

Enhancing Cell Type Annotation with Imperfect Single-Cell RNA-Seq References

A Siamese Network Pipeline for Equitable and Scalable Genomic Research

Mohammad Razib Mustafiz^{1*}, Chandan Kumar Roy^{2*}, Redwan Hossain³, Akram Hossain⁴, Khaled Mohsin⁵

¹ School of Computing, Dublin City University, Dublin 9, Ireland | ² Department of Statistics and Data Science, Jahangirnagar University, Bangladesh | ³ Department of Software Engineering, Daffodil International University, Bangladesh

⁴ Department of Statistics and Data Science, Jahangirnagar University, Bangladesh | ⁵ Department of Cardiology, St James's Hospital, Trinity College Dublin, Ireland

* Corresponding author: mohammad.mustafiz2@mail.dcu.ie

Article Info	ABSTRACT
Manuscript Status: Methodological & policy analysis. Empirical benchmarking against standard datasets (PBMC, pancreas, tumour micro-environment atlases) is underway; quantitative performance tables (accuracy, precision, recall, F1) will be appended to the final release version. Keywords: single-cell RNA-seq; cell type annotation; Siamese networks; deep learning; imperfect reference data; global health equity; evidence-based policy	Single-cell RNA sequencing (scRNA-seq) has transformed biomedical research by revealing cellular diversity at unprecedented resolution. However, the accuracy of computational cell type annotation is fundamentally constrained by the quality of reference datasets, which are frequently noisy, incomplete, or poorly representative of the populations under study. This paper presents a Siamese network-based pipeline that addresses this constraint through deep learning techniques specifically designed to operate under conditions of reference data imperfection. The methodology employs twin subnetworks with shared weights to learn similarity representations across paired gene expression profiles, enabling robust cell type classification even where reference annotations are suboptimal. The analytical framework encompasses systematic dataset selection, normalization and noise-reduction preprocessing, balanced batch training, and evaluation through accuracy, precision, recall, and F1 metrics benchmarked against established methods. The findings, framed analytically against current literature, indicate that Siamese architectures offer meaningful advantages in handling imperfect data, scaling to large datasets, and performing under limited labelled conditions. The paper's broader contribution lies in its policy relevance: by lowering the dependence on pristine, resource-intensive reference datasets, such pipelines can democratize access to advanced genomic analysis for research institutions in low- and middle-income countries, strengthening local disease surveillance, oncology research, and public health decision-making. The paper concludes with strategic recommendations for funders, policymakers, and practitioners seeking to translate deep learning genomic tools into equitable health outcomes.

Institutional publication of the Centre for Humanitarian Research and Social Development (CHRSD Foundation), Research Paper Series.

1. INTRODUCTION AND CONTEXT

1.1 Problem Statement and Background

Single-cell RNA sequencing (scRNA-seq) technology has revolutionized our understanding of cellular diversity, enabling researchers to profile the transcriptome of individual cells and thereby decode the composition of complex tissues (Shalek et al., 2013; Puram et al., 2017). A critical bottleneck in the scRNA-seq analytical workflow is cell type annotation — the assignment of biologically meaningful labels to clusters of cells. Annotation quality determines the reliability of downstream insights in oncology, immunology, cardiovascular research, and developmental biology (Paik et al., 2020; Karlsson et al., 2021).

Yet annotation remains fundamentally constrained by the quality of reference data. Reference atlases are frequently imperfect: they contain incomplete or erroneous labels, uneven coverage across cell types, batch effects, and systematic under-representation of non-European and non-model-organism populations (Kiselev et al., 2019; Zhang et al., 2024). Manual curation by domain experts, while accurate, is time-consuming, subjective, and difficult to scale (Díaz-Mejía et al., 2019). Computational and machine learning approaches — including scmap, scCATCH, CHETAH, and scSorter — offer speed but remain sensitive to the noise and incompleteness of the references on which they depend (Kiselev et al., 2018; Shao et al., 2020; de Kanter et al., 2019; Guo & Li, 2021). The emergence of foundation models such as scGPT and scFoundation marks substantial progress, but these approaches demand very large-scale training data and computational infrastructure that many institutions — particularly in low- and middle-income countries (LMICs) — cannot access (Cui et al., 2024; Hao et al., 2024; Hou & Ji, 2024).

