OXFORD

Briefings in Bioinformatics, 2025, **26(5)**, bbaf523
https://doi.org/10.1093/bib/bbaf523
Problem Solving Protocol

scHDeepInsight: a hierarchical deep learning framework for precise immune cell annotation in single-cell RNA-seq data

Shangru Jia[1], Artem Lysenko[2,*], Keith A Boroevich[3], Alok Sharma[2,3,4,5,‡,*], Tatsuhiko Tsunoda[1,2,‡,*]

[1]Laboratory for Medical Science Mathematics, Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[2]Laboratory for Medical Science Mathematics, Department of Biological Sciences, School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[3]RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi, Yokohama 230-0045, Japan
[4]Institute for Integrated and Intelligent Systems, Griffith University, 170 Kessels Rd, Nathan, Brisbane, QLD 4111, Australia
[5]College of Informatics, Korea University, 145 Anam-ro, Seongbuk District, Seoul 02841, South Korea

*Corresponding authors. Artem Lysenko, Laboratory for Medical Science Mathematics, Department of Biological Sciences, School of Science, The University of Tokyo. 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. E-mail: alysenko@g.ecc.u-tokyo.ac.jp; Alok Sharma, Laboratory for Medical Science Mathematics, Department of Biological Sciences, School of Science, The University of Tokyo. 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. E-mail: alok.fj@gmail.com; Tatsuhiko Tsunoda, Laboratory for Medical Science Mathematics, Department of Biological Sciences, School of Science, The University of Tokyo. 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan. Fax: +81-3-5841-4263; E-mail: tsunoda@bs.s.u-tokyo.ac.jp
‡Alok Sharma and Tatsuhiko Tsunoda are co-last authors

Abstract

Accurate classification of immune cells is crucial for elucidating their diverse roles in health and disease. However, this task remains very challenging in single-cell RNA sequencing (scRNA-seq) data due to the complex and hierarchical relationships of immune cell types. To address this, we introduce scHDeepInsight, a deep learning framework that extends our previous scDeepInsight model by integrating a biologically-informed classification architecture with an adaptive hierarchical focal loss (AHFL). The framework builds on our established method of converting gene expression data into two-dimensional structured images, enabling convolutional neural networks to effectively capture both global and fine-grained transcriptomic features. This design utilizes hierarchical relationships among immune cell types to enhance the classification ability beyond the flat classification approaches. scHDeepInsight dynamically adjusts loss contributions to balance performance across the hierarchy levels. Comprehensive benchmarking across seven diverse tissue datasets shows scHDeepInsight achieves an average accuracy of 93.2%, surpassing contemporary methods by 5.1 percentage points. The model successfully distinguishes 50 distinct immune cell subtypes with high accuracy, demonstrating proficiency for identifying rare and closely related cell subtypes. Additionally, SHAP-based interpretability quantifies individual gene contributions to reveal the biological basis of classification decisions. These qualities make scHDeepInsight a robust tool for high-resolution cell subtype characterization, well-suited for detailed profiling in immunological studies and extensible to nonimmune cell types.

Keywords: cell annotation; cell subtype; single-cell RNA sequencing; deep learning; transformers

Introduction

Comprehensive understanding of complex biological systems and their abnormal states, such as cancer and chronic diseases, requires accurate identification of cell types [2], particularly of immune cells. The advent of single-cell RNA sequencing (scRNA-seq) technology has revolutionized our ability to analyze gene expression profiles at an unprecedented resolution, enabling detailed characterization of cellular heterogeneity at the individual cell level. Widely used methods for cell type annotation, such as SingleR [1], Azimuth [2], and scmap [3], are reference-based, assigning cell type labels by comparing query cells to known reference profiles using similarity metrics. These comparisons are generally processed in a uniform manner, without explicit consideration of the hierarchical relationships among cell types. While such existing annotation tools achieve high accuracy in broad cell type identification, incorporating predefined immune cell hierarchies has the potential to further enhance their effectiveness for detailed subtype classification. By incorporating these hierarchies, it becomes possible to distinguish closely related subtypes, cells which often share high transcriptional similarity yet differ significantly in function and biological roles. Such refinement enhances the biological relevance and interpretability of downstream analyses. We posit that explicitly modeling the known biological hierarchy of immune cell populations in the annotation process is needed. Incorporating hierarchical modeling naturally captures functional relationships and widely accepted immune lineage structures, which are derived from curated ontological annotations and partially

Received: May 22, 2025. **Revised:** August 06, 2025. **Accepted:** September 10, 2025
© The Author(s) 2025. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

aligned with known differentiation patterns, thereby enhancing the biological accuracy and context-specific interpretation of annotation results.

To realize these important improvements, we introduce scHDeepInsight, an enhanced deep learning framework built upon the foundation of scDeepInsight [4] by explicitly employing a hierarchical annotation architecture. The core innovation of scHDeepInsight lies in the integration of three key components: (i) the transformation of high-dimensional gene expression data into spatially-organized two-dimensional (2D) images [5], making them suitable for convolutional neural network (CNN) [6] processing; (ii) the implementation of a multi level classification architecture that preserves the biological hierarchies of base-types and subtypes of cells; and (iii) the incorporation of an adaptive hierarchical focal loss (AHFL) function that automatically balances training priorities by adjusting the weights of base-type and subtype focal losses according to their relative performance. This approach enables researchers to precisely dissect the complexities of the immune system at a higher resolution. Furthermore, the multilevel classification structure enables stratified feature importance quantification through SHAP (SHapley Additive exPlanations) [7] analysis at both base-type and subtype cell levels, thereby providing biological insights into the distinct gene expression signatures that differentiate closely related immune cell populations.

While scHDeepInsight is primarily designed and trained for immune cell classification, its hierarchical architecture is inherently generalizable. The model's structure allows extension to other cellular contexts, including stromal and epithelial cells, which are particularly relevant in complex tissue environments. This adaptability broadens the utility beyond immunology, making it suitable for integrated annotation tasks in diverse biomedical applications.

Materials and methods

Data collection

The scHDeepInsight framework was developed and validated using scRNA-seq data from 10 published studies [2, 8–16], covering a wide range of tissues (e.g. blood, lung, intestine), as detailed in Supplementary Table S1.

In total, the reference dataset includes over 460,000 cells from healthy donors, providing a robust baseline for investigating immune cell heterogeneity. After rigorous preprocessing and STACAS-based [17] batch correction (as described in the Supplementary_Note), we constructed an integrated reference atlas centered on the top 5000 highly variable genes extracted via Scanpy [18] from the reference datasets, encompassing 15 base immune cell types and more than 50 subtypes. The data integration pipeline follows a systematic workflow from initial study selection through quality control, normalization, and batch correction (Fig. 1a, Supplementary_Note). The resulting low-dimensional embedding of the reference atlas shows distinct clusters corresponding to different immune cell types and subtypes (Fig. 1b). This hierarchical organization preserves immune cell lineage relationships, ensuring clear separation between major lineages while maintaining biological continuity among related subtypes.

The query (test) datasets utilized in this study for the evaluation were sourced from various public databases, each covering different tissues and cell types. The datasets include peripheral blood mononuclear cells (PBMCs), lung and liver tissue samples, and bone marrow cells (Table 1). They were generated using a range of scRNA-seq protocols, including multiple versions of 10× Genomics [19] (3′ v2, v3) and MARS-seq [20].

Overview of scHDeepInsight

scHDeepInsight is a computational framework for single-cell data that enables hierarchical immune cell annotation, rare cell-population detection, and biological interpretation through SHAP analysis. The framework applies a structured multilevel classification process that reflects known immune cell lineage relationships and supports fine-grained subtype resolution.

As illustrated in Fig. 2, the workflow consists of two main stages: (i) Training phase—gene expression vectors are transformed into 2D images via DeepInsight followed by random masking, and then processed by a CNN feature extractor and the hierarchical classification architecture with integrated loss function for model training. (ii) Application phase—scHDeepInsight performs batch effect correction on query datasets by transforming them using the analogous image conversion procedure followed by the hierarchically-trained CNN to predict both primary cell types and their subtypes.

During the training phase, scHDeepInsight transforms preprocessed gene expression profiles into two-dimensional gene expression images constructed from a reference atlas. These images are then used to train a CNN based on the EfficientNet-B5 architecture. Unlike the original scDeepInsight, which employs a flat classification approach, scHDeepInsight integrates a multilevel loss function that preserves and reinforces the hierarchical relationships between cell types during classification. The architecture enables a two-stage classification process: first identifying the primary immune cell types, then further refining subtype classification within each lineage.

In the application (test) phase, scHDeepInsight is applied to independent query datasets. The batch-corrected datasets are then fed into the pretrained CNN, which outputs predicted probabilities for both primary cell types and subtypes. For rare cell identification, scHDeepInsight employs a probability-based detection mechanism that analyzes the discrepancy patterns between base-type and subtype prediction confidences.

Conversion of tabular data into images

To transform high-dimensional scRNA-seq data into CNN-compatible 2D image representations, pyDeepInsight tool (https://github.com/alok-ai-lab/pyDeepInsight), based on the DeepInsight framework [23], is employed. In this approach, each gene is mapped to a specific pixel location, and pixel intensity reflects the corresponding gene expression level. Manifold techniques, such as t-SNE and Uniform Manifold Approximation and Projection (UMAP) [24], are applied with optimized perplexity settings to project the data into a 2D space, positioning genes with similar expression patterns in close proximity. These coordinates are then converted into pixel positions, creating images in which the intensities represent gene expression patterns. Considering that query datasets may lack certain genes present in the reference atlas used for training, random masking is introduced to the generated images, thereby injecting controlled noise to enhance robustness against missing gene features in query data. The resulting 2D representations ($224 \times 224 \times 3$) are then utilized by an EfficientNet-B5 CNN for feature extraction and cell type classification. Furthermore, to ensure an optimal gene-to-pixel assignment and eliminate potential collisions where multiple genes would be mapped to the same pixel location, a Linear Sum Assignment (LSA) algorithm was applied [25]. This image-based approach leverages the CNN's capacity for feature learning while

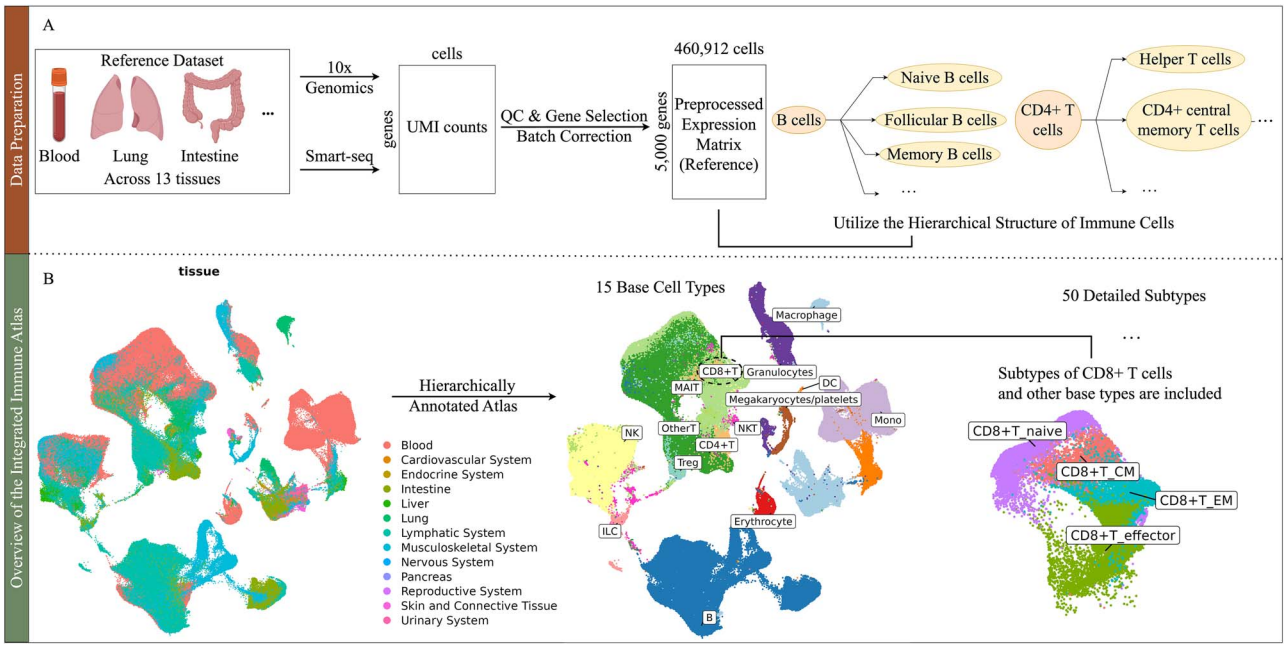

Figure 1. (a) Schematic of data integration, from study selection through quality control (QC), normalization, and batch correction. (b) Low-dimensional visualization of the final reference atlas comprising 15 base cell types and over 50 subtypes.

Table 1. Summary of datasets used for benchmarking scHDeepInsight and the other tools. Gene coverage percentage indicates the proportion of genes in each query dataset that overlaps with the 5000 highly variable genes used in the model training.

Dataset	Tissue	Protocol	Cell numbers	Gene coverage	Cell types
Arunachalam *et al.*, 2020	Blood	10× Multiome 3′ v2, v3	25,954	99.8%	22
Lee *et al.*, 2020	Blood	10× Multiome 3′ v3	16,298	99.5%	23
Schulte-Schrepping *et al.*, 2020	Blood	10× Multiome 3′ v2, v3	45,787	99.6%	24
Adams *et al.*, 2020	Lung	10× Multiome 3′ v2	10,934	96.4%	21
Travaglini *et al.*, 2020	Lung	10× Multiome 3′ v3	35,699	94.6%	24
MacParland *et al.*, 2018	Liver	10× Multiome 3′ v2	4,436	91.2%	18
Ledergor *et al.*, 2018	Bone Marrow	MARS-seq	6,367	96.3%	6

preserving the spatial structure among genes, thereby capturing the underlying gene–gene interaction patterns.

Adaptive hierarchical focal loss

For effective hierarchical classification, the model must address class imbalance, which is commonly observed in single-cell RNA-seq datasets where rare immune cell subtypes can be underrepresented by orders of magnitude compared to abundant populations, while maintaining predictive accuracy across different granularity levels. To achieve this, scHDeepInsight employs an AHFL that extends the focal loss framework [26] to optimize classification at multiple levels of the hierarchy. For both base-type and subtype classification levels, the focal loss $FL(p)$ is defined as:

$$FL(p) = -\sum_{i=1}^{K} y_i (1 - p_i)^{\gamma} \log(p_i) \quad (1)$$

In Equation 1, K represents the total number of possible classes at the respective classification level (15 base types or their corresponding subtypes), p is the predicted probability vector, where $p_i \in [0, 1]$ corresponds to the predicted probability for the ith class, $y_i \in \{0, 1\}$ denotes the ground truth binary label for class i (1 if the sample belongs to class i, 0 otherwise) and γ is the focusing parameter that modulates the influence of well-classified versus hard-to-classify examples. We set $\gamma = 2.0$ in all experiments, adopting the optimal value reported in the original focal loss paper [26] through empirical validation for addressing class imbalance in object detection tasks.

The AHFL adapts the focal loss to operate simultaneously at two hierarchical levels: base-type classification and subtype classification. The total loss function combines these two levels as:

$$L_{total} = \alpha \cdot FL^{base} + (1 - \alpha) \cdot FL^{sub} \quad (2)$$

The variables FL^{base} and FL^{sub} represent the focal losses computed for base-type and subtype classification respectively, calculated using Equation 1 with corresponding labels and predictions. The weighting parameter $\alpha \in [0, 1]$ balances their relative contributions. Unlike static weighting schemes, α is dynamically updated during training based on the relative performance of the two classification levels. The adaptive weighting mechanism is formalized as:

$$\alpha_{t+1} = \beta \cdot \alpha_t + (1 - \beta) \cdot \frac{FL^{base}}{FL^{base} + FL^{sub}} \quad (3)$$

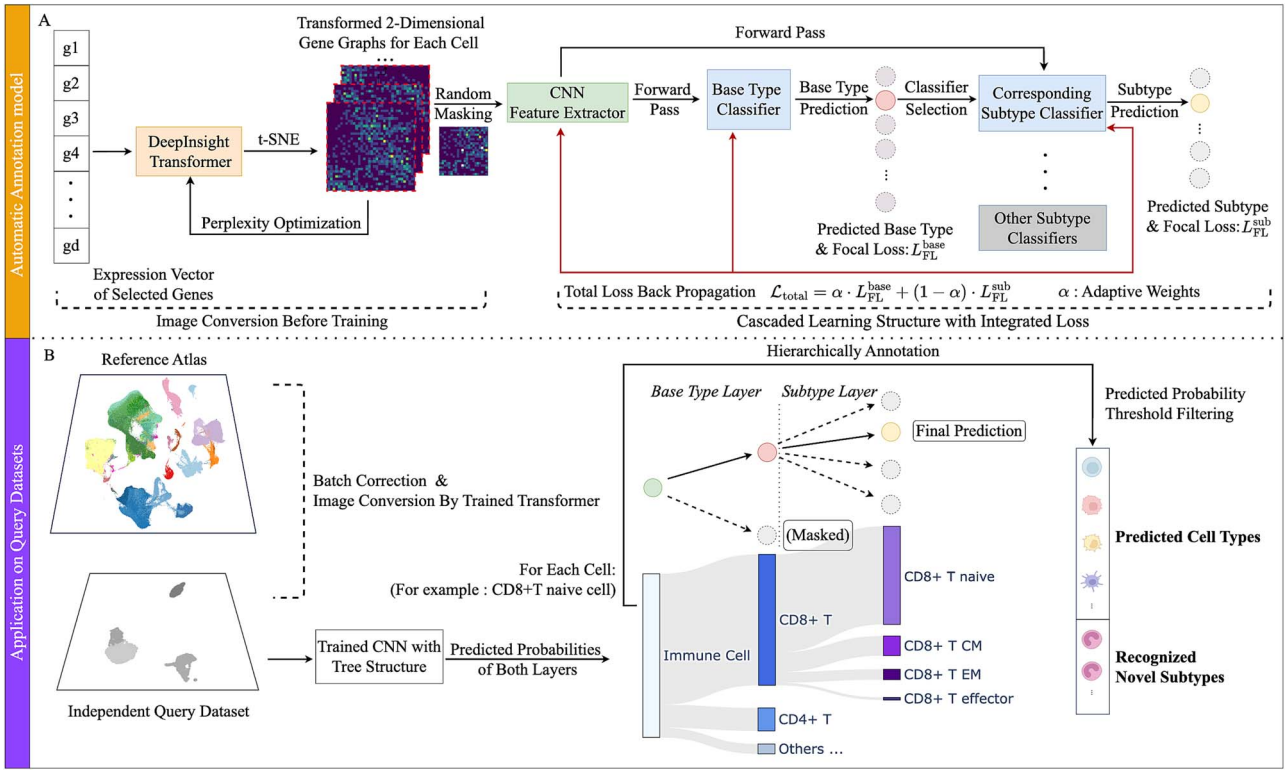

Figure 2. Overview of the scHDeepInsight framework. (a) each cell's gene expression vector is transformed into a 2D image via DeepInsight [5] (t-distributed stochastic neighbor embedding (t-SNE) [21] with perplexity optimization), optionally masked for data augmentation, then fed into an EfficientNet-B5 [22] CNN trained with a multilevel loss to classify cells hierarchically. (b) for query data, after batch correction and image conversion, the trained CNN outputs base-type and subtype probabilities. Masking excludes irrelevant subtypes, enabling hierarchical annotation and uncovering potentially new immune populations.

Here, α_t is the adaptive weight at training step t, and, $\alpha_{t+1}$is the updated weight for the next iteration. The momentum parameter $\beta \in [0, 1]$ controls the smoothness of weight adaptation and is set to 0.9, following standard practices in adaptive weighting schemes [27], to provide stable exponential smoothing that balances historical of α_t values with responsiveness to current loss ratios. The ratio of $FL^{base} / \left(FL^{base} + FL^{sub}\right)$serves as the adaptation signal that drives this dynamic adjustment. The adaptive weighting mechanism dynamically directs the model's focus towards the most challenging hierarchical level during training. When base-type classification becomes more accurate (indicated by lower FL^{base}), the algorithm shifts focus in the next iteration toward improving subtype classification by decreasing α_{t+1}. Conversely, when subtype classification improves substantially, the model allocates more weight to refining base-type predictions. Through joint backpropagation, the framework simultaneously optimizes both classification levels, improving overall accuracy while maintaining consistency with known immune cell hierarchies.

Hierarchical prediction model

scHDeepInsight integrates immune lineage structure directly into its classification architecture and training process. This design allows the model to make structured predictions that reflect known biological organization (the complete hierarchical organization of immune cell types and their relationships is visualized in Supplementary Fig. S1).

After image conversion and feature extraction with CNN, these extracted features are processed through a hierarchical classification pipeline (Fig. 3). The base-type classifier first identifies the broad immune cell category, followed by activation of the corresponding subtype classifier for refined identification within that lineage (detailed architecture illustrated in Supplementary Fig. S2).

During the training phase, a multilevel focal loss function optimizes predictions at both base-type and subtype levels simultaneously, with error signals propagated back through the network to capture hierarchical dependencies between cell types.

During prediction, probability masking sets the probabilities of subtypes outside the predicted primary category to zero, ensuring that downstream classifiers operate only within biologically relevant subtype spaces and reducing misclassification across unrelated lineages. In cases where a base-type has no defined subtypes, it is treated as a terminal leaf in the classification tree. The model bypasses the subtype classifier for such nodes, and the base-type prediction itself serves as the final output. Terminal node treatment preserves hierarchical consistency while accommodating the limited subtype resolution available in current immune reference annotations.

Results

Benchmarking with state-of-the-art methods

To evaluate the performance of scHDeepInsight (Methods), a series of benchmarking experiments were conducted using seven independent query datasets. These datasets were selected to represent a diverse array of tissues, cell types, and disease conditions, providing a testbed for assessing the accuracy, precision, and robustness of scHDeepInsight compared to other state-of-the-art (SOTA) cell annotation methods, including SingleR [1], Azimuth [2], scDeepInsight [4], CellTypist [8], Garnett

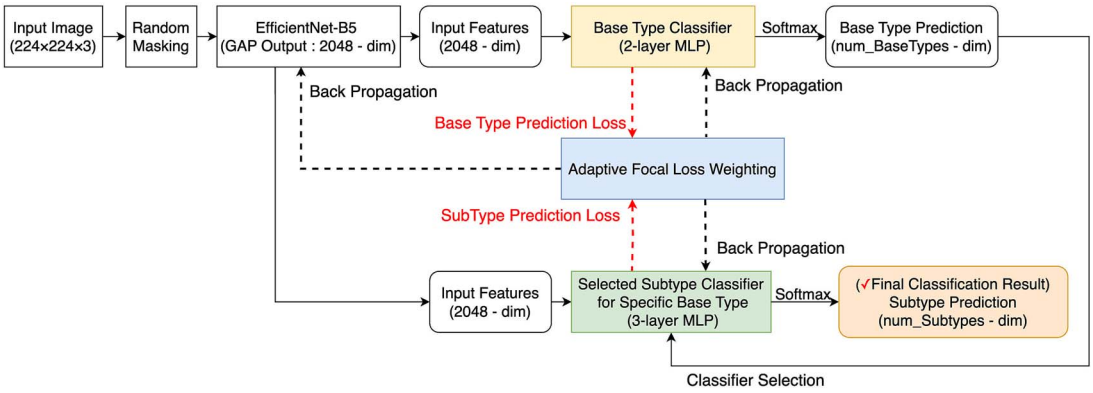

Figure 3. Hierarchical classification model for single-cell type identification. The model uses EfficientNet-B5 to extract features from transformed gene expression images for sequential classification of cell types (base-types) and subtypes. The adaptive focal loss weighting mechanism addresses class imbalance while balancing optimization between base-type and subtype levels.

[28], scType [29], and GPTCellType [30]. The evaluation metrics and technical summaries of these benchmarked methods are provided in the Supplementary_Note.

Benchmarking evaluation across multiple metrics

Comprehensive benchmarking was performed at both base-type and subtype classification levels. At the subtype level, scHDeepInsight demonstrated consistent robust performance, achieving an average accuracy of 93.2% and precision of 91.1% across diverse datasets (Fig. 4A and B; the detailed training and validation accuracy trends are provided in Supplementary Fig. S3). The method achieved an F1-score of 90.5%, reflecting a balance between precision and recall, alongside an area under the precision-recall curve (AUPRC) of 89.7%, a metric particularly suitable for evaluating classification performance on datasets with imbalanced class distributions (Fig. 4C and D). Comparative analysis with scDeepInsight (the next highest performing method) revealed improvements of 5.1% in accuracy, 3.3% in precision, 3.1% in F1-score, and 3.6% in AUPRC (detailed comparative results for all methods are shown in Supplementary Table S2). These improvements across all evaluation metrics indicate scHDeepInsight's enhanced classification capability across the diverse cell types and datasets tested. Technical strategies for rare subtype handling and overfitting prevention are detailed in Supplementary_Note.

At the base-type level, scHDeepInsight achieved 96.8% accuracy, 95.9% precision, 97.0% F1-score, and 93.5% AUPRC (Supplementary Fig. S4). Compared to the second-best method scDeepInsight, this represents improvements of 2.5, 1.6, 2.4, and 2.7 percentage points in accuracy, precision, F1-score, and AUPRC, respectively. The improvements were greater at the subtype level compared to the base-type level across all metrics, indicating that the hierarchical framework provides enhanced benefits for fine-grained classification tasks.

Fine-grained immune cell classification

Owing to the hierarchical classification design and comprehensive reference atlas covering 50 immune cell subtypes, scHDeepInsight demonstrates precise distinction of closely related subtypes. For instance, in the PBMC query dataset (Lee) [31], scHDeepInsight successfully distinguished between closely related subtypes, achieving high classification accuracy with minimal misclassification errors (Fig. 5A). In the labial gland dataset Pranzatelli [32], scHDeepInsight successfully recovered the distinct immunoglobulin-based subtypes (IgA+ and IgG+) of plasma cells originally labeled by experts (Fig. 5B and D), whereas, other supervised annotation methods failed to maintain this resolution, instead grouping all plasma cells into a single broad category (Fig. 5C). By delineating these subtle transcriptional differences, scHDeepInsight highlights the distinct advantage provided by the hierarchical classification framework in accurately resolving closely related immune cell populations.

Recognition of rare cell types

The hierarchical classification framework of scHDeepInsight enables novel cell population detection through analysis of prediction probability patterns: the model assigns probability scores at both base and subtype levels and creates a quantitative signature that reflects cellular identity with greater nuance than conventional binary classification approaches (Methods).

Validation using CITE-seq [33] data from a brain immune cell atlas [34] demonstrates the model's ability to detect novel cell populations through hierarchical probability patterns. Ground truth annotations reveal distinct microglia and macrophage populations (Fig. 6A), which are supported by differential surface protein expression: microglia exhibit higher TMEM119 expression while macrophages show elevated CD163, CD206, and CD86 (Fig. 6B). Our model predictions successfully identified these populations as macrophage lineage cells (Fig. 6C), with hierarchical probability difference analysis revealing the key signature of novel cell states: regions with high base-type probabilities for this lineage but low subtype-level confidence (Fig. 6D). This pattern, particularly evident in microglia-enriched regions, indicates cells that belong to the broader myeloid lineage but represent subtypes beyond the training atlas scope.

These findings illustrate that scHDeepInsight's probabilistic output supports a graded representation of cellular identity, capturing both canonical immune subtypes and cell populations that diverge from transcriptional profiles represented in the reference atlas.

Analysis of SHAP-based feature importance for immune cell classification

To identify discriminative gene features critical for accurate cell type classification, scHDeepInsight employs SHAP analysis separately for each immune cell type and subtype, whose absolute value quantitatively measures the contribution of each gene to the model's classification decisions. The resulting values represent the average feature importance across cells within each

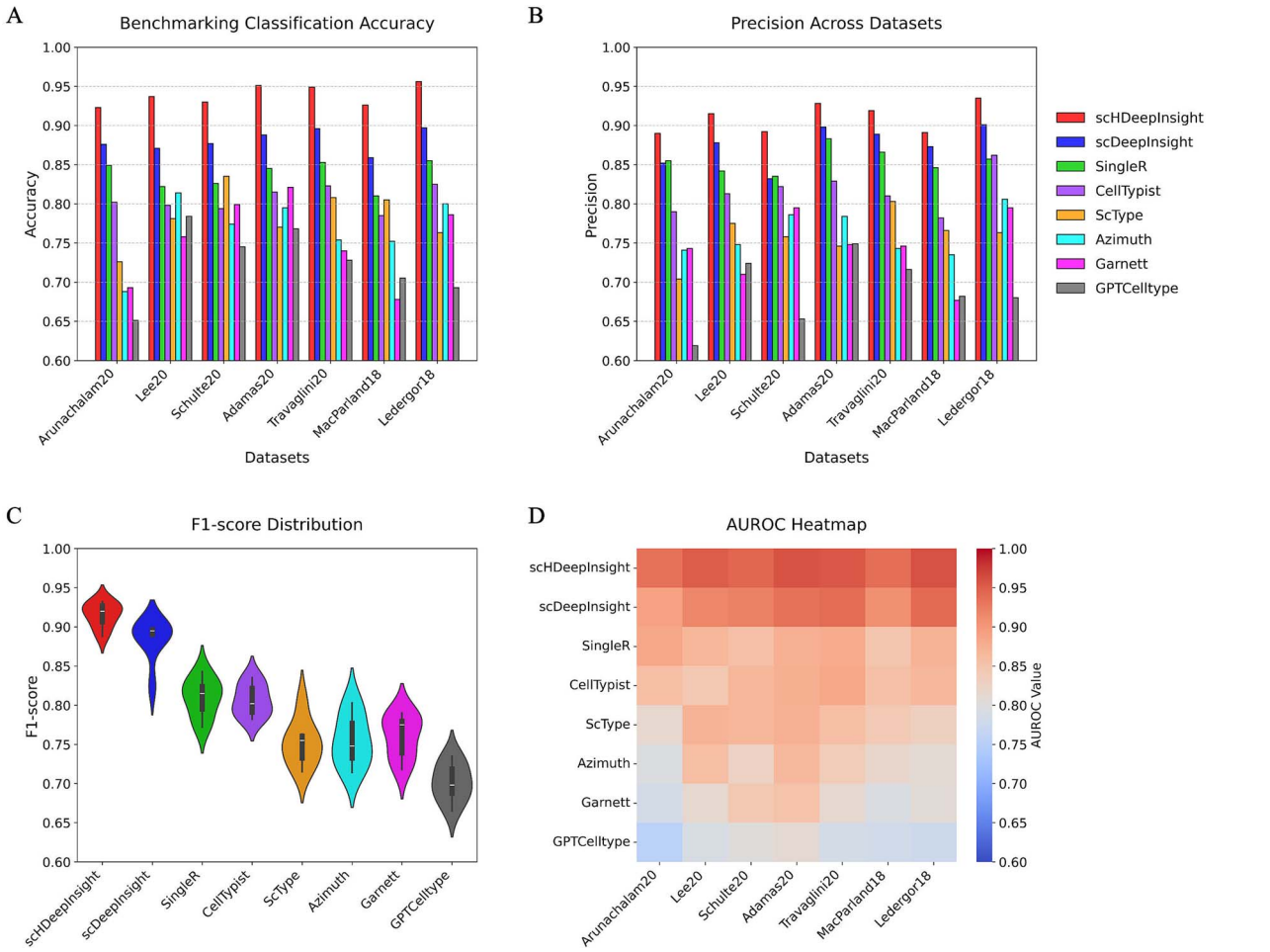

Figure 4. Benchmarking results. (a) Accuracies across seven datasets for all classification methods (scHDeepInsight in red). (b) Precisions for the same methods. (c) F1-score distributions as violin plots revealing median performance and variability across methods. (d) AUPRC heatmap displaying classification strength with color intensity corresponding to performance values.

class, thereby enabling identification of both lineage-defining and subtype-specific gene signatures. As shown in Fig. 7, the model identifies distinct gene importance patterns across different immune cell populations, with both expected canonical markers and subtype-specific genes. For instance CD8+ T-cells (Fig. 7A), canonical T-cell markers such as CD8A and CD3E demonstrate high positive SHAP values, reaffirming their expected roles in defining cytotoxic T-cell identity. Similarly, the cytotoxic effector molecule GNLY and transcription factors associated with T-cell activation contribute positively to classification decisions. Notably, SHAP analysis captures genes with reduced expression that characterize cell subtypes. For example, CD72, which is typically downregulated during plasma cell differentiation [35], is identified as an important discriminative feature in IgM-expressing plasma cells (Fig. 7B), consistent with known biological phenomena. The model also successfully identifies common gene features of plasma cells across subtypes, such as immunoglobulin light chain genes (IGLC2, IGLC3), while capturing the isotype-specific differences that biologically distinguish these populations: IGHA1 for IgA-expressing plasma cells (Fig. 7C), IGHM for IgM-expressing plasma cells (Fig. 7B), and IGHG1/IGHG2 for IgG-expressing plasma cells (Fig. 7D). These findings validate that the detailed subtype classifier within the hierarchical classification framework in scHDeepInsight can capture both lineage-shared and subtype-specific genes, further supporting its ability to discriminate between closely related cellular states from an immunological perspective.

Discussion

Accurate identification of broad cell types as well as subtypes of immune cells is a prerequisite for understanding their diverse roles in complex biological systems and diseases, such as cancer and chronic conditions. scHDeepInsight advances immune cell annotation in scRNA-seq data by implementing a hierarchical classification framework that reflects the biological organization of immune lineages. In contrast to conventional annotation strategies that treat all cell types as independent categories, scHDeepInsight leverages shared representations across hierarchical levels, thereby improving the resolution of both major lineages and transcriptionally similar subtypes.

Beyond improvements in predictive performance, scHDeepInsight incorporates several innovations designed to enhance biological interpretability. The AHFL dynamically balances optimization across classification levels, increasing sensitivity to both coarse-grained and fine-grained cell distinctions. By transforming gene expression profiles into spatially organized images, a CNN can learn complex expression patterns from inherently nonspatial data. Additionally, the model produces structured probability distributions across the hierarchy, providing a continuous and

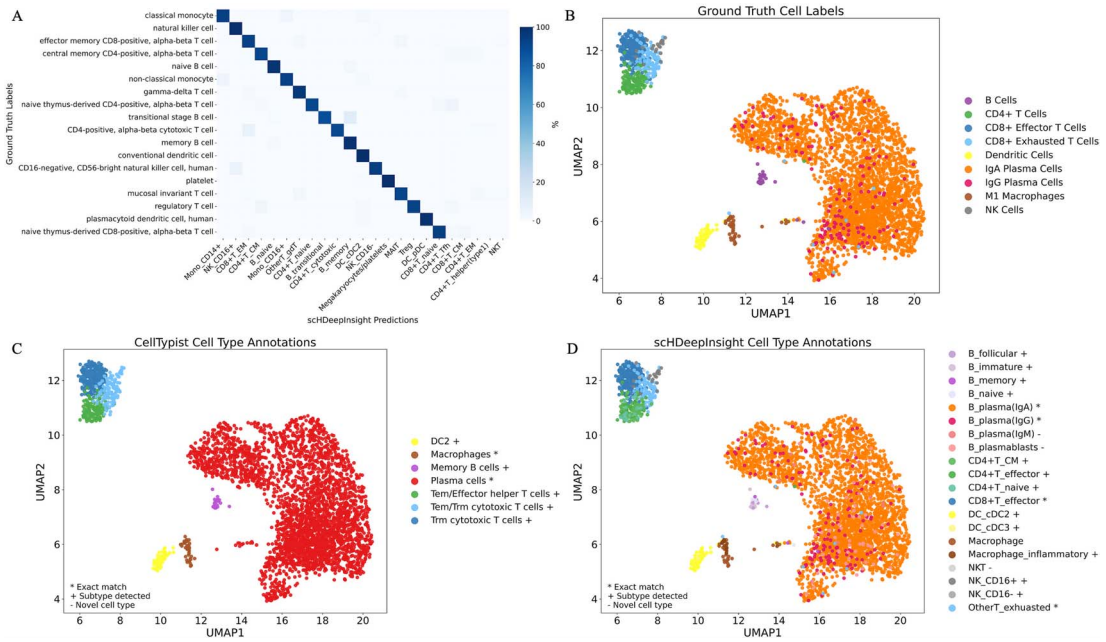

Figure 5. Classification results demonstrating immune cell subtype identification capabilities. (a) the confusion matrix of scHDeepInsight predictions on the Lee dataset. (b) UMAP visualization of the labial gland dataset with original expert annotations, including detailed plasma cell subtypes. (c) Annotation results with CellTypist on the labial gland dataset. (d) scHDeepInsight classification results on the labial gland dataset, with symbols indicating correspondence to the original annotations.

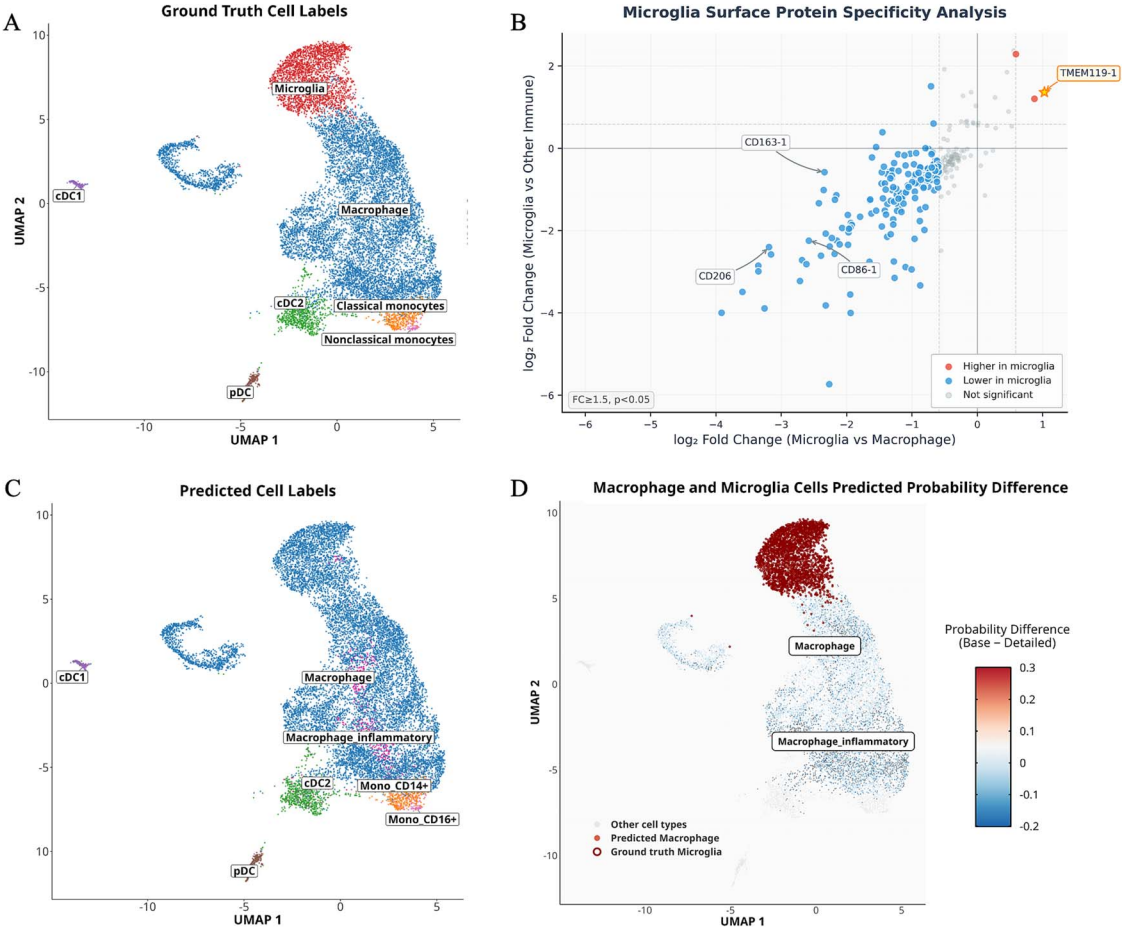

Figure 6. Identification of novel cell populations through hierarchical probability analysis validated with CITE-seq data. (a) UMAP visualization with ground truth cell annotations displaying microglia and macrophage populations. (b) Surface protein expression analysis demonstrating differential expression of microglia-specific (TMEM119) and macrophage-specific (CD163, CD206, CD86) markers. (c) scHDeepInsight classification results for the corresponding cell populations. (d) Hierarchical probability difference analysis (base-type minus subtype probabilities) highlighting potential novel cell states.

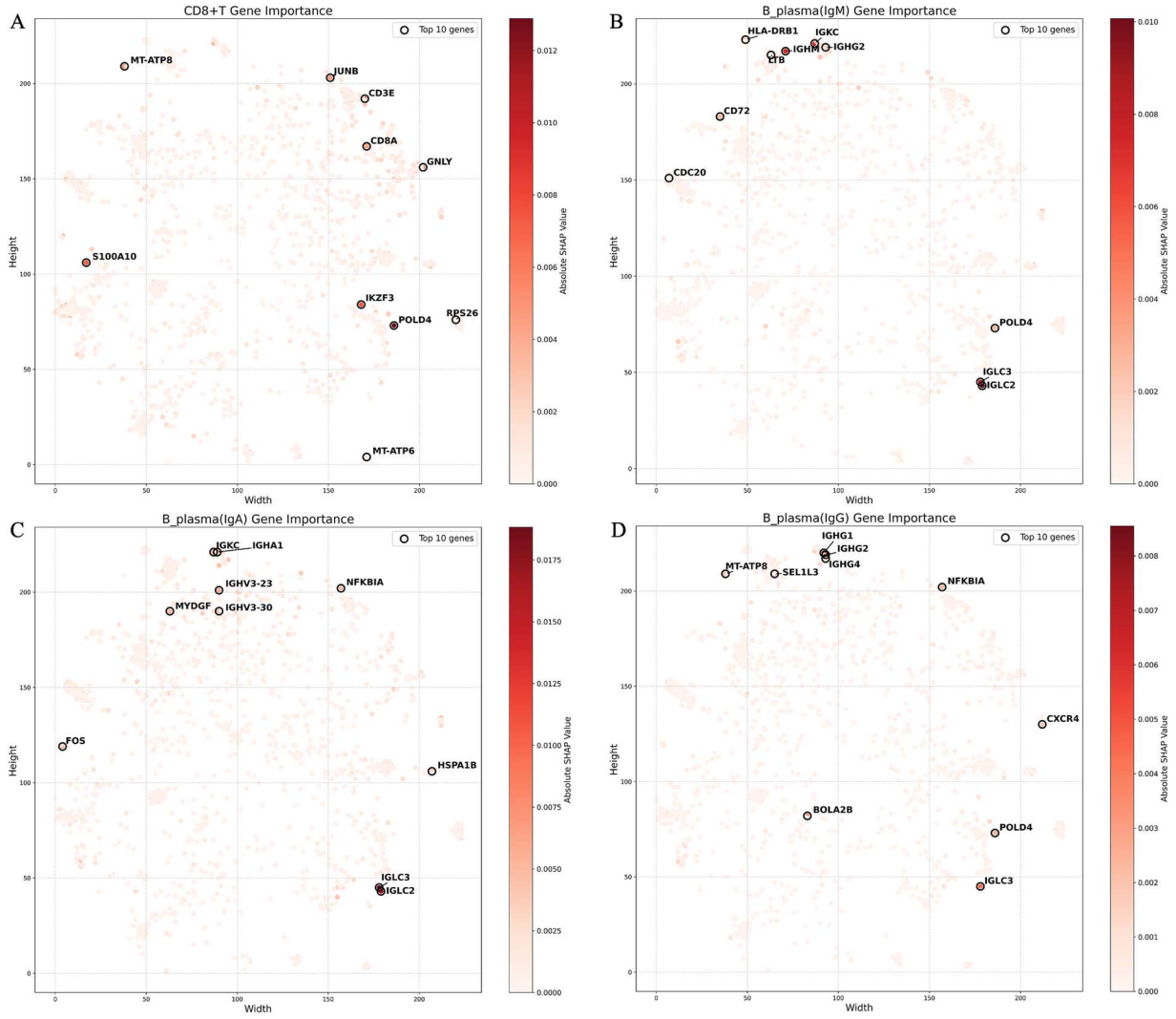

Figure 7. SHAP-based gene importance analysis for immune cell classification. (a) CD8$^+$ T-cell gene importance, highlighting CD8A, CD3E, and effector molecules like GNLY. (b) IgM-expressing plasma cell gene importance, with high values for IGHM. (c) IgA-expressing plasma cell gene importance, showing elevated SHAP values for IGHA1 and associated light chain genes. (d) IgG-expressing plasma cell gene importance, featuring IGHG1, IGHG2.

interpretable representation of cellular identity. This probabilistic framework enables identification of cells with ambiguous subtype assignments, which may represent novel cell populations, or context-specific immune phenotypes. Such cases are exemplified in the glioblastoma dataset, where cells demonstrate high confidence in base-type classification but exhibit uncertainty in subtype assignment. Model interpretability is further strengthened by SHAP-based feature analysis, which reveals gene expression patterns associated with both lineage specification and functional divergence. By identifying canonical markers alongside context-specific regulatory features, the model not only enhances confidence in subtype predictions but also provides insights into the gene expression patterns underlying immune heterogeneity. However, interpretation of SHAP-based feature importance should take into account the inherent gene dropout and sparsity in scRNA-seq data, particularly when evaluating negative feature contributions.

While scHDeepInsight presents significant advances, certain challenges remain. Classification performance is fundamentally linked to the resolution and completeness of the reference atlas, which may limit annotation in cases of sparse or incomplete reference data. Computational benchmarking on a 10,000-cell scRNA-seq dataset demonstrates that scHDeepInsight achieves 72 s runtime and 6.2 GB peak memory usage, which is comparable to scDeepInsight and more efficient than several baseline methods including SingleR and CellTypist (Supplementary Fig. S5). Despite its additional computational cost compared to lightweight methods such as scType, the substantial performance improvements justify the overhead for applications requiring high-resolution immune cell profiling.

In future work, further development of scHDeepInsight will focus on extending its generalizability and biological scope. While the current model is trained on immune cell types, the hierarchical classification architecture is generalizable to other cellular contexts. To demonstrate this potential, we trained a separate model on a breast cancer dataset [36] (GSE176078, $n = 100{,}064$ cells) containing both immune and nonimmune cell types. Cross-validation results show that the framework achieves robust performance across diverse cell populations including epithelial, stromal, and immune cells, with an overall accuracy of 90.7% (Supplementary Fig. S6), indicating the architectural adaptability of scHDeepInsight to diverse cellular systems beyond

the immune compartment. Additionally, integration of multi-omics modalities, such as CITE-seq and spatial transcriptomics, may improve subtype resolution by incorporating protein-level and spatial context information. These complementary data types would also enhance the reliability and interpretability of SHAP-based feature importance analysis by providing orthogonal validation of gene expression patterns identified as discriminative features. Also, self-supervised learning approaches may facilitate feature extraction from unlabeled data, enabling discovery of novel immune states without prior annotation. Incorporating pseudotime inference would allow dynamic modeling of differentiation trajectories, extending the framework beyond static classification. Finally, transfer learning strategies offer a promising path toward improving adaptability across tissues or species with minimal retraining, broadening the applicability of scHDeepInsight to diverse biological contexts. As a further extension, incorporating nonimmune cell populations into the hierarchical framework would enable unified modeling of diverse cellular systems beyond the immune compartment.

Conclusion

scHDeepInsight provides a deep learning framework for hierarchical immune cell classification using scRNA-seq data, incorporating CNN architectures and immune lineage structures to enhance accuracy and resolution of cell type annotation. Through the integrated use of the multilevel loss function and adaptive weighting approach, it achieves accurate identification of both common immune lineages and closely related subtypes while maintaining their hierarchical relationships. The integration of batch effect correction ensures consistent performance across datasets under diverse experimental conditions. The comprehensive benchmarking demonstrated that scHDeepInsight outperformed the existing annotation methods across the multiple performance metrics, particularly in distinguishing the closely related immune cell subtypes within complex tissues. In addition, scHDeepInsight revealed novel immune populations through hierarchical analysis on prediction probabilities, highlighting its potential for uncovering biologically relevant cellular diversity beyond canonical annotations. This improved resolution enables precise characterizations of cellular heterogeneities in immunological research.

As the single-cell technologies continue to evolve, the hierarchical classification approach implemented in scHDeepInsight will be increasingly valuable for advancing our understanding of immune cell diversity and functional specialization. The framework's proven adaptability to nonimmune cell types further extends its utility for broad single-cell annotation applications beyond immunological research.

Key Points

- scHDeepInsight introduces a deep learning framework that leverages biologically informed hierarchical classification and an adaptive hierarchical focal loss, which is demonstrated to deliver superior performance for cell type detection in single-cell RNA-seq data.
- This architecture addresses the limitations of flat classification, more specifically for the tasks of high-resolution annotation of transcriptionally similar or rare immune subtypes.
- Evaluation on seven independent datasets achieved the highest overall accuracy (93.2%) among seven alternative methods, surpassing the second-best performing method by 5.1 percentage points, and outperforming all others in precision, F1-score, and AUPRC.
- Feature importance analysis using SHAP identified both canonical and subtype-specific gene signatures consistent with established immune cell hierarchies, validating the biological relevance of these predictions and showcasing how this method can improve functional interpretation of immune heterogeneity.
- The complete pipeline is available as a fully functional Python package. The source code and reference datasets are also made available to facilitate reproducibility and reuse.

Acknowledgements

The results shown in this paper are in part based upon publicly available single-cell datasets from Gene Expression Omnibus (GEO), the European Genome-phenome Archive (EGA), and the CELLxGENE portal. We appreciate the researchers and consortia responsible for generating and openly sharing these valuable datasets, enabling comprehensive benchmarking and validation performed in this study.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Author contributions

SJ implemented the whole pipeline, evaluated the performance, and wrote the first draft and contributed to the subsequent versions of the manuscript. AL advised for the model and the evaluation, and contributed to the manuscript writeups. KAB checked the model, and helped in the manuscript writeup. AS and TT perceived, supervised, and contributed to the manuscript writeups. All authors read and approved the manuscript.

Conflict of interest: None declared.

Funding

This work was partly funded by JSPS KAKENHI Grant Numbers 24 K15175, 25KJ1104, JP20H03240 and JP25K02261, Japan and JST CREST Grant Number JPMJCR2231, Japan.

Data availability

All datasets used in this study were obtained from publicly available repositories. The Lee (GSE149689) [31], Pranzatelli (phs002446) [32], Adamas [37] (GSE134692), Travaglini [38] (EGAS00001004344), Arunachalam (GSE155673) [39], Schulte (EGAS00001004571) [40], MacParland (GSE115469) [41] and Ledergor (GSE117156) [42] datasets are also accessible via the CELLxGENE [43] portal: https://cellxgene.cziscience.com/

datasets. The integrated reference dataset used for model training is available at Figshare (https://doi.org/10.6084/m9.figshare.28831010).

Code availability

The source code is publicly available at the GitHub repository: https://github.com/shangruJia/scHDeepInsight. A packaged Python library is also accessible via PyPI (https://pypi.org/project/SCHdeepinsight/) for straightforward installation and use.

References

1. Aran D, Looney AP, Liu L, *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* 2019;**20**:163–72. https://doi.org/10.1038/s41590-018-0276-y.
2. Hao Y, Hao S, Andersen-Nissen E, *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573–3587.e29. https://doi.org/10.1016/j.cell.2021.04.048.
3. Kiselev VY, Yiu A, Hemberg M. Scmap: Projection of single-cell RNA-seq data across data sets. *Nat Methods* 2018;**15**:359–62. https://doi.org/10.1038/nmeth.4644.
4. Jia S, Lysenko A, Boroevich KA, *et al.* scDeepInsight: A supervised cell-type identification method for scRNA-seq data with deep learning. *Brief Bioinform* 2023;**24**:bbad266.
5. Sharma A, Vans E, Shigemizu D, *et al.* DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Sci Rep* 2019;**9**:11399. https://doi.org/10.1038/s41598-019-47765-6.
6. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJ, Bottou L, Weinberger KQ (eds.), *Adv Neural Inf Proces Syst 25 (NIPS)*. NY 12571 USA: Curran Associates, Inc., 2012, 1097–105.
7. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Guyon I, Von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds.), *Adv Neural Inf Proces Syst 30 (NIPS)*. NY 12571 USA: Curran Associates, Inc., 2017, 4768–77.
8. Domínguez Conde C, Xu C, Jarvis L, *et al.* Cross-tissue immune cell analysis reveals tissue-specific features in humans. *Science* 2022;**376**:eabl5197. https://doi.org/10.1126/science.abl5197.
9. Chan JM, Quintanal-Villalonga Á, Gao VR, *et al.* Signatures of plasticity, metastasis, and immunosuppression in an atlas of human small cell lung cancer. *Cancer Cell* 2021;**39**:1479–1496.e18. https://doi.org/10.1016/j.ccell.2021.09.008.
10. Elmentaite R, Kumasaka N, Roberts K, *et al.* Cells of the human intestinal tract mapped across space and time. *Nature* 2021;**597**:250–5. https://doi.org/10.1038/s41586-021-03852-1.
11. Rozenblatt-Rosen O, Stubbington MJT, Regev A, *et al.* The human cell atlas: From vision to reality. *Nature* 2017;**550**:451–3. https://doi.org/10.1038/550451a.
12. James KR, Gomes T, Elmentaite R, *et al.* Distinct microbial and immune niches of the human colon. *Nat Immunol* 2020;**21**:343–53. https://doi.org/10.1038/s41590-020-0602-z.
13. Jardine L, Webb S, Goh I, *et al.* Blood and immune development in human fetal bone marrow and down syndrome. *Nature* 2021;**598**:327–31. https://doi.org/10.1038/s41586-021-03929-x.
14. Madissoon E, Wilbrey-Clark A, Miragaia RJ, *et al.* scRNA-seq assessment of the human lung, spleen, and esophagus tissue stability after cold preservation. *Genome Biol* 2019;**21**:1. https://doi.org/10.1186/s13059-019-1906-x.
15. Popescu D-M, Botting RA, Stephenson E, *et al.* Decoding human fetal liver haematopoiesis. *Nature* 2019;**574**:365–71. https://doi.org/10.1038/s41586-019-1652-y.
16. The tabula sapiens consortium. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* 2022;**376**:eabl4896. https://doi.org/10.1126/science.abl4896.
17. Andreatta M, Hérault L, Gueguen P, *et al.* Semi-supervised integration of single-cell transcriptomics data. *Nat Commun* 2024;**15**:872. https://doi.org/10.1038/s41467-024-45240-z.
18. Wolf FA, Angerer P, Theis FJ. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol* 2018;**19**:15. https://doi.org/10.1186/s13059-017-1382-0.
19. Zheng GXY, Terry JM, Belgrader P, *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* 2017;**8**:14049. https://doi.org/10.1038/ncomms14049.
20. Jaitin DA, Kenigsberg E, Keren-Shaul H, *et al.* Massively parallel single-cell RNA-Seq for marker-free decomposition of tissues into cell types. *Science* 2014;**343**:776–9. https://doi.org/10.1126/science.1247651.
21. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605.
22. Tan M, Le QV. EfficientNet: Rethinking model scaling for convolutional neural networks. In: Chaudhuri K, Salakhutdinov R (eds.), *Proceedings of the 36th International Conference on Machine Learning*. PMLR. Cambridge MA: JMLR, 2019;**97**:6105–14.
23. Sharma A, Lysenko A, Jia S, *et al.* Advances in AI and machine learning for predictive medicine. *J Hum Genet* 2024;**69**:487–97. https://doi.org/10.1038/s10038-024-01231-y.
24. McInnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection. *J Open Source Softw* 2018;**3**:861. https://doi.org/10.21105/joss.00861.
25. Virtanen P, Gommers R, Oliphant TE, *et al.* SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nat Methods* 2020;**17**:261–72. https://doi.org/10.1038/s41592-019-0686-2.
26. Lin T-Y, Goyal P, Girshick R, *et al.* Focal loss for dense object detection. *IEEE Trans Pattern Anal Mach Intell* 2020;**42**:318–27. https://doi.org/10.1109/TPAMI.2018.2858826.
27. Morales D, Vogels T, Hendrikx H. Exponential moving average of weights in deep learning: Dynamics and benefits. *Trans Mach Learn Res* 2024.
28. Pliner HA, Shendure J, Trapnell C. Supervised classification enables rapid annotation of cell atlases. *Nat Methods* 2019;**16**:983–6. https://doi.org/10.1038/s41592-019-0535-3.
29. Ianevski A, Giri AK, Aittokallio T. Fully-automated and ultra-fast cell-type identification using specific marker combinations from single-cell transcriptomic data. *Nat Commun* 2022;**13**:1246. https://doi.org/10.1038/s41467-022-28803-w.
30. Hou W, Ji Z. Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nat Methods* 2024;**21**:1462–5. https://doi.org/10.1038/s41592-024-02235-4.
31. Lee JS, Park S, Jeong HW, *et al.* Immunophenotyping of COVID-19 and influenza highlights the role of type I interferons in development of severe COVID-19. *Sci Immunol* 2020;**5**:eabd1554. https://doi.org/10.1126/sciimmunol.abd1554.
32. Pranzatelli TJF, Perez P, Ku A, *et al.* GZMK+CD8+ T-cells target a specific acinar cell type in Sjögren's disease. *Res Sq* [Preprint] 2024 Jul 11:rs.3.rs-3601404. https://doi.org/10.21203/rs.3.rs-3601404/v2.
33. Stoeckius M, Hafemeister C, Stephenson W, *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* 2017;**14**:865–8. https://doi.org/10.1038/nmeth.4380.
34. Pombo Antunes AR, Scheyltjens I, Lodi F, *et al.* Single-cell profiling of myeloid cells in glioblastoma across species and

disease stage reveals macrophage competition and specialization. *Nat Neurosci* 2021;**24**:595–610. https://doi.org/10.1038/s41593-020-00789-y.
35. Wu HJ, Bondada S. Positive and negative roles of CD72 in B cell function. *Immunol Res* 2002;**25**:155–66. https://doi.org/10.1385/IR:25:2:155.
36. Wu SZ, Al-Eryani G, Roden D, *et al.* A single-cell and spatially resolved atlas of human breast cancers. *Nat Genet* 2021;**53**:1334–47.
37. Adams TS, Schupp JC, Poli S, *et al.* Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. *Sci Adv* 2020;**6**:eaba1983.
38. Travaglini KJ, Nabhan AN, Penland L, *et al.* A molecular cell atlas of the human lung from single cell RNA sequencing. *Nature* 2020;**587**:619–25. https://doi.org/10.1038/s41586-020-2922-4.
39. Arunachalam PS, Wimmers F, Mok CKP, *et al.* Systems biological assessment of immunity to mild versus severe COVID-19 infection in humans. *Science* 2020;**369**:1210–20. https://doi.org/10.1126/science.abc6261.
40. Schulte-Schrepping J, Reusch N, Paclik D, *et al.* Severe COVID-19 is marked by a dysregulated myeloid cell compartment. *Cell* 2020;**182**:1419–1440.e23. https://doi.org/10.1016/j.cell.2020.08.001.
41. MacParland SA, Liu JC, Ma X-Z, *et al.* Single cell RNA sequencing of human liver reveals distinct intrahepatic macrophage populations. *Nat Commun* 2018;**9**:4383.
42. Ledergor G, Weiner A, Zada M, *et al.* Single cell dissection of plasma cell heterogeneity in symptomatic and asymptomatic myeloma. *Nat Med* 2018;**24**:1867–76. https://doi.org/10.1038/s41591-018-0269-2.
43. Abdulla S, Aevermann B, Assis P, *et al.* CZ CELLxGENE discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res* 2025;**53**:D886–900. https://doi.org/10.1093/nar/gkae1142.

© The Author(s) 2025. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Briefings in Bioinformatics, 2025, **26(5)**, bbaf523
https://doi.org/10.1093/bib/bbaf523
Problem Solving Protocol