LucidProts: Controlled Protein Design using Diffusion Language Models

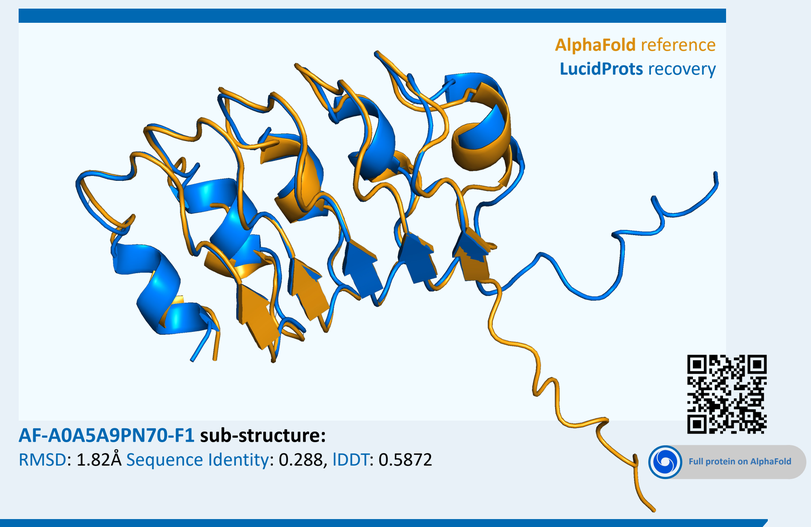
Adrian **Henkel***, Kyra **Erckert***, Burkhard **Rost**, Michael **Heinzinger***

Department for Bioinformatics | Computational Biology, Department of Informatics | **T**echnical **U**niversity of **M**unich

Contact: henkel [at] rostlab.org | erckert [at] rostlab.org | mheinzinger [at] rostlab.org

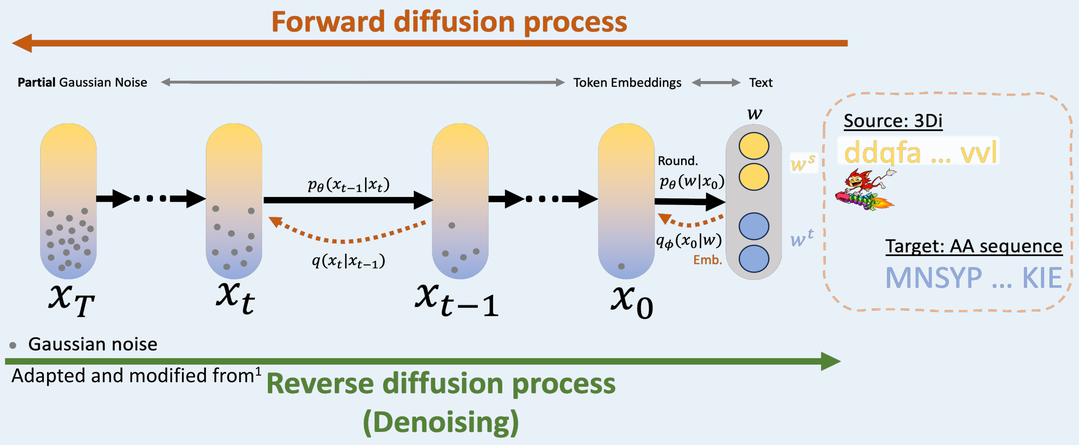
SHORT SUMMARY

LucidProts presents a **proof-of-concept** work proposing the application of **Diffusion Language Models** to effectively address the **inverse folding problem**.



