can identify these cell states across cell types, including annotated fibroblasts [76].

Supervised and Unsupervised Cell Type Prediction

Cell type classification performance was evaluated using six gene set scoring methods: ANS [6], JASMINE [37], Mean Attribution, Mean Expression, Scanpy [5,7], and UCell [36]. For each cell type from Adams *et al.* dataset [26] and each method, marker gene set scores (**Supplementary Table 2**) were used to train

univariate logistic regression classifiers using LogisticRegressionCV from scikit-learn (v. 1.2.0) [74] with balanced class weights and a random state of 114. Performance was assessed across 25 independent train/test splits (80%/20%) using F-score to account for class imbalance

Unsupervised classification employed Otsu's multi-thresholding method [41], using the "threshold_multiotsu" function from scikit-image (v. 0.19.3)[77]. For each cell type, marker gene set scores were calculated for each cell and optimal thresholds determined. Multiple threshold numbers (n=1,2,3) were evaluated, with optimal N selected by maximizing Cohen's d effect size between the top group and second-highest group, ensuring robust separation between populations. This approach provides an objective threshold selection method when true labels are unavailable. The cells above the chosen threshold were considered "hits" and used to calculate an F_1-score against the true labels.

MS1 Full Dataset Queries

The MS1 phenotype in monocytes was established through NMF analysis by Reyes *et al.* (2021) [10]. Attributions were queried across the SCimilarity cell database [12] using the top 99 genes associated with this factor to identify other cells activated for this phenotype. This search queries 22M cells across 412 datasets, which was then reduced to 2.3 cells from 244 studies that were confidently predicted as monocytes or macrophages (prediction_nn_dist < 0.02). For sample-level hit analysis (**Figure 4a**), samples were considered if they contained 25 monocytes/macrophages and were from a disease with 3 unique samples. Hits were called as the top 10% of monocytes or macrophages by MS1 score and the percentage of hits was calculated for each sample (**Figure 4a**). Other thresholds (5%, 15%, and 20%) were also tested and showed similar top diseases (**Extended Data. Fig 13**).

MS1 Cells in Individual Datasets

The MS1+ hits from the SCimilarity database were analyzed in the context of their respective datasets. Individual analyses were performed for multiple diseases using metadata from the original study: COVID-19 from Stephenson *et al.* [44], HLH from Liu *et al.* [49], SFTS from Park *et al.* [48] , and KD from Wang *et al.* [50].

Consensus NMF was run on scRNA-seq data from the 8282 monocytes/macrophages from Wang *et al.* [50] using the *cNMF* python package [32] in order to compare results against COVID-19 and sepsis samples processed in Reyes *et al.* [24]. The expression data from these cells were run through the *cNMF* package using the top 3000 variables genes, considering k values from 5 to 20, and repeating each trial 25 times. The optimal k was chosen as 14 based on the balance between high stability and relatively low error (**Extended Data Fig. 15**).

The standardized gene loadings from the KD monocytes were compared against the standardized gene loadings from Reyes et al. [24] by computing pairwise Pearson correlation coefficients. These values were clustered using the *seaborn* clustermap function with the "correlation" metric [66]. **Figure 4e** shows a subset of this clustered matrix, with the entire clustermap displayed in **Extended Data Fig. 16**. For each patient, a Mann Whitney U test was performed between the usage values for each cell collected before IVIG against those collected after IVIG using *scipy* [68].

MS1 Induction Experiments

Differentiation of myeloid cells from CD34+ bone marrow hematopoietic stem and progenitor cells (HSPCs) was performed as previously described [24]. HSPCs cells were cultured in SFEM II with 75 nM StemRegenin 1, 3.5 nM UM729, 1X CC110 (StemCell Technologies), and 1X penicillin-streptomycin (Gibco). To initiate differentiation, cells were cultured in the same medium supplemented with 20%

pooled healthy human serum (SeraCare) or serum from diseased patients. Serum samples from patients with multiple myeloma were obtained from BioIVT.

To assess MS1 induction, cells were stained with the following panel: CD14-FITC (clone M5E2), CD15-APC (clone W6D3), CD11b-AF700 (clone ICRF44), CD34-BV650 (clone 561), HLA-DR-PE/Cy7 (clone L243) (BioLegend) and IL1R2-PE (clone 34141; Thermo). Cells were resuspended in FACS buffer with 5% CountBright beads (Invitrogen) to allow determination of absolute counts during analysis. Flow cytometry data were acquired on an LSR Fortessa (BD Biosciences) and analyzed using FlowJo v10.10. For analysis of HLA^{low} monocytes, the average HLA intensity value was calculated for internal controls and then for each sample, the percentage of monocytes with HLA intensity below that value was calculated. Mixed linear analysis was performed using the *statsmodels* [69] package in python.

