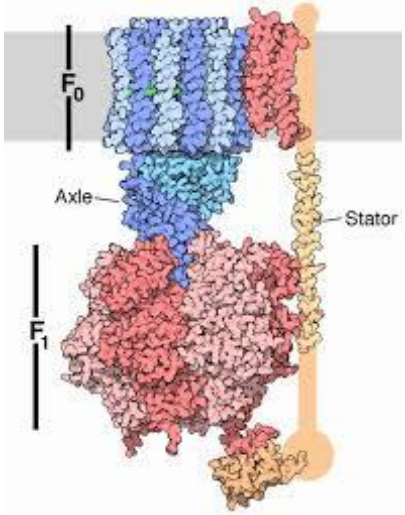
Protein Design and Engineering with Multi-Modal Generative AI Models

Yunan Luo

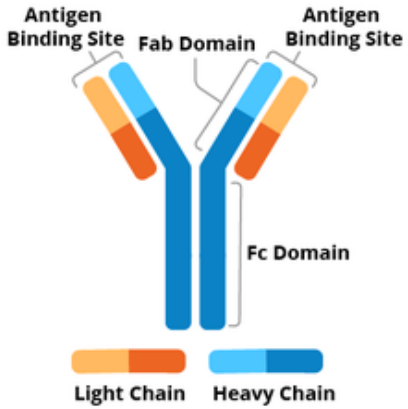
School of Computational Science and Engineering
College of Computing
Georgia Institute of Technology

Proteins perform diverse biological functions



ATP Synthase

Producing energy currency (ATP) of life

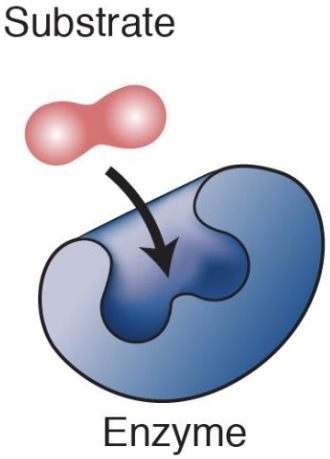


Antibody

Detecting and neutralizing viruses

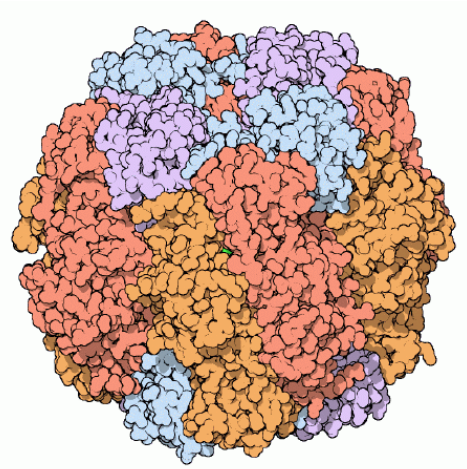
Insulin

Regulating blood sugar



Enzyme

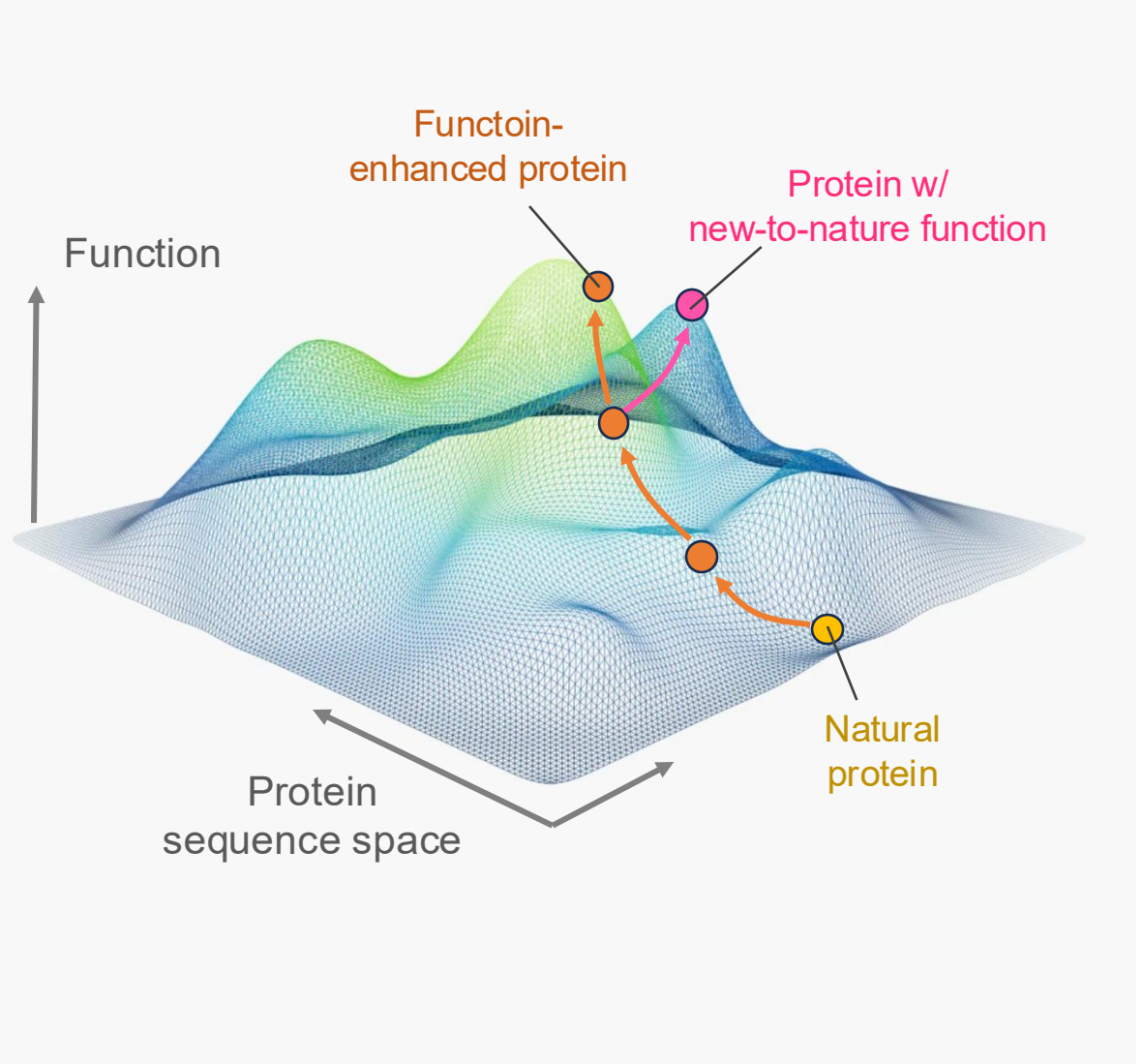
Speeding up chemical reactions



Rubisco

Capturing CO_2 in photosynthesis

Protein design and engineering

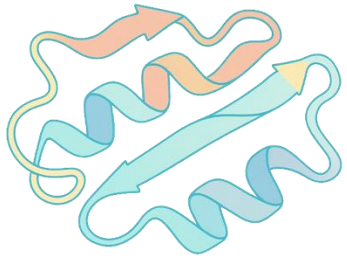


How is a protein's function determined?

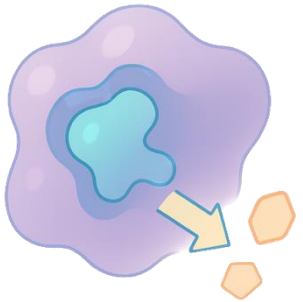
Protein's sequence-structure-function relationship

Protein sequence

Protein structure



Protein function



Nobel Prize in Chemistry 2024
for protein structure prediction (AlphaFold)

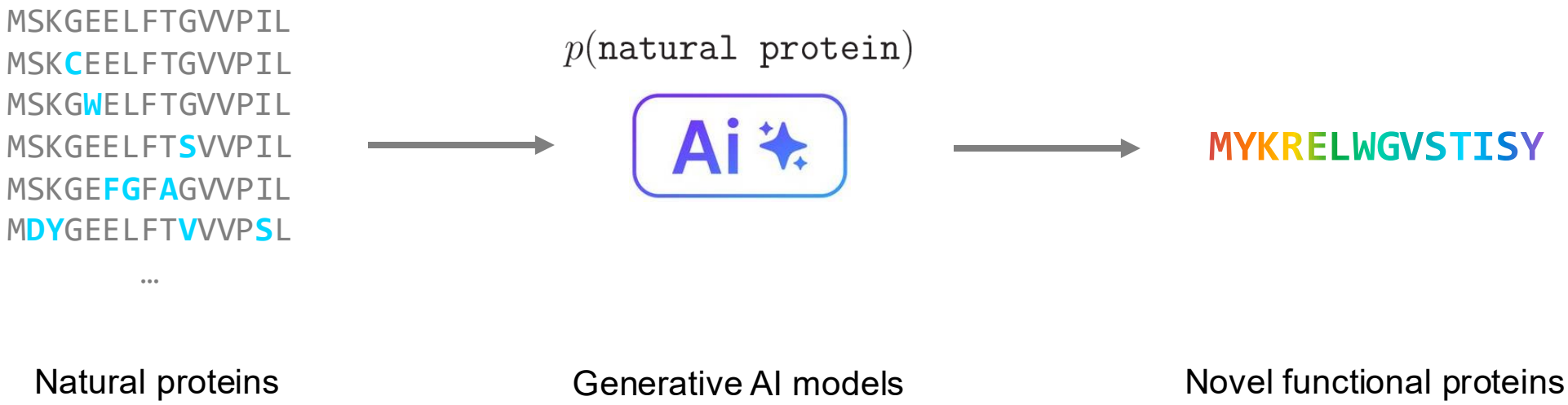
Demis Hassabis John Jumper

This project: *functional protein design*

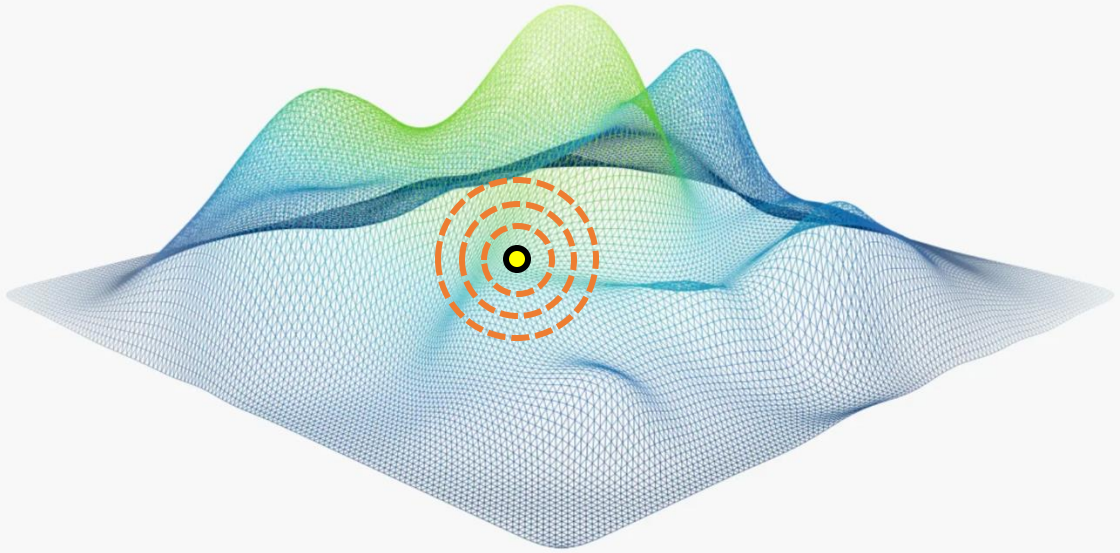
How to design functional proteins?

- Learning from natural protein evolution using GenAI

$$p(\texttt{functional protein}) \approx p(\texttt{natural protein})$$



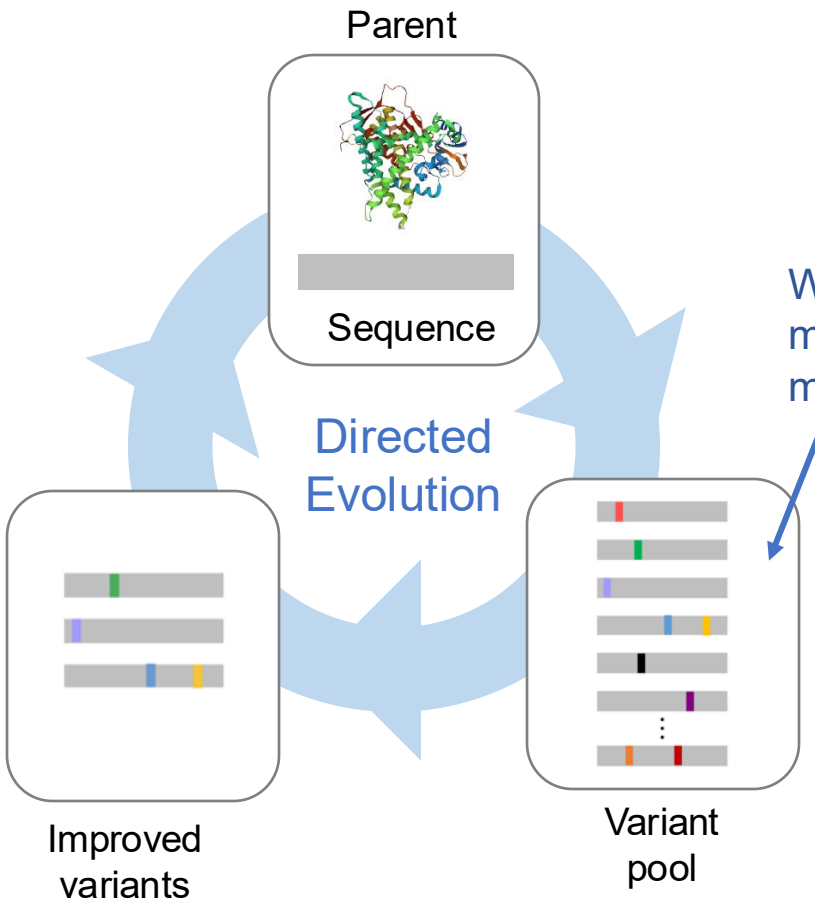
Protein fitness landscape



How to find functional proteins?

Directed evolution for protein engineering
(2018 Nobel Prize in Chemistry)

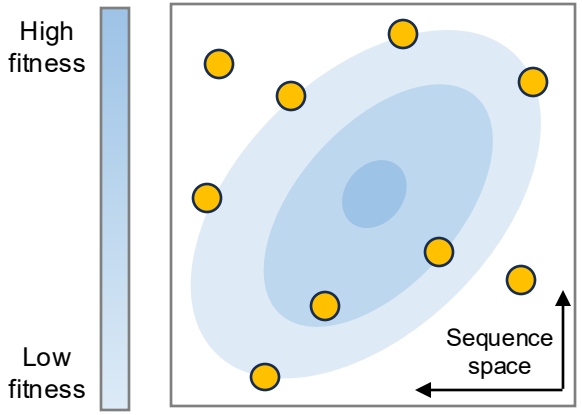
Frances Arnold
(Caltech)



Protein engineering for new-to-nature function

- **Task**: Design functional and diverse proteins
- **Challenge**: No existing fitness data (e.g., because the function is new-to-nature)

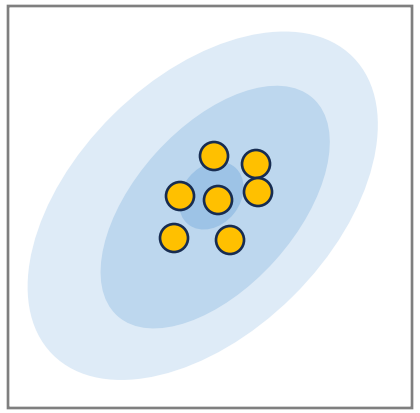
Conventional starting library design
(e.g., NNK library)



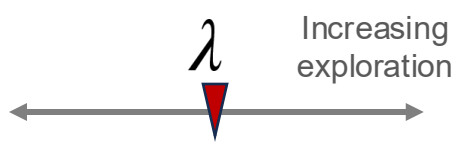
(random search, many are non-functional)

Our approach: Pareto-optimal library design (MODIFY)

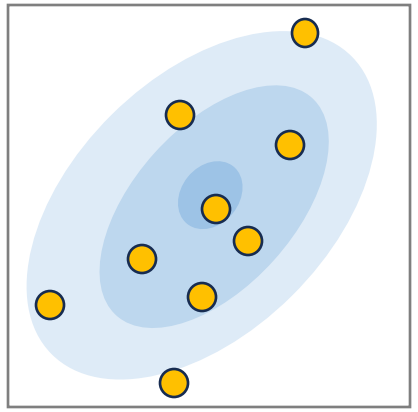
Goal 1: High ***fitness***



(most variants in the library are functional)



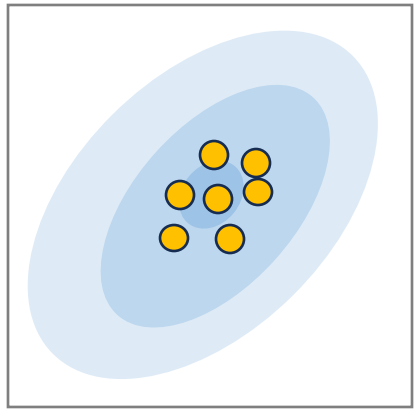
Goal 2: High ***diversity***



(the library contains novel protein variants)

Zero-Shot Fitness Prediction using Foundation Models

Goal 1: High ***fitness***



Q: Cold-start problem. No data to train a supervised fitness predictor.

A: Use pre-trained protein language models to make zero-shot fitness prediction

- ***Natural language model (LM):***

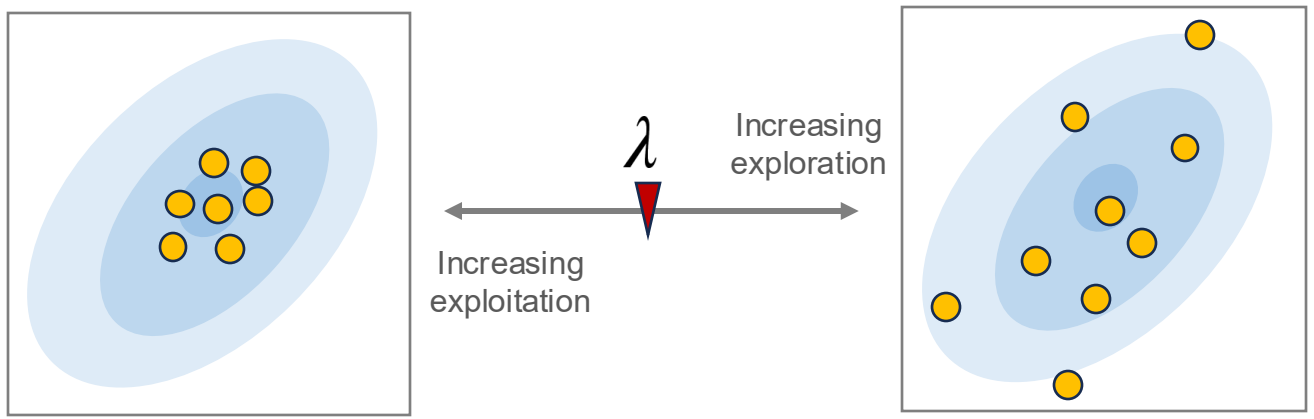
- ***Protein language model (PLM):***

- PLMs are trained on protein sequences we observed in nature
- Evolutionary plausibility correlates with fitness

Characterizing the library diversity

Goal 1: High ***fitness***

Goal 2: High ***diversity***



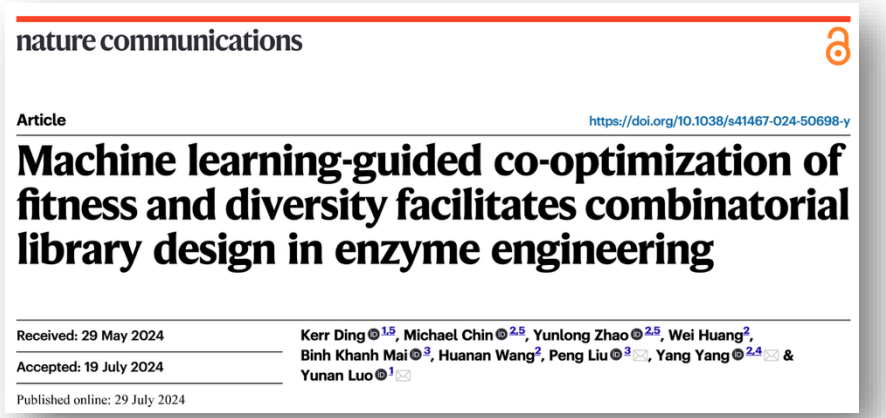
- **Goal:**

$$\max_{p \in \mathcal{P}} \mathbb{E}_{\boldsymbol{x} \sim p(\boldsymbol{x})} \, \text{fitness}\,(\boldsymbol{x}) + \lambda \cdot \, \text{diversity}\,(p)$$

Predicted by protein language model

Quantified by sequence entropy

- A Pareto optimization problem
- Learned a probability distribution over protein sequence


nature communications

Article

https://doi.org/10.1038/s41467-024-50698-y

Machine learning-guided co-optimization of fitness and diversity facilitates combinatorial library design in enzyme engineering

Received: 29 May 2024

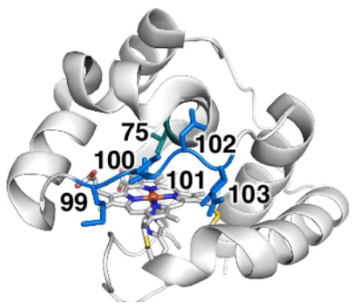
Accepted: 19 July 2024

Published online: 29 July 2024

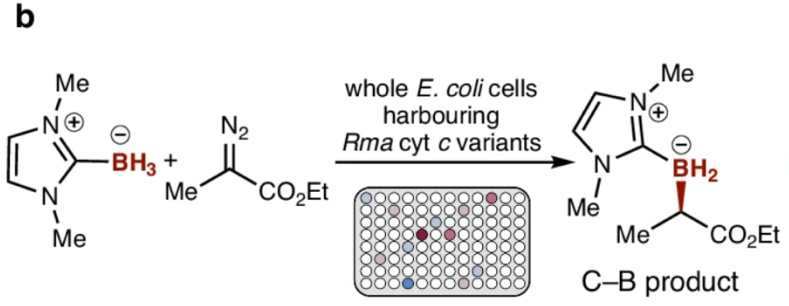
Kerr Ding[1,5], Michael Chin[2,5], Yunlong Zhao[2,5], Wei Huang[2], Binh Khanh Mai[3], Huanan Wang[2], Peng Liu[3] ✉, Yang Yang[2,4] ✉ & Yunan Luo[1] ✉

Ding, K., Chin, M., Zhao, Y. et al. "Machine learning-guided co-optimization of fitness and diversity facilitates combinatorial library design in enzyme engineering," *Nature Communications*, 2024

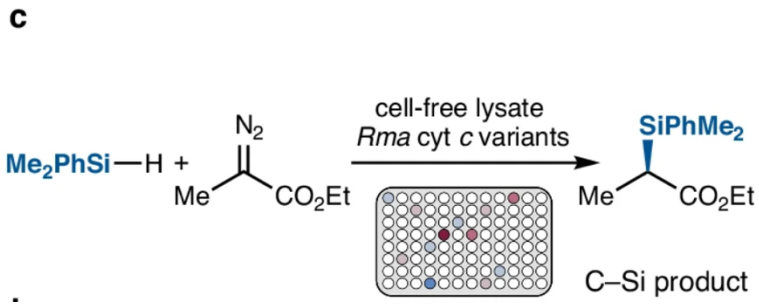
Using AI to Engineer generalist new-to-nature biocatalysts



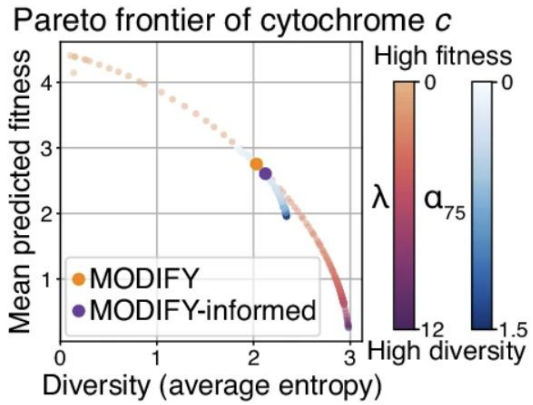
Rma cytochrome *c*

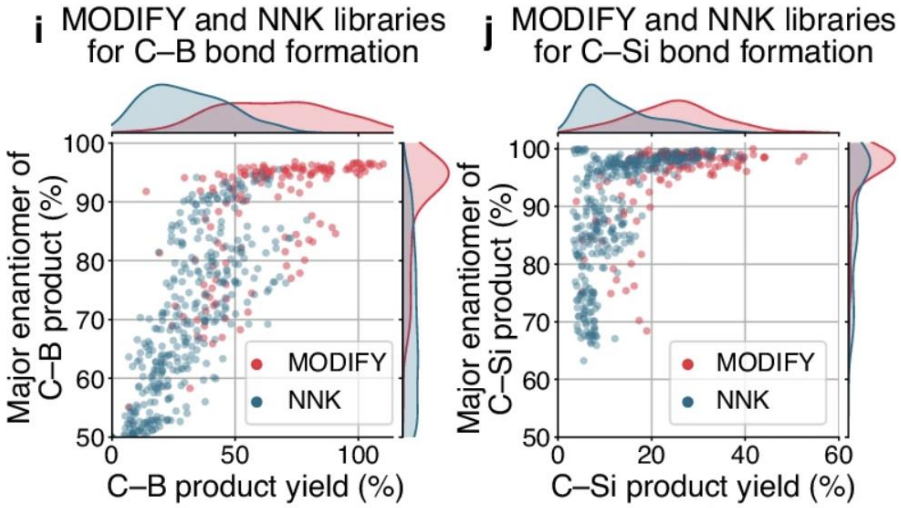


Carbon-boron (C-B) bond formation

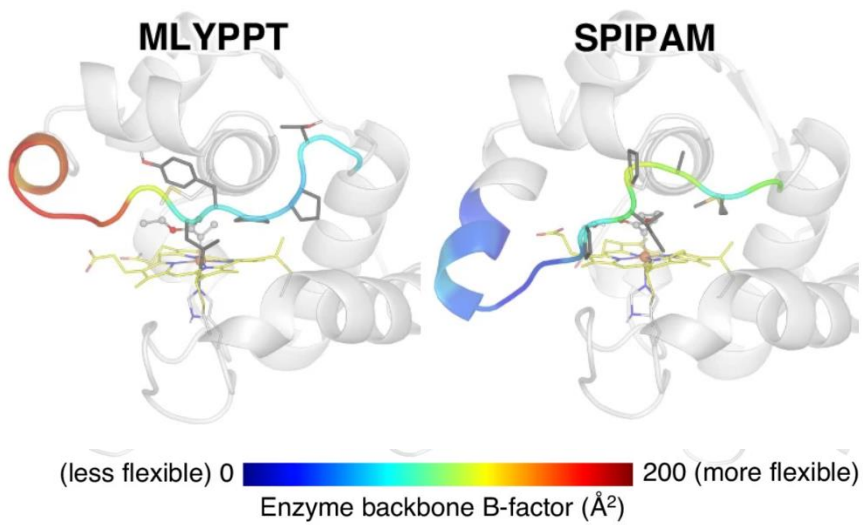


Carbon-silicon (C-Si) bond formation

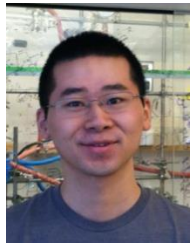




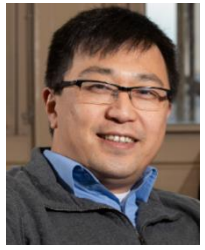
Improved two objectives: yield and selectivity



GenAI designed novel proteins 6-mutation away from human-designed proteins



Yang Yang (UCSB Chemistry)



Peng Liu (U of Pitt Chemistry)

Ding, K., Chin, M., Zhao, Y. et al. "Machine learning-guided co-optimization of fitness and diversity facilitates combinatorial library design in enzyme engineering," *Nature Communications*, 2024

How to design functional proteins?

- Learning from natural protein evolution using GenAI

$$p(\texttt{functional protein}) \approx p(\texttt{natural protein})$$

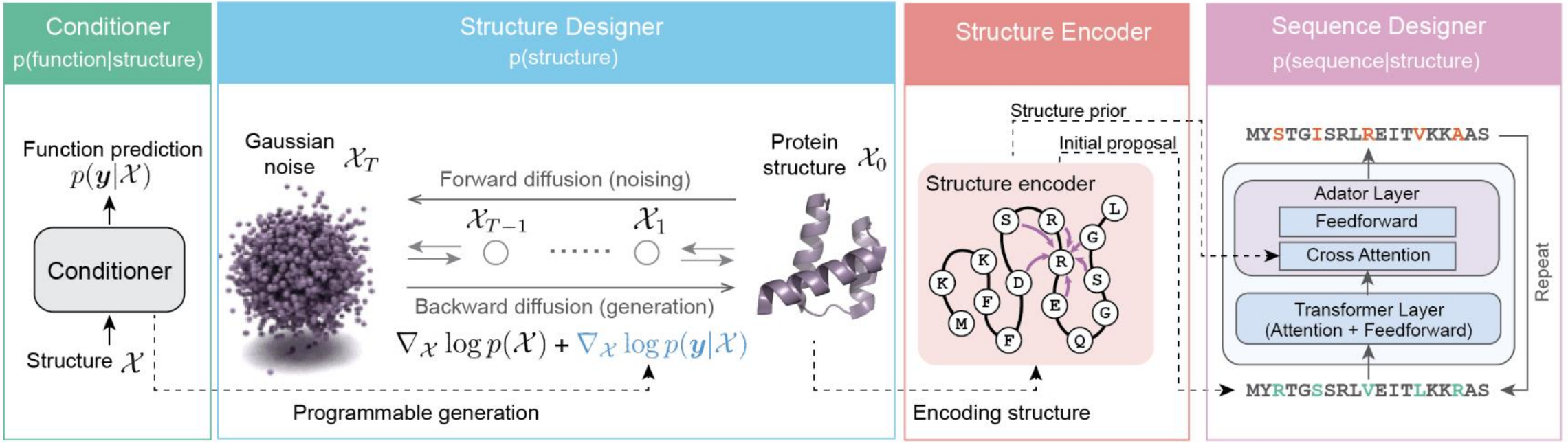
- Function-guided protein design with GenAI

$$p(\texttt{protein}|\texttt{function}) \propto p(\texttt{protein}) \times p(\texttt{function}|\texttt{protein})$$

