

Lecture 13:

AI Models and Biological Understanding

AI Revolution in Biology

Foundation Models

Scientific Discovery

Ho-min Park

<

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

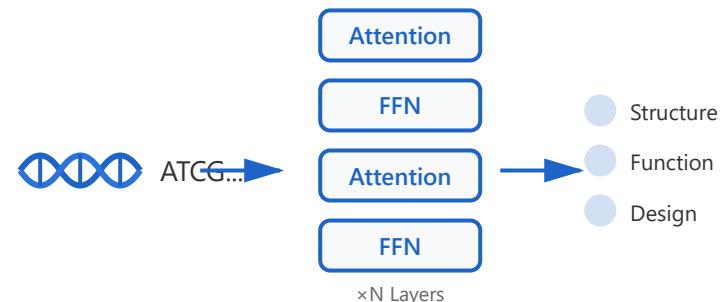
Masked LM, Next token prediction, Contrastive

Scale effects

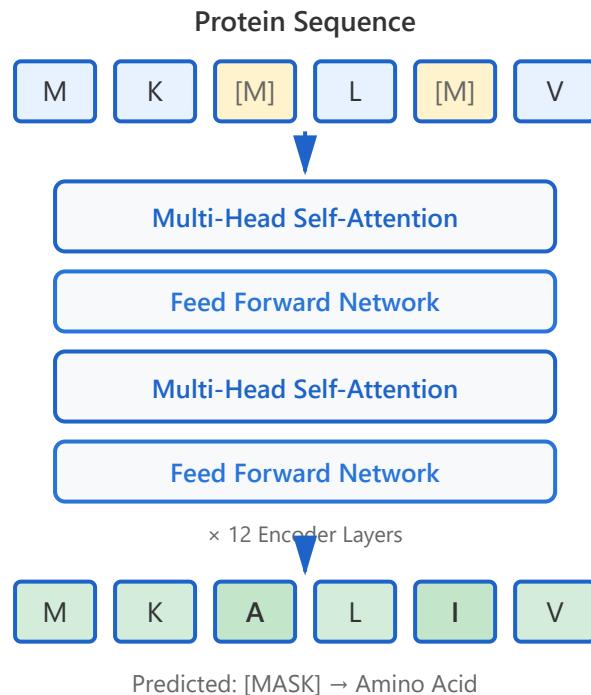
Model size vs. performance trade-offs

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

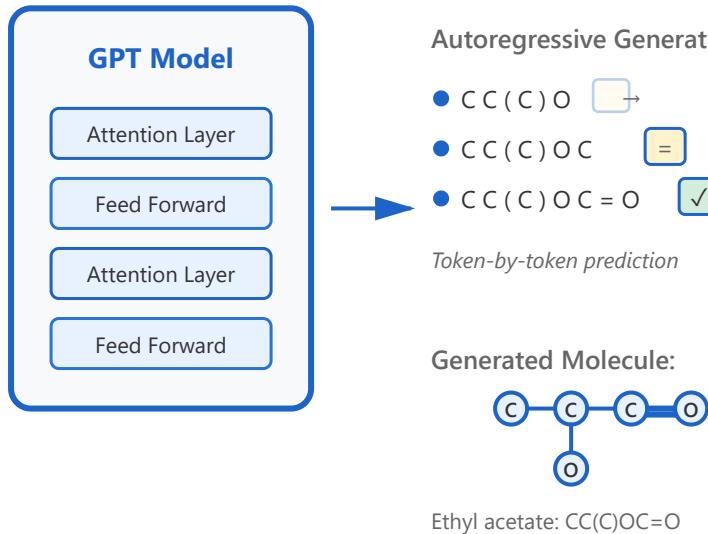
Captures 3D contact maps

Function prediction

GO terms, EC numbers

GPT for Molecules

SMILES Generation Process

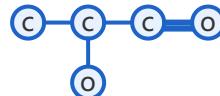


Autoregressive Generation:

