

# TOWARDS PROTEIN SEQUENCE & STRUCTURE CO-DESIGN WITH MULTI-MODAL LANGUAGE MODELS

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Proteins perform diverse biological functions, governed by the intricate relationship between their sequence and three-dimensional structure. While protein language models (PLMs) have demonstrated remarkable success in functional annotation and structure prediction, their potential for sequence-structure co-design remains underexplored. This limitation arises from pre-training objectives that favor masked token prediction over generative modeling. In this work, we systematically explore sampling strategies to enhance the generative capabilities of PLMs for co-design. Notably, we introduce a ranked iterative decoding with re-masking scheme, enabling PLMs to generate sequences and structures more effectively. Benchmarking ESM3 across multiple scales, we demonstrate that using PLMs effectively at sampling time for co-design tasks can outperform specialized architectures that lack comparable scaling properties. Our work advances the field of computational protein design by equipping PLMs with robust generative capabilities tailored to sequence-structure interdependence.

## 1 INTRODUCTION

Proteins are essential biomolecules that perform a wide array of functions in living organisms, driven by the intricate relationship between their sequences and three-dimensional structures. Advancements in computational protein design have significantly benefited from the development of protein language models (PLMs) (Rives et al., 2021; Lin et al., 2023b; Hayes et al., 2024; Ferruz et al., 2022; Wang et al., 2024a), which leverage the vast amount of available sequence data to learn meaningful representations. Such representations have shown promise in tasks ranging from functional annotation (Hu et al., 2022; Zhang et al., 2024) to structure prediction (Lin et al., 2023b), emphasizing the utility of PLMs in understanding the sequence-structure-function paradigm.

Recently, multimodal PLMs like SaProt (Su et al., 2023) and ESM3 (Lin et al., 2023b) have further extended this success by modeling both sequence and structure through the tokenization of protein backbones into discrete tokens (van den Oord et al., 2018). This multimodal approach offers new opportunities for capturing the sequence-structure interplay. However, most PLMs, including ESM3, are trained using masked language modeling (MLM) objectives (Devlin, 2018; Liu, 2019), which are designed for representation learning and not optimized for generative tasks. As a result, their potential for generating novel, biologically meaningful proteins remains under-explored.

In this work, we explore the generative potential of ESM3 on protein co-design, a task that demands the simultaneous generation of sequences and their corresponding structural representations. To address the challenges of jointly sampling both modalities, we introduce ranked iterative decoding with re-masking as an effective co-design sampling strategy. We benchmark ESM3’s performance against bespoke methods specifically designed for co-design and analyze how model size impacts key metrics. Our findings highlight that ESM3, despite being a general-purpose PLM, can deliver competitive results while remaining scalable and adaptable for large-scale protein design applications.

## 054 2 METHODS

### 055 2.1 BACKGROUND

056 **Notation.** A protein is represented by its amino acid sequence  $\mathcal{X} = (x_1; \dots; x_L) \in \mathcal{X}^L$ , where  $L$  is  
 057 the number of residues and  $\mathcal{X}$  denotes the set of 20 standard amino acids. The backbone structure  
 058 of the protein is given by  $\mathcal{Y} \in \mathbb{R}^{L \times 3}$ , encompassing all heavy atoms along the backbone. Using  
 059 a pre-trained structure encoder  $q(Z | \mathcal{Y})$ , the structure  $\mathcal{Y}$  is transformed into a sequence of latent  
 060 structure tokens  $Z = (z_1; \dots; z_L) \in \mathcal{Z}^L$ , where  $\mathcal{Z}$  is a predefined vocabulary of latent codes. These  
 061 structure tokens  $Z_i$  are then mapped into embedding vectors and decoded back to the 3D structure  
 062  $\mathcal{Y}$ . Throughout the following sections, we use the discrete random variables  $X$  and  $Z$  to denote the  
 063 sequence tokens and structure tokens with respective probability mass functions  $p(X)$  and  $p(Z)$ .  
 064

065 **Model.** Let  $m : \mathcal{X}^L \times \mathcal{Z}^L \rightarrow \mathbb{R}^L \times \mathbb{R}^L$  denote a *masked language model* (MLM) that  
 066 takes as input a tokenized protein sequence and structure  $(X; Z)$  of length  $L$ , possibly with some  
 067 positions replaced with [MASK], and outputs two matrices  $\hat{X}; \hat{Z}$  with shapes  $L \times 20$  and  $L \times |\mathcal{Z}|$   
 068 containing log probabilities over the set of amino acid and structure tokens in  $\mathcal{X}$  and  $\mathcal{Z}$  respectively.  
 069 The probability distribution over the residues and structure tokens at position  $i$  with temperature  $\beta$  is

$$070 \quad \beta(X^i | X; Z) = \text{softmax}(\hat{X}^i / \beta) \quad \& \quad \beta(Z^i | X; Z) = \text{softmax}(\hat{Z}^i / \beta) \quad (1)$$

### 071 2.2 CO-DESIGN VIA SAMPLING

072 **Concept.** In co-design, our goal is to sample from the joint sequence and structure distribution  
 073  $p(x; y)$ . Many existing methods simplify this by factorizing the joint distribution into  $p(y)p(x|y)$   
 074 or  $p(x)p(y|x)$ , which imposes a dependency order and ignores the full bidirectional relationship  
 075 between modalities. Instead, we frame  $p(x; y)$  as an undirected model, specifically a Markov random  
 076 field (MRF) (Wang & Cho, 2019), and use MCMC sampling methods to better approximate joint  
 077 samples from the target distribution.  
 078

#### 079 2.2.1 CHOICE OF SAMPLING METHOD

080 Our sampling algorithm proceeds as follows. For a protein of length  $L$ , we start with fully masked  
 081 amino acid and structure tokens  $(X; Z)^i = [\text{MASK}] \forall i \in [L]$ . For each step  $t$ , we pass the current  
 082 sample  $(X; Z)_t$  through the MLM  $m$  (potentially after masking), generating the posterior distribution  
 083 in eq. (1). From here, we selectively sample new residue or structure tokens to fill masked positions  
 084 and obtain  $(X; Z)_{t+1}$ . This process is repeated until all  $L$  positions are unmasked.  
 085

086 **Gibbs-like sampling.** At each iteration  $t$ , Gibbs-like sampling selects a masked position  $i$  and  
 087 unmask it by sampling from eq. (1). We follow the methodology introduced by the authors of the  
 088 ESM3 paper by performing Gibbs-like sampling for one modality at a time, starting with the sequence  
 089 followed by the structure. We use a constant temperature of 0.7 for both tracks.  
 090

091 **Chain-of-thought sampling.** Chain-of-thought sampling as introduced by the ESM3 authors is an  
 092 extension to Gibbs-like sampling that first unmask the secondary structure tokens (SS8), followed by  
 093 the structure track, and finally the amino acid tokens. The authors found that this ordering provided  
 094 higher quality unconditional samples than Gibbs-like sampling with sequence and structure alone.  
 095

#### 096 2.2.2 RANKED ITERATIVE DECODING W/ RE-MASKING

097 Although *Gibbs-like sampling* and *Chain-of-thought sampling* frame each modality as a Markov  
 098 random field, it remains that both methods impose an arbitrary ordering by fully unmasking one track  
 099 before moving onto the next. Instead, we propose a ranked iterative decoding algorithm that unmask  
 100 all modalities simultaneously, treating the full joint distribution as an MRF.  
 101

102 Formally, at each iteration  $t$ , iterative decoding selects a masked position *for each modality* to unmask.  
 103 We found that prioritizing locations based on the position-specific entropy yielded higher quality  
 104 samples. We generalize this to accommodate various ranking functions  $f(\cdot)$  applied over the masked  
 105 positions. The overall inference pipeline is detailed in algorithm 3.  
 106

107 **Entropy ranking.** We use the entropy of the posterior token distribution at unmasked locations as  
 a proxy of model uncertainty. Formally, the ranking score of the masked position (for structure) at

index  $i$  at time  $t$  is given by

$$f(i) = \sum_{z \in Z} p(Z_1^i = z | X_t; Z_t) \log p(Z_1^i = z | X_t; Z_t)$$

The same applies to other modalities such as sequence and secondary structure by swapping the conditioned variable. The positions in the sequence are then ranked and decoded according to their position-specific entropy in ascending order. Importantly, we perform a single inference to compute the posterior distributions for all modalities (e.g., structure, sequence, and secondary structure) in parallel. This allows decoding to be conducted simultaneously for each modality.

**Max-logit ranking.** Here, we rank positions according to the value of the maximum logit in the pre-softmax output of the MLM. We prioritize positions with the highest top logit score, interpreting this value as the model's confidence at a particular location in the sequence. Formally, the ranking function for max-logit at index  $i$  is defined as:

$$f(i) = \max_{z \in Z} p(Z_1^i = z | X_t; Z_t)$$

**Secondary structure ranking.** We consider a discrete ranking score that uses the secondary structure prediction head of ESM3. Given a user-defined function  $g: \{1, \dots, 8\} \rightarrow \mathbb{R}$  that maps the eight secondary structures (SS8) in the Dictionary of Protein Secondary Structure (DSSP) to an ordering score, we use the following ranking function to prioritize positions with preferred SS8 tokens.

$$f(i) = \begin{cases} g(s_t^i) & s_t^i \in [\text{MASK}] \\ 1 & s_t^i = [\text{MASK}] \end{cases}$$

where  $s_t^i$  is the identity of the structure token at position  $i$  and time step  $t$ .

**Ranked re-masking.** We introduce a ranked re-masking strategy inspired from corrector sampling (Gat et al., 2024) that re-masks locations that have already been decoded according to a ranking function as described above. Intuitively, we achieve this by performing  $t$  unmasking steps followed by re-masking steps where  $t \geq 1$ . In practice, we experimented with the following ranking functions: (1) the minimum per-location entropy and (2) the minimum logit per location. We provide experimental results for both choices in appendix D.2.

## 3 EXPERIMENTS

### 3.1 EXPERIMENTAL SETUP

**Metrics.** We follow the protocol proposed by previous protein co-design works (Wang et al., 2024b; Lu et al., 2024a; Campbell et al., 2024) and evaluated designability, quality, diversity, and novelty of generated proteins. We provide additional details for these metrics in appendix C.1.

**Baselines.** We consider multiple open-source protein generation models as evaluation baselines for the co-design task. These include models that first generate one modality, then predict the other using a separate model. ProteinGenerator (Lisanza et al., 2023) performs sequence space diffusion while predicting the structure at each step with RosettaFold (Baek et al., 2021). ConVerter (Chu et al., 2023) performs Euclidean diffusion on the structure while iteratively predicting the sequence with ProteinMPNN. More recently, we consider models that jointly sample both modalities: MultiFlow (Campbell et al., 2024) and LAID (Lu et al., 2024a).

### 3.2 RANKED RE-MASKING IMPROVES DESIGNABILITY

We begin by studying the effect of sampling strategy on the designability of generated proteins. Using ESM3-small, we sampled 100 proteins of each length 50, 100, 200, 500 using one of Gibbs-like sampling, chain-of-thought sampling, and ranked iterative decoding with and without re-masking. In fig. S1, we plot the pLDDT and pTM scores for each sample. We find that ranked iterative decoding provides by far the highest confidence samples and that our re-masking strategy further improves both of these metrics. In Hayes et al. (2024) it was shown that samples with pLDDT and pTM > 0.8 were highly designable. We reproduce this result in fig. S2 by plotting the cCRMSD of proteins passing this confidence threshold versus all proteins generated by the model.

Table 1: Evaluation of co-design performance across ESM3 size and sampling strategies

Method		ccRMSD (#)	ccRMSD < 2Å (")	# Seq. Clus. (")	# Str. Clus. (")	MMseqs Seq Id% (#)	Foldseek TM (#)	-Sheet (%)	-Helix (%)
Large (98B)	Gibbs-like	17.6	0.16	6	6	0.85	0.86	0.16	0.37
	CoT	5.67	0.54	50	30	0.56	0.64	0.29	0.42
	Rank	4.50	0.59	50	37	0.57	0.63	0.26	0.46
Medium (7B)	Gibbs-like	15.4	0.15	15	14	0.60	0.66	0.09	0.50
	CoT	9.81	0.38	26	21	0.60	0.64	0.21	0.52
	Rank	6.52	0.35	33	27	0.59	0.55	0.26	0.41
Small (1.4B)	Gibbs-like	30.5	0.15	6	14	0.67	0.70	0.00	0.91
	CoT	16.5	0.34	22	20	0.52	0.80	0.05	0.42
	Rank	8.96	0.28	22	19	0.59	0.65	0.22	0.50

### 3.3 CO-DESIGN PERFORMANCE SCALES WITH PLM SIZE

Next, we investigate the effect of model size on co-design performance. ESM3 offers three model sizes: ESM3-small (1.4B), ESM3-medium (7B), and ESM3-large (98B). The latter two are available via API calls, but are highly rate limited. Thus, we generate 100 proteins of length 100 for each model using each sampling strategy and plot the ccRMSD by method and model size in [g. 1](#). We observe that increasing the model size improves designability across sampling strategies and that ranked iterative decoding remains the superior method. We report comprehensive metrics for this experiment in [table 1](#), where we observe that increasing the model size also improves the diversity of sampled proteins with a greater number of unique sequence and structure clusters. Finally, we examine the proportion of alpha helix and beta sheet residues in the generated samples in [g. 1](#). Here, we additionally include samples obtained via the secondary structure ranking function introduced in [section 2.2](#) by prioritizing beta sheets, which successfully shifts the composition of our samples.

Figure 1: (left) ccRMSD of generated samples by ESM3 variants with different sampling strategies. (right) Proportion of alpha helix versus beta sheet residues for different generation methods.

### 3.4 COMPARISON TO BASELINE METHODS

Following previous work on unconditional co-design, we sample 100 proteins for each length in  $f \in \{50, 100, 200, 500\}$  for a total of 400 samples per method. We sample baseline methods with default parameters as provided by the authors. For the ranked iterative decoding methods, we use the min-entropy ranking function for unmasking and the max-entropy ranking for re-masking. These choices were made after a comprehensive sweep of ranking functions in [appendices D.1 and D.2](#).

The results in [tables 2 and 3](#) highlight key tradeoffs in protein co-design, where optimizing for one metric often comes at the cost of another. Designability, as measured by the proportion of structures

with ccRMSD < 2Å, reveals that bespoke methods like Multi ow (0.71) outperform the ESM3-based methods (0.60 for ESM3 (Rank w/ remask)). However, the fact that a general-purpose protein language model (PLM) like ESM3, without architectural modifications or fine-tuning, approaches this performance demonstrates its scalability and adaptability in tackling co-design tasks. When considering diversity, novelty, and quality together, ESM3 (Rank w/ remask) achieves a balanced performance. Although it generates fewer sequence clusters (82) compared to Multi ow (327), it maintains moderate novelty with a Foldseek TM-score of 0.84 and competitive quality metrics, including a perplexity of 4.23 and a scSR score of 0.45. By contrast, methods like PLAID offer higher novelty at the cost of lower designability and higher perplexity. This balance achieved by ESM3 underscores its ability to produce biologically coherent and novel sequences while remaining scalable, making it a promising approach for large-scale protein co-design tasks.

Table 2: Evaluation of protein Designability using cross-consistency (cc-\*) and self-consistency (sc-\*) metrics.

	ccTM (" )	scTM (" )	ccRMSD (#)	ccSR (" )	ccRMSD < 2Å (" )
Multi ow	0.87	0.87	3.19	0.51	0.71
PLAID	0.68	0.61	8.19	0.26	0.31
ProteinGenerator	0.60	0.66	10.4	0.28	0.18
Protpardelle	0.69	0.65	11.9	0.45	0.45
ESM3-small (Gibbs-like)	0.41	0.43	32.8	0.15	0.16
ESM3-small (CoT)	0.58	0.52	19.9	0.27	0.42
ESM3-small (Rank w/o remask)	0.66	0.62	27.3	0.43	0.52
ESM3-small (Rank w/ remask)	0.71	0.68	23.3	0.44	0.60

Table 3: Evaluation of Quality, Diversity and Novelty of co-designed proteins. pLDDT refers to the confidence score returned by the generative model; "-" is used for models which do not produce a pLDDT metric. Diversity and novelty metrics are computed on designable proteins (ccRMSD < 2Å).

	Diversity			Novelty		Quality				
	inner TM (#)	#Seq. Clus. (" )	#Str. Clus. (" )	MMseqs Seq Id% (#)	Foldseek TM (#)	pLDDT (" )	-Sheet (%)	-Helix (%)	scSR (" )	PPL. (#)
Multi ow	0.39	327	83	0.61	0.41	-	0.12	0.70	0.56	8.65
PLAID	0.32	168	88	0.50	0.79	0.59	0.14	0.44	0.26	15.37
ProteinGenerator	0.45	144	14	0.57	0.50	0.73	0.03	0.67	0.33	9.43
Protpardelle	0.49	97	73	0.56	0.51	-	0.12	0.54	0.46	9.34
ESM3 (Gibbs-like)	0.81	25	15	0.67	0.95	0.65	0.003	0.64	0.11	5.95
ESM3 (CoT)	0.55	55	105	0.58	0.71	0.77	0.06	0.53	0.31	5.6
ESM3 (Rank w/o remask)	0.34	67	81	0.63	0.82	0.79	0.14	0.42	0.44	4.12
ESM3 (Rank w/ remask)	0.39	82	79	0.66	0.84	0.81	0.16	0.39	0.45	4.23

## 4 CONCLUSION AND LIMITATIONS

In this work, we investigated the potential of multimodal protein language models like ESM3 for sequence and structure co-design, focusing on enhancing their generative capabilities through improved sampling strategies. Our benchmarking across model sizes further underscores the adaptability of foundational PLMs for protein co-design. Further, we conduct a comprehensive benchmark across different model sizes and against recent co-design methods, showcasing the potential of out-of-the-box foundational PLMs for this task. While our results highlight the promise of PLMs for co-design, they also reveal that ESM3 does not yet surpass state-of-the-art methods across all evaluation metrics. This suggests that additional refinements such as fine-tuning process after the MLM pre-training stage (Lu et al., 2024b) may be necessary to fully unlock the model's generative potential. Despite these limitations, our study highlights important challenges in multi-modal protein generation and lays the groundwork for future advancements to build upon. We hope that our findings will motivate further research into protein co-design, advancing the development of generative models that better capture the relationship between protein sequence and structure.

270 MEANINGFULNESS STATEMENT  
271

272 A meaningful representation of life requires understanding the intricate relationships between key  
273 biological modalities—sequence, structure, function, and dynamics—that drive cellular processes.  
274 Our work contributes to this direction by exploring how ESM3, a multimodal foundation protein  
275 language model, can jointly sample and co-design protein sequences and structures. By addressing  
276 the challenges of capturing these interconnected modalities, we take a step toward developing models  
277 capable of representing the complexity of life at the molecular level.

278 REFERENCES  
279

- 280 Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf  
281 Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure  
282 prediction of biomolecular interactions with alphafold3. *Nature*, pp. 1–3, 2024.
- 283 Sarah Alamdari, Nitya Thakkar, Rianne van den Berg, Neil Tenenholtz, Robert Strome, Alan M.  
284 Moses, Alex X. Lu, Nicolò Fusi, Ava P. Amini, and Kevin K. Yang. Protein generation with  
285 evolutionary diffusion: sequence is all you need. *bioRxiv*, 2024. doi: 10.1101/2023.09.11.  
286 556673. URL [https://www.biorxiv.org/content/early/2024/11/04/2023.](https://www.biorxiv.org/content/early/2024/11/04/2023.09.11.556673)  
287 [09.11.556673](https://www.biorxiv.org/content/early/2024/11/04/2023.09.11.556673).
- 288 Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie  
289 Lee, Jue Wang, Qian Cong, Lisa N. Kinch, R. Dustin Schaeffer, Claudia Millahn, Beom Park,  
290 Carson Adams, Caleb R. Glassman, Andy DeGiovanni, Jose H. Pereira, Andria V. Rodrigues,  
291 Alberdina A. van Dijk, Ana C. Ebrecht, Diederik J. Opperman, Theo Sagmeister, Christoph  
292 Buhlheller, Tea Pavkov-Keller, Manoj K. Rathinaswamy, Udit Dalwadi, Calvin K. Yip, John E.  
293 Burke, K. Christopher Garcia, Nick V. Grishin, Paul D. Adams, Randy J. Read, and David  
294 Baker. Accurate prediction of protein structures and interactions using a three-track neural  
295 network. *Science* 373(6557):871–876, 2021. doi: 10.1126/science.abj8754. URL  
296 <https://www.science.org/doi/abs/10.1126/science.abj8754>
- 297 Surojit Biswas, Grigory Khimulya, Ethan C. Alley, Kevin M. Esvelt, and George M. Church.  
298 Low-n protein engineering with data-efficient deep learning. *bioRxiv*, 2020. doi: 10.1101/  
299 2020.01.23.917682. URL [https://www.biorxiv.org/content/early/2020/08/](https://www.biorxiv.org/content/early/2020/08/31/2020.01.23.917682)  
300 [31/2020.01.23.917682](https://www.biorxiv.org/content/early/2020/08/31/2020.01.23.917682).
- 301 Avishek Joey Bose, Tara Akhound-Sadegh, Guillaume Huguet, Kilian Fatras, Jarrid Rector-Brooks,  
302 Cheng-Hao Liu, Andrei Cristian Nica, Maksym Korablyov, Michael Bronstein, and Alexander Tong.  
303 Se (3)-stochastic flow matching for protein backbone generation. *bioRxiv preprint arXiv:2310.02391*  
304 [2023](https://arxiv.org/abs/2310.02391).
- 305 Nadav Brandes, Dan Ofer, Yam Peleg, Nadav Rappoport, and Michal Linial. ProteinBERT: a universal  
306 deep-learning model of protein sequence and function. *Bioinformatics* 38(8):2102–2110, 02 2022.  
307 ISSN 1367-4803. doi: 10.1093/bioinformatics/btac020. URL [https://doi.org/10.1093/](https://doi.org/10.1093/bioinformatics/btac020)  
308 [bioinformatics/btac020](https://doi.org/10.1093/bioinformatics/btac020)
- 309 Andrew Campbell, Jason Yim, Regina Barzilay, Tom Rainforth, and Tommi Jaakkola. Generative  
310 flows on discrete state-spaces: Enabling multimodal flows with applications to protein co-design,  
311 2024. URL <https://arxiv.org/abs/2402.04997>
- 312 Ricky TQ Chen and Yaron Lipman. Riemannian flow matching on general geometries. *arXiv preprint*  
313 [arXiv:2302.03660](https://arxiv.org/abs/2302.03660) 2023.
- 314 Alexander E. Chu, Lucy Cheng, Gina El Nesr, Minkai Xu, and Po-Ssu Huang. An all-atom protein  
315 generative model. *bioRxiv*, 2023. doi: 10.1101/2023.05.24.542194. URL [https://www.](https://www.biorxiv.org/content/early/2023/05/25/2023.05.24.542194)  
316 [biorxiv.org/content/early/2023/05/25/2023.05.24.542194](https://www.biorxiv.org/content/early/2023/05/25/2023.05.24.542194)
- 317 J. Dauparas, I. Anishchenko, N. Bennett, H. Bai, R. J. Ragotte, L. F. Milles, B. I. M. Wicky,  
318 A. Courbet, R. J. de Haas, N. Bethel, P. J. Y. Leung, T. F. Huddy, S. Pellock, D. Tischer, F. Chan,  
319 B. Koepnick, H. Nguyen, A. Kang, B. Sankaran, A. K. Bera, N. P. King, and D. Baker. Robust  
320 deep learning-based protein sequence design using proteinGPT. *Science* 378(6615):49–56, 2022.  
321 doi: 10.1126/science.add2187. URL [https://www.science.org/doi/abs/10.1126/](https://www.science.org/doi/abs/10.1126/science.add2187)  
322 [science.add2187](https://www.science.org/doi/abs/10.1126/science.add2187)

- 324 Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding.  
325 preprint arXiv:1810.04805, 2018.
- 326  
327 Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep  
328 bidirectional transformers for language understanding, 2019. [https://arxiv.org/  
329 abs/1810.04805](https://arxiv.org/abs/1810.04805) .
- 330 Chai Discovery, Jacques Boitreaud, Jack Dent, Matthew McPartlon, Joshua Meier, Vinicius Reis,  
331 Alex Rogozhnikov, and Kevin Wu. Chai-1: Decoding the molecular interactions of life. [arXiv,  
332 pp. 2024–10, 2024.](https://arxiv.org/abs/2024.10.2024)
- 333 Noelia Ferruz, Steffen Schmidt, and Birte Öster. Protgpt2 is a deep unsupervised language model  
334 for protein design. *Nature communications*, 13(1):4348, 2022.
- 335  
336 Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky T. Q. Chen, Gabriel Synnaeve, Yossi Adi, and  
337 Yaron Lipman. Discrete flow matching. (arXiv:2407.15595), November 2024. doi: 10.48550/  
338 arXiv.2407.15595. URL <http://arxiv.org/abs/2407.15595> . arXiv:2407.15595.
- 339 Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian  
340 restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 6  
341 721–741, 1984.
- 342 W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.
- 343  
344 Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert  
345 Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat,  
346 Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra,  
347 Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido,  
348 and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*  
349 0(0):eads0018. doi: 10.1126/science.ads0018. [https://www.science.org/doi/  
350 abs/10.1126/science.ads0018](https://www.science.org/doi/abs/10.1126/science.ads0018) .
- 351 Tomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert  
352 Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. Simulating 500 million years of  
353 evolution with a language model. [arXiv, pp. 2024–07, 2024.](https://arxiv.org/abs/2024.07.2024)
- 354 Daniel Hesslow, Niccolò Zanichelli, Pascal Notin, Iacopo Poli, and Debora Marks. Rita: a study on  
355 scaling up generative protein sequence models. [arXiv preprint arXiv:2205.05789,  
356 2022.](https://arxiv.org/abs/2205.05789)
- 357 Mingyang Hu, Fajie Yuan, Kevin Yang, Fusong Ju, Jin Su, Hui Wang, Fei Yang, and Qiuyang  
358 Ding. Exploring evolution-aware &-free protein language models as protein function predictors.  
359 *Advances in Neural Information Processing Systems*, 35:38873–38884, 2022.
- 360 Guillaume Huguet, James Vuckovic, Kilian Fatras, Eric Thibodeau-Laufer, Pablo Lemos, Riashat  
361 Islam, Cheng-Hao Liu, Jarrid Rector-Brooks, Tara Akhound-Sadegh, Michael Bronstein, et al.  
362 Sequence-augmented self-supervised flow matching for conditional protein backbone generation.  
363 preprint arXiv:2405.20313, 2024.
- 364 John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent  
365 Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. Illuminating protein  
366 space with a programmable generative model. *Nature*, pp. 1–9, 2023.
- 367 John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger,  
368 Kathryn Tunyasuvunakool, Russ Bates, Augustin Zidek, Anna Potapenko, et al. Highly accurate  
369 protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.
- 370  
371 Yeqing Lin and Mohammed AlQuraishi. Generating novel, designable, and diverse protein structures  
372 by equivariantly diffusing oriented residue clouds. [arXiv preprint arXiv:2301.12485,  
373 2023.](https://arxiv.org/abs/2301.12485)
- 374 Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin,  
375 Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom  
376 Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level pro-  
377 tein structure with a language model. *Science* 379(6637):1123–1130, 2023a. doi: 10.1126/  
[science.  
ads2574](https://www.science.org/doi/abs/10.1126/science.ads2574) .

- 378 Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin,  
379 Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level  
380 protein structure with a language model. *Science* 379(6637):1123–1130, 2023b.
- 381
- 382 Sidney Lyayuga Lisanza, Jake Merle Gershon, Sam Tipps, Lucas Arnoldt, Samuel Hendel,  
383 Jeremiah Nelson Sims, Xinting Li, and David Baker. Joint generation of protein sequence and  
384 structure with rosettafold sequence space diffusion. *bioRxiv*, 2023. doi: 10.1101/2023.05.08.  
385 539766. URL [https://www.biorxiv.org/content/early/2023/05/10/2023.  
386 05.08.539766](https://www.biorxiv.org/content/early/2023/05/10/2023.05.08.539766).
- 387 Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint*  
388 [arXiv:1907.11692](https://arxiv.org/abs/1907.11692), 2019.
- 389
- 390 Amy X. Lu, Wilson Yan, Sarah A. Robinson, Kevin K. Yang, Vladimir Gligorijevic, Kyunghyun Cho,  
391 Richard Bonneau, Pieter Abbeel, and Nathan Frey. Generating all-atom protein structure from  
392 sequence-only training data. *bioRxiv*, 2024a. doi: 10.1101/2024.12.02.626353. URL <https://www.biorxiv.org/content/early/2024/12/05/2024.12.02.626353>.
- 393
- 394 Jiarui Lu, Xiaoyin Chen, Stephen Zhewen Lu, Chence Shi, Hongyu Guo, Yoshua Bengio, and  
395 Jian Tang. Structure language models for protein conformation generation. *arXiv preprint*  
396 [arXiv:2410.18403](https://arxiv.org/abs/2410.18403), 2024b.
- 397
- 398 Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton,  
399 Jose Luis Olmos, Caiming Xiong, Zachary Z Sun, Richard Socher, James S. Fraser, and Nikhil Vijay  
400 Naik. Large language models generate functional protein sequences across diverse features.  
401 *Biotechnology* pp. 1–8, 2023. URL [https://api.semanticscholar.org/CorpusID:  
402 256304602](https://api.semanticscholar.org/CorpusID:256304602).
- 403 Joshua Meier, Roshan Rao, Robert Verkuil, Jason Liu, Tom Sercu, and Alex Rives. Lan-  
404 guage models enable zero-shot prediction of the effects of mutations on protein function. In  
405 M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.),  
406 *Advances in Neural Information Processing Systems*, volume 34, pp. 29287–29303. Curran Asso-  
407 ciates, Inc., 2021. URL [https://proceedings.neurips.cc/paper\\_files/paper/  
408 2021/file/f51338d736f95dd42427296047067694-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2021/file/f51338d736f95dd42427296047067694-Paper.pdf).
- 409 Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward  
410 Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*  
411 21(6):1087–1092, 1953.
- 412
- 413 Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo,  
414 Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from  
415 scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National  
416 Academy of Sciences* 118(15):e2016239118, 2021.
- 417 Martin Steinegger and Johannes Song. Mmseqs2: sensitive protein sequence searching for the  
418 analysis of massive data sets. *bioRxiv*, 2017. doi: 10.1101/079681. URL [https://www.  
419 biorxiv.org/content/early/2017/06/07/079681](https://www.biorxiv.org/content/early/2017/06/07/079681).
- 420 Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein  
421 language modeling with structure-aware vocabulary. *bioRxiv*, pp. 2023–10, 2023.
- 422
- 423 Baris E. Suzek, Yuqi Wang, Hongzhan Huang, Peter B. McGarvey, Cathy H. Wu, and the UniProt Con-  
424 sortium. Uniref clusters: a comprehensive and scalable alternative for improving sequence sim-  
425 ilarity searches. *Bioinformatics* 31(6):926–932, 11 2014. ISSN 1367-4803. doi: 10.1093/  
426 bioinformatics/btu739. URL <https://doi.org/10.1093/bioinformatics/btu739>.
- 427 Brian L Trippe, Jason Yim, Doug Tischer, Tamara Broderick, David Baker, Regina Barzilay, and  
428 Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-  
429 scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022.
- 430
- 431 Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning,  
2018. URL <https://arxiv.org/abs/1711.00937>.



- 432 Michel van Kempen, Stephanie S. Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee,  
433 Cameron L.M. Gilchrist, Johannes Song, and Martin Steinegger. Fast and accurate protein  
434 structure search with foldseek. *bioRxiv*, 2023. doi: 10.1101/2022.02.07.479398. URL  
435 <https://www.biorxiv.org/content/early/2023/03/28/2022.02.07.479398>
- 436  
437 Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz  
438 Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL  
439 <https://arxiv.org/abs/1706.03762>
- 440 Alex Wang and Kyunghyun Cho. Bert has a mouth, and it must speak: Bert as a markov random field  
441 language model, 2019. URL  
442 <https://arxiv.org/abs/1902.04094>
- 443 Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion  
444 language models are versatile protein learners. *arXiv preprint arXiv:2402.18567*, 2024a.
- 445 Xinyou Wang, Zaixiang Zheng, Fei Ye, Dongyu Xue, Shujian Huang, and Quanquan Gu. Diffusion  
446 language models are versatile protein learners, 2024b. URL  
447 <https://arxiv.org/abs/2402.18567>
- 448 Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach,  
449 Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein  
450 structure and function with rdiffusion. *Nature*, 620(7976):1089–1100, 2023.
- 451  
452 Jeremy Wohlwend, Gabriele Corso, Saro Passaro, Mateo Reveiz, Ken Leidal, Wojtek Swiderski, Tally  
453 Portnoi, Itamar Chinn, Jacob Silterra, T. Jaakkola, and Regina Barzilay. Boltz-1 democratizing  
454 biomolecular interaction modeling. *bioRxiv*, 2024. URL  
455 <https://api.semanticscholar.org/CorpusID:274166333>
- 456 Ruidong Wu, Fan Ding, Rui Wang, Rui Shen, Xiwen Zhang, Shitong Luo, Chenpeng Su, Zuofan  
457 Wu, Qi Xie, Bonnie Berger, Jianzhu Ma, and Jian Peng. High-resolution de novo structure  
458 prediction from primary sequence. *bioRxiv*, 2022. doi: 10.1101/2022.07.21.500999. URL  
459 <https://www.biorxiv.org/content/early/2022/07/22/2022.07.21.500999>
- 460  
461 Jason Yim, Andrew Campbell, Andrew YK Foong, Michael Gastegger, José  
462 Lewis, Victor Garcia Satorras, Bastiaan S Veeling, Regina Barzilay, Tommi Jaakkola, et al. Fast  
463 protein backbone generation with se (3) now matching. *arXiv preprint arXiv:2310.05297*, 2023a.
- 464 Jason Yim, Brian L Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay,  
465 and Tommi Jaakkola. Se (3) diffusion model with application to protein backbone generation.  
466 *arXiv preprint arXiv:2302.02277*, 2023b.
- 467  
468 Zuobai Zhang, Jiarui Lu, Vijil Chenthamarakshan, ~~Ali~~ Lozano, Payel Das, and Jian Tang.  
469 Structure-informed protein language models. *arXiv preprint arXiv:2402.05856*, 2024.
- 470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485

## A EXTENDED RELATED WORK

Here we describe additional related work.

**Protein Language Foundation Models** In recent years, several language models of protein sequence have been built. Among these, the ESM series (Meier et al., 2021; Lin et al., 2023a) have become dominant with BERT-like pretraining objectives (Devlin et al., 2019), garnering great attention for successful application in downstream tasks such as property engineering (Biswas et al., 2020). On the other hand, models of protein structure have dramatically increased in scale, evolving into multipurpose foundation models that enable a wide range of downstream tasks. Structure prediction models like the AlphaFold series (Jumper et al., 2021; Abramson et al., 2024), Chai-1 (Discovery et al., 2024), and most recently Boltz-1 (Wohlwend et al., 2024) are now capable of modeling complex multimodal biomolecular interactions beyond just protein complexes. At the intersection of these directions lies the emerging trend of protein language foundation models (PLFM) such as ESM3 (Hayes et al.) that tokenize the protein structure and jointly model all modalities as a family of languages. These multipurpose models are capable of both representation learning and generative modeling tasks and benefit from the performance scaling of the transformer architecture (Vaswani et al., 2023).

**Generative Modeling for Proteins** State-of-the-art models for protein design have mainly focused on generating the backbone folds. Among these, diffusion-based (Ingraham et al., 2023; Watson et al., 2023) and flow-based (Chen & Lipman, 2023; Bose et al., 2023; Huguet et al., 2024) models operating on protein frames (Yim et al., 2023a;b) or simply the residues (Lin & AlQuraishi, 2023; Trippe et al., 2022) have proven successful. On the other hand, methods also exist for designing protein sequences (Brandes et al., 2022; Alamdari et al., 2024; Madani et al., 2023), usually followed by structure generation with a folding model. However, protein design is inherently a multimodal problem that requires sampling from the joint distribution of sequence and structure. Instead of factorizing this into a two-step procedure, MultiFlow (Campbell et al., 2024) builds on top of FrameFlow (Yim et al., 2023b) and uses discrete flow matching to simultaneously generate the sequence. More recently, PLAID (Lu et al., 2024a) performs Euclidean diffusion in the latent space of ESMFold (Lin et al., 2023a) and learns a sequence decoder to jointly sample both modalities.

**Sampling from Language Models** Protein language models estimate transition probabilities between sequences, making them suitable for interpretation as Markov random fields. When the system's mutational space stabilizes, various Markov Chain Monte Carlo (MCMC) techniques, including Gibbs sampling and Metropolis-Hastings can be applied to explore sequence space efficiently (Geman & Geman, 1984; Metropolis et al., 1953; Hastings, 1970). For example, Wang & Cho (2019) demonstrated how Gibbs sampling could be used to generate text by treating the English language model Bert as a Markov random field.

## B ADDITIONAL METHODS DETAILS

Algorithm descriptions. Here in algorithms 1 to 3, we provide the pseudocode for Gibbs-like sampling, Chain-of-thought sampling, and ranked iterative decoding with re-masking, respectively.

---

### Algorithm 1 Gibbs-like Sampling

---

```

1: Input: ESM3m (x; z), sequence length L, temperature t
2: for i = shuffle(1; ...; L) do
3:    $x^i \sim \exp(\log m(x^i | x^{[1:i-1]})) = t$ 
4: for i = shuffle(1; ...; L) do
5:    $z^i \sim \exp(\log m(z^i | z^{[1:i-1]}; x)) = t$ 
6: return (x; z)
```

---

---

**Algorithm 2 Chain-of-Thought Sampling**


---

```

1: Input: ESM3m (x; z; s), sequence length L, temperature t
2: for i = shuf e(f 1; :::; L) g do
3:   si = exp(log m (sj | s1:i-1) = t)
4: for i = shuf e(f 1; :::; L) g do
5:   zi = exp(log m (zj | z1:i-1; s) = t)
6: for i = shuf e(f 1; :::; L) g do
7:   xi = exp(log m (xj | x1:i-1; z; s) = t)
8: return (x; z)

```

---



---

**Algorithm 3 Rank-Informed Iterative Decoding w/ Remasking**


---

```

1: Input: ESM3m (x; z), forward ranking function ff(i) 8 i 2 L, backward ranking function
   b(i) 8 i 2 L, number of steps N, decoding schedule gn=1N, remasking schedule rn=1N,
   temperature schedule Tn=1N
2: Initialize x; z = mx; mz
3: for n = 1 to N do
4:   Rank i 2 L using ff(i)
5:   Rank i 2 L using b(i)
6:   Sample predictions xi; zi = exp(log m (xi; zi | x; z) = Tn); i 2 [L]
7:   Evaluate forward ranking scores ux; uz = f(xi); f(zi)
8:   Evaluate backward ranking scores vx; vz = b(xi); b(zi)
9:   Select positions to unmask Ux; Uz = argsort(ux)[ -rn :]; argsort(uz)[ -rn :]
10:  Assign xi; zi = xi; zi 8 i 2 Ux; j 2 Uz
11:  Select positions to re-mask Vx; Vz = argsort(vx)[ -rn :]; argsort(vz)[ -rn :]
12:  Assign xi; zi = [MASK]x; [MASK]z; 8 i 2 Vx; j 2 Vz
13: return (x; z)

```

---

## C ADDITIONAL EXPERIMENTAL DETAILS

### C.1 CO-DESIGN METRICS

Here, we provide descriptions of the metrics we use to evaluate the co-designed samples in section 3.

**Designability.** We fold the sequence using ESMFold (Lin et al., 2023a) to obtain a predicted structure  $\hat{y}(x)$ , then calculate the cross-consistency TM-score (ccTM) and cross-consistency RMSD (ccRMSD) between  $\hat{y}(x)$  and  $y(x)$ . Conversely, we use ProteinMPNN (Dauparas et al., 2022) to predict 8 sequences from the structure and report the best sequence recovery (SSR) with respect to  $x$ . Additionally, we use OmegaFold (Wu et al., 2022) to re-fold these 8 sequences and report the lowest TM-score (scTM) between these predicted structures and the original coordinates. Comparing ccTM against scTM tells us how consistent the co-designed sequence is compared to one obtained by inverse-folding the structure retroactively.

**Quality.** For structure quality, we report the percentage of alpha-helix and beta-strand residues in the secondary structure. For sequence quality, we report consistency sequence recovery (CSR) which is obtained by folding a generated sequence then inverse-folding the result. Finally, we evaluate the sequence perplexity under an autoregressive protein language model RITA-XL (Hesslow et al., 2022).

**Diversity & Novelty.** We first retain only the generated proteins that are designable (ccRMSD < 2Å). Among these samples, we report the number of sequence and structure clusters computed by MMseqs2 (Steinegger & Coling, 2017) and Foldseek (van Kempen et al., 2023) respectively. For novelty, we report the Foldseek TM-score between the designable structures and their closest match in PDB100. Similarly, for every designable sequence, we calculate the sequence identity with the closest homolog in UniRef50 (Suzek et al., 2014).

## D ADDITIONAL RESULTS

### D.1 CHOICE OF UN-MASKING RANKING FUNCTION

We provide here co-design results obtained by sweeping the ranking functions described in section 2.2. For each strategy, we follow the same protocol as section 3.4 by sampling 100 proteins at various sequence lengths, and report the following metrics: fraction designable (ccRMSD < 2Å), number of designable sequence and structure clusters (#Seq. Clus. & #Str. Clus), the sequence identity to UniRef50 (MMseqs Seq Id%), the Foldseek TM score (Foldseek TM), and the percentage of alpha and beta sheet residues (-Helix, -Sheet). We use sampling temperatures 0.25, 1:2 and 0:4 for the sequence, structure, and secondary structure tracks with temperature annealing.

Table S1: Evaluation of co-design performance using different un-masking ranking functions

	ccRMSD < 2Å (%)	# Seq. Clus. (#)	# Str. Clus. (#)	MMseqs Seq Id% (#)	Foldseek TM (#)	-Sheet (%)	-Helix (%)
Min Entropy	0.66	67	81	0.63	0.82	0.14	0.42
Max Logit	0.67	62	75	0.61	0.80	0.13	0.40
SS8	0.56	50	68	0.58	0.78	0.10	0.50
SS8	0.58	54	70	0.59	0.79	0.24	0.35

### D.2 CHOICE OF REMASKING RANKING FUNCTION

Similarly to appendix D.1, we provide co-design results for the ranking function used in the re-masking step described in section 2.2.2. For this experiment, we use minimum entropy ranking for the un-masking strategy and the same temperature scheme as described above.

Table S2: Evaluation of co-design performance using different re-masking ranking functions

	ccRMSD < 2Å (%)	# Seq. Clus. (#)	# Str. Clus. (#)	MMseqs Seq Id% (#)	Foldseek TM (#)	-Sheet (%)	-Helix (%)
Max Entropy	0.71	82	79	0.66	0.84	0.16	0.39
Min Logit	0.69	80	78	0.67	0.85	0.15	0.38

### D.3 EVALUATING DESIGNABILITY OF ESM3 SAMPLES BY CONFIDENCE

Here, we evaluate the choice of sampling strategy on the confidence and designability of samples generated by ESM3-open. In [fig. S1](#), we plot the pLDDT and pTM of samples generated by each method. Additionally, in [fig. S2](#), we plot the ccrMSD of samples obtained via ESM3. We see that the high confidence samples with pLDDT > 0.8 and pTM > 0.8 achieve higher designability (RMSD < 2Å).

Figure S1: pLDDT (left) and pTM (right) of unconditional generations using ESM3-open according to sampling method (CoT: Chain-of-Thought, Ranked Decoding, Iterative Decoding)

Figure S2: Cross-consistency RMSD of high confidence samples from ESM3 with pLDDT & pTM > 0.8 (left) and all ESM3 samples (right)

702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755

#### D.4 STRUCTURE GALLERY

Here we plot protein structures of varying lengths sampled from ESM3.

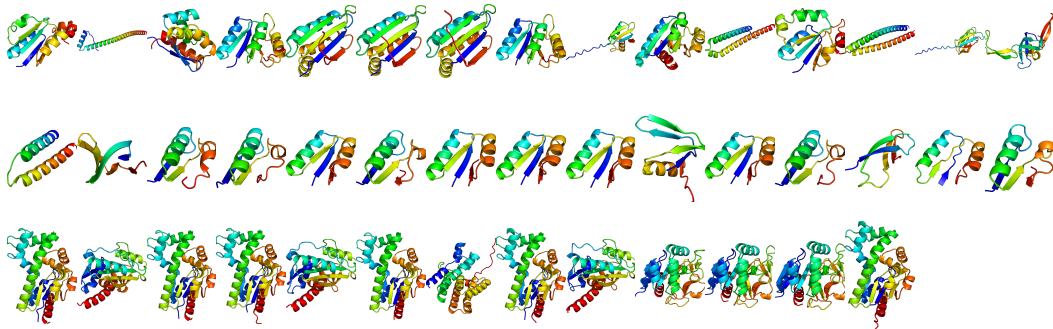


Figure S3: Unconditionally co-designed samples by ESM3 using ranked iterative sampling with re-masking 3. Rows are ordered from top to bottom by increasing sequence length in  $f_{50}$ , 100, 200, 500g.