scSemiProfiler: Advancing Large-scale Single-cell Studies through Semi-profiling with Deep Generative Models and Active Learning

Jingtao Wang[1,2], Gregory Fonseca[1,2,4], Jun Ding[1,2,3,4,5*]

[1]Meakins-Christe Laboratories, Research Institute of McGill University Health Centre, 1001 Decarie Blvd, Montreal, H4A 3J1, Quebec, Canada.
[2]Department of Medicine, Division of Experimental Medicine, McGill University, 1001 Decarie Blvd, Montreal, H4A 3J1, Quebec, Canada.
[3]School of Computer Science, McGill University, 3480 Rue University, Montreal, H3A 2A7, Quebec, Canada.
[4]Quantitative Life Sciences, McGill University, 845 Rue Sherbrooke Ouest, Montreal, H3A 0G4, Quebec, Canada.
[5]Mila-Quebec AI Institute, 6666 Rue Saint-Urbain, Montreal, H2S 3H1, Quebec, Canada.

*Corresponding author(s). E-mail(s): jun.ding@mcgill.ca;
Contributing authors: jingtao.wang@mail.mcgill.ca; gregory.fonseca@mcgill.ca;

Abstract

Single-cell sequencing is a crucial tool for dissecting the cellular intricacies of complex diseases. Its prohibitive cost, however, hampers its application in expansive biomedical studies. Traditional cellular deconvolution approaches can infer cell type proportions from more affordable bulk sequencing data, yet they fall short in providing the detailed resolution required for single-cell-level analyses. To overcome this challenge, we introduce "scSemiProfiler", an innovative computational framework that marries deep generative model with active learning strategies. This method adeptly infers single-cell profiles across large cohorts by fusing bulk sequencing data with targeted single-cell sequencing from a few carefully chosen representatives. Extensive validation across heterogeneous datasets verifies the precision of our semi-profiling approach, aligning closely with true single-cell profiling data and empowering refined cellular analyses. Originally developed for extensive disease cohorts, "scSemiProfiler" is adaptable for broad applications. It provides a scalable, cost-effective solution for single-cell profiling, facilitating in-depth cellular investigation in various biological domains.

Keywords: Deep Generative Learning, Active Learning, Bulk Deconvolution, Single-cell sequencing, Bulk sequencing, Cohort studies, scSemiProfiler

Introduction

The advent of single-cell sequencing has dramatically reshaped the landscape of biological research, providing an unparalleled view into the cellular complexities of organisms [1, 2]. This technique has unearthed the subtle distinctions among individual cells, enabling a richer understanding of cellular dynamics [3, 4]. High-resolution data from such analyses are essential for delineating and characterizing the myriad of cellular subpopulations within patient samples, leading to transformative developments in biomarker discovery and personalized therapy strategies [5–7]. Cohort studies,

which offer longitudinal insights by observing specific groups over time, are particularly poised to benefit from these advances [8–10]. However, the substantial financial cost associated with single-cell sequencing—such as the estimated $6,000 required to sequence 20,000 cells (costpercell)—can often be a limiting factor for extensive research endeavors.

A range of computational strategies, particularly deconvolution methods, are available for dissecting bulk data [11] into distinct cell populations. This approach enables a harmonious balance between affordability and data resolution. Prominent among these are deconvolution techniques such as CIBERSORTx [12] and Bisque [13], which have become particularly popular. These methods estimate the proportions of different cell types within bulk sequencing samples by utilizing signature profiles from single-cell reference datasets. Conventional bulk decomposition methods, while valuable for enhancing cohort study analyses, exhibit limitations in their resolution and accuracy. Capable of dissecting bulk samples into different cell types and ascertaining their gene expression patterns, they offer beneficial insights at the cell-type level. Yet, these methods fall short in delivering true single-cell resolution. Another significant challenge they face is the precise decomposition of cell types and accurate inference of gene expression. Additionally, these methods often overlook the considerable variability that exists within individual cell types. This variability is a critical component for deciphering the complexities of disease dynamics and their response to treatments [14–18]. Those limitations of conventional bulk decomposition methods may impede the thorough exploration of in-depth single-cell level analyses. These analyses are essential for understanding cellular heterogeneity and dynamics in datasets, including but not limited to cell pseudotime trajectory analysis [19–22], as well as various advanced machine learning techniques [23–27].

In response to the challenges previously highlighted and with the goal of offering a cost-effective approach for extensive single-cell sequencing, we introduce the single-cell Semi-profiler (*scSemiProfiler*). This innovative computational tool is crafted to significantly improve the precision and depth of single-cell analysis. It stands out as a more economical and scalable option for single-cell sequencing, thereby facilitating advanced single-cell analysis with greater accessibility. This tool effectively integrates active learning techniques [28] with deep generative neural network algorithms [29], aiming to provide single-cell resolution data at a more affordable price. *scSemiProfiler* aims to simultaneously achieve two fundamental goals in the semi-profiling process. On one hand, *scSemiProfiler*'s active learning module integrates information from the deep learning model and bulk data, intelligently selecting the most informative samples for actual single-cell sequencing. On the other hand, *scSemiProfiler*'s deep generative model component [30–33] effectively merges single-cell data from representative samples with the bulk sequencing data of the cohort. This process computationally infers the single-cell data for the remaining non-representative samples. This advanced approach leads to more detailed "deconvolution" of the target bulk data into precise single-cell level measurements. Consequently, with only the budget for bulk sequencing and single-cell sequencing of representatives, *scSemiProfiler* outputs single-cell data for all samples in the study. To the best of our knowledge, the *scSemiProfiler* is the first of its kind designed for such intricate single-cell level computational decomposition from bulk sequencing data.

Through comprehensive evaluations across a variety of datasets, *scSemiProfiler* has consistently delivered high-quality semi-profiled single-cell data, which not only accurately represents actual single-cell datasets but also mirrors the results of downstream tasks with remarkable fidelity. Therefore, *scSemiProfiler* has established itself as a pivotal tool in single-cell research. This innovative approach is poised to revolutionize access to single-cell data in large-scale studies, such as disease cohort investigations and beyond. By making large-scale single-cell studies more affordable, *scSemiProfiler* promises to catalyze the application of single-cell technologies in a wide array of expansive biomedical research. This advancement is set to expand the scope and enhance the depth of biological research globally.

Results

Method overview

The *scSemiProfiler* approach represents a novel method for decomposing broad-scope bulk sequencing data into detailed single-cell cohorts. It achieves this by conducting single-cell profiling on only a select few representative samples and then computationally inferring single-cell data for the remainder. This approach substantially lowers the costs associated with large-scale single-cell studies. As depicted in Fig. 1, our method is designed to deliver cost-effective, semi-profiled single-cell sequencing data,

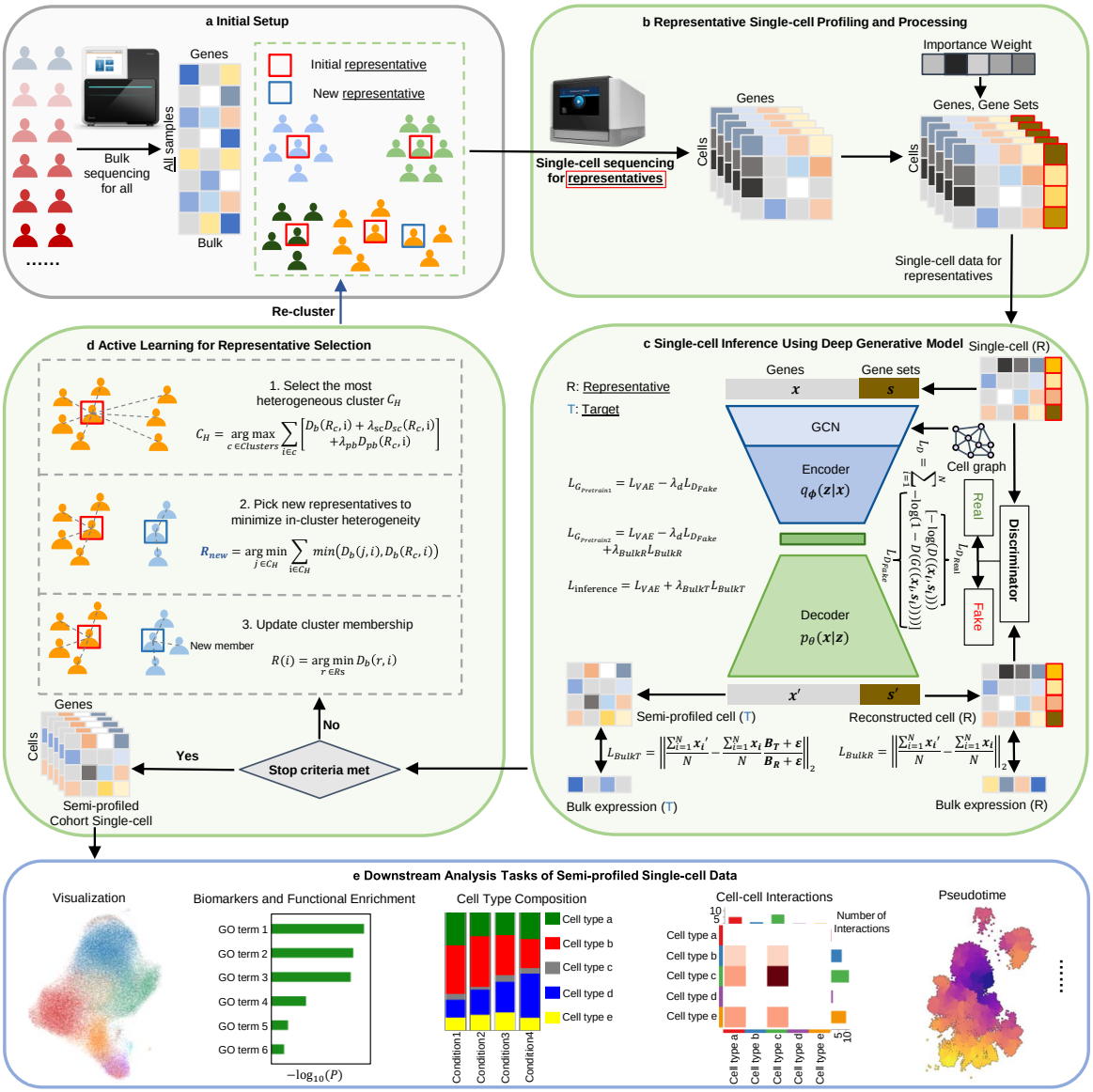

Fig. 1 Overview of the *scSemiProfiler* method. a, Initial Configuration: Bulk sequencing is first conducted on the entire cohort, followed by clustering analysis of this data. This analysis identifies representative samples, typically those closest to the cluster centroids. **b**, Representative Profiling: The identified representatives are then subjected to single-cell sequencing. The data obtained from this sequencing is further processed to determine gene set scores and feature importance weights, enriching the subsequent analysis steps. **c**, Deep Generative Inference: Utilizing a VAE-GAN-based model, the process integrates comprehensive bulk data from the cohort with the single-cell data derived from the representatives. During the model's 3-stage training, the generator aims to optimize losses $L_{G_{Pretrain1}}$, $L_{G_{Pretrain2}}$, and $L_{inference}$, respectively, whereas the discriminator focuses on minimizing the discriminator loss L_D. In L_D, G and D are the generator and discriminator respectively. The term $D((\mathbf{x_i}, \mathbf{s_i}))$ represents the discriminator's predicted probability that a given input cell is real, under the condition that it is indeed a real cell. Conversely, $D(G((\mathbf{x_i}, \mathbf{s_i})))$ denotes the discriminator's predicted probability that the input cell is real, when in fact it is a cell reconstructed by the generator. **d**, Representative Selection Decision: Decisions on further representative selection are made, taking into account budget constraints and the effectiveness of the current representatives. An active learning algorithm, which draws on insights from the bulk data and the generative models, is employed to pinpoint additional optimal representatives. These newly selected representatives then undergo further single-cell sequencing (**b**) and serve as new reference points for the ongoing *in silico* inference process (**c**). **e**, Comprehensive Downstream Analyses: The final panel shows extensive downstream analyses enabled by the semi-profiled single-cell data. This is pivotal in demonstrating the model's capacity to provide deep and wide-ranging insights, showcasing the full potential and applicability of the semi-profiled single-cell data.

enabling deep exploration of cellular dynamics in large cohorts. In this context, "semi-profiling" refers to the generation of single-cell data for an entire cohort, achieved either through direct single-cell sequencing of selected representative samples or via *in silico* inference using a deep generative model. This *in silico* inference process combines actual single-cell sequencing data from a representative

sample with bulk sequencing data, encompassing both the target and the representative sample. Thus, a "semi-profiled cohort" includes real single-cell data for representative samples and inferred data for the non-representative ones. This innovative approach facilitates a thorough examination of individual cellular profiles within a larger dataset, seamlessly linking the extensive scope of bulk sequencing with the granularity of single-cell analysis.

Initially, the semi-profiling pipeline commences with bulk sequencing of each cohort member (Fig. 1a), laying the foundational data layer for all subsequent analyses. Following this foundational step, the methodology employs a clustering analysis to form B (sample batch size) sample clusters utilizing the extensive data derived from the initial bulk sequencing and selecting a "representative" sample for each cluster (Fig. 1b). Single-cell profiling will be conducted on the selected representative samples in preparation for the following steps. The core of the *scSemiProfiler* involves an innovative deep generative learning model (Fig. 1c). This model is engineered to intricately meld actual single-cell data profiles with the gathered bulk sequencing data, thereby capturing complex biological patterns and nuances. Specifically, it uses a VAE-GAN [34] architecture initially pretrained on single-cell sequencing data of selected representatives for self-reconstruction. Subsequently, the VAE-GAN is further pretrained with a representative reconstruction bulk loss, aligning pseudobulk estimations from the reconstructed single-cell data with real pseudobulk. Finally, the model undergoes fine-tuning with another target bulk loss tied to the real bulk sequencing data of the target sample, facilitating precise *in silico* inference of the target's single-cell profile. Once the *in silico* single-cell inference is finished for all non-representative samples in the cohort, an active learning module can be used for selecting the next batch of potentially most informative representatives for single-cell sequencing to further improve the semi-profiling performance (Fig. 1d). When studying a smaller dataset, or when more cells per sample are required, a smaller batch size, such as 2, may be preferred. However, a batch size of 4 is set as default to maximize the usage of a 10x genomics single-cell toolkit, which can typically capture up to 20000 cells (4 samples if assuming 5,000 cells each). This dynamic, iterative process is continuously augmented with newly acquired single-cell data, ensuring that the most informative samples are selected for real single-cell profiling, leading to more accurate *in silico* single-cell inference for non-representative samples. This iterative process concludes when the budgetary constraint is met or when a sufficient number of representatives have been chosen to ensure satisfactory semi-profiling performance.

When any of the stop criteria is met, the semi-profiled single-cell data can be used for a broad spectrum of downstream single-cell analyses (Fig. 1e), such as cell feature visualization, biomarker, and function enrichment analysis, tracking cell type compositions in various tissues/conditions, cell-cell interaction analysis, and pseudotime analysis. Ultimately, *scSemiProfiler* offers a holistic sequencing perspective, delivering nuanced single-cell insights from bulk sequencing data. The method exhibits acceptable performance in terms of runtime and memory usage (refer to Supplementary Fig. S1), making it suitable for scaling to large-scale datasets.

