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;

Supplementary Figures

Runtime

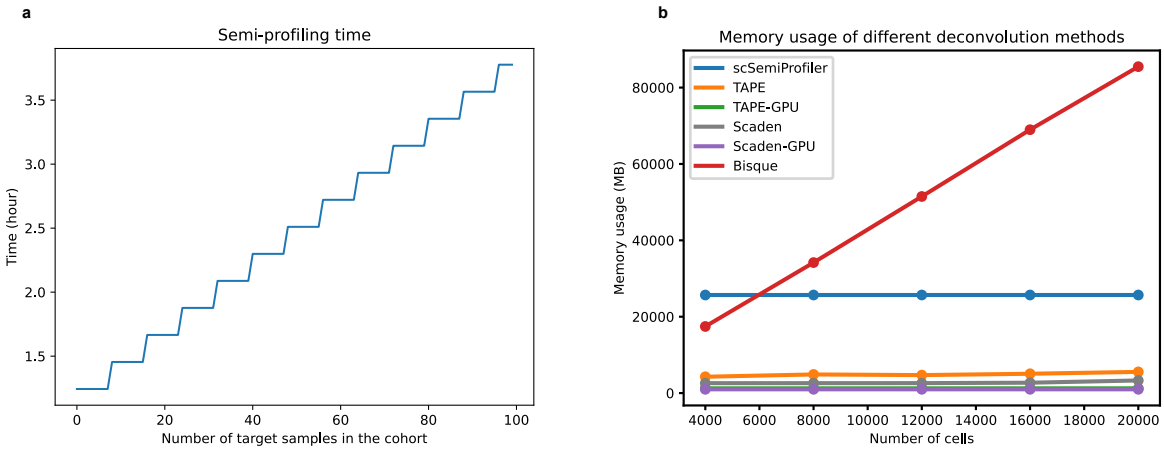

Fig. S1 Runtime and memory usage. **a**, Estimated runtime of scSemiProfiler when 4 representatives are sequenced and 8 GPUs (4 Tesla M10 and 4 NVIDIA A16 are used for semi-profiling the 3 cohorts in our study) are used. The x-axis represents the number of target samples in the cohort, and the y-axis indicates time in hours. **b**, Comparison of memory usage among various deconvolution methods. Memory usage was monitored across deconvolution methods when using 1 to 5 samples (each has 5000 cells) as single-cell references for deconvoluting 4 bulk samples. The cell data matrix is sampled from the COVID-19 cohort and contains 6030 gene features. Most deep learning methods have very low memory usage and only increase slightly for reading larger input data matrices. Our method's memory usage does not increase with the number of samples, as different samples are trained using separate models sequentially. Bisque exhibits nearly linear memory usage. CIBERSORTx, not being open-source software, offers its service exclusively through a website. Therefore, testing the exact memory usage is not feasible. The website permits a maximum upload of 1GB data per user, approximately equivalent to 42826 cells' gene expression data stored in TXT format, as required by the website.

COVID-19 dataset supplementary figures

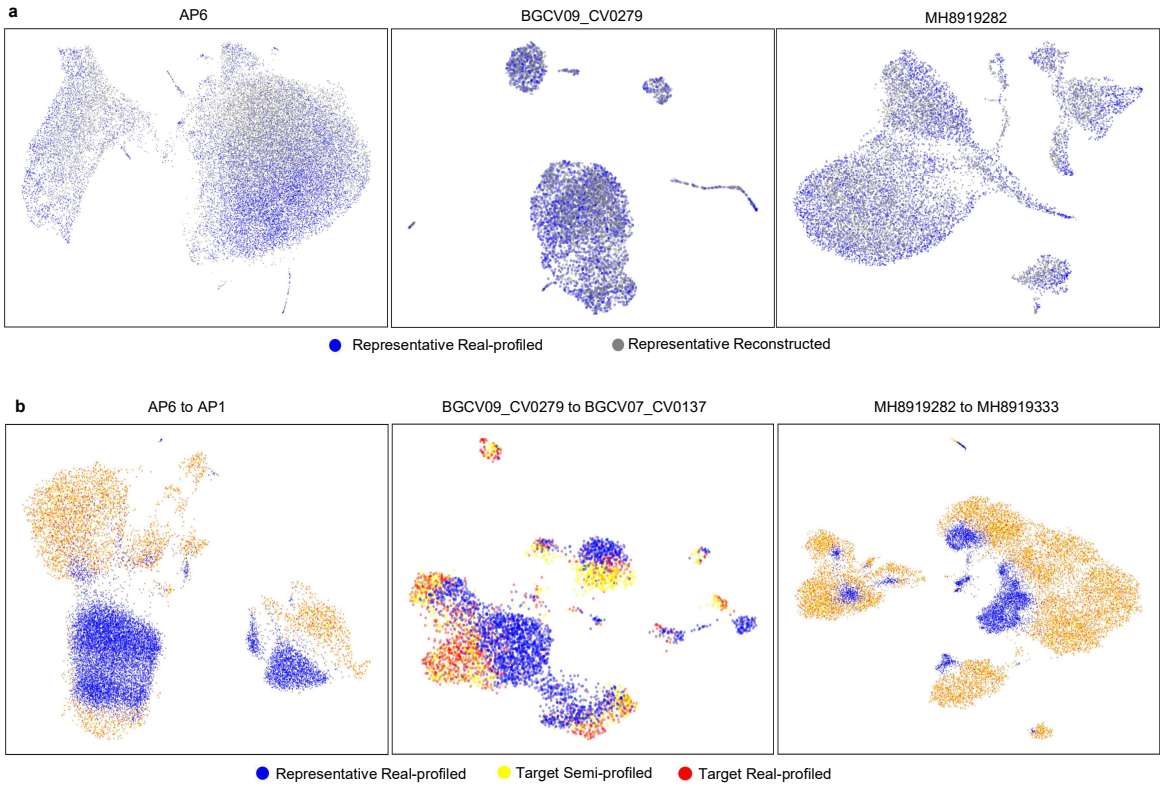

Fig. S2 Examples of the in silico single-cell data inference for target samples in the COVID-19 cohort. The first row shows the reconstruction of the single-cell data for 3 example representatives. The UMAP plots in the second row show the cell distribution of the representative, inferred target sample, and target sample ground truth single-cell data.

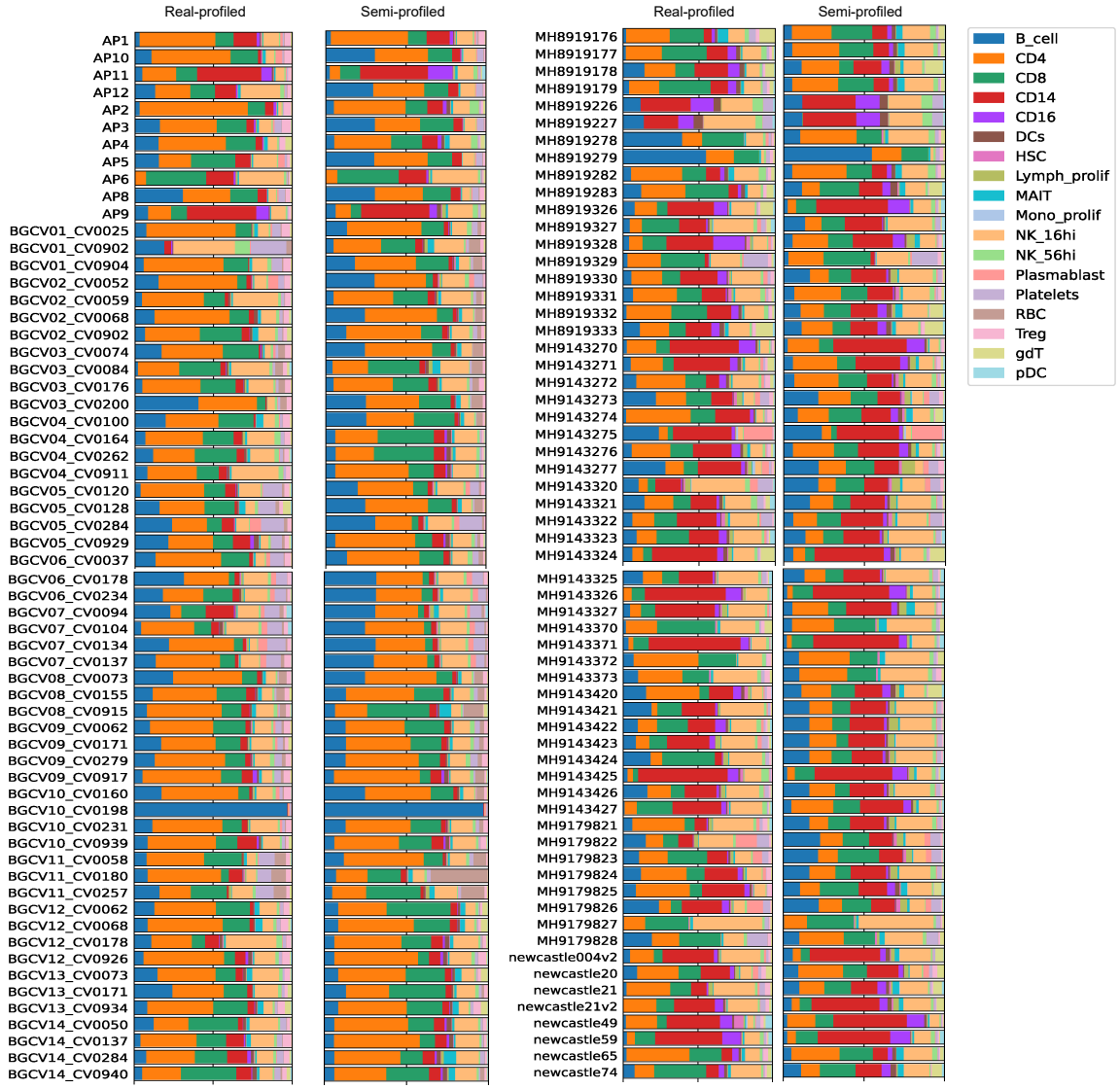

Fig. S3 Deconvolution results for all samples in the COVID-19 cohort. The average Pearson Correlation between the real-profiled and semi-profiled versions is 0.908, and the RMSE (Root Mean Square Error) is 0.146.

