A generalized protein design ML model enables generation of functional de novo proteins

Timothy P. Riley • Oleg Matusovsky • Mohammad S. Parsa • Pourya Kalantari • Ismail Naderi • Kooshiar Azimian • Kathy Y. Wei*

ABSTRACT

Despite significant advancements, the creation of functional proteins de novo remains a fundamental challenge. Although deep learning has revolutionized applications such as protein folding, a critical gap persists in integrating design objectives across structure and function. Here, we present MP4, a transformer-based AI model that generates novel sequences from functional text prompts that enables the design of fully folded, functional proteins from minimal input specifications. Our approach demonstrates the ability to generate entirely novel proteins with high experimental success rates or effectively redesign existing proteins. This transformer-based model highlights the potential of generalist AI to address complex challenges in protein design, offering a versatile alternative to specialized approaches.

HIGHLIGHTS

- Text-to-Protein Generation: MP4 translates functional text prompts into de novo protein sequences.
- Generalist vs. Specialized Models: Unlike structure-first approaches, MP4 directly generates functional proteins.
- High Success Rate: 84% of designed proteins were experimentally validated with stable expression and functional properties.
- Exploring Novel Sequence Space: MP4 generates proteins significantly different from natural sequences while maintaining stability.
- Thermostability & Expression: Many generated proteins exhibit robust thermostability (>62°C) and high expression yields.

310.AI

*kwei@310.ai

OVERVIEW OF THE MP4 MODEL

MP4 is a transformer-based generative model for de novo protein design. It accepts natural language prompts that encode comprehensive protein information—such as physical properties, source organism, and sequence-related properties—and generates sequences that meet the desired constraints. The model emphasizes molecule programmability by integrating data from diverse sources and synchronizing multiple design objectives during training.

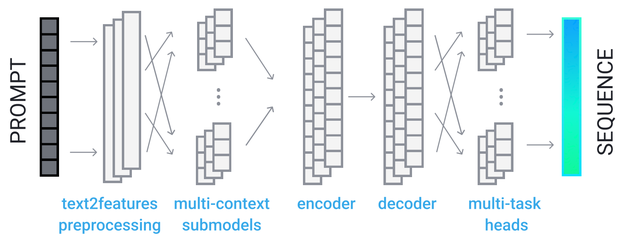


Figure 1: An overview of the MP4 architecture.

MP4 GENERATES PROTEIN SEQUENCES WITH HIGH PREDICTED FOLDABILITY AND FUNCTIONAL ACTIVITY

MP4 generates protein sequences with high foldability and biological relevance. The model was used on >1,000 diverse prompts specifying enzymatic activity, binding partners, and subcellular localization. The resulting sequences align with natural amino acid distributions, maintain structural integrity, and maintain functional alignment. Full repository: https://310.ai/mp/repo

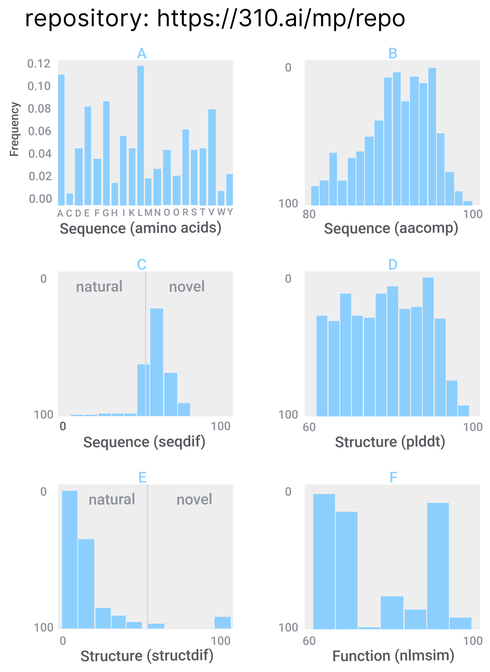


Figure 2:
Computational metrics of 1000+ AI designed proteins generated by MP4 model.

A) Frequency of 20 canonical amino acids.

B) Amino acid composition per sequence, normalized to UniProt database proteins.

C) Sequence comparison to NR/NT database proteins.

