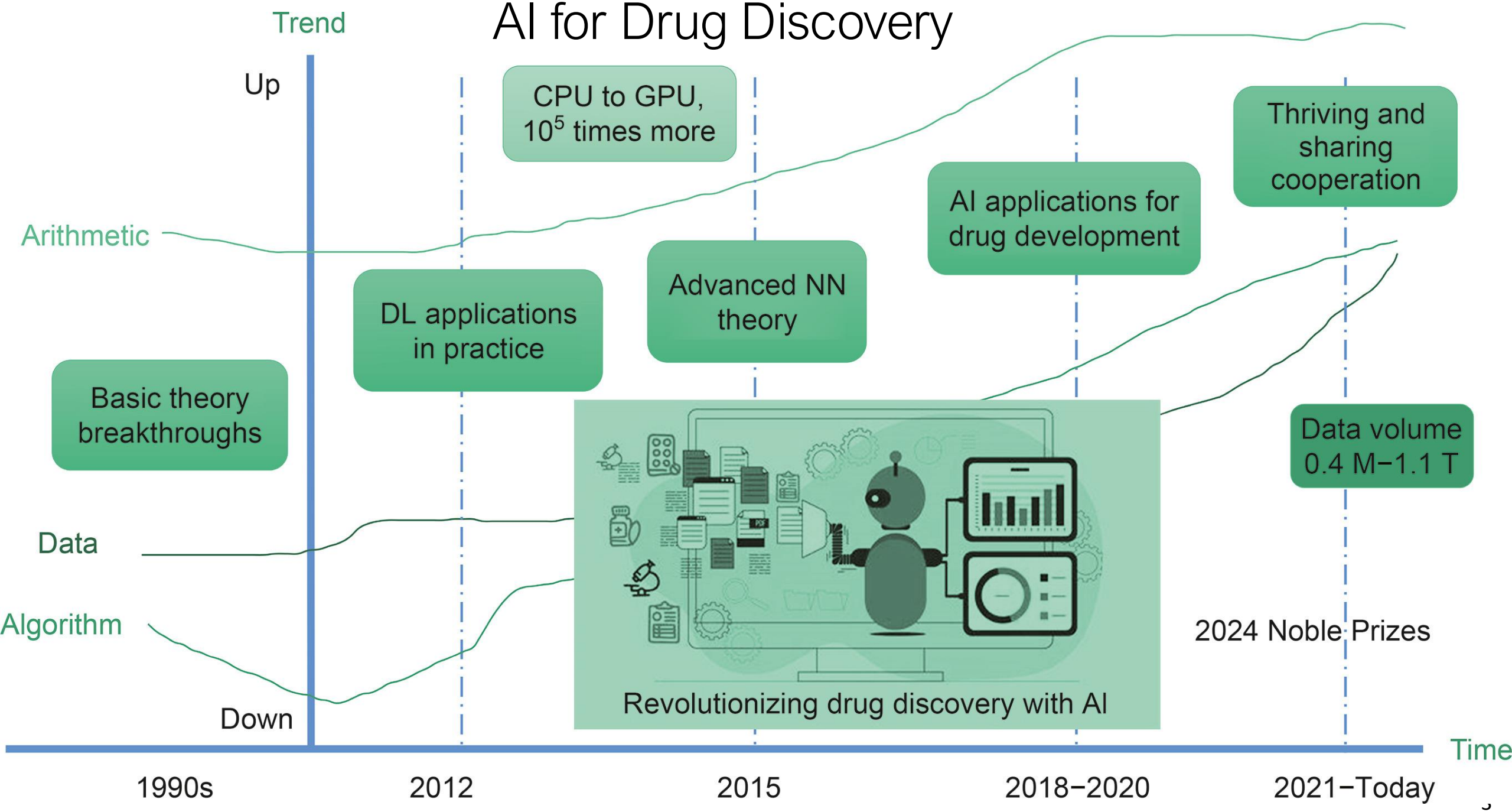
Accelerating Drug Design with Generative AI

Lei Li

November 26, 2025

Language Technologies Institute | Carnegie Mellon University School of Computer Science

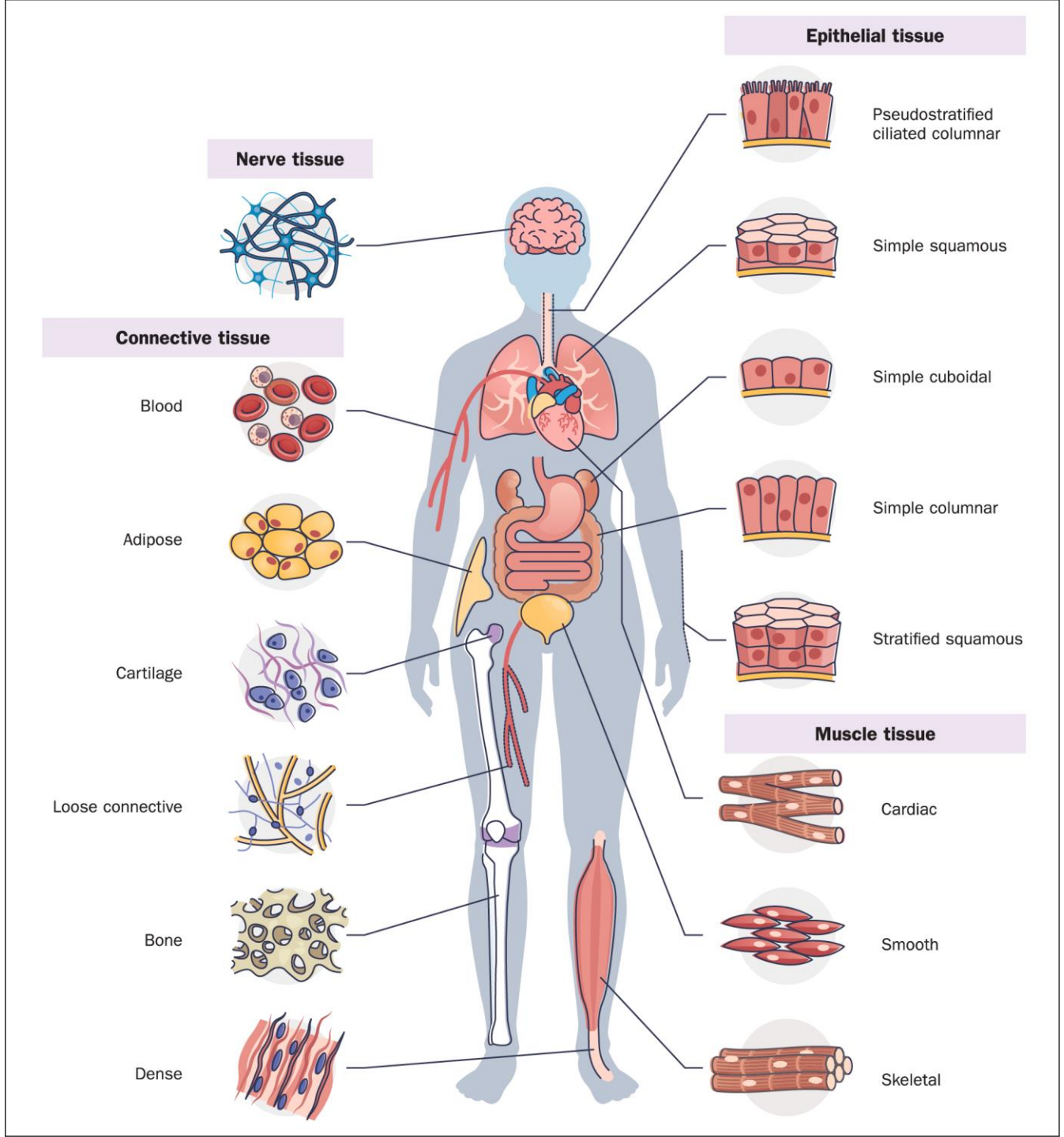
AI for Drug Discovery

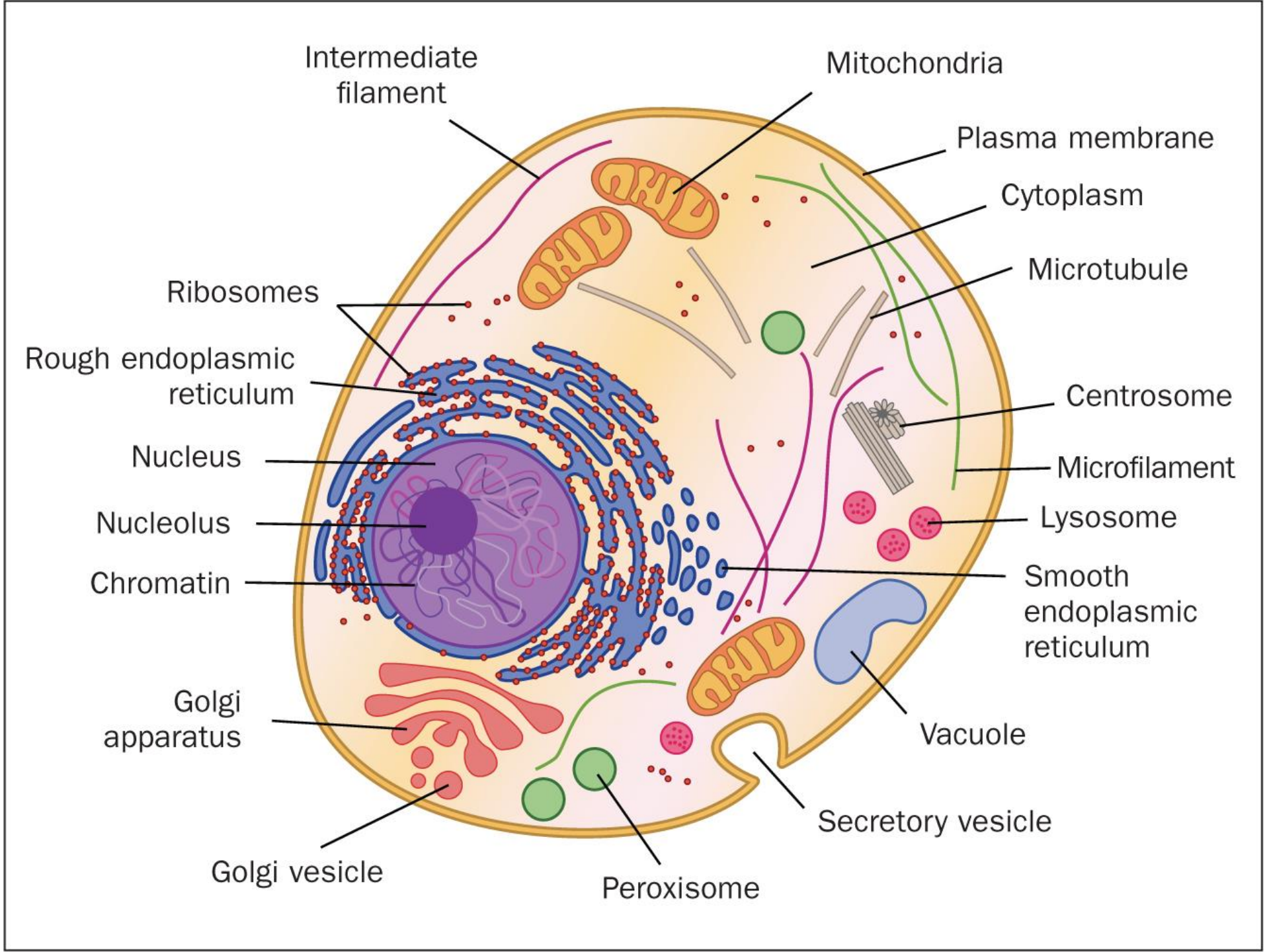


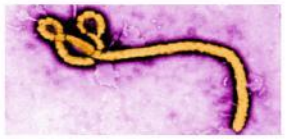
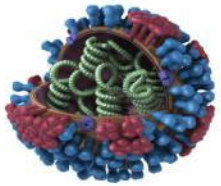
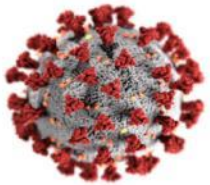
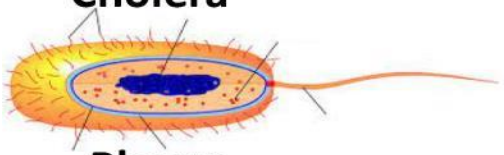
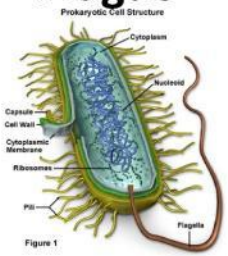
Fu et al, The future of pharmaceuticals: Artificial intelligence in drug discovery and development. JPA 2025.

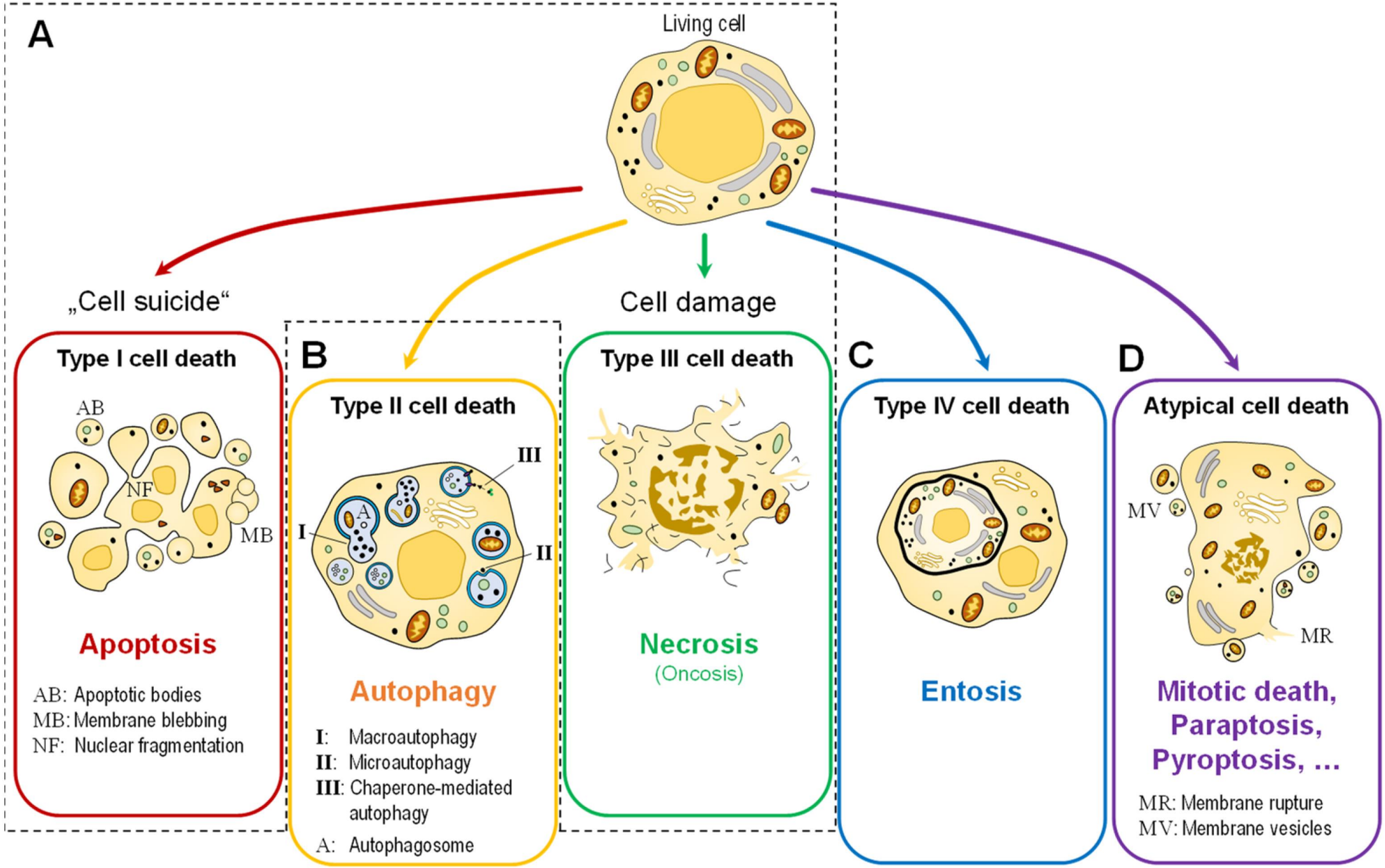
Outline

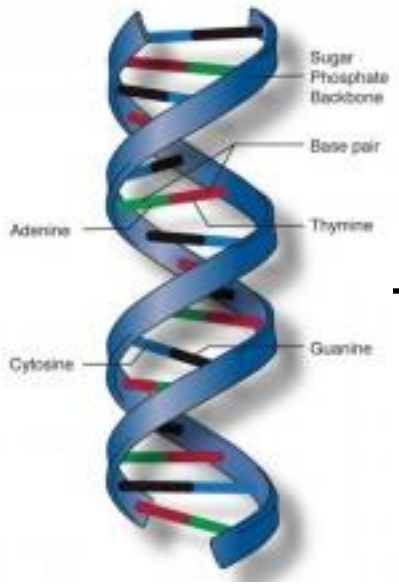
- Multi-scale Biological System and Drugs
  - molecule, cell, tissue, organs, cause of disease
- Basic AI Models for Biomecules
  - sequence, structure, generative model
- MARS: finding small molecule drugs with multiple properties
- EnzyGen: A general generative model for enzyme design
- PPDiff: protein-binding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

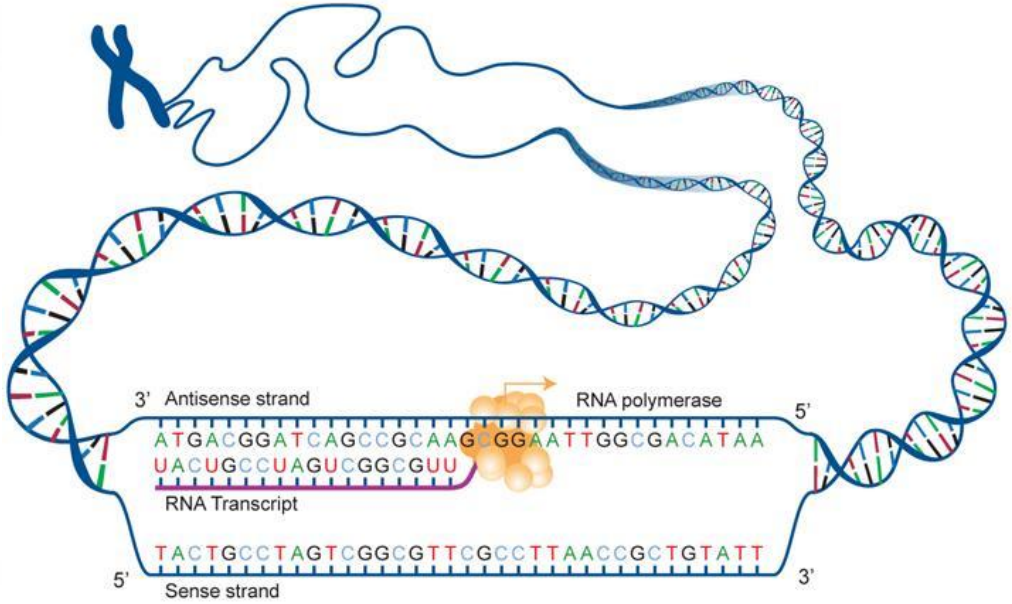

Epithelial tissue
Pseudostratified ciliated columnar
Simple squamous
Simple cuboidal
Simple columnar
Stratified squamous
Nerve tissue
Connective tissue
Blood
Adipose
Cartilage
Loose connective
Bone
Dense
Muscle tissue
Cardiac
Smooth
Skeletal

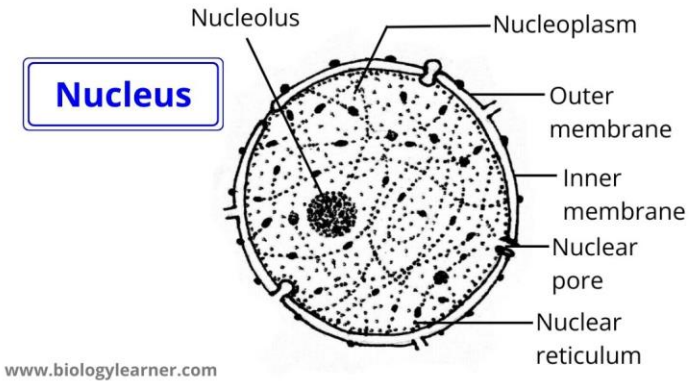

Intermediate filament
Mitochondria
Plasma membrane
Cytoplasm
Microtubule
Ribosomes
Rough endoplasmic reticulum
Centrosome
Nucleus
Microfilament
Nucleolus
Lysosome
Chromatin
Smooth endoplasmic reticulum
Golgi apparatus
Vacuole
Secretory vesicle
Golgi vesicle
Peroxisome

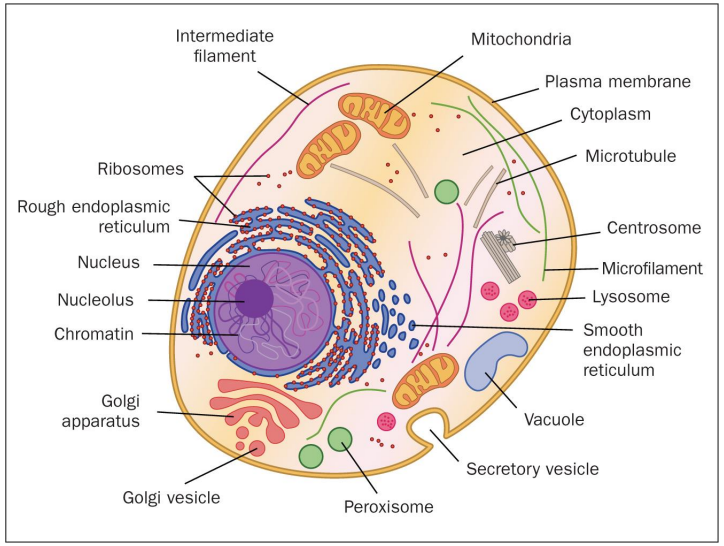
	Target Cells
Ebola Virus 	Hepatocytes, Endothelial cells, Phagocytes
Influenza Virus 	Ciliated Epithelium of conducting airways
Corona Virus 	Lungs and respiratory epithelial cells
Zika Virus 	From Brain, Placenta, skin to Testis, Kidney, Retina, During Pregnancy causes microcephaly
Cholera 	Intestinal Epithelial Cells
Plague 	Neutrophils

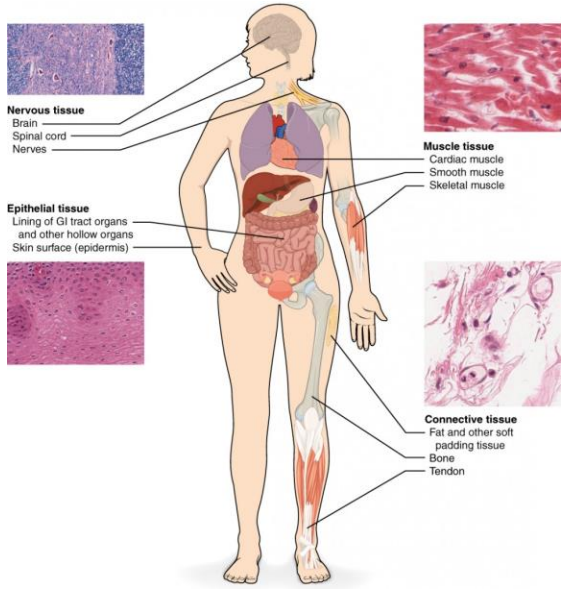

A
Living cell
„Cell suicide“
Cell damage
Type I cell death
AB
NF
MB
Apoptosis
AB: Apoptotic bodies
MB: Membrane blebbing
NF: Nuclear fragmentation
B
Type II cell death
III
I
A
II
Autophagy
I: Macroautophagy
II: Microautophagy
III: Chaperone-mediated autophagy
A: Autophagosome
Type III cell death
Necrosis
(Oncosis)
C
Type IV cell death
Entosis
D
Atypical cell death
MV
MR
Mitotic death, Paraptosis, Pyroptosis, ...
MR: Membrane rupture
MV: Membrane vesicles