Figures

Figure 1

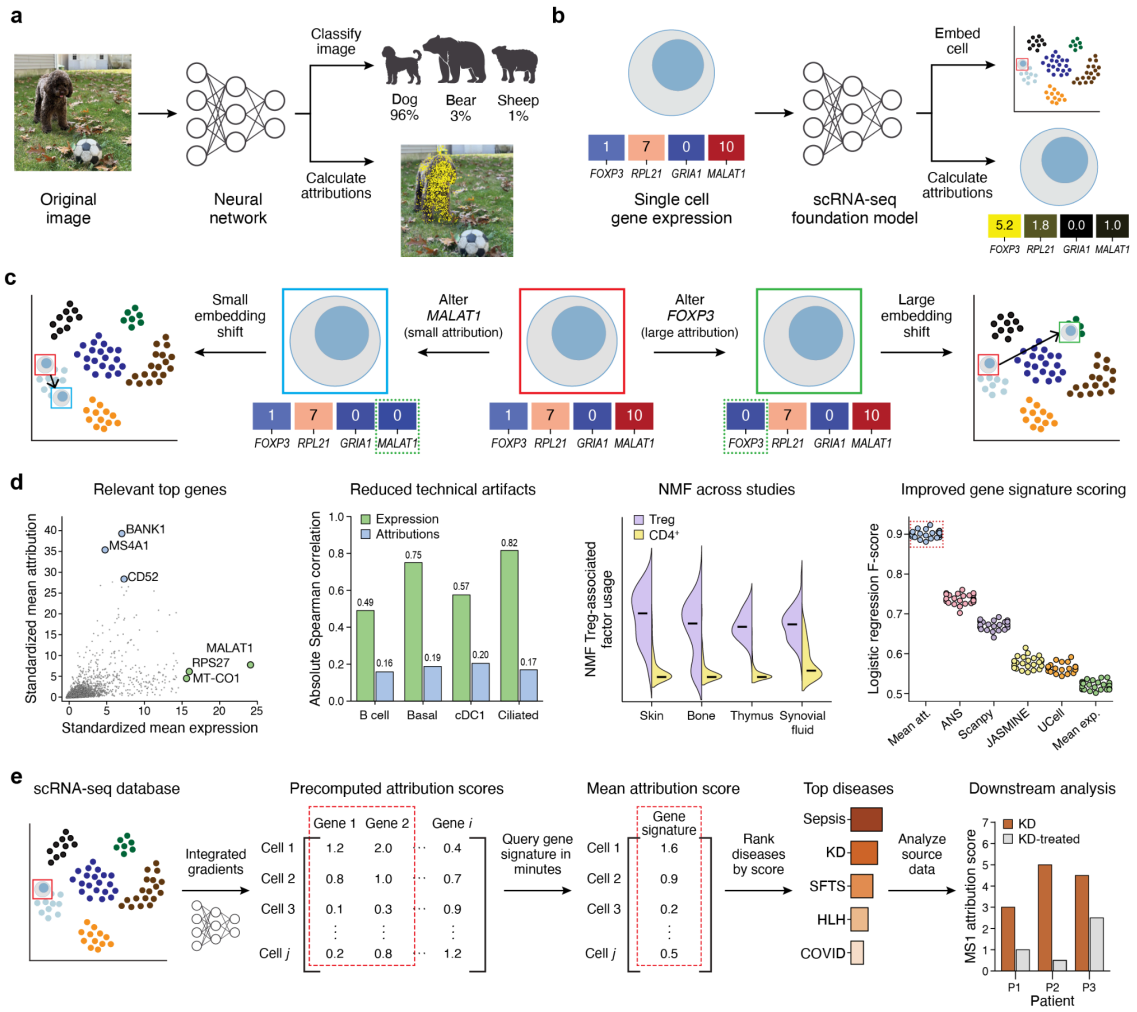
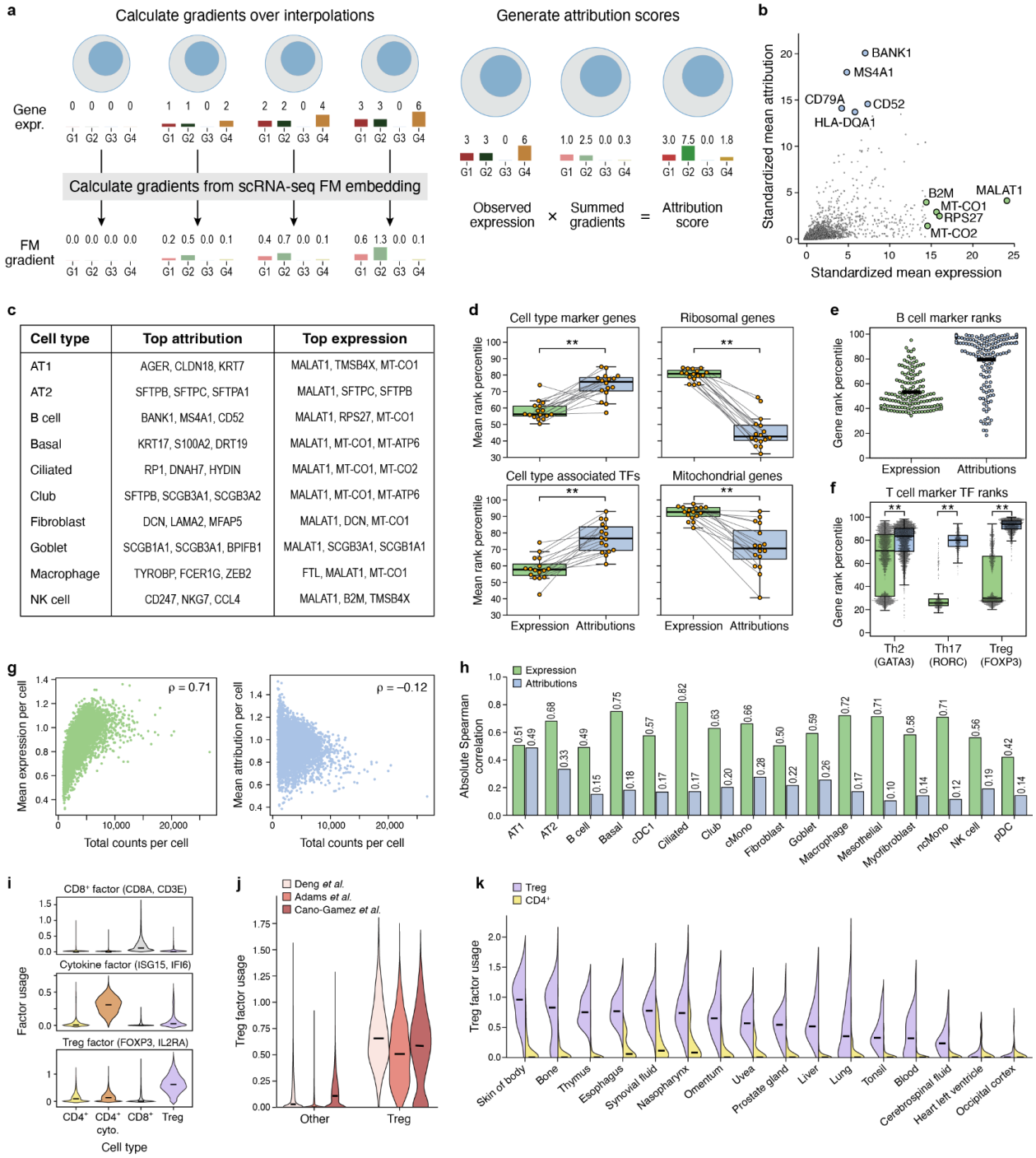


