Lecture 13:

AI Models and Biological Understanding

AI Revolution in Biology Foundation Models Scientific Discovery

Ho-min Park

[email protected]<

Lecture Contents

Part 1: Foundation Models

Part 2: Biological Applications

Part 3: Design and Engineering

Part 1/3 - Foundation Models

Large-scale pretraining

Transfer learning

Emergent capabilities

Language Models in Biology

Biological sequences as text

DNA, RNA, Protein sequences → Text format

Tokenization strategies

K-mers, BPE, Character-level encoding

Pretraining objectives

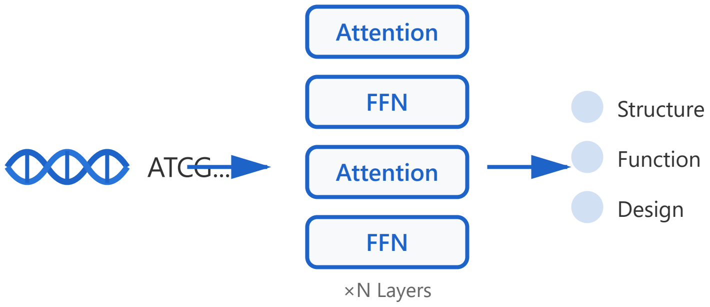
Masked LM, Next token prediction, Contrastive

Scale effects

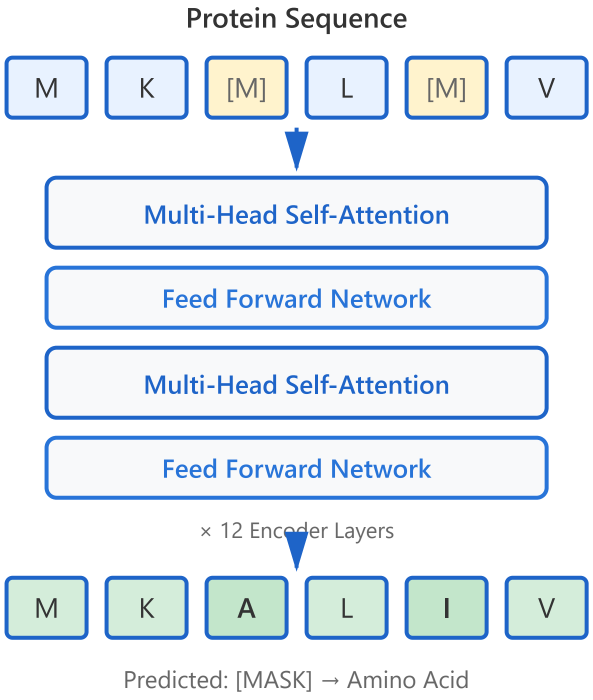
Model size vs. performance trade-offs

Downstream tasks

Structure, Function, Design applications



BERT for Proteins



ProtBERT architecture

12-layer bidirectional encoder

Masked language modeling

15% random masking strategy

Attention patterns

Learns residue interactions

Structural insights

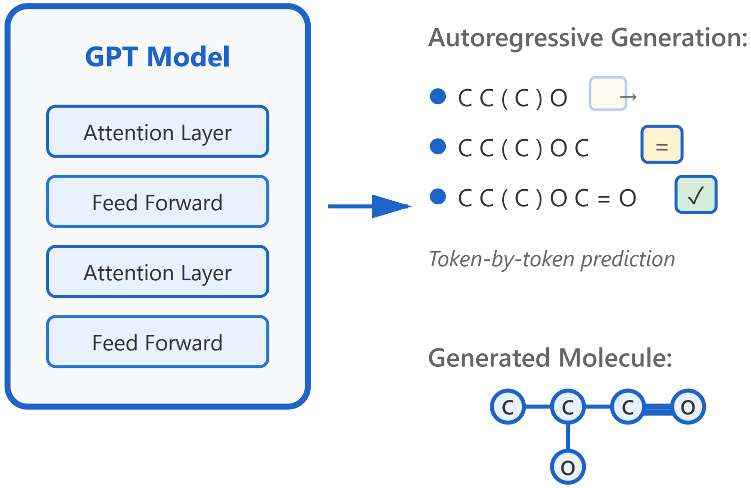
Captures 3D contact maps

Function prediction

GO terms, EC numbers

GPT for Molecules

SMILES Generation Process



GPT Model

Attention Layer

Feed Forward

Attention Layer

Feed Forward

Autoregressive Generation:

- ● C C (C) O □ →
- ● C C (C) O C =
- ● C C (C) O C = O ✓

Token-by-token prediction

Generated Molecule:

Ethyl acetate: CC(C)OC=O

Chemical language models

ChemGPT, MolGPT architectures

Property conditioning

Control molecular attributes

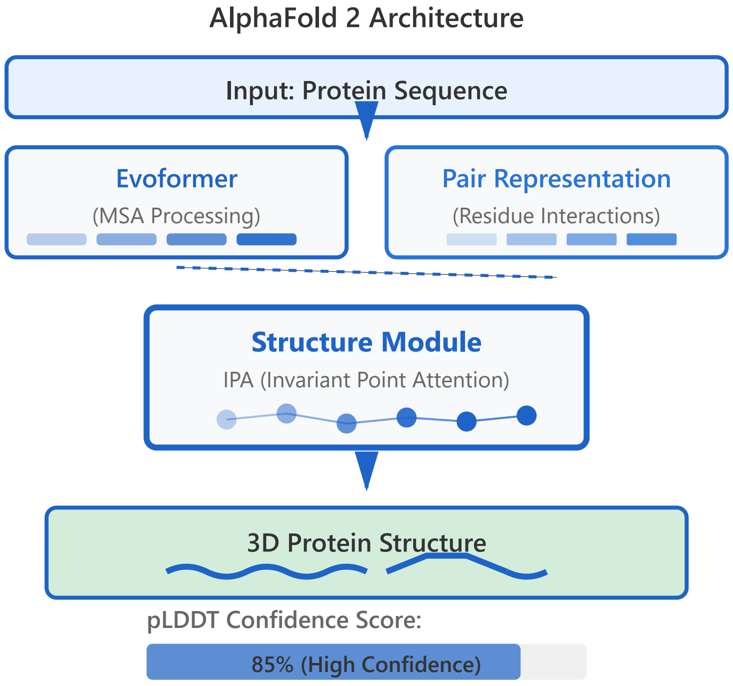
Reaction prediction

Reactants → Products mapping

Retrosynthesis

Backward synthesis planning

AlphaFold Revolution



Architecture innovations

Evoformer + Structure module

MSA processing

Evolutionary information extraction

Structure module

IPA: SE(3)-equivariant attention

Confidence metrics

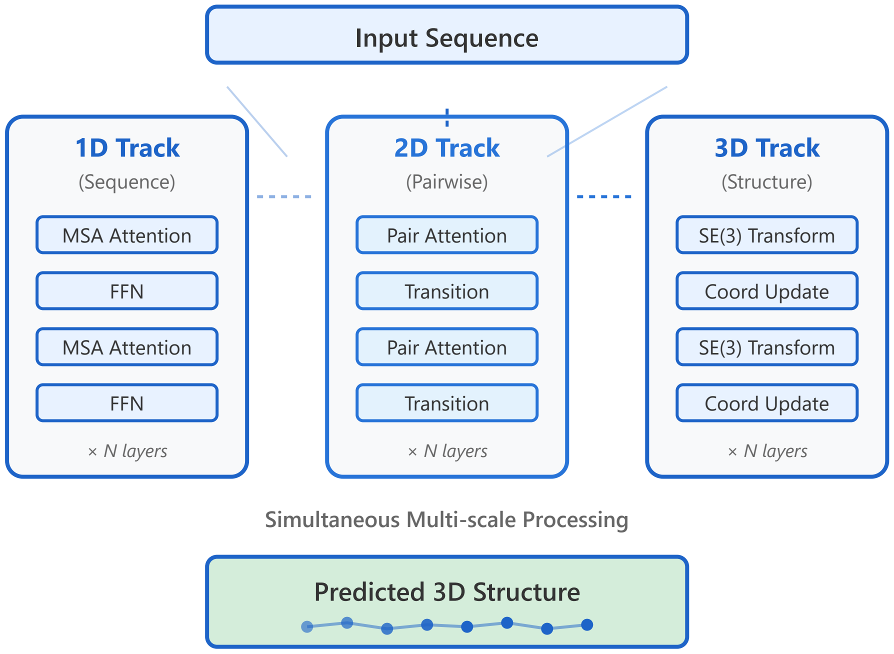
pLDDT per-residue scores

Database impact

200M+ structures predicted

RoseTTAFold

Three-Track Architecture



Input Sequence

1D Track
(Sequence)
MSA Attention
FFN
MSA Attention
FFN
× N layers

2D Track
(Pairwise)
Pair Attention
Transition
Pair Attention
Transition
× N layers

3D Track
(Structure)
SE(3) Transform
Coord Update
SE(3) Transform
Coord Update
× N layers

Simultaneous Multi-scale Processing

Predicted 3D Structure

Three-track architecture

1D, 2D, 3D parallel processing

End-to-end learning

Direct structure prediction

Complex prediction

Protein-protein interactions

Speed advantages

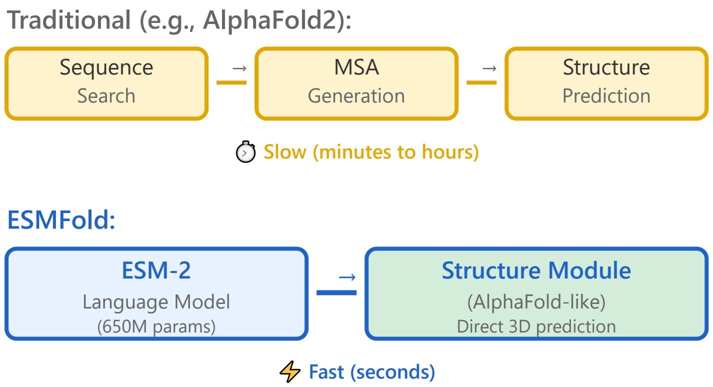
Faster than AlphaFold2

Applications

Structure, function, design

ESMFold

Language Model-Only Approach



Traditional (e.g., AlphaFold2):

Sequence Search → MSA Generation → Structure Prediction

⏱ Slow (minutes to hours)

ESMFold:

ESM-2 Language Model (650M params) → Structure Module (AlphaFold-like) Direct 3D prediction

⚡ Fast (seconds)

Key Innovation: No MSA Required

- Evolutionary info learned directly from 250M+ protein sequences
- 60× faster than AlphaFold2 (seconds vs minutes)
- Enables metagenomic-scale structure prediction

Language model only

ESM-2 pretrained transformer

No MSA required

Single sequence input

Speed benefits

60× faster inference

Metagenomic applications

Unknown protein discovery

Limitations

Lower accuracy on orphan proteins

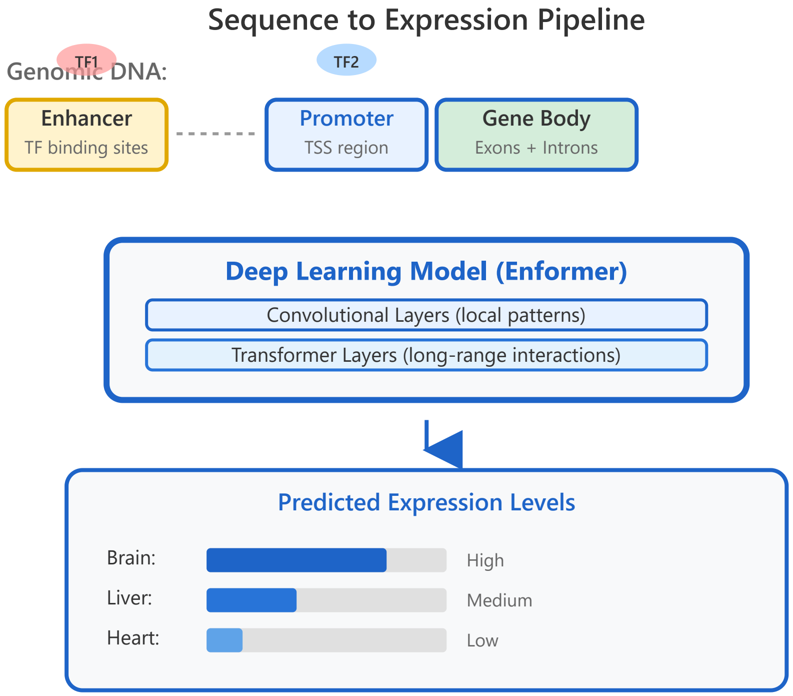
Part 2/3 - Biological AI

Predictive models

Interpretable AI

Biological insights

Gene Expression Prediction


Sequence to Expression Pipeline
Genomic DNA:
TF1
TF2
Enhancer
TF binding sites
Promoter
TSS region
Gene Body
Exons + Introns
Deep Learning Model (Enformer)
Convolutional Layers (local patterns)
Transformer Layers (long-range interactions)
Predicted Expression Levels
Brain:
High
Liver:
Medium
Heart:
Low

Cell-type specific expression prediction

Sequence to expression

DNA → RNA abundance mapping

Promoter models

TSS region activity prediction

Enhancer grammar

TF binding syntax learning

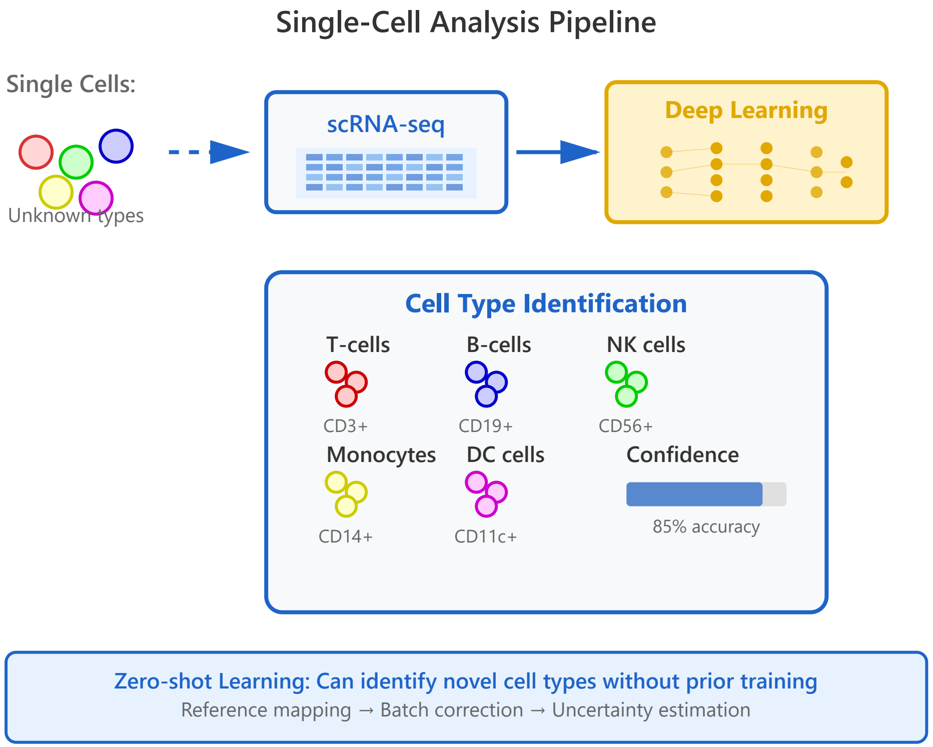
Cell type specificity

Context-dependent prediction

Enformer architecture

Transformer + CNN hybrid model

Cell Type Classification


Single-Cell Analysis Pipeline
Single Cells:
Unknown types
scRNA-seq
Deep Learning
Cell Type Identification
T-cells
CD3+
B-cells
CD19+
NK cells
CD56+
Monocytes
CD14+
DC cells
CD11c+
Confidence
85% accuracy
Zero-shot Learning: Can identify novel cell types without prior training
Reference mapping → Batch correction → Uncertainty estimation

