A GENERALIZED PROTEIN DESIGN ML MODEL ENABLES GENERATION OF FUNCTIONAL DE NOVO PROTEINS

Timothy P. Riley, Pourya Kalantari, Ismail Naderi, Kooshiar Azimian, Kathy Y. Wei
310 AI
San Francisco, CA, USA
{triley, p.kalantari, inaderi, koosh, kwei}@310.ai

Oleg Matusovsky
McGill University
Montreal, Quebec, Canada
oleg@310.ai

Mohammad S. Parsa
University of California, Berkeley
Berkeley, CA, USA
parsa@310.ai

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.

1 INTRODUCTION

Protein function is determined by the interplay between sequence and structure, making it essential when designing new proteins to account for both aspects. Traditional methods, such as Rosetta (Leaver-Fay et al. (2011)), employ empirical and physics-based approaches to link sequence with structure. More recently, deep learning based approaches, trained on extensive datasets, have demonstrated that large protein language models can learn sufficient information to accurately predict protein structures. Further advancements have shown that these deep-learning approaches can also capture some functional properties, such as protein-protein interactions and antibody complex structures (Abramson et al. (2024); Wohlwend et al. (2024)).

Most protein language models are trained on highly curated datasets and are designed to predict relatively narrow functions. For instance, some models can predict protein structures with atomic-level accuracy given a specific sequence Tunyasuvunakool et al. (2021). Others, like ProteinMPNN, focus on identifying sequences that will fold into a predefined backbone Dauparas et al. (2022). These models excel at tasks where the function is well defined, but they often require a large amount of a priori knowledge to generate meaningful results. While such approaches are highly effective for specific design goals, they limit the flexibility of these models in more generalist settings, where predicting novel protein functions or adapting to diverse design challenges is more complex. This restriction underscores the need for models that can handle broader design spaces, enabling de novo design of functional proteins across various applications.

Here, we present the molecular programming model version 4 (MP4), which utilizes broad and diverse datasets to generate protein sequences from minimal input. Trained on 138,000 tokens and 3.2 billion unique data points, MP4 incorporates a comprehensive range of protein-related information to learn the complex relationships between sequence, structure, and function. To evaluate the

models' capabilities, we randomly generated thousands of unique protein descriptions that specified various functional characteristics, such as binding partners, catalytic activity, and subcellular localization. These descriptions were used to design novel sequences that were evaluated for stable structural folds and functional matches. A subset of these de novo designed proteins was then explored experimentally, with the majority stably expressing and exhibiting favorable thermodynamic properties. Thus, MP4 not only generates novel protein sequences, but also optimizes key functional and structural features, making it a powerful tool for protein design.

2 RESULTS

2.1 OVERVIEW OF THE MP4 MODEL

MP4 is a transformer-based text-to-protein AI model designed to translate natural language prompts into de novo protein sequences that align with specified functions and properties. Unlike traditional methods that often follow a conventional pipeline - first defining a backbone structure and then generating sequences to match, MP4 utilizes a text-to-protein approach. This allows it to generate proteins directly from functional text prompts, making it more flexible and capable of addressing complex design objectives simultaneously.

MP4 is designed to tackle some of the primary challenges in protein science, particularly the programmability of proteins - creating proteins that can perform specific functions. One of the key innovations in the MP4 model is the integration of conditional language models, such as the conditional transformer language, which allows the model to generate sequences based on specific annotated functions or properties Keskar et al. (2019).

Each protein sequence generated by the MP4 model undergoes evaluation for amino acid composition, structural confidence, and functional similarity to ensure that the proteins are not only theoretically feasible but also practically functional. This method enables a joint sequence-function distribution, making it easier to tailor proteins for desired characteristics.

2.2 MP4 GENERATES PROTEIN SEQUENCES WITH HIGH PREDICTED FOLDABILITY AND FUNCTIONAL ACTIVITY

To evaluate MP4's ability to generate novel sequences from functional descriptions, we created over 1,000 prompts that specified diverse protein characteristics, including enzymatic activities, intracellular localizations, and binding partners. MP4 then generated diverse and unrelated sequences based on these prompts (Fig 1), which were subsequently analyzed to assess their plausibility and realism. This evaluation focused on key metrics such as amino acid composition, predicted foldability, and alignment with known biochemical principles, providing insights into the model's capacity to design biologically relevant proteins. The full repository can be explored at `https://310.ai/mp/repo`.

We began by examining the amino acid composition of the de novo sequences generated by MP4, comparing their distributions to verified sequences from the non-redundant protein (NR) databases

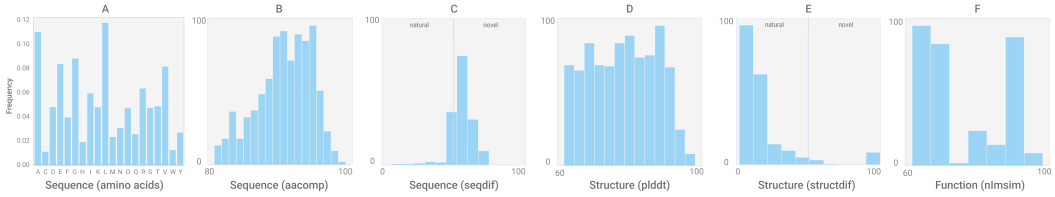


Figure 1: Computational metrics of 1000+ AI designed proteins generated by MP4 model. A) Frequency of 20 canonical amino acids. B) Amino acid composition per sequence, normalized to UniProt database proteins. C) Sequence comparison to NR/NT database proteins. D) Averaged ESMFold confidence pLDDT. E) Structure comparison to Protein Data Bank database proteins. F) Functional similarity based on prompt and predicted sequence function using ProtNLM model.

(Boratyn et al. (2013)). All amino acids were represented across the generated sequences, and their frequencies closely matched those observed in native UniProt sequences (Fig 1A,Bateman et al. (2024)). MP4 ensures the natural-like distribution of amino acids in the generated sequences, with amino acid composition (AAcomp) scores ranging from 80 to 100 (Fig1B). This metric identifies repetitive sequences, flagging potentially biologically implausible proteins.

A defining feature of the MP4 model is its ability to generate de novo protein sequences that significantly differ from natural sequences. Sequence novelty is assessed using the seqdif score, which quantifies how distinct a generated protein sequence is from known reference in the NR database. Seqdif scores range from 0 to 100, with higher values indicating greater novelty. According to the observed seqdif score distribution, the majority of the generated sequences cluster in the 50-60 score range, signifying sequences at least 50% different from any natural sequence (Fig1C). A smaller subset of proteins exhibits seqdif scores approaching 70-80 score, representing sequences that are highly divergent from natural proteins (Fig 1C), highlighting MP4's capacity to explore novel sequence space while maintaining a balance between sequence novelty and biological feasibility.

Next, we evaluated the structural stability of the generated sequences by predicting their folded structures using ESMFold (Lin et al. (2023)). For each sequence, we calculated the average predicted local distance difference test (pLDDT) as a measure of structural confidence (Mariani et al. (2013)). Similar to the amino acid distribution, most sequences were predicted to fold into stable protein structures (Fig 1D), with an average pLDDT of 82.6, indicating high local confidence in the predicted folds. Structural similarity was further evaluated using FoldSeek and the reported TM-score to compare the generated structures to those in the Protein Data Bank, reported as structdif (Fig 1E, van Kempen et al. (2023)). Despite their sequence novelty, most generated proteins adopted folds that are well-established in nature, consistent with the principle that structure is often tightly linked to function. These findings demonstrate that MP4 not only interprets intended functional descriptions but also designs novel sequences that adopt the necessary structural folds to perform those functions.

We evaluated how well the generated sequences aligned with their input prompts using ProtNLM, a UniProt-supported method that predicts protein functions from amino acid sequences (Gane et al. (2022)). Nlmsim, a ChatGPT-based similarity score, compares the input prompt with ProtNLM's output (OpenAI (2024)). Scores of 80–100 indicate exact or subset matches, while 60–80 suggests similar words, though synonyms or broader categories may score poorly. Many sequences showed keyword matches in ProtNLM outputs (Fig 1F), highlighting MP4's ability to translate functional descriptions into protein designs.

2.3 Proteins generated by MP4 have desirable experimental properties

To validate the experimental properties of the sequences generated by MP4, we characterized a subset of these designs to assess whether they possessed favorable traits beyond computational predictions. Specifically, we cloned a representative subset of 94 sequences, emphasizing those with stable predicted structural and diverse functional properties. This selected subset maintained sequence diversity (Fig 2A), highlighting that MP4 is not converging onto a single solution, nor replicating natural proteins. Each protein was expressed in a prokaryotic cell-free system using a split-GFP tag, and relative protein levels were quantified through a split-GFP assay (Bignon et al. (2022)). Notably, a significant proportion of the cloned sequences successfully translated into measurable protein yields, with 79 out of 94 sequences (84%) yielding detectable protein levels (Fig 2B). Full results can be explored at `https://310.ai/mp/lab/1`.

Thermostability, a key property of rationally designed proteins, was also assessed. This characteristic is defined by a protein's ability to maintain structural integrity under increasing temperatures (Vihinen (1987)). Material from each expression construct was subjected to differential scanning fluorimetry (DSF) to determine the melting temperature (Tm), representing the temperature at which 50% of the protein remains folded (Hellman et al. (2016)). However, due to small expression volumes and low tryptophan content for fluorescent detection (Wen et al. (2020)), reliable signals were obtained from only 17 protein samples (Table 1). Despite this limitation (which could be overcome by prioritizing buried tryptophans during design), the average thermostability measurement exceeding 62°C, with the most stable proteins approaching 90°C (Fig 2C). In addition to characterizing these 17 by DSF, we selected an additional 10 samples to quantify by dynamic light scattering

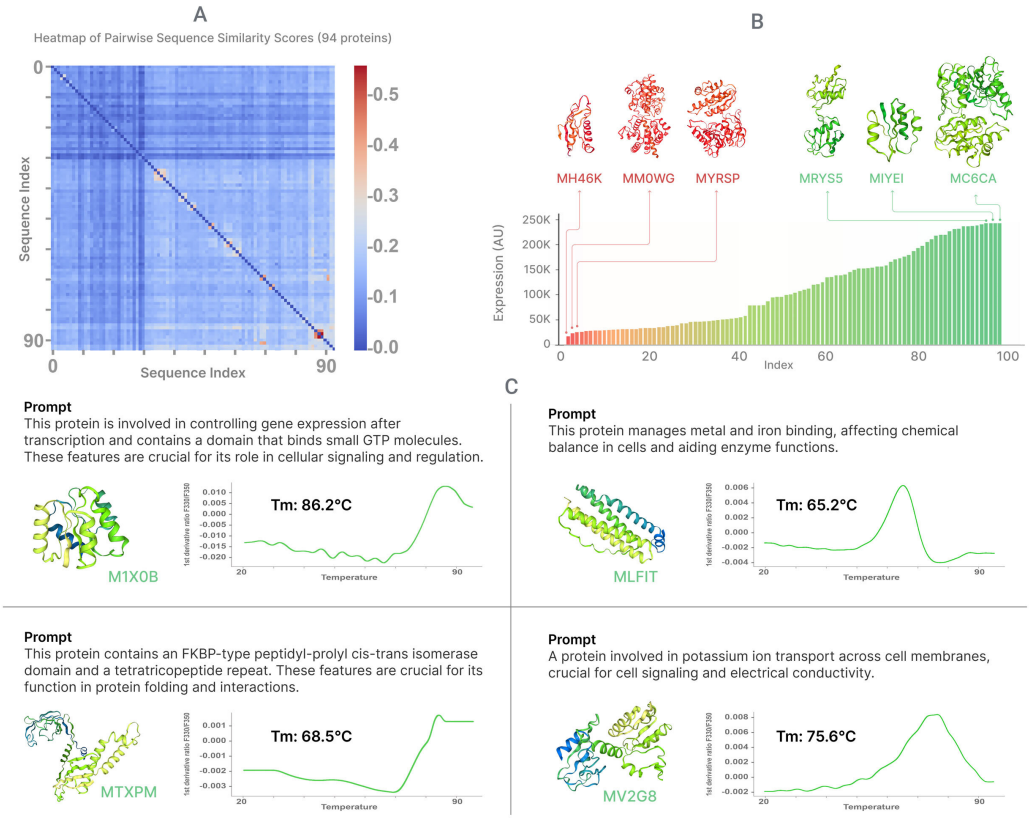


Figure 2: Experimental evaluation of 94 selected de novo designed proteins. A) Pairwise sequence similarity heatmap. B) Expression profile in a cell free expression system. C) Thermostability, measured by DSF, of 4 diverse proteins.

(DLS) (Stetefeld et al. (2016)). These samples, although providing no measurable signal by DSF, resulted in a uniform peak by DLS implying that stable protein characteristics were incorporated throughout the design panel.

These findings, although limited to a representative subset of the MP4-designed proteins, suggest that the MP4 model not only generates sequences with intended functional properties but also accounts for additional attributes such as expression efficiency and thermostability. The results highlight the model's capacity to design proteins with a high likelihood of successful experimental translation and robust structural properties.

2.4 Property interrogation of MP4 designed proteins

We next assessed how well commonly used computational metrics predict protein behavior and expression levels.

First, we examined the relationship between predicted secondary structure composition and expression levels. While no strong correlation was observed overall, MP4-designed proteins exhibited a broad range of alpha-helical content (20–90%, Fig 3A). Notably, well-expressing designs were found across this spectrum, including some with minimal alpha-helical content (traditionally considered difficult for computational design) and others composed almost entirely of alpha helices. However, the two designs with the highest predicted alpha-helical content failed to express, likely due to prediction biases from ESMFold. These findings indicate that MP4 does not impose a strong preference for specific protein folds and is capable of generating diverse, viable scaffolds.

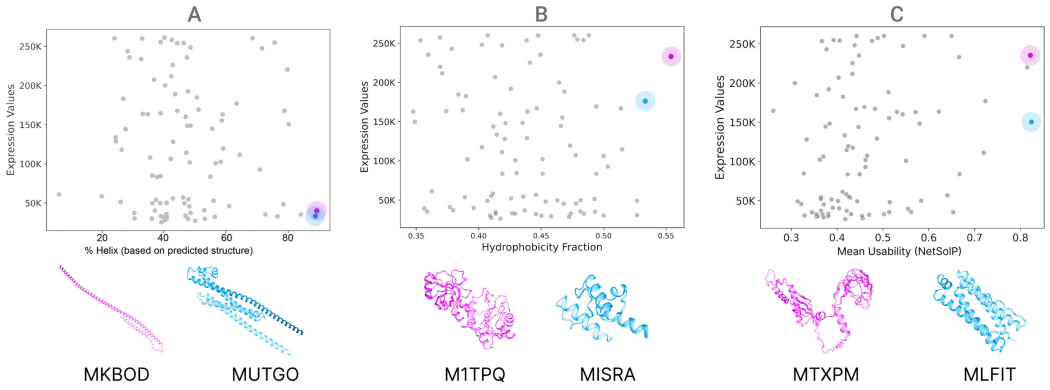


Figure 3: Structural and property analysis of designed proteins. A) Alpha-helical content vs relative expression levels. (B) Predicted hydrophobicity vs relative expression levels. (C) Predicted developability (usability score by NetSolP) vs relative expression levels.

Hydrophobicity is another commonly used metric for ranking and evaluating protein designs, as it is often linked to increased aggregation, which can negatively impact both expression levels and thermostability. A hydrophobicity prediction model (Malleshappa Gowder et al. (2014)) indicated that most MP4-designed proteins exhibited minimal hydrophobic content. However, only weak correlation was observed between predicted hydrophobicity and measured expression levels (Fig 3B), reinforcing the notion that while hydrophobicity plays a role in protein behavior, it is not the sole determinant of successful folding and expression.

Given the multifaceted nature of protein developability, we next evaluated a composite 'developability' predictor that integrates hydrophobicity, charge, and solubility into a weighted usability score, NetSolP (Thumuluri et al. (2021)) Unlike hydrophobicity alone, MP4-designed proteins spanned a broad range of predicted usability, indicating that some sequences may lack optimal characteristics for experimental expression. Despite this variation, the usability score showed only a modest improvement over hydrophobicity in correlating with expression levels (Fig 3C), suggesting that even multi-parameter predictors struggle to fully capture the complexity of factors influencing experimental protein expression.

3 Discussion

One of the key strengths of MP4 is its ability to generate protein sequences that can be translated into experimentally validated molecules. By optimizing multiple properties simultaneously, MP4 designs proteins that are structurally robust and stable under experimental conditions. This integrated approach highlights the potential of MP4 as a powerful tool to advance protein engineering and overcome practical challenges in the de novo protein design.

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.

The thermostability of the proteins designed by MP4 further underscores its utility for applications requiring robust protein performance under extreme conditions. Although only 17 proteins yielded reliable thermal melting curves due to a combination of low tryptophan content and technical constraints, the average melting temperature was 62°C, with the most stable protein nearing 90°C. These findings suggest that MP4 inherently considers stability as part of its design process. This is partic-

ularly significant for industrial and therapeutic applications, where proteins must remain functional under harsh environmental conditions.

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-grained control. The model will then generate optimized protein sequences in a single inference step, transforming biological design into a deterministic, programmable process.

4 REFERENCES

ACKNOWLEDGMENTS

Experimental lab work was done at Adaptyv, the cloud lab for proteins.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, May 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07487-w. URL http://dx.doi.org/10.1038/s41586-024-07487-w.

Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, Aduragbemi Adesina, Shadab Ahmad, Emily H Bowler-Barnett, Hema Bye-A-Jee, David Carpentier, Paul Denny, Jun Fan, Penelope Garmiri, Leonardo Jose da Costa Gonzales, Abdulrahman Hussein, Alexandr Ignatchenko, Giuseppe Insana, Rizwan Ishtiaq, Vishal Joshi, Dushyanth Jyothi, Swaathi Kandasaamy, Antonia Lock, Aurelien Luciani, Jie Luo, Yvonne Lussi, Juan Sebastian Martinez Marin, Pedro Raposo, Daniel L Rice, Rafael Santos, Elena Speretta, James Stephenson, Prabhat Totoo, Nidhi Tyagi, Nadya Urakova, Preethi Vasudev, Kate Warner, Supun Wijerathne, Conny Wing-Heng Yu, Rossana Zaru, Alan J Bridge, Lucila Aimo, Ghislaine Argoud-Puy, Andrea H Auchincloss, Kristian B Axelsen, Parit Bansal, Delphine Baratin, Teresa M Batista Neto, Marie-Claude Blatter, Jerven T Bolleman, Emmanuel Boutet, Lionel Breuza, Blanca Cabrera Gil, Cristina Casals-Casas, Kamal Chikh Echioukh, Elisabeth Coudert, Beatrice Cuche, Edouard de Castro, Anne Estreicher, Maria L Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Pascale Gaudet, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Arnaud Kerhornou, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Lucille Pourcel, Sylvain Poux, Monica Pozzato, Manuela Pruess, Nicole Redaschi, Catherine Rivoire, Christian J A Sigrist, Karin Sonesson, Shyamala Sundaram, Anastasia Sveshnikova, Cathy H Wu, Cecilia N Arighi, Chuming Chen, Yongxing Chen, Hongzhan Huang, Kati Laiho, Minna Lehvaslaiho, Peter McGarvey, Darren A Natale, Karen Ross, C R Vinayaka, Yuqi Wang, and Jian Zhang. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, November 2024. ISSN 1362-4962. doi: 10.1093/nar/gkae1010. URL http://dx.doi.org/10.1093/nar/gkae1010.

Christophe Bignon, Antoine Gruet, and Sonia Longhi. Split-gfp reassembly assay: Strengths and caveats from a multiparametric analysis. *International Journal of Molecular Sciences*, 23(21):

13167, October 2022. ISSN 1422-0067. doi: 10.3390/ijms232113167. URL http://dx.doi.org/10.3390/ijms232113167.

Grzegorz M. Boratyn, Christiam Camacho, Peter S. Cooper, George Coulouris, Amelia Fong, Ning Ma, Thomas L. Madden, Wayne T. Matten, Scott D. McGinnis, Yuri Merezhuk, Yan Raytselis, Eric W. Sayers, Tao Tao, Jian Ye, and Irena Zaretskaya. Blast: a more efficient report with usability improvements. *Nucleic Acids Research*, 41(W1):W29–W33, April 2013. ISSN 0305-1048. doi: 10.1093/nar/gkt282. URL http://dx.doi.org/10.1093/nar/gkt282.

Stéphanie Cabantous and Geoffrey S Waldo. In vivo and in vitro protein solubility assays using split gfp. *Nature Methods*, 3(10):845–854, September 2006. ISSN 1548-7105. doi: 10.1038/nmeth932. URL http://dx.doi.org/10.1038/nmeth932.

J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky, A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan, B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, October 2022. ISSN 1095-9203. doi: 10.1126/science.add2187. URL http://dx.doi.org/10.1126/science.add2187.

Andreea Gane, Maxwell L. Bileschi, David Dohan, Elena Speretta, Amélie Héliou, Laetitia Meng-Papaxanthos, Hermann Zellner, Eugene Brevdo, Ankur Parikh, Maria J. Martin, Sandra Orchard, UniProt Collaborators, and Lucy J. Colwell. Protnlm: Model-based natural language protein annotation. *Preprint*, 2022. Retrieved from https://storage.googleapis.com/brain-genomics-public/research/proteins/protnlm/uniprot_2022_04/protnlm_preprint_draft.pdf.

Lance M. Hellman, Liusong Yin, Yuan Wang, Sydney J. Blevins, Timothy P. Riley, Orrin S. Belden, Timothy T. Spear, Michael I. Nishimura, Lawrence J. Stern, and Brian M. Baker. Differential scanning fluorimetry based assessments of the thermal and kinetic stability of peptide–mhc complexes. *Journal of Immunological Methods*, 432:95–101, May 2016. ISSN 0022-1759. doi: 10.1016/j.jim.2016.02.016. URL http://dx.doi.org/10.1016/j.jim.2016.02.016.

Nitish Shirish Keskar, Bryan McCann, Lav Varshney, Caiming Xiong, and Richard Socher. CTRL - A Conditional Transformer Language Model for Controllable Generation. *arXiv preprint arXiv:1909.05858*, 2019.

Andrew Leaver-Fay, Michael Tyka, Steven M Lewis, Oliver F Lange, James Thompson, Ron Jacak, Kristian Kaufman, P Douglas Renfrew, Colin A Smith, Will Sheffler, Ian W Davis, Seth Cooper, Adrien Treuille, Daniel J Mandell, Florian Richter, Yih-En Andrew Ban, Sarel J Fleishman, Jacob E Corn, David E Kim, Sergey Lyskov, Monica Berrondo, Stuart Mentzer, Zoran Popović, James J Havranek, John Karanicolas, Rhiju Das, Jens Meiler, Tanja Kortemme, Jeffrey J Gray, Brian Kuhlman, David Baker, and Philip Bradley. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.*, 487:545–574, 2011.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. ISSN 1095-9203. doi: 10.1126/science.ade2574. URL http://dx.doi.org/10.1126/science.ade2574.

Shambhu Malleshappa Gowder, Jhinuk Chatterjee, Tanusree Chaudhuri, and Kusum Paul. Prediction and analysis of surface hydrophobic residues in tertiary structure of proteins. *The Scientific World Journal*, 2014:1–7, 2014. ISSN 1537-744X. doi: 10.1155/2014/971258. URL http://dx.doi.org/10.1155/2014/971258.

Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, August 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt473. URL http://dx.doi.org/10.1093/bioinformatics/btt473.

OpenAI. Chatgpt: A large language model, 2024. Available at https://openai.com/chatgpt.

Jörg Stetefeld, Sean A. McKenna, and Trushar R. Patel. Dynamic light scattering: a practical guide and applications in biomedical sciences. *Biophysical Reviews*, 8(4):409–427, October 2016. ISSN 1867-2469. doi: 10.1007/s12551-016-0218-6. URL http://dx.doi.org/10.1007/s12551-016-0218-6.

Vineet Thumuluri, Hannah-Marie Martiny, Jose J Almagro Armenteros, Jesper Salomon, Henrik Nielsen, and Alexander Rosenberg Johansen. Netsolp: predicting protein solubility in escherichia coli using language models. *Bioinformatics*, 38(4):941–946, November 2021. ISSN 1367-4811. doi: 10.1093/bioinformatics/btab801. URL http://dx.doi.org/10.1093/bioinformatics/btab801.

Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, Sameer Velankar, Gerard J. Kleywegt, Alex Bateman, Richard Evans, Alexander Pritzel, Michael Figurnov, Olaf Ronneberger, Russ Bates, Simon A. A. Kohl, Anna Potapenko, Andrew J. Ballard, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Ellen Clancy, David Reiman, Stig Petersen, Andrew W. Senior, Koray Kavukcuoglu, Ewan Birney, Pushmeet Kohli, John Jumper, and Demis Hassabis. Highly accurate protein structure prediction for the human proteome. *Nature*, 596(7873):590–596, July 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03828-1. URL http://dx.doi.org/10.1038/s41586-021-03828-1.

Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron L. M. Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, 42(2):243–246, May 2023. ISSN 1546-1696. doi: 10.1038/s41587-023-01773-0. URL http://dx.doi.org/10.1038/s41587-023-01773-0.

Mauno Vihinen. Relationship of protein flexibility to thermostability. *"Protein Engineering, Design and Selection"*, 1(6):477–480, 1987. ISSN 1741-0134. doi: 10.1093/protein/1.6.477. URL http://dx.doi.org/10.1093/protein/1.6.477.

Jie Wen, Harrison Lord, Nicholas Knutson, and Mats Wikström. Nano differential scanning fluorimetry for comparability studies of therapeutic proteins. *Analytical Biochemistry*, 593:113581, March 2020. ISSN 0003-2697. doi: 10.1016/j.ab.2020.113581. URL http://dx.doi.org/10.1016/j.ab.2020.113581.

Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Tally Portnoi, Itamar Chinn, Jacob Silterra, Tommi Jaakkola, and Regina Barzilay. Boltz-1: Democratizing biomolecular interaction modeling. November 2024. doi: 10.1101/2024.11.19.624167. URL http://dx.doi.org/10.1101/2024.11.19.624167.

A Appendix A: MP4 Architecture

This appendix provides a high-level overview of the MP4 model architecture and its training regime. Due to proprietary considerations, specific implementation details—including internal layer configurations, precise hyperparameters, and optimization strategies—are not disclosed. The following sections outline the key design choices that enable MP4 to translate natural language prompts into functional, de novo protein sequences.

A.1 High-Level Overview

MP4 is a transformer-based model specifically designed for de novo protein design. It accepts natural language prompts that encode comprehensive protein information—such as fitness criteria, physical properties, source organism, and sequence-related properties—and generates protein sequences that meet the desired functional and structural constraints. The model emphasizes molecule programmability by integrating data from diverse sources and synchronizing multiple design objectives during training. This capability enables MP4 to deliver state-of-the-art performance in generating novel proteins that are both experimentally feasible and functionally robust.

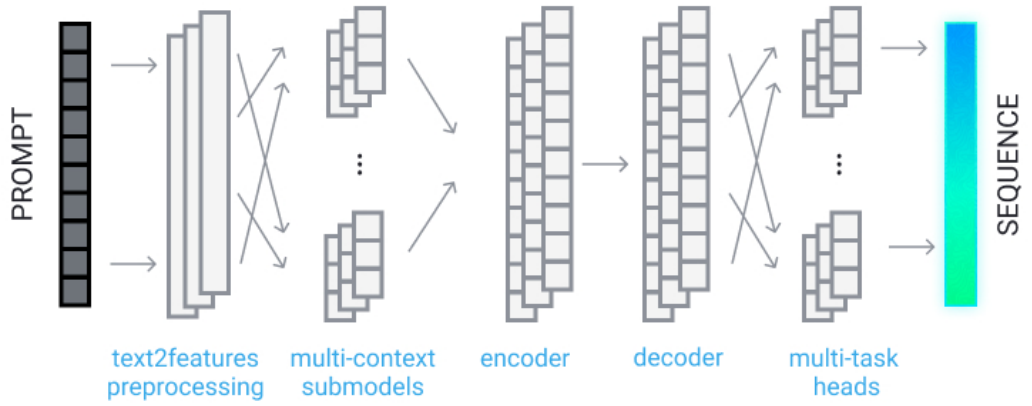


Figure A.1: An overview of the MP4 architecture.

A.2 Input and Output Representations

Input Representation: The input to MP4 is a detailed textual prompt. These prompts include various descriptors such as:

- **Fitness:** Desired functional and performance metrics.
- **Organism:** Source organism or related biological context.
- **Sequence Properties:** Attributes that might include partial sequences, motifs, or structural hints.

Before being processed by the core model, these textual inputs pass through a *text2feature preprocessing unit*, which tokenizes the prompt and converts it into a structured feature representation. This conversion ensures that all pertinent information is captured and made accessible to subsequent layers.

Output Representation: The final output of MP4 is a protein sequence composed of amino acids. The sequence is generated in a way that reflects the functional and structural requirements encoded in the input prompt. By balancing multiple design tasks, MP4 ensures that the generated sequence is consistent with both the raw input and the underlying biochemical principles.

A.3 Transformer Backbone and Conditional Generation

At the core of MP4 lies a multi-layer transformer architecture that has been adapted for the complexities of protein design. The architecture comprises several key components:

- **Multi-Context Sub Models:** These sub-models are designed to process different aspects of the input features. Each sub-model focuses on distinct information—derived from the prompt—ensuring that the model can attend to diverse functional, structural, and contextual cues.
- **Encoder:** After the multi-context sub models process the input, an encoder aggregates the information. It captures long-range dependencies and builds a rich, context-aware representation that summarizes the key aspects of the protein prompt.
- **Decoder:** The decoder translates the encoder's latent representation into a form that is directly amenable to sequence generation. In this phase, the model transforms abstract feature representations into sequential data.
- **Multi-Task Head:** The output from the decoder is routed through a set of task-specific heads. MP4 is trained to perform 70 synchronized tasks, where each head is responsible for interpreting the decoder vector with respect to a particular design objective. These

tasks include aspects such as structural fold determination, functional site prediction, and sequence novelty assessment. The collaborative output from these heads is then integrated to construct the final protein sequence.

Figure A.1 schematically illustrates the end-to-end architecture of MP4, highlighting the sequential processing from prompt to final protein sequence.

A.4 Training and Inference Overview

Training Data and Regime: MP4 has been trained on a highly diverse and extensive dataset, comprising over 1.8 billion datapoints which were collected from various repositories, including UniProt. The training process involved processing 138K tokens and was carried out across 70 synchronized tasks, with each task emphasizing a distinct aspect of protein design—from structural features to functional specificity. Overall, the training was carried out with approximately 3,800 AMD-Instinct GPU-days.

A.5 Proprietary Considerations and Future Directions

While this appendix outlines the overarching design and training strategy of MP4, many technical details remain confidential. In particular, specific modifications to the standard transformer framework, internal layer configurations, and fine-tuning strategies are proprietary. Future research will focus on:

- Refining the multi-context sub models to enhance the model's sensitivity to nuanced protein features.
- Expanding the range of synchronized tasks to capture an even broader spectrum of protein functionalities.
- Exploring alternative decoding strategies to further improve sequence fidelity and novelty.
- Integrating all-atom protein structure generation to enable the direct production of detailed, three-dimensional protein models.

This continued evolution aims to push the boundaries of molecule programmability in protein design.

B Appendix B: Experimental Methods

B.1 Construct Design

Constructs were designed with C-terminal tags (GFP11 for split-GFP solubility and Twin-Strep for purification) and sourced from Twist Bioscience. The DNA constructs were subsequently assembled into the appropriate expression vector using the NEBuilder HiFi DNA Assembly Kit (New England Biolabs) according to the manufacturer's protocol, with assembly reactions performed in 3µL volumes.

B.2 Protein Expression and Purification

Protein expression was carried out in a prokaryotic cell-free system at Adaptyv Bio. Briefly, expression reactions were prepared in a total volume of 60µL and incubated at 37°C for 12 hours to ensure robust protein synthesis. Expression was detected by the flourescent coexpression of the GFP11 tag and GFP1-10 marker as determined by a split-GFP solubility assay (Cabantous & Waldo (2006)). For protein purification, an affinity capture approach was employed using magnetic beads. Protein samples were mixed with the beads and incubated for 10 minutes at room temperature with gentle agitation to allow binding. Following binding, the beads were washed three times with a washing buffer composed of 1M Tris-Cl, 1.5M NaCl, and 10mM EDTA (pH 8) to remove unbound proteins. The concentration and yield of the purified proteins were quantified using an affinity-based assay, with normalization performed via a Qubit fluorometer assay.

B.3 BIOPHYSICAL EVALUATION

Protein thermostability was assessed using nano differential scanning fluorimetry (NanoDSF). Protein samples were diluted to a concentration of 100µg/mL in assay buffer (20mM sodium phosphate, 150mM NaCl, pH 7.0), and 10µL aliquots were pipetted into NanoDSF capillaries. The assay was performed with a temperature ramp of 1°C per minute while monitoring intrinsic fluorescence, specifically by tracking the ratio of fluorescence intensities at 350nm and 330nm. Fluorescence changes were recorded continuously during the temperature increase, and the melting temperature (T_m) was determined as the inflection point on the fluorescence change curve. T_m values obtained for different constructs were compared to evaluate relative thermostability. Dynamic light scattering (DLS) was used to evaluate the hydrodynamic radius and aggregation state of the expressed proteins. Prior to measurement, purified protein samples were allowed to equilibrate to room temperature. Approximately 20µL of each sample was loaded into a disposable cuvette, and measurements were performed at 25°C using the Unchained Labs Uncle instrument. A series of at least 10 runs per sample was acquired to ensure statistical reliability. The size distribution data were analyzed to determine the hydrodynamic radius and to assess the presence of protein aggregates.

C APPENDIX C: SUPPLEMENTAL DATA

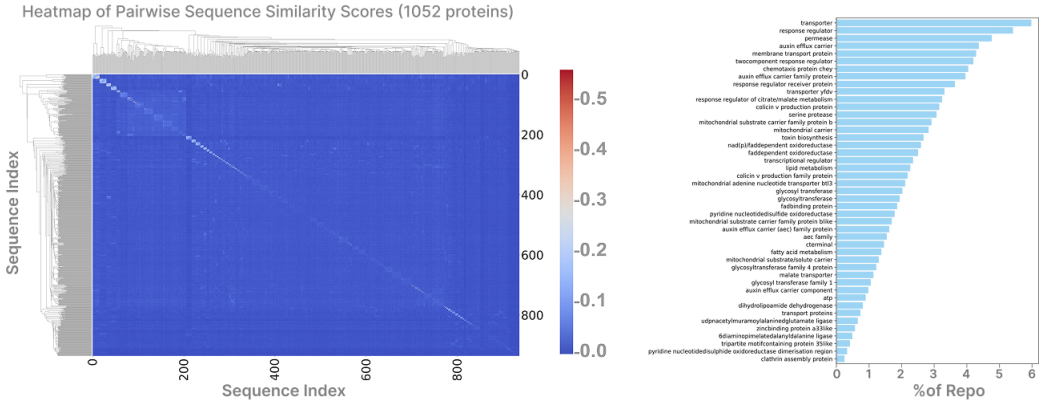
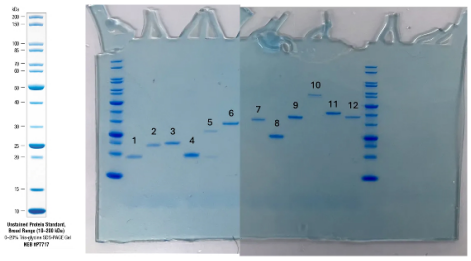


Figure A.2: Diversity of 1052 generated sequences



Lane	ID	Molecular Weight (kDa)	Expression (SplitGFP) [RFU]	Concentration [µg/ml]
1	MIYEI	7.88	7008	618
2	MYTDL	10.07	6634	444
3	M1X0B	13.37	7372	604
4	MISRA	9.04	8270	778
5	MRYS5	16.76	7462	566
6	MTSN0	22.71	6332	534
7	M76KJ	25.71	4321	446
8	MPCII	15.64	7101	676
9	MJ73M	29.73	5600	566
10	MCLNW	27.09	2503	382
11	MCMOZ	33.44	5637	580
12	MWG7X	30.36	6195	524

Figure A.3: SDS-PAGE gel characterization of select proteins.

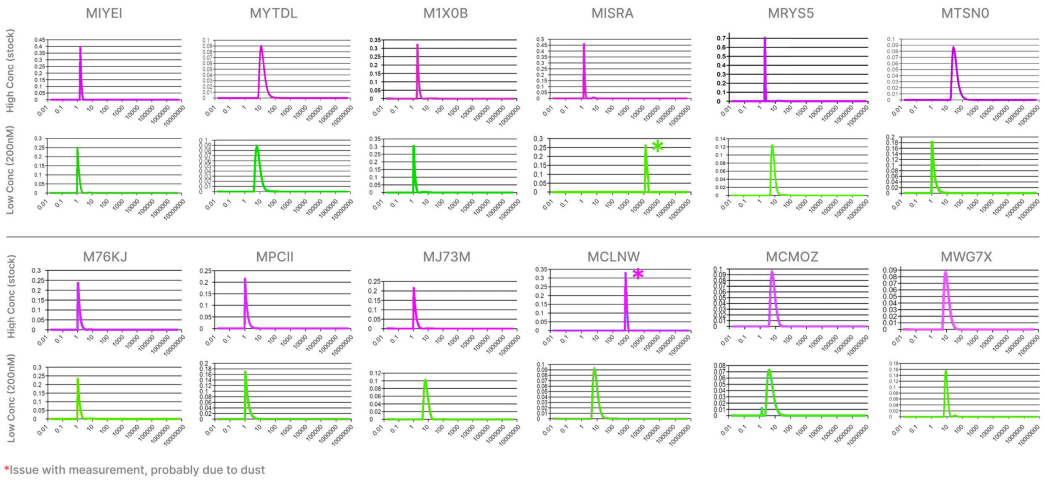


Figure A.4: Dynamic light scattering (DLS) characterization of select proteins.

Table 1: Measured thermostability

Repo ID	Prompt (snippet)	Length	Tm (°C)
M1X0B	...controlling gene expression after transcription...	119	86.2
MQLYM	...glycerol metabolism...in sperm cells	207	83.36
MJFIU	...kinase that binds ATP...	202	81.94
MW51C	...detection of external stimulus and a PAS fold.	490	79.89
M54FQ	...Periplasmic binding protein/LacI sugar binding...	296	76.8
MV2G8	...potassium ion transport across cell membranes...	214	75.64
M7PN6	...binds ferric iron...breakdown of carboxylic acids.	159	71.05
MTXPM	...FKBP-type peptidyl-prolyl cis-trans isomerase...	260	68.51
MT47E	...binding cations and DNA, regulates toxin...	127	67.05
MLFIT	...metal and iron binding...	157	65.18
MWG7X	...ACT-like domain...aromatic amino acid family...	277	62.22
MPCII	...hemerythrin-like and cation homeostasis...	133	55.97
MR1UH	...Aconitase A/isopropylmalate dehydratase...	234	47.97
M5CZB	...dephosphorylation and regulates metabolic processes...	252	45.43
MMUJG	...localization and transport within cells...	245	41.2
MJY78	DNA polymerase lambda lyase...	258	36.48
M7S72	...glycosyltransferase activity...	296	31.32