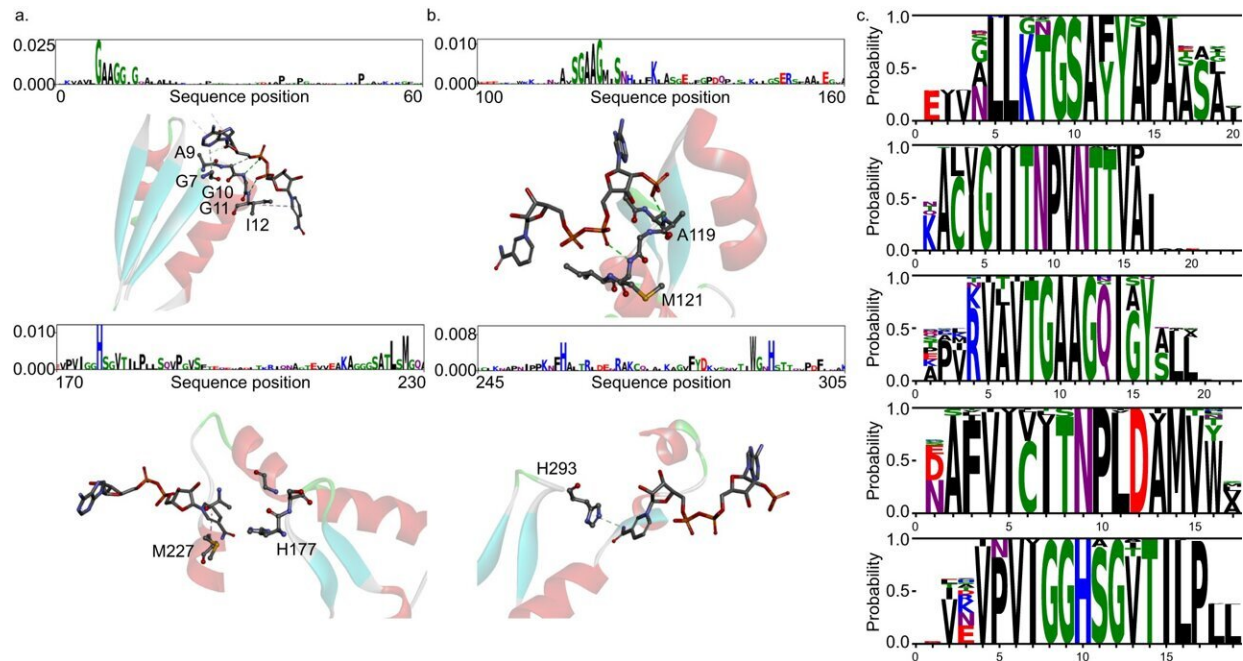# Researchers build an enzyme-discovering AI

November 24 2023

Highlighted amino acid residues by the DeepECtransformer neural network. Credit: *Nature Communications* (2023). DOI: 10.1038/s41467-023-43216-z

While E. coli is one of the most studied organisms, the function of 30% of proteins that make up E. coli has not yet been clearly revealed. For this, an artificial intelligence was used to discover 464 types of enzymes from the proteins that were unknown, and the researchers went on to verify the predictions of three types of proteins that were successfully identified through in vitro enzyme assay.

A joint research team, including Gi Bae Kim, Ji Yeon Kim, Dr. Jong An Lee and Distinguished Professor Sang Yup Lee of the Department of Chemical and Biomolecular Engineering at KAIST, and Dr. Charles J. Norsigian and Professor Bernhard O. Palsson of the Department of Bioengineering at UCSD, has developed DeepECtransformer, an artificial intelligence that can predict the enzyme functions from the protein sequence. In addition, the team has established a prediction system by utilizing the AI to quickly and accurately identify the enzyme function.

The team's work is described in the paper titled, "Functional annotation of enzyme-encoding genes using deep learning with transformer layers." The paper was published on 14 November in *Nature Communications*.

Enzymes are proteins that catalyze biological reactions, and identifying the function of each enzyme is essential to understanding the various chemical reactions that exist in living organisms and the metabolic characteristics of those organisms.

Enzyme Commission (EC) number is an enzyme function classification system designed by the International Union of Biochemistry and Molecular Biology, and in order to understand the metabolic characteristics of various organisms, it is necessary to develop a technology that can quickly analyze enzymes and EC numbers of the enzymes present in the genome.

Various methodologies based on deep learning have been developed to analyze the features of biological sequences, including protein function prediction, but most of them have a problem of a black box, where the inference process of AI cannot be interpreted.

Various prediction systems that utilize AI for enzyme function prediction have also been reported, but they do not solve this black box

problem, or cannot interpret the reasoning process at a fine-grained level (e.g., the level of amino acid residues in the enzyme sequence).

Output EC number

↑

| Linear |

↑

| Convolutional |

↑

| Convolutional |

↑
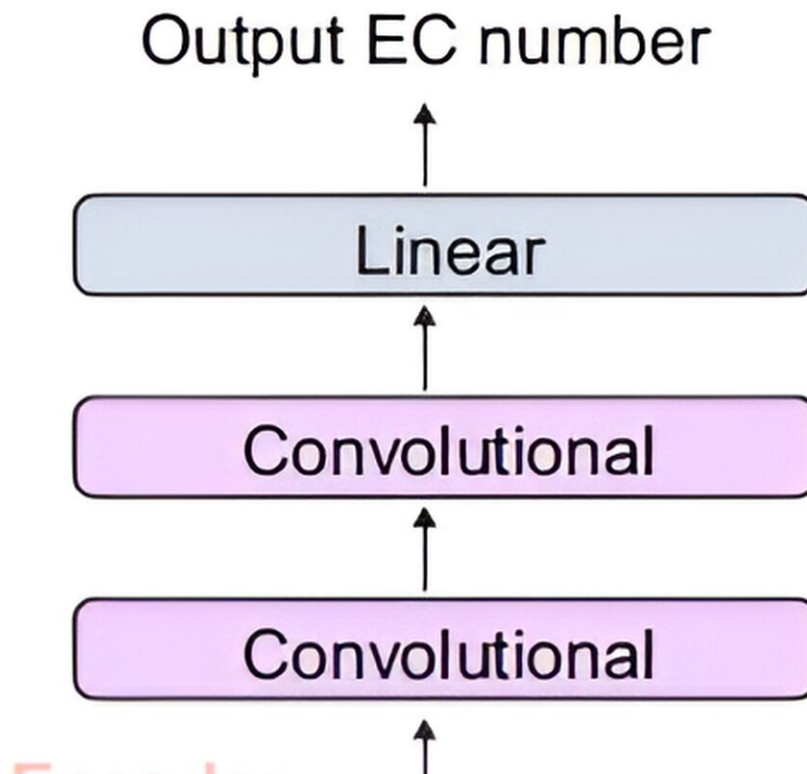
The structure of DeepECtransformer's artificial neural network. Credit: The Korea Advanced Institute of Science and Technology (KAIST)

The joint team developed DeepECtransformer, an AI that utilizes deep learning and a protein homology analysis module to predict the enzyme function of a given protein sequence.

To better understand the features of protein sequences, the transformer architecture, which is commonly used in natural language processing, was additionally used to extract important features about enzyme

functions in the context of the entire protein sequence, which enabled the team to accurately predict the EC number of the enzyme. The developed DeepECtransformer can predict a total of 5360 EC numbers.

The joint team further analyzed the transformer architecture to understand the inference process of DeepECtransformer, and found that in the inference process, the AI utilizes information on catalytic active sites and/or the cofactor binding sites which are important for enzyme function. By analyzing the black box of DeepECtransformer, it was confirmed that the AI was able to identify the features that are important for enzyme function on its own during the learning process.

"By utilizing the prediction system we developed, we were able to predict the functions of enzymes that had not yet been identified and verify them experimentally," said Gi Bae Kim, the first author of the paper.

"By using DeepECtransformer to identify previously unknown enzymes in living organisms, we will be able to more accurately analyze various facets involved in the metabolic processes of organisms, such as the enzymes needed to biosynthesize various useful compounds or the enzymes needed to biodegrade plastics," he added.

"DeepECtransformer, which quickly and accurately predicts enzyme functions, is a key technology in functional genomics, enabling us to analyze the function of entire enzymes at the systems level," said Professor Sang Yup Lee.

He added, "We will be able to use it to develop eco-friendly microbial factories based on comprehensive genome-scale metabolic models, potentially minimizing missing information of metabolism."

 More information: Gi Bae Kim et al, Functional annotation of