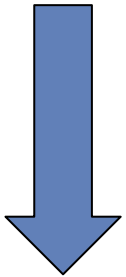
"Binds single-stranded DNA during recombination"



MKTLTFLAAAAAAAAMLGGCIVAP
PPHRHGPPPPPPPHWRGPPDHGR
GPWHR

MSPVQLSPLLLLGLAVIVLASAL
WLSRRLREHGFGPEGPPRRDED
DGARR

MAYGFFLPYSLLLTPYSLFPTPY
SLLPTPYSLLPTPYSLFPTPYSLL
PTP

Classifier-Guided Protein Generation via Natural Language Input

Michaela Harris, Sarah Jiang,
HyoJoo Kim, Yeonseo Kim

Sophia Vincoff, Benjamin Perry,
Zachary Quinn, Dr. Pranam Chatterjee

Motivation

Natural language captures **nuance** not encapsulated in other information types

Discrete diffusion is more accurate at generating **sequence**-based outputs

Other models have high **computational costs**

Implementation

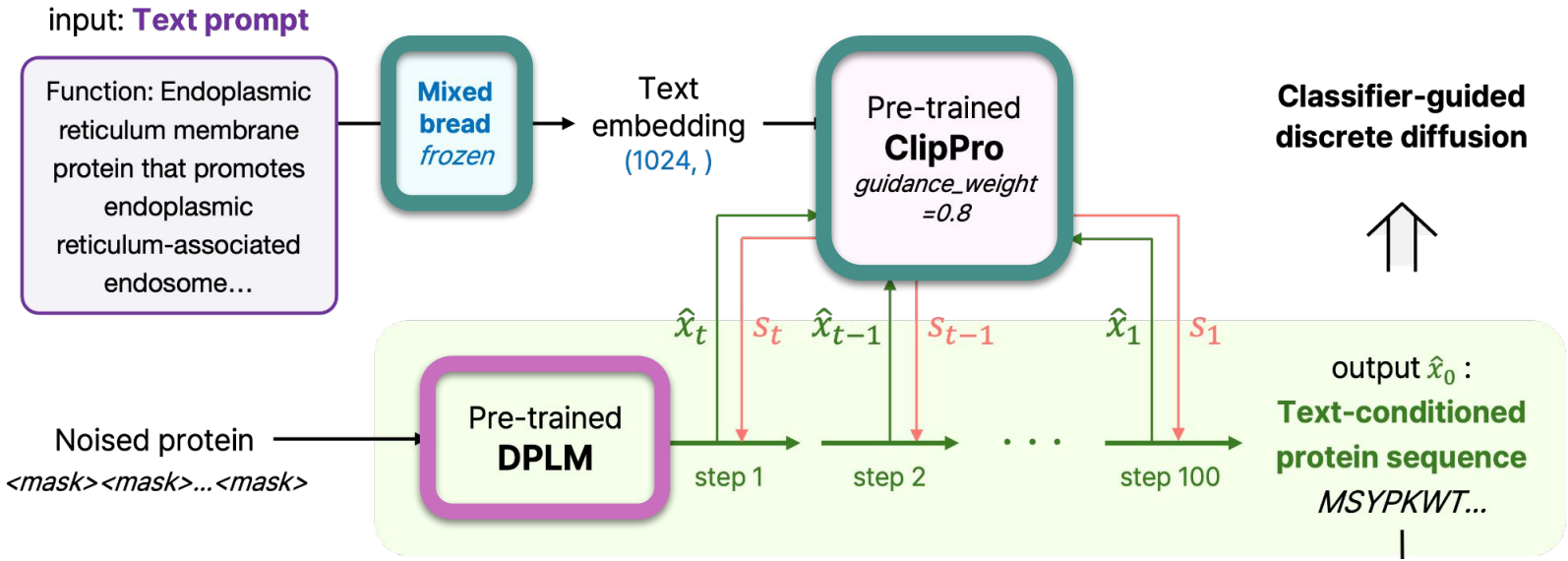
Focused on **outcomes** between **textual embeddings**
- *PubMedBERT vs Mixedbread*

ClipPro classifier task evaluated **functional description** alignment with **sequence**

