CellDiffusion: a generative model to annotate single-cell and spatial RNA-seq using bulk references

Xiaochen Zhang[1], Jiadong Mao[1,&], Kim-Anh Lê Cao[1,&,*]

[1]Melbourne Integrative Genomics, School of Mathematics and Statistics, The University of Melbourne, Australia

& indicates equal contribution

∗ corresponding author: kimanh.lecao@unimelb.edu.au

Abstract

Annotating single-cell and spatial RNA-seq data can be greatly enhanced by leveraging bulk RNA-seq, which remains a cost-effective and well-established benchmark for characterising transcriptional activity in immune cell populations. However, a major technical hurdle lies in the contrasting properties of these data types: single-cell and spatial data are inherently sparse due to its cell-level sampling scheme, leading to much lower sequencing depth compared to bulk RNA-seq.

We developed CellDiffusion, a generative machine learning (ML) tool that bridges this gap. CellDiffusion generates realistic virtual cells to augment the sparse single-cell and spatial data, improving signals and the representation of rare cell types. The augmented data are more comparable to bulk references, increasing the accuracy of cell type annotation using bulk references and automated ML classifiers.

We benchmarked CellDiffusion on single-cell and spatial datasets from human peripheral blood samples, white adipose tissues, and breast tumours. Our method significantly outperforms state-of-the-art methods such as SingleR, Seurat, and scVI. In addition, CellDiffusion provides critical biological insights, including the identification of novel cell subtypes and their function during cell state transition; the discovery of new marker genes for tissue-resident immune cells, revealing their functional shifts in myeloid populations; and the accurate characterisation of cell subtypes in spatial transcriptomics to decipher tumour microenvironment.

1 Introduction

The landscape of transcriptomic research has been fundamentally altered by technologies that profile gene expression at single-cell resolution [1]. Single-cell RNA sequencing (scRNA-seq) enables the high-throughput dissection of complex tissues into their constituent cell types, providing critical insights into cellular heterogeneity, rare cell populations, and pathological states [2]. Complementing this, image-based spatial transcriptomics overcomes the loss of positional information inherent in dissociated-cell methods [3]. By co-registering transcriptomic data with spatial coordinates, these techniques allow for the direct visualisation of gene expression patterns within the native tissue architecture, achieving true single-cell resolution upon computational cell segmentation [4]. Their application offers a comprehensive view of biological processes, from characterising the tumour microenvironment to identifying spatially resolved biomarkers [5].

Realising the full potential of this high-dimensional single-cell resolution data hinges on a critical subsequent step: assigning a precise biological identity to every cell [6]. This process, known as cell type annotation, is fundamental to extracting meaningful insights and has become the essential framework for characterizing tissue composition, elucidating disease mechanisms, and understanding cell-specific responses to therapies [7]. The exponential growth in dataset size and complexity has, in turn, driven a necessary shift from laborious manual annotation to scalable, automated computational approaches [8].

Currently, automated cell type annotation strategies predominantly rely on mapping query datasets to scRNA-seq reference atlases [9]. However, the efficacy of this approach remains critically dependent on both the quality and biological relevance of the selected reference atlas [10]. Generating new, application specific scRNA-seq references is not only prohibitively expensive but also demands extensive manual cell type curation [11]. Furthermore, these references must be precisely tissue matched to the query data, as they generally do not perform well across different tissue types [12]. These limitations are further compounded in spatial transcriptomics applications, where studies have shown that annotation based on scRNA-seq atlas can produce inadequate results due to inherent technical constraints, including low signal-to-noise ratios that compromise accurate and robust cell type

mapping [13].

An alternative strategy is to utilise the vast archives of bulk RNA sequencing data accumulated over decades. The high-throughput nature of bulk RNA-seq technology, coupled with numerous public repositories, renders this approach cost-effective and often eliminates the need for *de novo* reference atlas generation [14]. Furthermore, the higher signal-to-noise ratio and lower sparsity of bulk data make them inherently more robust for challenging tasks such as cross-tissue or cross-platform cell type annotation [15]. For years, these repositories have served as a cornerstone of clinical omics, creating a rich, biologically grounded standard coupled with meticulous clinical and pathological annotations from large patient cohorts [16]. Therefore, the integration of high-resolution single-cell data with established bulk RNA-seq resources offers a powerful analytical framework for aligning novel discoveries within existing clinical knowledge, effectively bridging the translational gap between laboratory findings and established clinical records [17, 18].

Various bioinformatics tools are available for the integration of single-cell resolution and bulk RNA-seq data. Prominent examples include SingleR, which uses correlation-based methods to provide accurate annotations, and Sincast, which enhances the sparse single-cell signal by imputation and pseudobulk aggregation to improve comparability with bulk data [19, 20]. Another strategy, used by methods such as Phi-space, involves projecting both data types into a common lower-dimensional space to enable integrated downstream analysis, including cell type annotation [21].

A key limitation of current methods for integrating single-cell and bulk RNA-seq data is their reliance on statistical methods alone [22]. These approaches never model the physical sampling process that causes the fundamental differences between single-cell and bulk RNA-seq data. The single-cell sequencing experiment is inherently an undersampling process. Due to practical and cost constraints, only a fraction of cells are captured from the tissue with lower sequencing depth [1]. Recent advances in generative artificial intelligence (AI), particularly the development of diffusion models for creating high-quality data, present a powerful opportunity to overcome this limitation [23]. Here, we introduce CellDiffusion, a method that applies a generative model to simulate the transcriptomes of cells likely missed during experimental sampling. By augmenting the query single-cell and spatial RNA-seq

dataset by using virtual cells, CellDiffusion makes it more comparable to comprehensive bulk RNA-seq atlases, thereby providing a more robust and cost-effective approach to cell type annotation. Our method aims to allow researchers to better utilise existing biological knowledge from bulk RNA-seq, potentially reducing the reliance on costly new reference atlases and improving the annotation of cell states that may be underrepresented in single-cell datasets.

2 Results

2.1 CellDiffusion to annotate cell types using bulk RNA-Seq references

Overview The design of CellDiffusion is directly inspired by the data generation process in single-cell experiments. During tissue dissociation and cell capture, only a fraction of the cells from the original sample are successfully sequenced, resulting in an inherently incomplete representation of the tissue's true cellular diversity. CellDiffusion addresses this undersampling issue by learning the underlying transcriptomic distributions from the observed cells to generate "virtual cells" with plausible transcriptional profiles of cells that were missed during the sequencing workflow. By aggregating virtual and observed cells, we create augmented cells that are directly comparable to bulk RNA-seq references while retaining local, single-cell resolution. This approach enables the use of robust classifiers for accurate cell type annotation. The complete conceptual model and modular workflow of CellDiffusion are presented in Figure 1.

A denoising diffusion generative model The core of our framework is based on a denoising diffusion probabilistic model (DDPM) trained on single-cell and spatial RNA-seq data. By learning the underlying data distribution (Figure 1B), the model generates a large population of synthetic, biologically plausible "virtual cells". This generative process enriches the dataset at a low computational cost, creating a more complete representation of the cellular landscape than initial, experimentally derived sample. Crucially, this approach

