Generative AI for Functional Protein Design

Lei Li

Carnegie Mellon University
Language Technologies Institute

Can GenAI design molecules with desired functions?

Medicine

Vaccine

Enzyme - Biocatalysts

Biosensors (e.g. GFP)

New materials

Commonality and Distinction in Language and Molecule Generation

- Modeling
  - Sequence of Discrete Tokens
  - Discrete Structures
  - Geometry (Unique for molecules)
- Training: direct, contrastive, PPO
- Generation
  - Score-conditional Generation
  - Iterative Editing

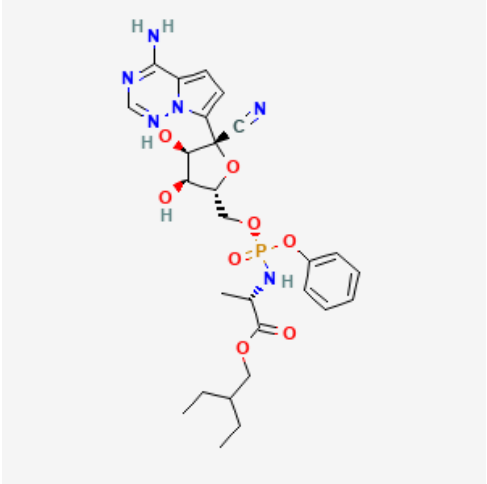
Discrete Sequences of Tokens

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, …

Remdesivir: $C_{27}H_{35}N_6O_8P$

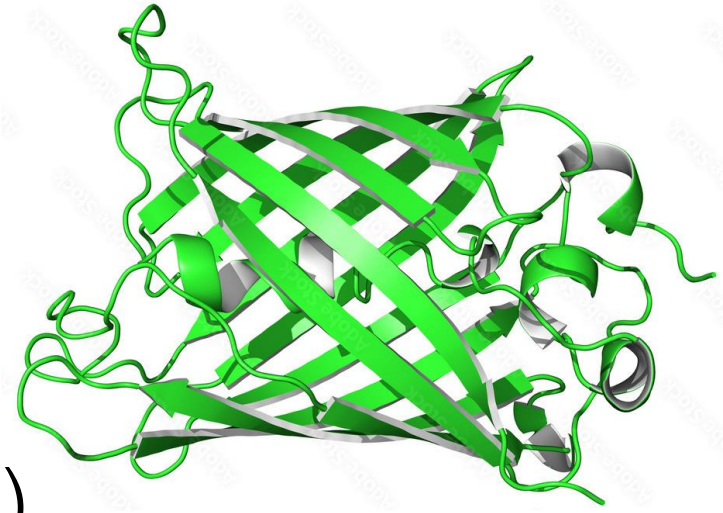
SMLES representation:

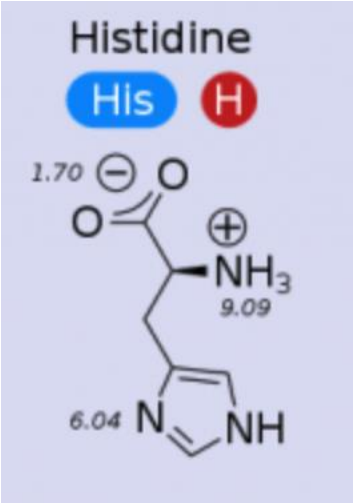
CCC(CC)COC(=O)C(C)NP(=O)(OCC1C(C(C(O1)(C#N)C2=CC=C3N2N=CN=C3N)O)O)OC4=CC=CC=C4

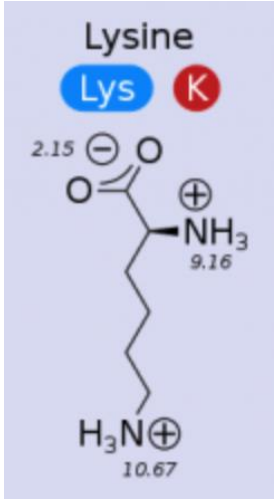


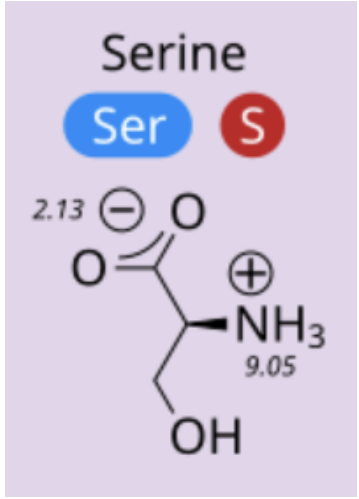
From Human Language to Protein Sequence

- Proteins are building blocks of life
- Important biological functions
- sequence of amino acid residues (20 types)



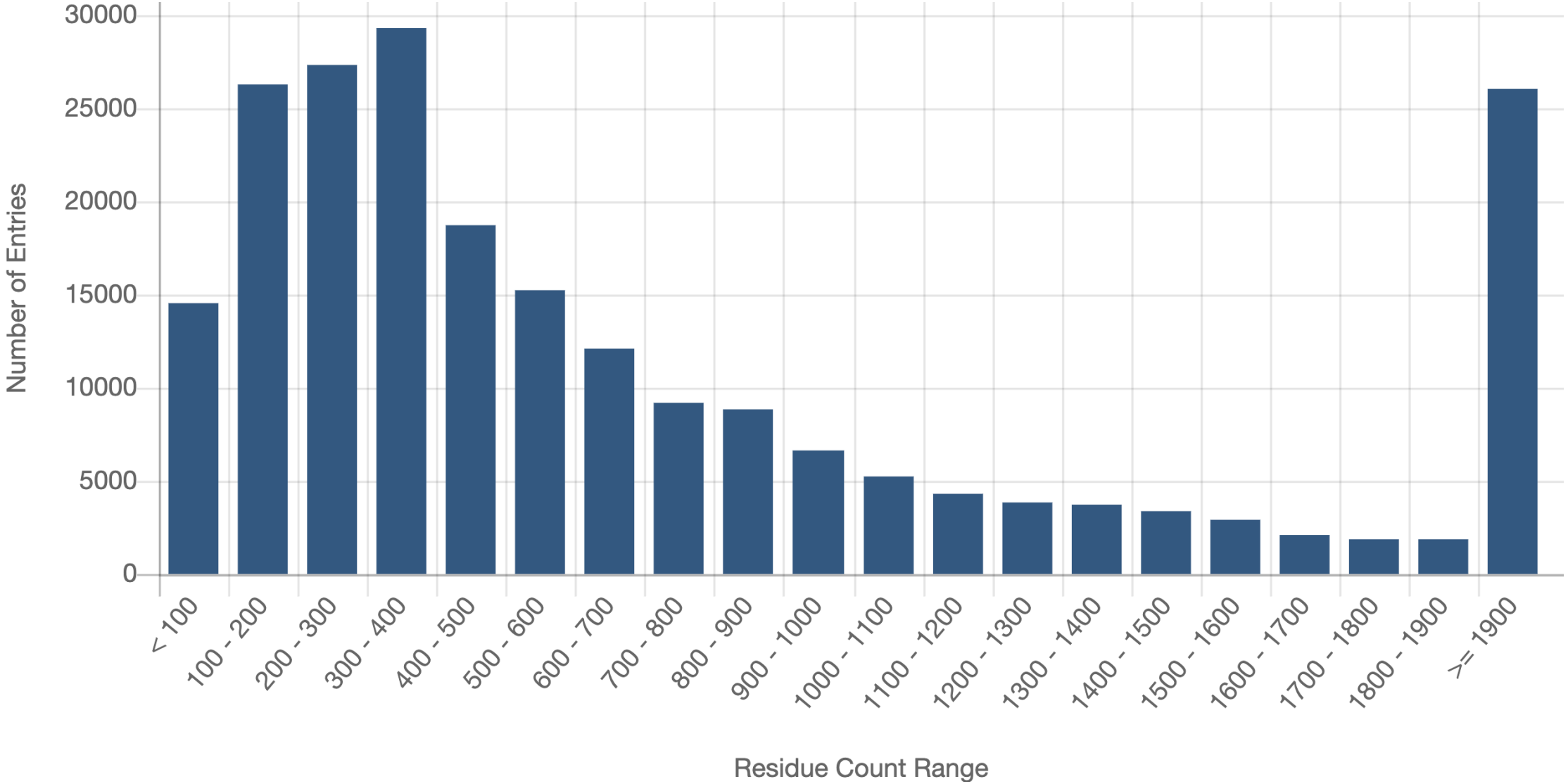




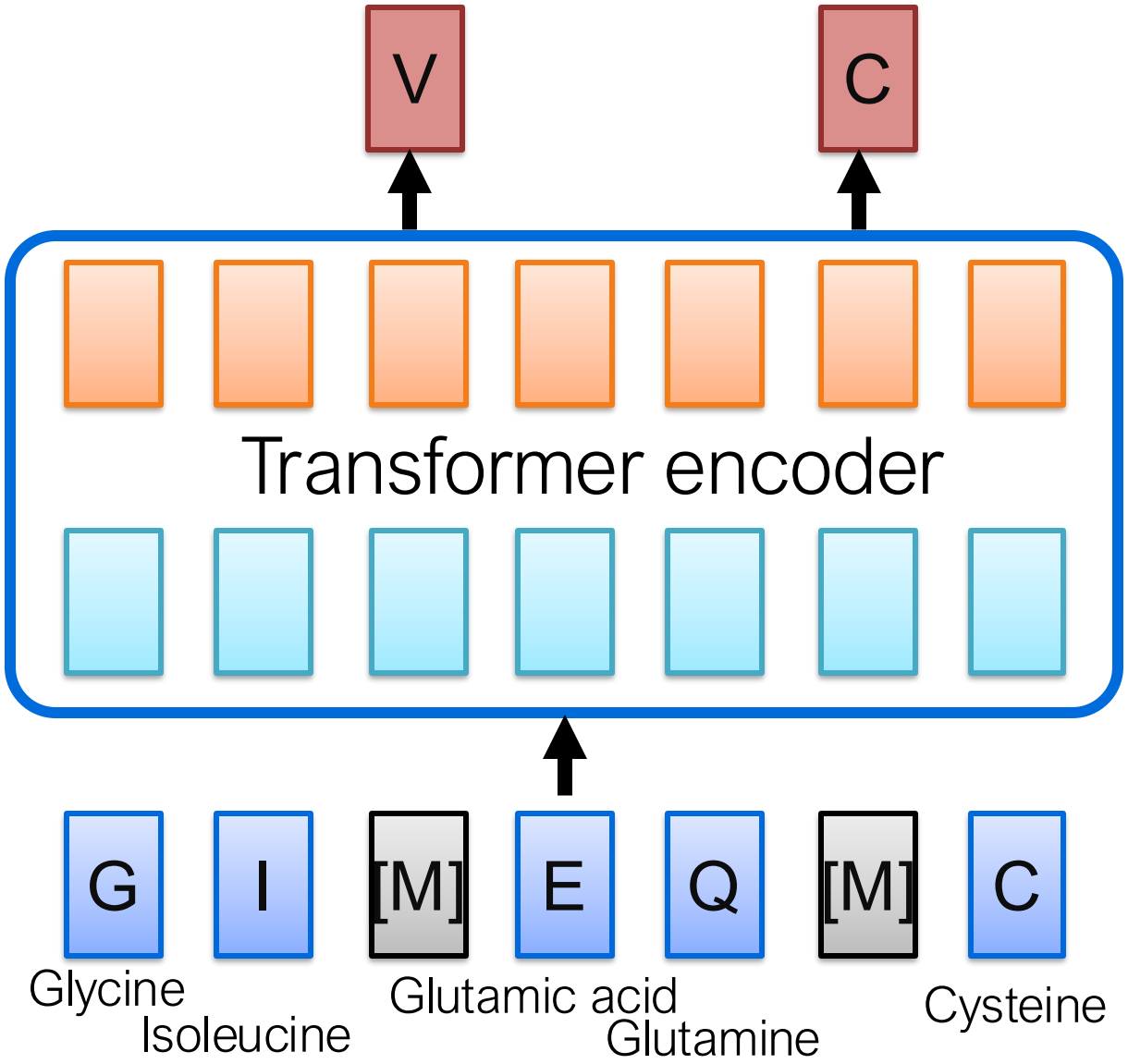


VLLPDNHYLSTQSALSKDPNE
KRDHMVLLEFVTAAGIT

Protein Sequences are much Longer than Text!

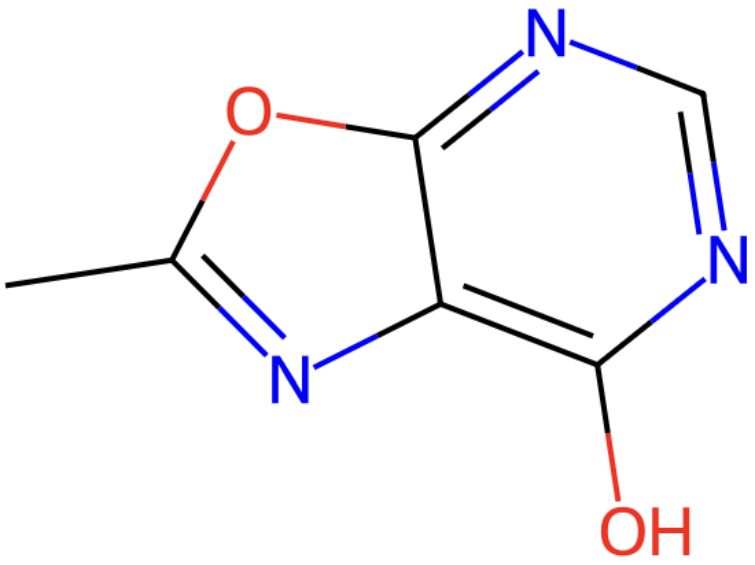


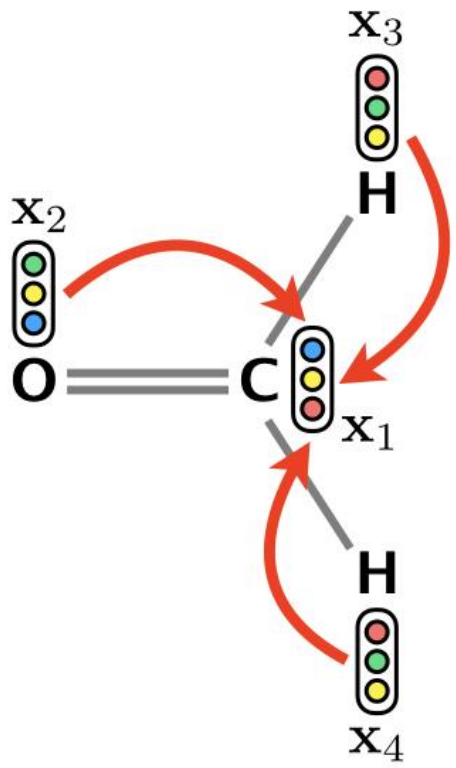
Protein Language Model 1: Mask LM



- Using raw protein sequences for pre-training
  - Training loss: predicting masked residues
- ESM [Meier et al 2021] and ESM-2 [Lin et al 2023]

Graph Neural Network





# of updates

Neural network
e.g., f(x) = ReLU(Wx+b)

$$\mathbf{x}_i^{(\ell+1)} = \mathbf{x}_i^{(\ell)} + \sum_j f\left(\mathbf{x}_j^{(\ell)}\right)$$

i-th atom vector

Neighboring atom vector

$$\mathbf{y} = \sum_i \mathbf{x}_i^{(L)}$$

Last layer

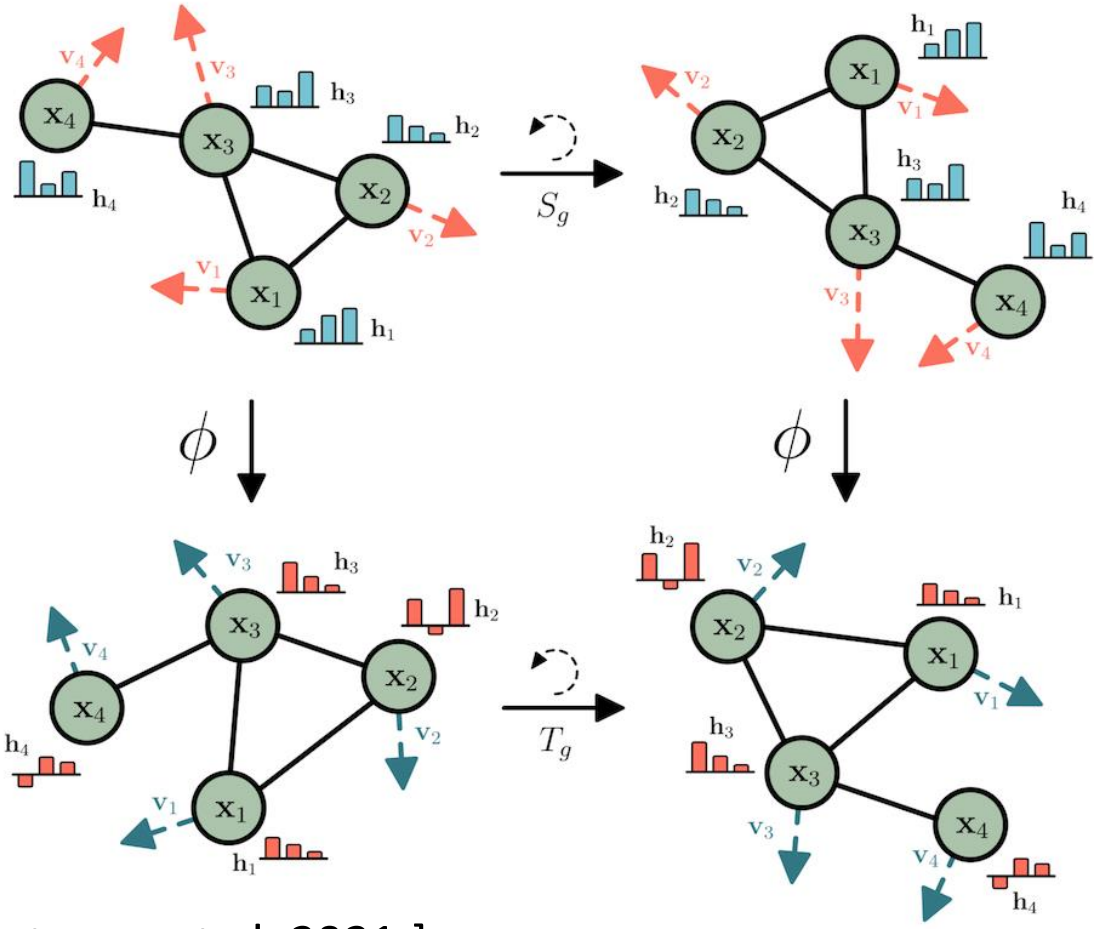
Molecular vector

Fig.1: The update function (or called transition, propagation, message passing, and convolution) in GNNs. On a molecular graph, the **GNN updates each atom vector with its neighboring atom vectors non-linear transformed by neural network.** The molecular vector is obtained by summing (or mean) the atom vectors.

