KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML) Workshop for Life and Medical Scientists

생명정보학 & 머신러닝 워크샵 (오프라인)

AI-based protein structure prediction and design

백민경 _ 서울대학교

KSBI KOREAN SOCIETY FOR BIOINFORMATICS | 한국생명정보학회

본 강의 자료는 한국생명정보학회가 주관하는 BIML 2024 워크샵 오프라인 수업을 목적으로 제작된 것으로 해당 목적 이외의 다른 용도로 사용할 수 없음을 분명하게 알립니다.

이를 다른 사람과 공유하거나 복제, 배포, 전송할 수 없으며 만약 이러한 사항을 위반할 경우 발생하는 **모든 법적 책임은 전적으로 불법 행위자 본인에게 있음을 경고**합니다.

KSBi-BIML 2024

Bioinformatics & Machine Learning(BIML)
Workshop for Life and Medical Scientists

안녕하십니까?

한국생명정보학회가 개최하는 동계 교육 워크샵인 BIML-2024에 여러분을 초대합니다. 생명정보학 분야의 연구자들에게 최신 동향의 데이터 분석기술을 이론과 실습을 겸비해 전달하고자 도입한 전문 교육 프로그램인 BIML 워크샵은 2015년에 시작하여 올해로 벌써 10년 차를 맞이하게 되었습니다. BIML 워크샵은 국내 생명정보학 분야의 최초이자 최고 수준의 교육프로그램으로 크게 인공지능과 생명정보분석 두 개의 분야로 구성되어 있습니다. 올해 인공지능 분야에서는 최근 생명정보 분석에서도 응용이 확대되고 있는 다양한 인공지능 기반 자료모델링 기법들에 대한 현장 강의가 진행될 예정이며, 관련하여 심층학습을 이용한 단백질구조예측, 유전체분석, 신약개발에 대한 이론과 실습 강의가 함께 제공될 예정입니다. 또한 단일세포오믹스, 공간오믹스, 메타오믹스, 그리고 롱리드염기서열 자료 분석에 대한 현장 강의는 많은 연구자의 연구 수월성 확보에 큰 도움을 줄 것으로 기대하고 있습니다.

올해 BIML의 가장 큰 변화는 최근 연구 수요가 급증하고 있는 의료정보자료 분석에 대한 현장 강의를 추가하였다는 것입니다. 특히 의료정보자료 분석을 많이 수행하시는 의과학자 및 의료정보 연구자들께서 본 강좌를 통해 많은 도움을 받으실 수 있기를 기대하고 있습니다. 또한 다양한 생명정보학 분야에 대한 온라인 강좌 프로그램도 점차 증가하고 있는 생명정보 분석기술의 다양화에 발맞추기 위해 작년과 비교해 5강좌 이상을 신규로 추가했습니다. 올해는 무료 강좌 5개를 포함하여 35개 이상의 온라인 강좌가 개설되어 제공되며, 연구 주제에 따른 연관된 강좌 추천 및 강연료 할인 프로그램도 제공되며, 온라인을 통한 Q&A 세션도 마련될 예정입니다. BIML-2024는 국내 주요 연구중심 대학의 전임 교원이자 각 분야 최고 전문가들의 강의로 구성되었기에 해당 분야의 기초부터 최신 연구 동향까지 포함하는 수준 높은 내용의 강의가 될 것이라 확신합니다.

BIML-2024을 준비하기까지 너무나 많은 수고를 해주신 운영위원회의 정성원, 우현구, 백대현, 김태민, 김준일, 김상우, 장혜식, 박종은 교수님과 KOBIC 이병욱 박사님께 커다란 감사를 드립니다. 마지막으로 부족한 시간에도 불구하고 강의 부탁을 흔쾌히 허락하시고 훌륭한 현장 강의와 온라인 강의를 준비하시는데 노고를 아끼지 않으신 모든 강사분들께 깊은 감사를 드립니다.

2024년 2월

한국생명정보학회장 이 인 석

강의 시간표

DAY1 : 2월 24일 (토)

시간	강 의 (자연과학대학 28동 101호)
12:30-12:50	등록
12:50-13:00	공지사항 전달
13:00-14:30	**의료빅데이터/인공지능 총론** 김헌성 교수(가톨릭대학교)
14:30-14:45	휴식
14:45-16:15	**의료영상 인공지능의 이해 및 의료영상 레이블링 실습** 백서연 교수(연석대학교)
16:15-16:30	휴식
16:30-18:00	**의료 정보처리 자동화 실습 / 독자적인 어플리케이션 만들기** 김선근 대표(원닥 주식회사), 서사도 조교

시간	강 의 (자연과학대학 28동 102호)
12:30-12:50	등록
12:50-13:00	공지사항 전달
13:00-14:20	**EMR 데이터를 활용한 머신러닝 기반 예후예측:** **Decision Tree-based Models + EMR 샘플 데이터 실습 (MIMIC sample dataset)** 고태훈 교수(가톨릭대학교)
14:20-14:40	휴식
14:40-16:00	**Chest X-ray 영상을 활용한 딥러닝 기반 폐질환 진단:** **Convolutional Neural Network + 의료영상 샘플 데이터 실습 (NIH Chest X-ray14)** 고태훈 교수(가톨릭대학교)
16:00-16:20	휴식
16:20-17:40	**심전도 데이터를 활용한 딥러닝 기반 부정맥 탐지: Recurrent Neural Network +** **Transformer + 심전도 샘플 데이터 실습 (MIT-BIH Arrhythmia Database)** 고태훈 교수(가톨릭대학교)

DAY1 : 2월 26일 (월)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	**DNN (이론)** 이상근 교수(고려대학교)
10:50-11:00	휴식
11:00-12:10	**CNN (이론)** 이상근 교수(고려대학교)
12:10-13:40	점심
13:40-15:10	**RNN, ChatGPT, XAI (이론)** 이상근 교수(고려대학교)
15:10-15:20	휴식
15:20-16:50	**CNN/RNN 모델 구조 정의, 학습 알고리즘 적용, 성능 평가, 시각화 방법 (Tensorflow 실습)** 이정현 조교, 한성민 조교

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	**Best practice for single-cell data analysis** 박종은 교수(KAIST)
11:00-11:10	휴식
11:10-12:40	**Practice1: Scanpy basic workflow** 정성민 조교, 고용준 조교
12:40-14:10	점심
14:10-15:30	**Public database, data integration, reference mapping, multiomics** 박종은 교수(KAIST)
15:30-15:40	휴식
15:40-16:50	**Practice2: Advanced single-cell analysis (siVI universe)** 정성민 조교, 고용준 조교

DAY1 : 2월 27일 (화)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	**AI-based protein structure prediction** **- Intro to protein structure prediction** **- Early AI-based approaches** **- AlphaFold and RoseTTAFold** 백민경 교수(서울대학교)
10:50-11:00	휴식
11:00-12:10	**단백질 구조 예측 실습** **- ColabFold를 활용한 단백질 구조 및 상호작용 예측** **- Tips &Tricks for better structure modeling** 백민경 교수(서울대학교)
12:10-13:40	점심
13:40-15:10	**AI-based protein design** **- Intro to protein design** **- Protein backbone design using RFdiffusion** **- Protein sequence design using ProteinMPNN** 백민경 교수(서울대학교)
15:10-15:20	휴식
15:20-16:50	**단백질 디자인 실습** **- RFdiffusion 및 ProteinMPNN의 활용법 실습** 백민경 교수(서울대학교)

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	**Introduction to Single-cell biology** 최정민 교수(고려대학교)
11:00-11:10	휴식
11:10-12:40	**i. Unsupervised Spatial transcriptome analysis** **ii. Tumor Boundary Determination in Spatial Transcriptomics** 유광민 조교, 이문영 조교
12:40-14:10	점심
14:10-15:30	**i. Deconvolution Analysis Using Single-cell RNA Sequencing and Spatial Transcriptomics** **ii. Cell-Cell Interaction Analysis in Spatial Transcriptomics** 김지현 조교, 최승지 조교
15:30-15:40	휴식
15:40-16:50	**i. Open Chromatin Region Analysis and Biological Interpretation of Using scATAC-seq Dataset** **ii. Construction of Gene Regulatory Networks Based on Integrated Analysis of scATAC-seq and scRNA-seq Datasets** 천하림 조교, 이호진 조교

DAY1 : 2월 28일 (수)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	**Introduction to Transformers (이론)** 전민지 교수 (고려대학교)
11:00-11:10	휴식
11:10-12:40	**Introduction to Transformers (실습)** 봉현수 조교, 임우택 조교
12:40-14:10	점심
14:10-15:40	**Deep learning in Bioinformatics** 노미나 교수(한양대학교)
15:40-15:50	휴식
15:50-17:20	**Deep learning model을 이용한 실습** 박예솔 조교

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	**마이크로바이옴 기본 이론** 이선재 교수(GIST)
10:50-11:00	휴식
11:00-12:10	**16S rRNA amplicon seq. - DADA2** 조준우 조교, 백재우 조교
12:10-13:40	점심
13:40-14:40	**최신 메타지놈 분석 기법의 현황** 이선재 교수(GIST)
14:40-14:50	휴식
14:50-16:50	**Shotgun metagenome 분석 (Linux)** 조준우 조교, 백재우 조교

DAY1 : 2월 29일 (목)

시간	강 의 (자연과학대학 28동 101호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-10:50	**화학정보학 기초(Cheminformatics) / 약물특성 및 약물다움(druglikeness)** **Molecular Notations &Descriptors / AI 신약개발을 위한 Databases** **AI 신약개발을 위한 Programming 기초** 김동섭 교수(KAIST)
10:50-11:00	휴식
11:00-12:10	**Google Colab에 RDKit 설치 / 화합물 정보 읽기 실습** **Bioactivity database 검색 및 정보 읽기 실습** **Molecular descriptor (fingerprint) 생성 및 similarity 계산 실습** 정수재 조교, 나민주 조교
12:10-13:40	점심
13:40-15:10	**AI 신약개발을 위한 기계학습법 기초 / QSAR 모델링 기초 / AI 신약개발을 위한 딥러닝 모델** **Virtual screening (ligand-based, structure-based) 및 de novo design** 김동섭 교수(KAIST)
15:10-15:20	휴식
15:20-16:50	**QSAR modeling 전체 과정 실습/ 화합물의 Bioactivity 예측 모델 개발** **Virtual screening 과정을 통한 신약후보물질 발굴 실습** 정수재 조교, 나민주 조교

시간	강 의 (자연과학대학 28동 102호)
09:00-09:20	등록
09:20-09:30	공지사항 전달
09:30-11:00	**Single cell multiomics 이론 / Gene regulatory network 이론** 김준일 교수(숭실대학교)
11:00-11:10	휴식
11:10-12:40	**Seurat/Signac, ArchR, TENET+ 실습** 김현규 조교, 정회빈 조교
12:40-14:10	점심
14:10-15:40	**롱리드 시퀀싱 소개 및 유전체 조립 실습** 김준 교수(충남대학교)
15:40-15:50	휴식
15:50-17:20	**변이 분석 및 시각화 실습** 김준 교수(충남대학교)

강의개요

AI-based protein structure prediction and design

단백질은 신호전달, 대사, 면역 등 우리 몸에서 일어나는 거의 모든 생명현상에 관여하고 있는 중요한 생체분자이다. 단백질은 각자의 기능을 수행하기 적합한 3차원 구조를 가지고 있으며, 이러한 구조는 단백질의 서열에 따라 결정되는 것으로 알려져 있다. 즉, 단백질의 기능을 잘 이해하기 위해서는 서열로부터 그 구조를 아는 것이 매우 중요하다. 단백질의 서열을 기반으로 그 3차원 구조를 정확하게 예측할 수 있다면 단백질과 연관된 수많은 생명현상에 대한 답을 찾는데 큰 도움을 주지 않을까? 본 강의에서는 단백질 구조 예측 방법이 어떻게 발전해왔는지 살펴보고, 인공지능이 단백질 구조 예측에 어떤 혁신을 가져왔는지 알아보고자 한다. 또한 인공지능 기반의 단백질 구조 예측이 단백질-단백질 상호작용 예측, 단백질 디자인과 같은 다른 연구분야에 어떤 영향을 주었는지 살펴본다. 강의에서 다루는 방법들을 실제 실습을 통해 사용해보고, 각 방법의 장단점을 알아보고자 한다.

강의는 다음의 내용을 포함한다:

- 단백질 구조 예측의 기본 원리
- 인공지능을 활용한 단백질 구조 예측
- 단백질-단백질 상호작용에의 응용
- 단백질 디자인으로의 응용

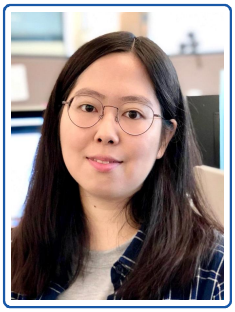
* 교육생준비물: 노트북

* 강의 난이도: 중급

* 강의: 백민경 교수 (서울대학교 생명과학부)

Curriculum Vitae

Speaker Name: Minkyung Baek, Ph.D.



▶ **Personal Info**

Name	Minkyung Baek
Title	Assistant Professor
Affiliation	Seoul National University

▶ **Contact Information**

Address	504-523, 1 Gwanak-ro, Gwanak-gu, Seoul 08826
Email	minkbaek@snu.ac.kr
Phone Number	02-880-6755

Research Interest

Structural bioinformatics, computational biology, protein structure prediction, artificial intelligence

Educational Experience

2013	B.S. in Chemistry, Seoul National University, Korea
2018	Ph.D. in Chemistry, Seoul National University, Korea

Professional Experience

2018-2019	Postdoctoral researcher, Seoul National University
2019-2022	Postdoctoral scholar, University of Washington, USA
2022-	Assistant Professor, Seoul National University

Selected Publications (5 maximum)

1. Minkyung Baek, et al., Accurate prediction of protein structures and interactions using a three-track neural network, Science, 373 (6557), 2021.
2. Ian R. Humphreys†, Jimin Pei†, Minkyung Baek†, Aditya Krishnakumar†, et al., Computed structures of core eukaryotic protein complexes, Science, 374 (6573), 2021. (†co-first authors)
3. Minkyung Baek, Ivan Anishchenko, Hahnbeom Park, Ian R. Humphreys, and David Baker, Protein oligomer modeling guided by predicted inter-chain contacts in CASP14, Proteins: Structure, Function, and Bioinformatics, 89 (12), 2021.
4. Ivan Anishchenko†, Minkyung Baek†, Hahnbeom Park†, et al., Protein tertiary structure prediction and refinement using deep learning and Rosetta in CASP14, Proteins: Structure, Function, and Bioinformatics, 89 (12), 2021. (†co-first authors)

KSBI | 한국생명정보학회

KSBi-BIML 2024

AI-based protein structure prediction and design

서울대학교 생명과학부
백민경
(minkbaek@snu.ac.kr)

제 1강.

Protein Structure Prediction in the era of AI

강의의 취지

ChatGPT가 표현한 "AlphaFold에 대한 연구자들의 반응"

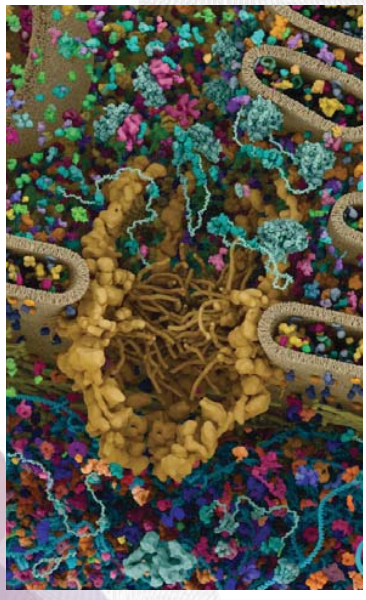
단백질 구조예측 AI의 원리를
이해하고 쓰자!!

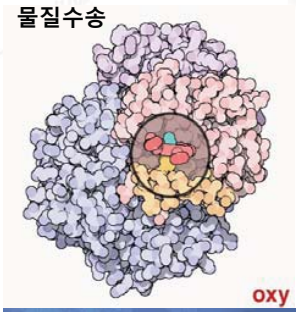
이 예측 결과를 믿어도 되나?

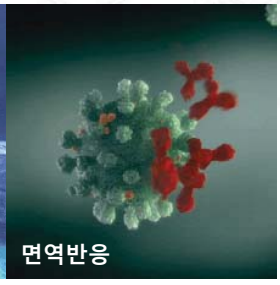
예측이 좀 아쉬운데..
어떻게 심폐소생을 할 수 있을까?

단백질?