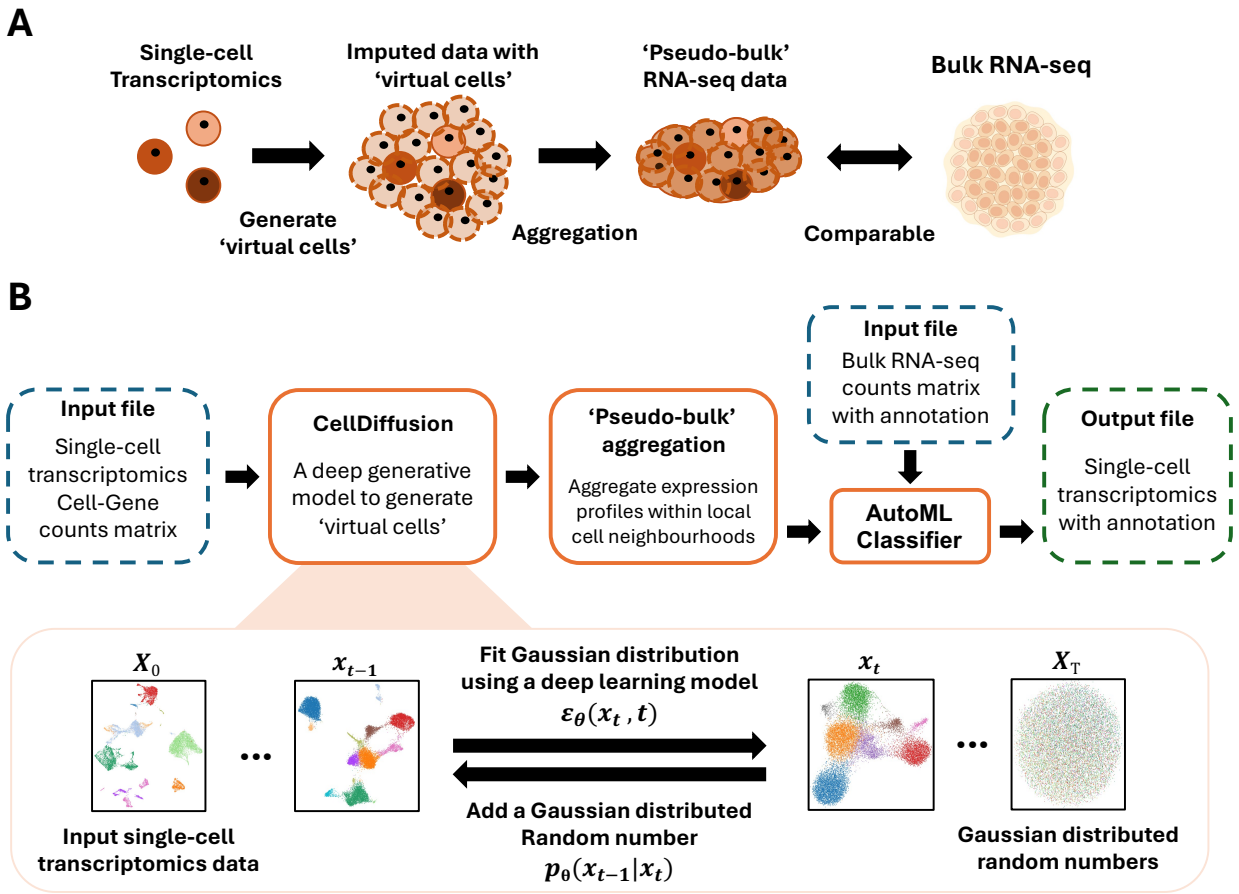

Fig 1: Overview of the CellDiffusion Framework. A: Conceptual model of CellDiffusion. Single-cell RNA-seq data represent an incomplete sample of the cellular diversity within tissues. CellDiffusion addresses this by computationally generating "virtual cells" to represent the unsampled portion of the population. **B:** The workflow starts with a core denoising diffusion probabilistic model (DDPM) that generates "virtual cells". The DDPM learns to reverse a forward noising process, allowing the generation of new, realistic virtual cells from a random input. The generated cells faithfully capture the underlying data distribution learned from the real cell population. Then, a cell augmentation module pools the original and virtual cells to create profiles comparable to bulk RNA-seq data. Finally, an automated machine learning classifier annotates these profiles using an established bulk RNA-seq reference atlas.

is highly effective for imputing the transcriptomic profiles of rare or transient cell states, which are often missed or underrepresented due to experimental sampling limitations.

Cell augmentation module To bridge the technical gap between single-cell and bulk sequencing modalities, this module pools small, similar groups of original and generated cells (typically around 15 cells) into local neighbourhoods (Figure 1B). By generating a unique augmented cell centred on each individual real cell, this method augments the sparse single-cell data, improving its signal quality to be comparable with robust bulk references. This

cell-centric strategy is critical: it enables the use of informative bulk atlases for annotation while preserving the single-cell resolution of the data.

Automated machine learning classifier For the final annotation step, we employ an automated machine learning (AutoML) classifier that operates on the high-quality augmented cells. This module systematically searches for the optimal classification model and hyperparameters for any given bulk reference dataset (Figure 1B). By automating model selection, this approach ensures robust and highly accurate annotations across diverse biological contexts and reference types, eliminating the need for laborious manual tuning. This design places the emphasis on the quality of the input signal rather than on a fixed choice of classifier, as the system automatically identifies a high-performing model when the biological signal is sufficient.

Validation and case studies We demonstrate the efficacy and versatility of our framework through rigorous benchmarking and three distinct case studies. First, we quantitatively assess CellDiffusion's performance against established annotation methods. We then showcase its ability to uncover novel biological insights through applications in profiling peripheral blood mononuclear cells, analysing white adipose tissue, and interrogating the tumour microenvironment with spatial transcriptomics.

2.2 Benchmark study

To systematically evaluate CellDiffusion's performance, we conducted a comprehensive benchmarking against four widely used annotation methods: SingleR, Seurat, CHETAH, and scVI (scANVI), across three distinct single-cell resolution datasets spanning different tissue types, technologies, and biological contexts [19, 22, 24, 25].

Uniform Manifold Approximation and Projection (UMAP) visualisation of cell type annotations (Figure 2A) demonstrates CellDiffusion's superior ability to recapitulate ground truth cell type distributions when using bulk RNA-seq references. While SingleR and Seurat also accommodated bulk references, they generated less coherent clustering patterns [19, 22]. In particular, CHETAH and scVI, designed primarily for scRNA-seq references, failed to gen-

erate meaningful annotations when provided with bulk RNA-seq references, highlighting the technical challenge that CellDiffusion effectively addresses [24, 25].

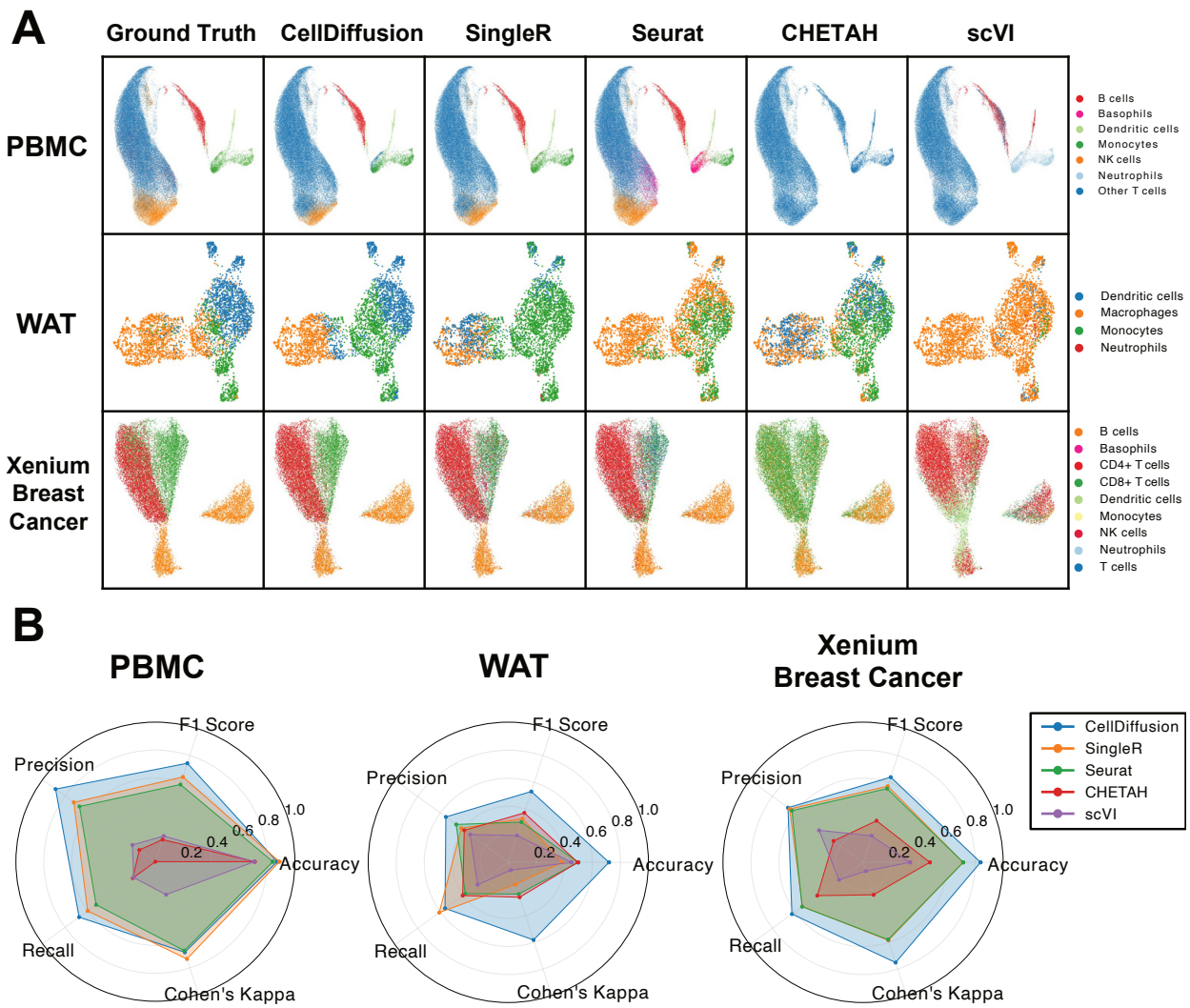

Fig 2: Benchmarking CellDiffusion against established cell type annotation methods. A: UMAP visualisations of cell type annotations on three distinct datasets, using a bulk RNA-seq reference. CellDiffusion's annotations closely recapitulate the ground truth distribution, generating coherent and well-separated cell clusters. In contrast, while SingleR and Seurat can process the bulk reference, they produce less distinct clustering patterns. Methods not designed for this task, CHETAH and scVI, fail to generate meaningful annotations. **B:** Radar plots comparing performance across five key metrics for the three datasets. CellDiffusion outperforms competing methods across the majority of evaluations. The larger area covered by CellDiffusion in each plot signifies its superior overall performance across all tested datasets.

Quantitative performance metrics further confirm CellDiffusion's advantages (Figure 2B). Across the three datasets, radar plots synthesising key metrics (accuracy, precision, recall, F1 score, and Cohen's Kappa) demonstrate that CellDiffusion consistently outperforms alternative methods. While SingleR showed slightly higher performance in terms of accuracy and Cohen's Kappa on the PBMC 68k dataset and for recall on the adipose dataset, CellDiffusion achieved the highest overall score across the vast majority of evaluations. CellDiffusion's superior performance was particularly pronounced when annotating rare cell populations, where traditional correlation-based (e.g. SingleR) and projection-based methods (e.g. Seurat) often struggle due to limited signal-to-noise ratios.

In summary, our benchmarking results established CellDiffusion as a competitive method for single-cell and spatial RNA-seq annotation using bulk RNA-seq references. Across diverse datasets, CellDiffusion consistently delivered more accurate and coherent cell type assignments than leading tools. Its key advantage lies in the robust identification of rare cell types, a task where conventional methods often fall short.

2.3 Identification of novel monocyte subtypes in human PBMC scRNA-seq data using CellDiffusion

We illustrate the identification of novel cell subtypes and cellular states with CellDiffusion to identify on a publicly available scRNA-seq dataset of human peripheral blood mononuclear cells (PBMCs) [26]. The complex and well-characterised nature of this immune cell population provides a compelling biological context to demonstrate the discovery power of our method.

While CellDiffusion successfully annotated all major immune lineages with high accuracy (Figure 3A), its superior resolution became evident in the analysis of monocytes. Focusing on this key myeloid population, CellDiffusion resolved three distinct subtypes: classical, intermediate, and non classical monocytes (Figure 3B1). This result provides a more granular view than standard analysis pipelines; for comparison, the widely used Azimuth reference mapped the same data to only two subtypes, classical and non classical monocytes, failing to distinguish the intermediate population (Figure 3B2).

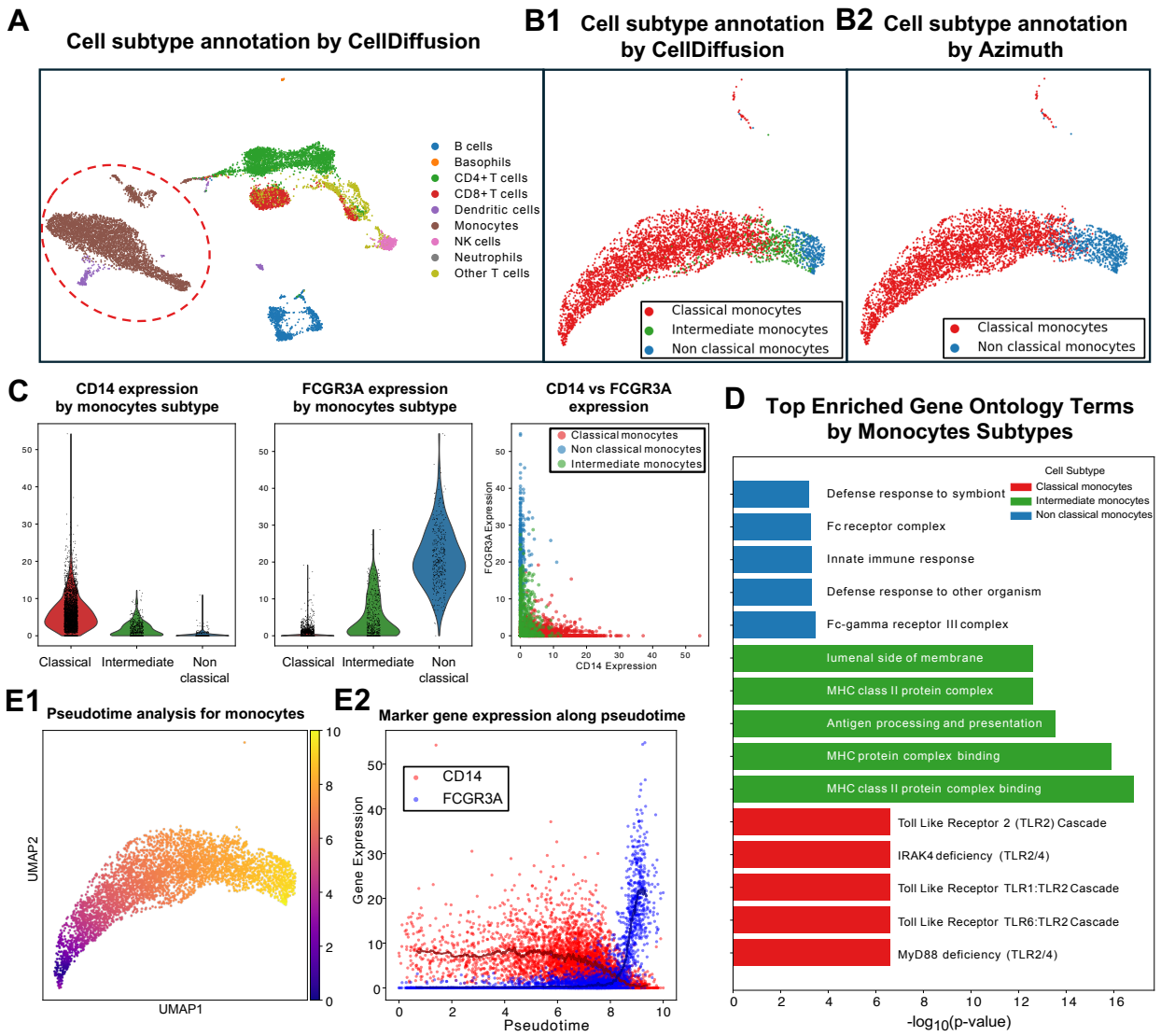

Fig 3: CellDiffusion identifies and characterises a distinct intermediate monocyte subtype in human PBMCs. A: UMAP visualisation of major cell types annotated by CellDiffusion in a public human PBMC dataset. Monocytes (highlighted) were isolated for detailed subtype analysis. **B:** Comparative UMAPs of monocyte subtypes. B1: CellDiffusion, using a bulk RNA-seq reference, resolves three populations: classical, intermediate, and nonclassical. B2: Azimuth, using a scRNA-seq reference, identifies only two populations: classical and non classical monocytes. **C:** Violin plots showing normalised expression of canonical markers CD14 and FCGR3A across the three monocyte subtypes, confirming the distinct expression profile of the intermediate population. **D:** Functional enrichment analysis of marker genes for intermediate monocytes. The top enriched terms are associated with antigen processing and presentation via MHC class II, suggesting a specialised immunological function. **E:** Pseudotime analysis models the differentiation trajectory from classical to nonclassical monocytes. E1: The trajectory clearly positions the intermediate population as a transitional state. E2: Gene expression dynamics along pseudotime show a decrease in CD14 and a concurrent increase in FCGR3A, consistent with this differentiation path.

The intermediate monocyte subtype identified by CellDiffusion was characterised by the simultaneous expression of CD14 and FCGR3A, as visualised in the UMAP plot of marker gene expression (Figure 3C). This unique expression pattern distinguishes intermediate monocytes from classical and non classical monocytes, suggesting that they represent a distinct cell population.

Importantly, this intermediate population represents more than a transitional state: this population has its own unique functional role. The enrichment analysis revealed that intermediate monocytes were significantly enriched for genes involved in the MHC class II antigen presentation pathways (Figure 3D). These pathways are fundamental to adaptive immunity, enabling specialised cells to present foreign antigens to helper T cells, which in turn orchestrate a targeted and lasting immune response [27]. Our analysis, therefore, pinpoints intermediate monocytes as professional antigen presenting cells (APCs), whose main function is to activate the adaptive immune system. This role is functionally distinct from the primary phagocytic duty of classical monocytes and the patrolling surveillance of non classical monocytes [28]. This conclusion is strongly supported by previous work that has identified intermediate monocytes as the most potent antigen-presenting subset in their constitutive state [29].

To place this functionally distinct subtype within its developmental context, we performed pseudotime trajectory analysis. The results inferred a continuous differentiation path from classical to non classical monocytes, with the intermediate cells positioned as a transitional state between them (Figure 3E). The smooth decrease in CD14 expression and the concurrent increase in FCGR3A expression along this trajectory provide strong evidence for a differentiation continuum.

In summary, by integrating its fine-grained annotations with downstream analyses, our study provides a more nuanced view of cellular differentiation, demonstrating that transitional cells within a continuum can be functionally specialized. These subtle cell states are often overlooked by standard scRNA-seq pipelines, which typically lack the robust high-resolution reference data needed to identify them. For instance, by leveraging bulk RNA-seq references in our analysis of PBMCs, CellDiffusion identified a functionally specialized intermediate monocyte population missed by other methods. This discovery underscores the

tool's ability to refine our understanding of cellular heterogeneity and the dynamic roles of transitional cells.

2.4 CellDiffusion reveals novel tissue-specific markers for immune cells in adipose tissue

We evaluated CellDiffusion's ability to characterise tissue-specific cellular features on a human white adipose tissue scRNA-seq dataset (Figure 4A1) [30]. This tissue is an ideal test case as it contains diverse populations of resident immune cells, which adapt to the local microenvironment and often diverge phenotypically from circulating immune cells [31].

Using a general bulk RNA-seq reference of myeloid cell lines, CellDiffusion accurately annotated the major adipose-resident myeloid cell types, showing high concordance with the original study's manual annotations (Figure 4A2) [30]. This performance notably exceeded that of a standard Seurat workflow with the Azimuth scRNA-seq reference (Figure 4A3). Importantly, neither the bulk reference nor Azimuth contained adipose-tissue-specific myeloid cells, yet Azimuth failed to correctly resolve the monocyte and macrophage populations. This discrepancy underscores a key challenge: tissue-resident myeloid cells diverge significantly from their counterparts in reference atlases due to tissue-specific adaptations, and CellDiffusion's approach appears more robust to this variation.

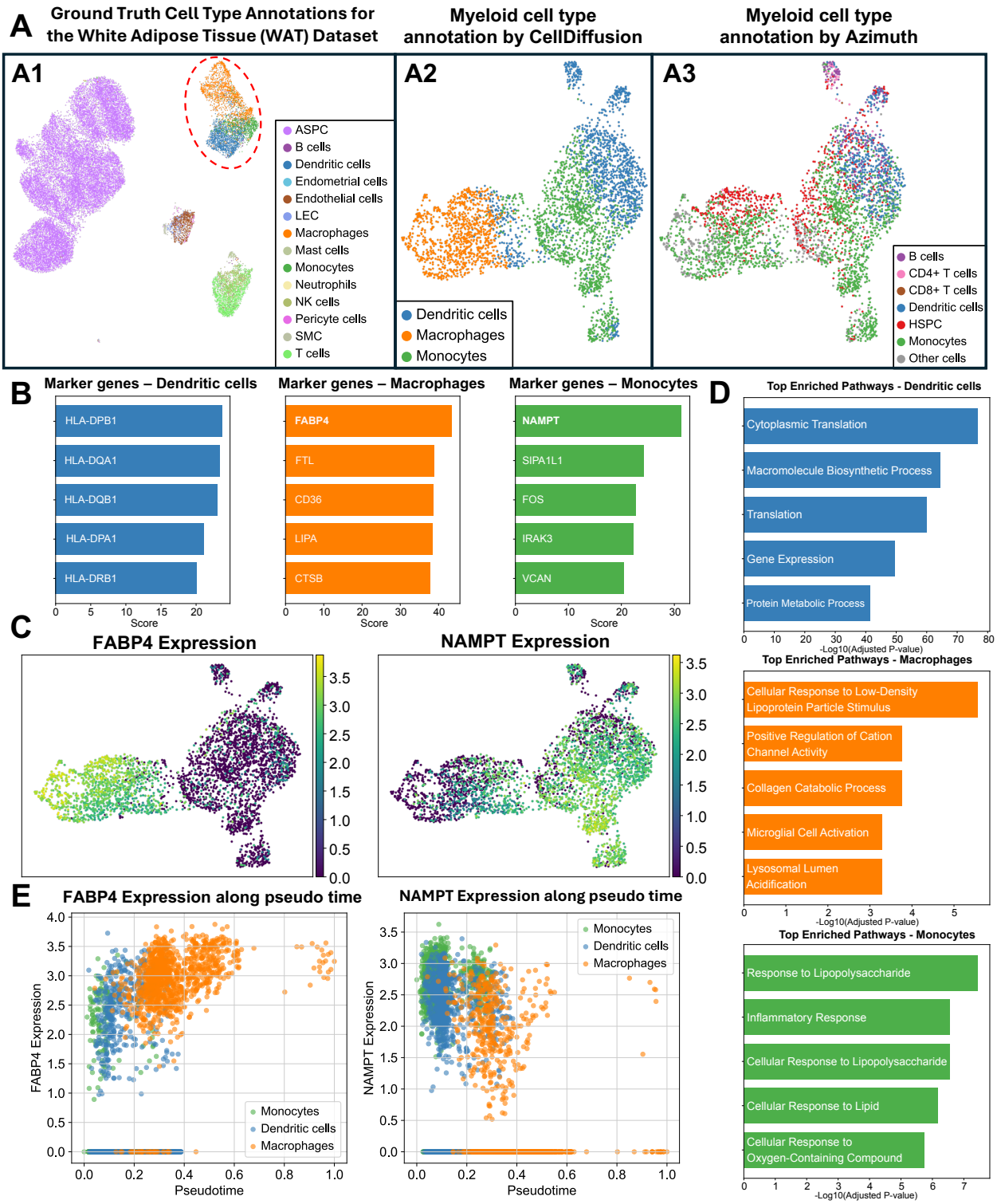

Fig 4: CellDiffusion accurately annotates tissue-resident immune cells and identifies novel markers in adipose tissue. A: UMAP projections of adipose tissue scRNA-seq data [30], focusing on myeloid cells. A1: Global view of all cell types with myeloid populations highlighted. A2: Cell type annotations of the isolated myeloid subset by CellDiffusion (using a bulk RNA-seq reference) A3: Annotations of the myeloid subset by Azimuth (using a scRNA-seq reference). [Continue next page.]

Fig 4: B: Bar plot of marker genes from differential expression analysis of CellDiffusion annotations. Analysis confirms known markers and identifies novel, highly significant markers such as FABP4 and NAMPT. **C:** Gene Ontology (GO) enrichment analysis for the three cell type populations. Enriched terms are consistent with known functions and highlight adipose-specific roles. **D:** UMAP feature plots showing distinct expression of the novel markers FABP4 (macrophages) and NAMPT (monocytes). **E:** Pseudotime trajectory analysis of monocyte to macrophage differentiation, revealing opposing expression dynamics where FABP4 increases and NAMPT decreases along the differentiation path.

In addition, the robust annotations provided by CellDiffusion enabled a detailed exploration of the adipose-resident myeloid scRNA-seq data. Differential expression analysis of the annotated populations first confirmed established markers, including HLA family genes in dendritic cells, CD36 in macrophages, and FOS in monocytes (Figure 4B) [32–34]. Beyond this validation, the analysis revealed FABP4 and NAMPT as highly significant and specific markers for adipose-resident macrophages and monocytes, respectively. Their expression was tightly restricted to these populations (Figure 4C), reflecting functional adaptation to the tissue microenvironment. FABP4 is a key protein in lipid metabolism, likely involved in processing lipids from surrounding adipocytes [35], while NAMPT is a rate-limiting enzyme in NAD^+ biosynthesis, pointing to specialised metabolic programming [36]. These findings are consistent with existing literature, which has demonstrated the function of FABP4 in adipose-resident macrophages [37] and has established the importance of NAMPT for both monocyte function and adipose tissue biology [38, 39].

This theme of metabolic adaptation was reinforced by gene ontology enrichment analysis. Both macrophages and monocytes showed significant enrichment for pathways related to lipid handling and metabolic responses, such as "cellular response to low-density lipoprotein particle stimulus" and "response to lipopolysaccharide" (Figure 4D). These functional profiles contrast with the primarily inflammatory signatures of circulating immune cells, highlighting a shift towards metabolic roles within the adipose tissue. To investigate the dynamics of these markers during monocyte to macrophage differentiation, we performed a pseudotime trajectory analysis. This analysis revealed opposing expression patterns: FABP4 expression increased along the differentiation trajectory, while NAMPT expression decreased (Figure 4E). This dynamic demonstrate the crucial role of FABP4 for macrophages and NAMPT for monocytes. It also suggests that these genes may play distinct regulatory roles during macrophage maturation within the adipose tissue microenvironment.

This case study highlights CellDiffusion's unique strength in leveraging general bulk RNA-seq references to accurately resolve tissue-specific cell populations that are challenging for conventional scRNA-seq atlases. The precision of CellDiffusion's annotations enabled both the validation of known markers and the discovery of novel, tissue-adapted marker genes like FABP4 and NAMPT. By providing a more accurate cellular map of the tissue, CellDiffusion directly facilitated deeper biological insights into the metabolic reprogramming of resident immune cells. Ultimately, this demonstrates CellDiffusion's value as a powerful discovery tool capable of revealing nuanced, tissue-specific biology that might otherwise be obscured when relying on canonical cell references [40].

2.5 CellDiffusion resolves the spatial architecture of the tumour microenvironment in breast cancer

We applied CellDiffusion to a Xenium spatial transcriptomics dataset from breast cancer samples to demonstrate its utility on image-based spatial transcriptomics data. We focused on resolving the immune landscape of the tumour microenvironment (TME), where the spatial arrangement of immune cells is crucial for understanding anti-tumour responses.

Using a bulk RNA-seq reference of immune cells, CellDiffusion successfully annotated the major immune populations within the TME. The resulting spatial cell type maps revealed distinct cellular clusters (Figure 5A), clearly identifying populations of $CD4^+$ T cells, $CD8^+$ T cells, and B cells in proximity to the tumour region (Figure 5B).

We next quantified the spatial relationships between the annotated immune populations and the tumour. This analysis revealed a clear spatial heterogeneity of lymphocytes: $CD4^+$ T cells were the most abundant population within the tumour region, followed by $CD8^+$ T cells, whereas B cells were sparse (Figure 5C). Consistent with this, both $CD4^+$ and $CD8^+$ T cells were significantly enriched and exhibited the highest local density near the tumour, in stark contrast to B cells, which showed no significant enrichment and the lowest density (Figures 5D-E).

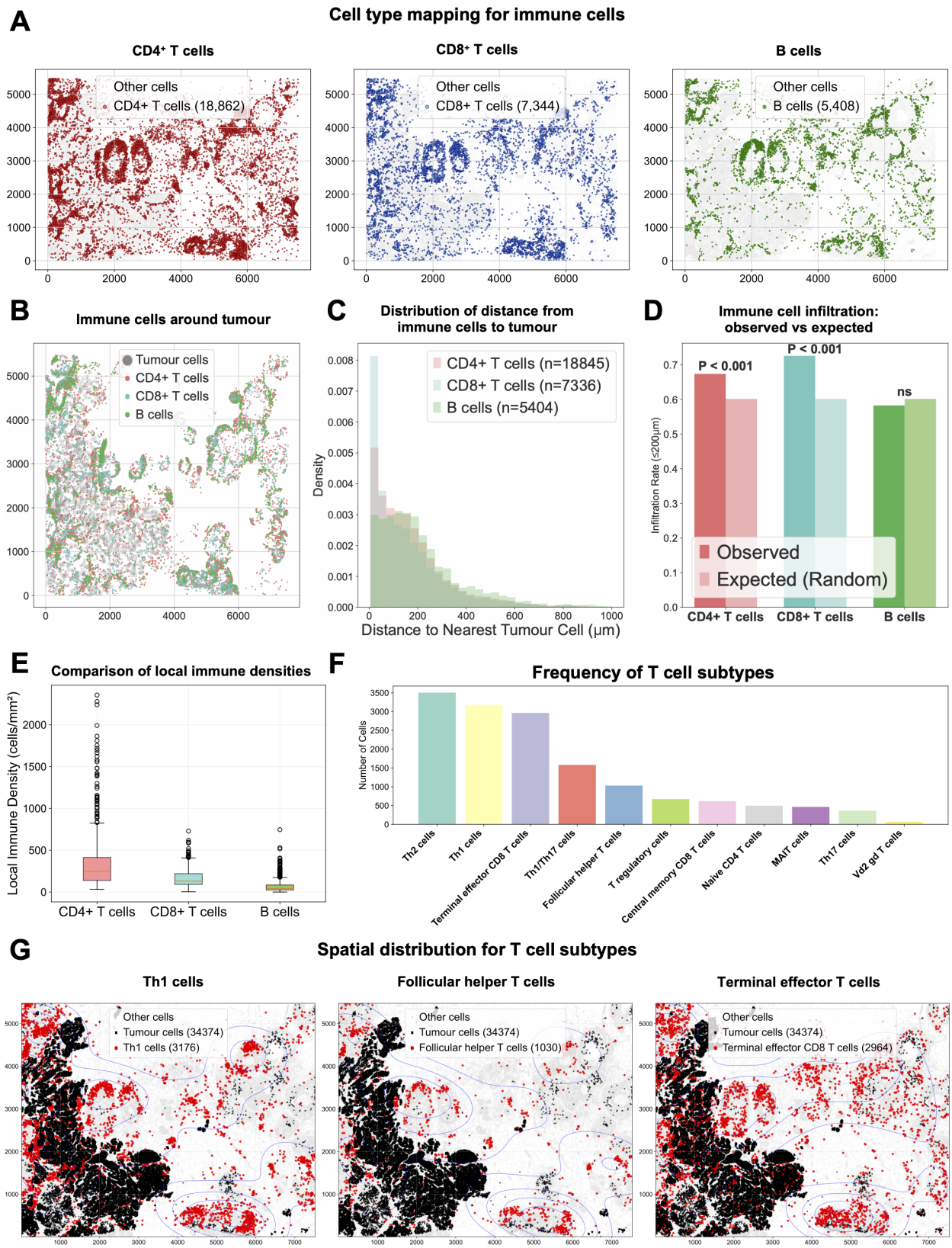

Fig 5: CellDiffusion annotate cell type for image-based spatial transcriptomics by using bulk RNA-seq reference. [Continue next page.]

Fig 5: A: Pseudo-colour from cell type annotation of spatial transcriptomics data for $CD4^+$ T cells, $CD8^+$ T cells, B cells from breast cancer tissue by using CellDiffusion framework with bulk RNA-seq reference, showing distinct clusters corresponding to major cell populations in the tumour microenvironment. **B:** Immune cell selected near tumour region for downstream analysis. **C:** The distribution of the distance between the annotated cell types and the tumour region, showing that $CD4^+$ T cells are the most abundant in the tumour region. $CD4^+$ T cells are the second most abundant, and B cells are the least abundant. **D:** Immune cells enrichment analyses show that $CD4^+$ T cells and $CD8^+$ T cells are significantly enriched around the tumour, but B cells are not. **E:** Local Immune cell density shows that $CD4^+$ T cells have the highest cell density around the tumour and $CD8^+$ T cells are the second highest, while B cells have the lowest cell density. **F:** The cell subtype distribution of T cells, showing that Th1, Th2 and terminal effector T cells are the most abundant T cells with the highest proportion, while Tfh cells have a lower proportion than we expect. **G:** Spatial map highlighting the distinct distributions of key T cell subtypes, including the peri-tumoural clustering of Th1 cells, the high local density of Terminal Effector $CD8^+$ T cells, and the exclusion of Tfh cells from the tumour interior.

Our analysis revealed a multi-faceted immune dysfunction within the tumor microenvironment. The most striking feature was a compromised cytotoxic response, characterized by a large population of terminally exhausted $CD8^+$ T cells (Figure 5F), indicating T cell exhaustion and a diminished anti-tumour effort [41, 42]. This weakened attack was compounded by a failure to mount an effective B cell response. We observed a significant scarcity of T follicular helper (Tfh) cells, which are essential for activating and sustaining B cell activity [43]. Spatially, these Tfh cells were almost entirely excluded from the tumor core, providing a clear mechanistic explanation for the observed lack of B cell infiltration 5G). Finally, the T helper landscape showed a slight polarization towards a Th2 phenotype, a known mechanism of tumor immune evasion [44].

This case study highlights CellDiffusion's power to resolve complex cellular architecture from highly sparse, image-based spatial transcriptomics data. By resolving these nuanced T cell subtypes and their spatial organization, our method overcomes significant data limitations. This enables the generation of functional hypotheses that directly link cellular organization to disease mechanisms, effectively bridging the gap between existing bulk RNA-seq references and new spatial transcriptomics data.

3 Discussion

CellDiffusion is a deep learning framework designed to address a central challenge in transcriptomics: the integration of large, well-annotated bulk RNA-seq datasets with high-

resolution single-cell and spatial data. By generating virtual cells via a diffusion model, our approach helps bridge these modalities, mitigating some of the technical and distributional disparities that have traditionally hindered their integration. Our contribution allows researchers to utilise decades of existing biological knowledge from bulk RNA-seq, reducing the need for costly new reference atlases and providing access to a richer diversity of cellular states that may be underrepresented in single-cell datasets.

Our framework facilitates the elucidation of nuanced biological processes across diverse contexts. For example, we identified subtle, intermediate monocyte states in PBMCs. In adipose tissue, CellDiffusion's ability to account for tissue specific expression patterns was essential for accurately characterising specialised immune cells. In the application to spatial transcriptomics, we highlighted CellDiffusion's capacity to handle highly sparse data while preserving spatial context. Our case studies demonstrate that CellDiffusion is more than a cell annotation tool, but also a hypothesis-generating framework to reveal complex cellular identities and states.

While CellDiffusion shows promise, it is important to acknowledge its current limitations. Firstly, the accuracy of annotations is fundamentally dependent on the quality and comprehensiveness of the reference bulk RNA-seq atlas. An incomplete or biased reference may limit the ability to identify certain cell populations. Furthermore, the reliance on high-purity bulk references often generated via cell sorting currently tailors the method's application primarily to immune cell types. Secondly, to make sure the generated virtual cells remain biologically accurate, our method currently uses a filtering step after generation. This process uses a k-NN graph to select the most realistic virtual cells. While effective, the necessity for this step indicates that the realism of the core generative model could be further improved. Finally, the training process for the diffusion model is computationally intensive and generally requires GPU acceleration, which may pose a challenge for users with limited computational resources, particularly when working with very large datasets.

This work paves the ways for future improvement, such as developing more efficient generative models to simultaneously enhance the quality of virtual cells and reduce running time. Creating models capable of predicting entire single-cell transcriptomic landscapes under specified biological conditions would represent a paradigm shift, enabling researchers

to supplement laboratory work with powerful *in silico* experiments to accelerate discovery.

4 Methods

4.1 CellDiffusion generative model

CellDiffusion is a denoising diffusion probabilistic model (DDPM) [23] trained to learn the underlying data distribution of single-cell resolution transcriptomics and generate synthetic data, which we term "virtual-cells". The model is versatile, capable of processing gene expression profiles from technologies like droplet-based scRNA-seq and image-based spatial transcriptomics.

The overall process involves three key stages: data preprocessing, model training, and virtual cell generation. First, raw gene expression data undergoes a standard preprocessing pipeline, including filtering, normalisation, and log transformation [45]. A set of highly variable genes is selected, and each cell's expression vector is reshaped into a square matrix, x_0, to be compatible with the model's convolutional architecture.

Conceptually, the DDPM framework operates through a forward process, where Gaussian noise is incrementally added to the input data, and a learned reverse process. The core of CellDiffusion is a U-Net model [46] trained to reverse this noising process. Starting from pure noise, the trained model can iteratively denoise a sample to generate a new, synthetic gene expression matrix, $\hat{x}_0$. To account for inherent limitations in model performance, we employ a generate and filter strategy. Initially, a large surplus of virtual cells is generated. Subsequently, during the cell augmentation stage, a filtering step is implemented to select a high-quality subset of virtual cells, using information from the real cell population, before they are aggregated. This approach is computationally feasible due to the low cost of synthesising virtual cells.

A detailed description of the DDPM theoretical foundation, specific model architecture, training parameters, and the mathematical formulation for data generation is provided at Supplemental Material S1.

4.2 Cell augmentation

We used a pseudobulk aggregation strategy to bridge the gap between single-cell and bulk RNA-seq data. This method synthesizes bulk-like expression profiles from local cellular neighborhoods, which augments the signal for each cell while retaining the transcriptional heterogeneity inherent to the single-cell data.

The process begins by constructing a k-nearest neighbour (k-NN) graph on the combined dataset of original cells and the synthetic cells generated by the diffusion model. We utilise the `scanpy` package for this purpose, with the number of neighbours (k) set to a default of 15. The local community, C_j, for a given cell j is defined as the set containing the cell itself and its k-nearest neighbours in the graph.

For each community, we generate a single augmented cell by averaging the gene expression vectors of all cells within that community. This aggregation step is defined by the formula:

$$P_j = \frac{1}{|C_j|} \sum_{i \in C_j} X_i$$

where P_j is the resulting augmented cell for community j, X_i is the expression vector of an individual cell i within that community, and $|C_j|$ is the total number of cells in the community (i.e., $k + 1$).

Finally, to ensure compatibility with real bulk RNA-seq data, the generated augmented cell are normalised to counts per million (CPM) and log-transformed. This neighbourhood-based aggregation effectively enhances the signal for distinct cellular states, including rare populations, and creates a well-founded dataset for direct comparison with bulk references.

4.3 Automated machine learning for cell type annotation

We employed an automated machine learning (AutoML) approach to classify our augmented cells using the bulk RNA-seq data as a reference. For this task, we utilised the Tree-based Pipeline Optimization Tool (TPOT, `version 0.12.2`), an AutoML framework that uses genetic programming to systematically discover the most effective machine learning pipeline [47]. TPOT was configured to search for an optimal classifier with a population size of

50 potential pipelines over 5 generations. This process automatically identified the best-performing model for assigning cell type labels to the augmented cells query data based on the patterns learned from the bulk reference.

4.4 Data and data preprocessing

We provide a comprehensive list of all datasets used in this study in Supplemental Material S2. We also specify the role of each dataset as either a query or a reference. We then detail the preprocessing pipelines for both data types. For query data, this includes the selection of highly variable genes, normalisation, log-transformation, and the crucial step of reshaping gene vectors into square matrices for compatibility with our model's architecture. A parallel preprocessing workflow is described for the bulk RNA-seq reference data to ensure consistency.

4.5 Methods for benchmark and case studies

To validate our model, we conducted a rigorous benchmark study against several state-of-the-art annotation tools, including SingleR, Seurat, CHETAH, and scVI (scANVI). The Supplemental Material S3 details the specific parameters used for each method and defines the suite of performance metrics, such as the macro F1-score and Cohen's Kappa, chosen for their robustness to class imbalance.

Furthermore, we outline the downstream analytical pipelines for our three biological case studies at Supplemental Material S4. These include methods for investigating cell subtype heterogeneity and differentiation dynamics in monocyte and adipose tissue data (e.g., UMAP, differential gene expression, trajectory inference) and techniques for characterising the tumour microenvironment in spatial transcriptomics data (e.g., proximity analysis with permutation testing).

Declarations

Code availability. The CellDiffusion Python package is available on GitHub (`https://github.com/ShiltonZhang/CellDiffusion`), along with the Python code for processing the data and reproducing our results.

Data availability. All datasets used in this study are publicly available. These include the PBMC 68k scRNA-seq dataset [48], the PBMC 10k scRNA-seq dataset [49], the human white adipose tissue data [30], the Xenium spatial transcriptomics dataset for human breast cancer [50], the bulk RNA-seq reference from Monaco et al. [51], and the FANTOM5 bulk RNA-seq reference atlas [52].

Author contributions. XZ developed the method, conducted the analysis, wrote the manuscript. KALC and JM supervised the work, wrote and edited the manuscript.

Competing interests. The authors declare they have no competing interests.

Acknowledgements. We would like to thank Dr Alexandr Garbali (University of Melbourne) for helpful discussions.

Funding. This research was supported by the Australian Research Council Centre of Excellence in Quantum Biotechnology (QUBIC) through project number CE230100021. XZ was supported by a QUBIC strategic Scholarship. KALC and JM were supported by the National Health and Medical Research Council (NHMRC) Investigator Grant (GNT2025648).

References

[1] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.

[2] Liyang Song, Wenhao Chen, Junren Hou, Minmin Guo, and Jian Yang. Spatially resolved mapping of cells associated with human complex traits. *Nature*, pages 1–10, 2025.

[3] Cameron G Williams, Hyun Jae Lee, Takahiro Asatsuma, Roser Vento-Tormo, and Ashraful Haque. An introduction to spatial transcriptomics for biomedical research. *Genome medicine*, 14(1):68, 2022.

[4] Sergio Marco Salas, Louis B Kuemmerle, Christoffer Mattsson-Langseth, Sebastian Tismeyer, Christophe Avenel, Taobo Hu, Habib Rehman, Marco Grillo, Paulo Czarnewski, Saga Helgadottir, et al. Optimizing xenium in situ data utility by quality assessment and best-practice analysis workflows. *Nature Methods*, pages 1–11, 2025.

[5] Yang Jin, Yuanli Zuo, Gang Li, Wenrong Liu, Yitong Pan, Ting Fan, Xin Fu, Xiaojun Yao, and Yong Peng. Advances in spatial transcriptomics and its applications in cancer research. *Molecular Cancer*, 23(1):129, 2024.

[6] Lukas Heumos, Anna C Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D Lücken, Daniel C Strobl, Juan Henao, Fabiola Curion, et al. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8):550–572, 2023.

[7] Benjamin D Hale, Yannik Severin, Fabienne Graebnitz, Dominique Stark, Daniel Guignard, Julien Mena, Yasmin Festl, Sohyon Lee, Jacob Hanimann, Nathan S Zangger, et al. Cellular architecture shapes the naïve t cell response. *Science*, 384(6700):eadh8697, 2024.

[8] Qiqing Fu, Chenyu Dong, Yunhe Liu, Xiaoqiong Xia, Gang Liu, Fan Zhong, and Lei Liu. A comparison of scrna-seq annotation methods based on experimentally labeled immune cell subtype dataset. *Briefings in Bioinformatics*, 25(5):bbae392, 2024.

[9] Tamim Abdelaal, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel JT Reinders, and Ahmed Mahfouz. A comparison of automatic cell identification methods for single-cell rna sequencing data. *Genome biology*, 20(1):194, 2019.

[10] Elisabetta Mereu, Atefeh Lafzi, Catia Moutinho, Christoph Ziegenhain, Davis J McCarthy, Adrián Álvarez-Varela, Eduard Batlle, N Sagar, Dominic Gruen, Julia K Lau, et al. Benchmarking single-cell rna-sequencing protocols for cell atlas projects. *Nature biotechnology*, 38

(6):747–755, 2020.

[11] Orit Rozenblatt-Rosen, Jay W Shin, Jennifer E Rood, Anna Hupalowska, Aviv Regev, and Holger Heyn. Building a high-quality human cell atlas. *Nature Biotechnology*, 39(2):149–153, 2021.

[12] Shuai He, Lin-He Wang, Yang Liu, Yi-Qi Li, Hai-Tian Chen, Jing-Hong Xu, Wan Peng, Guo-Wang Lin, Pan-Pan Wei, Bo Li, et al. Single-cell transcriptome profiling of an adult human cell atlas of 15 major organs. *Genome biology*, 21(1):294, 2020.

[13] Jinming Cheng, Xinyi Jin, Gordon K Smyth, and Yunshun Chen. Benchmarking cell type annotation methods for 10x xenium spatial transcriptomics data. *BMC bioinformatics*, 26(1): 22, 2025.

[14] John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[15] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47, 2015.

[16] Rory Stark, Marta Grzelak, and James Hadfield. Rna sequencing: the teenage years. *Nature reviews genetics*, 20(11):631–656, 2019.

[17] Xuanwen Bao, Run Shi, Tianyu Zhao, Yanfang Wang, Natasa Anastasov, Michael Rosemann, and Weijia Fang. Integrated analysis of single-cell rna-seq and bulk rna-seq unravels tumour heterogeneity plus m2-like tumour-associated macrophage infiltration and aggressiveness in tnbc. *Cancer Immunology, Immunotherapy*, 70(1):189–202, 2021.

[18] Xiaojun Zhang, Ran Feng, Junbin Guo, Lihui Pan, Yarong Yao, and Jinnan Gao. Integrated single-cell and bulk rna sequencing analysis identifies a neoadjuvant chemotherapy-related gene signature for predicting survival and therapy in breast cancer. *BMC Medical Genomics*, 16 (1):300, 2023.

[19] Dvir Aran, Agnieszka P Looney, Leqian Liu, Esther Wu, Valerie Fong, Austin Hsu, Suzanna Chak, Ram P Naikawadi, Paul J Wolters, Adam R Abate, et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nature immunology*, 20(2):163–172, 2019.

[20] Yidi Deng, Jarny Choi, and Kim-Anh Lê Cao. Sincast: a computational framework to predict cell identities in single-cell transcriptomes using bulk atlases as references. *Briefings in*

Bioinformatics, 23(3):bbac088, 2022.

[21] Jiadong Mao, Yidi Deng, and Kim-Anh Lê Cao. ϕ-space: Continuous phenotyping of single-cell multi-omics data. *Genome Biology*, 26(1):323, 2025.

[22] Yuhan Hao, Tim Stuart, Madeline H Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, Avi Srivastava, Gesmira Molla, Shaista Madad, Carlos Fernandez-Granda, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature biotechnology*, 42(2):293–304, 2024.

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[24] Jurrian K De Kanter, Philip Lijnzaad, Tito Candelli, Thanasis Margaritis, and Frank CP Holstege. Chetah: a selective, hierarchical cell type identification method for single-cell rna sequencing. *Nucleic acids research*, 47(16):e95–e95, 2019.

[25] Chenling Xu, Romain Lopez, Edouard Mehlman, Jeffrey Regier, Michael I Jordan, and Nir Yosef. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Molecular systems biology*, 17(1):e9620, 2021.

[26] 10x Genomics. 10k peripheral blood mononuclear cells (pbmcs) from a healthy donor. `https://support.10xgenomics.com/single-cell-gene-expression/datasets/3.0.0/pbmc_10k_v3`, 2018.

[27] Eilon Sharon, Leah V Sibener, Alexis Battle, Hunter B Fraser, K Christopher Garcia, and Jonathan K Pritchard. Genetic variation in mhc proteins is associated with t cell receptor expression biases. *Nature genetics*, 48(9):995–1002, 2016.

[28] Kolandaswamy Anbazhagan, Isabelle Duroux-Richard, Christian Jorgensen, and Florence Apparailly. Transcriptomic network support distinct roles of classical and non-classical monocytes in human. *International reviews of immunology*, 33(6):470–489, 2014.

[29] Justin Lee, Hanson Tam, Lital Adler, Alexandra Ilstad-Minnihan, Claudia Macaubas, and Elizabeth D Mellins. The mhc class ii antigen presentation pathway in human monocytes differs by subset and is regulated by cytokines. *PloS one*, 12(8):e0183594, 2017.

[30] Margo P. Emont, C. Jacobs, E. L. Adamson, D. J. Hughes, D. Korn, V. Bodò, S. L. Watson, A. C. Mower, M. E. Hmat, F. M. Loni, M. Takeda, J. B. Kang, E. Brown, J. H. Lee, D. R. Lemos, C. Lu, Y. H. Lee, S. Collins, C. Wolfrum, and S. Kajimura. A single-cell atlas of human and mouse white adipose tissue. *Nature*, 603(7903):926–933, 2022. doi: 10.1038/s41586-022-04518-2. URL `https://singlecell.broadinstitute.org/single_cell/study/`

SCP1376/a-single-cell-atlas-of-human-and-mouse-white-adipose-tissue. Dataset available at this URL. Accessed: 2025-10-16.

[31] Jia Li, Chu Xiao, Chunxiang Li, and Jie He. Tissue-resident immune cells: from defining characteristics to roles in diseases. *Signal Transduction and Targeted Therapy*, 10(1):12, 2025.

[32] Clara Bueno, Julia Almeida, Paulo Lucio, Josefa Marco, Raimundo Garcia, Jose Maria De Pablos, Antonio Parreira, Fernando Ramos, Francisco Ruiz-Cabello, Dimas Suarez-Vilela, et al. Incidence and characteristics of cd4 (+)/hla drhi dendritic cell malignancies. *haematologica*, 89(1):58–69, 2004.

[33] Roy L Silverstein and Maria Febbraio. Cd36, a scavenger receptor involved in immunity, metabolism, angiogenesis, and behavior. *Science signaling*, 2(72):re3–re3, 2009.

[34] Takashi Nakamura, R Datta, Surender Kharbanda, and Donald Kufe. Regulation of jun and fos gene expression in human monocytes by the macrophage colony-stimulating factor. *Cell Growth Differ*, 2(6):267–72, 1991.

[35] Kacey J Prentice, Jani Saksi, and Gökhan S Hotamisligil. Adipokine fabp4 integrates energy stores and counterregulatory metabolic responses. *Journal of Lipid Research*, 60(4):734–740, 2019.

[36] Masamichi Yano, Hiroshi Akazawa, Toru Oka, Chizuru Yabumoto, Yoko Kudo-Sakamoto, Takehiro Kamo, Yu Shimizu, Hiroki Yagi, Atsuhiko T Naito, Jong-Kook Lee, et al. Monocyte-derived extracellular nampt-dependent biosynthesis of nad+ protects the heart against pressure overload. *Scientific Reports*, 5(1):15857, 2015.

[37] Masato Furuhashi, Shigeyuki Saitoh, Kazuaki Shimamoto, and Tetsuji Miura. Fatty acid-binding protein 4 (fabp4): pathophysiological insights and potent clinical biomarker of metabolic and cardiovascular diseases. *Clinical medicine insights: cardiology*, 8:CMC–S17067, 2014.

[38] Marisela Rodriguez, Haifei Xu, Annie Hernandez, Julia Ingraham, Jason Canizales, Fernando Teran Arce, Sara M Camp, Skyler Briggs, Aikseng Ooi, James M Burke, et al. Nedd4 e3 ligase-catalyzed nampt ubiquitination and autophagy activation are essential for pyroptosis-independent nampt secretion in human monocytes. *Cell Communication and Signaling*, 23(1): 157, 2025.

[39] Kelly L Stromsdorfer, Shintaro Yamaguchi, Myeong Jin Yoon, Anna C Moseley, Michael P Franczyk, Shannon C Kelly, Nathan Qi, Shin-ichiro Imai, and Jun Yoshino. Nampt-mediated nad+ biosynthesis in adipocytes regulates adipose tissue function and multi-organ insulin

sensitivity in mice. *Cell reports*, 16(7):1851–1860, 2016.

[40] Shane Crotty. T follicular helper cell differentiation, function, and roles in disease. *Immunity*, 41(4):529–542, 2014.

[41] Jean-Christophe Beltra, Sasikanth Manne, Mohamed S Abdel-Hakeem, Makoto Kurachi, Josephine R Giles, Zeyu Chen, Valentina Casella, Shin Foong Ngiow, Omar Khan, Yinghui Jane Huang, et al. Developmental relationships of four exhausted cd8+ t cell subsets reveals underlying transcriptional and epigenetic landscape control mechanisms. *Immunity*, 52(5):825–841, 2020.

[42] Christian U Blank, W Nicholas Haining, Werner Held, Patrick G Hogan, Axel Kallies, Enrico Lugli, Rachel C Lynn, Mary Philip, Anjana Rao, Nicholas P Restifo, et al. Defining 't cell exhaustion'. *Nature Reviews Immunology*, 19(11):665–674, 2019.

[43] Adesola C Olatunde, J Scott Hale, and Tracey J Lamb. Cytokine-skewed tfh cells: functional consequences for b cell help. *Trends in immunology*, 42(6):536–550, 2021.

[44] Rafael Cardoso Maciel Costa Silva, Marcela Freitas Lopes, and Leonardo Holanda Travassos. Distinct t helper cell-mediated antitumor immunity: T helper 2 cells in focus. *Cancer Pathogenesis and Therapy*, 1(01):76–86, 2023.

[45] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19(1):15, 2018.

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[47] Pedro Ribeiro, Anil Saini, Jay Moran, Nicholas Matsumoto, Hyunjun Choi, Miguel Hernandez, and Jason H Moore. Tpot2: A new graph-based implementation of the tree-based pipeline optimization tool for automated machine learning. In *Genetic Programming Theory and Practice XX*, pages 1–17. Springer, 2024.

[48] Grace X. Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. T. Gregory, J. Shuga, L. Montesclaros, J. G. Underwood, D. A. Masquelier, S. Y. Nishimura, M. Schnall-Levin, P. W. Wyatt, C. M. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, 2017. doi: 10.1038/ncomms14049. URL `https:`

//www.10xgenomics.com/datasets/fresh-68-k-pbm-cs-donor-a-1-standard-1-1-0. Dataset available at this URL. Accessed: 2025-10-16.

[49] 10x Genomics. 10k peripheral blood mononuclear cells (pbmcs) from a healthy donor. https://www.10xgenomics.com/datasets/10-k-peripheral-blood-mononuclear-cells-pbm-cs-from-a-healthy-donor-single-indexed-3-1-standard-4-0-0, 2020. Accessed: 2025-10-16.

[50] 10x Genomics. Xenium in situ for human breast (preview dataset). https://www.10xgenomics.com/products/xenium-in-situ/preview-dataset-human-breast, 2025. Accessed: 2025-10-16.

[51] Gianni Monaco, Bernett Lee, Weili Xu, Kerrigan Tooke, Wilson Wen Bin Goh, Syu-Li Kong, Michael Zhen Mian Chow, Edwin Lim, David M. Kemeny, Ashley L. St. John, M. B. Zocca, A. Mortellaro, and F. Ginhoux. RNA-Seq Signatures of Group-Wise Cell-Type Proportions in Human Blood. *Cell Reports*, 26(7):1791–1805.e4, 2019. doi: 10.1016/j.celrep.2019.01.068. URL https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107019. Dataset available at this URL. Accessed: 2025-10-16.

[52] Robin Andersson, C. Gebhard, I. Miguel-Escalada, I. Hoof, J. Bornholdt, M. Boyd, Y. Chen, X. Zhao, C. Schmidl, T. Suzuki, E. Ntini, E. Arner, E. Valen, K. Männik, C. Sponholtz, J. Rador, J. Våges, A. Zhernakova, M. Jhamai, N. Bertin, J. Schein, L. J. Core, J. T. Lis, T. R. Gingeras, M. Hirst, M. A. Marra, C. Notredame, T. Arakawa, S. Kojima, H. Kawaji, A. R. R. Forrest, P. Carninci, Y. Hayashizaki, A. Sandelin, O. Hofmann, and FANTOM Consortium. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, 2014. doi: 10.1038/nature12787. URL http://fantom.gsc.riken.jp/5/datafiles/. Dataset available at this URL. Accessed: 2025-09-10.