

EvolutionaryScale releases ESM3 protein model and raises \$142m seed round

Source copy kept by wheresthe.ai for the story "EvolutionaryScale releases ESM3 protein model and raises \$142m seed round" (<https://wheresthe.ai/d/evolutionaryscale-esm3-2024-06/>).

Source: <https://www.evolutionaryscale.ai/blog/esm3-release>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: 61f6f99e8f8bb2b13096e97e197ff5e90a80e6d2ba4d093e12c62a4091c70e72

6.25.24 • Announcement •

ESM3: Simulating 500 million years of evolution with a language model

by Tom Sercu

[Preview our paper](#)

Update January 2025: We're thrilled to announce the publication of ESM3 in Science Magazine . We're opening our API for biological intelligence, now in public beta, allowing scientists in academia and industry a free limited time preview of the capabilities of some of our models through Forge .

EvolutionaryScale More than 3.5 billion years ago, life on Earth emerged from chemical reactions. Nature invented RNA, proteins, and DNA, the core molecules of life, and created the ribosome, a molecular factory that builds proteins from instructions in the genome. Proteins are wondrous dynamic molecules with incredible functions—from molecular engines that power motion, to photosynthetic machines that capture light and convert it to energy, scaffolding that builds the internal skeletons of cells, complex sensors that interact with the environment, and information processing systems that run the programs and operating system of life. Proteins underlie disease and health, and many life-saving medicines are proteins. Biology is the most advanced technology that has ever been created, far beyond anything that people have engineered. The ribosome is programmable—it takes the codes of proteins in the form of RNA and builds them up from scratch—fabrication at the atomic scale. Every cell in every organism on earth has thousands to millions of these molecular factories. But even the most sophisticated computational tools created to date barely scratch the surface: biology is written in a language we don't yet understand. If we could learn to read and write in the code of life it would make biology programmable. Trial and error would be replaced by logic, and painstaking experiments by simulation. As we introduce ourselves as a new company, we're excited to present ESM3—a frontier language model for the life sciences that advances our ability to program and create with the code of life. ESM3 takes a step towards the future where AI is a tool to engineer biology from first principles in the same way we engineer structures, machines and microchips, and write computer programs. In a new preprint, we describe the generation of a new green fluorescent protein (GFP). Fluorescent proteins are responsible for the glowing colors of jellyfish and corals, and are important tools in modern biotechnology. esmGFP, our new protein, has a sequence that is only 58% similar to the closest known fluorescent protein. From the rate of diversification of GFPs found in nature, we estimate that this generation of a new fluorescent protein is equivalent to simulating over 500 million years of evolution. The power and potential of these new technologies call for a commitment to principles of responsible development, including transparency and accountability from the start. To that end, drawing on our experience as scientists and researchers, we have crafted a responsible development framework that will guide our progress.

ESM3: A frontier language model for biology Today we are sharing ESM3, the first generative model for biology that simultaneously reasons over the sequence, structure, and function of proteins. ESM3 is trained across the natural diversity of the Earth—billions of proteins, from the Amazon rainforest, to the depths of the oceans, extreme environments like hydrothermal vents, and the microbes in a handful of soil. Trained on one of the highest throughput GPU clusters in the world today, ESM3 is a frontier generative model for biology created at the leading edge of parameters, computational power, and data. We believe that ESM3 is the most compute ever applied to training a biological model, trained with over 1×10^{24} FLOPS and 98B parameters. Across AI we see the power of scaling. As model scale increases in parameters, data, and compute, larger models gain new emergent capabilities that smaller models lack. In many different domains generalist models trained on diverse data are outperforming specialist models. The incredible pace of new AI advances is being driven by increasingly large models, increasingly large datasets, and increasing computational power. The same patterns hold true in biology. In research over the last five years, the ESM team has explored scaling in biology. We find that as language models scale they develop an understanding of the underlying principles of biology, and

discover biological structure and function. ESM3 represents a milestone model in the ESM family—the first created by our team at EvolutionaryScale, an order of magnitude larger than our previous model ESM2, and natively multimodal and generative. Reasoning over the sequence, structure, and function of proteins. Language models operate over discrete units, or tokens. To create one that can reason over three of the fundamental biological properties of proteins—sequence, structure, and function—we had to transform three dimensional structure and function into discrete alphabets, and construct a way to write every three dimensional structure as a sequence of letters. This allows ESM3 to be trained at scale, unlocking emergent generative capabilities. ESM3's vocabulary bridges sequence, structure, and function all within the same language model. ESM3 is trained with a simple objective. For each protein, its sequence, structure, and function are extracted, tokenized, and partially masked. ESM3's task is to predict the masked positions using the masked language modeling objective inspired by natural language processing models. In order to accomplish this task, ESM3 must learn a deep understanding of the connection between sequence, structure, and function across evolutionary-scale data. When scaled across billions of proteins and billions of parameters, ESM3 learns to simulate evolution. Given the limited volume of experimentally determined structure and function annotations, we augment ESM3's multimodal training dataset with hundreds of millions of synthetic data points, including predicted structures and functions for diverse sequences.

ESM3 is a multi-track transformer that jointly reasons over protein sequence, structure, and function. Programming biology. ESM3 is a generative model, and makes biology programmable. It can follow prompts to generate new proteins. Scientists can interact with ESM3, guiding it to create for a myriad of applications such as for medicine, biology research, and clean energy. Proteins can be generated by starting with a fully masked set of tokens and iteratively unmasking until all positions are filled. Because sequence, structure, and function are all masked and predicted during training, ESM3 can generate in all three modalities. This generation process can also be guided by any combination of partial or full specification of sequence, structure, and function. ESM3's multimodal reasoning power enables scientists to generate new proteins with an unprecedented degree of control. For example, the model can be prompted to combine structure, sequence, and function to propose a potential scaffold for the active site of PETase, an enzyme that degrades polyethylene terephthalate (PET), a target of interest to protein engineers for breaking down plastic waste.

ESM3 designs a scaffold for the PETase active site through multimodal prompting with sequence, structure, and function. We prompt ESM3 with the active site structure and amino acids, as well as a function keyword prompt for α/β hydrolase, a fold characteristic of hydrolytic enzymes. Emergence of capabilities with scale. ESM3's ability to solve challenging protein design tasks emerges with scale. One such task, atomic coordination, is designing a protein from prompts that specify the atomic positions of amino acids that are distant in the sequence but close in the structure. This measures the ability of the model to achieve atomic level accuracy in structure generation, which is critical for designing functional proteins. ESM3's ability to solve these tasks improves with scale, i.e., ESM3 solves harder generative problems as a function of scale. ESM3 further improves with feedback using alignment methods similar to Reinforcement Learning from Human Feedback (RLHF) applied in LLMs. Rather than receiving feedback from humans, ESM3 can self-improve, providing feedback on the quality of its own generations. Feedback from laboratory experiments or existing experimental data could also be applied to align ESM3's generations with biological success.

ESM3 models evaluated on the task of generating proteins that satisfy atomic coordination prompts. ESM3 solves harder generative problems as a function of scale, and the capabilities of larger models become more apparent after alignment. Simulating 500 million years of evolution Green fluorescent protein, commonly known as GFP, and its family of fluorescent proteins are some of the most beautiful proteins in nature. They occur in just a few branches of the tree of life. The discovery of GFP led to the award of a Nobel Prize and has become one of the most widely used tools in biology, allowing scientists to see proteins within cells. GFP contains a fluorescent chromophore—a molecular component that absorbs a single photon of one color at a short wavelength, captures some of its energy, and releases the rest as a new photon with a different color and longer wavelength. The natural GFP absorbs blue light and emits green light. GFP is a protein that transforms itself—its structure is an eleven stranded barrel with a helix that threads through its center—and after GFP folds, a spontaneous reaction occurs. At the center of the protein, the atoms that form the protein chain reform into a new configuration creating a fluorescent chromophore. This mechanism is unique. No other known protein spontaneously forms a fluorescent chromophore out of its own structure, suggesting that producing fluorescence is hard even for nature. Scientists have discovered many variants of GFP in nature, and created variants of those natural proteins in the lab. The very first artificial variants were found by making a handful of mutations that increased brightness or changed color. With more recent laboratory and machine learning techniques it has been possible to extend this search to

find more distant variants that differ by up to even 20% of the sequence. But still the bulk of variation of functional GFPs have come not from protein engineering but by prospecting the natural world. The process of evolution that gives rise to new fluorescent proteins takes epochs of time—the story of this protein family reaches back into depths of natural history and geologic time where somewhere in the distant past nature invented the first fluorescent protein. Natural fluorescent proteins have diverged over 100s of millions of years from ancestral sequences in deep history to become the proteins they are today. Prompted with the structure of a few residues in the core of natural GFP, ESM3 reasoned in a chain-of-thought to generate candidates of new GFPs. Generating one by pure chance from among an astronomical number of sequences and structures ($20\,229 \times 4096\,229$ to be precise—more possibilities than the number of atoms in the visible universe) would be virtually impossible. We tested 96 generations in a first experiment, and found a number of proteins that fluoresce, including one that was far from any protein in nature. This protein, located in well B8 of our experiment plate, was 50x less bright than natural GFPs and its chromophore matured over the course of a week, instead of in under a day, but it presented a signal of function in an unexplored portion of sequence space. Continuing the chain-of-thought starting from the sequence of B8, we generated another set of 96 proteins. We tested them and found several proteins that have similar brightness to natural GFPs, including the brightest one in well C10 which we call esmGFP. esmGFP differs by 96 mutations (out of 229 amino acids, 58% of the sequence is similar) to the closest fluorescent protein found in nature.

In a sequence of two experiments ESM3 generates B8, a dim GFP that is distant from all GFPs known in nature. Starting from B8, ESM3 generates esmGFP, a distant GFP with brightness similar to other natural GFPs. Unlike nature, protein language models do not explicitly work within evolutionary constraints. But in order for ESM3 to solve its training task of predicting the next masked token the model must learn how evolution moves through the space of potential proteins. In this sense, ESM3 can be thought of as an evolutionary simulator. A traditional evolutionary analysis of the ancestry of esmGFP is paradoxical as the protein was created outside natural processes, but still we can draw insight from the tools of evolutionary biology on the amount of time it would take for a protein to diverge from its closest sequence neighbor through natural evolution. We find naturally occurring GFPs with similar levels of sequence identity are separated by hundreds of millions of years of evolution. Using an analysis similar to one might perform on a new protein found in the natural world, we estimate that esmGFP represents an equivalent of over 500 million years of natural evolution performed by an evolutionary simulator.

A rendering of esmGFP, a new green fluorescent protein generated by ESM3 that is distant from other fluorescent proteins found in nature. Responsible Development EvolutionaryScale is a public benefit company. Our mission is to develop artificial intelligence to understand biology for the benefit of human health and society, through partnership with the scientific community, and open, safe, and responsible research. Molecular biology has already undergone one inflection point at the dawn of the recombinant DNA era in the 1970s, when scientists developed the technology of genetic engineering. The outcome of that technological inflection was a revolution in our understanding of genetics, decoding the human genome, and groundbreaking new medicines. To guide their work during a time of rapid technological development, the scientific community developed a set of principles and recommendations at the Asilomar Conference in 1975. These principles have led to robust frameworks that help to manage risk are in use by nucleotide synthesis companies, molecular biology vendors, and regulators. As we now enter an era where we can design and program novel biology, we look towards the history of our field as well as new principles and recommendations being proposed by the growing community of researchers exploring the frontiers of biological design. With these inspirations, we have created a Responsible Development Framework to guide our work towards our mission with transparency and clarity. Open Model Since inception the ESM project has committed to open science with code and model releases, and our commitment continues. We believe that sharing research and code accelerates progress and contributes to understanding and reducing risk, ultimately maximizing positive impact for the world. It has been incredible to see the imaginative and impactful applications of ESM models across research and industry. For example, Hie et al. used ESM-1v and ESM-1b to evolve antibodies, improving therapeutically relevant characteristics such as binding affinity, thermostability, and viral neutralization. BioNTech and InstaDeep fine-tuned an ESM language model on COVID spike proteins to detect variants that would pose a higher risk to public health, successfully flagging all 16 variants of concern before they were designated by the WHO. Brandes et al. used ESM-1b to predict the clinical effects of mutations, and is currently the most powerful method for this important task. Marsiglia et al. used ESM-1v to engineer novel anti-CRISPR protein variants that maintain on-target editing functionality while reducing off-target side effects. Shanker et al. used ESM-IF1 to guide evolution of diverse proteins, including lab-validated high-potency antibodies against SARS-CoV-2. Yu et al. fine-tuned ESM-1b to predict function for enzymes, including rare and understudied enzymes, and experimentally validated the predictions. Rosen et al. used ESM2 embeddings to build

representations of genes in a single-cell foundation model . Høie et al. fine-tuned ESM-IF1 on antibody structures to achieve state-of-the-art performance in sequence recovery across CDR regions , to design antibodies with high binding affinity. This is just a small fraction of the amazing work building on the ESM platform! We will continue to develop and release open models to accelerate research and empower the scientific community. This starts with releasing the weights and code for an ESM3 1.4B open model under MIT license, to allow scientists and developers to build on the ideas and architecture of ESM3. We're excited to see what you create! Where We Are Headed We believe in a future where AI can help us to understand the complex systems of life at the most basic level, make new scientific discoveries that change our understanding of biology, help us to find cures to disease, and build a more sustainable world. ESM3 is a tool for scientists. Our API and open model allow scientists to explore the frontiers of protein design and synthetic biology, and invent new solutions for some of the most important problems facing our world. We are also developing specialized versions of ESM3 to unlock applications at the frontier of drug design. The same capabilities that can be used to design one of the most complex and beautiful proteins in nature will help scientists to create new medicines. ESM3 is only the first step on our roadmap for programming biology. We think the future will be increasingly multimodal models that learn from biological data and integrate across the scales of life from individual molecules to cells, which will contribute to humanity's ability to understand and program biology to build a better world.

Table of Contents

{links are generated when publishing the site}

Related Posts