The semi-profiled COVID-19 single-cell cohort exhibits significant similarity to its real counterpart

To test the performance of *scSemiProfiler* in generating semi-profiled single-cell data that resonates with the granularity and details of actual single-cell sequencing, we utilized a COVID-19 cohort single-cell sequencing dataset [35]. After quality controls, this dataset includes 124 samples, including healthy controls and infected patients of different severity levels: asymptomatic, mild, moderate, severe, or critical. Here, we produced pseudobulk data by taking the average of the normalized count single-cell data. We then tested *scSemiProfiler*'s ability to regenerate the single-cell cohort from the pseudo-bulk data and real-profiled single-cell representatives using semi-profiling, as the actual bulk data for those samples in the COVID-19 cohort is absent. In generating our semi-profiled dataset, 28 representatives were selected in batches of 4 using our active learning algorithm for real single-cell profiling. Deep generative models then inferred the single-cell profiles for the remaining samples based on the representatives' real-profiled single-cell data and the bulk data of all samples. The estimated total cost of these bulk and single-cell sequencing is \$62,640. This is only 33.7% of the estimated price, \$186,000, for actually conducting single-cell sequencing for the entire cohort. Additionally, our approach offers the advantage of generating extra bulk data for the cohort, a benefit not provided by single-cell sequencing of the entire cohort. The price of bulk sequencing is estimated based on the cost at McGill Genome Centre in the year 2023. This estimation assumes the use of one NovaSeq 6000 S2 system (capable of sequencing up to 4 billion reads per run) at an approximate cost of \$7,000,

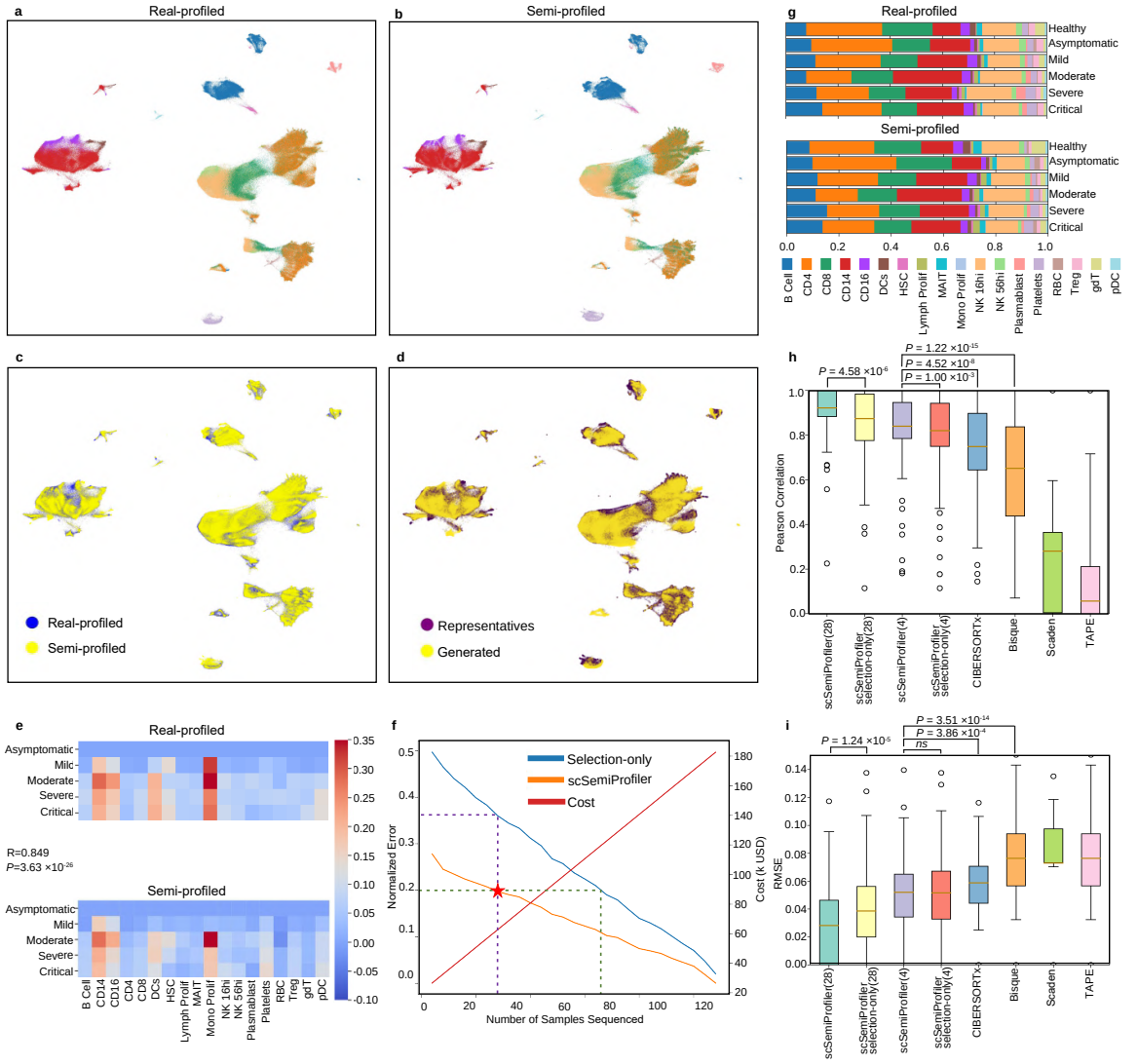

Fig. 2 Overall comparisons of the semi-profiled and real-profiled COVID-19 dataset. a, UMAP visualization of the real-profiled data. Colors correspond to cell types and are consistent with (**g**). **b**, UMAP visualization of the semi-profiled data. **c**, UMAP visualization of semi-profiled data and real-profiled data together. The color differentiation signifies whether cells originate from the semi-profiled or the real-profiled dataset. Areas of overlap between the two indicate where the semi-profiled data closely resembles the real-profiled data. **d**, UMAP visualization of the semi-profiled cohort, displaying different colors to distinguish cells produced by a deep generative model (labeled as "Generated") from the representative cells obtained through real-profiling (labeled as "Representatives"). **e**, Visualization illustrates the relative activation patterns of the interferon pathway. The comparison of these values between the semi-profiled and real-profiled matrices yields a Pearson correlation coefficient of 0.849 and a p-value of 3.63×10^{-26}. **f**, Graph depicting the normalized error in semi-profiled data with an increasing number of representatives. The terms 'scSemiProfiler' and 'Selection-only' represent our semi-profiling method and a method that only selects representatives using an active learning algorithm, respectively. It is important to note that actual costs may vary based on the sequencing technology and specific cells sequenced. **g**, Stacked bar plot illustrating the proportions of cell types across various disease conditions. The upper portion represents the real-profiled data, while the lower portion depicts the semi-profiled data. Pearson correlation coefficients comparing cell type proportions between the real-profiled and semi-profiled datasets are provided for different conditions: Healthy (0.987), Asymptomatic (0.970), Mild (0.996), Moderate (0.992), Severe (0.978), and Critical (0.989), indicating a high degree of similarity between the two datasets across these conditions. **h-i**, Cell type deconvolution benchmarking. **h**, Figure displaying Pearson correlation coefficients between actual (ground truth) cell type proportions and those estimated by various deconvolution methods. Except for the first two columns, all other columns' results are based on 4 representatives' single-cell data as reference. **i**, Comparison of Root Mean Square Error (RMSE) across various deconvolution methods.

plus an additional \$110 for library preparation per bulk sample. The cost for single-cell sequencing is based on the tool (costpercell). This tool provides a cost estimate for capturing 0.8 billion total reads from 20,000 cells across four samples in one 10x lane, equating to \$0.3 per cell. Consequently, the estimated cost for each sample (5,000 cells) is \$1,500.

Using UMAP [36] visualizations, we show a significant alignment between the semi-profiled and the real-profiled in Fig. 2a and b. We also annotated the semi-profiled COVID-19 cohort using an unbiased approach. We trained a Multi-layer Perceptron (MLP) classifier using the annotated representatives' cells and used it to predict the cell types of the rest of the non-representative samples' cells generated by the deep learning model. As shown in Fig. 2b, cell clustering remains intact in the semi-profiled cohort, from the distinctive clusters of B cells, plasmablasts, and platelets to the nuanced similarities of the CD14 and CD16 cells. The semi-profiled dataset illustrates its remarkable fidelity to the real-profiled version. This fidelity extends to capturing even the subtle batch effects, as observed in the twin CD4 clusters, further accentuating *scSemiProfiler*'s robust *in silico* inference capabilities. The UMAP plots in Fig. 2c weave together directly sequenced samples with the semi-profiled ones, showcasing the tool's finesse. The overlapping data points found in both sequencing techniques resonate with the transformative nature of *scSemiProfiler*—it harmonizes accuracy and cost-efficiency seamlessly. This prompts an inquiry: Does this alignment owe its credit to the active learning mechanism that identifies the most informative representatives, or does it also hail from the deep generative model's prowess in inferring target samples' cells with finesse? To answer this, Fig.2d uses distinct colors to delineate real-profiled cells from representatives and generated cells for target samples. It shows that our representative selection strategy is able to select representatives such that their cells have a relatively good coverage of the overall cell distribution. Meanwhile, the deep learning model managed to generate cells to complement the rest and make the overall semi-profiled cohort almost identical to the original real-profiled cohort. The effectiveness of the deep learning model in the semi-profiling process is further demonstrated in Supplementary Fig. S2. This figure illustrates the model's capability in reconstructing data from single representative samples and its proficiency in inferring data for individual target samples.

To test the fidelity of single-cell gene expression semi-profiling, we tested the interferon (IFN) pathway gene set—crucial for the innate immune response against COVID-19 [37–40]. Through the prism of our semi-profiled dataset, Fig. 2e reveals IFN activation patterns that harmonize with the real-profiled dataset. The uniformity in IFN activation patterns across various key cell types and severity levels, as highlighted by similar heatmaps, confirms the effectiveness of the semi-profiling technique. This uniformity indicates that the critical disease-related pathways were effectively captured and maintained in the semi-profiled data.

Further, we explored the quantitative metrics of efficacy and cost-effectiveness in semi-profiling, Fig. 2f. Our analysis centered on understanding the relationship between the number of representative samples used for single-cell sequencing and the associated semi-profiling error. While an increase in representative samples intuitively raises the costs of single-cell sequencing, *scSemiProfiler* effectively leverages the single-cell data of these representatives for accurate inference of target samples. As more representatives are selected, our semi-profiling method effectively reduced the normalized error. We also compared our method to a selection-only method, which uses the same representatives chosen by *scSemiProfiler* using the active learning algorithm. For each target sample, it performs the single-cell data inference in a naive manner: merely copying the corresponding representative's data. The dashed line shows that our semi-profiling method has a huge lead over the selection-only method. The star symbol denotes the number of representatives selected for the specific semi-profiled cohort that was utilized for subsequent analyses, along with the associated error. The vertical dashed line underscores that by using the same representatives, our method achieves substantially lower errors compared to the selection-only method. Moreover, the horizontal dashed line demonstrates that the selection-only method requires a considerably larger pool of representatives to attain the same level of semi-profiling accuracy as our method. The comparative analysis between *scSemiProfiler* and the selection-only method highlights the deep learning model's efficacy in reducing costs and minimizing errors.

In further evaluating the effectiveness of semi-profiling, defined as a single-cell granularity bulk decomposition, we turned to the conventional realm of cell type proportion metrics, as anchored by Fig. 2g. Existing research [35] elucidates that PBMC cell type proportions undergo dynamic shifts with evolving disease conditions. True to this, our real-profiled dataset indicates a pronounced expansion of B cells and CD14 cells under aggravated conditions—a pattern mirrored in the semi-profiled dataset. The Pearson correlation coefficient [41] of cell type composition associated with different disease conditions between the semi-profiled and the real single-cell datasets consistently surpasses 0.9.

In our comprehensive analysis of cell deconvolution methods, we meticulously compared *scSemiProfiler* with several leading-edge techniques, including CIBERSORTx [12], Bisque [13], Scaden [42], TAPE [43], and DWLS [44], as depicted in Fig. 2h and i. Each method was tested under identical conditions, using the same bulk data and single-cell reference data. Notably, DWLS was not included in our final results due to its failure to yield results within a week. A key challenge in this analysis was the memory constraints encountered by most methods, elaborated in Supplementary Fig. S1b, which hindered their ability to process the full set of 28 representative single-cell data. To maintain a level playing field, we limited the single-cell reference to an initial batch of 4 representative samples for all methods. Our results unequivocally demonstrate the superior deconvolution performance of *scSemiProfiler* over all benchmarked methods. Its effectiveness is not only apparent when utilizing a smaller reference set of 4 samples but also becomes increasingly pronounced with a larger set of 28 samples. This distinct superiority of *scSemiProfiler* is a testament to its remarkable efficiency and versatility. Capable of excelling with both compact and extensive single-cell reference datasets, *scSemiProfiler* stands out as the most adept tool in cell deconvolution, surpassing its contemporaries in handling diverse data scales with unparalleled precision and reliability. In Supplementary Fig. S3, we also show the high deconvolution accuracy using a side-by-side comparison between each sample's predicted cell type proportion using 28 representatives with the ground truth.

The semi-profiled COVID-19 single-cell cohort proves reliable for single-cell downstream analyses

We have previously demonstrated the capability of *scSemiProfiler* in accurately generating semi-profiled single-cell data that closely aligns with its real-profiled counterpart. Moving beyond basic cell type proportion predictions, which is the primary focus of other methods, *scSemiProfiler* excels in predicting gene expression for each cell within a population, thereby more authentically mimicking true single-cell data. This advancement is crucial for more complex downstream single-cell analysis tasks.

To illustrate the effectiveness of semi-profiled data in standard downstream single-cell analyses, we conducted a series of evaluations. A key task in these analyses is the identification of biomarkers within distinct cell clusters, highlighting genes with elevated expression levels. Utilizing the semi-profiled data generated by *scSemiProfiler*, we performed various single-cell level downstream analyses. The results from these analyses, as depicted in Fig. 3, demonstrate a remarkable consistency with outcomes derived from real-profiled data. For instance, we identified top cell type signature genes using the real-profiled cohort. When comparing their expression patterns in both real-profiled and semi-profiled datasets, the similarities were striking. The dot plots in Fig. 3a display these patterns, showcasing an almost indiscernible difference between the datasets. The semi-profiling at the single-cell level provides high-resolution expression data for marker genes within each cell population (cell type). Our approach reveals the distribution of marker gene expression across all cells within a specific type. While existing bulk deconvolution methods can offer average gene expression data for cell type markers, they fall short in depicting detailed gene expression distribution and variations among individual cells within the same population. This high level of cell type biomarker concordance underscores the *scSemiProfiler*'s robustness not only in replicating single-cell data but also in ensuring the fidelity of downstream analytical processes.

We further explored the similarities between biomarkers identified using semi-profiled data and those from real-profiled data. Biomarker discovery is possible with lower-resolution data, such as the average cell type gene expression data provided by current bulk deconvolution tools. However, these methods are less reliable compared to single-cell data. The limitation of average cell type gene expression data is its lack of replicates for each cell type, leading to reduced statistical power. The absence of replicates makes it challenging to estimate variance within a cell type, which is essential for standard differential expression tests. In contrast, our semi-profiled single-cell data supports robust biomarker discovery through rigorous statistical testing, as it includes multiple samples (i.e. cells) for each cell type. We demonstrate the similarity of biomarkers identified using real-profiled and semi-profiled datasets through a rank–rank hypergeometric overlap (RRHO) plot. An RRHO plot [45] visualizes the overlap between two ranked gene lists, highlighting the degree of similarity and the significance of the overlap between them (see the Methods section for more details). Leveraging the RRHO plots, we compared the top 50 positive and top 50 negative gene lists associated with different cell types (Fig. 3b for CD4 cells and Supplementary Fig. S4 for all other cell types. The plots show positive marker and negative marker lists from both datasets are highly similar. A

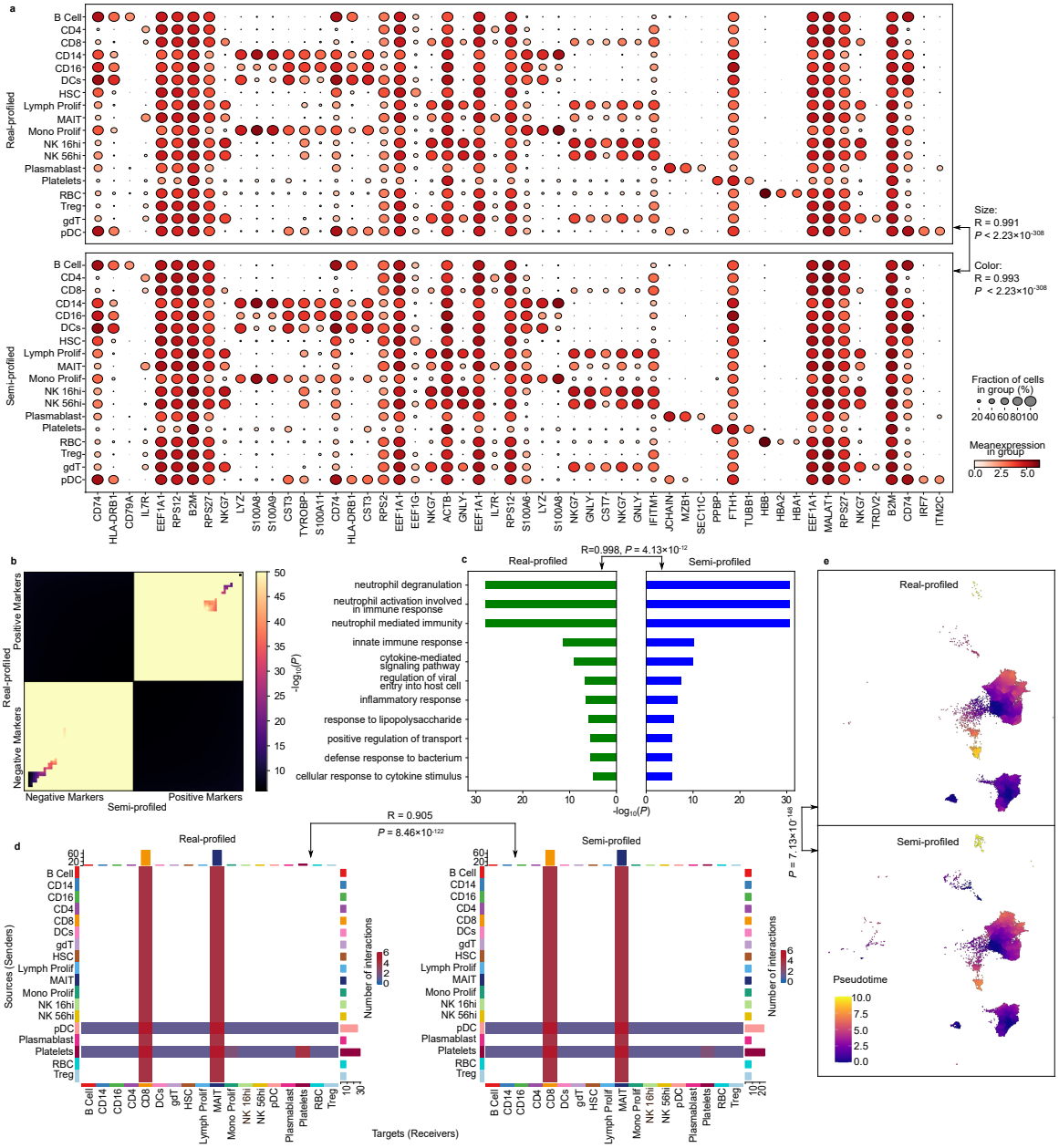

Fig. 3 Comparative analyses of single-cell level downstream analysis tasks using real-profiled and semi-profiled COVID-19 datasets. a, Dot plots elucidating the expression proportion and intensity of discerned cell type signature genes. The top half showcases the real-profiled dataset, while the bottom delineates the semi-profiled version. **b**, RRHO plot emphasizing the congruence between the CD4 positive and negative markers in both datasets. **c**, Visualization of the GO term enrichment outcomes rooted in CD4 signature genes from both dataset versions. The plot accentuates the union of the top 10 enriched terms, with the Pearson correlation coefficients of between the bar lengths, which is based on the corresponding p-value and therefore represents the significant level. **d**, A juxtaposition of cell-cell interaction analyses stemming from real-profiled and semi-profiled cells from moderate COVID-19 patients, underscoring the similarity in interaction types and counts. **e**, Comparative depiction of pseudotime trajectories for CD4 cells across both datasets, highlighting their striking similarity in reconstructing dynamic cellular processes.

marked dissimilarity was evident between the positive and negative marker lists, which is intuitively anticipated. By definition, positive markers are genes that are higher expressed in the corresponding cell types, and negative markers are the opposite - lower expressed. Therefore, they should have no overlap. The compelling concordance demonstrated in Fig. 3a and b bolsters our claim that the semi-profiled data can viably supplant real-profiled data for the pivotal task of biomarker discovery using *scSemiProfiler*.

Next, we used the biomarkers derived from the semi-profiled dataset and those from the real-profiled dataset for gene functional enrichment analysis, assessing whether the two versions of the

analysis yield consistent results. Fig. 3c compares the Gene Ontology (GO) [46, 47] enrichment [48, 49] outcomes derived from real-profiled and semi-profiled datasets. The top 100 signature genes from both datasets are used for the enrichment analysis. We observed an overlap of 95 genes between the two lists, yielding a highly significant hypergeometric test p-value of 4.00×10^{-200} (the population size of the hypergeometric test is the number of highly variable genes used for this dataset, 6030). A comparison of the top 10 overlapping terms from both versions reveals nearly identical significance (Pearson correlation coefficient of 0.998 with a p-value of 4.13×10^{-12} for comparing the significant levels). The results for other cell types are in Supplementary Fig. S5. Reactome pathway [50] enrichment analysis results are in Supplementary Fig. S6. This further corroborates the reliability of semi-profiled data in downstream analyses.

Progressing to yet another pivotal single-cell level downstream analysis task, we evaluated the congruence in cell-cell interaction analyses derived from real-profiled and semi-profiled datasets. Given the paramount role of cell-cell interactions in orchestrating a myriad of multicellular processes, their analysis often unveils pivotal biological insights [51, 52]. Fig.3d juxtaposes cell-cell interaction analyses rooted in real-profiled and semi-profiled cells from moderate COVID-19 patients (see results for other severity levels in Supplementary Fig. S7). The evident concordance in types and counts of interactions in both renditions reinforces the reliability of our semi-profiled data ($R = 0.905, P = 8.46 \times 10^{-122}$). We also show a comparison of partition-based graph abstraction (PAGA) plots generated using real-profiled cohort and semi-profiled cohort in Supplementary Fig. S8, which demonstrates that the semi-profiled data can accurately capture the cellular trajectories and relationships between cell types. Given that such analyses intrinsically require single-cell data, *scSemiProfiler* emerges as the sole contender capable of producing data apt for this task from bulk sources.

Delving further into the capacity of semi-profiled data for other downstream single-cell level analysis tasks, we turned our attention to pseudotime analysis [19, 20]. Pseudotime is a pivotal tool in reconstructing dynamic cellular processes, ranging from differentiation pathways to developmental timelines or disease trajectories. As depicted in Fig. 3e, the pseudotime trajectories derived from real-profiled and semi-profiled CD4 cells are strikingly similar (Consistent results for the pseudotime analysis of other cell types can be found in Supplementary Fig. S9). Such compelling evidence underscores that the semi-profiled data retains its reliability even for intricate biological explorations like cell trajectory and differentiation analyses.

Semi-profiling maintains accuracy on a heterogeneous colorectal cancer dataset

We further tested the effectiveness of *scSemiProfiler* by validating it against a notably heterogeneous colorectal cancer dataset [53], which encompassed 112 single-cell sequencing samples that passed the quality control. This collection comprised 19 normal tissues, 86 tumor tissues (including colorectal cancer subtype iCMS2 and iCMS3), and 7 lymph node tissues. Considering the inherent diversity of this dataset, achieving accurate semi-profiling could ostensibly test the limits of *scSemiProfiler*. Again, this study does not include paired bulk sequencing data. Therefore, we have utilized the pseudo-bulk data derived from single-cell analysis as a surrogate for actual bulk sequencing. Nevertheless, following a consistent data processing and semi-profiling protocol, and selecting 36 representatives in batches of 4, the semi-profiled dataset mirrored its real-profiled counterpart. This congruence manifested not only in visual similarity but also in cell type distributions and subsequent analyses outcomes. Using the same estimation method as applied to the COVID-19 cohort, the total cost for both bulk and single-cell sequencing to obtain this highly similar semi-profiled single-cell cohort is approximately \$73,320. This price also includes the cost of bulk data for the cohort and represents only 43.6% of the \$168,000 estimated for conducting single-cell sequencing on the entire cohort.

Figure 4a and b graphically highlight the remarkable similarity between the semi-profiled and real-profiled data using UMAP visualizations. These visualizations, color-coded according to cell types, show a substantial alignment between the datasets. The semi-profiled data mirrors the real-profiled data in terms of the location and shape of each cell type cluster. Notably, both datasets effectively segregate cell types such as plasma B, enteric glial, Mast, and epithelial. Additionally, a nuanced connection between fibroblast and endothelial cells is evident in both versions. Immune-centric cells like McDC, T_NK, and B cells are also accurately positioned in close proximity in both datasets, underscoring the precision of the semi-profiled data. This similarity is further emphasized

in Figure 4c, which showcases the significant overlap between the two datasets. Moreover, Figure 4d employs distinct color schemes to differentiate between cells generated by the deep generative learning model and those from real-profiled representative data. The cells from the representatives cover a substantial portion, indicating their well-chosen representative selection. However, numerous cells that fall outside the representatives' distribution are accurately generated by the deep learning model, highlighting the critical role of both active learning and the deep generative model in achieving effective semi-profiling. The accuracy of the deep learning model's generation can be further shown in Supplementary Fig. S10, where we present the model's single-cell inference for individual samples.

To further justify the semi-profiled gene expression values are accurate and can be used for biological analysis, we compute the gene set activation pattern of the GO term "activation of immune response"(GO:0002253) for the two tumor tissue types the same way as we did for the COVID-19 cohort (Fig. 4e). We chose this term because the immune response plays a significant role in the body's defense against cancer, and its activation or suppression can influence cancer progression and patient outcomes [54, 55]. The gene set activation scores are calculated and then adjusted by subtracting the score of the "Normal" tissue. The activation patterns in the real-profiled and semi-profiled datasets are highly similar, leading to a high Pearson correlation coefficient of 0.919 between them.

We also quantitatively examined the overall performance of *scSemiProfiler* as different numbers of representatives are selected. As shown in Fig. 4f, while our approach trumps the selection-only method, the gap narrows in comparison to results on the COVID-19 dataset—owing largely to the colorectal cancer dataset's inherent heterogeneity. Despite this, the deep generative model's efficacy remains conspicuous, ensuring cost-effective error reduction. Also, if we aim to achieve an error as low as the previous COVID-19 cohort, which leads to almost identical analysis results as the real data, only half of the samples need to be selected as representatives. This still reduces the cost significantly.

Diving deeper into the cell type proportions within the colorectal cancer cohort, one discerns variations across different tissue types—"Lymph Node", "Normal", "iCMS2", and "iCMS3". Fig. 4g illustrates these differences. For example, "Lymph Node" contains an expanded population of B cells compared with other tissue types, "Normal" is enriched with PlasmaB cells, and the two tumor subtypes have a pronounced epithelial presence. Remarkably, the semi-profiled dataset captures these nuances with precision, underlining its capability to replicate intricate analyses with fidelity.

Lastly, the benchmarking of deconvolution results presented in Fig. 4h and i positions *scSemiProfiler* at the forefront, significantly outperforming existing methods such as Bisque, TAPE, and Scaden. While its performance with four representatives is on par with CIBERSORTx, *scSemiProfiler* excels in computational memory efficiency. Further extending the representatives to 36 dramatically boosts the deconvolution accuracy. In contrast, existing methods such as CIBERSORT falter when tasked with handling a large reference set like 36 representatives, mainly due to their computational inefficiencies. This distinction underscores the *scSemiProfiler*'s distinct advantage unshared by its peers. To provide a more detailed perspective on our deconvolution outcomes, Supplementary Fig. S11 showcases a comparative analysis. It displays our predicted cell type proportions for each individual sample alongside the ground truth, enabling a side-by-side evaluation

scSemiProfiler Ensures consistent downstream analyses between semi-profiled and real single-cell data in heterogeneous colorectal cancer cohorts

In the context of a heterogeneous dataset like this colorectal cancer one, the semi-profiled dataset stands robust, offering downstream analysis results that mirror the real-profiled data. This close resemblance is consolidated in Fig. 5.

A notable observation is the accuracy in analyzing biomarker expression pattern and their intra-cluster variation using semi-profiled data. Fig. 5a showcases dot plots for top cell type signature genes derived from the real dataset. These plots reflect an identical pattern in both the real-profiled and semi-profiled datasets. Further affirmation comes from the strong Pearson correlation coefficient between the colors (0.994) and sizes (0.996) of the dots. Notably, these correlation coefficients even surpass those observed in the more homogeneous COVID-19 dataset. The semi-profiled dataset also reproduces biomarker discovery results, establishing its credibility as a suitable stand-in for the real-profiled data in such analyses. This consistency is exemplified by genes like KRT18 and KRT8, exclusive to epithelial cells, corroborated by existing literature [56, 57]. Another illustration is the unique expression of CPA3 and TPSB2 in Mast across both datasets. Beyond these top cell type signature genes, a granular examination of epithelial cells—encompassing the top 50 positive and

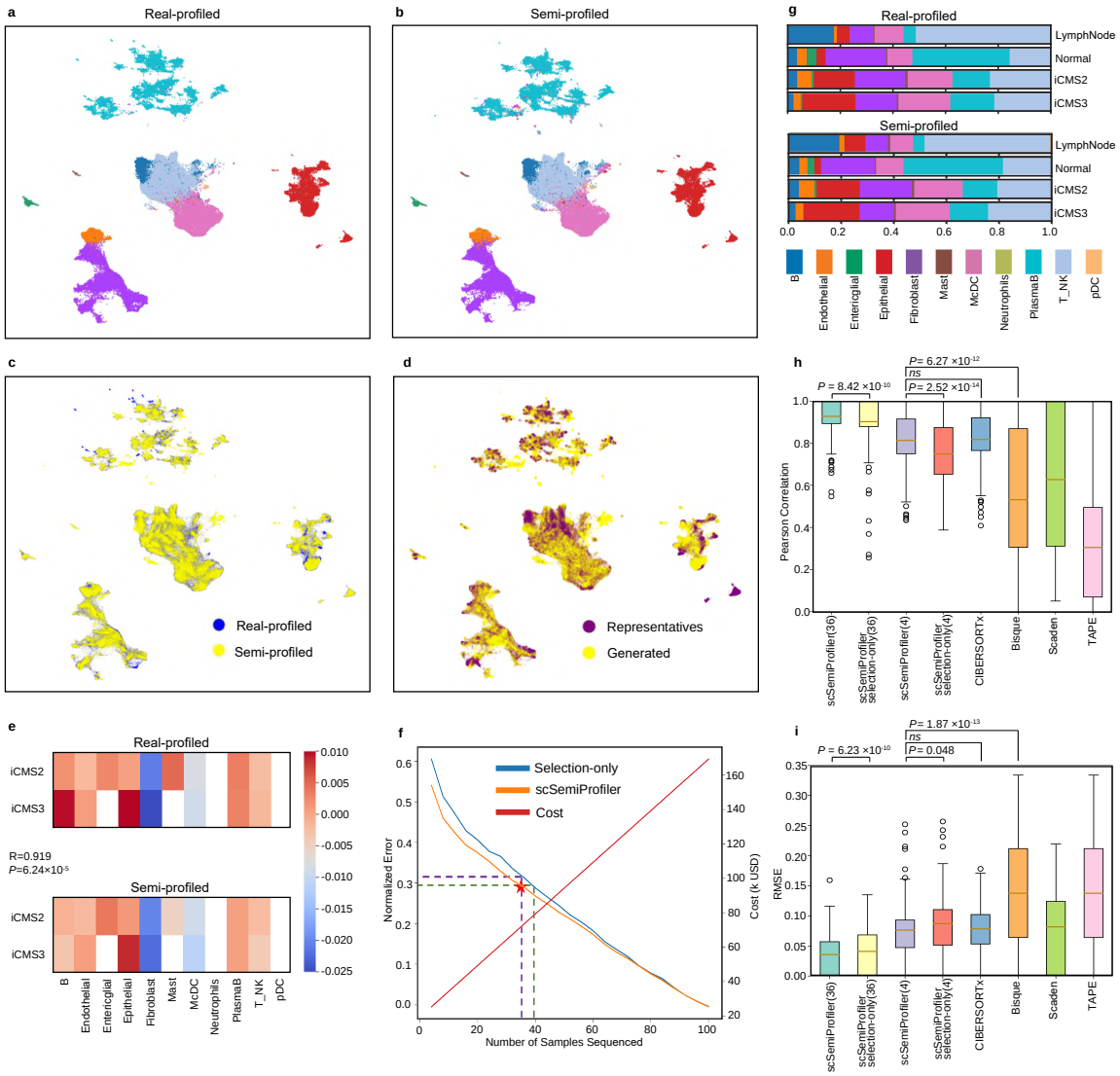

Fig. 4 Detailed comparisons between the semi-profiled and real-profiled data in the heterogeneous colorectal cancer dataset. a, UMAP visualization of the real-profiled data, with colors denoting distinct cell types. Colors are consistent with (**g**). **b**, UMAP visualization of the semi-profiled data, with colors denoting distinct cell types. **c**, Joint UMAP visualization highlighting the close resemblance between the semi-profiled and real-profiled data. **d**, UMAP plot of the semi-profiled dataset, with color-coding distinguishing cells from the actual sequenced representatives and the ones generated through semi-profiling. **e**, "Activation of immune response" gene set relative activation pattern calculated for different tumor tissue types as compared to the "Normal" type in the real-profiled and semi-profiled datasets. Entries with fewer than 500 cells are left blank. **f**, Performance trajectory of the *scSemiProfiler* on the colorectal cancer dataset, showcasing its superiority over the selection-only approach, with costs computed similarly to Fig. 2d. **g**, Stacked bar plots comparing cell type compositions between the semi-profiled and real-profiled datasets across different tissues. The Pearson correlation coefficients between the real-profiled and semi-profiled tissues are LymphNode: 0.995, Normal: 0.993, iCMS2: 0.994, iCMS3: 0.988. **h-i**, Cell type deconvolution benchmarking. **h**, Pearson correlation coefficients between the actual cell type proportions and those estimated by various deconvolution methodologies. Note: DWLS failed to yield results after a week-long computation. **i**, Root Mean Square Error (RMSE) comparisons among different deconvolution techniques, highlighting the computational efficiency and accuracy of the *scSemiProfiler*, especially with an extended set of representatives.

negative markers—reinforces the congruence. All markers were identified using data at single-cell resolution through thorough statistical testing, a process unachievable with decomposed cell type-level data from standard bulk decomposition methods. As depicted in Fig. 5b, the preponderance of highly significant entries, with many p-values lower than 10^{-50}, strongly indicates a high degree of similarity in marker lists between the two datasets. RRHO plots for additional cell types can be found in Supplementary Fig. S12.

Diving deeper into the gene functional enrichment analysis, Fig. 5c offers further validation. Analyzing the top 100 signature genes across both datasets reveals a staggering 96 common markers,

yielding a hypergeometric test p-value of 8.88×10^{-188} (population size = 4053). GO terms from both datasets, along with their respective p-values, showcase pronounced similarity. Although the heterogeneity of the dataset leads to a relatively lower Pearson correlation coefficient of 0.593, the overall patterns in the two plots remain statistically similar ($P = 0.042$), leading to the same scientific conclusions. Moreover, when considering the union of the top 10 Gene Ontology (GO) terms from both the semi-profiled and actual datasets (comprising a total of 12 terms), there is a significant similarity in terms of enrichment p-values. Notably, the two versions of the top 10 GO terms have 9 overlap terms. More comprehensive GO term and pathway enrichment analysis results for other cell types' signature genes are also consistent for the real-profiled and semi-profiled versions (Supplementary Figs. S13 and S14).

