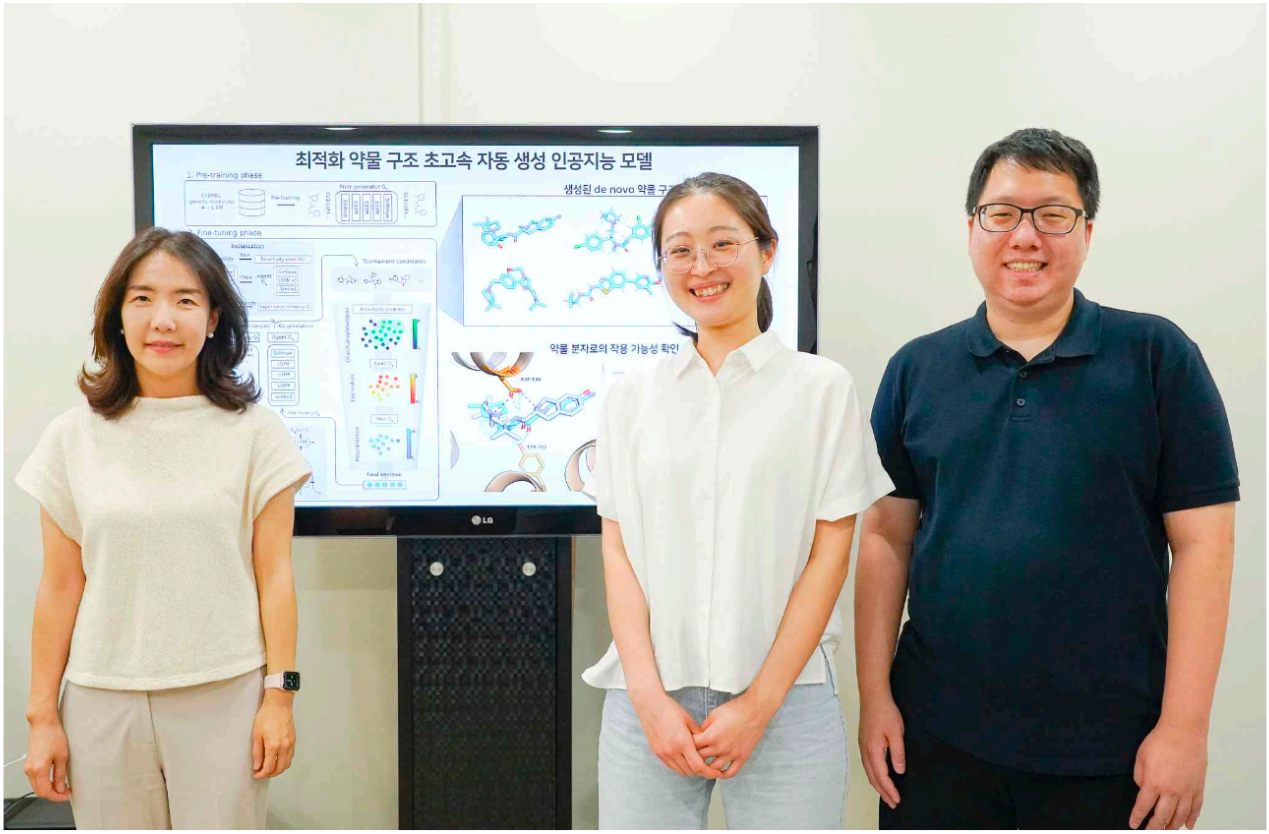
Design new drugs faster and more effectively... Development of AI platform for generating small molecule compounds

- Overcoming the problems of existing compound generation AI research... Expected to automatically generate new drug structures with improved effectiveness



▲ (From the left) Professor Hojung Nam, doctoral student Haelee Bae, and integrated master's and doctoral student Bongsung Bae

Professor Hojung Nam's research team in the School of Electrical Engineering and Computer Science at the Gwangju Institute of Science and Technology (GIST, President Kichul Lim) developed an artificial intelligence model that generates new drug structures.

Recently, various artificial intelligence-based drug design platforms have emerged, and generative artificial intelligence models based on large-scale compound data are trained to design drugs with new structures that did not exist before (*de novo* drug design) or optimize lead substances. Lead optimization* research is attracting attention.

In particular, research is actively underway to introduce transfer learning*, which can be used even in situations where data to be used for teaching artificial intelligence models is insufficient. This method first interprets the compound structure into linguistic grammar to pre-train a large amount of compound data (pre-trained model), and then it performs fine-tuning* using a small number of drug activity data.

* Lead optimization: To develop new drugs, it is the process of finding lead substances (new drug candidates that can control disease-causing substances) and developing them into substances with excellent biological activity.

* Transfer learning: To teach an artificial intelligence model suitable for a specific purpose, it means taking a part of a model that has already been trained with data previously refined for another purpose and retraining it for a new purpose.

* Fine-tuning: Classifies existing data as before and learns to fine-tune the weights again with new data. This model is more effective when the given problem is very different from the compound on which the model was trained. This is advantageous when the data set is small and can be quickly optimized.

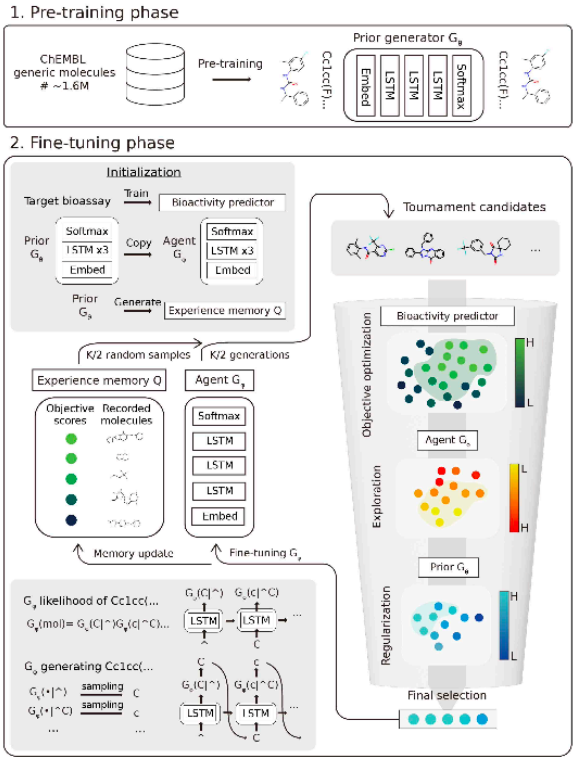
In studies using existing transfer learning, the mode collapse* phenomenon occurs due to a lack of structural diversity of the product in the fine-tuning stage, making it difficult for the generative model to learn the optimal distribution. As a result, this acts as a limitation in creating various new drug structures.

* Mode collapse is a phenomenon that can occur when training generative artificial intelligence. It refers to a phenomenon in which the generative model is unable to produce products of various shapes/ structures and repeatedly outputs only similar products.

In the traditional new drug development process, the optimization stage of the lead material requires a long research process that can take anywhere from several months to several years. However, if artificial intelligence is introduced at this stage, the development time can be greatly reduced from several weeks to several months, so the use of artificial intelligence technology is attracting great attention.

In particular, the artificial intelligence model developed this time has the great advantage of being designed to easily change target protein information so that it can be applied to the development of a variety of therapeutics, rather than a model specialized for a specific target protein, so it is expected to be applicable to the general drug development market.

The research team developed an optimal generative distribution learning technology (LOGICS: Learning optimal generative distribution for designing *de novo* chemical structures) for designing new drug structures.



▲ LOGICS research overview. This shows the data used in the study and the proposed model training method.

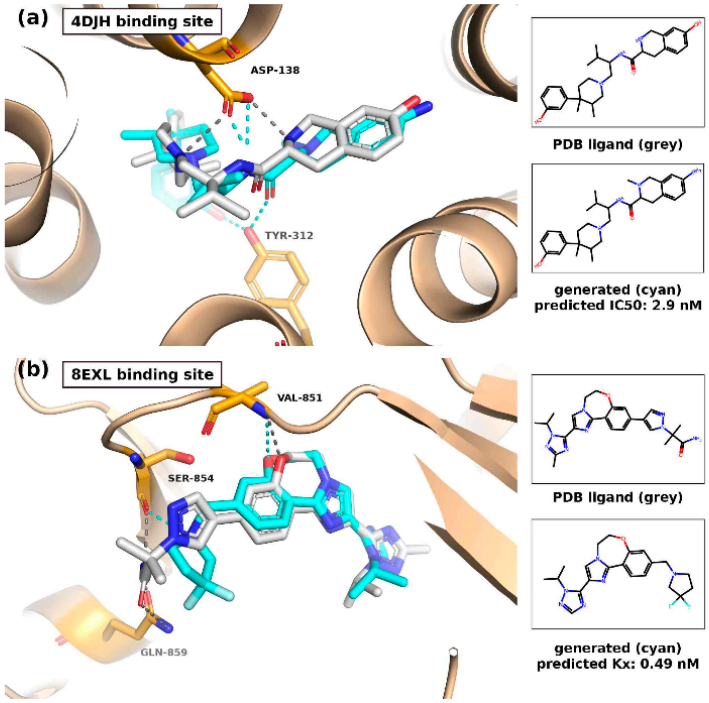
To solve the problems of existing transfer learning models, the research team proposed a training algorithm and devised a method that uses experience memory and

tournament selection in the fine-tuning stage, allowing the generative model to explore more diverse compound structures.

This study demonstrated superior performance compared to previously introduced representative new drug development platforms (REINVENT, DrugEx, etc.) based on FCD*, an indicator that measures the chemical distance from the actual drug molecule (29.4 FCD compared to REINVENT's 32.7 FCD), confirming improved performance with FCD, and it was verified that the new compounds have properties more similar to actual active drugs.

* FCD (Fréchet ChemNet Distance): An indicator of the chemical distance from actual drug molecules. The smaller the value, the higher the probability of generating molecules with high similarity to actual drugs set as a standard.

As a result of analyzing the molecular structure presented by the artificial intelligence model developed by the research team, it was revealed that many of the products predicted as drug candidates were new structures with low similarity to molecules in existing data.



▲ Example of a new molecular structure generated from the model. The resulting molecules show high binding affinity in the binding pocket of each target protein.

In addition, it was confirmed that a drug optimization method that maintains the main framework of the existing compound but improves target activity and three-dimensional structural binding to the target protein through detailed changes in the structure can be applied.

Professor Hojung Nam said, "This research outcome can induce stable learning by resolving the problems of existing transfer learning models, and it is possible to propose a variety of new high-quality molecular structures even in situations where data is limited. By applying it in the early stages of new drug development, it is expected to dramatically shorten the process of discovering candidate and lead materials and increase efficiency."

This research was led by Professor Hojung Nam and conducted by integrated master's and doctoral student Bongsung Bae and doctoral student Haelee Bae with support from the National Research Foundation of Korea's 'Mid-career Researcher Support Project' and the 'Source Technology Development Project' project and the Information and Communication Technology Promotion Center of the Ministry of

Science, ICT and Future Planning and was published online on September 7, 2023, in *Journal of Cheminformatics*, a renowned academic journal in the field of chemical informatics.

▲ Professor Hojung Nam's research team discuss research results while developing an artificial intelligence model that generates new drug structures.

GIST
Since 1993