Figure 1: **Gene importance is quantified through attributions**. **a**. In image classification, attribution scores highlight key pixels (yellow) used by the model for the highest probability prediction (dog). **b**. In a scRNAseq foundation model, attributions quantify each gene's importance to the cell's latent embedding position. Attribution values do not necessarily correlate with expression (*e.g. MALAT1 vs. FOXP3*). **c**. Impact of gene expression alteration on embedding: altering *MALAT1* (low attribution) has little effect; altering *FOXP3* (high attribution) causes major shifts. **d**. Key features of attributions in scRNA-seq analysis. Attributions enhance marker gene detection, reduce technical artifacts compared to log-normalized expression, and improve both cross-study NMF and gene signature scoring. **e**. SIGnature search workflow. Pre-computed attributions enable fast cell-level scoring for the genes in the signature (red); atlas data reveal signature enrichment in cell types and diseases.

Figure 2



c

Cell type	Top attribution	Top expression
AT1	AGER, CLDN18, KRT7	MALAT1, TMSB4X, MT-CO1
AT2	SFTPB, SFTPC, SFTPA1	MALAT1, SFTPC, SFTPB
B cell	BANK1, MS4A1, CD52	MALAT1, RPS27, MT-CO1
Basal	KRT17, S100A2, DRT19	MALAT1, MT-CO1, MT-ATP6
Ciliated	RP1, DNAH7, HYDIN	MALAT1, MT-CO1, MT-CO2
Club	SFTPB, SCGB3A1, SCGB3A2	MALAT1, MT-CO1, MT-ATP6
Fibroblast	DCN, LAMA2, MFAP5	MALAT1, DCN, MT-CO1
Goblet	SCGB1A1, SCGB3A1, BPIFB1	MALAT1, SCGB3A1, SCGB1A1
Macrophage	TYROBP, FCER1G, ZEB2	FTL, MALAT1, MT-CO1
NK cell	CD247, NKG7, CCL4	MALAT1, B2M, TMSB4X

Figure 2: **Attributions are biologically meaningful and useful for cross-study analyses**. **a**. Attribution scores using a scRNA-seq FM are computed using gradients from interpolated expression profiles. **b**. Standardized mean expression (x-axis) and attribution (y-axis) for genes (dots) across B cells [26]. Larger dots, labeled with corresponding genes, indicate top expression (green) and attribution (blue). **c**. Top genes by mean attribution (left) and expression (right) across lung cell types. **d**. Gene family rank comparison across lung cell types for four gene families (panel titles). Each dot shows a cell type's average rank percentile for genes of that family; green and blue boxplots show expression and attribution quartiles, respectively. Same cell types are connected by thin black lines. (**p<0.01; Wilcoxon). **e**. Marker gene comparisons for B cells. Dots represent a marker gene's average rank percentile across B cells. Thick black lines show mean values. **f**. In cytokine-treated T cells [31] (Th2, Th17, Treg), lineage TFs (*GATA3*, *RORC*, *FOXP3*) rank higher by attribution than expression (**p<0.01; Wilcoxon). **g**. Correlation of marker gene (dot) in non-classical monocytes (n=5,616): expression correlates strongly (ϱ=0.71), attributions minimally (ϱ=-0.12). **h**. Marker gene score correlations (y-axis) between expression (green) or attribution (blue) *vs.* sequencing depth across lung cell types (x-axis). Expression shows stronger correlation than attribution. **i**. NMF factor usage (y-axis) by cell type (x-axis). Factors shown correspond to CD8+ T cells (top), interferon response (middle), and Tregs (bottom). **j**. Treg factor usage (y-axis) across datasets used for NMF (legend). **k**. Treg NMF factor usage across tissues. Cross-study NMF factor applied to 100 predicted Tregs cells (purple) and 100 CD4+ T cells (yellow) not included in the original NMF. Treg-specific factor scores are significantly elevated in predicted Tregs across all tissues (**p<0.01; paired t-test).