INTRODUCTION

- Generative diffusion models like DALL-E and Imagen are effective in controlled continuous data generation.
- *DiffuSeq*[1] by Gong et al. (2023) employs partial noising and conditional denoising in a classifier-free manner on sequence embeddings.
- LucidProts utilizes *DiffuSeq* with the RoFormer architecture for the inverse folding problem.
- **Related work**: ProteinMPNN[4] uses an encoder-decoder architecture on protein backbone coordinates
- **Goal:** Generation of amino acid sequences with a desired structure



Adapted and modified from[1]

METHODS

Diffuseq:

- Map discrete data into continuous token embeddings
- Sequentially add noise to the target sequence and train a model on removing the noise.
- **Inference:** Start with pure gaussian noise and the condition tag (in this case the embedded 3DI residues)
  → recover the target target embeddings

Evaluation:

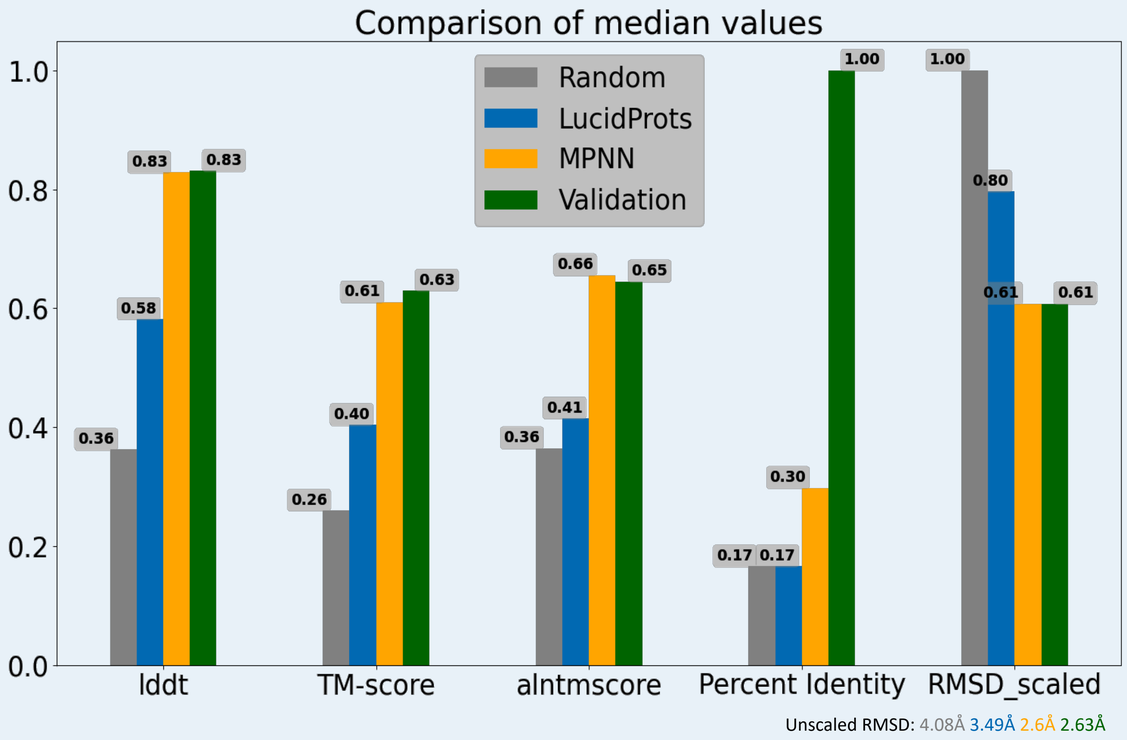
- Predict structure using ESMFold[2]
- Aling structures with Foldseek[4] to ground truth

Data:

- AlphaFold database
- 1M Sequences pLDDT > 0.7 and L >= 30
- 20 most diverse structures per cluster
- Tokenized 3Di[2] structure representation