- $\text{CC}(\text{C})\text{O}$
- $\text{CC}(\text{C})\text{O C}$
- $\text{CC}(\text{C})\text{O C}=\text{O}$

Token-by-token prediction

Generated Molecule:



Ethyl acetate: $\text{CC}(\text{C})\text{OC}=\text{O}$

Chemical language models

ChemGPT, MolGPT architectures

Property conditioning

Control molecular attributes

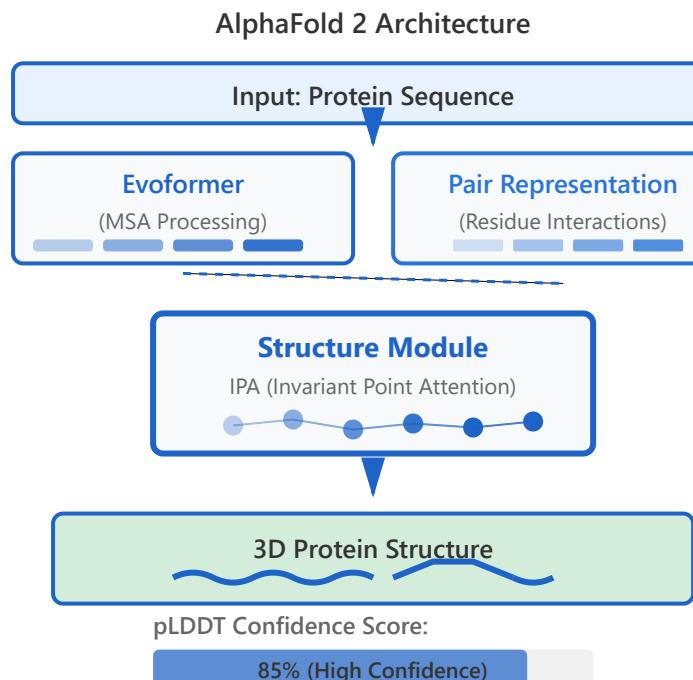
Reaction prediction

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

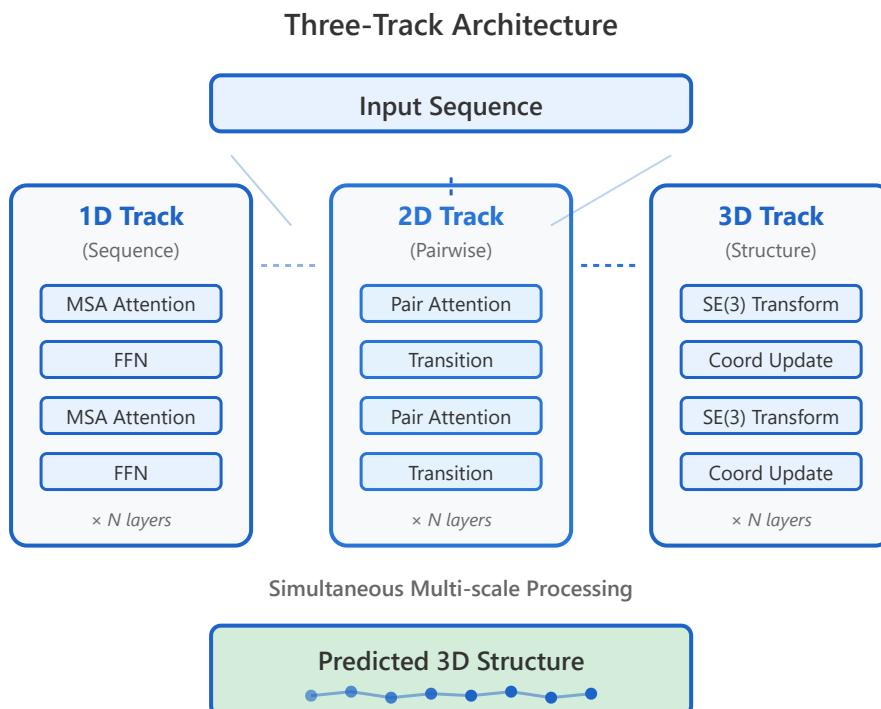
Confidence metrics

pLDDT per-residue scores

Database impact

200M+ structures predicted

RoseTTAFold



Three-track architecture

1D, 2D, 3D parallel processing

End-to-end learning

Direct structure prediction

Complex prediction

Protein-protein interactions

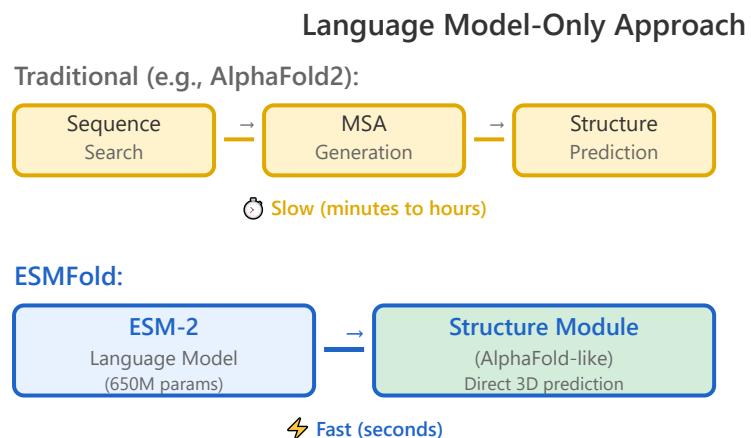
Speed advantages

Faster than AlphaFold2

Applications

Structure, function, design

ESMFold



Key Innovation: No MSA Required

- Evolutionary info learned directly from 250M+ protein sequences
- 60x 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

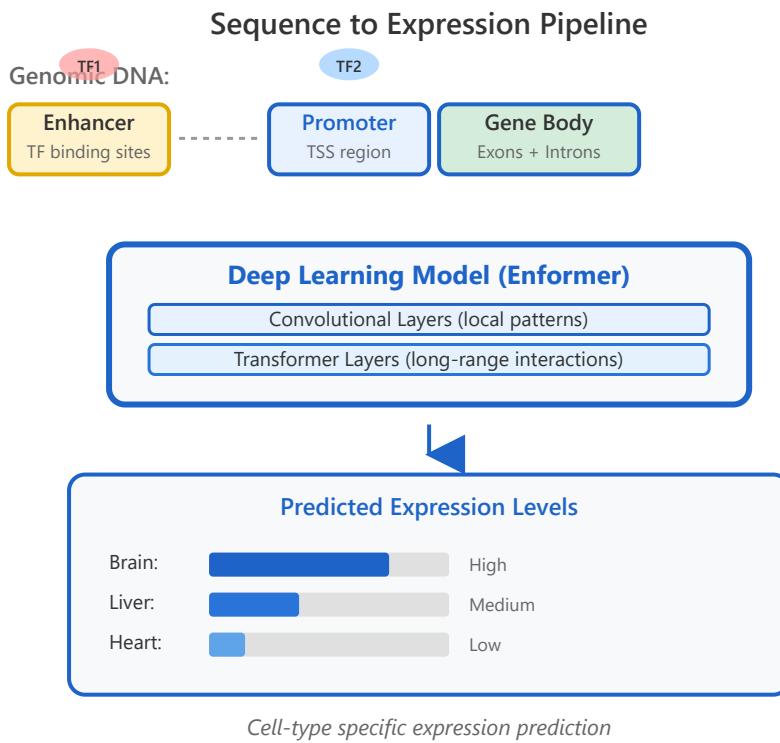
Part 2/3 - Biological AI

Predictive models

Interpretable AI

Biological insights

Gene Expression Prediction



Sequence to expression

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 models
- Reference mapping
- Zero-shot learning
- Batch correction
- Uncertainty estimation

Protein Function

- GO term prediction
- EC number classification
- Domain annotation
- Interaction prediction
- Evolutionary insights

Drug-Target Affinity

- Binding prediction
- Kinase selectivity
- Allosteric sites
- Cryptic pockets
- Residence time

Mutation Effects

- Pathogenicity prediction
- Stability changes
- Function impact
- Evolutionary constraints
- Clinical interpretation

Evolution Modeling

- Sequence evolution
- Phylogenetic inference
- Ancestral reconstruction
- Coevolution
- Fitness landscapes

Part 3/3 - Applications

Design problems

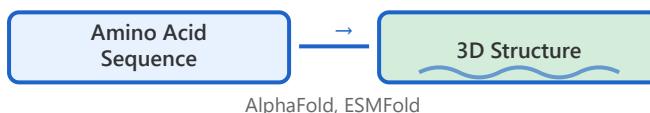
Engineering solutions

Therapeutic development

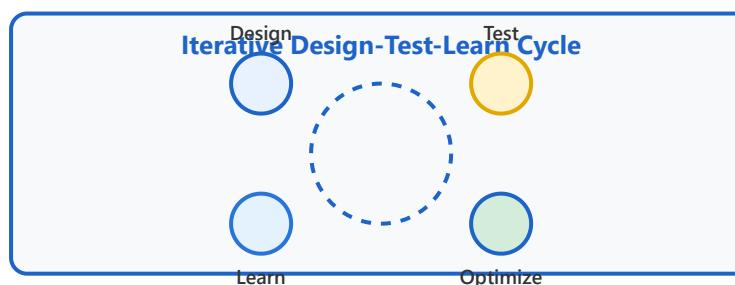
Protein Design

Inverse Folding: Structure → Sequence

Forward Problem:



Inverse Problem (Design):



Inverse folding

Structure → sequence prediction

Scaffold design

De novo backbone generation

Interface design

Protein-protein interactions

De novo binders

Target-specific protein design

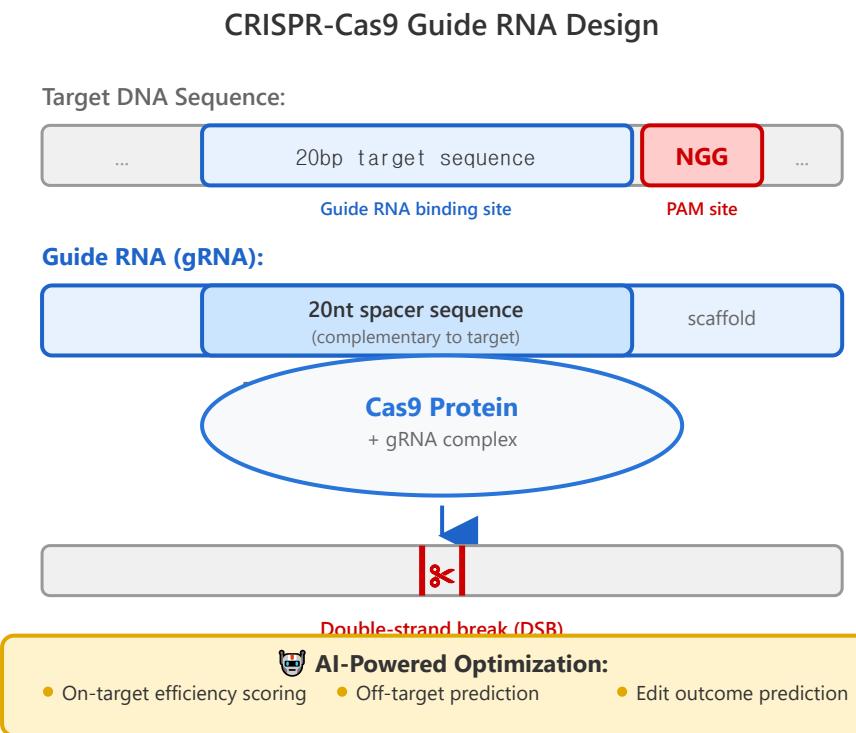
Stability optimization

Thermostability enhancement

Antibody Design

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

CRISPR Optimization



Guide RNA design

20nt spacer + scaffold optimization

Off-target prediction

Minimize unintended cuts

Efficiency scoring

On-target activity models

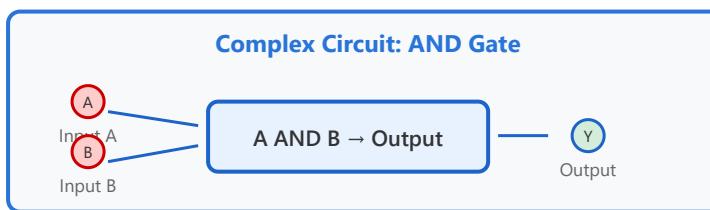
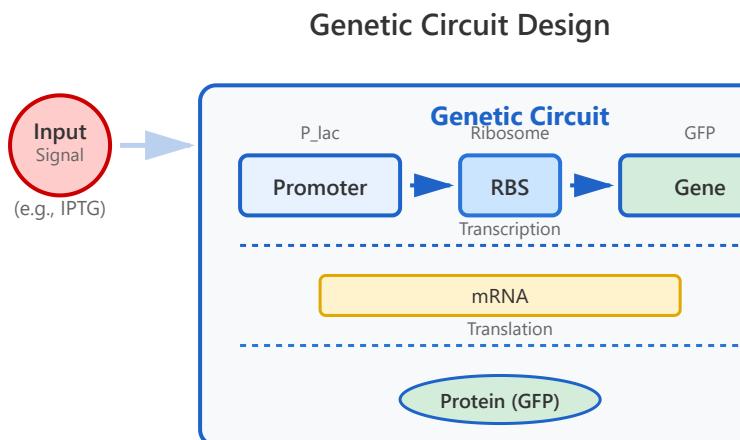
Prime editing

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

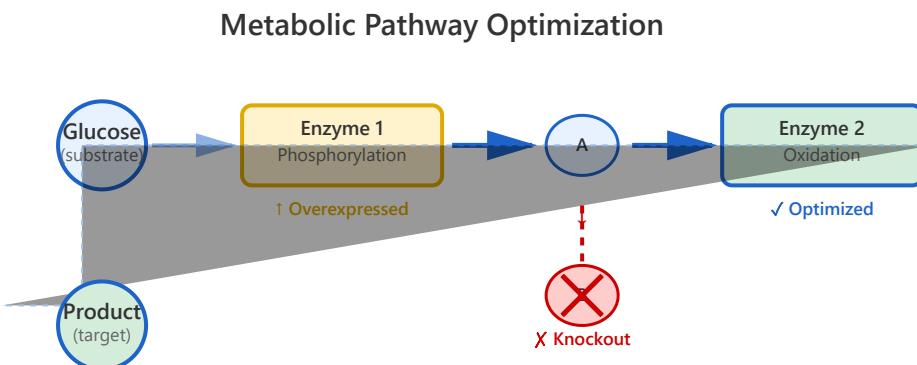
Orthogonal systems

Independent control modules

Predictive models

AI-guided circuit optimization

Metabolic Engineering



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
- Immunogenicity
- Coverage optimization
- Adjuvant selection
- mRNA design

Therapeutic Proteins

- Stability engineering
- Half-life extension
- Immunogenicity reduction
- Formulation prediction
- Manufacturing optimization

Enzyme Engineering

- Activity improvement
- Substrate specificity
- Thermostability
- Solvent tolerance
- Directed evolution

Future Perspectives

- Larger models
- Multi-modal learning
- Active learning
- Automated labs
- Closed-loop discovery

Limitations

- Data biases
- Generalization gaps
- Interpretability challenges
- Experimental validation
- Computational costs

Hands-on: AlphaFold Usage



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

Hands-on: Bio Transformers



- 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: