# AI successfully simulates molecular evolution to create new glowing protein



An artificial intelligence (AI) model has successfully simulated half a billion years of molecular evolution to generate the code for a previously unidentified protein. This breakthrough, which was detailed in a study published in the journal Science on January 16, unveils a glowing protein, known as esmGFP, that bears similarities to proteins found in jellyfish and corals. Researchers suggest that this newly created protein may play a significant role in the advancement of new medicinal therapies.

Proteins are essential biomolecules that perform various critical functions in the body, such as muscle construction and disease inhibition. The esmGFP, while only manifesting as a computer code, features the genetic blueprint for an entirely new type of green fluorescent protein. In nature, green fluorescent proteins are responsible for the luminescent qualities of certain jellyfish and corals.

The sequence of instructions that constitute the esmGFP is 58% similar to the closest known fluorescent protein, a modified version derived from a bubble-tip sea anemone (Entacmaea quadricolor). The remaining sequence is distinct, necessitating 96 genetic mutations, which would have taken over 500 million years to evolve naturally, as per the researchers involved in the study.

The AI model responsible for this development is named ESM3, created by a team at EvolutionaryScale. The unveiled protein, along with the AI model, was initially presented in a preprint study last year. The findings have since been peer-reviewed by independent scientists. Alex Rives, co-founder and chief scientist of EvolutionaryScale, elaborated on the functionality of ESM3, stating that it operates beyond the conventional constraints of evolutionary design. "We've found that ESM3 learns fundamental biology, and can generate functional proteins outside the space explored by evolution," he explained in correspondence with Live Science.

The research conducted by Rives and his colleagues traces back to their time at Meta, the parent company of Facebook and Instagram. The ESM3 model can be likened to generative language models such as OpenAI's GPT-4, albeit focused on biological constructs rather than textual data.

Proteins are composed of chains of molecules called amino acids, which are sequenced according to genetic instructions. Variations in the amino acid sequence yield different proteins, each characterised by a unique three-dimensional structure that facilitates its specific functions. To enable ESM3 to understand the intricacies of proteins, researchers supplied the model with extensive data on the properties of proteins, including their sequence, structure, and function.

For its training, the team utilised data from approximately 2.78 billion naturally occurring proteins. Subsequently, they introduced random omissions in a protein blueprint for ESM3 to decrypt, expertly filling in the blanks with learned information. "The same way a person can fill in the blanks in the soliloquy 'to \_ or not to *, that is the* ,' we can train a language model to fill in the blanks in proteins," Rives noted.

The applications of ESM3's capabilities extend to various domains of protein engineering, including the design of new pharmaceuticals. Green fluorescent proteins, which are frequently employed in scientific research to trace protein activity, exemplify the utility of engineering proteins for specific functions.

Tiffany Taylor, an evolutionary biologist at the University of Bath in the U.K. who was not involved with the research, commented on the potential impact of AI models like ESM3. In her review of the preprint study, published in Live Science, she asserted that such AI technologies enable innovations in protein engineering that traditional evolution cannot achieve. However, she expressed caution about the assumption that these models can surpass the intricate processes shaped by natural selection over millions of years. "AI-driven protein engineering is intriguing, but I can't help feeling we might be overly confident in assuming we can outsmart the intricate processes honed by millions of years of natural selection," Taylor stated.

Source: [Noah Wire Services](https://www.noahwire.com)

## References

* <https://www.biorxiv.org/content/10.1101/2024.07.01.600583v1.full> - This preprint study discusses the generation of a new green fluorescent protein, esmGFP, using the AI model ESM3, simulating over 500 million years of evolution.
* <https://english.elpais.com/science-tech/2025-01-20/ai-simulates-500-million-years-of-evolution-to-discover-artificial-fluorescent-protein.html> - This article explains how AI simulated 500 million years of evolution to create a new fluorescent protein, highlighting its potential in biotechnology.
* <https://www.noahwire.com> - The source article provides an overview of the breakthrough in AI-driven protein engineering and its implications for future research.
* <https://www.biorxiv.org/content/10.1101/2024.07.01.600583v1.full.pdf> - This PDF version of the preprint study offers detailed insights into the ESM3 model and its role in generating the esmGFP protein.
* <https://www.science.org> - Although not directly linked to the specific study, Science.org is a journal where such breakthroughs in molecular evolution and protein engineering are often published.
* <https://www.livescience.com> - Live Science features reviews and discussions about AI-driven protein engineering, including comments from experts like Tiffany Taylor.
* <https://www.meta.com> - Meta, the parent company of Facebook and Instagram, is mentioned as the background for the researchers involved in developing ESM3.
* <https://www.openai.com> - OpenAI's GPT-4 is mentioned as a comparable generative language model, though ESM3 focuses on biological constructs.
* <https://www.bath.ac.uk> - The University of Bath is mentioned as the affiliation of evolutionary biologist Tiffany Taylor, who commented on the study.