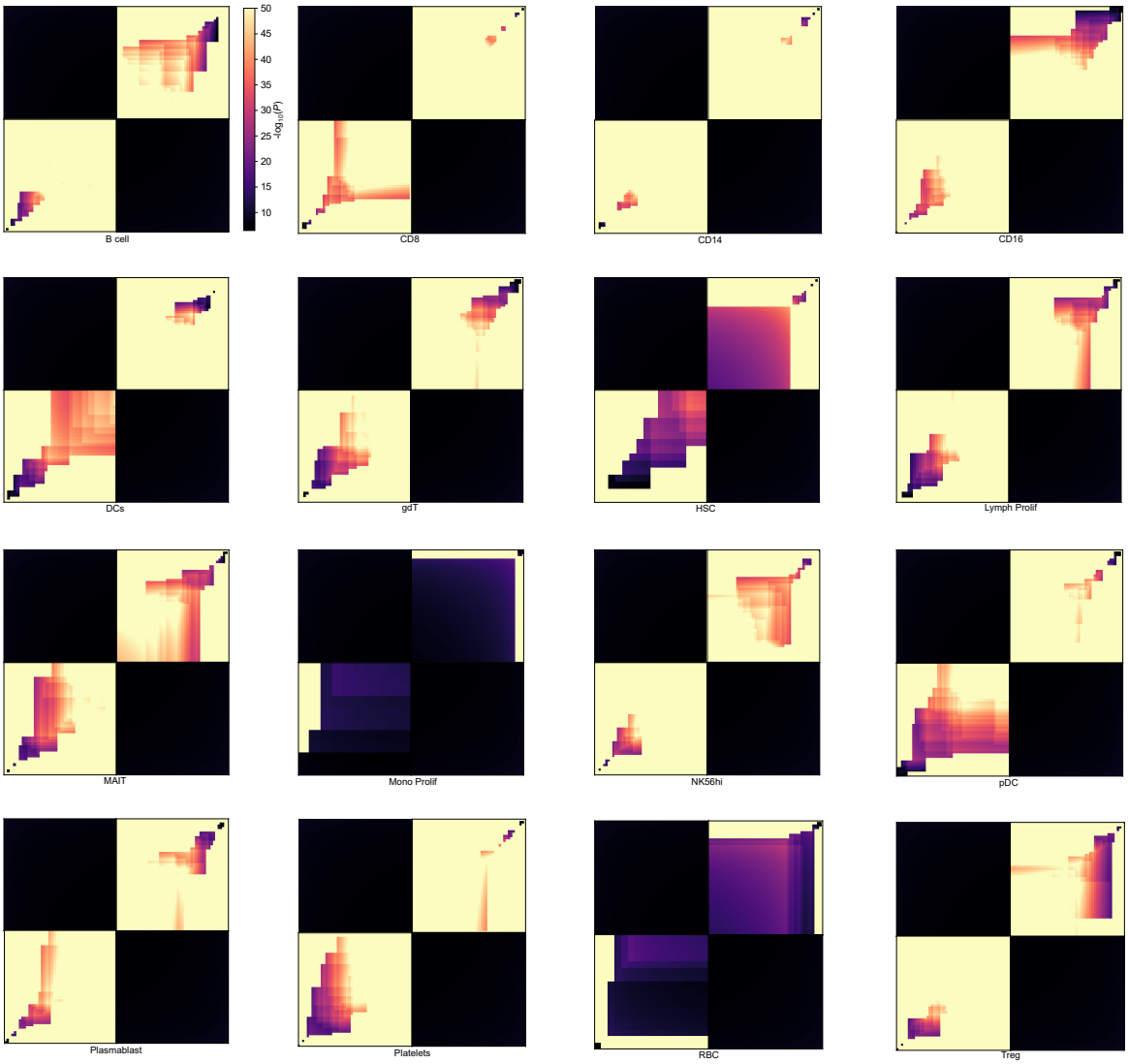

Fig. S4 The Rank–rank hypergeometric overlap (RRHO) plots for other cell types in the COVID-19 dataset. An RRHO plot visualizes the overlap between two ranked gene lists, highlighting the degree of similarity and the significance of the overlap between them. See more details about the RRHO plot in the Methods section. The top 50 positive and negative markers, identified using real-profiled and semi-profiled datasets, are utilized for the plots.

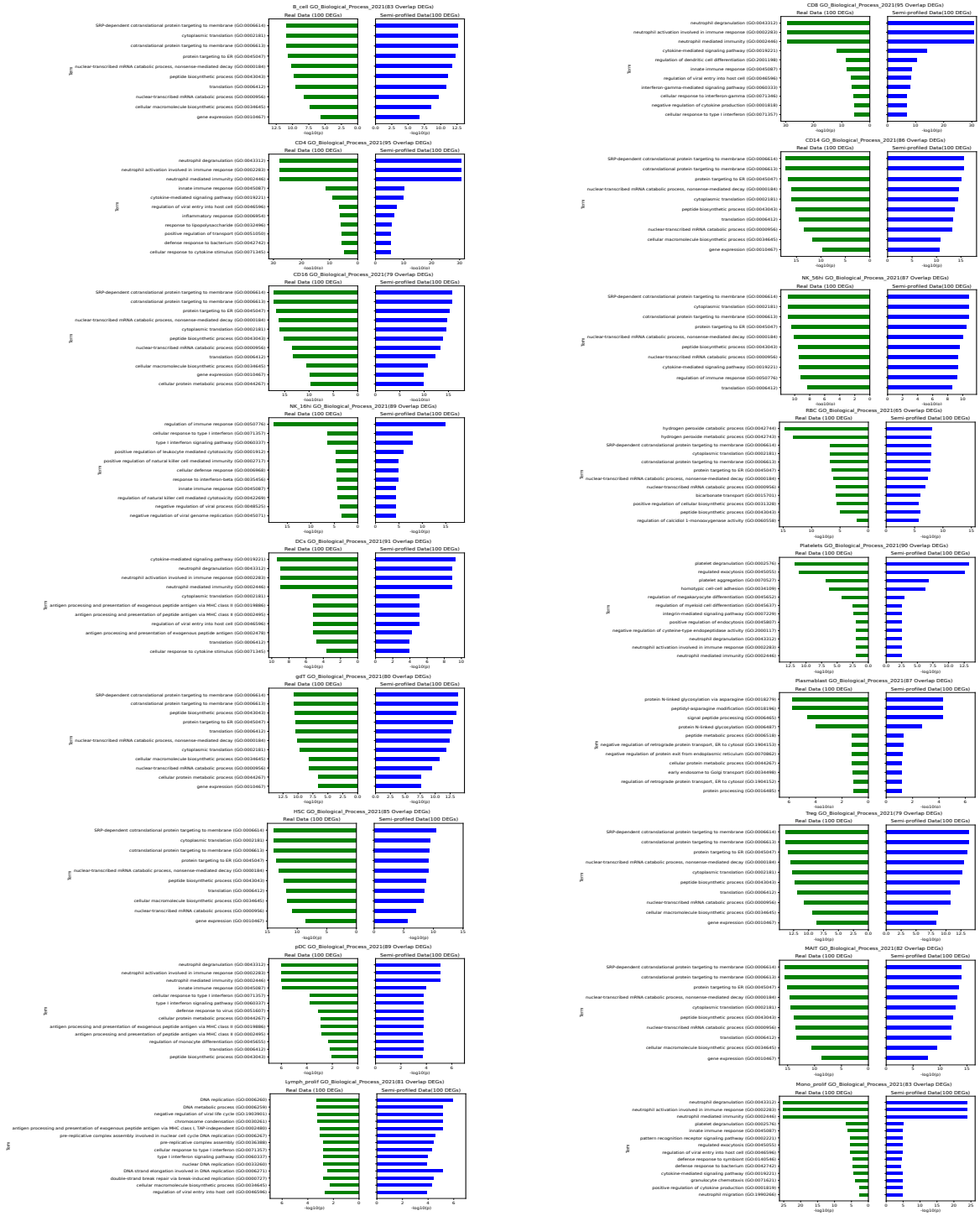

Fig. S5 GO enrichment analysis for additional cell types in the COVID-19 dataset. The top 100 signature genes are used for the analysis. The Pearson Correlations between the negative logged p-values (bar lengths) in the real-profiled and semi-profiled versions for each cell type are as follows: B_cell: 0.999, CD8: 0.999, CD14: 0.980, CD16: 0.994, DCs: 0.987, gdT: 0.985, HSC: 0.964, Lymph_prolif: 0.592, MAIT: 0.999, Mono_prolif: 0.993, NK_16hi: 0.974, NK_56hi: 0.996, pDC: 0.839, Plasmablast: 0.982, Platelets: 0.986, RBC: 0.656, Treg: 0.984