Figure 3

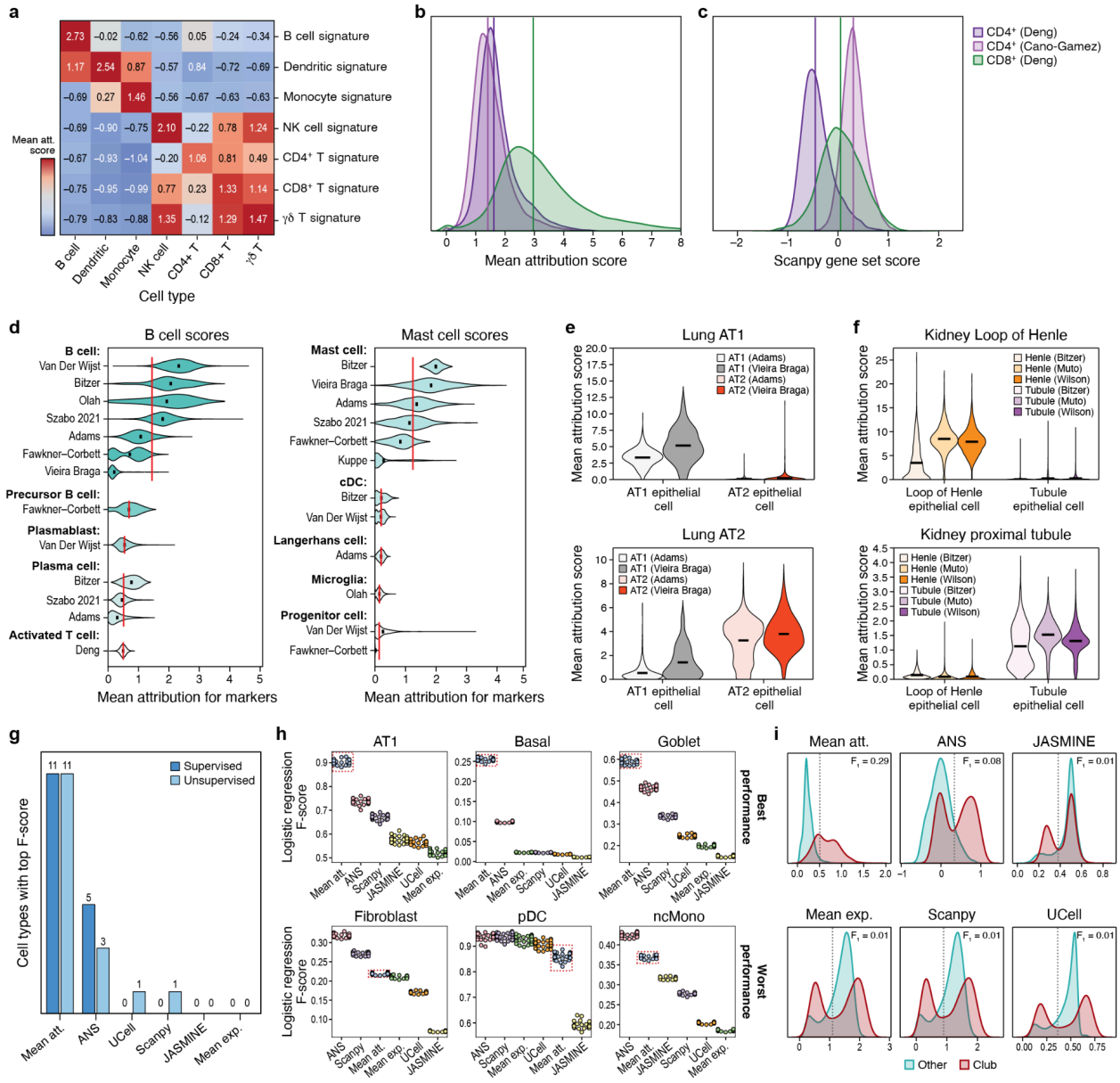

a
Mean att. score
B cell signature
Dendritic signature
Monocyte signature
NK cell signature
CD4+ T signature
CD8+ T signature
γδ T signature
B cell
Dendritic
Monocyte
NK cell
CD4+ T
CD8+ T
γδ T
Cell type
b
Mean attribution score
c
CD4+ (Deng)
CD4+ (Cano-Gamez)
CD8+ (Deng)
Scanpy gene set score
d
B cell scores
Mast cell scores
Mean attribution for markers
e
Lung AT1
Lung AT2
AT1 (Adams)
AT1 (Vieira Braga)
AT2 (Adams)
AT2 (Vieira Braga)
Mean attribution score
AT1 epithelial cell
AT2 epithelial cell
f
Kidney Loop of Henle
Kidney proximal tubule
Henle (Bitzer)
Henle (Muto)
Henle (Wilson)
Tubule (Bitzer)
Tubule (Muto)
Tubule (Wilson)
Loop of Henle epithelial cell
Tubule epithelial cell
g
Supervised
Unsupervised
Cell types with top F-score
Mean att.
ANS
UCell
Scanpy
JASMINE
Mean exp.
h
AT1
Basal
Goblet
Fibroblast
pDC
ncMono
Logistic regression F-score
Best performance
Worst performance
i
Mean att.
ANS
JASMINE
Mean exp.
Scanpy
UCell
Other
Club

