GIST develops the world's first AI for creating genetically tailored anticancer drugs... Personal genotype analysis, optimal treatment design

- Professor Hojung Nam's research team from the Department of Electrical Engineering and Computer Science develops AI model 'G2D-Diff' for generating customized anticancer drug candidates based on cancer genes... Learning 1.5 million chemical structures and 1.2 million drug response data

- Shows generation ability with significantly higher accuracy than existing models (error rate 1%, 35-44% improvement in suitability), suggesting possibility of precision treatment for intractable cancer... Published in international academic journal 《Nature Communications》

▲ (From left) Professor Hojung Nam of the Department of Electrical Engineering and Computer Science at GIST, Dr. Hyunho Kim (currently a senior researcher at the National Institute of Toxicology)

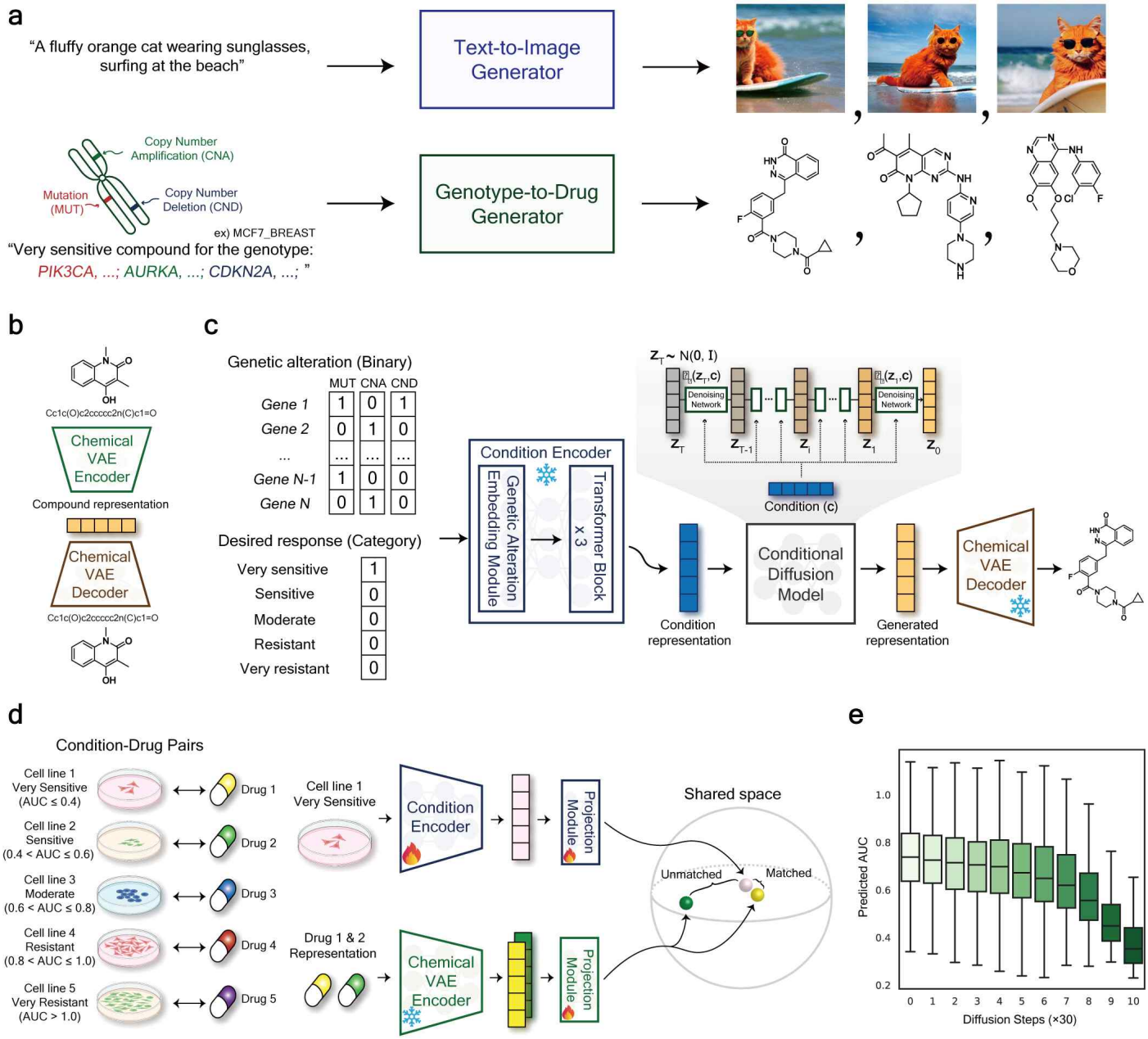
Even if cancer is the same type of disease, the genotype* of each patient is different, showing a large difference in the treatment effect. In particular, incurable cancer such as 'triple-negative breast cancer (TNBC)'* has no clear target, making it difficult to expect sufficient effects with existing treatments.

* genotype: A form that synthesizes mutation information and copy number variation information of various genes.

* triple-negative breast cancer (TNBC): Refers to breast cancer in which estrogen receptors, progesterone receptors, and HER2 proteins are all negative. It is classified as an incurable cancer in which existing hormone therapy or targeted therapy does not work because these three major receptors are not expressed. Therefore, TNBC is very difficult to treat, and the prognosis is often poor, and the development of new customized treatments is urgently needed.

The Gwangju Institute of Science and Technology (GIST, President Kichul Lim) announced that the research team led by Professor Hojung Nam of the Department of Electrical Engineering and Computer Science has developed the world's first generative artificial intelligence (AI) model that analyzes the genotypes of cancer patients and proposes personalized anticancer drug candidates.

The AI model developed by the research team can learn genotype information and drug response data that are different for each cancer cell, and generate new anticancer drug candidates optimized for each patient. This will enable personalized precision medicine as well as a new solution for incurable cancer that does not respond well to existing treatments.



▲ Overview of the G2D-Diff model. a) Concept of how G2D-Diff works, based on the text-image generator. b-c) Structure of the model d) Learning the Condition encoder, which is the main part of the model. e) Example of model operation - Conditional suitability increases as the diffusion step progresses.

Previous generative AI-based anticancer drug development studies had several limitations. First, in complex diseases such as cancer, the treatment target is often unclear, so the effectiveness of the generated drugs was limited. Second, it often relied on special data that was difficult to obtain in clinical settings, so the possibility of practical use was low.

To overcome these limitations, the research team developed a generative AI model, 'G2D-Diff,' that learned about 1.5 million chemical structures and 1.2 million drug response data. This model automatically designs an anticancer drug candidate optimized for the input of genetic information (mutations and copy number variations) that can be obtained in actual clinical practice and the target drug response level.

G2D-Diff works in a similar way to AI models that generate text-to-image. For example, if you input a condition such as 'a drug that is highly sensitive to a specific cancer genotype,' it will generate an anticancer drug molecule structure that matches that condition.

This model consists of ▲ a 'chemical variational autoencoder (VAE)' that numerically expresses molecular structures, ▲ a 'conditional encoder' that numerically transforms input conditions (such as genotypes and drug response targets), and ▲ a 'conditional diffusion model' that generates new molecular structures that meet the conditions.

* sensitivity: This means that the drug is sensitive to cancer cell death and promotes death.

G2D-Diff showed overwhelming performance in all performance indicators compared to existing generative AI models. In particular, compared to IBM's 'PaccMannRL*', known as the model with the highest performance, it showed superior performance in diversity, feasibility, and condition fitness.

In particular, in the 'condition fitness' item that evaluates how well the generated compounds match the input genotype conditions, the existing model showed an average error rate of about 51% in drug response prediction, while G2D-Diff recorded an average error rate of about 1%.

In addition, the generated molecular structure showed an average 35-44% higher distribution similarity* with actual drug groups in terms of drug similarity (QED) and synthetic accessibility (SAS) than the existing model, proving that it is more likely to be developed into an actual new drug.

* PaccMannRL: A generative AI model that predicts the efficacy of anticancer drugs by integrating gene expression information and chemical structure data and generates new drug candidates based on reinforcement learning. It focuses on genes or molecular structures important to cancer cells through an attention mechanism and is optimized to design compounds that exhibit the desired drug response. This enables efficient exploration of new drug candidates in the pre-experimental stage.

* average error rate (%): $\frac{|\text{Product prediction sensitivity average} - \text{Control group prediction sensitivity average}|}{\text{Control group prediction sensitivity average}} \times 100$

* distribution similarity: Jensen-Shannon divergence (distribution difference) from the actual drug group distribution is used as an evaluation index. This measure is commonly used as a measure of the distance between two distributions.

The research team applied the G2D-Diff model to triple-negative breast cancer, a representative case of intractable cancer, to verify its practical applicability.

The candidate substances generated by inputting the patients' genetic mutation information accurately targeted PI3K, HDAC, CDK*, and other key proteins that inhibit cancer cell proliferation.

In addition, the chemical structures of these compounds are completely different from existing treatments, but they can produce the same therapeutic effects.

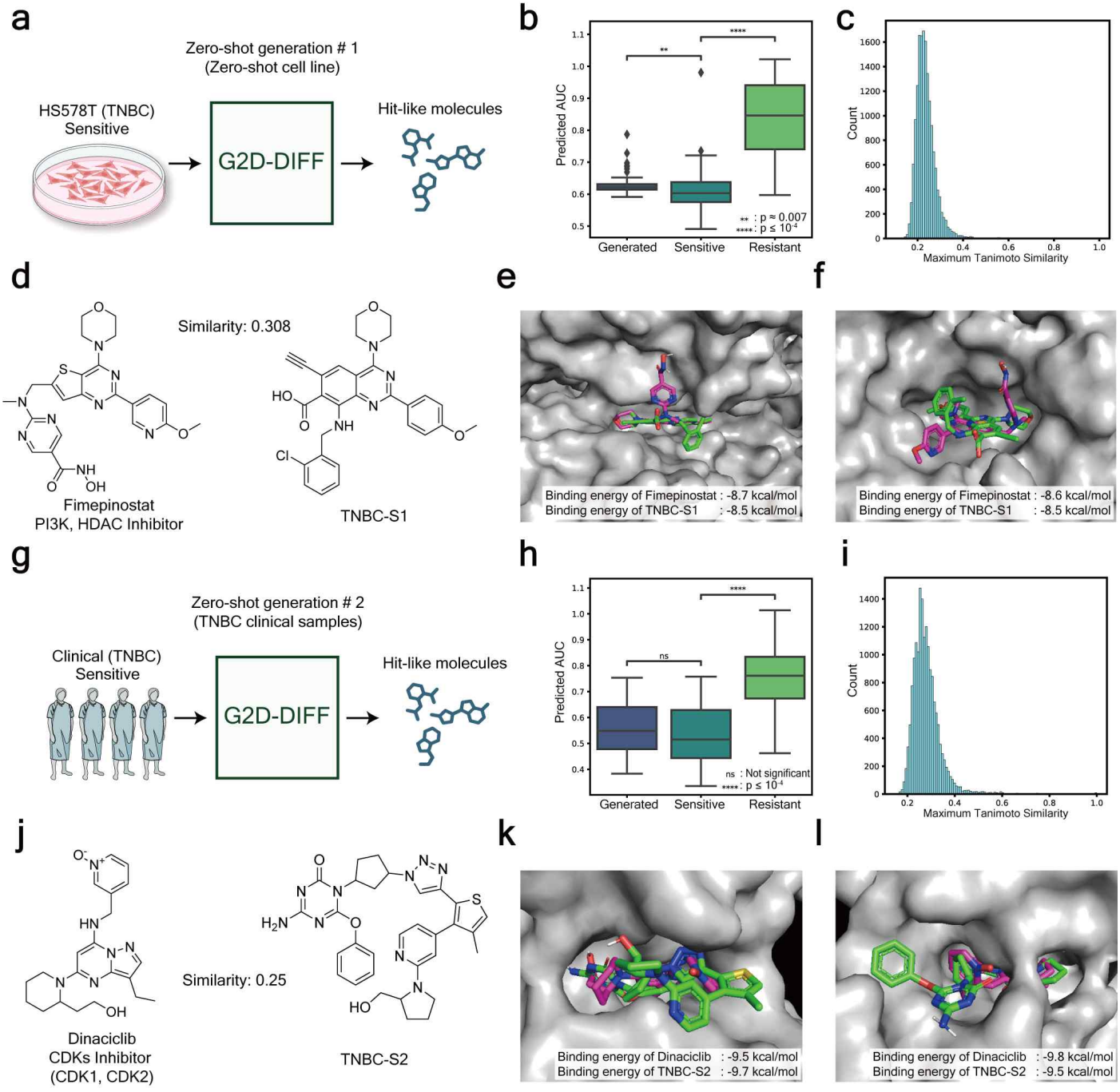
Through computer docking simulation*, it was also confirmed that these compounds can actually bind to target proteins of cancer cells.

This can be seen as a groundbreaking achievement that shows that AI can not only imitate existing drugs but can analyze the genetic characteristics of each patient and design completely new treatments optimized for them.

* PI3K, HDAC, CDK: These are key proteins involved in the growth, differentiation, survival, and cell cycle control of cancer cells, and are attracting attention as major targets for the development of anticancer drugs. PI3K regulates the intracellular survival signaling

pathway, HDAC is an epigenetic enzyme that regulates gene expression, and CDK is an enzyme that regulates cell cycle progression, and their abnormal activity is closely related to tumor occurrence and progression.

* docking simulation: This is a method that uses computers to predict how drug molecules bind to target proteins. By analyzing the interaction between drugs and proteins through this simulation, it is possible to evaluate in advance whether the drug can effectively bind to the target protein.



▲ Contents of the study to verify the applicability of G2D-Diff. a) Overview of the generation experiment for the triple-negative breast cancer cell line HS578T. b-c) Results of confirming the condition suitability and high structural diversity of the generated molecules. d) Analysis of structural similarity between the generated molecules and existing drugs. e-f) Results of docking simulation with the target protein. g) Overview of the generation experiment for the triple-negative breast cancer patient group. h-i) Results of confirming the condition suitability and high structural diversity of the generated molecules. j) Analysis of structural similarity between the generated molecules and existing drugs. k-l) Results of docking simulation with the target protein.

Another strength of G2D-Diff is its ‘interpretability’. This model can use the attention mechanism* to identify which genes or biological pathways are important for drug design in cancers with specific genotypes.

This allows it to go beyond simply generating new molecules and provide scientific evidence to support the validity of the treatment by explaining why the molecule is effective at the genetic and biological pathway level.

Unlike existing 'black box*' AI models, this is an important technological advancement that helps researchers understand why specific molecules were generated.

* attention mechanism: A technique in which AI focuses on important parts of various information

* black box: A system whose internal working principles are unknown

Professor Hojung Nam said, "This study opens up new possibilities for personalized medicine, and we expect that AI technology will provide new hope to patients with incurable cancer."

Dr. Hyunho Kim (first author) emphasized that "G2D-Diff can drastically improve the efficiency of the early candidate substance exploration stage, which is the most difficult in the new drug development process, and thus can significantly shorten the development period of anticancer drugs."

This study, supervised by Professor Hojung Nam of the Department of Electrical Engineering and Computer Science at GIST and conducted by Dr. Hyunho Kim (currently a senior researcher at the National Institute of Toxicology), Bongsung Bae, Minsu Park, and Yewon Shin, who are integrated master's and doctoral students, and Professor Trey Ideker (University of California, San Diego), was supported by the ▲ Ministry of Science and ICT and the National Research Foundation of Korea's Mid-career Researcher Support Project, ▲ Ministry of Health and Welfare and the Ministry of Science and ICT's Joint Learning-based New Drug Development Acceleration Project (K-MELLODDY), ▲ National Institute of Toxicology's Basic Project, and ▲ National Institute of Toxicology's Bridge2AI Program. The results of the study were published online in the international academic journal 《Nature Communications》 on July 1, 2025.

GIST
Since 1993