생명현상의 핵심 분자

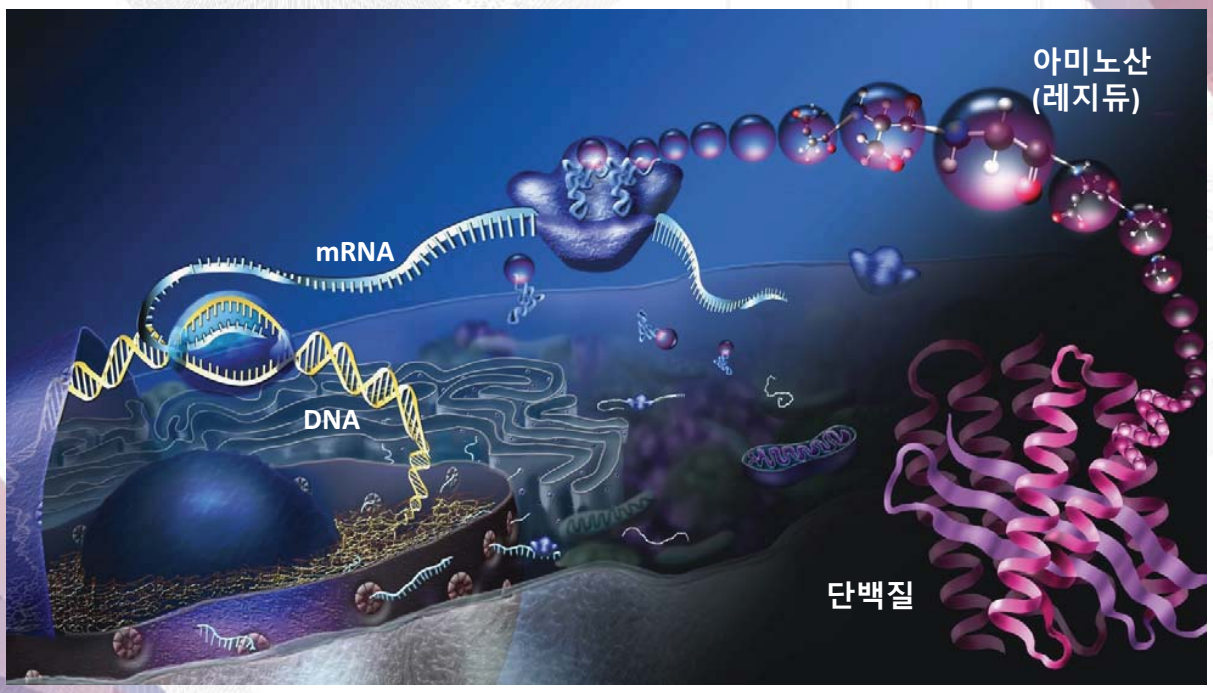






6

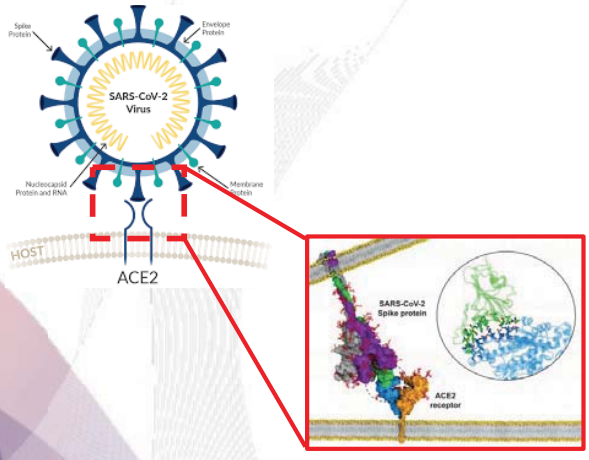
Central Dogma



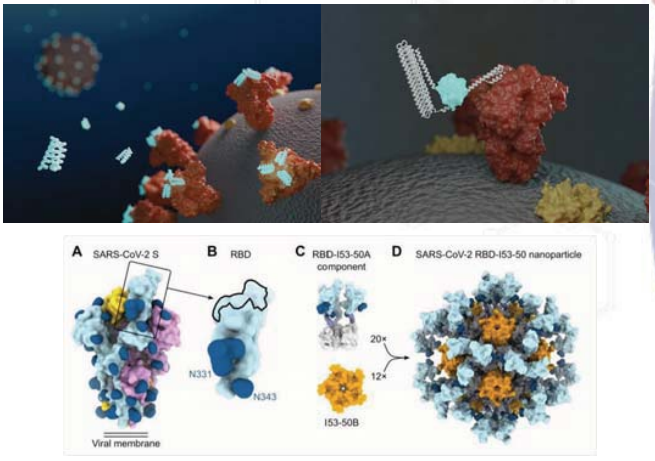
7

단백질의 구조를 아는 것이 중요한 이유?

생명현상에 대한 더욱 깊은 이해



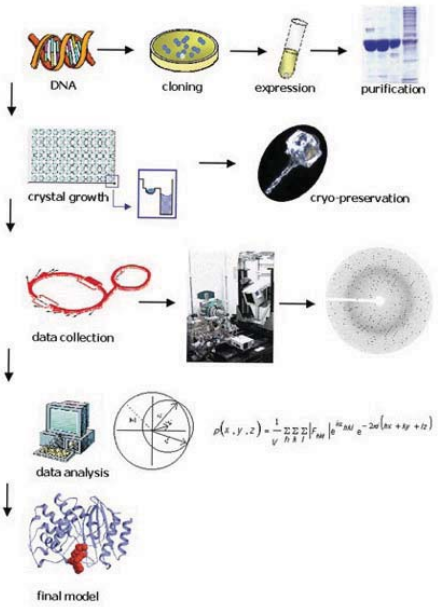
신약/백신/바이오센서 개발로의 응용



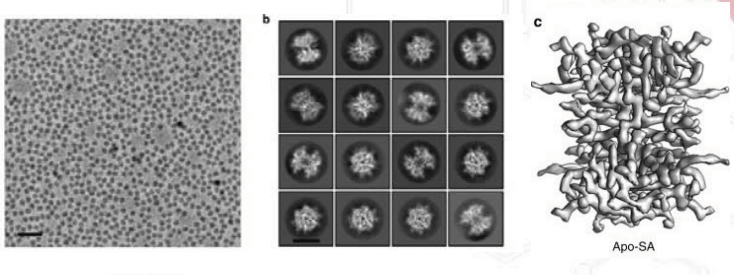
8

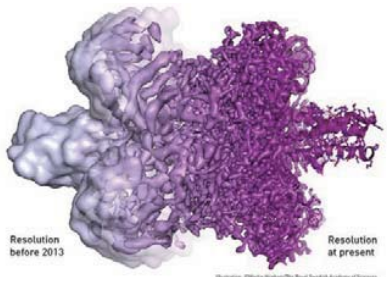
실험을 통한 단백질 구조 결정

X-ray 결정법 (노벨화학상, 1962년)



극저온전자현미경 (노벨화학상, 2017년)

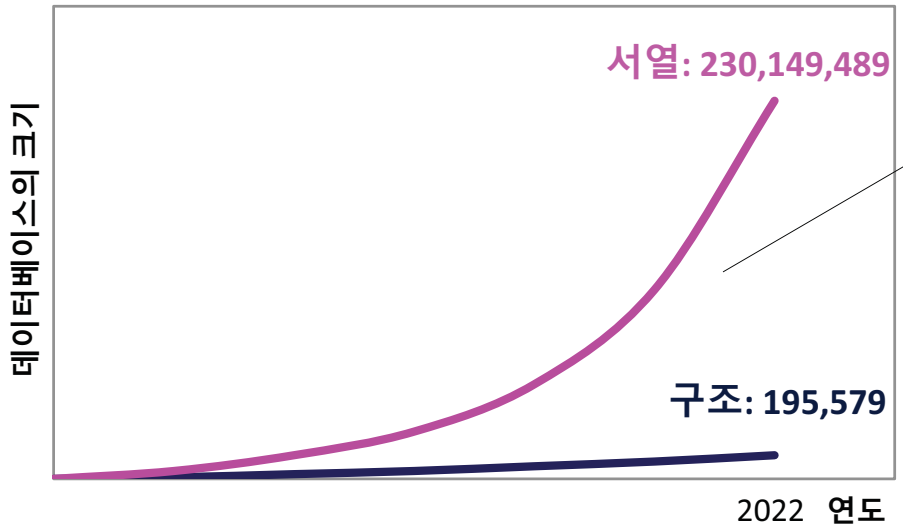




비싸다
시간이 오래 걸린다
(수개월 ~ 수년)

알고있는 단백질의 서열 >> 단백질 구조

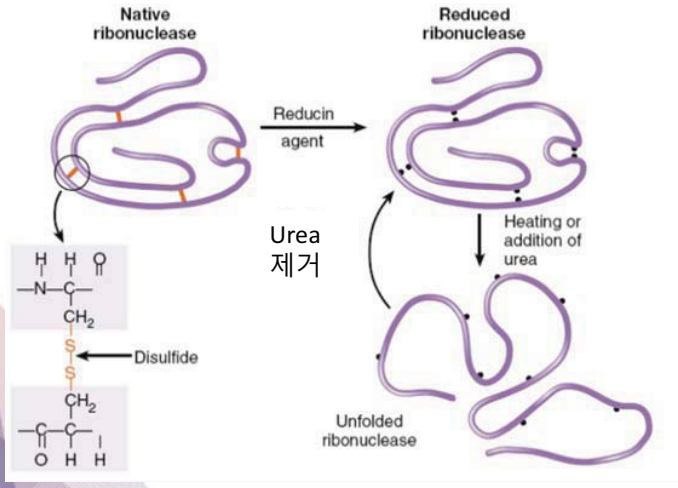
단백질 서열/구조 데이터베이스 크기 비교



컴퓨터 계산을 통한
단백질 구조 예측?

컴퓨터 계산을 통해 예측할 수 있을까?

Anfinsen's experiment



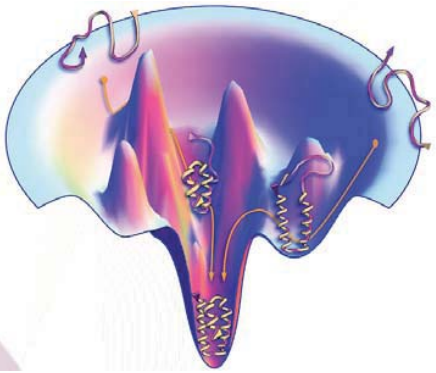
Christian B. Anfinsen
(노벨화학상, 1972)

"단백질의 3차원 구조는 단백질의 서열에 의해 결정된다"

단백질 구조 예측의 가능성!
(50여년간 다양한 예측방법 개발)

Protein folding simulation with physical principle

The view of chemist: The most stable (Free Energy minimum) state



주어진 환경에서 가장 안정한 구조

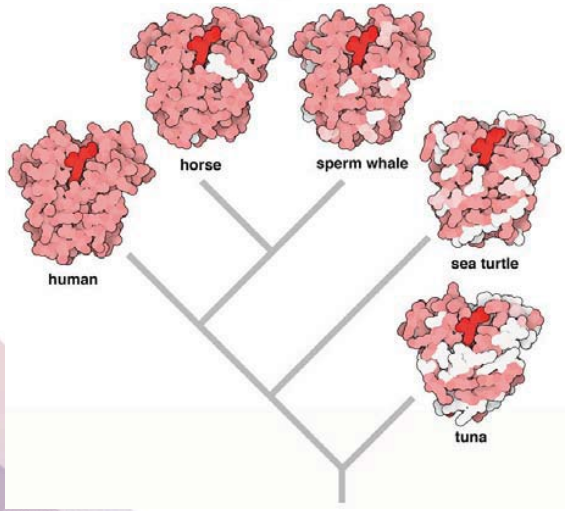
Protein Folding MD simulation

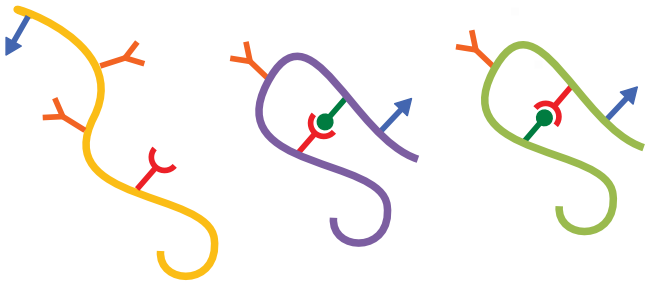
Only works for very small proteins (<50 aa)
Too slow (days to months)
Energy function (force field) also has errors

https://youtu.be/jVOyaT56LEU?si=I2Kr2BNdSW1qnyiv

Protein structure information in evolutionary history

Structure of Myoglobin

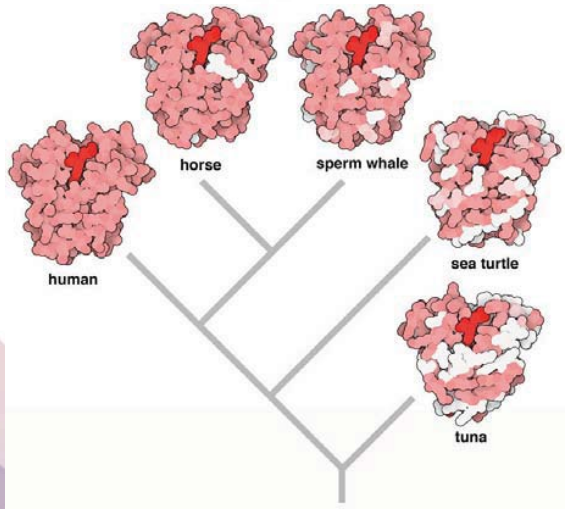




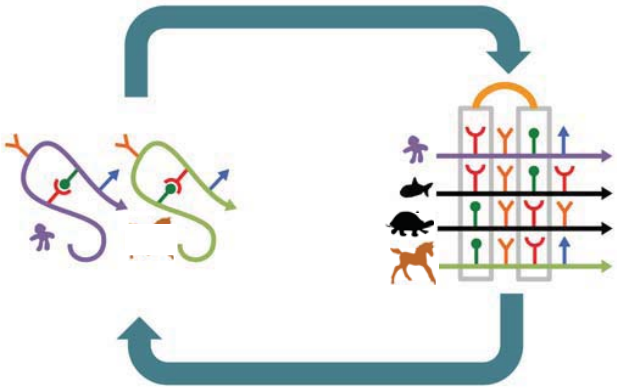
Loss of function | **Normal Protein** | **Maintain its function**

Protein structure information in evolutionary history

Structure of Myoglobin



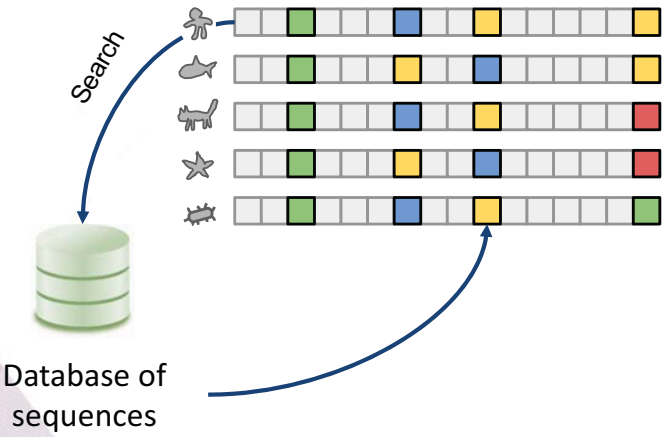
Restraints to maintain
its function & structures



Structural patterns in MSA

"Coevolution" of residue pairs

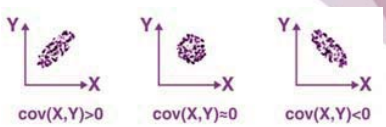
Coevolution-guided protein structure modeling

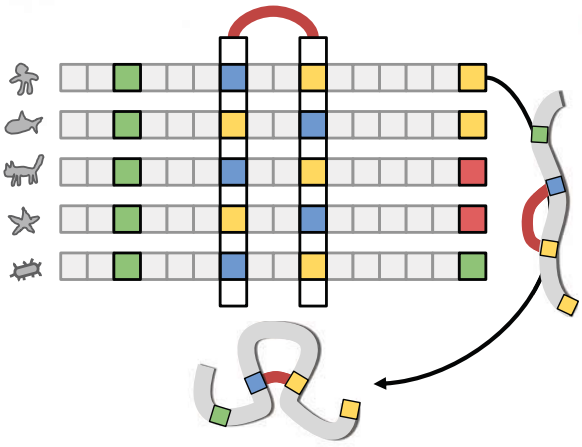


Slide credit: Sergey Ovchinnikov
citations: tinyurl.com/coevopapers

Coevolution-guided protein structure modeling

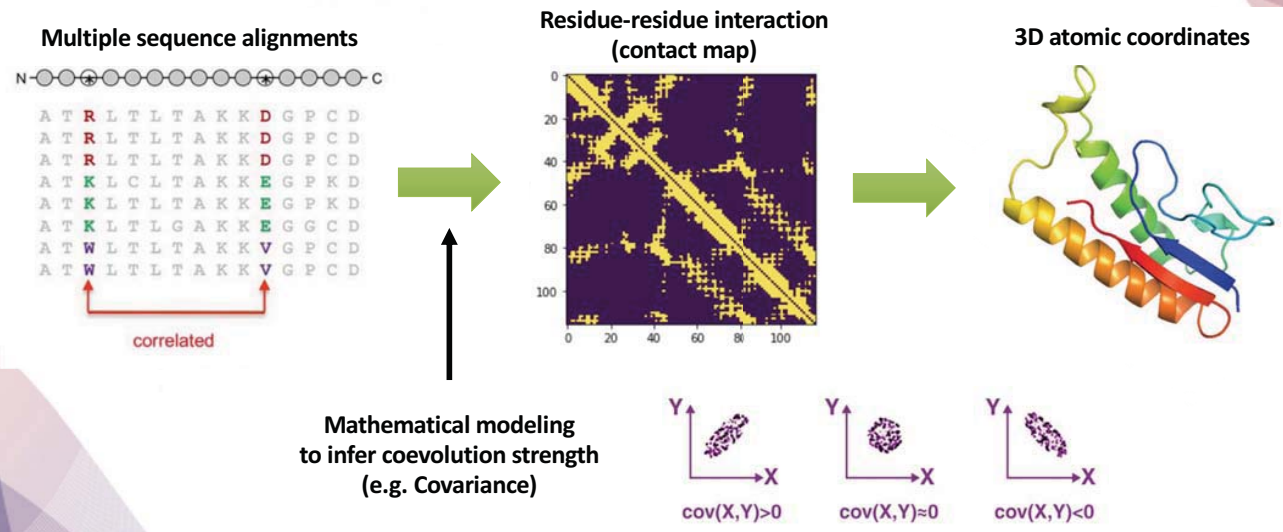
Mathematical modeling to infer Coevolution (e.g. covariance)