[Tsubaki et al, 2018.]

Modelling Geometry of Molecules

- Equivariant Graph Neural Network (EGNN)



[Satorres et al, 2021.]

Equivariance:

$$f(x) + z = f(x + z)$$

$$\mathbf{m}_{ij} = \phi_e\left(\mathbf{h}_i^l, \mathbf{h}_j^l, \left\|\mathbf{x}_i^l - \mathbf{x}_j^l\right\|^2, a_{ij}\right)$$

$$\mathbf{x}_i^{l+1} = \mathbf{x}_i^l + C\sum_{j\neq i}\left(\mathbf{x}_i^l - \mathbf{x}_j^l\right)\phi_x\left(\mathbf{m}_{ij}\right)$$

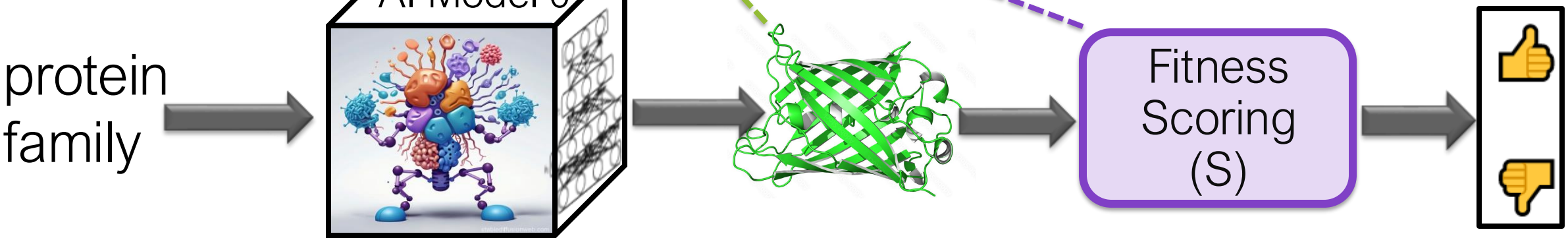
$$\mathbf{m}_i = \sum_{j\neq i}\mathbf{m}_{ij}$$

$$\mathbf{h}_i^{l+1} = \phi_h\left(\mathbf{h}_i^l, \mathbf{m}_i\right)$$

Guiding Protein Generation with Function Fitness

- Fitness functions $P(S|x)$ can be trained using lab data
  - e.g. Green Fluorescent Protein (avGFP) [Sakisyan et al 2016]

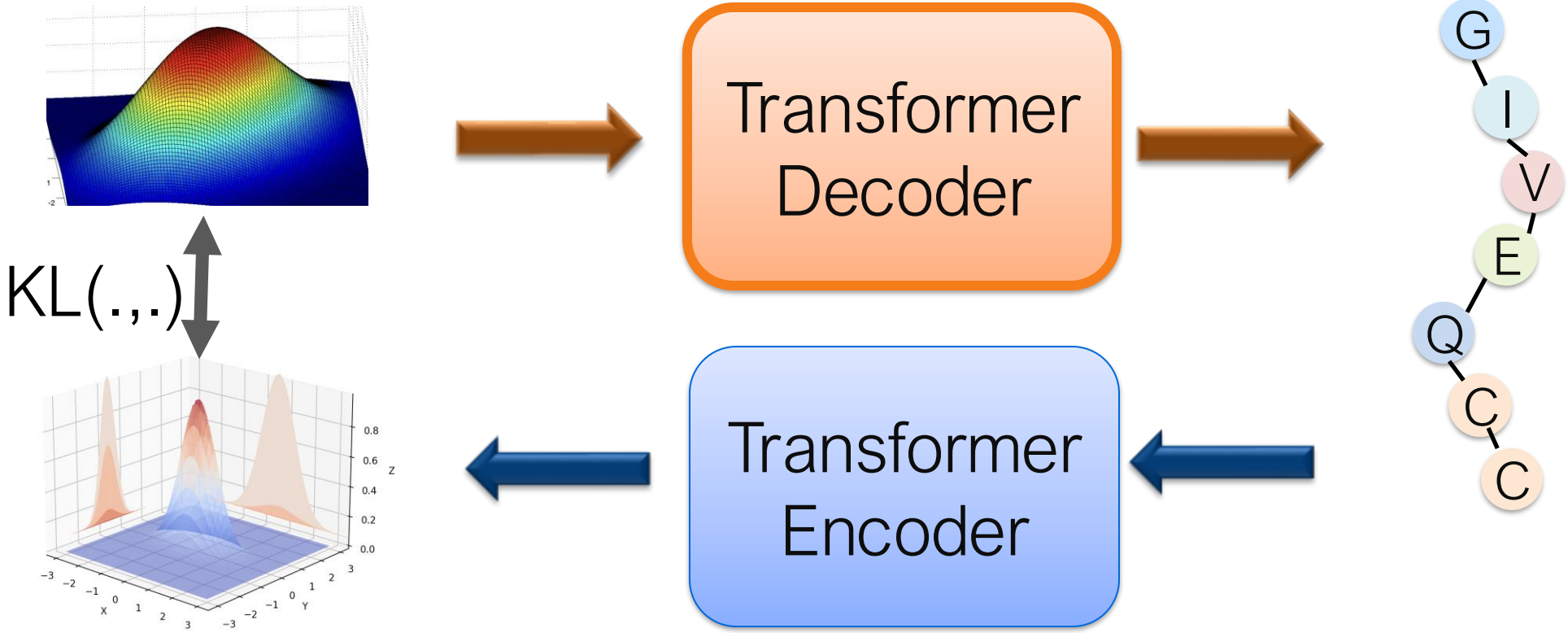
$$\max_x P_\theta(x|S) = \frac{P_\theta(x)P(S|x)}{P_\theta(S)}$$



Brooks et al. Conditioning by adaptive sampling for robust design. ICML 2019.

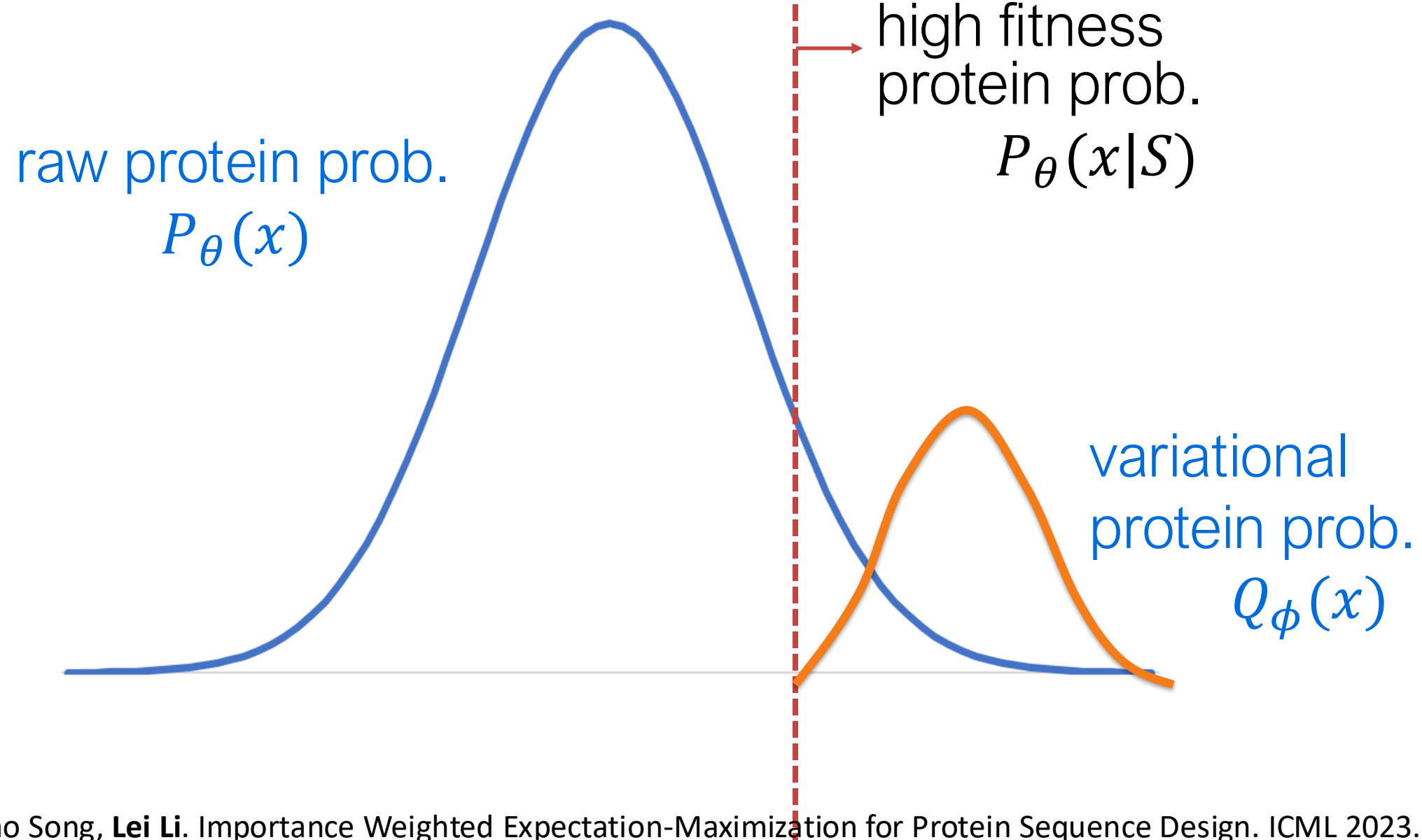
Pre-training Protein Generative Model

$$\max_x P_\theta(x|S) \propto P_\theta(x)P(S|x)$$



- But the generated proteins will have very low fitness score!

IsEMPro: Intuition



Zhenqiao Song, **Lei Li**. Importance Weighted Expectation-Maximization for Protein Sequence Design. ICML 2023.

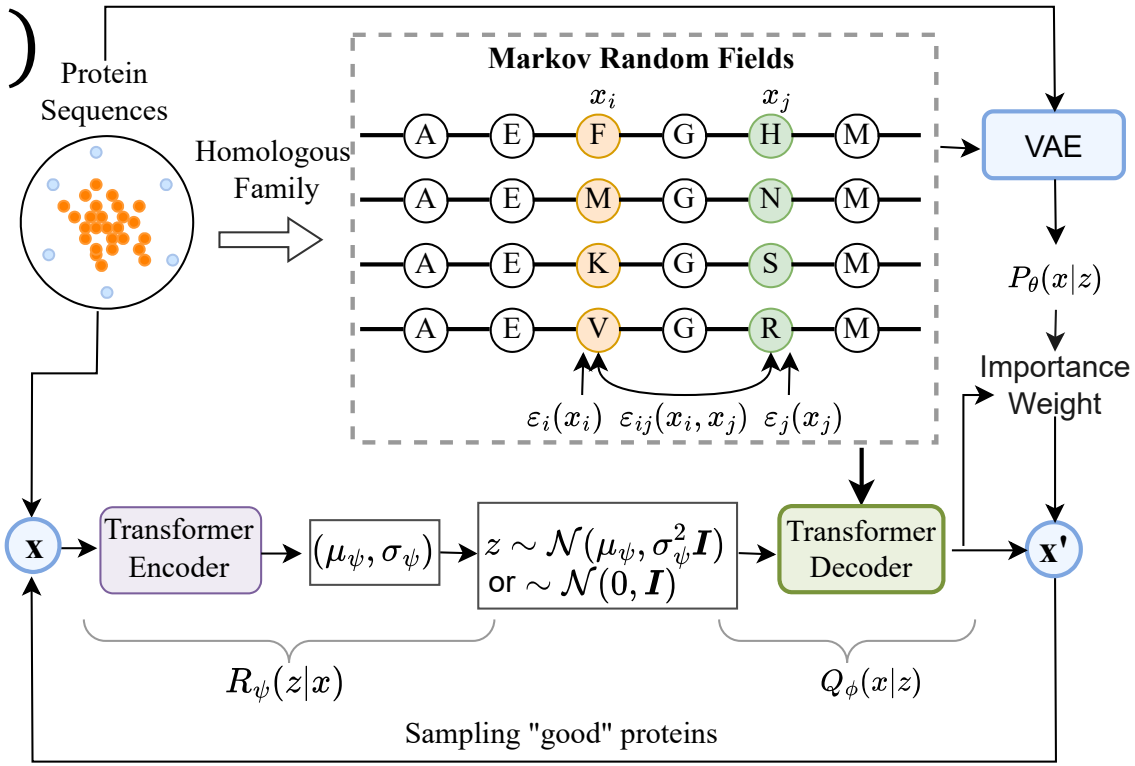
IsEMPro Method

- Intuition:
  - Learning a proposal $Q_\phi(x)$ to approximate distribution of "good" proteins $P_\theta(x|S)$

$$\phi^* = \underset{\phi}{\operatorname{argmax}} -D_{KL}(P_\theta(x|S)||Q_\phi(x))$$

- Model architecture:
  - two VAEs
  - Augmented with MRF features

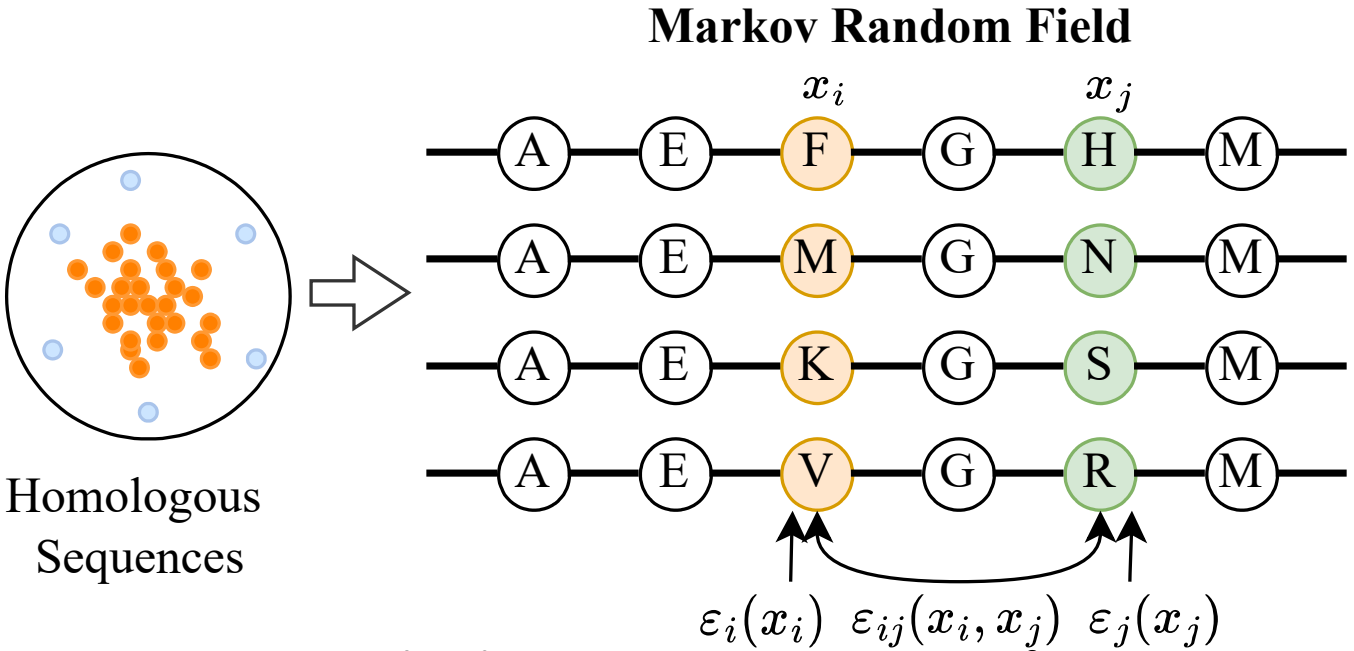
- Expectation-Maximization with Importance Sampling (self-learning)



Zhenqiao Song, **Lei Li**. Importance Weighted Expectation-Maximization for Protein Sequence Design. ICML 2023.

MRF: Learning the Combinatorial Structures of Amino Acids

- These structure constraints are the results of evolutionary process under nature selection
  - Favorable amino-acid combinations
  - Guiding model toward higher fitness landscape



Zhenqiao Song, **Lei Li**. Importance Weighted Expectation-Maximization for Protein Sequence Design. ICML 2023.

