Designing New and Diverse Proteins with Generative AI

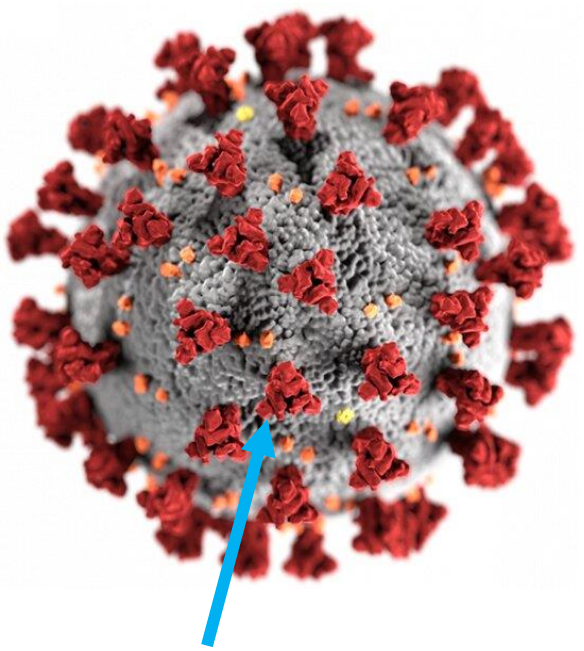
Yunan Luo

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

Proteins

A sequence s of length L over the alphabet Σ of 20 characters

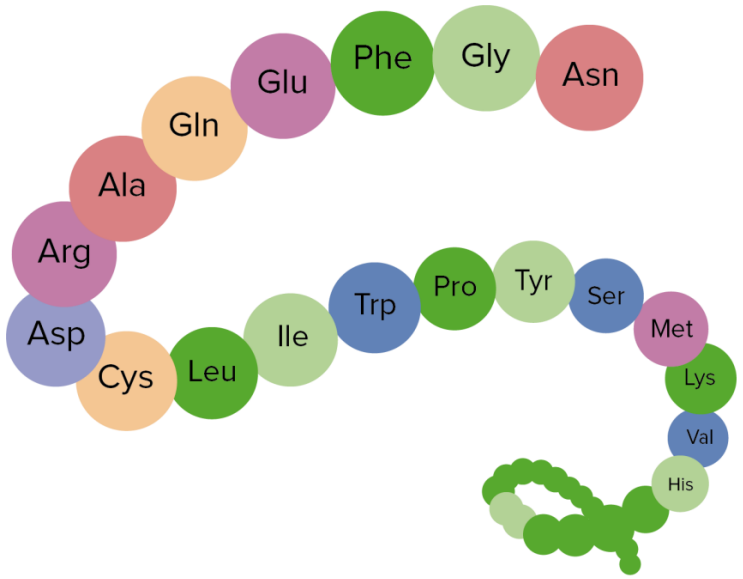
```
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFS
NVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIV
NNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLE
GKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQT
LLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETK
CTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISN
CVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIAD
YNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPC
NGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVN
FNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITP
GTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSY
ECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI
SVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE
VFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDC
LGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAM
QMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALN
TLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA
SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
ICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDP
LQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDL
QELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDD
```



SARS-CoV-2 Spike protein

Protein Universe

The arrangement of amino acids encodes the astonishing functional diversity of proteins



# of possible 100-residue proteins > # atoms in the universe

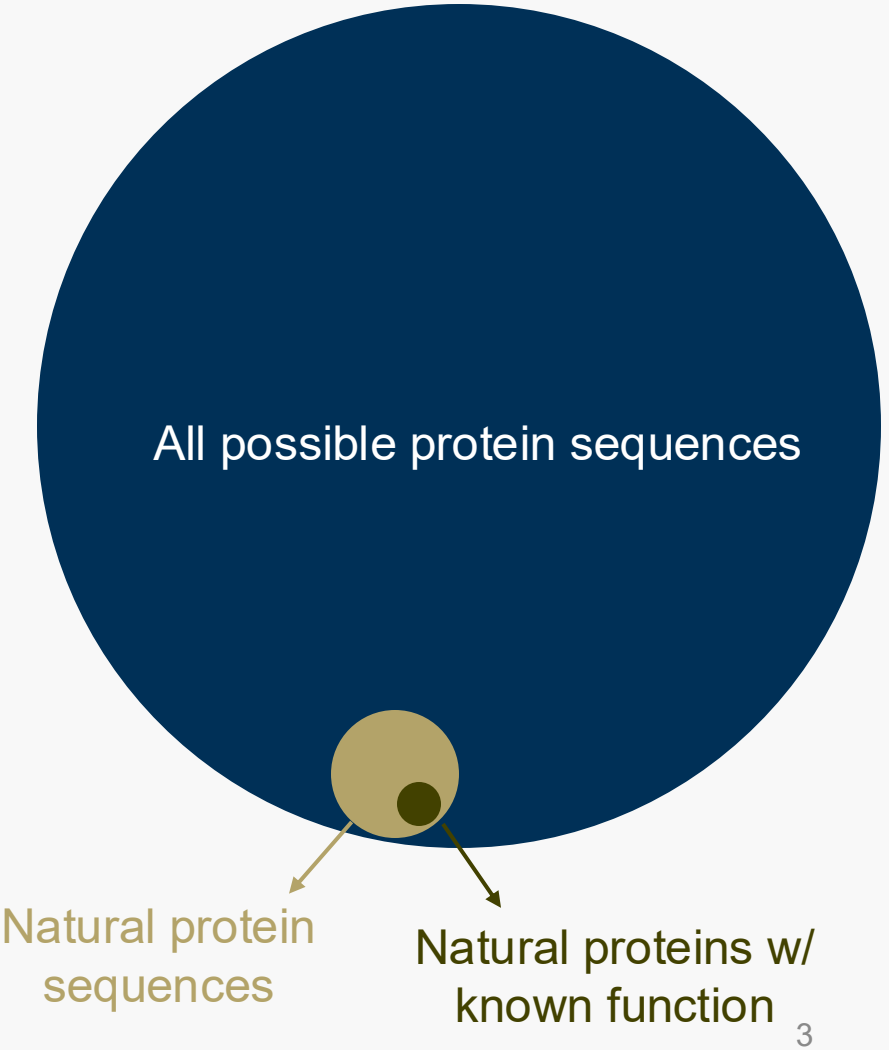
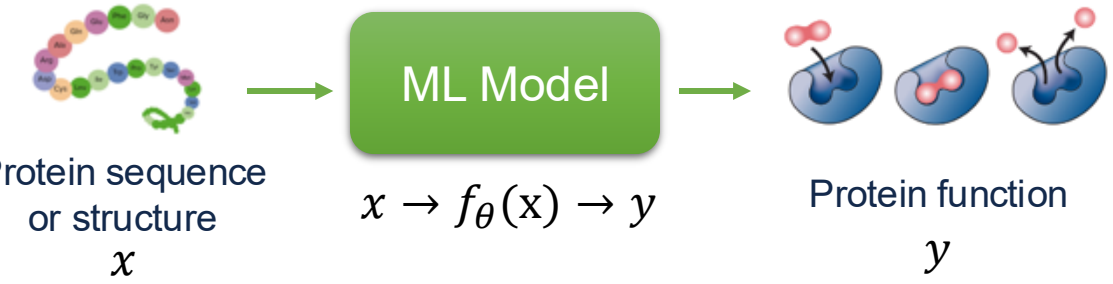


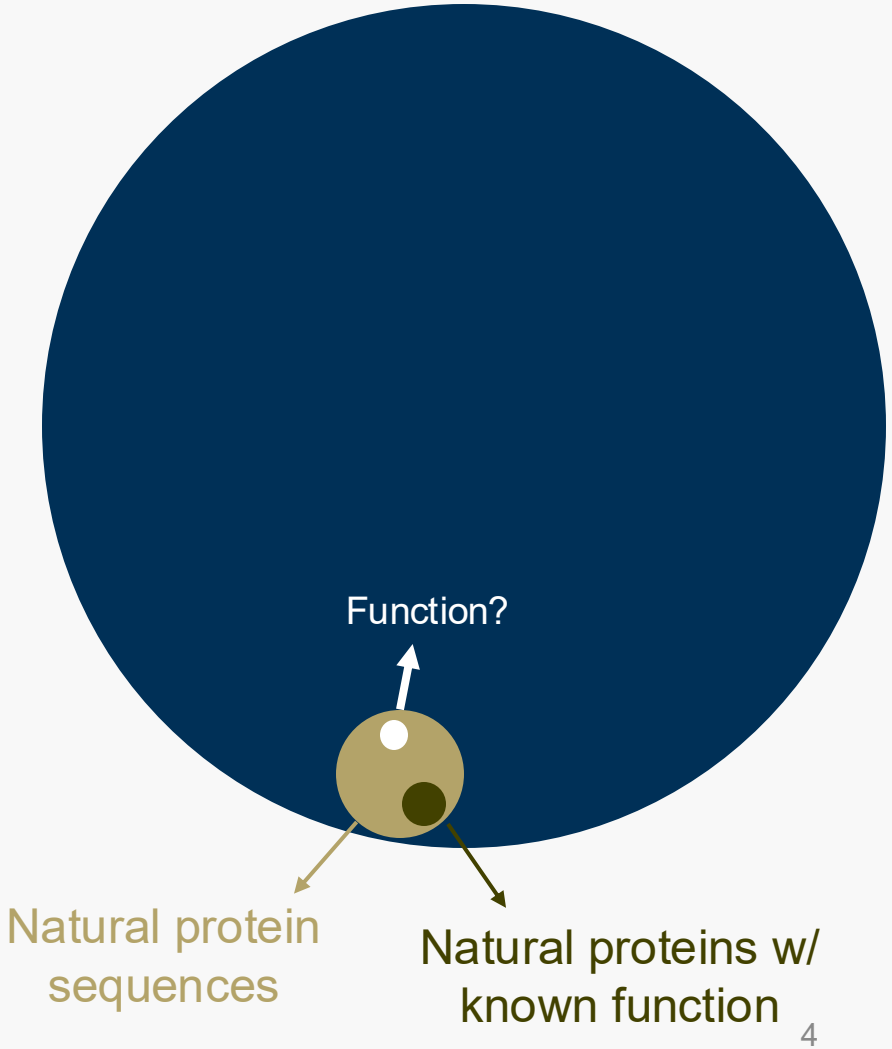
Figure: lecturio.com

Protein Universe

- Annotating functions for existing proteins

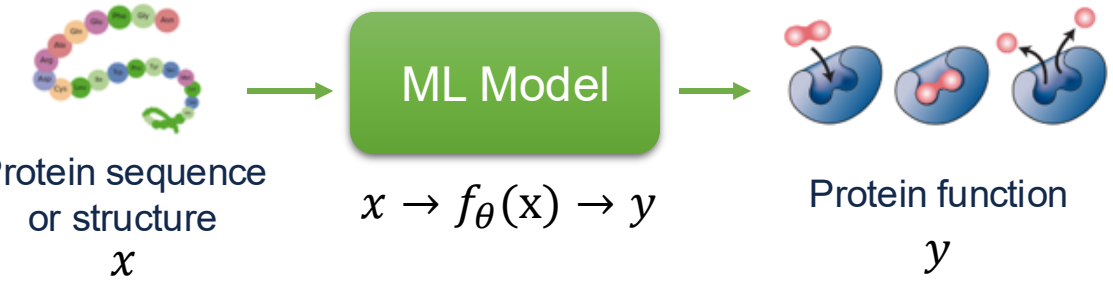


CLEAN: [Yu et al., *Science*, 2023];
PenLight: [Luo et al., *PSB*, 2023]; **MSRep** [Luo et al., *RECOMB*, 2025]