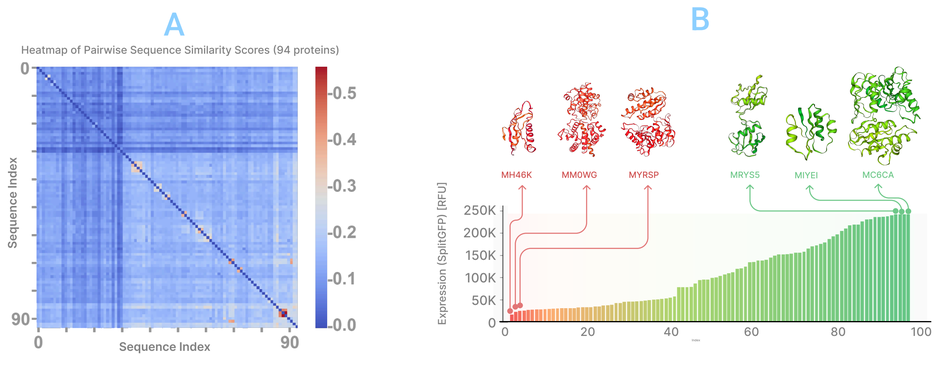
D) Averaged ESMFold confidence pLDDT.

E) Structure comparison to Protein Data Bank database proteins.

F) Functional similarity based on prompt and predicted sequence function using ProtNLM model.

PROTEINS GENERATED BY MP4 HAVE DESIRABLE EXPERIMENTAL PROPERTIES

MP4-generated proteins were experimentally validated to assess their stability and expression efficiency. A subset of 94 diverse sequences was cloned and expressed in a prokaryotic cell-free system, with 84% successfully producing measurable protein levels. Thermostability tests using differential scanning fluorimetry (DSF) revealed that many proteins remained stable above 62°C. Dynamic light scattering (DLS) confirmed minimal aggregation for selected samples. These results demonstrate that MP4 not only generates novel sequences but also optimizes key properties for experimental viability, highlighting its potential for real-world protein engineering applications. Experimental repository details: https://310.ai/mp/lab/1



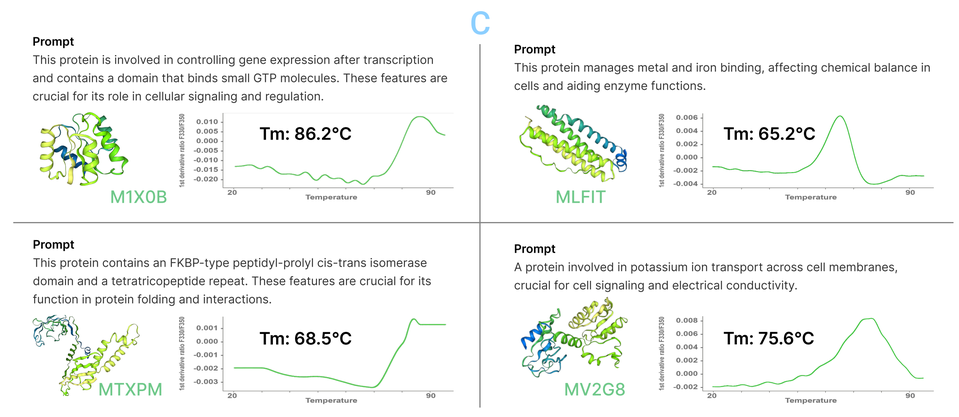


Figure 3: Experimental evaluation of 94 selected de novo designed proteins.
A) Pairwise sequence similarity heatmap.
B) Expression profile in a cell free expression system.
C) Thermostability, measured by DSF, of 4 diverse proteins.

PROPERTY INTERROGATION OF MP4 DESIGNED PROTEINS

MP4-designed proteins were analyzed to understand how computational metrics predict experimental behavior. While no strong correlation was found between secondary structure composition and expression levels, MP4-generated proteins exhibited diverse and viable scaffolds. Hydrophobicity, often linked to aggregation, showed only a weak correlation with expression success. A composite "developability" score incorporating hydrophobicity, charge, and solubility provided modest improvements in predicting expression levels.