Despite the increased heterogeneity of the dataset, the analysis of cell-cell interactions with the colorectal cancer semi-profiled cohort remains promising. As illustrated in Fig. 5d, the cell-cell interaction analysis, when executed on the real-profiled tumor tissue cells and the semi-profiled counterpart, reveals substantial consistency. Navigating the intricate interaction patterns characteristic of tumor tissues, the semi-profiled data astonishingly replicates the intricate layout with a robust Pearson correlation coefficient of 0.933. Both versions highlight enteric glial and fibroblast as the primary senders, while neutrophils emerge as the predominant receivers. Significantly, the most intense interactions identified across both sets involve the enteric glial with itself, the enteric glial with neutrophils, and the fibroblast with neutrophils. The cell-cell interaction results for other tissues also exhibit a high degree of similarity between the real-profiled and semi-profiled versions, as shown in Supplementary Fig. S15. Additionally, the strikingly similar PAGA plots presented in Supplementary Fig. S16 further demonstrate the utility of semi-profiled data in studying cellular trajectories and relationships between different cell types.

Shifting the focus to pseudotime analysis, we evaluated epithelial cells across all tissues as an example demonstration (Fig. 5e). Given the presence of cells from tumor tissues, this might introduce elevated heterogeneity within the cell type. Yet, the consistency between pseudotime analysis results from both versions is significant. Both versions discern lower pseudotime values concentrated at the base of the cluster, culminating in larger values towards the upper regions, with the pinnacle being the top-right quadrant. The statistical significance of the similarity is further validated by a Mann-Whitney U test [58, 59], yielding a compelling p-value of 2.84×10^{-197}. The pseudotime analyses for other cell types also demonstrate a high degree of similarity, as evidenced in Supplementary Fig. S17. Such a finding underscores the capability of the semi-profiled data to adeptly capture intra-cluster nuances in detailed analyses.

Semi-profiling with real bulk measurements yields a dataset nearly identical to the original single-cell data

To further illustrate the adaptability of *scSemiProfiler* in real-world applications, we directed our analysis towards the iMGL dataset [60], which uniquely profiles both single-cell and real bulk RNA-seq measurements for the human inducible pluripotent stem cell (iPSC)-derived microglia-like (iMGL) cells, differing from the pseudobulk datasets previously used. There are 25 samples having both single-cell data and bulk RNA sequencing data. Samples are of different conditions (grown in cell culture for 0-4 days and under various treatments). The availability of such datasets, which include both single-cell and bulk sequencing data on a large scale, remains very limited, partially due to the unnecessity of doing both sequencing for the same large-scale cohort and its prohibitive cost. This deviation from pseudobulk data is challenging. Pseudobulk, created by averaging out the noisy single-cell data [61], often fails to encapsulate the subtle features of actual bulk RNA-seq measurements. These real bulk measurements are inherently less noisy and often exhibit systematic differences when compared to single-cell RNA-seq data.

To navigate this complexity, we devised a method to infer pseudobulk directly from real bulk data (refer to the Methods section for comprehensive details). This approach enables us to more effectively utilize our deep learning model in the pseudobulk data space, which often aligns more closely with single-cell data. Through this method, despite the intricate challenges of the iMGL dataset, the results in Fig. 6 illustrate that our semi-profiled data parallels the real-profiled data quite closely. By selecting only eight representative samples out of a total of 25 samples, we could notably reduce the overall cost without compromising on accuracy. For this smaller dataset, we employed active learning to select 8 representative samples in batches of 2. Using the same estimation method as in the other two studies, the total cost for acquiring both bulk and semi-profiled single-cell data through

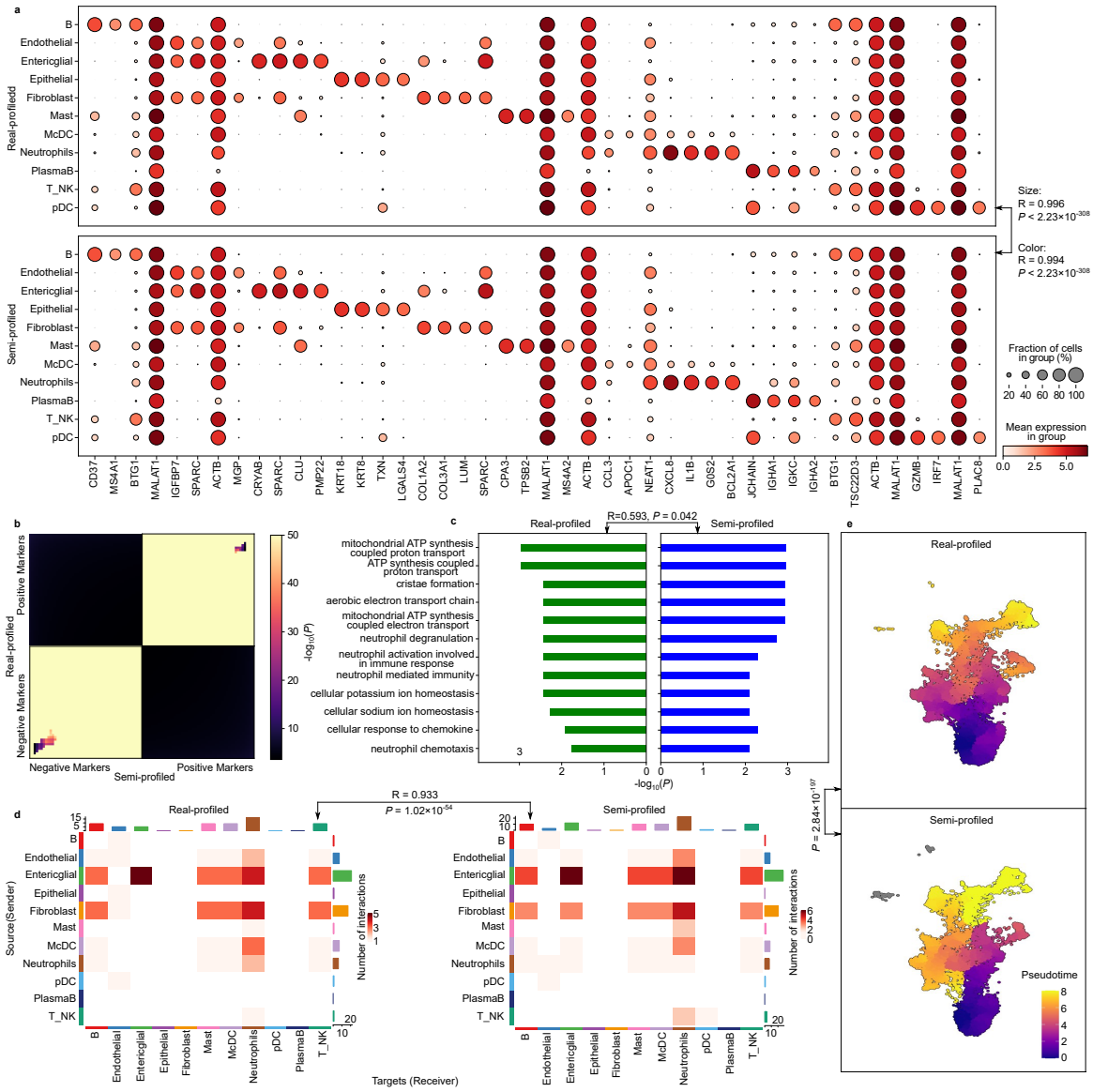

Fig. 5 Downstream analysis results comparisons for the colorectal cancer dataset. a, Dot plots visualizing the cell type signature genes. **b**, RRHO plot visualizing the comparison between semi-profiled and real-profiled markers of epithelial cells. **c**, Epithelial cell type signature genes GO enrichment analysis results comparison. **d**, Cell-cell interaction results comparison between the real-profiled tumor tissue cells and semi-profiled tumor tissue cells. **e**, Pseudotime results comparison using the epithelial cells.

our method is approximately $21,750. This amount is just 58% of the estimated $37,500 required for conducting single-cell sequencing across the entire cohort.

The UMAP visualization, as presented in Fig. 6a and b, solidifies our findings. Here, the two versions — semi-profiled and real-profiled — show remarkable consistency. The cell distributions of various iMGL subtypes (as delineated by the dataset provider) in the UMAP follow nearly identical patterns. The detailed observation showcases that clusters like "C2: Activated, immediate-early", "C4: Activated, non-immediate-early", and "C5: Activated, immediate-early" are interconnected and primarily located on the right-hand side. Likewise, "C3: Homeostatic, proliferative" finds its position at the upper left, with "C1: Homeostatic, non-proliferative" and "C6: Freshly thawed" lying at the bottom left. This consistency transcends to Fig. 6c, emphasizing a consistent cell distribution across the two versions. Furthermore, Fig. 6d underlines the precision of *scSemiProfiler*, where a majority of cells in the semi-profiled version were accurately generated. The efficient coverage of representative cells in this figure also highlights that our active learning strategy remains robust

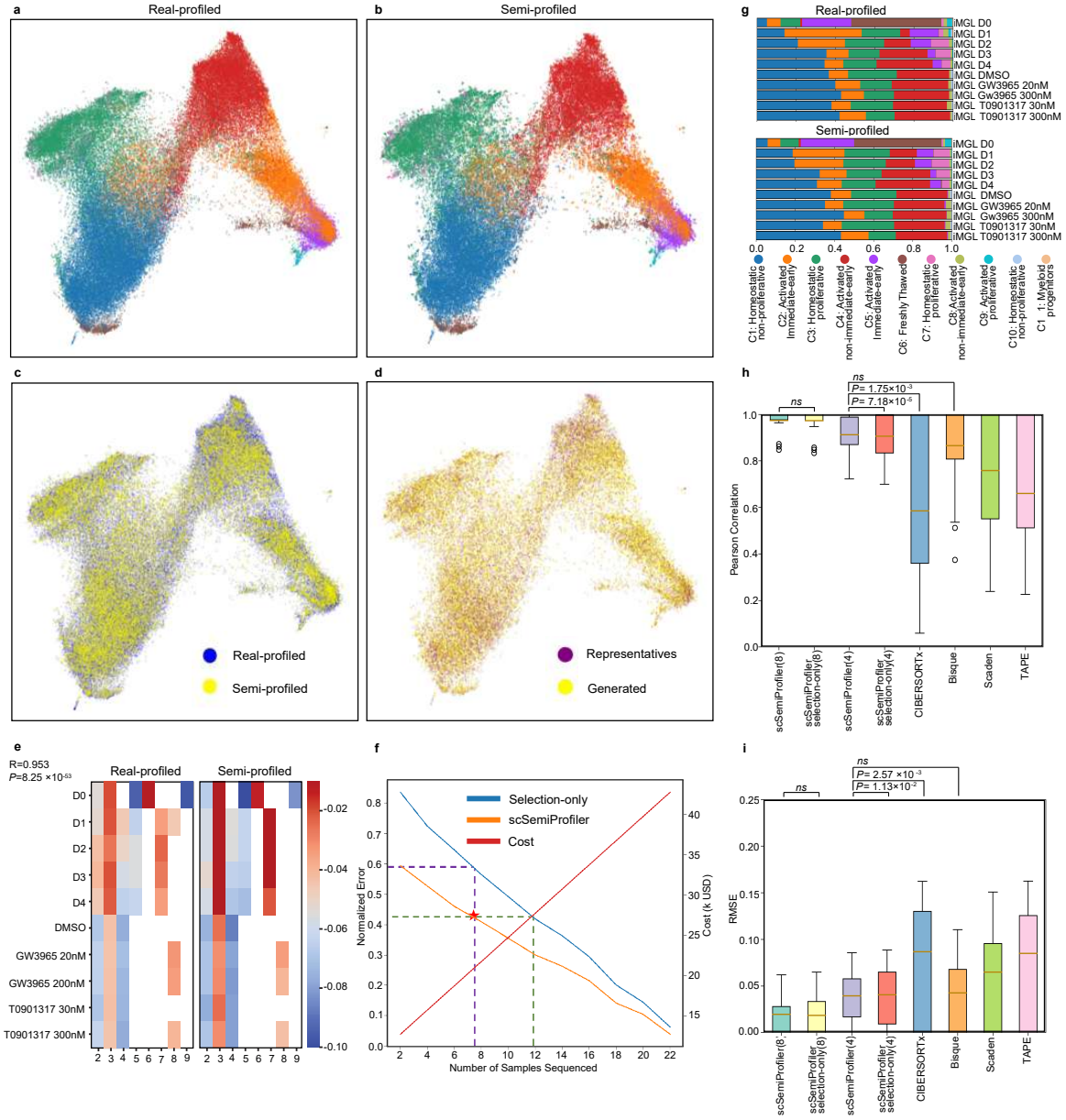

Fig. 6 Comparative analyses between semi-profiled and real-profiled iMGL datasets. a, UMAP visualization of the real-profiled iMGL cohort. Different colors represent different cell types. Colors are consistent with (**g**). **b**, UMAP visualization of the real-profiled iMGL cohort. **c**, Combined UMAP visualization showcasing the consistent cell distribution across both data versions. **d**, UMAP visualization highlighting the representatives' cells alongside the semi-profiled cells within the semi-profiled dataset. **e**, Relative pathway activation pattern of the GO term "activation of immune response" calculated for cells of different treatments as compared to cell type "C1: Homeostatic non-proliferative" in the real-profiled and semi-profiled cohorts. Entries with fewer than 100 cells are left as blank. **f**, Performance evaluation of the *scSemiProfiler* on the iMGL dataset, emphasizing its efficiency in error reduction. **g**, A comparative illustration of cell type proportions under varying experimental conditions, accentuating the similarity in patterns between datasets. The Pearson correlation coefficients between the real-profiled and semi-profiled versions of cell type proportions under different conditions are: iMGL_D0: 0.999, iMGL_D1: 0.871, iMGL_D2: 0.993, iMGL_D3: 0.989, iMGL_D4: 0.992, iMGL_DMSO: 0.998, iMGL_GW30: 0.960, iMGL_GW_300: 0.999, iMGL_T_30: 0.987, iMGL_T_300: 0.999. **h,i**, Deconvolution performance benchmarking using Pearson correlation and RMSE.

even when navigating the challenges of real bulk measurements. The supplementary Fig. S18 further illustrates the precise single-cell inference achieved by our deep learning model on this dataset.

Fig. 6e provides an in-depth look at the accuracy of semi-profiled gene expression values. Since microglia are the resident immune cells in the brain, we checked the activation pattern of GO term "activation of immune response" in each cell type under each treatment. The semi-profiled dataset presents a highly similar activation pattern as the real-profiled dataset.

Fig. 6f offers further evidence of the *scSemiProfiler*'s effectiveness. Here, the performance curve of the semi-profiled approach significantly undercuts the selection-only one, demonstrating its capability to limit semi-profiling errors while optimizing on costs.

An intriguing observation from the iMGL dataset is the cell type proportion's dynamic shifts under various experimental conditions (Fig. 6g). The transitions from iMGL_D0 to iMGL_D4, for instance, reveal a progressive increase in the proportions of C1 and C4 cells. Contrastingly, "C2: Activated, immediate-early" cells peak at iMGL_D1 and then decrease steadily. Although the intricate effects of drugs iMGL_GW and iMGL_T need further investigation, preliminary data suggests that elevated doses result in a surge of "C1: Homeostatic, non-proliferative" cells. Impressively, these intricate variations are mirrored in the semi-profiled dataset.

A juxtaposition of our deconvolution method against others on the iMGL dataset offers illuminating insights. The performance of CIBERSORTx, which was one of the best in other datasets, dramatically decreases, potentially due to the inherent challenges posed by real bulk data and the nuanced similarities among cell types. Other methods, such as Bisque, TAPE, and Scaden, are more robust to those challenges, showing decent deconvolution performance. Despite these challenges, *scSemiProfiler* showcases resilience and consistently outperforms all its peers except Bisque, a fact further corroborated by the Wilcoxon test [62] (see p-values in Fig. 6h and i). The marginal difference between our *scSemiProfiler* and the selection-only method is a testament to both approaches nearing optimal performance in this specific context. The supplementary Fig. S19 further presents our accurate deconvolution for individual samples.

Semi-profiling using real bulk measurement also leads to reliable downstream results

In this more realistic setting, where scSemiProfiler is tasked with semi-profiling using real bulk data for downstream analyses, the semi-profiled data consistently mirrored results from the real-profiled version. This is particularly remarkable given the unique challenges presented by the real bulk data. We present these downstream analysis results in Fig. 7.

The markers identified using single-cell resolution data in the real-profiled dataset were almost identical in expression patterns to those in the semi-profiled dataset. Fig. 7a visually reinforces this, showing "C1: Homeostatic, non-proliferative" and "C10: Homeostatic, non-proliferative" with virtually indistinguishable expression patterns across both data types. Additionally, unique expressions in "C11: Myeloid, progenitors", such as GAPA2 and HPGDS, were consistently observed. The overarching similarities were further quantified with impressive Pearson correlation coefficients for both dot sizes and colors, clocking in at 0.980 and 0.989, respectively.