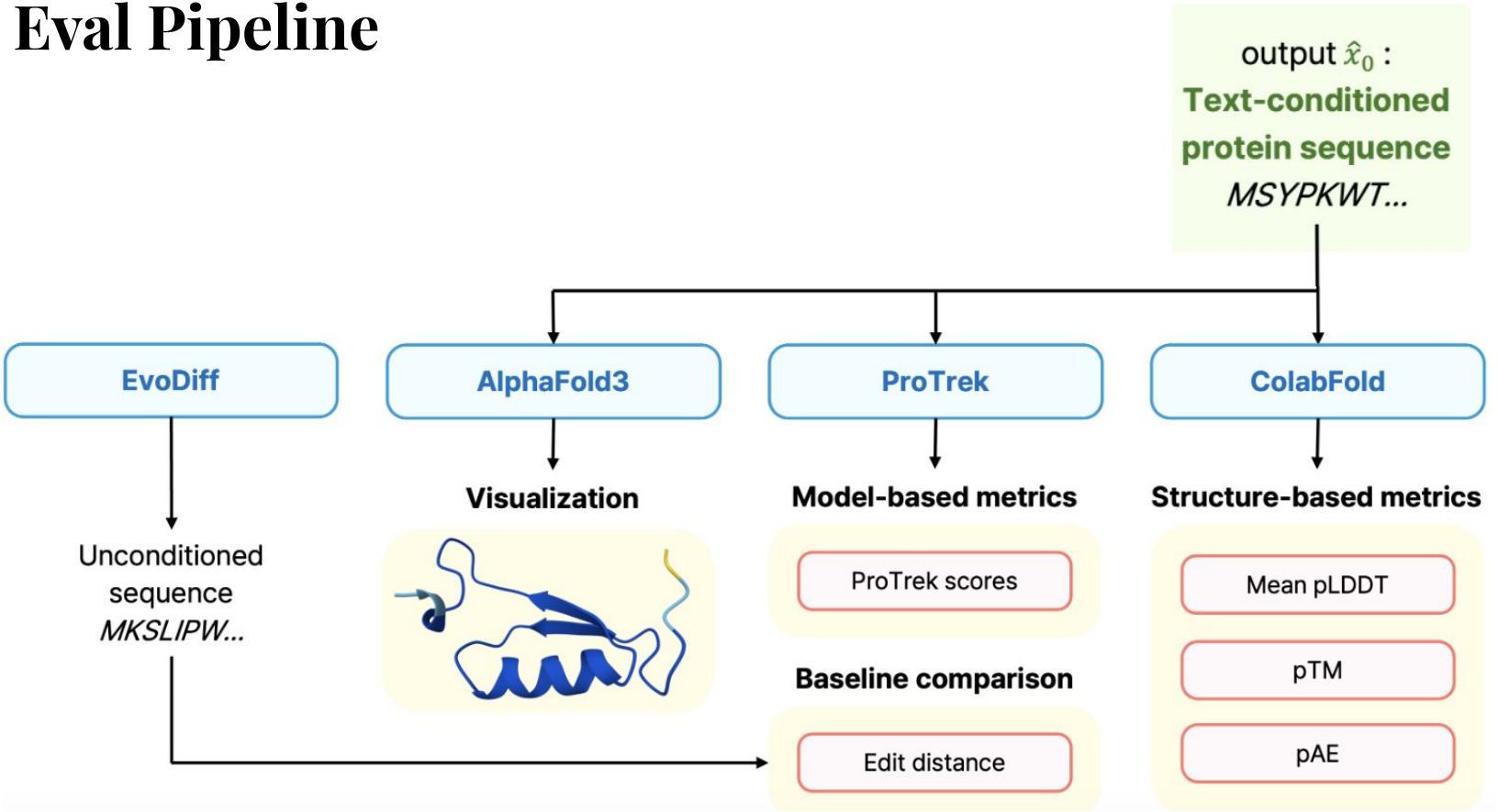
Used **DPLM** for generating novel sequences