Sugar
Phosphate
Backbone
Base pair
Adenine
Thymine
Cytosine
Guanine

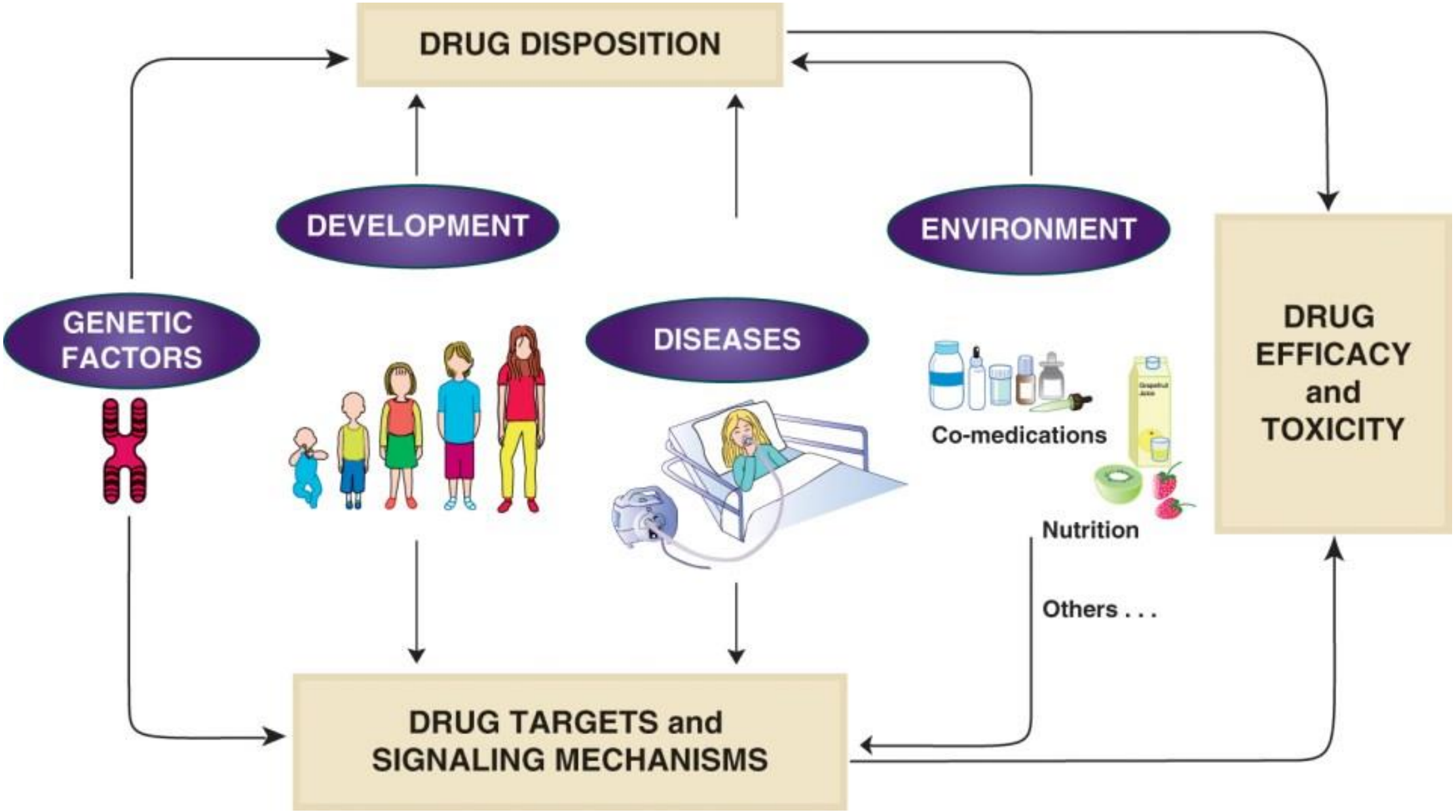

3' Antisense strand
RNA polymerase
5'
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAA
UACUGCCUAGUCGGCGUU
RNA Transcript
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT
5'
Sense strand
3'


Nucleus
Nucleolus
Nucleoplasm
Outer membrane
Inner membrane
Nuclear pore
Nuclear reticulum
www.biologylearner.com

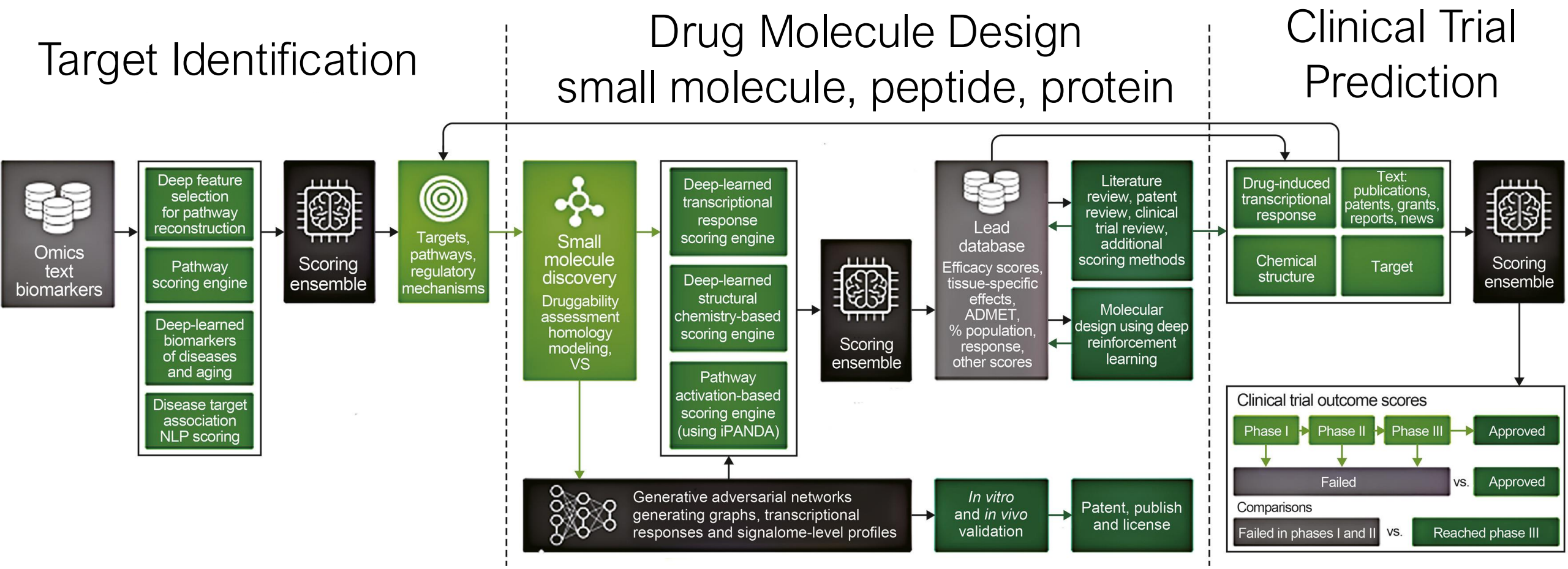

Intermediate filament
Mitochondria
Plasma membrane
Cytoplasm
Microtubule
Ribosomes
Rough endoplasmic reticulum
Centrosome
Nucleus
Microfilament
Nucleolus
Lysosome
Chromatin
Smooth endoplasmic reticulum
Golgi apparatus
Vacuole
Secretory vesicle
Golgi vesicle
Peroxisome


Nervous tissue
Brain
Spinal cord
Nerves
Muscle tissue
Cardiac muscle
Smooth muscle
Skeletal muscle
Epithelial tissue
Lining of GI tract organs and other hollow organs
Skin surface (epidermis)
Connective tissue
Fat and other soft padding tissue
Bone
Tendon

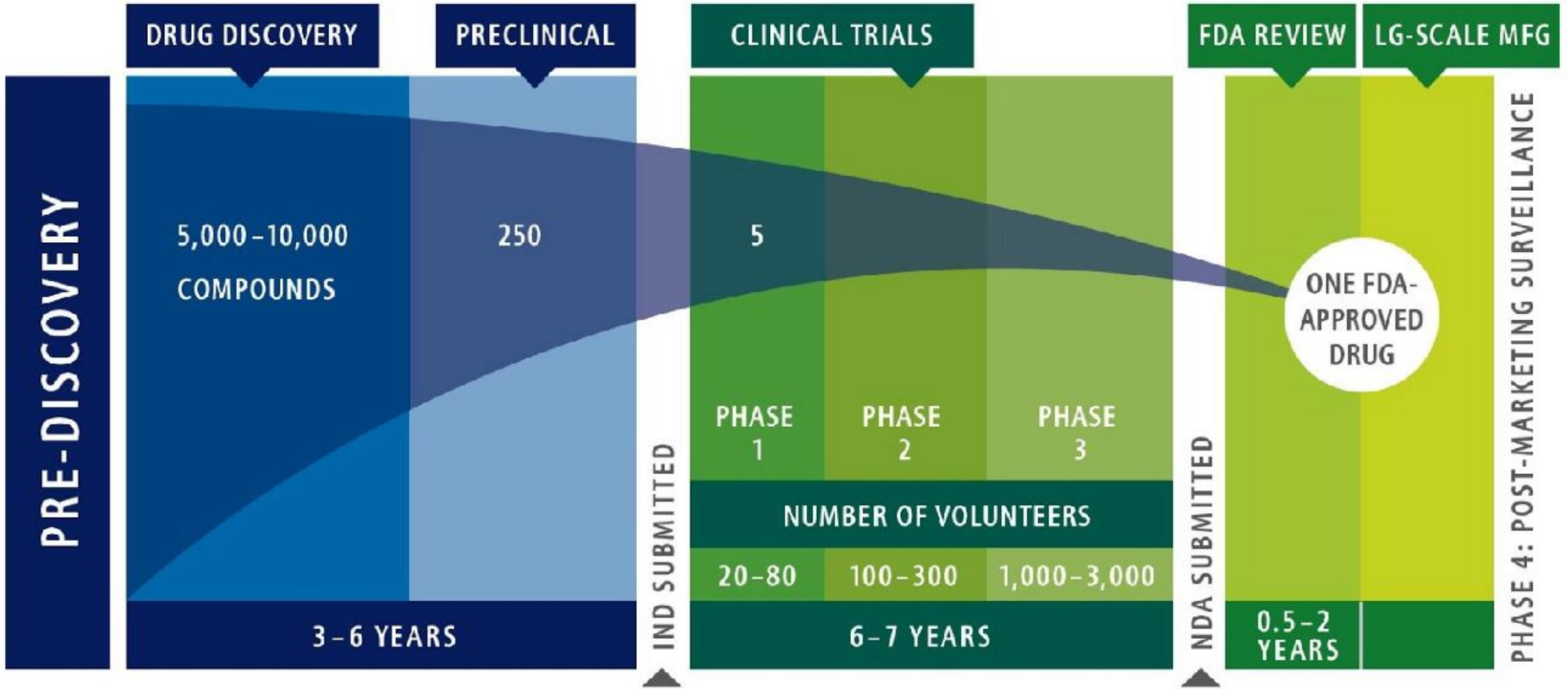
Finding Effective Drugs



AI for Drug Discovery



Drug Discovery and Development

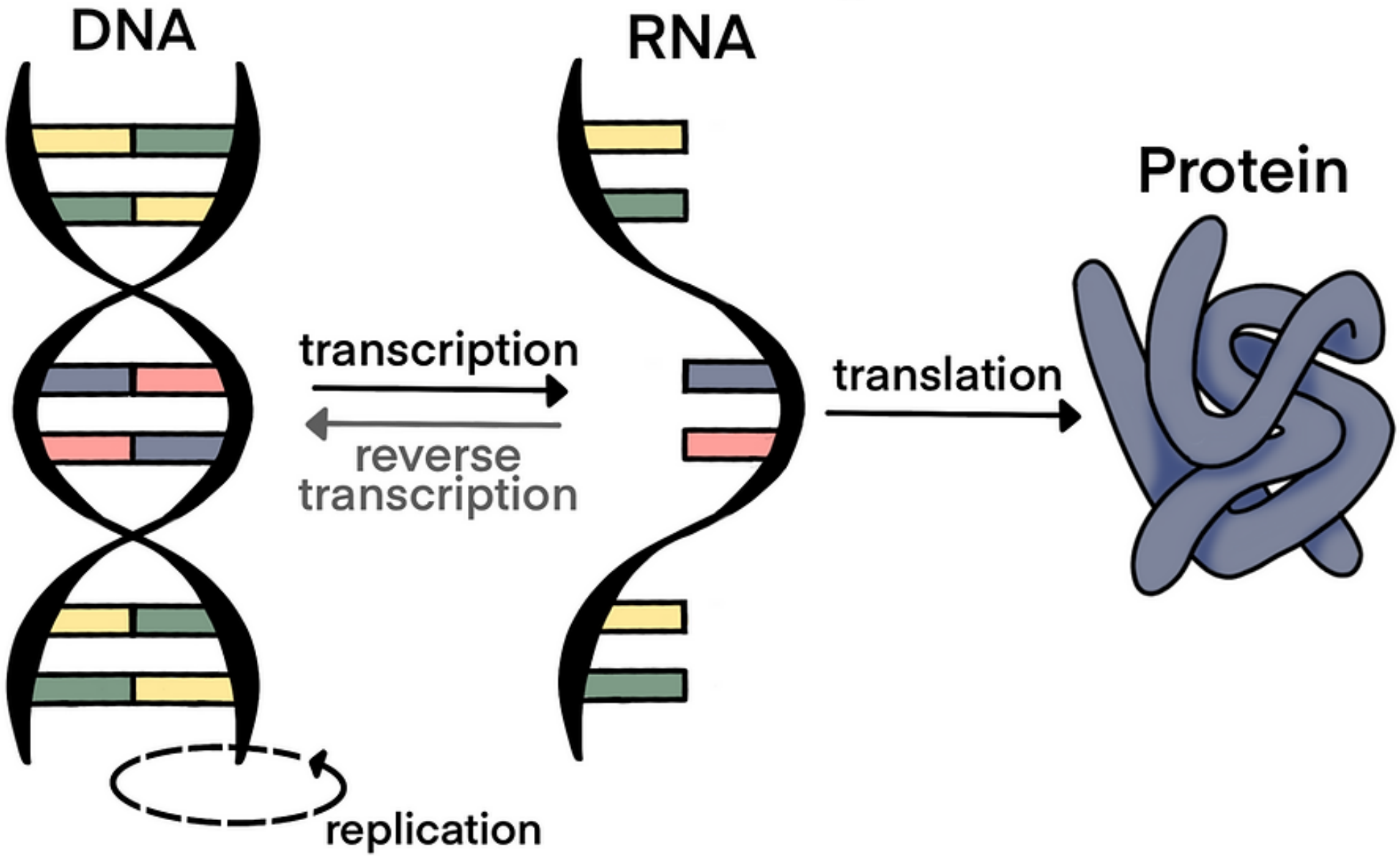


- Drug discovery and development is a long and risky process.
- The successful launch of a new drug cost $1.3 billion on average and more than 10 years.

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

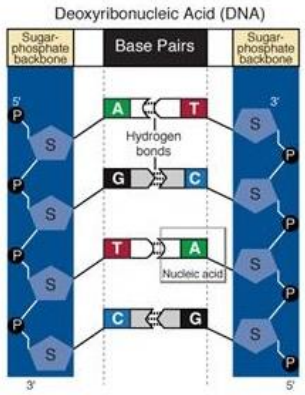
central dogma

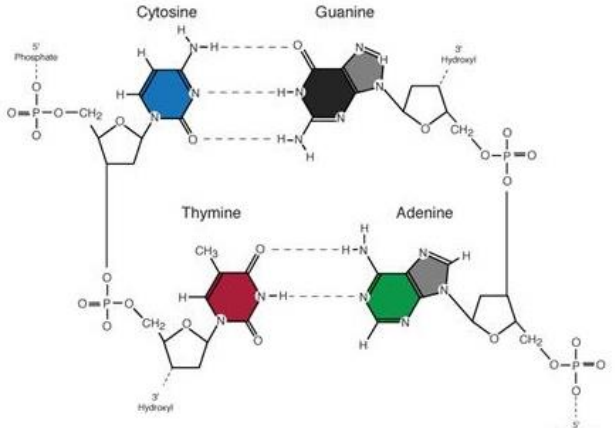


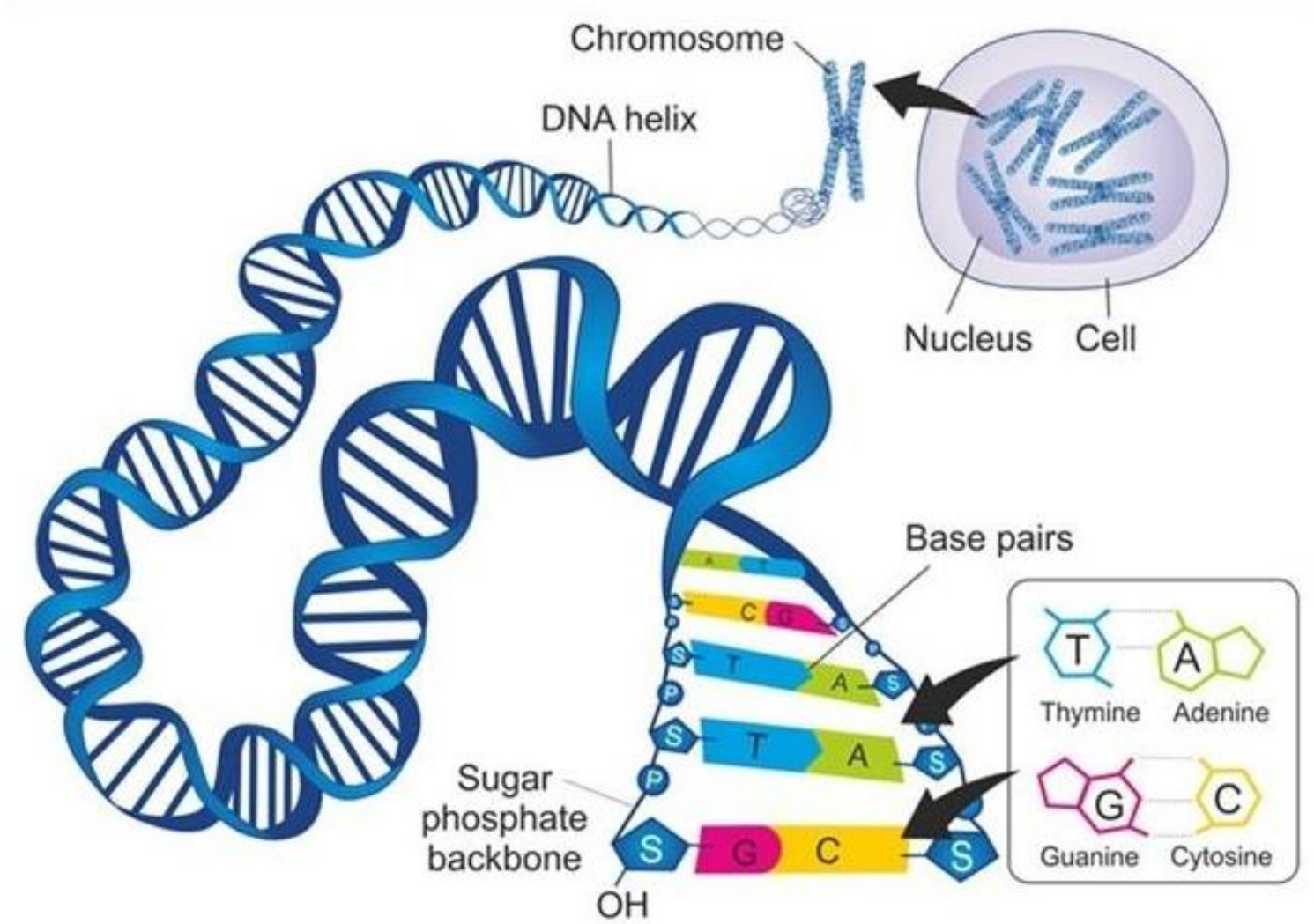
DNA

Deoxyribonucleic acid (DNA): encode and store genetic information.

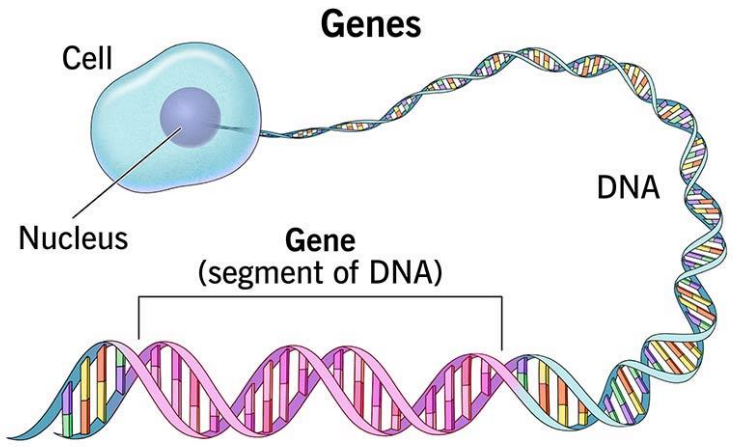
composed of base pairs: A, T, G, C





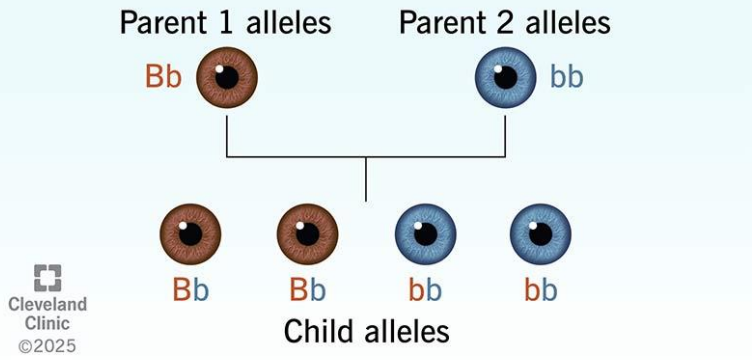


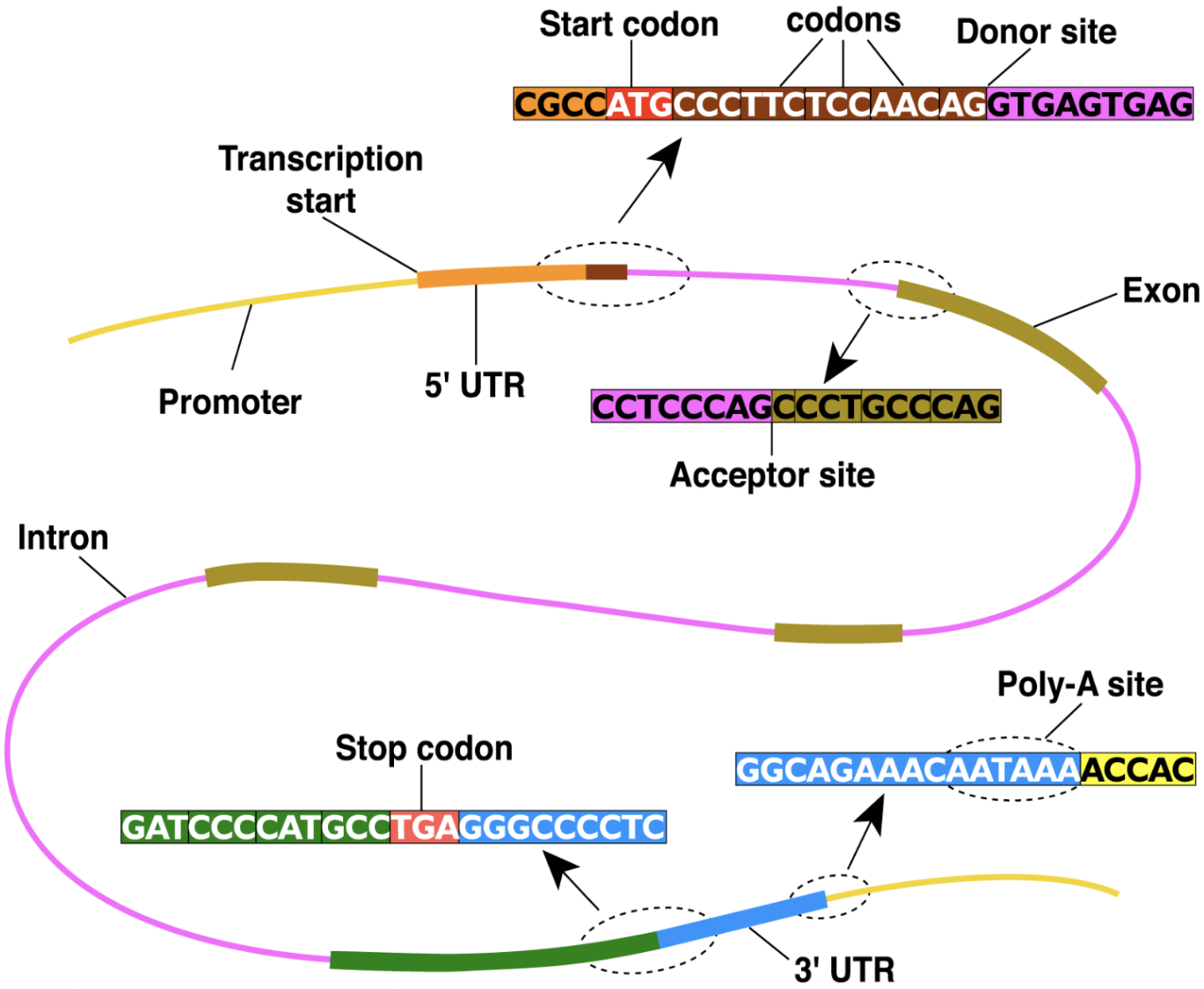
Gene



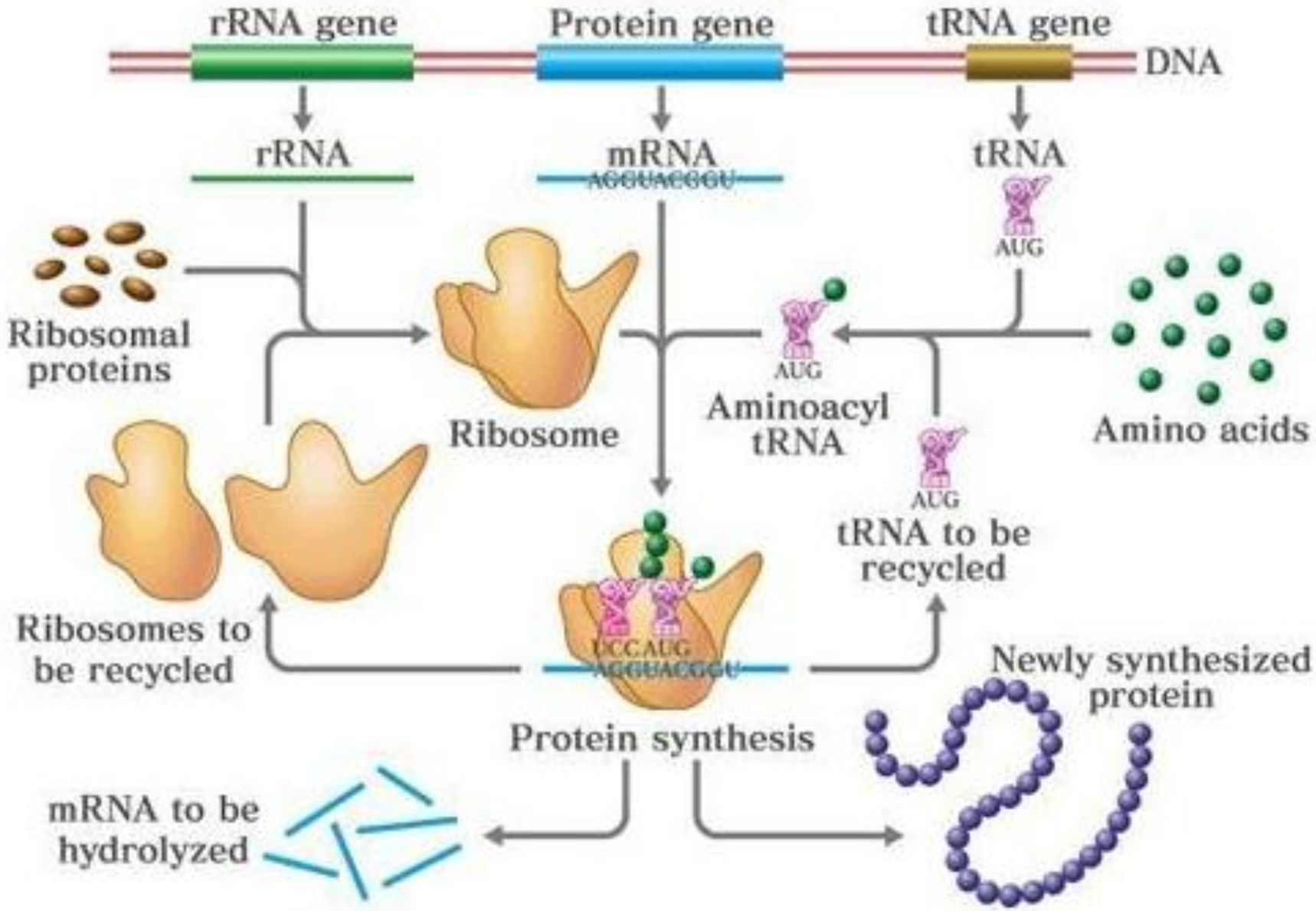
Alleles - different versions of genes (B/b)

Ex. Child's possible eye color depends on the allele received from each parent

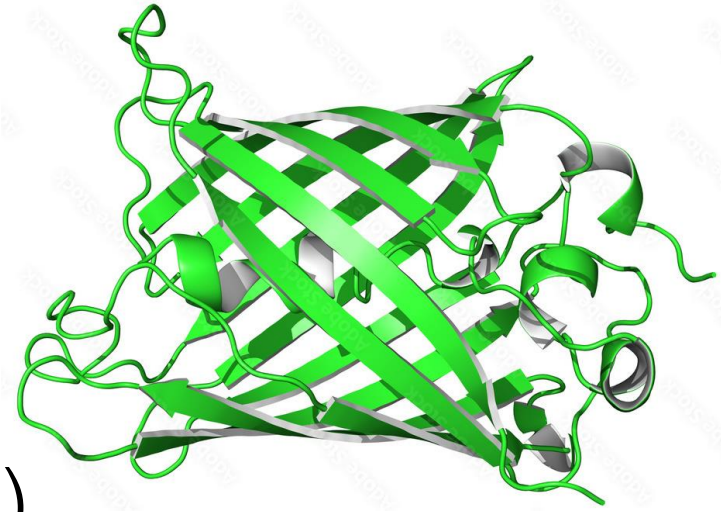




RNA

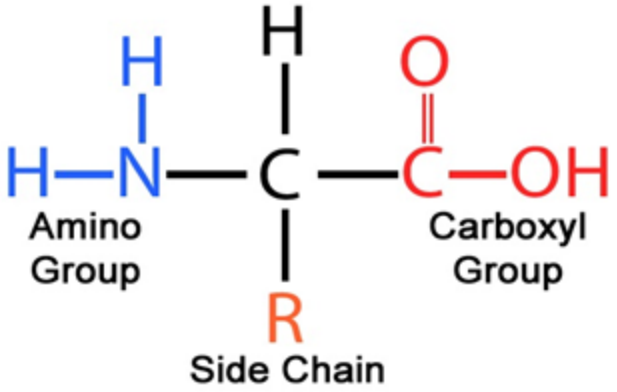

rRNA gene
Protein gene
tRNA gene
DNA
rRNA
mRNA
tRNA
AUG
Ribosomal proteins
Ribosome
AUG
Aminoacyl tRNA
Amino acids
AUG
tRNA to be recycled
Ribosomes to be recycled
UCCAUG
Protein synthesis
Newly synthesized protein
mRNA to be hydrolyzed

Protein



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

Amino Acids

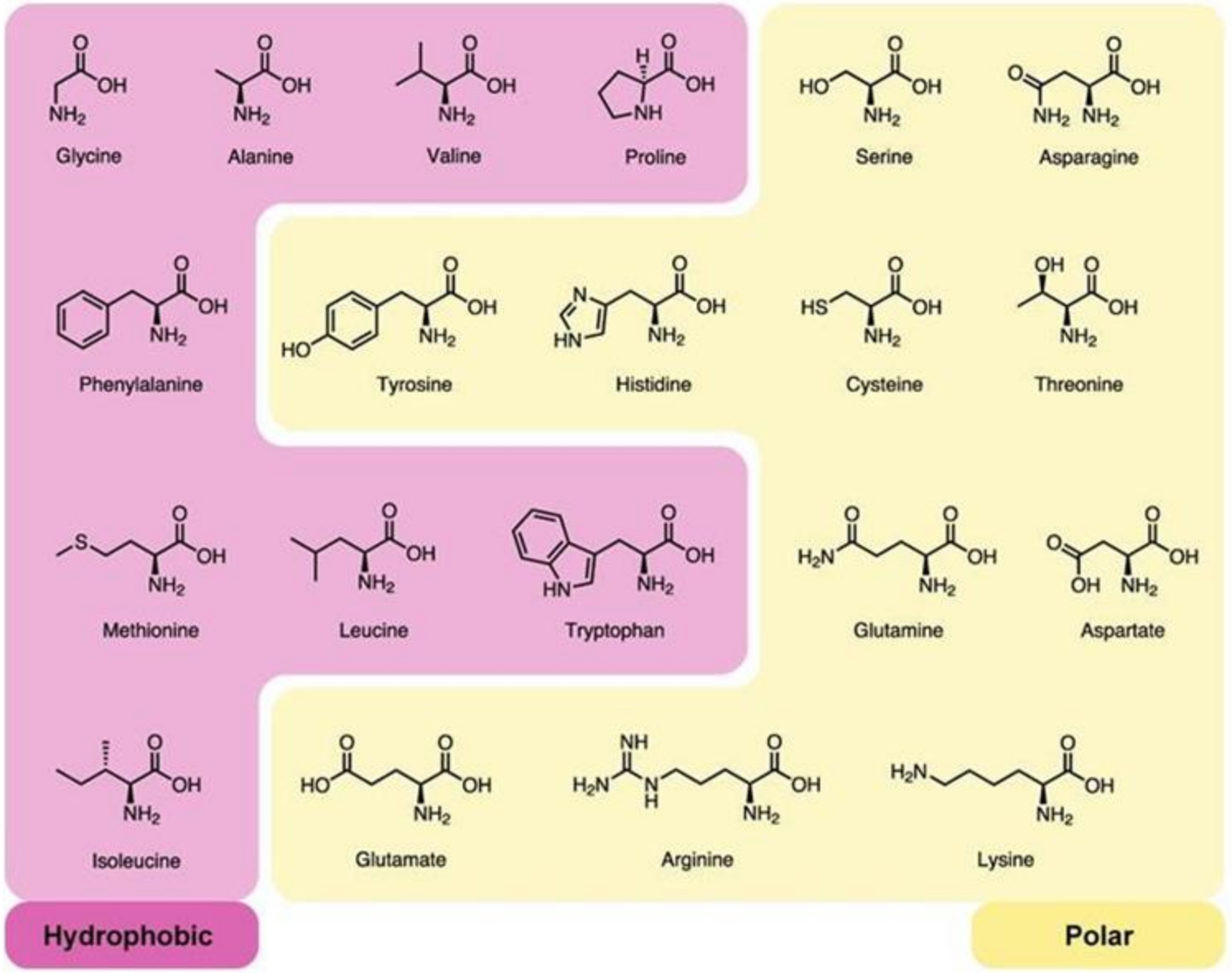


VLLPDNHYLSTQSALSKDPNEKRD
HMVLLEFVTAAGIT

Amino Acids

Amino acids could be hydrophobic or hydrophilic.

Amino acids are basic units of proteins.



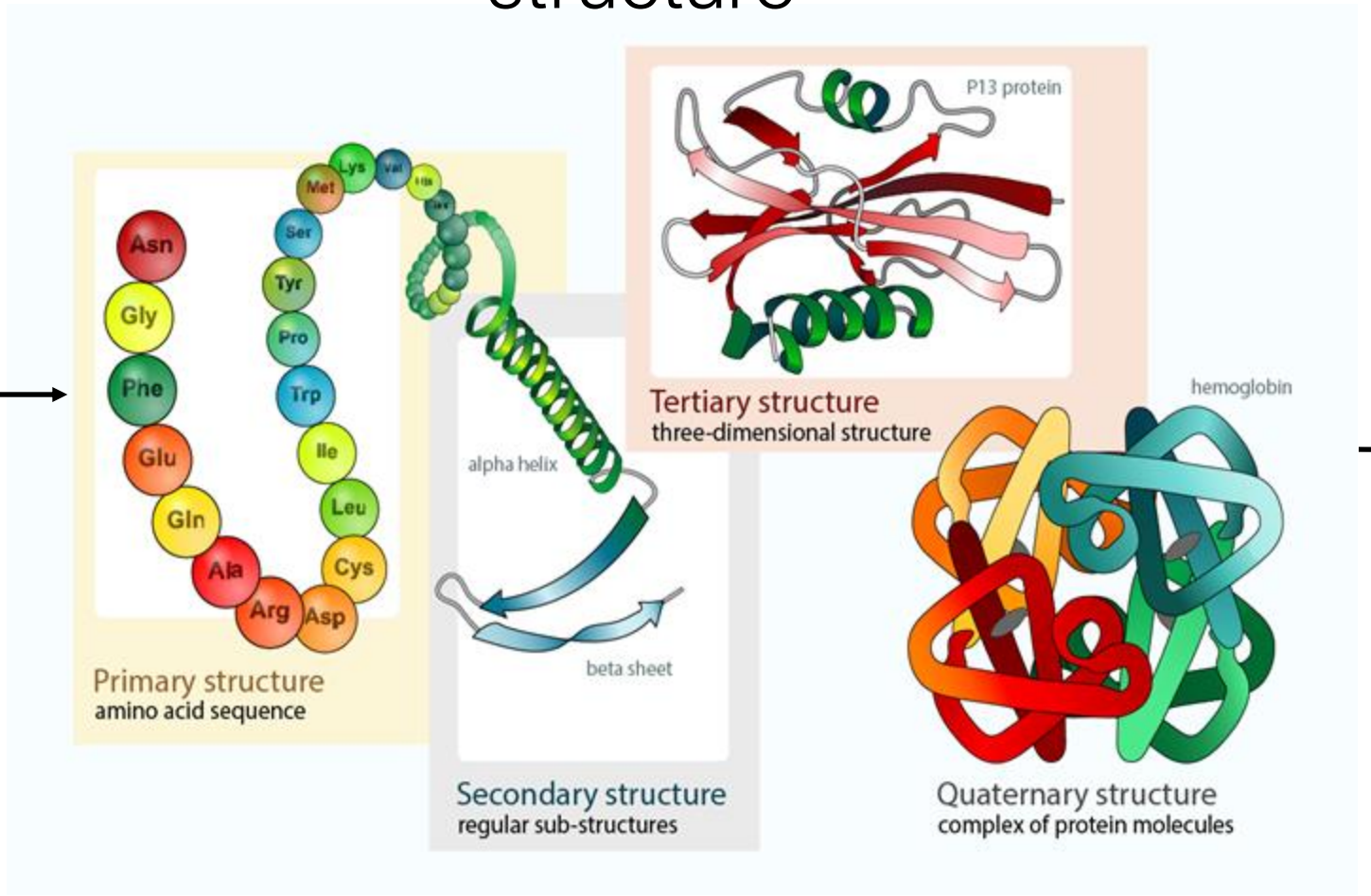
[Amino acid Table]

Protein: sequence to structure to function

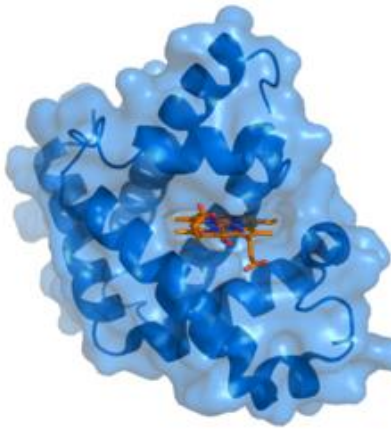
structure

sequence

VLLPDNHYLS
TQSALSKD



function



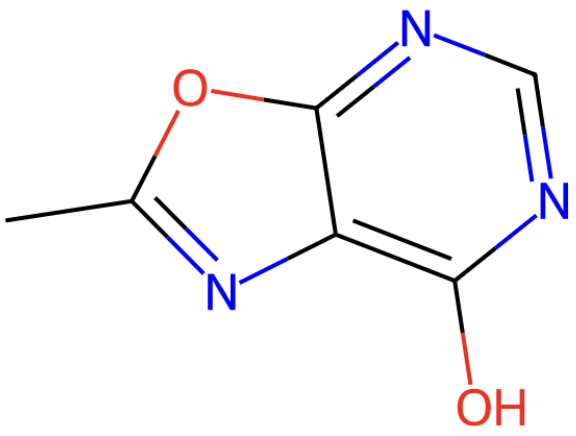
Drug Discovery as AI Tasks

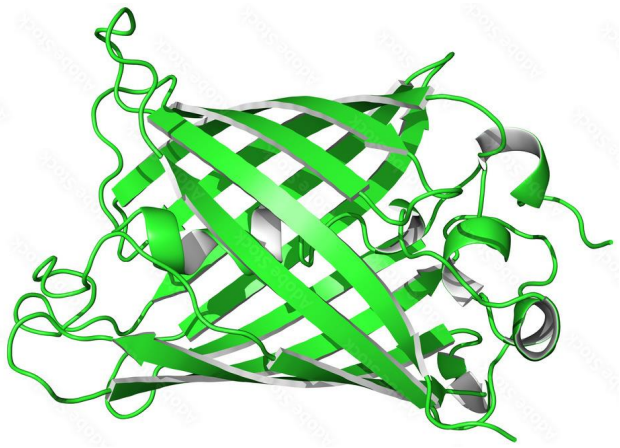
- Target Identification:
  - which gene/protein is critical on the pathway of a disease context
  - Feature/Network-based mining/ranking/prediction task
- Molecule generation
  - search or generate highly possible drug molecule leads
- Structure/Property prediction
  - predicting structures, protein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

Basic AI Modelling Tools

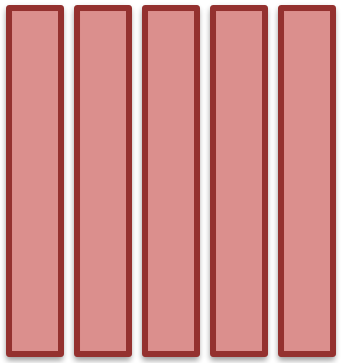
- Molecule Sequence
  - Language model: Transformer
- Molecule structure
  - Graph neural networks
  - Equivariant GNN for 3D structures
- Modeling interactions
  - molecule docking: how ligand binds to protein
  - molecular dynamics simulation

Goal of Molecule Modelling





AI Model
(Neural Net)



structure

property

binding

generation

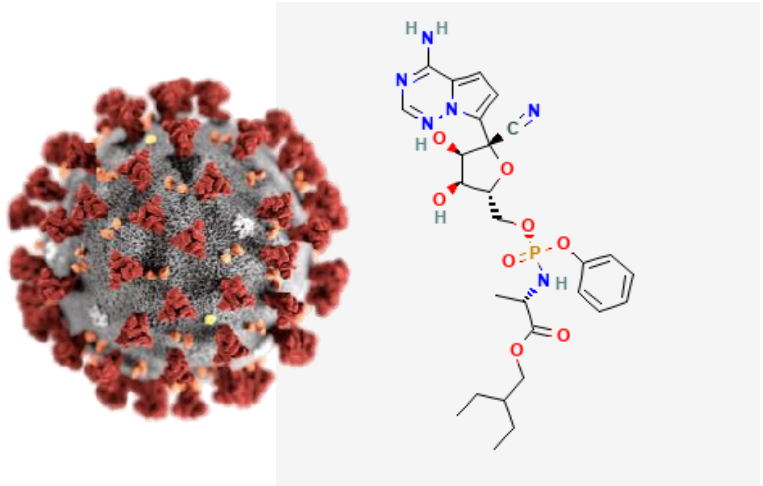
Molecule as Discrete Sequences of Tokens

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

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



Language Model

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

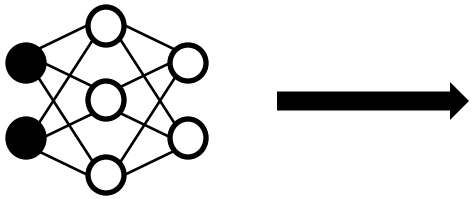
$\mathbf{Probability}$(“Pittsburgh is a city of bridges”)

$\mathbf{Probability}$(“VLLPDNHYLSTQSALSKDPN”)

Probability Model for Next Token

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

Santa Barbara has very nice ____



beach	0.5
weather	0.4
snow	0.01

Pittsburgh is a city of ____

bridges	0.6
corn	0.02

Mathematics of Language Model

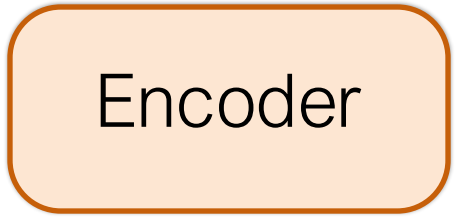
Probability("Pittsburgh is a city of bridges")
$= P(\text{“}Pittsburgh\text{”}) \cdot P(\text{“}is\text{”}|\text{“}Pittsburgh\text{”})$
$\cdot P(\text{“}a\text{”}|\text{”}Pittsburgh\ is\text{”}) \cdot P(\text{“}city\text{”}|\ldots) \cdot P(\text{“}of\text{”}|\ldots)$
$\cdot P(\text{“}bridges\text{”}|\ldots)$

$$\text{Prob.}(x_{1..T}) = \prod_{t=1}^{T} P(x_{t+1}|x_{1..t})$$

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

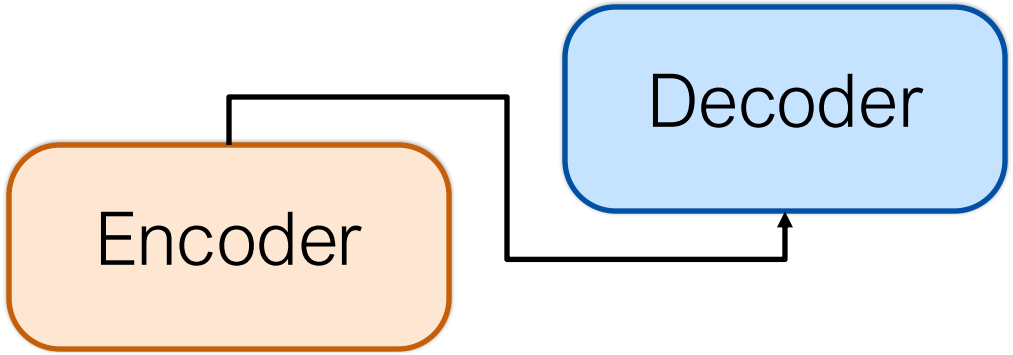
Type of Language Models

Encoder-only
Masked LM
Non-autoregressive



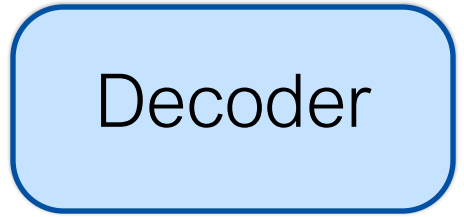
e.g. BERT
RoBERTa
ESM (for protein)

Encoder-decoder



e.g. T5

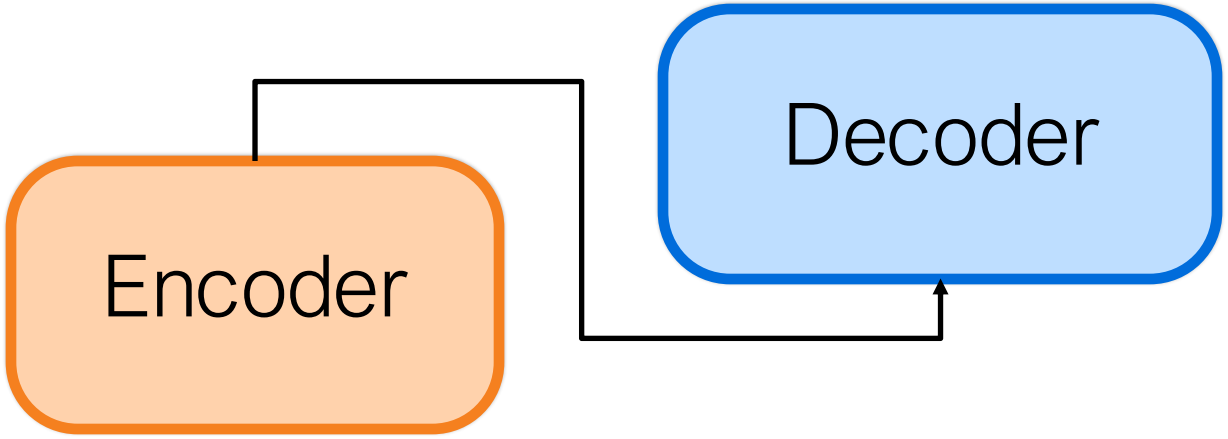
Decoder-only
Autoregressive



e.g. GPT
LLaMA
ProGen (for protein)

Encoder-Decoder Paradigm

target:
I like singing and dancing.



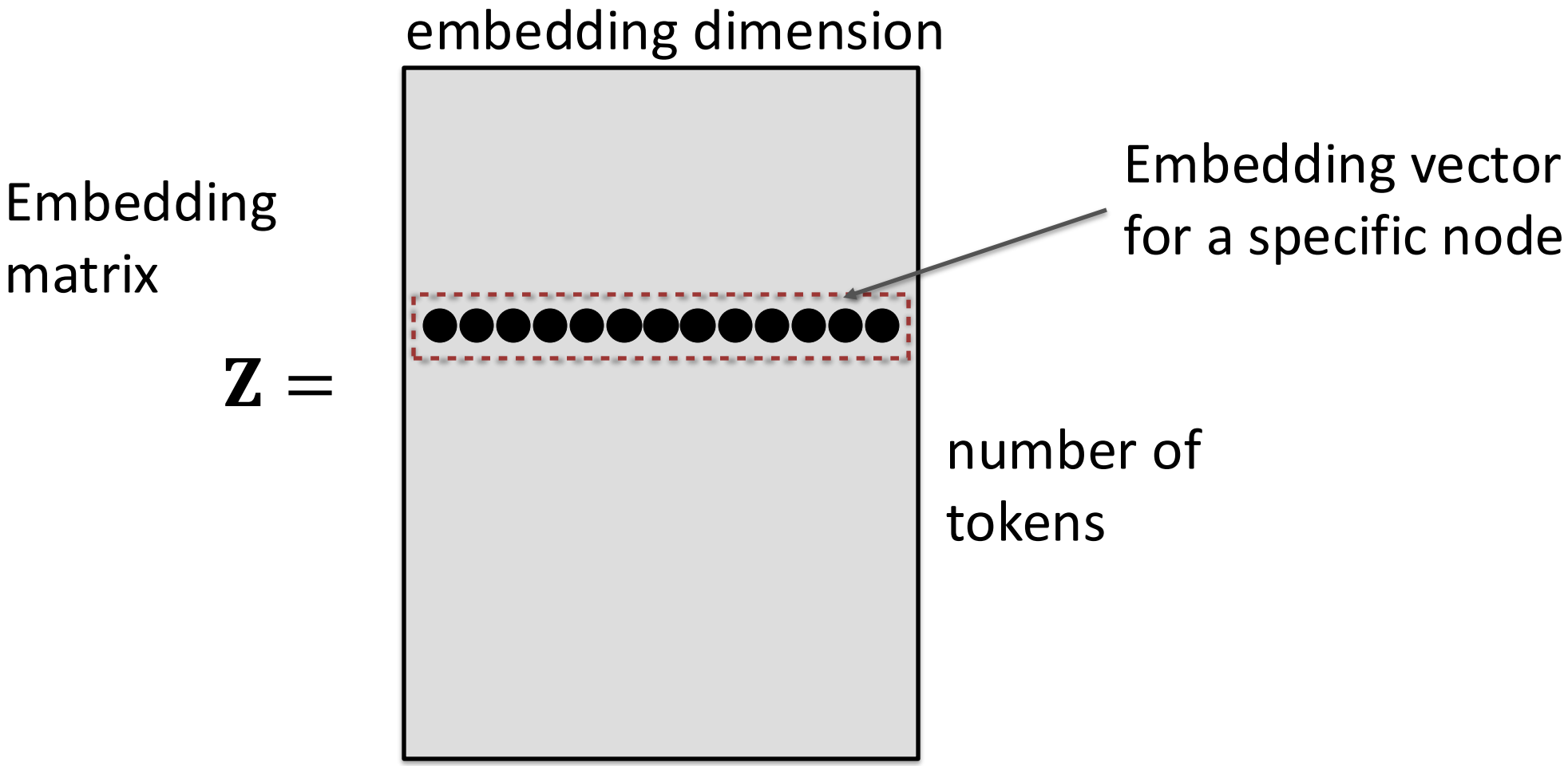
Source: 我喜欢唱歌和跳舞。

$$p_\theta(y|x) = \prod_i p(y_i|x, y_{1:i-1})$$

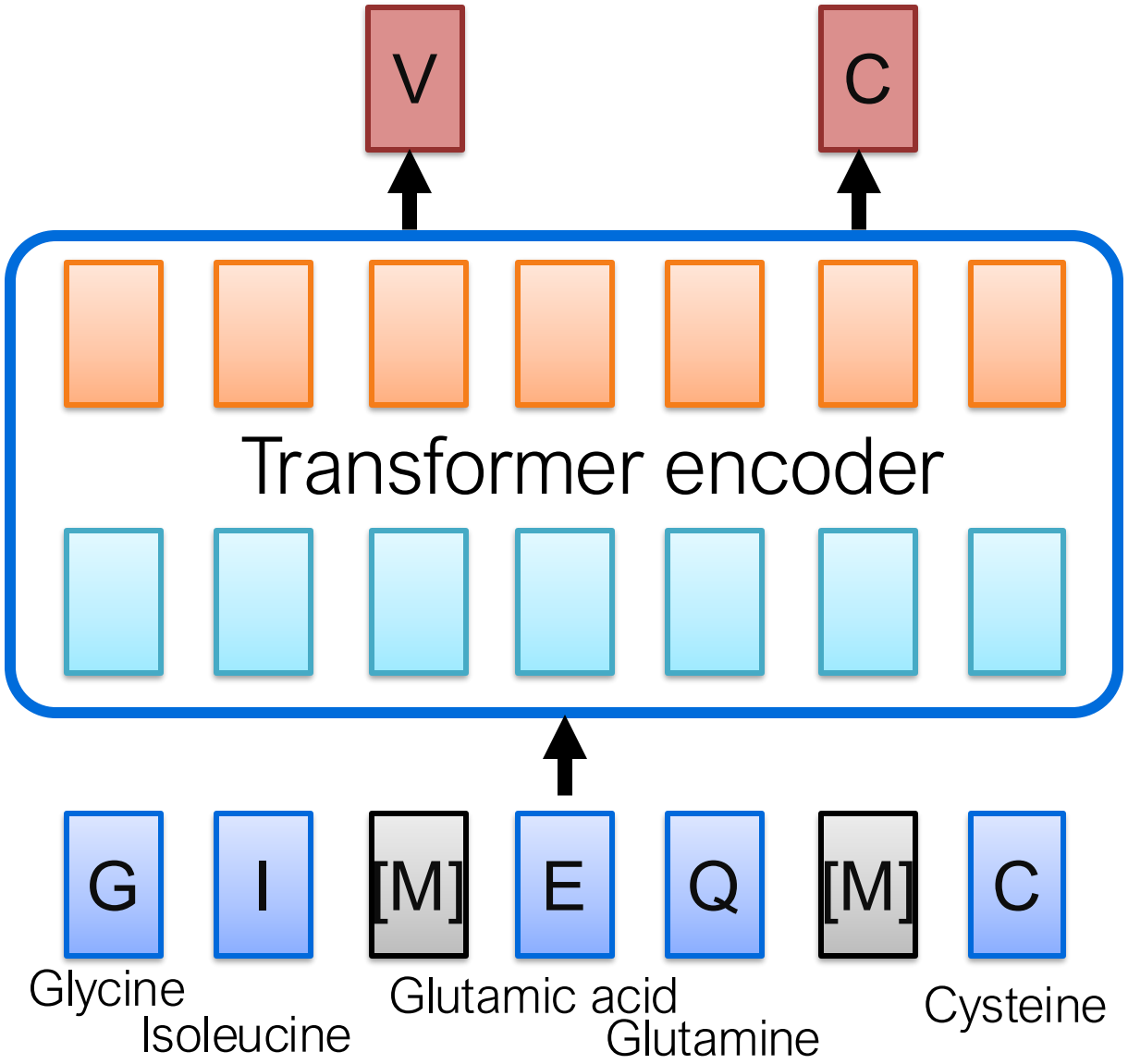
conditional prob. modeled by neural networks (Transformer)

Mapping Token to Embedding

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

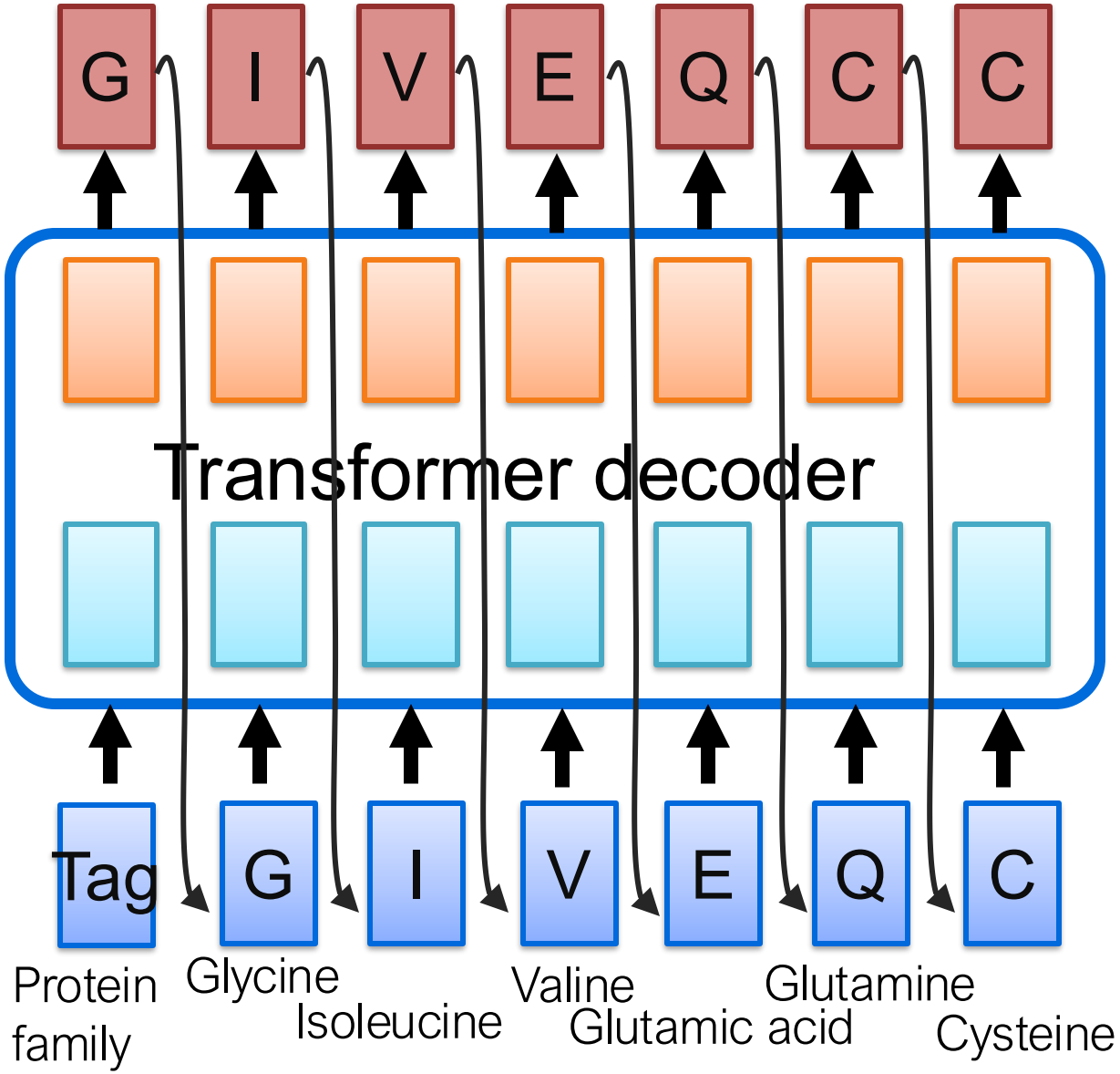


Protein Language Model 1: Mask LM



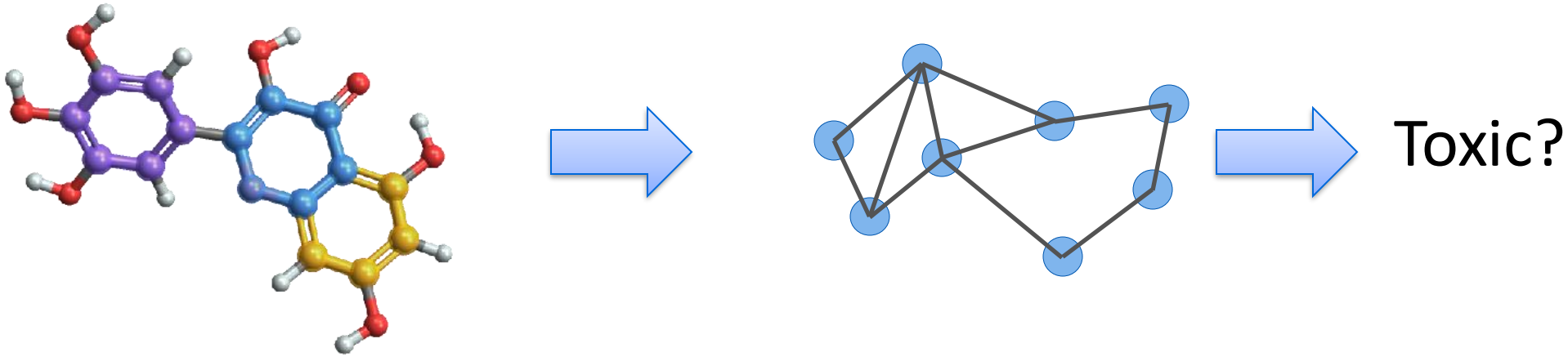
- Using raw protein sequences for pre-training
  - Training loss: predicting masked residues
- ESM [Meier et al 2021] and ESM-2 [Lin et al 2023]

Protein Language Model 2: Casual LM

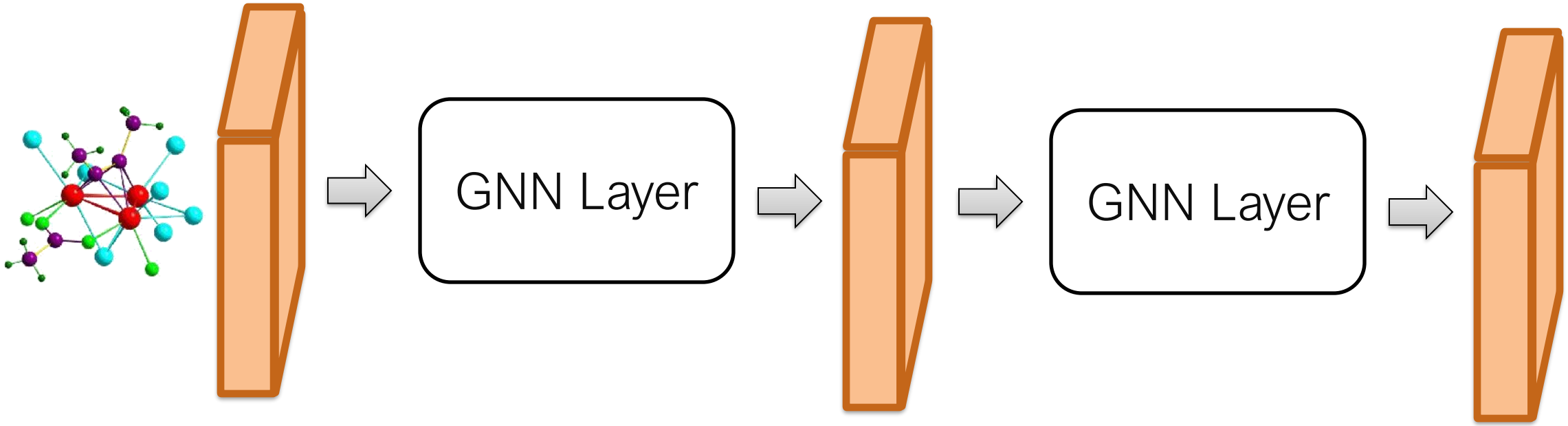


- Using raw protein sequences and their category tags for pre-training
  - training loss: predicting next residue
- ProGen [Madani et al 2023] and ProGen2 [Nijkamp et al 2023]
- Protein Tag is insufficient!

Molecule as a Graph

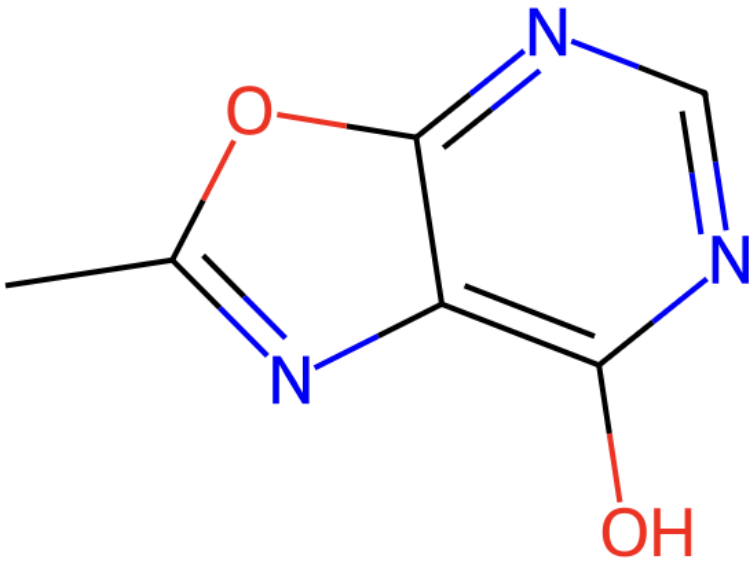


Graph Neural Network



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

Graph Neural Network (MPNN) to model molecule graphs



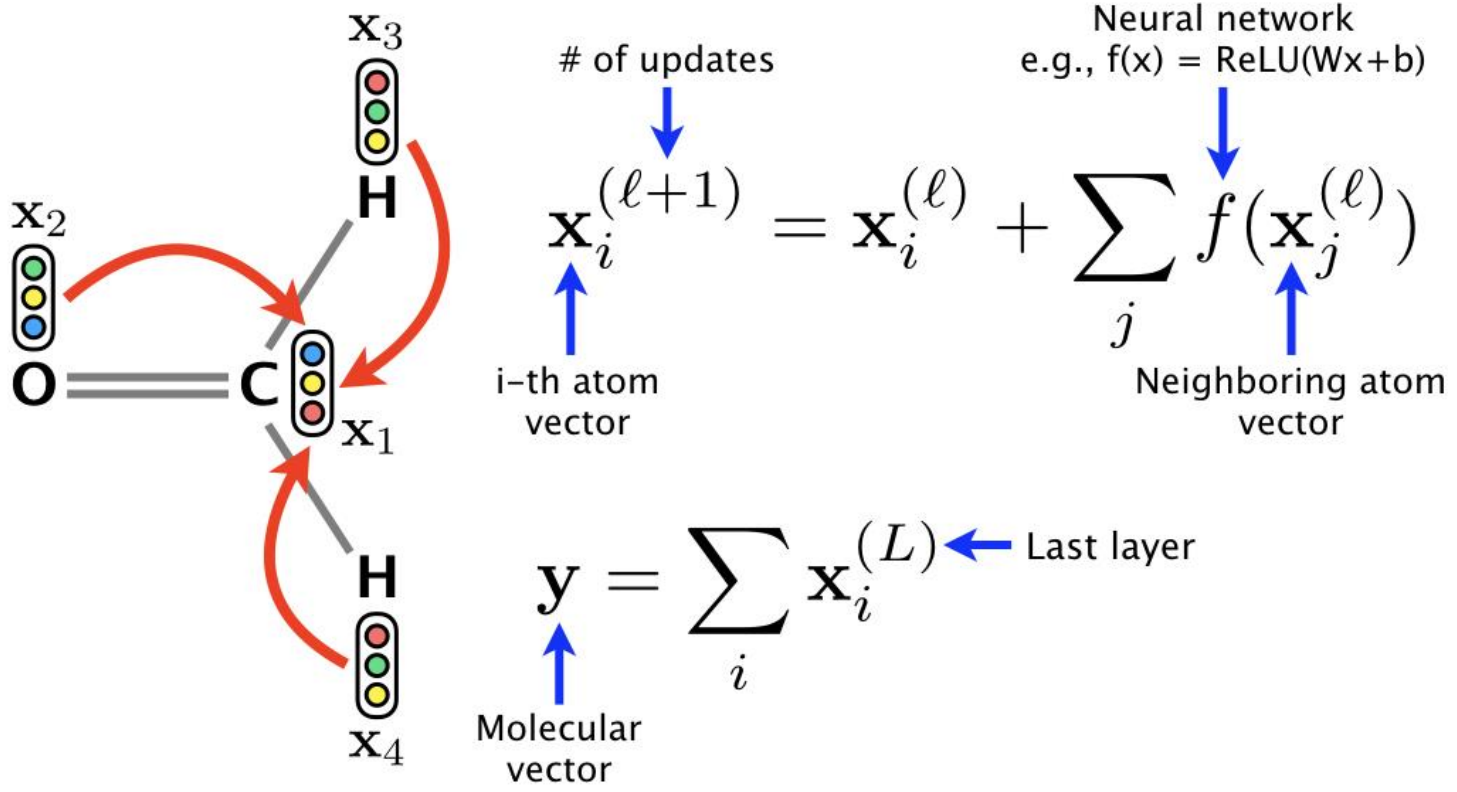


Fig.1: The update function (or called transition, propagation, message passing, and convolution) in GNNs. On a molecular graph, the **GNN updates each atom vector with its neighboring atom vectors non-linear transformed by neural network.** The molecular vector is obtained by summing (or mean) the atom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

[Tsubaki et al, 2018.]

Graph of Fragments

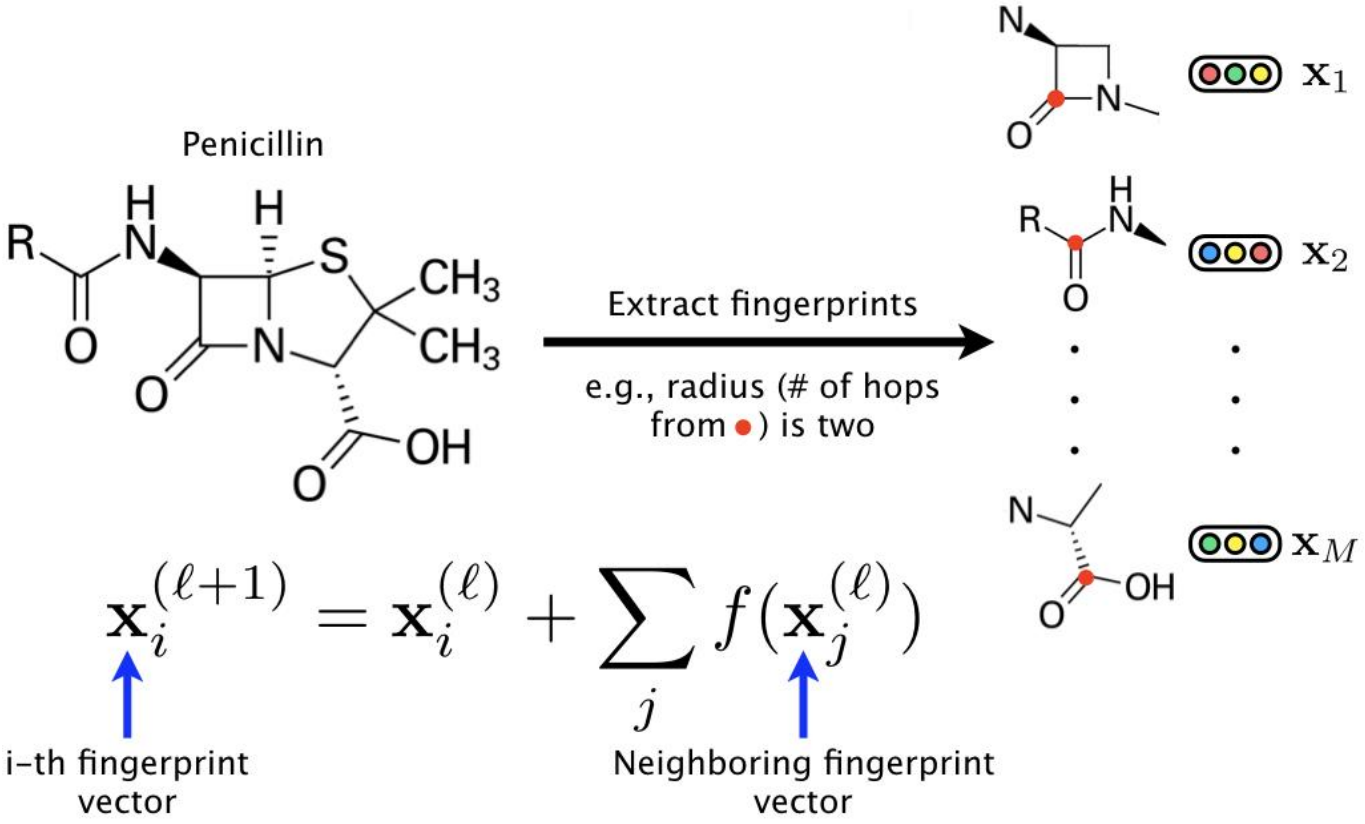
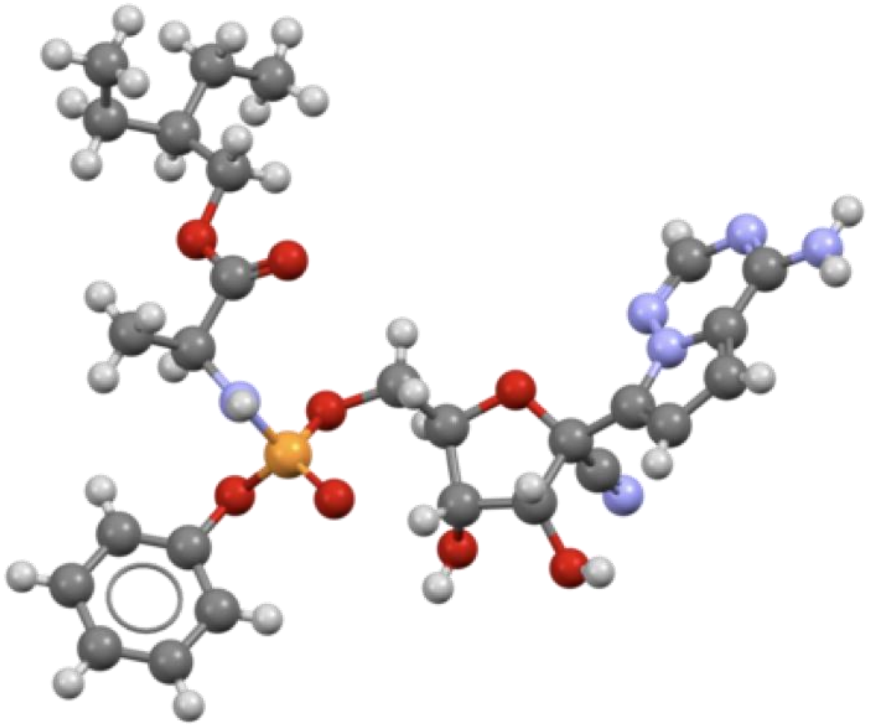


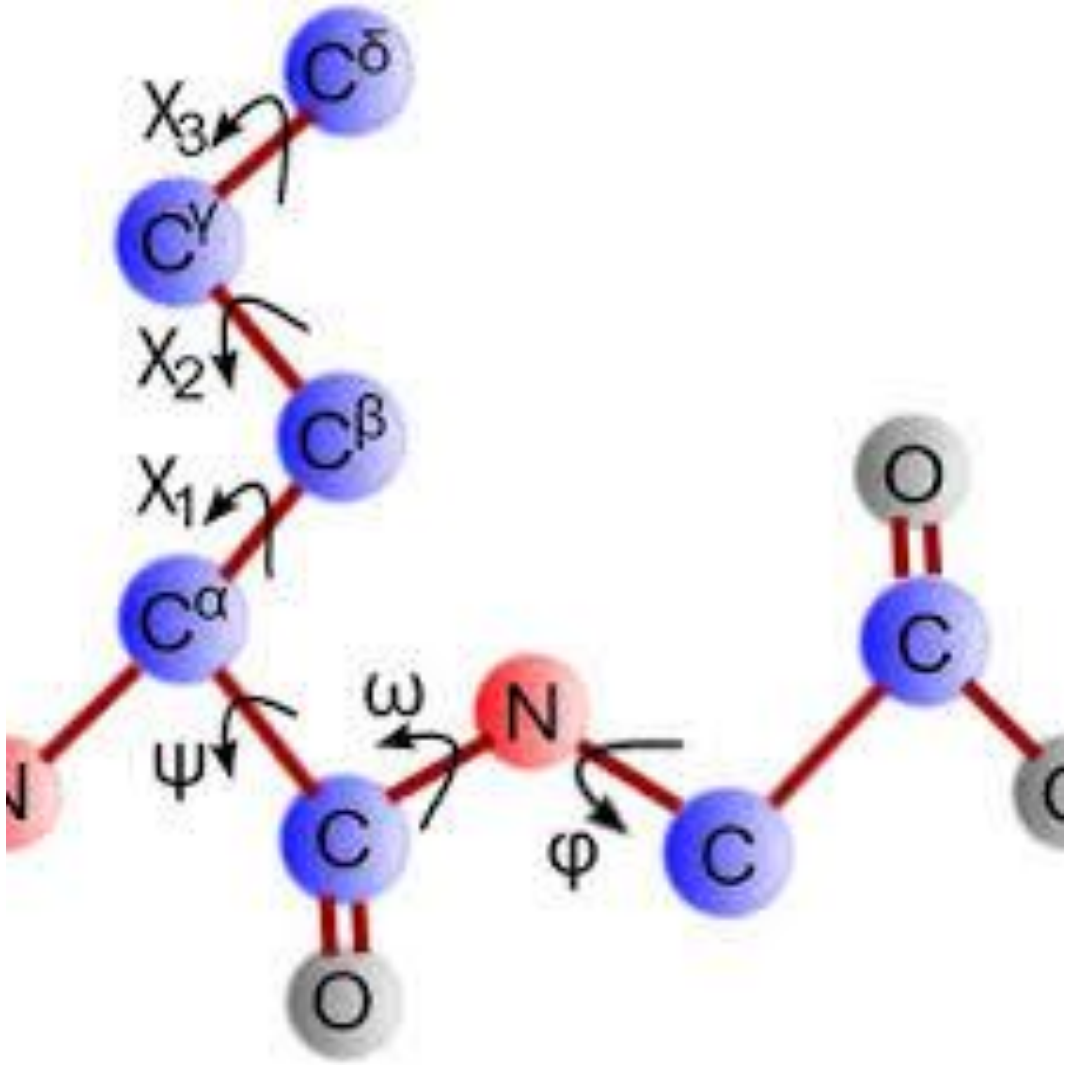
Fig.2: The update function based on radius-based subgraphs, i.e., molecular fingerprints. **Each fingerprint is initialized with a random vector**. The following procedure is the same as that of basic GNN.

[Tsubaki et al, 2018.]

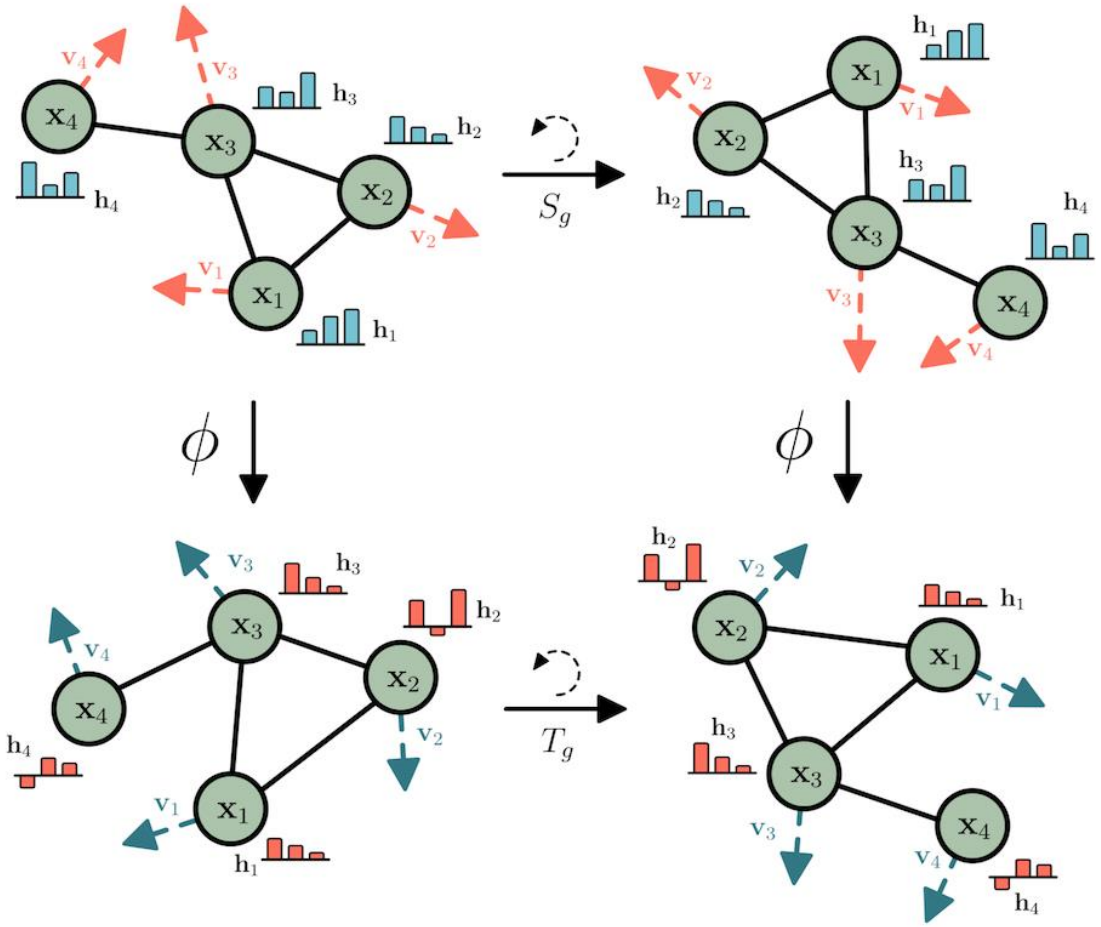
Modelling 3D Structure of Molecules

- 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

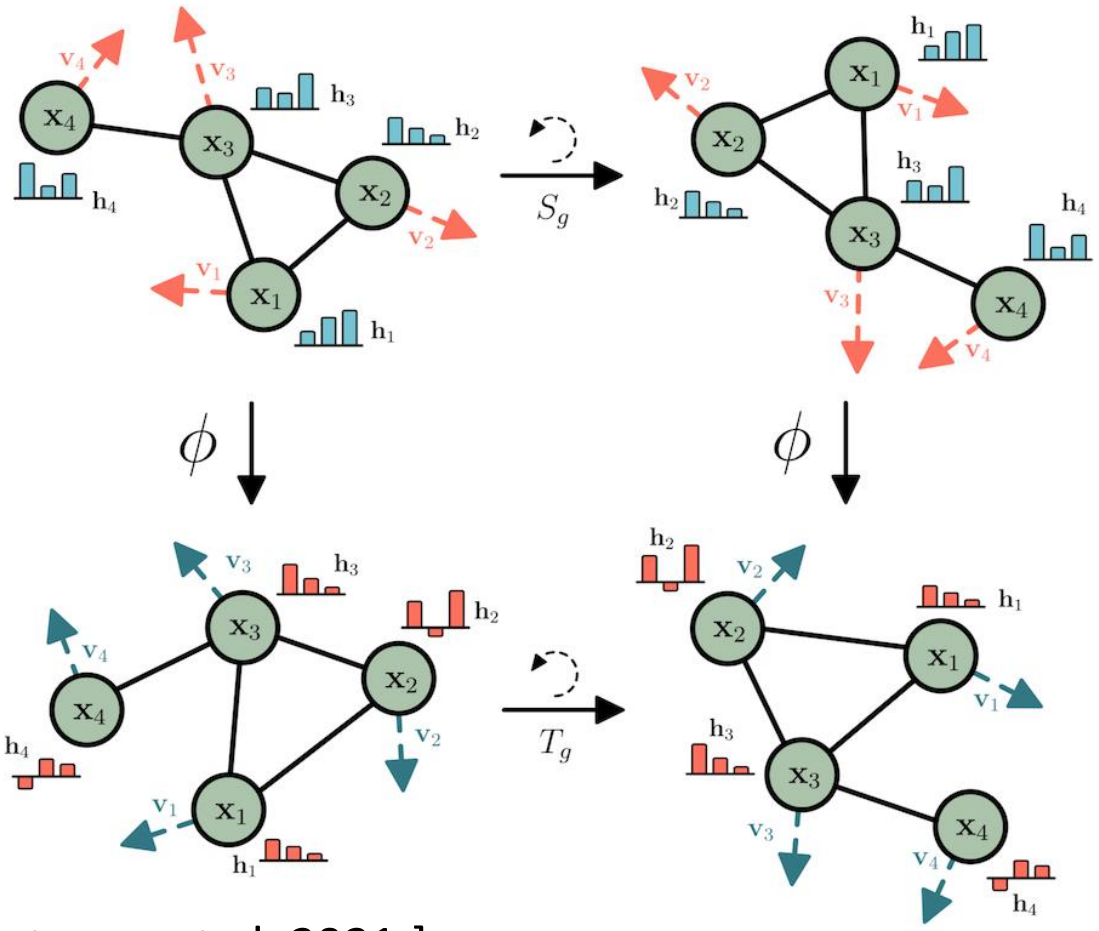
g is the rotation/translation

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

[Satorres et al, 2021.]

Equivariant Graph Neural Network (EGNN)

- model the 3D geometry of a molecule



[Satorres et al, 2021.]

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

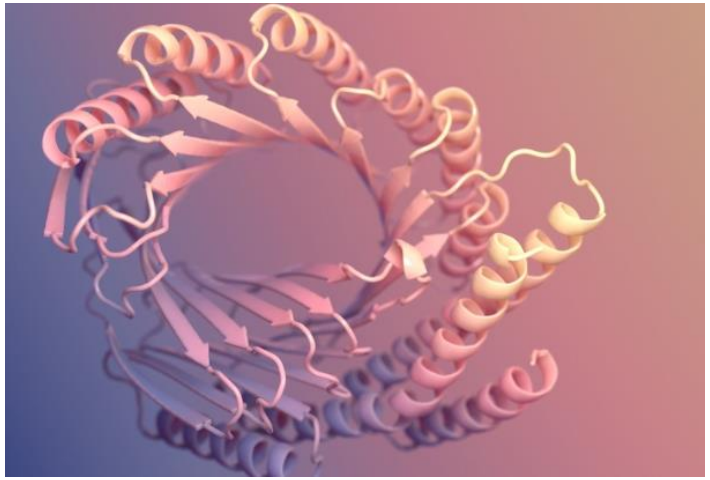
$$m_{j\to i} = M^k\left(h_i^k, h_j^k, d_{ij}\right)$$

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

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

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

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

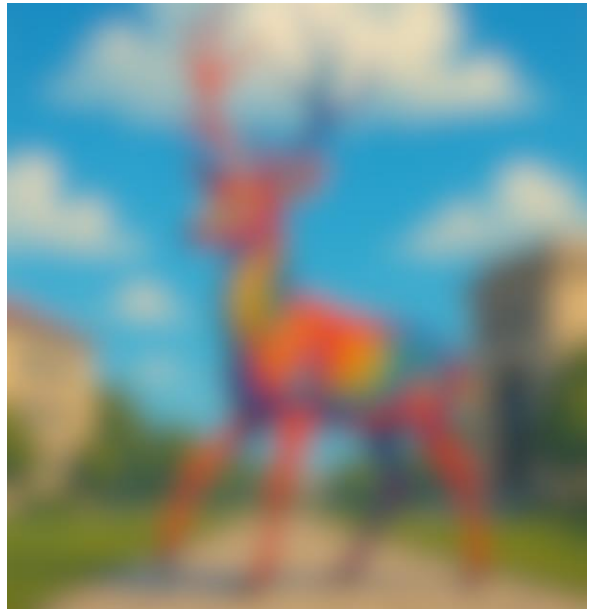


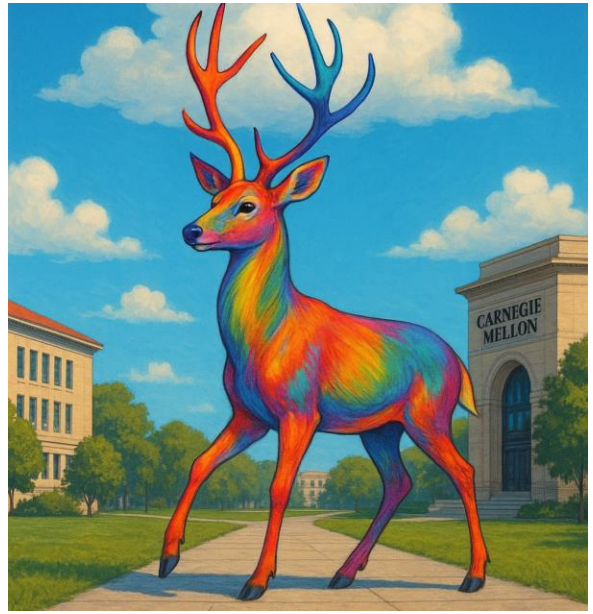
Probabilistic Generative Process

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 $\mathrm{x_t} \sim q\left(x_t | x_{t-1}\right) = N(\sqrt{\alpha_t} x_{t-1}, \beta_t I)$, $\alpha_t = 1 - \beta_t$

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

t=0 t=1 t=T

(Reverse) Generation/Denoising Process

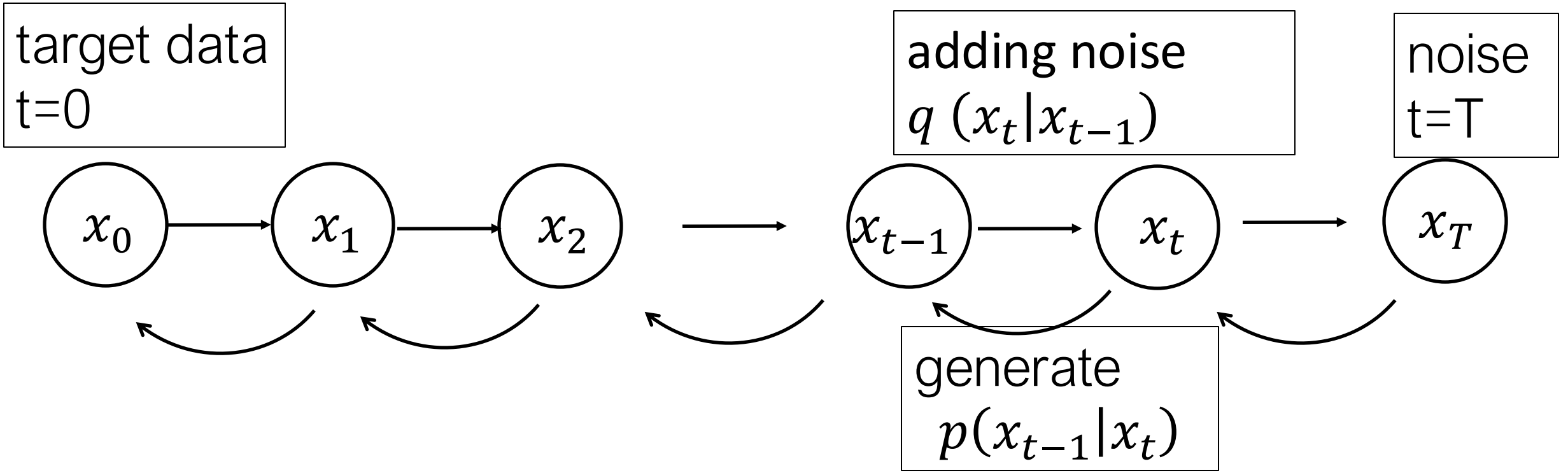
adding noise $q\ (x_t|x_{t-1})$

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

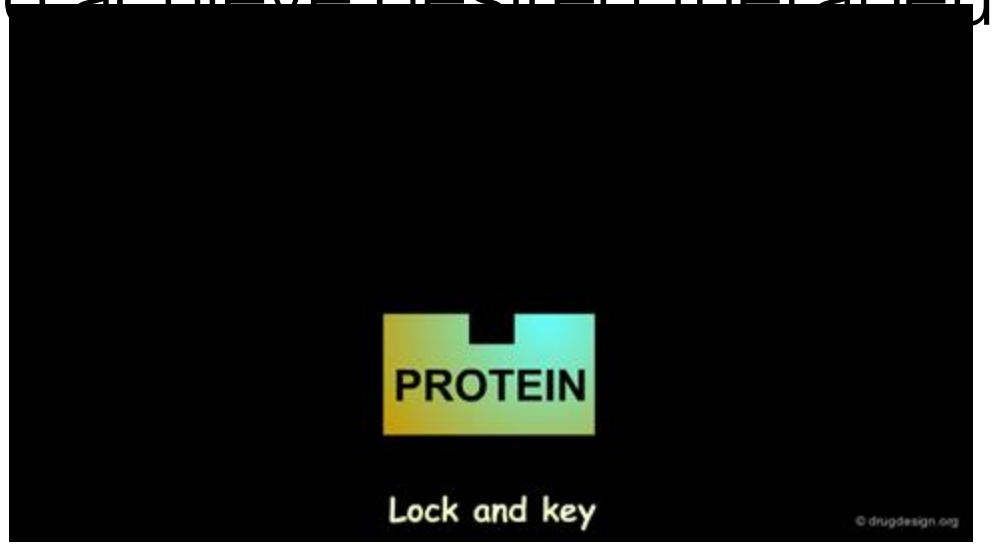
- Multi-scale Biological System and Drugs
  - molecule, cell, tissue, organs, cause of disease
- Basic AI Models for Biomecules
  - sequence, structure, generative model
- MARS: finding small molecule drugs with multiple properties
- EnzyGen: A general generative model for enzyme design
- PPDiff: protein-binding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

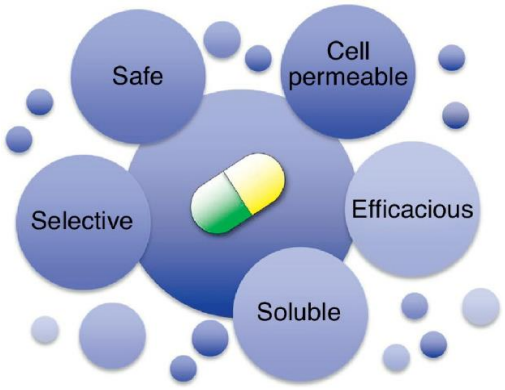
Key Terminology for Drug Design

- **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.

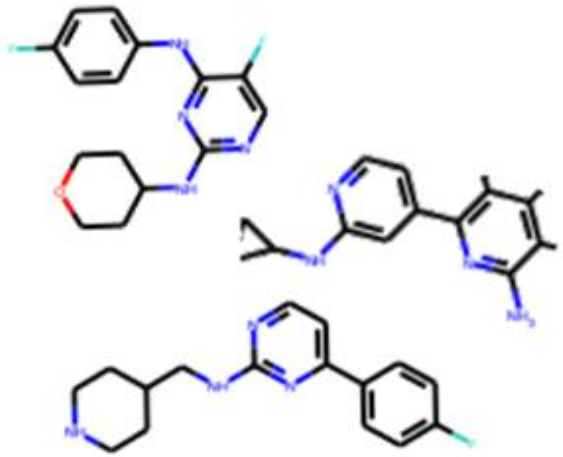


Designing Small-molecule Drugs needs to Meet Multiple Objectives

Satisfy multiple properties with high scores



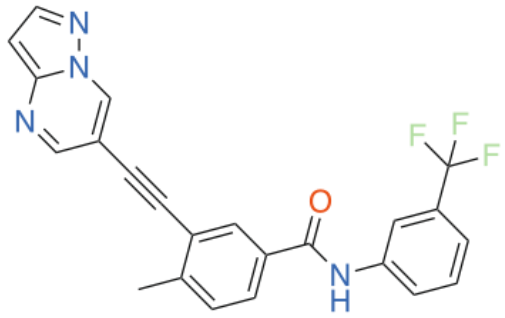
Produce diverse and novel molecules



scarce lab measured data

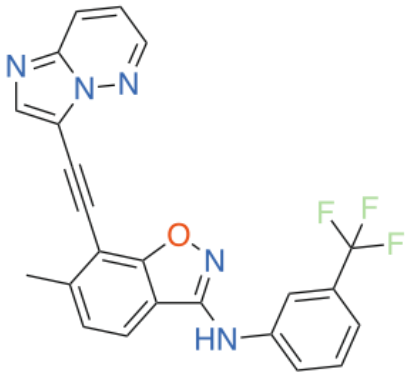
Finding Novel Drugs is Important

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



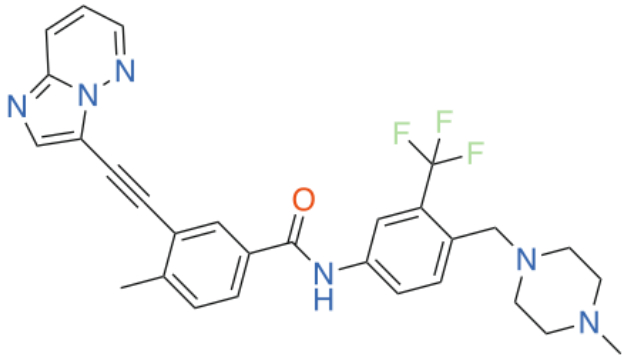
Gao et al.
Compound 7r
6 nM

Previously Reported Drug Candidate



Zhavoronkov et al.
Compound 1
10 nM

AI Designed Drug Candidate

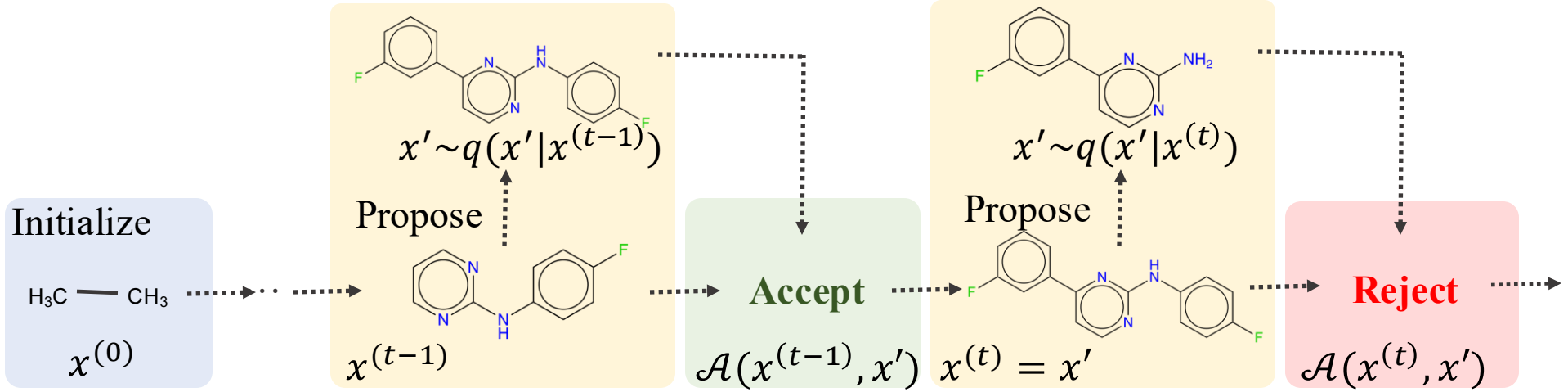


Ponatinib
9 nM

FDA Approved Drug

Walters et al. Nature Biotechnology ***38****, 143–145(2020)*

MARS: Iterative Graph Editing for Drug Generation

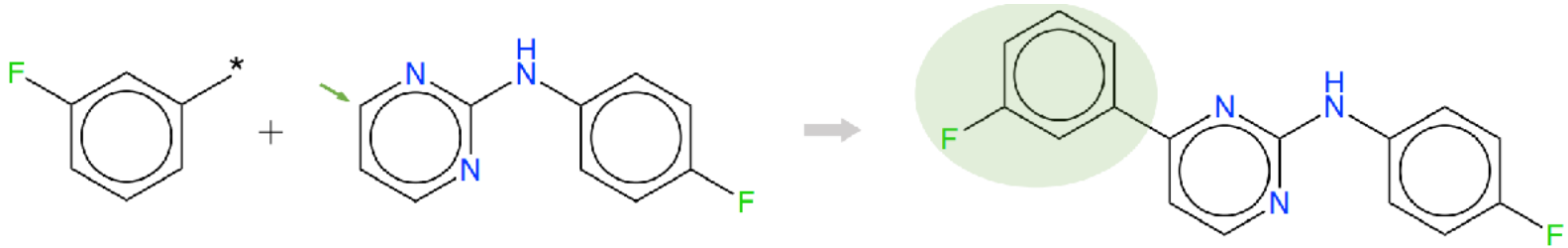


- Start from initial molecule (e.g. C2H6)
- For each step, propose a new molecule x' by modifying existing one, $x' \sim q(x'|x^{(t-1)})$
- Accept wrt ratio: $$\mathcal{A}(x, x') = min\left\{1, \frac{\pi^{\alpha}(x')q(x|x')}{\pi^{\alpha}(x)q(x'|x)}\right\}$$

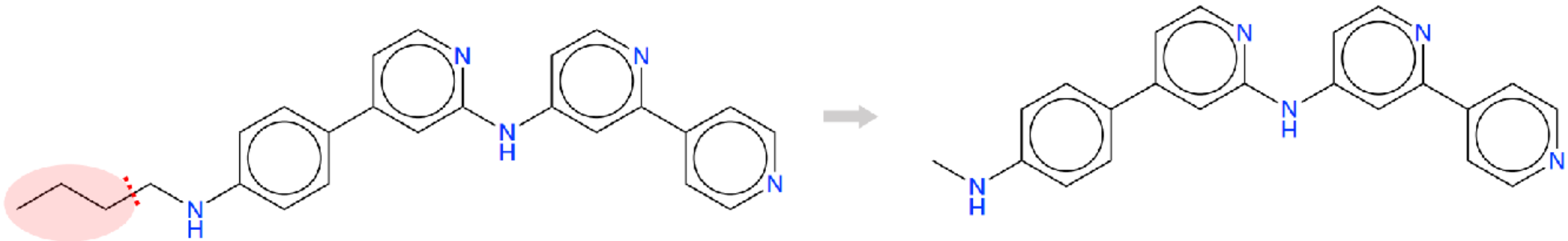
Xie et al. MARS: Markov Molecular Sampling for Multi-objective Drug Discovery. ICLR 2021.

Molecular Graph Editing

- Adding fragment

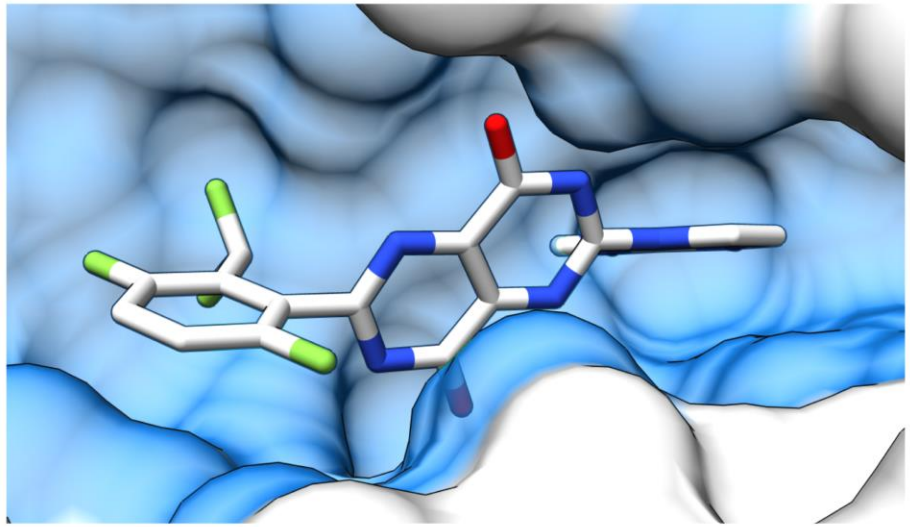


- Deleting fragment

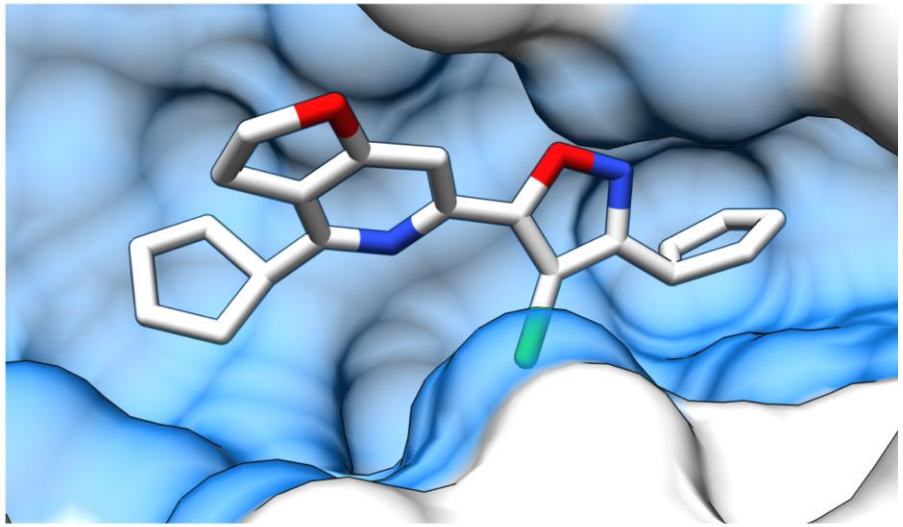


Xie et al. MARS: Markov Molecular Sampling for Multi-objective Drug Discovery. ICLR 2021.

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

Xie et al. MARS: Markov Molecular Sampling for Multi-objective Drug Discovery. ICLR 2021.

Highlights of MARS

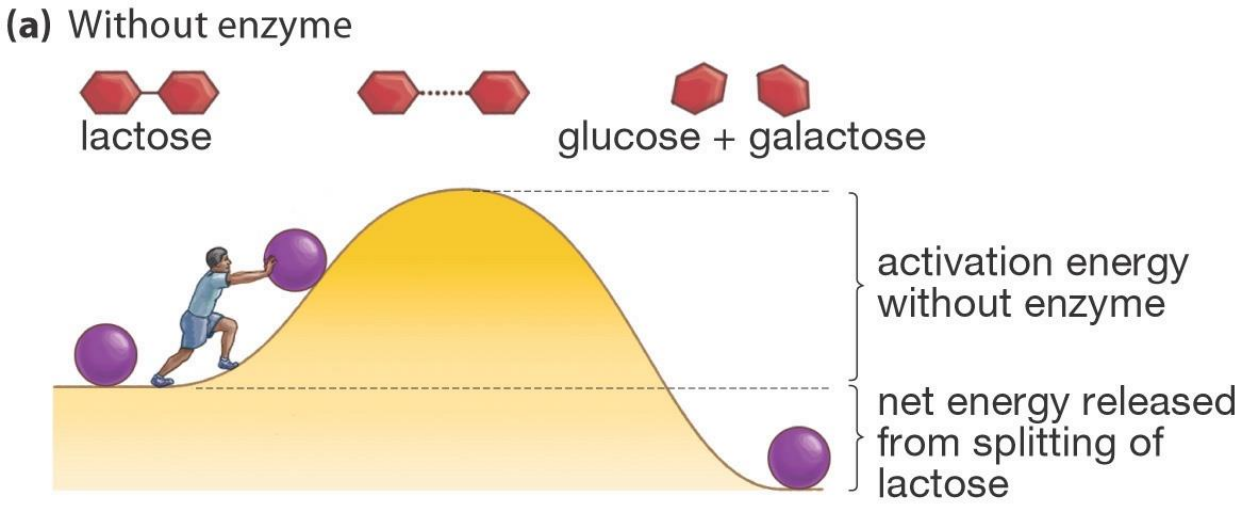
- Target-based molecule generation
- MARS, a simple yet flexible framework for multi-objective
  - Based on MCMC sampling
  - Self-adaptive proposal trained on the fly => no need for data
  - Generates better molecules and explores lager chemical space
- => can discover novel and diverse drug-like molecules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

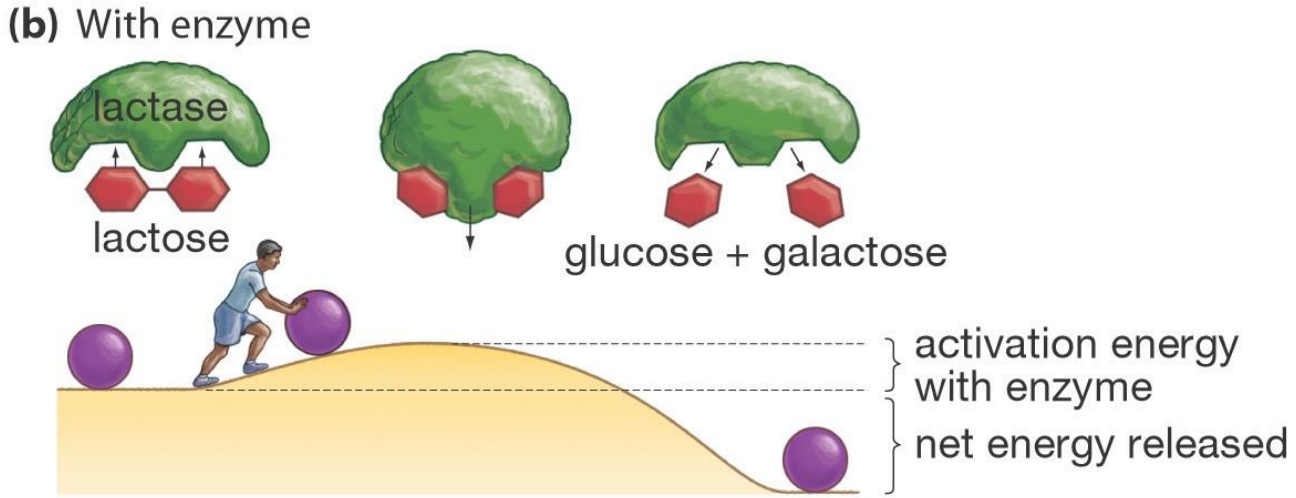
Outline

- 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

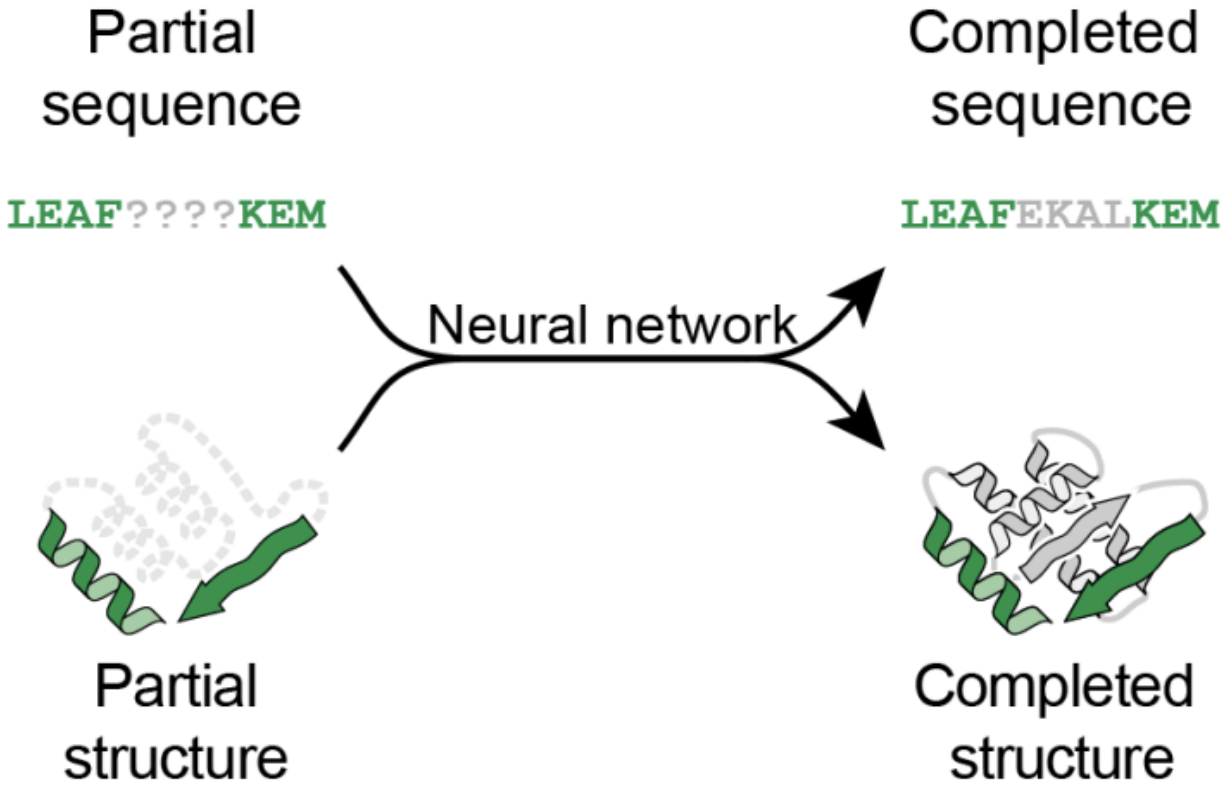
- biological catalyst to accelerate chemical reactions
  - Enzymes reduce a reaction's activation energy





Motivation 1: How to design desired enzymes?

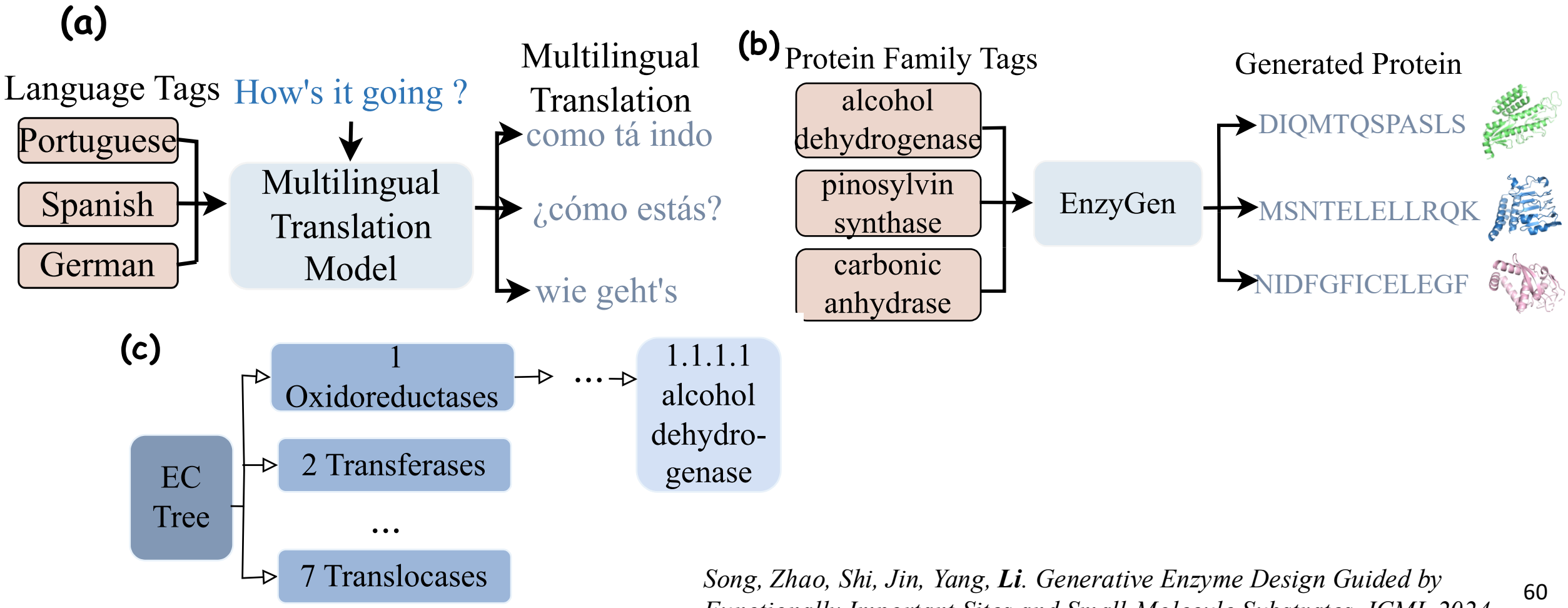
- Functional Important Sites (Motif)
  - Active sites – Binding to substrates



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

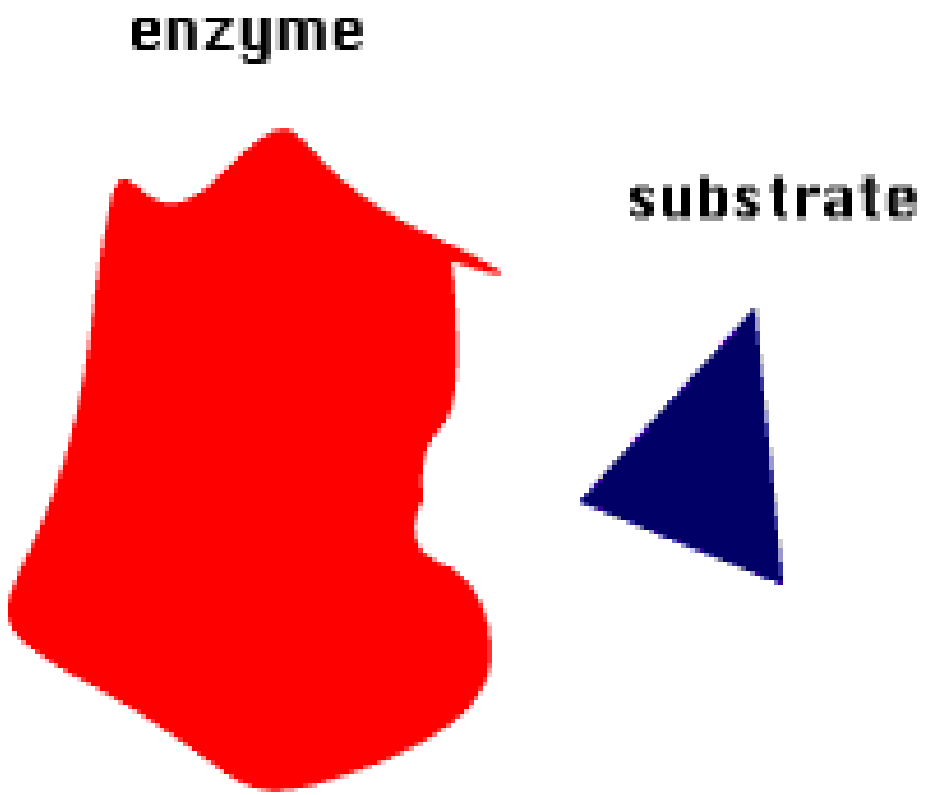


*Song, Zhao, Shi, Jin, Yang, **Li**. Generative Enzyme Design Guided by Functionally Important Sites and Small-Molecule Substrates. ICML 2024.*

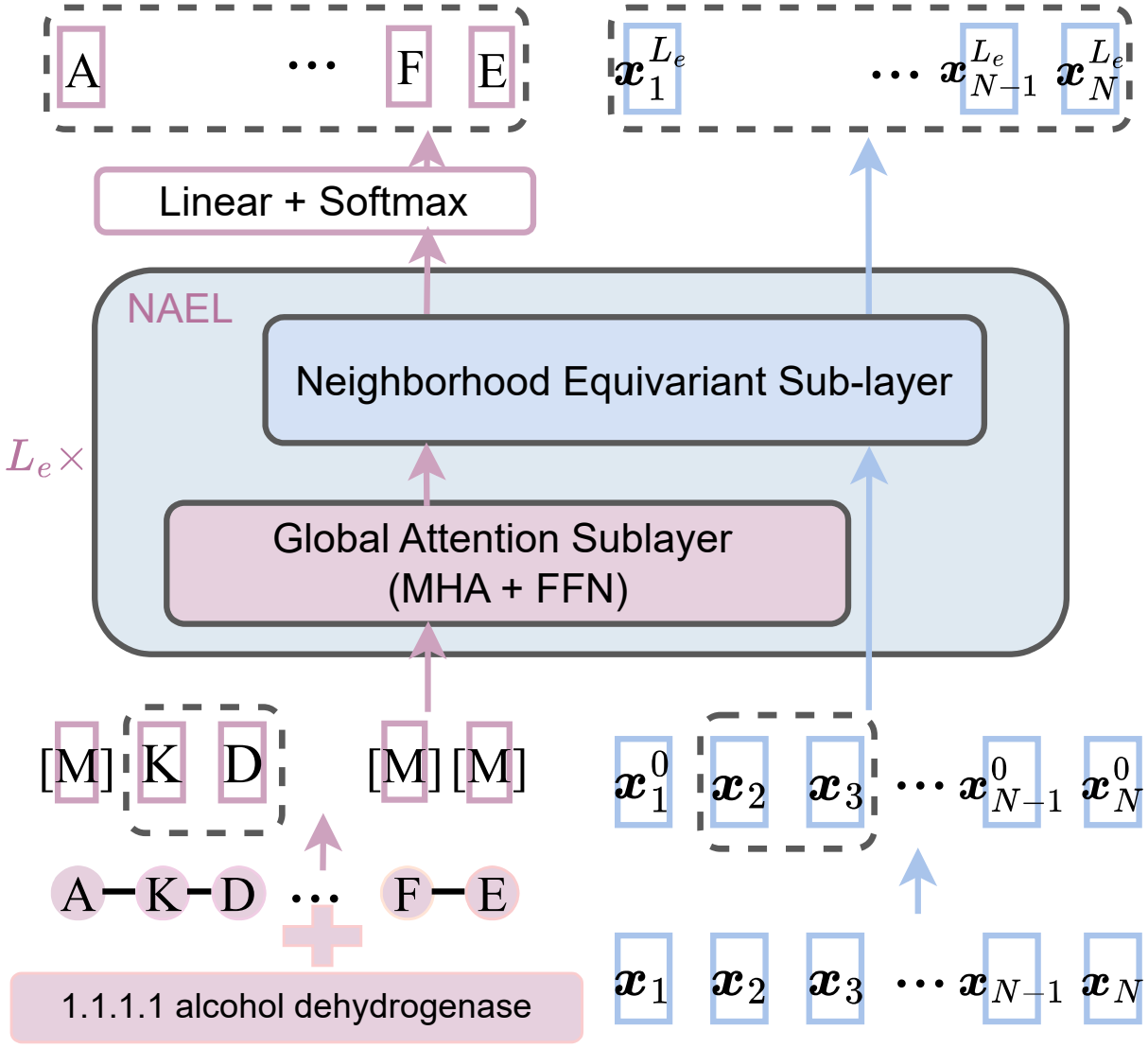
Motivation 3: How to design desired enzymes?

- Substrate Specificity:

Different enzymes binding to specific substrates to speedup enzymatic reactions



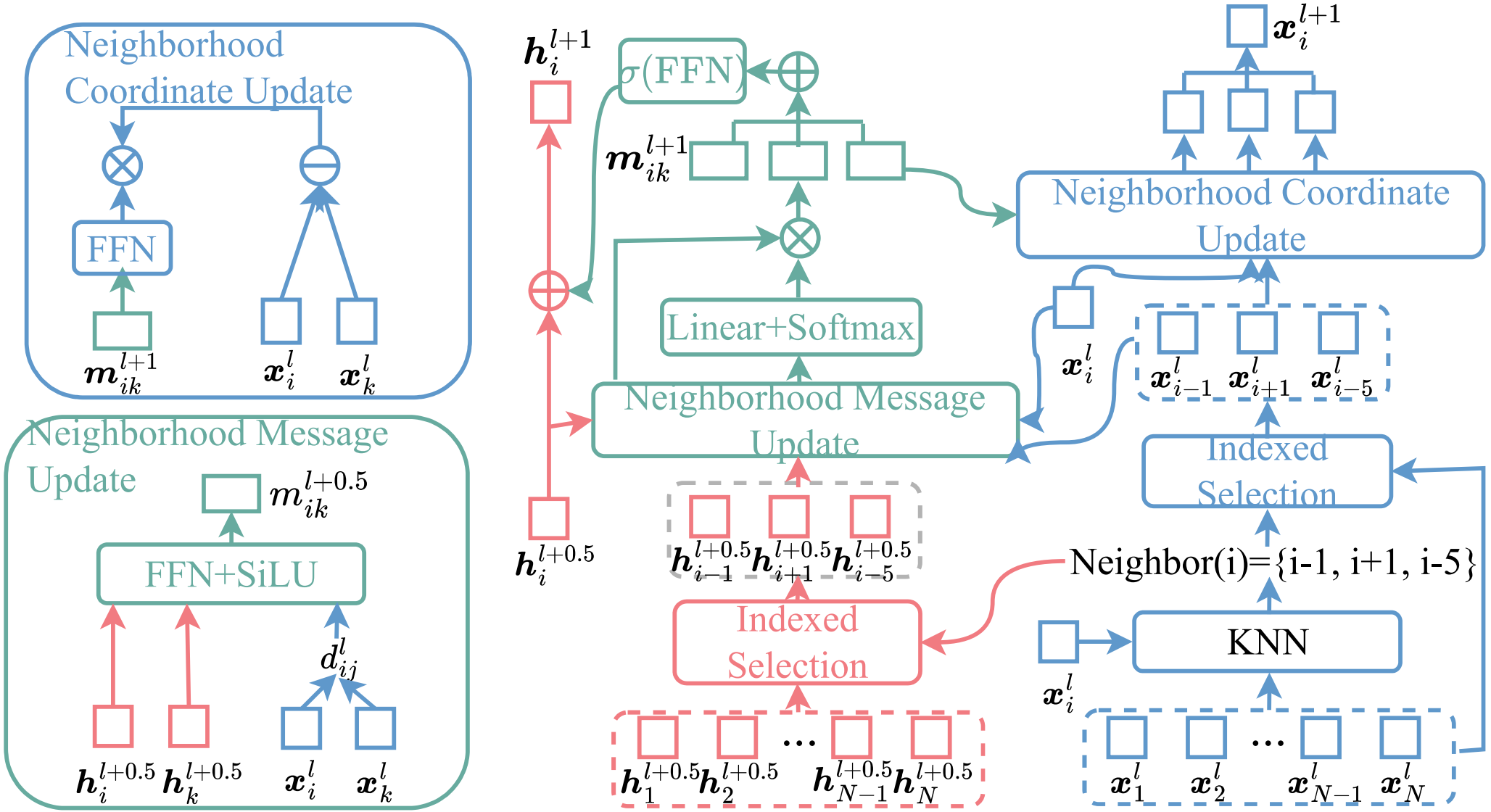
EnzyGen Model – NAEL backbone



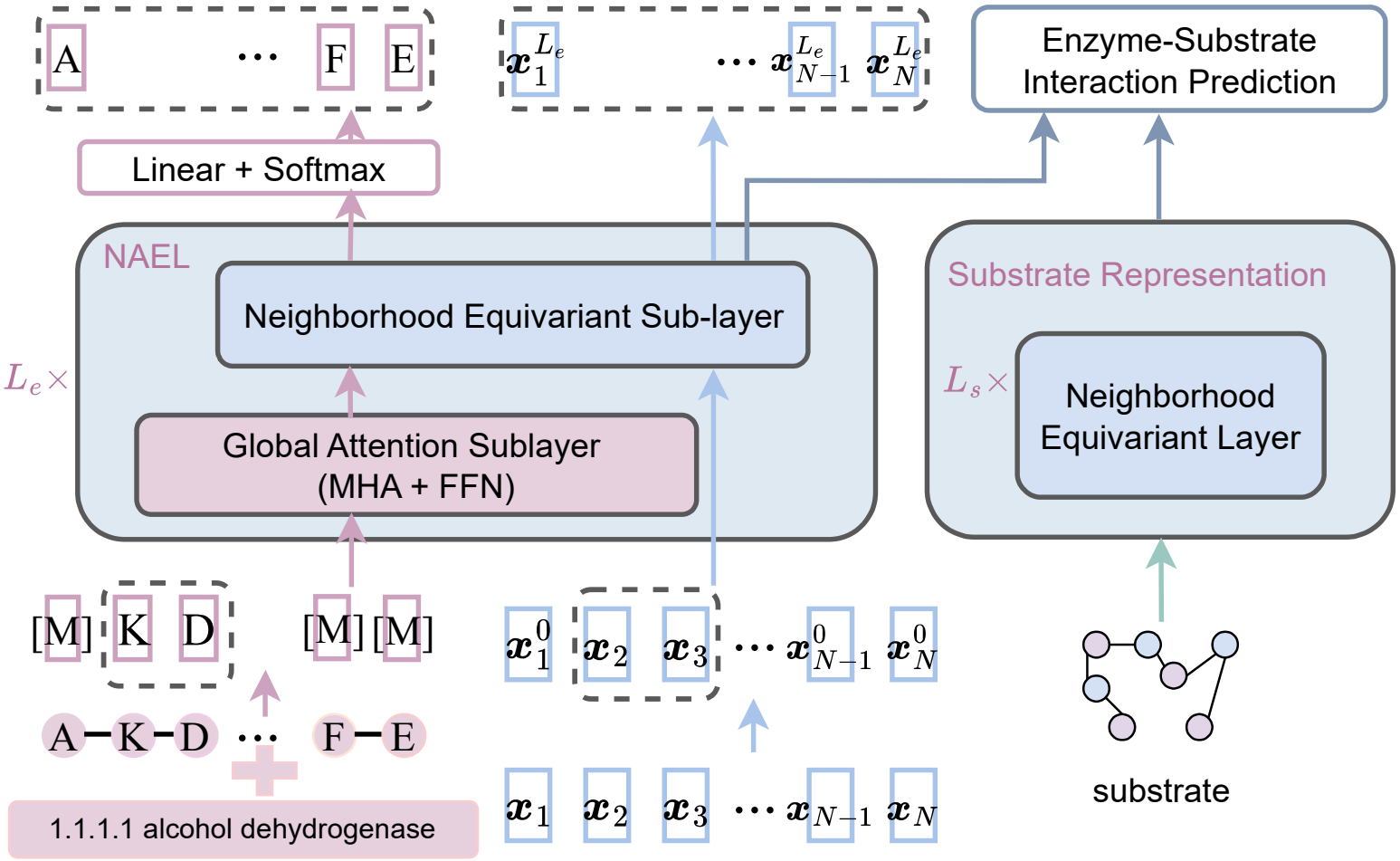
- Controllable Design
  - Functional Sites
  - Enzyme family category

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

Neighborhood Attentive Equivariant Layer



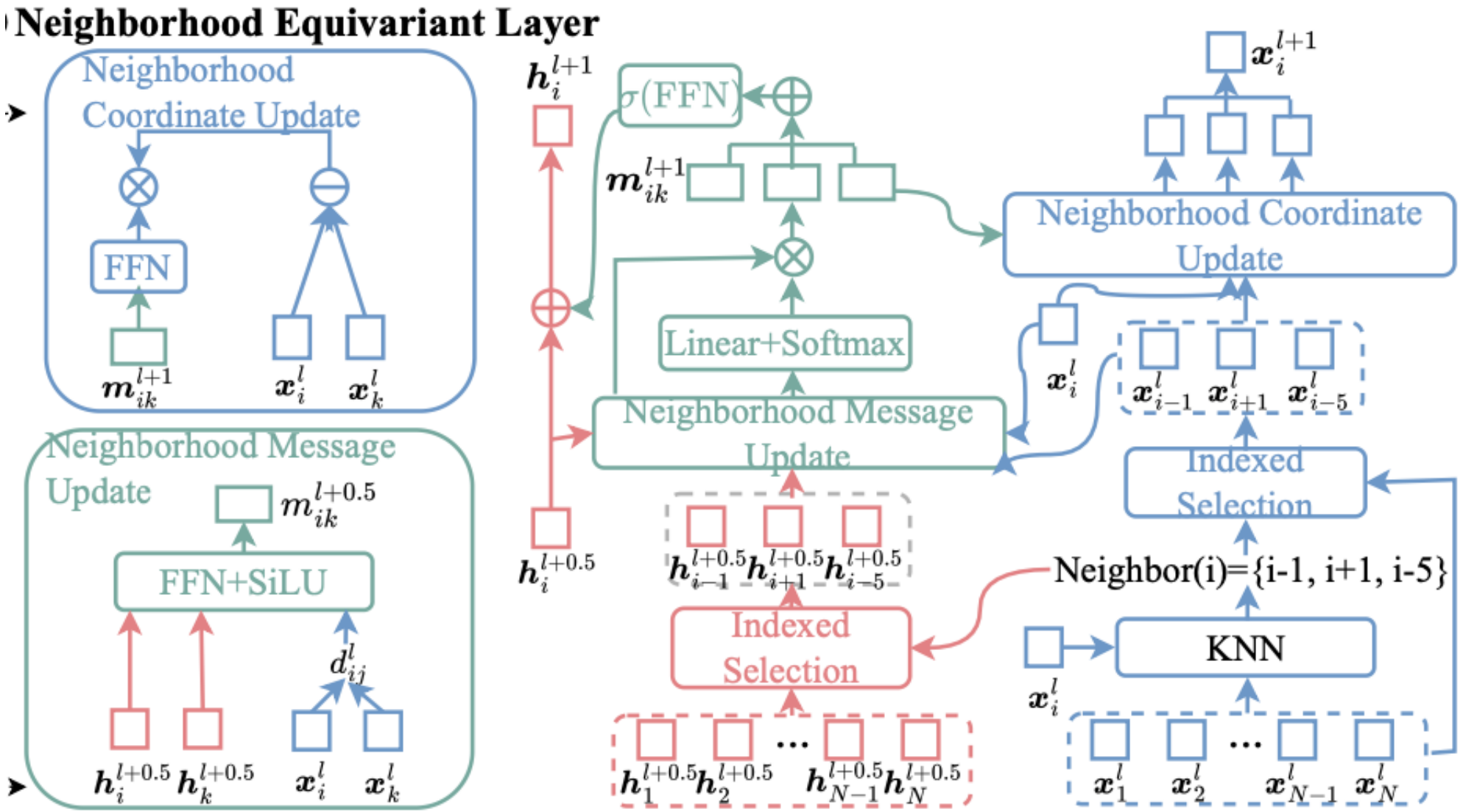
EnzyGen Learning



- 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.*

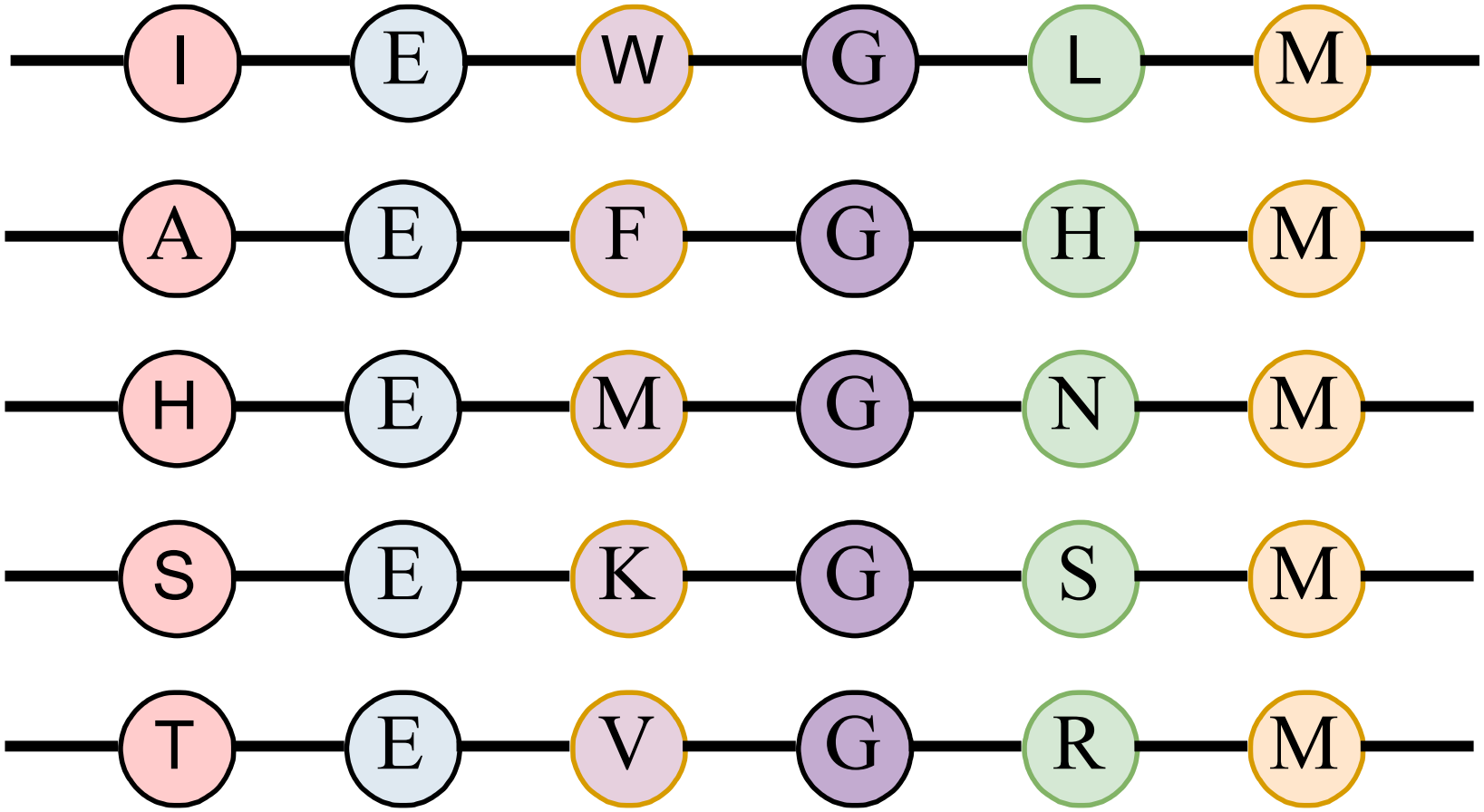
Neighborhood Attentive Equivariant Layer (NAEL)



*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.*

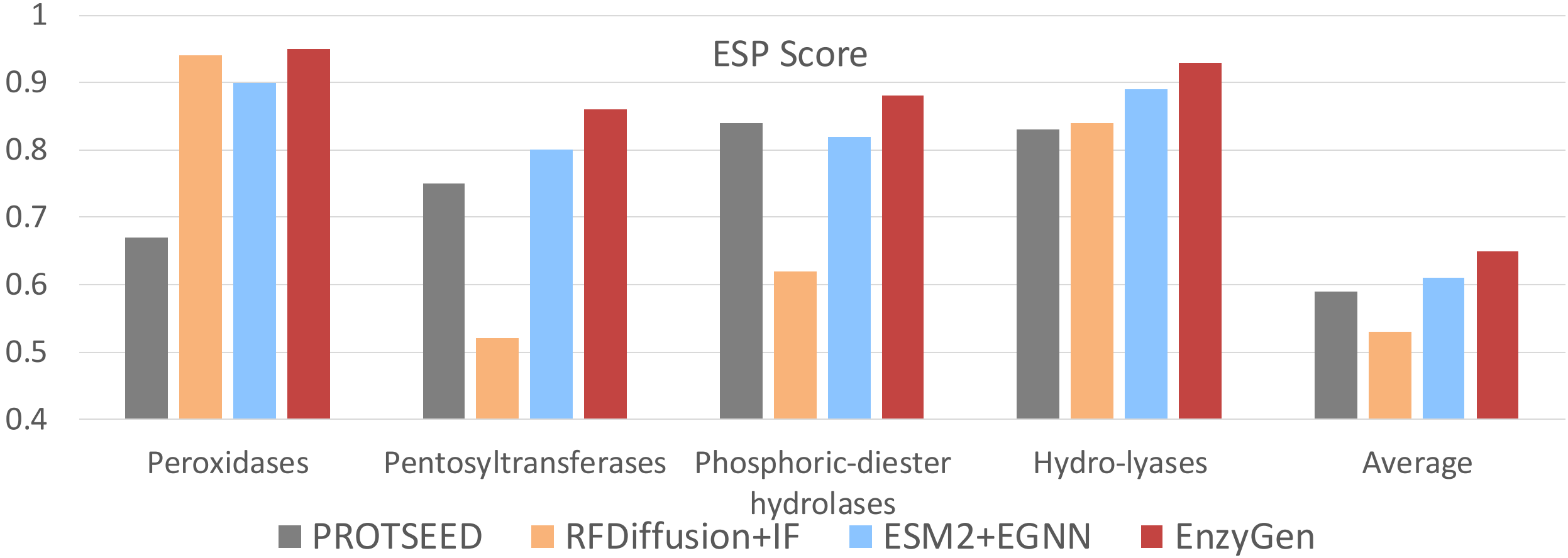
EnzyBench Dataset

- 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.*

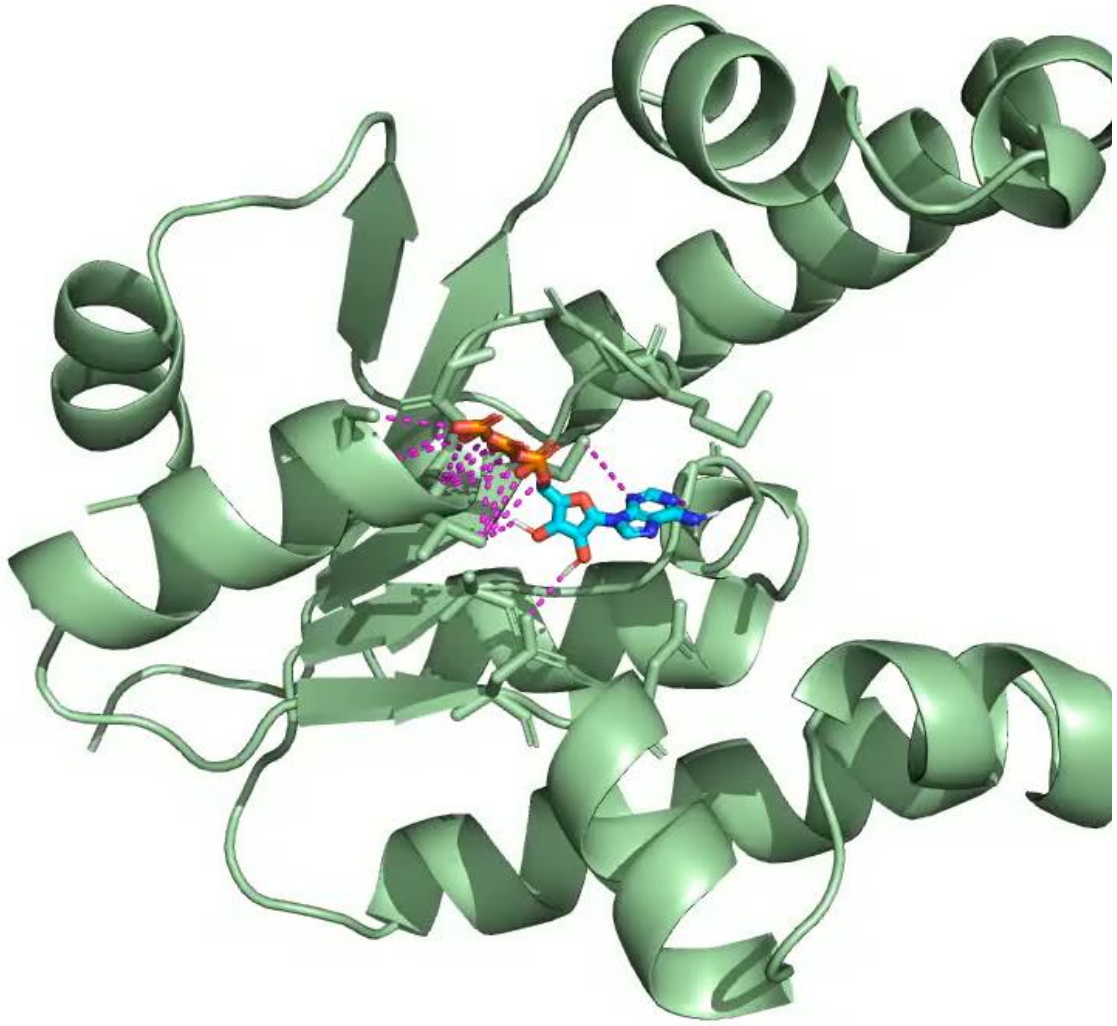
EnzyGen generates enzymes with higher function scores

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

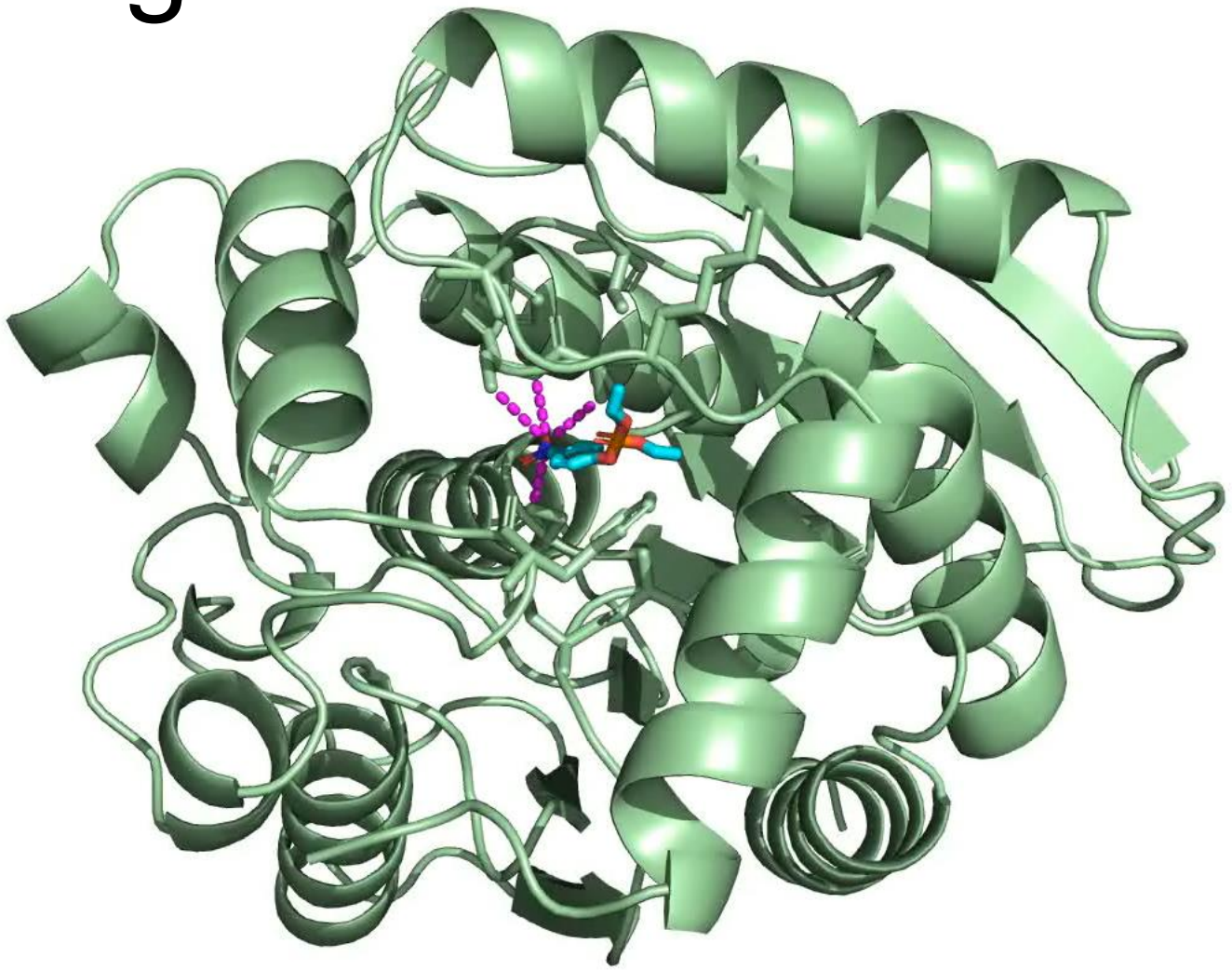


*Song, Zhao, Shi, Jin, Yang, **Li**. Generative Enzyme Design Guided by Functionally Important Sites and Small-Molecule Substrates. ICML 2024.*

EnzyGen designs “good” enzymes in zero-shot categories



Shikimate kinase
(ATP:shikimate 3-phosphotransferase)



Arylesterase
(substrate paraoxon)

Highlights of EnzyGen

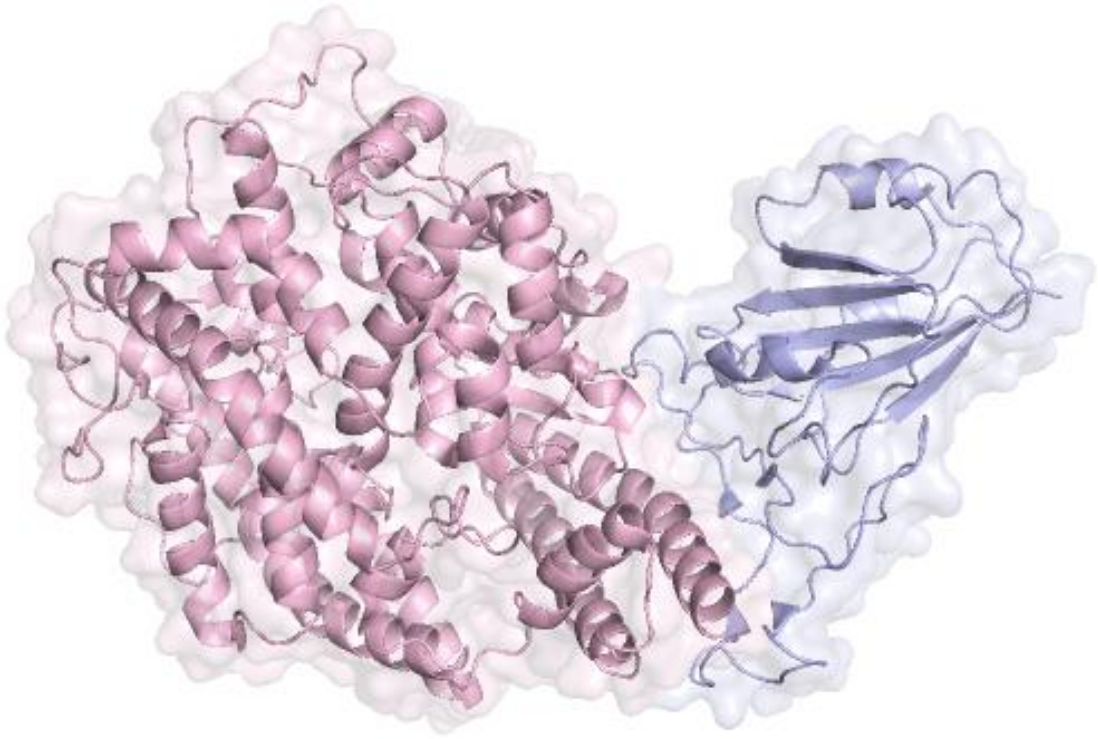
- A unified model for 3k enzyme families
- Guided Generation
  - Functional Important Sites, automatically mined from PDB
  - Enzymy category tags (BRENDA)
- Sequence and Structure Co-design
  - Neighborhood Attentive Equivariant Layer
- Trained takes substrate binding into consideration

Outline

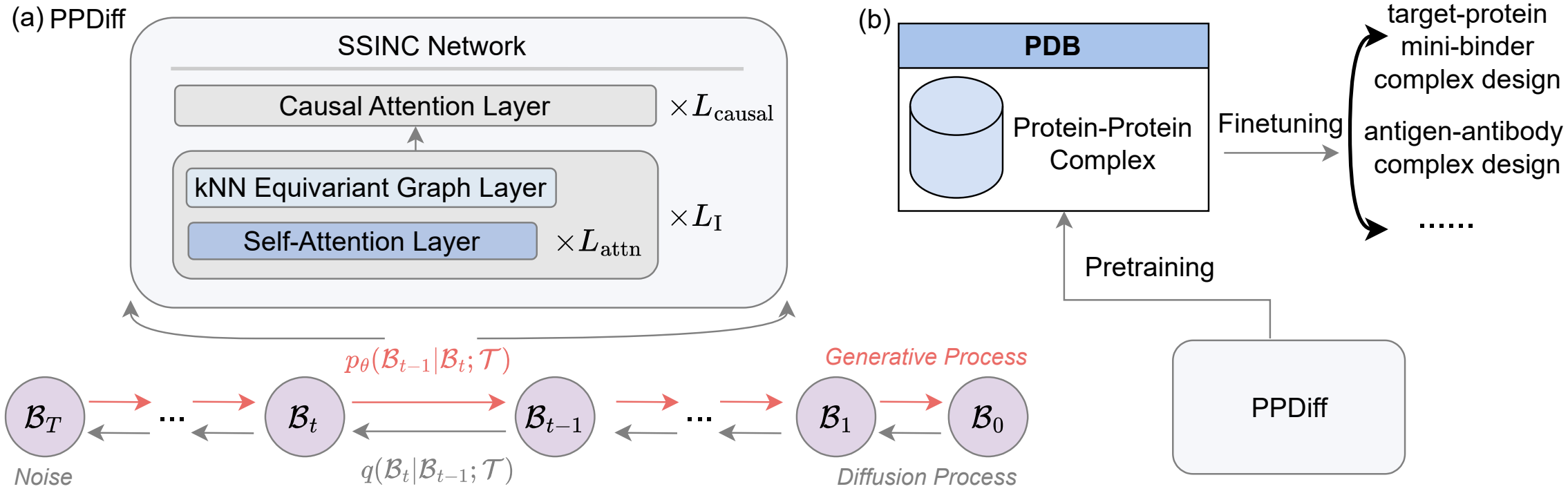
- Multi-scale Biological System and Drugs
  - molecule, cell, tissue, organs, cause of disease
- Basic AI Models for Biomecules
  - sequence, structure, generative model
- MARS: finding small molecule drugs with multiple properties
- EnzyGen: A general generative model for enzyme design
- PPDiff: protein-binding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

Protein-Protein Complex Design

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



PPDiff

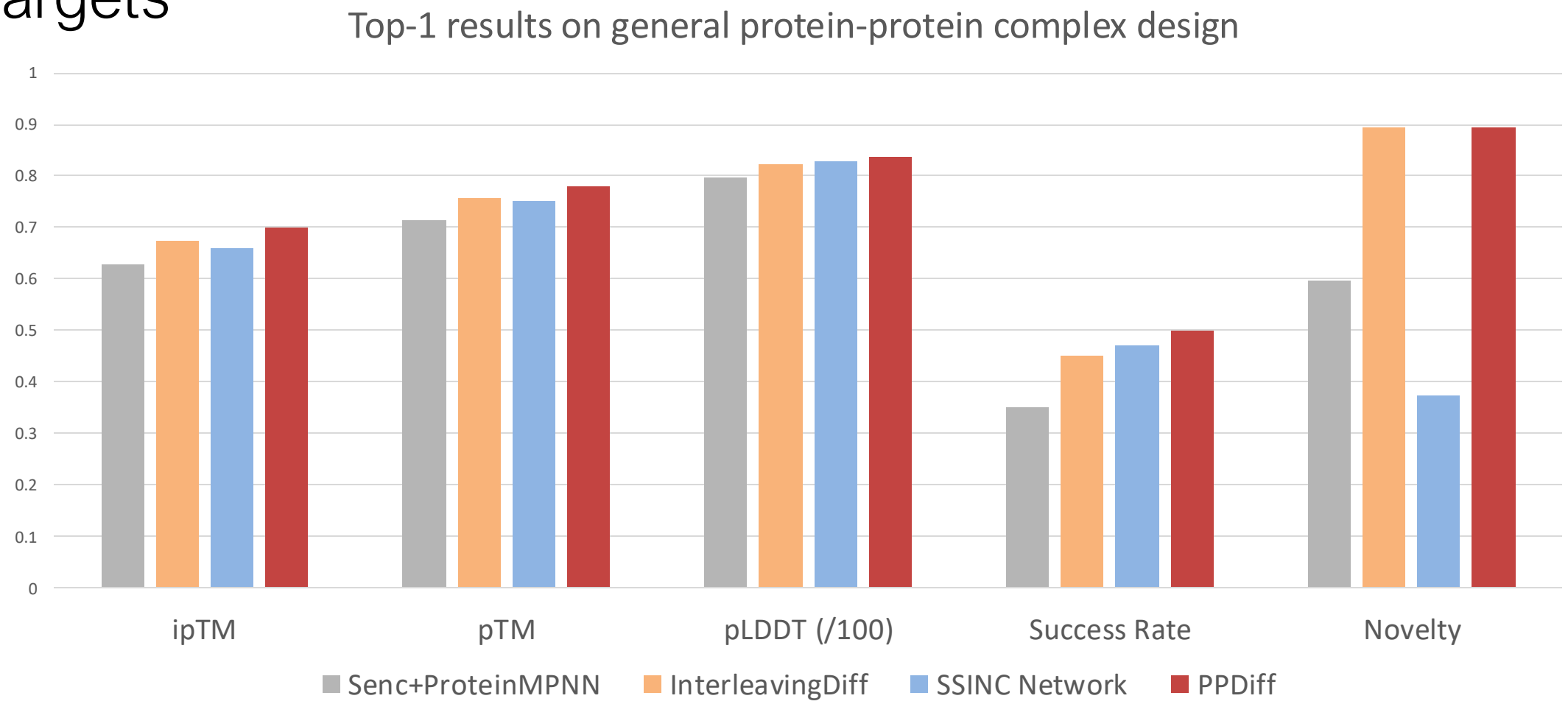


- Diffusing in hybrid space
  - Discrete sequence diffusion
  - Continuous structure diffusion
- SSINC Network
  - Interleaving network (NAEL)
  - Casual attention layers

*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

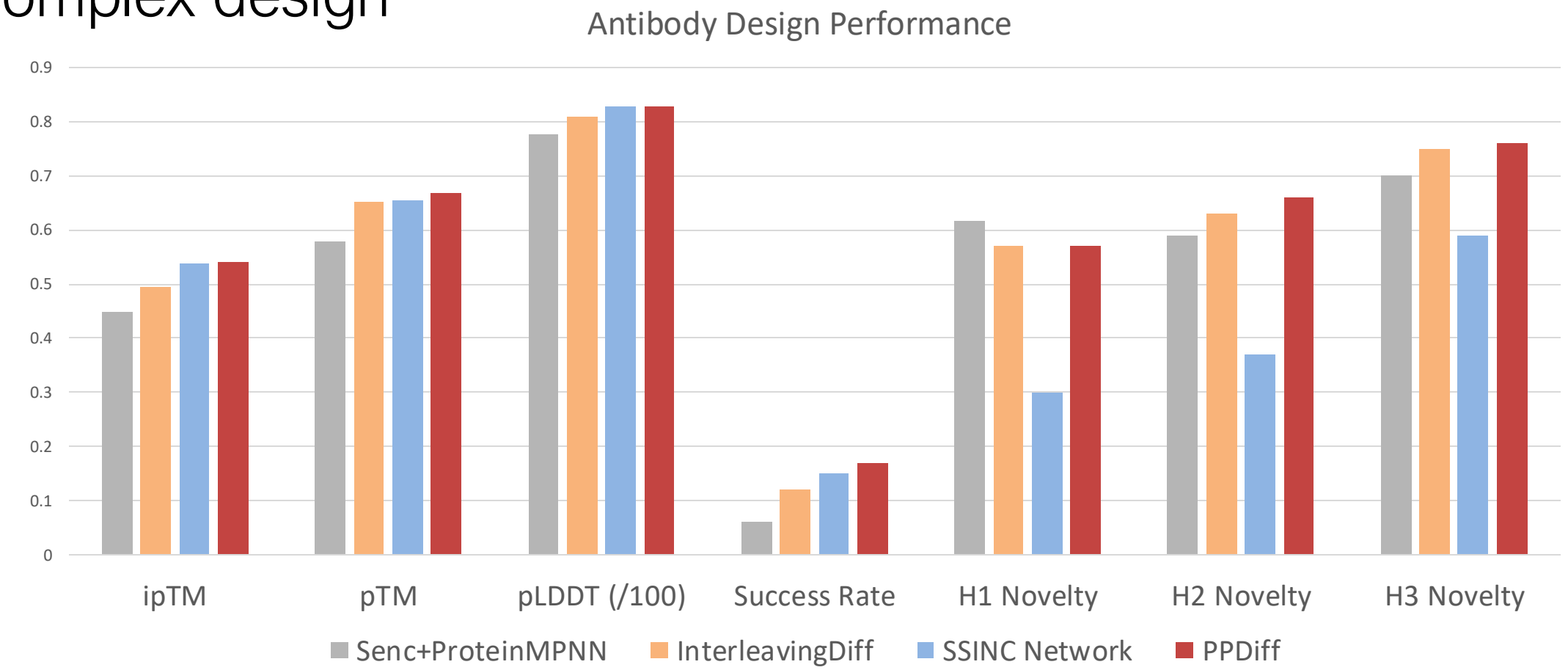
PPDiff achieves **50%** success rate across diverse protein targets



*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

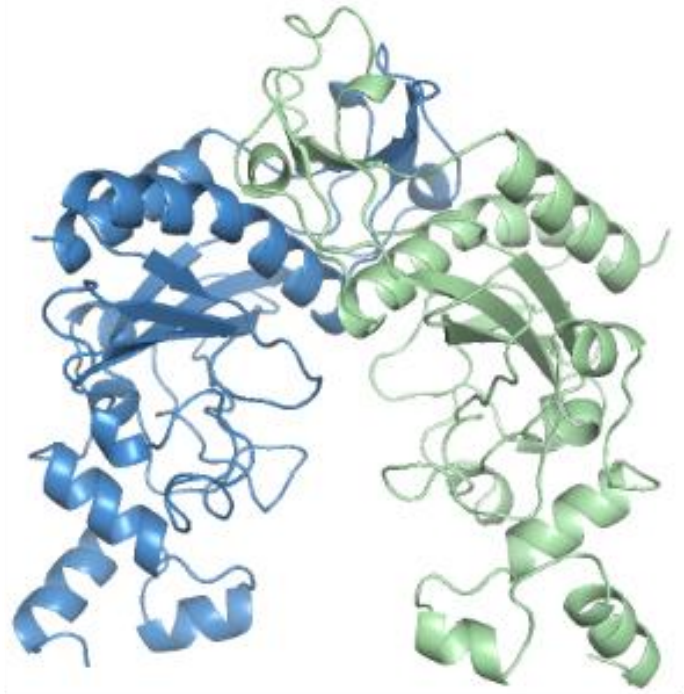
PPDiff achieves 16.89% success rate on antigen-antbody complex design



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

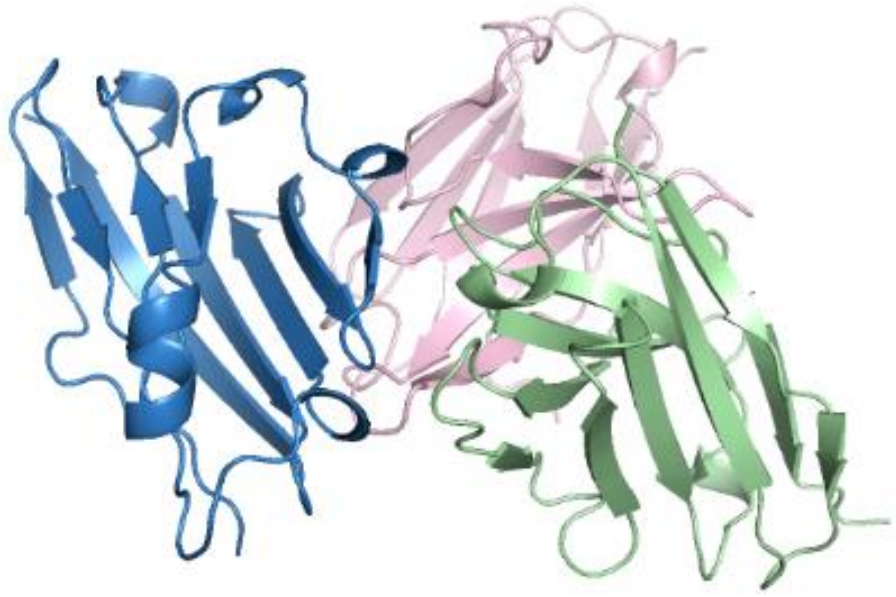
PPDiff designs high-affinity binders/antibody across diverse target proteins

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

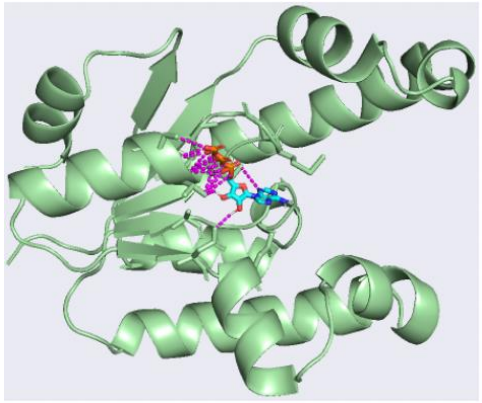
Summary

- 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

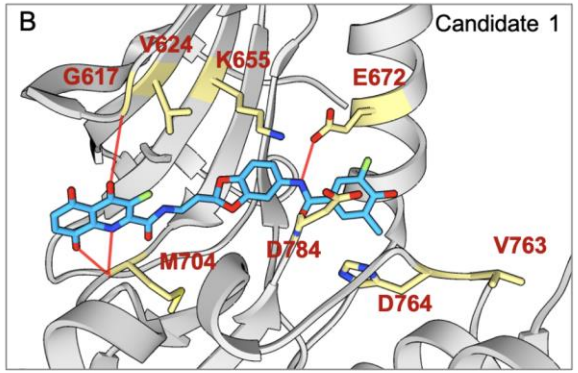
Takeaway of AI Drug Design

- 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