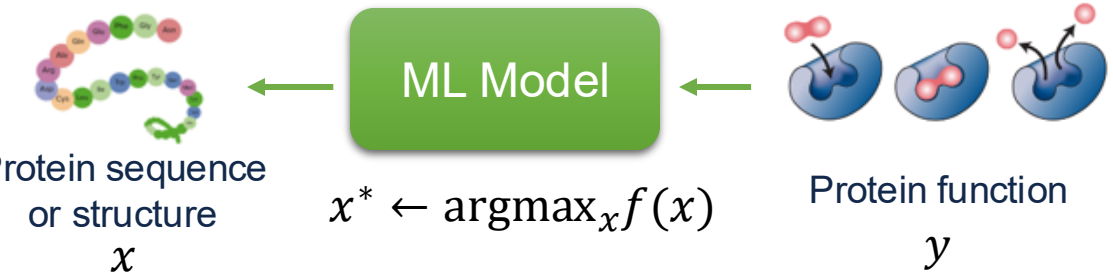
Protein Universe

- Annotating functions for existing proteins

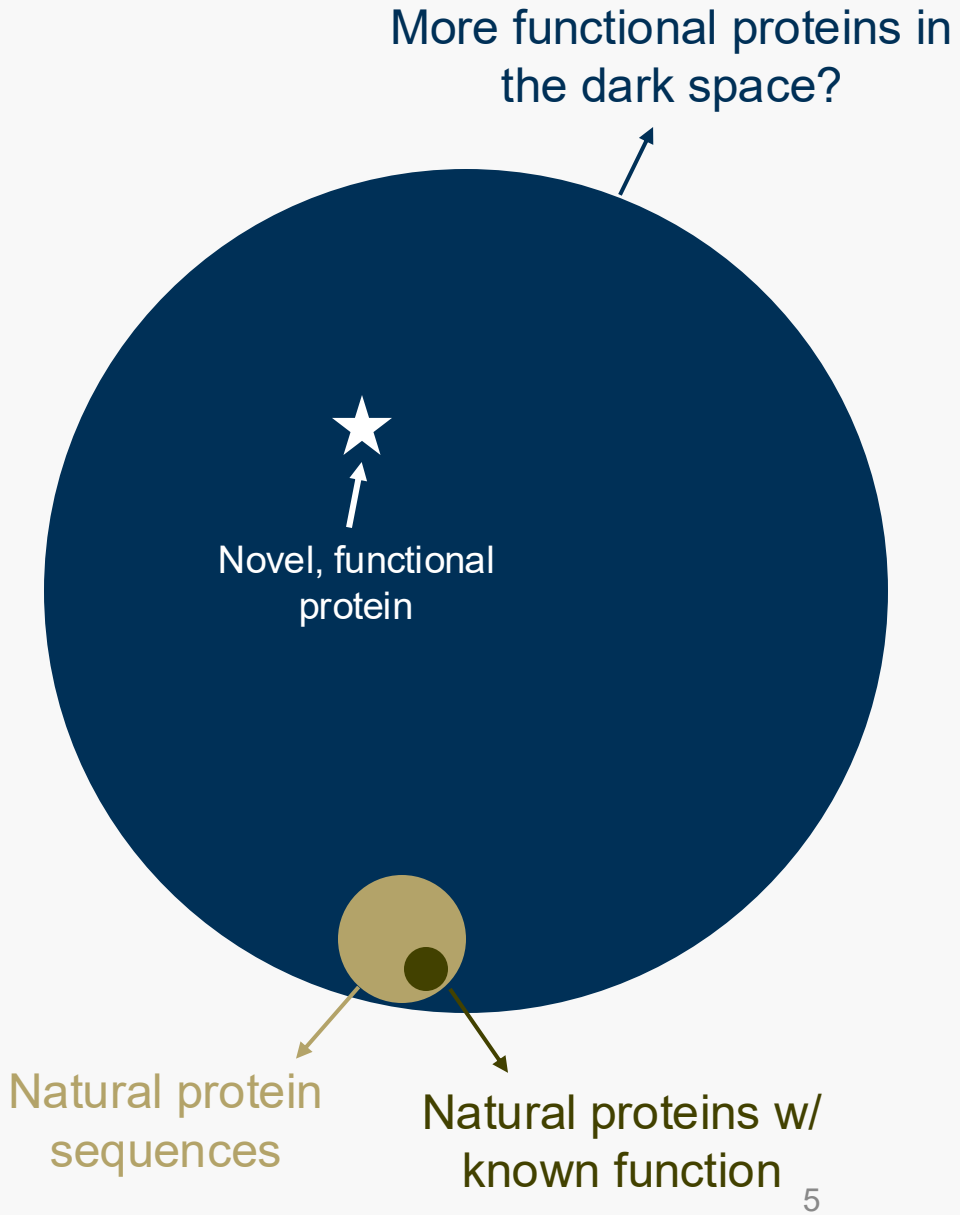


CLEAN: [Yu et al., *Science*, 2023];
PenLight: [Luo et al., *PSB*, 2023]; **MSRep** [Luo et al., *RECOMB*, 2025]

- Designing protein to achieve novel function



MODIFY: [Ding et al., *Nat Commun*, 2024];
ConFit: [Zhao et al., *RECOMB*, 2024]; **SPURS**: [Li et al., *RECOMB*, 2025]



Learning the language of protein sequences

Inspired by self-supervised learning of large language models (LLMs)

Masked token (word/amino acid) prediction

- **Natural language (e.g., GPT):**

```
To be, or not to __, that is the question
```

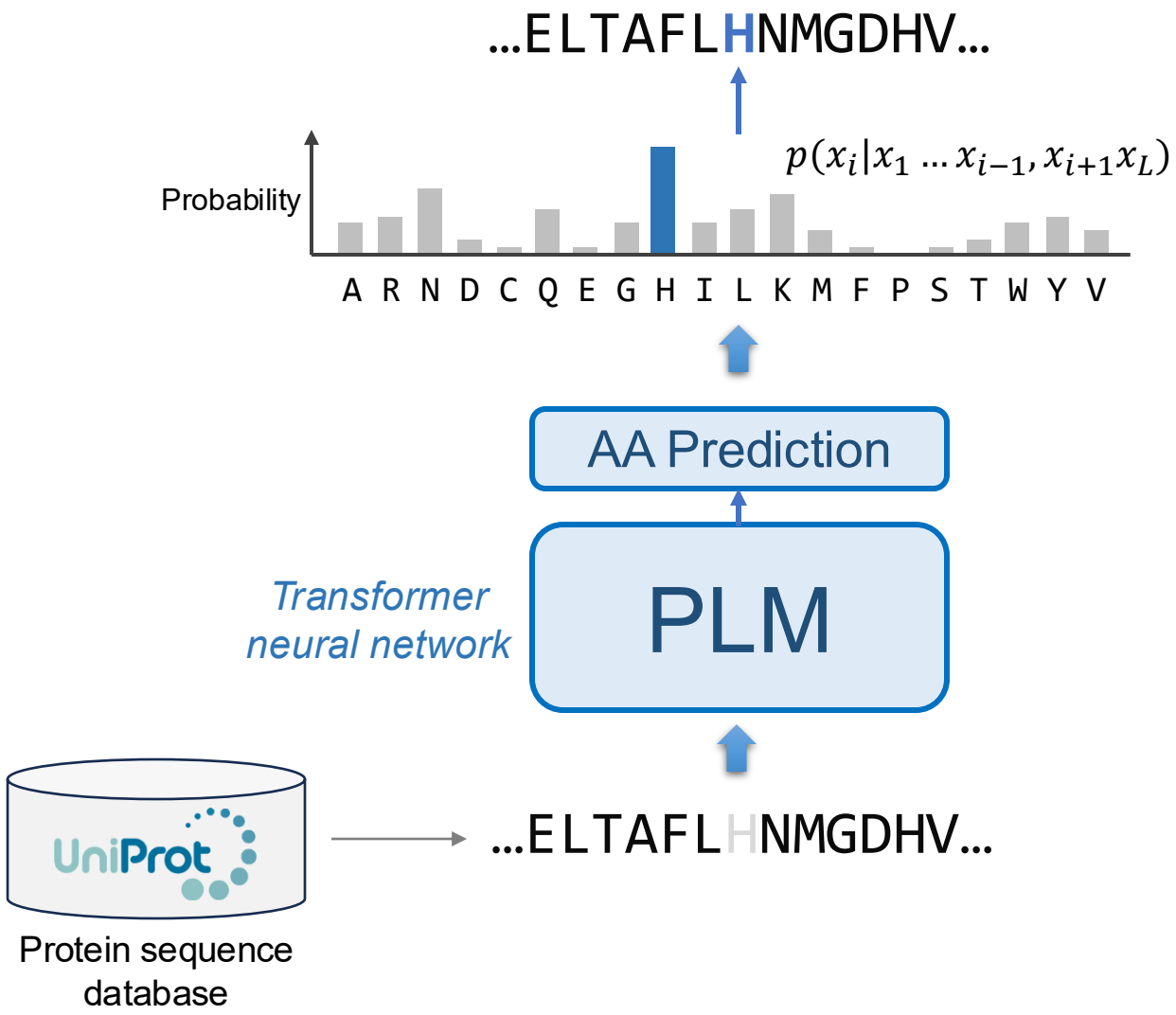
- **Protein language:**

```
MSKGEELFTGVVPILVE_DGDVNGHKFSVS
```


Adapted from Alex Rives

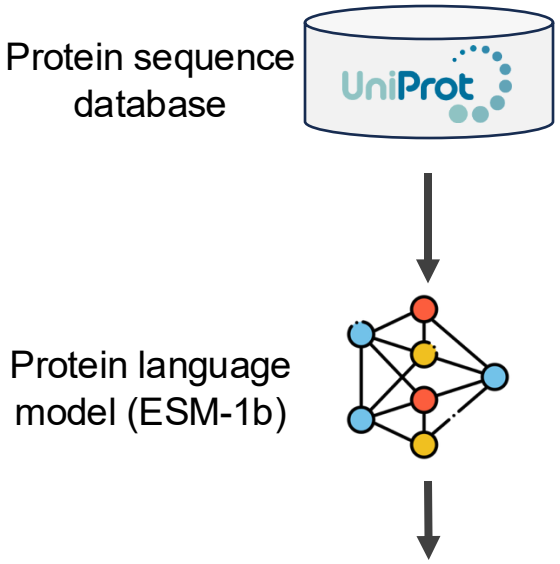
Protein Language Models (PLMs)

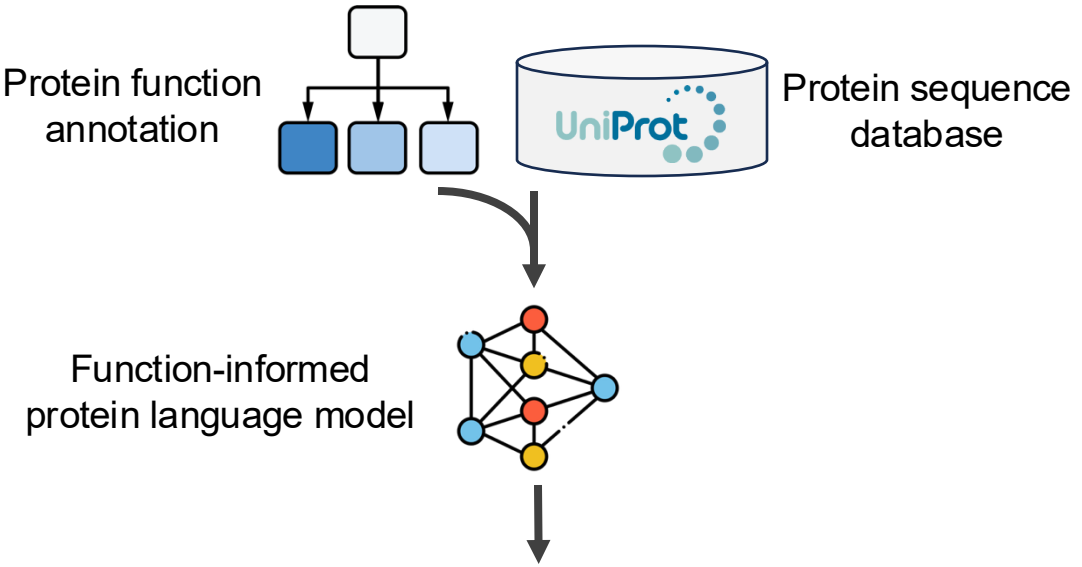
Inspired by large language models (LLMs) in natural language processing (NLP)



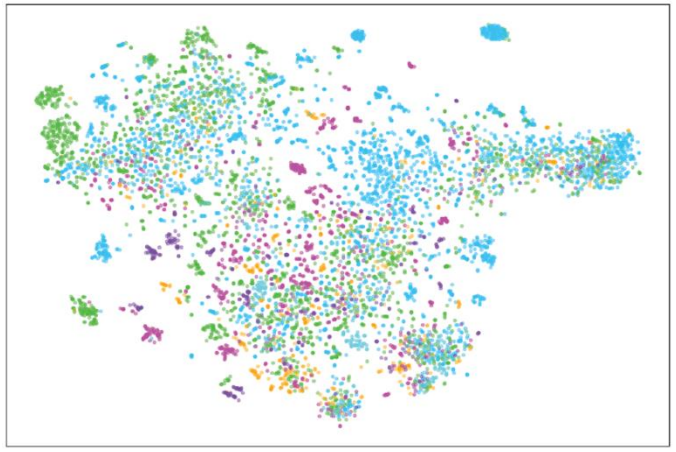
Bepler & Berger, *ICLR*, 2019
Rao et al, *NeurIPS*, 2019
Alley et al., *Nat. Methods*, 2019
Elnaggar et al., *TPAMI*, 2021
Rives et al., *PNAS*, 2021
Lin et al., *Science*, 2023
Madani et al., *Nat. Biotech*, 2023
Hayes et al., *Science*, 2025

Injecting function knowledge to protein language models





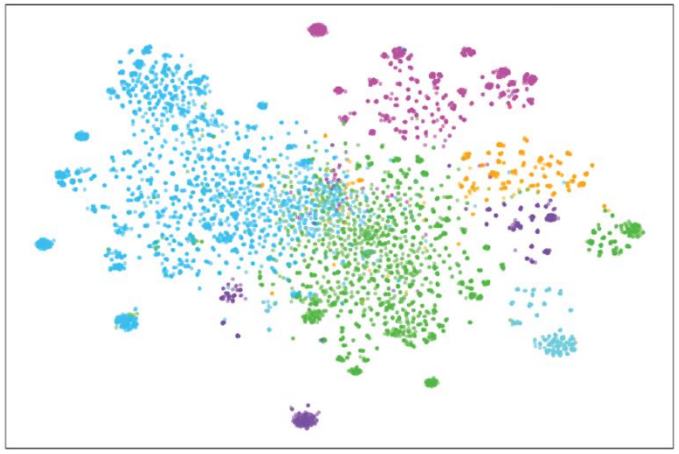
Sequence representation (tSNE)



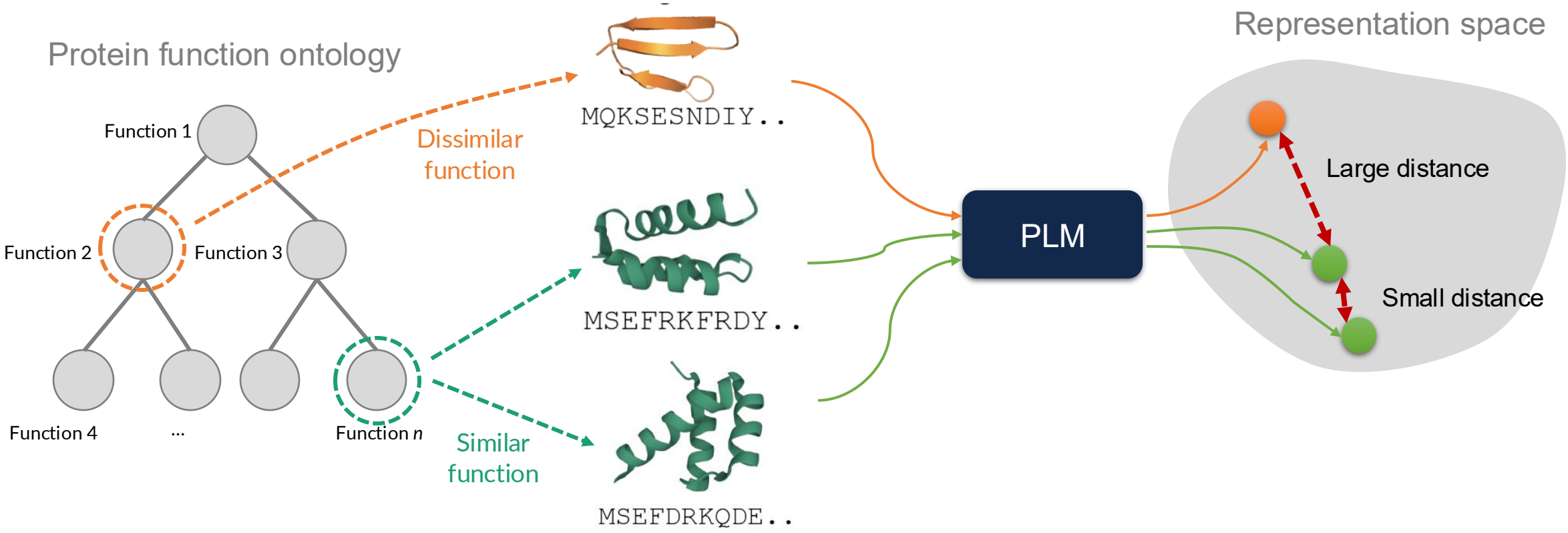
6 major enzyme functions

- Oxidoreductases
- Transferases
- Hydrolases
- Lyases
- Isomerases
- Ligases

Sequence representation (tSNE)



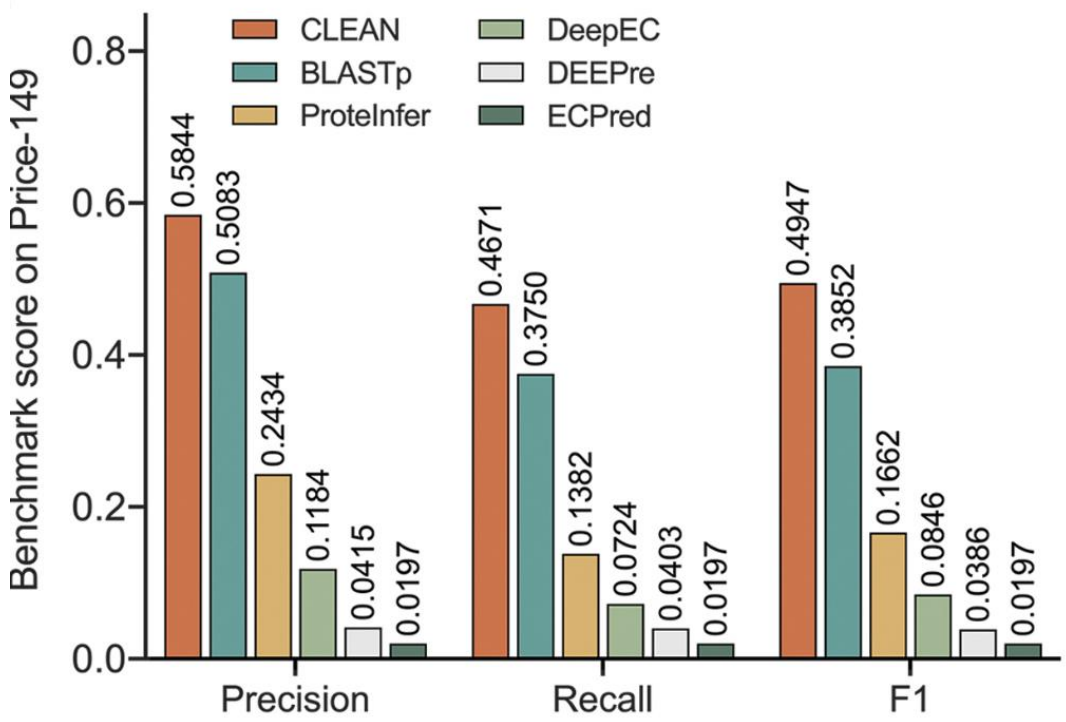
CLEAN: contrastive learning for enzyme function annotation



Yu et al., *Science* (2023)

State-of-the-art enzyme function annotation

(In collaboration with Prof. Huimin Zhao's lab at University of Illinois)



Science

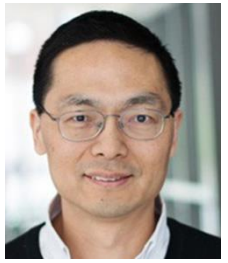
RESEARCH

FUNCTION PREDICTION

Enzyme function prediction using contrastive learning

Tianhao Yu[1,2,3]†, Haiyang Cui[1,2,3]†, Jianan Canal Li[3,4], Yunan Luo[5], Guangde Jiang[1,2], Huimin Zhao[1,2,3,6]*

Yu et al., *Science* (2023)



Huimin Zhao
(UIUC ChemE)

CLEAN annotates understudied enzymes

(In collaboration with Prof. Huimin Zhao's lab at University of Illinois)

Enzyme SsFIA (UniProt: W0W999)

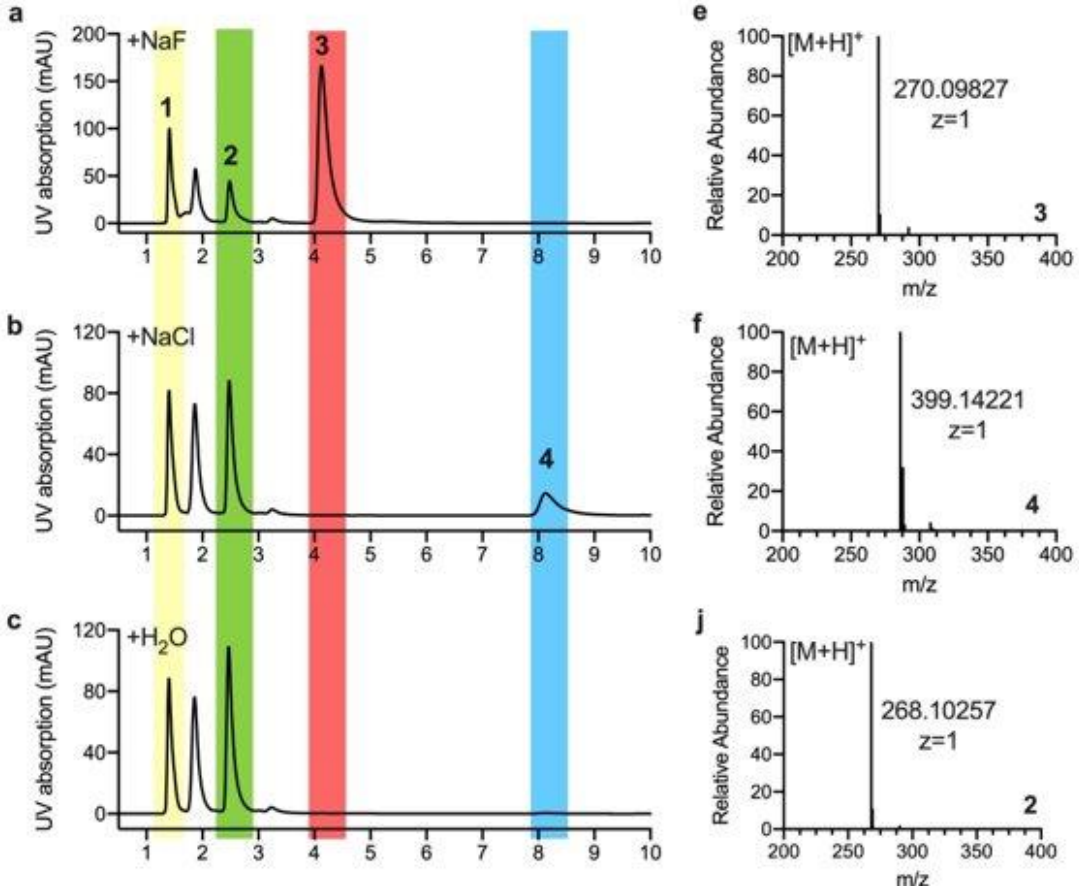
- In UniProt/TrEMBL: labeled as Fluorinase (EC: 2.5.1.63)

Baseline predictions:

- BLAST: Fluorinase (EC: 2.5.1.63)
- DeepEC: Fluorinase (EC: 2.5.1.63)
- ProtInfer: Fluorinase (EC: 2.5.1.63)

Our predictions:

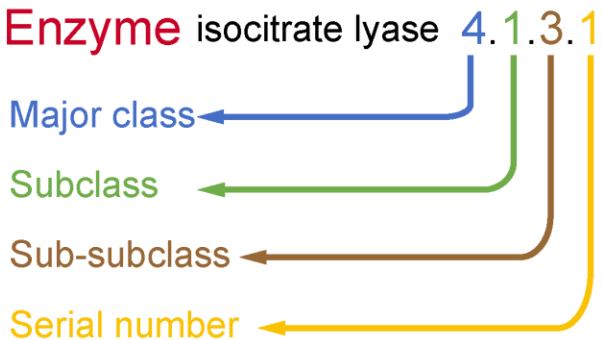
- Fluorinase (EC: 2.5.1.63)
- Chlorinase (EC: 2.5.1.94)
- Hydrolase (EC: 3.13.1.8)