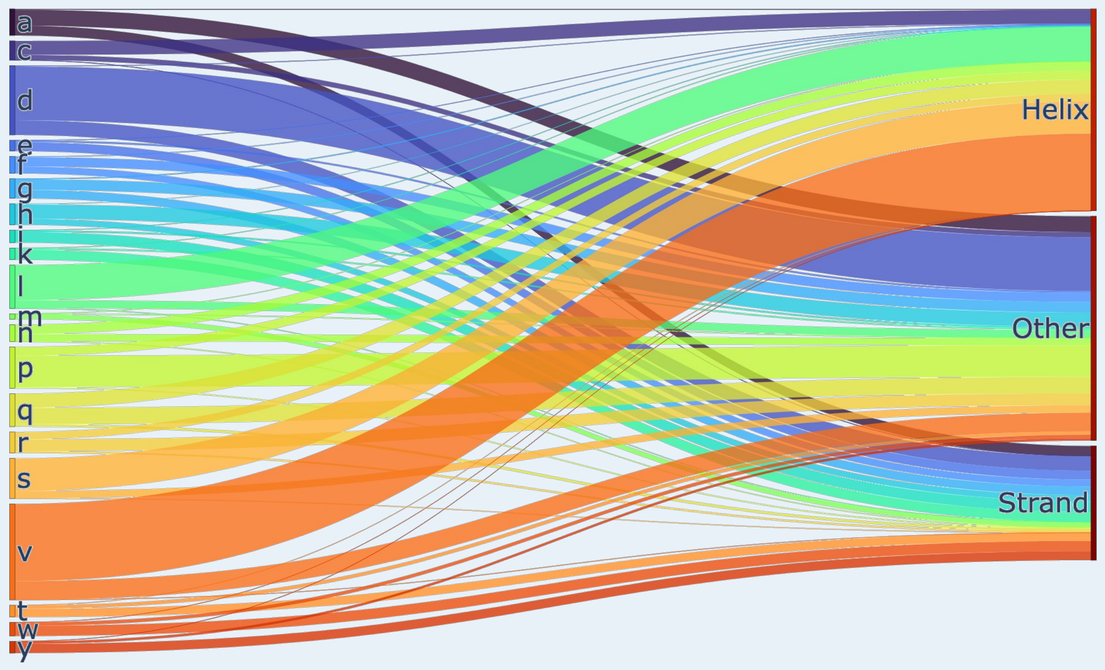
PRELIMINARY RESULTS & DISCUSSION

- **Generation results:**



Unscaled RMSD: 4.08Å 3.49Å 2.6Å 2.63Å

- **Dataset:** 3Di residues tend to have 3-state secondary structure preferences. 12 / 20 residues with >70% prevalence the others with at least >50%



CONCLUSION

- Applying diffusion language models on inverse folding **gives promising results.**
- Generation of smaller sequences is more precise, and sequences were limited to a length of 125
- Inference time can be sped-up by reducing diffusion steps[1]
- **Easy application**: Proof-of-concept suggests that the same concept can be transferred to arbitrary condition tags s.a. function or binding partner

ACKNOWLEDGMENTS:

Shansan Gong (Shanghai AI Lab) for patiently answering my questions and Timothy Karl (TUM) for invaluable help with software and hardware issues.

AVAILABILITY:

Code : https://github.com/mainpyp/Prot-DiffuSeq
Data : https://huggingface.co/datasets/adrianhenkel/lucidprots_full_data
Model : https://huggingface.co/adrianhenkel/lucid-prots-model

REFERENCES:

[1] Gong, Shansan, et al. "Diffuseq: Sequence to sequence text generation with diffusion models." arXiv preprint arXiv:2210.08933 (2022).
[2] van Kempen, Michel, et al. "Fast and accurate protein structure search with Foldseek." Nature Biotechnology (2023): 1-4.
[3] Dauparas, Justas, et al. "Robust deep learning–based protein sequence design using ProteinMPNN." Science 378.6615 (2022): 49-56.
[4] Lin, Zeming, et al. "Evolutionary-scale prediction of atomic-level protein structure with a language model." Science 379.6637 (2023): 1123-1130.

Scan for more information

ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY