Token-Level Guided Discrete Diffusion For Membrane Protein Design

Shrey Goel,[1] Peregrine Schray,[2] Yinuo Zhang,[3] Sophia Vincoff,[4] Huong T. Kratochvil,[2] **Pranam Chatterjee**[4,†]

[1]Duke · [2]The University of North Carolina at Chapel Hill · [3]DukeNUS Medical School · [4]Penn

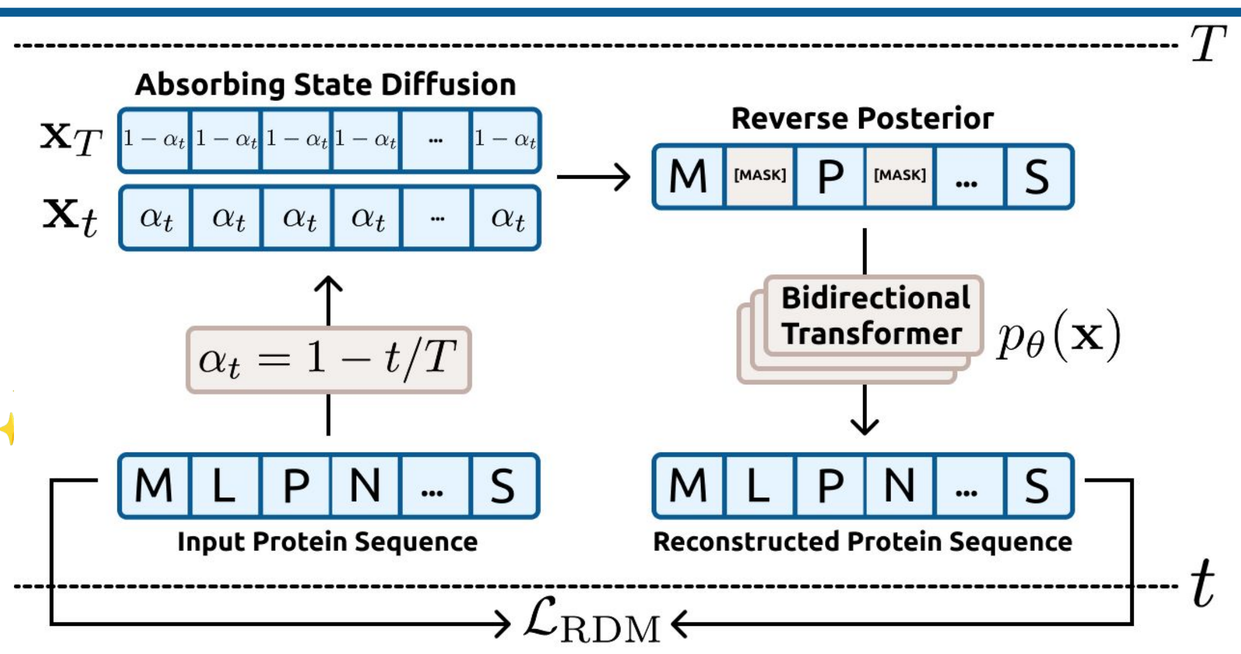
[†]Correspondence to pranam@seas.upenn.edu

Motivation

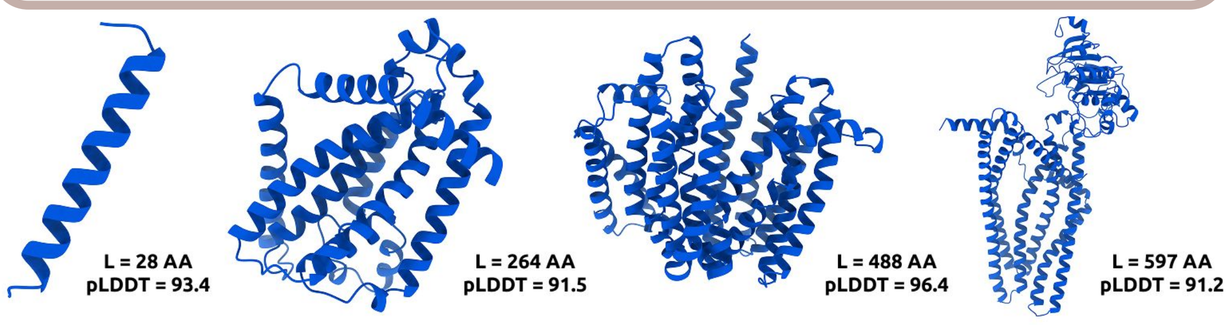
Membrane proteins regulate signaling and transport, rendering them **prime targets for therapeutic intervention**. However, they are challenging to design:

- Structure-based membrane protein design methods are severely limited by the **scarcity of high-resolution membrane protein structures** (~1% of the PDB).
- Autoregressive language models **struggle to capture the long-range** dependencies of interleaved transmembrane and soluble regions.
- Existing diffusion guidance strategies **lack the token-level precision** required to solubilize exposed loops while strictly preserving critical transmembrane domains.

Membrane Diffusion LM Architecture

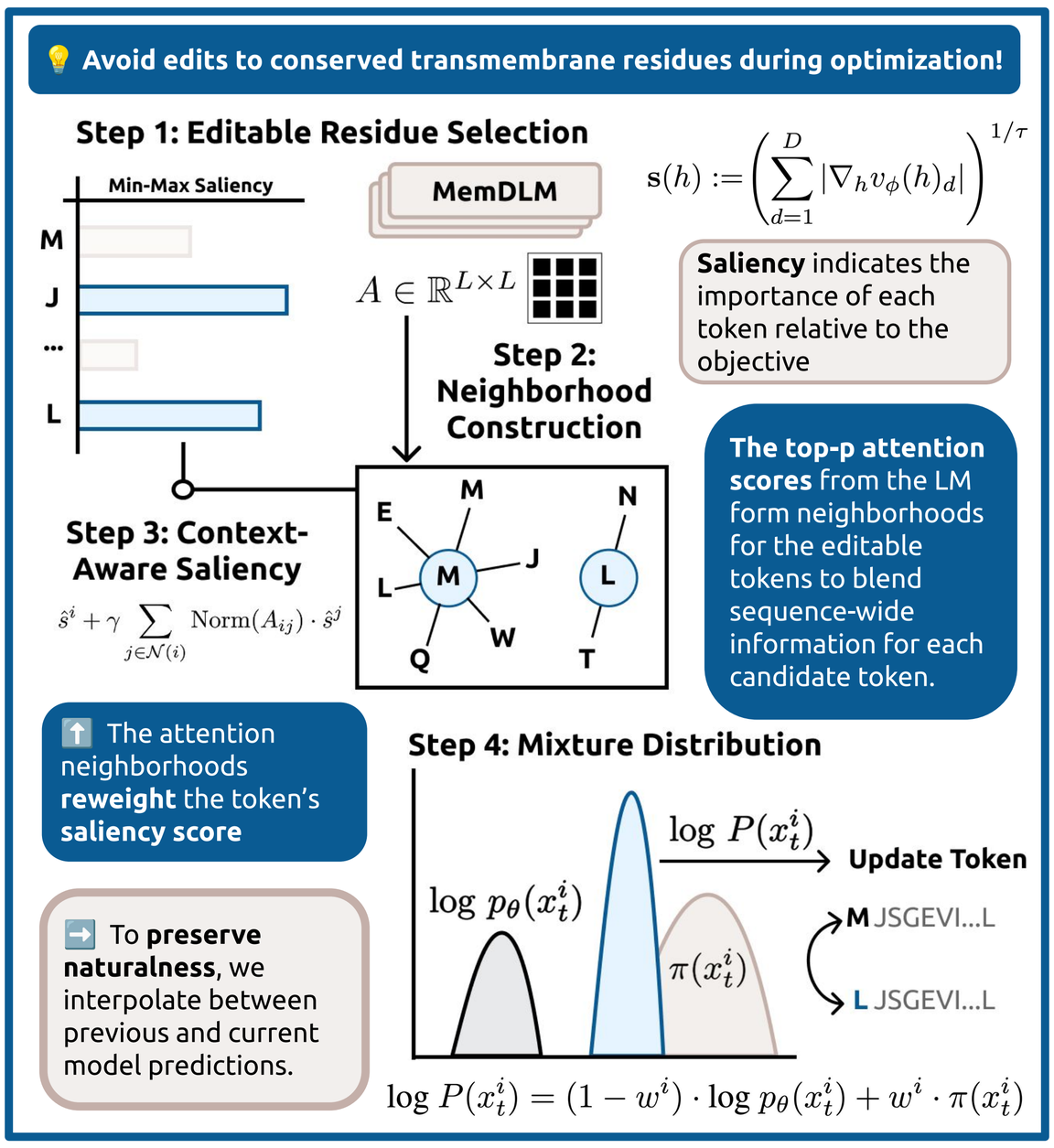


- We fine-tune the EvoFlow protein LM on a **reparameterized diffusion model (RDM) training objective**, forcing our model to learn denoising over higher masking rates
- During inference, the trained MemDLM can ***de novo* generate high-quality, plausible membrane protein sequences** from a fully masked prior



Per-Token Discrete Classifier Guidance

💡 **Avoid edits to conserved transmembrane residues during optimization!**

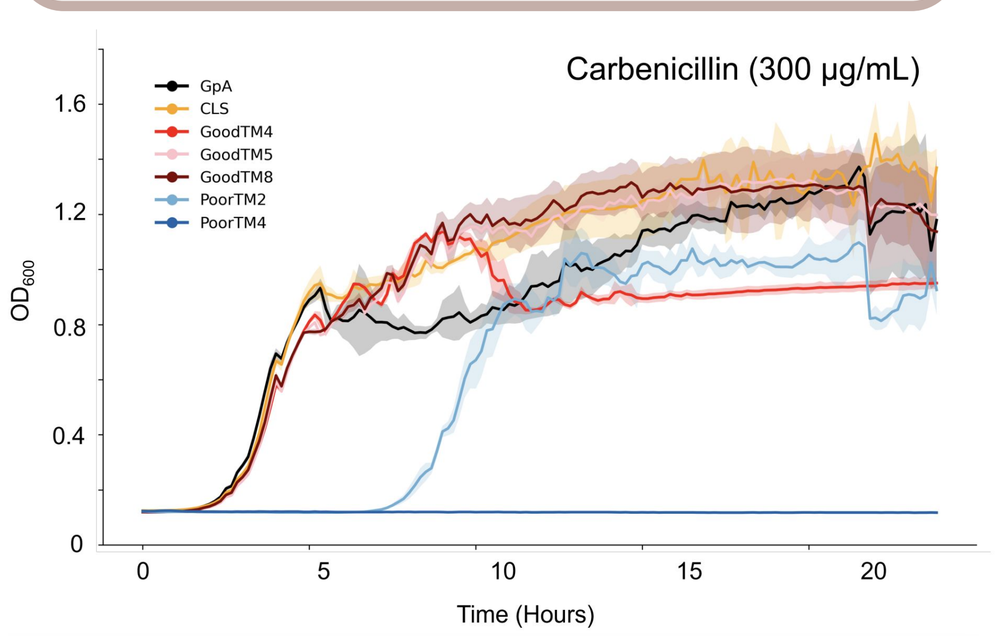


Unconditional Sequence Generation

	pLDDT (↑)	TMRD	PPL (↓)	Entropy (↑)
Test Set	76.637±10.676	0.294±0.219	5.707±3.435	3.918±0.253
ProGen2	54.998±19.235	0.048±0.153	126.646±1415.166	2.622±1.290
DPLM	62.318±20.669	0.310±0.264	**6.323**±10.317	3.179±0.812
MemDLM	**67.410**±14.828	**0.311**±0.250	6.344±3.278	**3.743**±0.326

⬆️ **MemDLM** more effectively **captured the underlying distribution** of membrane protein sequences

⬇️ In the TOXCAT β–lactamase growth assay, MemDLM-generated sequences exhibit **growth kinetics similar to natural** membrane proteins



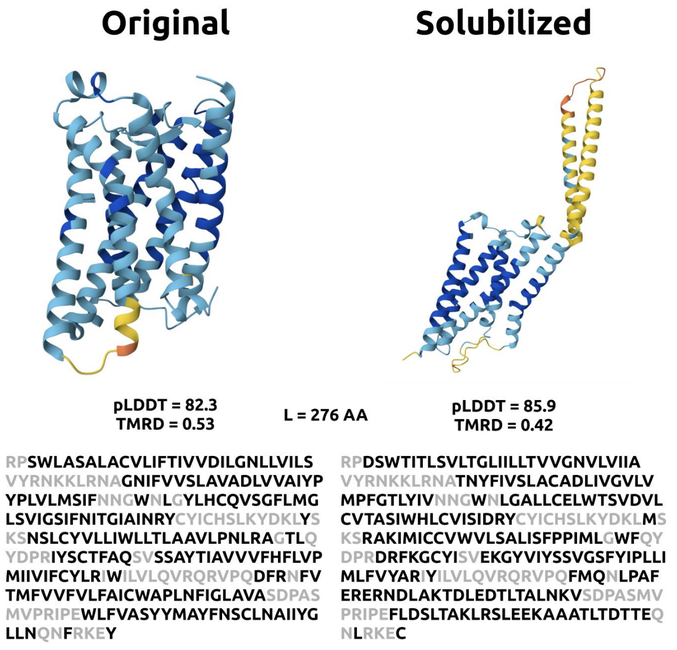
Motif Scaffolding & Representation Learning

Model	Solubility (↑)	Membrane Localization (↑)
ESM-2-650M	0.9383	0.6011
Fine-Tuned ESM-2	0.9375	0.6000
MemDLM	0.9375	0.5964

	Motif	pLDDT (↑)	PPL (↓)	BLOSUM (↑)	Entropy (↑)
Test Set	Insol	76.637±10.676	5.707±3.435	–	3.918±0.253
	Sol	76.637±10.676	5.707±3.435	–	3.918±0.253
EvoDiff	Insol	**64.058**±19.229	9.841±4.091	2.176±1.587	3.841±0.268
	Sol	64.036±19.145	4.632±3.271	-0.188±1.134	**3.841**±0.268
MemDLM	Insol	62.762±21.212	**8.748**±14.777	**2.964**±1.559	**3.876**±0.341
	Sol	**70.112**±16.912	**3.242**±2.362	**0.512**±1.556	3.803±0.321

MemDLM **scaffolds** over transmembrane and soluble motifs and **encodes important membrane protein features** in its learned latent space

Solubilization & Scaffold Retention



Metric	Test Set	MemDLM
pLDDT (↑)	76.637±10.676	62.979±17.906
TMRD (↓)	0.294±0.219	0.181±0.192
PPL (↓)	5.707±3.435	8.472±4.879
BLOSUM (↑)	–	0.495±2.346
Entropy (↑)	3.918±0.253	3.870±0.268

- PET solubilizes targeted transmembrane residues while retaining a sequence scaffold
- Solubilized sequences have positive BLOSUM scores, indicating PET's edits are evolutionarily conserved

Conclusions

Our results establish MemDLM as the first experimentally validated, classifier-guided discrete diffusion framework for membrane protein design, operating without structural templates. Per-Token Guidance (PET) enables solubilization while preserving functional TM scaffolds, offering controllability unattainable by traditional diffusion guidance methods. Benchmarking against SOTA models (DPLM, EvoDiff), MemDLM demonstrates superior motif scaffolding and *de novo* generation quality. Finally, the TOXCAT-β-lactamase assay shows our designs successfully insert into membranes , establishing MemDLM as a versatile platform for engineering programmable membrane protein therapeutics.

Acknowledgments

We would like to thank the entire Chatterjee Lab for their guidance and support. We further thank Mark III Systems for hardware support and Vishrut Thoutam and Samuel Scheinbach for technical help. We finally thank the Kuleshov Group at Cornell, specifically Aaron Gokaslan, Edgar Mariano Marroquin, Subham Sahoo and Yair Schiff, for reviewing earlier versions of MemDLM.

Read the full paper!

Programmable Biology Group

ChatterjeeLab