Figure 3: Average attributions measure gene set activation. a. Heatmap displaying standardized average attributions for marker genes (y-axis) in PBMC cell types (x-axis) (570,249 cells) [38]. **b.** Density plots showing average attribution scores CD8+ T cell marker genes. CD8+ T cells from Deng *et al.* [33] (n=56,337, green) have higher scores compared to CD4+ T cells from Deng *et al.* [33] (n=16,794, purple) and Cano-Gomez *et al.* [31] (n=40,284, light purple). Medians indicated by vertical lines. **c.** Scanpy scoring of the same CD8+ gene signature assigns CD4+ T cells from Cano-Gomez *et al.* [31] higher scores than CD8+ T cells from Deng *et al.* [33] **d.** Marker gene attributions for B cells (left) and mast cells (right) across studies. Top five cell types by average study-level score are shown. Bolded text indicates cell type and y-axis label shows source study. X-axis indicates mean attribution score for marker genes. Black lines indicate mean values for each dataset, and red lines show the average of the study means for each cell type. **e.** Marker gene attribution scores for lung epithelial cell subtypes across multiple studies [26,78]. **f.** Marker gene attribution scores for kidney epithelial cell types across multiple studies [79–81] **g.** Classification performance for each scoring method (x-axis) across lung cell types (y-axis). **h.** Supervised classification F_1-scores for top 3 best and worst performing cell types for attributions. Each dot represents the F_1-score for one of 25 cross-validation runs. Red boxes highlight results for mean attributions. **i.** Distribution plots comparing club cell marker gene scores in club cells (maroon, n=2,492) *vs.* other lung cells (teal, n=279,302). Dashed lines indicate the Otsu threshold for determining F_1-score.

Figure 4

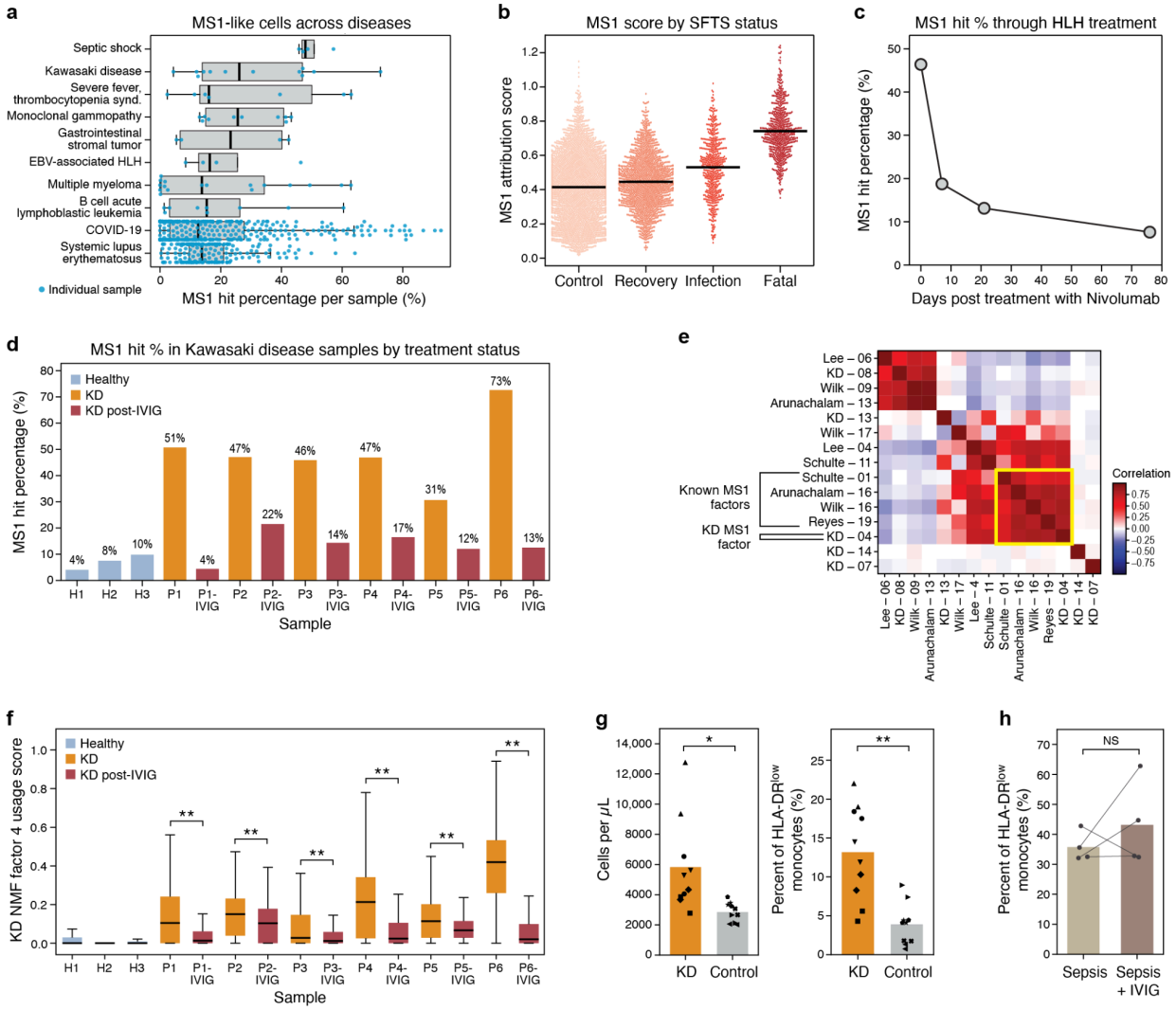

a
MS1-like cells across diseases
Septic shock
Kawasaki disease
Severe fever, thrombocytopenia synd.
Monoclonal gammopathy
Gastrointestinal stromal tumor
EBV-associated HLH
Multiple myeloma
B cell acute lymphoblastic leukemia
COVID-19
Systemic lupus erythematosus
Individual sample
MS1 hit percentage per sample (%)
b
MS1 score by SFTS status
MS1 attribution score
Control
Recovery
Infection
Fatal
c
MS1 hit % through HLH treatment
MS1 hit percentage (%)
Days post treatment with Nivolumab
d
MS1 hit % in Kawasaki disease samples by treatment status
Healthy
KD
KD post-IVIG
MS1 hit percentage (%)
Sample
e
Known MS1 factors
KD MS1 factor
Correlation
f
KD NMF factor 4 usage score
Healthy
KD
KD post-IVIG
Sample
g
Cells per μL
KD
Control
Percent of HLA-DR low monocytes (%)
h
Percent of HLA-DR low monocytes (%)
NS
Sepsis
Sepsis + IVIG

Figure 4: MS1 attribution scores are elevated in hyperinflammatory conditions. a. MS1 hit percentage by disease. The x-axis displays the percentage of monocytes/macrophages classified as MS1 hits per sample, while the y-axis lists the top 10 diseases, ordered by mean hit percentage. Each dot represents a sample, and boxplots display quartiles, excluding outliers greater than 1.5 IQR. **b.** MS1 attribution scores across SFTS disease phases. Black lines indicate group averages. **c.** Decrease in MS1-like monocytes during Nivolumab treatment in an HLH patient. **d.** MS1 hit percentage in Kawasaki disease pre- and post-IVIG, showing reductions post-treatment. **e.** Clustering of NMF coefficients for monocyte gene programs. KD programs cluster with COVID-19 and sepsis [24]; KD Factor 4 correlates with MS1 signatures (Arunachalam-16, Reyes-19, Schulte-01, Wilk-16) (**Extended Data Fig. 16**). **f.** KD Factor 4 scores by treatment status, with decreased usage post-IVIG (**$p<0.01$, Mann-Whitney U test). Boxplots exclude outliers > 1.5 IQR; cell counts: H1 (n=442), H2 (n=200), H3 (n=355), P1 (n=418), P1-IVIG (n=1,767), P2 (n=576), P2-IVIG (n=144), P3 (n=586), P3-IVIG (n=181), P4 (n=640), P4-IVIG (n=121), P5 (n=1,222), P5-IVIG (n=390), P6 (n=833), P6-IVIG (n=407). **g.** KD serum induces emergency myelopoiesis and increases MS1-like cells. Left: Elevated monocyte concentration with KD serum compared to febrile controls (*$p<0.05$, mixed linear model). Right: Significant increase in HLA-DR^{low} monocytes (**$p<0.01$, mixed linear model). IgG has no effect on percentage of HLA-DR^{low} monocytes using sepsis serum (NS $p=0.47$, paired t-test). Thin lines connect the same patient between untreated (left) and treated (right) conditions.

Data Availability

The datasets used in this manuscript are summarized in **Supplementary Table 1**. Precomputed attribution scores for the SIGnature package can be found on Zenodo (https://zenodo.org/communities/signature).

Code Availability

The SIGnature package is available on Github (https://github.com/Genentech/SIGnature).

Acknowledgements & Funding

We thank J. Collier, Y. Lee, A. Andersson, B. Chen, S. Malek, and M. Crow for their suggestions on the manuscript and L. Gaffney for her assistance in figure generation. We also thank S. de Ferranti and K. Bonello for their help in biological interpretation and sample procurement.

MBS is supported by the Samara Jan Turkel Clinical Center for Pediatric Autoimmune Disease.

Author Contributions

MPG, GH, and TB designed the study; MPG developed the method with ND, GH, AT, HCB, TK, TB, and EH; Model implementation was performed by MPG, ND, TK, and EH; Python API was developed by TK, with help from MPG; MPG performed the data analysis with guidance from GH, TB, AB, AT, HCB, MR, and SBK; MPG and MR conceived of the biological validation, with input from SBK, PYL, MBFS, and JWN; PYL, MBFS, and JWN provided advice on biological interpretation and coordinated sample procurement; MR performed the experimental validation. MPG, GH, and TB wrote the manuscript with input from EH and HCB; All authors reviewed the manuscript.

Competing interests

AB, TB, HCB, ND, MPG, EH, GH, TK, MR, GS, & AT are employed by Genentech or Roche.

References

1. Love, M., Anders, S. & Huber, W. Differential analysis of count data – the DESeq2 package. *Genome Biol.* (2013).
2. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **102**, 15545–15550 (2005).
3. Hänzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).
4. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
5. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
6. Ciernik, L., Kraft, A., Barkmann, F., Yates, J. & Boeva, V. ANS: Adjusted Neighborhood Scoring to improve assessment of gene signatures in single-cell RNA-seq data. *bioRxiv* 2023.09.20.558114 (2023).
7. Tirosh, I. *et al.* Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189–196 (2016).
8. Wang, B. *et al.* SIMLR: A Tool for Large-Scale Genomic Analyses by Multi-Kernel Learning. *arXiv* **18**, 1700232 (2017).
9. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).

10. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

11. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep Generative Modeling for Single-cell Transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).

12. Heimberg, G. *et al.* A cell atlas foundation model for scalable search of similar human cells. *Nature* (2024) doi:10.1038/s41586-024-08411-y.

13. Hao, M. *et al.* Large Scale Foundation Model on Single-cell Transcriptomics. *bioRxiv* 2023.05.29.542705 (2023).

14. Yang, F. *et al.* scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence* **4**, 852–866 (2022).

15. Cui, H. *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat. Methods* **21**, 1470–1480 (2024).

16. Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2022).

17. Bian, H. *et al.* scMulan: a multitask generative pre-trained language model for single-cell analysis. *bioRxivorg* (2024) doi:10.1101/2024.01.25.577152.

18. Lu, Y.-C. *et al.* ScChat: A large language model-powered co-pilot for contextualized single-cell RNA sequencing analysis. *bioRxiv* (2024) doi:10.1101/2024.10.01.616063.

19. Kalfon, J., Samaran, J., Peyré, G. & Cantini, L. scPRINT: pre-training on 50 million cells allows robust gene network predictions. *Nat. Commun.* **16**, 3607 (2025).

20. Wen, H. *et al.* CellPLM: Pre-training of Cell Language Model Beyond Single Cells. *bioRxiv* 2023.10.03.560734 (2023) doi:10.1101/2023.10.03.560734.

21. Rosen, Y. *et al.* Universal cell embeddings: A foundation model for cell biology. *bioRxiv* 2023.11.28.568918 (2023) doi:10.1101/2023.11.28.568918.

22. Shen, H. *et al.* Generative pretraining from large-scale transcriptomes for single-cell deciphering.

iScience **26**, 106536 (2023).

23. Reyes, M. *et al.* An immune-cell signature of bacterial sepsis. *Nat. Med.* **26**, 333–340 (2020).
24. Reyes, M. *et al.* Plasma from patients with bacterial sepsis or severe COVID-19 induces suppressive myeloid cell production from hematopoietic progenitors in vitro. *Sci. Transl. Med.* **13**, eabe9599 (2021).
25. Hajiramezanali, E. *et al.* Score-based Explainability for Graph Representations. *Transactions on Machine Learning Research* (2024).
26. Adams, T. S. *et al.* Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Science Advances* **6**, eaba1983 (2020).
27. Yokoyama, K. *et al.* BANK regulates BCR-induced calcium mobilization by promoting tyrosine phosphorylation of IP(3) receptor. *EMBO J.* **21**, 83–92 (2002).
28. Patterson, H. C. *et al.* Cytoplasmic Ig alpha serine/threonines fine-tune Ig alpha tyrosine phosphorylation and limit bone marrow plasma cell formation. *J. Immunol.* **187**, 2853–2858 (2011).
29. Tedder, T. F., Klejman, G., Schlossman, S. F. & Saito, H. Structure of the gene encoding the human B lymphocyte differentiation antigen CD20 (B1). *The Journal of Immunology* **142**, 2560–2568 (1989).
30. Lovering, R. C. *et al.* A GO catalogue of human DNA-binding transcription factors. *Biochim. Biophys. Acta Gene Regul. Mech.* **1864**, 194765 (2021).
31. Cano-Gamez, E. *et al.* Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4+ T cells to cytokines. *Nat. Commun.* **11**, 1801 (2020).
32. Kotliar, D. *et al.* Identifying gene expression programs of cell-type identity and cellular activity with single-cell RNA-Seq. *Elife* **8**, e43803 (2019).
33. Deng, Q. *et al.* Characteristics of anti-CD19 CAR T cell infusion products associated with efficacy and toxicity in patients with large B cell lymphomas. *Nat. Med.* **26**, 1878–1887 (2020).
34. Fontenot, J. D., Gavin, M. A. & Rudensky, A. Y. Foxp3 programs the development and function of

CD4+CD25+ regulatory T cells. *Nat. Immunol.* **4**, 330–336 (2003).

35. Chinen, T. *et al.* An essential role for the IL-2 receptor in Treg cell function. *Nat. Immunol.* **17**, 1322–1333 (2016).
36. Andreatta, M. & Carmona, S. J. UCell: Robust and scalable single-cell gene signature scoring. *Comput. Struct. Biotechnol. J.* **19**, 3796–3798 (2021).
37. Noureen, N., Ye, Z., Chen, Y., Wang, X. & Zheng, S. Signature-scoring methods developed for bulk samples are not adequate for cancer single-cell RNA sequencing data. *Elife* **11**, e71994 (2022).
38. van der Wijst, M. G. P. *et al.* Type I interferon autoantibodies are associated with systemic immune alterations in patients with COVID-19. *Sci. Transl. Med.* **13**, eabh2624 (2021).
39. Hu, C. *et al.* CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Res.* **51**, D870–D876 (2023).
40. Newman, A. M. *et al.* Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
41. Otsu, N. A Threshold Selection Method from Gray-Level Histograms. *IEEE Trans. Syst. Man Cybern.* **9**, 62–66 (1979).
42. Heimberg, G. *et al.* Scalable querying of human cell atlases via a foundational model reveals commonalities across fibrosis-associated macrophages. (2023) doi:10.1101/2023.07.18.549537.
43. Veglia, F., Sanseviero, E. & Gabrilovich, D. I. Myeloid-derived suppressor cells in the era of increasing myeloid cell diversity. *Nat. Rev. Immunol.* **21**, 485–498 (2021).
44. Stephenson, E. *et al.* Single-cell multi-omics analysis of the immune response in COVID-19. *Nat. Med.* **27**, 904–916 (2021).
45. Newburger, J. W., Takahashi, M. & Burns, J. C. Kawasaki Disease. *J. Am. Coll. Cardiol.* **67**, 1738–1749 (2016).
46. La Rosée, P. *et al.* Recommendations for the management of hemophagocytic lymphohistiocytosis in adults. *Blood* **133**, 2465–2477 (2019).

47. Takahashi T. Severe fever with thrombocytopenia syndrome (SFTS) and SFTS virus. *Uirusu* **65**, 7–16 (2015).

48. Park, A. *et al.* Molecular Signatures of Inflammatory Profile and B-Cell Function in Patients with Severe Fever with Thrombocytopenia Syndrome. *MBio* **12**, (2021).

49. Liu, P. *et al.* Nivolumab treatment of relapsed/refractory Epstein-Barr virus–associated hemophagocytic lymphohistiocytosis in adults. *Blood* **135**, 826–833 (2020).

50. Wang, Z. *et al.* Single-cell RNA sequencing of peripheral blood mononuclear cells from acute Kawasaki disease patients. *Nat. Commun.* **12**, 5444 (2021).

51. Lo, M. S. & Newburger, J. W. Role of intravenous immunoglobulin in the treatment of Kawasaki disease. *Int. J. Rheum. Dis.* **21**, 64–69 (2018).

52. Charytonowicz, D., Brody, R. & Sebra, R. Interpretable and context-free deconvolution of multi-scale whole transcriptomic data with UniCell deconvolve. *Nat. Commun.* **14**, 1350 (2023).

53. Liao, K. *et al.* X-scPAE: An explainable deep learning model for embryonic lineage allocation prediction based on single-cell transcriptomics revealing key genes in embryonic cell development. *Comput. Biol. Med.* **188**, 109787 (2025).

54. Chen, J. *et al.* Deep transfer learning of cancer drug responses by integrating bulk and single-cell RNA-seq data. *Nat. Commun.* **13**, 6494 (2022).

55. Levine, D. *et al.* Cell2Sentence: Teaching large language models the language of biology. *bioRxivorg* (2024) doi:10.1101/2023.09.11.557287.

56. Schaar, A. C. *et al.* Nicheformer: a foundation model for single-cell and spatial omics. *Bioinformatics* (2024).

57. Cui, H. *et al.* Towards multimodal foundation models in molecular cell biology. *Nature* **640**, 623–633 (2025).

58. Mao, W., Zaslavsky, E., Hartmann, B. M., Sealfon, S. C. & Chikina, M. Pathway-level information extractor (PLIER) for gene expression data. *Nat. Methods* **16**, 607–610 (2019).

59. Kunes, R. Z., Walle, T., Land, M., Nawy, T. & Pe'er, D. Supervised discovery of interpretable gene programs from single-cell data. *Nat. Biotechnol.* **42**, 1084–1095 (2024).

60. Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. *arXiv [cs.CV]* (2013).

61. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. *arXiv* (2017) doi:10.48550/arxiv.1703.01365.

62. Lundberg, S. & Lee, S.-I. A unified approach to interpreting model predictions. *arXiv [cs.AI]* (2017).

63. Ribeiro, M. T., Singh, S. & Guestrin, C. Why should I trust you? in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, USA, 2016). doi:10.1145/2939672.2939778.

64. Shrikumar, A., Greenside, P. & Kundaje, A. Learning Important Features Through Propagating Activation Differences. *arXiv* (2017) doi:10.48550/arxiv.1704.02685.

65. Kokhlikyan, N. *et al.* Captum: A unified and generic model interpretability library for PyTorch. *arXiv [cs.LG]* (2020).

66. Waskom, M. seaborn: statistical data visualization. *Journal of Open Source Software* **6**, 3021 (2021).

67. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

68. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

69. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with python. in *Proceedings of the Python in Science Conference* 92–96 (SciPy, 2010).

70. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

71. Travaglini, K. J. *et al.* A molecular cell atlas of the human lung from single cell RNA sequencing. *Nature* **587**, 619–625 (2020).

72. Ogata, H. *et al.* KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **27**, 29–34

(1999).

73. Rouillard, A. D. *et al.* The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database* **2016**, (2016).

74. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *arXiv* (2012) doi:10.48550/arxiv.1201.0490.

75. Collin, M. & Bigley, V. Human dendritic cell subsets: an update. *Immunology* **154**, 3–20 (2018).

76. Riba, A. *et al.* Cell cycle gene regulation dynamics revealed by RNA velocity and deep-learning. *Nat. Commun.* **13**, 2865 (2022).

77. Walt, S. van der *et al.* scikit-image: Image processing in Python. *PeerJ PrePrints* **2**, e336v1 (2014).

78. Vieira Braga, F. A. *et al.* A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat. Med.* **25**, 1153–1163 (2019).

79. Muto, Y. *et al.* Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney. *Nat. Commun.* **12**, 2190 (2021).

80. Wilson, P. C. *et al.* Multimodal single cell sequencing implicates chromatin accessibility and genetic background in diabetic kidney disease progression. *Nat. Commun.* **13**, 5253 (2022).

81. HCA seed network precise tumor-nephrectomy samples. https://explore.data.humancellatlas.org/projects/29ed827b-c539-4f4c-bb6b-ce8f9173dfb7.