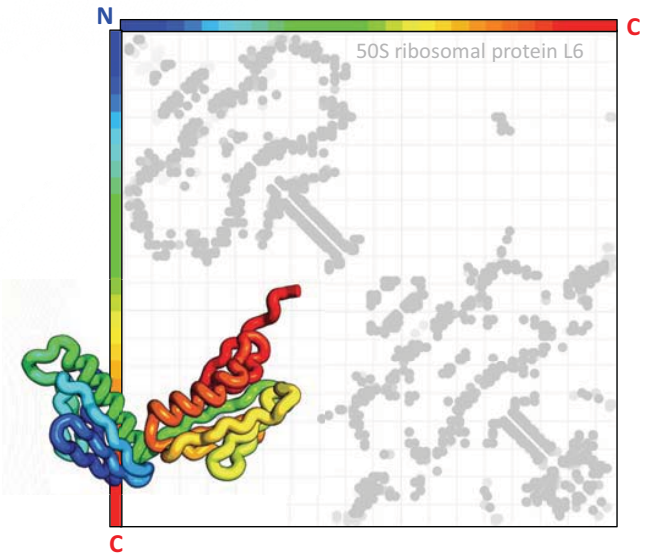


Slide credit: Sergey Ovchinnikov
citations: tinyurl.com/coevopapers

Coevolution-guided protein structure modeling



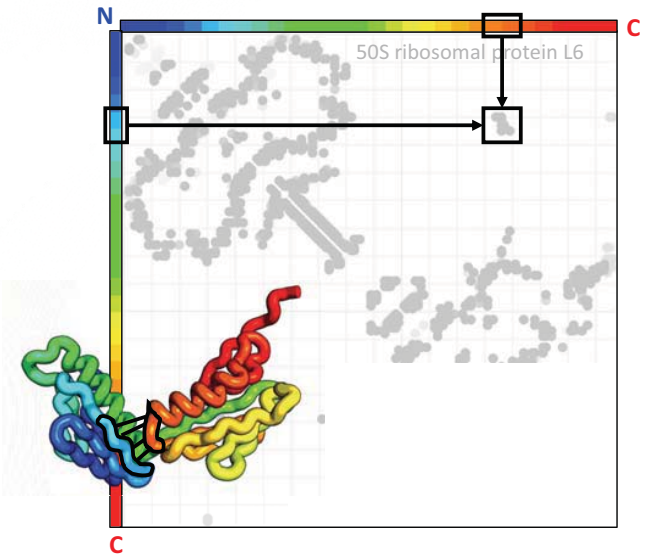
How to read a contact map?



XRAY Contacts

Slide credit: Sergey Ovchinnikov

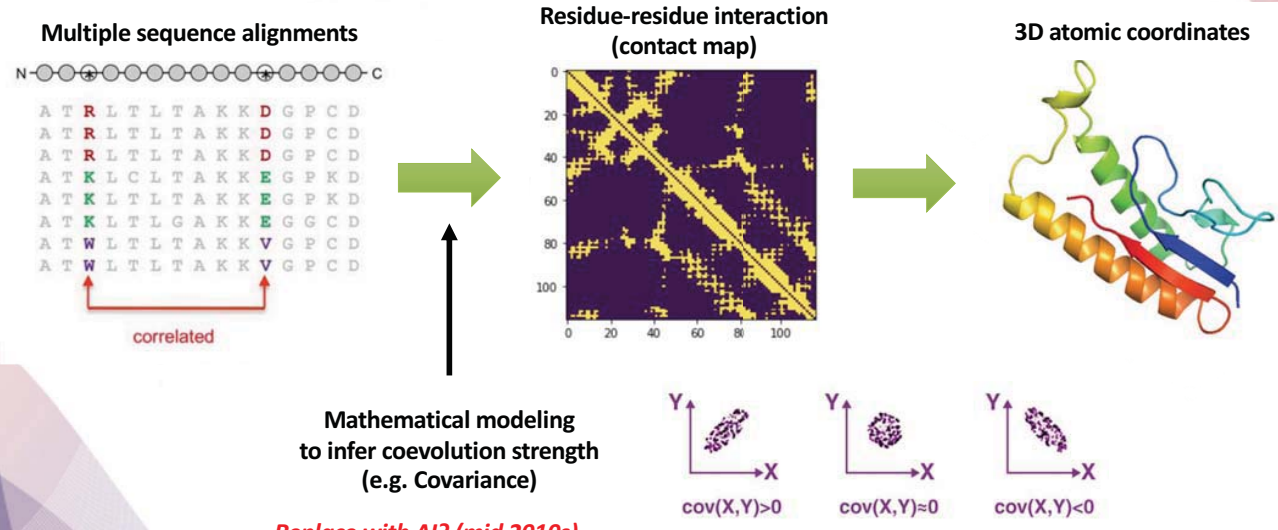
How to read a contact map?



XRAY Contacts

Slide credit: Sergey Ovchinnikov

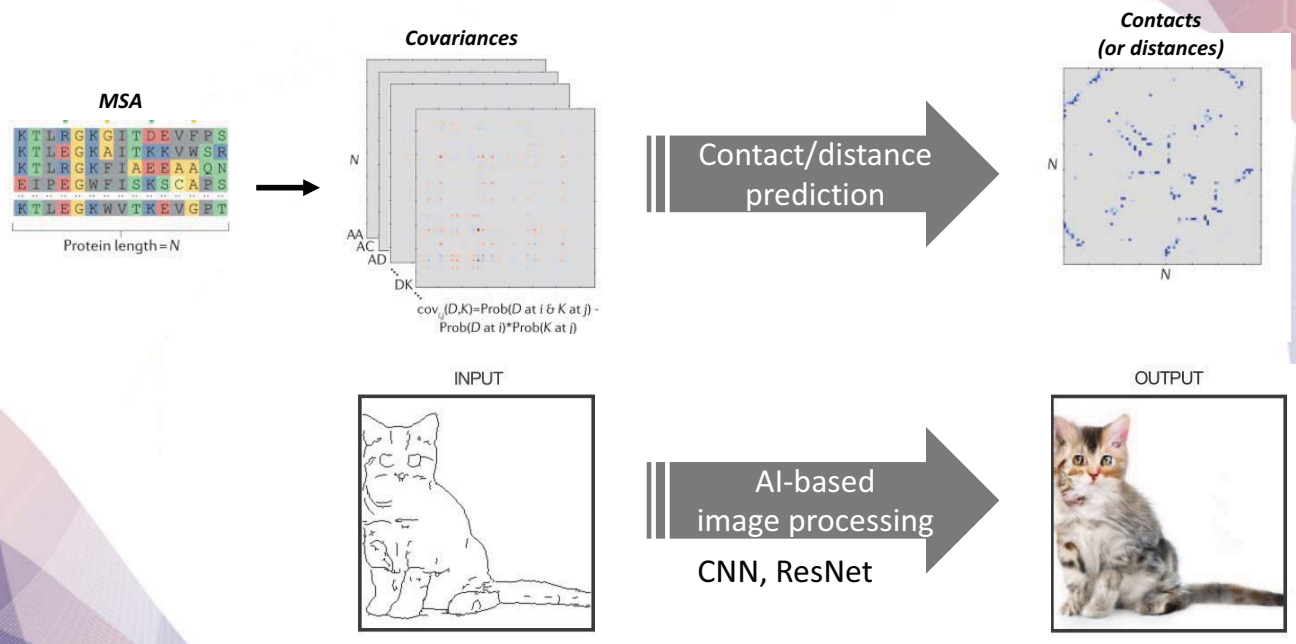
Coevolution-guided protein structure modeling



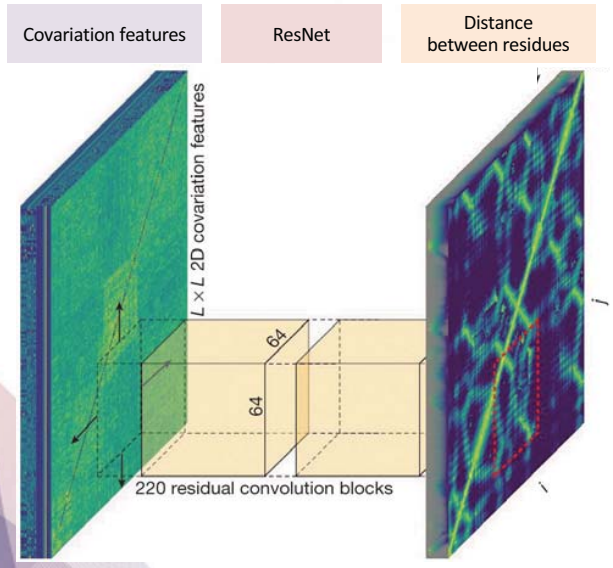
Applying AI to protein structure prediction

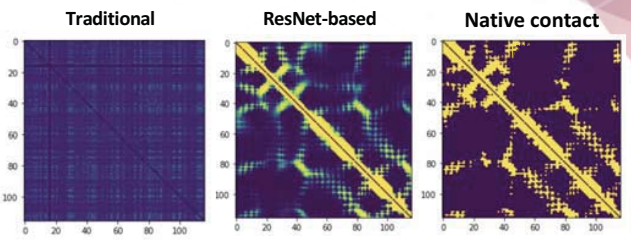
- AI tries to **mimic human's learning/thinking process**
  - Better implementation of human's intuition → better performance!
- Exceptionally effective at **learning patterns**
- If you provide the system **tons of training examples**, it begins to understand hidden patterns in data and respond in useful ways

Early attempt: Contact prediction = Image processing

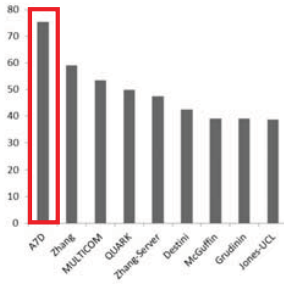


The beginning of legend: AlphaFold (2018)

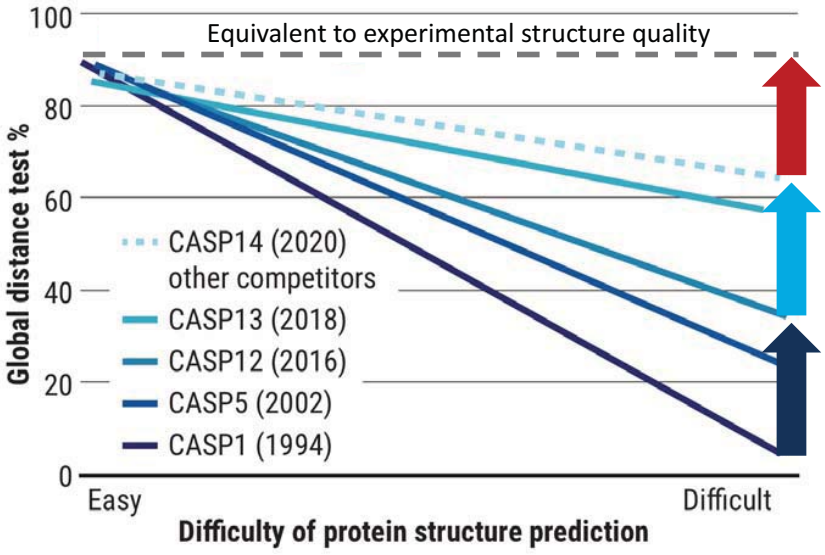




Protein structure prediction performance in CASP13 (2018)



Progress in protein structure prediction



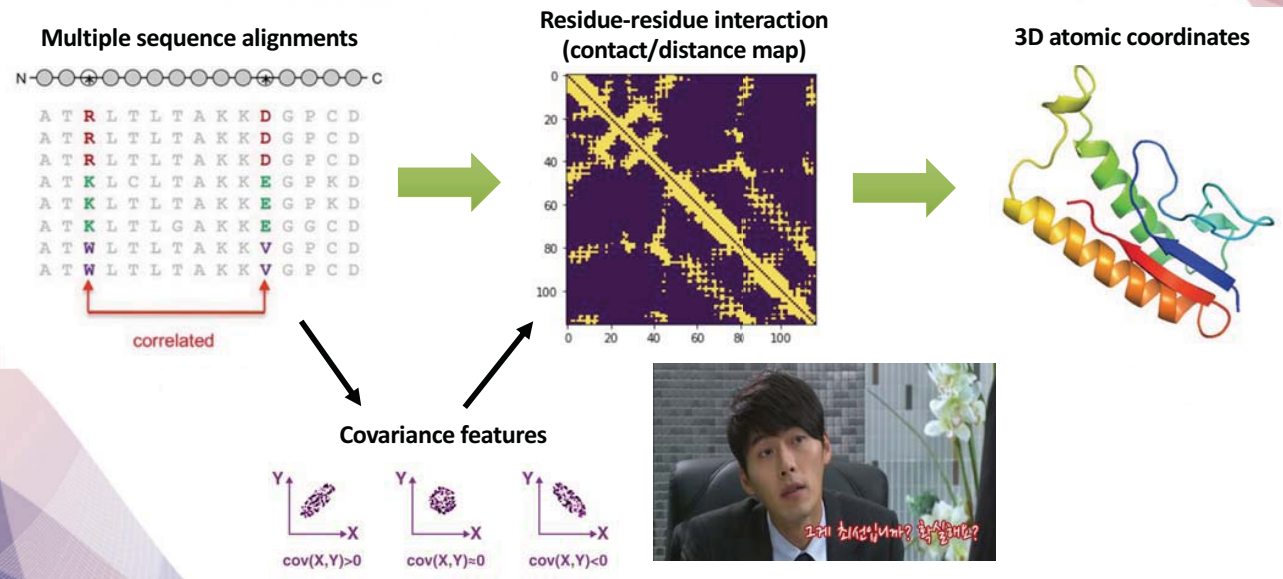
+ ????

+ Early AI methods (mainly ResNet)

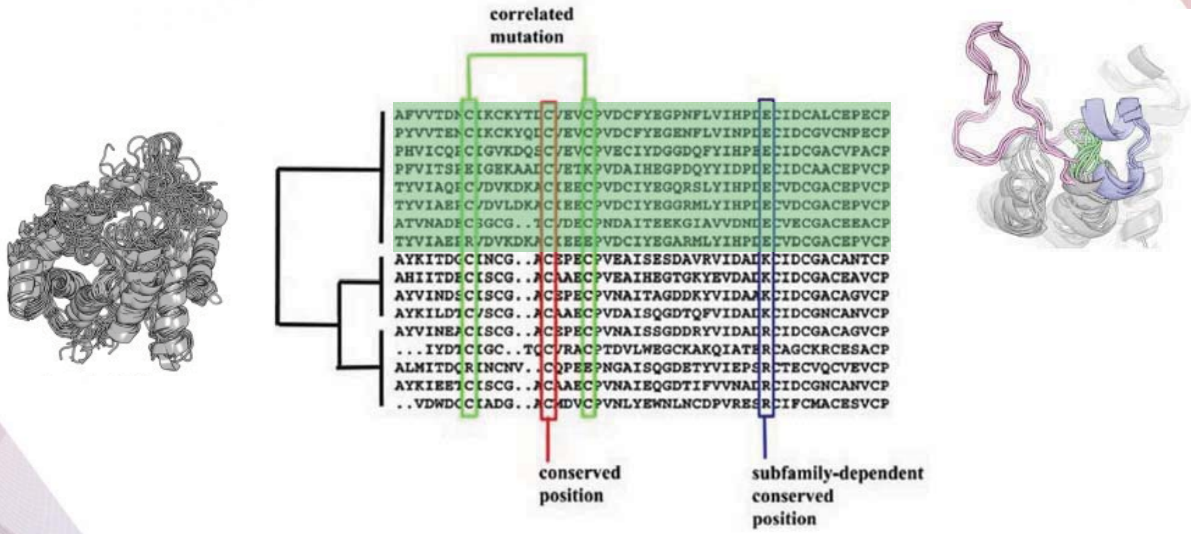
Improvements in traditional methods + Coevolution

Service, Robert F. "'The game has changed.'AI triumphs at protein folding." (2020): 1144-1145.

