

News Release

Removing Experimental Noise Without Losing Disease Signals

Kanade: An AI Method for Separating Experimental Noise from Biological Signals

Key Points

- A major problem in single-cell analysis is the unintended loss of disease- or treatment-related differences during correcting experimental variation.
- We developed Kanade, a machine-learning method that separates experimental variation from information related to disease, treatment, and cell type.
- Kanade removed experimental variation while preserving disease-related differences in cell states in both simulated and mouse pancreas data.
- Kanade also reconstructs corrected gene expression levels, enabling downstream analyses such as identifying disease-related genes.

Summary

Designated Assistant Professor Shunta Sakaguchi (Graduate School of Medicine, Nagoya University), together with Assistant Professor Masato Tsutsumi (Life Science Center for Survival Dynamics, University of Tsukuba), Ph.D. student Kentaro Nishi (Graduate School of Integrated Sciences for Life, Hiroshima University), and Professor Naoki Honda (Graduate School of Medicine, Nagoya University; Graduate School of Integrated Sciences for Life, Hiroshima University) has developed a new machine-learning method named Kanade.

Single-cell RNA sequencing (scRNA-seq) measures how strongly genes are expressed in individual cells and is widely used to study cellular changes associated with disease and treatment. However, data collected in different experiments often contain technical variation caused by differences in instruments, reagents, or experimental procedures. When researchers try to remove these “batch effects,” they may also unintentionally remove disease- or treatment-related differences of interest. This is particularly problematic when healthy and disease samples are measured in separate experiments, making it difficult to distinguish experimental variation from true differences between conditions. Because comparing conditions is fundamental to scientific research, there has been a need for a method that can separate these effects and selectively correct batch effects.

Kanade is based on a variational autoencoder, a type of deep generative model.

It learns different types of information in gene expression data separately, including experimental variation, differences associated with disease or treatment conditions, and differences among cell types. Kanade then removes only the experiment-related information and reconstructs gene expression data while preserving the features needed for biological comparison. In other words, rather than simply making datasets look more similar, Kanade distinguishes between differences that should be removed and differences that should be preserved.

When tested on simulated data, Kanade removed technical variation between experiments while preserving cellular features associated with disease or treatment. Kanade was then applied to real pancreatic single-cell data from diabetic mouse models, and disease-related differences in cell states remained detectable after batch correction.

Because Kanade reconstructs corrected gene expression levels, its output can be used not only for cell classification and visualization, but also for downstream analyses such as identifying disease-related genes. In the future, Kanade may provide a useful framework for comparing and integrating single-cell datasets collected across different experiments and research institutions while preserving important differences between conditions.

The study was published in the international journal *Patterns* on October 9, 2026.

Research Background

In recent years, single-cell RNA sequencing (scRNA-seq), which measures gene expression in individual cells, has advanced rapidly. It has made it possible to examine differences in gene expression at the level of individual cells, rather than only as averages across populations of cells. This technology plays an important role in understanding which types of cells change, and how they change, during development, disease, and drug response.

However, when data from multiple experiments are compared or integrated, technical variation caused by differences in experimental timing, instruments, reagents, experimenters, or research institutions can become a problem. These differences are known as “batch effects.” If batch effects remain in the data, variation caused by the experimental environment may be mistaken for differences in cell properties or changes associated with disease. The process

of removing these effects is called “batch correction.”

Simply reducing differences between experiments, however, is not enough. For example, when healthy and disease groups are measured in separate experiments, the observed differences in the data may reflect both disease-related changes and effects of the experimental environment. If these two sources of variation are not distinguished, correcting differences between experiments may also remove important disease-related information. In studies of disease and treatment, the goal is not merely to make datasets look similar, but to accurately compare what has changed. Therefore, there has been a need for batch-correction methods that can separate “experimental variation that should be removed” from “biological differences that should be preserved for comparison,” while retaining the information needed for analysis.

Research Results

In this study, we developed a new machine-learning method, **Kanade**, that learns different sources of information in scRNA-seq data separately and selectively corrects batch effects. Kanade is based on a variational autoencoder (VAE), a type of deep generative model. A VAE compresses complex, high-dimensional data, such as the expression levels of many genes, into a smaller number of variables and then reconstructs the original data from them. In a conventional VAE, different types of information, including experimental variation, can become mixed together during this compression. Kanade is designed instead to separate information related to batch effects, conditions such as disease or treatment, and cell types, and to store them separately. This allows Kanade to distinguish between batch effects that should be removed and differences between conditions that should be preserved. During correction, only the information corresponding to batch effects is removed, and gene expression data are reconstructed while retaining condition-related differences and cell-type features.

Kanade was first evaluated using simulated data. The results showed that it could remove batch effects while preserving differences between healthy and disease conditions, as well as features of different cell types. With conventional methods, differences between conditions could become less distinct during batch correction, whereas Kanade preserved condition-related information more clearly.

Kanade was then applied to pancreatic scRNA-seq data containing multiple

conditions, including healthy mice and diabetic mouse models. It corrected technical differences between datasets while preserving cellular changes associated with diabetes. These results show that Kanade does not simply make datasets more uniform but can distinguish experimental noise from important disease-related differences.

The source code for Kanade is publicly available on GitHub and can be used in any suitable computing environment.

(<https://github.com/ShuntaSakaguchi/kanade>)

Research Summary and Future Perspective

Kanade is expected to have broad applications in single-cell omics analyses that compare multiple conditions, such as disease, drug response, and developmental stages, particularly in large-scale datasets. In the future, we aim to develop Kanade into an analytical framework that supports global data integration by connecting single-cell datasets collected across different research institutions and experimental platforms.

Kanade is also highly extensible because it is based on a relatively simple VAE architecture. Beyond the scRNA-seq data tested in this study, the same framework could potentially be extended to other types of biomedical and life-science data, such as brain imaging. More broadly, the concept of separating “measurement-related differences that should be removed” from “biological differences that should be preserved for comparison” may be useful for integrating a wide range of complex datasets.

Moonshot Goal 2, “Realization of ultra-early disease prediction and intervention by 2050,” aims to advance research and development toward detecting disease onset and progression at an early stage and enabling their prediction and prevention. Kanade is useful for integrating and analyzing data across different disease states, including healthy and diseased conditions as well as early and late stages of disease. By facilitating the detection of early disease-associated signals from such datasets, this study is expected to contribute to the achievement of Moonshot Goal 2.

Publication

Journal: Patterns

Title:

Disentangling batch effects and biological signals across conditions in

single-cell transcriptome

Authors:

Shunta Sakaguchi¹, Masato Tsutsumi², Kentaro Nishi³, and Naoki Honda^{1,3,4}

Affiliations:

¹ Graduate School of Medicine, Nagoya University

² Life Science Center for Survival Dynamics, University of Tsukuba

³ Graduate School of Integrated Sciences for Life, Hiroshima University

⁴ Center for One Medicine Innovative Translational Research (COMIT), Nagoya University

DOI: 10.1016/j.patter.2026.101638

Publication date: Oct. 9th, 2026

Japanese ver.

https://www.med.nagoya-u.ac.jp/medicalJ/research/pdf/Pat_261013.pdf