Further validating the congruency of our method, Fig.7b presented the RRHO plot of the top 50 positive and negative C3 markers from both single-cell datasets (see RRHO plots for other cell types in Supplementary Fig. S20). The degree of similarity is substantial, with most of the markers showcasing p-values less than 10^{-50}. Such findings strongly suggest that the *scSemiProfiler* is adept at producing reliable data for biomarker discovery.

Proceeding to more in-depth downstream analysis using the top 100 signature genes for GO enrichment, we observed an overlap of 90 genes between the semi-profiled and the real single-cell datasets (Fig. 7c). The hypergeometric test revealed a significant p-value of 1.81×10^{-183}. The enriched terms identified from both the semi-profiled and real datasets matched closely. The Pearson correlation coefficient between the two versions' significance is 0.995, underscoring the consistency in their analytical outcomes. Extended GO and Reactome enrichment analysis in Supplementary Figs. S21 and S22 further confirm the accuracy of finding signature genes using semi-profiled data.

In the case of cell-cell interactions, both real and semi-profiled single-cell data did not capture any significant interactions, probably due to the similarity between cell clusters in the iMGL data. Instead, the partition-based graph abstraction (PAGA) analysis [63] showcased in Fig.7d highlighted that major cell type links were consistent across datasets. Further cementing this was the strong Pearson correlation coefficient of 0.865 between the adjacency matrices of the two networks.

Pseudotime analysis in Fig.7e also affirmed the alignment between the two datasets. The topographical pseudotime alignment between them is almost congruent, to the extent that the Mann-Whitney U test implemented in SciPy returned a p-value smaller than the smallest float number our computer can represent, which is 2.23×10^{-308}.

In conclusion, our findings robustly demonstrate that *scSemiProfiler* seamlessly adapts to real-world scenarios employing real bulk data. The downstream analytical outcomes derived from the semi-profiled data are significantly consistent with those based on real data.

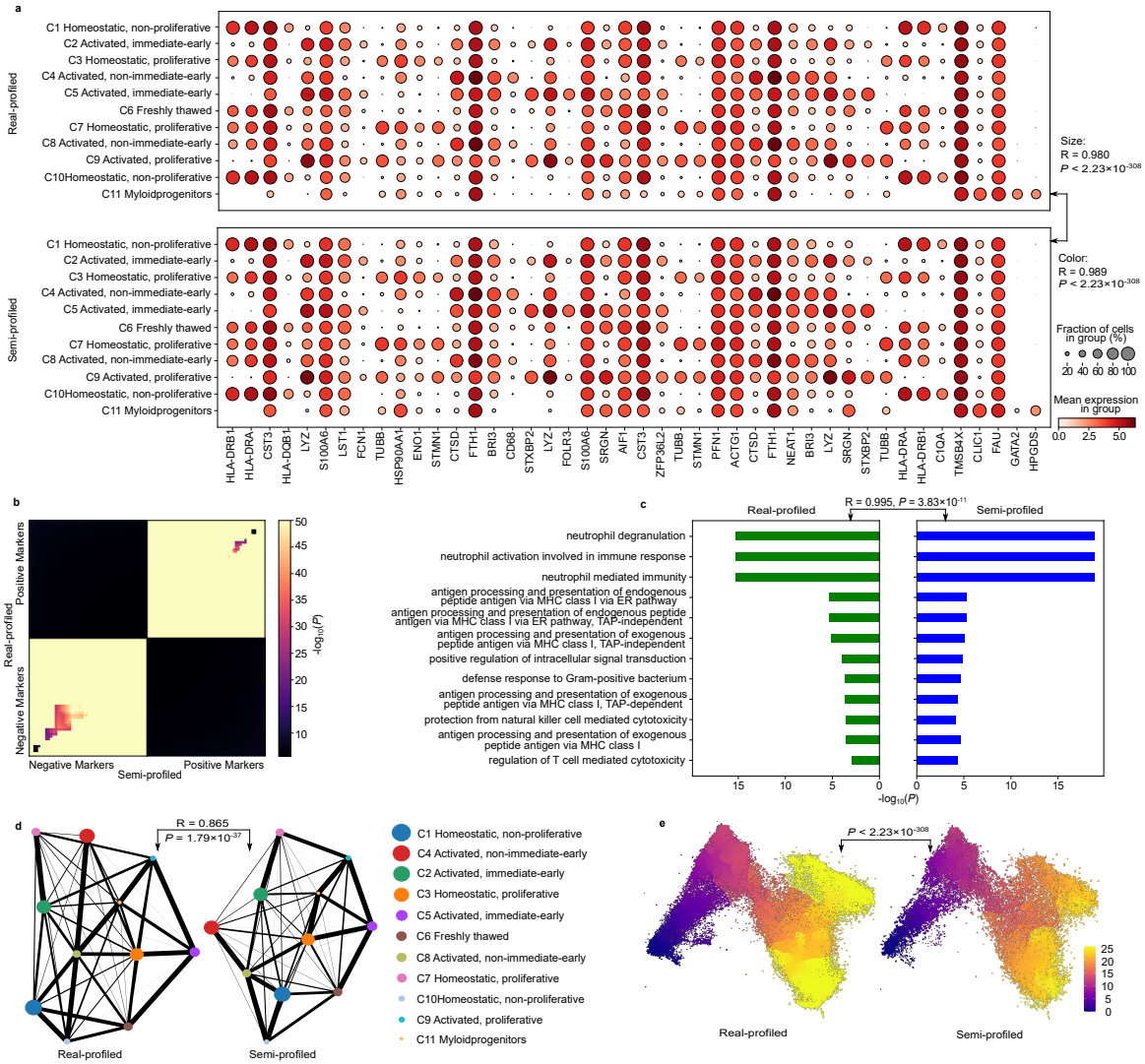

Fig. 7 Similarities in downstream single-cell analyses between real-profiled and semi-profiled data for the iMGL cohort. a, Dot plots visualizing the nearly identical expression patterns of cell type signature genes across both datasets. **b**, RRHO plots highlighting the striking similarities between the top 50 positive and negative C3 markers from both real and semi-profiled datasets. **c**, Overlapping GO enrichment analysis results for the top C3 signature genes, emphasizing consistent analytical outcomes between the datasets. **d**, PAGA plots illustrating the consistent major cell type links observed in both datasets. **e**, Pseudotime plots affirming the topographical alignment between the real-profiled and semi-profiled cohort. The p-value of the Mann-Whitney U test is smaller than the smallest float number that the system can represent, which is 2.23×10^{-308}.

Active learning demonstrates its prowess in selecting the most informative samples for enhanced single-cell profiling

The crux of *scSemiProfiler*'s strategy revolves around judiciously selecting representative samples. The rationale is straightforward: the more informative the chosen representatives, the better the semi-profiling performance. This not only enhances the fidelity of the generated profiles but also provides a cost-effective approach by minimizing the number of necessary representatives.

Initially, our methodology bore similarities to uncertainty sampling [28, 64, 65], a heuristic active learning technique. Here, the intuition is to query samples with the most uncertainty, thereby maximizing the incremental information acquired. In our algorithm, we employed bulk data to pick out

batches of samples that exhibited the most variance from their designated representatives. Since this method does not use information from the base learner (the deep generative model), it is still a passive learning algorithm.

To improve the representative selection, we then turn the algorithm into an active learning algorithm by incorporating the information from the deep generative models. The algorithm also utilizes the clustering information in the cohort and is thus a type II active learning algorithm [66]. Combining these two ideas, the algorithm aims to reduce the total heterogeneity of each sample cluster, ensuring that each target sample has a similar representative, thus optimizing semi-profiling performance.

As depicted in Fig. 8, we juxtaposed our advanced active learning algorithm against the rudimentary passive learning approach. Each panel, from Fig. 8a to Fig. 8c, encapsulates the comparative analyses derived from distinct datasets. While the x-axis maps the representatives earmarked for single-cell sequencing, the y-axis portrays the single-cell *in silico* inference difficulty, which is quantified using the average single-cell-level difference (see Equation 16 in the Methods section) between target samples and their representatives. This premise holds that as the dissimilarity between the target and the representatives increases, the complexity of the *in silico* inference task also rises. This metric is crucial as it highlights the challenges encountered by the deep generative learning model in semi-profiling, thereby illuminating the effectiveness of the strategies used for selecting representatives.

The empirical evidence is resounding. Across all datasets, the active learning algorithm showcased its mettle by consistently pinpointing representatives that considerably reduced the total distance to other samples. This underscores the algorithm's capability to foster superior representative selection for semi-profiling. In Fig. 8a, when applying our method to the COVID-19 dataset, active learning shows better performance than passive learning, especially in the beginning iterations of the semi-profiling. Active learning reduces the inference difficulty when the same number of representatives are selected. Also, consider the marked point where 28 representatives are selected, which is consistent with representatives we selected for our analyses in previous sections. To reach the same level of inference difficulty, passive learning needs to select 4 more representatives, i.e. the cost of 4 samples' single-cell sequencing experiment is saved. For the colorectal cancer dataset (Fig. 8b), active learning continues to perform better than passive learning, and at the point we selected, the cost for more than two batches (8) of representatives can be saved using active learning. Fig. 8c shows the results for the iMGL dataset, in which active learning is also significantly better than passive learning. In all experiments, active learning consistently outperforms passive learning in selecting representatives. With the same budget, active learning achieves lower inference difficulty. Furthermore, to reach a comparable level of inference difficulty, active learning requires a lower cost.

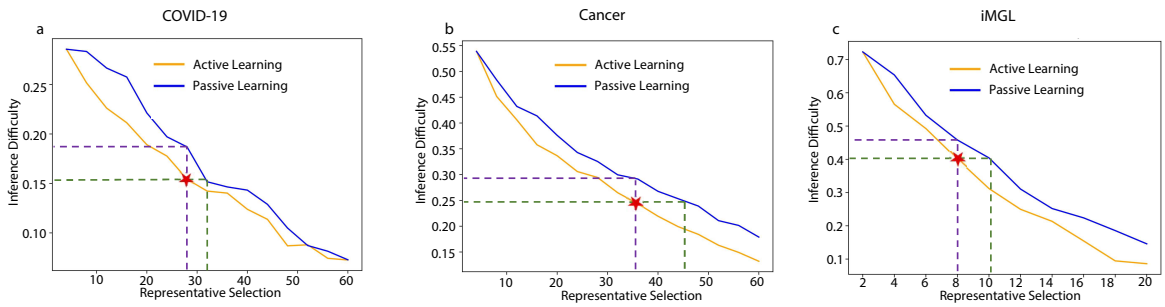

Fig. 8 Active learning demonstrates its prowess in selectively profiling the most informative samples at the single-cell level. The x-axis represents the number of samples selected for single-cell profiling (representatives). The y-axis shows the single-cell *in silico* inference difficulty of the dataset, which is quantified by the average single-cell difference from each sample to its representative, showcasing the efficiency of representative selection strategies. The marked stars signify the iterations chosen for our methodology, with the generated data underpinning the analyses detailed in previous sections. **a**, Results from the COVID-19 dataset. Active learning shows significantly better performance, especially in the beginning when a few representatives are selected. **b**, Observations derived from the colorectal cancer dataset. Active learning continues to show significantly better performance even when more representatives are selected. **c**, Insights from the iMGL dataset with real bulk measurements. Active learning still manages to outperform passive learning significantly.

Discussion

In the present work, we introduced *scSemiProfiler*, an innovative computational framework designed to produce affordable single-cell measurements for a given large cohort. The allure of *scSemiProfiler* stems from its unparalleled capacity to mirror the outcomes of conventional single-cell sequencing, yet with remarkable cost-effectiveness. Marrying expansive sequencing with pinpoint profiling and optimizing data utilization, it charts a robust, intricate, and economical path for large-scale single-cell exploration. This tool is not merely a means of generating data but ensures the derived single-cell information remains dependable for an array of downstream analyses. The pipeline ingests bulk data, leverages an active learning module to judiciously select representatives for authentic single-cell sequencing, and employs a deep generative learning model, VAE-GAN, to infer the single-cell data of the remainder of the cohort. Through this approach, every sample in the cohort has access to single-cell data, either semi-profiled or real-profiled. We subjected *scSemiProfiler* to rigorous evaluation using three distinct cohorts, simulating a scenario devoid of single-cell data and then juxtaposing the output of our tool with the authentic single-cell datasets. The striking resemblance between semi-profiled and real-profiled data in facets like UMAP visualization and cell type proportions ascertains the robustness of *scSemiProfiler*. Furthermore, both datasets align almost perfectly in subsequent analyses, encompassing biomarkers, enrichment analyses, cell-cell interactions, and pseudotime trajectories, offering a cost-effective means of integrating single-cell data in cohort studies without compromising on analytical precision.

The magnitude and uniqueness of our study's contributions can be primarily attributed to two game-changing innovations that stand apart in their novelty and transformative potential. **Precision-driven Computational Modeling:** The crowning achievement of our study lies in the design and implementation of a groundbreaking computational model for single-cell level bulk decomposition. While traditional methodologies offer a mere cell type level decomposition of bulk data, *scSemiProfiler* shatters these bounds with an unprecedented capability. What we've pioneered isn't just a method to interpret bulk data, but a mechanism to deconvolute it into authentic single-cell data. This offers a myriad of sophisticated and meaningful downstream analytic possibilities, previously unattainable with conventional techniques. To the best of our knowledge, *scSemiProfiler* is the very first in its class to provide a true single-cell level decomposition. Furthermore, when placed under rigorous scrutiny, the semi-profiled data generated by our model consistently showcases remarkable similarity to real-profiled single-cell data, underscoring its reliability and potential to reshape the paradigm of single-cell analytics, particularly for large-scale cohort studies in which real single-cell sequencing cost would be prohibitive. **Intelligent Active Learning Mechanism:** Our second salient innovation hinges on the power of active learning. Instead of passively selecting representatives, our model embarks on an intelligent, iterative journey. By constantly evaluating the data and learning from previous rounds of selection, this module discerns which representatives would offer the most informative insights for semi-profiling. This isn't a mere addition; it's a reinvention of how representative selection operates. The synchronization of insights from the deep learning model with this active learning module ensures that the selection process is not just data-driven, but also insight-driven. The implications of this are twofold: firstly, it enhances the quality and relevance of the selected representatives, ensuring that they truly resonate with the cohort's cellular composition. Secondly, it offers an economic advantage. By intelligently choosing representatives, the financial overheads associated with single-cell experiments can be minimized, ensuring maximum return on investment both in terms of data quality and budgetary constraints. Together, these innovations not only elevate the efficacy of *scSemiProfiler* but also pioneer a new direction in how single-cell data can be derived, analyzed, and utilized, holding profound implications for future biomedical research endeavors.

Future work on *scSemiProfiler* is focused on expanding its capabilities from single-cell RNA sequencing (scRNA-seq) to other biological modalities, such as proteomics. The adaptability of the tool's current methodology will require adjustments in algorithms to suit the unique data characteristics of each new modality, along with comprehensive performance evaluations benchmarked against established standards. A pivotal element of this development is the utilization of cross-modal information exchange, which aims to improve the accuracy of semi-profiling. This approach could significantly reduce the costs associated with single-cell multi-omics studies beyond the scope of the current study, by minimizing the need for extensive representative samples from each modality and leveraging the inherent similarities across various biological data types.

The *scSemiProfiler* represents a groundbreaking shift in single-cell analysis, particularly for large-scale and cohort studies, by offering a cost-effective yet comprehensive approach. It achieves this by selectively performing single-cell experiments on a few samples using active learning, coupled with deep generative models for *in silico* inference of single-cell profiles for the remaining samples, thereby creating similar semi-profiled cohorts at a substantially lower cost. This approach not only alleviates financial constraints in expansive biomedical research but also extends its utility beyond cohort studies. The tool's deep generative model enables "single-cell level deconvolution" across various studies, provided there is a single-cell reference. Initially centered on RNA sequencing datasets, *scSemiProfiler*'s adaptability to other single-cell and bulk data modalities opens new avenues in personalized medicine and can significantly enrich global data repositories. Significantly, scSemiProfiler is designed to complement, rather than compete with, large-scale single-cell sequencing technologies, such as WT Mega by Parse Biosciences. WT Mega is notable for its ability to process up to 1 million cells across 96 samples. When scSemiProfiler is integrated with these advanced sequencing platforms, it further reduces costs, enabling single-cell level profiling of thousands of samples in large-scale studies at a more manageable expense. This synergy significantly broadens the scope and depth of possible research, enhancing the efficiency and affordability of large-scale single-cell studies.

Methods

The initial setup of *scSemiProfiler*

The initial setup of the scSemiProfiler method plays a pivotal role (depicted in Fig. 1 a), serving as the foundation for subsequent semi-profiling iterations. Initially, each individual sample within the designated cohort undergoes bulk sequencing. Upon obtaining the raw count data from this sequencing, we perform library size normalization followed by a log1p transformation. From the processed data, we identify and select highly variable genes and relevant markers for further analysis. Subsequently, we employ Principal Component Analysis (PCA) [67–69], implemented in the Python package Scikit-learn [70], to reduce the dimensionality of the processed bulk data. We then determine the batch size of representatives (B) representing the number of selected representatives for actual single-cell profiling in each iteration. Once the representative batch size is defined, we cluster the dimensionality-reduced bulk data to select the initial batch of representatives. The dimensionality-reduced data is clustered into B distinct clusters using the KMeans algorithm [71], the sample that is closest to the cluster centroid is designated as the representative for that cluster. This well-prepared initial setup provides the clustered bulk sequencing data and identified representatives for high-resolution single-cell sequencing.

Representative single-cell profiling and processing

After the selection of representatives, the scSemiProfiler method proceeds to the single-cell profiling phase, focusing specifically on the chosen representatives (Fig. 1b). The raw count data obtained from single-cell profiling undergoes a sequence of preprocessing steps. These steps encompass traditional quality controls as well as advanced processing techniques aimed at enhancing the learning of the deep generative model.

To initiate, expression values that are extremely low (below 0.1% of the library size) are adjusted to zero. This adjustment is based on insights from prior studies [72, 73], which have demonstrated that these minimal values are often representative of background noise and, therefore, should be excluded. The threshold for this processing is empirically determined by selecting the value that yields the most high-quality UMAP visualizations consistent with the original cell type annotations. Subsequently, a series of standard preprocessing steps commonly used by existing methods [74, 75] is applied, including the removal of low-quality cells (expressing fewer than 200 genes), genes (expressed in fewer than 3 cells), dead cells (with a high proportion of mitochondrial reads), and doublets (expressing an excessive number of genes). Afterward, the data is normalized to a library size (total count of each cell) of 10,000, followed by a log1p transformation. The data matrix is then cropped to include only the highly variable genes identified from the bulk data. Our preprocessing approach is carefully tailored to meet the unique requirements of each dataset. For instance, the gene expression data of the COVID-19 dataset arrive in a preprocessed state, including cell quality control, normalization, and log1p transformation. We retrieved the data as normalized RNA counts, eliminated the background noise, and subsequently reapplied normalization and log1p transformation. Finally, we selected the columns corresponding to highly variable genes and important cell type markers. In contrast, the

colorectal cancer dataset is provided in raw counts, necessitating a full spectrum of standard single-cell preprocessing procedures. With the iMGL dataset, our focus is on cells annotated by the dataset provider. We identify genes common to both bulk and single-cell datasets and apply standard preprocessing steps, excluding the removal of low-quality cells, as this task has already been performed by the dataset provider.

After the initial preprocessing, the single-cell data is further refined by integrating two types of prior knowledge to enhance cell representations. Firstly, gene set scores are calculated using all curated gene sets from MSigDB [48] and Ernst et al [76], sourced from UNIFAN [77]. Gene set scores are determined by averaging the expression values for each gene set, followed by a log1p transformation, and then combined with the preprocessed single-cell gene expression matrix. Secondly, we adopt a feature weighting method, giving more weight to features with higher variance in each cell's input when calculating reconstruction loss. More comprehensive details of this strategy are available in the section focusing on single-cell inference.

Pretrain the deep generative learning model for reconstructing the single-cell data of the selected representatives

The next phase, depicted in Fig.1c, involves training a deep generative learning model for the *in silico* single-cell data inference of a non-representative target sample using processed representative single-cell data and bulk data of both samples. This will be executed for all non-representative samples to ensure everyone in the study cohort will have single-cell data (real-profiled or semi-profiled) in the end. Firstly, our deep generative model aims to reconstruct the single-cell profiles of representatives. This reconstruction lays the foundation of the single-cell inference of the target samples, which can be viewed as a modified single-cell data reconstruction task. When performing the single-cell inference, the initial single-cell data reconstruction of the representative is modified by introducing the difference between the target sample and the representative based on their distinct bulk data profiling, aligning the synthetic single-cell data for the target sample closer to its actual single-cell profiling counterpart. The fundamental reasoning is that the single-cell data matrix of a specific target sample should bear resemblance to the representatives chosen from the same study. Significant differences between them can be adjusted and guided by the disparity as delineated by the bulk sequencing data.

We designed a VAE-GAN-based model for the representative reconstruction and the target single-cell generation guided by their expression difference at the bulk level. It ingests gene expression data into an MLP encoder, which subsequently outputs parameters of a multivariate Gaussian. A random variable, $\mathbf{z}$, is sampled from this Gaussian and processed through an MLP decoder, resulting in parameters of a Zero-Inflated Negative Binomial (ZINB) distribution - optimal for modeling single-cell RNA-seq data [78–80]. Training the VAE parameters involves maximizing the data likelihood and minimizing the KL divergence [81] between the latent variable distribution and a standard Gaussian distribution, following the Evidence Lower BOund (ELBO) loss framework [30].

Innovating upon this foundational VAE structure, we integrated the following four techniques to enhance effective cell representation learning. **Gene Set Score Inclusion**: As mentioned in the data preprocessing section, beyond mere gene expression reconstruction, we compute gene set scores for cells and concatenate them to the gene expression data. During the two stages of pretrain, we also compute the reconstruction loss for gene set scores. Such inclusion furnishes the model with an enriched biological context, fostering a more comprehensive learning of the input cells. Hence, input for each cell becomes the concatenation of gene expression and gene set scores: $(\mathbf{x_i}, \mathbf{s_i})$. **Feature Importance Weight**: We also compute a weight vector $\mathbf{w}$ to weight the contribution of each feature to the VAE's reconstruction loss based on the feature's variance. **Graph Convolutional Networks (GCN)** [82, 83]: A GCN layer assimilates adjacent cells' information, mitigating dropout concerns and the inherent noise of single-cell sequencing. **Generative Adversarial Network (GAN)** [31] **Dynamics**: We employ a discriminator network, resonating with GAN structures, which discerns between genuine and generated cell data. This guides the generator towards producing more authentic single cell data. The resultant loss function for the generator discussed above is:

$$L_{G_{Pretrain1}} = L_{VAE} - \lambda_d L_{D_{Fake}} \tag{1}$$

$$L_{VAE} = \sum_{i=1}^{N} [-\mathbf{w} \cdot log(ZINB((\mathbf{x_i}, \mathbf{s_i}) | \boldsymbol{\rho}(\boldsymbol{z_i}), \boldsymbol{\pi}(\boldsymbol{z_i}), \boldsymbol{\theta}))) + KL(q(\mathbf{z}|(\mathbf{x_i}, \mathbf{s_i})) || \mathcal{N}(\mathbf{0}, I))] \tag{2}$$

$$L_{D_{Fake}} = \sum_{i=1}^{N} -log(1 - D(G((\mathbf{x_i}, \mathbf{s_i})))) \tag{3}$$

$L_{G_{Pretrain1}}$ is the loss used to train the VAE-based generator network during the first pertrain stage. It has two components, a VAE loss L_{VAE} and another loss $L_{D_{Fake}}$ including the feedback from the discriminator. In L_{VAE}, N is the number of cells in the representative's single-cell dataset. $\mathbf{x_i}$ is the vector representing the gene expression value of the i-th cell in the dataset and $\mathbf{s_i}$ is the corresponding gene set score. $\mathcal{N}(\mathbf{0}, I)$ is the standard Gaussian distribution. $\mathbf{z_i}$ is a low-dimensional latent variable sampled from the Gaussian distribution $q(\mathbf{z}|(\mathbf{x_i}, \mathbf{s_i}))$ whose parameters $\boldsymbol{\mu_i}$ and $\boldsymbol{\Sigma_i}$ are generated by the encoder network: $MLP_{Encoder}(GCN((\mathbf{x_i}, \mathbf{s_i})))$. We follow the ZINB distribution design of SCVI [78] and generate ZINB parameter using the decoder network: $\boldsymbol{\rho(z_i)} = MLP_{Decoder,\rho}(\mathbf{z_i})$, $\boldsymbol{\pi(z_i)} = MLP_{Decoder,\pi}(\mathbf{z_i})$ and $\boldsymbol{\theta}$ is a free parameter vector that are learned in the training process. In $L_{D_{\text{Fake}}}$, G represents the VAE-based generator so $G((\mathbf{x_i}, \mathbf{s_i}))$ represents the i-th cell reconstructed by the generator. The reconstructed cell data is the mean of the generator's $ZINB$ distribution. And D is our discriminator network, so $D(G((\mathbf{x_i}, \mathbf{s_i})))$ is the discriminator's predicted probability of the i-th reconstructed cell being a real-profiled cell. λ_d denotes an empirically determined scaling factor. Meanwhile, the discriminator network will be trained using cross-entropy loss to reach higher classification performance.

$$L_D = \sum_{i=1}^{N} [-log(D((\mathbf{x_i}, \mathbf{s_i}))) - log(1 - D(G((\mathbf{x_i}, \mathbf{s_i}))))] \tag{4}$$

Here, $D((\mathbf{x_i}, \mathbf{s_i}))$ represents the discriminator's predicted probability of the i-th input cell being real when it is indeed real, and $D(G((\mathbf{x_i}, \mathbf{s_i})))$, as mentioned in a previous paragraph, represents the discriminator's predicted probability of the i-th cell being real when it is, in fact, a cell reconstructed by the generator network. When trained together, the generator will first be trained until convergence, followed by alternating training with the discriminator every three epochs. Together, they form a GAN with a min-max objective:

$$min_G max_D \sum_{i=1}^{N} [log(D((\mathbf{x_i}, \mathbf{s_i}))) + log(1 - D(G((\mathbf{x_i}, \mathbf{s_i}))))] \tag{5}$$

In this context, the discriminator's objective is to maximize its accuracy, aiming for precise classification, while the generator's goal is to minimize discrepancies, generating realistic cells to challenge the discriminator.

Overall, our approach involves two pretrain stages and one fine-tune stage for performing single-cell inference. In the initial pretrain stage, we employ $L_{G_{Pretrain1}}$ and L_D to train the generator G and discriminator$D$, respectively, forming a GAN. The primary objective here is to train a generator capable of reconstructing genuine cell data from the representatives. Subsequently, the second pretraining stage closely resembles the first, but it functions in full-batch mode. This modification enables the inclusion of an extra bulk loss term in the generator's loss function. Such an addition improves the model's capacity to leverage insights from bulk data and better aligns the single-cell data with its bulk counterpart. This enhancement is crucial for priming the model for the *in silico* inference of single-cell data from bulk data, while also strengthening the data reconstruction process.

$$L_{G_{Pretrain2}} = L_{VAE} - \lambda_d L_{D_{Fake}} + \lambda_{BulkR} L_{BulkR} \tag{6}$$

$$L_{BulkR} = \left\| \frac{\sum_{i=1}^{N} \mathbf{x'_i}}{N} - \frac{\sum_{i=1}^{N} \mathbf{x_i}}{N} \right\|_2 \tag{7}$$

The term λ_{BulkR} represents another empirical hyperparameter for regularization weight. $\mathbf{x'_i}$ represents the reconstructed i-th cell. The bulk loss is defined as the disparity between the pseudobulk of the real representative dataset and that of the reconstructed representative dataset.

Fine-tune the deep generative learning model to infer the single-cell measurements for the target samples

After successfully pretraining our deep generative model so that it can perform accurate single-cell data reconstruction, the subsequent stage of our method leverages this foundation. In this phase,

we introduce an extra term to the loss function that accounts for the difference between the bulk data of the representative and that of the target sample. This fine-tuning process refines the model's capabilities, enabling it to perform *in silico* generation of cells from the target sample while maintaining consistency with the observed bulk expression patterns. Consequently, our deep generative model transitions from being solely focused on reconstruction to generating single-cell data for the target sample. In this crucial stage, we retire the discriminator from our model architecture, as our focus shifts from generating cells that resemble the representatives' cells to adapting the model for the target samples. Consequently, the training process exclusively involves the generator. The loss function for the generator adopted in this stage bears resemblance to that of the second pretrain phase, with the distinction of the bulk loss transitioning from L_{BulkR} to L_{BulkT}:

$$L_{\text{Inference}} = L_{\text{VAE}} + \lambda_{\text{BulkT}} L_{\text{BulkT}} \tag{8}$$

$$L_{\text{BulkT}} = \left\| \frac{\sum_{i=1}^{N} \mathbf{x_i'}}{N} - \frac{\sum_{i=1}^{N} \mathbf{x_i}}{N} \cdot \frac{\mathbf{B_T} + \boldsymbol{\epsilon}}{\mathbf{B_R} + \boldsymbol{\epsilon}} \right\|_2 \tag{9}$$

Here, $\mathbf{B_T}$ and $\mathbf{B_R}$ represent the real bulk data of the target and representative samples, respectively, while $\boldsymbol{\epsilon}$, whose elements are all set to 0.1, is introduced as a pseudocount vector to avoid division by zero. The second term of equation 9, $\frac{\sum_{i=1}^{N} \mathbf{x_i}}{N} \frac{\mathbf{B_T}+\boldsymbol{\epsilon}}{\mathbf{B_R}+\boldsymbol{\epsilon}}$, functions as an estimation of the pseudobulk value for the target sample, derived from the actual bulk data. This term is incorporated to reduce the systematic discrepancy between the pseudobulk and the actual bulk data corresponding to the same sample.

By incorporating the actual bulk difference alongside the pseudobulk data of the representative, this term effectively simulates the pseudobulk data for the target sample, ensuring that any disparities between the pseudobulk and actual bulk do not distort the model's performance. In cases such as the COVID-19 and colorectal cancer dataset, where real bulk data is unavailable, the entire second term simplifies to the target's pseudobulk. Furthermore, as the bulk loss L_{BulkT} inherently propagates more gradient to higher gene expression values (given the target bulk loss is in a squared error form), we implement a strategy to gradually reduce the gradient received by larger values. See details in the "Deep Generative Learning Model Training Details" section. Concurrently, we increase the weight of the bulk loss in the loss function over the course of training. This refined approach ensures that, after adequately training for larger values, the model can fine-tune smaller values, resulting in a comprehensive and precise single-cell data generation for the target samples. To counteract the influence of potentially noisy gene expression with minuscule values, we introduced a neuron activation threshold. This threshold converts values that are exceedingly small, specifically those below 0.1% of the input library size (a threshold consistent with that used for background noise removal in preprocessing), to zero.

Deep generative learning model architecture

In this study, our model integrates a Variational Autoencoder (VAE) as the generator and a Multi-Layer Perceptron (MLP) as the discriminator. We developed both components using `PyTorch` [84]. A concise overview of the model architecture is provided below, and for an in-depth understanding, we encourage consulting the complete details available in our GitHub repository (refer to the code availability section).

Before sending the cell data into the model, we performed some precomputation for the cells, which is part of a Graph Convolutional Network (GCN) layer. We aggregated information from the K (15 by default) nearest neighboring cells for each input cell. The neighbor finding for each cell was based on the 'neighbors.kneighbors_graph' tool in the Scikit-learn package. We preprocessed this message-passing step on the log1p transformed cell expression data, subsequently inputting the aggregated cell data into the deep learning model. The encoder of the VAE then employs linear layers and activation functions, leading to a hidden layer comprising 256 neurons. The message passing, the first linear layer, and the first activation function in the encoder form a GCN layer together. The design of the rest of the VAE structure, inclusive of both the encoder and decoder, was adapted from SCVI. We chose 32 for the latent space dimension and 256 for the decoder's hidden layer size. For the discriminator, we constructed a 4-layer MLP, with Leaky ReLU activations [85] between linear layers. The hidden layers are configured with 256, 128, and 10 neurons, respectively. The output layer employs a Sigmoid activation to produce the prediction probability of the input cell being real. Weight initialization in the model utilizes the Kaiming method [86].

Deep Generative Learning Model Training Details

Following the detailed outline of the VAE-GAN model structure, we next describe the detailed training settings, which were meticulously designed to optimize the model's performance. The training process was structured in a three-stage sequence to ensure comprehensive learning and fine-tuning of the model parameters.

Initially, in the first pretraining stage, we focus on training the VAE generator independently until convergence (setting 100 epochs as a default value). Then, the generator and discriminator are trained jointly until the generator cannot further reduce its loss. When trained together, the generator and the discriminator are trained alternatively for 3 epochs each, and this will go for 100 iterations by default. During this phase, we employed the default SCVI learning rate of 1×10^{-3}.

The second pretraining stage mirrors the first in its basic approach but incorporates notable modifications. It is executed in full-batch mode, with an additional representative bulk loss integrated into the training process. Here, the generator is again trained until convergence (set to 50 epochs by default), and then trained jointly with the discriminator until convergence. The default setting is training the two networks for 50 iterations, where each network is trained for 3 epochs in each iteration. To facilitate more refined adjustments during this stage, the learning rate is reduced to 1×10^{-4}.

In the final stage of training, the single-cell inference fine-tuning stage, the discriminator is set aside, and training is executed through a series of mini-stages. Each mini-stage involves setting the gradients for highly expressed genes above specific thresholds to zero, a strategy aimed at ensuring equitable optimization for genes with smaller expression values. These thresholds are calculated relative to the peak expression value (max_expr) from the representative's normalized count matrix, set at No threshold, $\frac{1}{2}$max_expr, $\frac{1}{4}$max_expr, $\frac{1}{6}$max_expr, and $\frac{1}{8}$max_expr. Training is continued until convergence in each mini-stage, typically not exceeding 150 epochs, with a learning rate of 2×10^{-4}. As the thresholds are lowered, the weight for the target bulk loss in the loss function is progressively quadrupled, ensuring a consistent total magnitude for this loss component. Upon completion of all mini-stages, the trained model is then utilized to generate single-cell data for the target sample.

Incrementally select representatives using active learning to improve single-cell inference

Upon executing the deep generative learning models for *in silico* inference, every sample in the cohort possesses single-cell data, either real-profiled or semi-profiled. While researchers can opt to conclude here and proceed with downstream analysis using the single-cell data, when budget permits, our active learning module can further refine the process. By pinpointing additional informative representatives for single-cell sequencing, some target samples will be assigned a more similar representative, leading to enhanced single-cell inference performance.

Our initial representative selection strategy is predicated on the belief that selecting the most "uncertain" samples as the representatives results in the most inference improvement and yields the richest insights for our model. Consequently, we select the sample with the largest bulk data difference compared to its representative to serve as the new representative:

$$R_{new} = \underset{i \in Cohort}{\arg\max}\, D_b(R(i), i) \tag{10}$$

Where $R(i)$ is the representative of i, and D_b is the Euclidean distance between the PCA-reduced bulk data of two samples. This process is executed B (representative batch size) times. Upon choosing new representatives, we update the membership for all samples. This ensures that, if a sample aligns more closely with a new representative than their previous one, their affiliation shifts:

$$R(i) = \underset{r \in Rs}{\arg\min}\, D_b(r, i) \tag{11}$$

where $R(i)$ is the representative of the sample i and Rs is the set of representatives. Conceptually, this approach is similar to the uncertainty sampling [64] algorithm, a subtype of active learning. However, this method is passive learning since it does not utilize the knowledge acquired by the base learners, our deep generative learning models.

To improve the representative selection, we developed our active learning algorithm by merging the information gained by the deep generative model base learners into the selection of new

representatives. To commence, we pinpoint the B sample clusters exhibiting the highest in-cluster heterogeneity, which is the aggregate distance from each sample to the cluster's representative. The heterogeneity for a cluster c is expressed as: $\sum_{i \in c}[D_b(R_c, i) + \lambda_{sc} D_{sc}(R_c, i) + \lambda_{pb} D_{pb}(R_c, i)$, where i can be any sample in the cluster and R_c is the representative. Essentially, a sample's heterogeneity is the combined distance—across three distance metrics—between the sample and its representative. The total cluster heterogeneity is derived from the summation of all sample heterogeneity. D_b denotes the Euclidean distance between the PCA-reduced bulk data of two samples. The term $D_{sc}(R_c, i)$ represents the difficulty of transforming the representative single-cell data to that of a target. This is quantified by averaging the Euclidean distances between the representative cells and their K nearest neighbors in the target sample, all within a PCA-reduced space. Formally:

$$D_{sc}(a, b) = \frac{1}{N_a} \sum_{i=1}^{N_a} \sum_{j \in KNN(i)} ED(\mathbf{v_{a,i}}, \mathbf{v_{b,j}}) \tag{12}$$

with a, b being two samples in the semi-profiled cohort, v_a, v_b being PCA reduced single-cell data matrices for the two samples, $ED(x, y)$ computing the Euclidean distance between x and y, and $KNN(i)$ identifying the cell's K (1 by default) nearest neighbor in another sample's data matrix. Subscript represents rows, e.g. $\mathbf{v_{a,i}}$ is the ith row (ith cell) of the PCA reduced data of sample a. D_{pb} is the Euclidean distance of the log1p transformed pseudobulk data of two samples, which is positively related to the amount of bulk loss and thus quantifies the deep learning model's difficulty in performing the *in silico* inference. Finally, λ_{sc} and λ_{pb} are empirical scaling factors.

The most heterogeneous cluster C_H is chosen as:

$$C_H = \underset{c \in clusters}{\arg\max} \sum_{i \in c} [D_b(R_c, i) + \lambda_{sc} D_{sc}(R_c, i) + \lambda_{pb} D_{pb}(R_c, i)] \tag{13}$$

After selecting the B most heterogeneous clusters to split, the subsequent step involves selecting a new representative for each of these B clusters to minimize total in-cluster heterogeneity. Within each cluster, every non-representative sample is a potential new representative, each representing a different way of splitting the cluster. In the cluster, samples that are closer in bulk distance D_b to the potential new representative than to the original representative can be reassigned to the new representative, splitting the cluster into two. The total heterogeneity of these two new clusters is then calculated based on the bulk distance. For each of the B most heterogeneous clusters, we select the sample whose corresponding split results in the minimal total in-cluster heterogeneity to be the new representative R_{new}, as shown in the equation below.

$$R_{new} = \underset{j \in C_H}{\arg\min} \sum_{i \in C_H} min(D_b(j, i), D_b(R_c, i)) \tag{14}$$

Where R_c denotes the original representative of this cluster. Finally, the cluster membership updating will be executed to make sure each sample i is assigned to the closest representative $R(i)$.

$$R(i) = \underset{r \in Rs}{\arg\min} D_b(r, i) \tag{15}$$

With the identification of new representatives, they are subjected to real single-cell data profiling and appended to the existing collection. In subsequent semi-profiling iterations, certain target samples will be assigned with more analogous representatives, enhancing *in silico* single-cell inference accuracy. This iterative procedure persists until either the budget is exhausted or a sufficient number of representatives are ascertained, ensuring optimal semi-profiling results.

Semi-profiling pipeline stop criteria

For our semi-profiling pipeline that iteratively selects representatives and performs *in silico* single-cell inference, we recommend two types of stop criteria for the users to choose according to their own needs. First, in the case that the user knows their budget and aims to achieve the best possible semi-profiling performance, they can run the pipeline and keep choosing representatives until they run out of budget. Second, in the case that the user aims to achieve an acceptable semi-profiling performance with the least amount of budget, the iterative cycle persists until an acceptable performance is reached. The performance is measured by comparing the representatives' actual single-cell data and

the inferred data for them in previous iterations. Based on the user's needs, we recommend three levels of criteria. First, when the user is mainly studying the cell type proportion, we recommend stopping when the Pearson correlation between the ground truth and the estimated proportion reaches 0.85. Second, when the user's study includes more detailed analysis, such as biomarker discovery, we recommend stopping when the Pearson correlations between the biomarker expression patterns (dot colors and sizes in the dot plots) discovered in real-profiled data and those in semi-profiled data reach 0.85. Finally, for studies requiring more intricate single-cell level analysis, it is advised to halt when the Pearson correlation between data matrices, generated from advanced analysis tasks like trajectory inference and cell-cell interaction, for semi-profiled and real-profiled data, reaches 0.85. These guidelines are merely suggestions, and users have the flexibility to modify the stopping criteria to suit their specific requirements. The effectiveness of the semi-profiling process can be consistently evaluated by comparing the semi-profiled single-cell data (from the previous round) with the actual single-cell profiling (in the subsequent round). If the actual single-cell profiling aligns with the semi-profiling generated by our model from the previous round, this indicates a robust semi-profiling model has been effectively learned, and the iterative process can thus be concluded.

Semi-profiling performance evaluation

Cell type Annotation for Semi-Profiled Data: Utilizing the annotated single-cell data from the representatives, we trained an MLP classifier available in the Scikit-learn package [70]. This classifier was then employed to categorize the cell types for cells in the semi-profiled data.

Dimensionality Reduction: In order to juxtapose the real-profiled and semi-profiled cohorts, we merged them and executed Principal Component Analysis (PCA) to reduce it to 100 principal components. This PCA-processed data serves as the foundation for computing single-cell level difference metrics. For visualization purposes, we first used SCANPY [75] to compute the neighbor graph using the PCA-reduced data, with the local neighborhood size parameter set to 50. Subsequently, the Uniform Manifold Approximation and Projection (UMAP) functionality within SCANPY was used for the visualization process.

Semi-profiling error quantification: To assess the accuracy of the semi-profiled cohort, we performed PCA to reduce its dimensionality, in conjunction with the real-profiled data, to 100 dimensions. For each sample within the semi-profiled dataset, we calculated a single-cell difference metric in comparison to the real-profiled single-cell data corresponding to that sample. This involves determining the Euclidean distance of each cell to its K-nearest neighbors in another dataset within the PCA-reduced space, followed by averaging these distances across all cells. The formula below outlines the difference computation between single-cell matrices $\mathbf{m_a}$ and $\mathbf{m_b}$.

$$Dif_{sc}(\mathbf{m_a}, \mathbf{m_b}) = \frac{1}{N_a + N_b}\left(\sum_{i=1}^{N_a} \sum_{j \in KNN(i)} ED(\mathbf{v_{a,i}}, \mathbf{v_{b,j}}) + \sum_{i=1}^{N_b} \sum_{j \in KNN(i)} ED(\mathbf{v_{b,i}}, \mathbf{v_{a,j}})\right) \tag{16}$$

Where Dif_{sc} denotes single-cell level difference. N_a and N_b signify the cell counts in matrices $\mathbf{m_a}$ and $\mathbf{m_b}$ respectively. The $\mathbf{v}$s represent the PCA-reduced matrices. The terms $\mathbf{v_{a,i}}$, $\mathbf{v_{b,j}}$ represent the ith and jth cell (rows) in the PCA-reduced matrices of sample a and b respectively. The function KNN identifies the K nearest neighbors within the alternate sample's PCA-reduced data matrix. We chose $K = 1$ for our experiments. ED computes the Euclidean distance between two vectors.

Referring to the comprehensive performance curves depicted in (Fig.2d, Fig.4d, Fig.6d, the normalized error E_N displayed on the left y-axis represents the mean single-cell difference between each sample's real-profiled version and semi-profiled version. This error is normalized using a theoretical upper bound (UB) and a lower bound (LB) of an expected single-cell difference between samples in the study.

$$E_N = \frac{1}{UB - LB}\left(\frac{1}{P}\sum_{i=1}^{P} Dif_{sc}(\mathbf{m_i}, \tilde{\mathbf{m}}_\mathbf{i}) - LB\right) \tag{17}$$

and if i itself is a representative, the single-cell difference equals the lower bound.

$$Dif_{sc}(\mathbf{m_i}, \tilde{\mathbf{m}}_\mathbf{i}) = LB, \quad \text{if } i \in Rs$$

where P stands for the total count of samples within the cohort and $\tilde{\mathbf{m}}_\mathbf{i}$ refers to the semi-profiled version of $\mathbf{m}_\mathbf{i}$. Rs denotes the set of representatives. The upper bound UB was derived by averaging the single-cell difference of randomly paired samples based on their PCA-reduced real-profiled data. This represents the performance of the worst possible representative selection, which is random selection. For computing the lower bound LB, we first divided the comprehensive PCA matrix, which contains all single-cell samples, into two random halves. Due to the randomness of the split, these two halves have the same data probabilistic distribution. Furthermore, given the large sample size, these halves are expected to have approximately the same actual cell distribution and thus can be regarded as two replicates of the same study. The difference between the replicates should be regarded as the lower bound, as it represents the best performance that an optimal selection can achieve [87].

Deconvolution benchmarking: We benchmarked the accuracy of our deconvolution method by comparing its performance against established methods: CIBERSORTx, Bisque, TAPE, Scaden, and DWLS. We ran CIBERSORTx using the official web portal. Bisque, TAPE, and DWLS were installed according to the official GitHub repositories provided in the corresponding publications. We used the PyTorch version of Scaden implemented in TAPE's GitHub repository for its testing. First, we annotated our cells generated by the deep learning model using the MLP classifier, which was trained on the representatives' annotated data, as discussed previously. Using this annotated data, we computed the cell type proportion. Subsequently, we employed the Pearson correlation coefficient (calculated using the Python package SciPy [59]) and RMSE to assess the consistency with the ground truth. However, when applying the aforementioned methods to the three datasets, we encountered both time and space complexity constraints for many benchmarked methods. Specifically, CIBERSORTx and Bisque were unable to utilize as many representative samples for single-cell references as our approach. Scaden and TAPE always only sample 5,000 cells for training their models, thereby also failing to fully exploit all available representatives. To ensure a fair comparison, we also offered an alternative version of our results for these datasets. In this version, the single-cell reference is confined to just the first 4 representatives, aligning with the capacity of all other benchmarked methods. Across all three datasets, DWLS failed to produce results even after running for a week and was thus excluded from the comparison. To statistically ascertain if one method notably surpasses another in performance, we calculated the p-values using a one-sided Wilcoxon test [62], as implemented in SciPy. This test enables us to determine whether the metric of one method is significantly higher or lower compared to that of another.

Gene set activation scores computation and heatmap plotting: We computed the interferon pathway activation scores following the same procedure in the study [35] from which we acquired the COVID-19 dataset. To compute the activation scores depicted in the heatmaps (refer to Fig.2c), we first computed the average expression for each cell type and severity combination using the code from the COVID-19 dataset provider, which uses the 'tl.score_genes' tool in SCANPY. Then, for each cell type, we computed the activation score of each severity level as the fold changes from the healthy condition to it. For the "activation of immune response" activation pattern in the colorectal cancer dataset depicted in Fig.4c, we first collected the genes of this GO term (GO:0002253) and used the same SCANPY tool to compute the average expressions for each cell type and tissue combination. For each cell type, we calculated the activation scores by determining the fold changes in comparison to the "Normal" tissue type across all other tissue types. For the "activation of immune response" activation pattern in the iMGL dataset depicted in Fig.6c, we first computed the average gene expression for each cell type and tissue combination similarly. The activation score was determined by calculating the fold change for each cell type, using "C1: Homeostatic non-proliferative" as the reference background. Considering the relatively smaller size of the colorectal cancer and iMGL dataset and the fact that some entries have very few cells, the values were only computed for entries with more than 500 cells for the colorectal cancer dataset and entries with more than 100 cells for the iMGL dataset.

Biomarker discovery and enrichment analysis: For making the RRHO plots, we used SCANPY to identify each cell type's top 50 positive cell type markers and top 50 negative cell type markers for the real-profiled and semi-profiled datasets. RRHO plots are used for visualizing the overlap between two ranked gene lists. Each entry in the plot corresponds to a negative logged p-value of the hypergeometric test between two marker lists. The bottom left quadrant corresponds to the

comparison between two positive marker lists. In this quadrant, the plot entry in the i-th row from bottom to top and j-th column from left to right corresponds to the comparison between the top i genes in the real-profiled version top negative markers and top j gene in the semi-profiled version top negative markers. Other quadrants are plotted similarly, with the marker lists always starting from the respective corners of the plot.

For the GO and Reactome enrichment analysis, we identified the top 100 cell type signature genes for both real-profiled and semi-profiled datasets using SCANPY. The identified markers were subsequently employed for enrichment analysis with the Python package GSEApy [88]. For enrichment analysis visualization in Fig.3c, Fig.5c, Fig.7c, we employed the adjusted p-value from GSEApy's output to indicate significance.

Cell-cell interaction analysis: We conducted cell-cell interaction analyses on both real-profiled and semi-profiled datasets using the R package CellChat [89]. For the COVID-19 dataset, we analyzed interactions of cells from patients with each disease severity. In the colorectal cancer dataset, our focus was on interactions of cells originating from different tissues, and in the iMGL dataset, we evaluated cells across various condition groups.

Pseudotime analysis: We carried out pseudotime analysis utilizing the Monocle3 R package [90]. The UMAP coordinates, previously generated in our UMAP comparison between real-profiled and semi-profiled datasets, served as the input for pseudotime computation. To ensure a fair comparison between the real-profiled and semi-profiled data, we consistently selected the roots at the same positions within the UMAP space. To compare the similarity between the pseudotime values computed for the two datasets, we employed the Mann-Whitney U test, as implemented in SciPy.

Cell trajectory analysis using partition-based graph abstraction (PAGA): The PAGA plots were generated using SCANPY. First, PCA was applied independently to each dataset to reduce them to 100 principal components. Subsequently, the PCA-reduced data was utilized to compute the neighbor graphs, with the size of the local neighborhood set to 50, aligning with our UMAP settings. Based on the neighbor graphs, we generated the PAGA plots using SCANPY.

Data availability

Three datasets are used for evaluating our method. The preprocessed COVID-19 dataset is from Stephenson et al.'s study [35] and can be downloaded from Array Express under accession number E-MTAB-10026 The colorectal cancer dataset is from Joanito et al.'s study [53]. The count expression matrices are available through Synapse under the accession codes syn26844071 The iMGL dataset is from Ramaswami et al.'s study [60]. The raw count iMGL bulk and single-cell data can be downloaded from the Gene Expression Omnibus (GEO) repository under accession number GSE226081.

Code availability

Code for the models and results reproduction is publically available on GitHub: https://github.com/mcgilldinglab/scSemiProfiler.

Acknowledgments

This work was funded in part by grants awarded to [JD]. We gratefully acknowledge the support from the Canadian Institutes of Health Research (CIHR) under Grant Nos. PJT-180505; the Funds de recherche du Québec - Santé (FRQS) under Grant Nos. 295298 and 295299; the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant No. RGPIN2022-04399; and the Meakins-Christie Chair in Respiratory Research. We also extend our gratitude to Yanshuo Chen, the lead author of TAPE, for his indispensable support in the deconvolution benchmarking experiments. This work is part of the HCA publication bundle.

References

[1] Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., Wang, X., Bodeau, J., Tuch, B.B., Siddiqui, A., *et al.*: mrna-seq whole-transcriptome analysis of a single cell. Nature methods

6(5), 377–382 (2009)

[2] Shapiro, E., Biezuner, T., Linnarsson, S.: Single-cell sequencing-based technologies will revolutionize whole-organism science. Nature Reviews Genetics **14**(9), 618–630 (2013)

[3] Li, Y., LeMaire, S.A., Shen, Y.H.: Molecular and cellular dynamics of aortic aneurysms revealed by single-cell transcriptomics. Arteriosclerosis, Thrombosis, and Vascular Biology **41**(11), 2671–2680 (2021)

[4] Kazer, S.W., Aicher, T.P., Muema, D.M., Carroll, S.L., Ordovas-Montanes, J., Miao, V.N., Tu, A.A., Ziegler, C.G., Nyquist, S.K., Wong, E.B., *et al.*: Integrated single-cell analysis of multicellular immune dynamics during hyperacute hiv-1 infection. Nature medicine **26**(4), 511–518 (2020)

[5] Chai, R.L., Grandis, J.R.: Advances in molecular diagnostics and therapeutics in head and neck cancer. Current treatment options in oncology **7**, 3–11 (2006)

[6] Marquette, C.-H., Boutros, J., Benzaquen, J., Ferreira, M., Pastre, J., Pison, C., Padovani, B., Bettayeb, F., Fallet, V., Guibert, N., *et al.*: Circulating tumour cells as a potential biomarker for lung cancer screening: a prospective cohort study. The Lancet Respiratory Medicine **8**(7), 709–716 (2020)

[7] Wedge, D.C., Gundem, G., Mitchell, T., Woodcock, D.J., Martincorena, I., Ghori, M., Zamora, J., Butler, A., Whitaker, H., Kote-Jarai, Z., *et al.*: Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets. Nature genetics **50**(5), 682–692 (2018)

[8] Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J.Z., Menon, M., He, L., Abdurrob, F., Jiang, X., *et al.*: Single-cell transcriptomic analysis of alzheimer's disease. Nature **570**(7761), 332–337 (2019)

[9] Pal, B., Chen, Y., Vaillant, F., Capaldo, B.D., Joyce, R., Song, X., Bryant, V.L., Penington, J.S., Di Stefano, L., Tubau Ribera, N., *et al.*: A single-cell rna expression atlas of normal, preneoplastic and tumorigenic states in the human breast. The EMBO journal **40**(11), 107333 (2021)

[10] Depuydt, M.A., Prange, K.H., Slenders, L., Örd, T., Elbersen, D., Boltjes, A., De Jager, S.C., Asselbergs, F.W., De Borst, G.J., Aavik, E., *et al.*: Microanatomy of the human atherosclerotic plaque by single-cell transcriptomics. Circulation research **127**(11), 1437–1455 (2020)

[11] Wang, Z., Gerstein, M., Snyder, M.: Rna-seq: a revolutionary tool for transcriptomics. Nature reviews genetics **10**(1), 57–63 (2009)

[12] Newman, A.M., Steen, C.B., Liu, C.L., Gentles, A.J., Chaudhuri, A.A., Scherer, F., Khodadoust, M.S., Esfahani, M.S., Luca, B.A., Steiner, D., *et al.*: Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature biotechnology **37**(7), 773–782 (2019)

[13] Jew, B., Alvarez, M., Rahmani, E., Miao, Z., Ko, A., Garske, K.M., Sul, J.H., Pietiläinen, K.H., Pajukanta, P., Halperin, E.: Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nature communications **11**(1), 1971 (2020)

[14] Nel, S., Durandt, C., Murdoch, C., Pepper, M.S.: Determinants of dental pulp stem cell heterogeneity. Journal of Endodontics **48**(10), 1232–1240 (2022)

[15] Papalexi, E., Satija, R.: Single-cell rna sequencing to explore immune cell heterogeneity. Nature Reviews Immunology **18**(1), 35–45 (2018)

[16] Oliveira, G., Wu, C.J.: Dynamics and specificities of t cells in cancer immunotherapy. Nature Reviews Cancer, 1–22 (2023)

[17] Rahimi, A.M., Cai, M., Hoyer-Fender, S.: Heterogeneity of the nih3t3 fibroblast cell line. Cells **11**(17), 2677 (2022)

[18] Liu, Y., Mi, Y., Mueller, T., Kreibich, S., Williams, E.G., Van Drogen, A., Borel, C., Frank, M., Germain, P.-L., Bludau, I., *et al.*: Multi-omic measurements of heterogeneity in hela cells across laboratories. Nature biotechnology **37**(3), 314–322 (2019)

[19] Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., Rinn, J.L.: The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature biotechnology **32**(4), 381–386 (2014)

[20] Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., Trapnell, C.: Reversed graph embedding resolves complex single-cell trajectories. Nature methods **14**(10), 979–982 (2017)

[21] Song, Q., Wang, J., Bar-Joseph, Z.: scstem: clustering pseudotime ordered single-cell data. Genome Biology **23**(1), 1–17 (2022)

[22] Ding, J., Lin, C., Bar-Joseph, Z.: Cell lineage inference from snp and scrna-seq data. Nucleic Acids Research **47**(10), 56–56 (2019)

[23] Ma, Q., Xu, D.: Deep learning shapes single-cell data analysis. Nature Reviews Molecular Cell Biology **23**(5), 303–304 (2022)

[24] Amodio, M., Van Dijk, D., Srinivasan, K., Chen, W.S., Mohsen, H., Moon, K.R., Campbell, A., Zhao, Y., Wang, X., Venkataswamy, M., *et al.*: Exploring single-cell data with deep multitasking neural networks. Nature methods **16**(11), 1139–1145 (2019)

[25] Lin, X., Tian, T., Wei, Z., Hakonarson, H.: Clustering of single-cell multi-omics data with a multimodal deep learning method. Nature communications **13**(1), 7705 (2022)

[26] Wang, J., Ma, A., Chang, Y., Gong, J., Jiang, Y., Qi, R., Wang, C., Fu, H., Ma, Q., Xu, D.: scgnn is a novel graph neural network framework for single-cell rna-seq analyses. Nature communications **12**(1), 1882 (2021)

[27] Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D.J., Lee, E.B., Shinohara, R.T., Li, M.: Spagcn: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. Nature methods **18**(11), 1342–1351 (2021)

[28] Settles, B.: Active learning literature survey (2009)

[29] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)

[30] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

[31] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A.: Generative adversarial networks: An overview. IEEE signal processing magazine **35**(1), 53–65 (2018)

[32] Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1798–1828 (2013)

[33] Bengio, Y., Courville, A.C., Vincent, P.: Unsupervised feature learning and deep learning: A review and new perspectives. CoRR, abs/1206.5538 **1**(2665), 2012 (2012)

[34] Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. In: International Conference on Machine Learning, pp. 1558–1566 (2016). PMLR

[35] Stephenson, E., Reynolds, G., Botting, R.A., Calero-Nieto, F.J., Morgan, M.D., Tuong, Z.K., Bach, K., Sungnak, W., Worlock, K.B., Yoshida, M., *et al.*: Single-cell multi-omics analysis of the immune response in covid-19. Nature medicine **27**(5), 904–916 (2021)

[36] McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)

[37] Hadjadj, J., Yatim, N., Barnabei, L., Corneau, A., Boussier, J., Smith, N., Péré, H., Charbit, B., Bondet, V., Chenevier-Gobeaux, C., *et al.*: Impaired type i interferon activity and inflammatory responses in severe covid-19 patients. Science **369**(6504), 718–724 (2020)

[38] Ramasamy, S., Subbian, S.: Critical determinants of cytokine storm and type i interferon response in covid-19 pathogenesis. Clinical microbiology reviews **34**(3), 10–1128 (2021)

[39] Lee, J.S., Park, S., Jeong, H.W., Ahn, J.Y., Choi, S.J., Lee, H., Choi, B., Nam, S.K., Sa, M., Kwon, J.-S., *et al.*: Immunophenotyping of covid-19 and influenza highlights the role of type i interferons in development of severe covid-19. Science immunology **5**(49), 1554 (2020)

[40] Lee, J.S., Shin, E.-C.: The type i interferon response in covid-19: implications for treatment. Nature Reviews Immunology **20**(10), 585–586 (2020)

[41] Pearson, K.: Liii. on lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin philosophical magazine and journal of science **2**(11), 559–572 (1901)

[42] Menden, K., Marouf, M., Oller, S., Dalmia, A., Magruder, D.S., Kloiber, K., Heutink, P., Bonn, S.: Deep learning–based cell composition analysis from tissue expression profiles. Science advances **6**(30), 2619 (2020)

[43] Chen, Y., Wang, Y., Chen, Y., Cheng, Y., Wei, Y., Li, Y., Wang, J., Wei, Y., Chan, T.-F., Li, Y.: Deep autoencoder for interpretable tissue-adaptive deconvolution and cell-type-specific gene analysis. Nature Communications **13**(1), 6735 (2022)

[44] Tsoucas, D., Dong, R., Chen, H., Zhu, Q., Guo, G., Yuan, G.-C.: Accurate estimation of cell-type composition from gene expression data. Nature communications **10**(1), 2975 (2019)

[45] Plaisier, S.B., Taschereau, R., Wong, J.A., Graeber, T.G.: Rank–rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. Nucleic acids research **38**(17), 169–169 (2010)

[46] Thomas, P.D., Ebert, D., Muruganujan, A., Mushayahama, T., Albou, L.-P., Mi, H.: Panther: Making genome-scale phylogenetics accessible to all. Protein Science **31**(1), 8–22 (2022)

[47] Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., Lewis, S., Hub, A., Group, W.P.W.: Amigo: online access to ontology and annotation data. Bioinformatics **25**(2), 288–289 (2009)

[48] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., *et al.*: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences **102**(43), 15545–15550 (2005)

[49] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., *et al.*: Gene ontology: tool for the unification of biology. Nature genetics **25**(1), 25–29 (2000)

[50] Gillespie, M., Jassal, B., Stephan, R., Milacic, M., Rothfels, K., Senff-Ribeiro, A., Griss, J., Sevilla, C., Matthews, L., Gong, C., *et al.*: The reactome pathway knowledgebase 2022. Nucleic acids research **50**(D1), 687–692 (2022)

[51] Nishida-Aoki, N., Gujral, T.S.: Emerging approaches to study cell–cell interactions in tumor microenvironment. Oncotarget **10**(7), 785 (2019)

[52] Armingol, E., Officer, A., Harismendy, O., Lewis, N.E.: Deciphering cell–cell interactions and communication from gene expression. Nature Reviews Genetics **22**(2), 71–88 (2021)

[53] Joanito, I., Wirapati, P., Zhao, N., Nawaz, Z., Yeo, G., Lee, F., Eng, C.L., Macalinao, D.C.,

Kahraman, M., Srinivasan, H., *et al.*: Single-cell and bulk transcriptome sequencing identifies two epithelial tumor cell states and refines the consensus molecular classification of colorectal cancer. Nature genetics **54**(7), 963–975 (2022)

[54] Gao, J., Wu, Z., Zhao, M., Zhang, R., Li, M., Sun, D., Cheng, H., Qi, X., Shen, Y., Xu, Q., *et al.*: Allosteric inhibition reveals shp2-mediated tumor immunosuppression in colon cancer by single-cell transcriptomics. Acta Pharmaceutica Sinica B **12**(1), 149–166 (2022)

[55] Cao, Y., Tang, W., Tang, W.: Immune cell infiltration characteristics and related core genes in lupus nephritis: results from bioinformatic analysis. BMC immunology **20**, 1–12 (2019)

[56] Fortier, A.-M., Asselin, E., Cadrin, M.: Keratin 8 and 18 loss in epithelial cancer cells increases collective cell migration and cisplatin sensitivity through claudin1 up-regulation. Journal of Biological Chemistry **288**(16), 11555–11571 (2013)

[57] Karantza, V.: Keratins in health and cancer: more than mere epithelial cell markers. Oncogene **30**(2), 127–138 (2011)

[58] Mann, H.B., Whitney, D.R.: On a test of whether one of two random variables is stochastically larger than the other. The annals of mathematical statistics, 50–60 (1947)

[59] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., *et al.*: Scipy 1.0: fundamental algorithms for scientific computing in python. Nature methods **17**(3), 261–272 (2020)

[60] Ramaswami, G., Yuva-Aydemir, Y., Akerberg, B., Matthews, B., Williams, J., Golczer, G., Huang, J., Huh, D., Burkly, L.C., Engle, S.J., et al.: Transcriptional characterization of ipsc-derived microglia as a model for therapeutic development in neurodegeneration. bioRxiv, 2023–03 (2023)

[61] Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., *et al.*: Eleven grand challenges in single-cell data science. Genome biology **21**(1), 1–35 (2020)

[62] Pratt, J.W.: Remarks on zeros and ties in the wilcoxon signed rank procedures. Journal of the American Statistical Association **54**(287), 655–667 (1959)

[63] Wolf, F.A., Hamey, F.K., Plass, M., Solana, J., Dahlin, J.S., Göttgens, B., Rajewsky, N., Simon, L., Theis, F.J.: Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome biology **20**, 1–9 (2019)

[64] Nguyen, V.-L., Shaker, M.H., Hüllermeier, E.: How to measure uncertainty in uncertainty sampling for active learning. Machine Learning **111**(1), 89–122 (2022)

[65] Lewis, D.D.: A sequential algorithm for training text classifiers: Corrigendum and additional data. In: Acm Sigir Forum, vol. 29, pp. 13–19 (1995). ACM New York, NY, USA

[66] Dasgupta, S.: Two faces of active learning. Theoretical computer science **412**(19), 1767–1781 (2011)

[67] Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. Journal of the Royal Statistical Society Series B: Statistical Methodology **61**(3), 611–622 (1999)

[68] Halko, N., Martinsson, P.-G., Tropp, J.A.: Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM review **53**(2), 217–288 (2011)

[69] Martinsson, P.-G., Rokhlin, V., Tygert, M.: A randomized algorithm for the decomposition of matrices. Applied and Computational Harmonic Analysis **30**(1), 47–68 (2011)

[70] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M.,

Prettenhofer, P., Weiss, R., Dubourg, V., *et al.*: Scikit-learn: Machine learning in python. the Journal of machine Learning research **12**, 2825–2830 (2011)

[71] MacQueen, J., *et al.*: Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 281–297 (1967). Oakland, CA, USA

[72] Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M., *et al.*: Mast: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell rna sequencing data. Genome biology **16**(1), 1–13 (2015)

[73] Shalek, A.K., Satija, R., Shuga, J., Trombetta, J.J., Gennert, D., Lu, D., Chen, P., Gertner, R.S., Gaublomme, J.T., Yosef, N., *et al.*: Single-cell rna-seq reveals dynamic paracrine control of cellular variation. Nature **510**(7505), 363–369 (2014)

[74] Hasanaj, E., Wang, J., Sarathi, A., Ding, J., Bar-Joseph, Z.: Interactive single-cell data analysis using cellar. Nature communications **13**(1), 1998 (2022)

[75] Wolf, F.A., Angerer, P., Theis, F.J.: Scanpy: large-scale single-cell gene expression data analysis. Genome biology **19**, 1–5 (2018)

[76] Ernst, J., Vainas, O., Harbison, C.T., Simon, I., Bar-Joseph, Z.: Reconstructing dynamic regulatory maps. Molecular systems biology **3**(1), 74 (2007)

[77] Li, D., Ding, J., Bar-Joseph, Z.: Unifan: A tool for unsupervised single-cell clustering and annotation. Journal of Computational Biology **29**(11), 1229–1232 (2022)

[78] Lopez, R., Regier, J., Cole, M.B., Jordan, M.I., Yosef, N.: Deep generative modeling for single-cell transcriptomics. Nature methods **15**(12), 1053–1058 (2018)

[79] Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., Vert, J.-P.: A general and flexible method for signal extraction from single-cell rna-seq data. Nature communications **9**(1), 284 (2018)

[80] Berge, K., Perraudeau, F., Soneson, C., Love, M.I., Risso, D., Vert, J.-P., Robinson, M.D., Dudoit, S., Clement, L.: Observation weights unlock bulk rna-seq tools for zero inflation and single-cell applications. Genome biology **19**, 1–17 (2018)

[81] Kullback, S., Leibler, R.A.: On information and sufficiency. The annals of mathematical statistics **22**(1), 79–86 (1951)

[82] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., Philip, S.Y.: A comprehensive survey on graph neural networks. IEEE transactions on neural networks and learning systems **32**(1), 4–24 (2020)

[83] Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. IEEE transactions on neural networks **20**(1), 61–80 (2008)

[84] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)

[85] Maas, A.L., Hannun, A.Y., Ng, A.Y., *et al.*: Rectifier nonlinearities improve neural network acoustic models. In: Proc. Icml, vol. 30, p. 3 (2013). Atlanta, GA

[86] He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034 (2015)

[87] Kleyman, M., Sefer, E., Nicola, T., Espinoza, C., Chhabra, D., Hagood, J.S., Kaminski, N., Ambalavanan, N., Bar-Joseph, Z.: Selecting the most appropriate time points to profile in high-throughput studies. Elife **6**, 18541 (2017)

[88] Fang, Z., Liu, X., Peltz, G.: Gseapy: a comprehensive package for performing gene set enrichment analysis in python. Bioinformatics **39**(1), 757 (2023)

[89] Jin, S., Guerrero-Juarez, C., Zhang, L., Chang, I., Ramos, R., Kuan, C., Myung, P., Plikus, M., Nie, Q.: Inference and analysis of cell-cell communication using CellChat. Nat. Commun. 12, 1088 (2021)

[90] Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D.M., Hill, A.J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F.J., *et al.*: The single-cell transcriptional landscape of mammalian organogenesis. Nature **566**(7745), 496–502 (2019)