Protein structure prediction: the past, the present, and the future

Dec. 11, 2022

CASP15 meeting

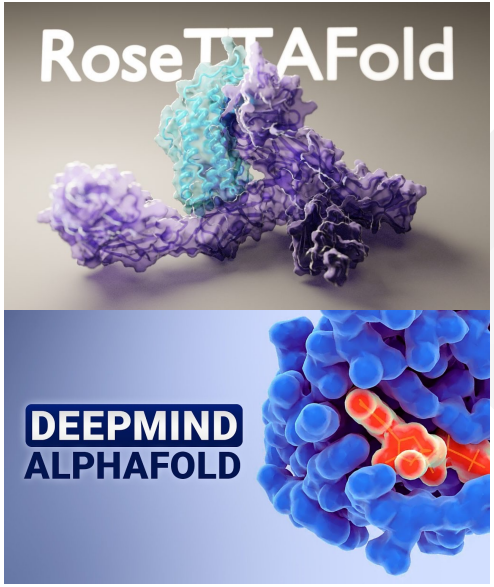
Minkyung Baek

minkbaek@snu.ac.kr

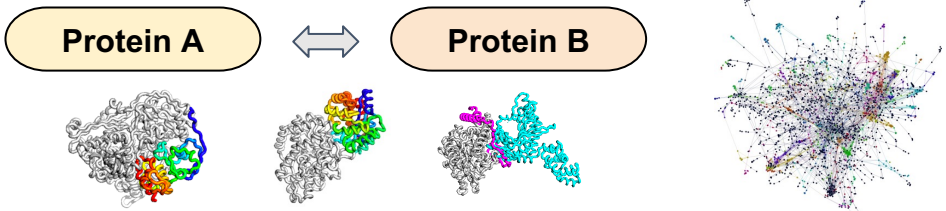
INSTITUTE FOR Protein Design

UNIVERSITY of WASHINGTON

AI-based protein modeling

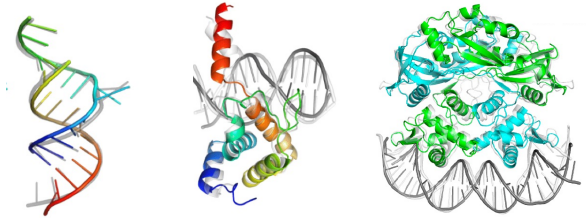


Large-scale *in silico* PPI screening



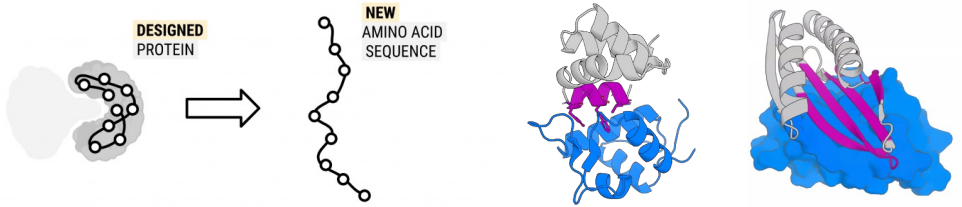
Humphreys, I., Pei, J., Baek, M., Krishnakumar, A., et al, *Science* (2021)

Nucleic acid structure & interaction prediction



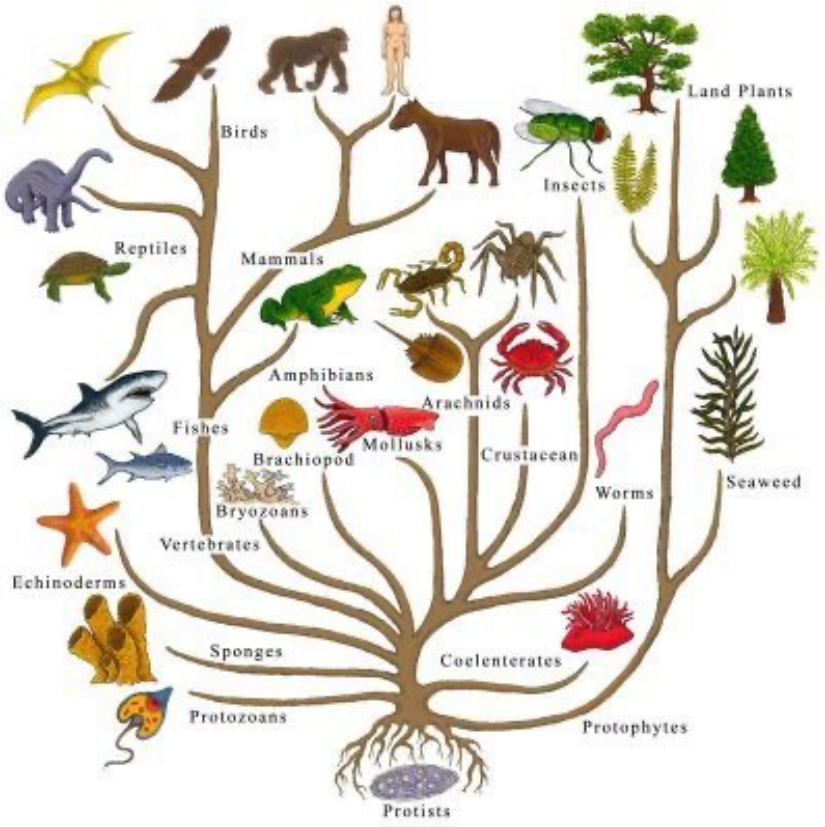
Baek, M., McHugh, R., Anishchenko, I., Baker, D., DiMaio, F., et al, *biorxiv* (2022)

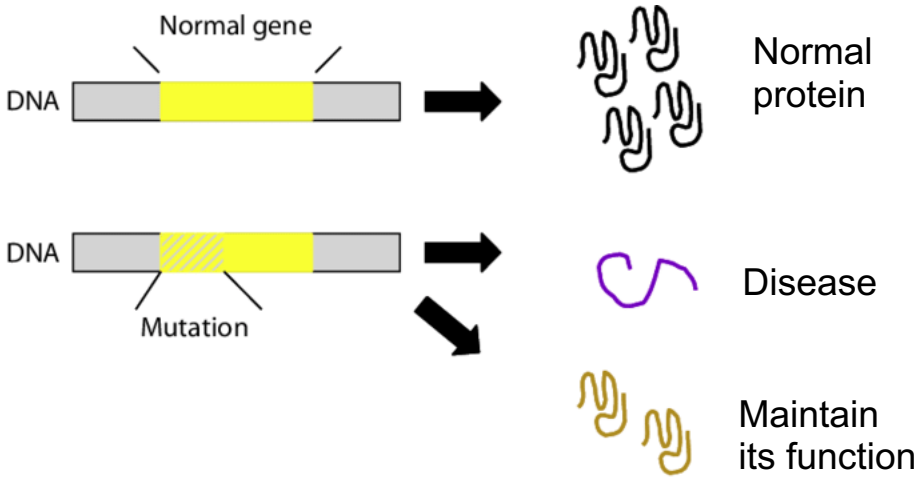
De novo functional protein design

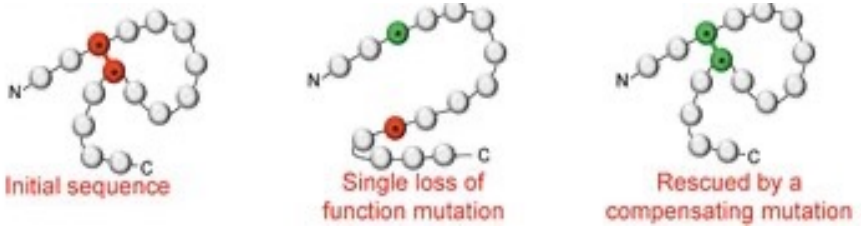


Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J., et al, *Science* (2022)

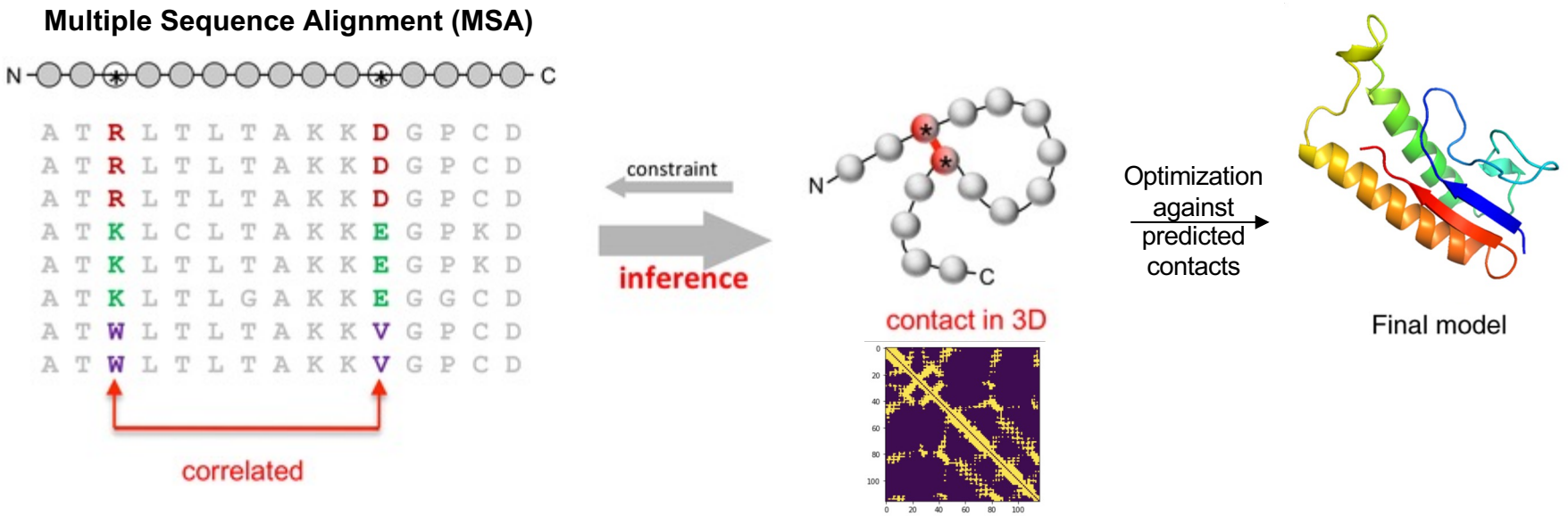
Protein structure prediction using evolution history







Coevolution guided modeling



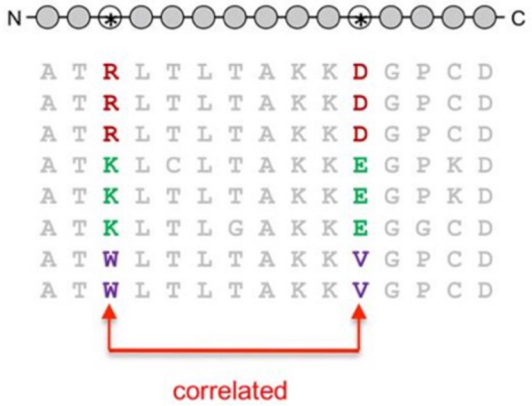
Q: How can we find the coevolution pattern from MSA?