Yu et al., *Science* (2023)

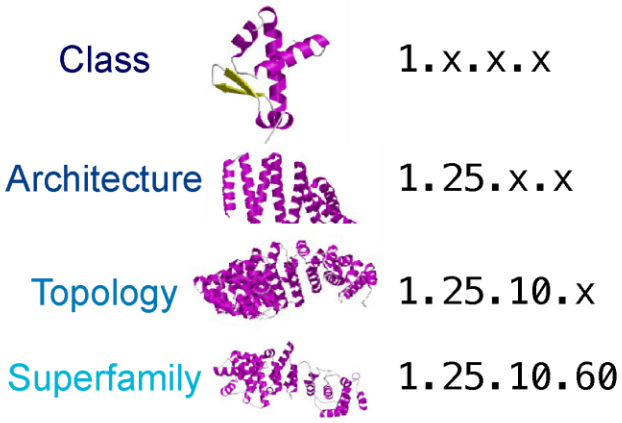
Protein function annotation schemes

EC number



Enzyme catalysis reaction classification

Gene3D

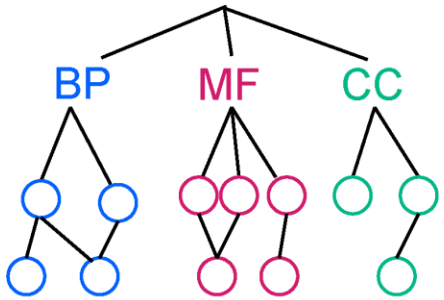


Protein domain structure annotations

Pfam

Protein family classification

Gene Ontology

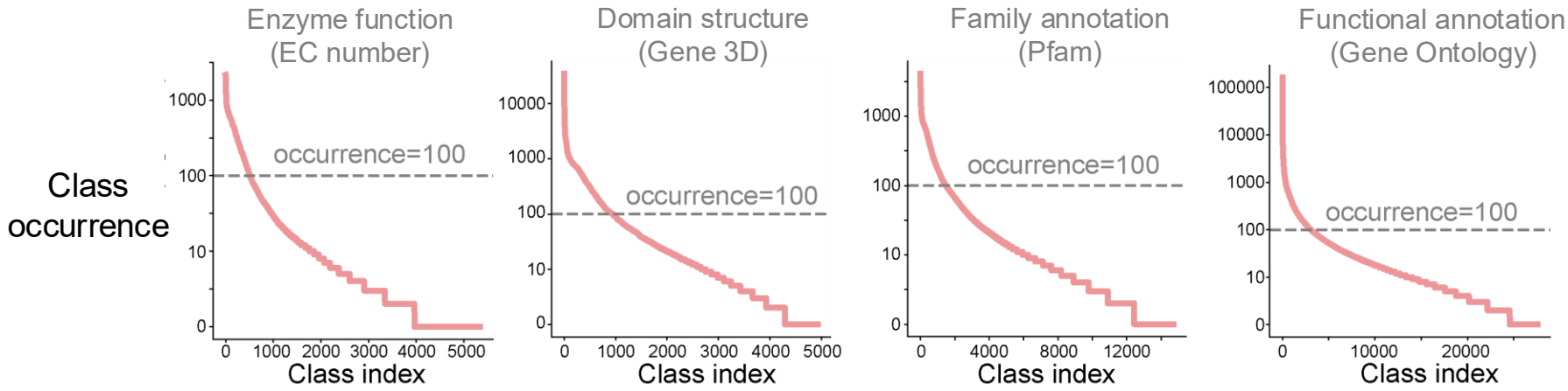


Hierarchical function annotation

Supported by GaTech IDEaS x Microsoft CloudHub Seed Grant

Annotation gap of protein functions

- Databases of protein function annotations follow a long-tail distribution



  - A few function classes are associated with the majority of proteins (‘head classes’), while many others are sparsely annotated (‘tail classes’)

Unequal representativeness in ML parameter space

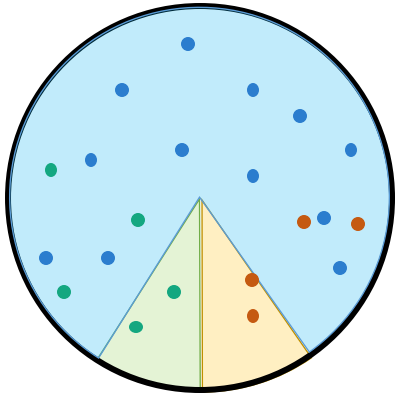
$$\mathcal{L}_{all} = \sum_{i=1}^{N} \mathcal{L}(y_{\text{pred}}^{(i)}, y_{\text{true}}^{(i)}) = \sum_{y_{\text{true}}^{(i)} \in H} \mathcal{L}(y_{\text{pred}}^{(i)}, y_{\text{true}}^{(i)}) + \sum_{y_{\text{true}}^{(i)} \in T} \mathcal{L}(y_{\text{pred}}^{(i)}, y_{\text{true}}^{(i)})$$

Head classes
(many samples)

Tail classes
(very few samples)

Biased representation space

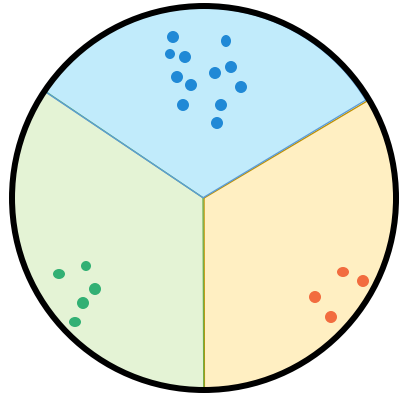
Biased ML model:



Most of the model's "learning capacity" overly focuses on optimizing prediction accuracy for head classes

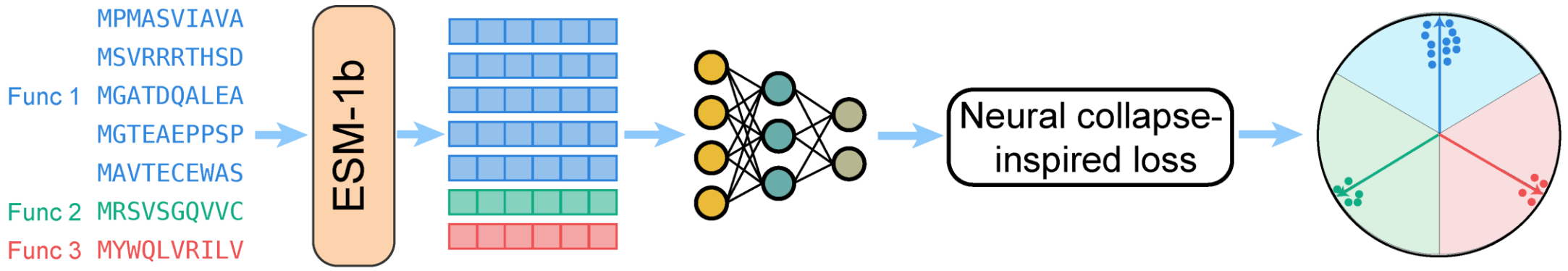
Equally distributed representation

What we want:



Allocating "parameter budget" equally to all function classes

MSRep: learning maximally spanning embedding space



Imbalanced data | Pretrained protein language model representations | Projection network | Loss | Representation space

Property 1: Class centers are maximally separated.

$$L_{inter} = \max \min_{i \neq j} \arccos\left(\frac{\langle \boldsymbol{\mu}_i - \boldsymbol{\mu}_C, \boldsymbol{\mu}_j - \boldsymbol{\mu}_C \rangle}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_C\| \cdot \|\boldsymbol{\mu}_j - \boldsymbol{\mu}_C\|}\right)$$

Property 2. Embeddings of the same class collapse to class center.

$$L_{intra} = \sum_{j=1}^{K} \frac{1}{n_j^{\gamma}} \sum_{\boldsymbol{x}_i : \psi(\boldsymbol{x}_i) = j} \frac{\langle \boldsymbol{x}_i, \boldsymbol{\mu}_i \rangle}{\|\boldsymbol{x}_i\| \cdot \|\boldsymbol{\mu}_i\|}$$

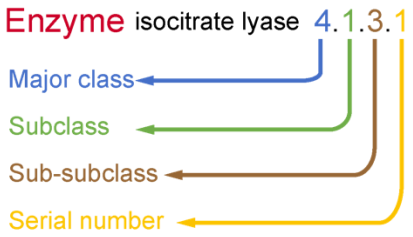
Final loss: $Loss = L_{intra} + \beta \cdot L_{inter}$

Luo, Jiaqi, and Yunan Luo. "Learning maximally spanning representations improves protein function annotation." *RECOMB*, 2025.

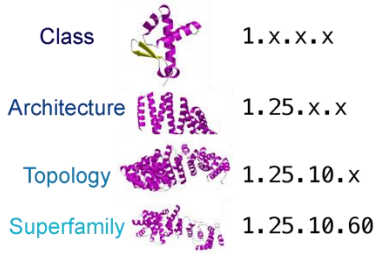
Supported by GaTech IDEaS x Microsoft CloudHub Seed Grant

Improved performance across protein function annotation tasks

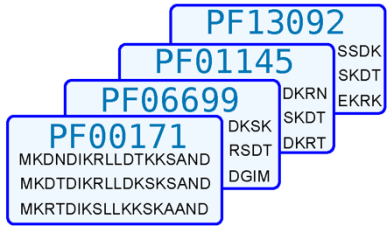
Enzyme function annotation (EC number)



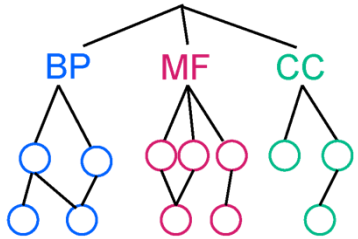
Protein domain structure classification (Gene3D)

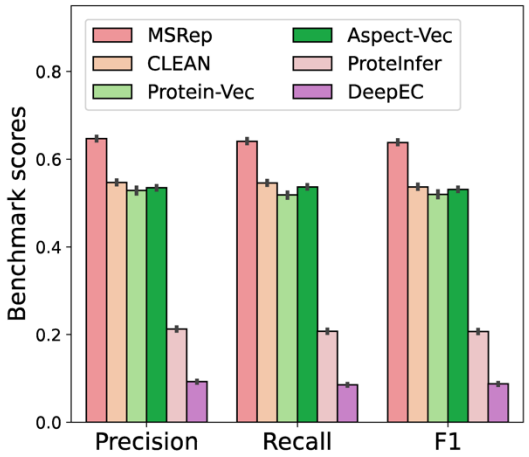


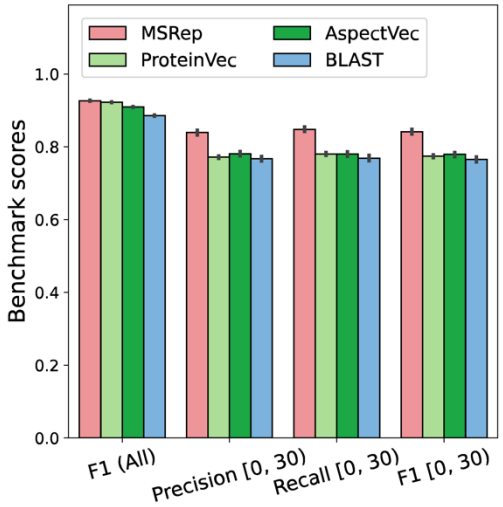
Protein family classification (Pfam)

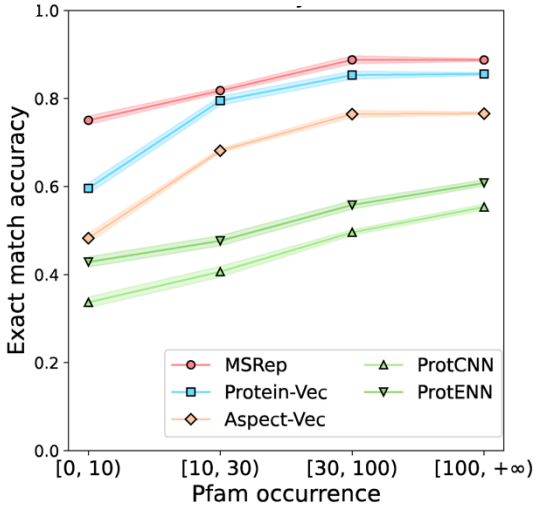


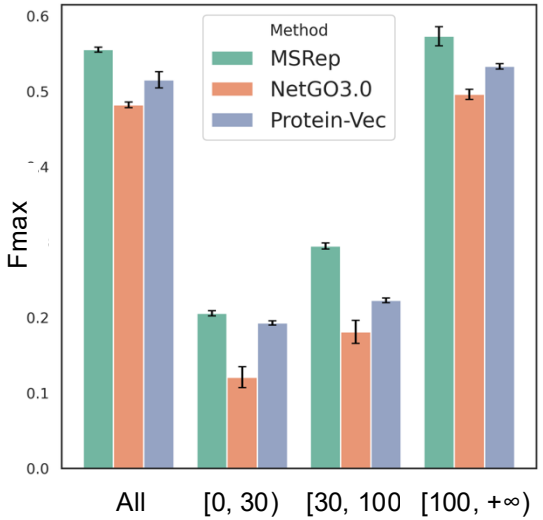
Gene function annotation (Gene ontology)











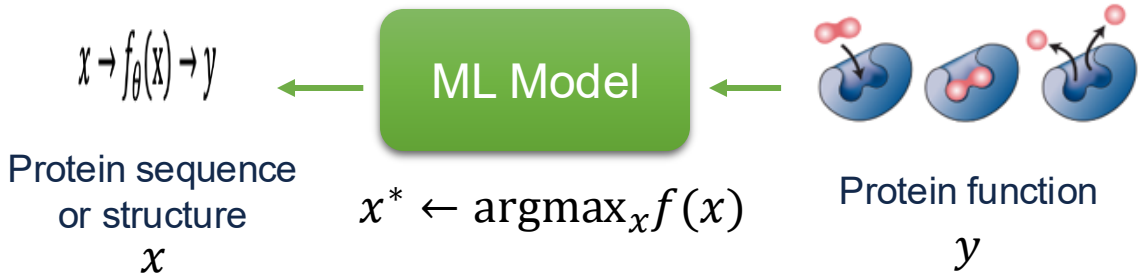
Luo, Jiaqi, and Yunan Luo. "Learning maximally spanning representations improves protein function annotation." *RECOMB*, 2025.

Supported by GaTech IDEaS x Microsoft CloudHub Seed Grant

- Annotating functions for existing proteins

CLEAN: [Yu et al., *Science*, 2023]; **PenLight**: [Luo et al., *PSB*, 2023];
MSRep [Luo et al., *RECOMB*, 2025]

- Designing protein to achieve novel function



MODIFY: [Ding et al., *Nat Commun*, 2024];
ConFit: [Zhao et al., *RECOMB*, 2024]; **SPURS**: [Li et al., *RECOMB*, 2025]

Protein fitness landscape

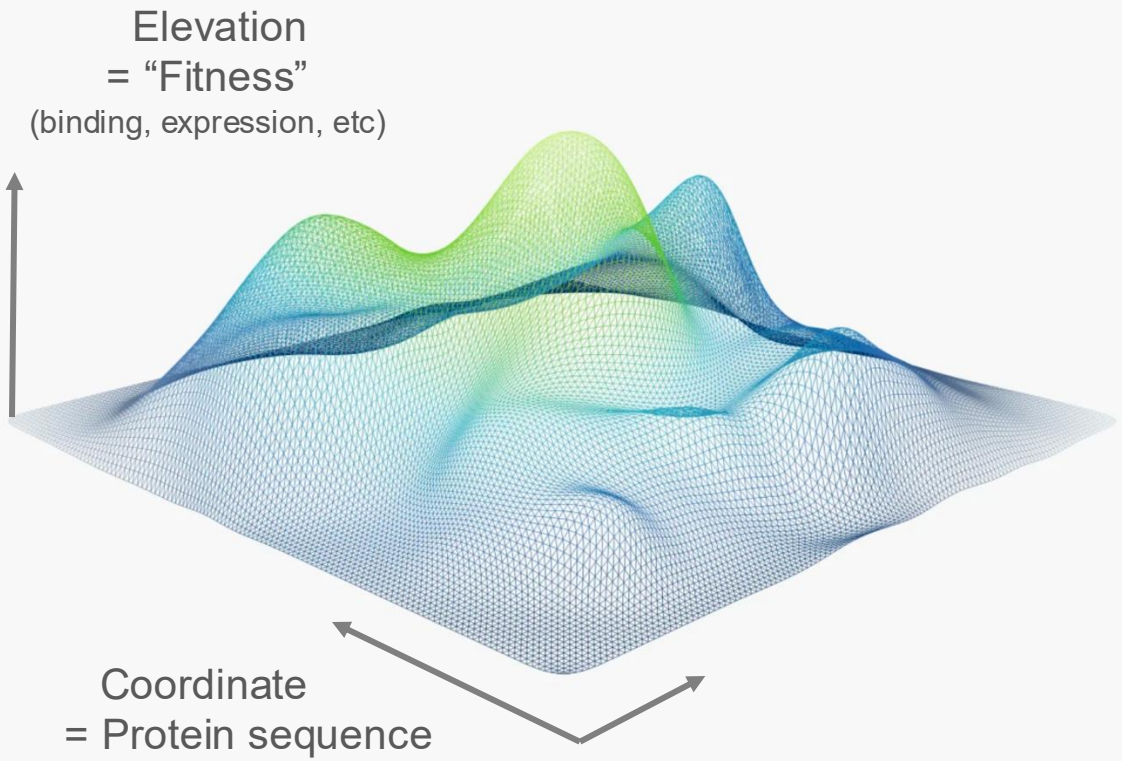


Figure: cradle.bio

How to find functional proteins?

Protein fitness landscape

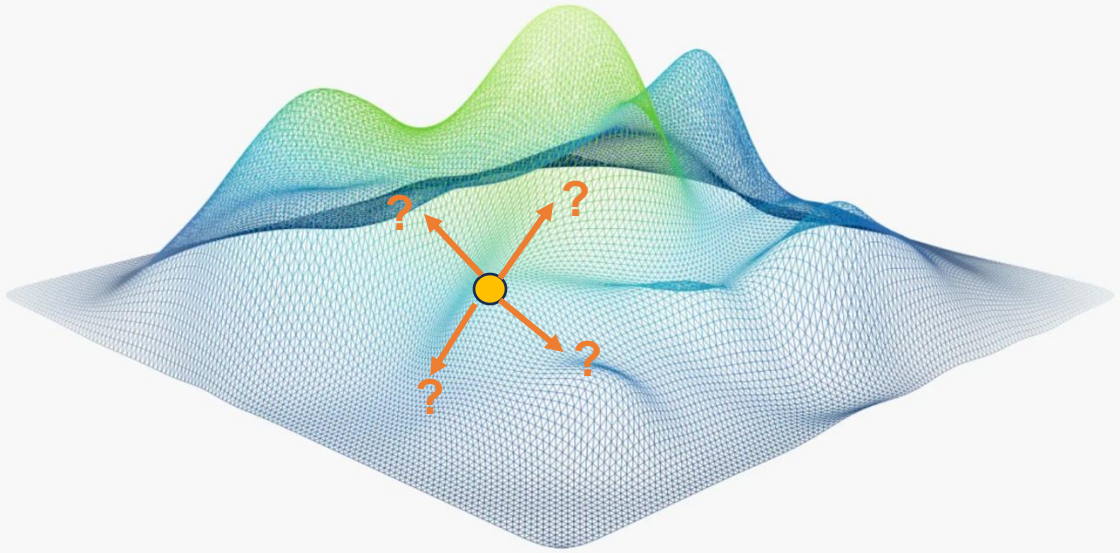


Figure: cradle.bio

How to find functional proteins?

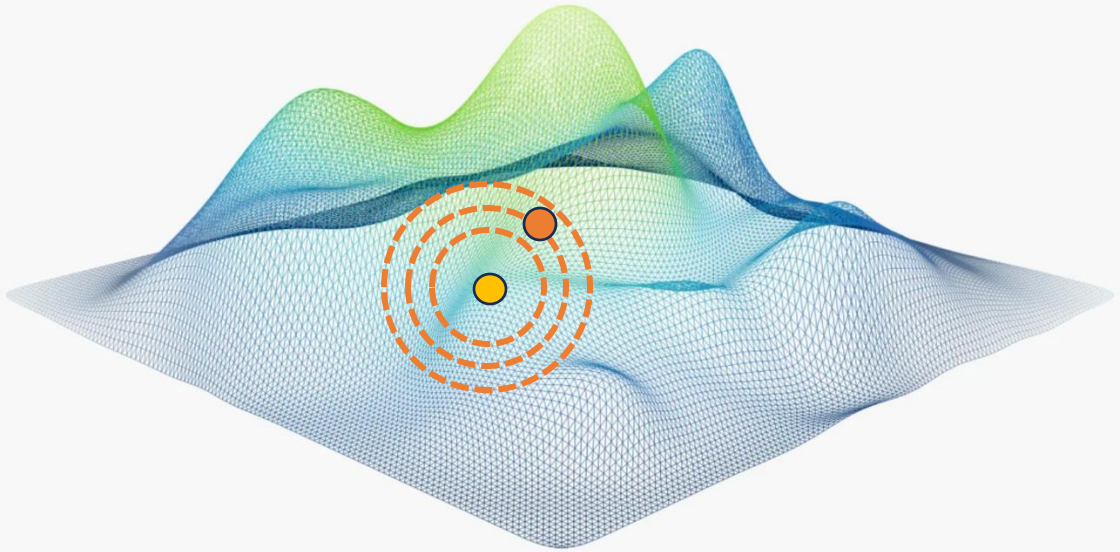
Natural Selection and the Concept of a Protein Space

JOHN MAYNARD SMITH

School of Biological Sciences,
University of Sussex.

Maynard Smith.
Nature (1970): 563-564.

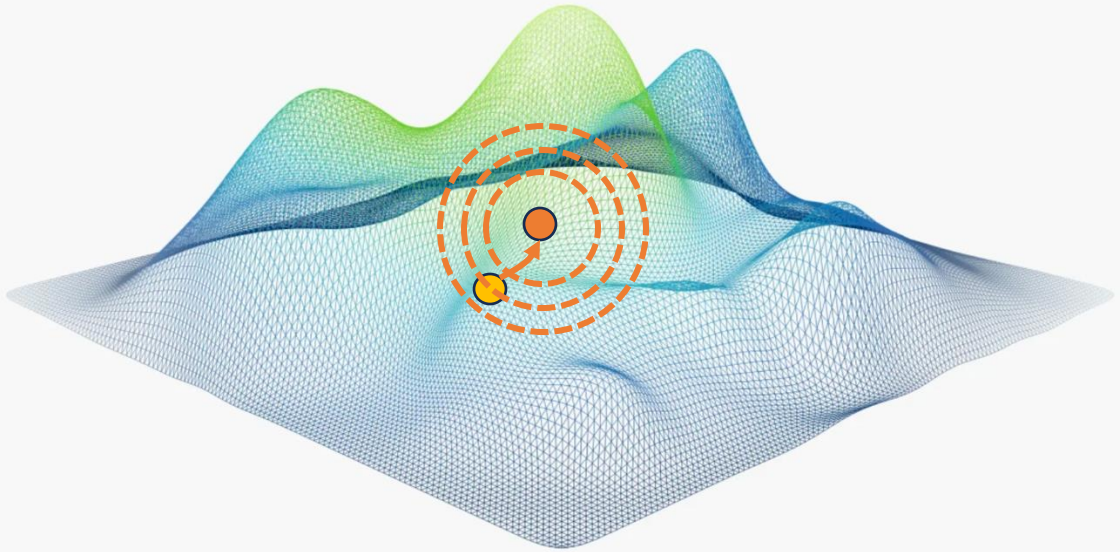
If evolution is to occur, functional proteins must be surrounded by other functional proteins



Protein engineering

- Mutagenesis experiments

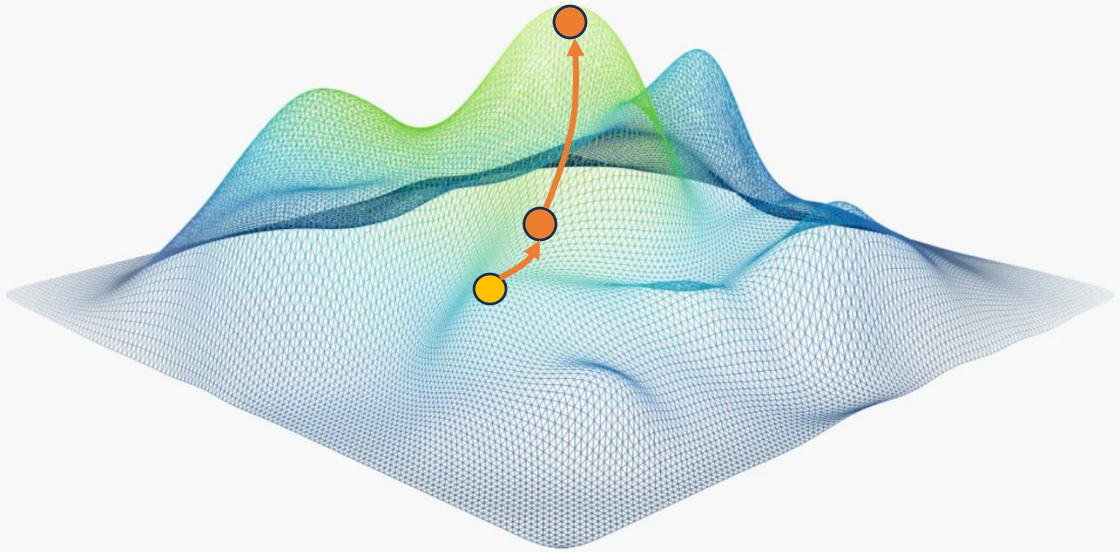
Sequence	Fitness of interest (e.g., binding affinity, stability, expression)
MSKGEELFTGVVPIL	1.0
MSKCEELFTGVVPIL	0.5
MSKGWELFTGVVPIL	5.4
MSKGEELFTSVVPIL	0.6
MSKGEELFTVVVPIL	1.1
MSYGEEVFTGVVPIL	2.5
MSKGEFGFAGVVPIL	0.1
MDYGEELFTVVVPSL	1.3
MSKGWELFTGVAPIL	6.1



Protein engineering

- Mutagenesis experiments

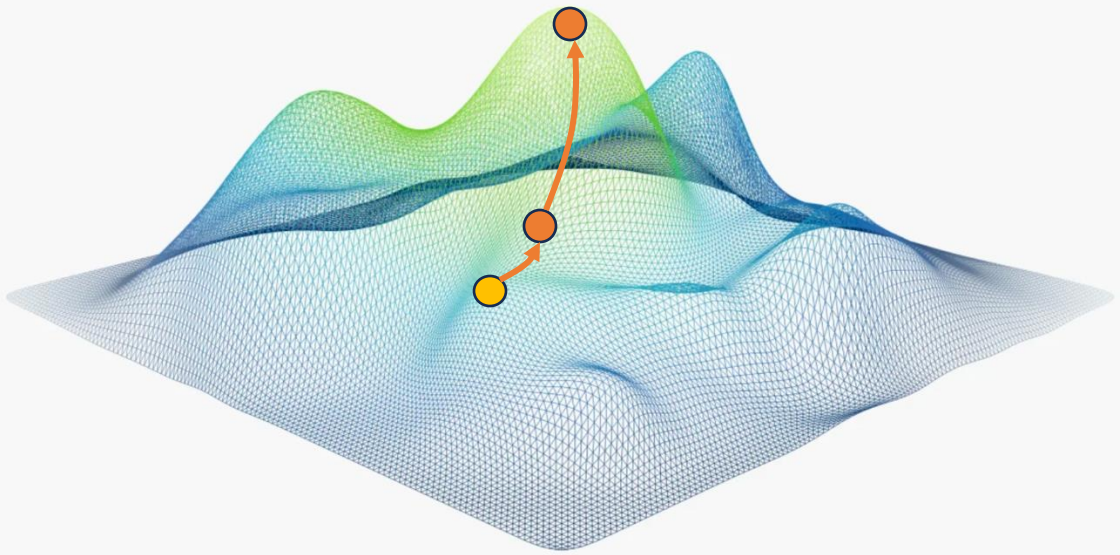
Sequence	Fitness of interest (e.g., binding affinity, stability, expression)
MSKGEELFTGVVPIL	1.0
MSKCEELFTGVVPIL	0.5
MSKGWELFTGVVPIL	5.4
MSKGEELFTSVVPIL	0.6
MSKGEELFTVVVPIL	1.1
MSYGEEVFTGVVPIL	2.5
MSKGEFGFAGVVPIL	0.1
MDYGEELFTVVVPSL	1.3
MSKGWELFTGVAPIL	6.1



Protein engineering

- Mutagenesis experiments

Sequence	Fitness of interest (e.g., binding affinity, stability, expression)
MSKGEELFTGVVPIL	1.0
MSKCEELFTGVVPIL	0.5
MSKGWELFTGVVPIL	5.4
MSKGEELFTSVVPIL	0.6
MSKGEELFTVVVPIL	1.1
MSYGEEVFTGVVPIL	2.5
MSKGEFGFAGVVPIL	0.1
MDYGEELFTVVVPSL	1.3
MSKGWELFTGVAPIL	6.1

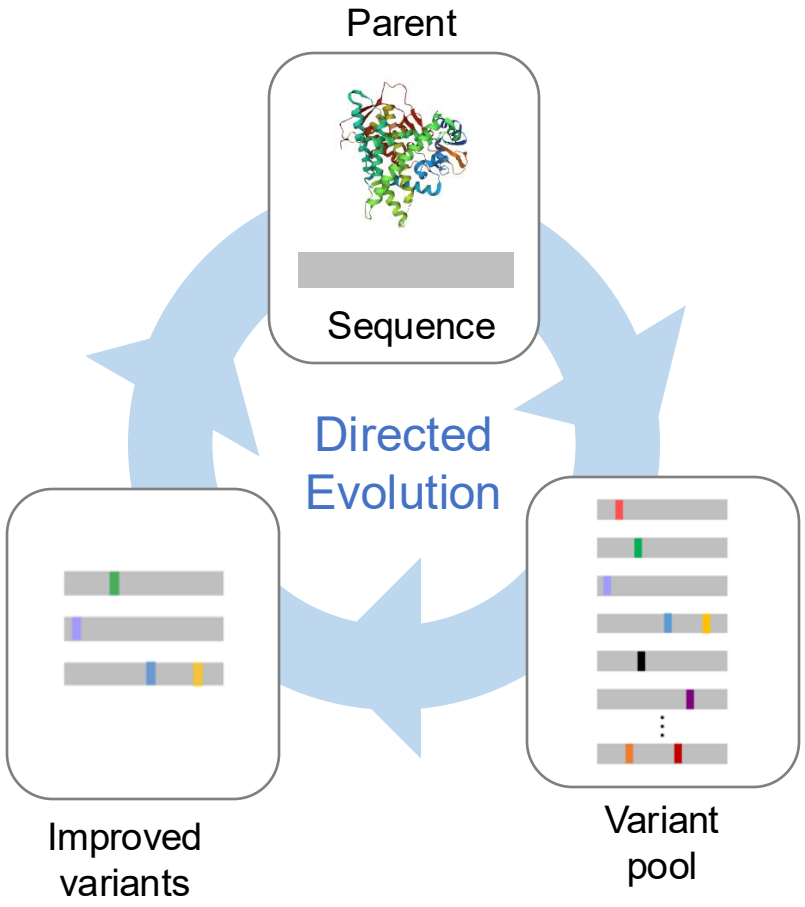


Protein engineering

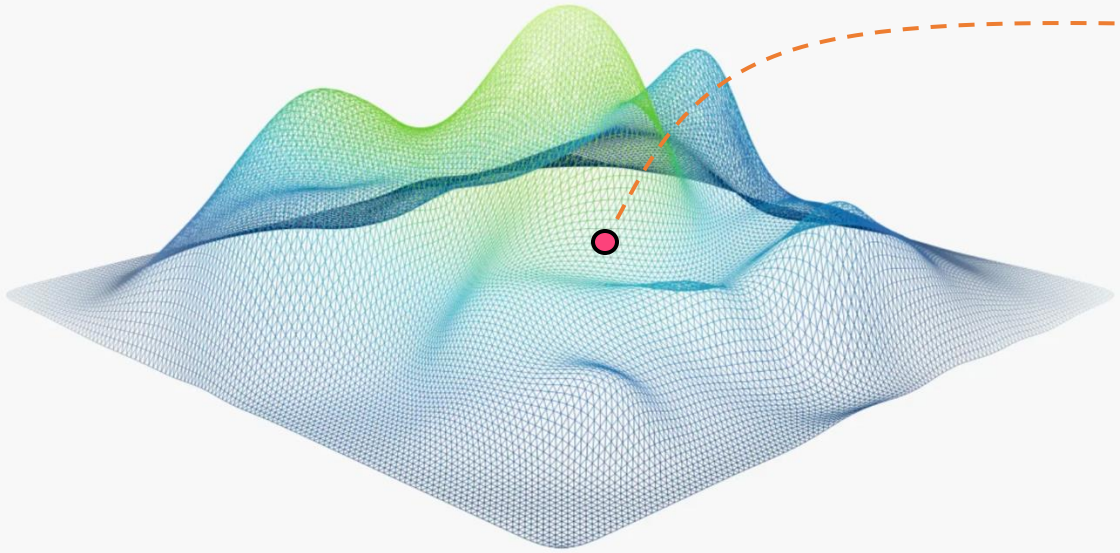
Directed evolution for protein engineering

(2018 Nobel Prize in Chemistry)

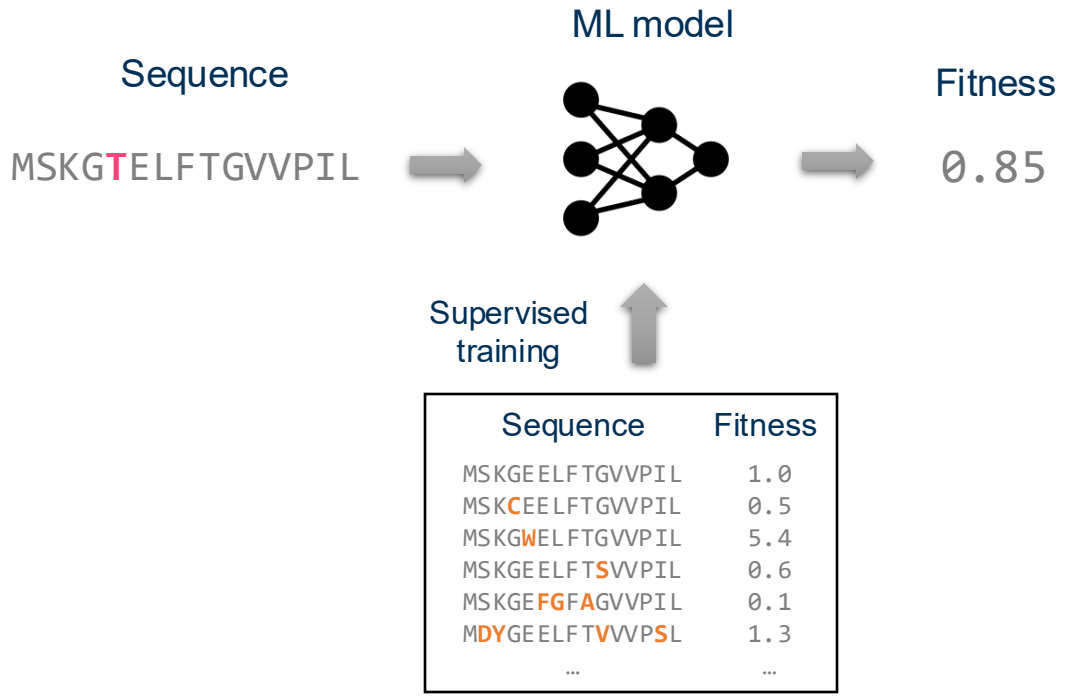
Frances Arnold
(Caltech)



ML-guided directed evolution



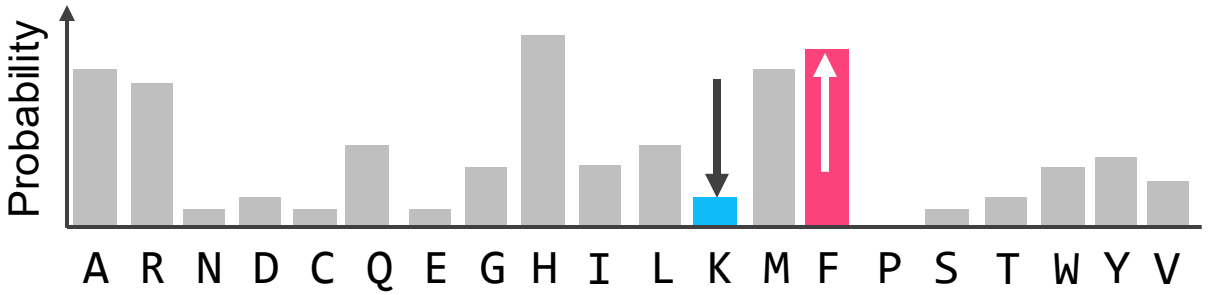
- **Sequence-to-fitness prediction** with machine learning

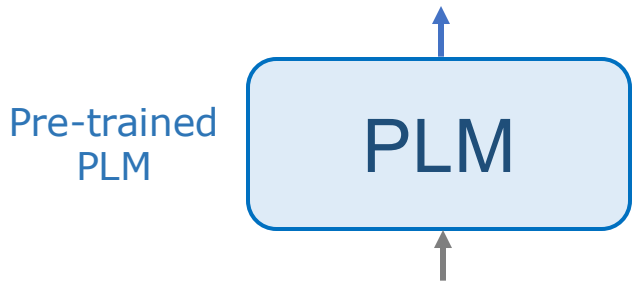


- **Small-data (low-*N*) problem**: labeled data is limited in biology
  - Experiment data: 10^1-10^2 characterized variants
  - ML models: often require >10^5 training samples

Our idea: **Calibrating** PLMs

- **Observation**: PLM already has strong unsupervised fitness prediction performance
- **Hypothesis**: Calibrate PLM with fitness data → more accurate supervised prediction!





x^+ = ELTAFLFNMGDHV y^+ = 4.25

x^- = ELTAFLKNMGDHV y^- = 1.06

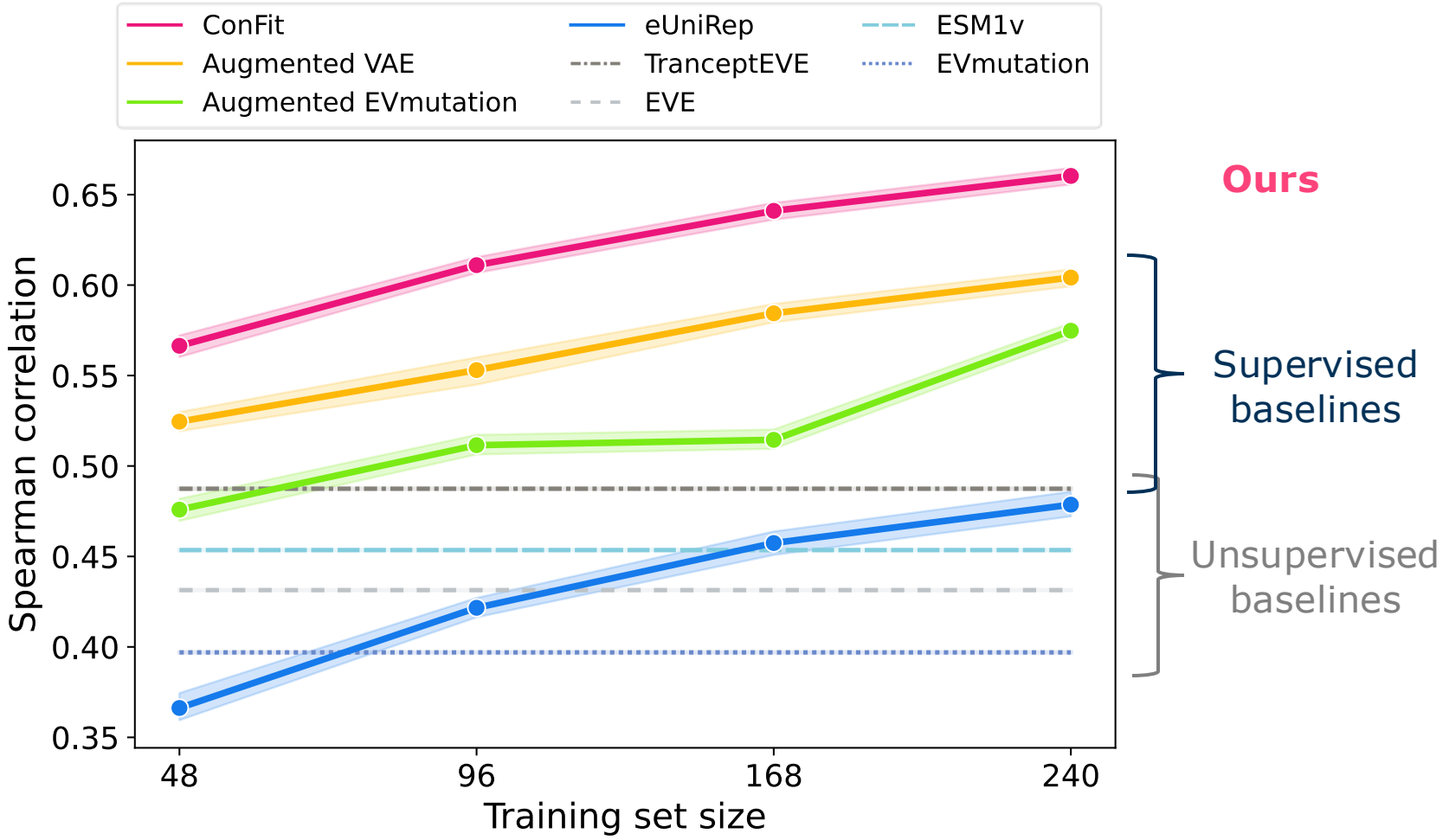
- A **contrastive learning** method:
  Given two samples (x^+, y^+) and (x^-, y^-). When $y^+ > y^-$, we want to have $f_\theta(x^+) > f_\theta(x^-)$

PLM's calibrated probability

Zhao et al., *RECOMB* (2024)

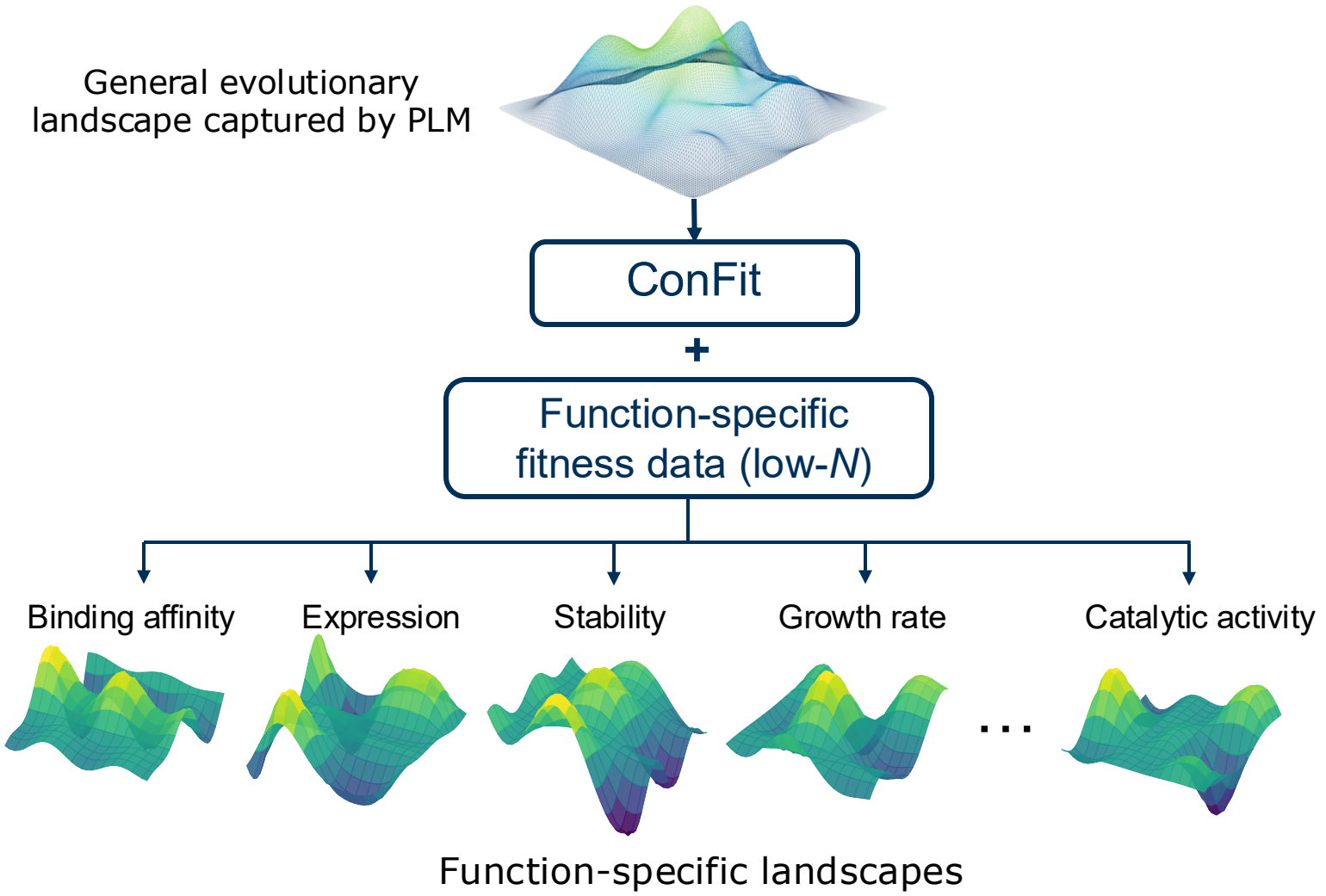
Accurate low-*N* fitness prediction

ConFit outperformed baseline methods across 34 datasets



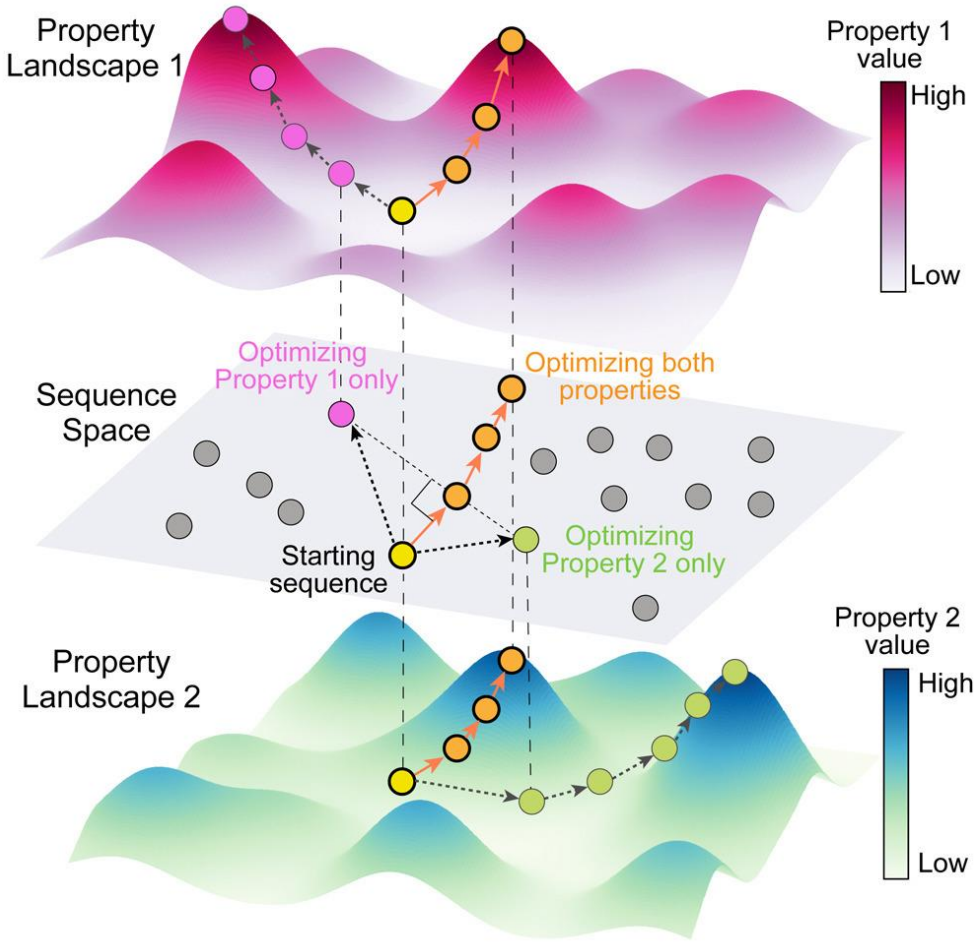
Zhao et al., *RECOMB* (2024)

From general evolutionary landscape to function-specific landscapes



Multi-objective protein sequence design

Recent results



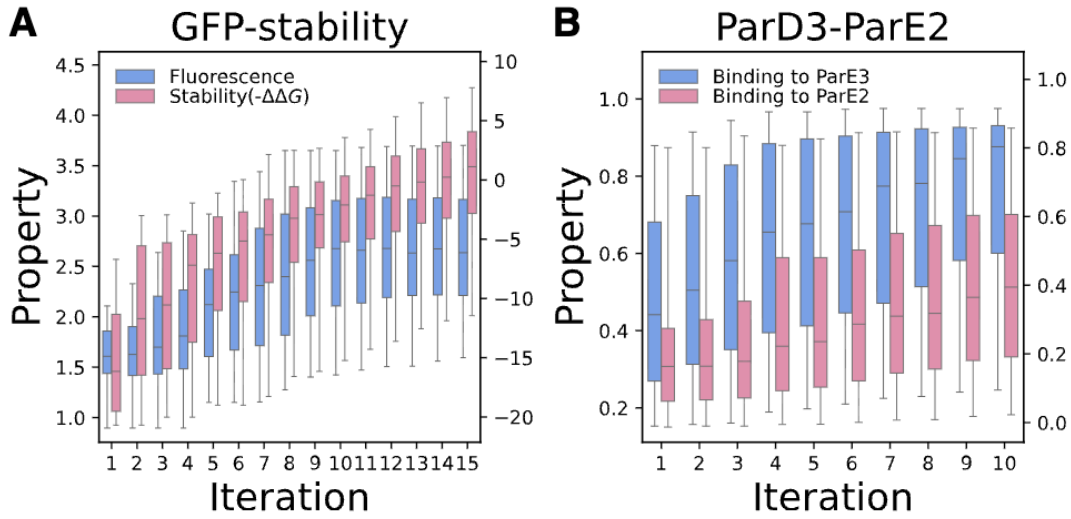
iScience

CellPress
OPEN ACCESS

Article

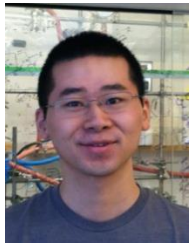
Pareto-optimal sampling for multi-objective protein sequence design

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



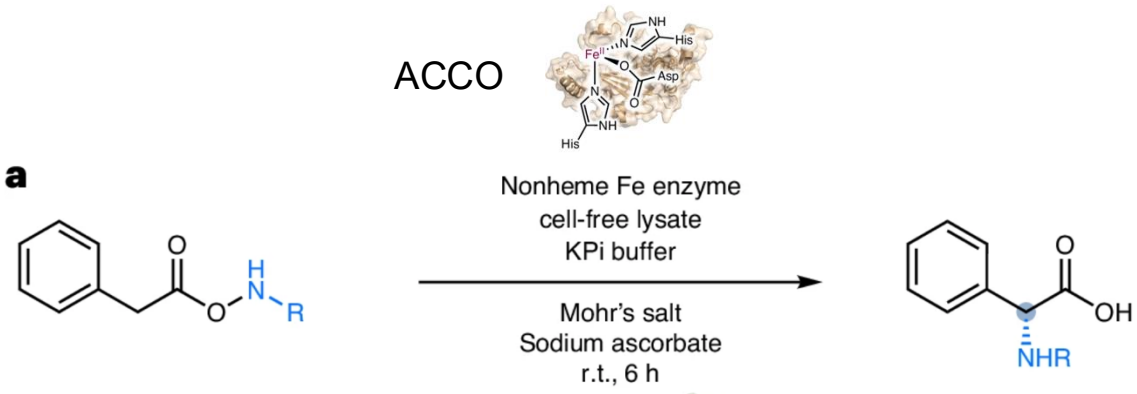
Supported by GaTech IDEaS x Microsoft CloudHub Seed Grant 2024

Experimental validation in biocatalysis

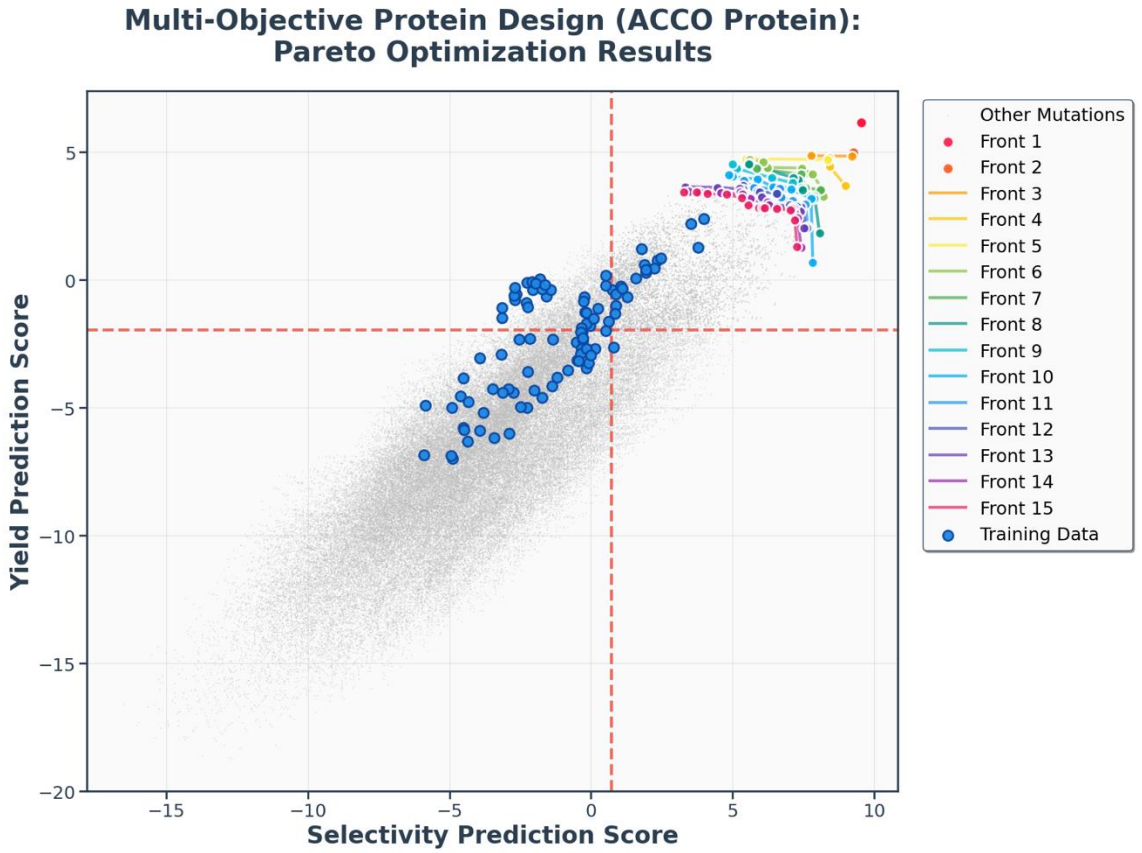


Yang Yang (UCSB Chemistry)

- New-to-nature enzymatic reaction

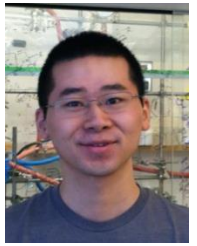


- Two key enzymatic performance metrics:
  - Yield
  - Selectivity

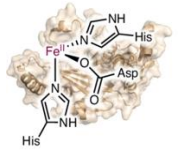


Work in Progress;
Supported by GaTech IDEaS x Microsoft CloudHub Seed Grant

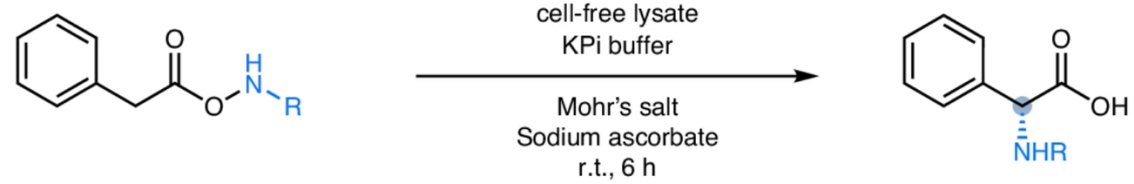
Experimental validation in biocatalysis

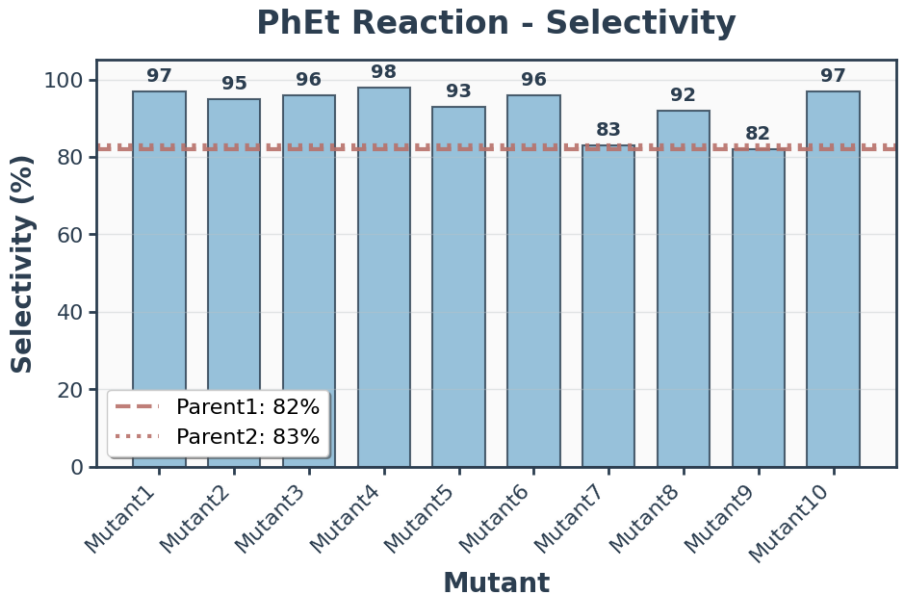


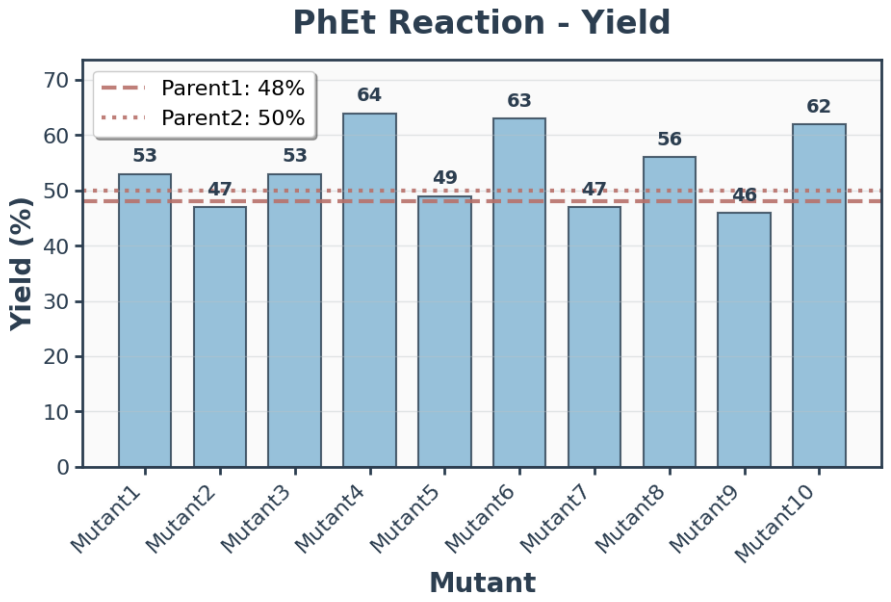
Yang Yang (UCSB Chemistry)



a







Work in Progress;
Supported by GaTech IDEaS x Microsoft CloudHub Seed Grant

Conclusion

- Protein generative AI models capture evolutionary patterns of functional proteins
- Enhanced protein function annotation
- Enabled the design of novel, diverse functional proteins

Acknowledgements

- **Group:** Jiaqi Luo, Ziang Li, Tony Tu, Kerr Ding, Shitong Dai
- **Collaborators:** Huimin Zhao, Yang Yang, Peng Liu, Chao Zhang, Tianhao Yu, Junming Zhao, Liupeng Zhao, 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