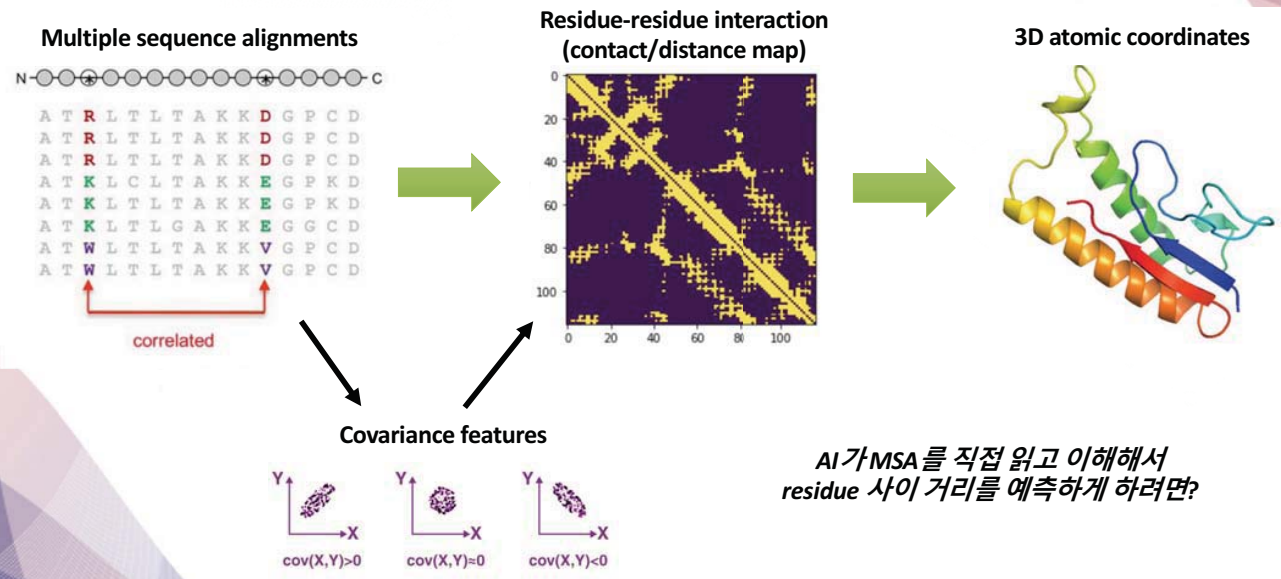
From MSA to 3D structures with AI



There is more structural information besides coevolution

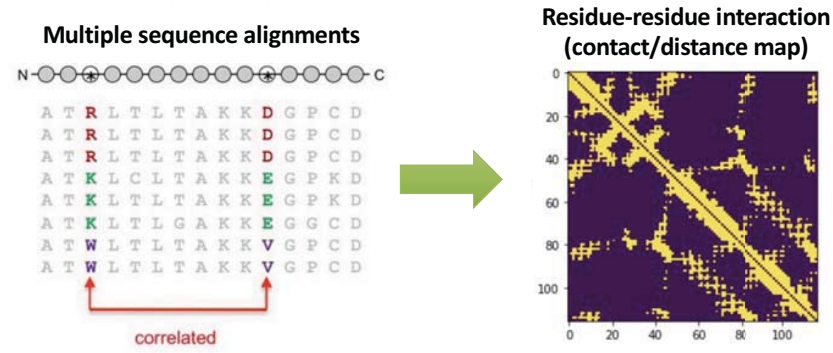


From MSA to 3D structures with AI



AI가 MSA를 직접 읽고 이해해서
residue 사이 거리를 예측하게 하려면?

Finding a similar, but simpler problem



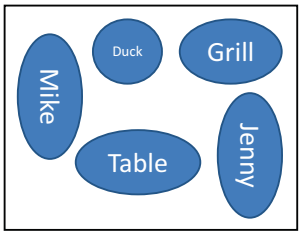
Mike is surprised at the duck.
The duck is standing on the grill.
Jenny is running towards Mike
and the duck. There is yellow
table between Mike and Jenny.

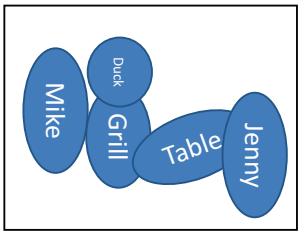
Analogy to text-to-image generation?

Mike is surprised at the duck. The duck is standing on the grill. Jenny is running towards Mike and the duck. There is yellow table between Mike and Jenny.

Mike is surprised at the duck. The duck is standing on the grill. Jenny is running towards Mike and the duck. There is yellow table between Mike and Jenny.

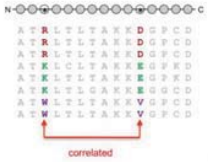
Mike is surprised at the duck. The duck is standing on the grill. Jenny is running towards Mike and the duck. There is yellow table between Mike and Jenny.





From MSA to 3D structures with AI

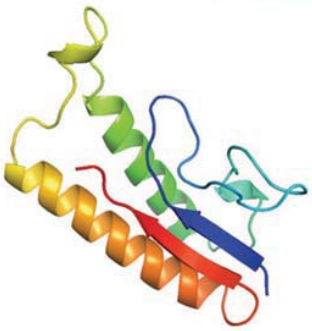
Extract structural information from MSA

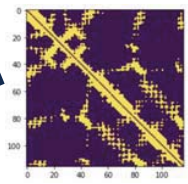


Transformer-based network
to understand information in MSA

Transformer-based network
to refine contact signal
& get better distance prediction

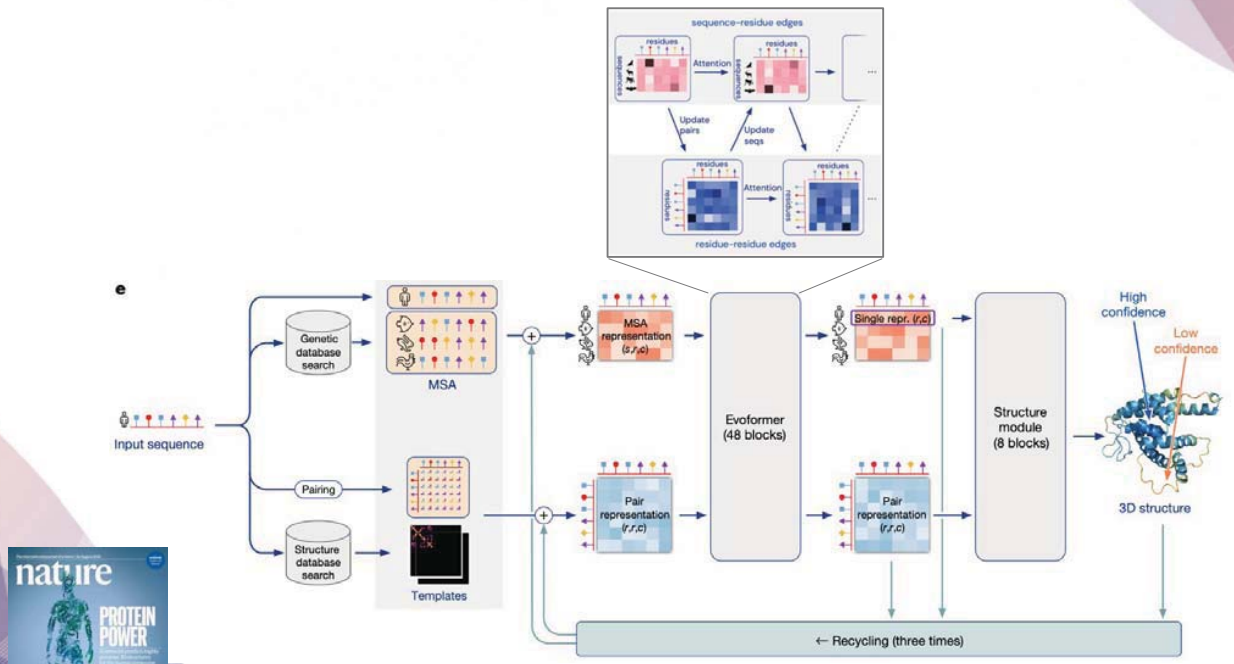
Structure Generator





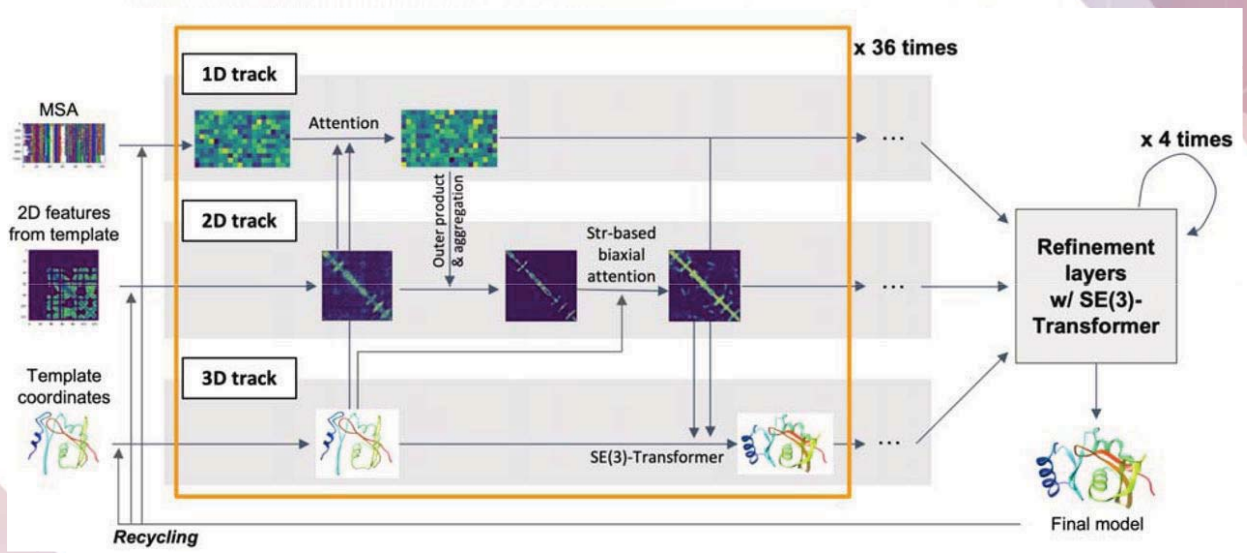
30

AI architecture of AlphaFold2



Jumper, J., et al, *Nature* (2021)

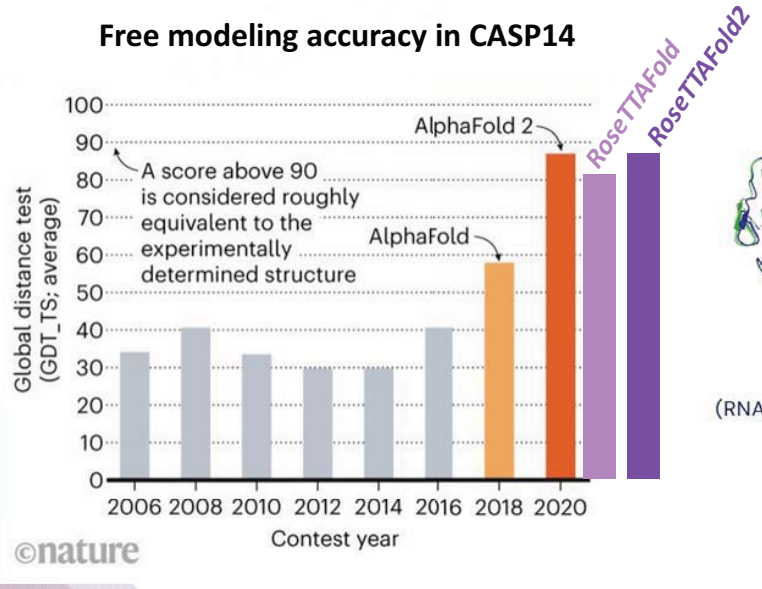
RoseTTAFold: A tighter connection of 1D, 2D, and 3D info



Baek, M., et al, *Science* (2021)
Baek, M. et al, *Under revision*

High accuracy protein structure prediction using AI

Free modeling accuracy in CASP14



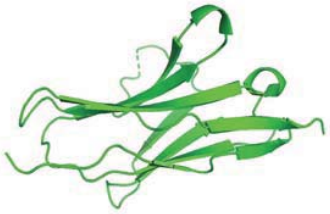
T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

T1049 / 6y4f
93.3 GDT
(adhesin tip)

Experimental result
Computational prediction

단백질의 구조만 알면 되나?

암세포에서
유난히 많이 발현되는 단백질



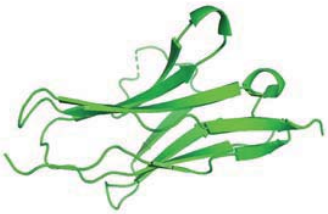
살해T세포에
존재하는 단백질

34

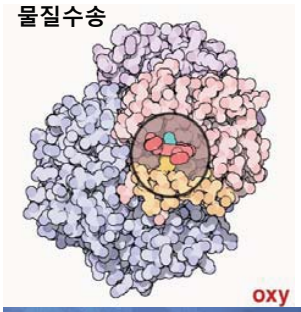
단백질의 구조만 알면 되나?

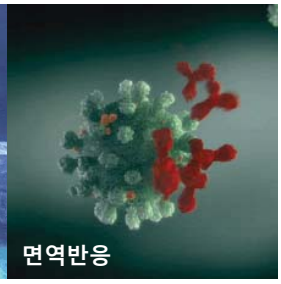
상호작용 예측이 더 중요!

암세포에서
유난히 많이 발현되는 단백질



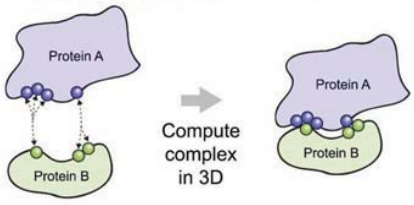
살해T세포에 존재하는 단백질

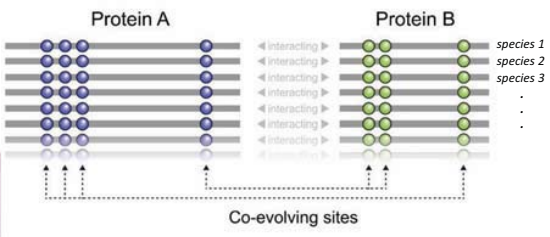


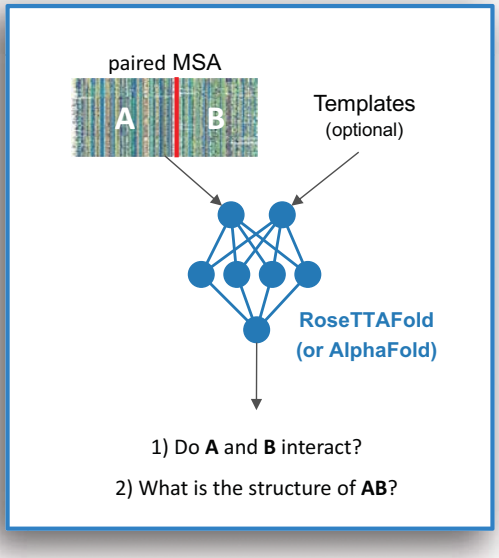


35

Protein-protein interaction prediction with protein structure prediction AI

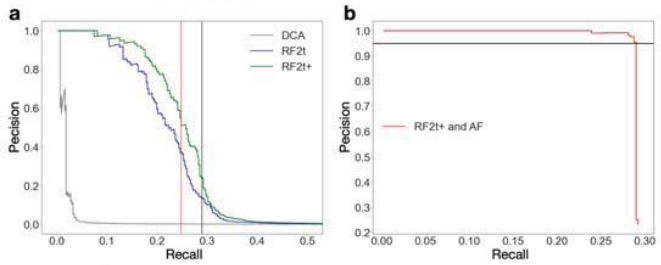




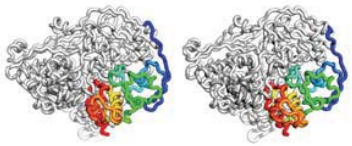


Protein-protein interaction prediction with protein structure prediction AI

PPI prediction performance

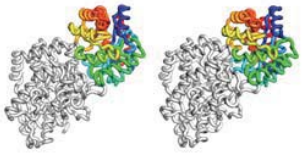


Aldehyde oxidoreductase

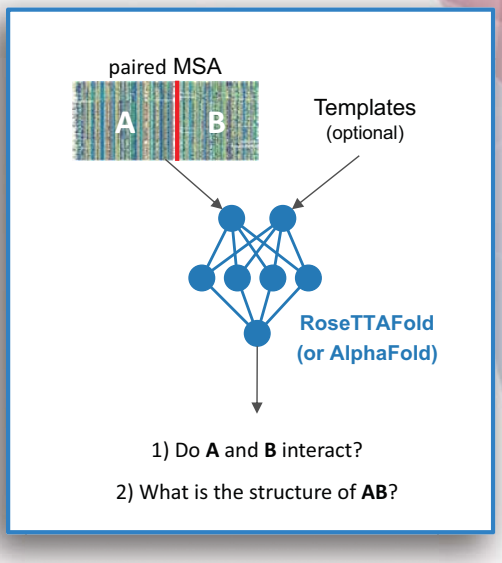


TM-score: 95

Tryptophan synthase

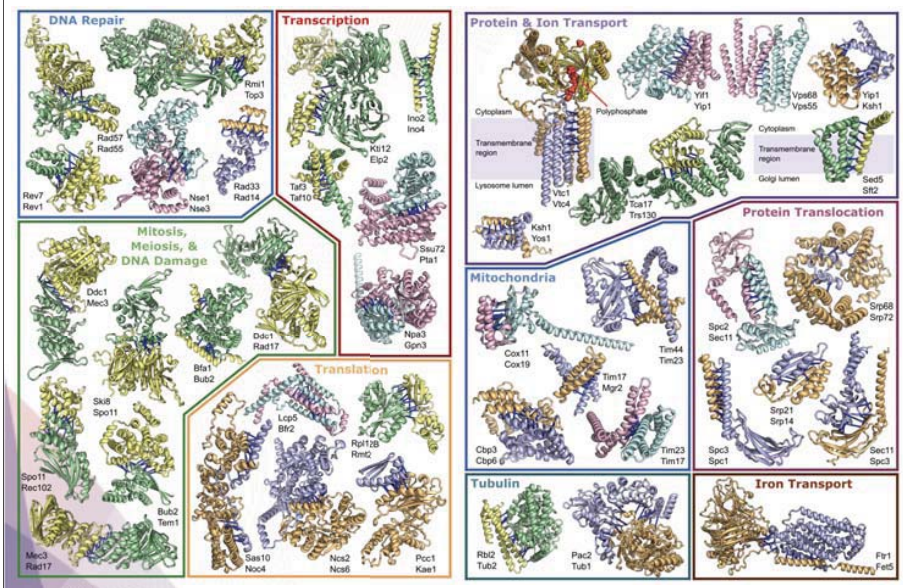


TM-score: 92

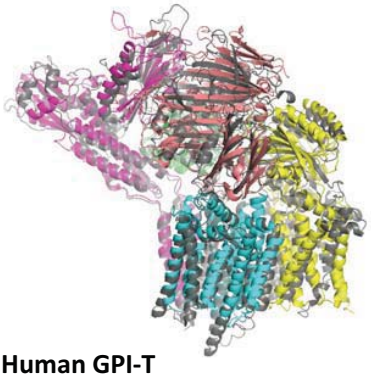


Humphreys, I.R., Pei, J., Baek, M., Krishnakumar, A., et al, *Science*, (2021)

Research example: *in silico* interactome screening



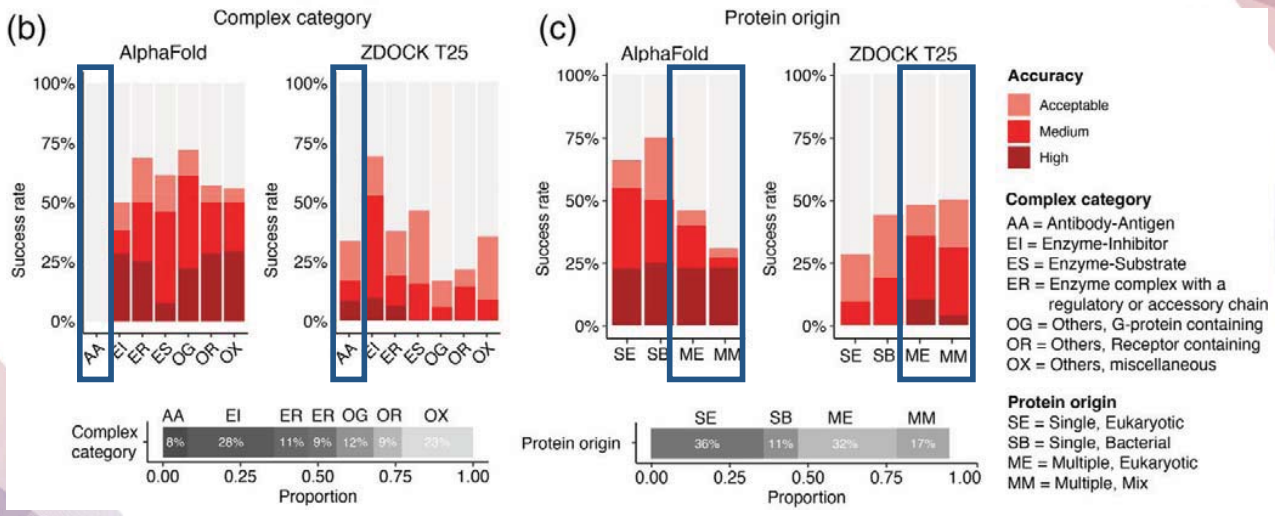
Yeast GPI-T
(our prediction, Oct 2021)



Human GPI-T
(PDB: 7W72, published Feb 2022)

Humphreys, I.R., Pei, J., Baek, M., Krishnakumar, A., et al, *Science*, (2021)

But, some interactions are still hard to predict…



R. Yin (2023), *Protein Science*

Let’s do some prediction with ColabFold!

ColabFold: Anyone can use protein structure prediction AI

https://github.com/sokrypton/ColabFold (Credit: Sergey Ovchinnikov, Martin Steinegger)

Making Protein folding accessible to all via Google Colab!

Notebooks	monomers	complexes	mmseqs2	jackhmmer	templates
AlphaFold2_mmseqs2	Yes	Yes	Yes	No	Yes
AlphaFold2_batch	Yes	Yes	Yes	No	Yes
AlphaFold2 (from Deepmind)	Yes	Yes	No	Yes	No
relax_amber (relax input structure)					
ESMFold	Yes	Maybe	No	No	No

ColabFold: Anyone can use protein structure prediction AI

https://github.com/sokrypton/ColabFold (Credit: Sergey Ovchinnikov, Martin Steinegger)

ColabFold: AlphaFold2 using MMseqs2

Easy to use protein structure and complex prediction using AlphaFold2 and Alphafold2-multimer. Sequence alignments/templates are generated through MMseqs2 and HHsearch. For more details, see bottom of the notebook, checkout the ColabFold GitHub and read our manuscript. Old versions: v1.0, v1.1, v1.2, v1.3

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: Making protein folding accessible to all. *Nature Methods*, 2022

Input protein sequence(s), then hit `Runtime -> Run all`

query_sequence: PIAQIHILEGRSDEQKETLIREVSEAISRSLDAPLTSVRVIITEMAKGHFGIGGELASK

- Use `:` to specify inter-protein chainbreaks for **modeling complexes** (supports homo- and hetro-oligomers). For example **PI...SK:PI...SK** for a homodimer

jobname: test

use_amber:

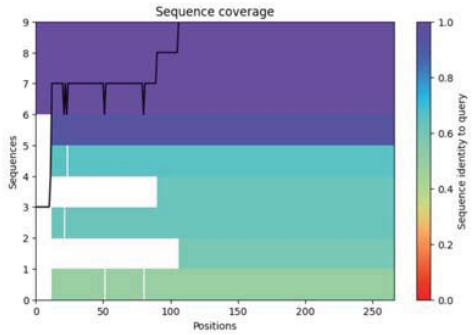
template_mode: none

- "none" = no template information is used, "pdb70" = detect templates in pdb70, "custom" - upload and search own templates (PDB or mmCIF format, see notes below)

Show code

How to interpret prediction results

1. Quality of MSAs:
# of sequences, diversity, etc

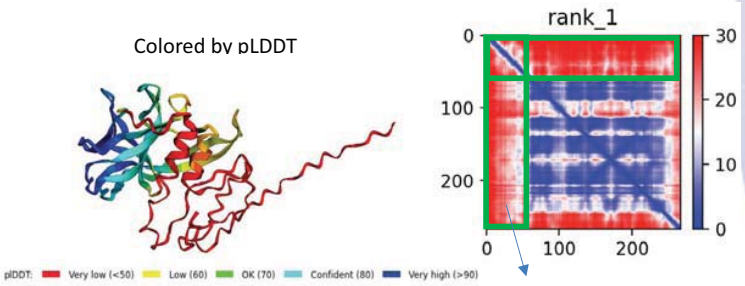


Too shallow MSA, less diversity
→ Less reliable prediction

2. Quality of predicted structure:
global pLDDT, pTM, residue-wise pLDDT, pAE

Global pLDDT = 58.2 / pTM = 0.611

cf) Actual LDDT value to experimental structure = 52.8



Do Not believe orientation of N-terminal region

The CASP15 rankings

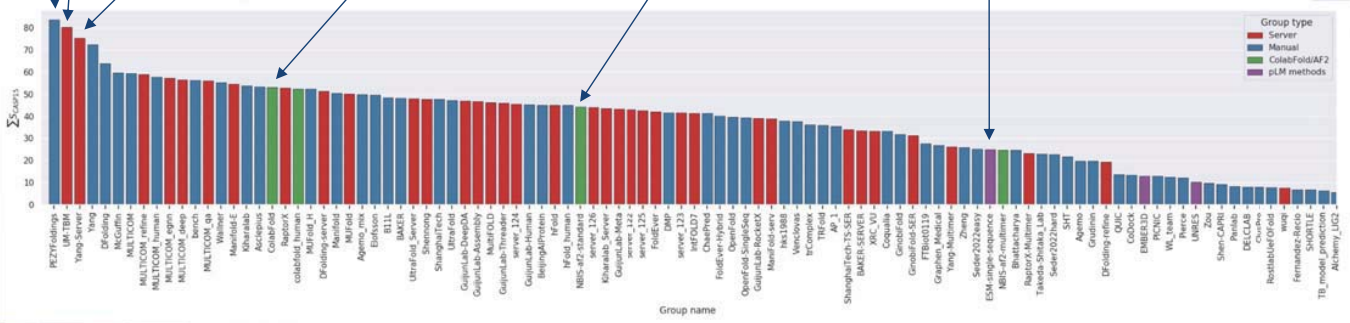
#1 PEZYFoldings AF2-based. Diverse MSAs. Custom, fine-tuned AF2 refinement

#2 UM-TBM Diverse MSAs. Threading then AF2 predictions guide I-TASSER REMC

#3 Yang-Server Diverse MSAs. AF2 predictions fed to trRosettaX2

ColabFold and NBIS-af2-standard

ESM-singlesequence is the top pure pLM method by this metric



Slides from CASP15 assessment talk given by D. Rigden

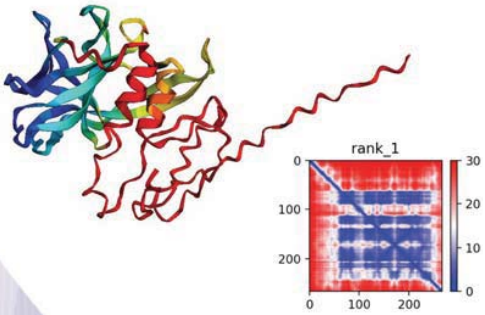
Tips & tricks to improve your prediction

- When your MSA is shallow…
  - More recycling (e.g. *num_recycles = 12*)

num_recycles=3

Global pLDDT = 58.2 / pTM = 0.611

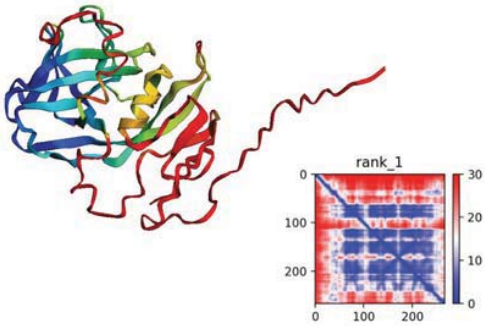
cf) Actual LDDT value to experimental structure = 52.8



num_recycles=12

Global pLDDT = 61.4 / pTM = 0.64

cf) Actual LDDT value to experimental structure = 53.6



Tips & tricks to improve your prediction

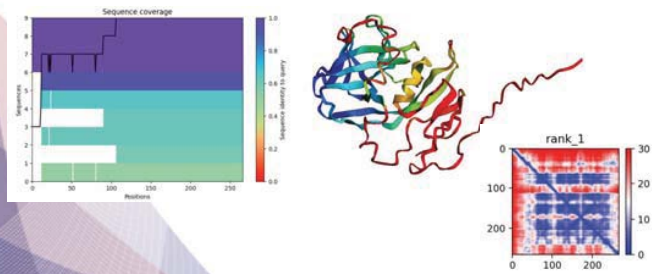
- When your MSA is shallow…
  - More recycling (e.g. *num_recycles = 12*)

  - Custom MSA search (utilize more sequence DB)

Number of seqs in MSA = 9

Global pLDDT = 61.4 / pTM = 0.64

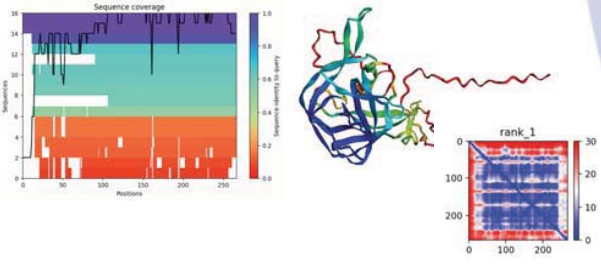
cf) Actual LDDT value to experimental structure = 53.6



Number of seqs in MSA = 16

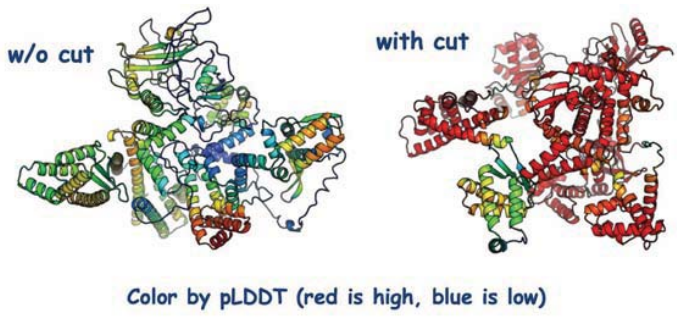
Global pLDDT = 70.1 / pTM = 0.717

cf) Actual LDDT value to experimental structure = 67.9



Tips & tricks to improve your prediction

- When your MSA is shallow…
  - More recycling (e.g. *num_recycles = 12*)
  - Custom MSA search (utilize more sequence DB)
- When your target protein has multiple domains
  - Split targets into domains & perform domain-wise MSA search
  - Merge domain-wise MSAs & do structure modeling



Tips & tricks to improve your prediction

- When your MSA is shallow…
  - More recycling (e.g. *num_recycles = 12*)
  - Custom MSA search (utilize more sequence DB)
- When your target protein has multiple domains
  - Split targets into domains & perform domain-wise MSA search
  - Merge domain-wise MSAs & do structure modeling
- When you predict complexes
  - Try predictions w/ paired MSA as well as w/ unpaired MSA
  - More sampling (e.g. *num_seeds = 5*) – pick one having highest pTM or ipTM

Take home messages

- **현재의 구조 예측 AI는 진화정보에 의존!**
  - 진화정보를 못찾는다? (shallow MSA)
    → 구조예측이 틀릴 가능성이 높다
  - 추가적인 sequence DB, MSA search tool을 활용하는 것이 도움이 될 수 있다
- **pLDDT, pTM, pAE 등 예측 결과에 대한 confidence metric을 반드시 체크하자**
  - 때론 confidence가 잘못 추정될 수도 있다. 구조도 같이 꼭 들여다보자.

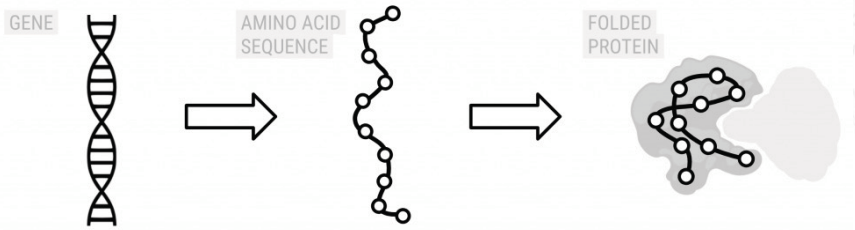
이 그림은 "AlphaFold는 신이 아닙니다. AlphaFold도 틀릴 수 있습니다"라는 문구를 개념적으로 표현합니다. AlphaFold AI가 복잡한 기계로 나타나 있으며, 과학자가 오류를 나타내는 부분이 강조된 단백질 구조를 분석하고 있습니다. 이 장면은 AlphaFold가 강력한 도구임에도 불구하고 완벽하지 않으며 인간의 감독이 필요함을 강조합니다.

제 2강.
Protein Design with AI

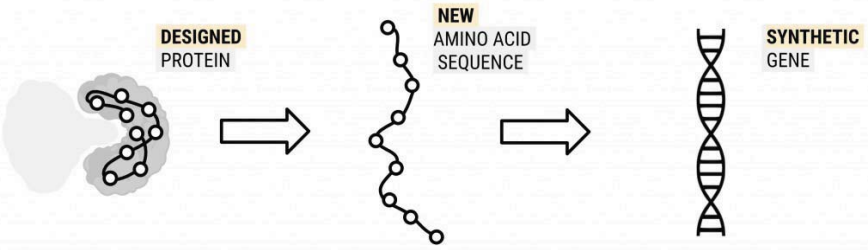
50

Protein Design

Protein structure prediction



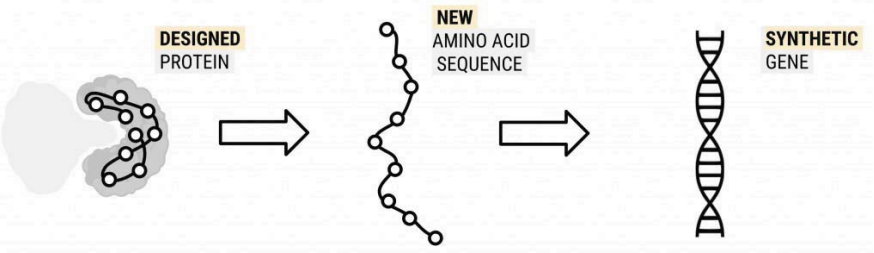
Protein design



51

Protein design before AI era

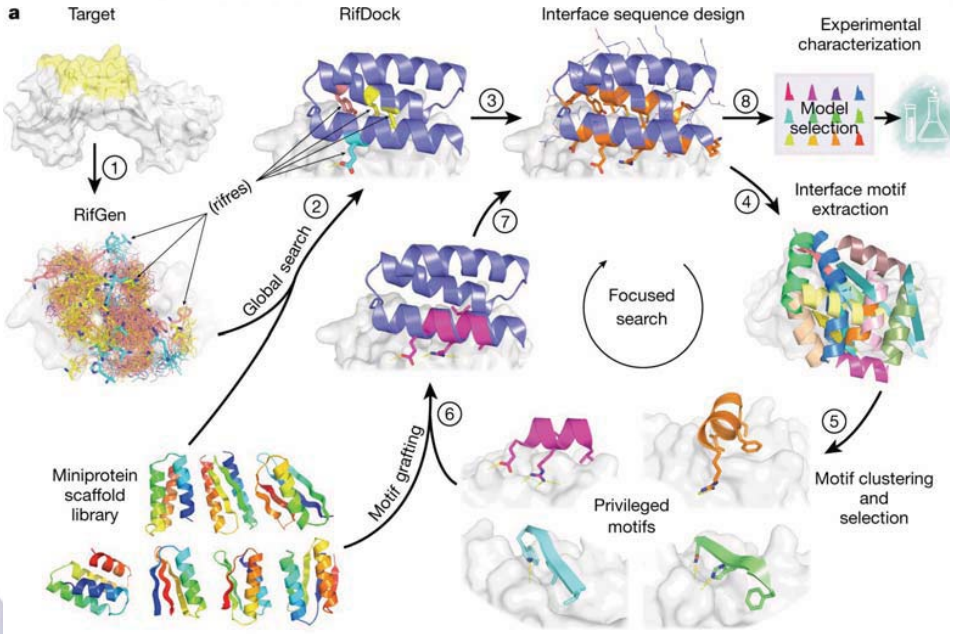
Protein Design



1. Sample backbone structures to have a desired function
2. Amino acid sequence design on sampled backbone structures
3. Filter designs (whether it will have designed structure, have low enough energies, etc)
4. Select candidates for experimental validation
5. Experimental validation (feedback & iteration)

Protein binder design before AI era

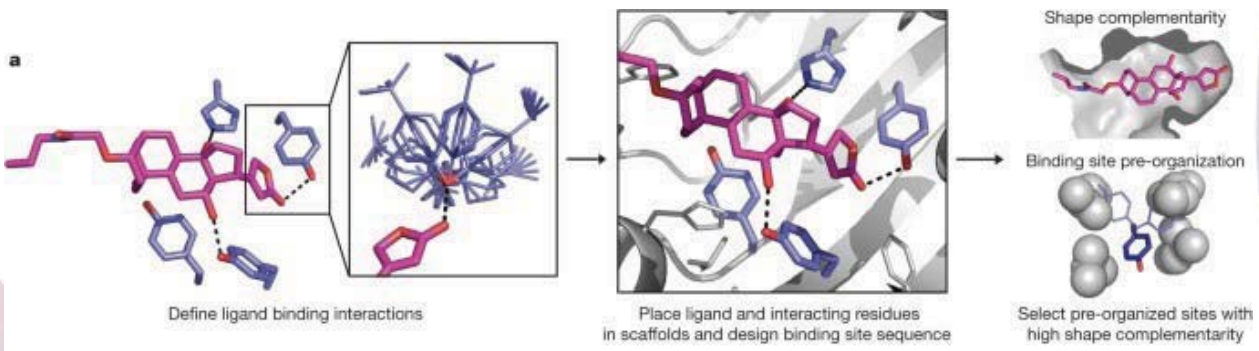
Binding motif sampling → find protein scaffold to accommodate motif → scoring/refining



Cao, Longxing, et al. "Design of protein-binding proteins from the target structure alone." *Nature* 605.7910 (2022): 551-560.

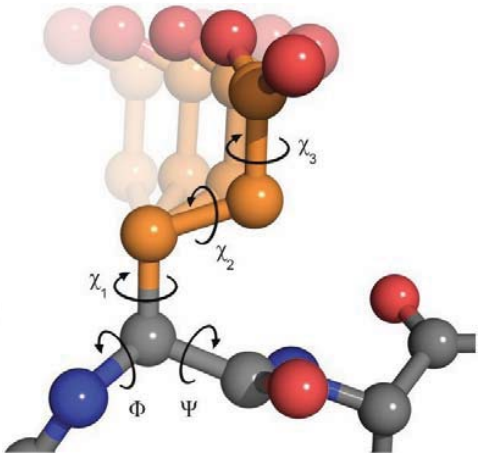
Protein binder design before AI era

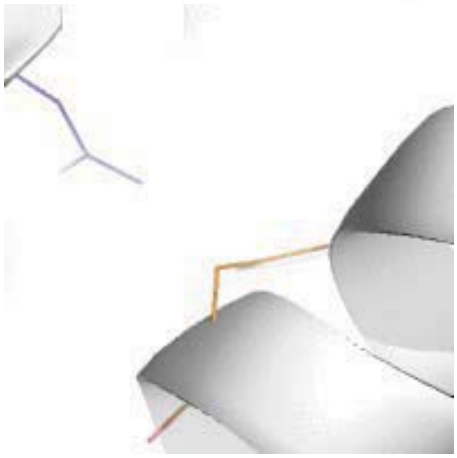
Binding motif sampling → find protein scaffold to accommodate motif → scoring/refining



Tinberg, Christine E., et al. "Computational design of ligand-binding proteins with high affinity and selectivity." *Nature* 501.7466 (2013): 212-216.

Protein sequence design before AI era

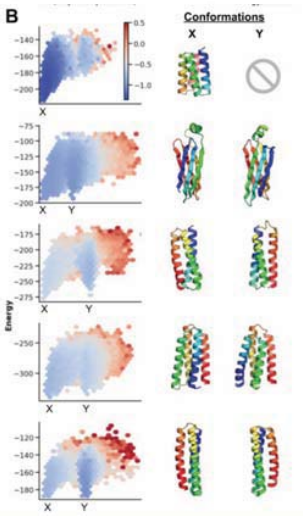




Find a best combination of amino acids and rotamers having the lowest energy

Selection criteria (filters) for protein design

Folding energy landscape (for monomers / scaffolds)



For binders..

Docking energy landscape

Binding energy (ddG) calculation

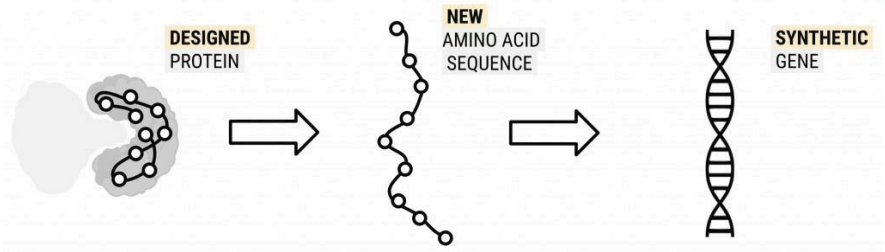
Buried interface area

Unsatisfied polar groups

....

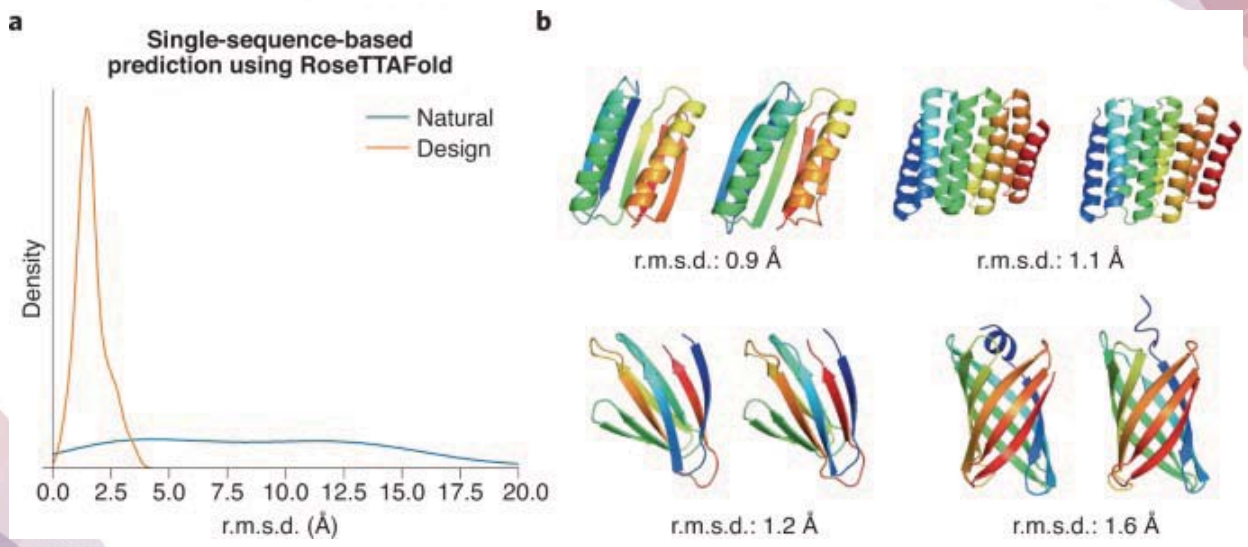
Protein design in AI era

Protein Design

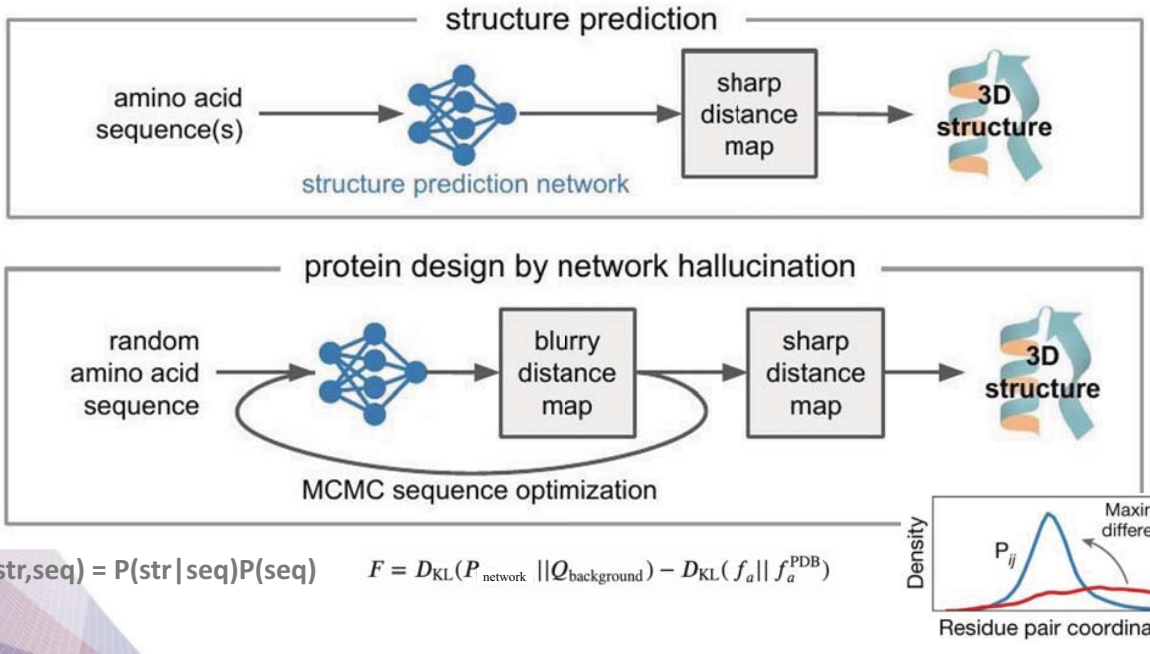


1. Sample backbone structures to have a desired function using AI
2. Amino acid sequence design on sampled backbone structures using AI
3. Filter designs using AI
4. Select candidates for experimental validation
5. Experimental validation (feedback & iteration)

Utilization of protein structure prediction AI as filter



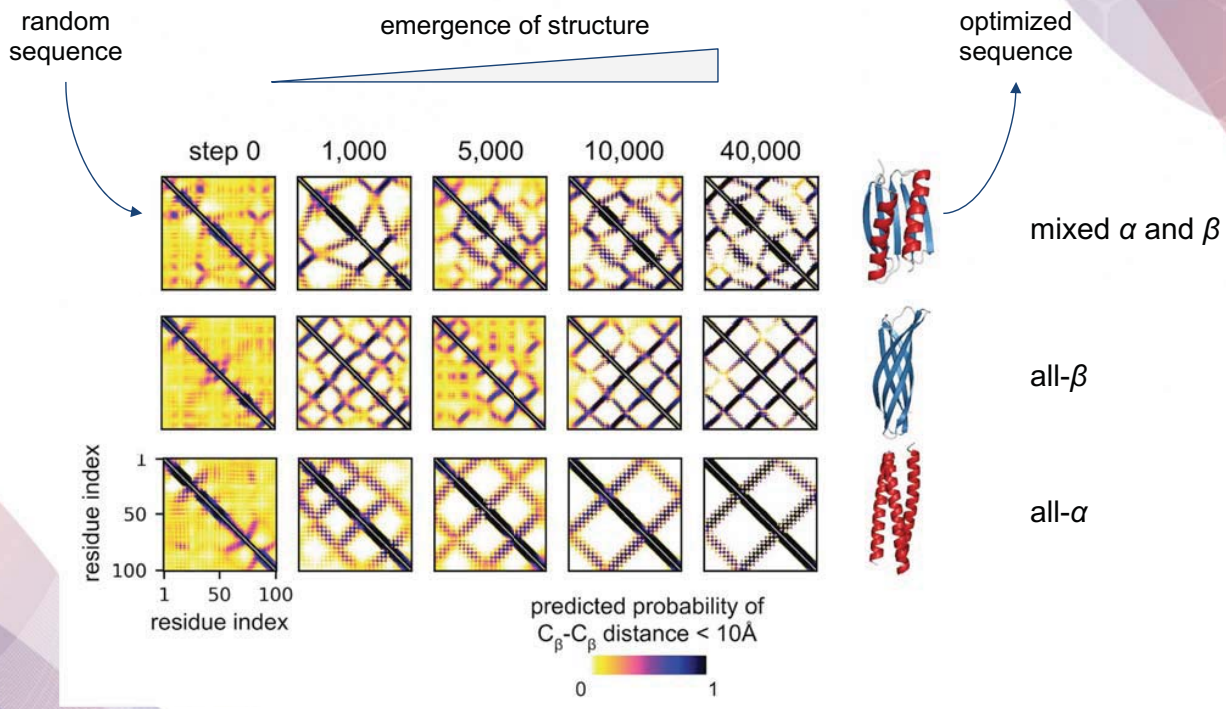
Backbone sampling using protein modeling AI: Network Hallucination



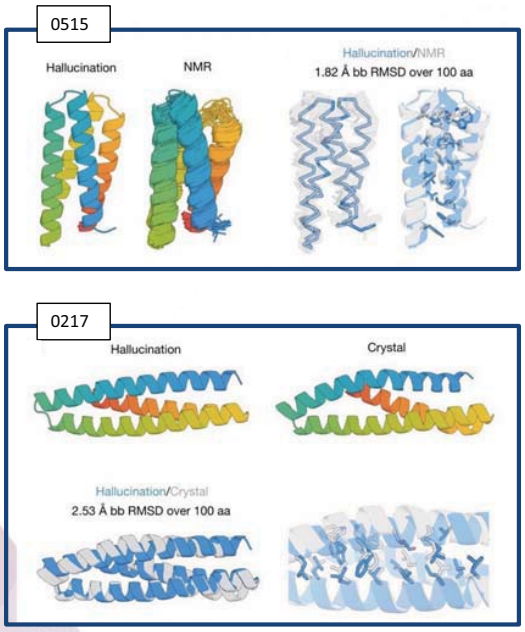
$P(str,seq) = P(str|seq)P(seq)$

$$F = D_{KL}(P_{network} || Q_{background}) - D_{KL}(f_a || f_a^{PDB})$$

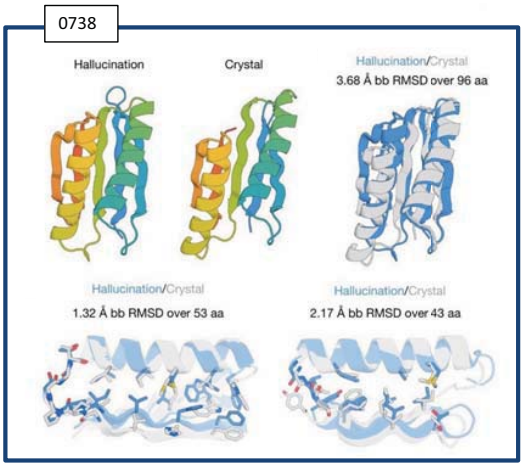
Anishchenko, I., Pellock, S., et al. *Nature* (2021)



Anishchenko, I., Pellock, S., et al. *Nature* (2021)

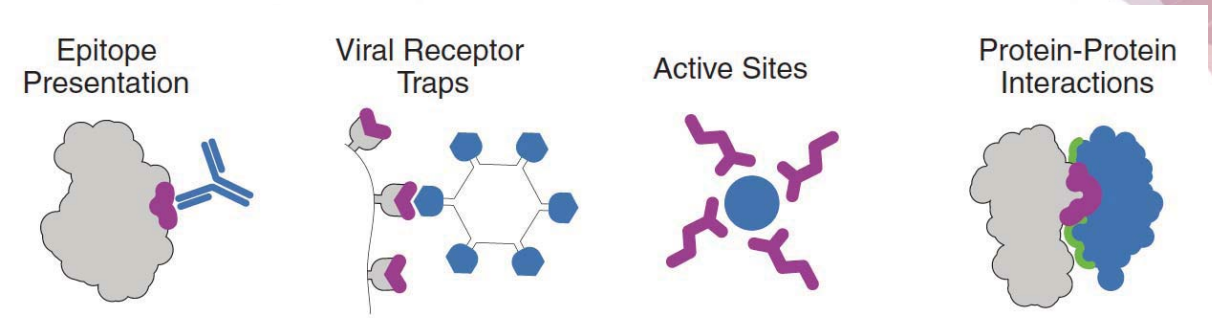


Structures of 3 hallucinations were confirmed experimentally



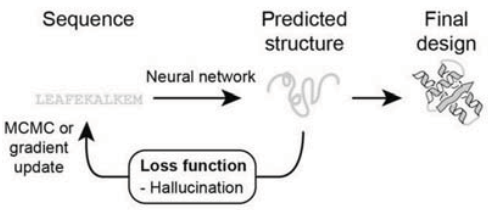
Anishchenko, I., Pellock, S., et al. *Nature* (2021)