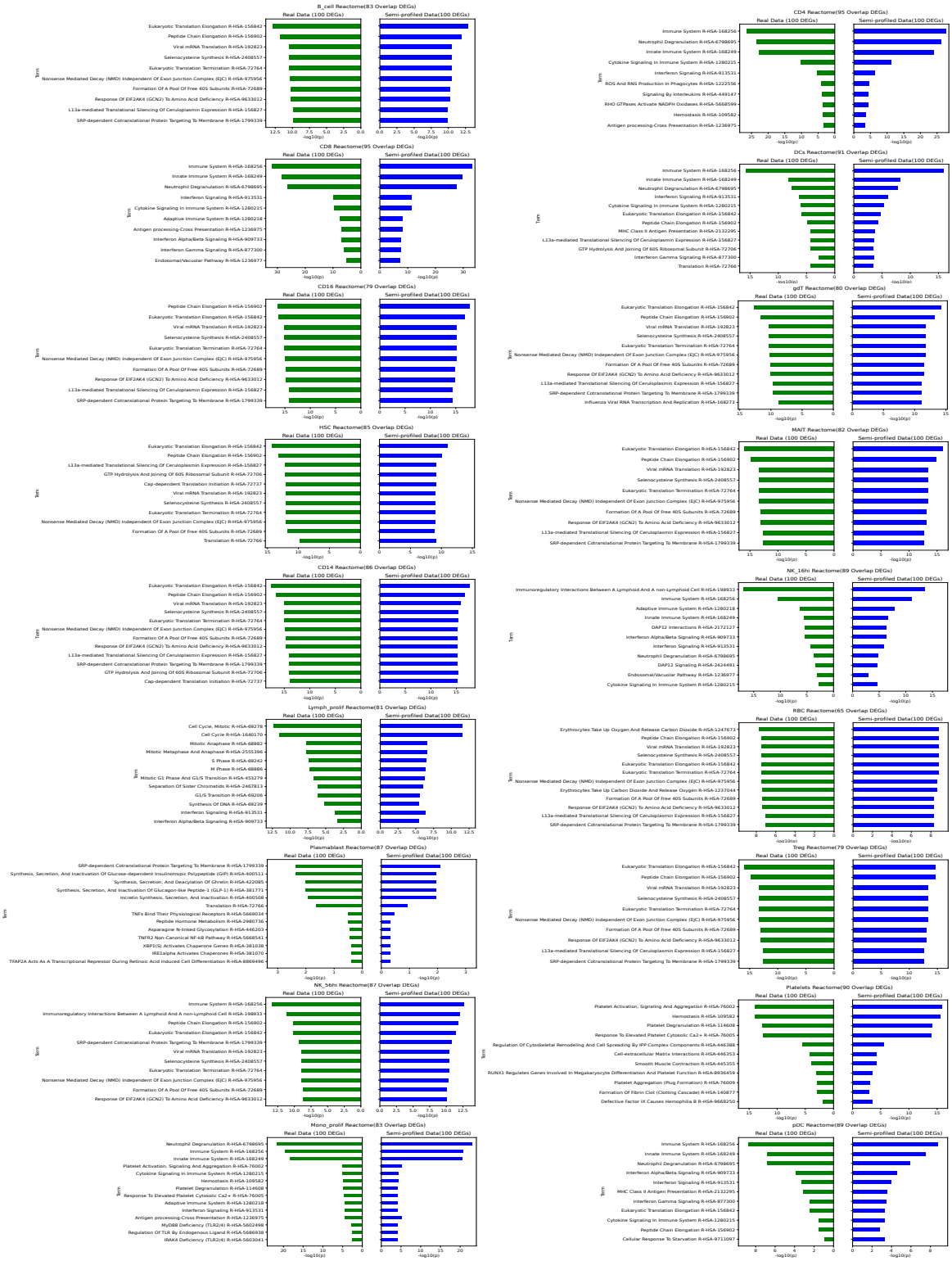

Fig. S6 Reactome enrichment analysis for additional cell types in the COVID-19 dataset. The top 100 signature genes are used for the analysis. The Pearson Correlations between the negative logged p-values (bar lengths) in the real-profiled and semi-profiled versions for each cell type are as follows: B_cell: 0.999, CD4:0.999, CD8: 0.999, CD14: 0.942, CD16: 0.963, DCs: 0.989, gdT: 0.961, HSC: 0.754, Lymph_prolif: 0.891, MAIT: 0.999, Mono_prolif: 0.993, NK_16hi: 0.968, NK_56hi: 0.952, pDC: 0.962, Plasmablast: 0.974, Platelets: 0.996, RBC: 0.767, Treg: 0.961

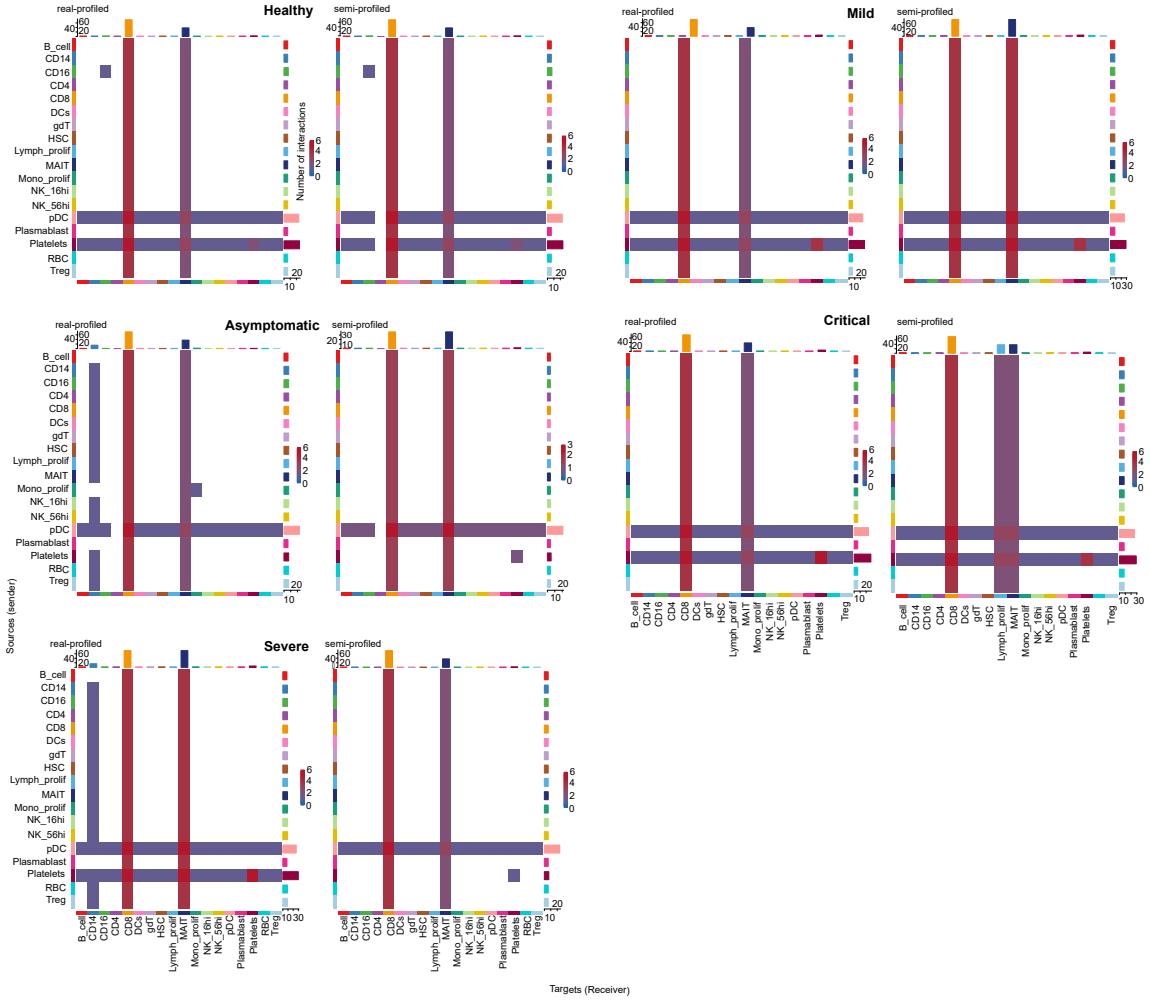

Fig. S7 Comparison of cell-cell interaction analysis results across all severity levels in the COVID-19 dataset. Pearson Correlations between the real and semi-profiled versions for each severity level are as follows: Healthy: 0.997, Asymptomatic: 0.915, Mild: 0.946, Moderate: 0.996, Severe: 0.908, Critical: 0.913.

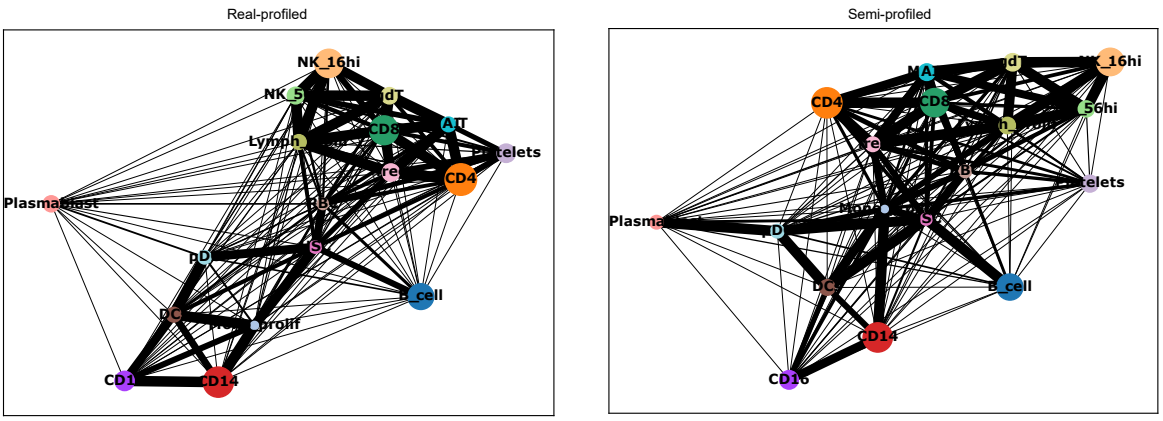

Fig. S8 Comparison of PAGA results for the COVID-19 dataset. The Pearson Correlation between the connectivity matrices of the PAGA plots derived from the real and semi-profiled datasets is 0.893.

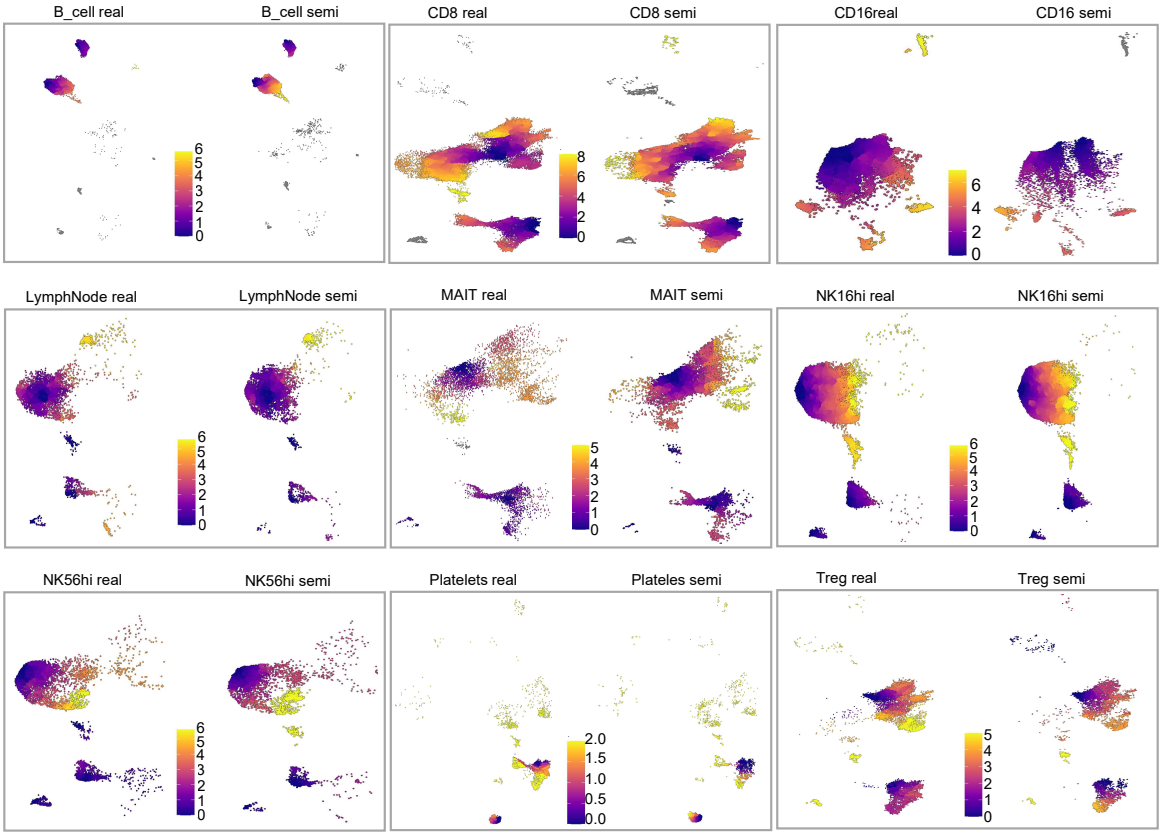

Fig. S9 Pseudotime analysis results comparison for other cell types in the COVID-19 dataset. Cell types with insufficient cells to form a cluster structure were excluded. The p-values of the Mann-Whitney U test between the real and semi-profiled versions for each cell type are: B_cell: $< 2.23 \times 10^{-308}$, CD8: 2.97×10^{-111}, CD16: 2.00×10^{-110}, Lymph_prolif: 3.71×10^{-222}, MAIT: 2.13×10^{-72}, NK_16hi: $< 2.23 \times 10^{-308}$, NK_56hi: 4.63×10^{-22}, Platelets: 4.50×10^{-176}, Treg: 3.76×10^{-111}

Colorectal cancer dataset supplementary figures

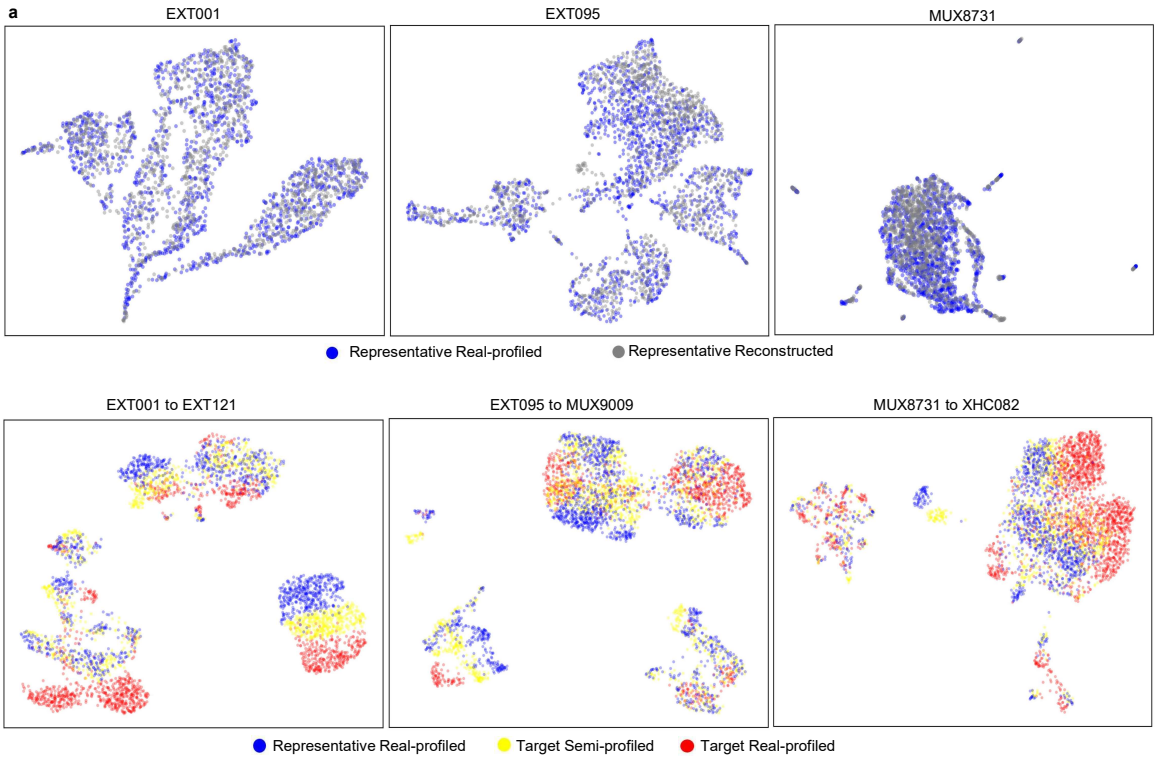

Fig. S10 Examples of the in silico single-cell data inference for target samples in the colorectal cancer dataset. The first row shows the reconstruction of the corresponding representatives. The second row shows the cell distribution of the representative, inferred target sample, and target sample ground truth.

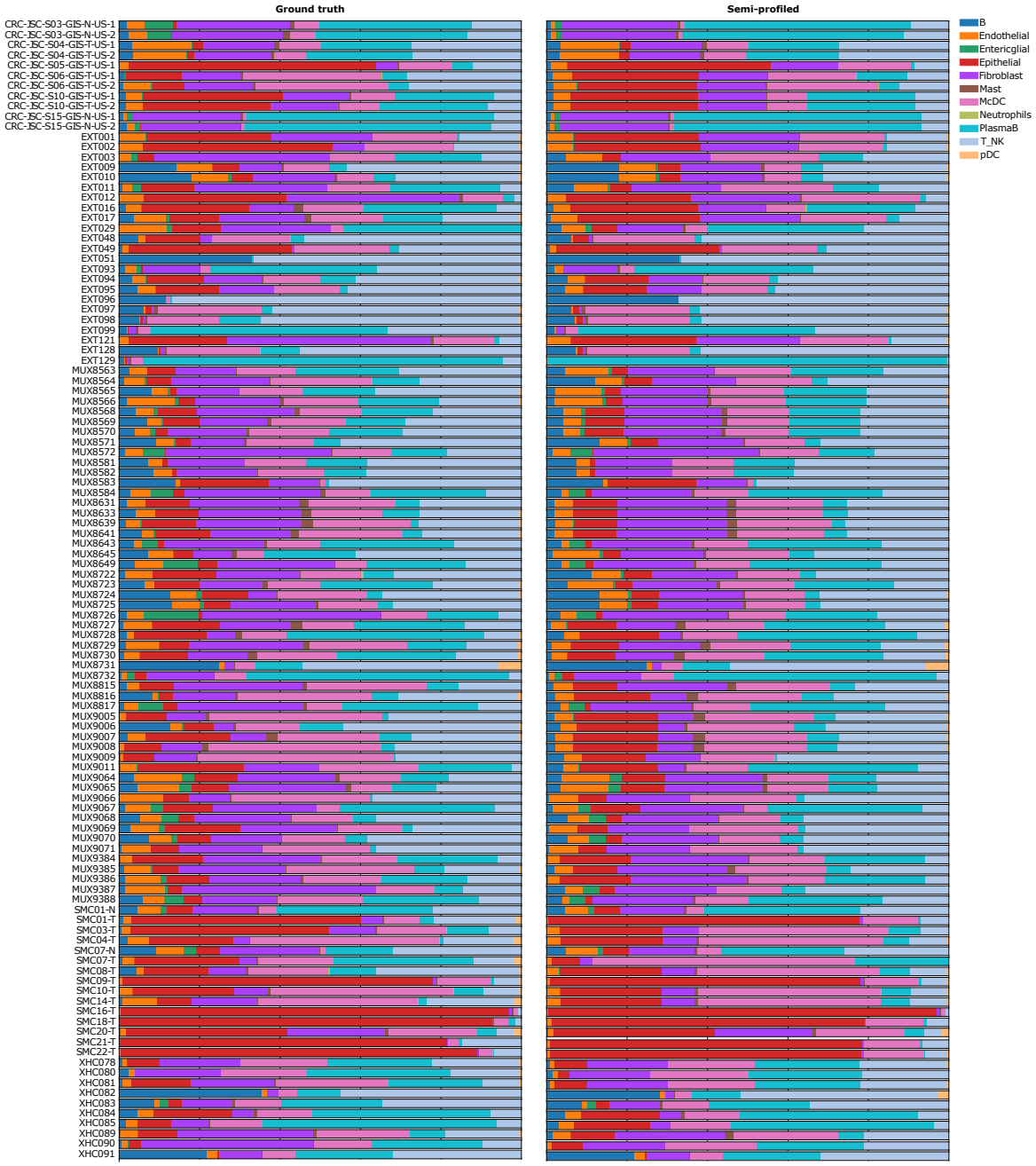

Fig. S11 Deconvolution results for every sample in the colorectal cancer cohort. The average Pearson Correlation between the real and semi-profiled versions is 0.928, and the RMSE is 0.118.

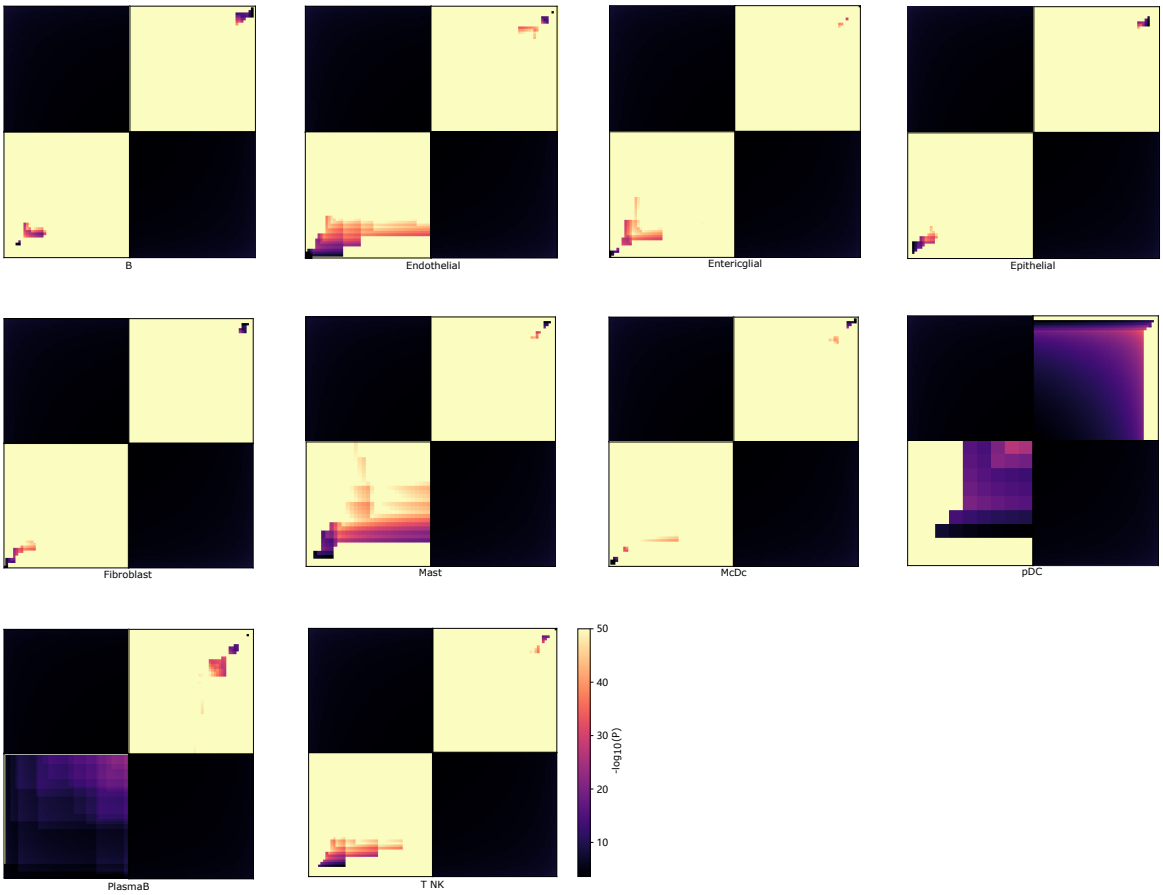

Fig. S12 The RRHO plots for other cell types in the colorectal cancer dataset. The top 50 positive and negative markers, identified using real-profiled and semi-profiled datasets, are utilized for the plots.

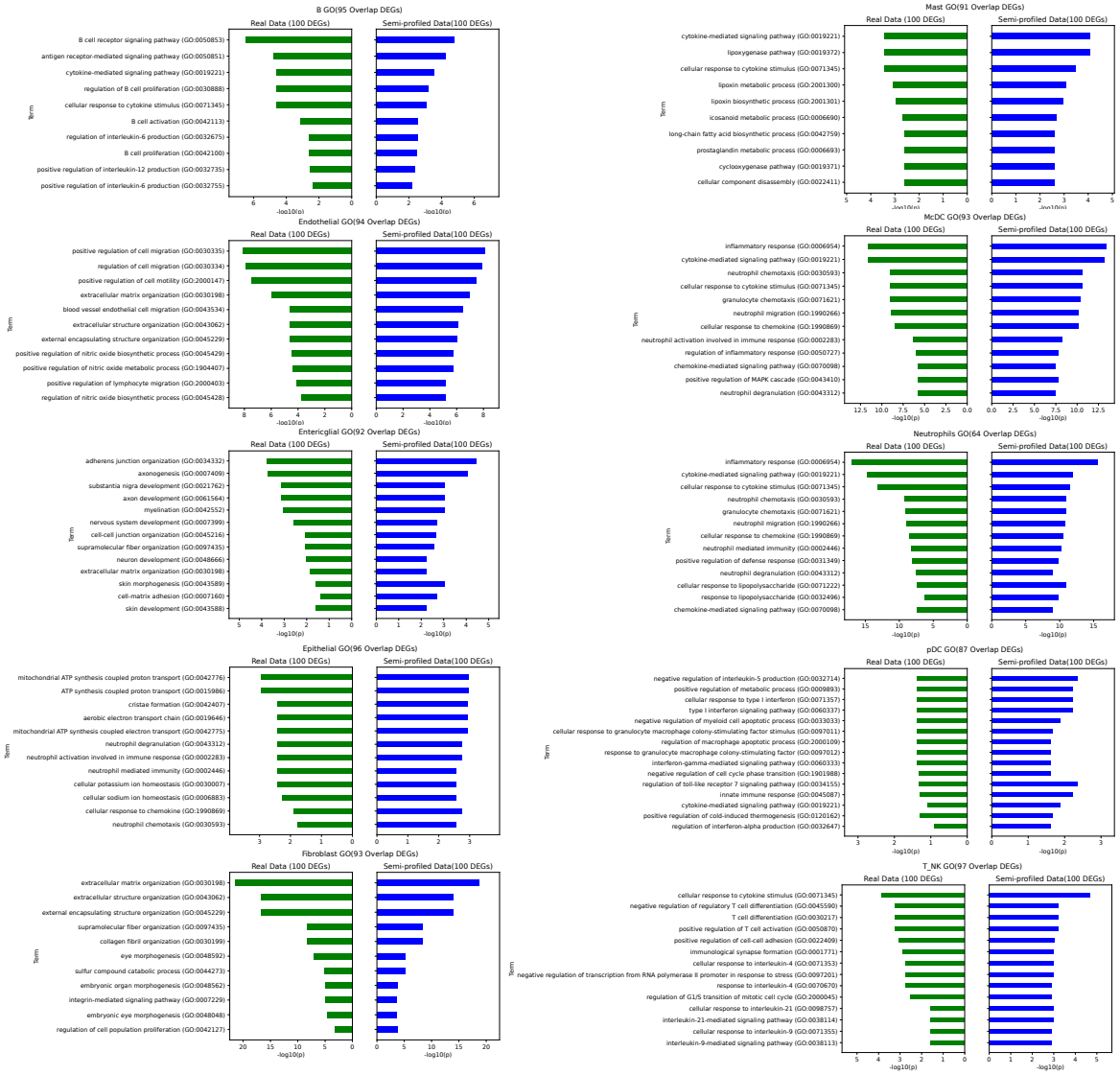

Fig. S13 GO enrichment analysis for additional cell types in the colorectal cancer dataset. The top 100 signature genes are used for the analysis. The Pearson Correlations between the negative logged p-values (bar lengths) in the real-profiled and semi-profiled versions for each cell type are as follows: B: 0.940, Endothelial: 0.969, Entericglial: 0.806, Fibroblast: 0.989, Mast: 0.946, McDC: 0.995, Neutrophils: 0.870, pDC: 0.245, T_NK: 0.603.

