GIST presents a new paradigm for single-cell analysis overcoming the limitations of single-cell RNA sequencing technology with AI technology

- AI Graduate School Professor Hyunju Lee's team develops 'scRobust', which greatly improves prediction performance for cell types through a self-supervised learning methodology that can learn all genes from single-cell RNA data

- "It is now possible to compare and analyze minute features occurring in cell types, such as genes expressed only in a small number of cells" Published in the international academic journal 《Briefings in Bioinformatics》

▲ (From left) Professor Hyunju Lee of the AI Graduate School and Sejin Park, a doctoral student in the School of Electrical Engineering and Computer Science

Single-cell ribonucleic acid (RNA)* sequencing*, which can measure gene expression levels at the individual cell level, has recently been gaining attention for its rapid development in various fields such as biology, new drug development, and clinical research.

Single-cell RNA sequencing technology is suitable for providing customized medical services such as new diagnosis and prognosis prediction for diseases because it allows for gene analysis according to cell type, but it has a limitation in that only a portion of all genes are detected due to its low accuracy compared to multi-cell RNA sequencing* technology, which measures the expression levels of multiple cells by adding them together.

* RNA: A material that copies DNA and contains the genetic information of a cell.

* RNA sequencing (single-cell RNA sequencing): A technology that measures the amount of RNA in a cell, which can be used to estimate the level of gene expression (activation). (For example, in the case of diabetic patients, the expression of genes related to insulin secretion decreases.)

The Gwangju Institute of Science and Technology (GIST, President Kichul Lim) announced that Professor Hyunju Lee's research team from the AI Graduate School has developed a self-supervised learning* methodology that can overcome the fundamental limitations of single-cell RNA sequencing technology.

As a result of applying this, it was possible to discover detailed features that differentiate even the same cell type according to the degree of diabetes. In addition, it showed the highest F1 score* in 12 out of 15 single-cell RNA datasets in cell type classification tests.

* self-supervised learning: To train a general AI model, data and labels (information that can explain the data) are required, and this learning method is called supervised learning. On the other hand, self-supervised learning refers to a methodology for training an AI model without separate labels. For example, to train a chatbot model, only conversation text is required, and a separate label (whether the conversation is positive or negative) is not required.

* F1 score: This is a score that evaluates the balance between precision (how accurately it was corrected: precision) and recall (how much it did not miss: recall). This score does not simply focus on how many correct answers were given, but measures how accurately the correct answers were given. For example, when a test is conducted to distinguish between 99 normal people and 1 cancer patient, if it is always judged as normal, it will have a 99% correct answer rate, but the F1 score will be 0. This is because the normal people were judged as normal, but the cancer patients were not diagnosed as cancer patients.

Unlike multicellular RNA sequencing technology that can measure gene expression levels mixed with RNA from multiple cells, single-cell RNA sequencing has low measurement accuracy because it targets only single cells. Therefore, single-cell RNA data often obtains expression levels of only 2,000 to 3,000 genes out of more than 30,000 genes.

In other words, only 10% of all genes have high resolution that can be measured, and the remaining 90% of information cannot be measured due to low resolution.

Accordingly, previous studies have mainly predicted and analyzed cell types using only about 10% of genes commonly expressed in multiple cells.

However, genes expressed only in certain cell types often describe the cell in more detail, and the single-cell RNA sequencing technology currently in use has a fundamental problem of not being able to use about 90% of the genetic information.

The research team developed 'scRobust', a technology that can identify both universal and detailed features of each single cell with less than 5% of the genetic information by utilizing a contrastive learning (one of the self-supervised learning methods)* methodology suitable for single-cell RNA sequencing data.

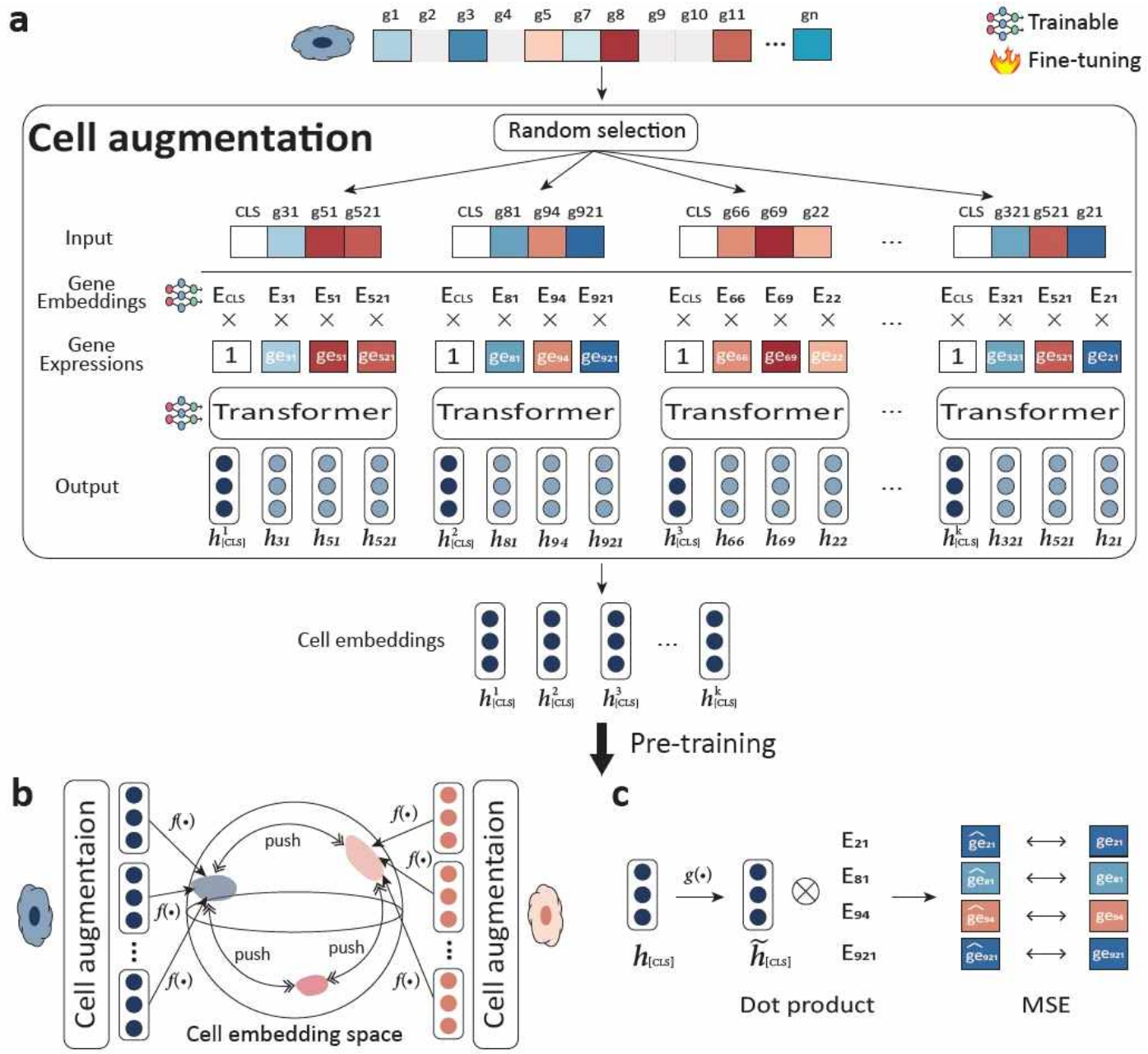
Through this, it became possible to use up to 90% of the genetic information that was not used in the previous method, which not only improved the prediction performance for cell types, but also enabled more precise analysis within the same cell.

* contrastive learning: Contrastive learning is a method that creates multiple data by applying various changes to the given data, and trains the AI model to find data from the same original among the numerous data created in this way.

This technology is based on a methodology that can create multiple cell representation vectors* by creating various gene combinations from a single cell, and is suitable for data augmentation* for single-cell RNA sequencing data.

* representation vector: This refers to the form in which an arbitrary data sample is converted into a vector form. For example, a cell representation vector converts a cell into a vector form so that an AI model can understand the concept of a cell.

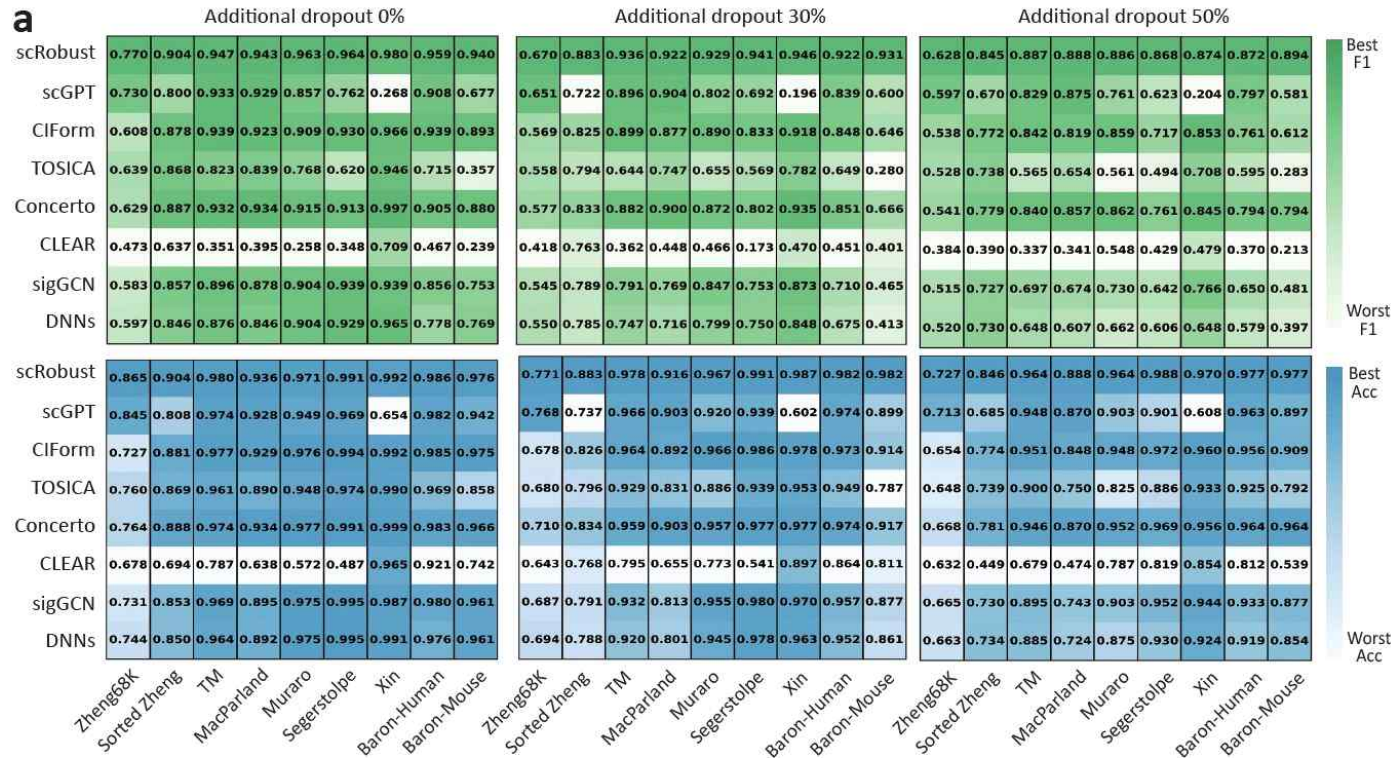
* data augmentation: This is a technology that creates multiple data by applying various transformations to given data (e.g., rotating a photo or changing it to black and white). Using this technology increases the number of data that can be used to train an AI model.



▲ Schematic diagram of the self-supervised learning developed in this study: (a) data (cell) augmentation, various cell expression vectors are generated from random cells. (b) contrastive learning, cell expression vectors (cell embeddings) generated from the same cell exist in similar locations. (c) Using cell expression vectors, the expression level of a random gene is predicted.

By training an AI model through contrastive learning, it is possible to distinguish whether cell representation vectors generated with different gene combinations come from the same cell or different cells, and through this process, cell representation vectors (local cell representation vectors) generated with various gene combinations converge to a single unified cell representation vector (global cell representation vector).

As a result, even if only a small number of genes are used, a cell expression vector similar to that utilizing all genes can be obtained, so the effect of using all genes can be expected.



▲ Cell type prediction results. The model developed in this study (scRobust) shows the highest performance (F1 score) in 8 out of 9 different data sets. This shows that scRobust does not only work on specific data sets, but also shows good performance in most data sets.

Professor Hyunju Lee said, "The algorithm developed in this study enables AI models to learn about all genes, not just a subset of genes. This makes it possible to compare and analyze even subtle features that occur in cell types, such as genes that are expressed in only a small number of cells."

She also said, "It is expected that the paradigm of single-cell analysis will change in the future as it can extract marker genes of various cell types as well as marker genes* related to drug resistance."

* Marker gene: A gene that is actively expressed only in certain cells or tissues and helps distinguish the cells or tissues.

This study, supervised by Professor Hyunju Lee of the GIST AI Graduate School and conducted by doctoral student Sejin Park, was supported by the Information and Communications Technology Planning and Evaluation Institute (IITP) and was published on November 16, 2024 in the international academic journal 《Briefings in Bioinformatics》, a top 4% JCR journal in the bioinformatics field.

GIST
Since 1993