Integrating MRF into IsEMPro Generation

- MRFs features (i-th residue)

$$\varepsilon_i(x_i) = [\varepsilon_i(x_i), \varepsilon_{i1}(x_i, a_1.), \ldots, \varepsilon_{iM}(x_i, a_M.)]$$

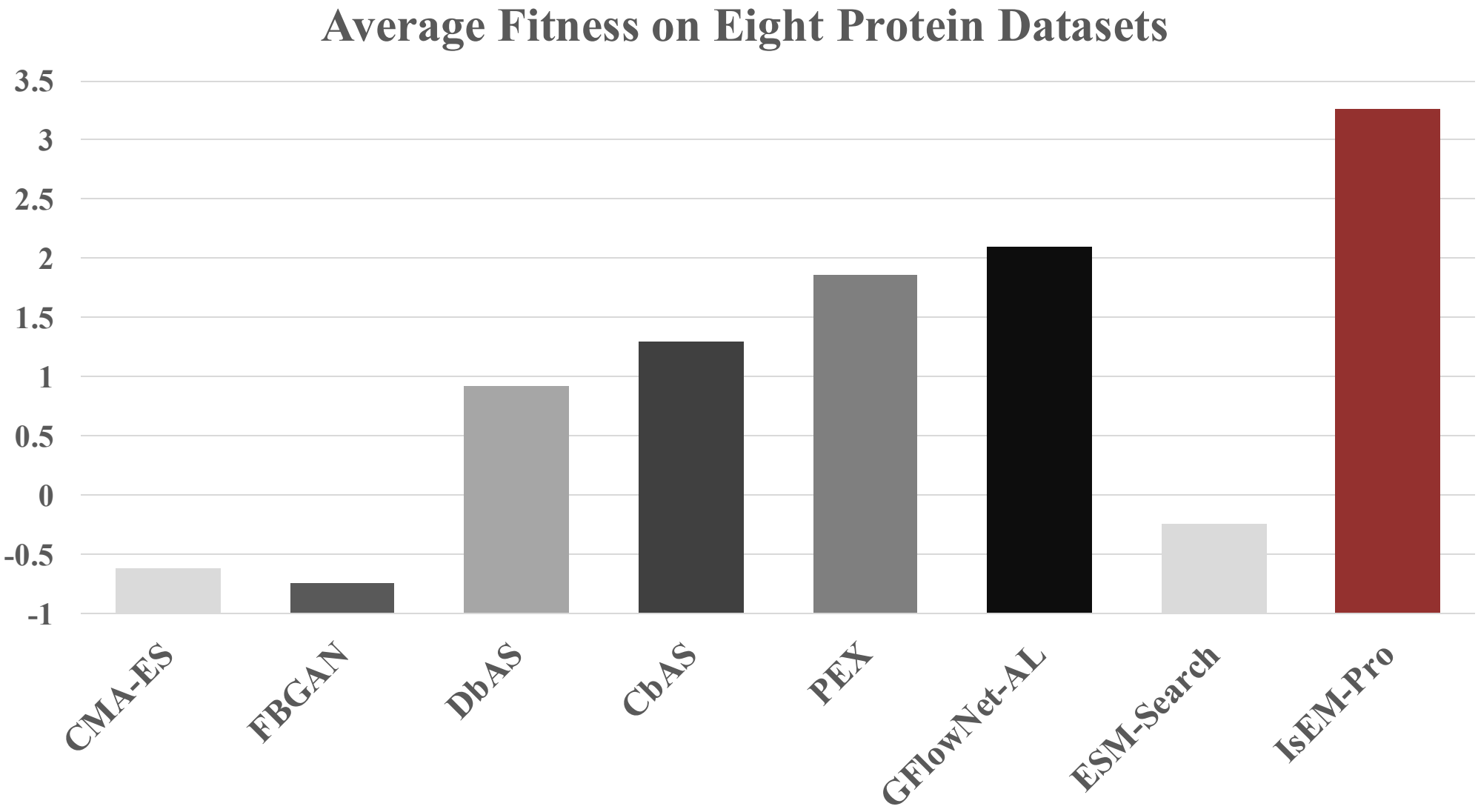
$$\varepsilon_{ij}(x_i, a_j.) = [\varepsilon_{ij}(x_i, a_1), \varepsilon_{ij}(x_i, a_2), \ldots, \varepsilon_{ij}(x_i, a_{20})]$$

- Transformer decoder (autoregressive)
  - First token input: latent vector (learned) $H_0 = \tilde{z}$
  - Other input: combinatorial structure enhanced feature vector

$$H_i = emb(x_{i-1}) + W * \varepsilon_{i-1}(x_{i-1}), 1 \leq i \leq M$$

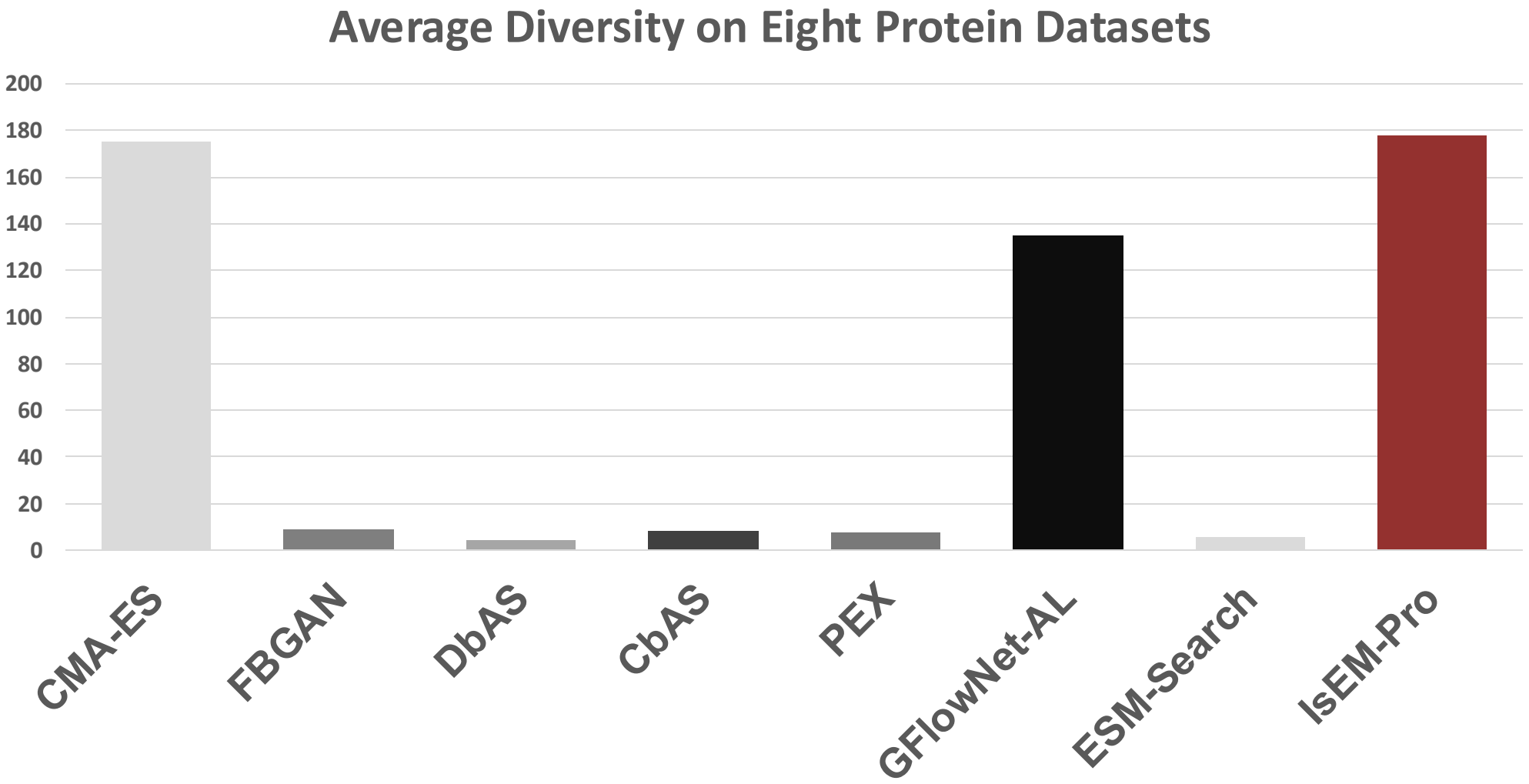
Zhenqiao Song, **Lei Li**. Importance Weighted Expectation-Maximization for Protein Sequence Design. ICML 2023.

IsEM-Pro generates higher-fitness proteins



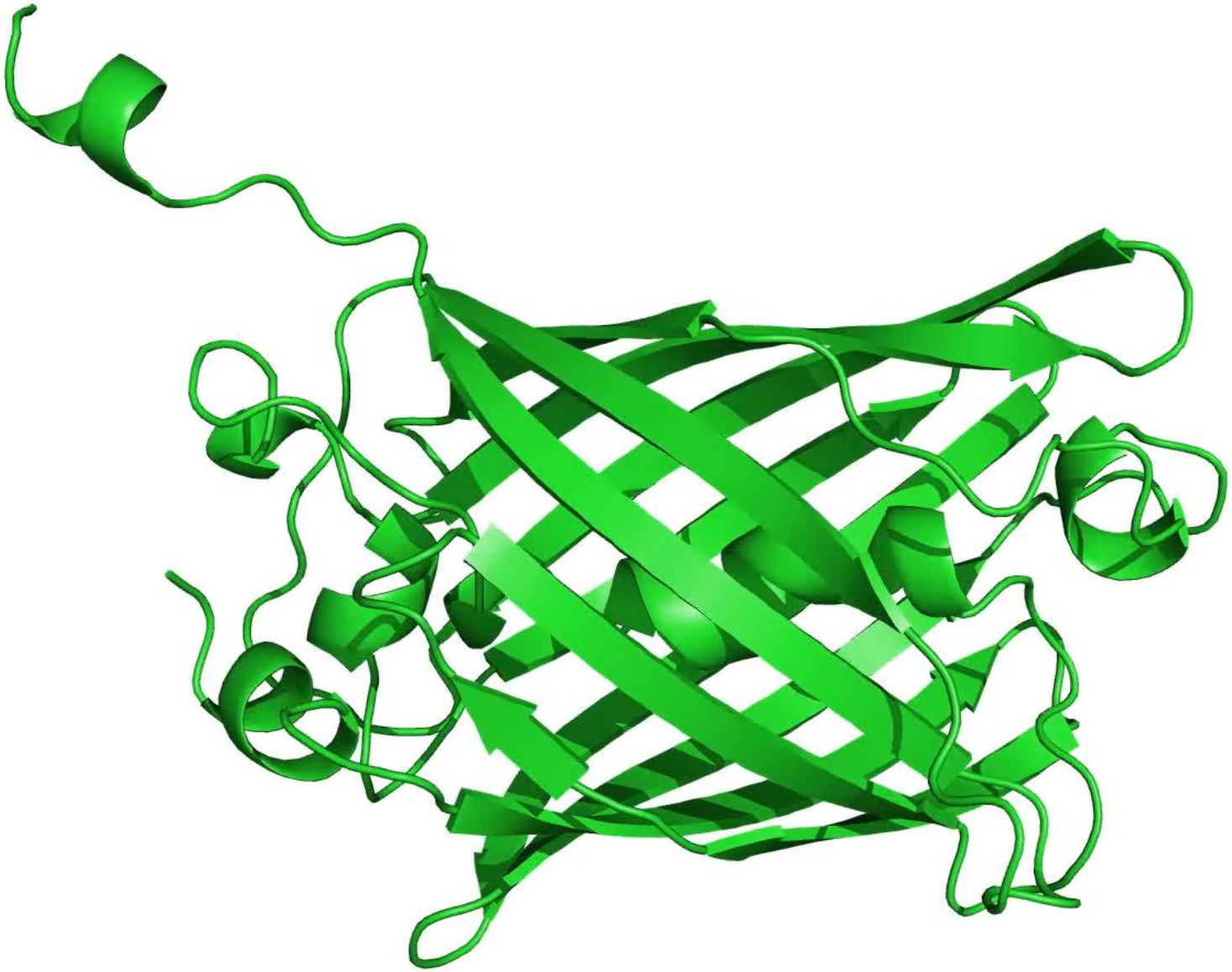
Zhenqiao Song, **Lei Li**. Importance Weighted Expectation-Maximization for Protein Sequence Design. ICML 2023.

IsEM-Pro generates more diverse proteins



Zhenqiao Song, **Lei Li**. Importance Weighted Expectation-Maximization for Protein Sequence Design. ICML 2023.

Green Fluorescent Protein designed by IsEMPro



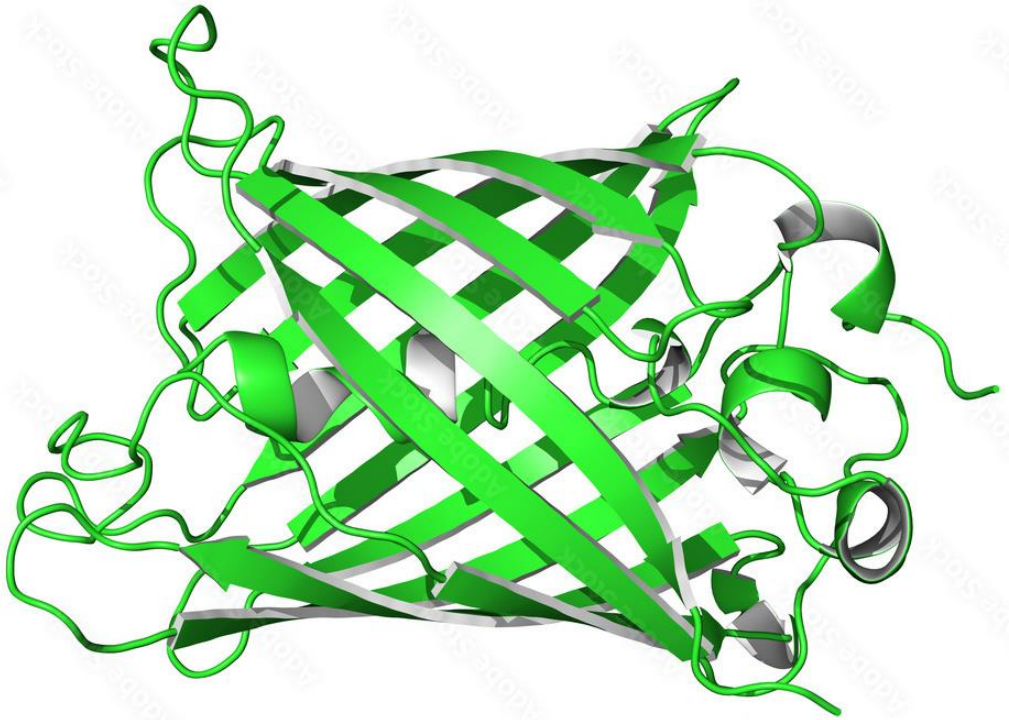
Highlights of IsEM-Pro

- Using importance sampling inside the EM is efficient to generate functional proteins
- The combinatorial enhanced latent generative model boosts diverse and novel protein sequences
- The self-learning process helps to find proteins with higher fitness scores

Zhenqiao Song, **Lei Li**. Importance Weighted Expectation-Maximization for Protein Sequence Design. ICML 2023.

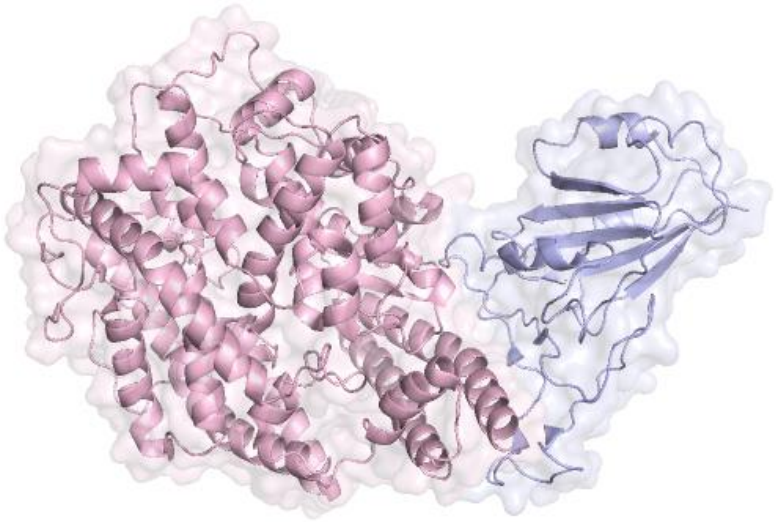
Protein Design Approaches

- Sequence-based Generation
- Structure-based Generation
  - Secondary structure-based
  - Inverse Folding
  - Surface geometry
- Sequence-Structure Co-design
  - Protein monomer
  - Protein complex



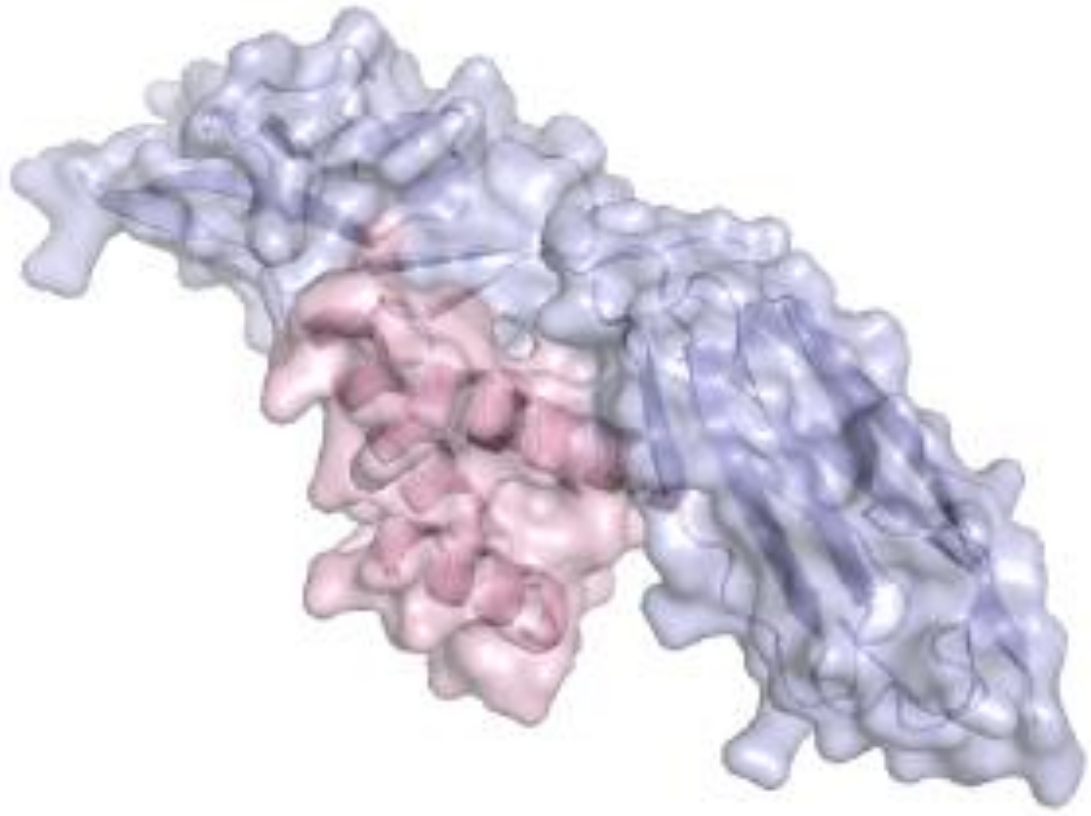
Protein-Protein Complex

- A molecular assembly formed when two or more protein molecules interact and bind together
  - Covid19 Sars-Cov2 ACE2 complex
  - Biomedicine

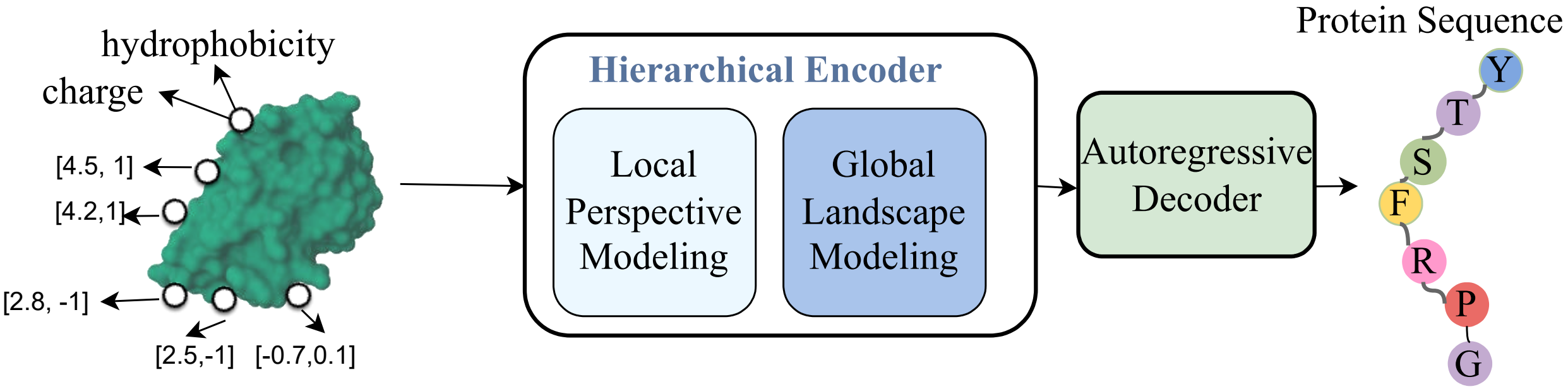


Surface-based Protein Design

- Intuition: fill in the content given an outline
- Complementary shapes
- Poorly placed charges, polarity or hydrophobicity prevents molecule binding



SurfPro Method



Song, Huang, **Li**, Jin. SurfPro: Functional Protein Design Based on Continuous Surface. ICML 2024.

Protein Surface Construction

- MSMS tool
  - Transform a PDB file into a point cloud ➔ molecule surface
  - Each vertex contains
    - 3D coordinates
    - [hydrophobicity, charge]

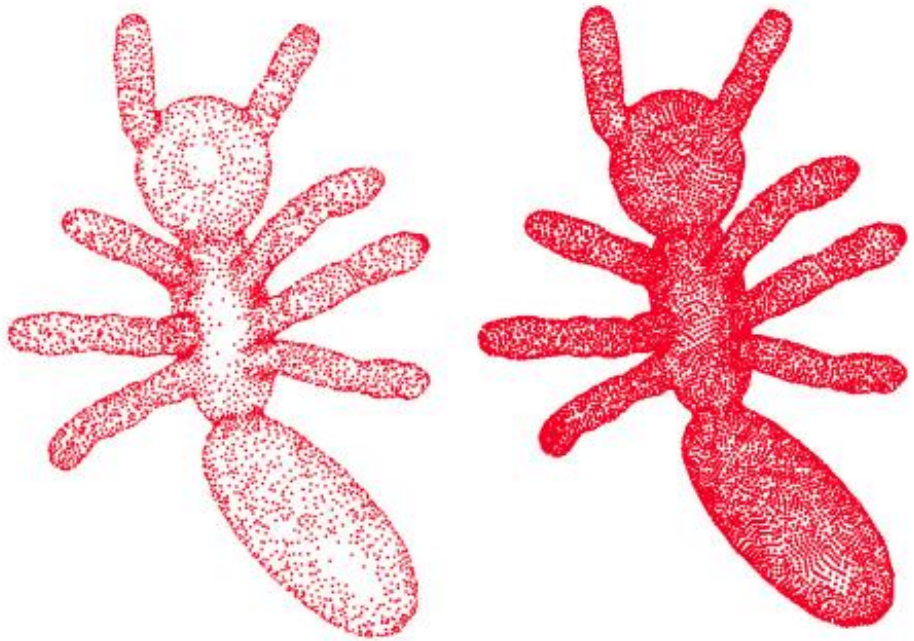
From IMGT physicochemical classes

Feature	Description	Value
hydrophobicity	The hydrophobicity level of a residue, the higher the hydrophobicity, the more hydrophobic the residue	I: 4.5 V: 4.2 L: 3.8 F: 2.8 C: 2.5 M: 1.9 A: 1.8 W: -0.9 G: -0.4 T: -0.7 S: -0.8 Y: -1.3 P: -1.6 H: -3.2 N: -3.5 D: -3.5 Q: -3.5 E: -3.5 K: -3.9 R: -4.5
charge	The charge value of a residue	R: 1 K: 1 D: -1 E: -1 H: 0.1 Others: 0

Surface Construction

- Surface smoothing ➔ Compression using octree
  - Gaussian kernel smoothing – higher expressiveness

$$x_i' = \sum_{x_j \in N(x_i)} \frac{\kappa(x_i, x_j) x_i}{\sum_{x_t \in N(x_i)} \kappa(x_i, x_t)}, \qquad \kappa(x, y) = e^{-\frac{(x-y)^2}{\eta}}$$



Song, Huang, **Li**, Jin. SurfPro: Functional Protein Design Based on Continuous Surface. ICML 2024.

Hierarchical Encoder: Local Perspective Modelling

- K-nearest equivariant graph convolutional layers
  - Local Message

$$m_{ij} = SiLU(\phi_e([h_i^l; h_j^l; ||x_i' - x_j'||_2]))$$

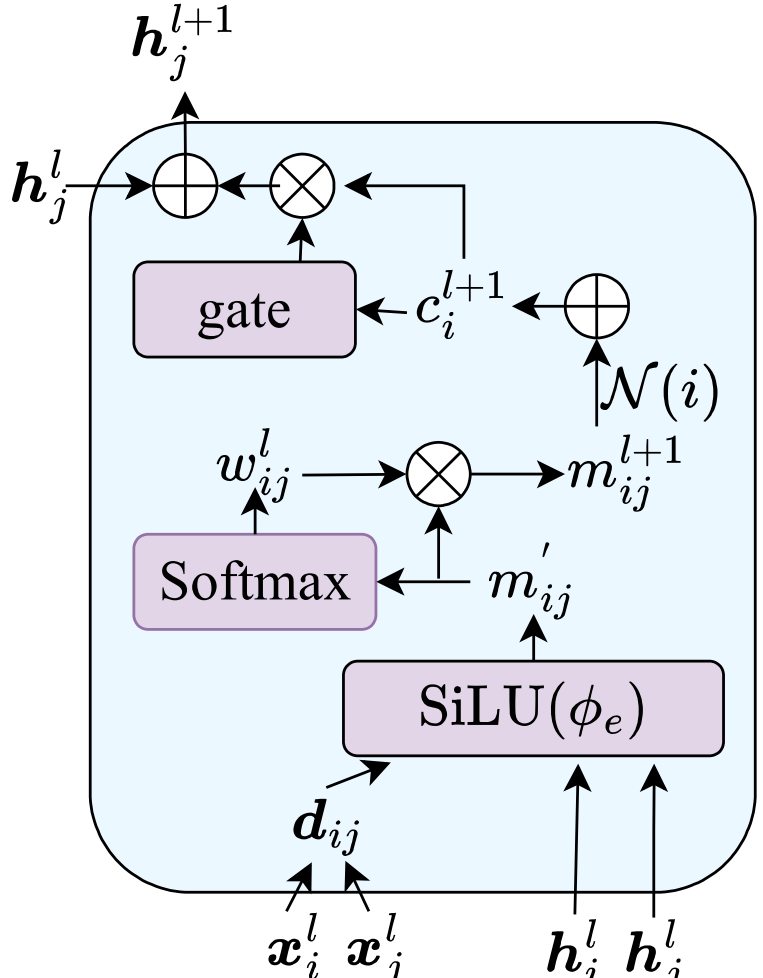
$$w_{ij}^l = \frac{\exp(W_s^l m_{ij}' + b_s^l)}{\sum_{k \in N(x_i)} \exp(W_s^l m_{ik}' + b_s^l)}$$

$$m_{ij}^{l+1} = w_{ij}^l * m_{ij}'$$

  - Vertex feature representation

$$c_i^{l+1} = \sum_{j \in N(x_i)} m_{ij}^{l+1}$$

$$h_i^{l+1} = h_i^l + gate(c_i^{l+1}) \odot c_i^{l+1}$$



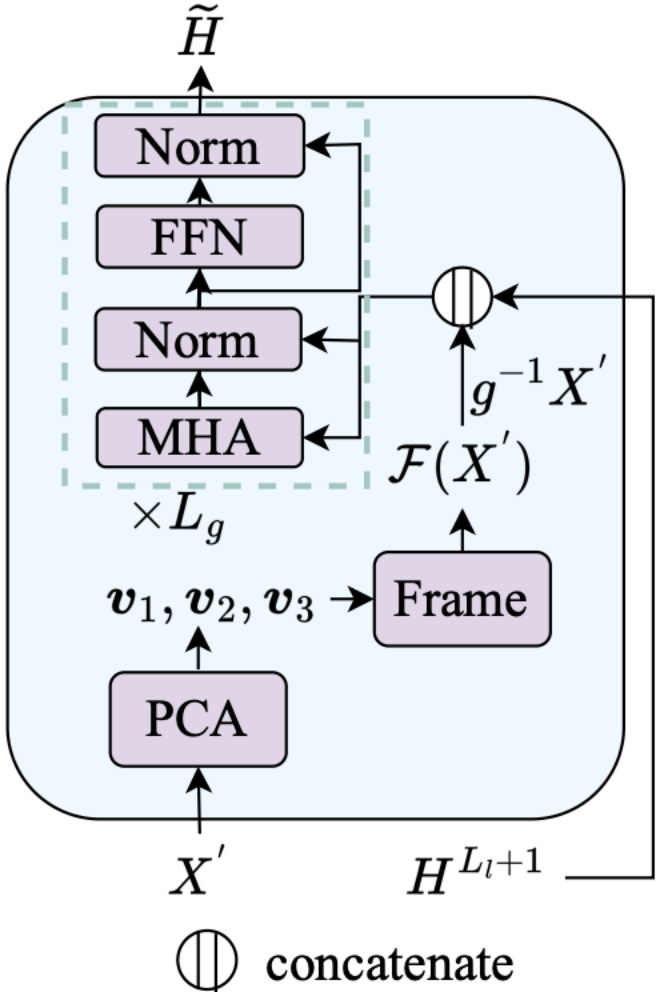
Song, Huang, **Li**, Jin. SurfPro: Functional Protein Design Based on Continuous Surface. ICML 2024.

Hierarchical Encoder: Global Landscape Modelling

- Frame Calculation
  - Point cloud X – PCA – three principle componen
    - Map a 3D molecule into the 8 coordinate systems

$$\mathrm{F}(X') = \{([\alpha_1 v_1, \alpha_2 v_2, \alpha_3 v_3], t) | \alpha_i \in \{-1, +1\}\}$$

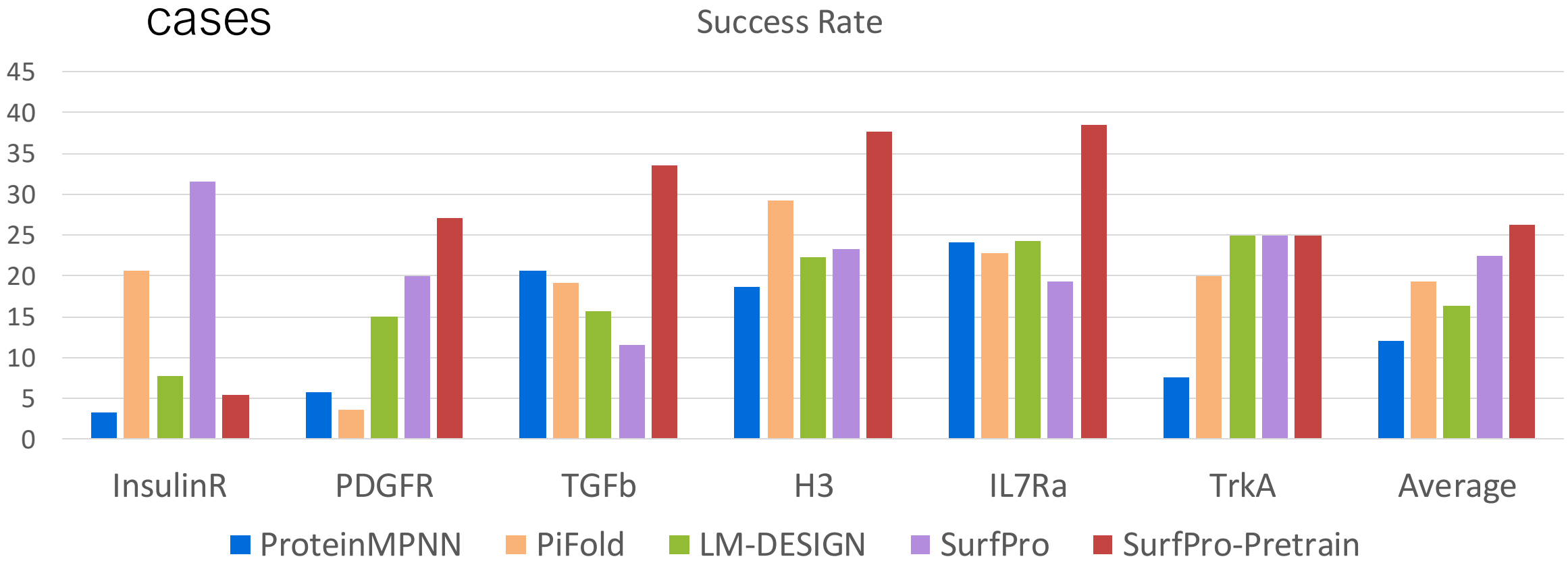
    - Average the representations across 8 frames
    - Equal to any translation + Rotation operation theoretic
- Global Landscape Modeling - MHA
- Autoregressive decoder
  - Maximum likelihood optimization



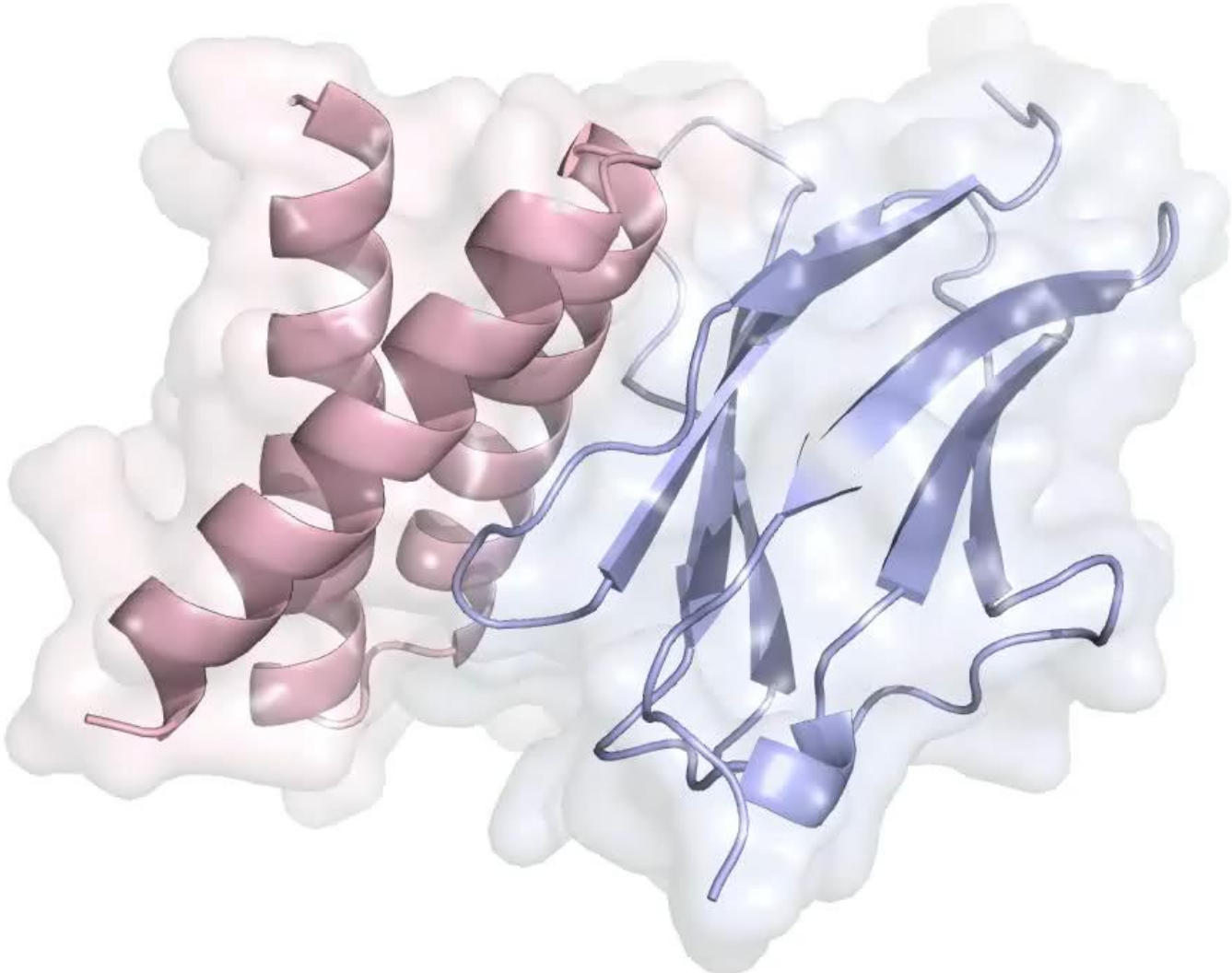
Song, Huang, **Li**, Jin. SurfPro: Functional Protein Design Based on Continuous Surface. ICML 2024.

SurfPro generates more successful binders

- Six target proteins
  - Three are used as supervised cases; three are used as zero-shot cases



Song, Huang, **Li**, Jin. SurfPro: Functional Protein Design Based on Continuous Surface. ICML 2024.



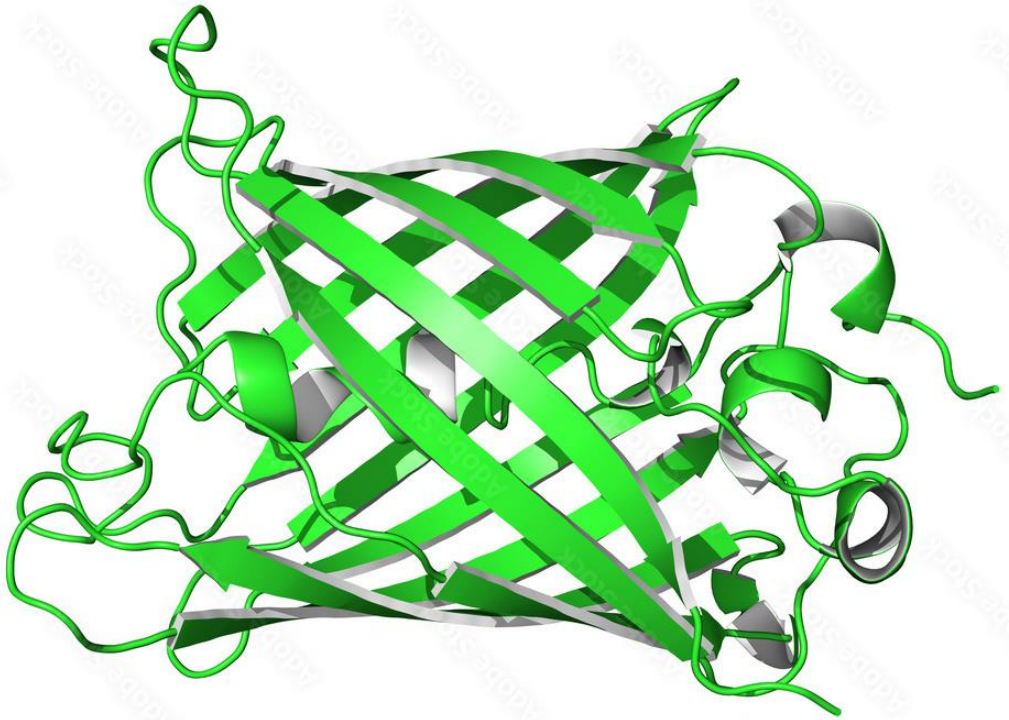
Highlights of SurfPro🏄

- Designing proteins based on
  - surface geometry
  - chemical property on the surface
- Effective in Binder-design, inverse-folding, and enzyme design tasks

Song, Huang, **Li**, Jin. SurfPro: Functional Protein Design Based on Continuous Surface. ICML 2024.

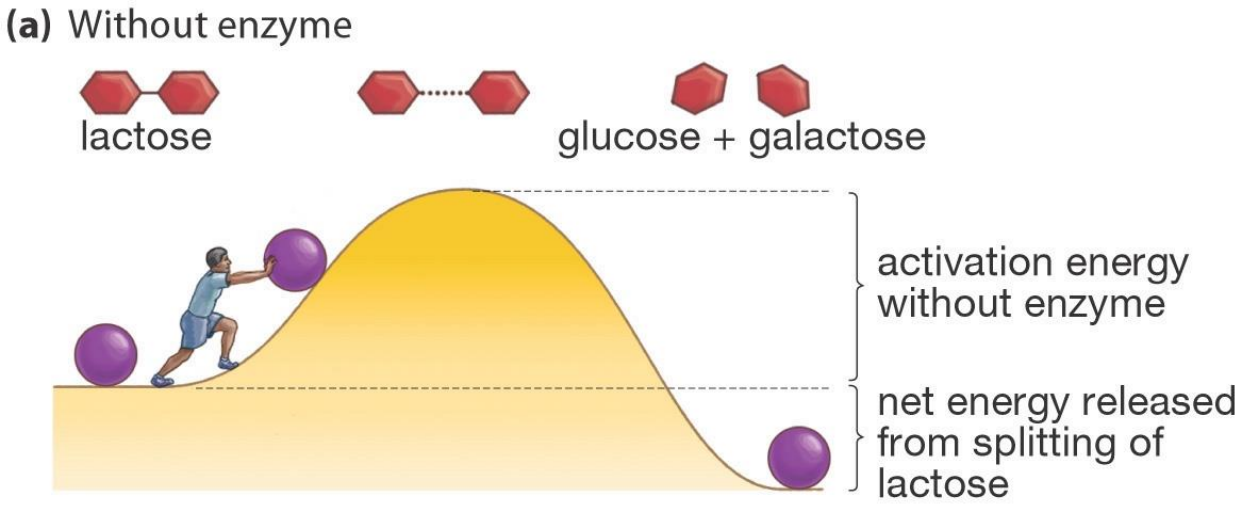
Protein Design Approaches

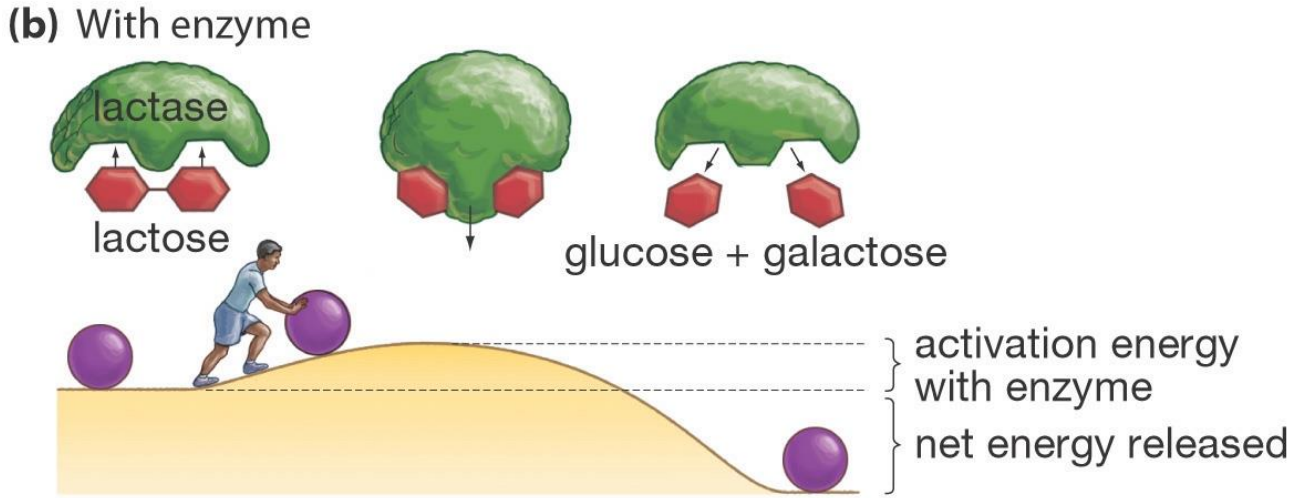
- Sequence-based Generation
- Structure-based Generation
  - Secondary structure-based
  - Inverse Folding
  - Surface geometry
- Sequence-Structure Co-design
  - Protein monomer
  - Protein complex



Enzyme

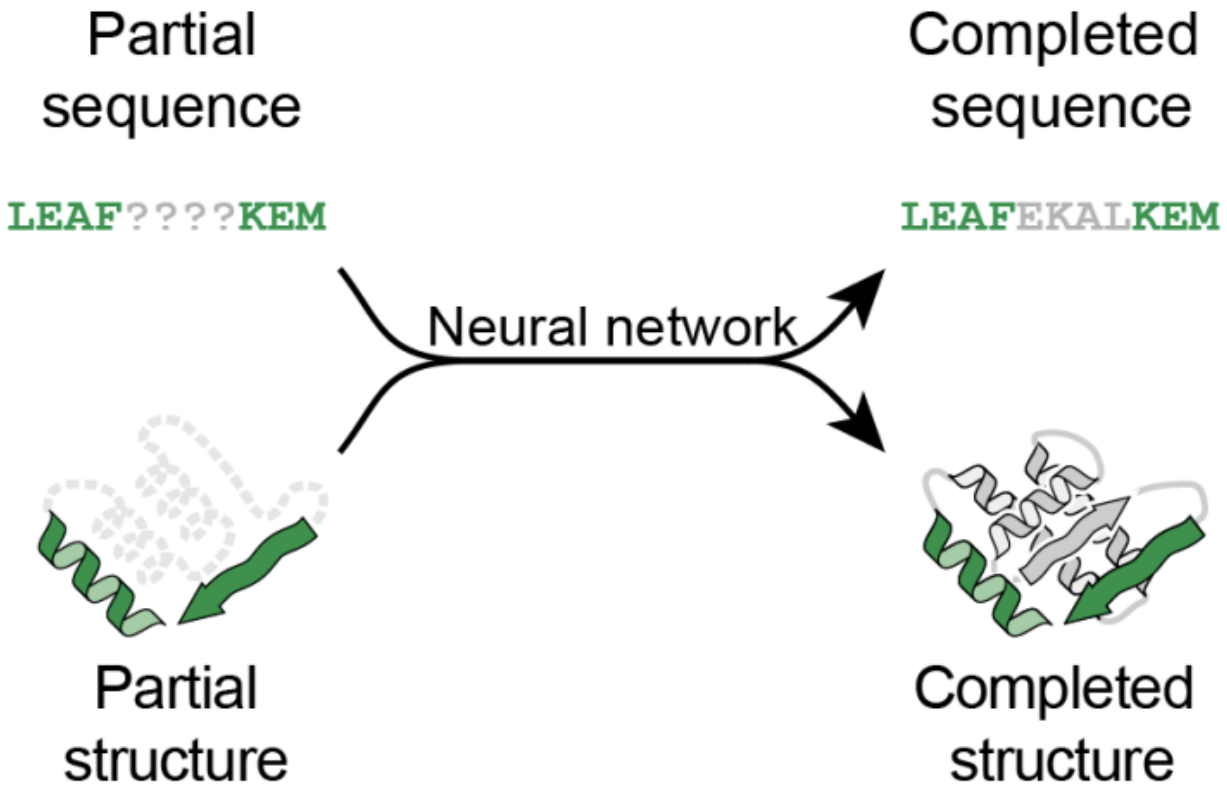
- biological catalyst to accelerate chemical reactions
  - Enzymes reduce a reaction's activation energy





Motivation 1: How to design desired enzymes?

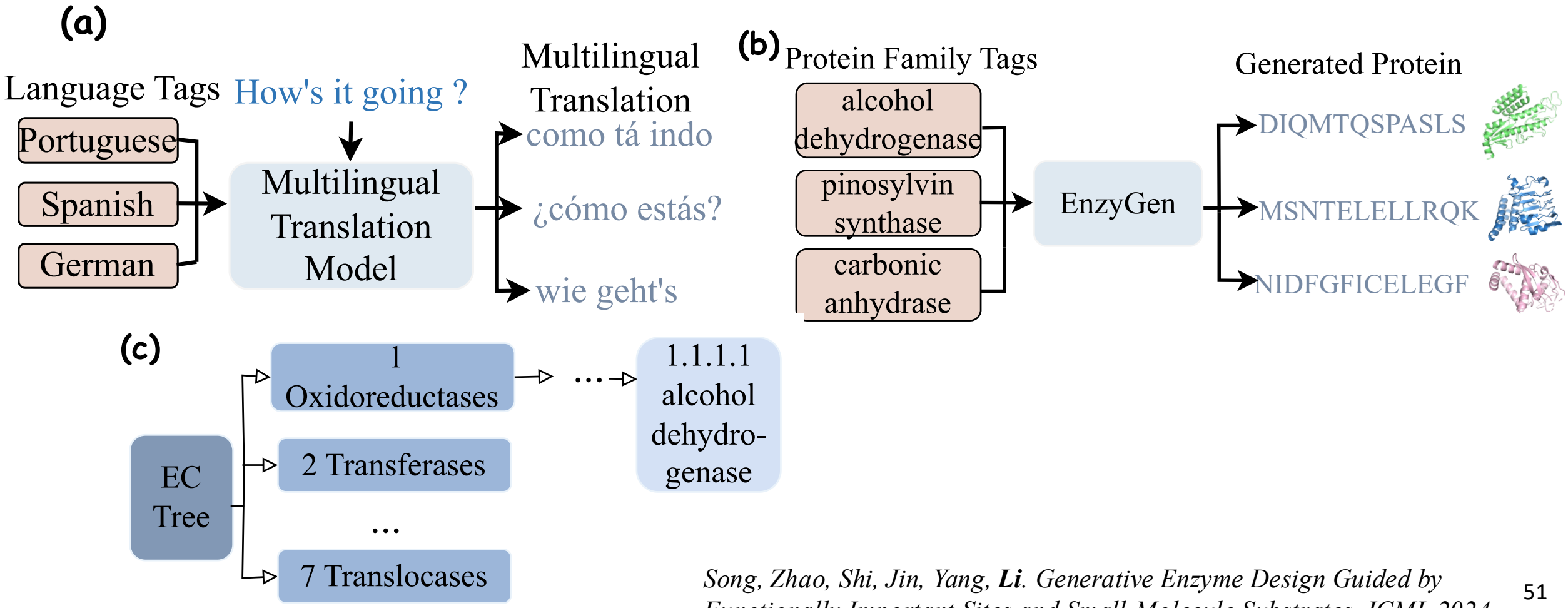
- Functional Important Sites (Motif)
  - Active sites – Binding to substrates



Scaffolding protein functional sites using deep learning. Wang et al, Science 2022.

Motivation 2: How to design desired enzymes?

Enzyme classification tree indicates enzymatic reaction type

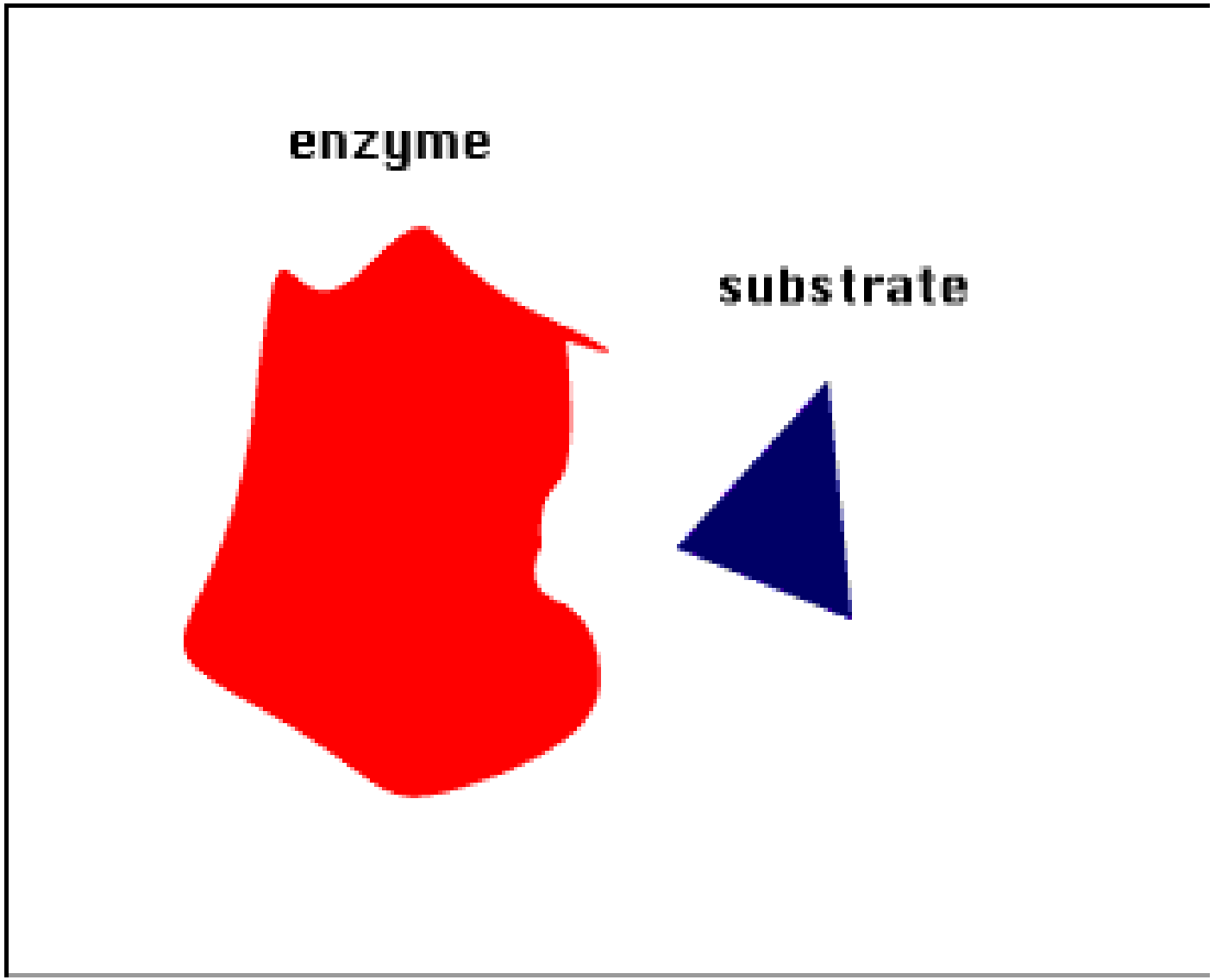


*Song, Zhao, Shi, Jin, Yang, **Li**. Generative Enzyme Design Guided by Functionally Important Sites and Small-Molecule Substrates. ICML 2024.*

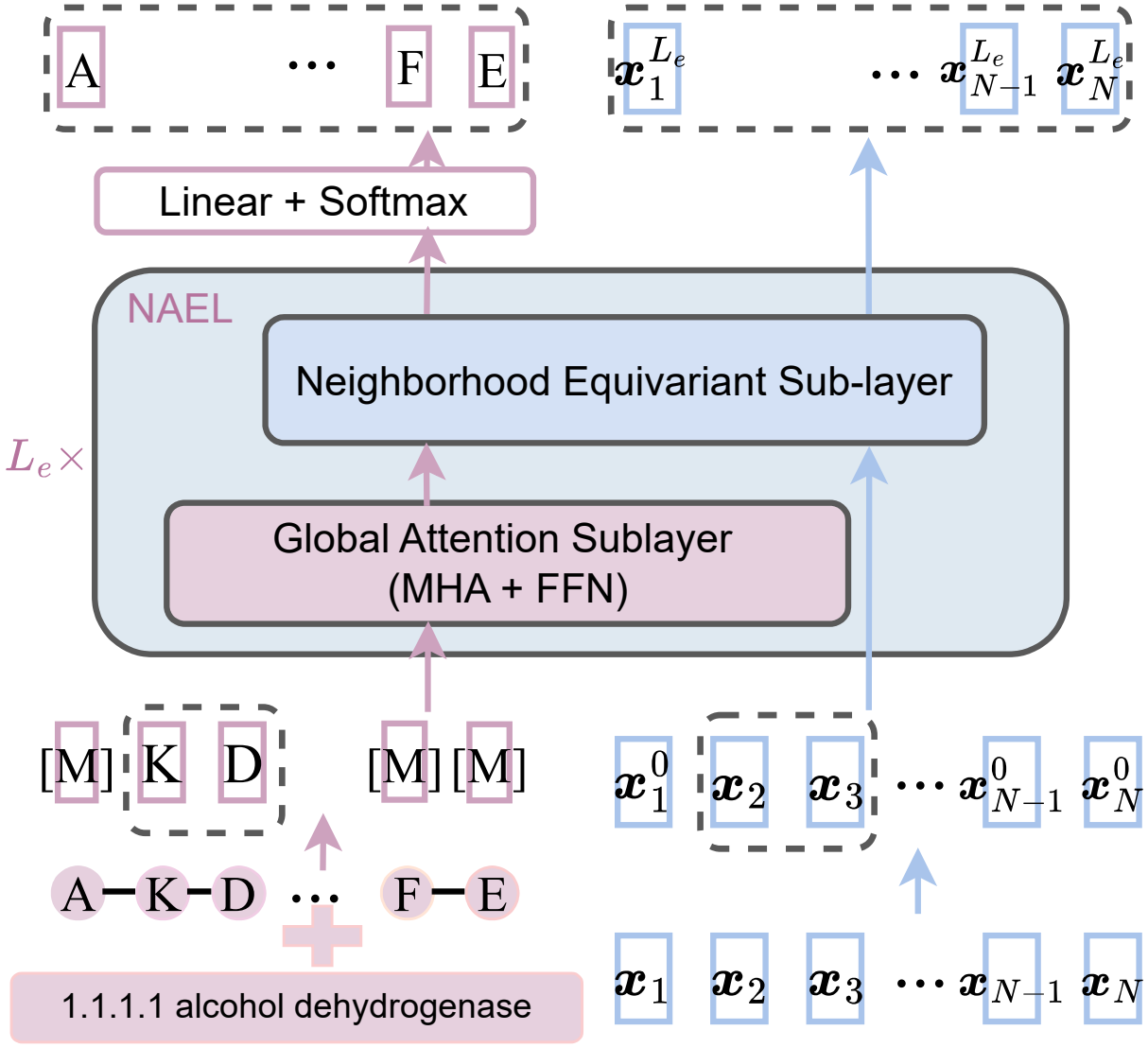
Motivation 3: How to design desired enzymes?