Single-cell models

scBERT, Geneformer architectures

Reference mapping

Atlas-based annotation

Zero-shot learning

Novel cell type discovery

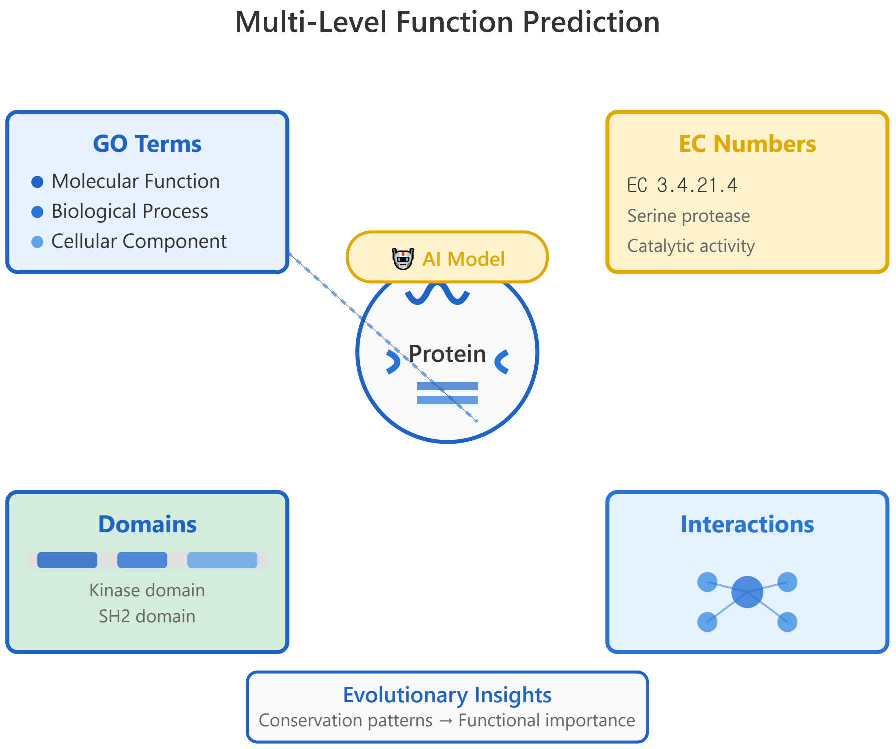
Batch correction

Remove technical variation

Uncertainty estimation

Confidence scoring

Protein Function



GO term prediction

Molecular/biological/cellular

EC number classification

Enzyme commission numbers

Domain annotation

Functional regions identification

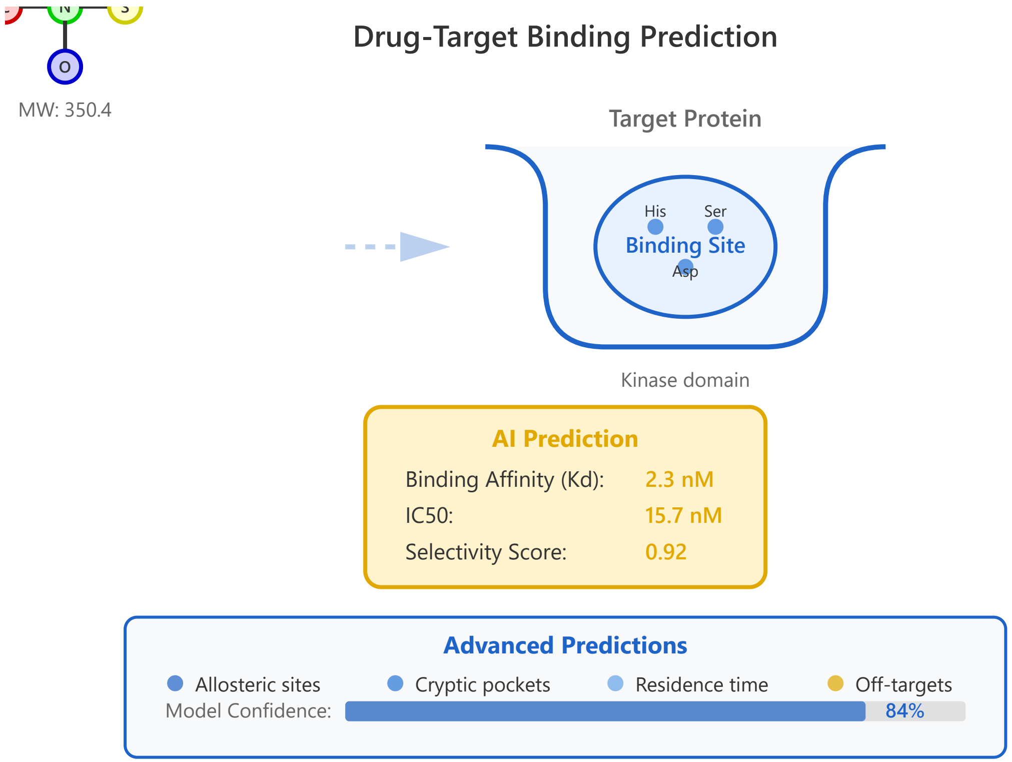
Interaction prediction

Protein-protein networks

Evolutionary insights

Conservation-function mapping

Drug-Target Affinity



Drug-Target Binding Prediction

AI Prediction

Binding Affinity (Kd):	2.3 nM
IC50:	15.7 nM
Selectivity Score:	0.92

Advanced Predictions

● Allosteric sites ● Cryptic pockets ● Residence time ● Off-targets

Model Confidence: 84%

Binding prediction

Kd, Ki, IC50 values

Kinase selectivity

Off-target profiling

Allosteric sites

Non-competitive binding

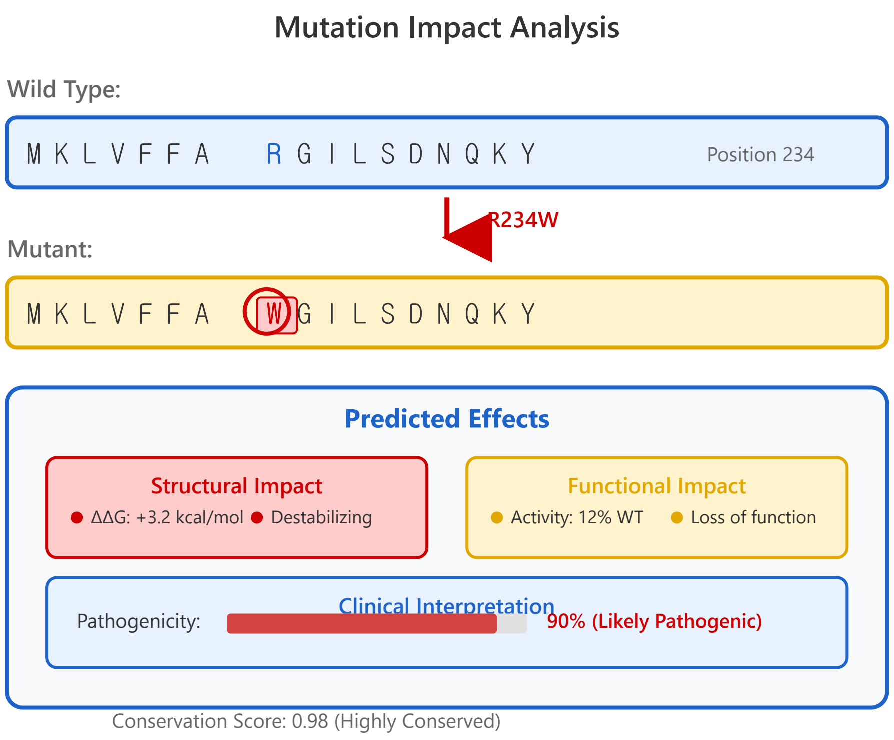
Cryptic pockets

Hidden binding sites

Residence time

Drug-target kinetics

Mutation Effects



Mutation Impact Analysis

Wild Type:

M K L V F F A R G I L S D N Q K Y

Position 234

R234W

Mutant:

