

Accelerating Drug Design with Generative AI

Lei Li

March 31, 2026

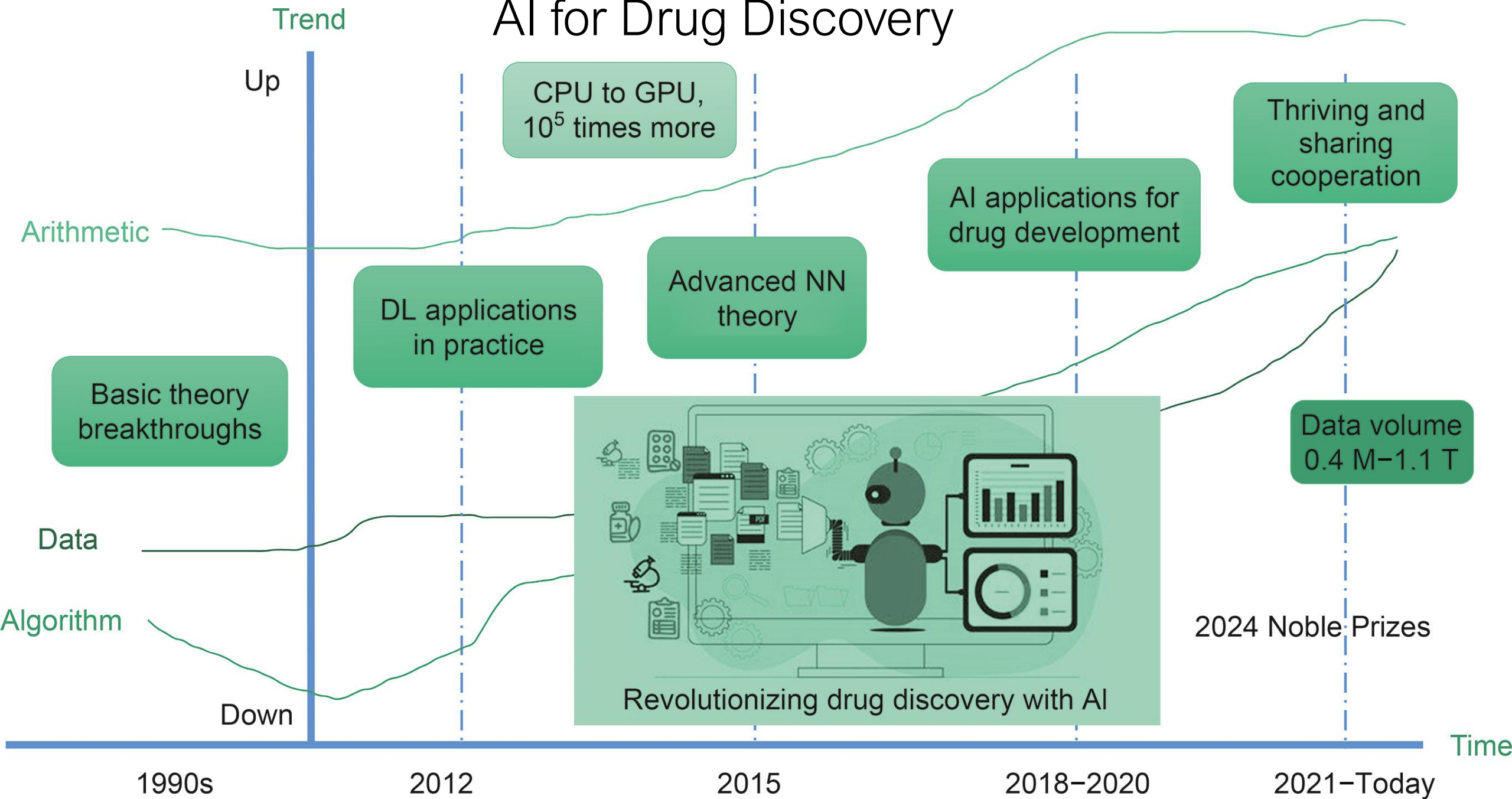


Language
Technologies
Institute

Carnegie Mellon University

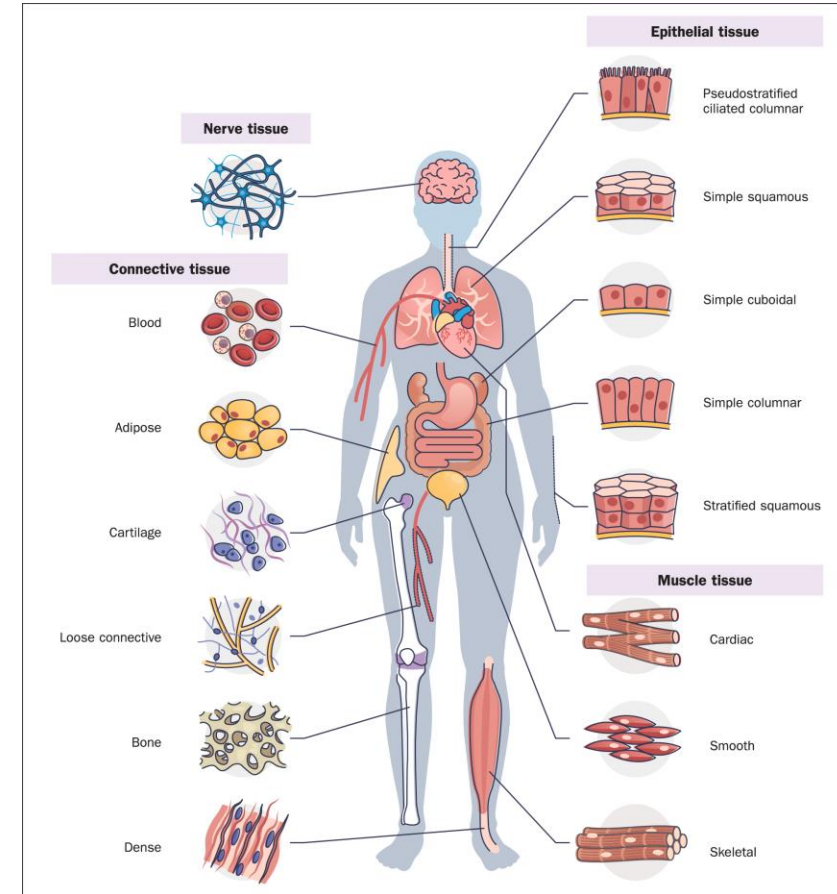
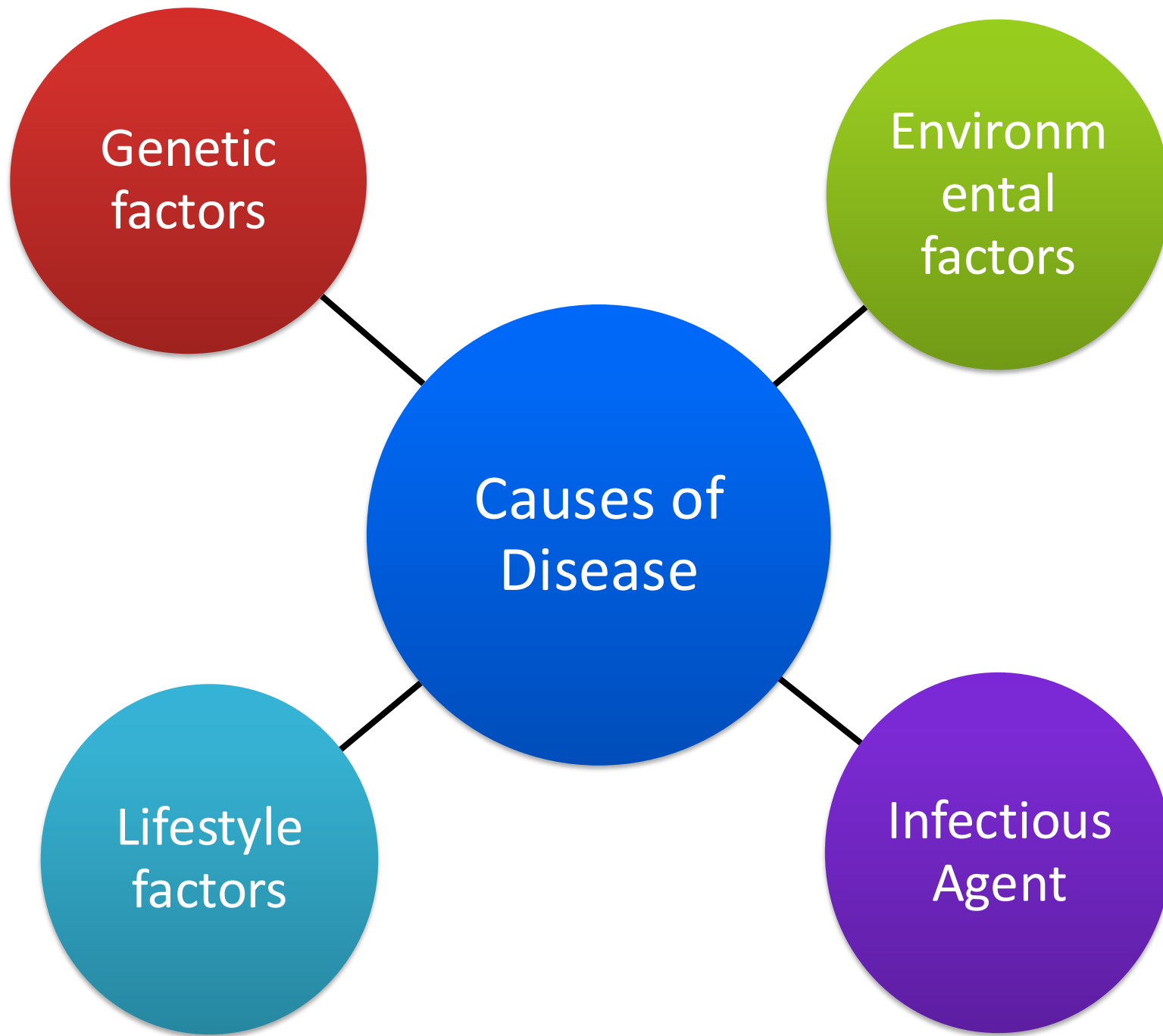
School of Computer Science

AI for Drug Discovery

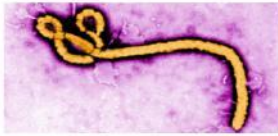


Outline

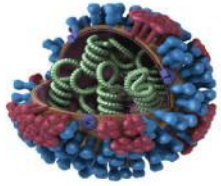
- Multi-scale Biological System and Drugs
 - molecule, cell, tissue, organs, cause of disease
- Basic AI Models for Biomecules
 - sequence, structure, generative model
- MARS: finding small molecule drugs with multiple properties
- EnzyGen: A general generative model for enzyme design
- PPDiff: protein-binding complex design



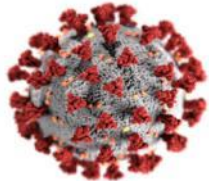
Ebola Virus



Influenza Virus



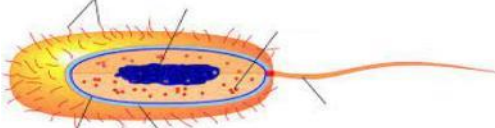
Corona Virus



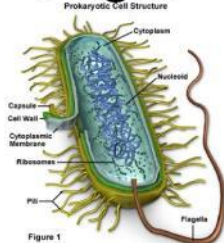
Zika Virus



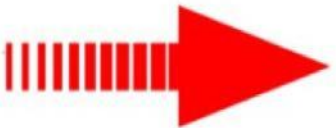
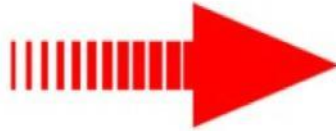
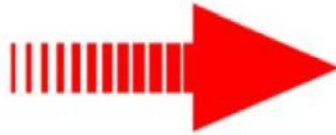
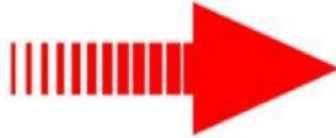
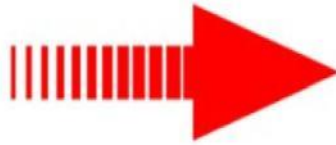
Cholera



Plague



Target Cells



Hepatocytes, Endothelial cells,
Phagocytes

Ciliated Epithelium of conducting
airways

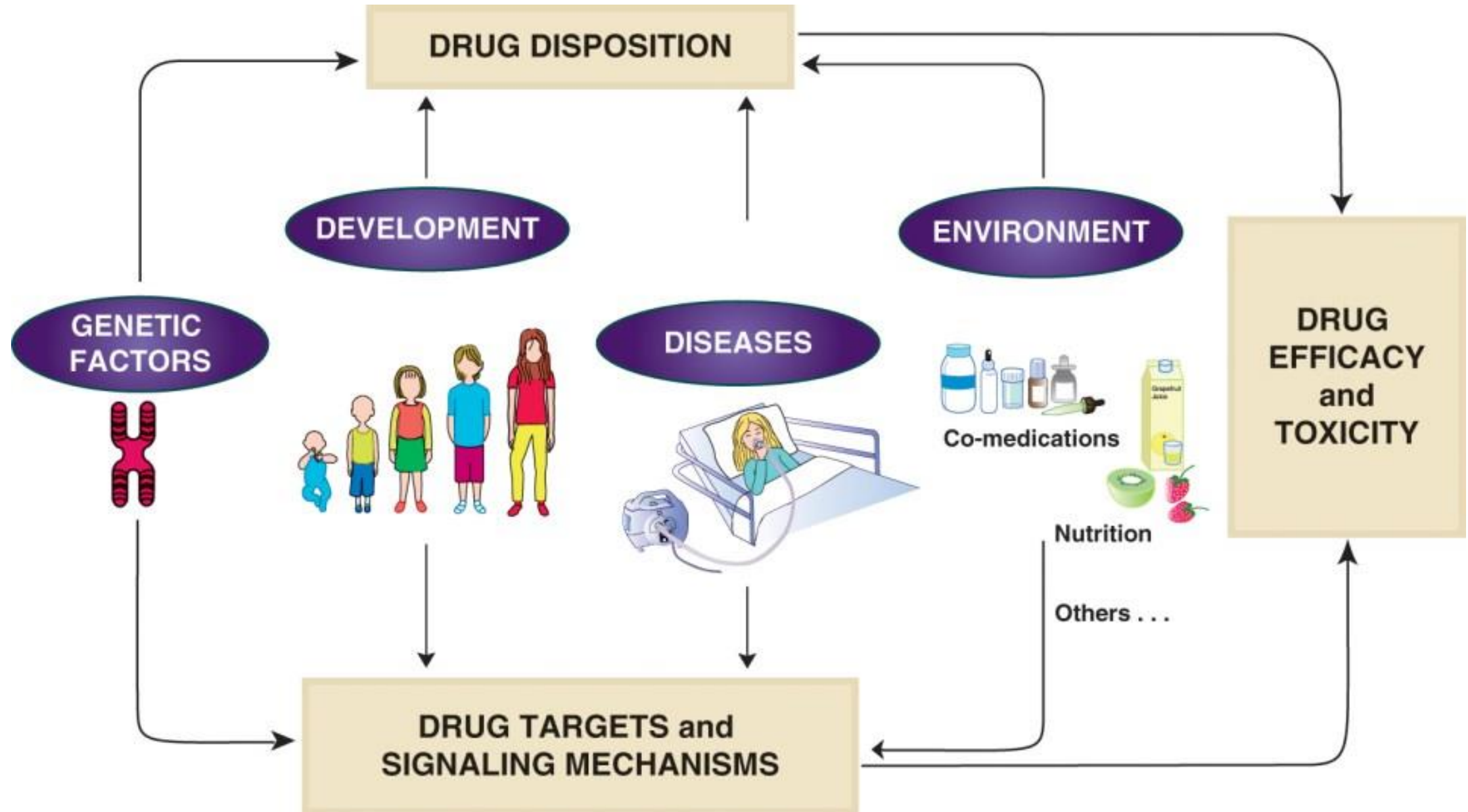
Lungs and respiratory epithelial cells

From Brain, Placenta, skin to Testis,
Kidney, Retina, During Pregnancy
causes microcephaly

Intestinal Epithelial Cells

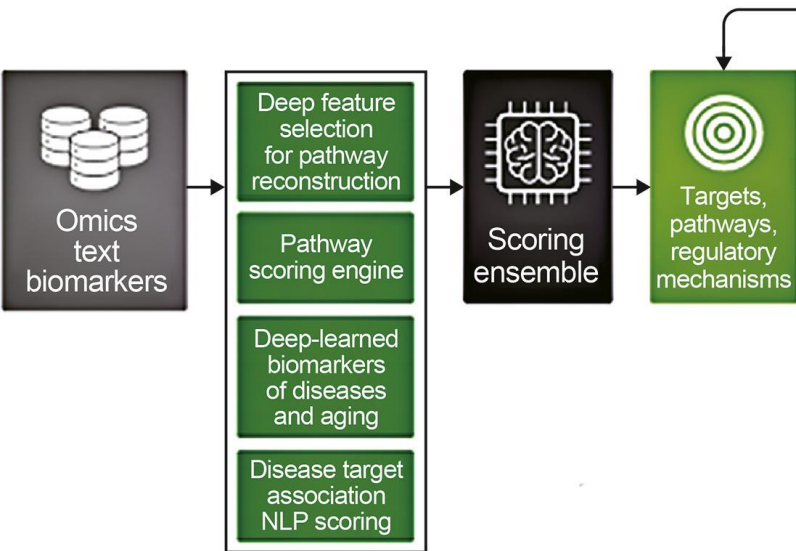
Neutrophils

Finding Effective Drugs

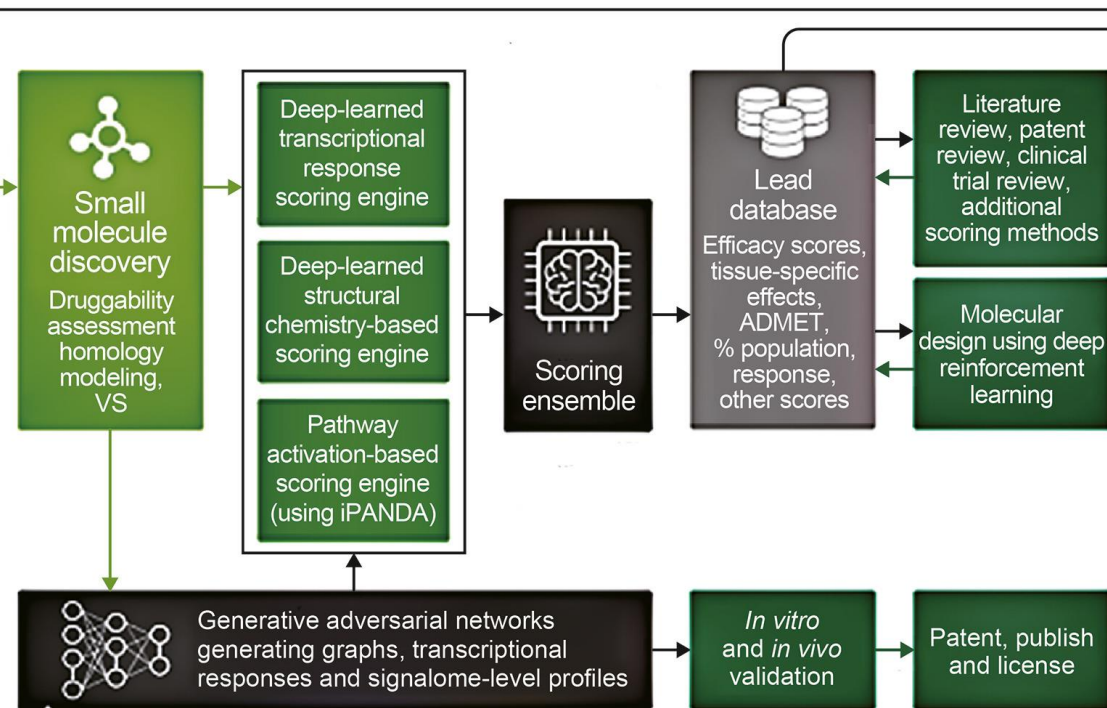


AI for Drug Discovery

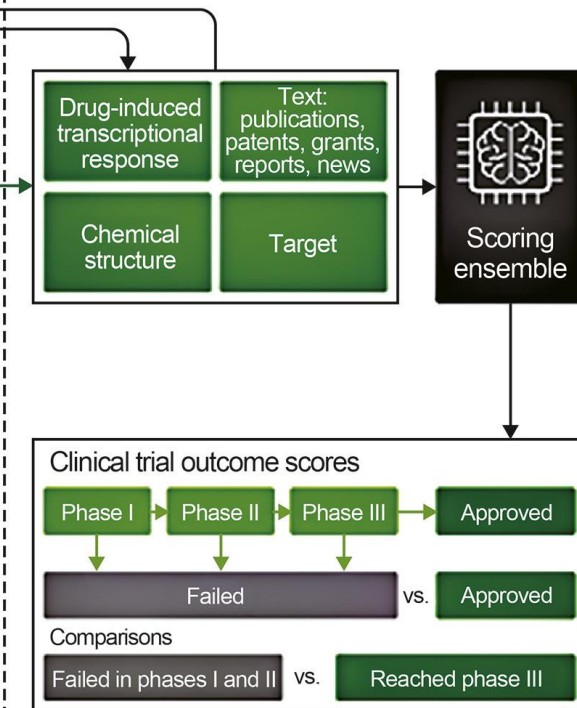
Target Identification



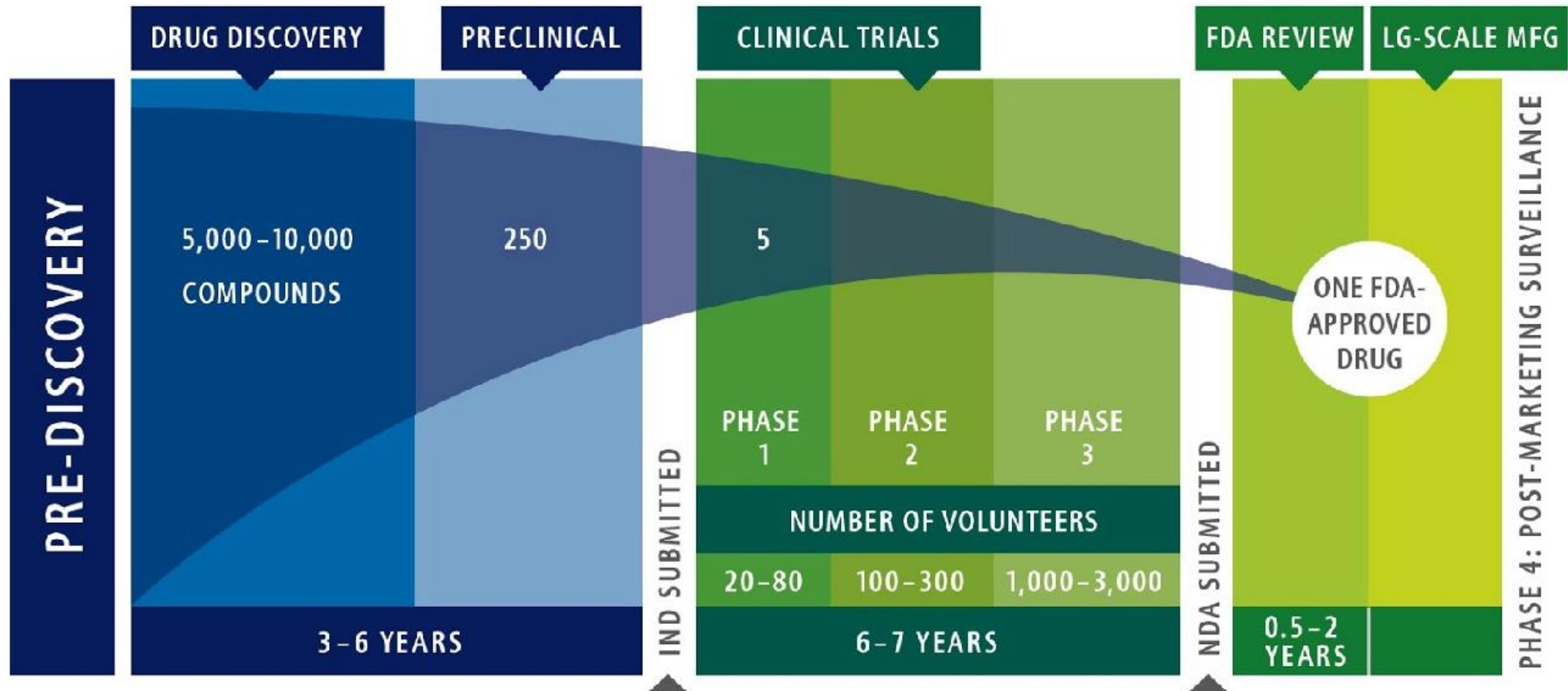
Drug Molecule Design small molecule, peptide, protein



Clinical Trial Prediction



Drug Discovery and Development

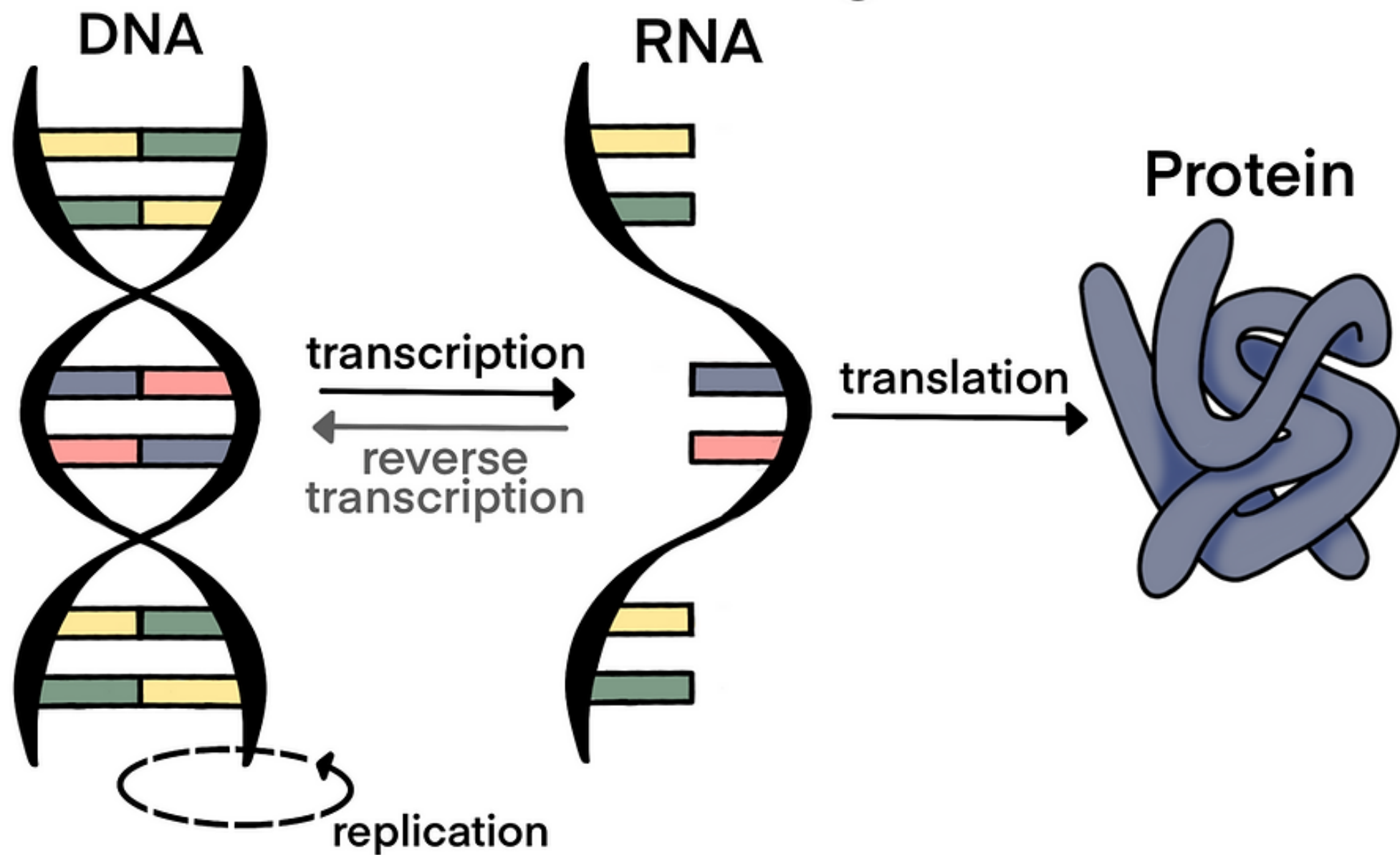


- Drug discovery and development is a long and risky process.
- The successful launch of a new drug cost \$1.3 billion on average and more than 10 years.

Outline

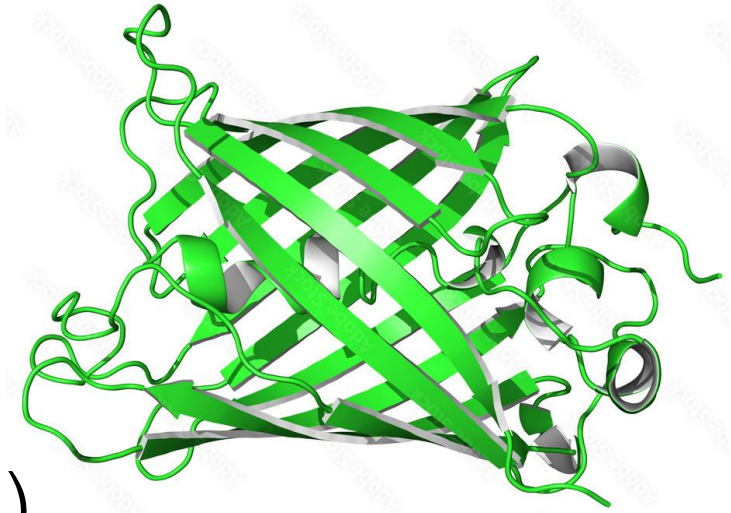
- Multi-scale Biological System and Drugs
 - molecule, cell, tissue, organs, cause of disease
-  • Basic AI Models for Biomecules
 - sequence, structure, generative model
- MARS: finding small molecule drugs with multiple properties
- EnzyGen: A general generative model for enzyme design
- PPDiff: protein-binding complex design

central dogma

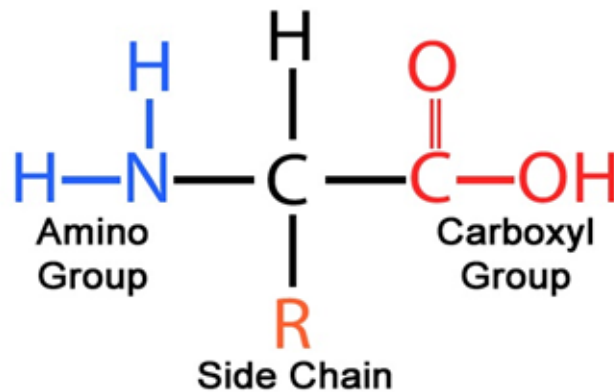


Protein

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



Amino Acids

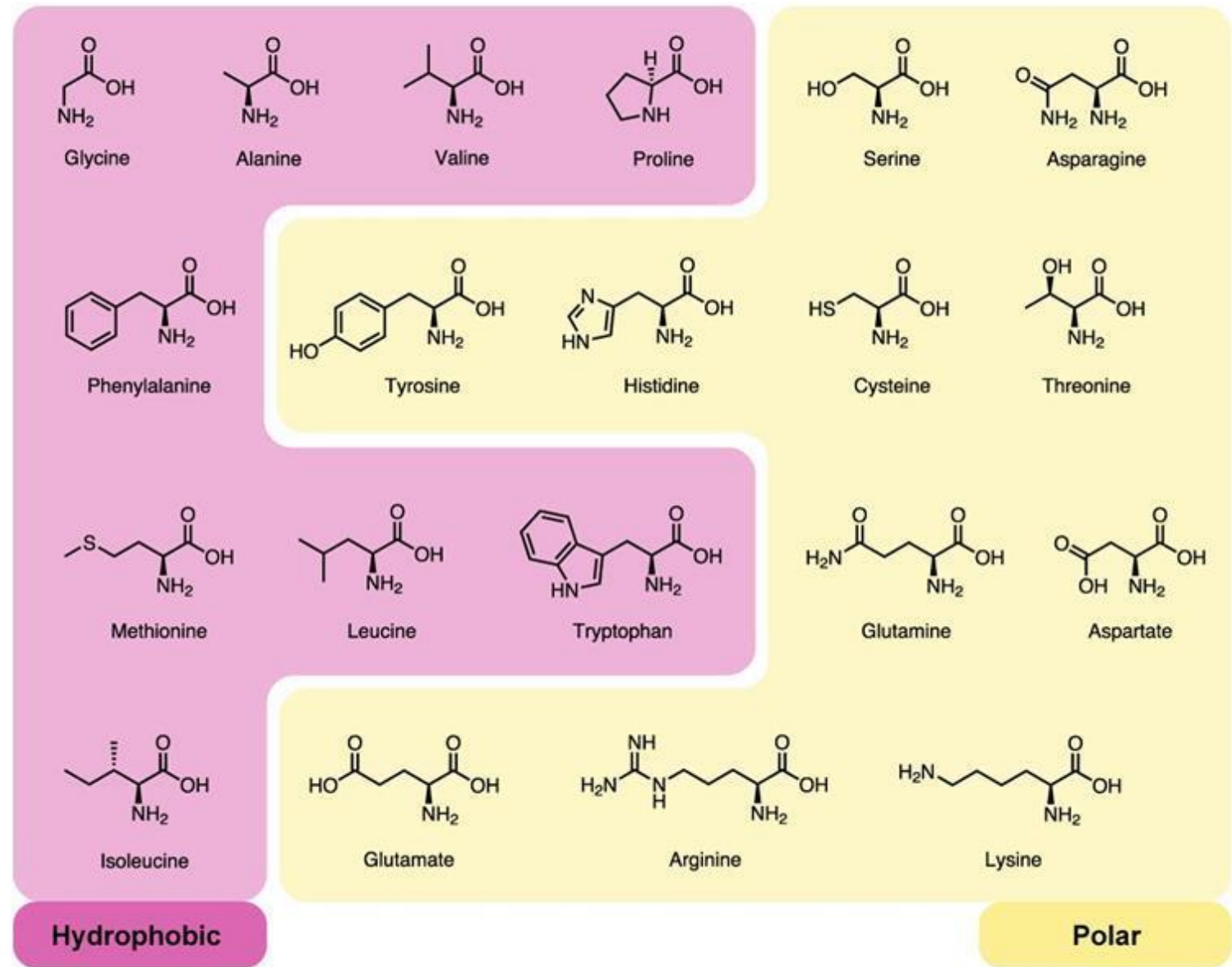


VLLPDNHYLSTQSAISKDPNEKRD
HMLLEFVTAAGIT

Amino Acids

Amino acids could be hydrophobic or hydrophilic.

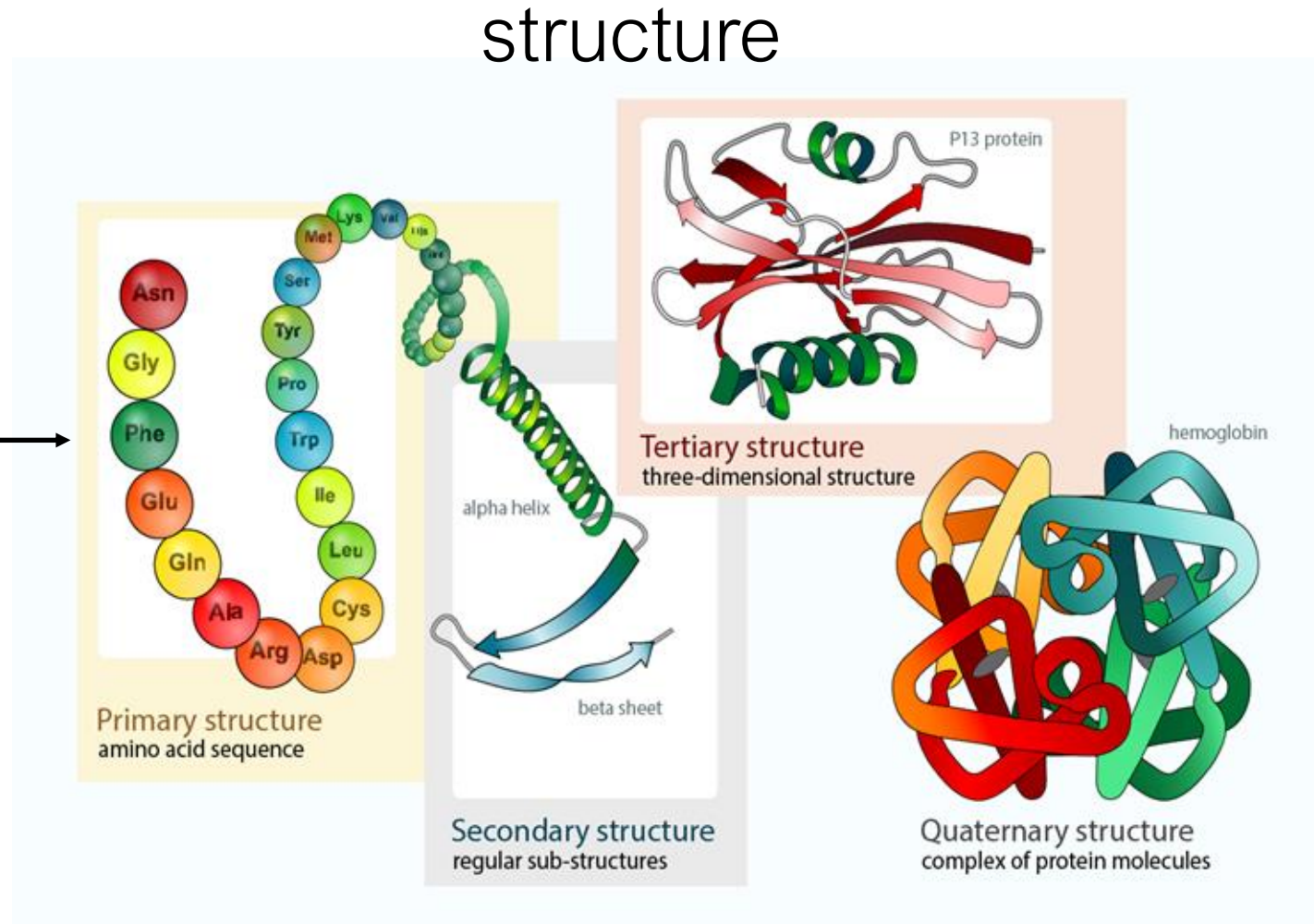
Amino acids are basic units of proteins.



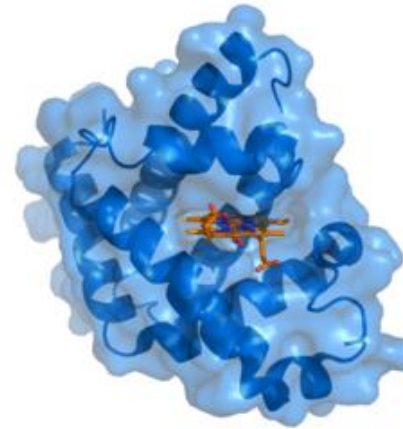
Protein: sequence to structure to function

sequence

VLLPDNHYLS
TQSALSKD



function



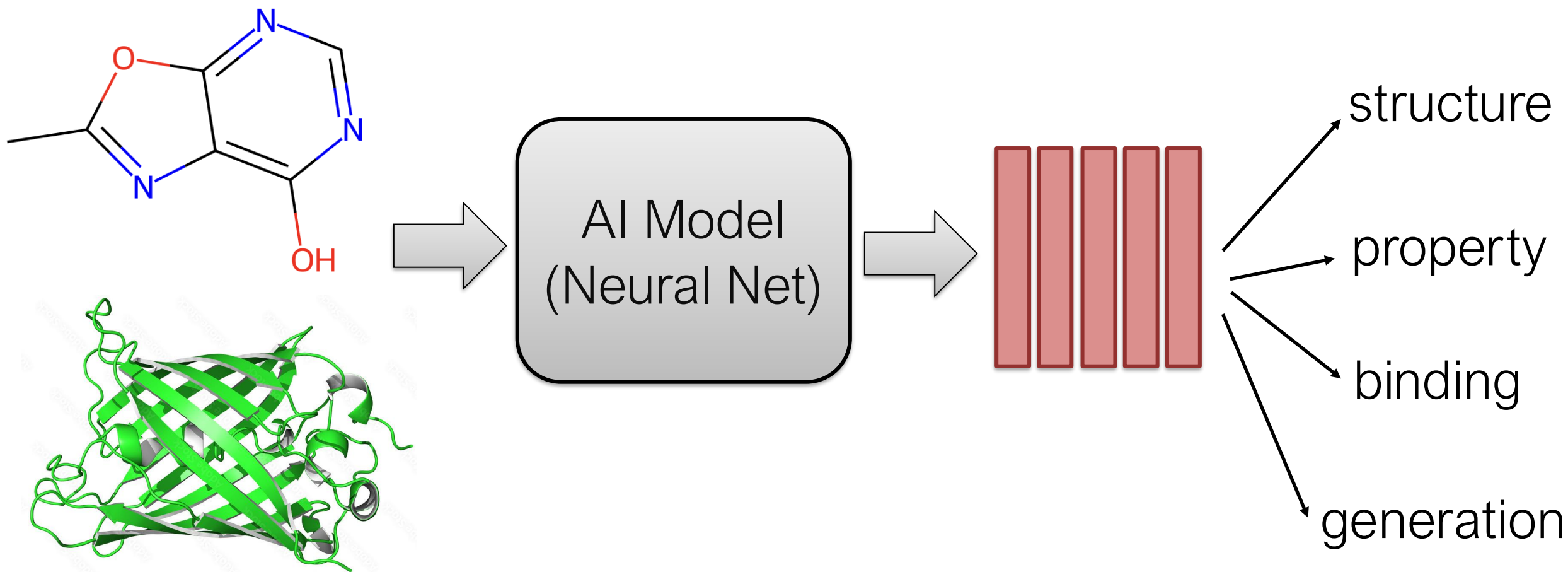
Drug Discovery as AI Tasks

- Target Identification:
 - which gene/protein is critical on the pathway of a disease context
 - Feature/Network-based mining/ranking/prediction task
- Molecule generation
 - search or generate highly possible drug molecule leads
- Structure/Property prediction
 - predicting structures, protein folding
 - binding to target, toxic, absorbable, synthesizable, side effect
- Clinical trial outcome prediction
 - consider cell and patient environment

Basic AI Modelling Tools

- Molecule Sequence
 - Language model: Transformer
- Molecule structure
 - Graph neural networks
 - Equivariant GNN for 3D structures
- Modeling interactions
 - molecule docking: how ligand binds to protein
 - molecular dynamics simulation

Goal of Molecule Modelling



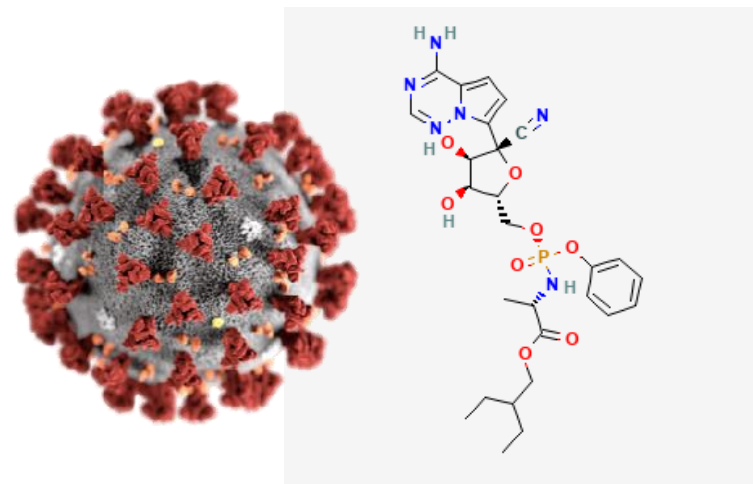
Molecule as Discrete Sequences of Tokens

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)(OCC1  
C(C(C(O1)(C#N)C2=CC=C3N2N=CN  
=C3N)O)O)OC4=CC=CC=C4
```



Language Model

- A probabilistic model of discrete sequences, including human languages and biological languages

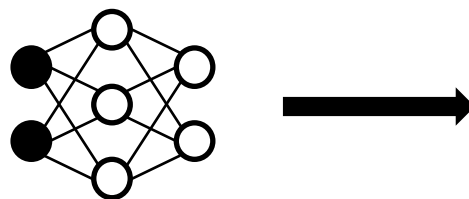
Probability("Pittsburgh is a city of bridges")

Probability("VLLPDNHYLSTQSALSKDPN")

Probability Model for Next Token

$P(\text{next word } y_t \mid \text{Prompt } x, \text{previous words } y_{1:t-1})$

Santa Barbara has very nice _____



Pittsburgh is a city of _____

beach	0.5
weather	0.4
snow	0.01
bridges	0.6
corn	0.02

Mathematics of Language Model

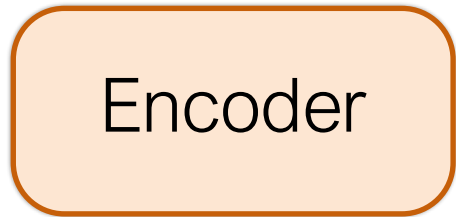
Probability("Pittsburgh is a city of bridges")
= $P(\text{"Pittsburgh"}) \cdot P(\text{"is"} | \text{"Pittsburgh"})$
· $P(\text{"a"} | \text{"Pittsburgh is"}) \cdot P(\text{"city"} | \dots) \cdot P(\text{"of"} | \dots)$
· $P(\text{"bridges"} | \dots)$

$$\text{Prob.}(x_{1..T}) = \prod_{t=1}^T \underbrace{P(x_{t+1} | x_{1..t})}_{\text{Predicting using Neural Nets}}$$

Predicting using Neural Nets
(Transformer network, CNN, RNN)

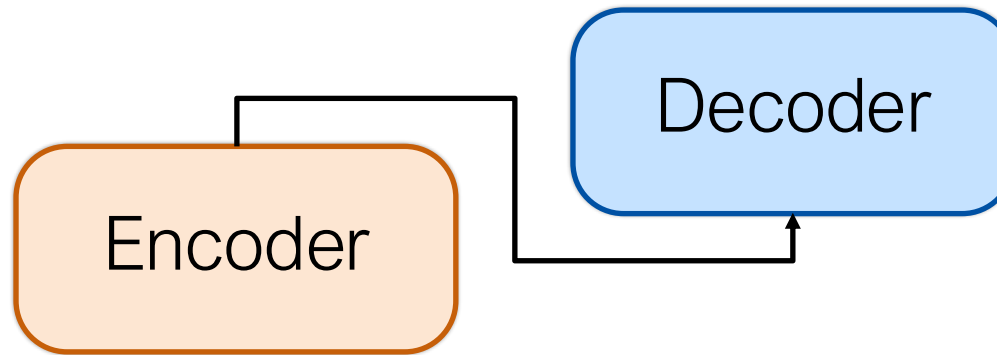
Type of Language Models

Encoder-only
Masked LM
Non-autoregressive



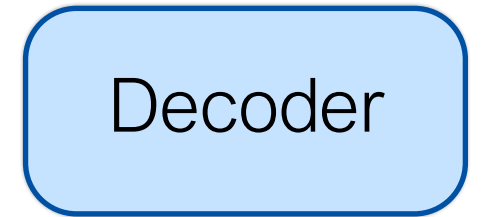
e.g. BERT
RoBERTa
ESM (for protein)

Encoder-decoder



e.g. T5

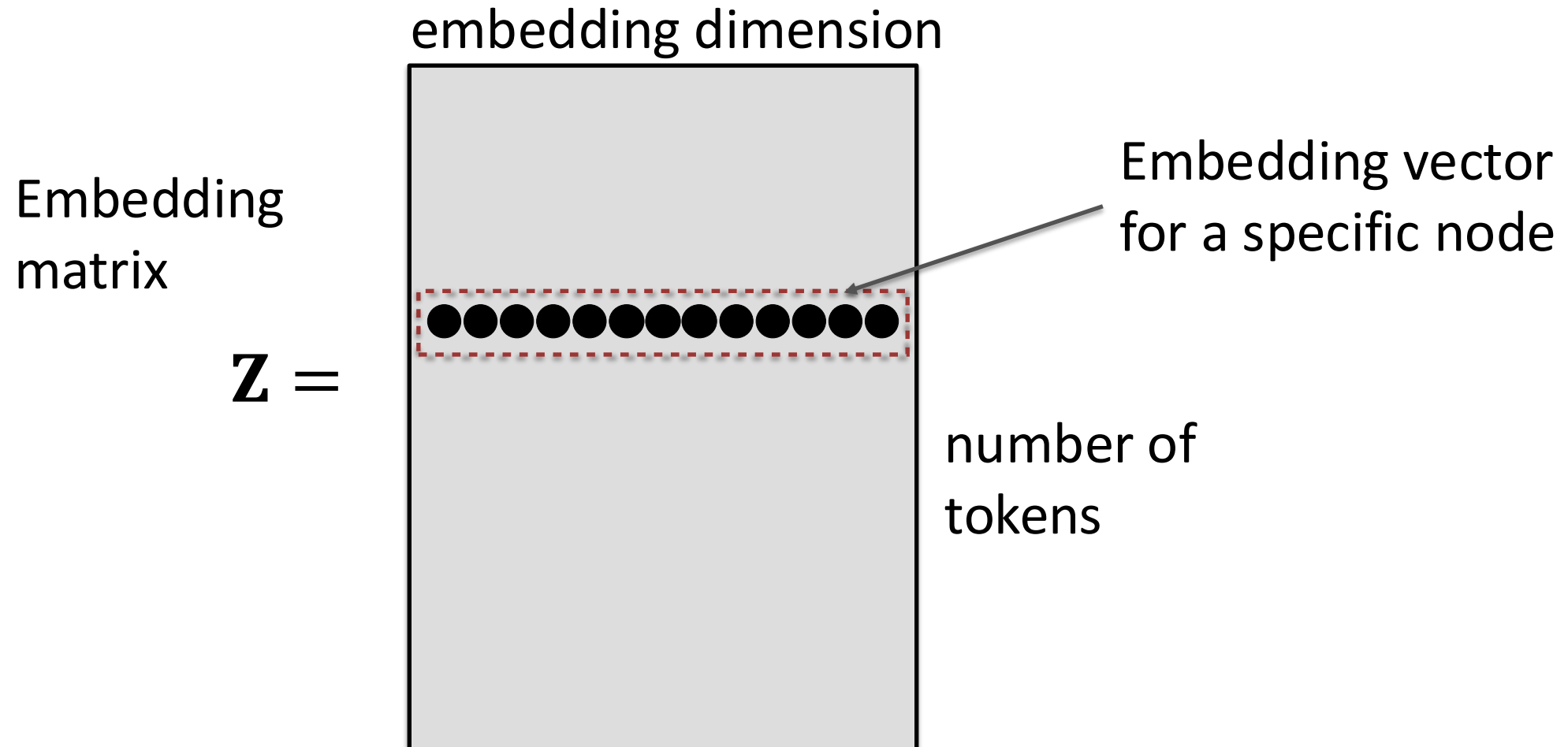
Decoder-only
Autoregressive



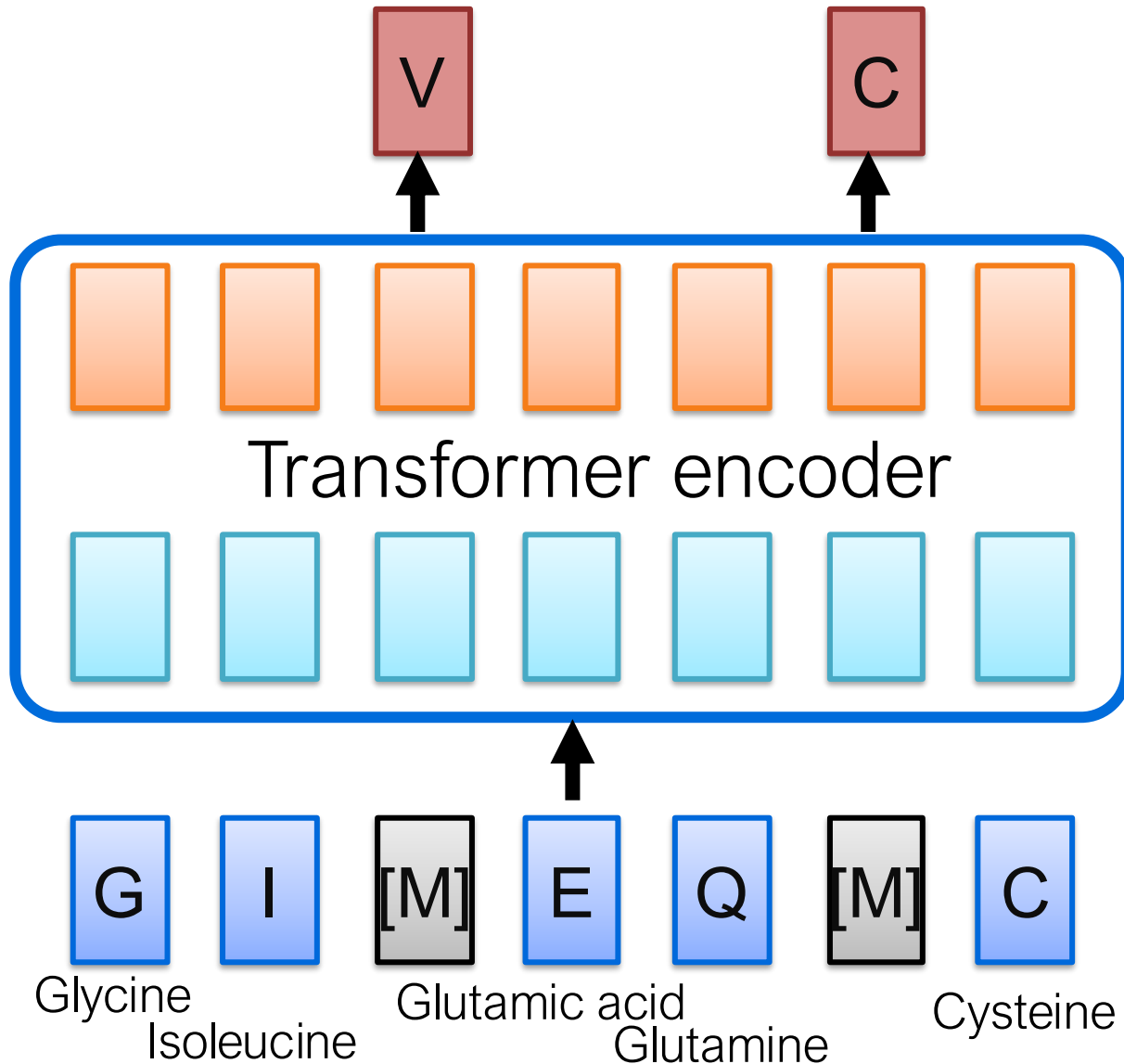
e.g. GPT
LLaMA
ProGen (for protein)

Mapping Token to Embedding

- Token: a basic unit, atom, amino acid, nucleotide

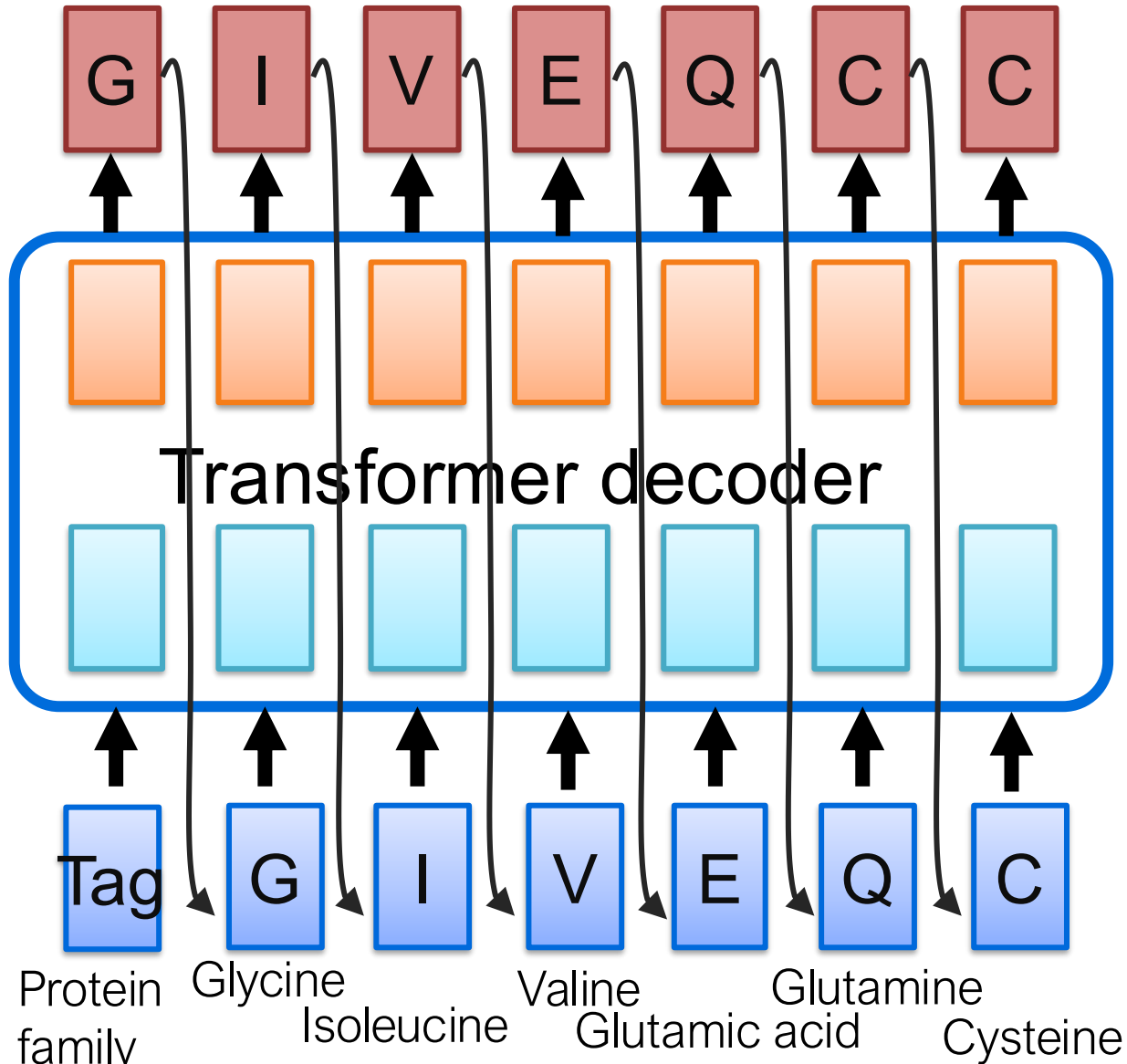


Protein Language Model 1: Mask LM



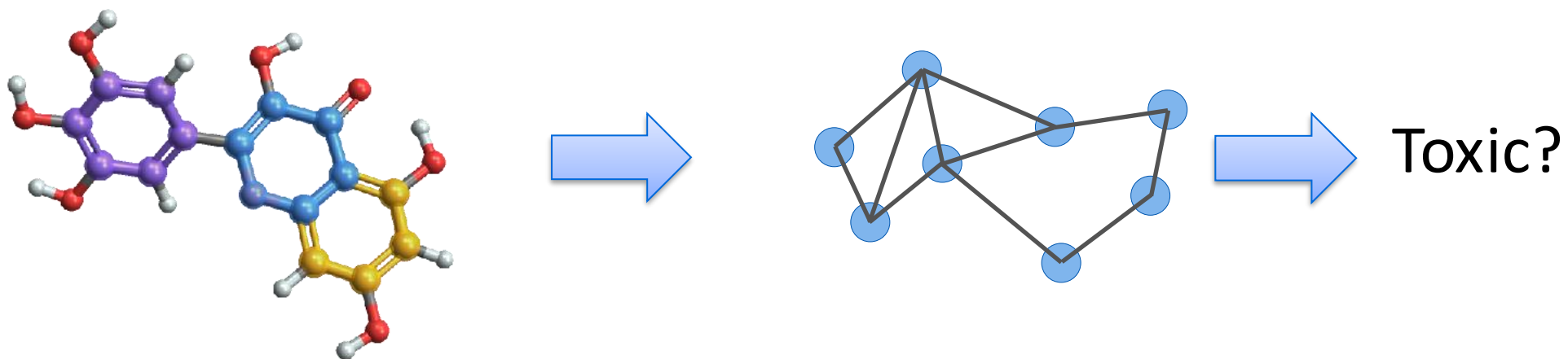
- Using raw protein sequences for pre-training
 - Training loss: predicting masked residues
- ESM [Meier et al 2021] and ESM-2 [Lin et al 2023]

Protein Language Model 2: Casual LM

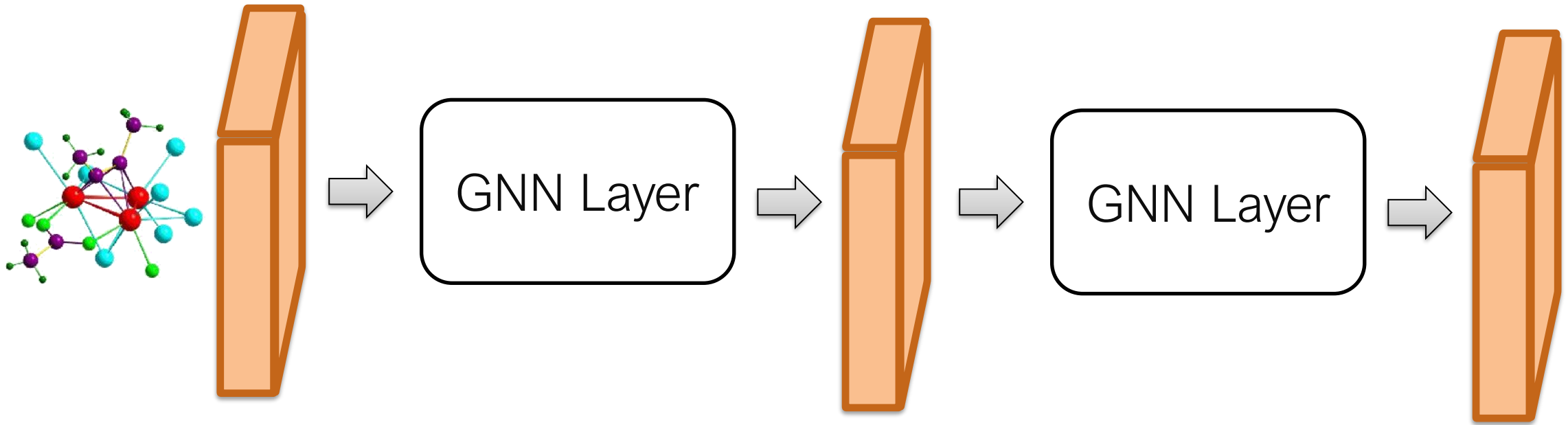


- Using raw protein sequences and their category tags for pre-training
 - training loss: predicting next residue
- ProGen [Madani et al 2023] and ProGen2 [Nijkamp et al 2023]
- Protein Tag is insufficient!

Molecule as a Graph



Graph Neural Network



Output is an embedding matrix for nodes
for further downstream tasks: e.g. node property prediction

Graph Neural Network (MPNN) to model molecule graphs

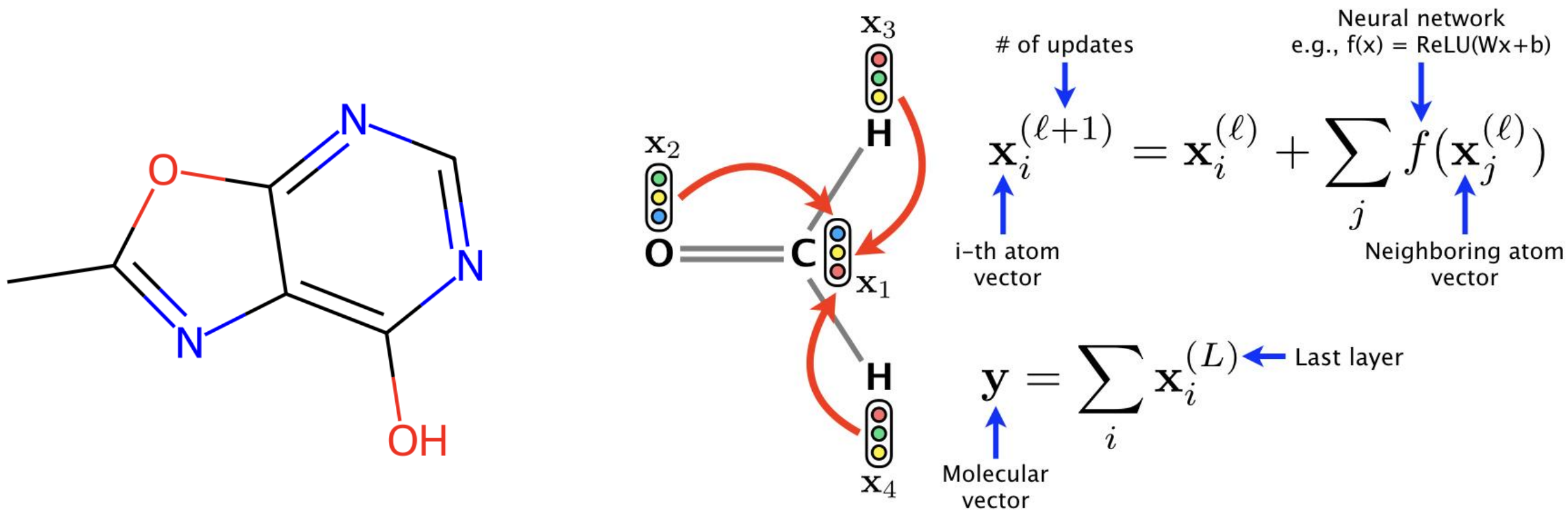


Fig.1: The update function (or called transition, propagation, message passing, and convolution) in GNNs. On a molecular graph, the **GNN updates each atom vector with its neighboring atom vectors non-linear transformed by neural network**. The molecular vector is obtained by summing (or mean) the atom vectors.
[Tsubaki et al, 2018.]

Graph of Fragments

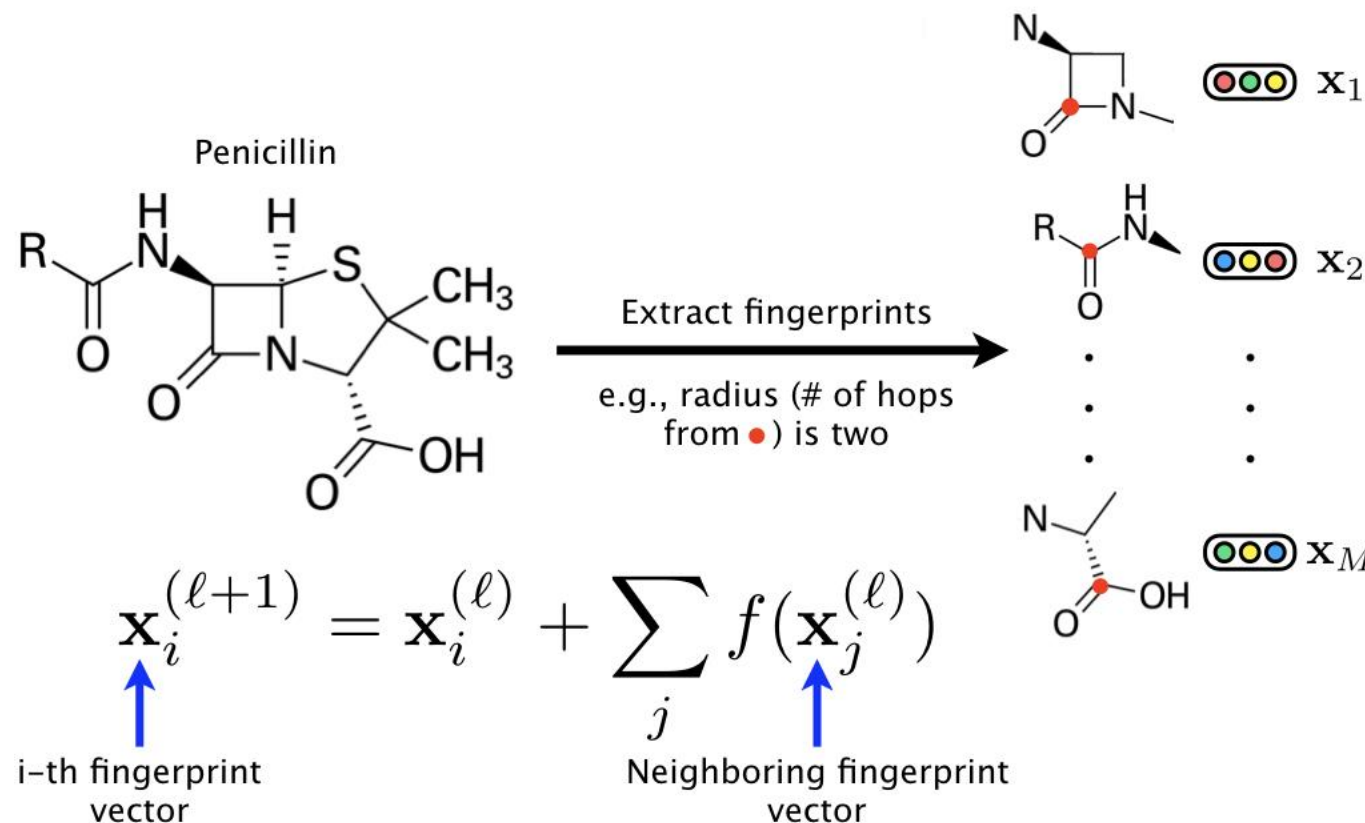
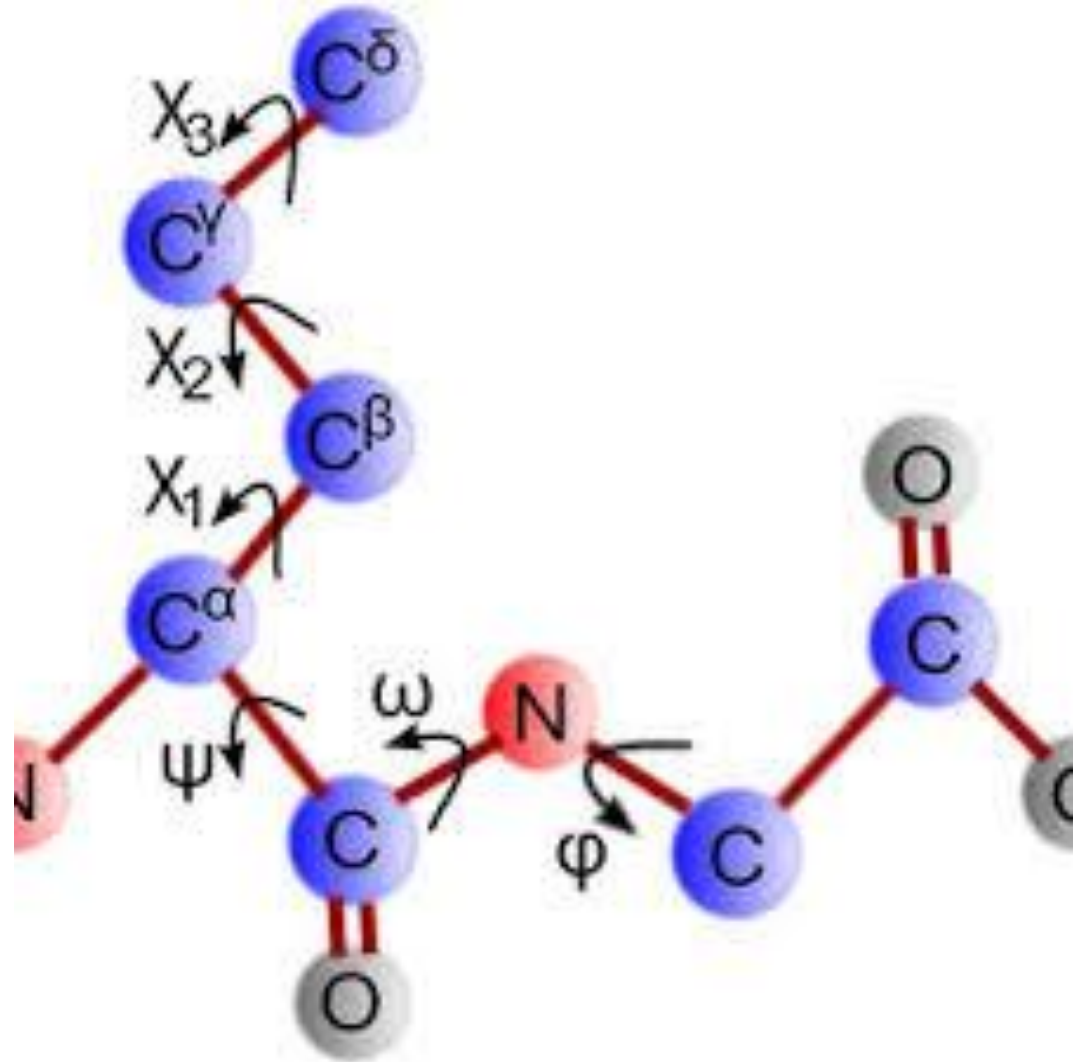
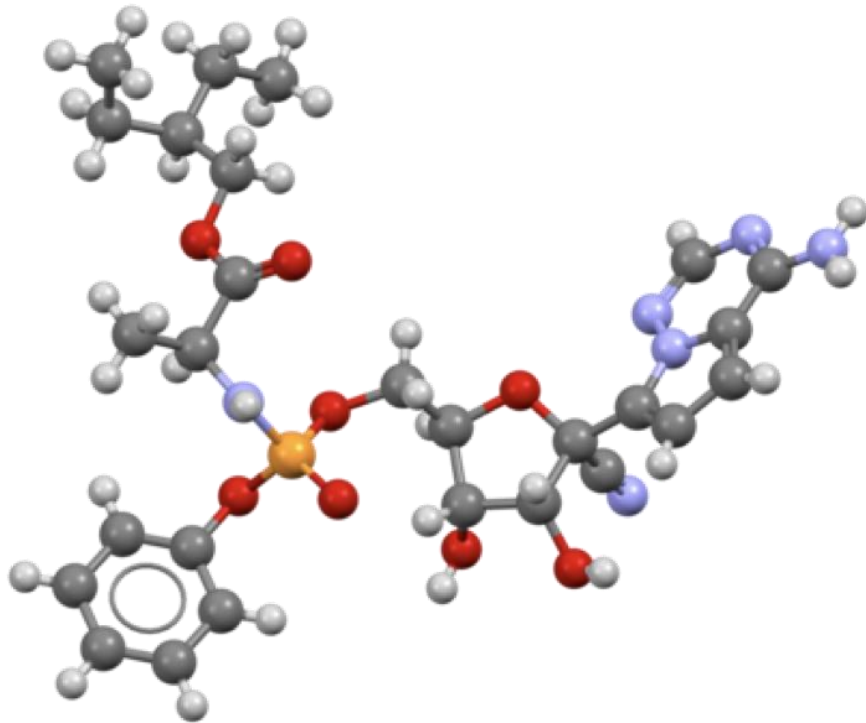


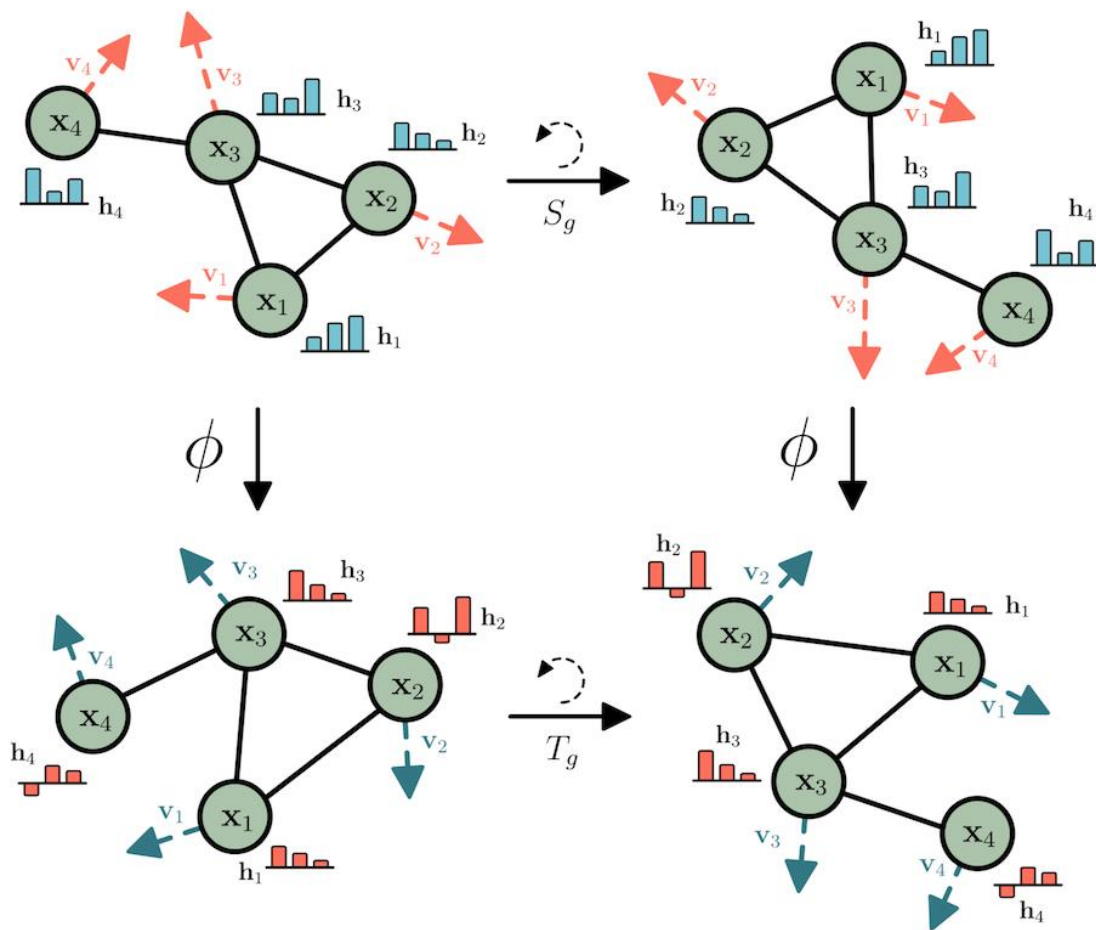
Fig.2: The update function based on radius-based subgraphs, i.e., molecular fingerprints. **Each fingerprint is initialized with a random vector.** The following procedure is the same as that of basic GNN.

Modelling 3D Structure of Molecules

- Matrix of 3D coordinates
- Matrix of angles



Equivariance in Embedding Transformation



f is Equivariant (SE3 invariant):

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

No matter how we rotate and move the molecule, the NN will compute the same embeddings
X are coordinates

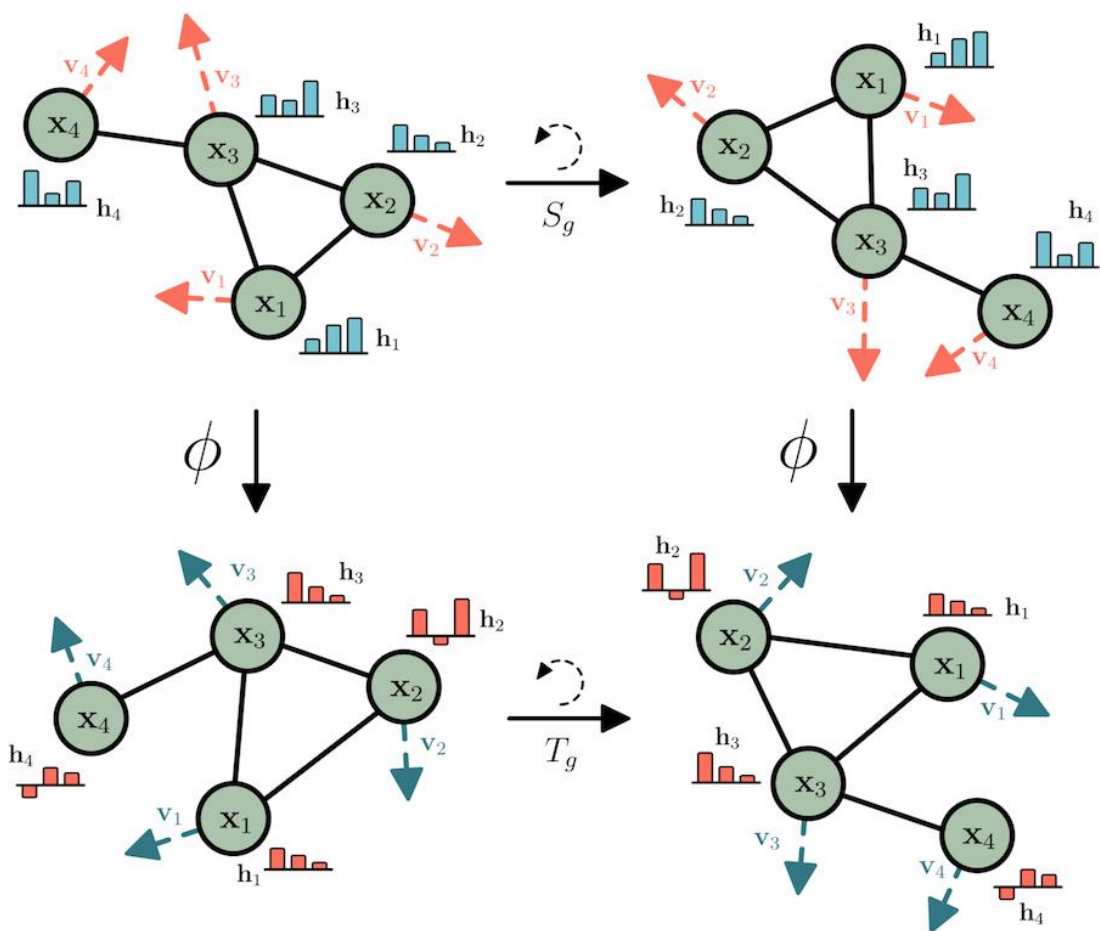
H are features

g is the rotation/translation

$$F(H, g(X)) = F(H, X)$$

Equivariant Graph Neural Network (EGNN)

- model the 3D geometry of a molecule



$$d_{ij} = \|x_i - x_j\|_2$$

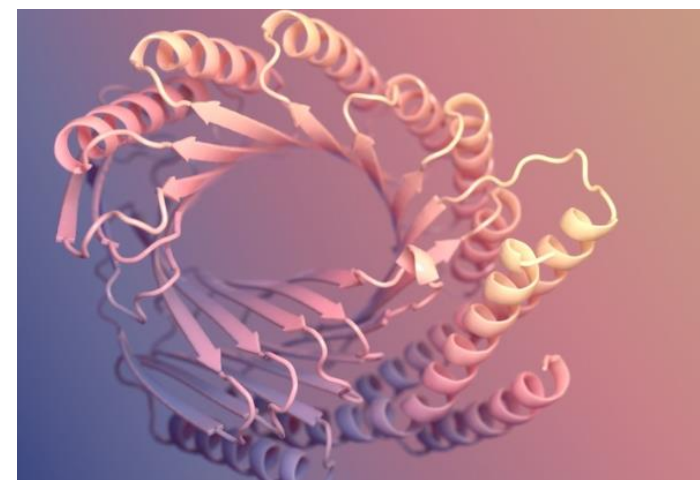
$$m_{j \rightarrow i} = M^k(h_i^k, h_j^k, d_{ij})$$

$$m_i^{k+1} = \sum_{v_j \in N(v_i)} m_{j \rightarrow i}$$

$$h_i^{k+1} = U^k(h_i^k, m_i^{k+1})$$

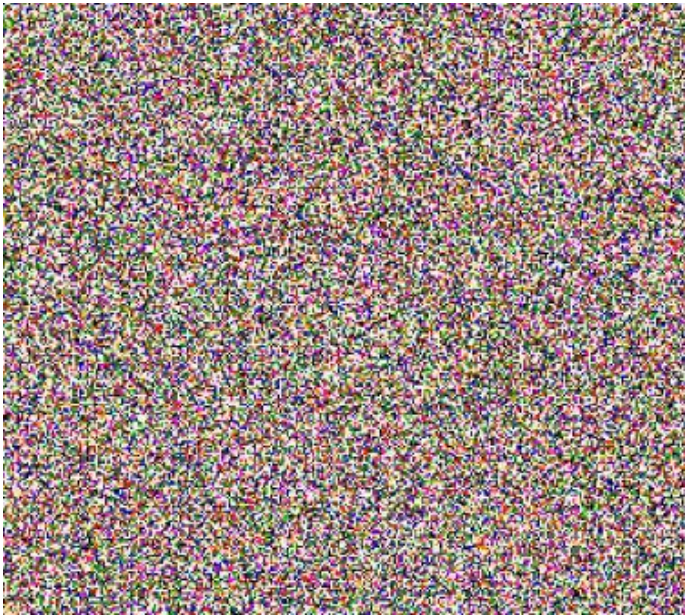
Diffusion Models are state-of-the-art models for (Continuous) Data Generation

- 3D Objects (coordinates)
 - Molecule structures
- Discrete sequences

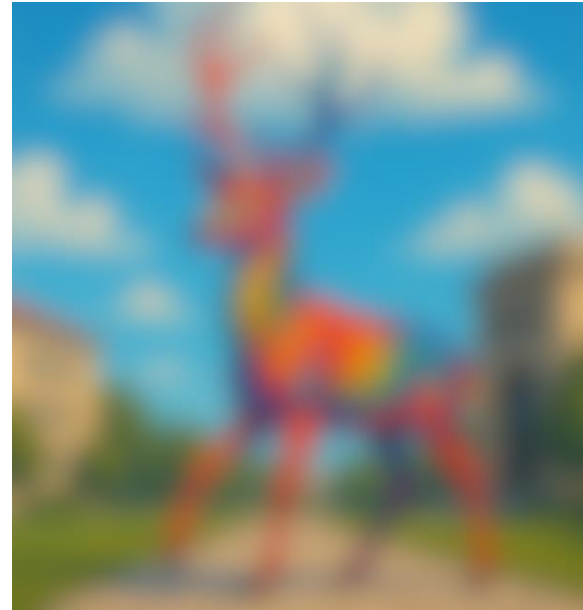


Probabilistic Generative Process

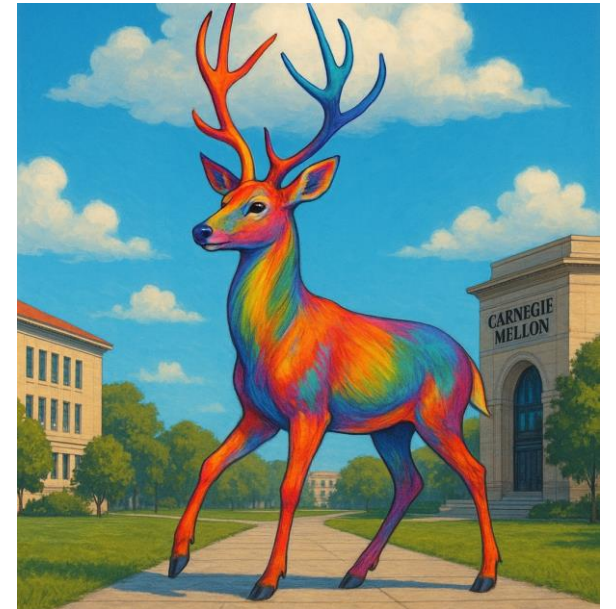
Start from initial
distribution $p_T(x_T)$
e.g. Gaussian $N(0, I)$



Generate slightly
improved data
 $x_{t-1} \sim p_{t-1}(x_{t-1} | x_t)$



final data
 $x_0 \sim p_0(x_0 | x_1)$



The generation process is a Markov chain

Learning the generative model
= learning the parameters for
each $p_{t-1}(x_{t-1}|x_t)$

It is difficult to directly construct a series of probability distributions

→ Diffusion Models

(Forward) Noising Process

- adding Gaussian noise at each time step t $x_t \sim q(x_t | x_{t-1}) = N(\sqrt{\alpha_t}x_{t-1}, \beta_t I)$, $\alpha_t = 1 - \beta_t$

equivalently $x_t = \sqrt{\alpha_t}x_{t-1} + \sqrt{\beta_t}\epsilon$, $\epsilon \sim N(0, I)$

$t=0$

$t=1$

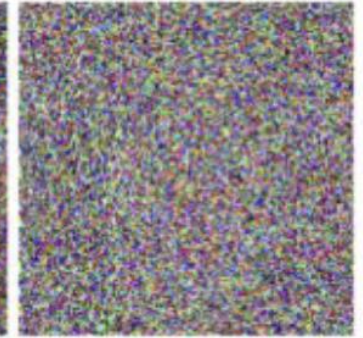
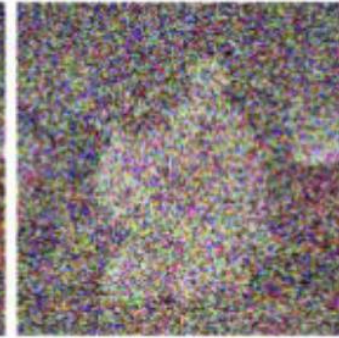
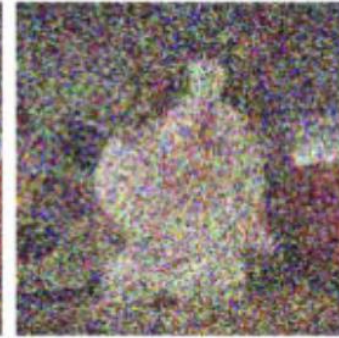


(Reverse) Generation/Denoising Process

adding noise $q(x_t|x_{t-1})$

t=0

t=1



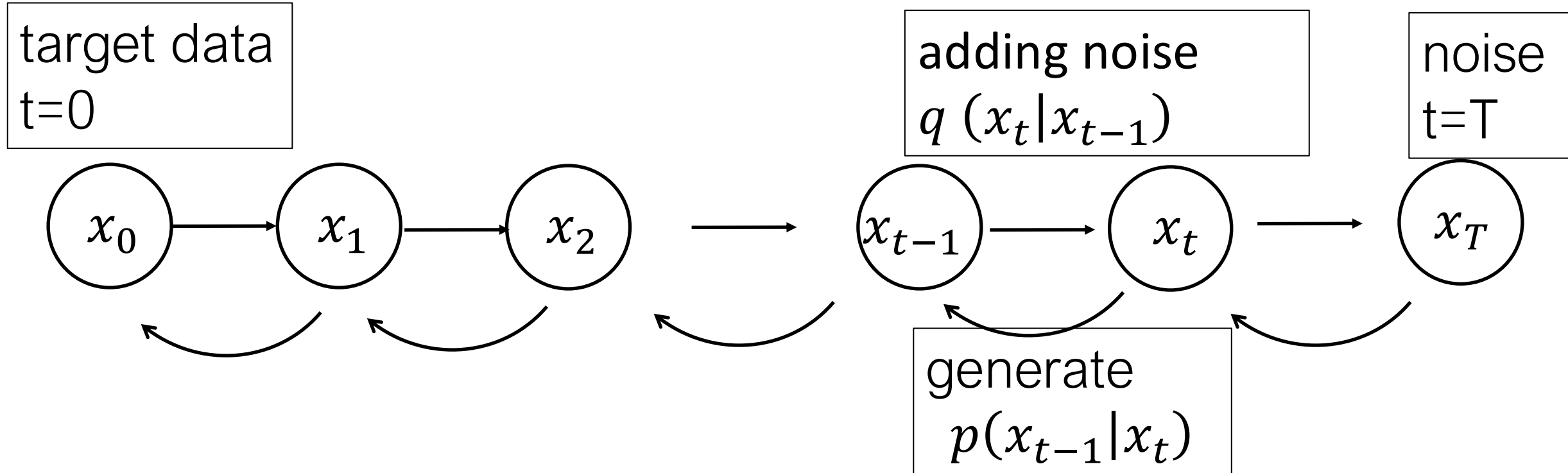
t=T

generate

$p(x_{t-1}|x_t)$

Learning problem: how to find parameters of p to match the sequence of noised images?

Diffusion Model



Learning problem: how to find parameters of p to match the sequence of noised images?

Outline

- Multi-scale Biological System and Drugs
 - molecule, cell, tissue, organs, cause of disease
- Basic AI Models for Biomecules
 - sequence, structure, generative model
- ➔ • MARS: finding small molecule drugs with multiple properties
- EnzyGen: A general generative model for enzyme design
- PPDiff: protein-binding complex design

Key Concepts for Drug Design

- **Target:** A large, complex protein that performs various functions in the body.
- **Ligand (Small Molecule):** A small chemical compound that binds to a protein to produce a biological effect. The "drug" we want to design.
- **Binding Pocket:** The specific region on a protein's surface where a ligand binds.
- **Conformation:** The specific 3D arrangement of atoms in a molecule. Proteins can be flexible and change their shape

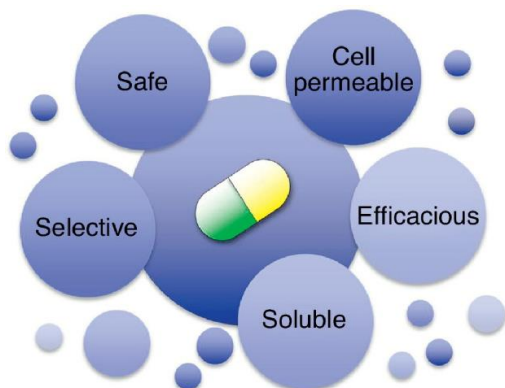
Key Terminology for Drug Design

- **Side Chains:** The flexible parts of a protein's building blocks (amino acids) that line the binding pocket.
- **De Novo Drug Design:** The process of creating novel molecular structures from scratch, rather than modifying existing ones, to achieve desired therapeutic properties.

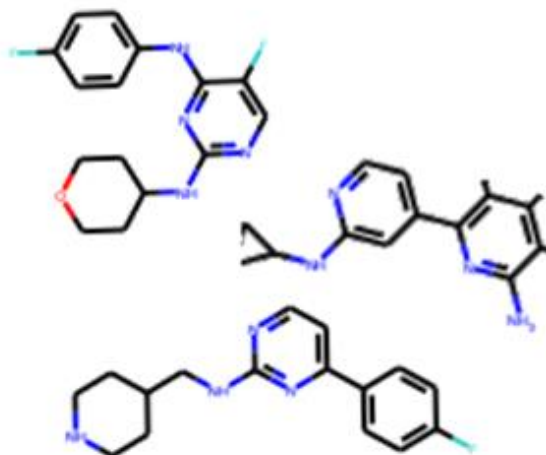


Designing Small-molecule Drugs needs to Meet Multiple Objectives

Satisfy multiple properties with high scores



Produce diverse and novel molecules

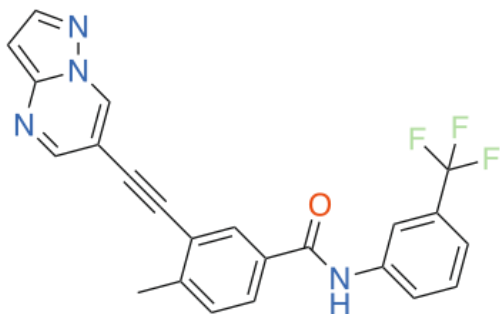


scarce lab measured data



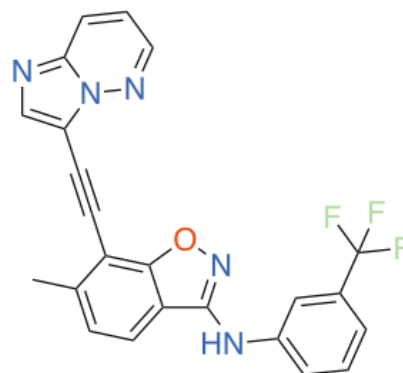
Finding Novel Drugs is Important

- Current generative model can only find molecules with high similarities comparing to existing drugs/potential drug candidates.



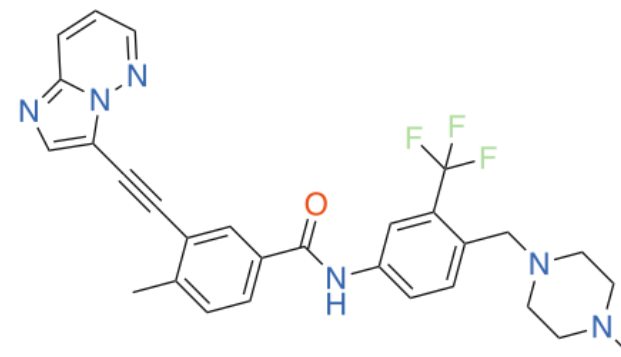
Gao et al.
Compound 7r
6 nM

Previously Reported
Drug Candidate



Zhavoronkov et al.
Compound 1
10 nM

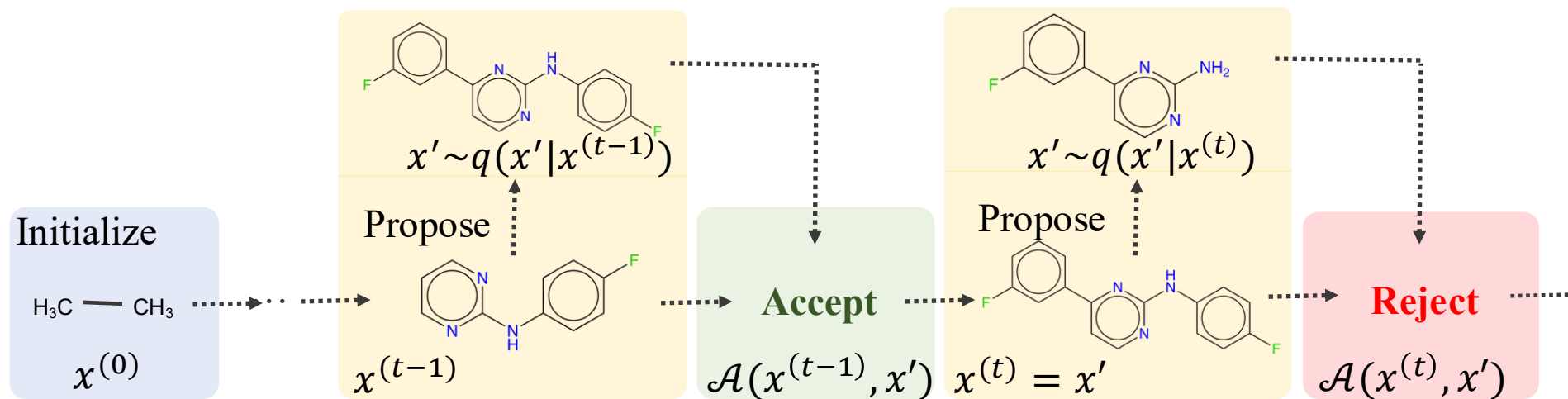
AI Designed
Drug Candidate



Ponatinib
9 nM

FDA Approved Drug

MARS: Iterative Graph Editing for Drug Generation

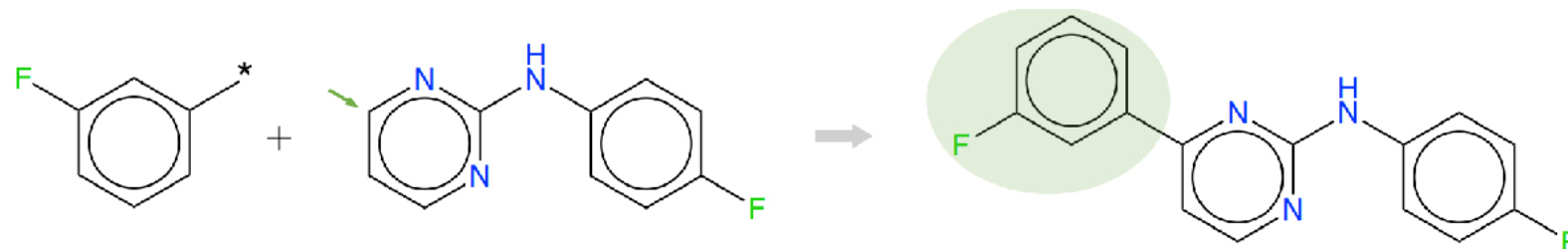


- Start from initial molecule (e.g. C_2H_6)
- For each step, propose a new molecule x' by modifying existing one, $x' \sim q(x'|x^{(t-1)})$

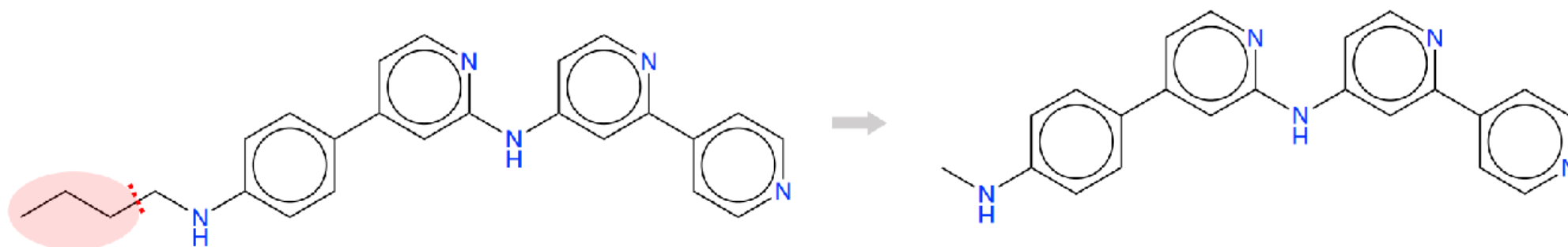
- Accept wrt ratio:
$$\mathcal{A}(x, x') = \min \left\{ 1, \frac{\pi^\alpha(x')q(x|x')}{\pi^\alpha(x)q(x'|x)} \right\}$$

Molecular Graph Editing

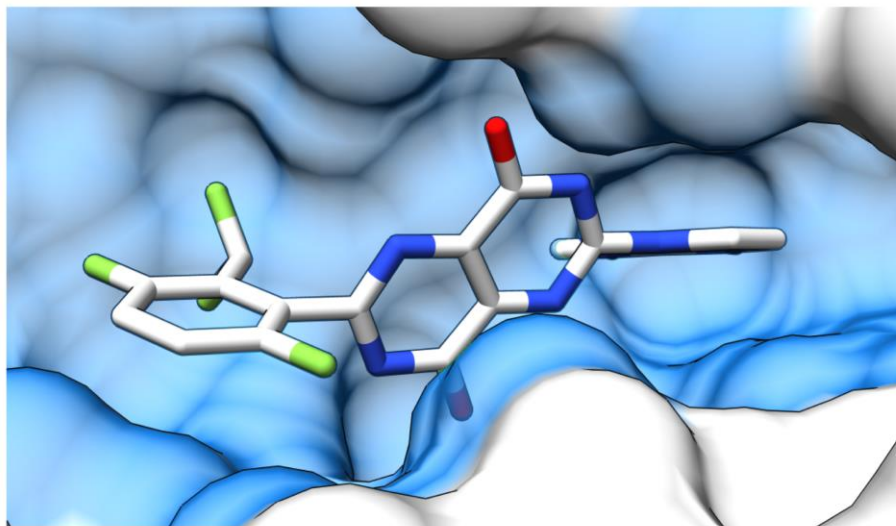
- Adding fragment



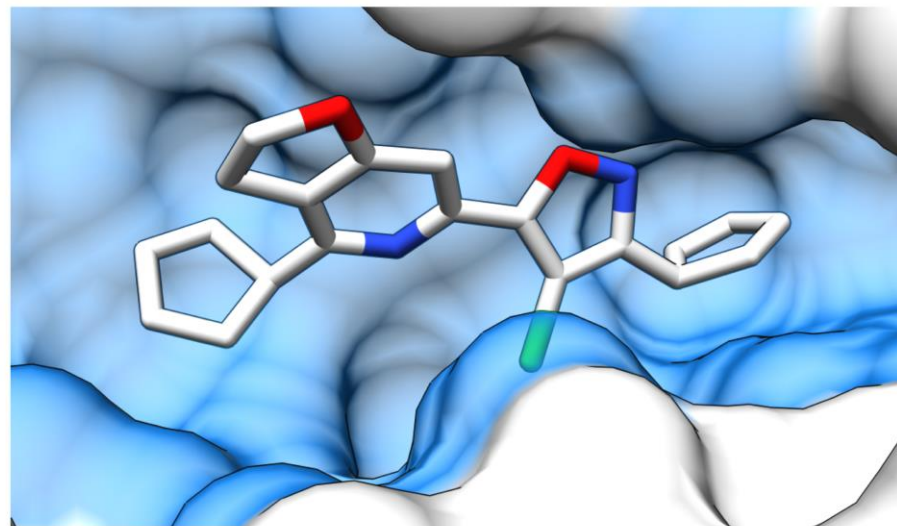
- Deleting fragment



MARS with 3D Structures Finds better Drugs fitting the Protein Pocket



Vina: -12.49 QED: 0.58 SAscore: 0.63



Vina: -12.41 QED: 0.80 SAscore: 0.68

Highlights of MARS

- Target-based molecule generation
- MARS, a **simple yet flexible** framework for **multi-objective**
 - Based on **MCMC** sampling
 - **Self-adaptive** proposal trained on the fly => **no need for data**
 - Generates better molecules and explores larger chemical space
- => can discover **novel and diverse** drug-like molecules
- Challenges remaining:
 - more properties
 - larger molecule, peptide, protein

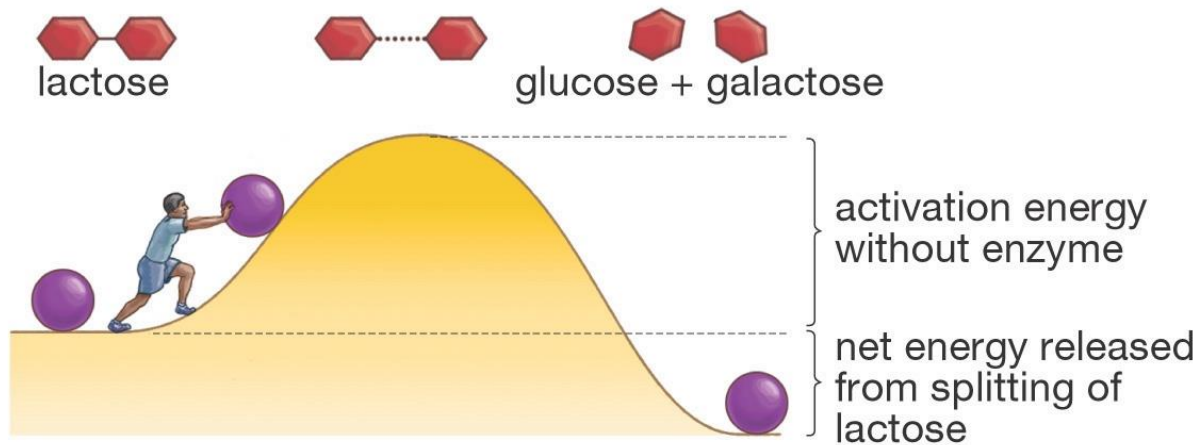
Outline

- Multi-scale Biological System and Drugs
 - molecule, cell, tissue, organs, cause of disease
- Basic AI Models for Biomecules
 - sequence, structure, generative model
- MARS: finding small molecule drugs with multiple properties
- ➔ • EnzyGen: A general generative model for enzyme design
- PPDiff: protein-binding complex design

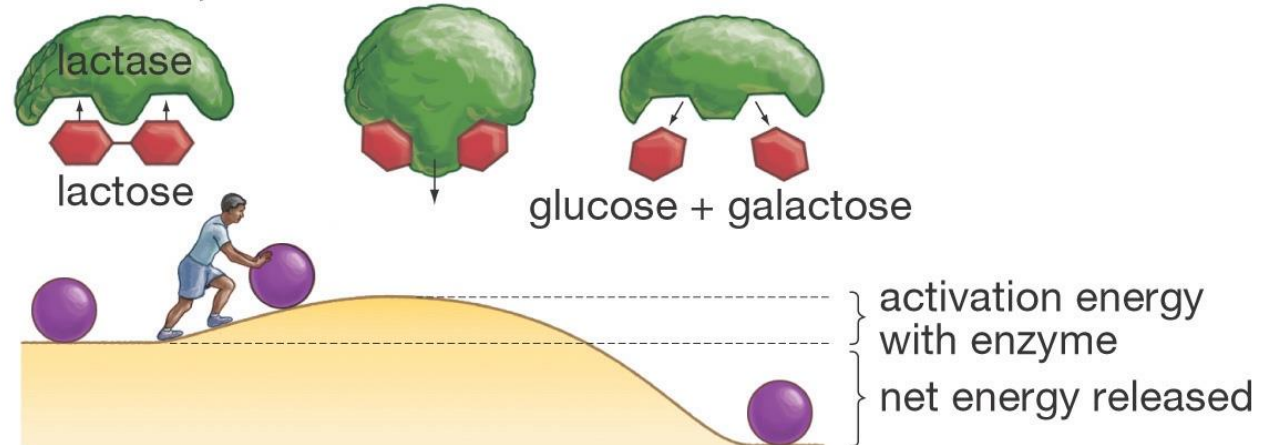
Designing Effective Enzymes

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

(a) Without enzyme

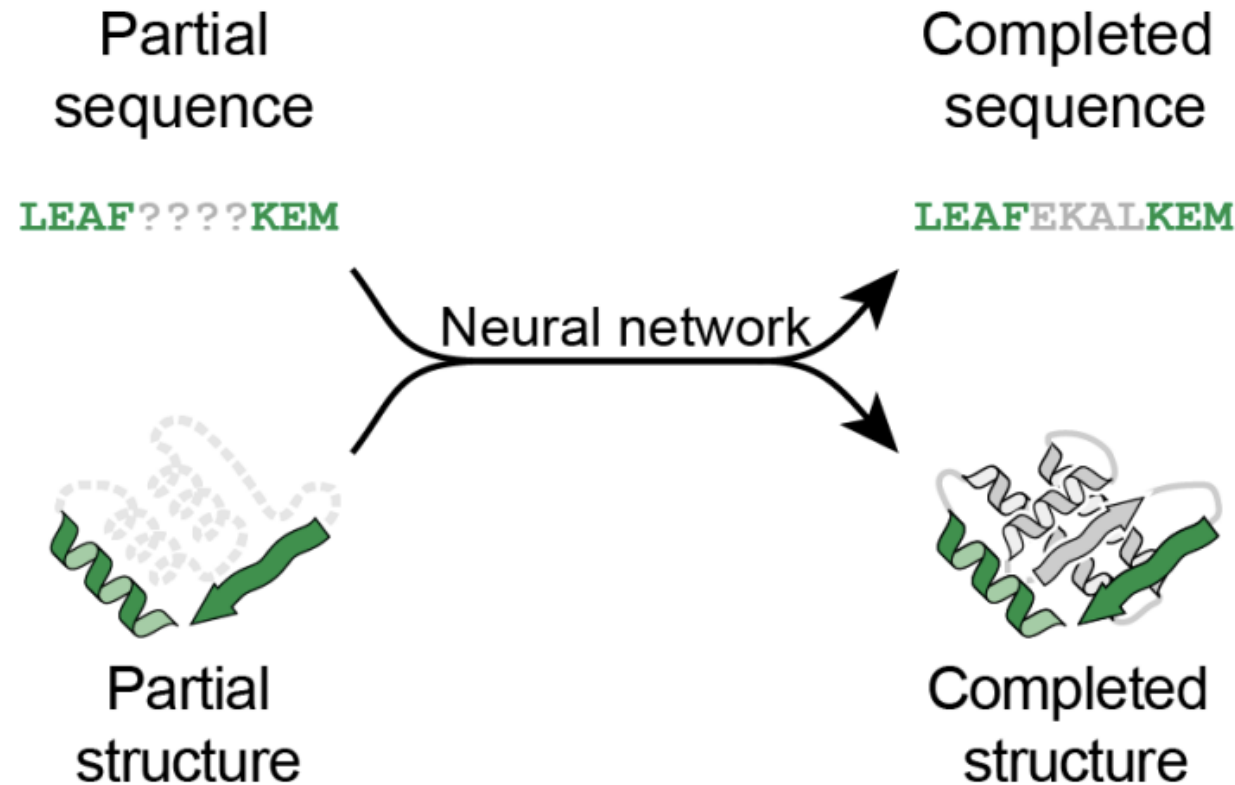


(b) With enzyme



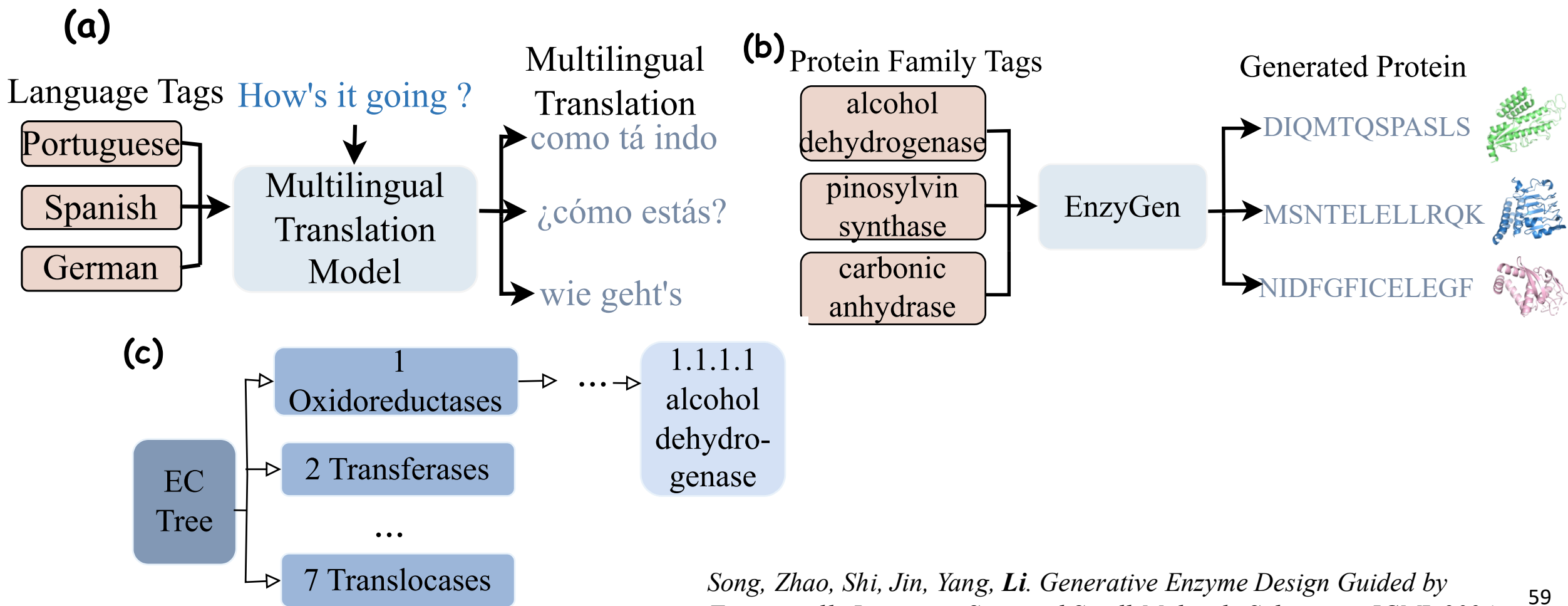
Motivation 1: How to design desired enzymes?

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



Motivation 2: How to design desired enzymes?

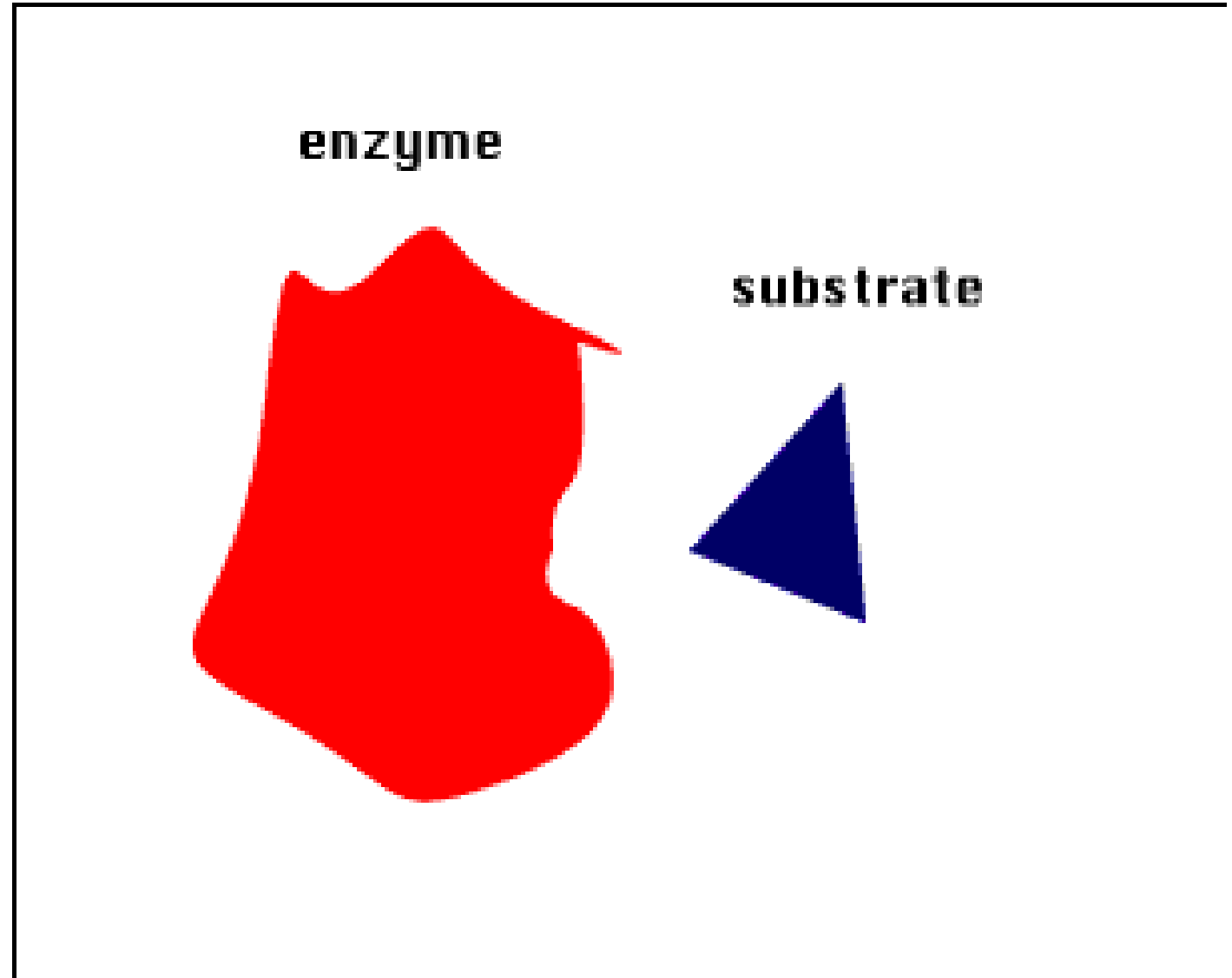
Enzyme classification tree indicates enzymatic reaction type



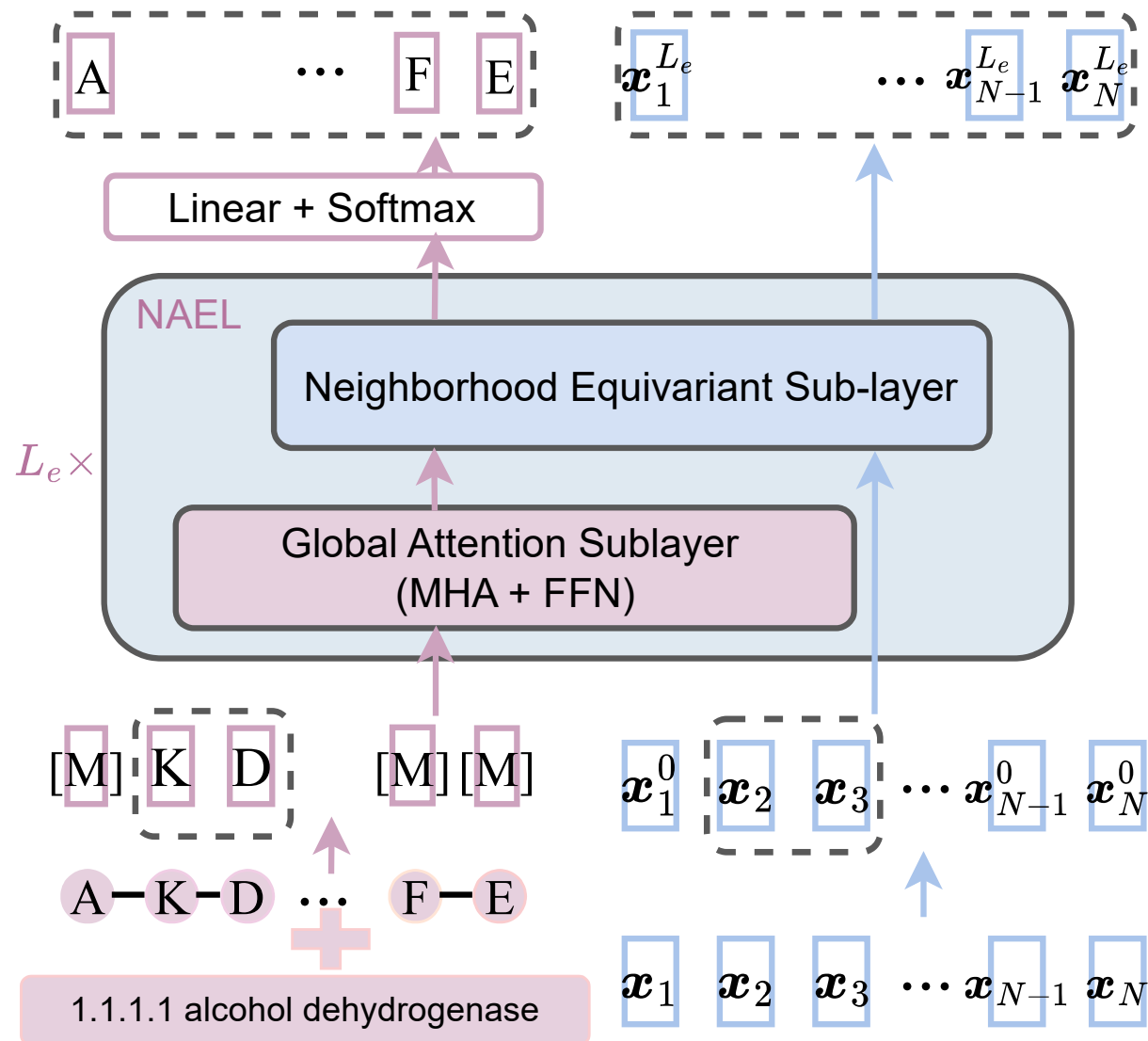
Motivation 3: How to design desired enzymes?

- Substrate Specificity:

Different enzymes binding to specific substrates to speedup enzymatic reactions

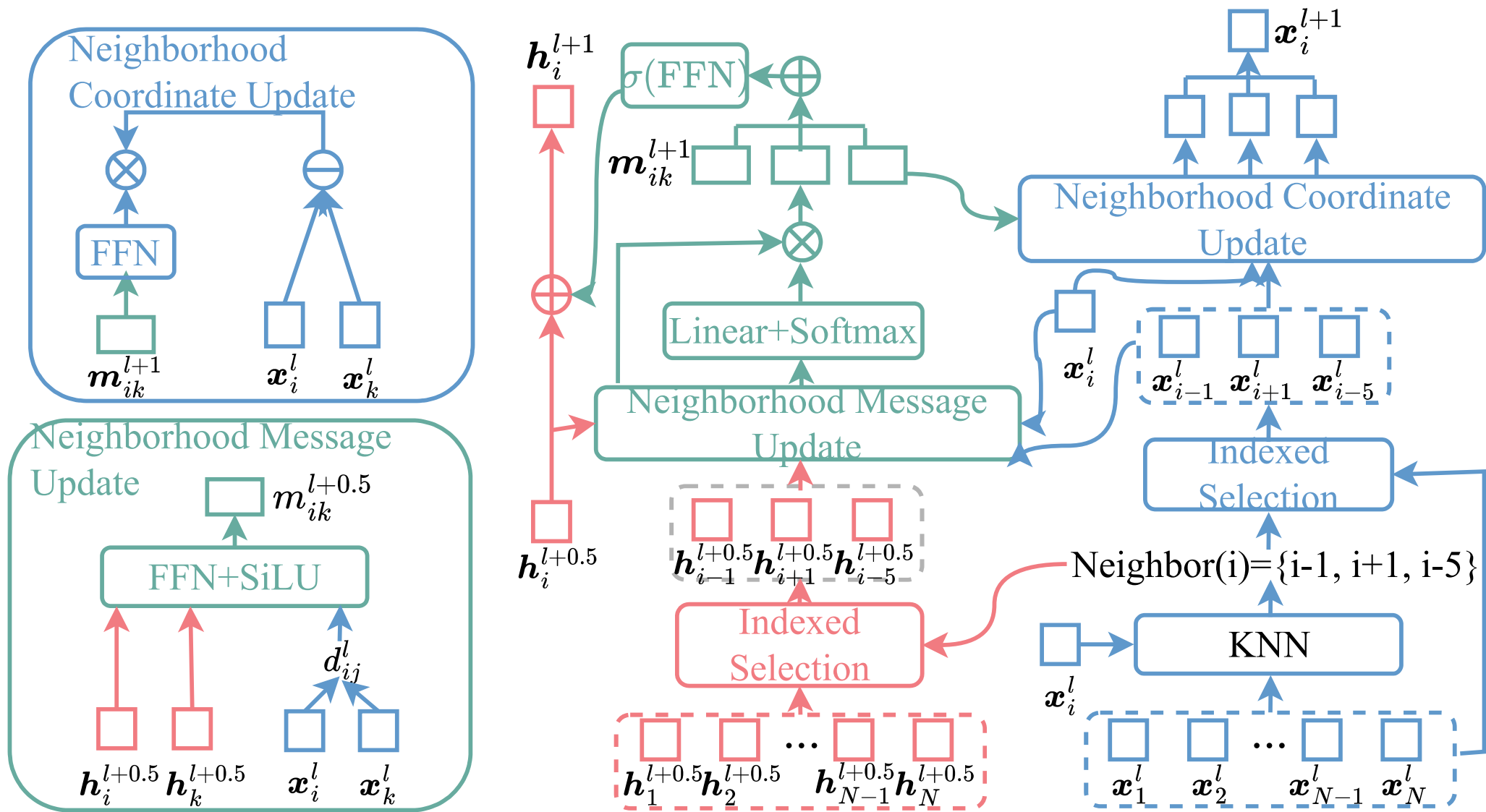


EnzyGen Model – NAEEL backbone

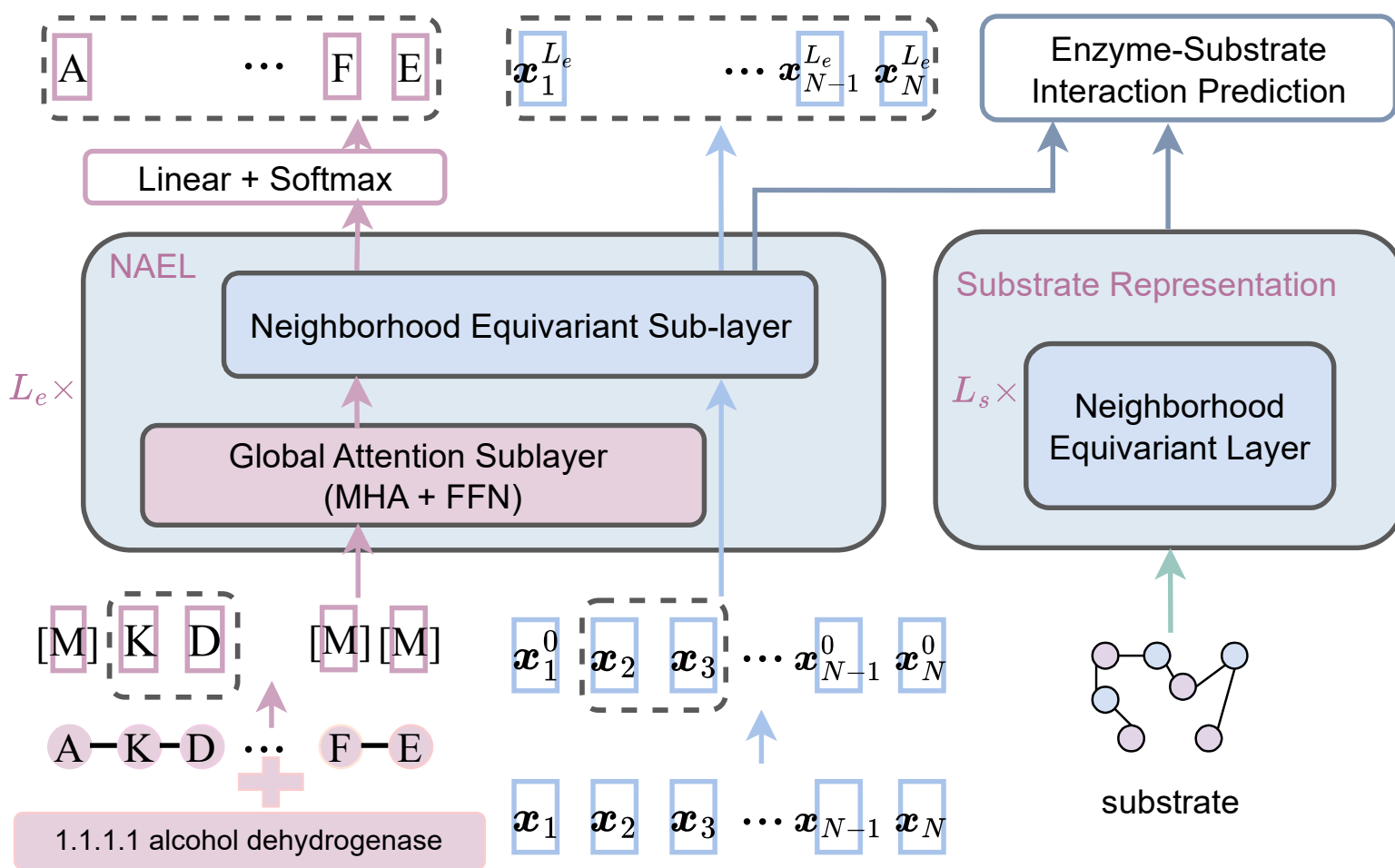


- Controllable Design
 - Functional Sites
 - Enzyme family category

Neighborhood Attentive Equivariant Layer



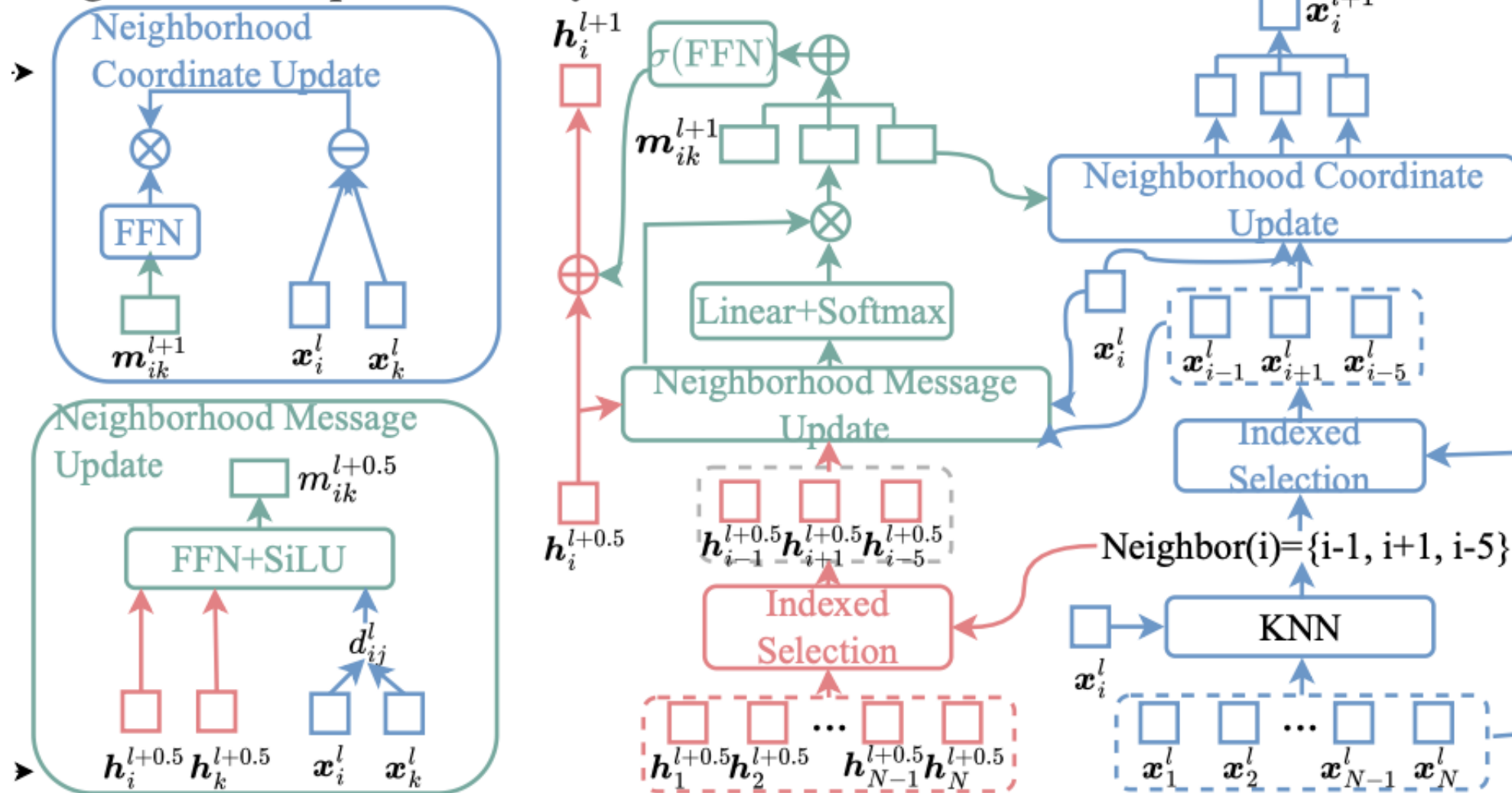
EnzyGen Learning



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

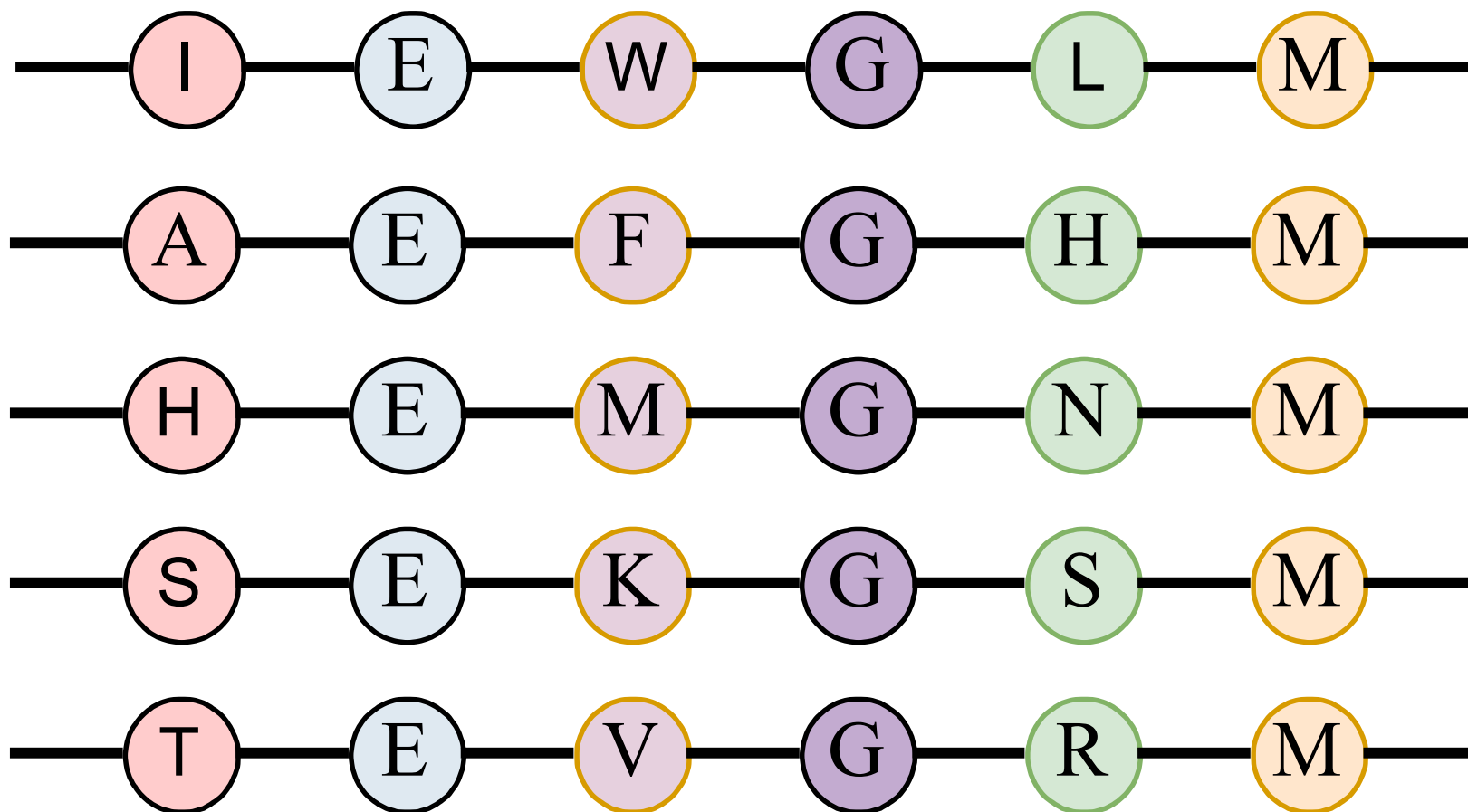
Neighborhood Attentive Equivariant Layer (NAEL)

Neighborhood Equivariant Layer



Functional Site Discovery

mining common sites within one family

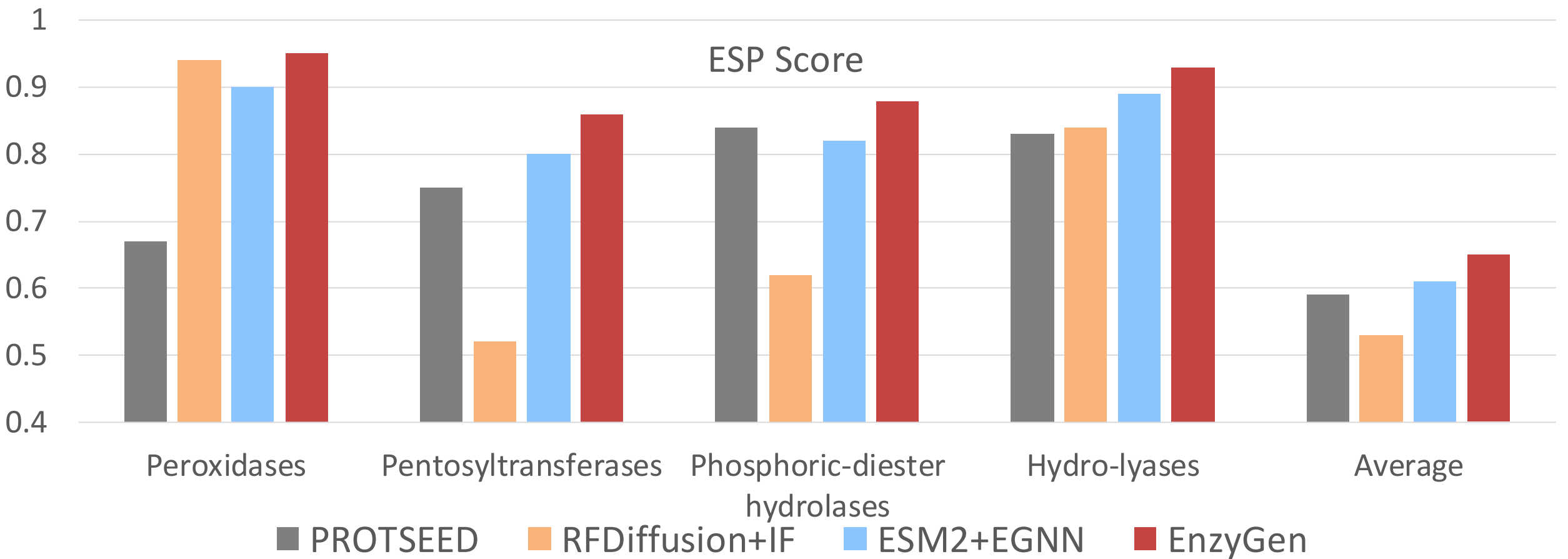


EnzyBench Dataset

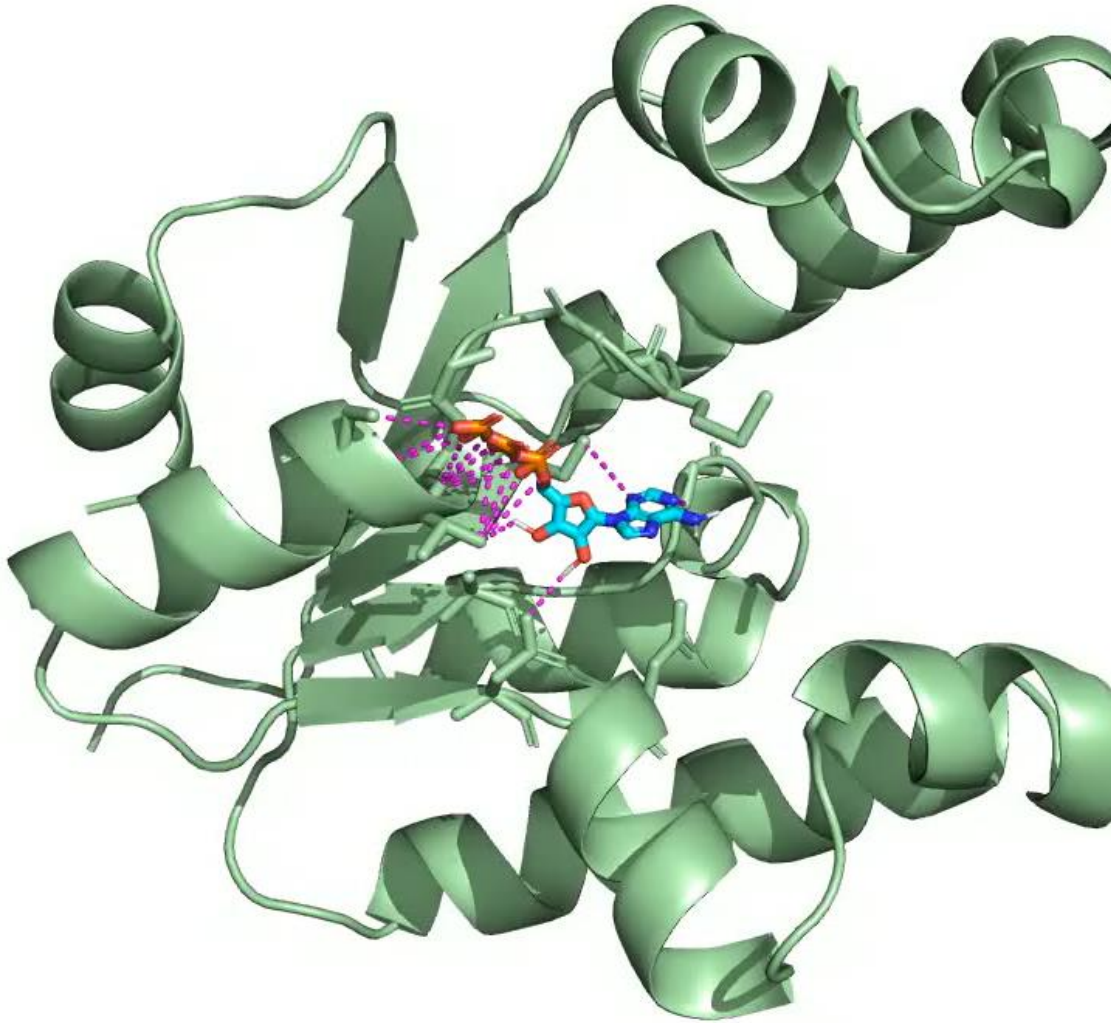
- Extracted from BRENDA
 - 8422 fourth-level enzyme classes (enzymatic reaction types)
- Selected PDB entries: 101974
 - 3157 fourth-level enzyme classes
 - discover functional sites for each class
 - Merging into third-level categories: 256
 - 30 largest categories
 - Split 50 for validation & 50 for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

EnzyGen generates enzymes with higher function scores

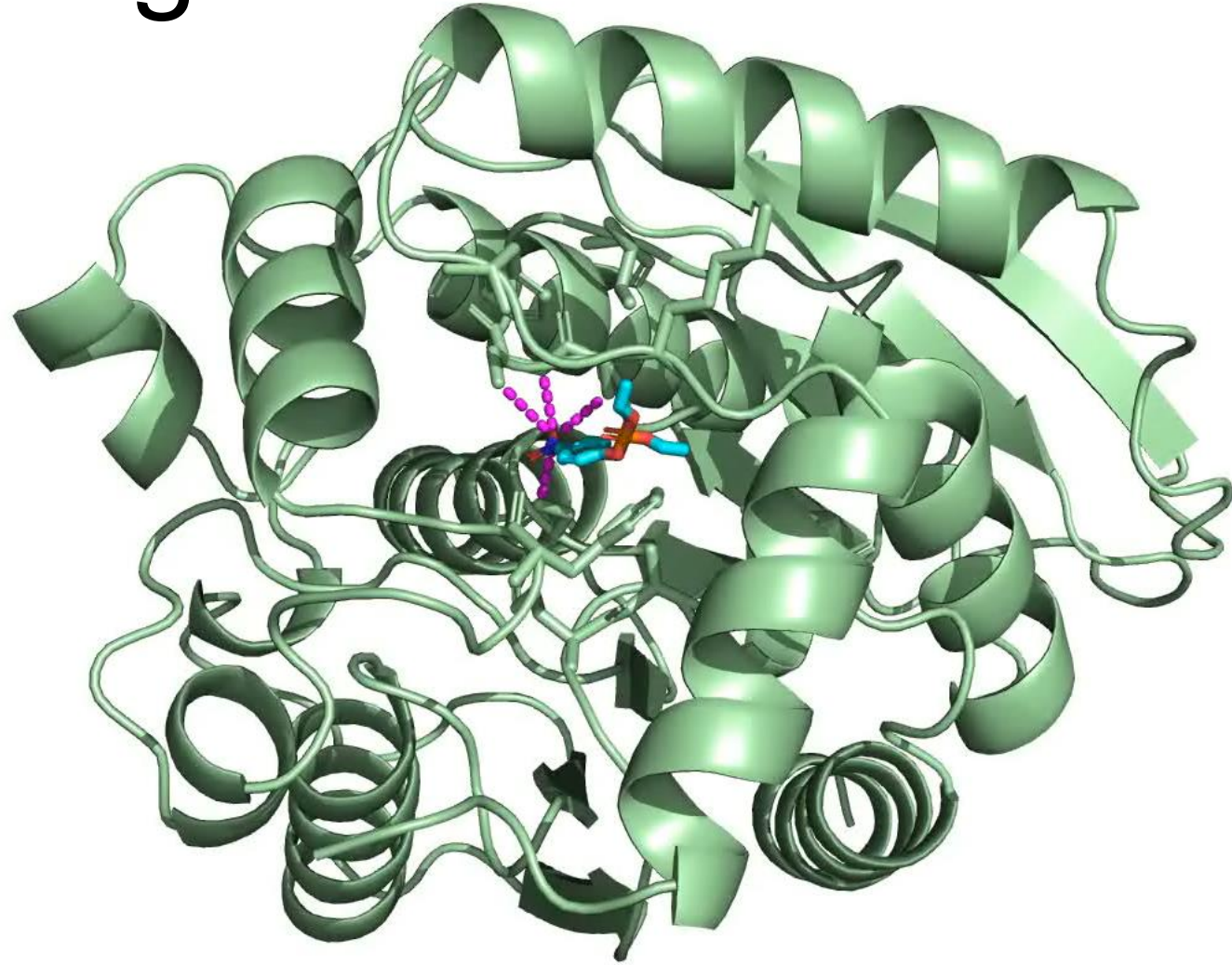
EnzyGen achieves higher enzyme-substrate interaction score in 20 out of 30 categories



EnzyGen designs “good” enzymes in zero-shot categories



Shikimate kinase
(ATP:shikimate 3-phosphotransferase)



Arylesterase
(substrate paraoxon)

Highlights of EnzyGen

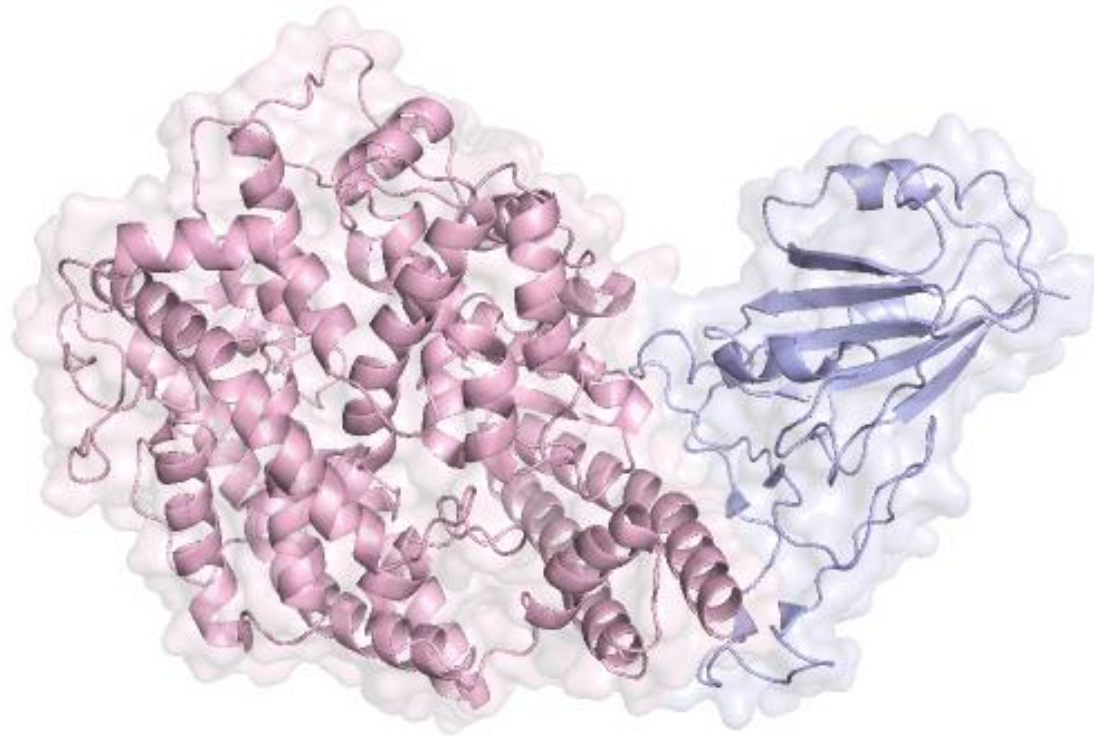
- A unified model for 3k enzyme families
- Guided Generation
 - Functional Important Sites, automatically mined from PDB
 - Enzymy category tags (BRENDA)
- Sequence and Structure Co-design
 - Neighborhood Attentive Equivariant Layer
- Trained takes substrate binding into consideration

Outline

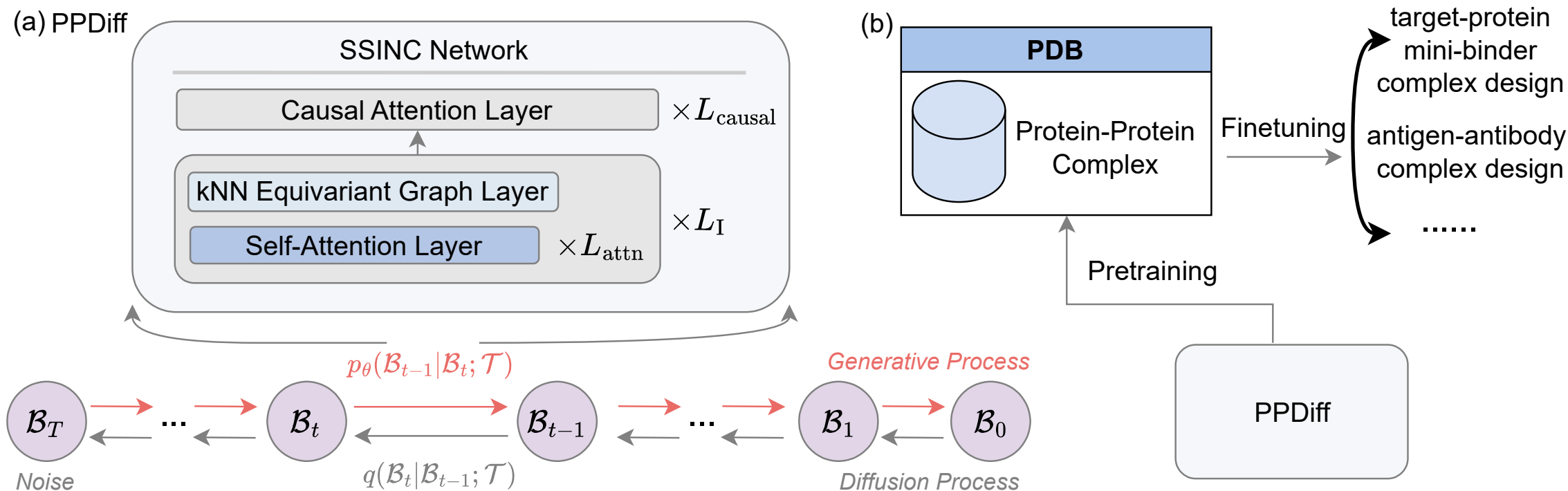
- Multi-scale Biological System and Drugs
 - molecule, cell, tissue, organs, cause of disease
- Basic AI Models for Biomecules
 - sequence, structure, generative model
- MARS: finding small molecule drugs with multiple properties
- EnzyGen: A general generative model for enzyme design
- ➔ • PPDiff: protein-binding complex design

Protein-Protein Complex Design

- Goal: design a protein to bind to another target protein



PPDiff

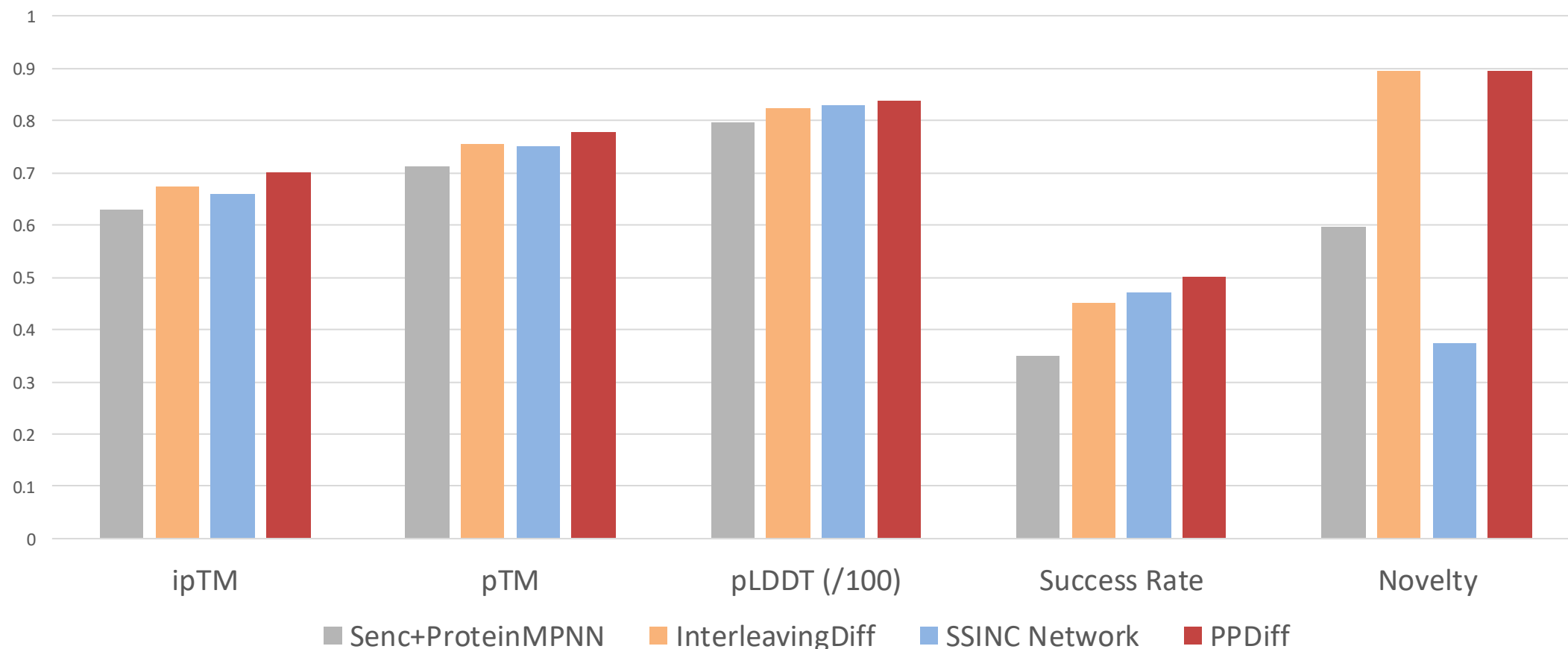


- Diffusing in hybrid space
 - Discrete sequence diffusion
 - Continuous structure diffusion
- SSINC Network
 - Interleaving network (NAEL)
 - Casual attention layers

PPDiff generates novel binders with higher success rate

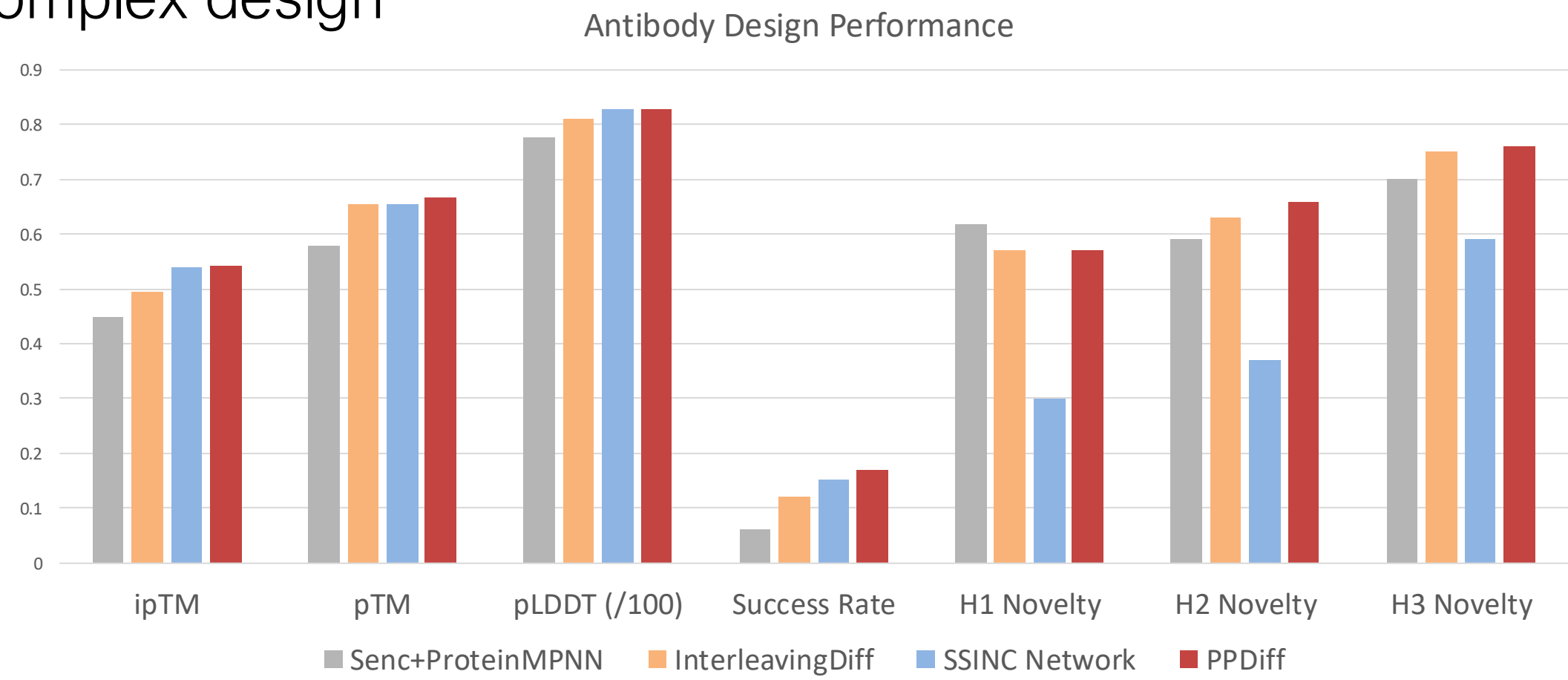
PPDiff achieves **50%** success rate across diverse protein targets

Top-1 results on general protein-protein complex design



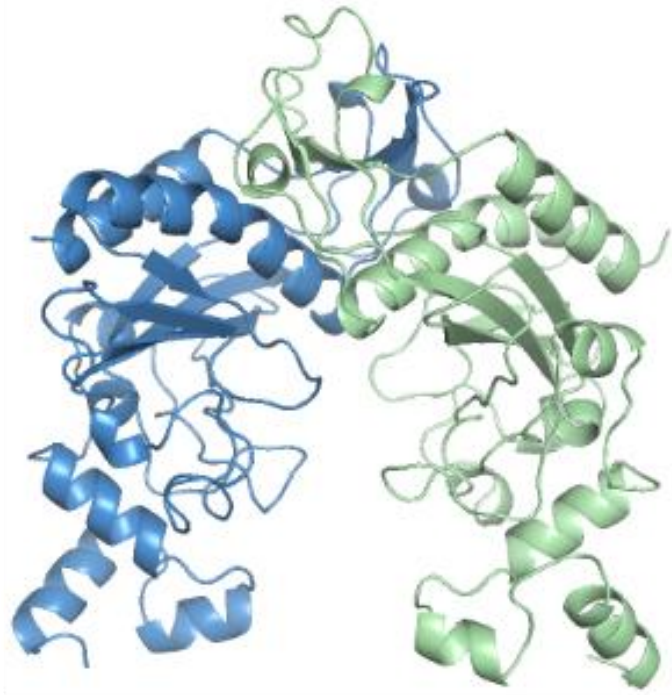
PPDiff generates novel antibodies with higher success rate

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

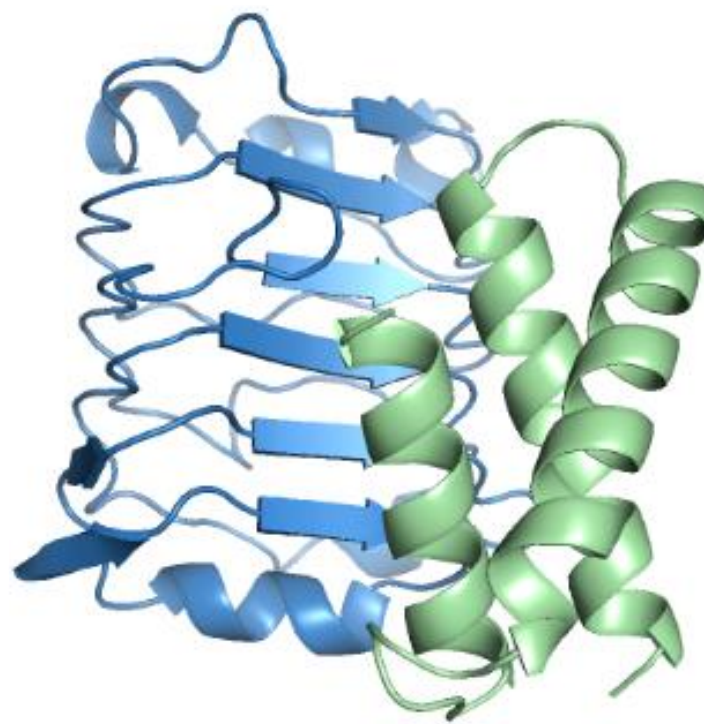


PPDiff designs high-affinity binders/antibody across diverse target proteins

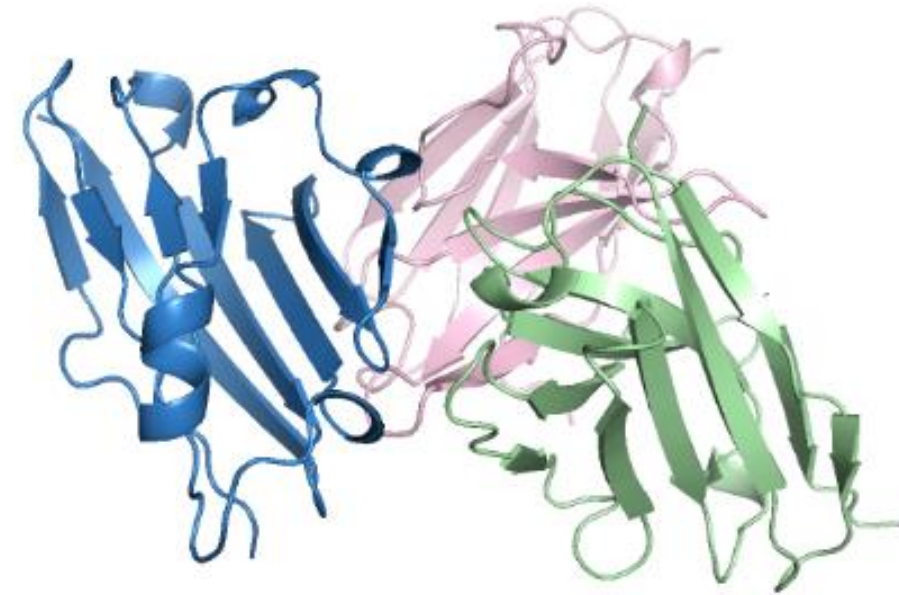
influenza A H3 haemagglutinin



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%

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

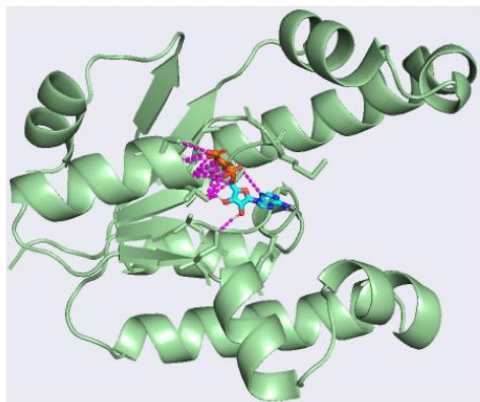
Summary

- Multi-scale Biological System and Drugs
 - molecule, cell, tissue, organs, cause of disease
- Basic AI Models for Biomecules
 - Tokenization, Transformer, MPNN, EGNN, Diffusion Model
- MARS: finding small molecule drugs with multiple properties
- EnzyGen: A general generative model for enzyme design
- PPDiff: protein-binding complex design

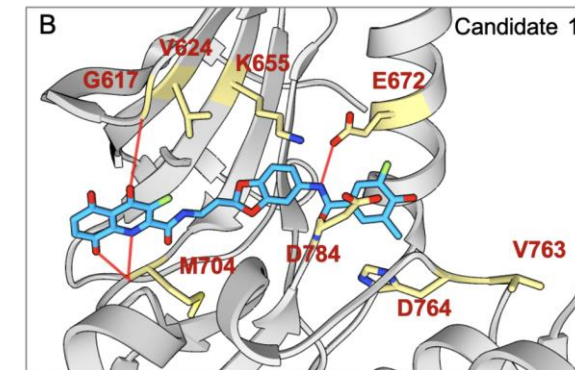
Takeaway of AI Drug Design

- Formulating as an AI problem
 - Input info: properties, tags, motifs, ligand, 3D structure
 - Generate: sequence, graph, structure
- Modelling Structure/Geometry is critical for molecules and function
- Modeling the mutual constraints between sequence and structure is useful
- Modelling Interaction between protein-ligand complex

Molecule Design at CMU Li lab



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



Protein

EnzyGen

SurfPro

IsEMPro

LSSAMP

PPDiff

InstructPro

Small Molecule

MARS

MolEdit3D

RLHEX