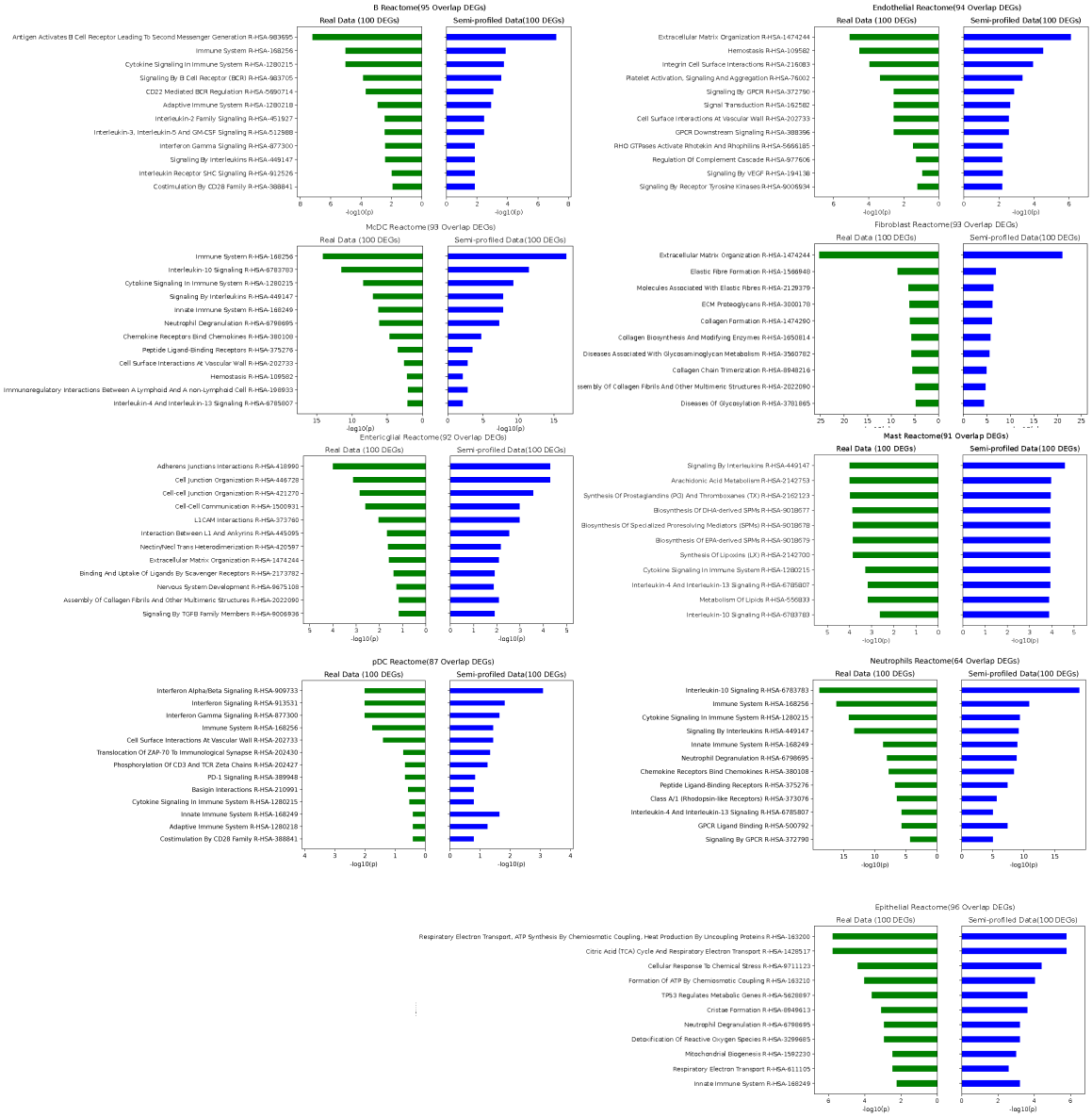

Fig. S14 Reactome enrichment analysis for additional cell types in the colorectal cancer dataset. The top 100 signature genes are used for the analysis. The Pearson Correlations between the negative logged p-values (bar lengths) in the real-profiled and semi-profiled versions for each cell type are as follows: B: 0.959, Endothelial: 0.929, Entericglial: 0.959, Epithelial: 0.996, Fibroblast: 0.997, Mast: 0.368, McDC: 0.990, Neutrophils: 0.864, pDC: 0.697, T_NK: 0.912.

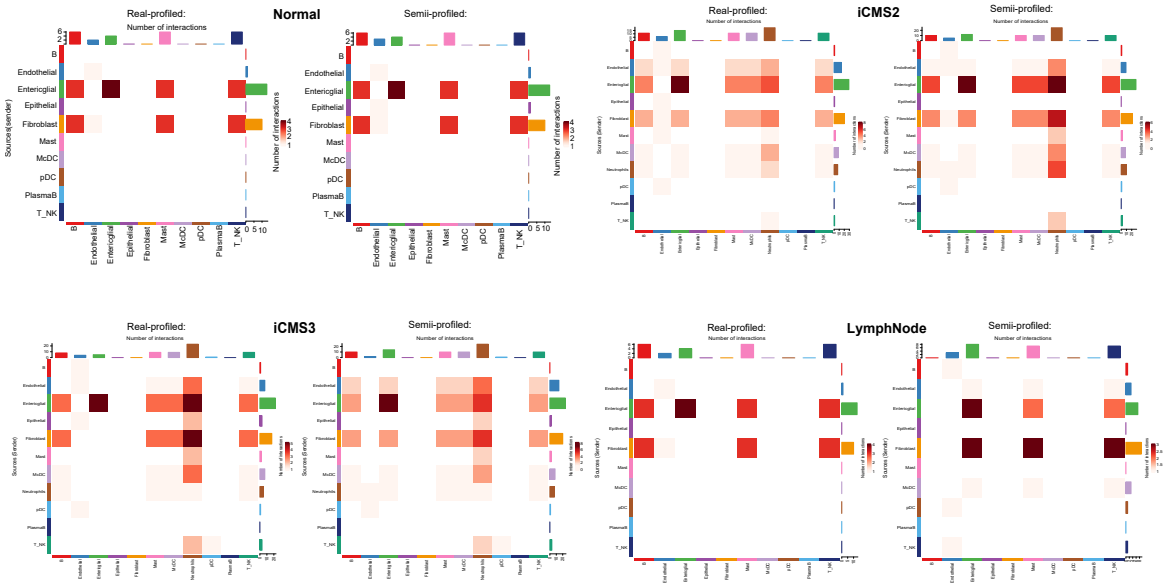

Fig. S15 Cell-cell interaction analysis results comparison for all tissue types in the colorectal cancer dataset. The Pearson Correlations between the real and semi-profiled versions of each tissue type are: Normal: 0.993, LymphNode: 0.700, iCMS2: 0.953, iCMS3: 0.922.

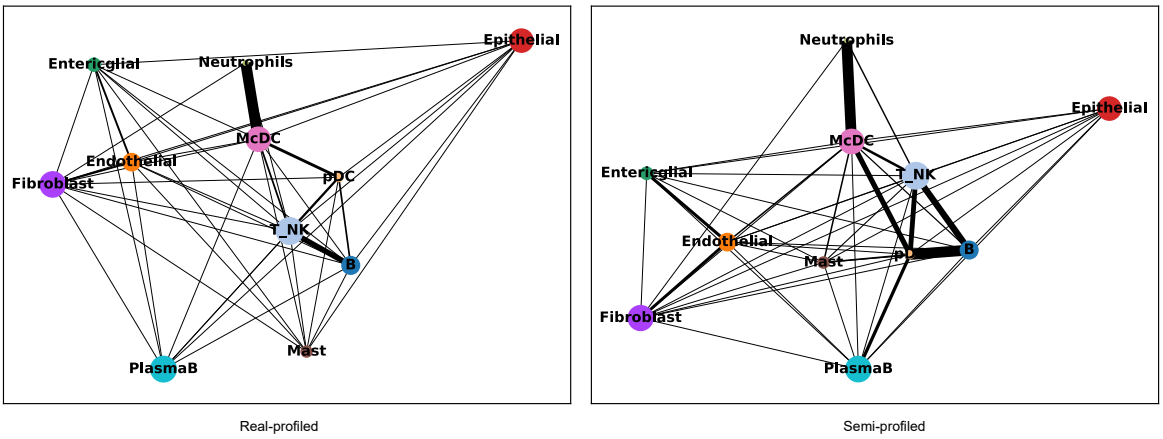

Fig. S16 PAGA results comparison for the colorectal cancer dataset. The Pearson Correlation between the two connectivity matrices is 0.830.

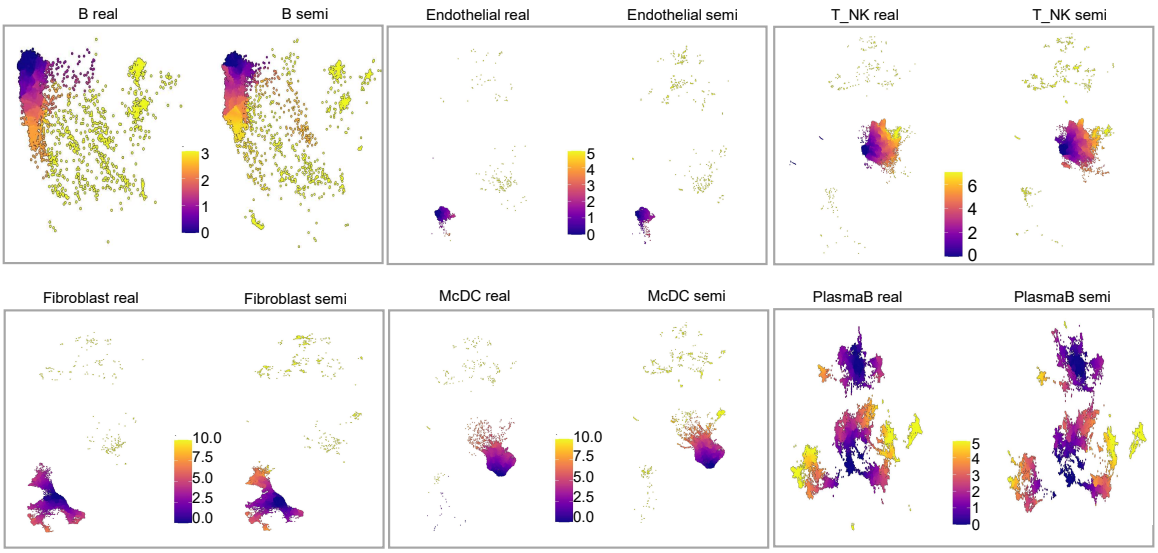

Fig. S17 Pseudotime analysis results comparison for other cell types in the colorectal cancer dataset. Cell types with insufficient cells to form a cluster structure were excluded. The p-values of the Mann-Whitney U test between the real and semi-profiled versions for each cell type are: B_cell: B: 1.37×10^{-101}, Endothelial: 1.66×10^{-35}, T_NK: 5.80×10^{-98}, Fibroblast: 5.71×10^{-253}, McDC: 2.52×10^{-78}, PlasmaB: 5.18×10^{-14}

iMGL dataset supplementary figures

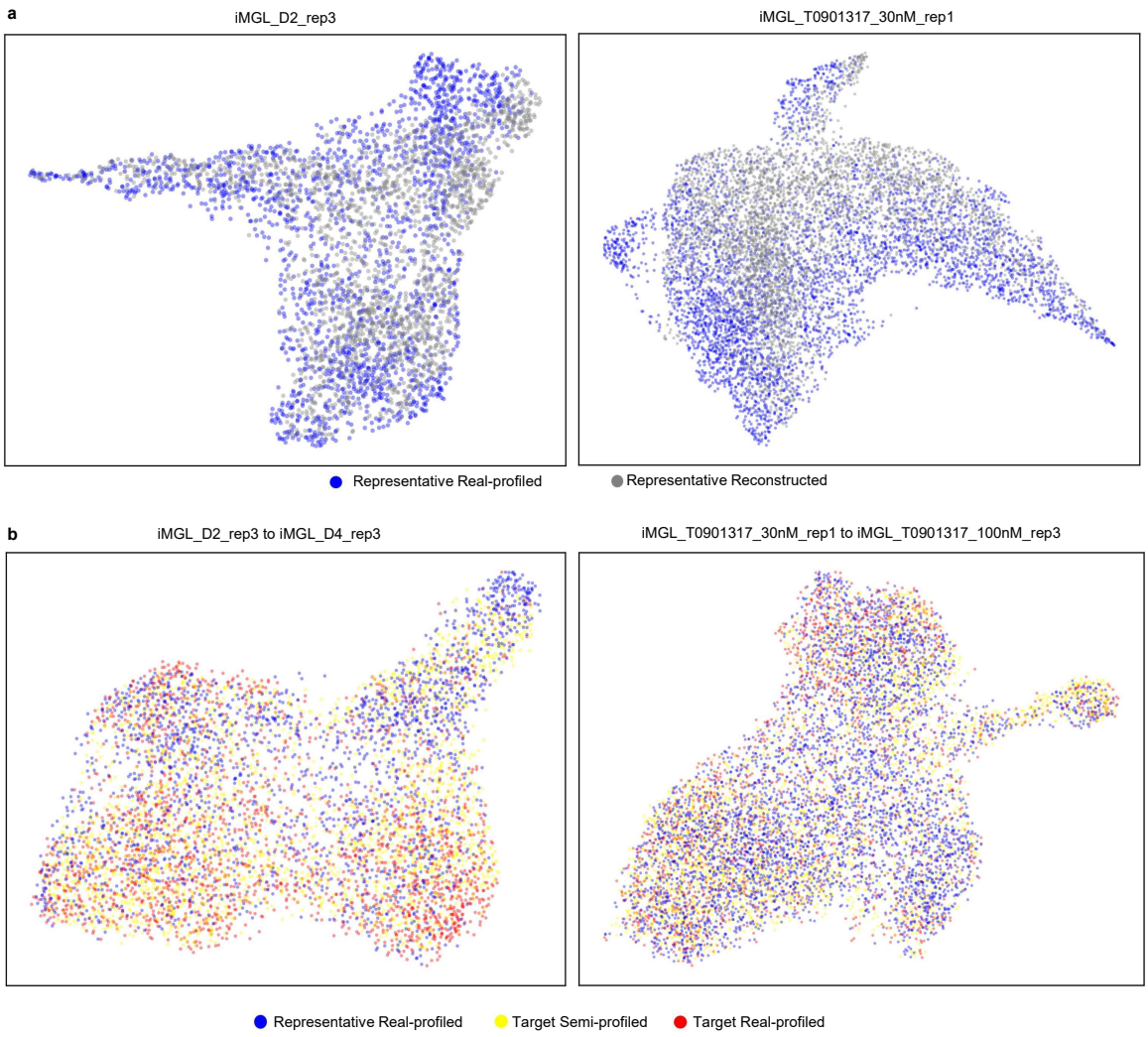

Fig. S18 Examples of the in silico single-cell data inference for target samples in the iMGL cohort. The first row shows the reconstruction of the corresponding representatives. The second row shows the cell distribution of the representative, inferred target sample, and target sample ground truth.

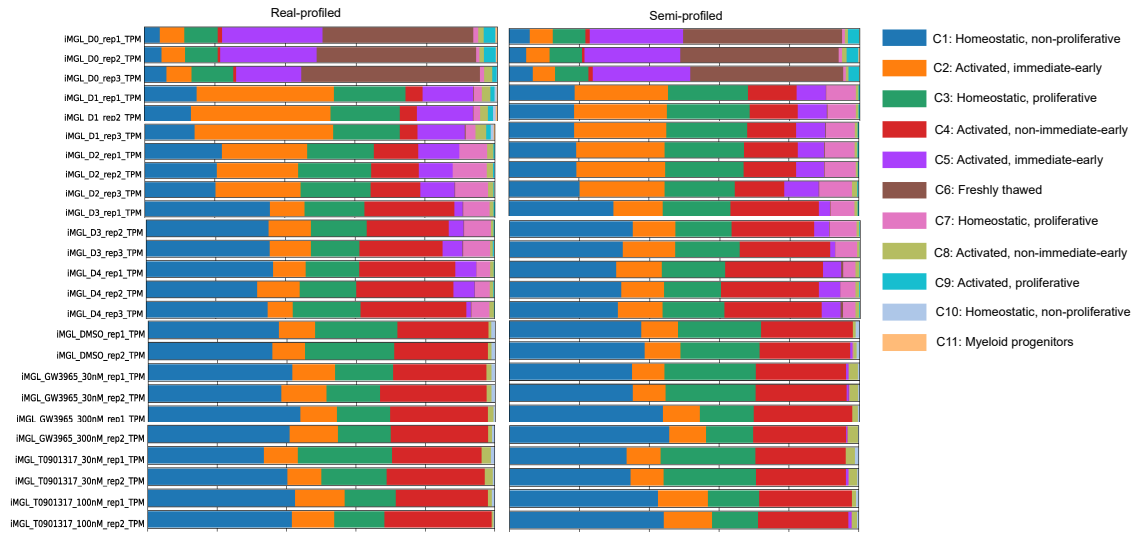

Fig. S19 Deconvolution results for every sample in the iMGL cohort. The average Pearson Correlation is 0.973, and the RMSE is 0.069.

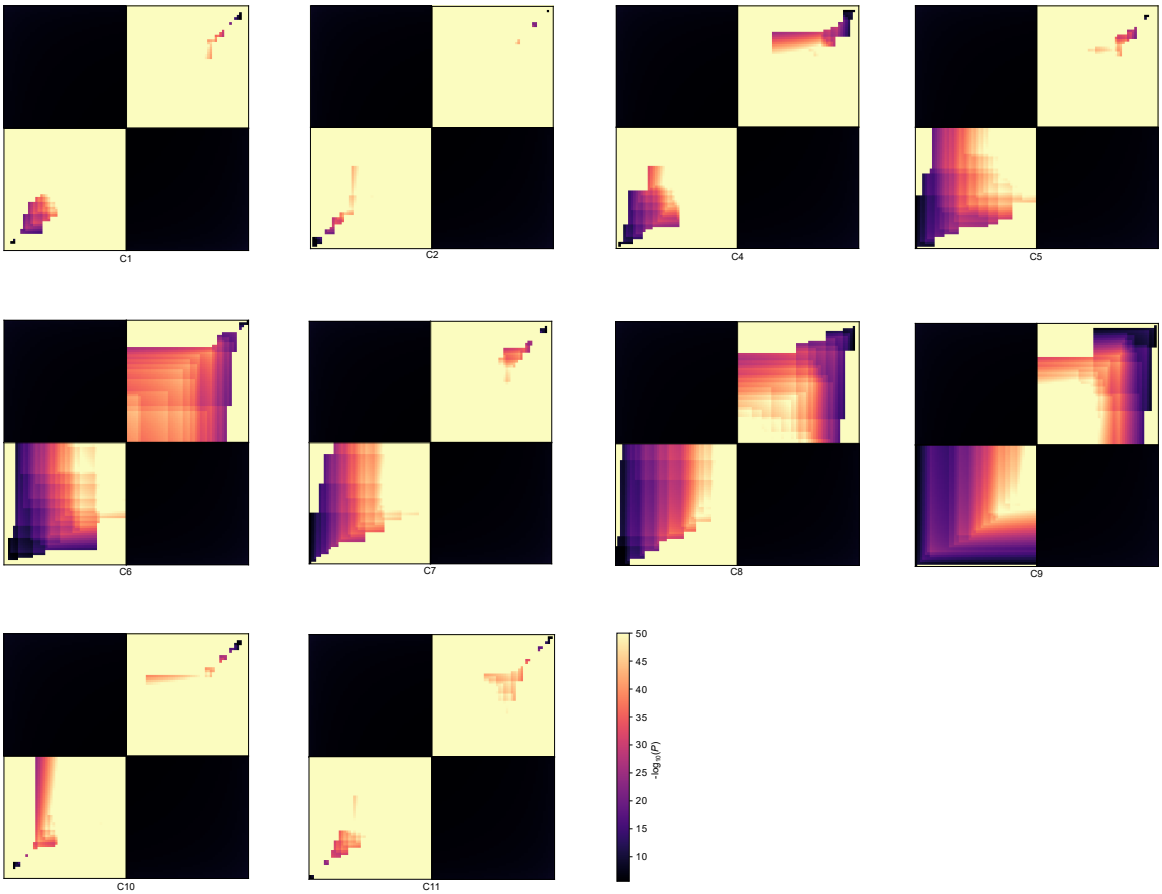

Fig. S20 The RRHO plots for other cell types in the iMGL dataset. The top 50 positive and negative markers, identified using real-profiled and semi-profiled datasets, are utilized for the plots.

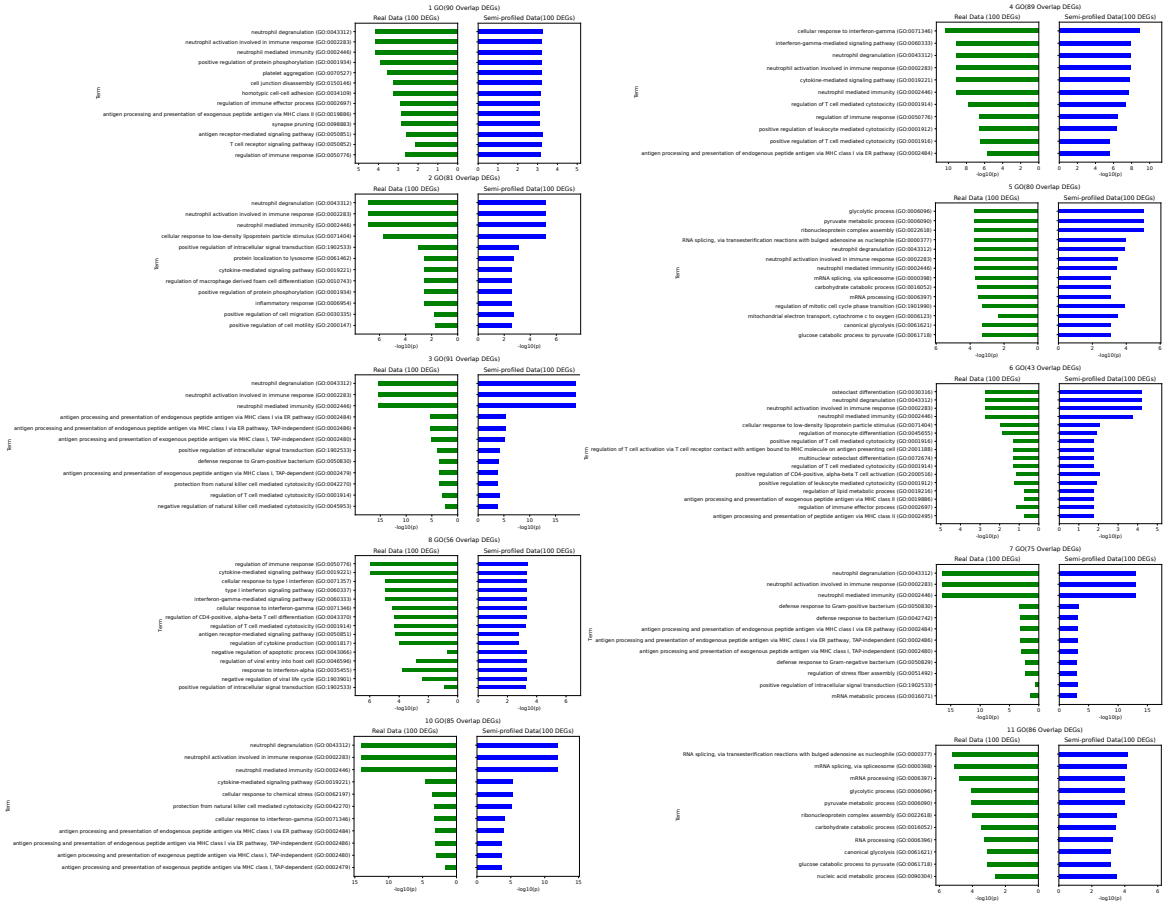

Fig. S21 GO enrichment analysis for additional cell types in the iMGL dataset. The top 100 signature genes are used for the analysis. Cell types without at least 10 GO terms are not included. The Pearson Correlations between the significance level in real-profiled and semi-profiled version for each cell type are: C1 Homeostatic, non-proliferative: 0.352, C2 Activated, immediate-early: 0.978, C3 Homeostatic, proliferative: 0.995, C4 Activated, non-immediate-early: 0.974, C5 Activated, immediate-early: 0.353, C6 Freshly thawed: 0.909, C7 Homeostatic, proliferative: 0.993, C8 Activated, non-immediate-early: 0.04, C9 Activated, proliferative: 0.992, C11 Myloidprogenitors: 0.855

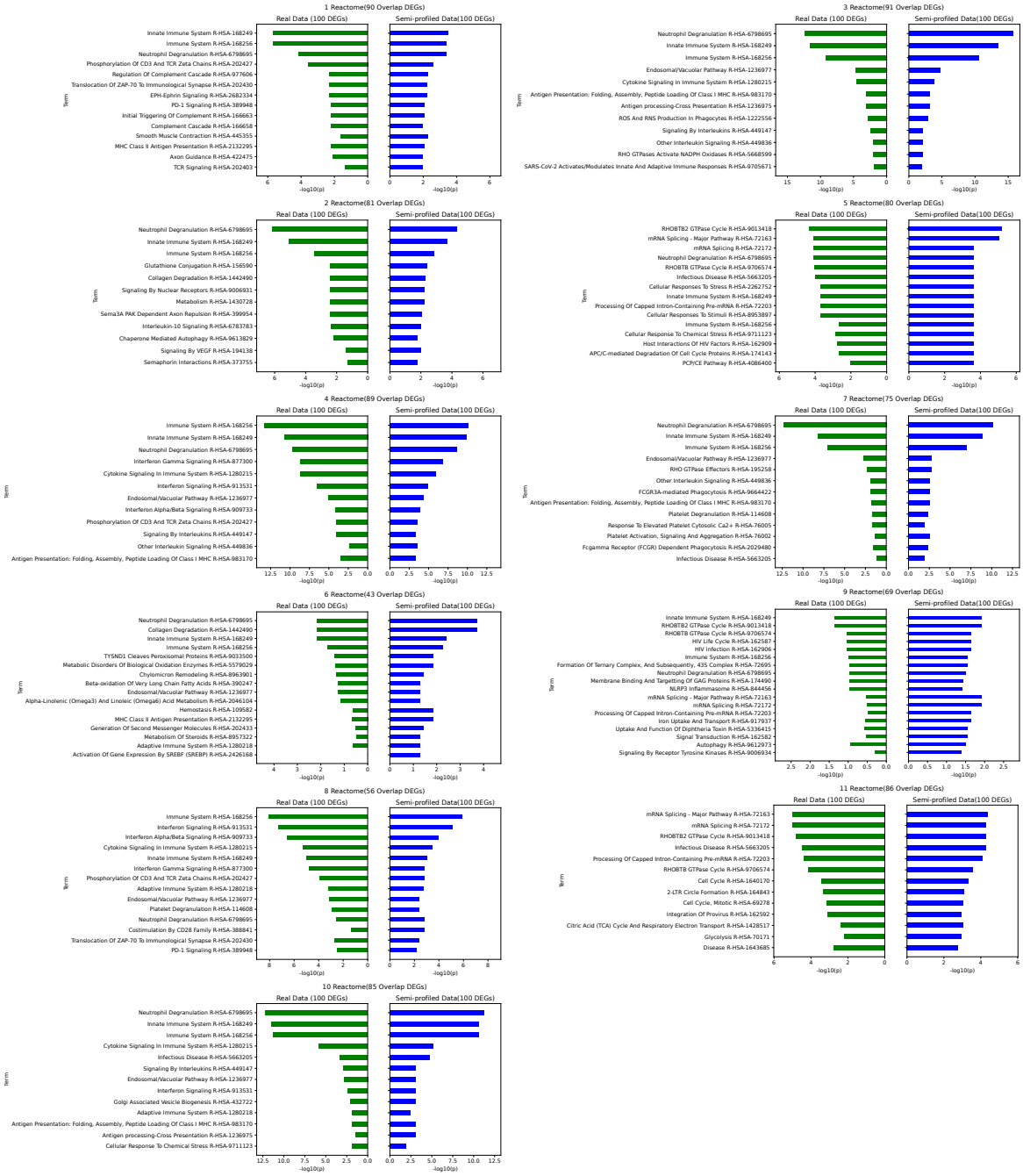

Fig. S22 Reactome enrichment analysis for other cell types in the iMGL dataset. The top 100 signature genes are used for the analysis. The Pearson Correlations between the significance level in real-profiled and semi-profiled version for each cell type are: C1 Homeostatic, non-proliferative: 0.940, C2 Activated, immediate-early: 0.974, C3 Homeostatic, proliferative: 0.995, C4 Activated, non-immediate-early: 0.961, C5 Activated, immediate-early: 0.414, C6 Freshly thawed: 0.755, C7 Homeostatic, proliferative: 0.984, C8 Activated, non-immediate-early: 0.889, C9 Activated, proliferative: 0.199, C10 Homeostatic, non-proliferative: 0.987, C11 Myloidprogenitors: 0.943