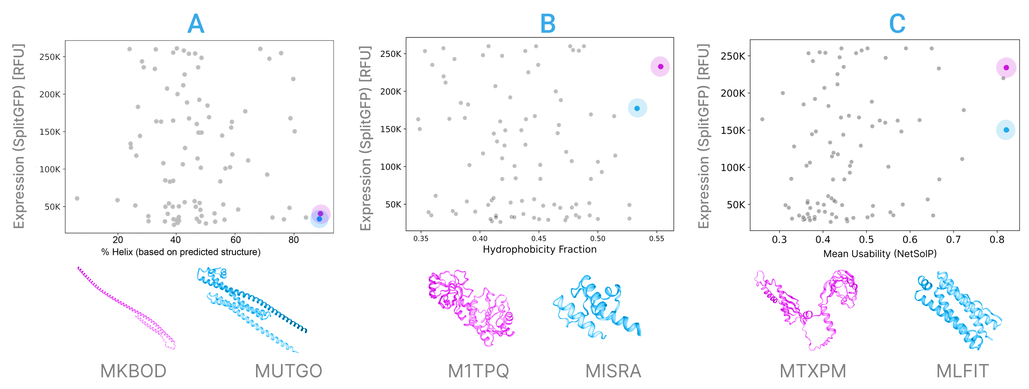


Figure 4: Structural and property analysis of designed proteins.
A) Alpha-helical content vs relative expression levels. B) Predicted hydrophobicity vs relative expression levels.
C) Predicted developability (usability score by NetSolP) vs relative expression levels.

AI DESIGNED A FUNCTIONAL PROTEIN FROM TEXT - NO TEMPLATES, 52 MUTS

Using the prompt "This protein has adenylate kinase activity and is involved in the metabolic process of ribonucleoside monophosphates", MP4 designed a 214-amino-acid enzyme that expressed solubly bacteria and purified to >95% purity. A measurable thermostability shift upon addition of ATP is detected, confirming ATP binding as designed. The AI design diverges from natural adenylate kinases by substituting leucine for the universally conserved phenylalanine at position 35. This bold mutation, unlikely to be attempted by a human, may signal that the AI has discovered a viable alternative unrecognized by evolution.

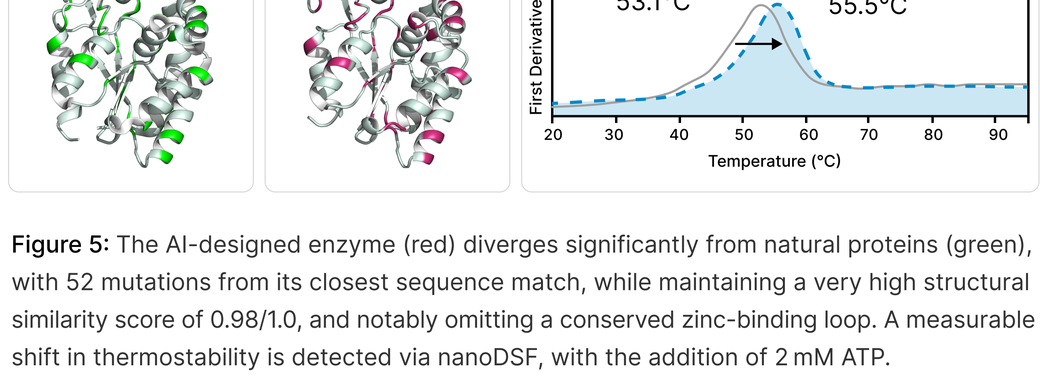


Figure 5: The AI-designed enzyme (red) diverges significantly from natural proteins (green), with 52 mutations from its closest sequence match, while maintaining a very high structural similarity score of 0.98/1.0, and notably omitting a conserved zinc-binding loop. A measurable shift in thermostability is detected via nanoDSF, with the addition of 2 mM ATP.

DISCUSSION

- This study demonstrates the capability of MP4 to generate protein sequences that exhibit desirable experimental properties, such as efficient expression and thermostability, while maintaining a high success rate in translation. The findings underscore the value of generalist protein design models, which consider a range of structural and functional properties simultaneously. By achieving measurable protein expression in 84% of the tested sequences and identifying several proteins with thermostability exceeding 65°C, MP4 highlights its potential as a versatile tool for rational protein design.
- Due to the diversity of functions in this set, it would be difficult to test each protein experimentally to verify its function. Instead, a separate set of designs was created focused on the function of ATP binding. These will be tested experimentally.
- While the current vocabulary understood by MP4 is constrained, future iterations will incorporate an expanded, precise, and technically sophisticated lexicon. This advancement will enable true molecular programming, where users can specify target protein properties—function, stability, binding affinity, and more—with fine control. The model will then generate optimized protein sequences in a single inference step, transforming biological design into a deterministic, programmable process.