- Substrate Specificity:

Different enzymes binding to specific substrates to speedup enzymatic reactions



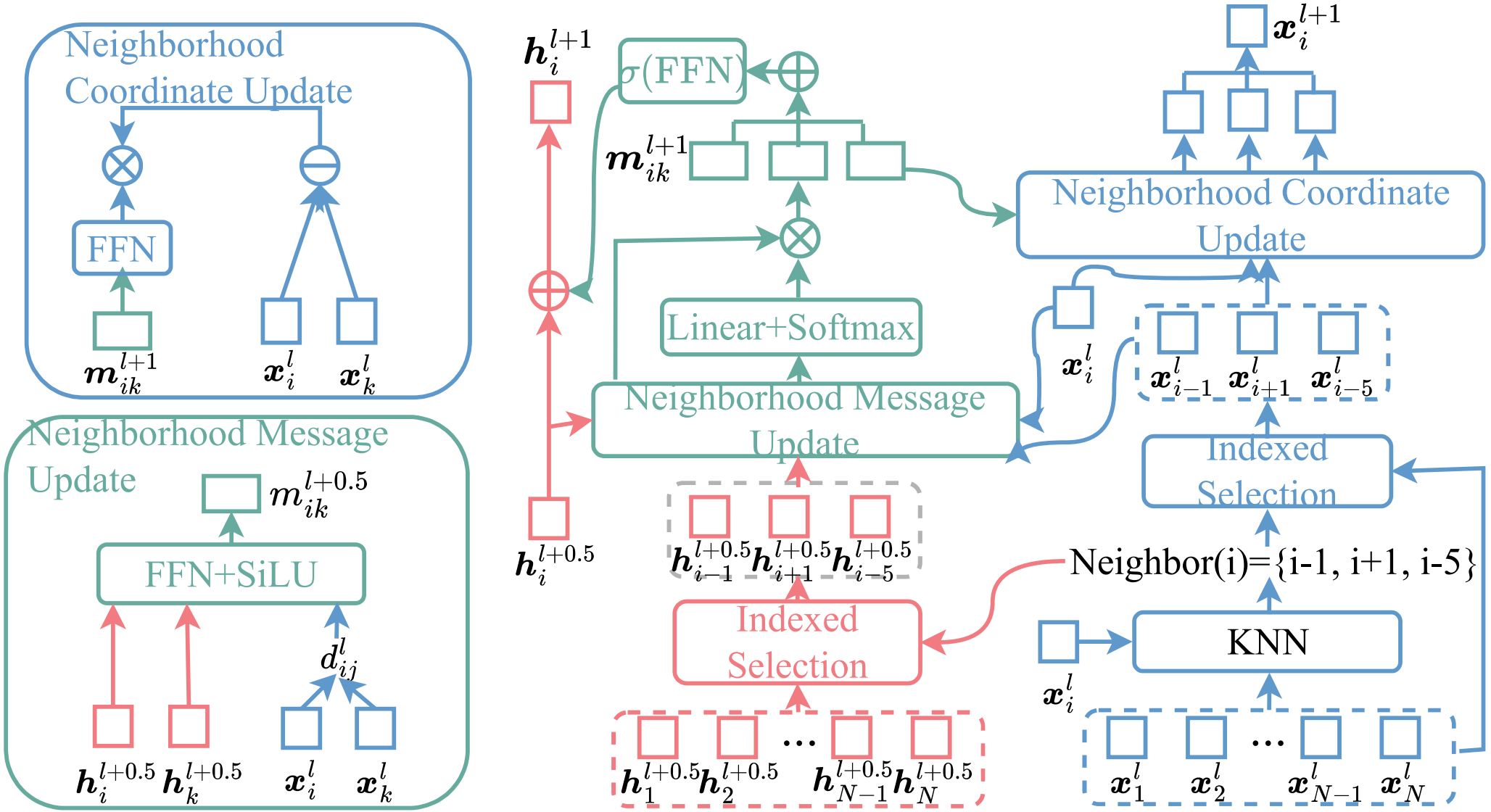
EnzyGen Model – NAEL backbone



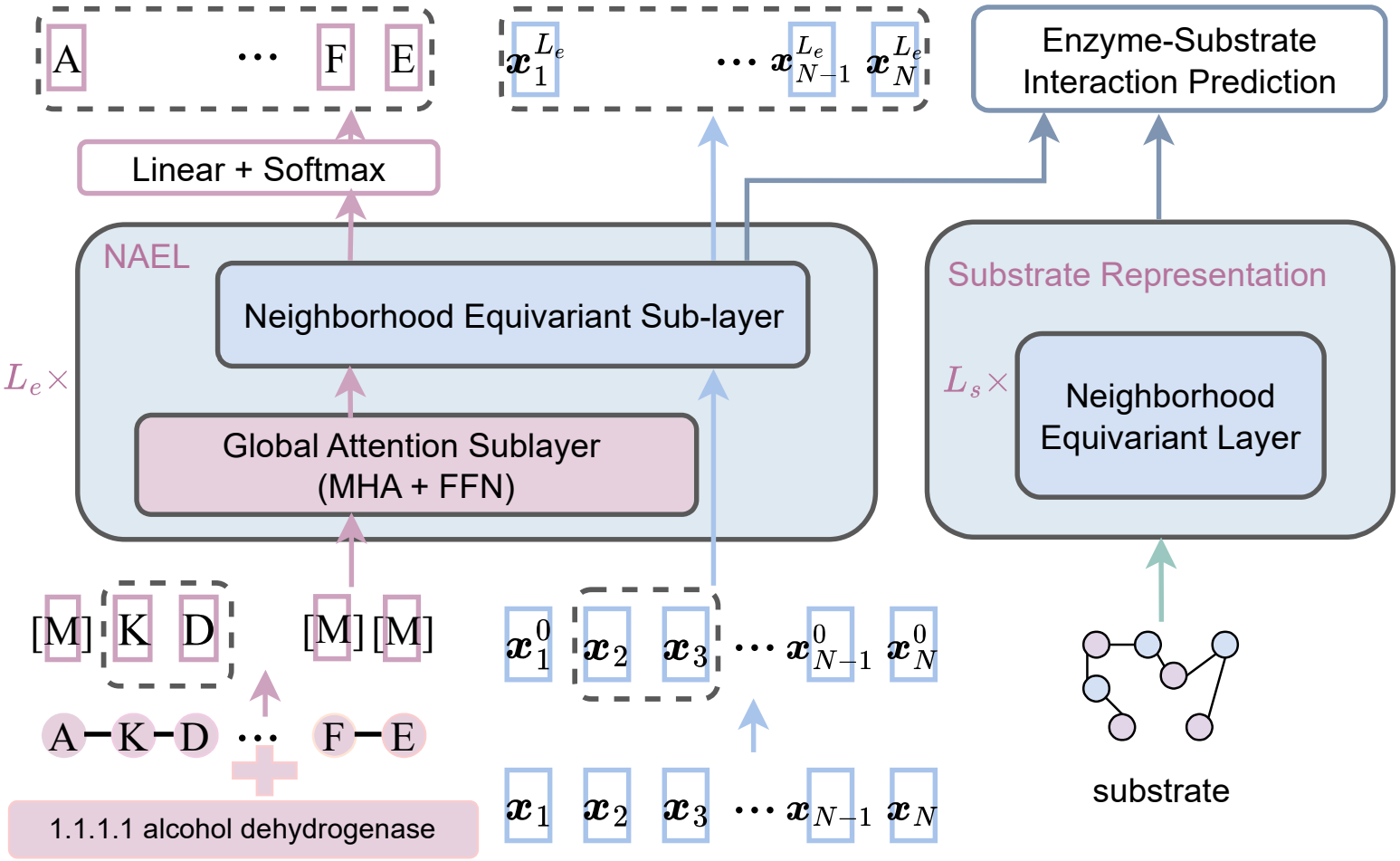
- Controllable Design
  - Functional Sites
  - Enzyme family category

*Song, Zhao, Shi, Jin, Yang, **Li**. Generative Enzyme Design Guided by Functionally Important Sites and Small-Molecule Substrates. ICML 2024.*

Neighborhood Attentive Equivariant Layer



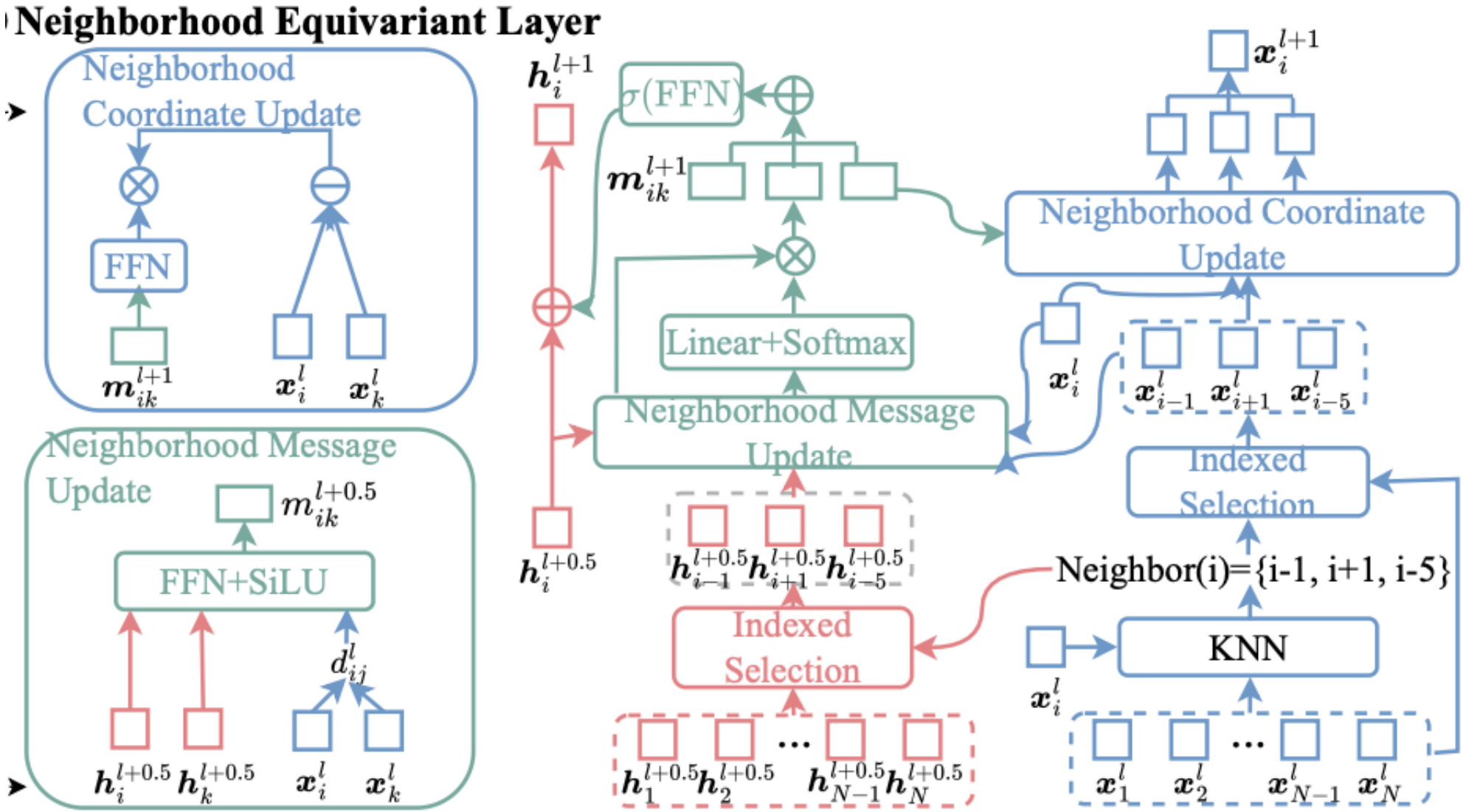
EnzyGen Learning



- Training Objective
  - Predict whole protein sequence
  - Predict whole structure
  - Predict enzyme-substrate binding

*Song, Zhao, Shi, Jin, Yang, **Li**. Generative Enzyme Design Guided by Functionally Important Sites and Small-Molecule Substrates. ICML 2024.*

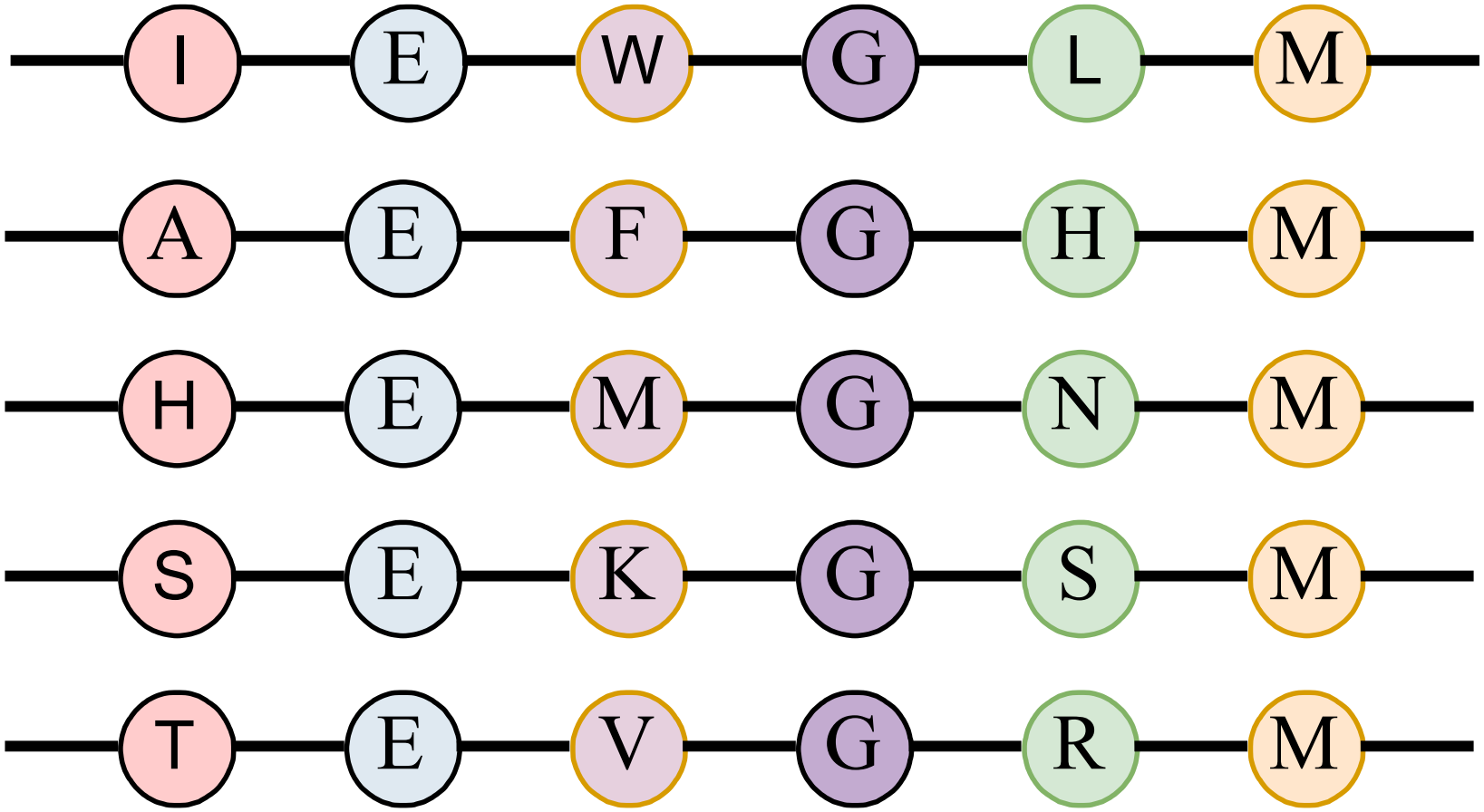
Neighborhood Attentive Equivariant Layer (NAEL)



*Song, Zhao, Shi, Jin, Yang, **Li**. Generative Enzyme Design Guided by Functionally Important Sites and Small-Molecule Substrates. ICML 2024.*

Functional Site Discovery

mining common sites within one family



*Song, Zhao, Shi, Jin, Yang, **Li**. Generative Enzyme Design Guided by Functionally Important Sites and Small-Molecule Substrates. ICML 2024.*

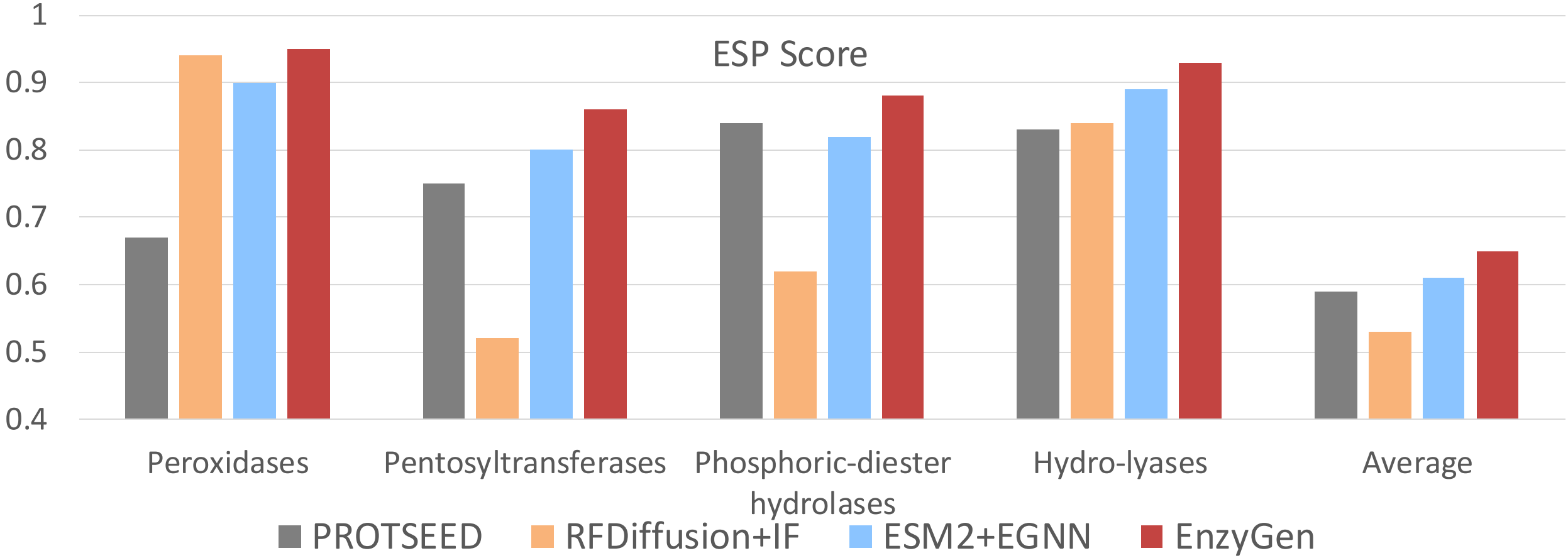
EnzyBench Dataset

- Extracted from BRENDA
  - 8422 fourth-level enzyme classes (enzymatic reaction types)
- Selected PDB entries: 101974
  - 3157 fourth-level enzyme classes
  - discover functional sites for each class
  - Merging into third-level categories: 256
  - 30 largest categories
    - Split 50 for validation & 50 for testing

*Song, Zhao, Shi, Jin, Yang, **Li**. Generative Enzyme Design Guided by Functionally Important Sites and Small-Molecule Substrates. ICML 2024.*

EnzyGen generates enzymes with higher function scores

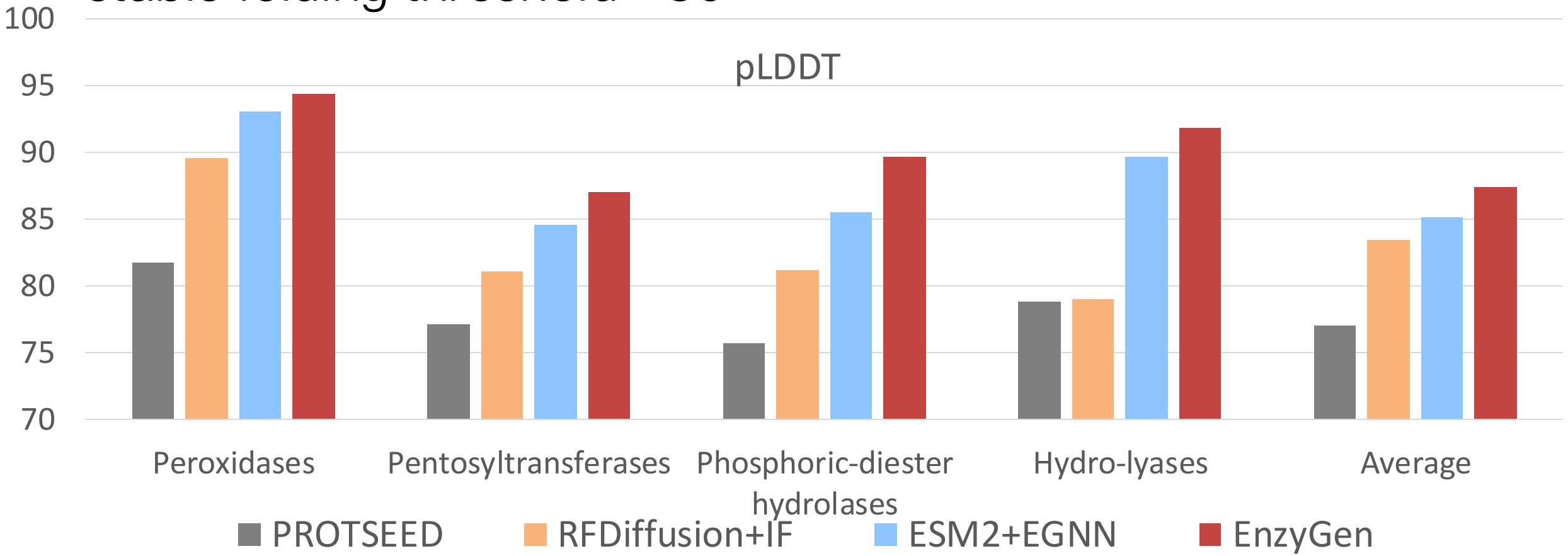
EnzyGen achieves higher enzyme-substrate interaction score in 20 out of 30 categories



*Song, Zhao, Shi, Jin, Yang, **Li**. Generative Enzyme Design Guided by Functionally Important Sites and Small-Molecule Substrates. ICML 2024.*

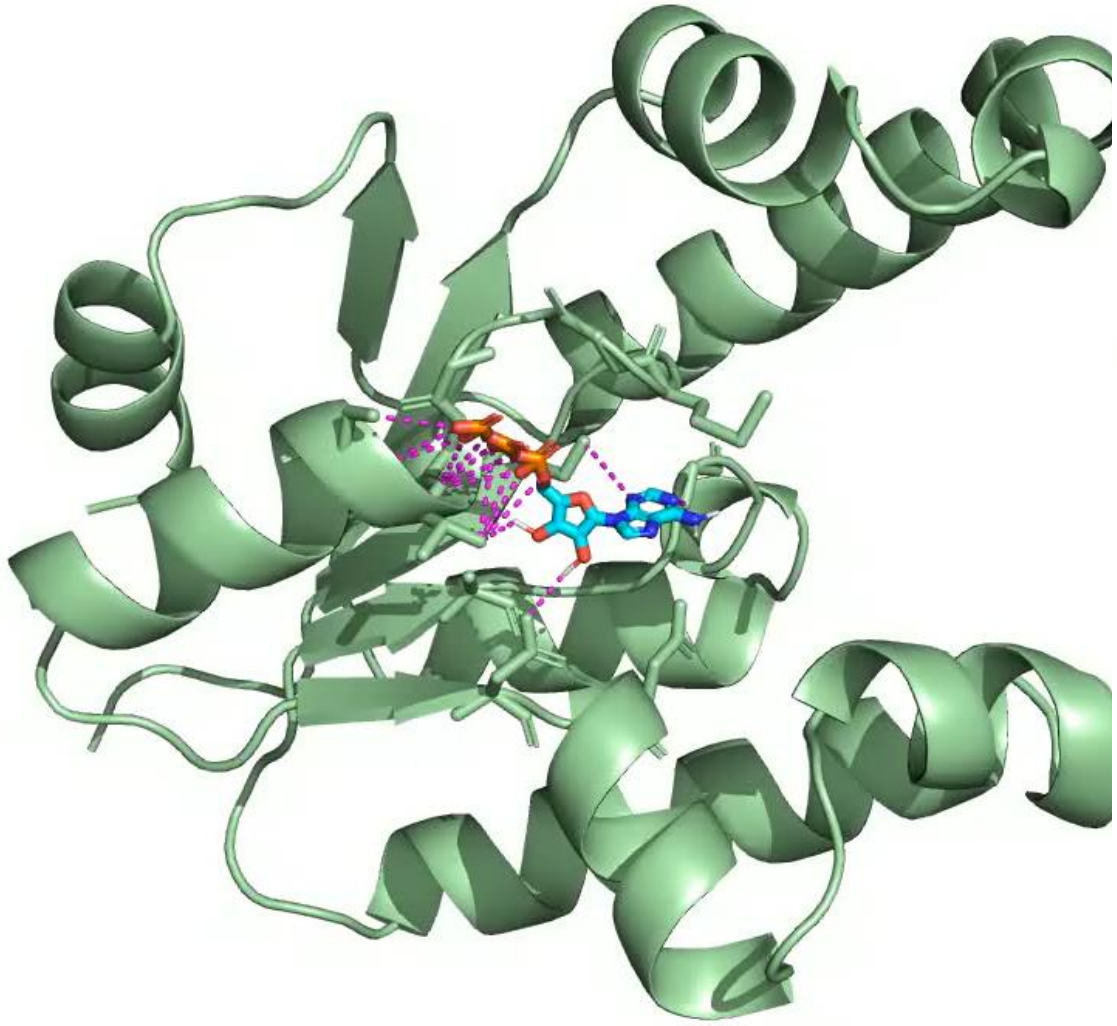
EnzyGen generates enzymes with more stable structures

Average pLDDT across 30 categories is higher than suggested stable folding threshold - 80

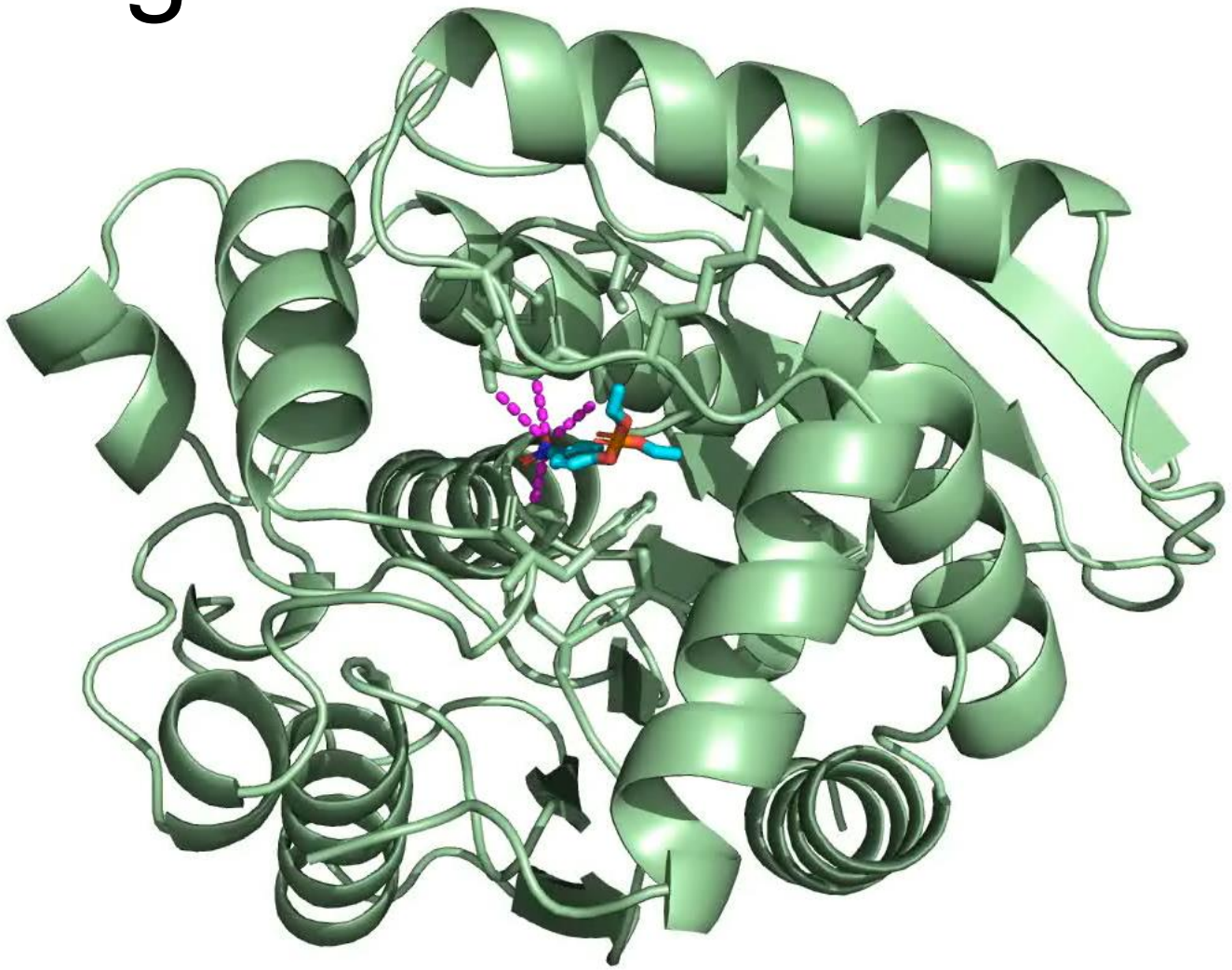


*Song, Zhao, Shi, Jin, Yang, **Li**. Generative Enzyme Design Guided by Functionally Important Sites and Small-Molecule Substrates. ICML 2024.*

EnzyGen designs “good” enzymes in zero-shot categories



Shikimate kinase
(ATP:shikimate 3-phosphotransferase)



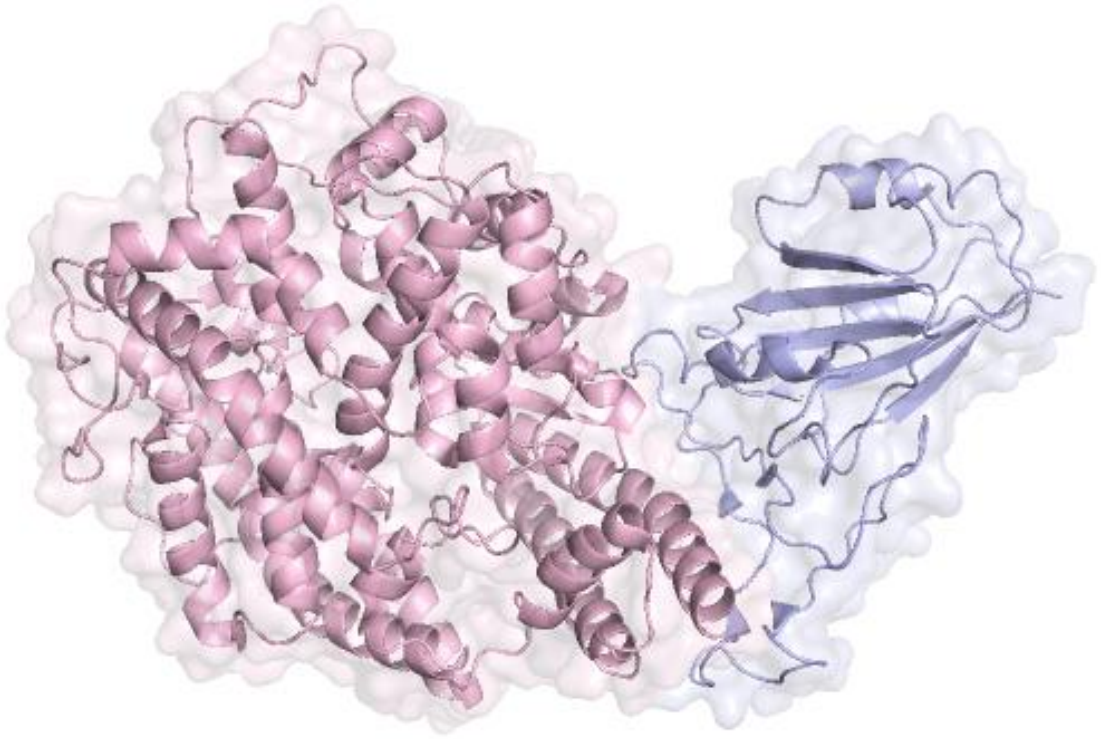
Arylesterase
(substrate paraoxon)

Highlights of EnzyGen

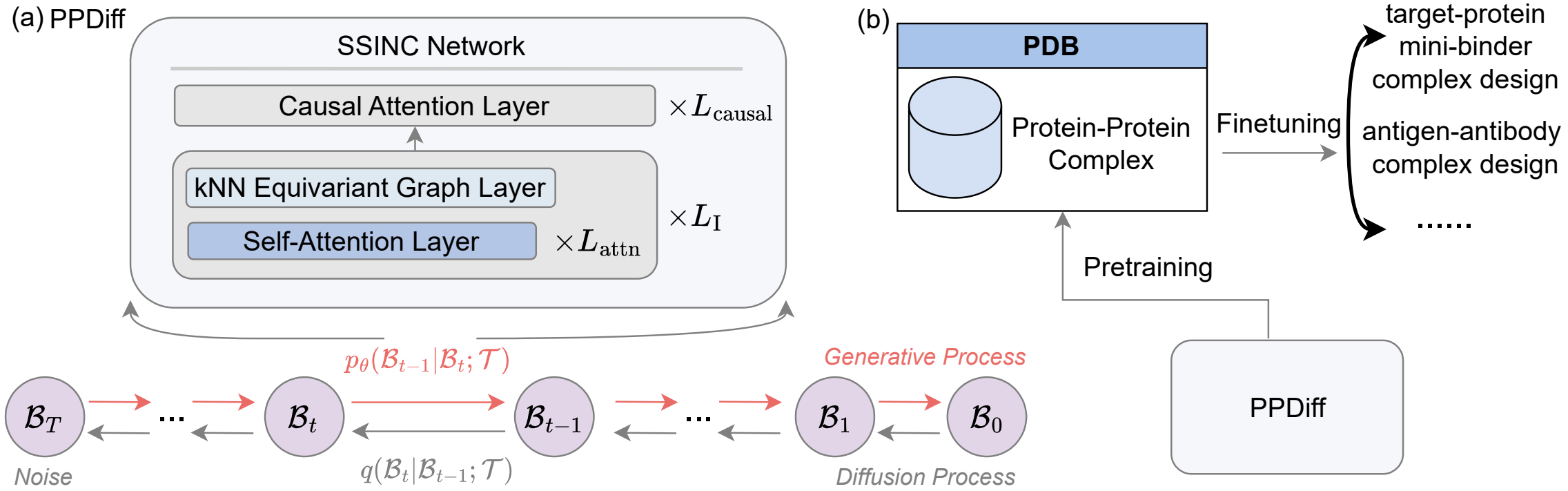
- A unified model for 3k enzyme families
- Guided Generation
  - Functional Important Sites, automatically mined from PDB
  - Enzymy category tags (BRENDA)
- Sequence and Structure Co-design
  - Neighborhood Attentive Equivariant Layer
- Trained takes substrate binding into consideration

Protein-Protein Complex Generation

- Sequence-structure co-design in iterative refinement procedure



PPDiff

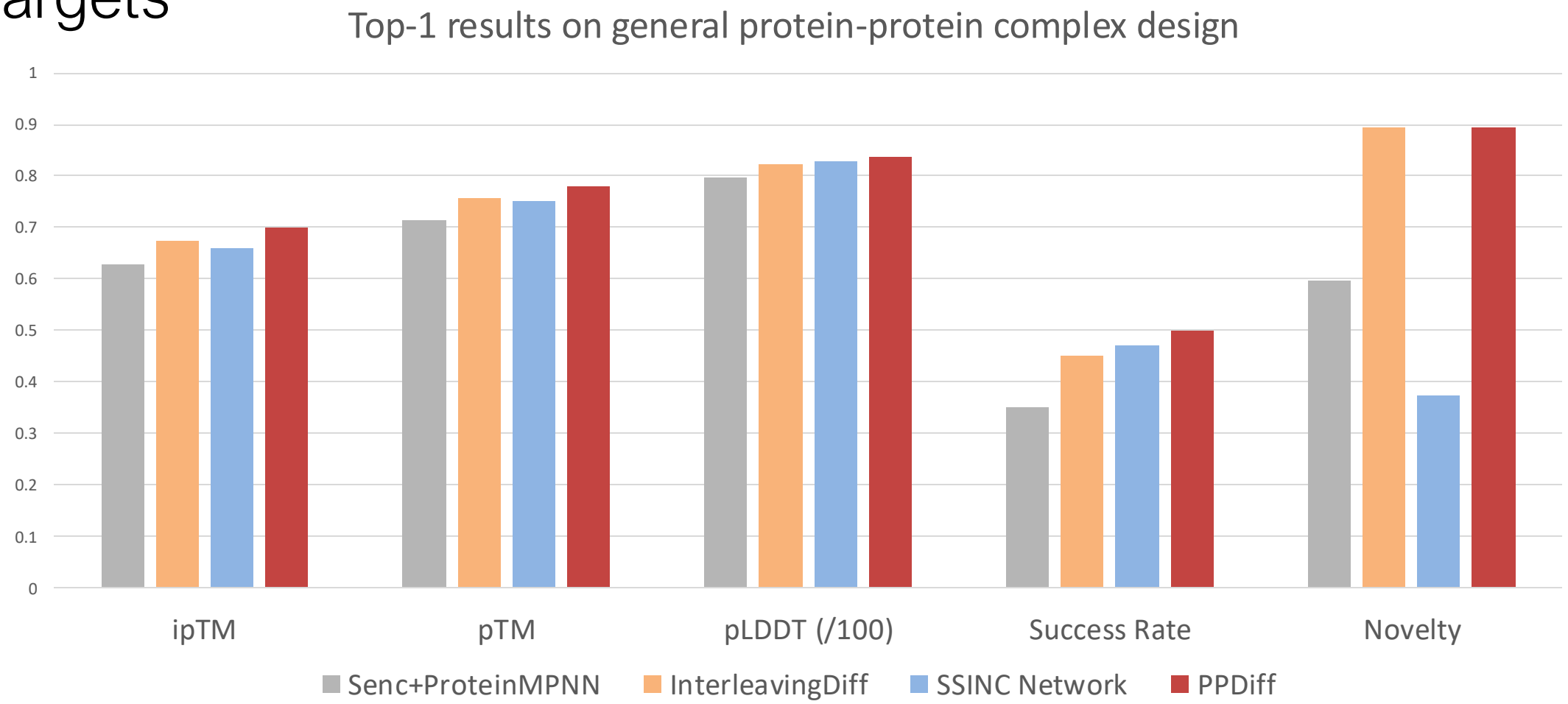


- Diffusing in hybrid space
  - Discrete sequence diffusion
  - Continuous structure diffusion
- SSINC Network
  - Interleaving network (NAEL)
  - Casual attention layers

*Song, Li, **Li**, Min. PPDiff: Diffusing in Hybrid Sequence-Structure Space for Protein-Protein Complex Design. ICML 2025.*

PPDiff generates novel binders with higher success rate

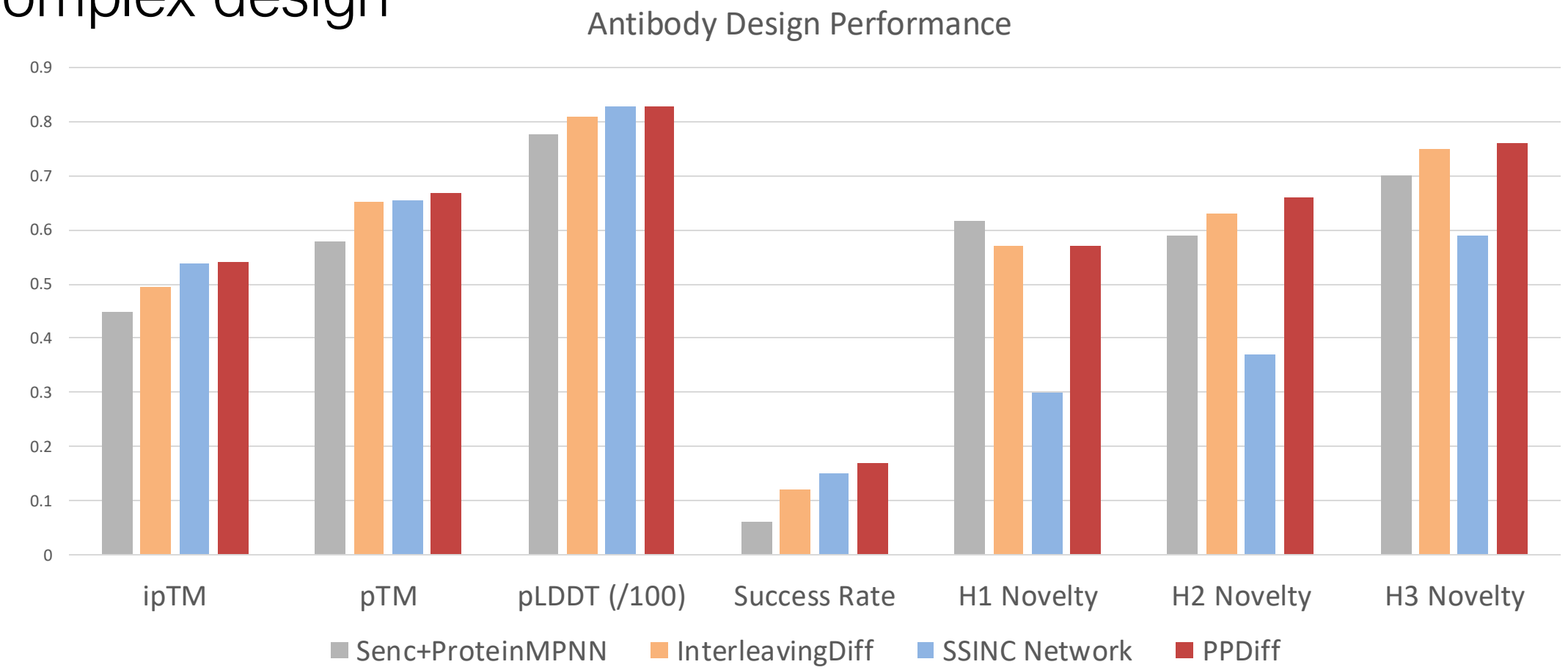
PPDiff achieves **50%** success rate across diverse protein targets



*Song, Li, **Li**, Min. PPDiff: Diffusing in Hybrid Sequence-Structure Space for Protein-Protein Complex Design. ICML 2025.*

PPDiff generates novel antibodies with higher success rate

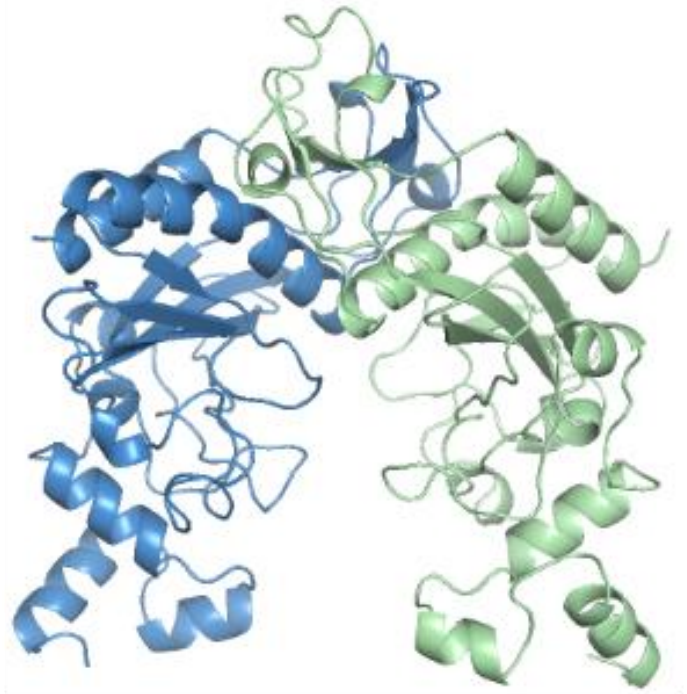
PPDiff achieves 16.89% success rate on antigen-antbody complex design



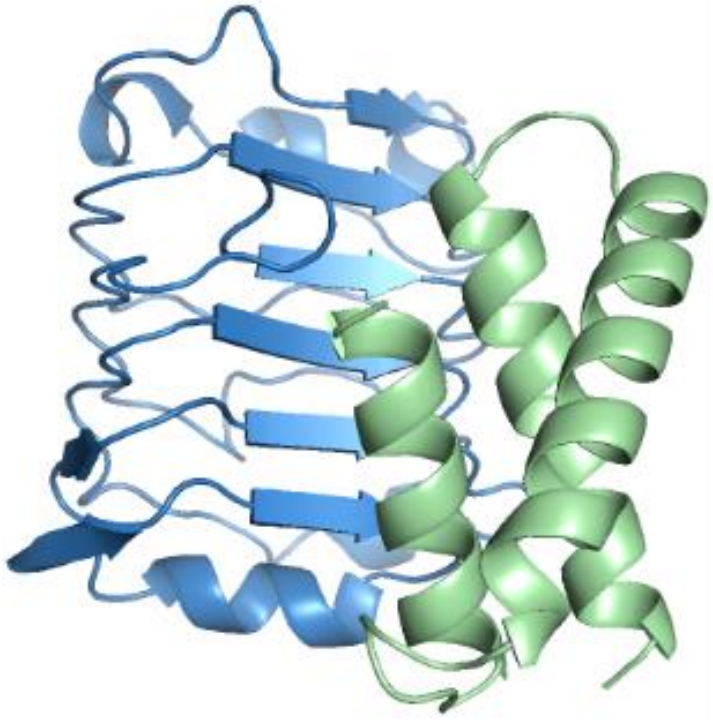
*Song, Li, **Li**, Min. PPDiff: Diffusing in Hybrid Sequence-Structure Space for Protein-Protein Complex Design. ICML 2025.*

PPDiff designs high-affinity binders/antibody across diverse target proteins

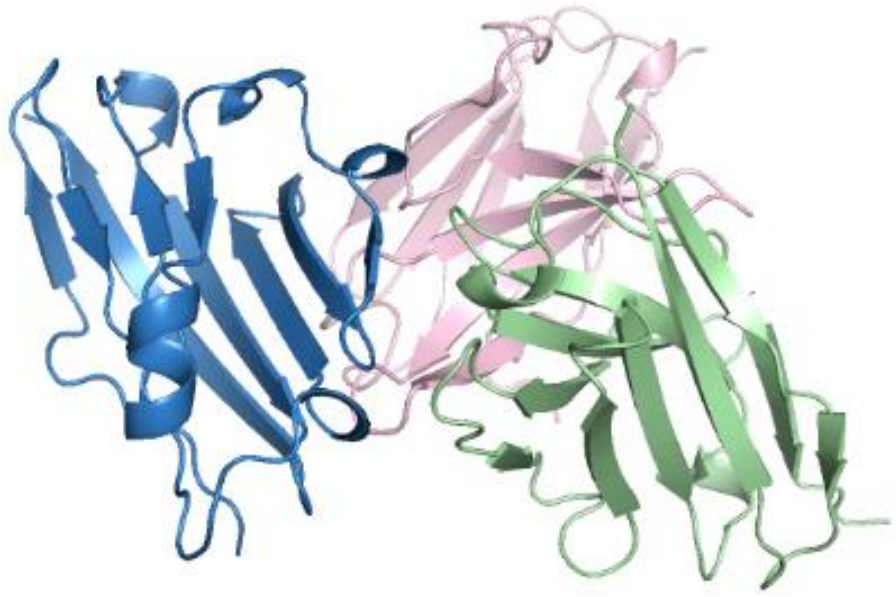
influenza A H3 haemagglutinin



ipTM=0.89, pLDDT=90.12,
pTM=0.88, Novelty=77%



ipTM=0.85, pLDDT=87.21,
pTM=0.87, Novelty=92%



ipTM=0.83, pLDDT=90.80,
pTM=0.87, CDRH3 novelty=55%

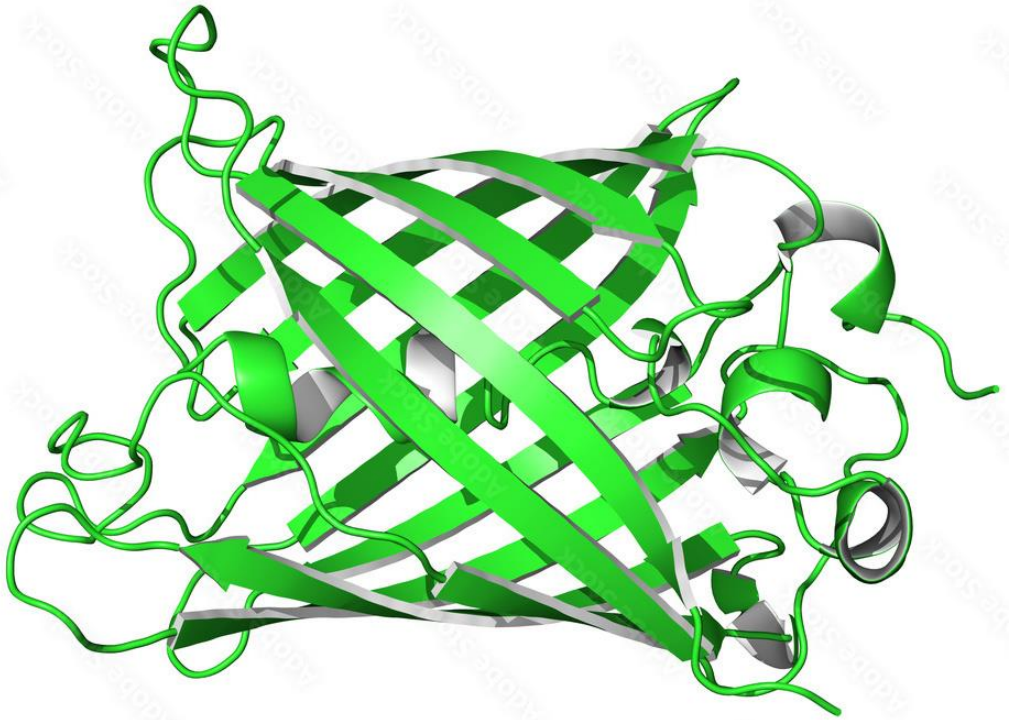
*Song, Li, **Li**, Min. PPDiff: Diffusing in Hybrid Sequence-Structure Space for Protein-Protein Complex Design. ICML 2025.*

Highlights of PPDiff

- A unified model for protein complex sequence-structure co-design
- Diffusion in hybrid space
  - Discrete sequence diffusion
  - Continuous structure diffusion
- Performs well in wide applications
  - Generation protein-protein complex design
  - Target protein-mini binder complex design
  - Antigen-antibody complex design

Protein Design Approaches

- Sequence-based Generation
- Structure-based Generation
  - Secondary structure-based
  - Inverse Folding
  - Surface geometry
- Sequence-Structure Co-design
  - Protein monomer
  - Protein complex



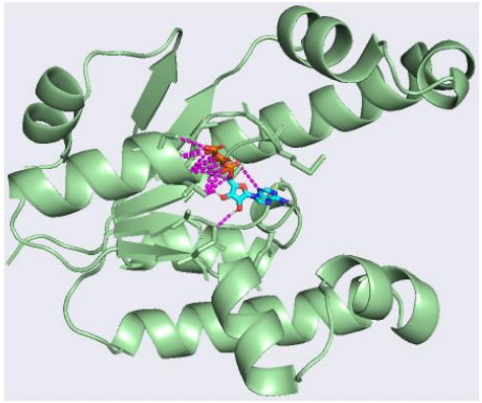
Takeaway of Protein Design

- Problem formulation: Guiding information is important
  - fitness scores, chemical properties, tags, motifs
- Modelling Structure/Geometry is critical for molecules
  - Keeping SE(3) equivariance implicitly augments training data
- Modeling the mutual constraints between sequence and structure is useful
- Interaction between protein-ligand complex
- Diffusion method to further iteratively refine: discrete+continuous

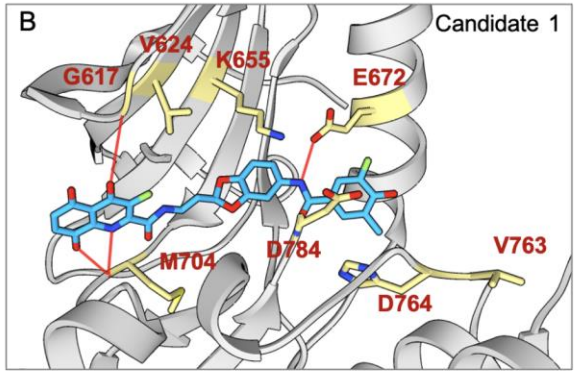
Commonality and Distinction in Generating Language and Molecules

Distribution	Sequence	BERT, GPT	ESM, ProGen
	2D Structure	Tree-LSTM	MPNN
	3D Geometry		EGNN, EnzyGen [ICML24], SurfPro [ICML24], PPDiff
Generation	Score-guided	C-VAE	IsEMPro [ICML 23]
	Editing	CGMH[AAAI19]	MARS [ICLR21], MolEdit3D

Molecule Design at CMU Li lab



https://leililab.github.io/



Protein

EnzyGen

SurfPro

IsEMPro

LSSAMP

PPDiff

Small Molecule

MARS

MolEdit3D

RLHEX