Functional protein design based on motifs

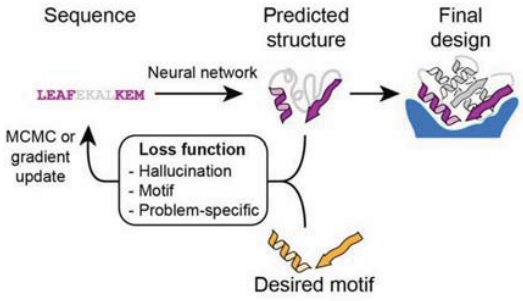


Design proteins scaffolding a given functional motif

Constrained hallucination:
Design a new protein having a given functional motif



Free hallucination:
generate novel protein folds

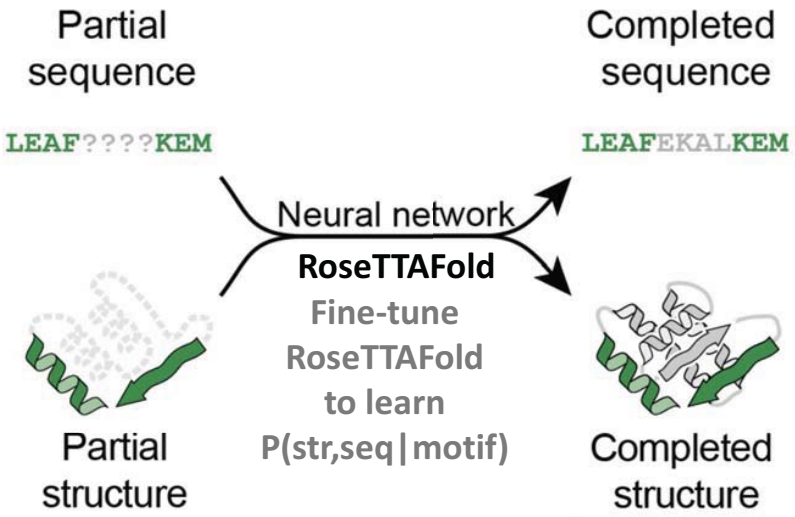


Constrained hallucination:
generate scaffolds harboring
pre-specified functional sites

Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J., *Science* (2022)

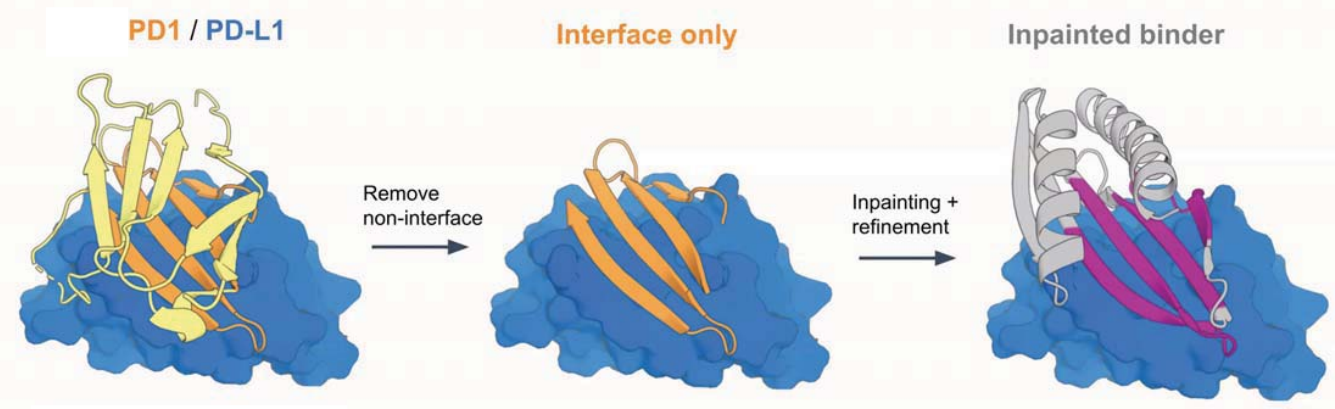
Protein design via inpainting

Formulate motif-based protein design as information completion (or inpainting)



Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J., Science (2022)

Design PD-L1/PD-1 binding inhibitor



Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J., Science (2022)

Pros/Cons of hallucination & inpainting approaches

Hallucination

Pros

- No further training
- Can solve diverse design problem

Cons

- Slow (1k~10k structure prediction per design)
- **Unusual sequences** e.g. poly P

Inpainting

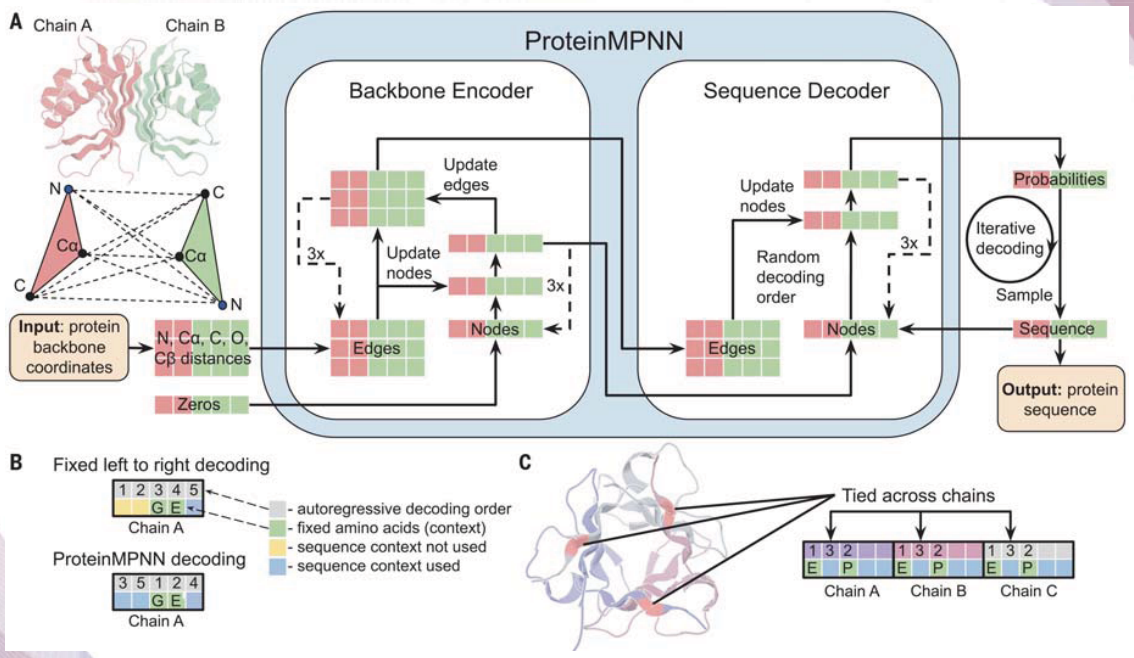
Pros

- Fast
- Can design larger protein

Cons

- Need further training for each design problem
- **Unusual sequences**
- Hard to get diverse structures

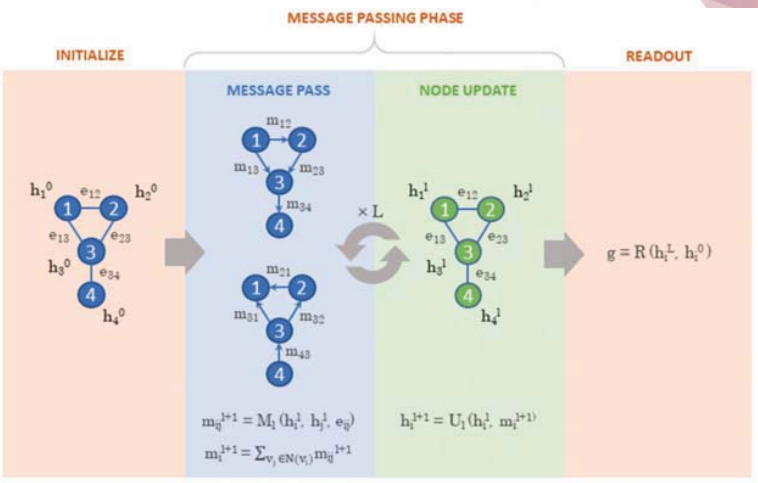
AI-based protein sequence design: ProteinMPNN



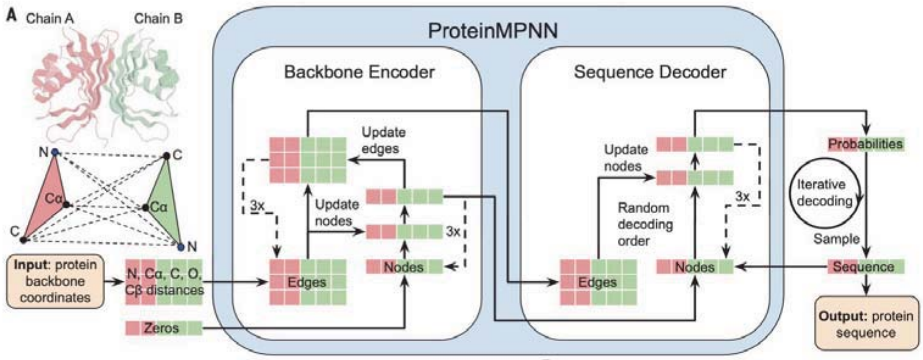
Robust deep learning–based protein sequence design using ProteinMPNN, Science (2022)

MPNN: Message Passing Neural Network

- Graph Neural Network
  - Node: Residue
  - Edge: neighboring residue pairs
- Each node has information (message)
- Messages are updated using information from direct neighbors connected by edges



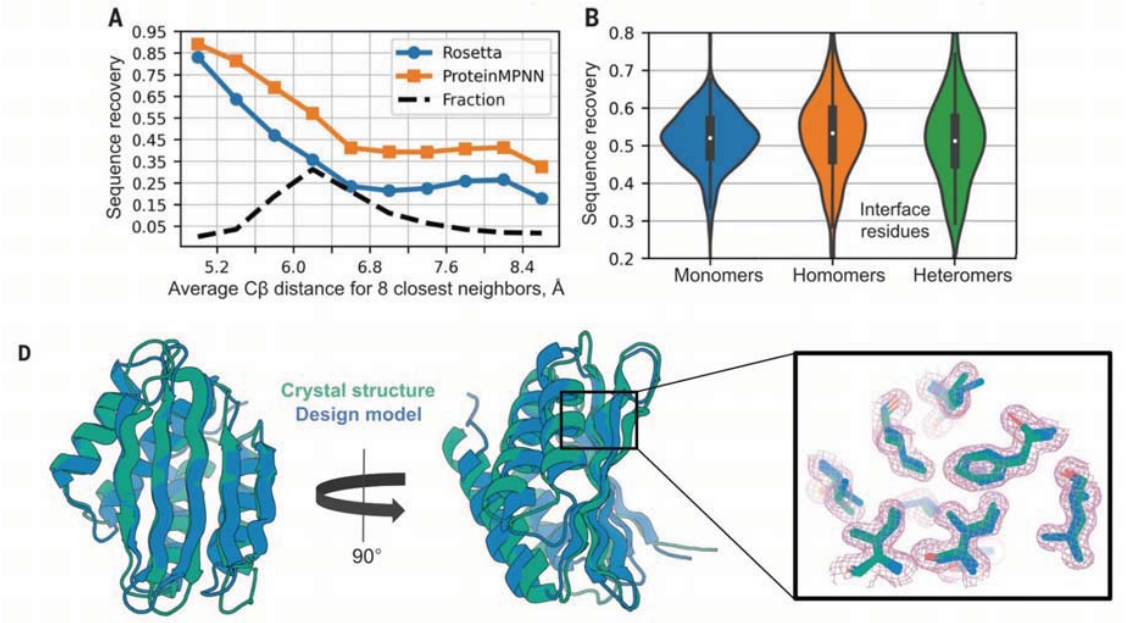
What information is important for ProteinMPNN?



Noise level when training: 0.00 Å/0.02 Å	Modification	Number of parameters in millions	PDB test accuracy (%)	PDB test perplexity	AlphaFold model accuracy (%)
Baseline model	None	1.381	41.2/40.1	6.51/6.77	41.4/41.4
Experiment 1	Add N, Cα, C, Cβ, O distances	1.430	49.0/46.1	5.03/5.54	45.7/47.4
Experiment 2	Update encoder edges	1.629	43.1/42.0	6.12/6.37	43.3/43.0
Experiment 3	Combine 1 and 2	1.678	50.5/47.3	4.82/5.36	46.3/47.9
Experiment 4	Experiment 3 with random decoding	1.678	50.8/47.9	4.74/5.25	46.9/48.5

Robust deep learning–based protein sequence design using ProteinMPNN, Science (2022)

ProteinMPNN generates highly probable sequences



Robust deep learning–based protein sequence design using ProteinMPNN, Science (2022)

Pros/Cons of hallucination & inpainting approaches

Hallucination

Pros

- No further training
- Can solve diverse design problem

Cons

- **Slow (1k~10k structure prediction per design)**
- ~~Unusual sequences~~ e.g. poly P

Inpainting

Pros

- Fast
- Can design larger protein

Cons

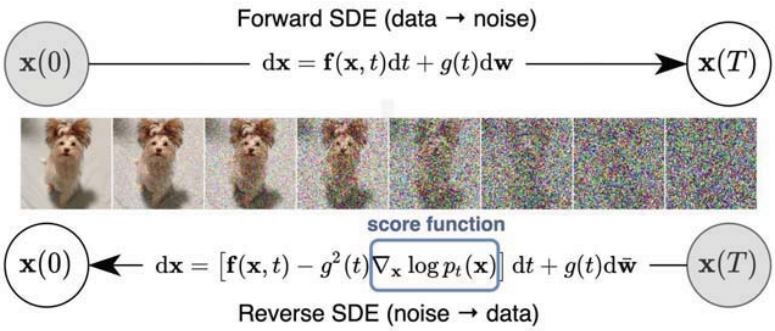
- Need further training for each design problem
- ~~Unusual sequences~~
- **Hard to get diverse structures**

Can we make a "generative AI" for protein design?

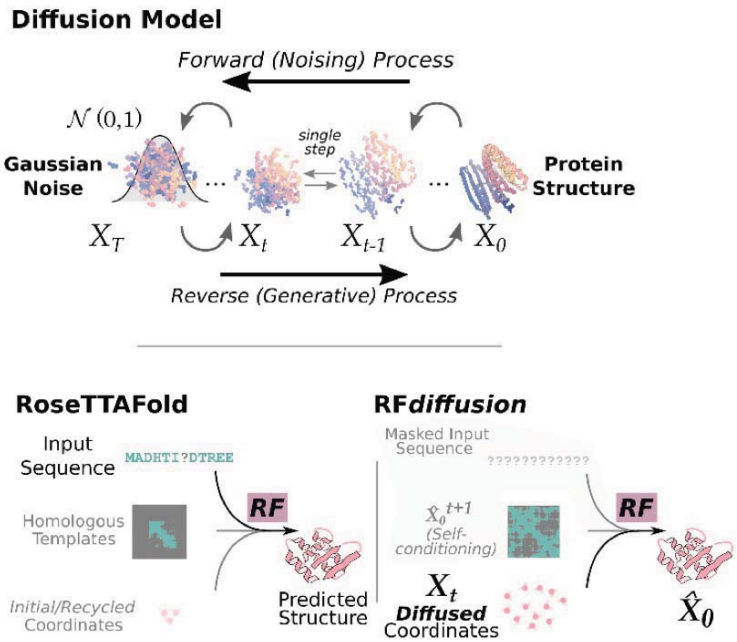
- Goal: Generate diverse structures having desired function
- Cutting edge AI technology: Image generation w/ "Diffusion Model"

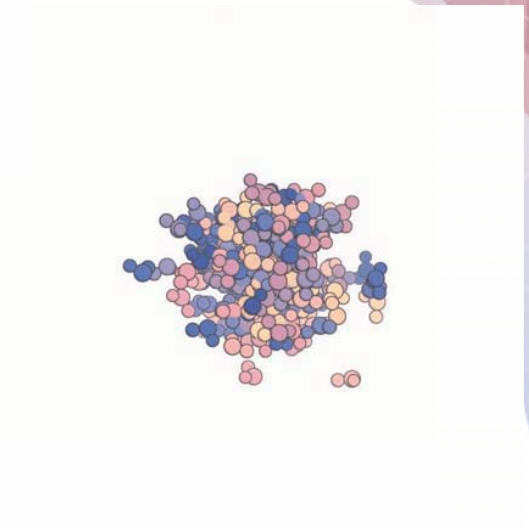
a painting of a fox in the style of Starry Night

Diffusion Model



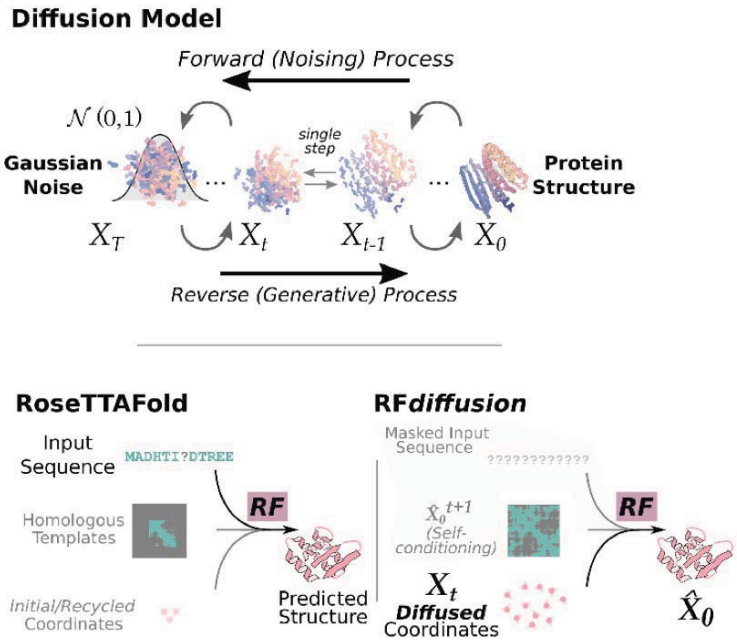
Generative model for protein design (RFdiffusion)

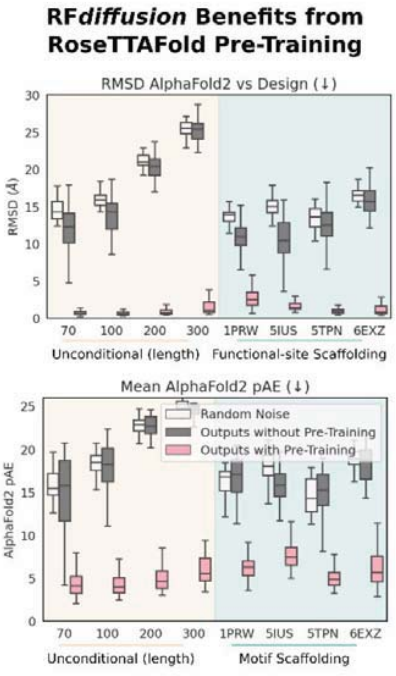




Watson, J., Juergens, D., Bennett, N., Trippe, B., Yim, J., Eisenach, H., Ahern, W., *Nature* (2023)

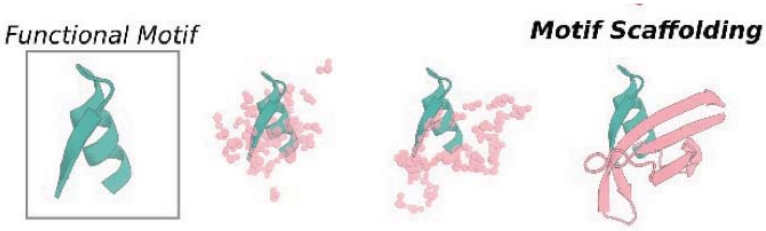
Generative model for protein design (RFdiffusion)

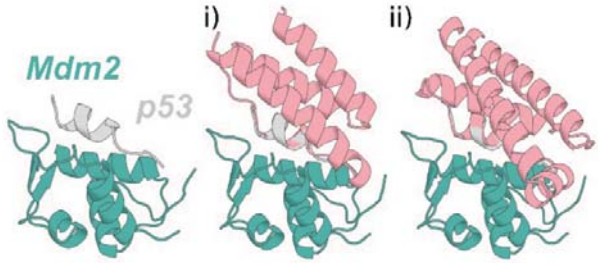




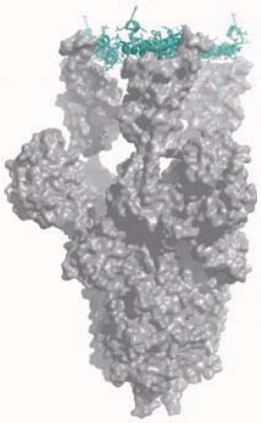
Watson, J., Juergens, D., Bennett, N., Trippe, B., Yim, J., Eisenach, H., Ahern, W., *Nature* (2023)

Functional motif scaffolding w/ RFdiffusion





K_D=0.7nM K_D=0.5nM

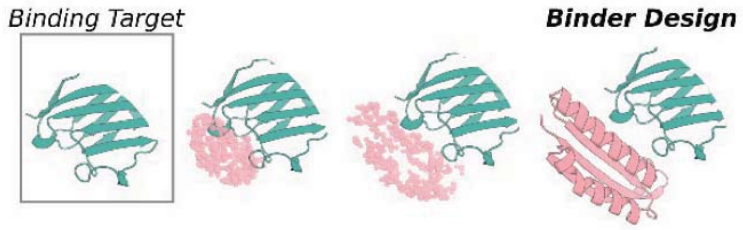


Designed antiviral proteins

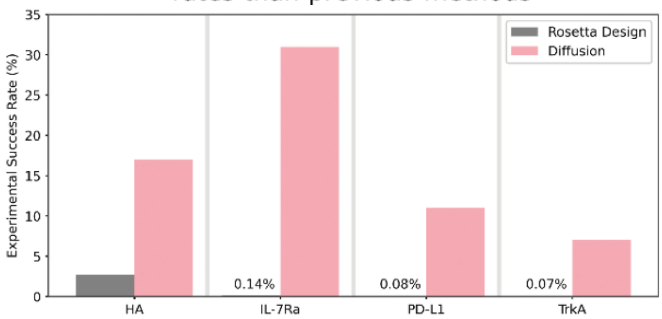
SARS-CoV-2 Spike protein

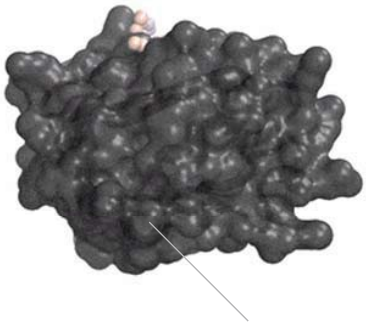
Watson, J., Juergens, D., Bennett, N., Trippe, B., Yim, J., Eisenach, H., Ahern, W., *Nature* (2023)

Protein binder design w/ RFdiffusion



RFdiffusion has orders-of-magnitude higher **experimental** success rates than previous methods





Insulin Receptor

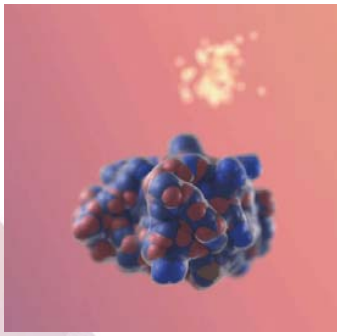
Watson, J., Juergens, D., Bennett, N., Trippe, B., Yim, J., Eisenach, H., Ahern, W., *Nature* (2023)

Current state-of-the-art protein design pipeline

Generation of backbone structures that may have desired function

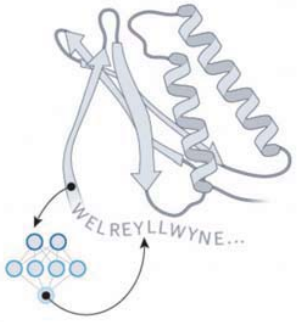
Design of protein sequence for given backbone structures

Computational & experimental validation



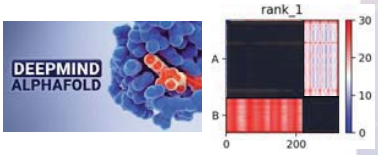
RFdiffusion

J. Watson et al, Nature, 2023



ProteinMPNN

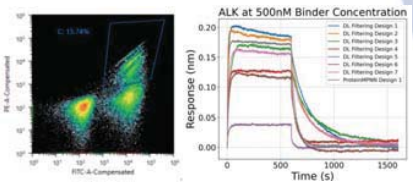
J. Dauparas et al, Science, 2022



AI-based filters

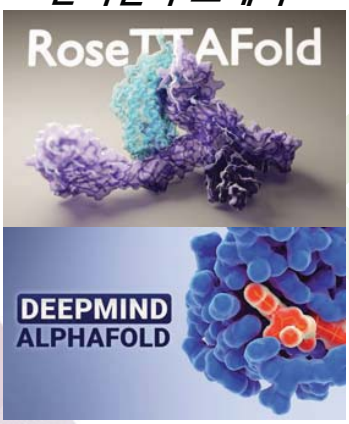
e.g. Interchain PAE < 10

N. Bennett, et al, Nat. Comm., 2023

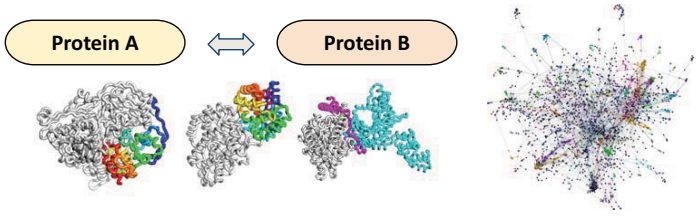


단백질 구조 예측의 무궁무진한 응용가능성

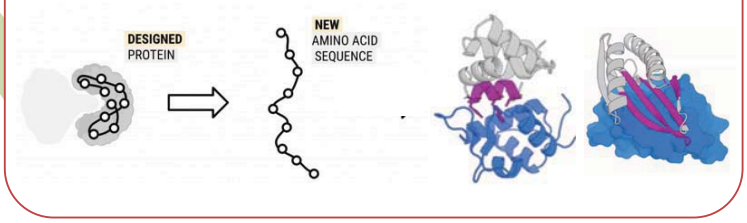
인공지능 기반 단백질 구조예측



단백질 사이의 상호작용 예측

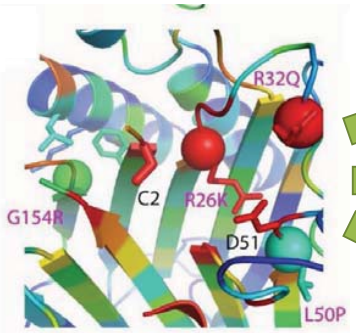


인공지능 기반 단백질 디자인



Remaining challenges

- Mutation effect prediction
  - Better understanding on diseases / Protein engineering

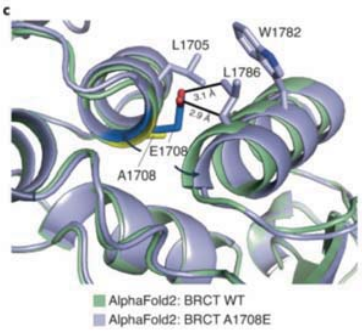


Structural changes

Stability changes

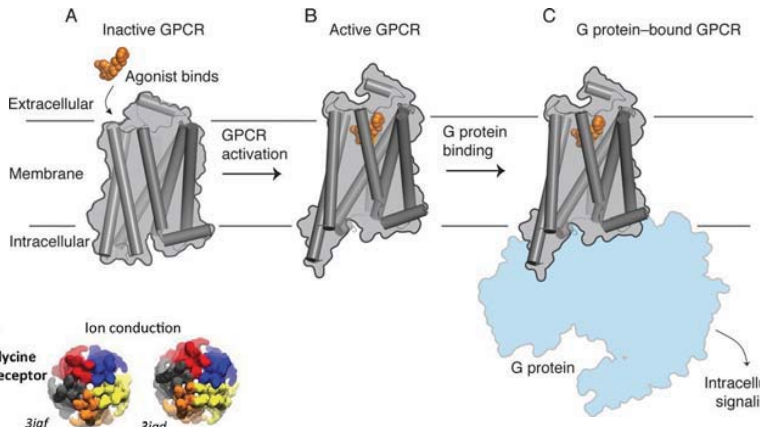
Activity changes

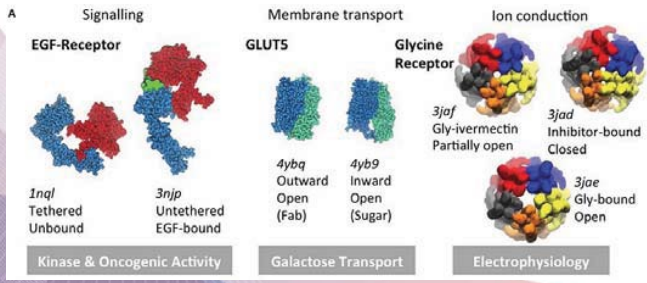
구조 변화 예측 실패



Remaining challenges

- Mutation effect prediction
- Multi-state modeling





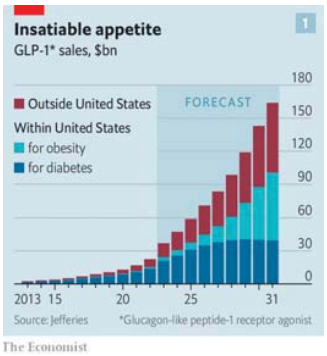
Remaining challenges

- Mutation effect prediction
- Multi-state modeling
- Interaction with other molecules
  - Antigen-antibody, host-pathogen (no coevolution)
  - Protein-DNA/RNA/small molecules
  - Lack of dataset

Let’s do some design
using Rfdiffusion & ProteinMPNN!

Tutorial: Designing novel GLP-1 ligand

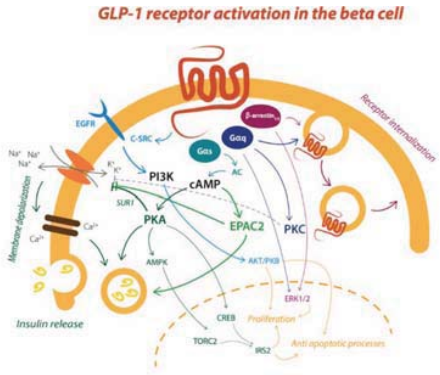
- Semaglutide (Ozempic) is an antidiabetic medication used for the treatment of type 2 diabetes and an anti-obesity medication
- Its maker, Novo nordisk, became the most valuable company in Europe
  - Market cap over $400bn
  - More than total GDP of Denmark

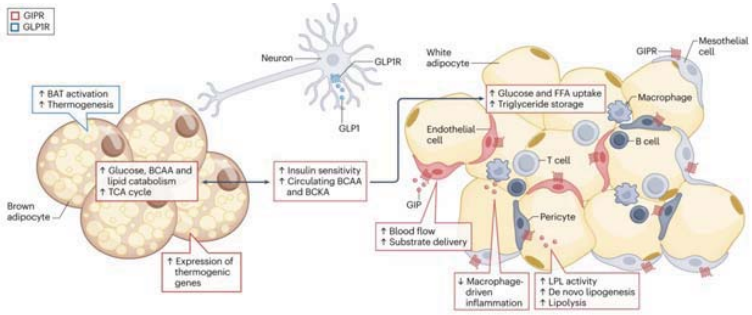


Goals of this tutorial

- Goal 1
  - Redesign alternative peptide sequences based on the GLP-1:GLP-1 receptor complex crystal structure using ProteinMPNN
- Goal 2
  - Binder design of GLP-1 receptor using RFdiffusion

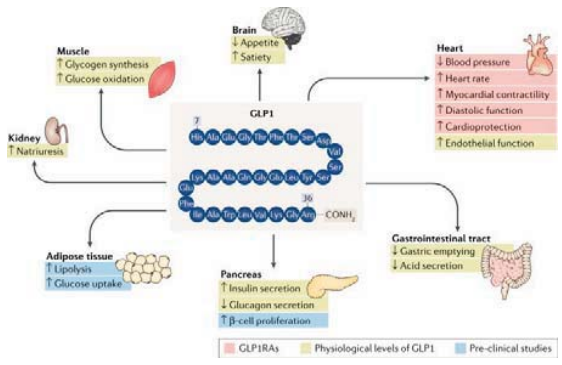
Background: Mechanism of GLP-1



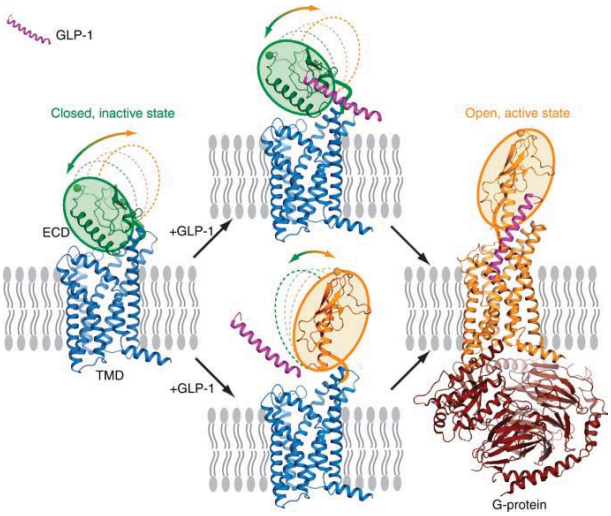


- GLP-1 receptor activation leads to increase in insulin release
- Promote brown adipose tissue activation
  - Promotes resting metabolic rate, fatty acid clearance & thermogenesis

Background:
The complex structure of GLP-1 and GLP-1 receptor



Andersen, A., Lund, A., Knop, F.K. *et al. Nat Rev Endocrinol* **14**, 390–403 (2018)

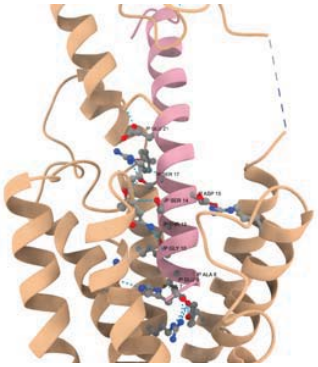


- GLP-1 is a natural hormone, a peptide

Wu, F., Yang, L., Hang, K. *et al.* Full-length human GLP-1 receptor structure without orthosteric ligands. *Nat Commun* **11**, 1272 (2020)

Target PDB: 6X18

GLP-1 peptide hormone bound to Glucagon-Like peptide-1 (GLP-1) Receptor



- Receptor domain: R chain
- Peptide domain: P chain

6X18