Protein sequence/structure GenAI model

Protein function prediction model

Programmable multi-modal protein design



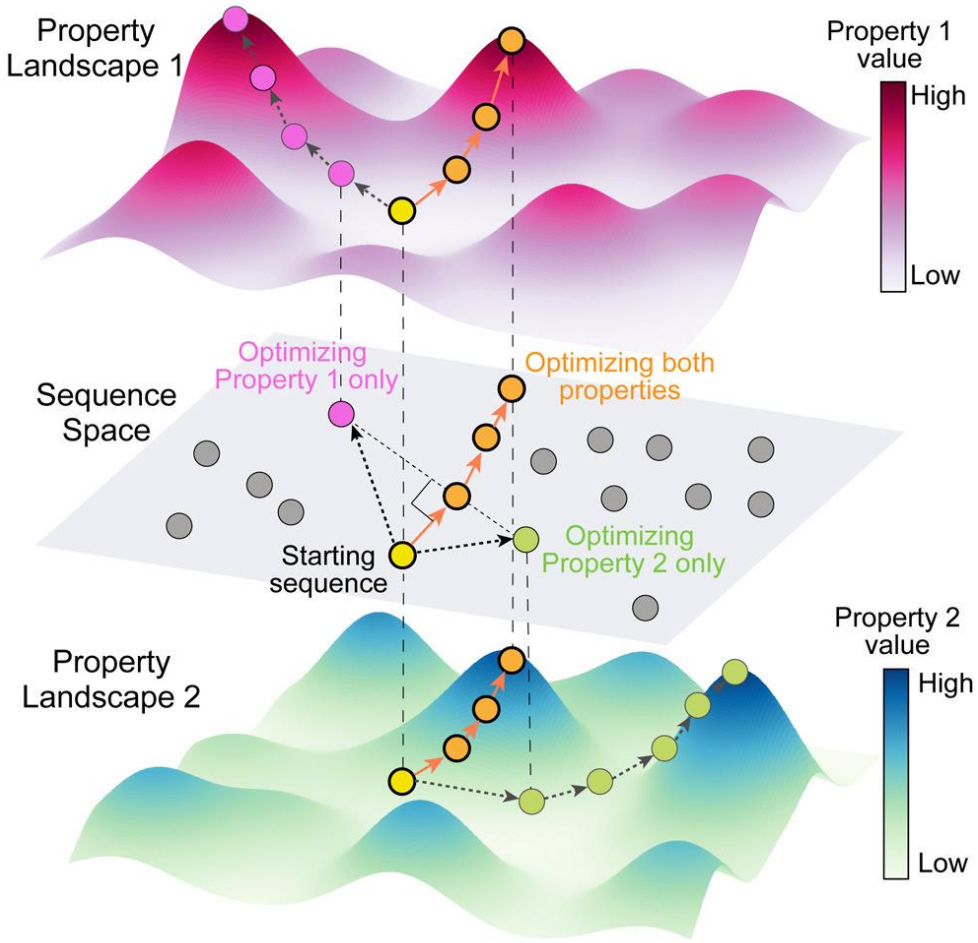
Protein function prediction model

Protein structure generative model

Protein sequence generative model

Multi-objective protein design

Recent results



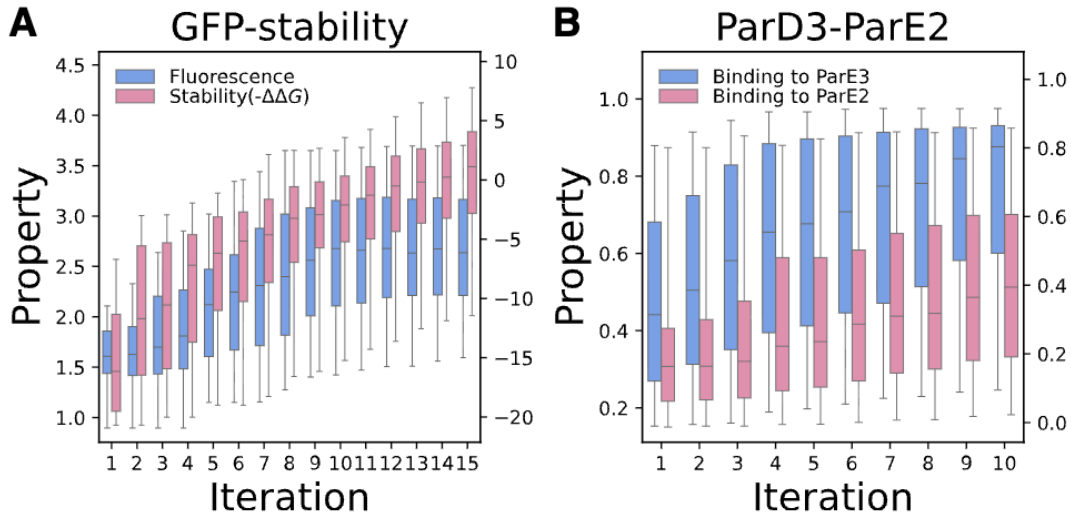
iScience

CellPress
OPEN ACCESS

Article

Pareto-optimal sampling for multi-objective protein sequence design

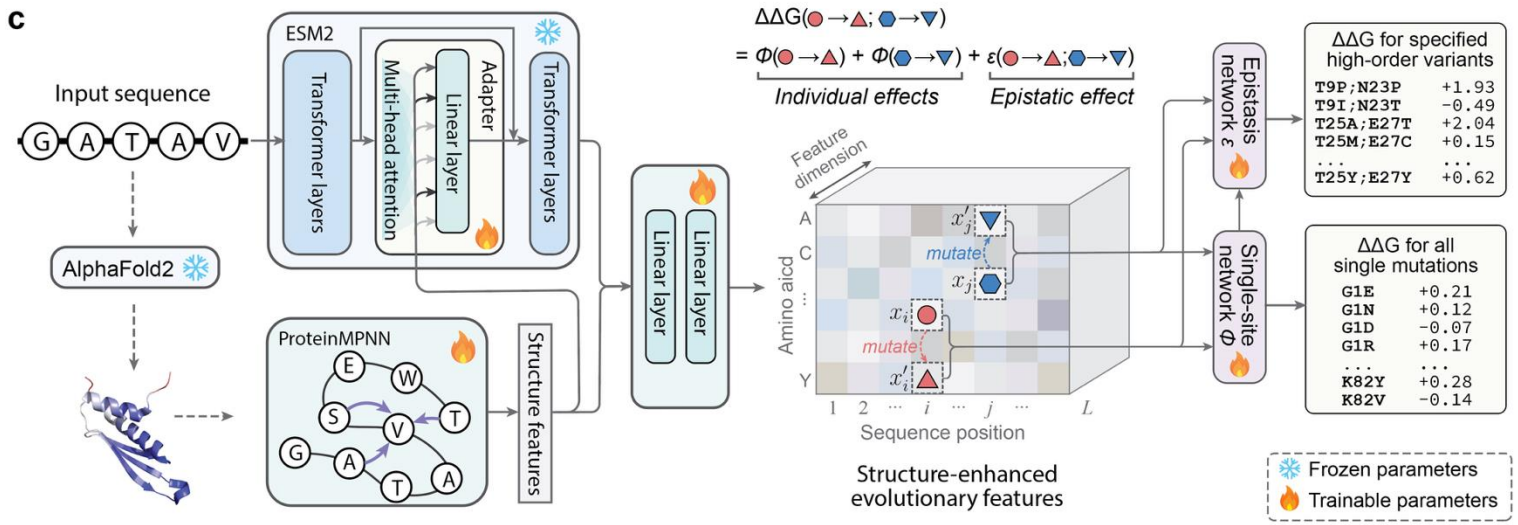
Jiaqi Luo,[1] Kerr Ding,[1] and Yunan Luo[1,2,*]
[1]School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, GA 30308, USA
[2]Lead contact
*Correspondence: yunan@gatech.edu
https://doi.org/10.1016/j.isci.2025.112119

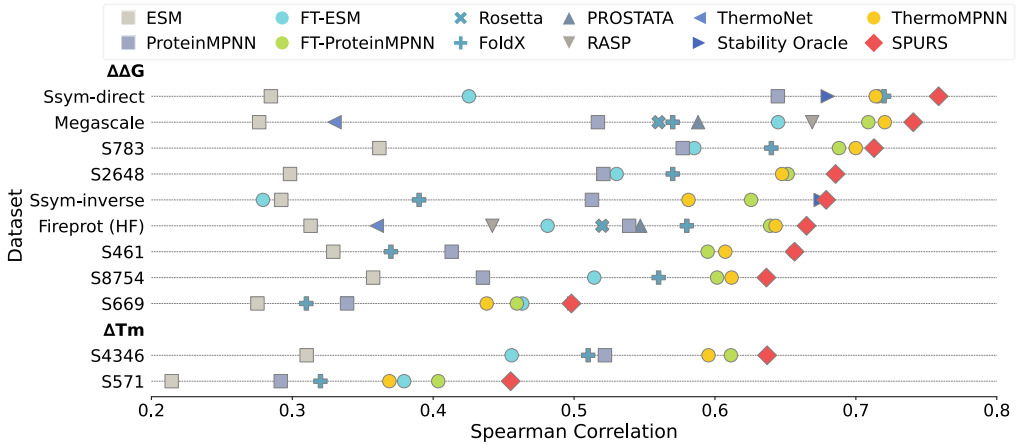


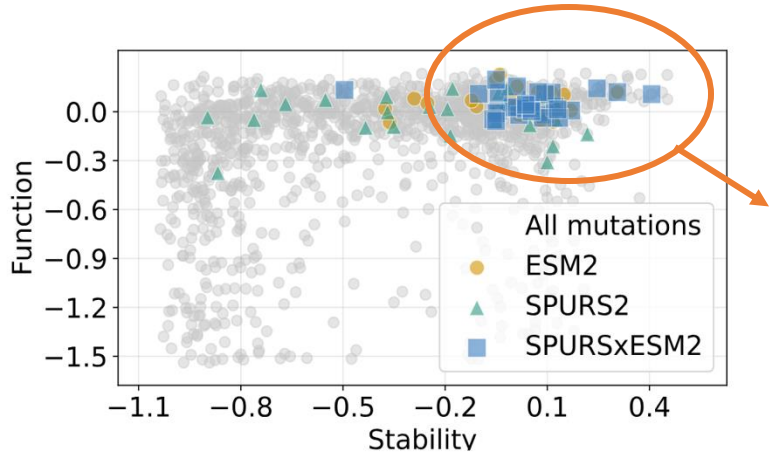
Supported by GaTech IDEaS x Microsoft CloudHub Seed Grant 2024

Multi-modal, multi-objective protein design

Preliminary results







Functional and stable proteins

Li, Ziang, and Yunan Luo. "Rewiring protein sequence and structure generative models to enhance protein stability prediction." *RECOMB*, 2025.

Conclusion

- Protein generative AI models capture evolutionary patterns of functional proteins
- Function-guided protein design improves hits rate
- Enables the design of novel, diverse functional proteins

Acknowledgements

- **Group:** Kerr Ding, Ziang Li, Jiaqi Luo, Tony Tu, Shitong Dai
- **Collaborators:** Huimin Zhao, Yang Yang, Peng Liu, Tianhao Yu, Junming Zhao, Liupeng Zhao, Michael Chin, Yunlong Zhao, Wei Huang, Binh Khanh Mai, Huanan Wang, et al.
- **Fundings:** GaTech IDEaS x Microsoft CloudHub Seed Grant, NIH MIRA (R35GM150890), NSF CAREER (2442063)

Georgia Tech Institute for Data Engineering and Science

Microsoft Azure

NIH National Institutes of Health

NSF