M K L V F F A W G I L S D N Q K Y

Predicted Effects

Structural Impact

● ΔΔG: +3.2 kcal/mol ● Destabilizing

Functional Impact

● Activity: 12% WT ● Loss of function

Clinical Interpretation

Pathogenicity: 90% (Likely Pathogenic)

Conservation Score: 0.98 (Highly Conserved)

Pathogenicity prediction

Disease association scoring

Stability changes

ΔΔG calculation

Function impact

Activity & binding changes

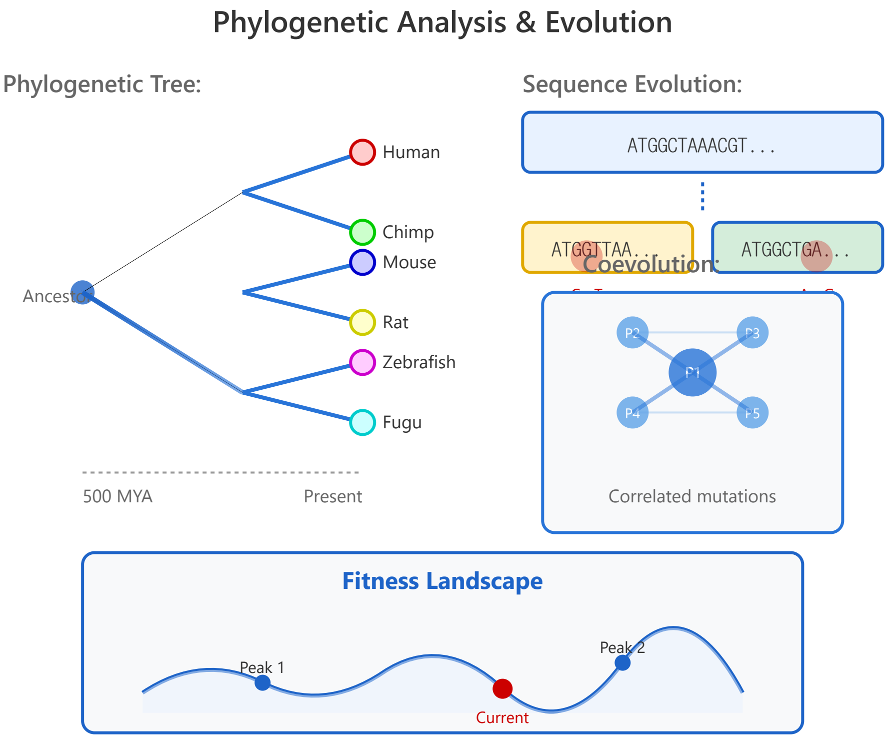
Evolutionary constraints

Conservation analysis

Clinical interpretation

Variant classification

Evolution Modeling



Sequence evolution

Substitution models & rates

Phylogenetic inference

Tree reconstruction methods

Ancestral reconstruction

Ancient sequence prediction

Coevolution

Correlated mutations analysis

Fitness landscapes

Adaptive evolution mapping

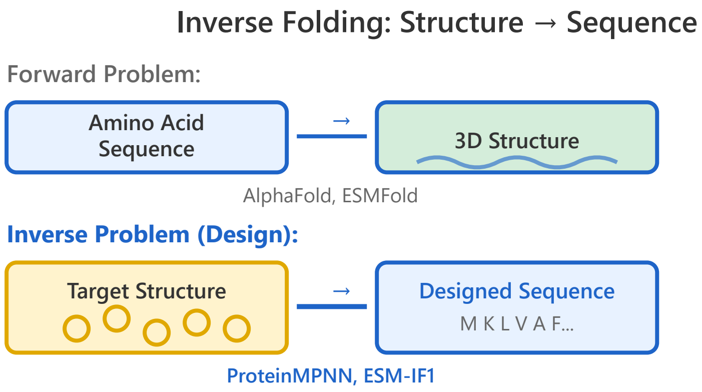
Part 3/3 - Applications

Design problems

Engineering solutions

Therapeutic development

Protein Design



Inverse Folding: Structure → Sequence

Forward Problem:

Amino Acid Sequence → 3D Structure

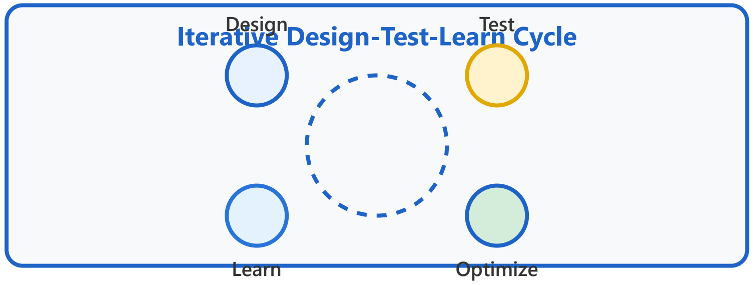
AlphaFold, ESMFold

Inverse Problem (Design):

Target Structure → Designed Sequence

M K L V A F...