1.2 Objectives and Scope

This paper has four interrelated objectives: (i) to present a Siamese network-based pipeline for cell type annotation engineered specifically to tolerate imperfect scRNA-seq reference data; (ii) to articulate the methodological pipeline from dataset selection and preprocessing through training and evaluation; (iii) to analyze the expected performance characteristics of the approach relative to existing annotation methods, grounded in the architecture's design properties and the current evidence base; and (iv) to assess the humanitarian and policy implications of deploying robust, low-reference-dependent annotation tools in resource-constrained research environments. The scope is methodological and policy-analytical; comprehensive empirical benchmarking is identified as a priority for the next phase of this research programme.

1.3 Relevance to Humanitarian and Social Development Goals

The pipeline addresses an equity problem as much as a technical one. Where reference data are scarce or of low quality, research capacity concentrates in well-resourced laboratories, widening the global gap in genomic science. A robust annotation method that performs reliably with imperfect references reduces the dependence on expensive, pristine datasets and high-performance computing clusters. This has direct implications for Sustainable Development Goals 3 (Good Health and Well-Being), 9 (Industry, Innovation and Infrastructure), and 10 (Reduced Inequalities): it enables hospitals, universities, and public health agencies in LMICs to conduct credible single-cell research on locally relevant diseases — from infectious disease surveillance to cancer biology — using the data they actually have, rather than the data they ideally need.

2. METHODOLOGY AND ANALYTICAL FRAMEWORK

2.1 The Siamese Network Architecture

The Siamese network, first introduced by Bromley et al. (1993) and later popularized for one-shot learning by Koch et al. (2015), consists of two identical subnetworks that share the same weights and parameters. Each subnetwork processes one input of a paired sample (e.g., two single-cell gene expression profiles) and extracts a feature representation. A distance metric then compares the two representations, and the network learns — via contrastive or triplet loss — to embed similar cell types close together and dissimilar ones far apart. Because learning is anchored in pairwise similarity rather than absolute classification against a fixed label set, the architecture is intrinsically tolerant of noisy, incomplete, or partially mislabelled reference data.

Component	Function
Input Layer	Receives paired single-cell RNA-seq expression profiles
Convolutional Layers	Extracts salient features from high-dimensional expression data
Fully Connected Layers	Processes and refines learned feature representations
Output (Distance) Layer	Computes similarity scores to discriminate between cell types

2.2 Data Sources and Dataset Selection

The pipeline is designed for publicly available scRNA-seq reference datasets sourced from repositories such as the Gene Expression Omnibus (GEO) and ArrayExpress. Selection criteria emphasize: (i) coverage of a broad range of cell types; (ii) documented imperfections — including noise, missing labels, and batch effects — so that model training reflects real-world conditions; and (iii) biological and demographic diversity to mitigate the representational biases that characterize many existing atlases (Kiselev et al., 2019). Where available, community-curated annotation resources such as CellSTAR (Zhang et al., 2024) are used to cross-validate reference labels.

2.3 Preprocessing Protocol

Three preprocessing stages prepare raw count data for the network: (i) normalization, scaling expression values across cells to reduce technical variability; (ii) noise reduction, applying filtering algorithms to improve the signal-to-noise ratio characteristic of sparse scRNA-seq data; and (iii) feature selection, retaining the most informative highly variable genes to reduce dimensionality and computational burden. A representative normalization routine (z-score per gene) is implemented as follows:

```
def normalize_data(data):
    mean = np.mean(data, axis=0)
    std = np.std(data, axis=0)
    return (data - mean) / std
```

2.4 Training and Optimization Strategy

Training follows a structured protocol: (i) normalization and preprocessing of the RNA-seq data; (ii) balanced batch construction, pairing similar and dissimilar cell types in controlled proportions to prevent class bias; (iii) model initialization with random weights; (iv) iterative training cycles with validation; and (v) hyperparameter management through dynamic learning-rate adjustment, dropout-based regularization, gradient clipping to prevent exploding gradients, and early stopping when validation performance plateaus.

2.5 Evaluation Framework

Model performance is assessed through a standard metric suite: accuracy, precision, recall, and the F1 score, supplemented by confusion matrices to expose per-class behaviour. Comparative analysis is conducted against established annotation methods — scmap, scCATCH, CHETAH, scSorter, and, at the frontier, foundation models such as scGPT — following the benchmarking logic of Díaz-Mejía et al. (2019).

3. KEY FINDINGS AND ANALYSIS

Editorial note: This section presents analytically grounded expectations derived from the architecture’s design properties and the published evidence base, not yet original empirical results. Systematic benchmarking on standard datasets (PBMC, pancreas, tumour micro-environment atlases) is underway; quantitative tables (accuracy, precision, recall, F1 per cell type) will be appended to the final release version of this paper, per the authors’ own note in Section 3.5.

3.1 Robustness to Imperfect Reference Data

The defining strength of the Siamese design is its tolerance of reference imperfection. Because the network learns a similarity function rather than a fixed classifier, individual mislabelled or noisy reference examples degrade performance gracefully rather than catastrophically. This property is directly relevant to real-world datasets, where even curated atlases contain contested labels — a documented challenge in neuronal cell-type classification (Zeng & Sanes, 2017). The twin-subnetwork structure, by enforcing weight sharing, further regularizes the model and reduces the risk of overfitting to idiosyncrasies of a flawed reference.

3.2 Scalability and Computational Efficiency

Siamese pipelines train efficiently on large datasets without proportional degradation in performance, and inference scales linearly with dataset size once embeddings are learned. Relative to foundation models — which require very large training corpora and substantial GPU infrastructure (Cui et al., 2024; Hao et al., 2024) — the Siamese approach offers a materially lighter computational footprint, making deployment feasible on institutional hardware common in LMIC research settings.

3.3 Performance with Limited Labelled Data

A central finding of the similarity-learning literature is strong performance under few-shot conditions (Koch et al., 2015). For cell annotation, this translates into credible classification even where labelled reference cells for a given type are scarce — a common situation for rare cell populations, newly characterized subtypes, and understudied tissues. This directly addresses the data-scarcity barrier that excludes many research programmes from single-cell analysis.

3.4 Comparative Positioning Against Existing Methods

Against conventional reference-based tools, the pipeline trades absolute dependence on label completeness for robustness: methods such as scmap and CHETAH perform well with clean references but degrade as reference quality falls (Kiselev et al., 2018; de Kanter et al., 2019; Díaz-Mejía et al., 2019). Against marker-based tools such as scCATCH and scSorter, the Siamese approach captures expression-pattern similarity beyond canonical marker genes (Shao et al., 2020; Guo & Li, 2021). Against generative foundation models, it offers interpretability, lower resource requirements, and easier auditability — attributes of particular value for clinical and policy applications where model transparency matters.

3.5 Illustrative Applications

- **Neuronal subtypes:** annotation of closely related neuron subtypes, where imperfect reference atlases have historically impeded classification (Zeng & Sanes, 2017; Fincher et al., 2018).
- **T-cell subtypes:** precise identification of immune cell states with clinical relevance to infectious and autoimmune disease (Szabo et al., 2019).
- **Stem and progenitor cells:** robust annotation of regenerative cell populations relevant to cardiovascular and tissue-repair research (Paik et al., 2020).
- **Disease-marker discovery:** cell-type-resolved marker identification supporting oncology research, including multiple myeloma and head-and-neck cancer studies (Puram et al., 2017; Liu et al., 2021).

Note for release: quantitative results tables — accuracy, precision, recall, and F1 scores per cell type, with comparative benchmarks — will be inserted upon completion of the experimental validation phase.

4. STRATEGIC IMPLICATIONS AND POLICY RECOMMENDATIONS

4.1 Democratizing Genomic Research Capacity

Because the pipeline performs with imperfect references and modest infrastructure, it lowers the effective cost of credible single-cell research. Policymakers and funders should treat annotation-software robustness as capacity-building infrastructure: support for deployment of such tools in national research institutions yields higher marginal returns than hardware alone.

4.2 Data Governance and Open Science

Reference data quality is a governance issue. Governments and research funders should invest in nationally and regionally representative reference atlases, mandate open deposition of annotated scRNA-seq data in repositories such as GEO and ArrayExpress, and support curation initiatives like CellSTAR that improve the shared reference layer on which all annotation methods depend.

4.3 Integration into Public Health Surveillance

Cell-type-resolved transcriptomics is increasingly relevant to infectious disease monitoring, tumour profiling, and immune-response assessment. Health ministries should explore the integration of validated annotation pipelines into national genomics programmes and hospital-based molecular diagnostics, with appropriate regulatory oversight and clinical validation.

4.4 Research Funding Priorities

- Fund empirical benchmarking of robust annotation methods — including this pipeline — on diverse, LMIC-sourced datasets, where reference imperfections are most acute.
- Prioritize research on rare and understudied cell populations, for which labelled data are inherently scarce and few-shot architectures are best suited.
- Support interdisciplinary training at the intersection of deep learning and molecular biology, particularly in LMIC institutions.

4.5 Ethical and Equity Safeguards

Automated annotation can silently propagate the biases of imperfect reference data. Deployment should be accompanied by human expert review, transparent reporting of model confidence, and monitoring for

systematic under-annotation of populations under-represented in reference atlases — consistent with CHRSD’s commitment to evidence that serves all communities equitably.

5. CONCLUSION

This paper has presented a Siamese network-based pipeline for cell type annotation that is deliberately engineered for the imperfect reference data that characterize real-world single-cell research. Its twin-shared-weight architecture offers robustness to noise and mislabelling, scalability, and strong few-shot performance — properties that simultaneously address a technical challenge and a structural inequity in global genomic science. By reducing dependence on pristine, resource-intensive reference datasets, the approach opens credible single-cell analysis to a far wider community of researchers and clinicians, with direct relevance to disease surveillance, oncology, and health-system strengthening in resource-constrained settings. Future work will complete large-scale empirical benchmarking, extend the framework to multi-omics integration and real-time annotation, and evaluate cloud-based deployment models. Embracing such robust, accessible methods will help ensure that the genomic revolution delivers benefits equitably — a goal squarely aligned with the CHRSD Foundation’s mission of evidence-based research for human development and social progress.

REFERENCES (APA 7TH EDITION)

- Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1993). Signature verification using a ‘Siamese’ time delay neural network. *Advances in Neural Information Processing Systems*, 6, 737–744.
- Cortral, A., Martignetti, L., Six, E., et al. (2021). Gene signature extraction and cell identity recognition at the single-cell level with Cell-ID. *Nature Biotechnology*, 39(10), 1269–1277.
- Cui, H., Wang, C., Maan, H., et al. (2024). scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8), 1470–1480.
- de Kanter, J. K., Lijnzaad, P., Candelli, T., et al. (2019). CHETAH: A selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Research*, 47(16), e95–e95.
- De Veirman, K., De Beule, N., Maes, K., et al. (2017). Extracellular S100A9 protein in bone marrow supports multiple myeloma survival by stimulating angiogenesis and cytokine secretion. *Cancer Immunology Research*, 5(10), 839–846.
- Díaz-Mejía, J. J., Meng, E. C., Pico, A. R., et al. (2019). Evaluation of methods to assign cell type labels to cell clusters from single-cell RNA-sequencing data. *F1000Research*, 8, 296.
- Duan, B., Zhu, C., Chuai, G., et al. (2020). Learning for single-cell assignment. *Science Advances*, 6(48), eabd0855.
- Fincher, C. T., Wurtzel, O., de Hoog, T., et al. (2018). Cell type transcriptome atlas for the planarian *Schmidtea mediterranea*. *Science*, 360(6391), eaaq1736.
- Ge, F., Zhang, L., Tao, S.-C., et al. (2011). Quantitative proteomic analysis of tumor reversion in multiple myeloma cells. *Journal of Proteome Research*, 10(2), 845–855.
- Guo, H., & Li, J. (2021). scSorter: Assigning cells to known cell types according to marker genes. *Genome Biology*, 22(1), 98.
- Hao, M., Gong, J., Zeng, X., et al. (2024). Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8), 1471–1480.
- Hou, W., & Ji, Z. (2024). Assessing GPT-4 for cell type annotation in single-cell RNA-seq analysis. *Nature Methods*, 21(8), 1486–1490.
- Karlsson, M., Zhang, C., Mear, L., et al. (2021). A single-cell type transcriptomics map of human tissues. *Science Advances*, 7(31), eabd0855.

-
- Kiselev, V. Y., Andrews, T. S., & Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, 20(5), 273–282.
- Kiselev, V. Y., Yiu, A., & Hemberg, M. (2018). scmap: Projection of single-cell RNA-seq data across data sets. *Nature Methods*, 15(5), 359–362.
- Koch, G., Zemel, R., & Salakhutdinov, R. (2015). Siamese neural networks for one-shot image recognition. *Proceedings of the ICML Deep Learning Workshop*, Lille, France.
- Liu, M., Wang, Y., Miettinen, J. J., et al. (2021). S100 calcium binding protein family members associate with poor patient outcome and response to proteasome inhibition in multiple myeloma. *Frontiers in Cell and Developmental Biology*, 9, 723016.
- Paik, D. T., Cho, S., Tian, L., et al. (2020). Single-cell RNA sequencing in cardiovascular development, disease and medicine. *Nature Reviews Cardiology*, 17(7), 457–473.
- Puram, S. V., Tirosh, I., Park, A. S., et al. (2017). Single-cell transcriptomic analysis of primary and metastatic tumor ecosystems in head and neck cancer. *Cell*, 171(7), 1611–1624.e24.
- Shalek, A. K., Satija, R., Adiconis, X., et al. (2013). Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. *Nature*, 498(7453), 236–240.
- Shao, X., Liao, J., Lu, X., et al. (2020). scCATCH: Automatic annotation on cell types of clusters from single-cell RNA sequencing data. *iScience*, 23(3), 100882.
- Szabo, P. A., Levitin, H. M., Miron, M., et al. (2019). Single-cell transcriptomics of human T cells reveals tissue and activation signatures in health and disease. *Nature Communications*, 10, 4706.
- Theodoris, C. V., Xiao, L., Chopra, A., et al. (2023). Transfer learning enables predictions in network biology. *Nature*, 618(7965), 616–624.
- Zeng, H., & Sanes, J. R. (2017). Neuronal cell-type classification: Challenges, opportunities and the path forward. *Nature Reviews Neuroscience*, 18(9), 530–546.
- Zhang, Y., Sun, H., Zhang, W., et al. (2024). CellSTAR: A comprehensive resource for single-cell transcriptomic annotation. *Nucleic Acids Research*, 52(D1), D859–D870.
- Suggested citation:** Mustafiz, M. R., Roy, C. K., Hossain, R., Hossain, A., & Mohsin, K. (2026). *Enhancing cell type annotation with imperfect single-cell RNA-seq references: A Siamese network pipeline for equitable and scalable genomic research*. CHRSD Foundation Research Paper Series. <https://chrds.org/research-papers>