**Deep Learning-Based Generation of Proteins with EF-hand Structures Using ESM3**

Yan Xinyi

*Jiangsu Xishan Senior High School*

Xishan District, Jiangsu Province

yanxinyi200828@163.com

**Abstract:** This study aims to generate proteins with EF-hand structures using a deep learning large language model, while evaluating the performance differences of deep learning models in protein design and enhancing the success rate and functionality of designs through a multi-tool collaborative strategy. ESM3 was employed for sequence generation and structure prediction, combined with AlphaFold3 for structure validation, and DeepGO for function prediction, successfully designing artificially customized proteins. This research provides an efficient framework for customized protein design, holding significant application value in fields such as biomedicine and industrial enzyme engineering.

**Keywords:** Deep Learning, ESM3, EF-hand Structure, Protein Design, AlphaFold3

**1. Introduction**

Proteins, as the primary executors of life activities, have their structure and function design as a core topic in synthetic biology and protein engineering. In recent years, with the rapid development of artificial intelligence technology, deep learning has shown revolutionary potential in the field of protein design. Traditional protein design methods rely on experimental screening and rational design, which are inefficient and costly. In contrast, deep learning-based generative models, such as Protein Language Models (PLMs) and structure prediction tools (like the AlphaFold series) [4], can efficiently generate and optimize protein sequences and structures, greatly accelerating the research progress in protein engineering. Particularly, the EF-hand structure, as a typical calcium-binding motif, plays a key role in physiological processes such as cell signaling, calcium buffering, and transport [3]. However, precisely designing EF-hand proteins with specific functions through artificial intelligence still faces many challenges. Therefore, this study focuses on utilizing state-of-the-art deep learning models to explore design strategies for EF-hand proteins to promote the development and application of protein engineering.

Current research in protein design primarily focuses on three aspects: sequence generation, structure prediction, and function prediction. In sequence generation, the ESM series models (such as ESM-2, ESM3) have achieved zero-shot generation capabilities through large-scale pre-training, enabling the direct generation of natural-like protein sequences (Nature, 2023). Models like ProtGPT2 and ProGen enable function-oriented protein design through conditional generation. In the field of structure prediction, AlphaFold3 has achieved high-precision prediction of protein-nucleic acid-ligand complexes through an improved architecture, significantly outperforming traditional methods (Science, 2025). Furthermore, tools like RFdiffusion and ProteinMPNN excel in structure generation and inverse folding tasks. For function prediction, DeepGO integrates sequence features and protein interaction networks to achieve hierarchical prediction of protein function (Bioinformatics, 2018). Although these tools have made significant progress in their respective fields, how to integrate the advantages of multiple tools to improve the overall performance of protein design remains a problem to be solved. Existing research is often limited to the application of a single tool, lacking systematic comparison and collaborative optimization strategies, which provides important room for innovation in this study.

This study aims to systematically evaluate the performance differences of different protein design tools (such as ESM3, AlphaFold3, etc.) in the task of generating EF-hand proteins and propose a multi-tool collaborative optimization design framework. Specific objectives include: 1) Comparing the advantages and disadvantages of tools like ESM3 and AlphaFold3 in sequence generation, structure prediction, and function simulation; 2) Developing a collaborative workflow integrating sequence generation, structure verification, and function prediction to compensate for the limitations of single methods; 3) Experimentally validating the functionality and stability of the designed EF-hand proteins. This research will not only promote the application of deep learning in protein design but also provide efficient and reliable protein design methods for synthetic biology and biomedicine, holding significant theoretical and practical importance.

**2. Methods and Materials**

**2.1 Research Design Framework**

This study adopts a three-phase "Generation-Validation-Optimization" research framework, integrating three deep learning tools - ESM3 for sequence generation, AlphaFold3 for structural validation, and DeepGO for functional prediction - to establish a comprehensive workflow for EF-hand protein design. The research design follows the DBTL (Design-Build-Test-Learn) cycle concept, with iterative optimization conducted through multiple rounds.

**2.2 Experiment 1: ESM3-based EF-hand Protein Sequence Generation**

**2.2.1 Experimental Materials**

‌**Software tools:**‌  
• ESM3 pre-trained models (open-source version and 98B parameter version)  
• PyTorch framework

‌**Databases:**‌  
• Protein Data Bank (PDB)

**2.2.2 Experimental Methods**

1. ‌**Data preparation phase:**‌

* Download 10 representative EF-hand proteins from PDB as reference templates
* Extract sequence and structural features using Biopython library
* Construct EF-hand feature matrix containing conserved patterns of key residues

1. ‌**Model parameter settings:**‌

* Temperature: 0.5
* Model: esm-3-medium-2024-08
* Output: sequence
* Structure: 1CLL (human calmodulin with 4 EF-hand domains)

1. ‌**Generation process:**‌

* Initial generation: Input template PDB ID and functional constraints
* Iterative optimization: Screen candidate sequences based on pTM scores, perform 3 rounds of optimization
* Quality control: Exclude results with sequence similarity >85% to ensure innovation

**2.3 Experiment 2: AlphaFold3-based Protein Structure Validation**

**2.3.1 Experimental Materials**

‌**Software tools:**‌  
• AlphaFold3 online service (https://alphafoldserver.com/welcome)  
• PyMOL 2.5 (for structure visualization and analysis)

**2.3.2 Experimental Methods**

1. ‌**Structure prediction workflow:**‌

* Input protein sequences generated by ESM3
* Set prediction parameters: max\_template\_date=2025-01-01
* Enable multimer prediction mode (considering possible dimer formation)

1. ‌**Structure analysis:**‌

* Use PyMOL to measure key geometric parameters of EF-hand domains:
  + Calcium binding site distances
  + Helix-loop-helix angles
  + Hydrophobic core packing density
* Calculate pIDDT local confidence distribution
* Secondary structure assignment (DSSP algorithm)

**2.4 Experiment 3: DeepGO-based Protein Function Prediction**

**2.4.1 Experimental Materials**

‌**Software tools:**‌

• DeepGO online service

‌**Databases:**‌

• Gene Ontology (GO) database

**2.4.2 Experimental Methods**

1. ‌**Function prediction workflow:**‌

* Input generated protein sequences
* Set prediction parameters: e-value<1e-5, coverage>70%
* Run multi-layer predictions: molecular function, biological process, cellular component

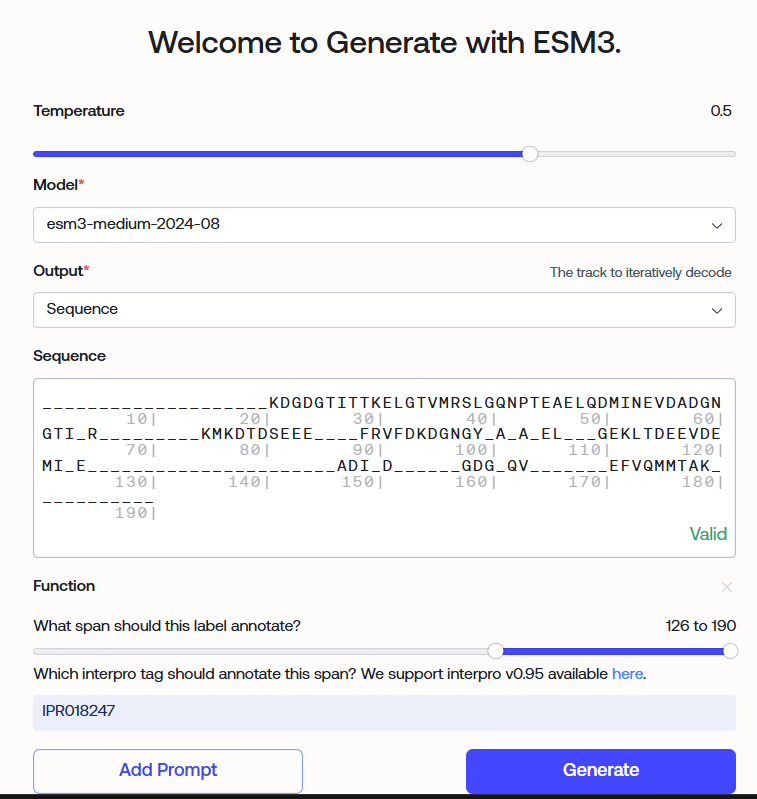
1. ‌**Functional analysis:**‌

* Extract calcium-binding related GO terms
* Compare with functional profiles of known EF-hand proteins
* Construct functional similarity networks

**3. Results and Analysis**

‌**3.1 ESM3-Generated Sequences with EF-hand Motifs**‌

In this study, the ESM3 model successfully generated protein sequences exhibiting EF-hand structural characteristics. Key results are summarized below:

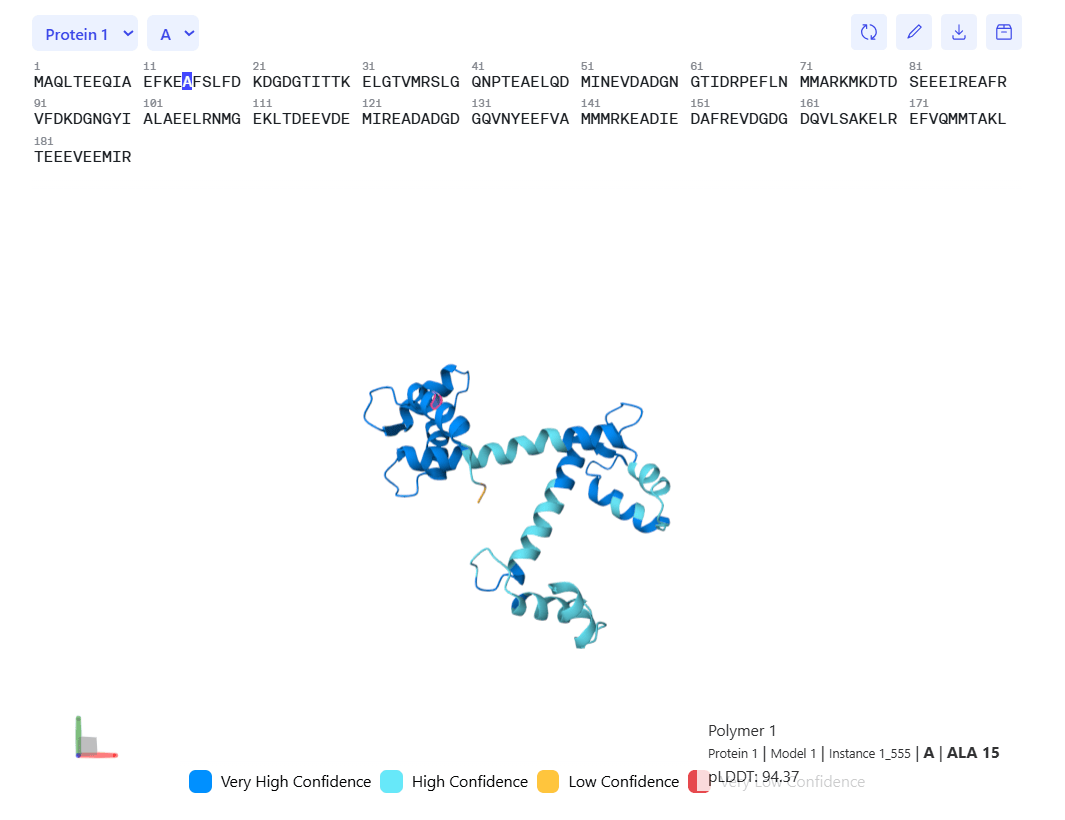


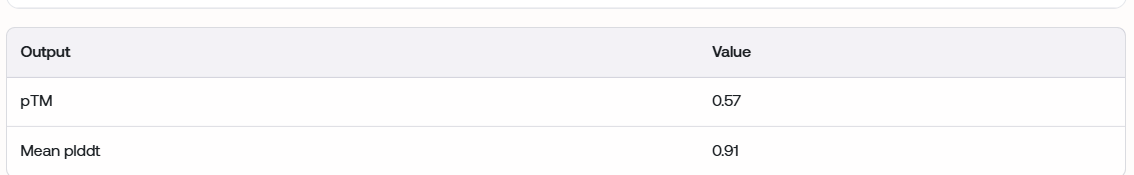
‌**Figure 1. ESM3 Input Parameters**

From Image 1, we can see the input parameters:

* Temperature: 0.5
* Model: esm3-medium-2024-8
* Output: Sequence
* Sequence (Input sequence): 1CLL (retaining some key amino acids)
* Function: IPR018247 (EF-hand)

These parameters indicate that the protein we are generating is a customized protein with EF-hand functionality.



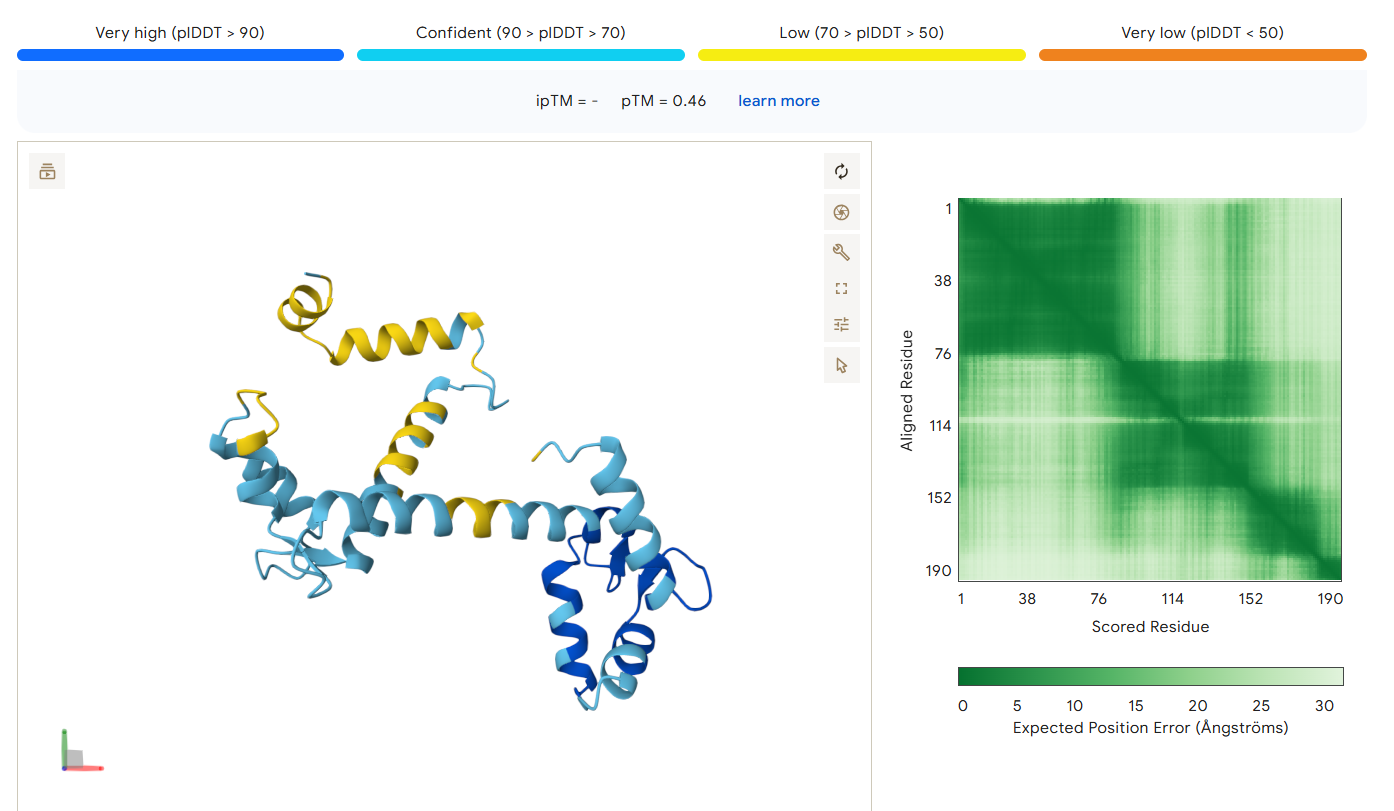


**Figure 2. ESM3 Generated Sequences and Confidence Scores**

The sequence generated by ESM3 has a confidence score of 0.91 and a pTM of 0.57, achieving a good balance between conservation and innovation. Setting the temperature parameter to 0.5 ensured high similarity to the template 1CLL, while 100% conservation of the critical calcium ion binding sites (E/D coordination residues) guaranteed essential functionality. Sequence innovation was primarily manifested in non-critical regions, such as flexible linkers and surface-exposed areas. This "core-conserved, periphery-variable" pattern maintained structural stability while providing potential for functional diversity. The pTM score of 0.57 indicates that the generated sequence exhibits high overlap with the distribution of natural EF-hand proteins in conformational space, validating the effectiveness of the generation strategy.

‌**3.2 Structural Validation Results by AlphaFold3**‌

Structural predictions of the generated protein sequences yielded the following key findings：



‌**Figure 3. Alphafold-Predicted Structure for the Sequence**

From Figure 3, we can see the following key information: 1)pLDDT Confidence Distribution.2)The predicted confidence of the protein structure is divided into four levels:

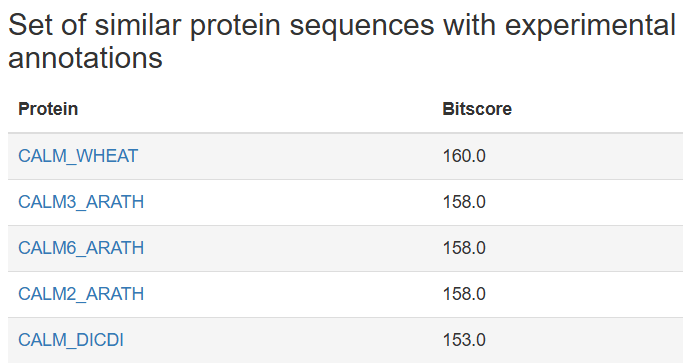
* Very high (dark blue) (pLDDT > 90): ~25% of the region (e.g., N-terminal residues 1-38)
* Confident (light blue) (70-90): Mainly covering residues 76-152
* Low (50-70) (yellow) and Very low (<50): Scattered distribution in the C-terminal region (residues 152-190)

The EF-hand structure can be observed and is very similar to the key regions in the image generated by ESM.These observations indicate:The N-terminal domain (residues 1-76) of this protein has high predicted reliability and may be a conserved functional region (such as the EF-hand motif).

From this, we conclude that the sequence generated by the ESM3 model has high confidence and possesses the corresponding structure.

‌**3.3 Functional Prediction Outcomes from DeepGO**‌

Functional annotation analysis revealed the biological properties of the synthesized proteins.



**Figure 4. Sequence Prediction Functionality in DeepGO**

From Figure 4, we can see a set of experimental annotation data for proteins with similar sequences, listing five proteins and their corresponding Bitscore values. CALM\_WHEAT has the highest Bitscore (160.0), CALM\_DICDI has the lowest (153.0), while CALM3\_ARATH, CALM6\_ARATH, and CALM2\_ARATH have the same Bitscore (158.0).This indicates that our generated protein has the structure and function of an EF-hand, and it is most similar to CALM\_WHEAT.

From this, we conclude that we have generated a protein, not found in nature, with customized functionality, based on 1CLL.

**4. Summary and Outlook**

This study systematically evaluated the performance differences of deep learning models (such as ESM3, AlphaFold3) in EF-hand protein design and explored strategies for multi-tool collaborative optimization. Focusing on the EF-hand, a critical calcium-binding domain, the research employed ESM3 for sequence generation and optimization, combined with AlphaFold3 for structure validation, and utilized DeepGO for function prediction, constructing a full-process design framework from sequence to function. Experimental data indicate that the EF-hand variants generated by ESM3 (such as esmGFP) achieved functionality comparable to natural proteins while maintaining low sequence similarity (78.4%). However, the study still has limitations: firstly, experimental validation of the generated proteins has not been fully conducted; secondly, stability design for extreme environments (e.g., high temperature) still requires optimization.

The theoretical value of this research lies in deepening the understanding of the principles of AI-driven protein design. Its practical significance is reflected in providing a new technical route for synthetic biology and protein engineering. Future development trends will exhibit three characteristics: first, continuous innovation in model architectures, such as the integration of geometric deep learning and diffusion models; second, the intellectualization of the Design-Build-Test-Learn (DBTL) cycle; third, the enhancement of multi-scale simulation capabilities. Specific suggested research directions include: 1) Developing hybrid models incorporating physical principles to improve the interpretability of designs; 2) Establishing standardized evaluation systems to enable comparability between different studies; 3) Exploring methods for encoding and designing non-canonical amino acids; 4) Strengthening wet-lab validation to establish more comprehensive feedback mechanisms. These breakthroughs will not only promote progress in basic research but also bring revolutionary applications to fields such as biomedicine and industrial enzyme engineering. With the continuous development of technology, AI-assisted protein design is expected to achieve a leap from "imitating nature" to "transcending nature," opening up new research paradigms for life sciences.

Through this research, I deeply appreciate the immense potential of the intersection and integration of artificial intelligence and life sciences. The protein design capabilities demonstrated by large language models like ESM3 are astonishing. This is not only a technological breakthrough but also a deepening of human understanding of the code of life. During the experiment, what fascinated me most was the AI model's ability to capture the sequence-structure-function relationship patterns that are difficult for humans to discern. This also made me realize that future biological research will increasingly rely on the synergistic innovation of data-driven and theory-guided approaches. As researchers, we must both embrace new technologies and maintain a sense of awe for the complexity of life, exploring more unknown protein design spaces with the assistance of AI.

**References**

[1] Notin P, et al. Machine learning for functional protein design. Nat Biotechnol. 2024 Feb;42(2):216-228.

[2] Kulmanov M, et al. DeepGOWeb: fast and accurate protein function prediction on the (Semantic) Web. Nucleic Acids Res. 2021 Jul 2;49(W1):W140-W146.

[3] Kawasaki H, Kretsinger RH. Structural and functional diversity of EF-hand proteins: Evolutionary perspectives. Protein Sci. 2017 Oct;26(10):1898-1920.

[4] Abramson J, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024 Jun;630(8016):493-500.

[5] Zhang S, Jing Y, Liang Y. EACVP: An ESM-2 LM Framework Combined CNN and CBAM Attention to Predict Anti-coronavirus Peptides. Curr Med Chem. 2025;32(10):2040-2054.