ProteinMPNN, ESM-IF1



Iterative Design-Test-Learn Cycle

Inverse folding

Structure → sequence prediction

Scaffold design

De novo backbone generation

Interface design

Protein-protein interactions

De novo binders

Target-specific protein design

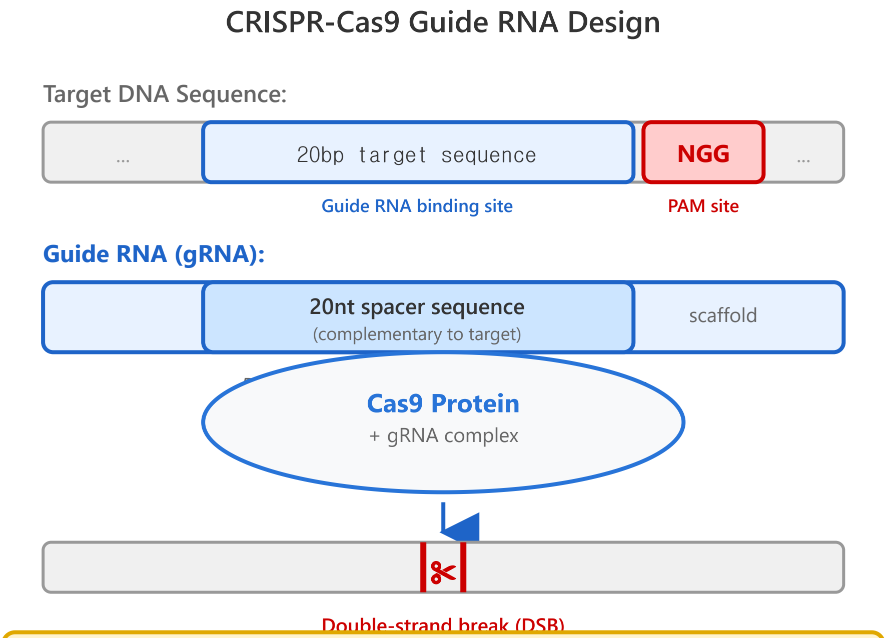
Stability optimization

Thermostability enhancement

Antibody Design

- CDR optimization
- Humanization
- Affinity maturation
- Specificity engineering
- Developability

CRISPR Optimization



CRISPR-Cas9 Guide RNA Design

Target DNA Sequence:

... | 20bp target sequence | NGG | ...

Guide RNA binding site

PAM site

Guide RNA (gRNA):

20nt spacer sequence (complementary to target) | scaffold

Cas9 Protein

+ gRNA complex

✂

Double-strand break (DSB)

AI-Powered Optimization:

- On-target efficiency scoring
- Off-target prediction
- Edit outcome prediction

Guide RNA design

20nt spacer + scaffold optimization

Off-target prediction

Minimize unintended cuts

Efficiency scoring

On-target activity models

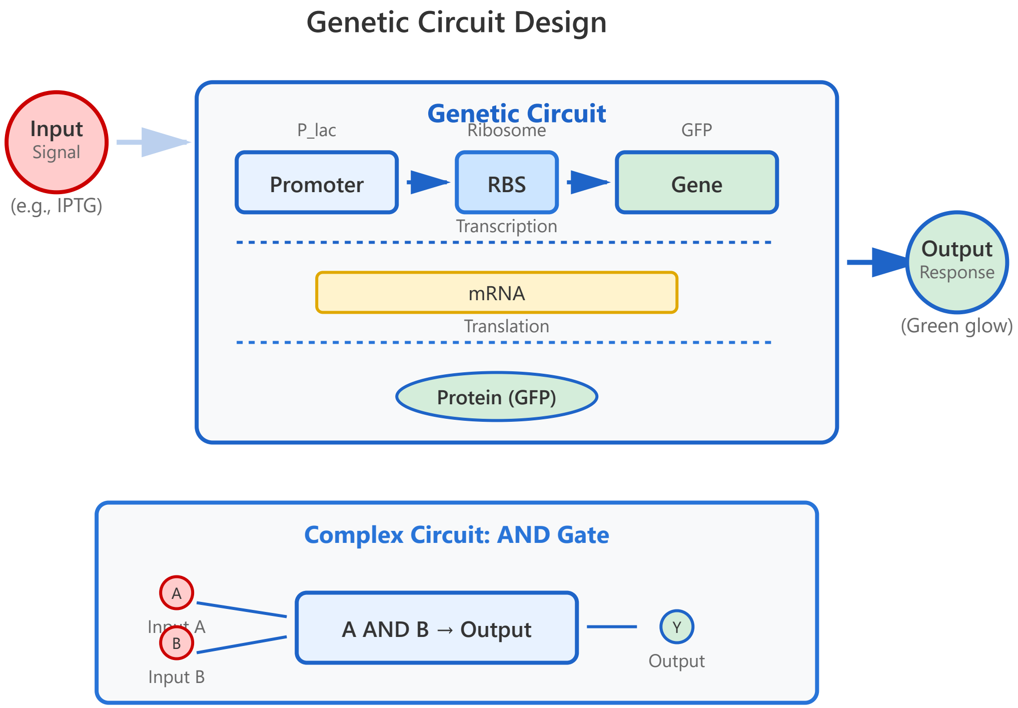
Prime editing

Precise base substitutions

Base editing

C→T, A→G conversions

Synthetic Biology



Circuit design

Logic gates & regulatory networks

Part optimization

Promoters, RBS, terminators

Metabolic pathways

Multi-enzyme cascades

Orthogonal systems

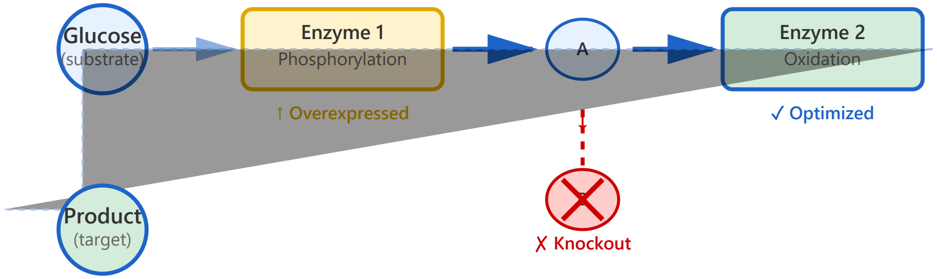
Independent control modules

Predictive models

AI-guided circuit optimization

Metabolic Engineering

Metabolic Pathway Optimization



AI-Guided Metabolic Engineering

Flux Balance Analysis (FBA)

- Identify bottlenecks
- Predict knockouts
- Optimize expression levels

Machine Learning Models

- Enzyme activity prediction
- Strain design
- Growth prediction

Flux optimization
Balance metabolic flow

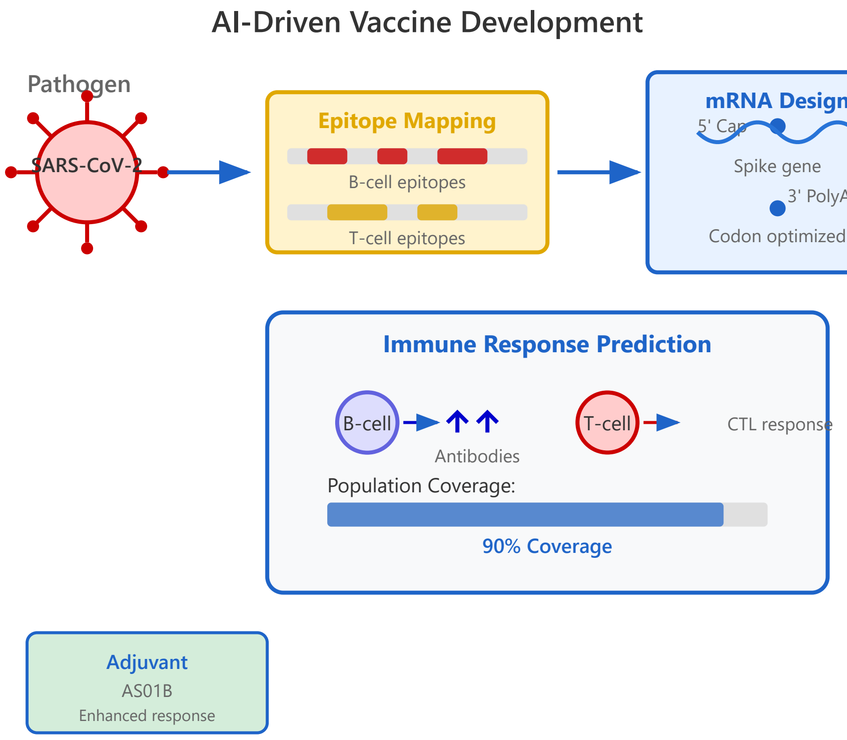
Enzyme engineering
Improve catalytic efficiency

Pathway design
Novel biosynthetic routes

Strain optimization
Host organism engineering

Scale-up prediction
Lab → production modeling

Vaccine Design



Epitope prediction

B-cell & T-cell epitopes

Immunogenicity

Immune response modeling

Coverage optimization

Population HLA diversity

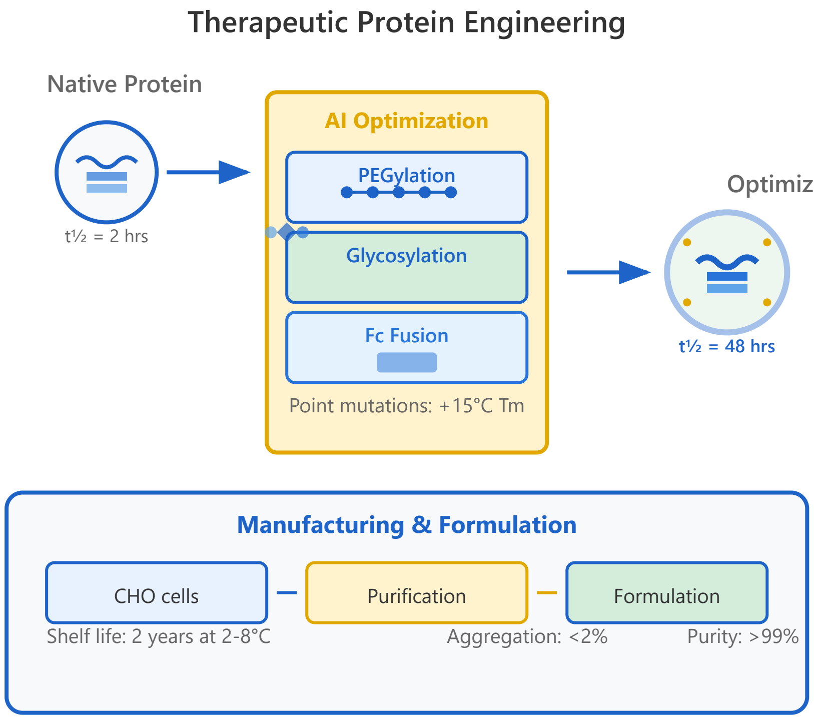
Adjuvant selection

Enhance immune response

mRNA design

Codon optimization & stability

Therapeutic Proteins



Stability engineering
Thermal & chemical stability

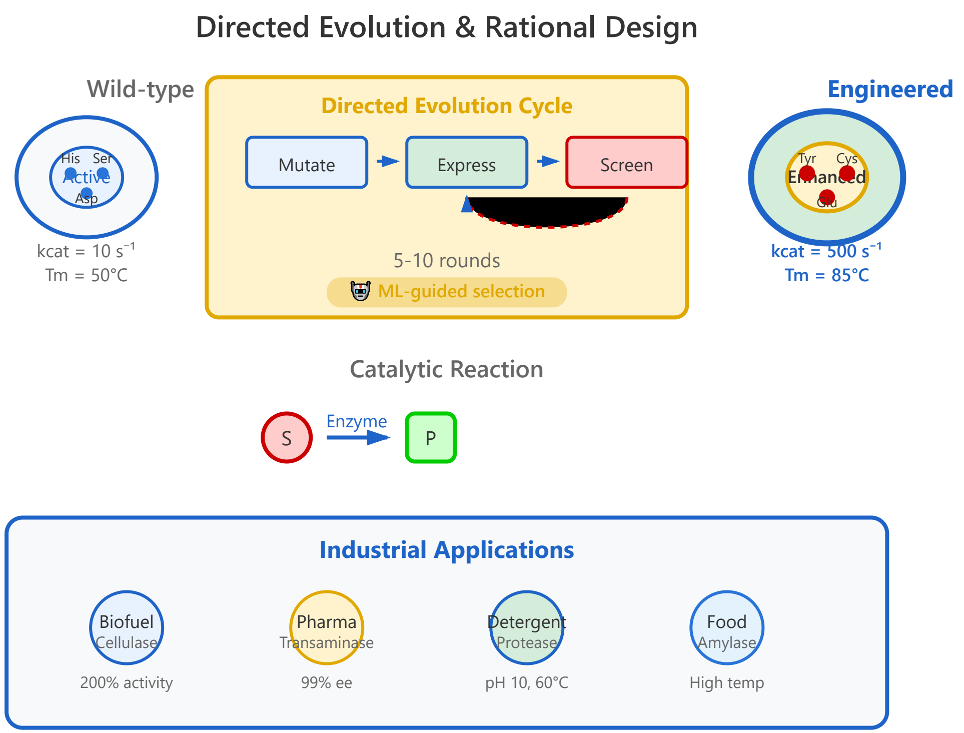
Half-life extension
PEGylation, Fc fusion

Immunogenicity reduction
T-cell epitope removal

Formulation prediction
Aggregation prevention

Manufacturing optimization
Yield & quality improvement

Enzyme Engineering



Activity improvement

kcat/Km optimization

Substrate specificity

Promiscuity engineering

Thermostability

High temperature operation

Solvent tolerance

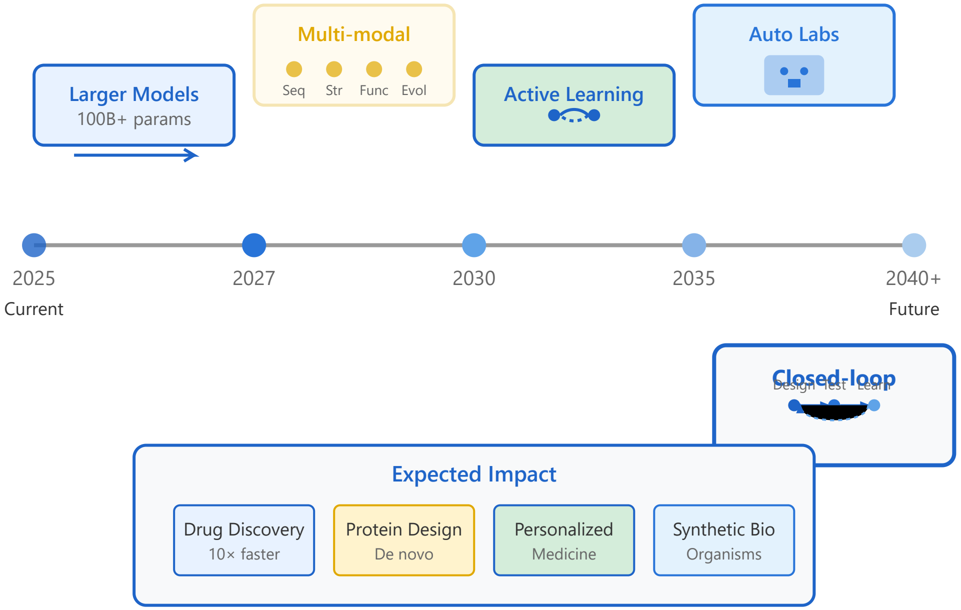
Organic solvent resistance

Directed evolution

Iterative improvement cycles

Future Perspectives

Evolution of Biological AI Systems



Larger Models
100B+ params

Multi-modal
Seq Str Func Evol

Active Learning

Auto Labs

2025
Current

2027

2030

2035

2040+
Future

Closed-loop
Design Test Learn

Expected Impact

Drug Discovery
10× faster

Protein Design
De novo

Personalized
Medicine

Synthetic Bio
Organisms

Larger models

100B+ parameter systems

Multi-modal learning

Seq + structure + function

Active learning

Experimental feedback loops

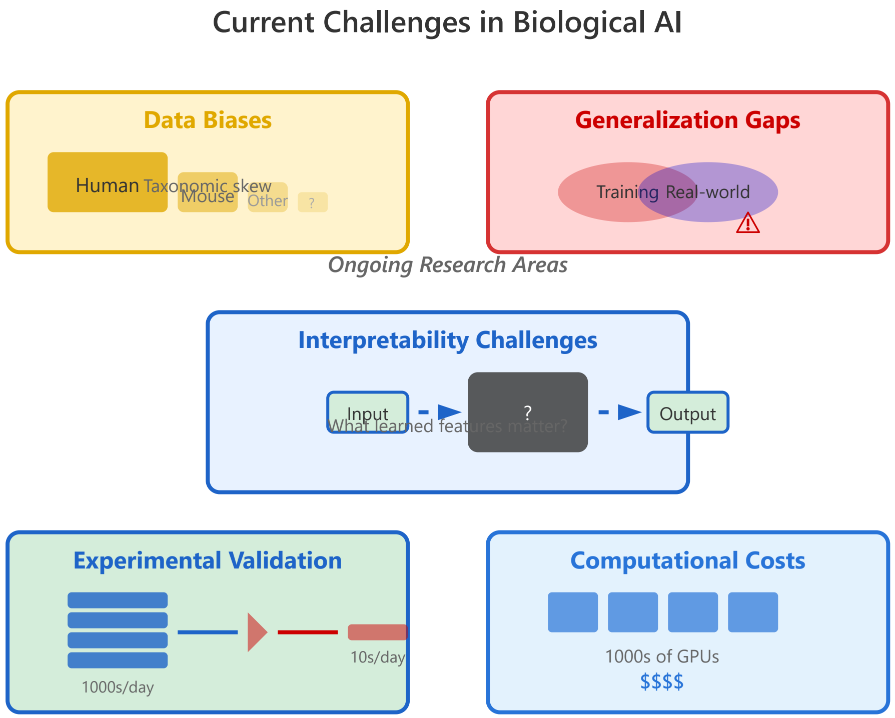
Automated labs

Robot-driven experiments

Closed-loop discovery

End-to-end automation

Limitations



Data biases

Taxonomic & functional skew

Generalization gaps

Out-of-distribution failures

Interpretability challenges

Black box models

Experimental validation

Lab throughput bottleneck

Computational costs

Training & inference expense

Hands-on: AlphaFold Usage

💻 Practical Exercise

- Structure prediction
- Confidence interpretation
- Complex modeling
- Mutation analysis
- Drug discovery applications

Hands-on: Bio Transformers

💻 Practical Exercise

- Model loading
- Sequence encoding
- Fine-tuning
- Embedding extraction
- Downstream tasks

Thank You!

Scientific breakthroughs

Drug discoveries

Future potential

Career opportunities

Questions? Contact: homin.park@ghent.ac.kr

Thank You!

Scientific breakthroughs

Drug discoveries

Future potential

Career opportunities

Questions? Contact: