MeMDLM: *De Novo* Membrane Protein Design with Masked Discrete Diffusion Protein Language Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Masked Diffusion Language Models (MDLMs) have recently emerged as a strong class of generative models, paralleling state-of-the-art (SOTA) autoregressive (AR) performance across natural language modeling domains. While there have been advances in AR as well as both latent and discrete diffusion-based approaches for protein sequence design, masked diffusion language modeling with protein language models (pLMs) is unexplored. In this work, we introduce MeMDLM, an MDLM tailored for membrane protein design, harnessing the SOTA pLM ESM-2 to *de novo* generate realistic membrane proteins for downstream experimental applications. Our evaluations demonstrate that MeMDLM-generated proteins exceed AR-based methods by generating sequences with greater transmembrane (TM) character. We further apply our design framework to scaffold soluble and TM motifs in sequences, demonstrating that MeMDLM-reconstructed sequences achieve greater biological similarity to their original counterparts compared to SOTA inpainting methods. Finally, we show that MeMDLM captures physicochemical membrane protein properties with similar fidelity as SOTA pLMs, paving the way for experimental applications. In total, our pipeline motivates future exploration of MDLM-based pLMs for protein design.

1 Introduction

1.1 Background

Membrane proteins play a crucial role in biological systems, regulating molecular transport, signal transduction, and cellular communication (1). Their capacity to bind specific ligands or undergo conformational changes renders them essential targets for drug development and therapeutics for various diseases (2). Even more interestingly, *de novo* design and engineering of membrane proteins offers a powerful therapeutic modality by enabling the creation of highly-specific and stable proteins that can precisely modulate cell signaling pathways, transport processes, and immune responses, making them ideal for targeting diseases such as cancer and neurological disorders (1). Current methods for designing new protein sequences or scaffolds rely on pre-trained structure-prediction networks (3; 4; 5), which remains a particularly challenging prerequisite for membrane protein targets. The scarcity of high-resolution structures hinders the training of high-fidelity DL structure prediction models for membrane proteins: only ~1% of the current PDB structures are annotated as membrane proteins. Further, energy functions underlying physics-based computational models are suboptimal for membrane proteins and often fail to accurately capture the interactions of membrane proteins within the lipid bilayer. As a result, current methods in de novo membrane protein design are limited to simple helical barrel or beta-barrel folds with low sequence complexity (6). The pitfalls of structure-based protein design methods and the clinical viability of membrane proteins necessitate a sequence-first design platform.

Submitted to the NeurIPS 2024 AIDrugsX Workshop.

Current methods for protein sequence generation leverage protein language models (pLMs) that capture physicochemical, structural, and functional properties of proteins on a per-residue basis from their sequence alone (7; 8). Although pLMs produce rich protein sequence embeddings, they are trained on the masked language modeling (MLM) objective, where a backbone model learns to reconstruct only a minor fraction of tokens (15%) across a sequence, making complete *de novo* generation difficult (9). However, recent advancements in generative language models have displayed the effectiveness of leveraging diffusion and autoregressive (AR) models for protein design tasks (10; 11; 12) as well as span masking (13) and progressive masking rate strategies (14). Still, there exists a significant gap between AR and diffusion language modeling. Notably, the MDLM objective has recently closed this performance gap: training BERT transformer encoder-style DNA language models on the MDLM objective significantly outperforms AR perplexity and sampling speed (15). This fusion between foundational biological models and MDLM offers a promising new frontier for protein design.

In this study, we introduce the first masked diffusion protein language model, MeMDLM. Specifically, this model uniquely leverages the MDLM framework to generate novel membrane protein sequences. MeMDLM introduces discrete noise to protein sequences by replacing amino acid tokens with `<mask>` tokens during the forward pass and reverses this corruption to *de novo* generate novel sequences or scaffolds. Overall, we introduce principled generative capabilities into BERT pLMs with the MDLM formulation by first pre-training the SOTA ESM-2 pLM on a comprehensive protein sequence space, then fine-tuning it on membrane protein sequences. After fine-tuning, MeMDLM is able to scaffold membrane protein domains and unconditionally generate diverse membrane protein sequences that capture the complexity of natural membrane proteins.

1.2 Related Works

Recent advancements in protein sequence generation have leveraged AR and diffusion-based approaches to produce naturalistic proteins. Specifically, Ferruz, et al., demonstrated the ability of a decoder-only transformer architecture along with the AR modeling objective to *de novo* generate biologically plausible sequences (12). Furthermore, Alamdari, et al., highlighted the effectiveness of leveraging discrete diffusion models and evolutionary information to accurately scaffold over functional motifs (10). These methods have showcased high novelty in protein generation while retaining physicochemical information.

2 Methods

2.1 Masked Diffusion Language Model (MDLM)

The MDLM training task leverages the absorbing-state forward diffusion process along with specific reverse diffusion parameterization rules to simplify the computation of the loss function and increase model accuracy. The absorbing state diffusion process, $q(\mathbf{z_t}, \mathbf{x})$ is a distribution parameterized by a time-conditioned noise schedule $\{\alpha_t\}$ that determines the probability of replacing a token with a mask token m at each timestep:

$$q(\mathbf{z_t}, \mathbf{x}) = \text{Cat}(\mathbf{z_t}; \alpha_t \mathbf{x} + (1 - \alpha_t)m) \tag{1}$$

Noise schedules are selected such that all tokens are masked by the end of all timesteps of the forward diffusion process, ensuring masked tokens are not unmasked during the forward diffusion process.

The reverse diffusion process, matching the estimated forward diffusion posterior $p(\mathbf{z_s} \mid \mathbf{z_t})$, is parameterized by a categorical distribution ("SUBS") that enforces restrictions on the original discrete diffusion formulation specific to absorbing state diffusion methods. During the SUBS-parameterized reverse diffusion process, unmasked tokens are unchanged and masked tokens are guaranteed to be unmasked.

$$p_\theta(\mathbf{z_s}|\mathbf{z_t}, \mathbf{x}) = \begin{cases} \text{Cat}(\mathbf{z_s}; \mathbf{z_t}) & \mathbf{z_t} \neq m \\ \text{Cat}\left(\mathbf{z_s}; \dfrac{(1 - a_s)m + (a_s - a_t)\mathbf{x}}{1 - a_t}\right) & \mathbf{z_t} = m \end{cases} \tag{2}$$

We utilize the ESM-2-150M pLM as the backbone model for learning the denoising network $x_\theta(\mathbf{z_t})$ that reconstructs the original sequence from its masked counterpart (7). Because SUBS "carries-over" unmasked tokens and masking rates are scheduled in a log-linear fashion, batches with 100% masking rates are problematic because x_θ does not have contextual information to guide the denoising process. Thus, we employ a maximal masking rate of 75% to ensure our denoising network learns long-range sequence dependencies while still training on higher masking rates to aid with *de novo* generation.

With the SUBS parameterization, we minimize a modified NELBO loss function, a Rao-Blackwellized form of the original D3PM loss (16) that eliminates the reconstruction loss term:

$$\mathcal{L}_T = \mathbb{E}_{q,t}\left[-\log p_\theta(\mathbf{x}|\mathbf{z}_{t(0)}) + T\left[\frac{a_t - a_s}{1 - a_t}\log\langle x_\theta(\mathbf{z_t}), \mathbf{x}\rangle\right]\right] \tag{3}$$

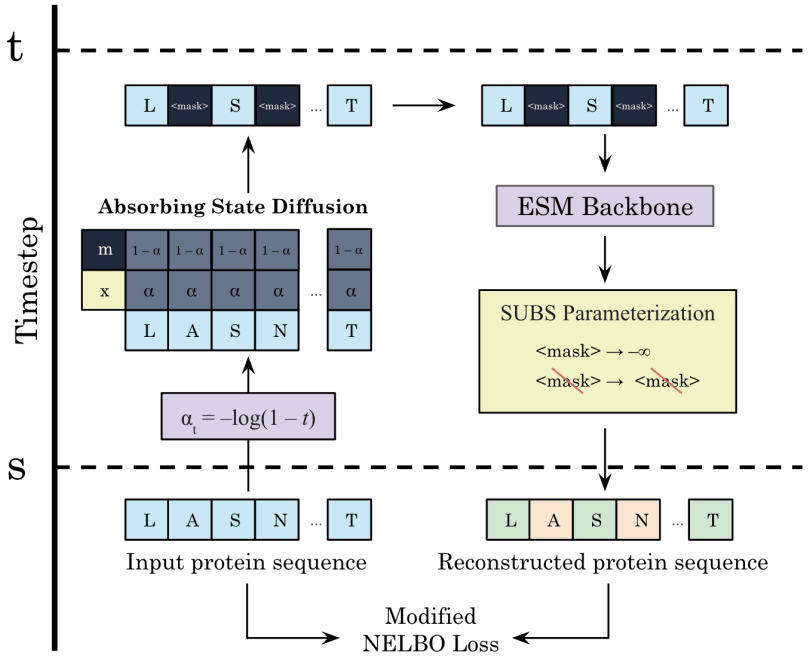

Figure 1: Denoising and noising processes guided by SUBS parameterization in MeMDLM. Protein sequences are corrupted according to the noising scheduler a_t and denoised via x_θ (ESM-2), calculating loss between the true and reconstructed sequence.

Overall, MeMDLM is a fine-tuned encoder that unconditionally generates membrane-like protein sequences and produces membrane-aware protein sequence embedding (Figure 1). To enable the ESM-2 pLM with principled generation capabilities, we first pre-train ESM-2-150M on the MDLM task using protein sequences that span the entire protein space. Then, we fine-tune this model using a MDLM head with only membrane protein sequences to facilitate *de novo* generation of membrane protein sequences.

2.2 Data

Pre-training data was sourced from UniRef50, sampling random sequences that span the entire protein space. Fine-tuning data was obtained from TM protein databases and included *de novo* generated sequences with varying sequence identity thresholds to introduce diversity; TM and soluble residues were also annotated within the sequences for downstream evaluations (17; 18; 19). The MMSeqs2 easy clustering module was used for homology-based sequence clustering into an 80-10-10 split of training, validation, and testing sequences. See Supplementary Section 5.6 for full data curation details.

2.3 Evaluation

TM Residue Prediction TM residues in MeMDLM, ProtGPT2, and experimentally annotated membrane protein sequences (from the test set) were predicted using Phobius (https://phobius.sbc.su.se/) (20; 21). We normalized the TM residue counts to sequence length and reported the frequency per 100 residues.

ESM-2 Pseudo Perplexity The model's generation quality was assessed using the ESM-2-650M pseudo-perplexity metric (7). Typically, a lower pseudo-perplexity value indicates higher confidence. Specifically, the pseudo-perplexity is computed as the exponential of the negative pseudo-log-likelihood of a sequence as in 4. This metric yields a deterministic value for each sequence but necessitates L forward passes for computation, where L represents the input sequence length.

$$\text{PPL}(\mathbf{x}) = \exp\left\{-\frac{1}{L}\sum_{i=1}^{L}\log p(\mathbf{x}_i|\mathbf{x}_{j\neq i})\right\} \tag{4}$$

Cosine Similarity To assess if sequences reconstructed from motif-scaffolding retained physicochemical properties of membrane proteins, we computed the cosine similarity, $\frac{\mathbf{x}\cdot\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|}$, between ESM-2-650M embeddings of the original sequence, $\mathbf{x}$, and the reconstructed sequence, $\mathbf{y}$. In this context, cosine similarity values closer to 1 indicate a strong biological similarity while values closer to 0 indicate dissimilarity. We utilized ESM-2-650M over ESM-2-150M to generate more expressive sequence embeddings, ensuring the evaluation was minimally influenced by the embedding quality.

Physicochemical Property Evaluation To determine if MeMDLM-generated sequences encode the physicochemical properties of membrane proteins, we evaluated the performance of MeMDLM latent embeddings on predicting per-residue solubility and membrane localization (22). In each case, we compared the predictive performance of MeMDLM embeddings against wild-type ESM-2-150M embeddings and ESM-MLM (ESM-2-150M fine-tuned on an MLM task using only membrane protein sequences; see Supplementary Section 5.4)

3 Results

***De Novo* Generation Quality** Given the limited availability of experimentally verified membrane structures, we focused on the overall TM character of the generated sequences by predicting TM residues with Phobius (21). Figure 2 represents a comparison of the TM residue frequency between experimentally annotated membrane proteins and *de novo* generated sequences. In this context, 927 experimental sequences were derived from the MeMDLM model test set, yielding a realistic evaluation of TM residue density (Supplementary Section 5.6). Specifically, Table 1 shows that the difference in mean predicted TM residues between MeMDLM and the test set is significantly lower than ProtGPT2 and the test set. These results suggest that the sequences generated from MeMDLM exhibit a density of TM residues much closer to experimentally verified membrane proteins, demonstrating that MeMDLM has successfully learned the underlying distribution of these proteins. In contrast, ProtGPT2 tends to severely under-generate TM residues, indicating a critical lack of understanding of some of the fundamental characteristics of functional membrane proteins. We further visualized randomly selected *de novo*-generated MeMDLM sequences with AlphaFold 3 (23) (Supplementary Figure 4) and observed alpha-helical bundles, the hallmark structural features of membrane proteins (24).

	Experimental	**MeMDLM**	**ProtGPT2**
Average TM Residue Frequency	29.193	25.737	15.562

Table 1: TM residue frequency (number of TM residues per 100 residues) in experimentally annotated, MeMDLM-generated, and ProtGPT2-generated protein sequences.

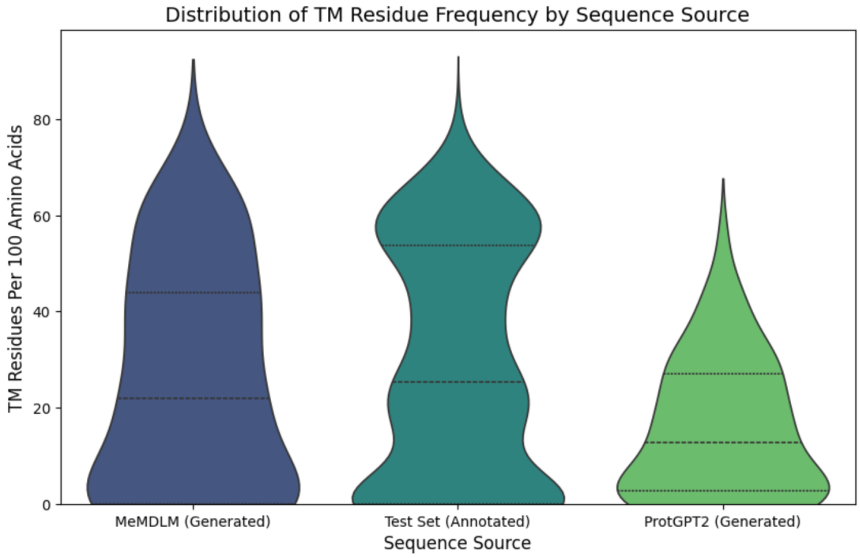

Figure 2: Distribution comparison of TM residue frequency predicted by Phobius for 100 MeMDLM-generated, 100 ProtGPT2-generated, and 927 experimentally annotated membrane protein sequences.

Scaffolded Generation Quality As a natural extension of *de novo* design, we scaffolded around TM and soluble motifs of experimentally annotated membrane proteins. We take the entire test set—comprising 927 experimentally verified membrane protein sequences with annotated TM and soluble motifs (Supplementary Section 5.6)—and we mask out all residues except those in the TM or soluble motif(s). We use these partially masked sequences as input to the models to assay their capability to generate scaffolds conditioned on known TM or soluble motifs. For this study, we focused on these domains due to their distinct hydrophilic and hydrophobic regions that govern the folding and thus function of the overall protein. Figure 3 compares MeMDLM and EvoDiff's reconstruction quality for TM and soluble domains of experimentally annotated membrane proteins.

	Transmembrane		**Soluble**	
	MeMDLM	**EvoDiff**	**MeMDLM**	**EvoDiff**
Pseudo Perplexity	3.819	20.554	7.029	16.991
Cosine Similarity	0.768	0.742	0.778	0.777

Table 2: Reconstruction quality comparison of models scaffolding around TM and soluble motifs of 927 experimental membrane protein sequences that reprsent the MeMDLM model test set.

This comparison considers the cosine similarity between ESM-2-650M embeddings of test set sequences and their reconstructed counterparts, along with the ESM-2-650M pseudo-perplexity of the reconstructed sequence. Table 1 shows that MeMDLM-inpainted sequences not only achieve lower average pseudo-perplexities but also exhibit cosine similarities closely aligned with EvoDiff-based scaffolds across both soluble and TM domains. These results suggest that MeMDLM scaffolds functional motifs with greater confidence while preserving biological relevance comparable to SOTA-generated scaffolds.

Representation Quality We finally assessed if the generated sequences retain physicochemical information critical to membrane protein function by predicting per-residue solubility and membrane localization (Table 3). MeMDLM latent embeddings achieve predictive performance that closely parallels SOTA pLM embeddings, which are designed specifically for delivering precise representa-

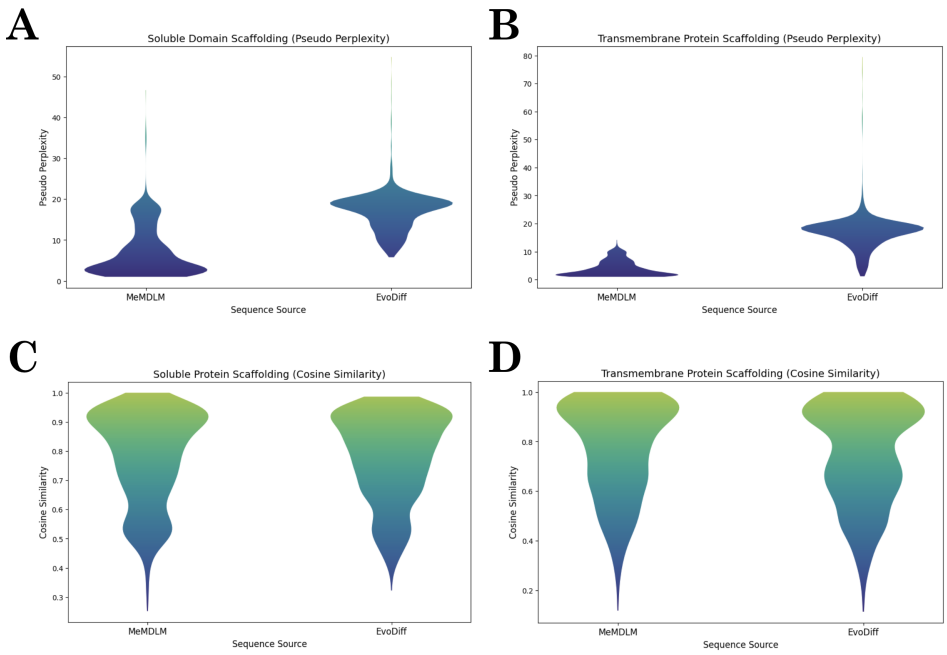

Figure 3: Distribution comparison of reconstruction quality. Scaffolding was performed over the test set sequences (927 experimentally annotated membrane proteins). **A, B** Pseudo perplexity of soluble and TM regions scaffolded by MeMDLM and EvoDiff. **C, D** Cosine similarity between embeddings of true and reconstructed sequences from MeMDLM and EvoDiff.

tions. In total, these results demonstrate that MeMDLM accurately captures the biological features underpinning functional membrane proteins.

	ESM-2-150M	**ESM-MLM**	**MeMDLM**
Solubility	0.966	0.897	0.949
Membrane Localization	0.576	0.584	0.541

Table 3: Performance comparison (AUROC) of embeddings in predicting physicochemical properties of MeMDLM test set sequences.

4 Conclusion

In this work, we introduce MeMDLM, a fine-tuned encoder that *de novo* generates and provides feature-rich representations of membrane protein sequences. By pre-training and fine-tuning the rich embedding space of the ESM-2-150M pLM on membrane protein sequences, we *de novo* generate membrane protein sequences with TM-character similar to experimentally annotated membrane proteins. We further apply our generative capabilities to scaffold soluble and TM domains of natural membrane protein sequences with lower pseudo perplexity compared to SOTA methods while maintaining the physicochemical features of membrane proteins. This indicates the potential use of MeMDLM-designed membrane proteins for applications in drug discovery where designing stable and functional membrane proteins is critical for therapeutic targets, biosensors, selective channels, and enzymes.

Still, current *in silico* structural prediction methods such as AlphaFold3 are constrained for certain protein classes (23) due to the complex interactions between membrane proteins, the lipid bilayer,

and the bulk aqueous phase. While our model generates membrane proteins with significant TM-character and relevant structural features such as alpha-helical bundles, accurately assessing binding and docking for drug development purposes is crucial. To address this, we are building experimental validation platforms to quantify binding affinity and structural stability.

In summary, MeMDLM provides a promising platform for designing novel, realistic membrane proteins. With MeMDLM, we introduce a new dimension to protein research by enriching encoder-only pLMs with powerful generative capabilities. MeMDLM further motivates future usage of training BERT-style models with the MDLM objective for *de novo* protein sequence design. Future work will focus on integrating experimental assays to screen *de novo* membrane protein sequences as we aim to produce scalable and effective tools to facilitate drug discovery.

Acknowledgements

We thank Andrei Lomize for his technical assistance with the PPM software, Duke Compute Cluster and Mark III Systems for providing computational support for this project, and the Kuleshov Group at Cornell Tech, specifically Subham Sahoo and Yair Schiff, with technical assistance for this project.

References

[1] M. Jelokhani-Niaraki, "Membrane proteins: structure, function and motion," 2022.

[2] R. R. Sanganna Gari, J. J. Montalvo-Acosta, G. R. Heath, Y. Jiang, X. Gao, C. M. Nimigean, C. Chipot, and S. Scheuring, "Correlation of membrane protein conformational and functional dynamics," *Nature Communications*, vol. 12, no. 1, p. 4363, 2021.

[3] J. Wang, S. Lisanza, D. Juergens, D. Tischer, J. L. Watson, K. M. Castro, R. Ragotte, A. Saragovi, L. F. Milles, M. Baek, *et al.*, "Scaffolding protein functional sites using deep learning," *Science*, vol. 377, no. 6604, pp. 387–394, 2022.

[4] H. Yin, J. S. Slusky, B. W. Berger, R. S. Walters, G. Vilaire, R. I. Litvinov, J. D. Lear, G. A. Caputo, J. S. Bennett, and W. F. DeGrado, "Computational design of peptides that target transmembrane helices," *Science*, vol. 315, p. 1817–1822, Mar. 2007.

[5] A. Elazar, N. J. Chandler, A. S. Davey, J. Y. Weinstein, J. V. Nguyen, R. Trenker, R. S. Cross, M. R. Jenkins, M. J. Call, M. E. Call, and S. J. Fleishman, "De novo-designed transmembrane domains tune engineered receptor functions," *eLife*, vol. 11, May 2022.

[6] J. Zhu and P. Lu, "Computational design of transmembrane proteins," *Current Opinion in Structural Biology*, vol. 74, p. 102381, June 2022.

[7] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, *et al.*, "Evolutionary-scale prediction of atomic-level protein structure with a language model," *Science*, vol. 379, no. 6637, pp. 1123–1130, 2023.

[8] E. Ahmed, M. Heinzinger, C. Dallago, G. Rihawi, Y. Wang, L. Jones, T. Gibbs, T. Feher, C. Angerer, S. Martin, *et al.*, "Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing," *bioRxiv*, 2020.

[9] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[10] S. Alamdari, N. Thakkar, R. van den Berg, A. Lu, N. Fusi, A. Amini, and K. Yang, "Protein generation with evolutionary diffusion: Sequence is all you need. biorxiv 2023," *Google Scholar*.

[11] T. Chen, P. Vure, R. Pulugurta, and P. Chatterjee, "Amp-diffusion: Integrating latent diffusion with protein language models for antimicrobial peptide generation," Mar. 2024.

[12] N. Ferruz, S. Schmidt, and B. Höcker, "Protgpt2 is a deep unsupervised language model for protein design," *Nature communications*, vol. 13, no. 1, p. 4348, 2022.

[13] T. Chen, M. Dumas, R. Watson, S. Vincoff, C. Peng, L. Zhao, L. Hong, S. Pertsemlidis, M. Shaepers-Cheu, T. Z. Wang, D. Srijay, C. Monticello, P. Vure, R. Pulugurta, K. Kholina, S. Goel, M. P. DeLisa, R. Truant, H. C. Aguilar, and P. Chatterjee, "Pepmlm: Target sequence-conditioned generation of therapeutic peptide binders via span masked language modeling," 2023.

[14] T. Hayes, R. Rao, H. Akin, N. J. Sofroniew, D. Oktay, Z. Lin, R. Verkuil, V. Q. Tran, J. Deaton, M. Wiggert, R. Badkundri, I. Shafkat, J. Gong, A. Derry, R. S. Molina, N. Thomas, Y. Khan, C. Mishra, C. Kim, L. J. Bartie, M. Nemeth, P. D. Hsu, T. Sercu, S. Candido, and A. Rives, "Simulating 500 million years of evolution with a language model," July 2024.

[15] S. S. Sahoo, M. Arriola, Y. Schiff, A. Gokaslan, E. Marroquin, J. T. Chiu, A. Rush, and V. Kuleshov, "Simple and effective masked diffusion language models," *arXiv preprint arXiv:2406.07524*, 2024.

[16] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[17] D. Kozma, I. Simon, and G. E. Tusnady, "Pdbtm: Protein data bank of transmembrane proteins after 8 years," *Nucleic acids research*, vol. 41, no. D1, pp. D524–D529, 2012.

[18] M. A. Lomize, I. D. Pogozheva, H. Joo, H. I. Mosberg, and A. L. Lomize, "Opm database and ppm web server: resources for positioning of proteins in membranes," *Nucleic acids research*, vol. 40, no. D1, pp. D370–D376, 2012.

[19] T. D. Newport, M. S. P. Sansom, and P. J. Stansfeld, "The memprotmd database: a resource for membrane-embedded protein structures and their lipid interactions," *Nucleic acids research*, vol. 47, no. D1, pp. D390–D397, 2019.

[20] L. Käll, A. Krogh, and E. L. Sonnhammer, "A combined transmembrane topology and signal peptide prediction method," *Journal of molecular biology*, vol. 338, no. 5, pp. 1027–1036, 2004.

[21] L. Käll, A. Krogh, and E. L. Sonnhammer, "Advantages of combined transmembrane topology and signal peptide prediction—the phobius web server," *Nucleic acids research*, vol. 35, no. suppl_2, pp. W429–W432, 2007.

[22] V. Thumuluri, J. J. Almagro Armenteros, A. R. Johansen, H. Nielsen, and O. Winther, "Deeploc 2.0: multi-label subcellular localization prediction using protein language models," *Nucleic acids research*, vol. 50, no. W1, pp. W228–W234, 2022.

[23] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, *et al.*, "Accurate structure prediction of biomolecular interactions with alphafold 3," *Nature*, pp. 1–3, 2024.

[24] S.-Q. Zhang, D. W. Kulp, C. A. Schramm, M. Mravic, I. Samish, and W. F. DeGrado, "The membrane-and soluble-protein helix-helix interactome: similar geometry via different interactions," *Structure*, vol. 23, no. 3, pp. 527–541, 2015.

5 Supplementary Material

5.1 De Novo Generation Visualizations

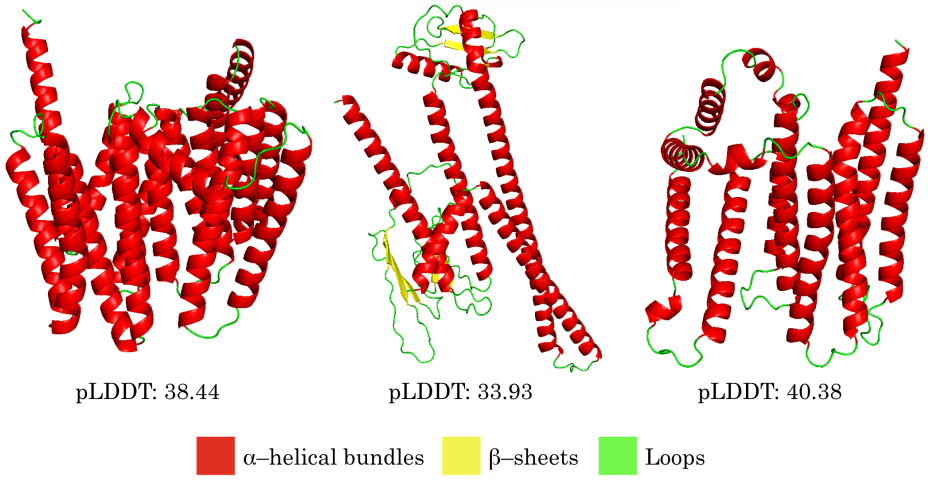

Figure 4: AlphaFold3-predicted structures of *de novo* MeMDLM-generated protein sequences.

5.2 Perplexity and Loss During pLM Training and Evaluation

	Loss			Perplexity		
	Train	**Validation**	**Test**	**Train**	**Validation**	**Test**
ESM-MLM	0.072	0.072	0.072	1.074	1.074	1.074
ProtGPT2	1.585	–	3.392	4.879	–	29.730
MeMDLM	0.695	2.479	2.230	2.002	7.722	9.285

Table 4: Loss and perplexity comparison across models

5.3 MeMDLM Training

MeMDLM was pre-trained for 7 epochs and fine-tuned for 60 epochs on 4xA6000 NVIDIA GPUs each with 48 GB of VRAM. A batch size of 16, learning rate of 3e-4 with linear warmup of 2,500 steps, and the AdamW optimizer with a weight decay of 0.075 was used. All model training and implementation was done with Python 3.10 and PyTorch 2.2.2.

5.4 Masked Language Model (MLM)

ESM-MLM is a fine-tuned encoder that produces membrane-aware protein sequence embedding used as a baseline comparison for the MDLM training task. We trained a MLM head on top of ESM-2-150M using membrane protein sequences to force comprehension of membrane protein properties. 15% of amino acid tokens were randomly masked and passed into ESM-2-150M to retrieve their output embeddings. The MLM loss function is defined as:

$$L_{\text{MLM}} = -\sum_{i \in \mathcal{M}} \log P(x_i | x_{\setminus \mathcal{M}}) \tag{5}$$

where $\mathcal{M}$ represents the set of masked positions in the input sequence, x_i is the true amino acid token at position i, and $x_{\setminus\mathcal{M}}$ denotes the sequence with the masked tokens excluded.

During training, we unfroze the key, query, and value weights in the attention heads of the final three encoder layers. With this training recipe, we augment the pre-existing ESM-2-150M latent space with physicochemical properties of membrane proteins without overfitting on the new sequences.

ESM-MLM was trained on one NVIDIA A6000 GPU with 48 GB of VRAM over 10 epochs with a batch size of 2 and a learning rate of 5e-5. The Adam optimizer was used with no weight decay. Membrane protein sequences were padded to match the length of the sequence.

5.5 Physicochemical Property Prediction

Solubility Prediction We first predicted TM and soluble residues, a hallmark characteristic of membrane protein sequences. We utilized each embedding type as inputs to train a two-layer perceptron classifier in PyTorch that minimized the standard binary cross-entropy (BCE) loss to compute the probability that each residue in the sequence is either soluble (probability < 0.5, class 0) or TM (probability > 0.5, class 1). The BCE loss is formally defined as: $\text{BCE}(y, \hat{y}) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}))$

Membrane Localization Prediction Proteins originating from the endomembrane system and localizing in the plasma membrane differ in conformation and function from those in the cytosol and other cellular organelles. We predicted the subcellular localization of protein sequences by using each embedding type to train a XGBoost classifier that minimized the standard BCE loss (above) to compute the probability that a protein sequence localizes in the plasma membrane (probability > 0.5, class 1) or in other regions (probability < 0.5, class 0).

5.6 Data Curation

Pre-training We queried the UniRef50 database for a random set of 100,000 unique protein sequences containing only the 20 natural amino acids; we only considered sequences shorter than 1,024 residues due to GPU memory limits, and shorter sequences were padded to this maximal length. Sequences were split using the MMSeqs2 easy clustering module with a minimum sequence identity of 30% and a coverage threshold of 50%. The resulting clusters were split to a 80-10-10 ratio into the training set (80,231 sequences, 80.23%), validation set (9,904 sequences, 9.90%), and the testing set (9,865 sequences, 9.87%).

Fine-tuning Bioassembly structures from X-ray scattering or electron microscopy with better than 3.5 Å resolution, annotated by PDBTM1, mpstruc2, OPM3, or MemProtMD4, were used to curate membrane protein sequences for fine-tuning. *de novo* designed membrane proteins were added manually to the database. The proteins were culled at 100% sequence identity and 30% sequence identity to result in a non-redundant set and a sequence-diverse set, respectively. Integral membrane residues, defined as residues with at least one atom within the bilayer, were parsed from the resulting bioassembly structures using the membrane boundaries predicted by PPM 3.05. From the dataset of integral membrane residues, only structures with at least one TM chain spanning the entire membrane bilayer were included in the dataset. Additionally, chains without integral membrane residues were removed from the structure. All peripheral membrane proteins, defined as proteins with no TM chain, were filtered out. The remaining 9,329 TM sequences were then split using the MMSeqs2 easy clustering module with a minimum sequence identity of 80% and a coverage threshold of 50%. The resulting clusters were split to an 80-10-10 ratio into the training set (7,632 sequences, 81.81%), validation set (770 sequences, 8.25%), and the testing set (927 sequences, 9.94%).

5.7 Benchmarking Data Curation

Solubility We leveraged the same set of 9,329 membrane sequences from the MeMDLM training dataset to develop a binary classifier that predicts the solubility of each amino acid within a protein sequence. Each sequence was annotated on a per-residue basis, with TM (class 1) and soluble (class 0) labels assigned according to the sequence's uppercase and lowercase residues, respectively. The same training, testing, and validation data splits used to train MeMDLM were also utilized to train and evaluate this classifier.

Membrane Localization We collected 30,020 protein sequences from DeepLoc 2.0 to build a binary classifier that predicts a protein sequence's cellular localization. The authors of the dataset provided a multi-label label for each sequence indicating its localization(s). We used the authors' provided data splits, with training sequences having 11 labels and testing sequences having 8 labels.

5.8 Protein Sequence Generation

ProtGPT2 Prepared sequences—split to contain 60 amino acids per line with beginning- and end-of-sequence tags—were passed into the run_clm.py script (https://huggingface.co/nferruz/ProtGPT2) to fine-tune the pre-trained ProtGPT2 pLM. Fine-tuning was performed over 100 epochs with a learning rate of 3e-4 and batch size of 2, calculating training loss at every step as the negative log-likelihood loss between logits and labels. The fine-tuned model was used to generate 100 *de novo* membrane protein sequences.

MeMDLM We generated 100 *de novo* protein sequences of random lengths by inputting sequences consisting of only `<mask>` tokens into the forward pass of MeMDLM. Next, we scaffolded around TM or soluble motifs by masking specific residues; partially masked sequences were passed through the model for generation. We evaluated MeMDLM against EvoDiff's reconstruction quality.