

From **Words** to **Molecules**: Harnessing Generative AI for Breakthroughs in Language and Molecular Design

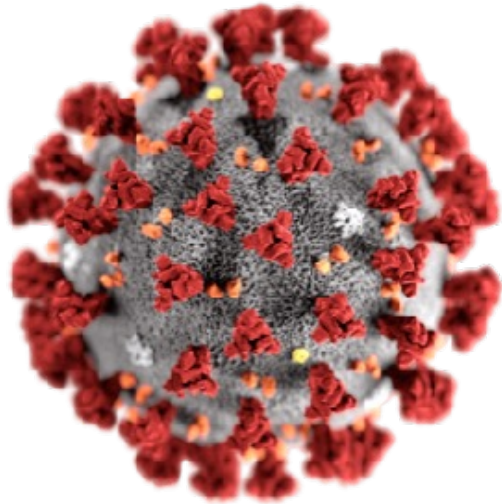
Lei Li



Carnegie Mellon University

Language Technologies Institute

February 2020



MIT AI powered Drug Discovery
and Manufacturing Conference.
Boston, 2020.2

Large Language Models drive the Productivity

Translate

Summarize

Editing

Write email



ChatGPT



LLaMA

Chat

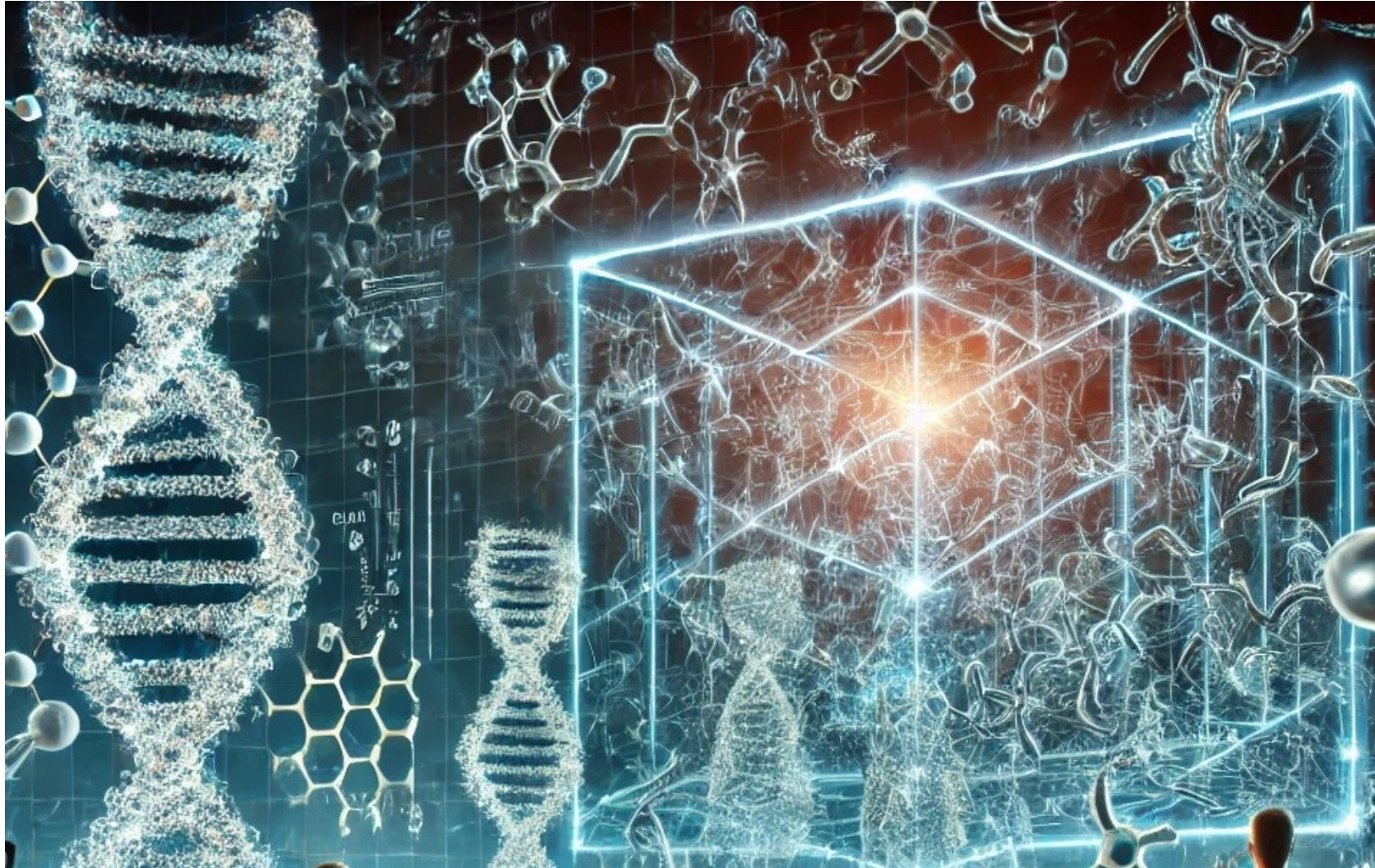
Answer questions

Suggest names

Write code

Recommend restaurants

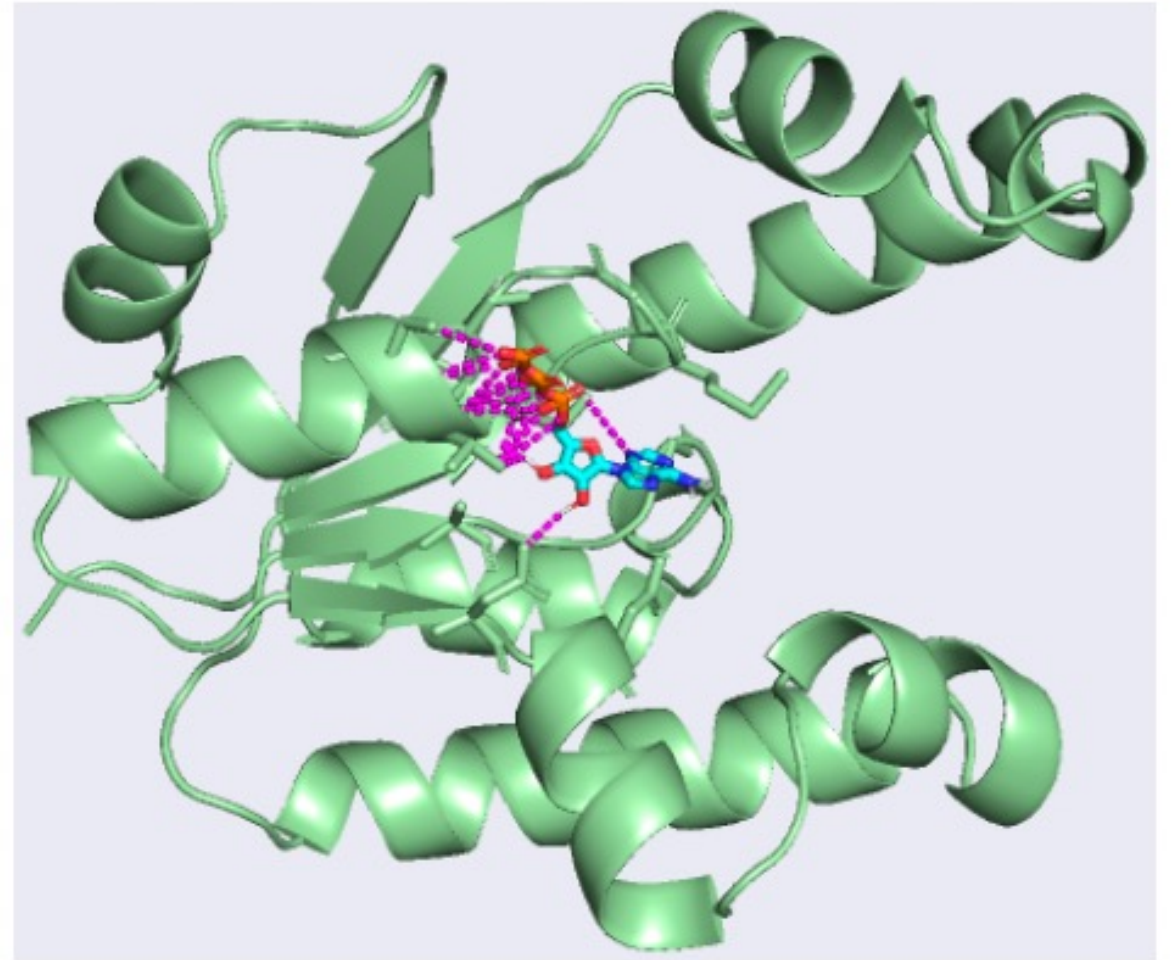
Generative AI powers Imagination



generated by DALL-E with prompt “generate an image using generative AI for protein design”

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

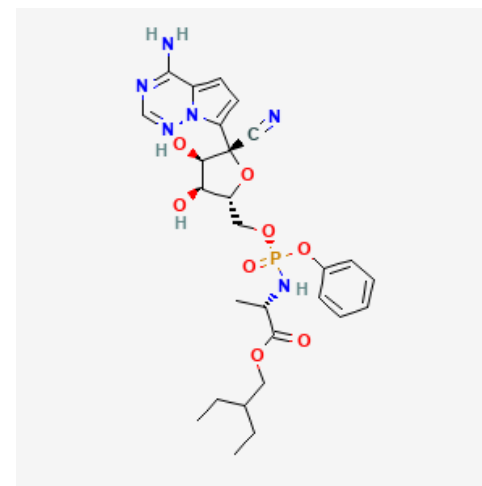
Discrete Sequences of Tokens

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

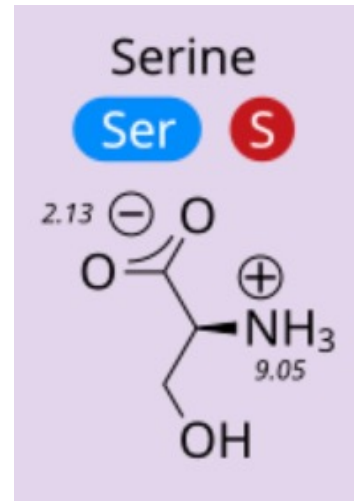
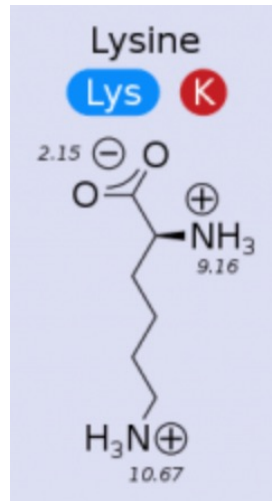
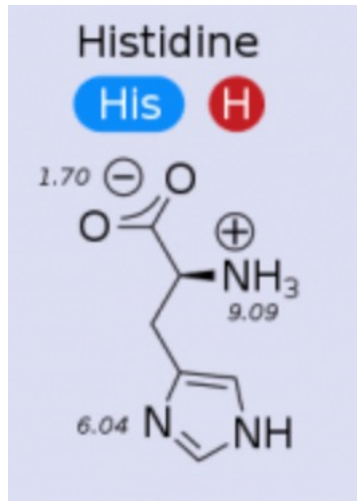
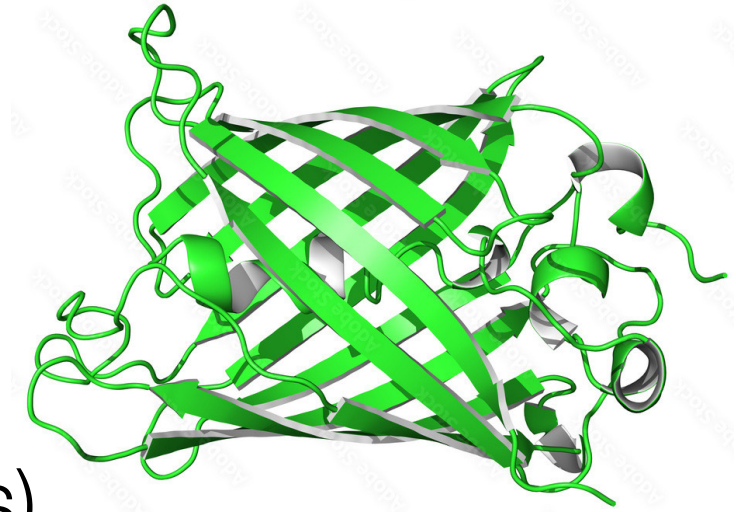
SMILES representation:

```
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
```



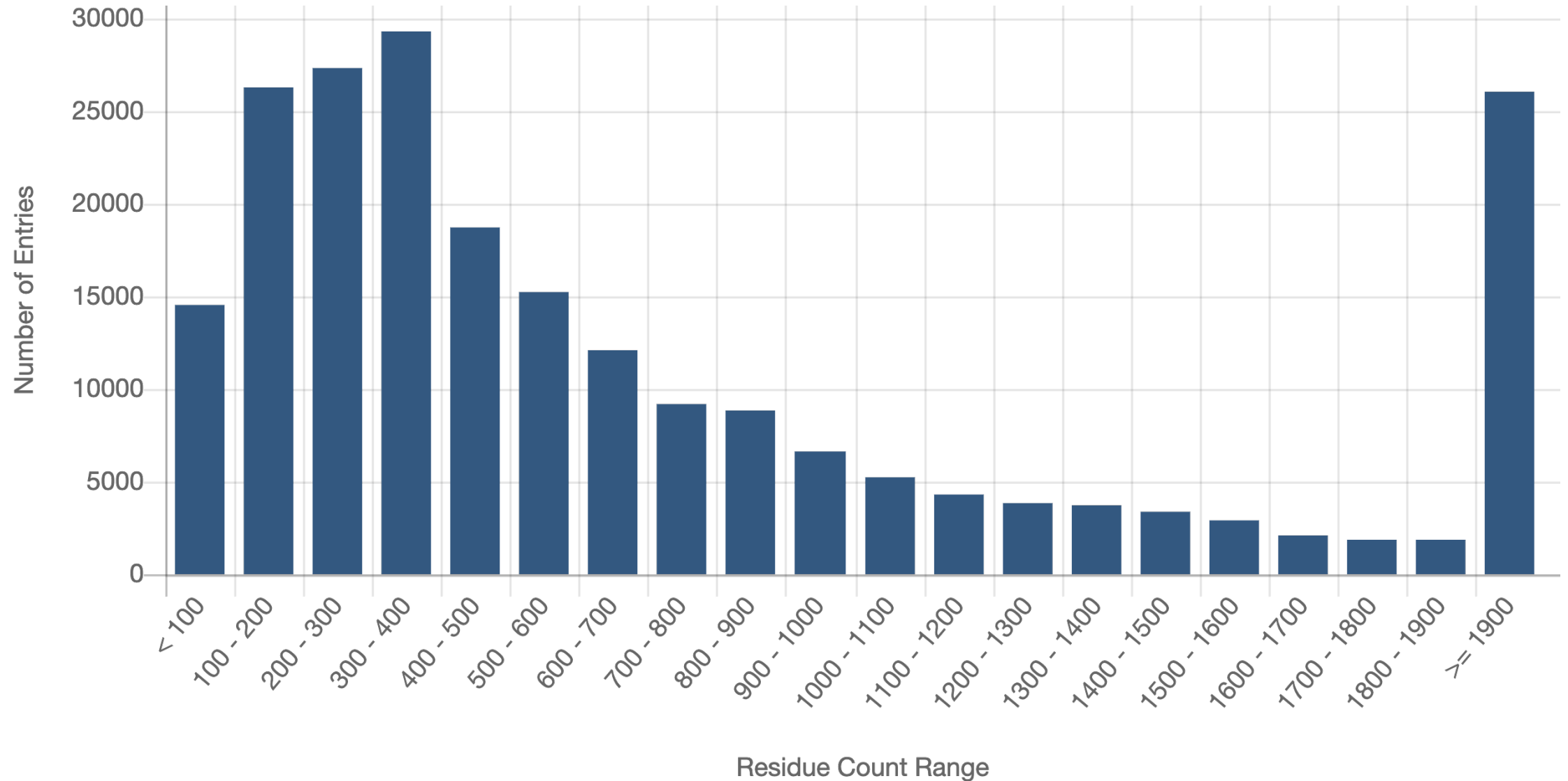
From Human Language to Protein Sequence

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



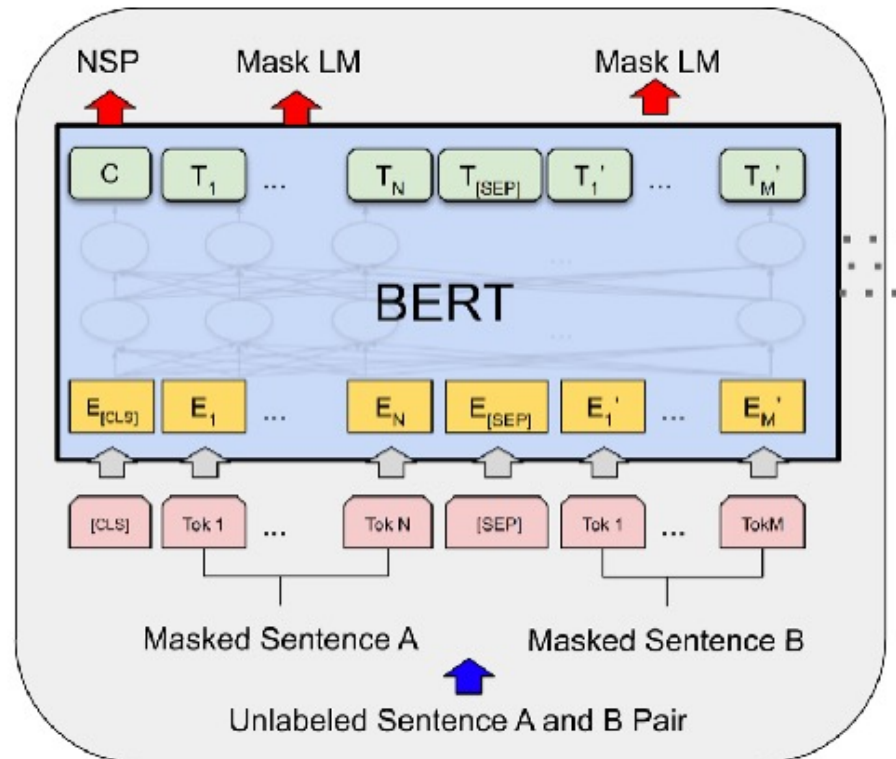
VLLPDNHYLSTQSALS KDPN
EKRDHMLLEFVTAAGIT

Protein Sequences are much Longer than Text!

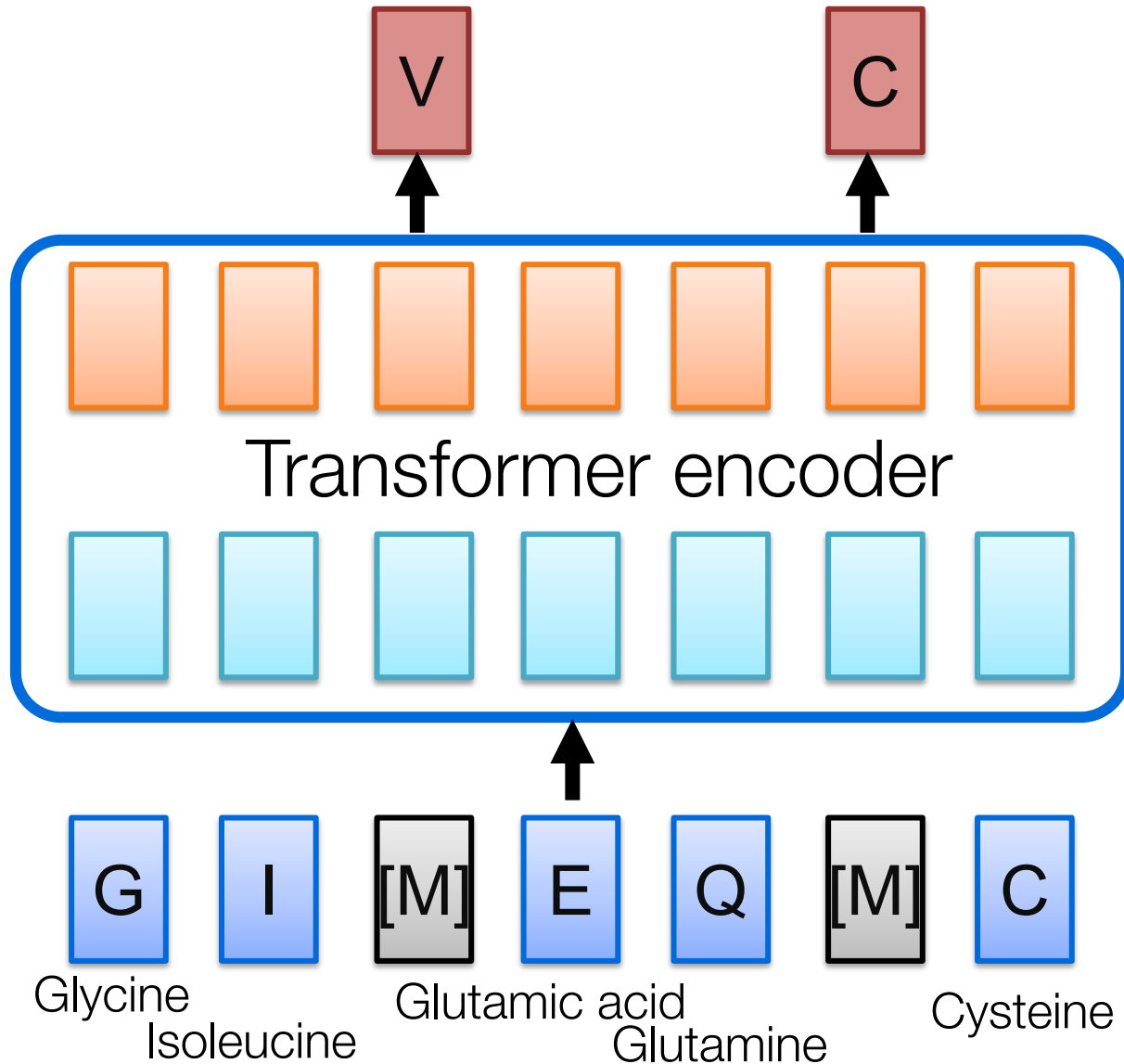


Modeling Distribution of Sequences

- BERT
- GPT
- ESM, ESM-2
- ProGen

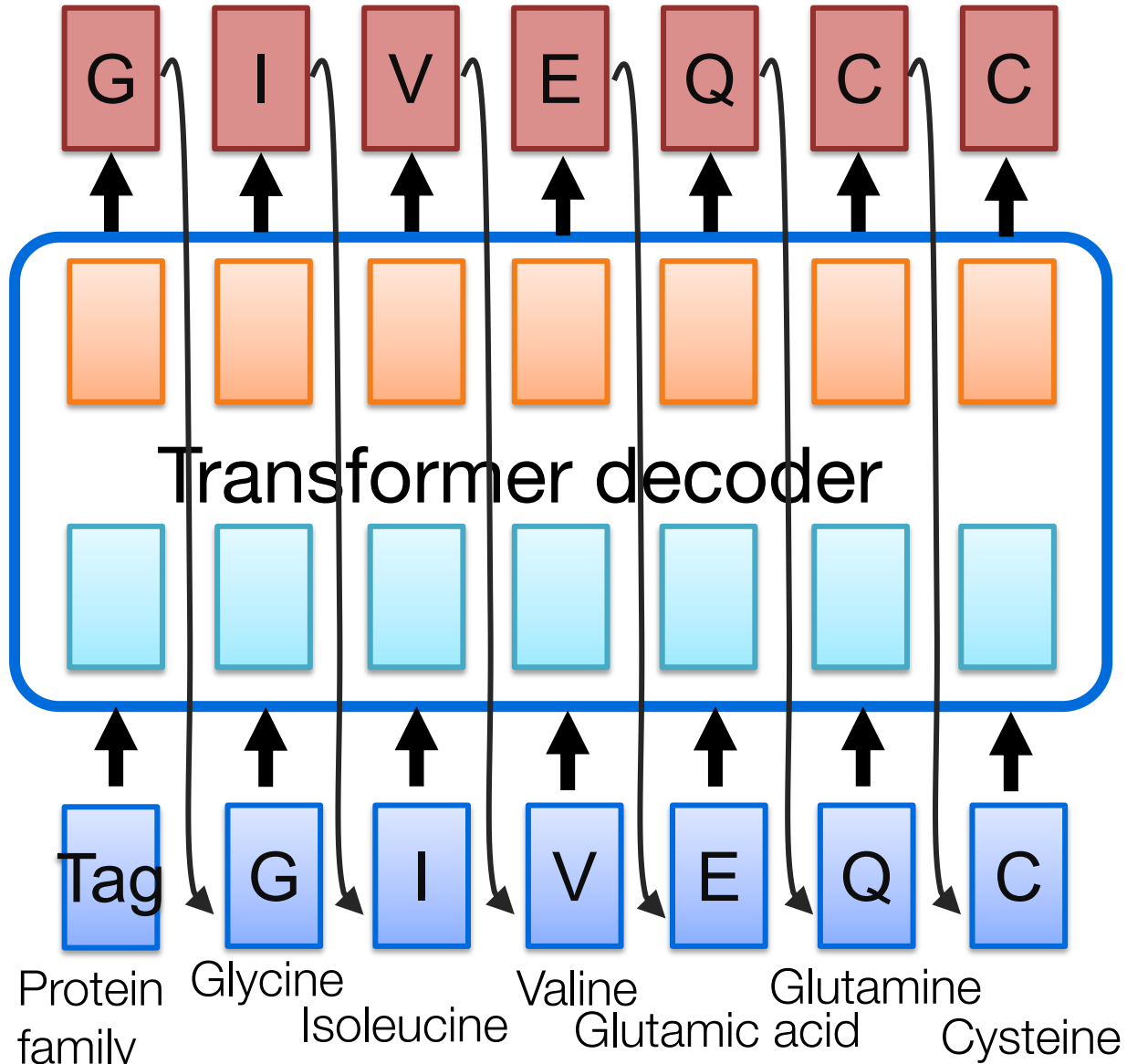


Protein Language Model 1: Mask LM



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

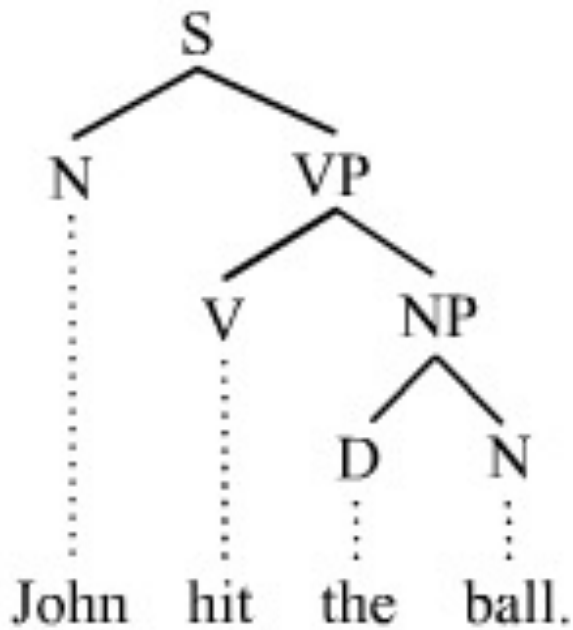
Protein Language Model 2: Casual LM



- Using raw protein sequences and their category tags for pre-training
 - training loss: predicting next residue
- ProGen [Madani et al 2023] and ProGen2 [Nijkamp et al 2023]
- Protein Tag is insufficient!

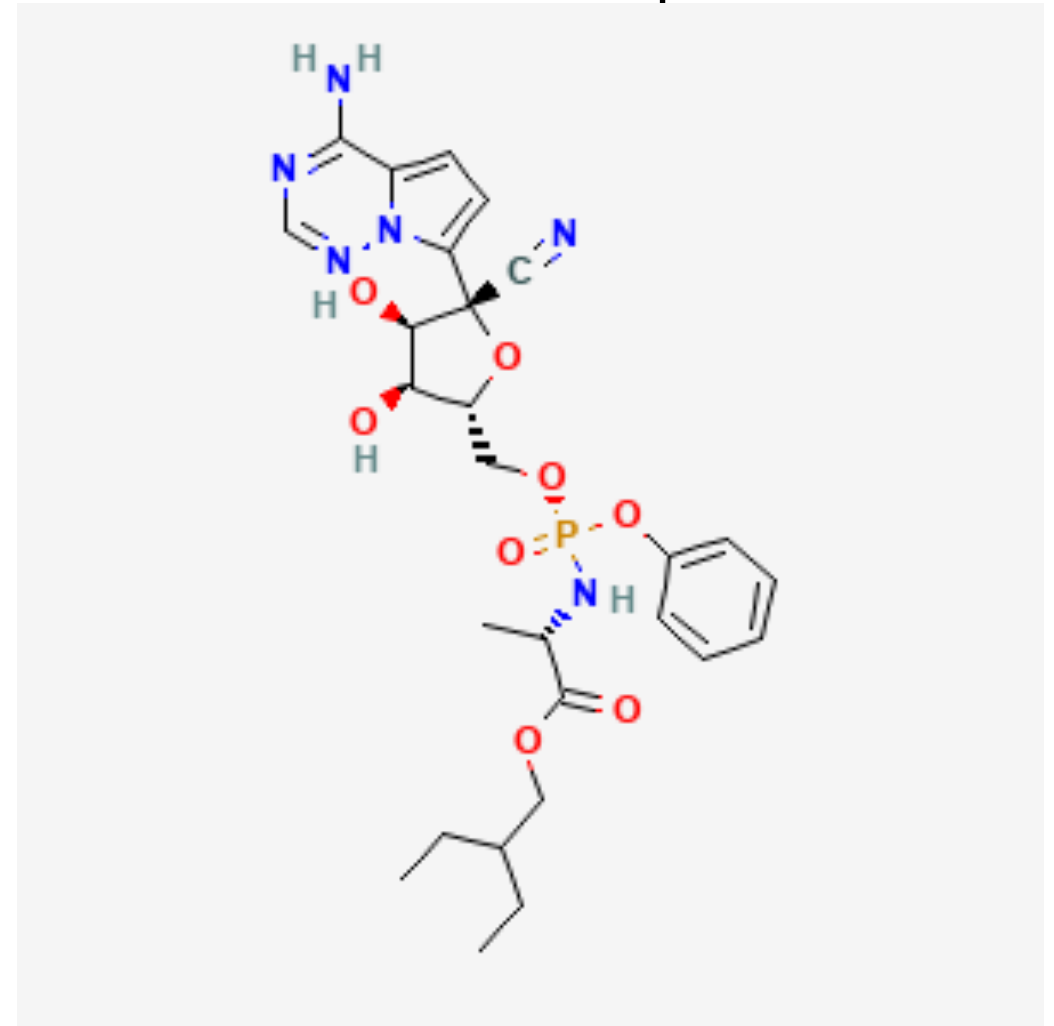
Discrete Structure

- Trees
 - Dependency tree
 - Constituency tree



Constituency-based parse tree

- 2D Molecule Graph



Graph Neural Network

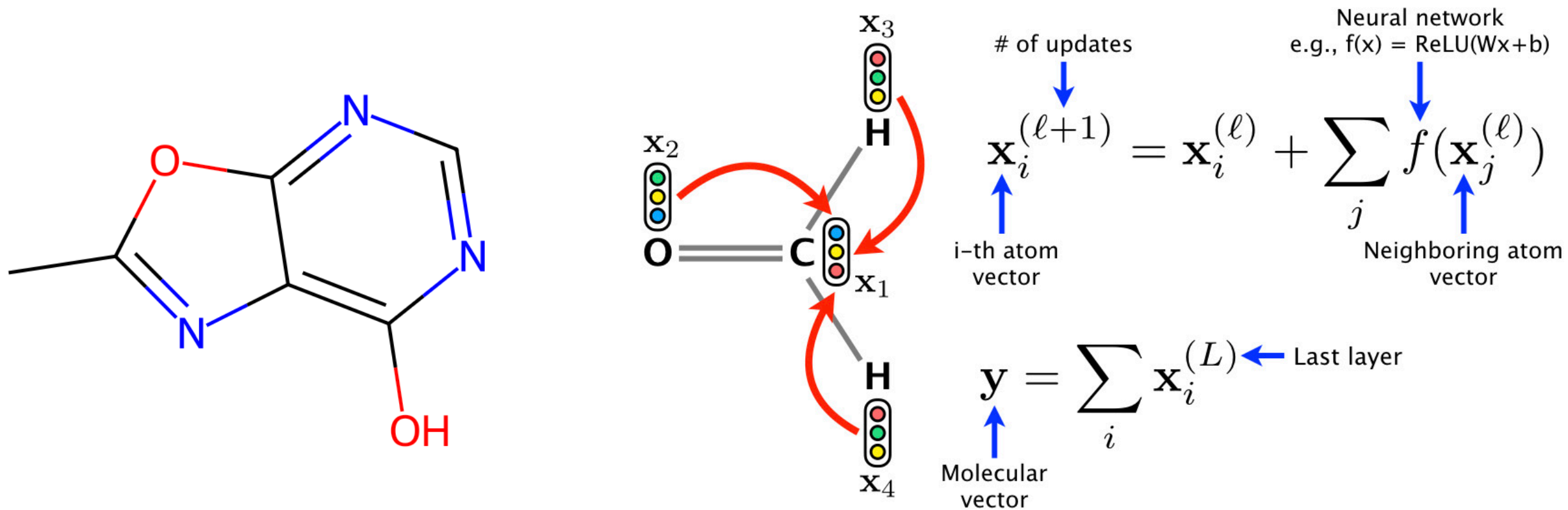


Fig.1: The update function (or called transition, propagation, message passing, and convolution) in GNNs. On a molecular graph, the **GNN updates each atom vector with its neighboring atom vectors non-linear transformed by neural network**. The molecular vector is obtained by summing (or mean) the atom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

Graph of Fragments

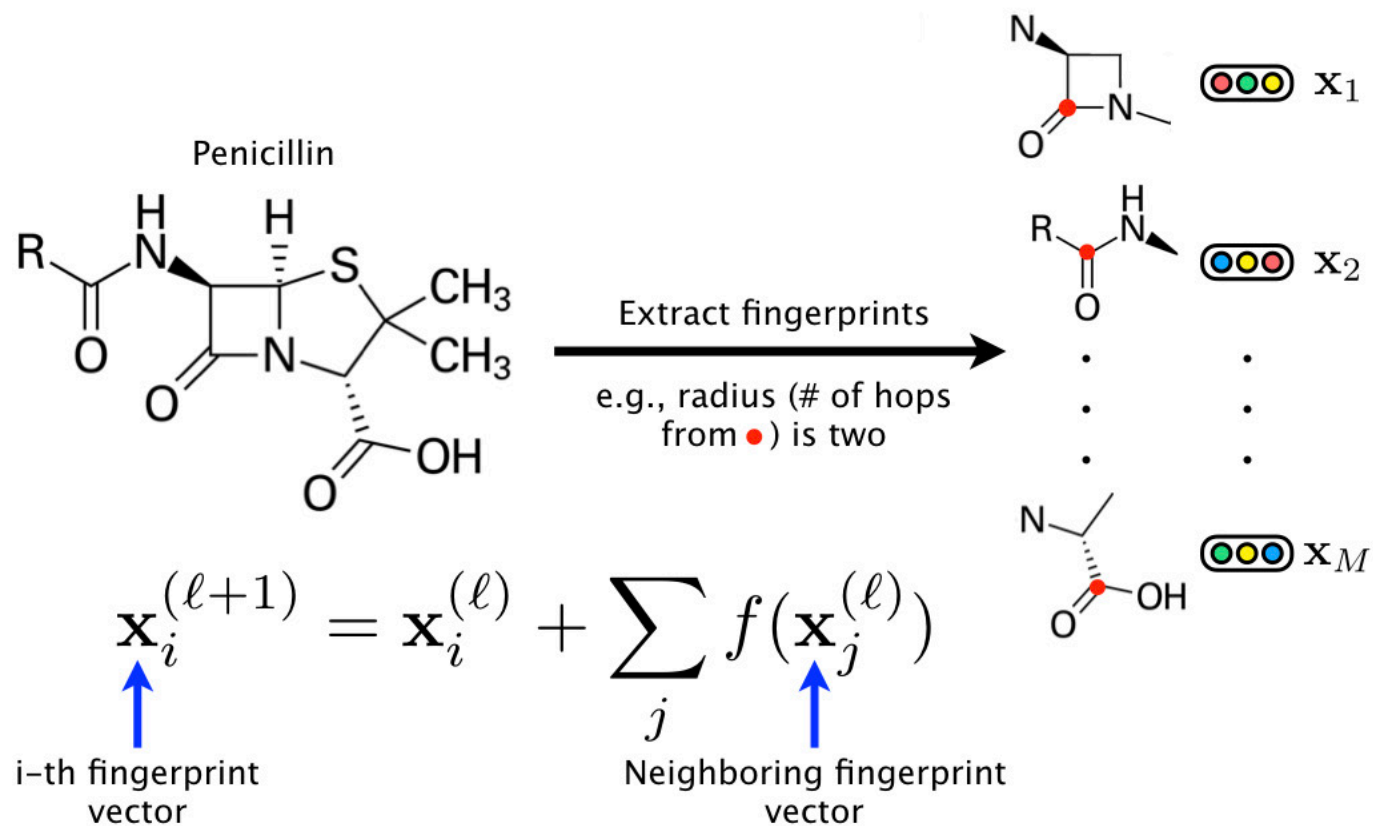
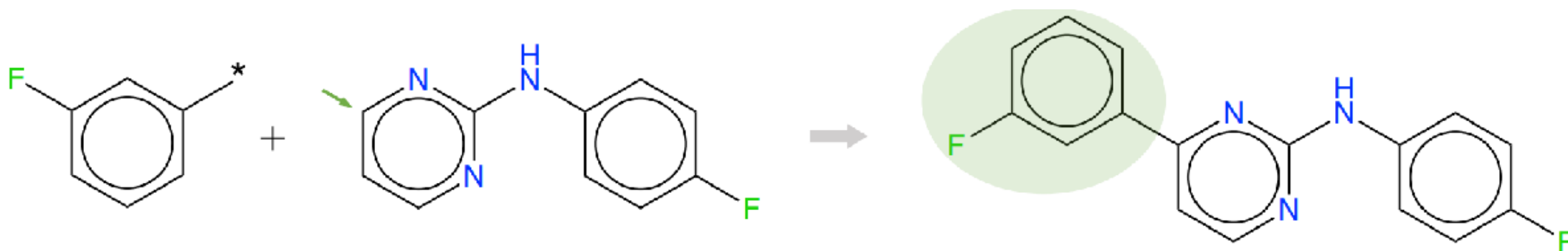


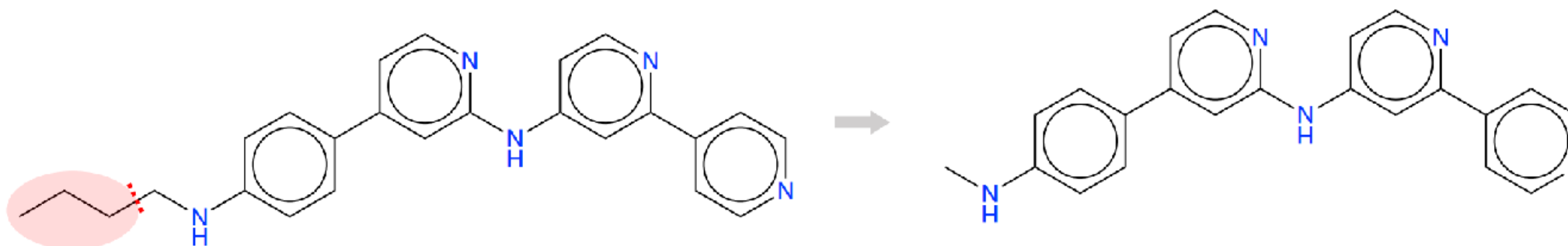
Fig.2: The update function based on radius-based subgraphs, i.e., molecular fingerprints. **Each fingerprint is initialized with a random vector.** The following procedure is the same as that of basic GNN.

Iterative Editing based Molecule Generation

- Adding fragment

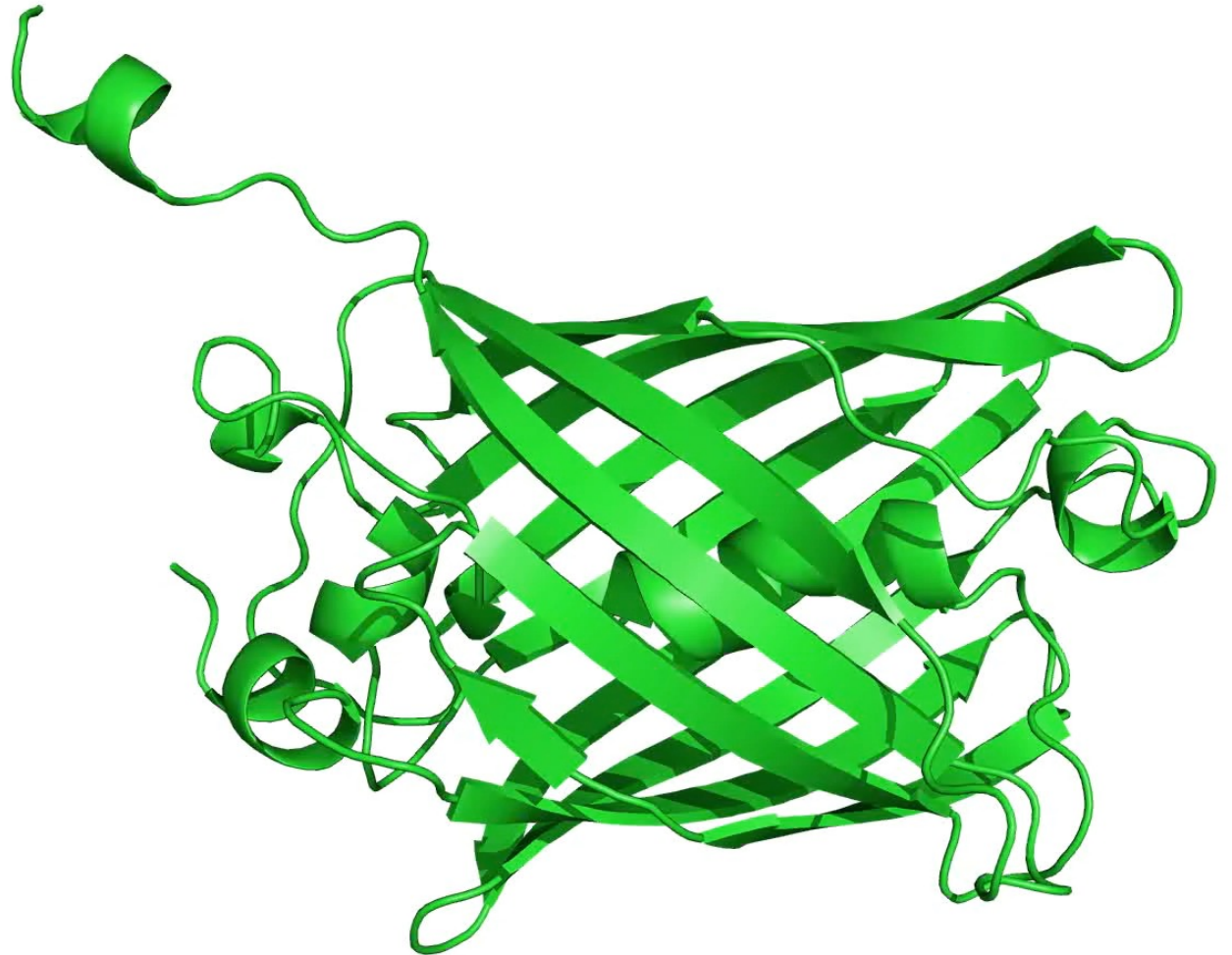
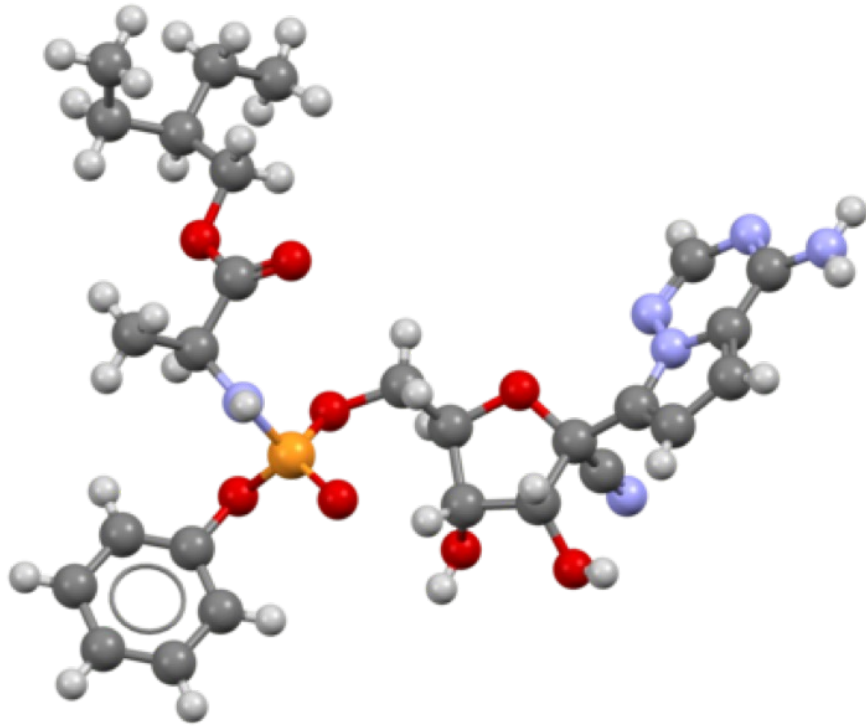


- Deleting a fragment



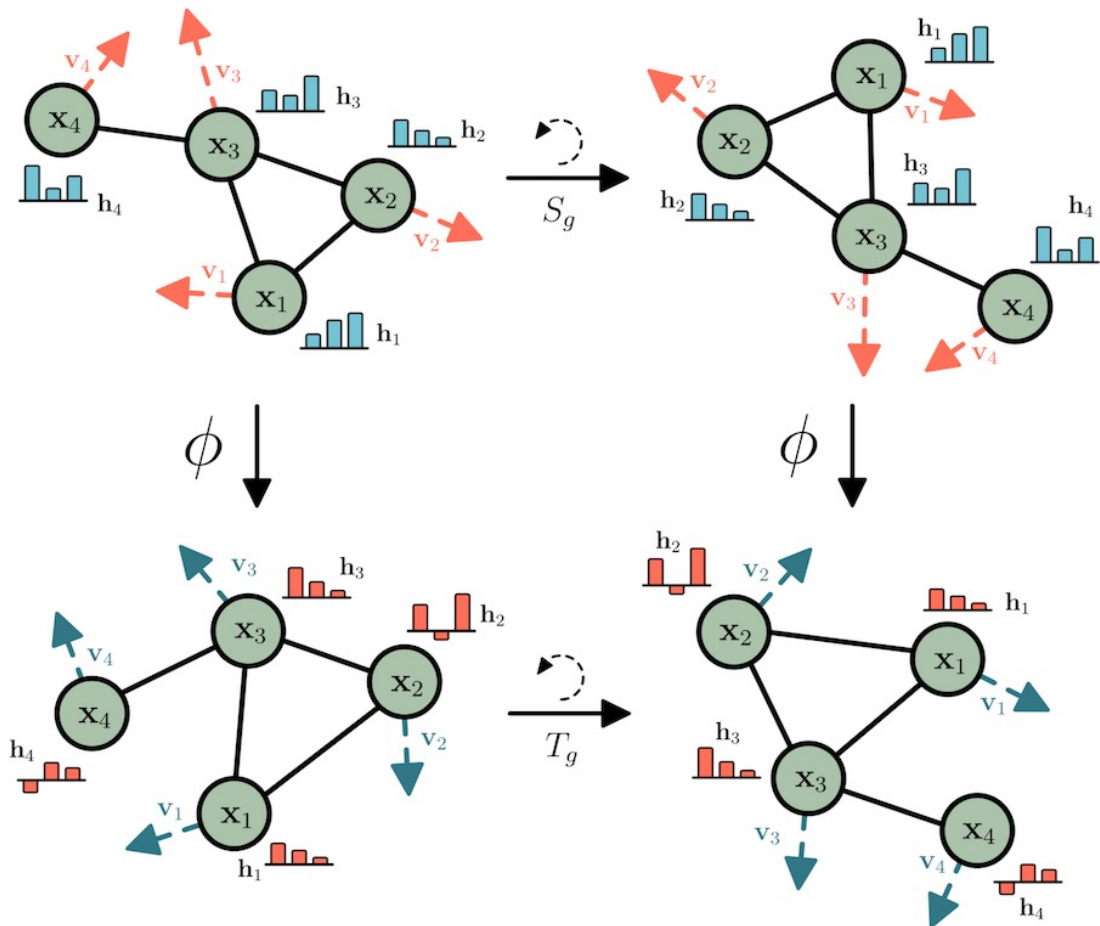
Geometry of Molecule

- Matrix of 3D coordinates
- Matrix of angles



Modelling Geometry

- Equivariant Graph Neural Network (EGNN)



Equivariance:

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

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

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

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

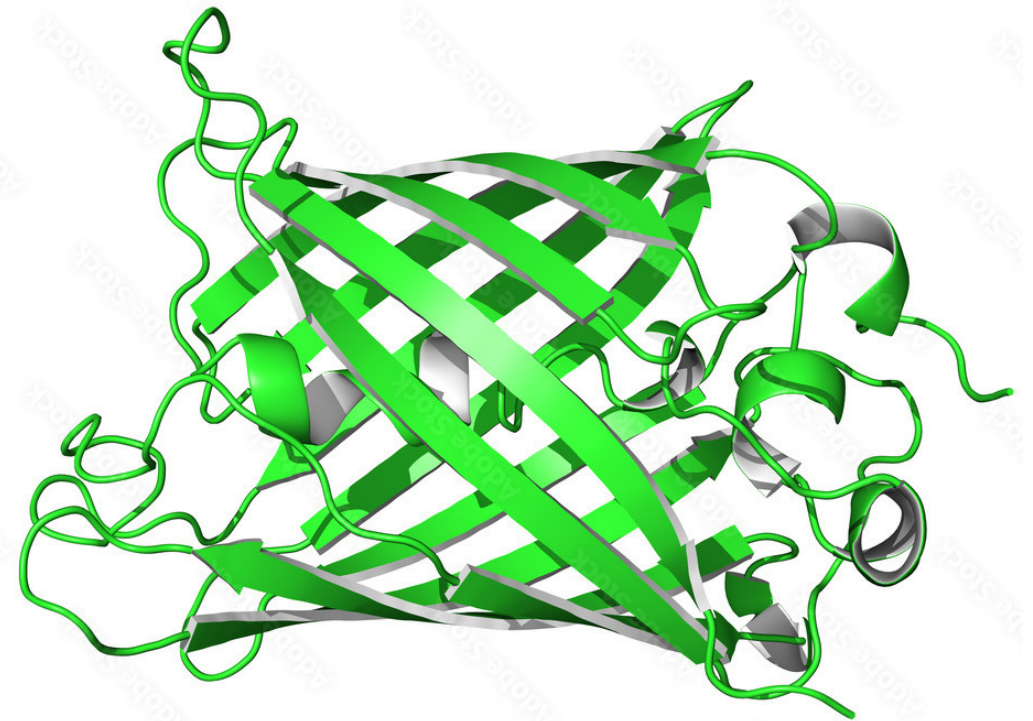
$$\mathbf{h}_i^{l+1} = \phi_h(\mathbf{h}_i^l, \mathbf{m}_i)$$

Molecular Design Problems

- Small-molecule drug design
 - bind to a protein target
 - with desired properties: toxicity, synthesizability, drug likeness (QED), ...
- Protein design:
 - Enzyme
 - Protein Binder
 - Biosensor: GFP
 - Antibody
- RNA design
- ...

Protein Design Approaches

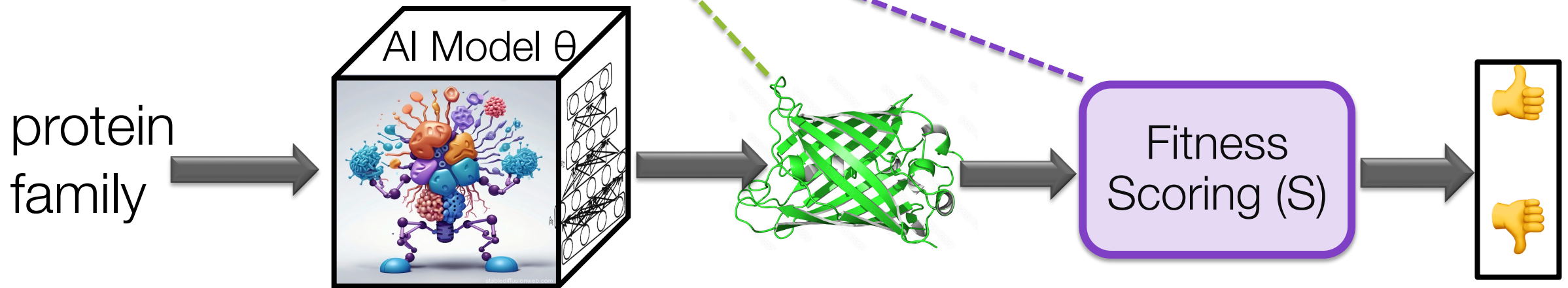
- ➔ • Sequence-based Generation
- Structure-based Generation
 - Secondary structure-based
 - Inverse Folding
 - Surface geometry
- Sequence-Structure Co-design



Guiding Protein Generation with Function Fitness

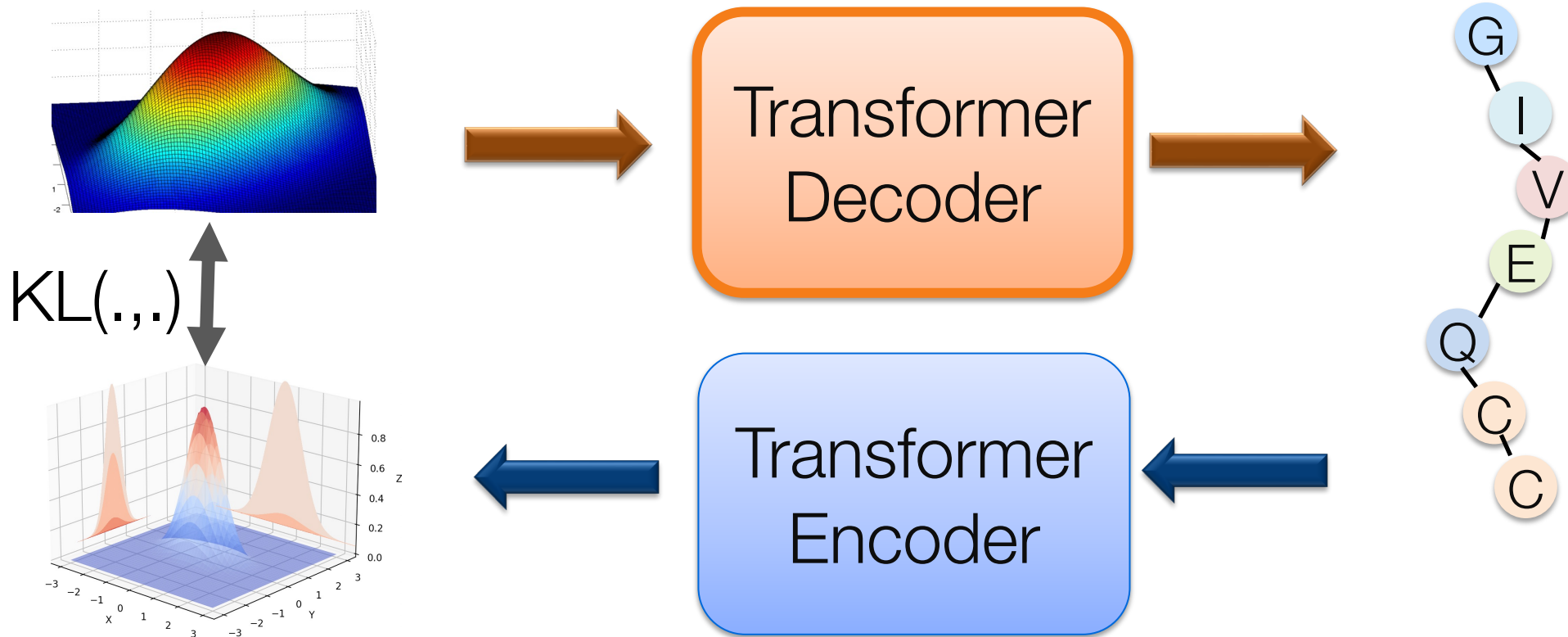
- Fitness functions $P(S|x)$ can be trained using lab data
 - e.g. Green Fluorescent Protein (avGFP) [Sakisyan et al 2016]

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



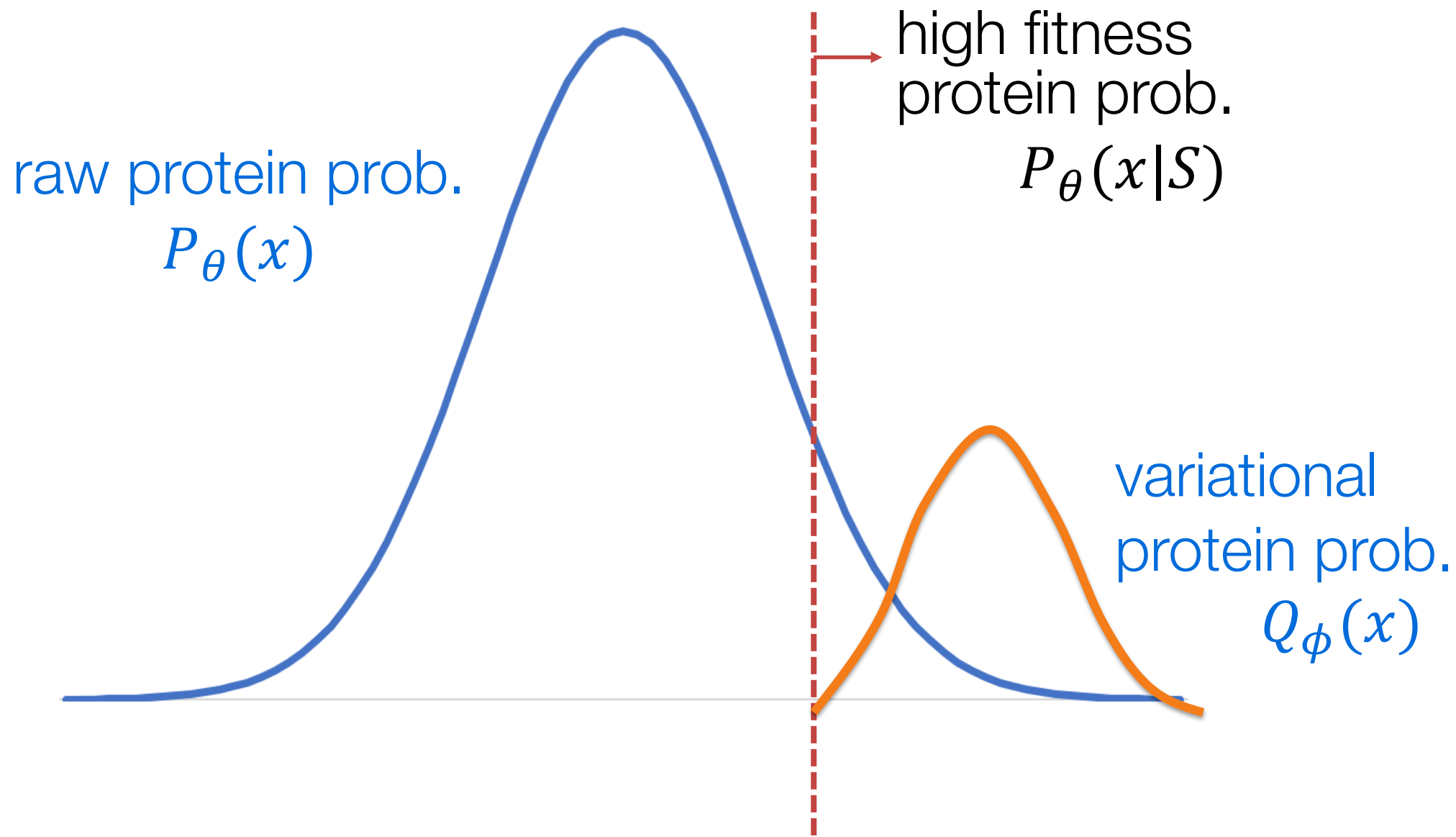
Pre-training Protein Generative Model

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



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

IsEMPro: Intuition

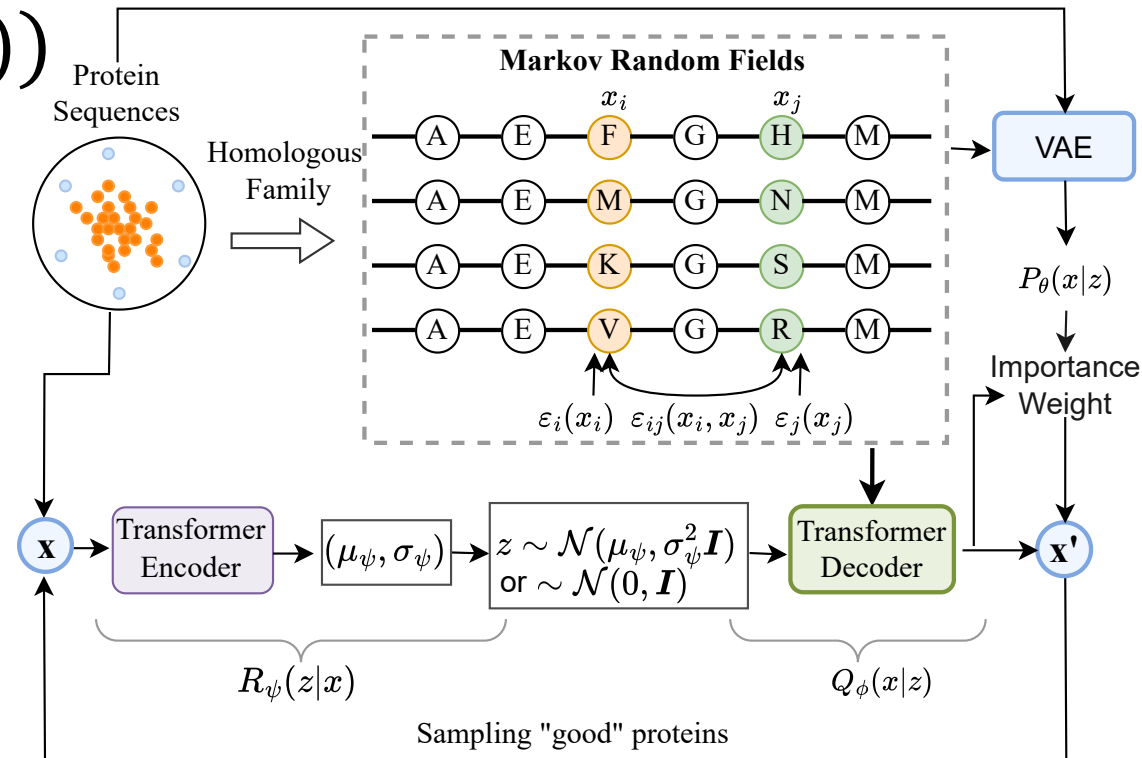


IsEMPro Method

- Intuition:
 - Learning a proposal $Q_\phi(x)$ to approximate distribution of "good" proteins $P_\theta(x|S)$
$$\phi^* = \operatorname{argmax}_\phi -D_{KL}(P_\theta(x|S) || Q_\phi(x))$$

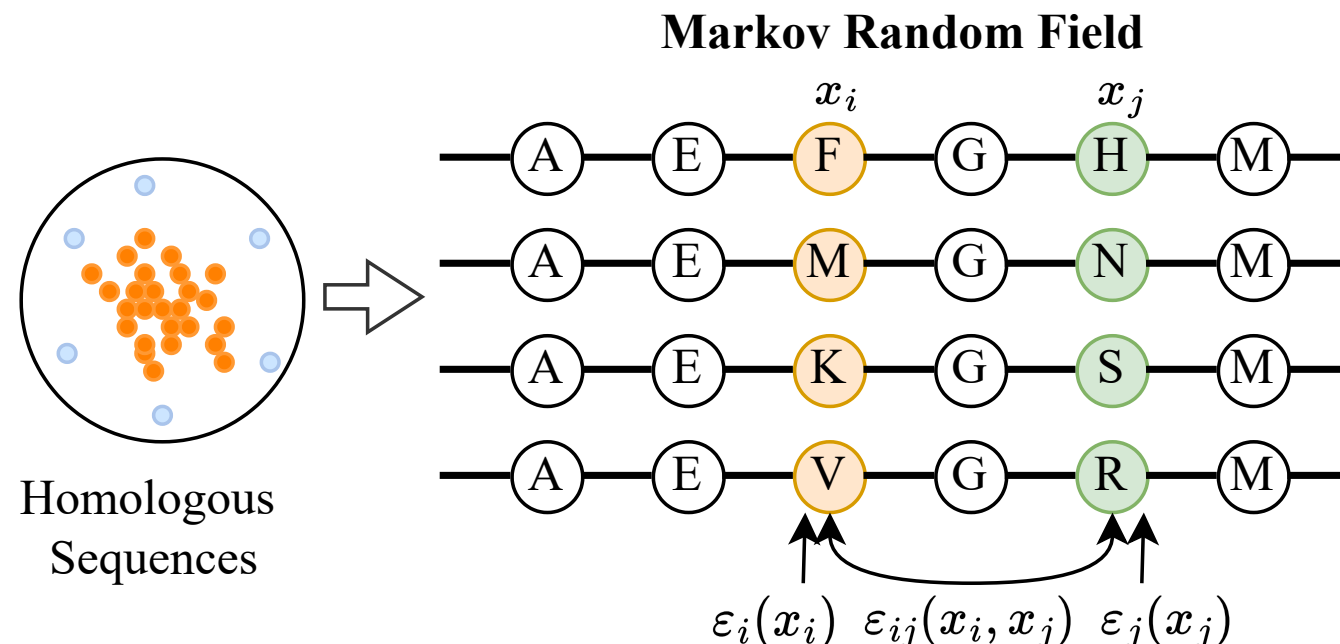
- Model architecture:
 - two VAEs
 - Augmented with MRF features

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



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



Integrating MRF into IsEMPro Generation

- MRFs features (i-th residue)

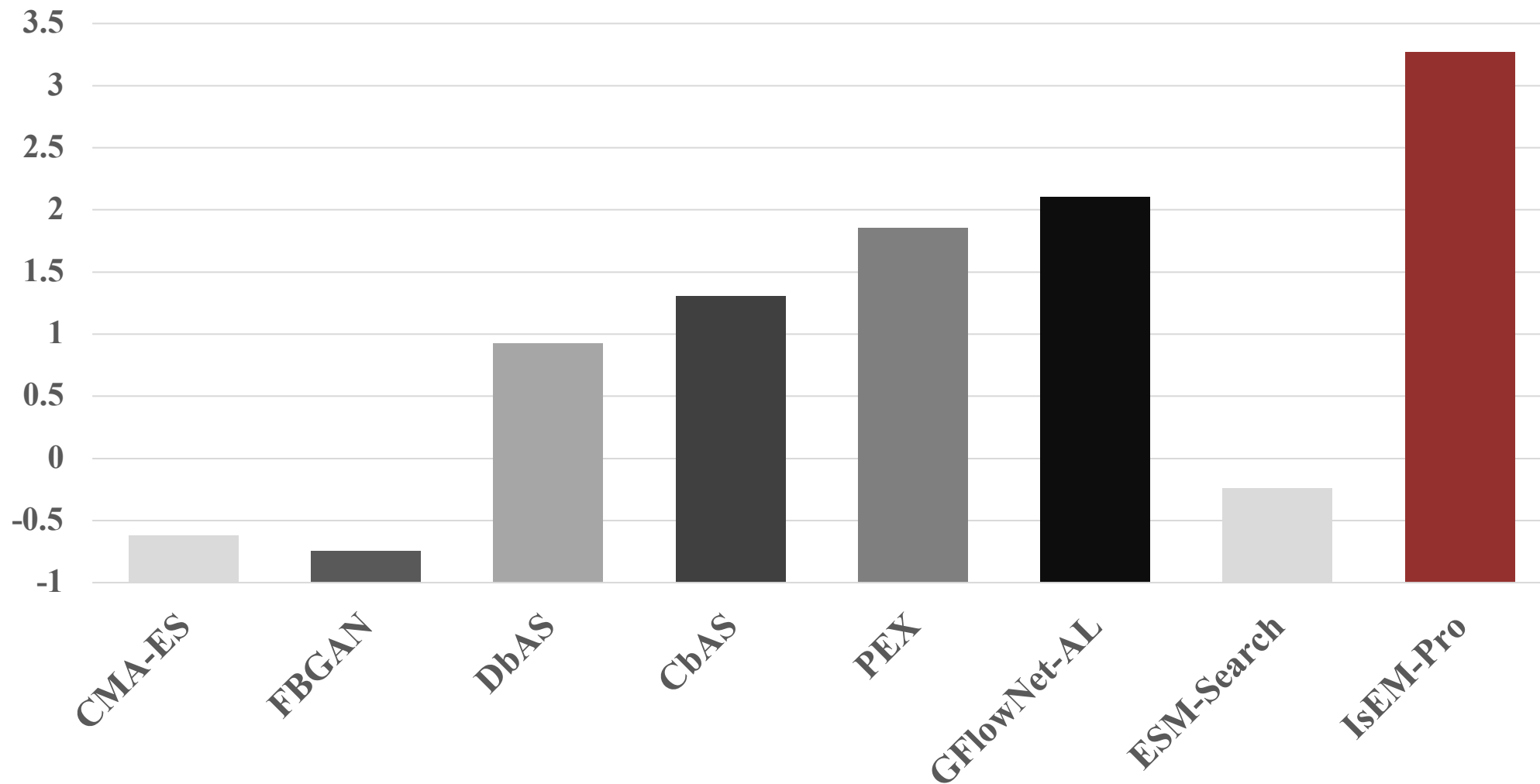
$$\begin{aligned}\varepsilon_i(x_i) &= [\varepsilon_i(x_i), \varepsilon_{i1}(x_i, a_{1.}), \dots, \varepsilon_{iM}(x_i, a_{M.})] \\ \varepsilon_{ij}(x_i, a_{j.}) &= [\varepsilon_{ij}(x_i, a_1), \varepsilon_{ij}(x_i, a_2), \dots, \varepsilon_{ij}(x_i, a_{20})]\end{aligned}$$

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

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

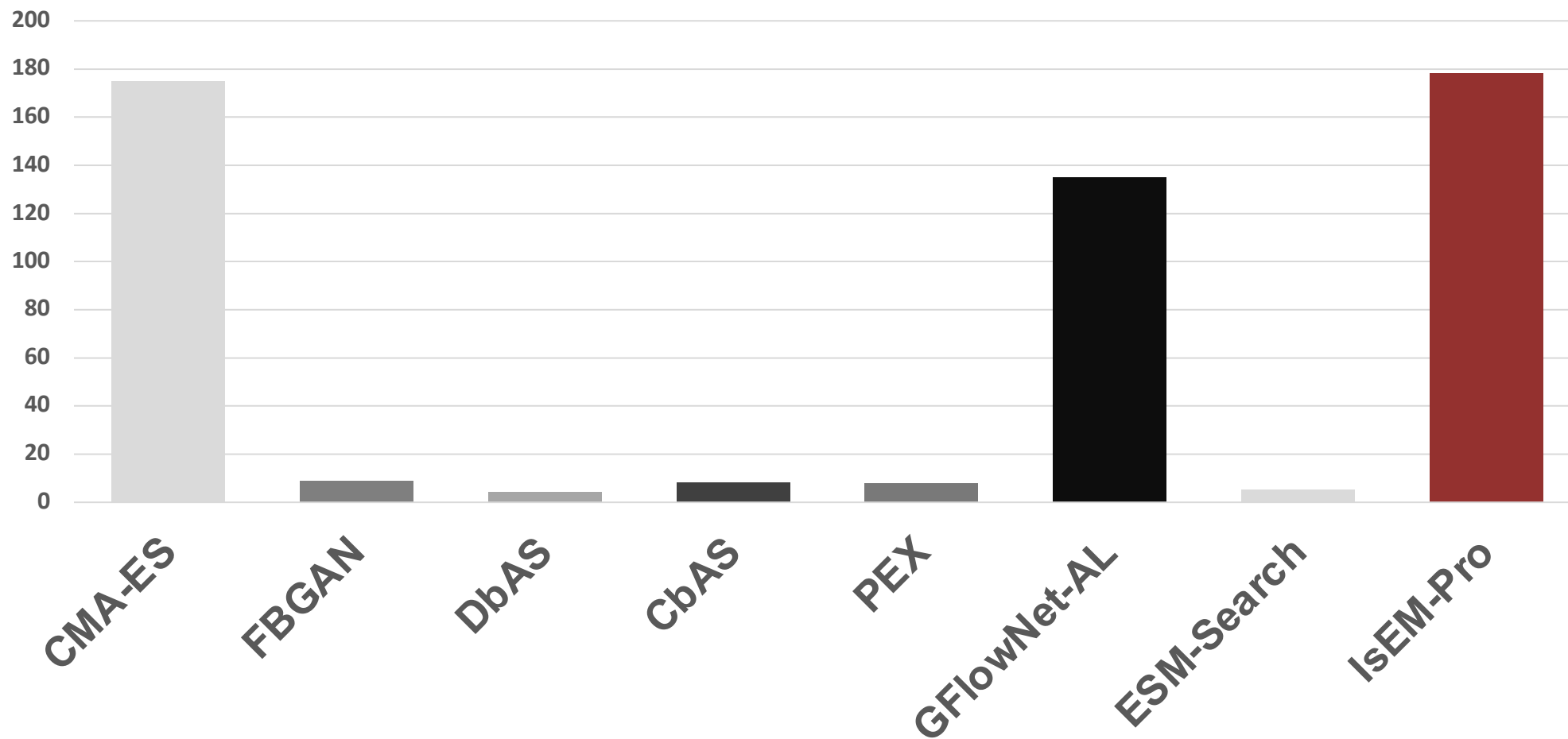
IsEM-Pro generates higher-fitness proteins

Average Fitness on Eight Protein Datasets



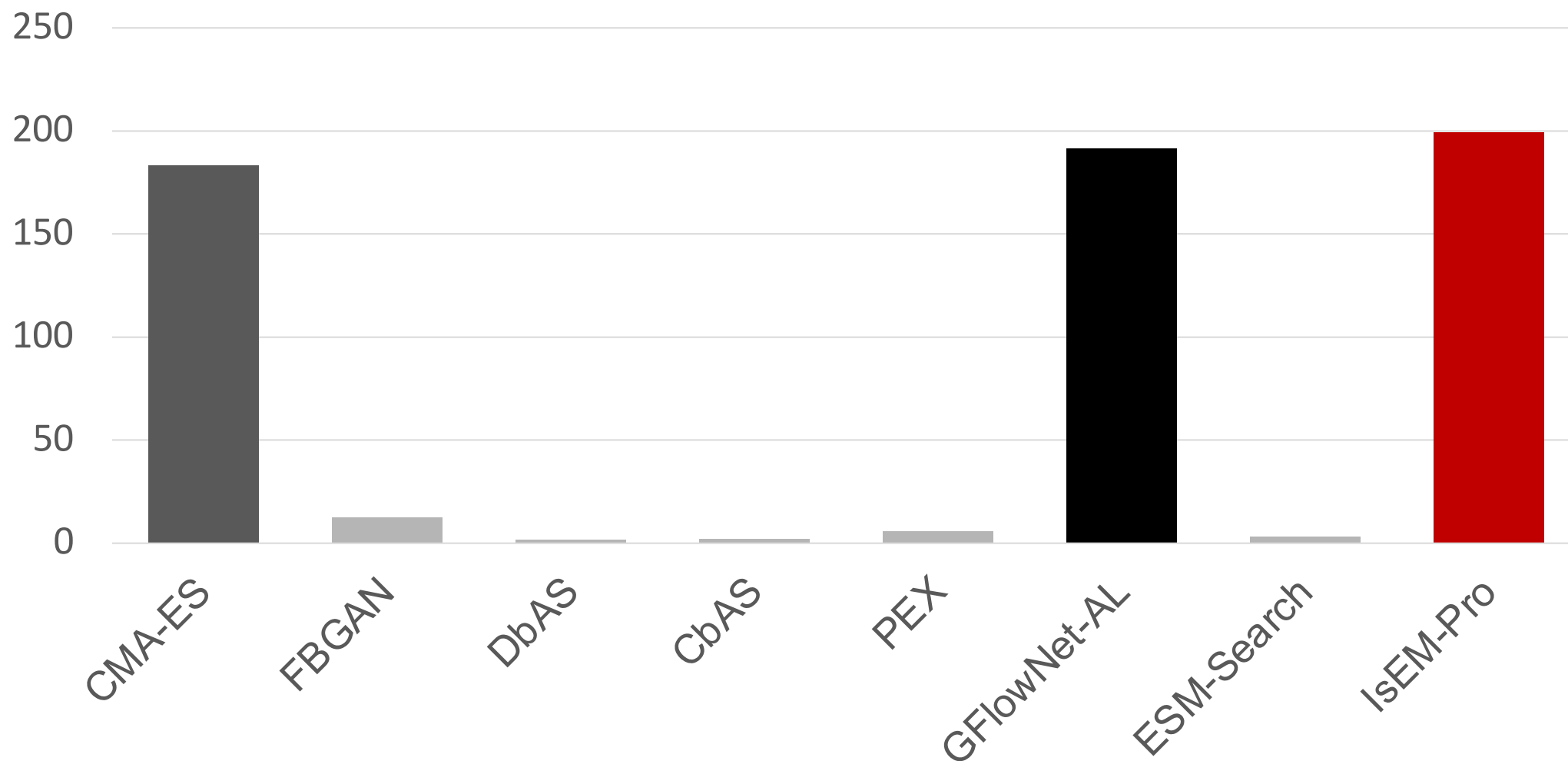
IsEM-Pro generates more diverse proteins

Average Diversity on Eight Protein Datasets

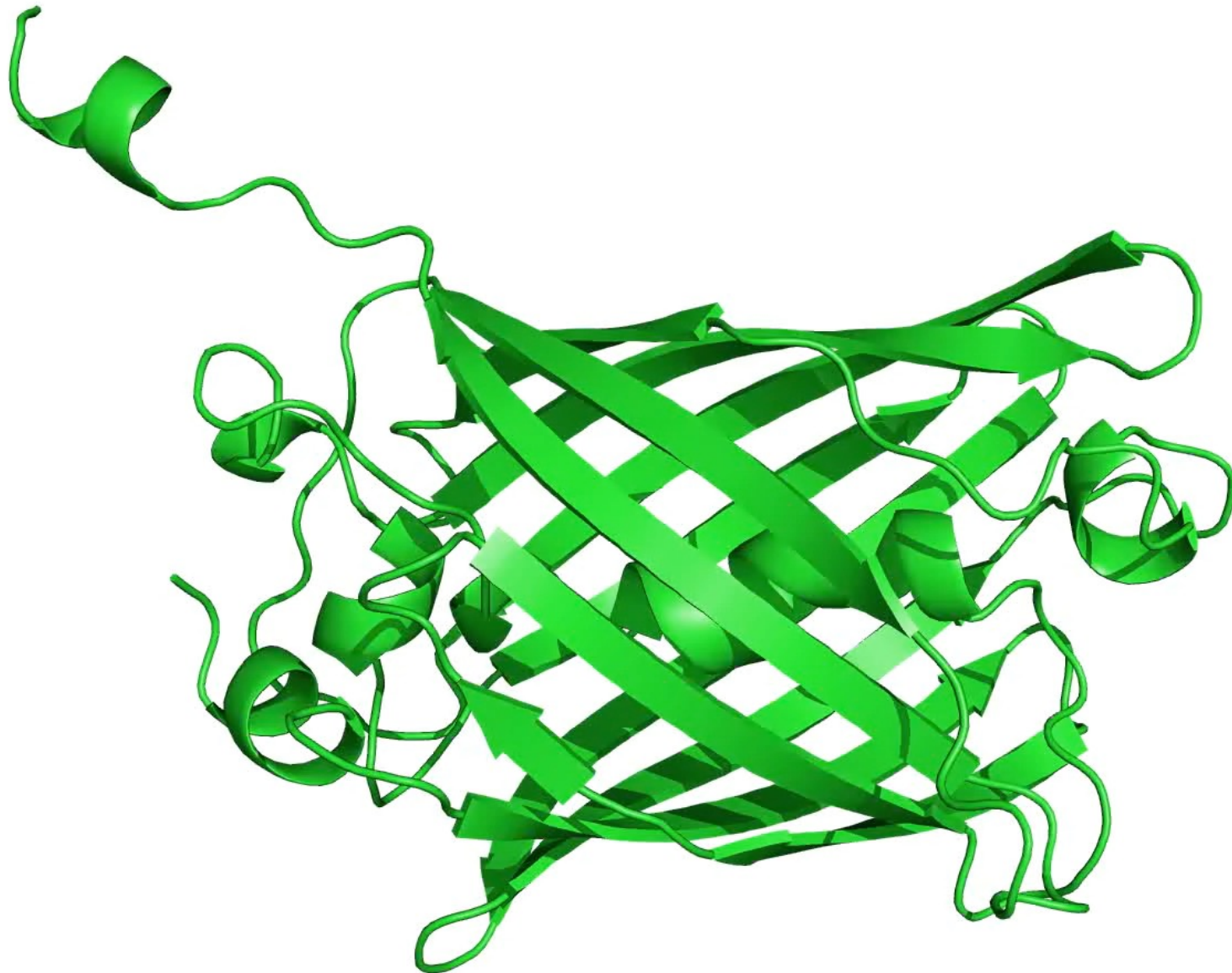


IsEM-Pro achieves the highest average novelty score

Average Novelty Score on Eight Protein Dataset



Green Fluorescent Protein designed by IsEMPro

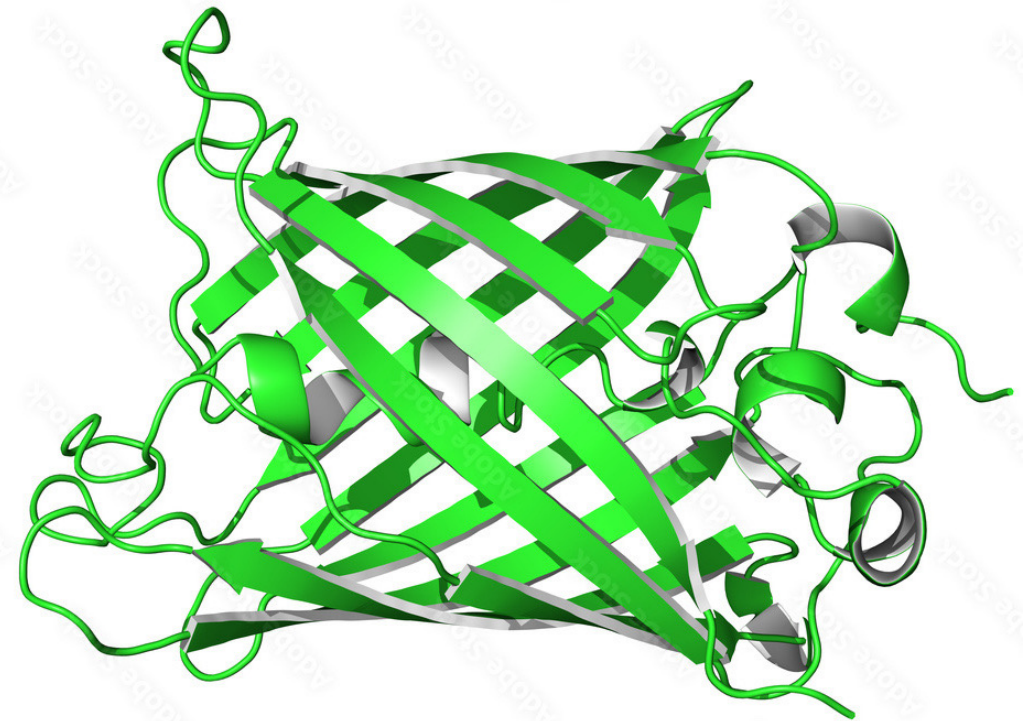


Highlights of IsEM-Pro

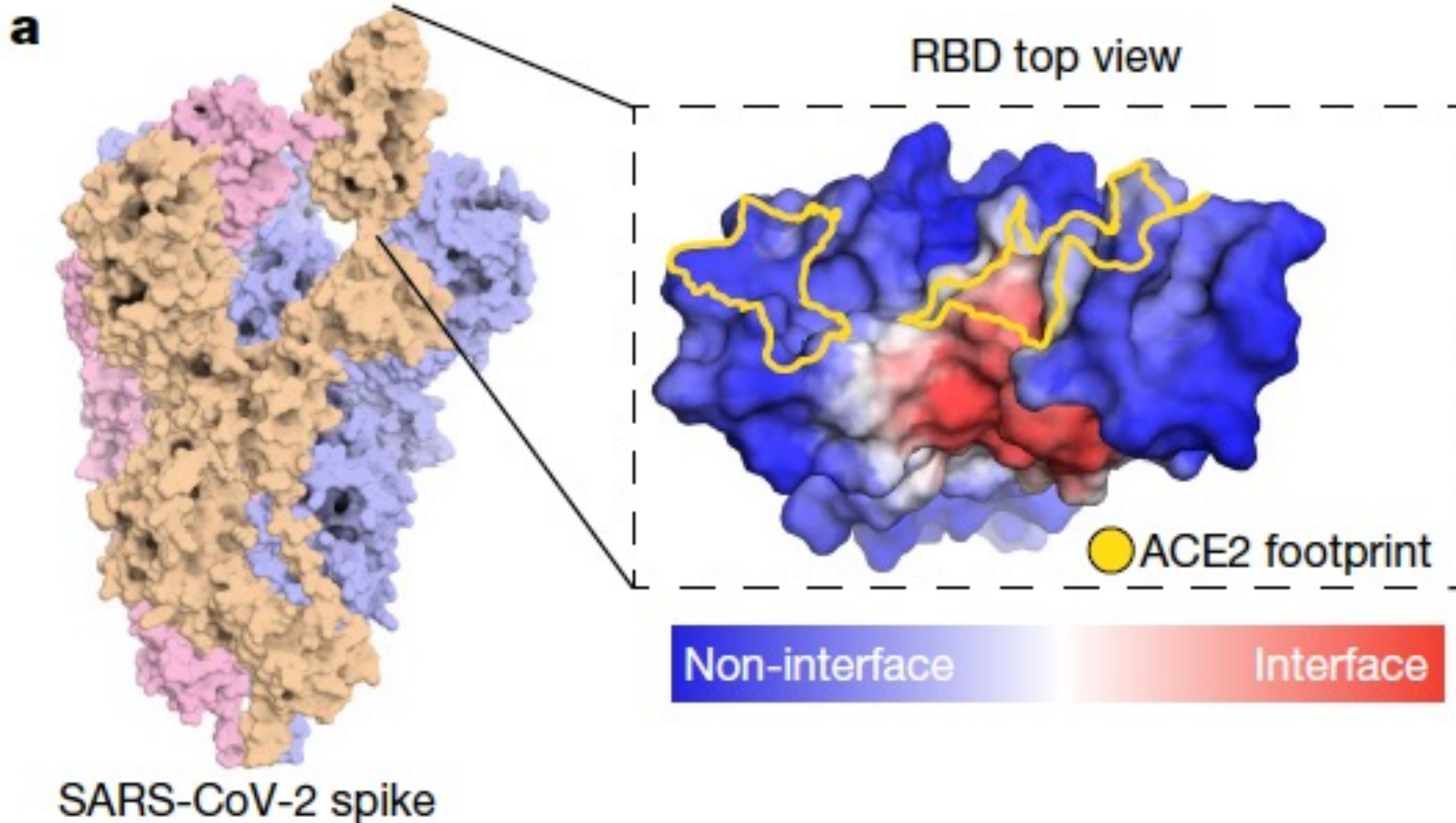
- Using importance sampling inside the EM is efficient to generate functional proteins
- The combinatorial enhanced latent generative model boosts diverse and novel protein sequences
- The self-learning process helps to find proteins with higher fitness scores

Protein Design Approaches

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

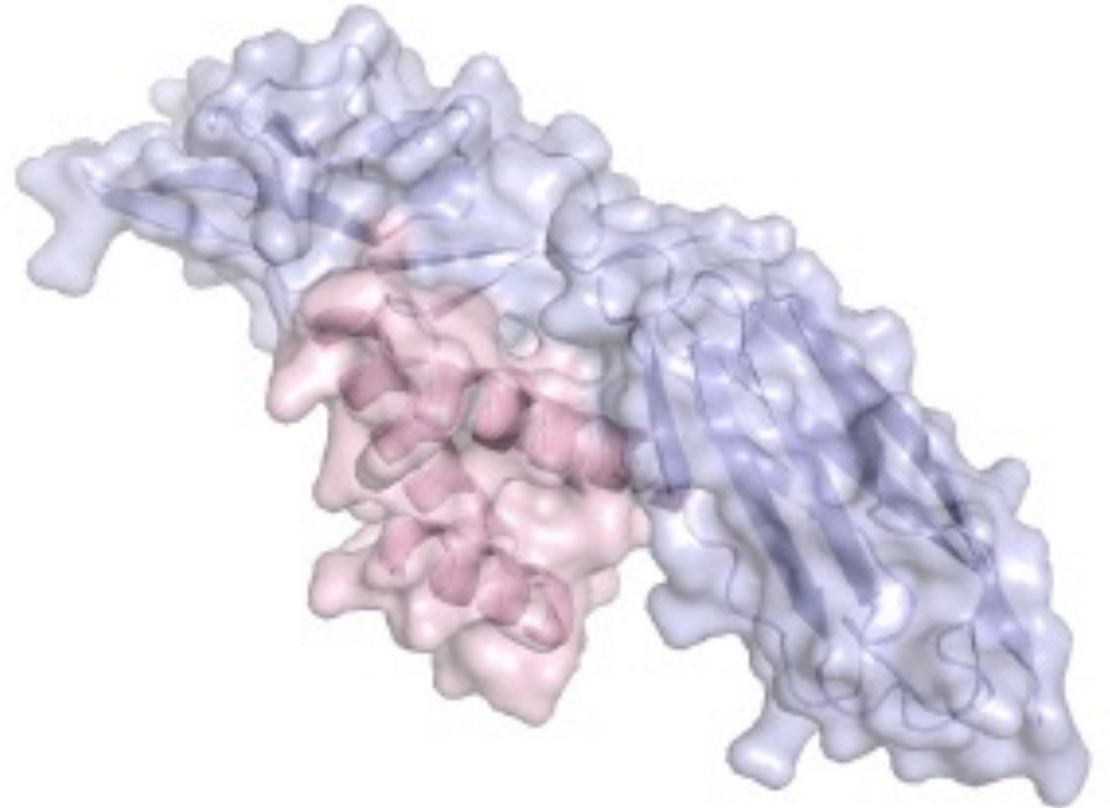


Protein Surface



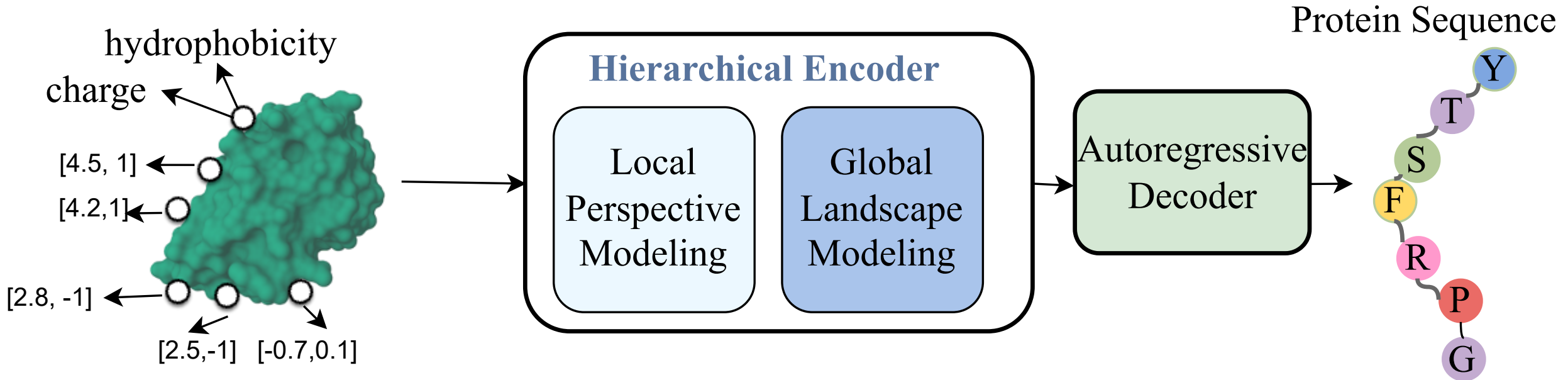
Surface-based Protein Design

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





SurfPro Method



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
 - Gaussian kernel smoothing – higher expressiveness

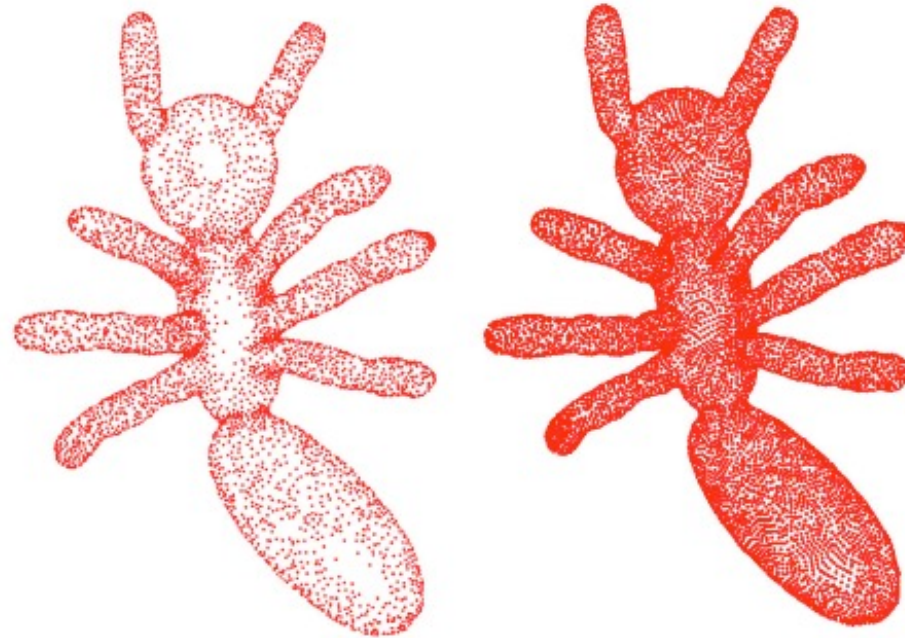
$$x'_i = \sum_{x_j \in N(x_i)} \frac{\kappa(x_i, x_j) x_j}{\sum_{x_t \in N(x_i)} \kappa(x_i, x_t)}, \quad \kappa(x, y) = e^{-\frac{(x-y)^2}{\eta}}$$



Surface Construction

- Surface compression
 - Octree-based downsampling
 - Convert the surface into small cubes
 - Each cube is recursively divided into 8 octants – minimum points

$$N = V \cdot r$$



Hierarchical Encoder: Local Perspective Modelling

- K-nearest equivariant graph convolutional layers

- Local Message

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

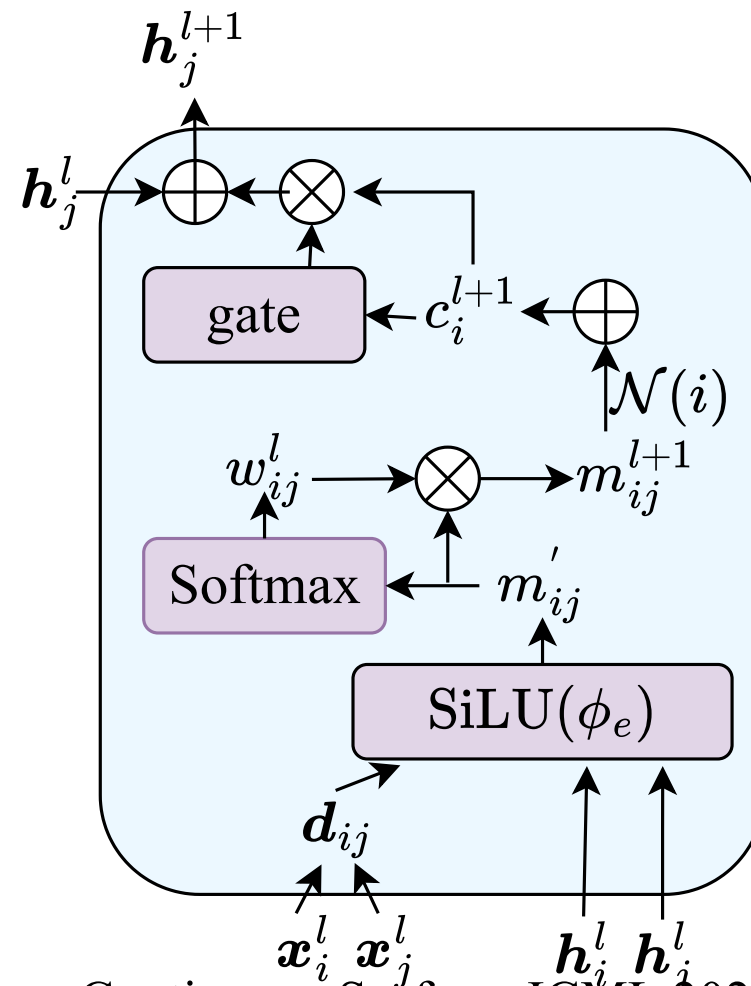
$$w_{ij}^l = \frac{\exp(W_s^l m'_{ij} + b_s^l)}{\sum_{k \in N(x_i)} \exp(W_s^l m'_{ik} + b_s^l)}$$

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

- Vertex feature representation

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

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



Hierarchical Encoder: Global Landscape Modelling

- Frame Calculation

- Point cloud X – PCA – three principle componen

- Map a 3D molecule into the 8 coordinate systems

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

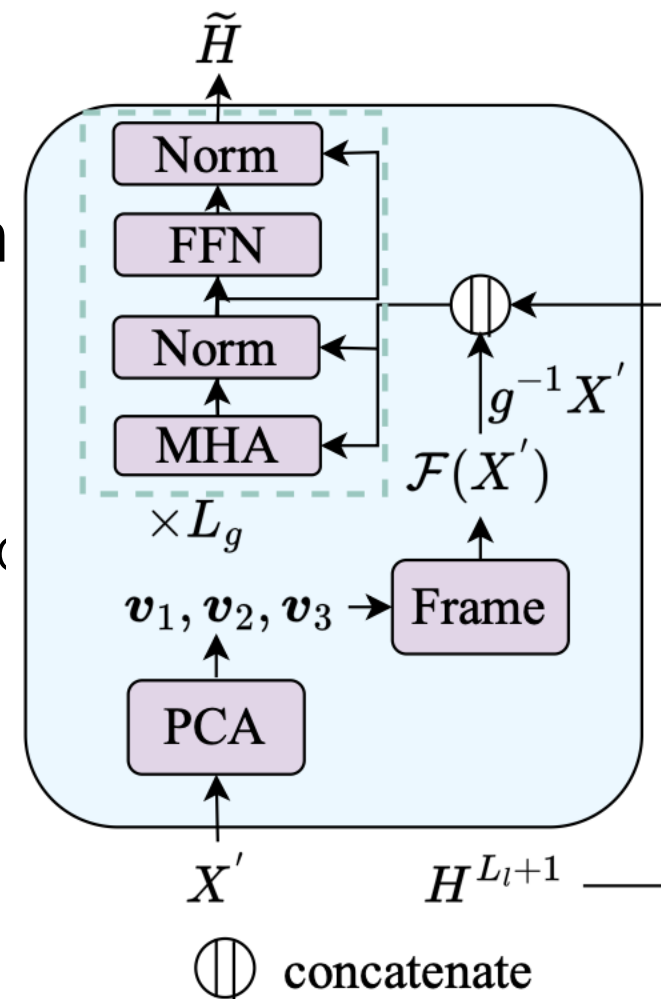
- Average the representations across 8 frames

- Equal to any translation + Rotation operation theoretic

- Global Landscape Modeling - MHA

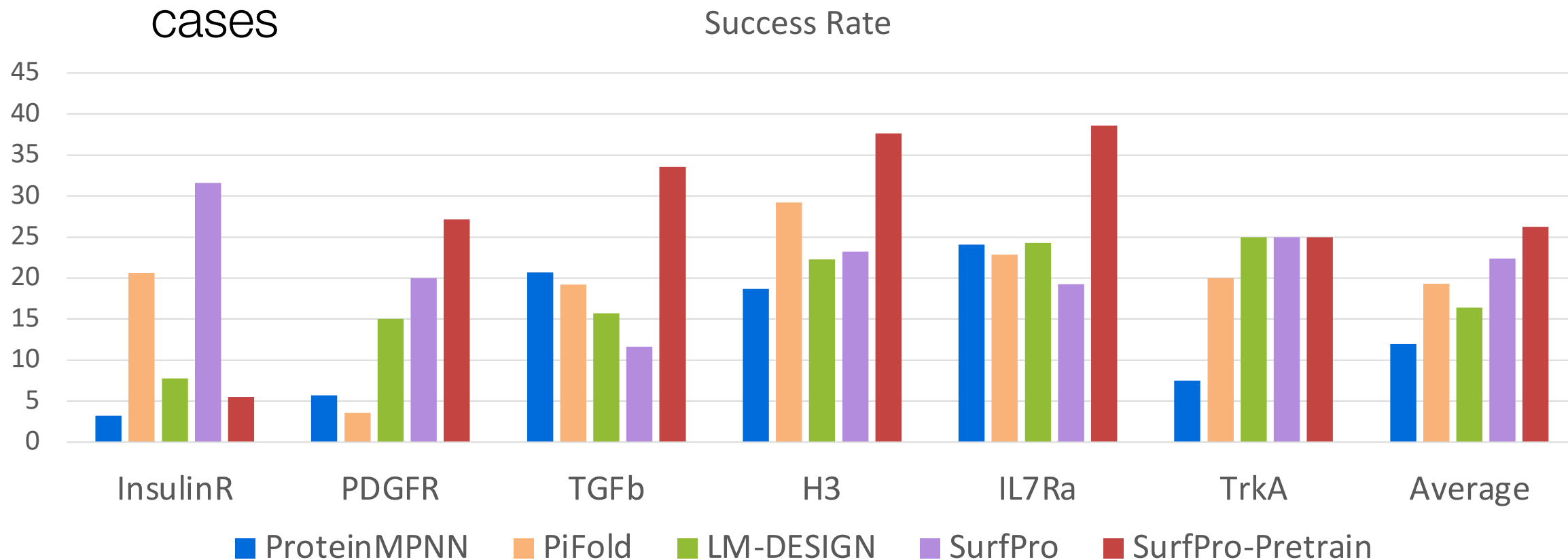
- Autoregressive decoder

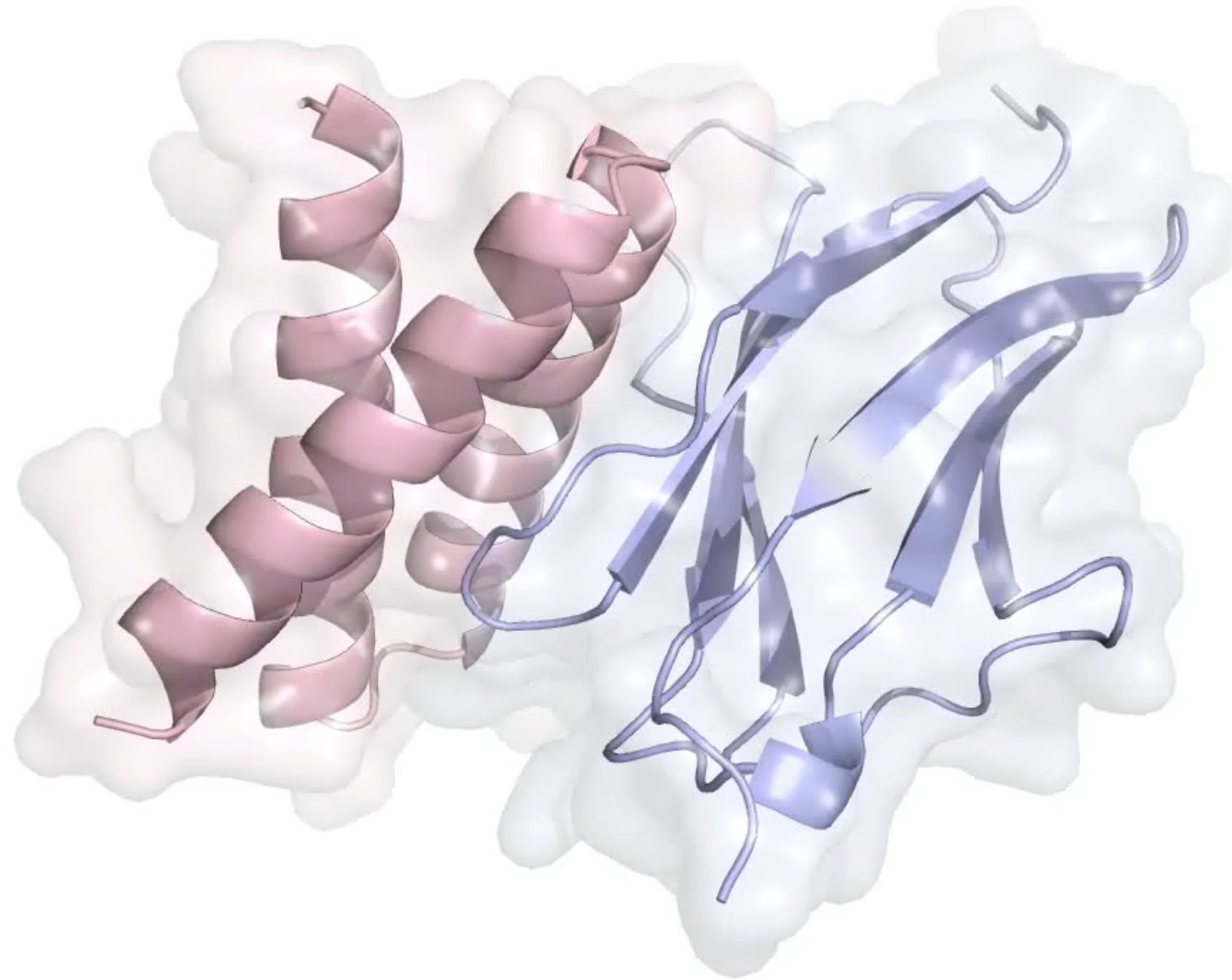
- Maximum likelihood optimization



SurfPro generates more successful binders

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



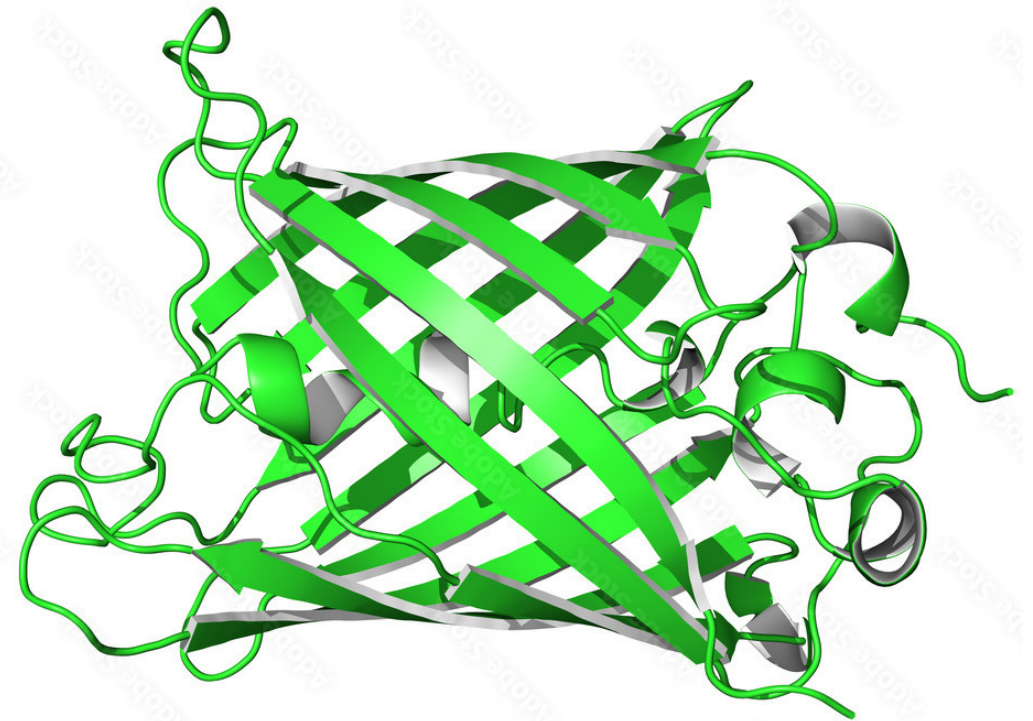


Highlights of SurfPro

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

Protein Design Approaches

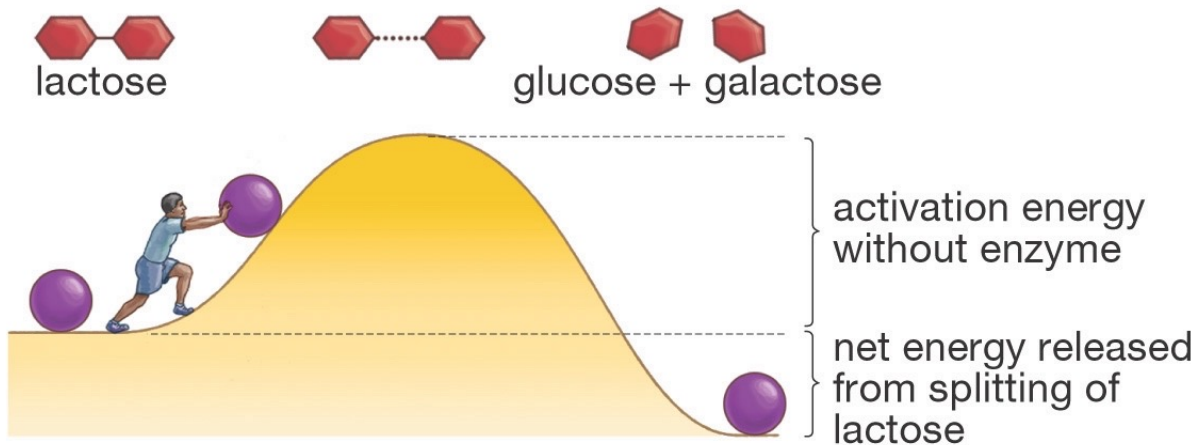
- Sequence-based Generation
- Structure-based Generation
 - Secondary structure-based
 - Inverse Folding
 - Surface geometry
- ➔ • Sequence-Structure Co-design



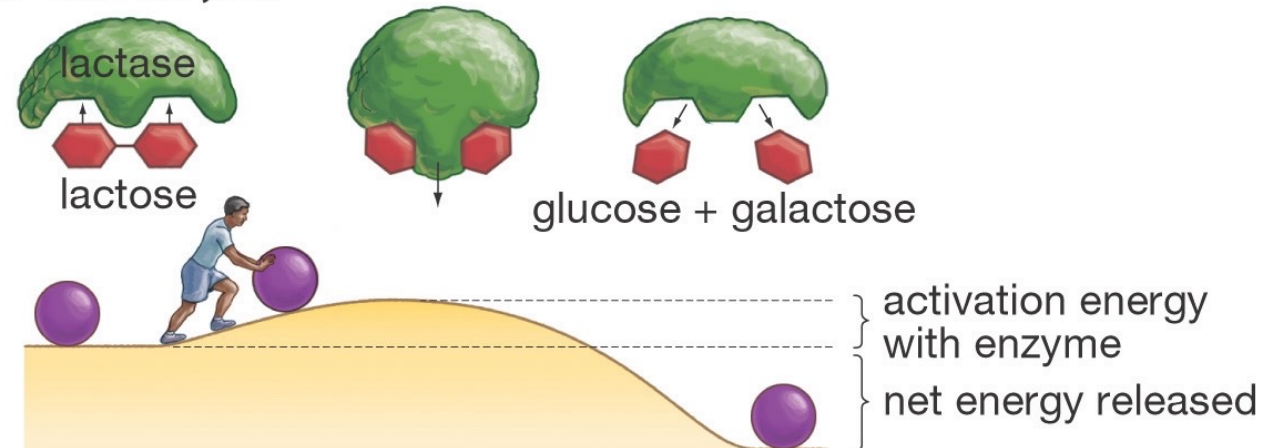
Enzyme

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

(a) Without enzyme

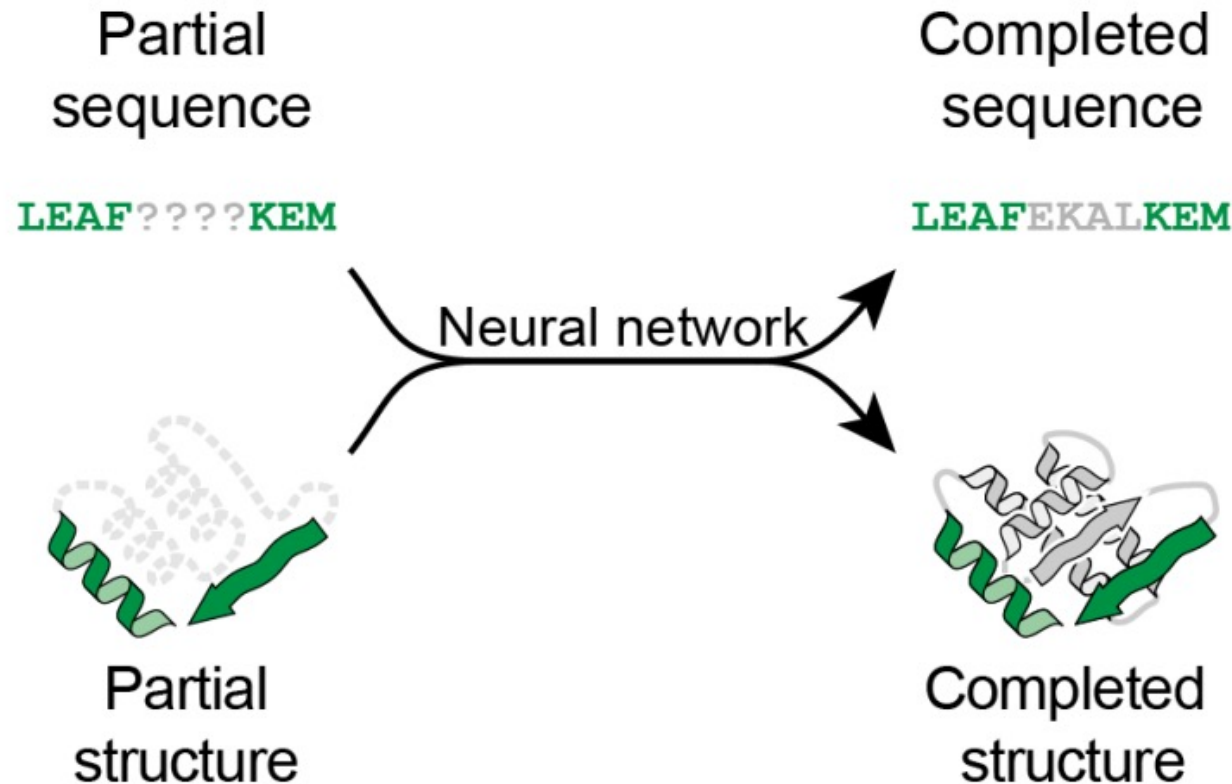


(b) With enzyme



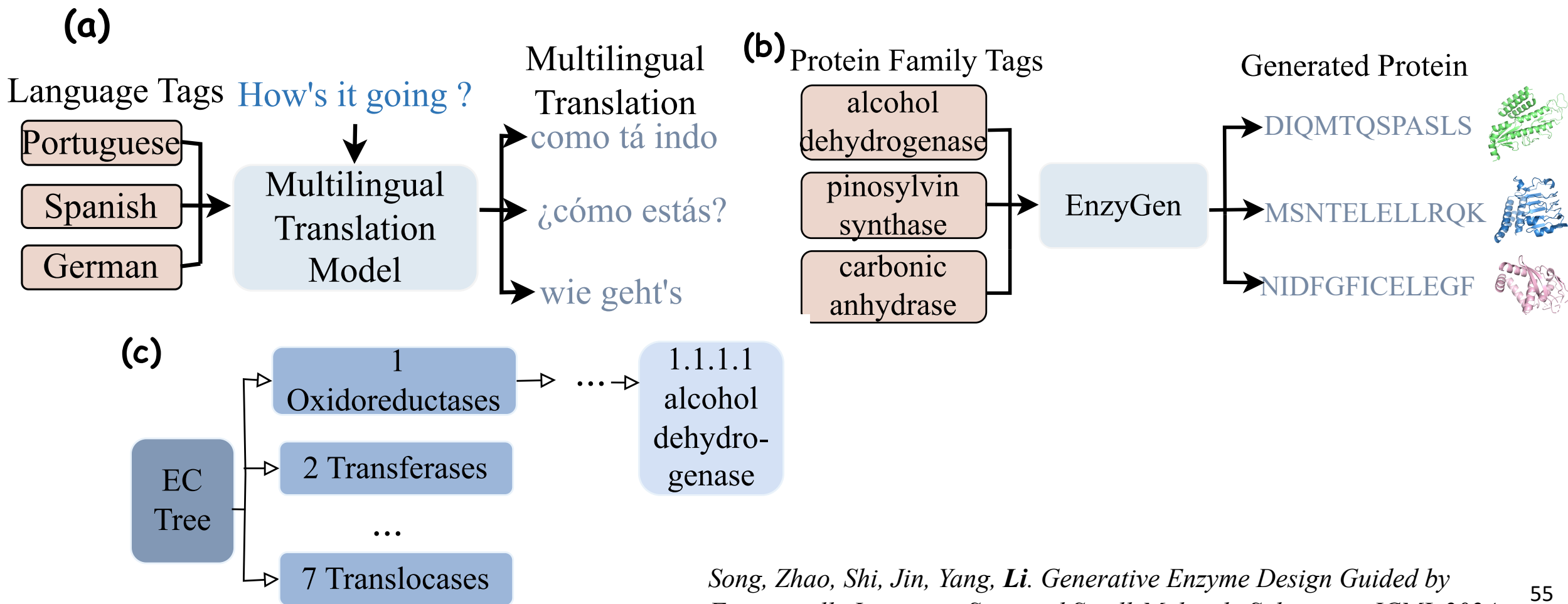
Motivation 1: How to design desired enzymes?

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



Motivation 2: How to design desired enzymes?

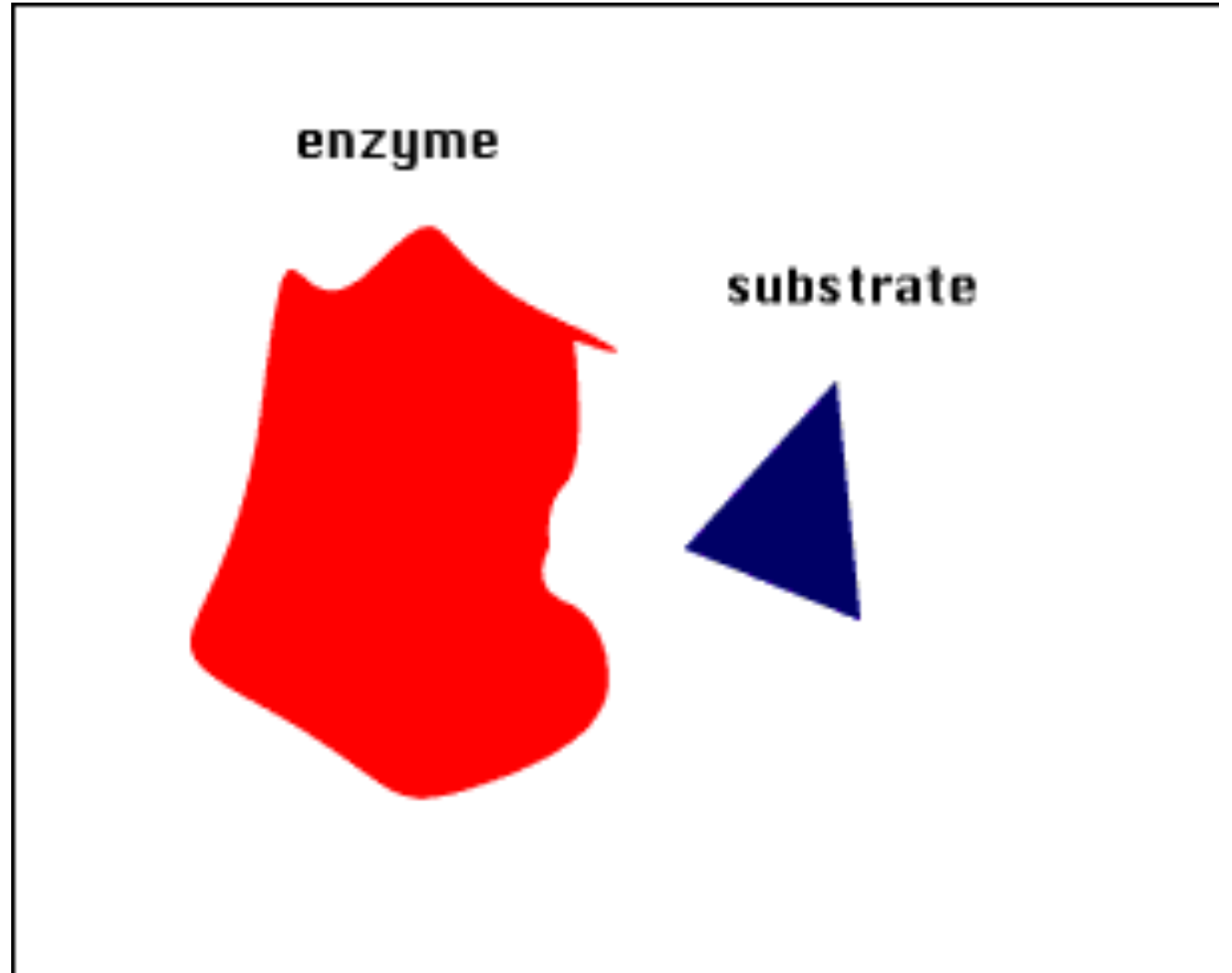
Enzyme classification tree indicates enzymatic reaction type



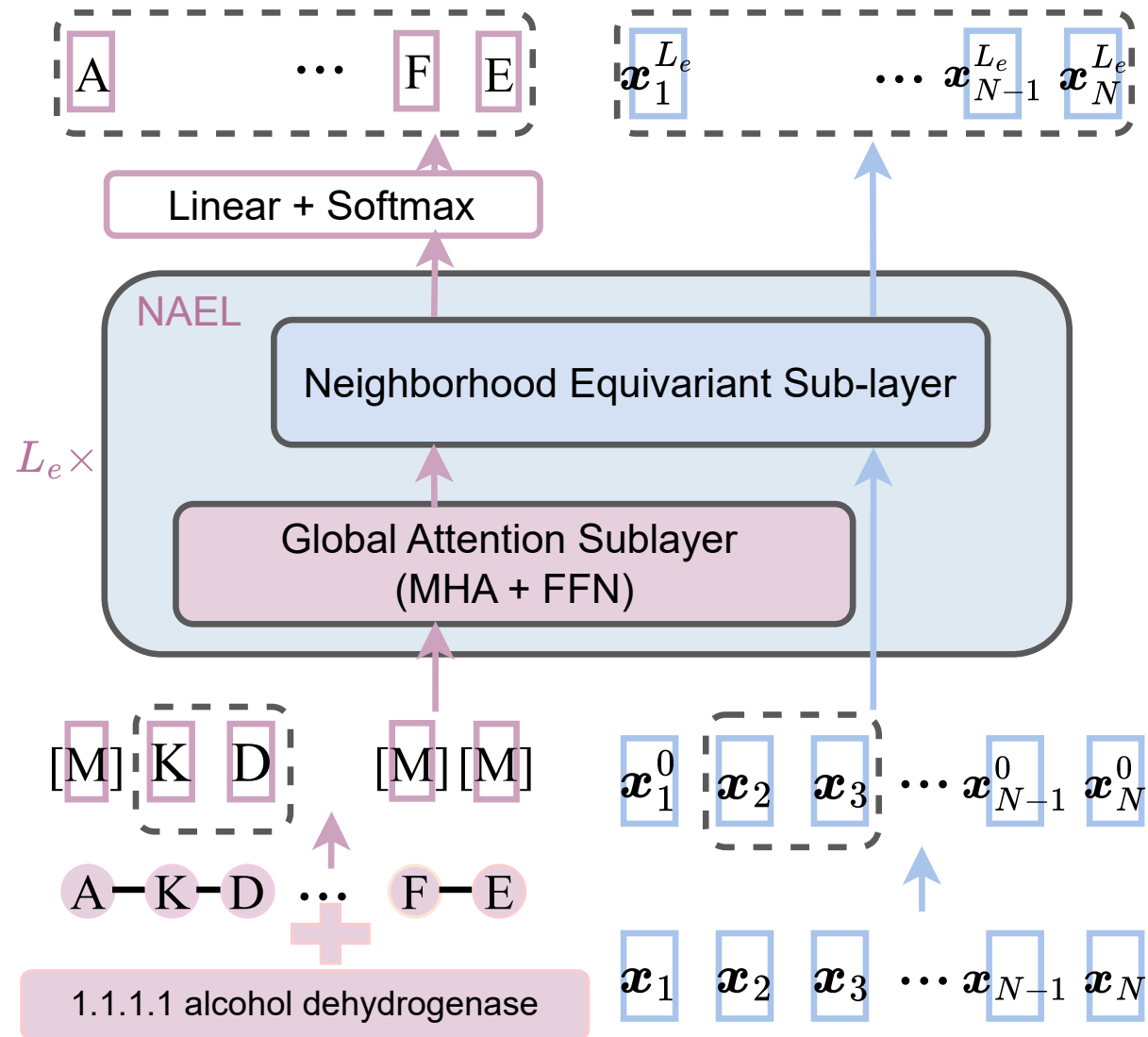
Motivation 3: How to design desired enzymes?

- Substrate Specificity:

Different enzymes binding to specific substrates to speedup enzymatic reactions

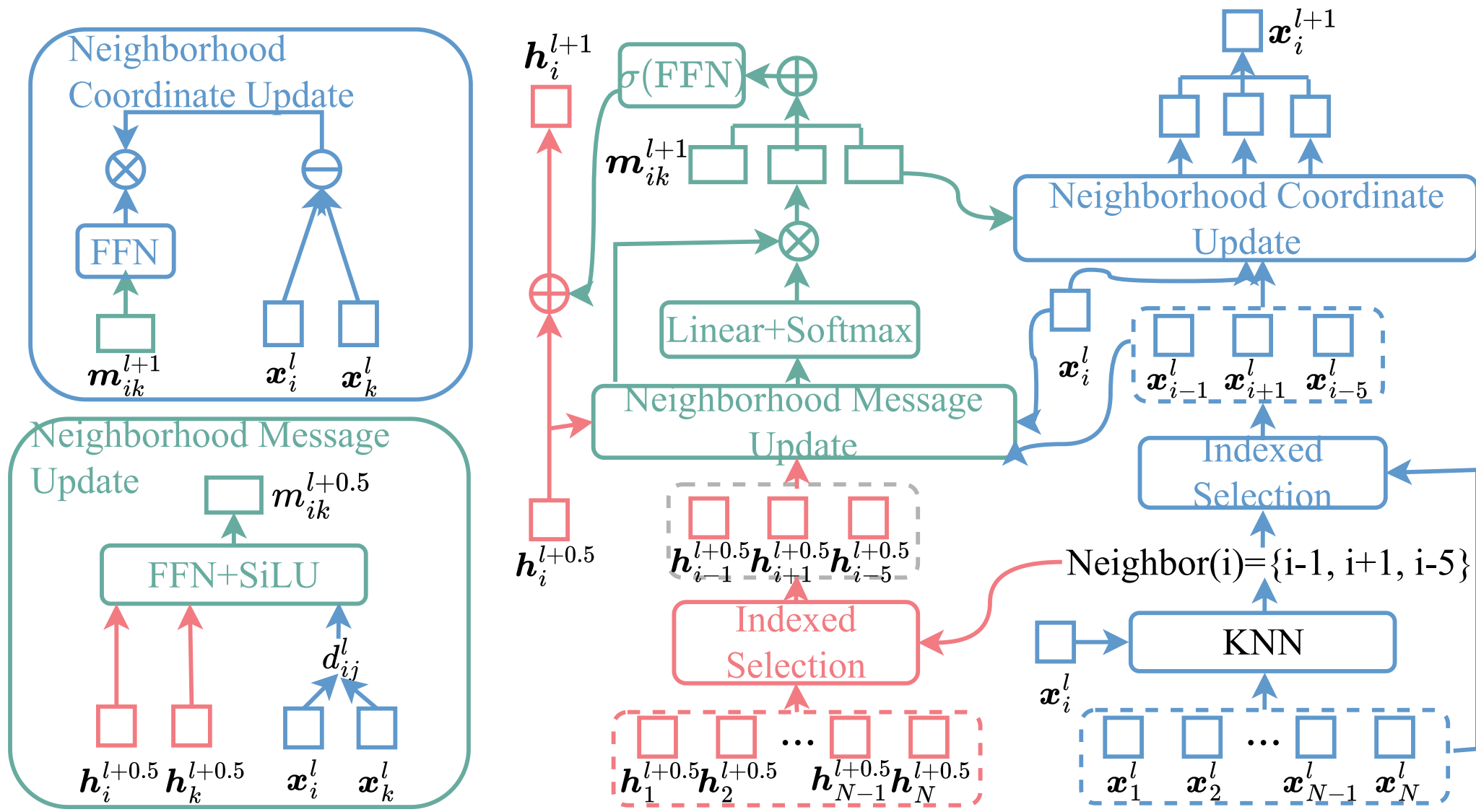


EnzyGen Model – NAEL backbone

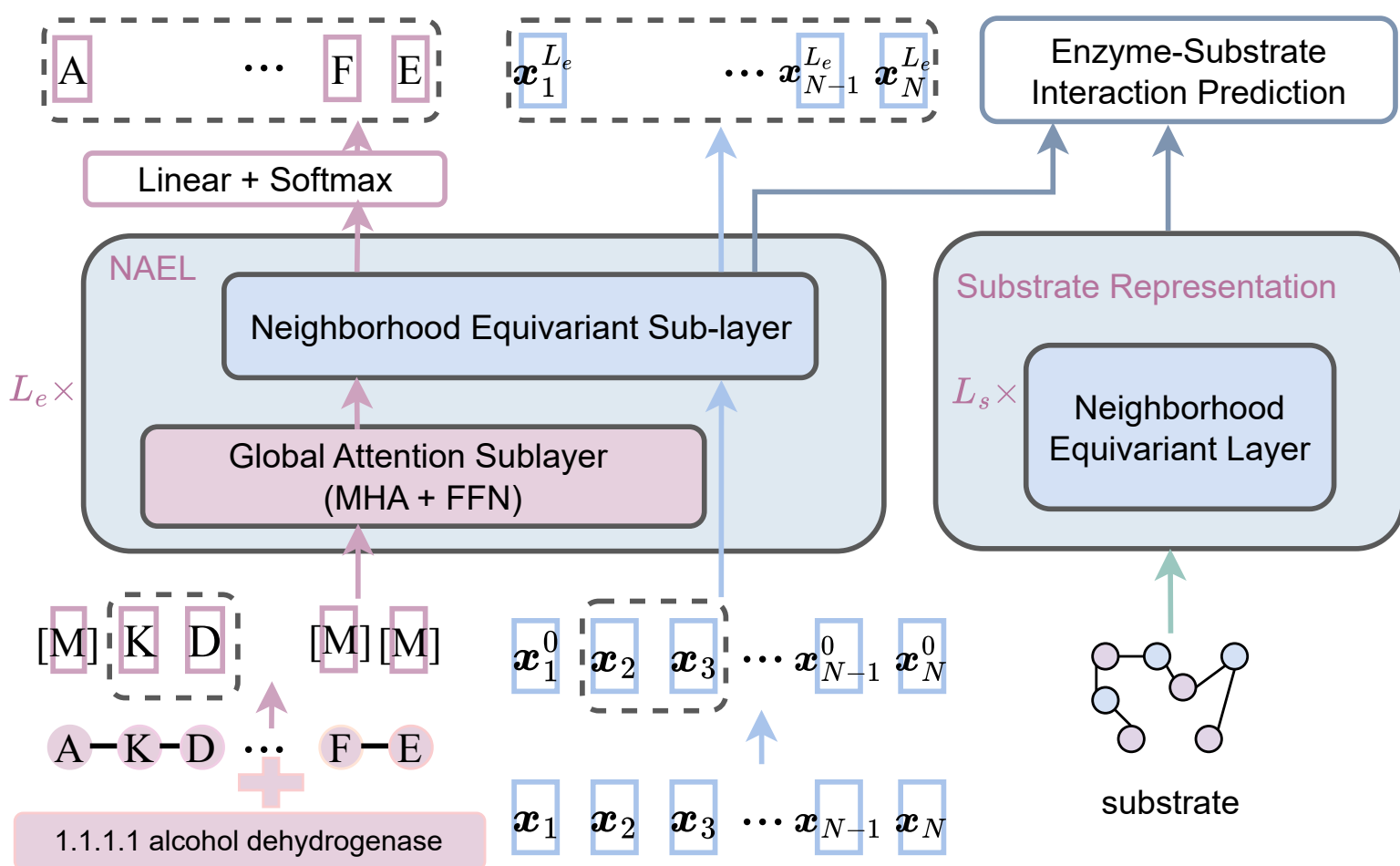


- Controllable Design
 - Functional Sites
 - Enzyme family category

Neighborhood Attentive Equivariant Layer



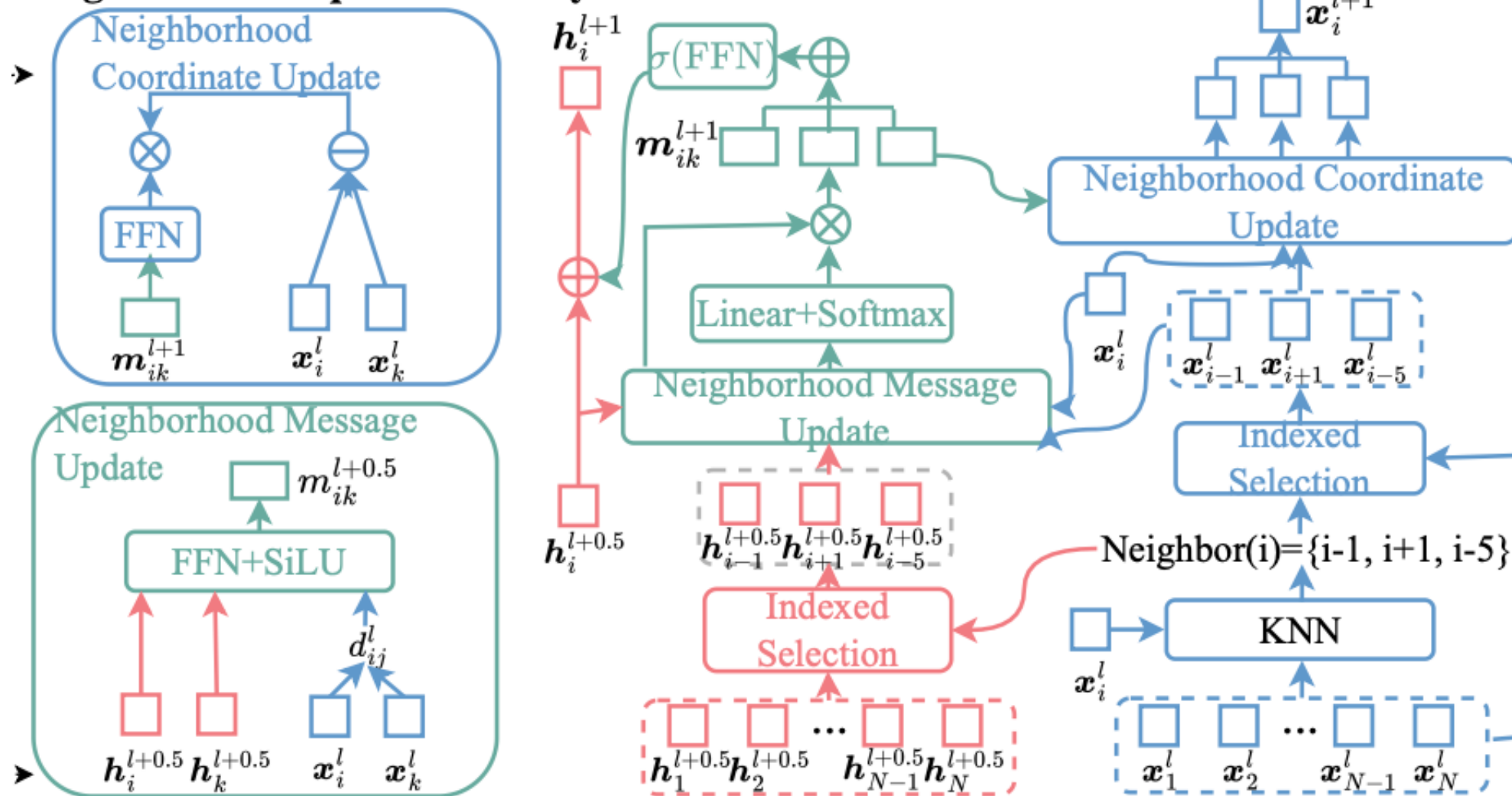
EnzyGen Learning



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

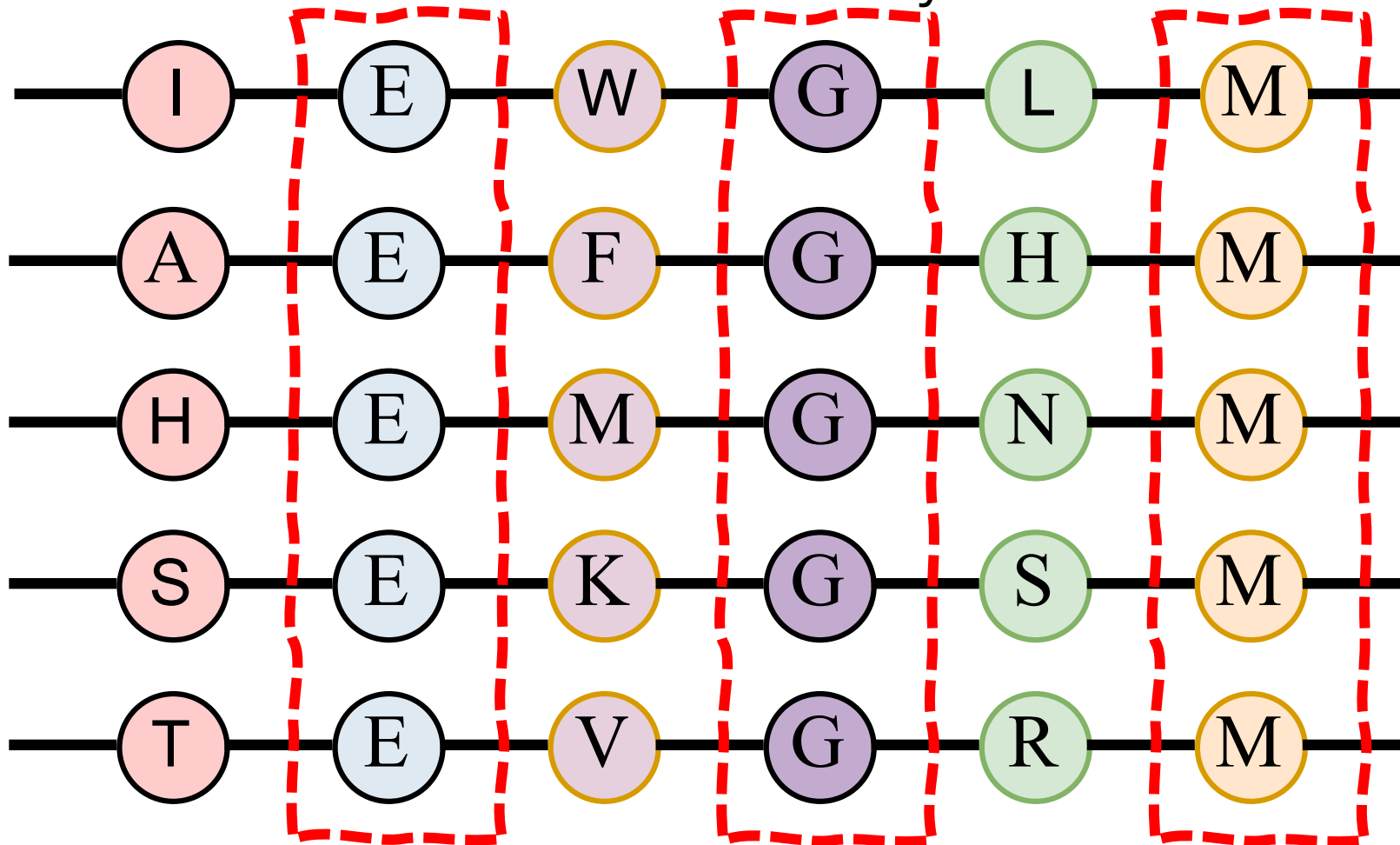
Neighborhood Attentive Equivariant Layer (

Neighborhood Equivariant Layer



Functional Site Discovery

mining common sites within one family

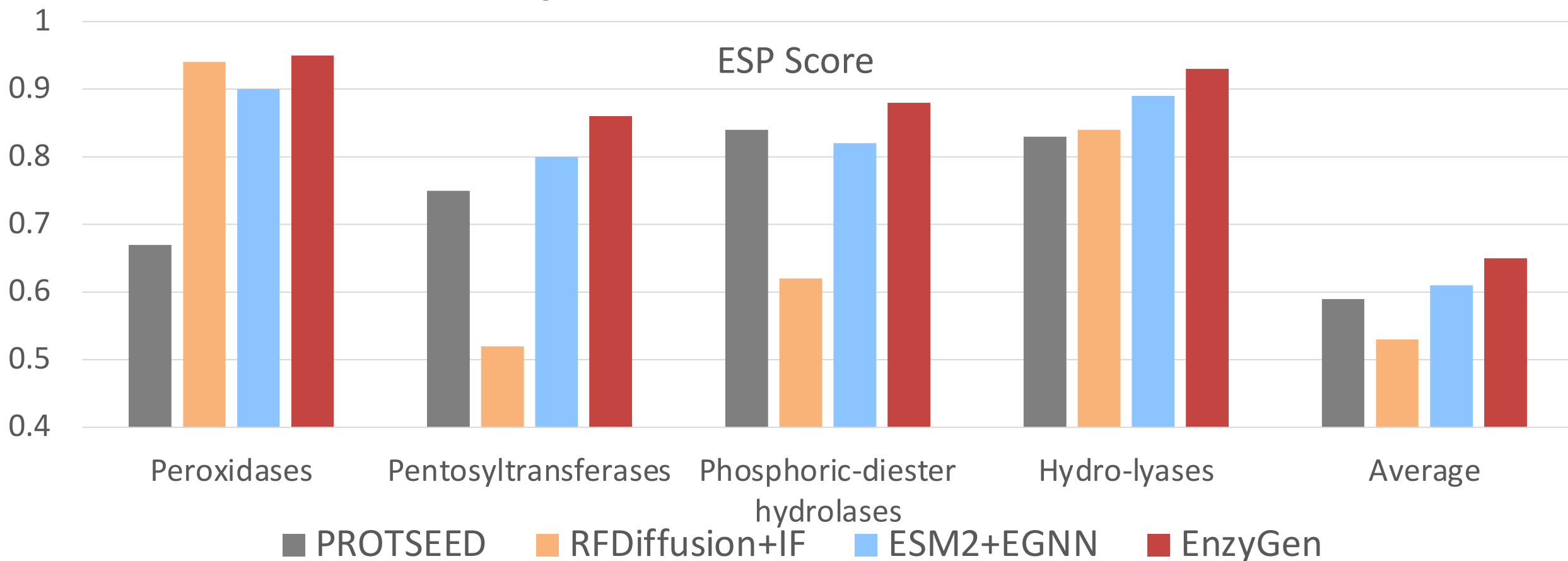


EnzyBench Dataset

- 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

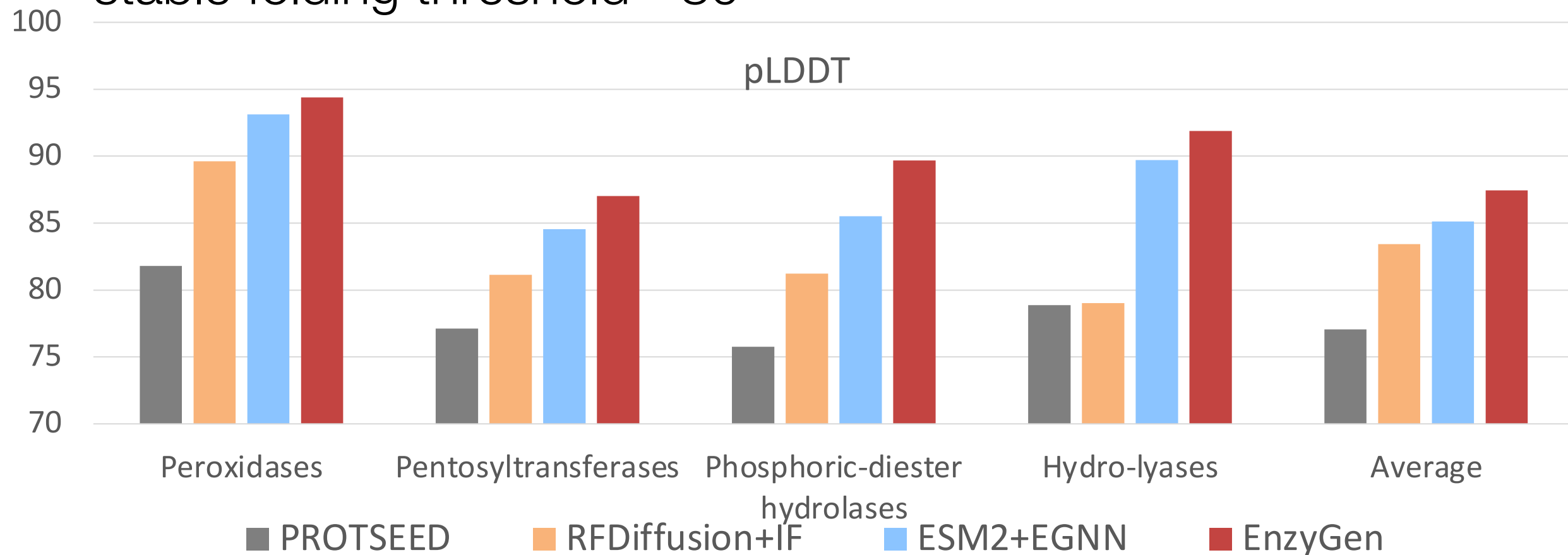
EnzyGen generates enzymes with higher function scores

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

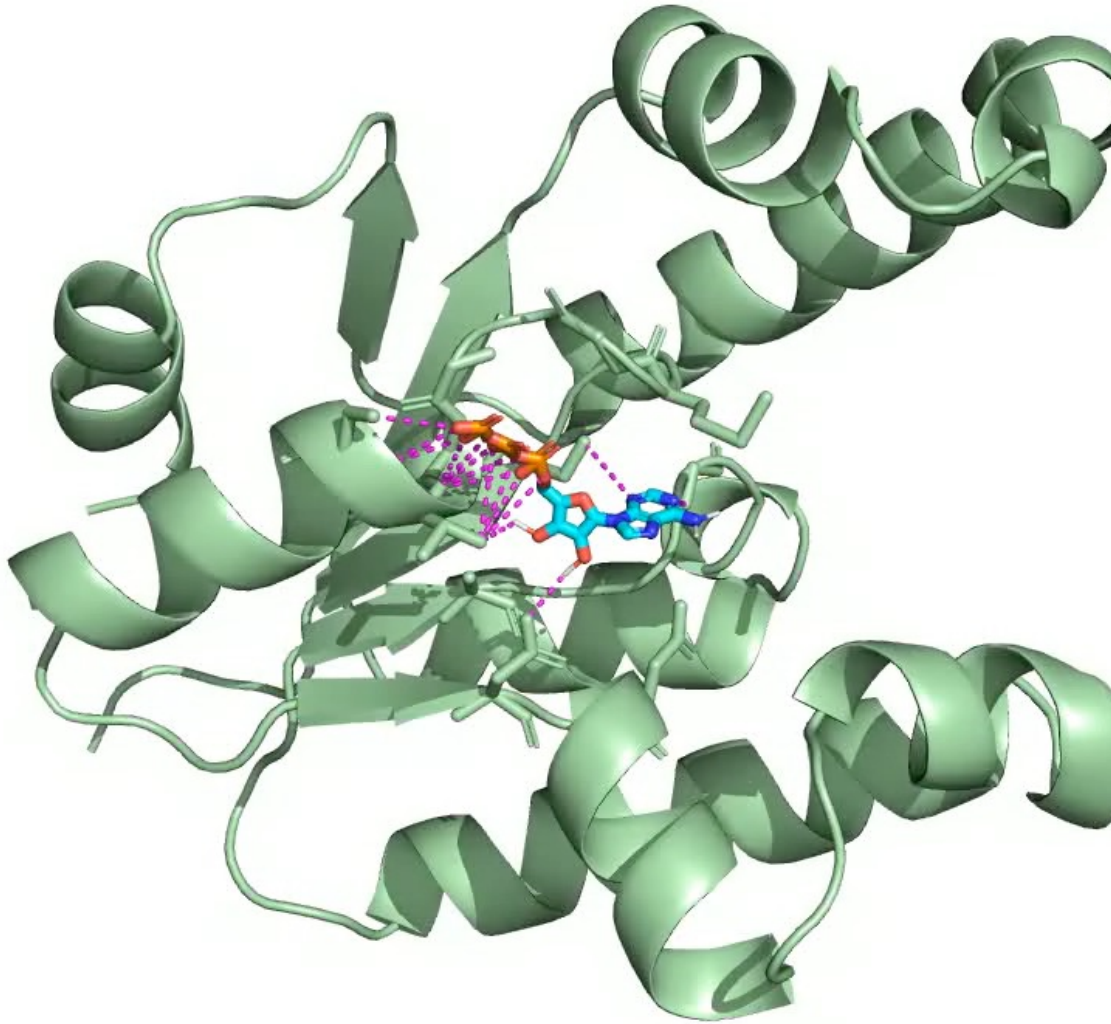


EnzyGen generates enzymes with more stable structures

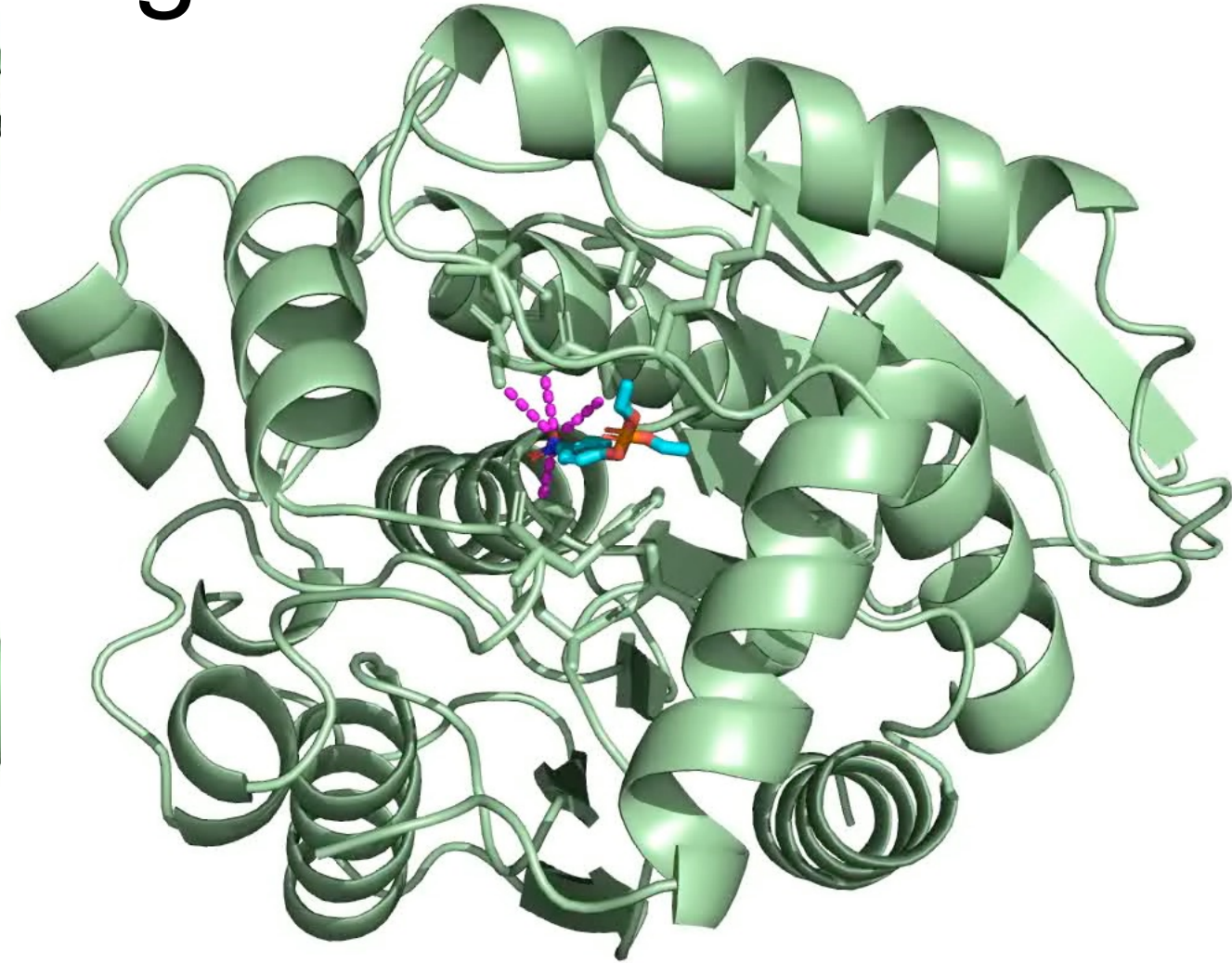
Average pLDDT across 30 categories is higher than suggested stable folding threshold - 80



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

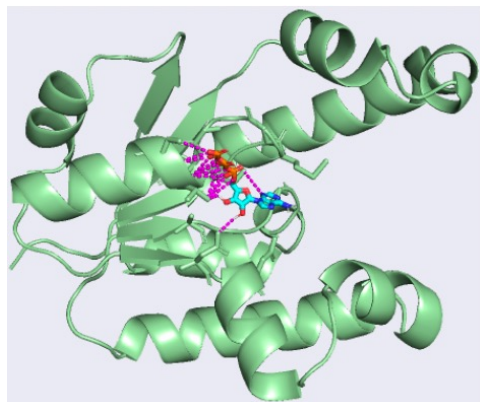
Commonality and Distinction in Generating Language and Molecules

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

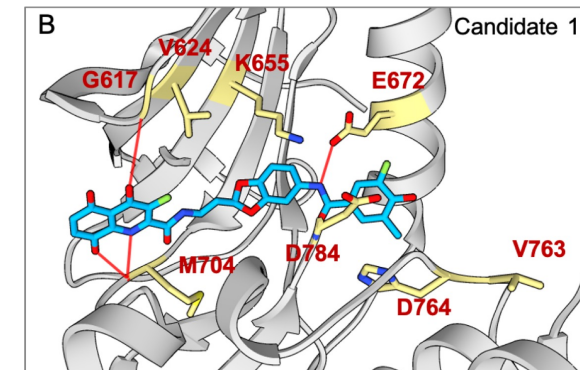
Takeaway of Molecular Design

- Problem formulation: Guiding information is important
 - fitness scores, chemical properties, tags, motifs
- Modeling 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

Molecule Design at CMU Li lab



<https://leililab.github.io/>



Protein

EnzyGen SurfPro

IsEMPro LSSAMP

Small Molecule

MARS MolEdit3D

RLHEX