Leverage **pre-trained** models

Model Pipeline



Eval Pipeline

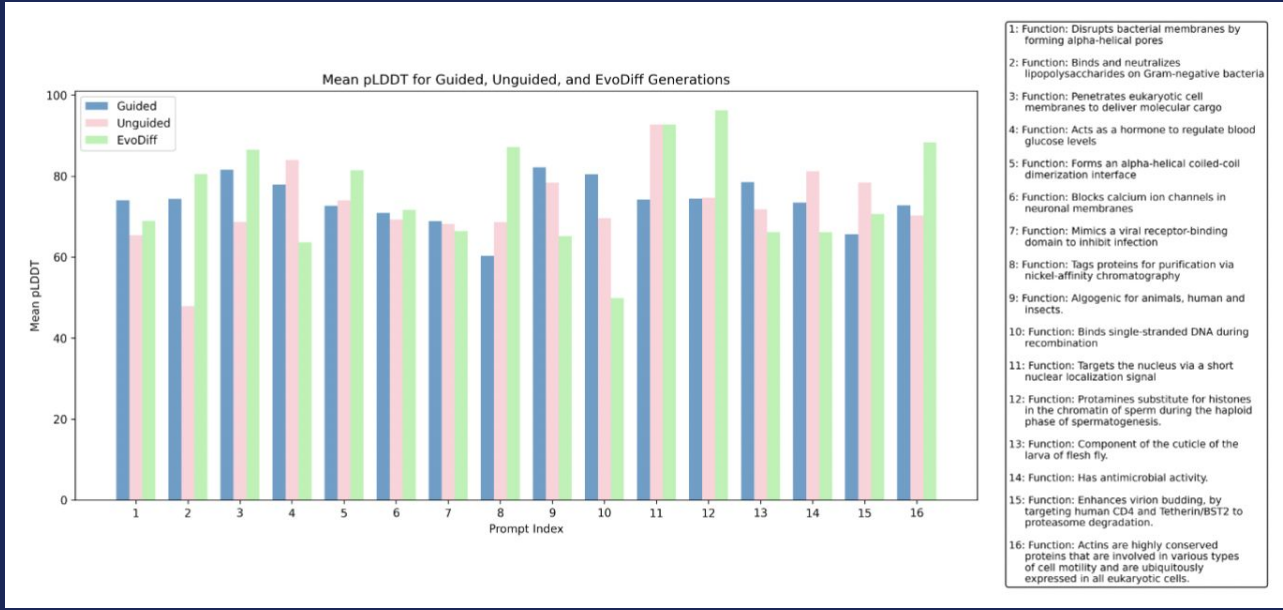


Key Results

Condition	Mean pLDDT	Mean pTM	Mean pAE	Shannon Entropy	Levenshtein Similarity
Guided	73.90	0.3923	12.61	2.6824	0.2088
Random Guidance	70.13	0.3421	3.1201	3.2022	0.1011
Unguided	72.70	0.3881	11.76	2.6619	0.2040
EvoDiff	75.11	0.3694	12.45	3.3009	–

- Guided model behaved comparably to EvoDiff for pLDDT and pAE
- Random guidance consistently had the worst outcomes (except shannon entropy)
- Guided model generally outperformed un and randomly guided implementations

Ablation Study



- Guided model consistently outperformed unguided model in pLDDT and was comparable to EvoDiff across a variety of prompts

Conclusions

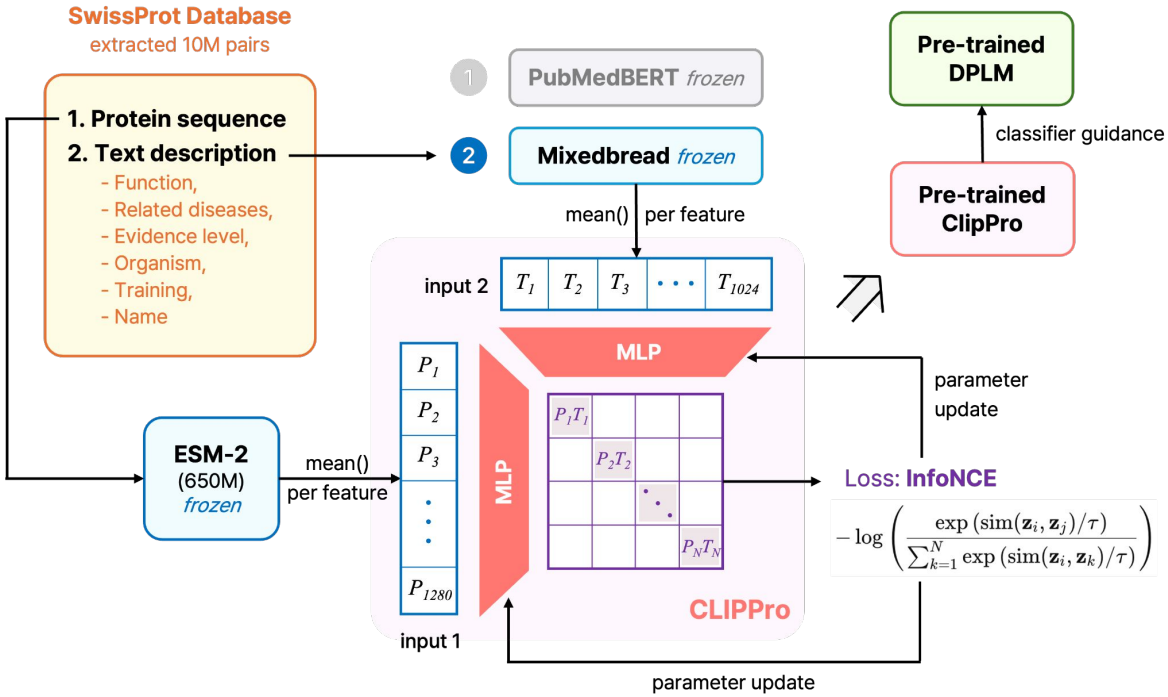
- Guidance must be semantically meaningful to improve generation quality
- Mixedbread text embeddings outperformed the more commonly used PubMedBERT embeddings

- Future work should focus on increasing length of generated sequences and further enhancing language-sequence relationships

- Natural-language-only informed models can generate sequences that perform comparably to gold standard models

Thank you!

Supplementary Slides: Training



Supplementary Slides: AlphaFold + pAE Heatmaps