CSIAM Trans. Life Sci.
doi: 10.4208/csiam-ls.SO-2025-0025

Vol. **2**, No. 1, pp. 177-202
March 2026

Effectively Preserving Biological Variations in Multi-Batch and Multi-Condition Single-Cell Data Integration

Qingbin Zhou[1], Tao Ren[2,3], Fan Yuan[4], Jiating Yu[5], Jiacheng Leng[6], Jiahao Song[1], Duanchen Sun[1,7,*] and Ling-Yun Wu[2,3,*]

[1] *School of Mathematics, Shandong University, Jinan 250100, China.*
[2] *State Key Laboratory of Mathematical Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.*
[3] *School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China.*
[4] *School of Mathematics and Information Science, Yantai University, Yantai 264005, China.*
[5] *School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing 210044, China.*
[6] *Zhejiang Lab, Hangzhou 311121, China.*
[7] *Shandong Key Laboratory of Cancer Digital Medicine, Jinan 250033, China.*

Received 13 August 2025; Accepted 17 October 2025

Abstract. Understanding phenotypic differences at the cell level is critical for comprehending the underlying pathogenesis of related complex diseases. However, the biological variations are obscured by batch effects, posing a challenge for integrating multi-batch and multi-condition single-cell datasets. Here, we present scFLASH, a deep learning-based model specially designed to explore single-cell biological variations while correcting undesired batch effects. scFLASH employs a conditional variational autoencoder with adversarial training to separate biological variations from technical noise and introduces a penalized condition classifier to preserve condition-specific biological signals. Through comprehensive benchmarking evaluations, scFLASH shows superior integration performances compared to other state-of-the-art methods. Applied to datasets such as Alzheimer's disease, COVID-19, and diabetes, we demonstrate that scFLASH is applicable to various scenarios, effectively integrating datasets with two or more conditions and different batch sources. scFLASH can enhance the gene expression profiles and identify the condition-related cell subpopulations, facilitating downstream analyses and offering biological insights into the cellular mechanisms of disease pathology.

AMS subject classifications: 92D10, 92-08, 68T05

Key words: Biological variations, data integration, batch correction, deep learning.

*Corresponding author. *Email addresses:* `dcsun@sdu.edu.cn` (D. Sun), `lywu@amss.ac.cn` (L.-Y. Wu)

1 Background

Single-cell RNA sequencing (scRNA-seq) has become an indispensable tool for understanding cellular heterogeneity and functionality, paving the way for personalized medicine and drug development [21, 28, 44, 52]. Among the single-cell analyses, deciphering phenotypic differences at the single-cell level is critical for uncovering the underlying pathogenesis of related complex diseases [4, 40, 55]. One practical approach to achieving this goal is integrating scRNA-seq from diverse sources and conditions [56,57]. However, the biological variations across phenotypes are entangled with batch effects, posing a challenge for integrating multi-batch and multi-condition (MBMC) single-cell datasets [6,29].

Several MBMC integration methods have recently been developed to remove technical variations while preserving phenotypic differences. scINSIGHT [37] utilized non-negative matrix factorization to learn common and specific gene expression modules under different conditions. It employed linear assumptions to expression decomposition, which may hinder its model from effectively analyzing the complex nonlinear data. scMerge2 [24] used factor analysis to remove unwanted technical variations, whereas it only introduced pseudo-replicates within each condition and did not consider necessary connections between different conditions. Besides, scMerge2 also requires specific model inputs and thus cannot flexibly handle a single dataset under the same conditions. scDisInFact [56] was a deep learning-based model that used multiple encoders to disentangle gene expressions into shared and un-shared biological factors. Its established independent encoders may overlook the interactions between condition-specific biological signals. scDisco [25] was a recently proposed approach that employed condition and domain-specific layers in variational autoencoders to enhance the condition representations.

In addition to the inherent limitations of the methods discussed above, which affect their utility and versatility, the MBMC integration problem faces several key challenges. First, differences between conditions and cellular heterogeneity are both necessary biological variations that should be preserved. It is important to decipher condition differences without sacrificing the accurate representation of cell type heterogeneity characterization. Second, overly emphasizing batch correction can inadvertently distort or suppress biological variations across conditions, hindering the extraction of meaningful biological insights. Overcorrection should be avoided to prevent excessive data homogenization and misleading downstream analyses. Last, an advanced MBMC integration algorithm should be appropriate for various scenarios, effectively handling tasks with multiple conditions and diverse batch sources.

To this end, we developed a deep learning-based model, scFLASH, to explore phenotypic differences in MBMC integration. scFLASH utilizes a conditional variational autoencoder (CVAE) [48] framework to learn a disentangled latent space, capturing embeddings for known conditional attributes and unknown biological attributes. To remove batch effects, scFLASH applies adversarial training in the latent space. Furthermore,