Applying deep learning to protein structure prediction

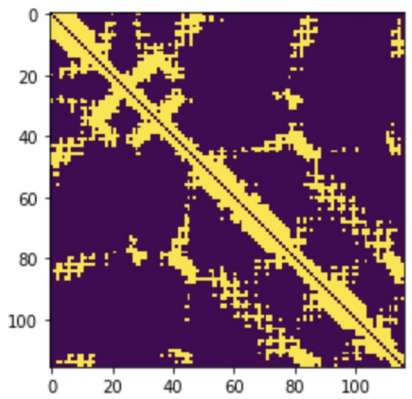
- Deep learning algorithms attempt to learn (multiple levels of) representation by using a **hierarchy of multiple layers**
- Exceptional effective at **learning patterns**
- If you provide the system **tons of training examples**, it begins to understand hidden patterns in data and respond in useful ways

From MSA to 3D structures with deep learning

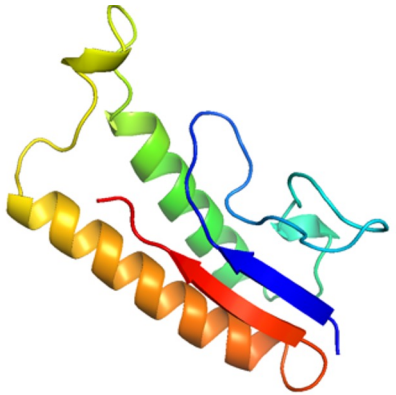
Multiple sequence alignments



Residue pairwise interaction



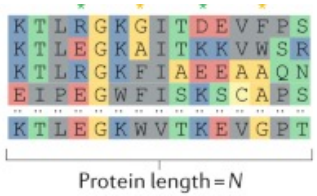
3D atomic coordinates



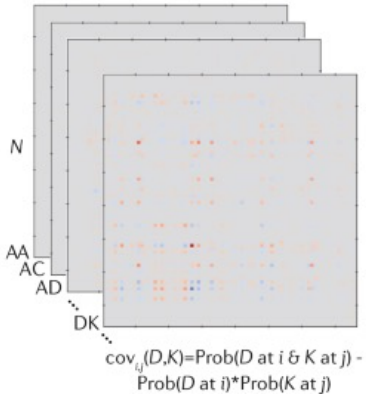
How to extract residue pairwise interactions & build 3D models based on that?

Early attempts w/ DL: ResNet-based approach

MSA

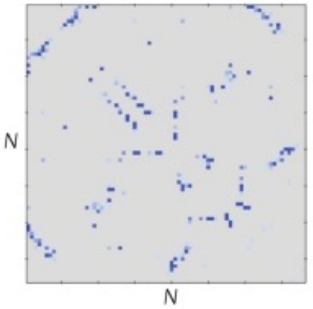


Covariances



Contact prediction

Contacts



INPUT

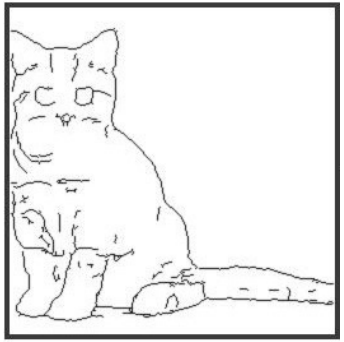
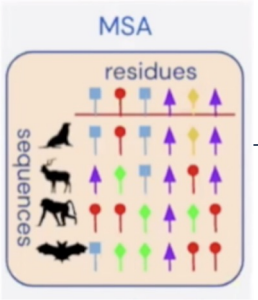


Image processing
(ResNet)

OUTPUT

Early attempts w/ DL: ResNet-based approach



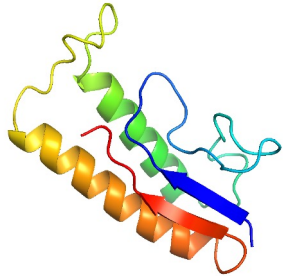
Extract structural information from MSA (one-shot prediction)

MRF (AlphaFold1)
Inverse covariance matrix (trRosetta)
MSA encoder (CopulaNet)

templates

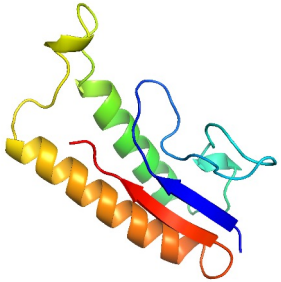
2D ResNet to refine contact signal & get better distance prediction

Folding



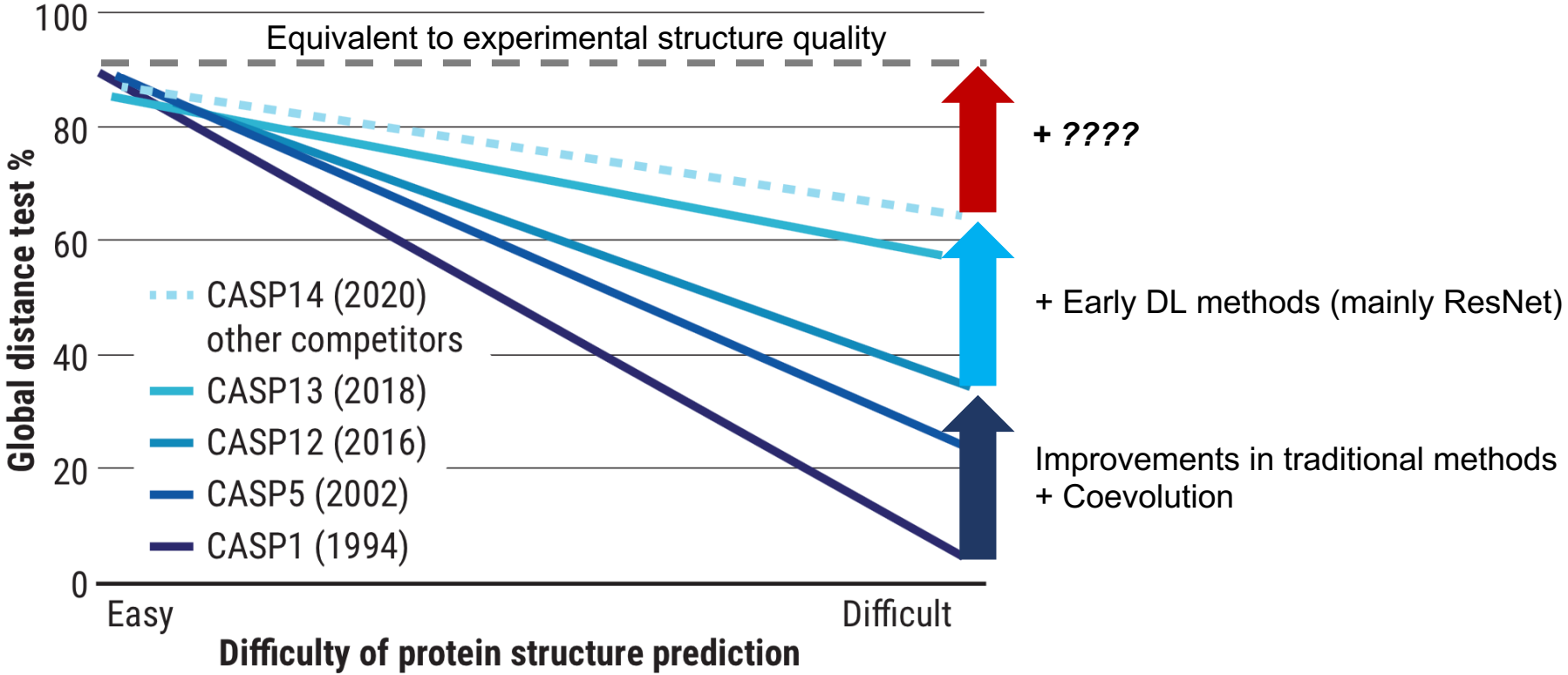
Initial model

(classical) Refinement



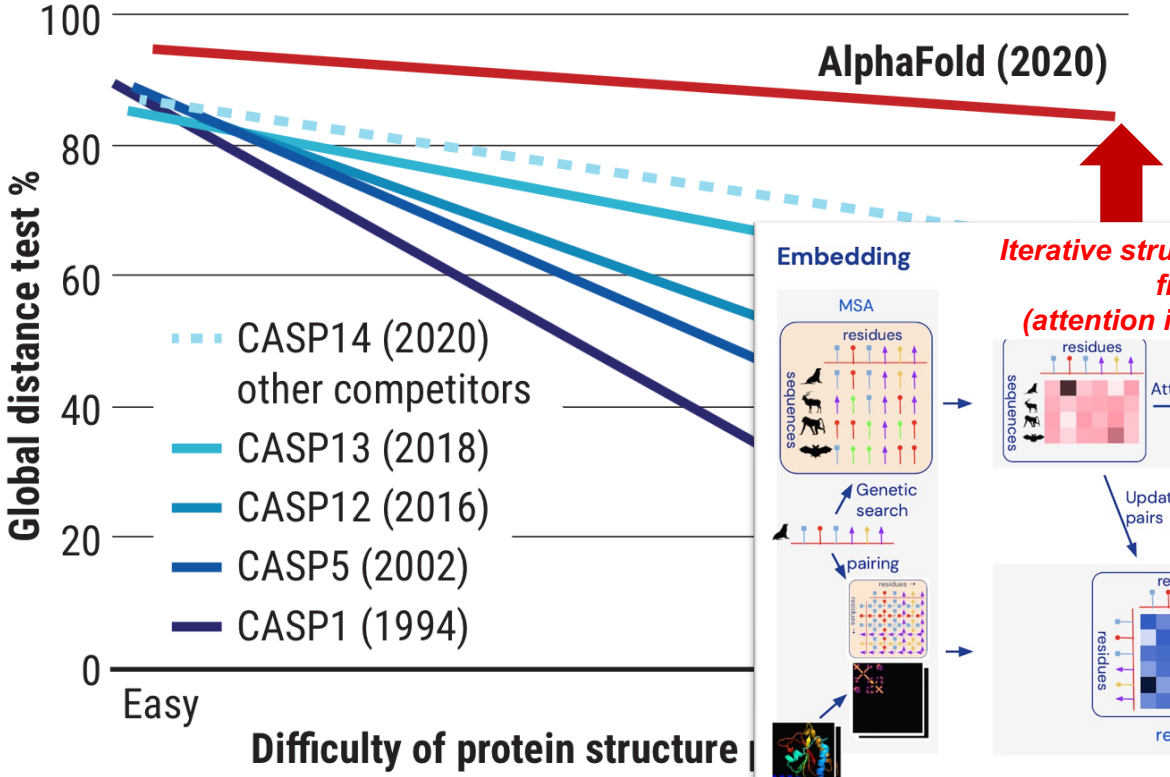
Refined model

Progress in protein structure prediction



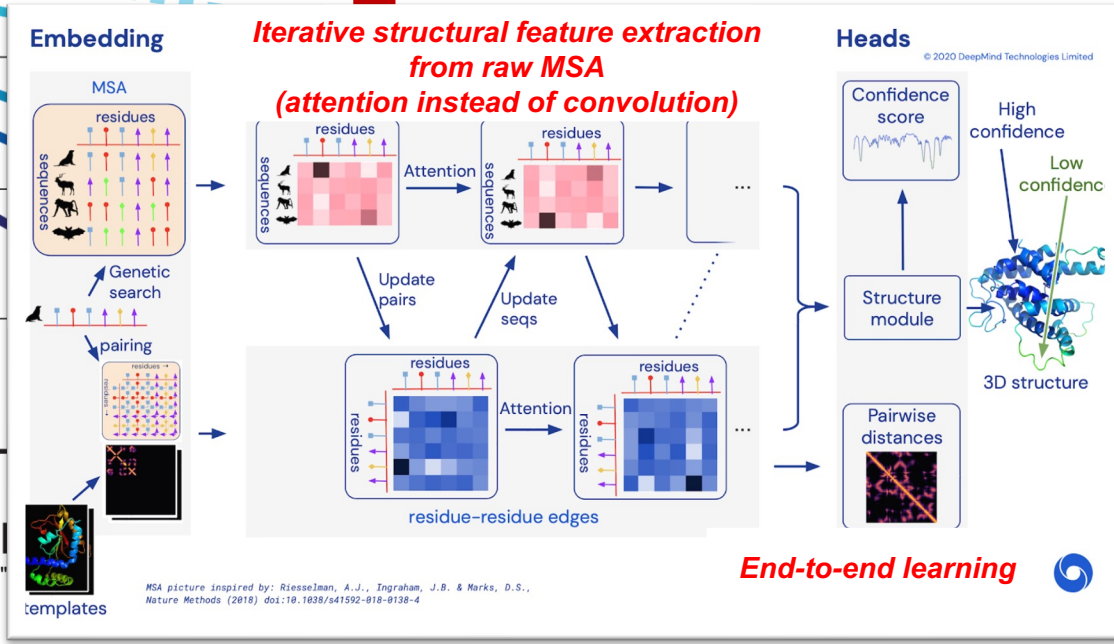
Service, Robert F. "'The game has changed.'AI triumphs at protein folding." (2020): 1144-1145.

AlphaFold2: A game changer



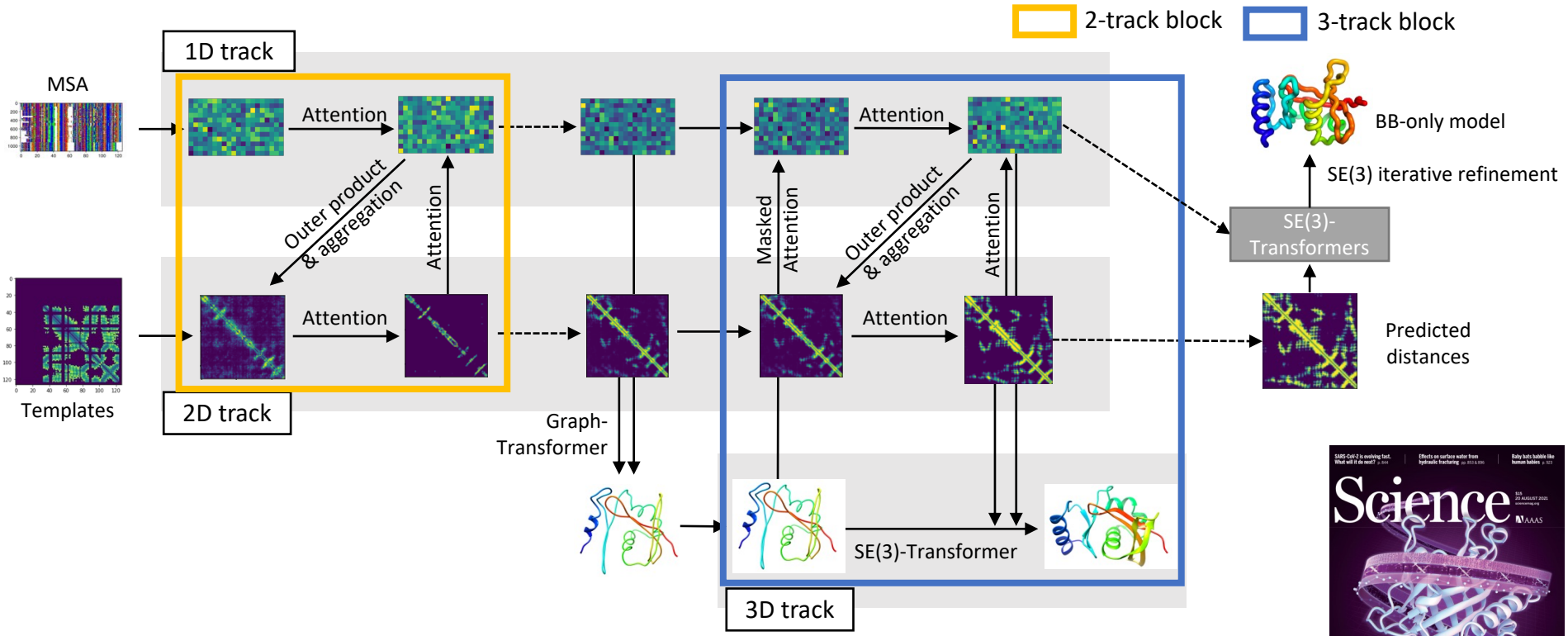
Service, Robert F. "'The game has changed.'AI triumphs at protein folding."

+ Proper DL architectures



End-to-end learning

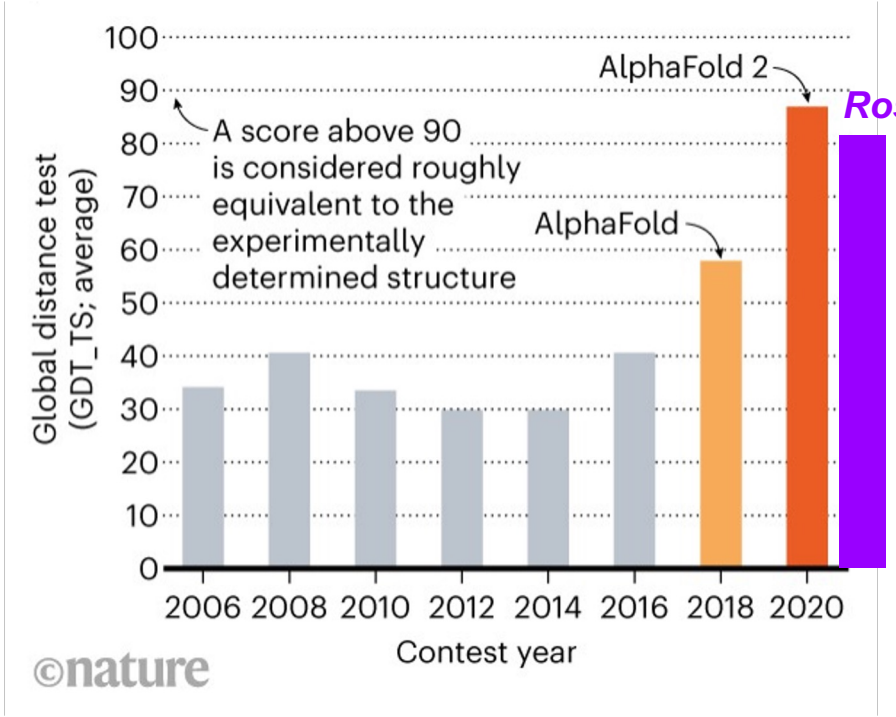
RoseTTAFold: Can academia make something like AlphaFold?

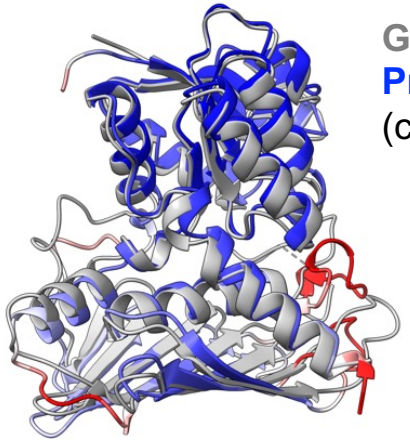


Baek, M., et al, *Science* (2021)

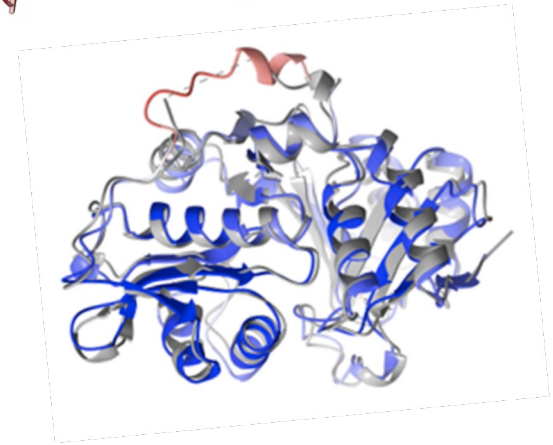
RoseTTAFold: Academia can do!

Free modeling accuracy in CASP14





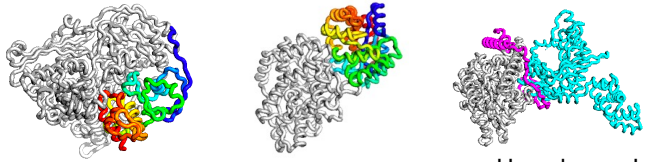
Ground-truth
Predicted
(colored by pLDDT)



Beyond accurate modeling of protein structures

Large-scale *in silico* PPI screening

Protein A ⟺ **Protein B**

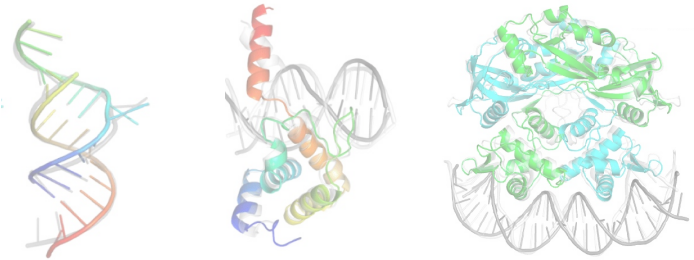


1) Do **A** and **B** interact?

2) What is the structure of **AB**?

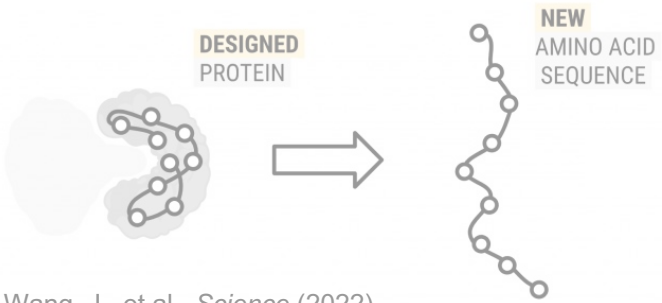
Humphreys, I., Pei, J., Baek, M., Krishnakumar, A., et al, *Science* (2021)

Nucleic acid structure and interaction modeling



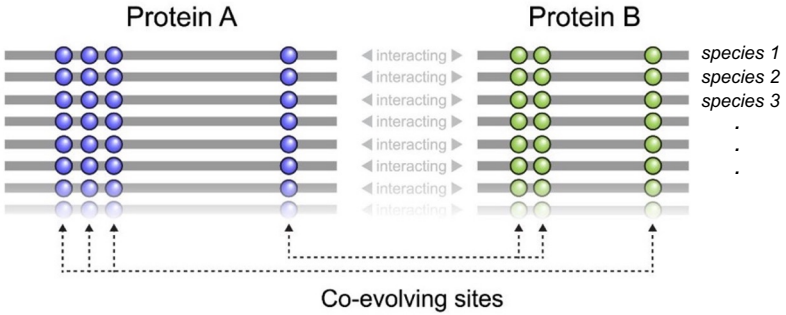
Baek, M., et al, *biorxiv* (2022)

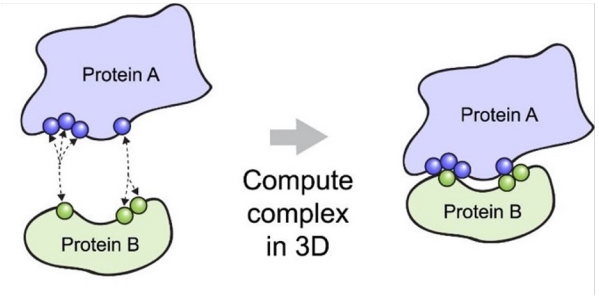
De novo functional protein design

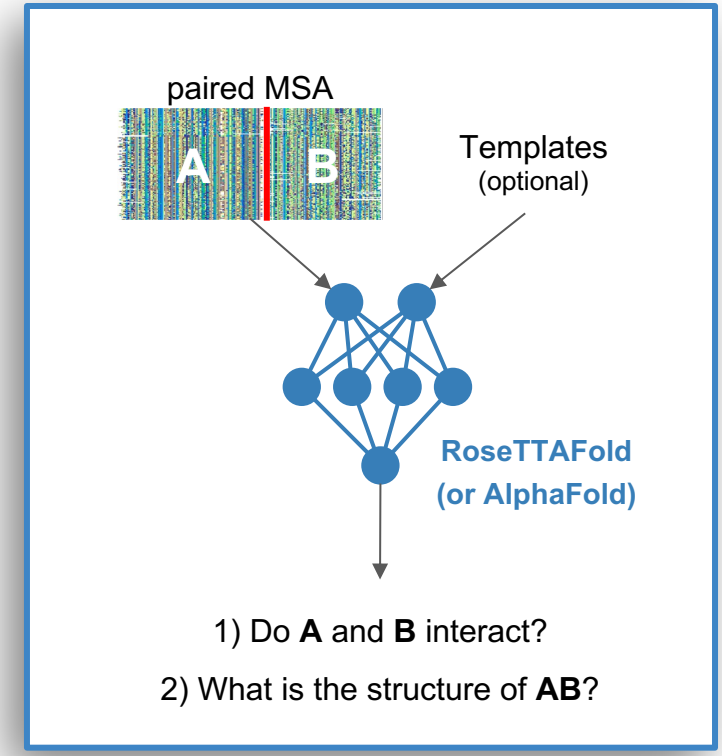


Wang, J., et al., *Science* (2022)

Protein-protein complex structure prediction

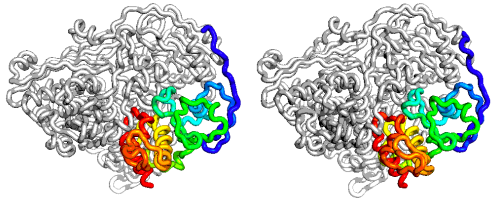






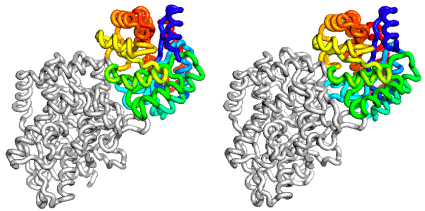
Protein-protein complex structure prediction

Aldehyde oxidoreductase



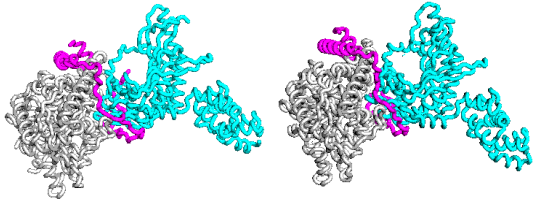
TM-score: 95

Tryptophan synthase

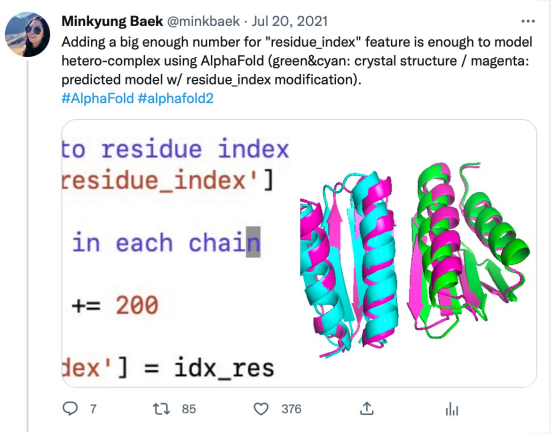


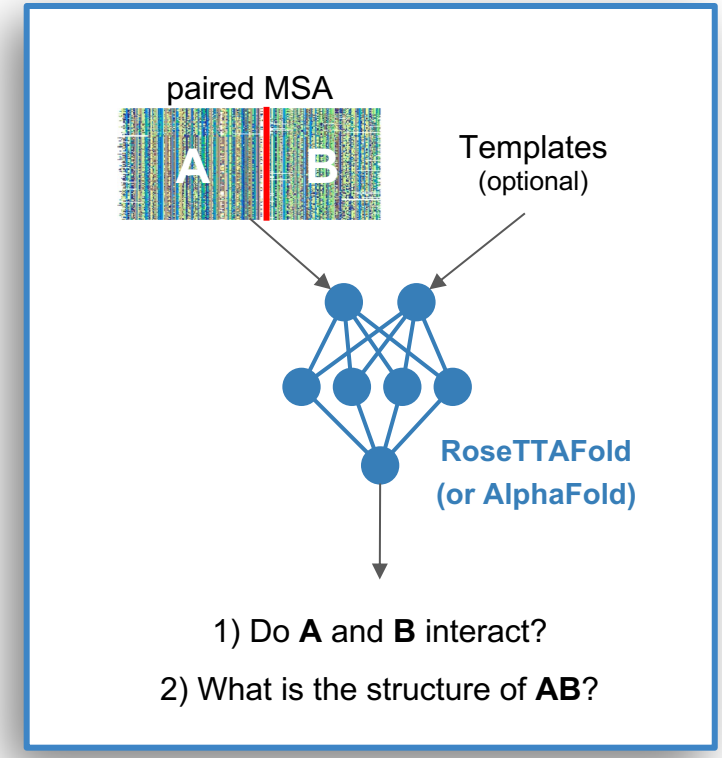
TM-score: 92

tRNA-dependent amidotransferase

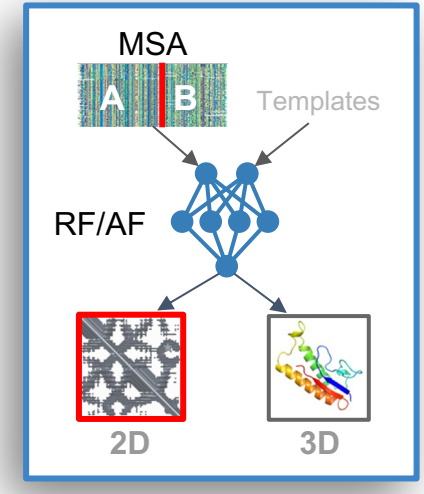


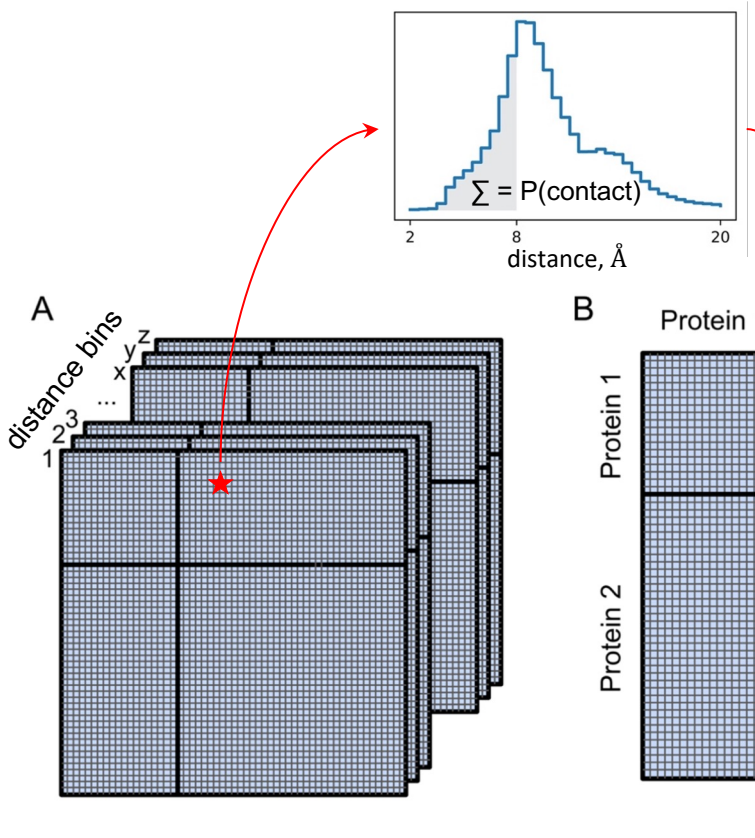
TM-score: 89





Predicting protein-protein interactions

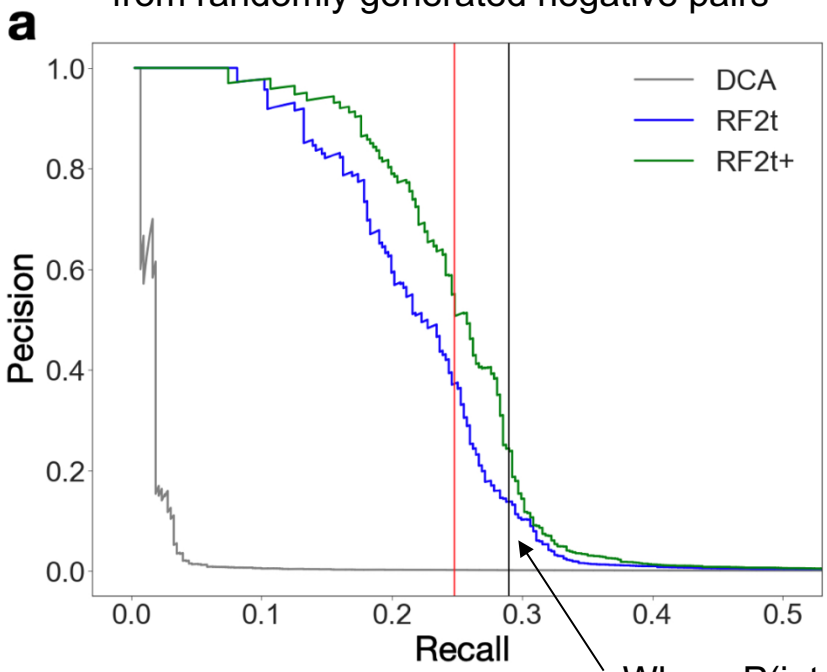




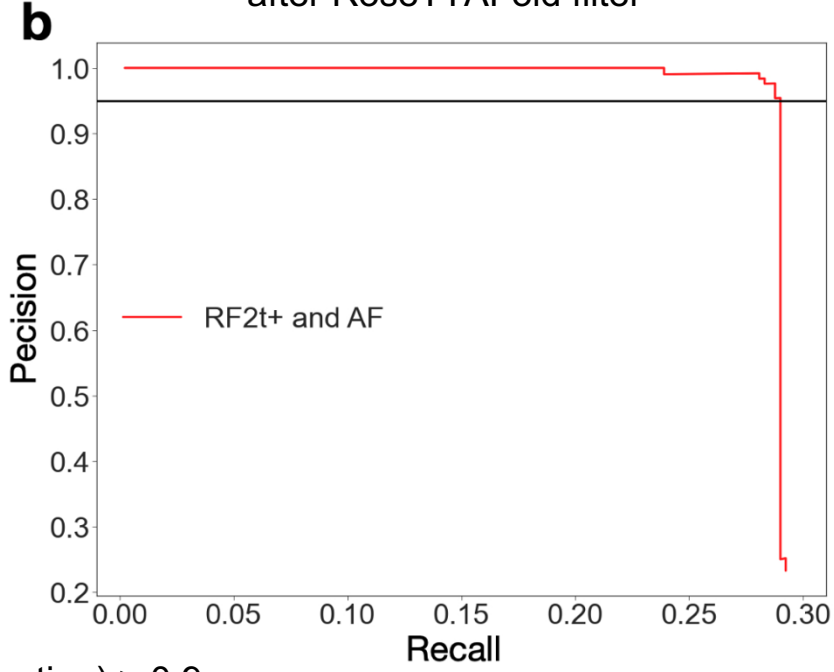
highest prob. contact
≡
prob. of interaction

Predicting protein-protein interactions

Distinguish 768 gold-standard pairs
from randomly generated negative pairs



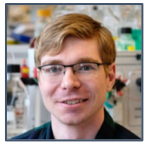
Distinguish 717 gold-standard pairs
after RoseTTAFold filter



Where P(interaction) > 0.9

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

In silico PPI screening: Yeast interactome

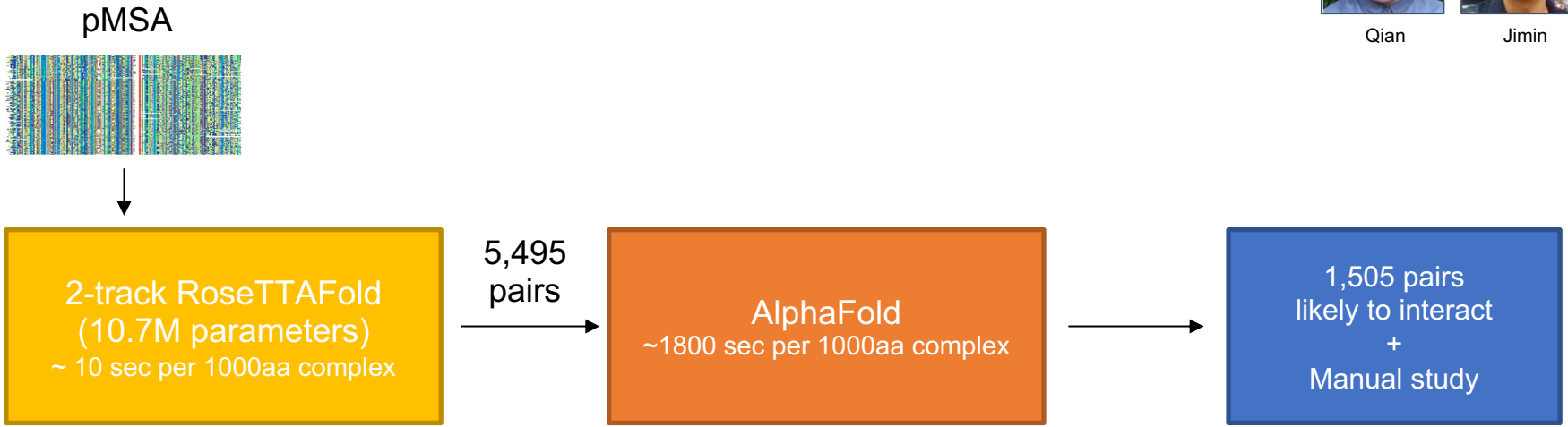

Ian


Aditya


Qian

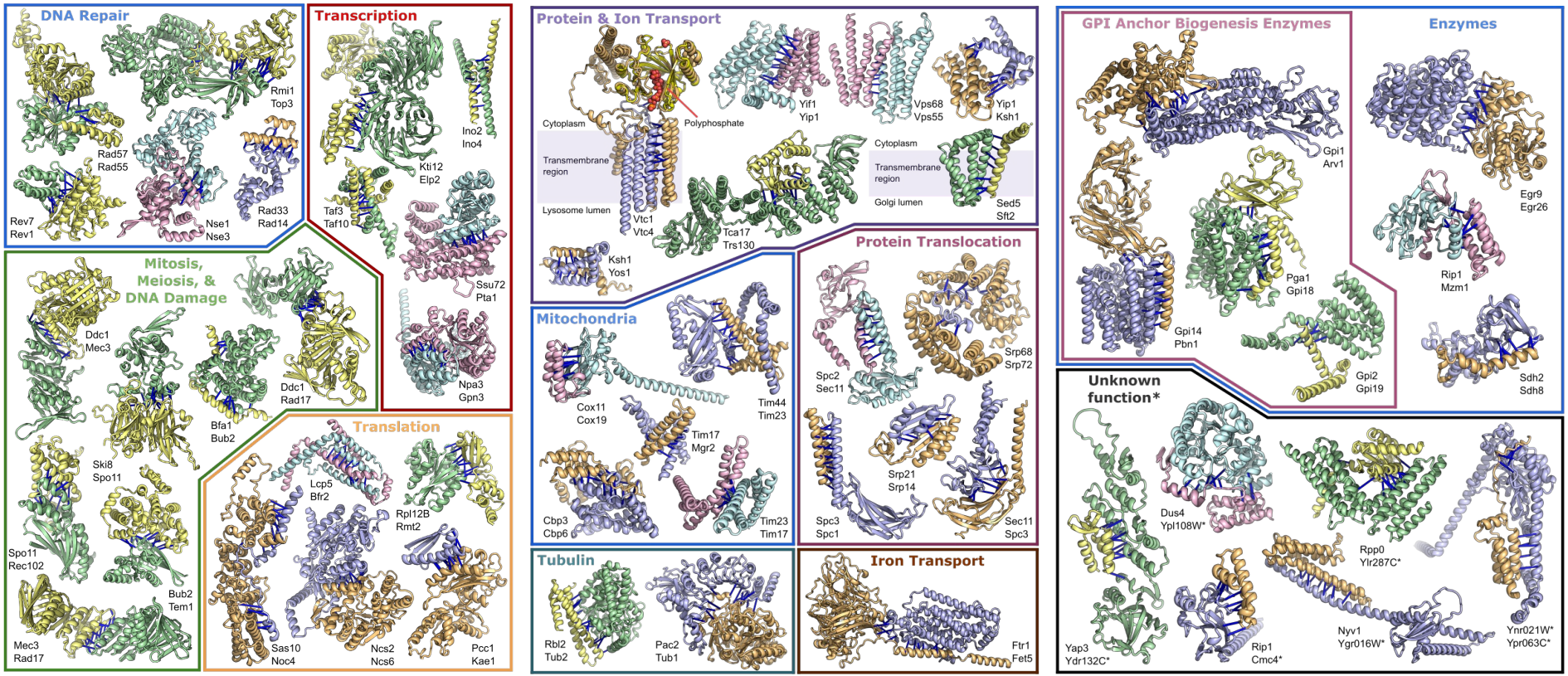

Jimin

- 4.3 million protein pairs



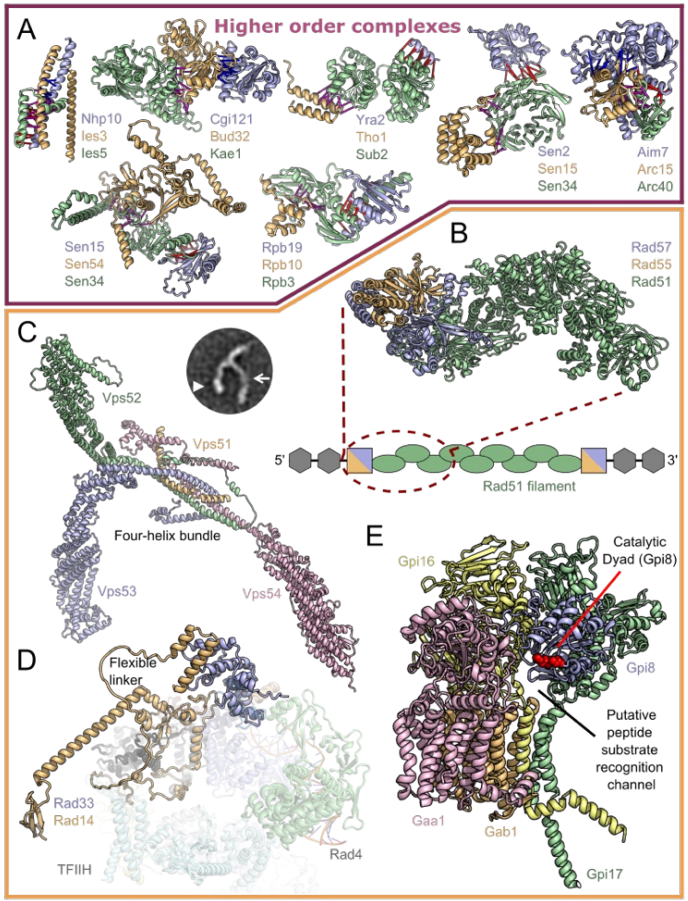
Humphreys, I.R., Pei, J., Baek, M., Krishnakumar, A., et al, *Science*, (2021)

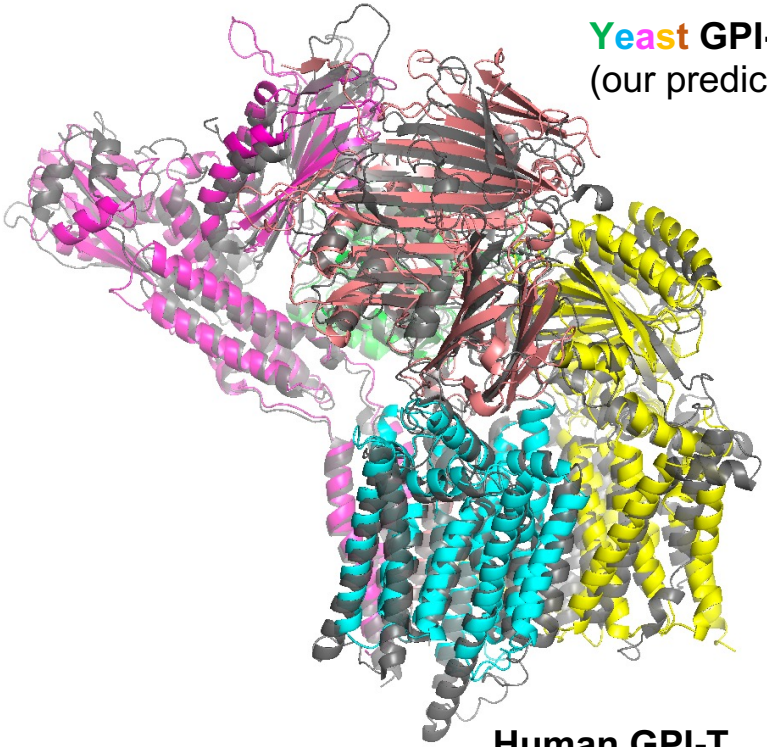
In silico PPI screening: Yeast interactome



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

In silico PPI screening: Yeast interactome

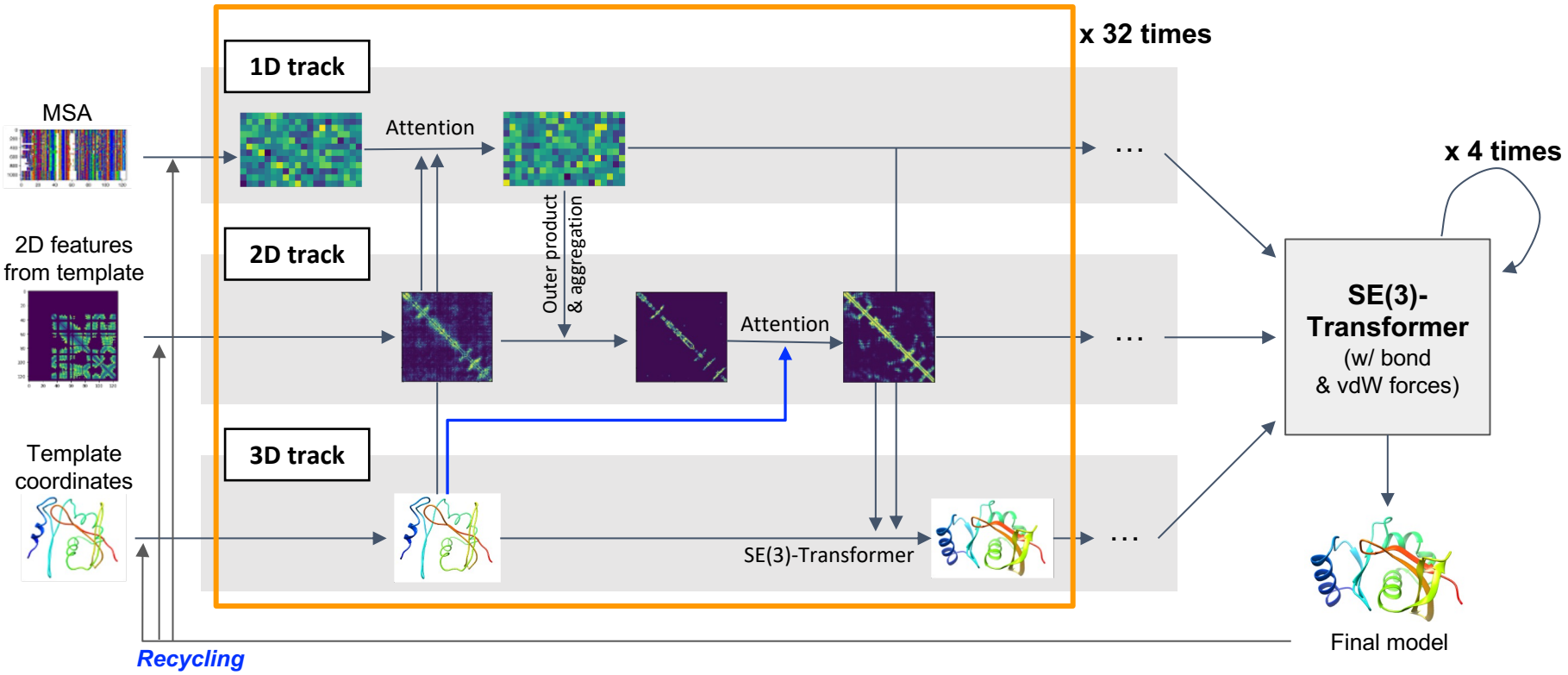




Yeast GPI-T
(our prediction, Oct 2021)

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

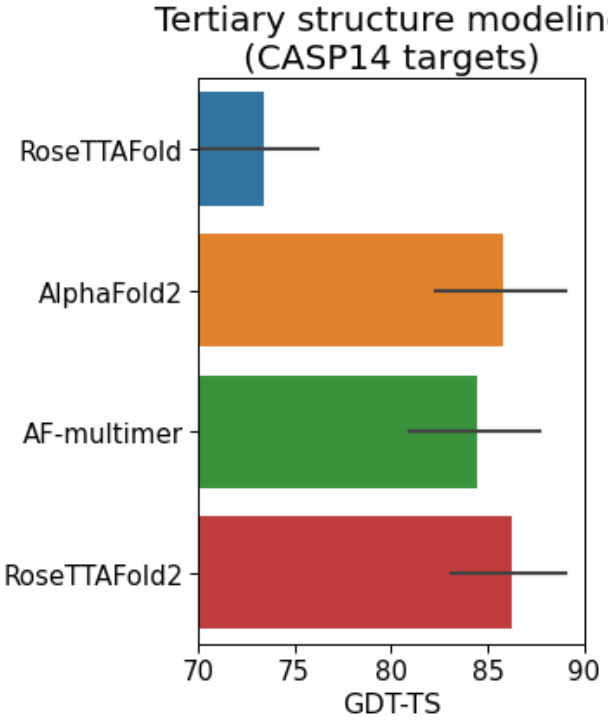
RoseTTAFold2: improving RF for better modeling & screening

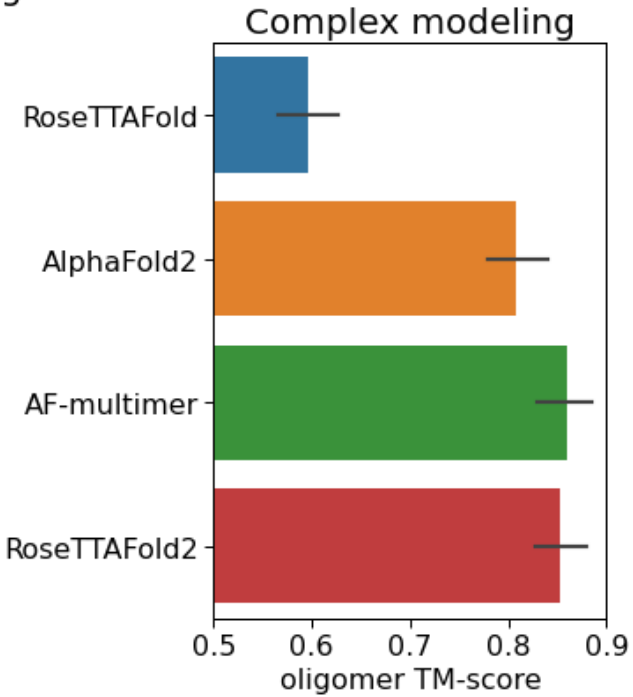


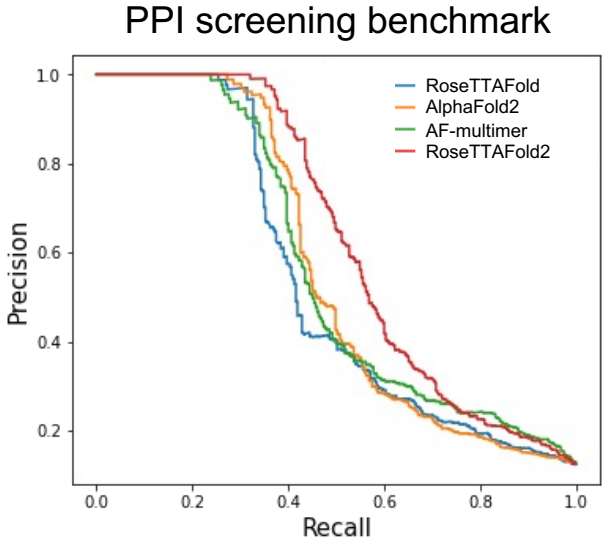
Baek, M. et al., *In preparation*

Trained on PDB (monomer+**multimer**) + **UniRef50 models** + **negative PPI sets**

RoseTTAFold2: improving RF for better modeling & screening





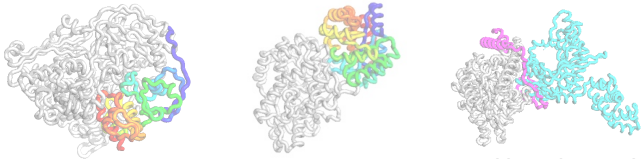


Baek, M. et al., *In preparation*

Beyond accurate modeling of protein structures

Large-scale *in silico* PPI screening

Protein A ⟺ Protein B

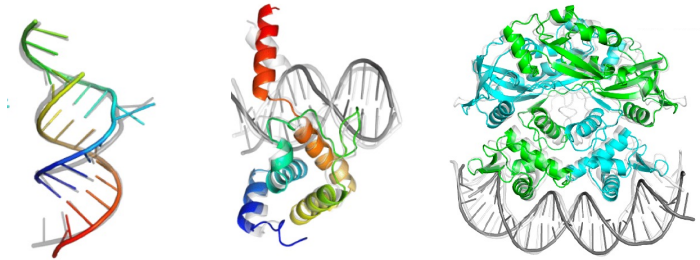


1) Do **A** and **B** interact?

2) What is the structure of **AB**?

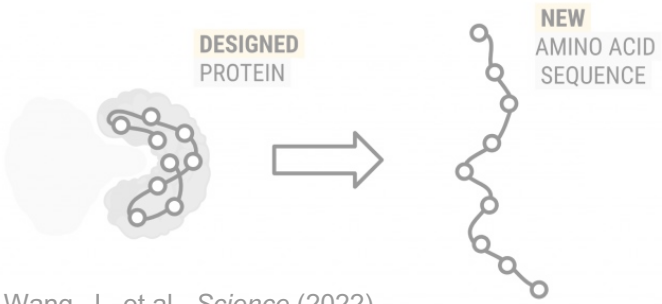
Humphreys, I., Pei, J., Baek, M., Krishnakumar, A., et al, *Science* (2021)

Nucleic acid structure and interaction modeling



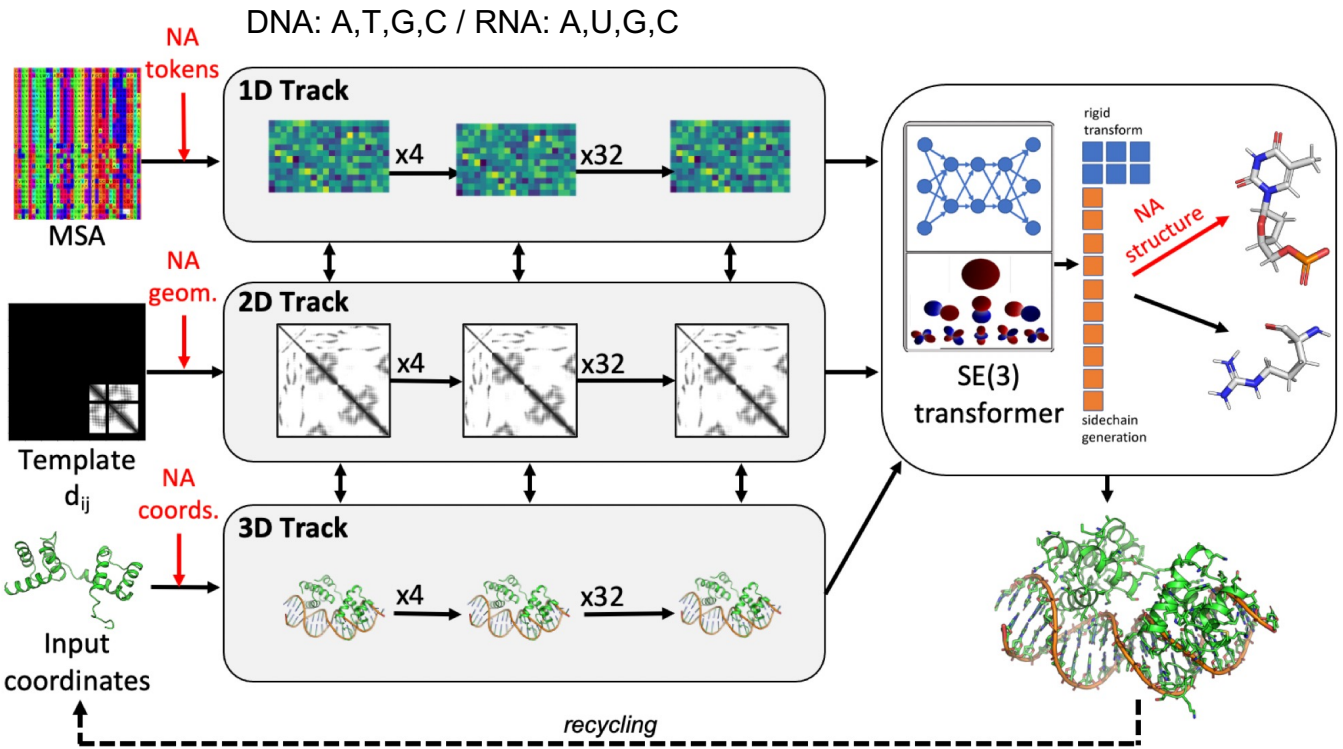
Baek, M., et al, *biorxiv* (2022)

De novo functional protein design



Wang, J., et al., *Science* (2022)

Extending RoseTTAFold – nucleic acid prediction



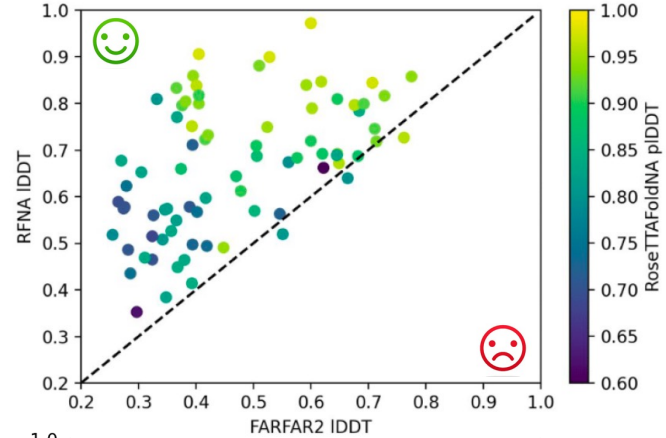
Frank DiMaio

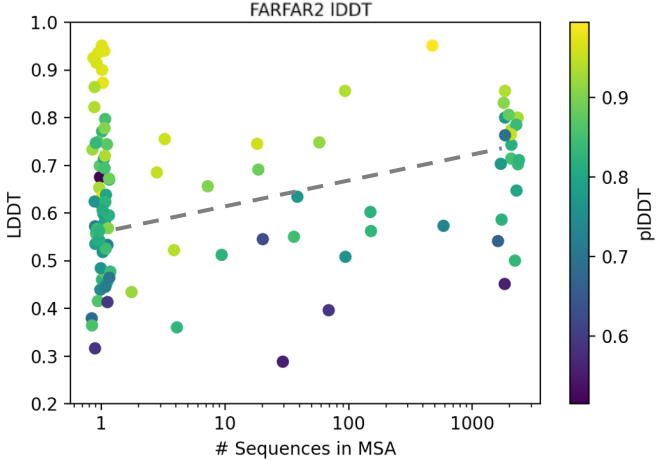
Trained on protein only set (>26k clusters) + ***nucleic acid included structures (~3k clusters)***

Baek, M., McHugh, R., Anishchenko, I., Baker, D., DiMaio, F., et al, *biorxiv* (2022)

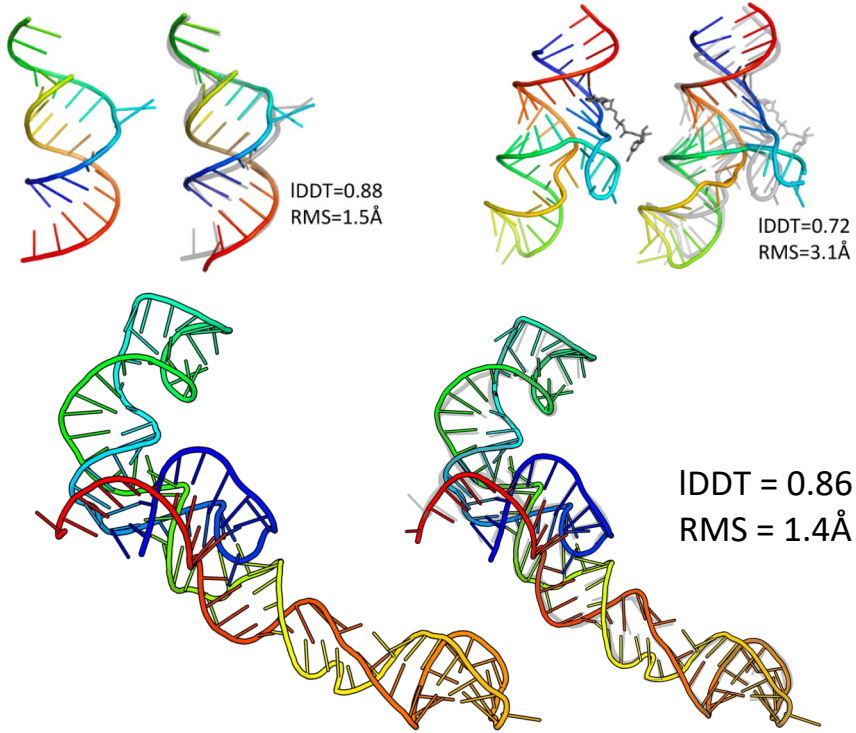
RoseTTAFoldNA shows promising results

RNA structure prediction





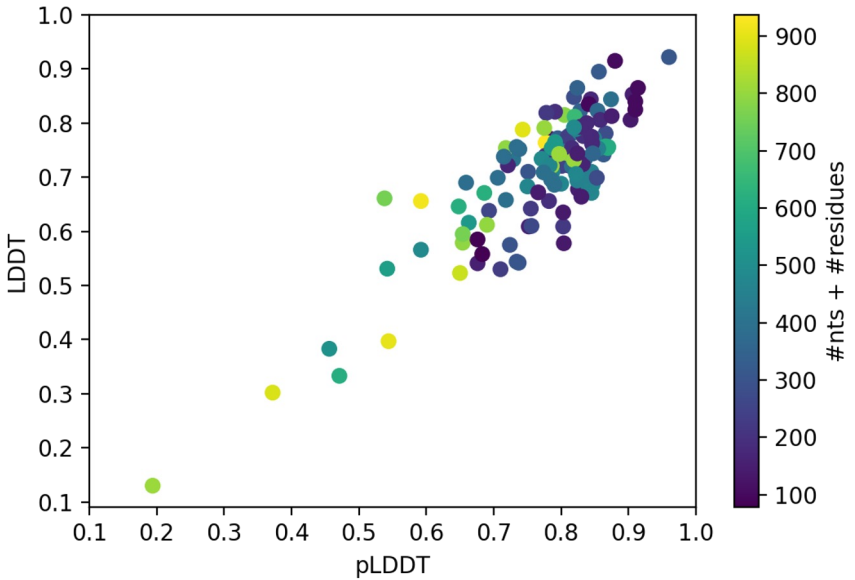
Ground-truth (left) vs Predicted (right)



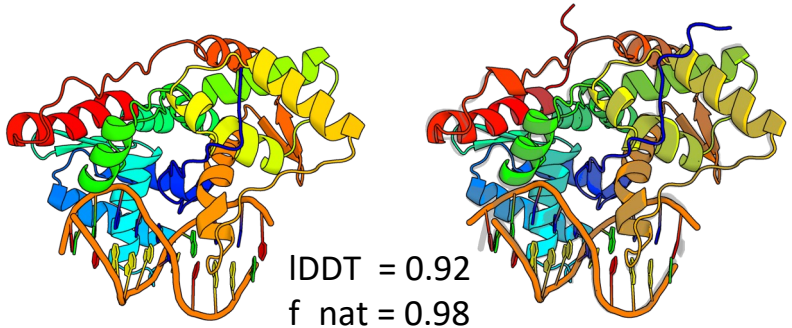
Baek, M., McHugh, R., Anishchenko, I., Baker, D., DiMaio, F., et al, *biorxiv* (2022)

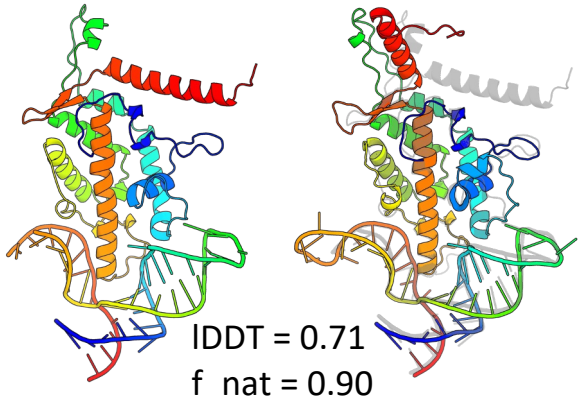
RoseTTAFoldNA shows promising results

Protein-nucleic acid complex prediction



Ground-truth (left) vs Predicted (right)



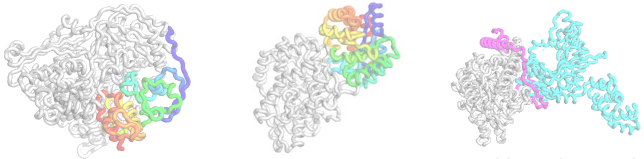


Baek, M., McHugh, R., Anishchenko, I., Baker, D., DiMaio, F., et al, *biorxiv* (2022)

Beyond accurate modeling of protein structures

Large-scale *in silico* PPI screening

Protein A ⟺ Protein B

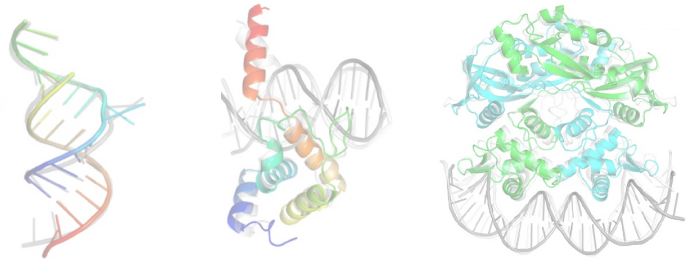


1) Do **A** and **B** interact?

2) What is the structure of **AB**?

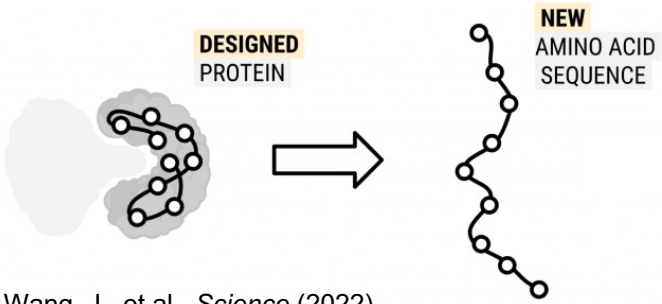
Humphreys, I., Pei, J., Baek, M., Krishnakumar, A., et al, *Science* (2021)

Nucleic acid structure and interaction modeling



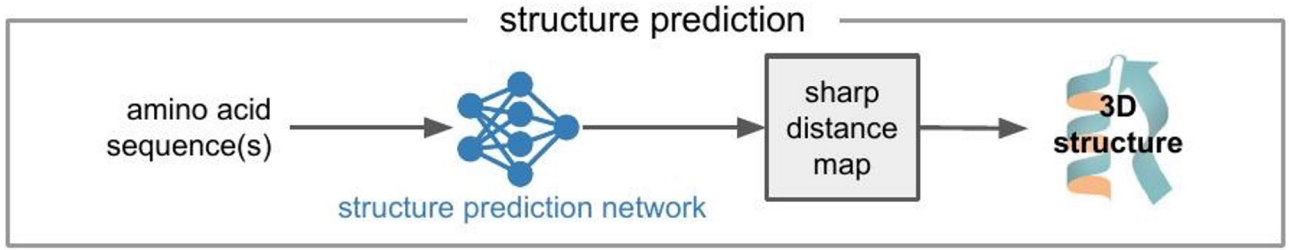
Baek, M., et al, *biorxiv* (2022)

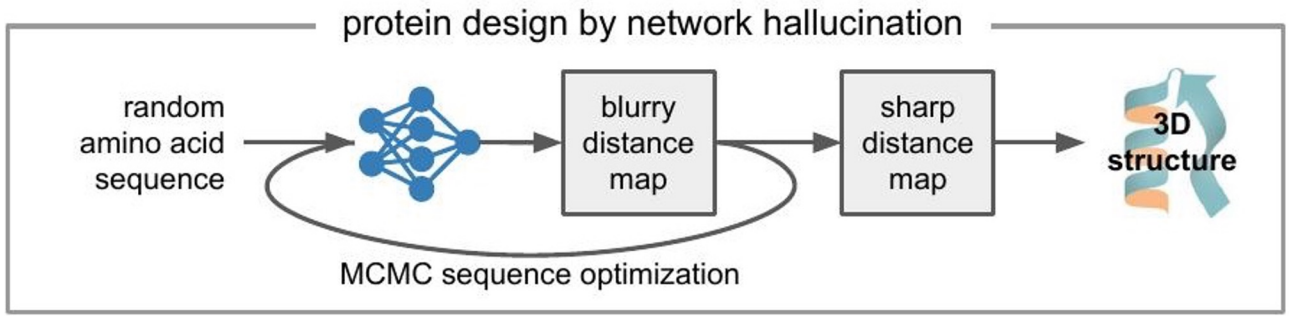
De novo functional protein design



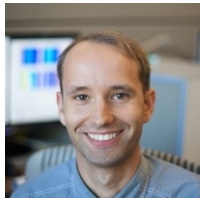
Wang, J., et al., *Science* (2022)

Hallucination: Optimize sequence to have a structure





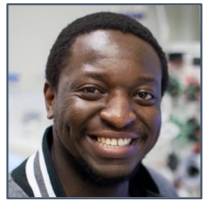
method development



Ivan

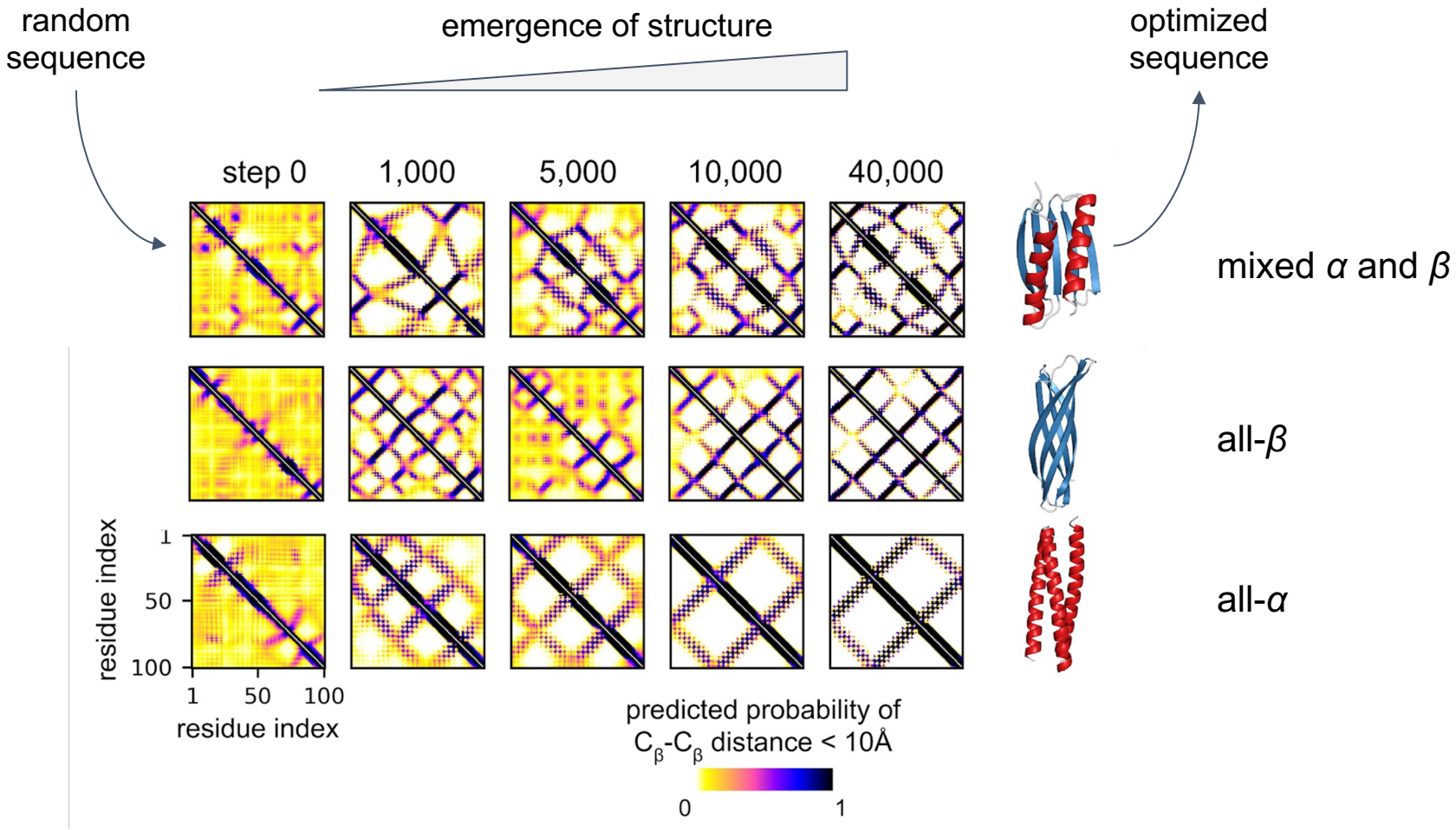
experimental validation

Sam



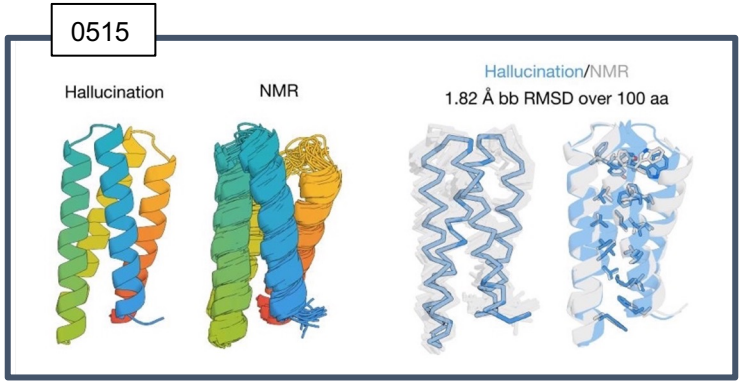
Tamuka

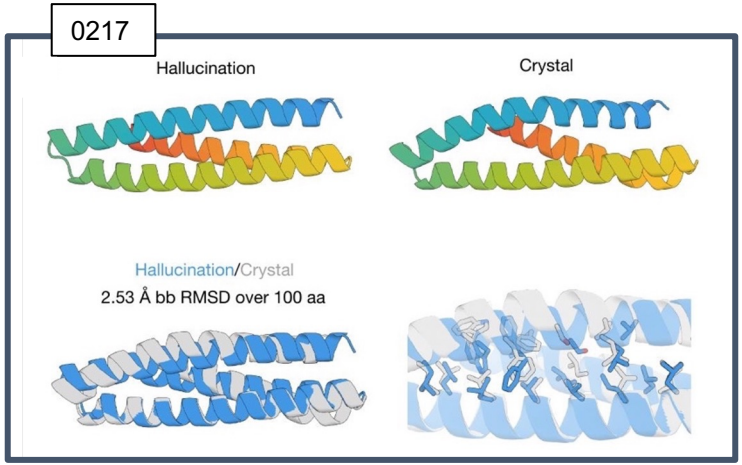
Anishchenko, I., Pellock, S., et al. *Nature* (2021)

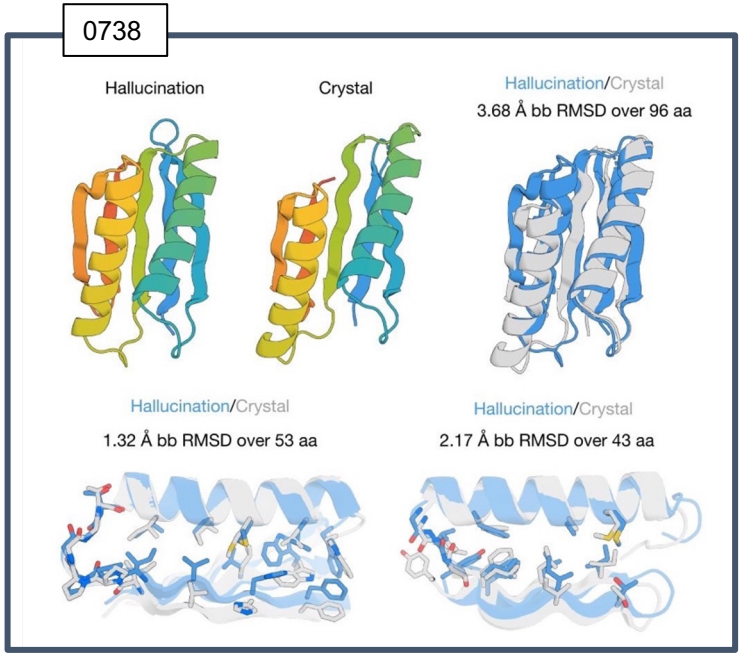


Anishchenko, I., Pellock, S., et al. *Nature* (2021)

Structures of 3 hallucinations were confirmed experimentally

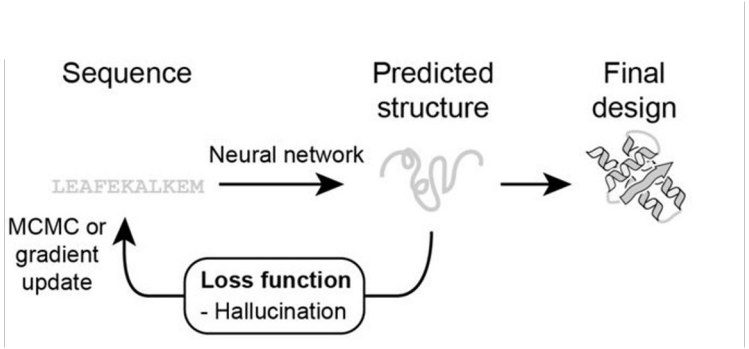




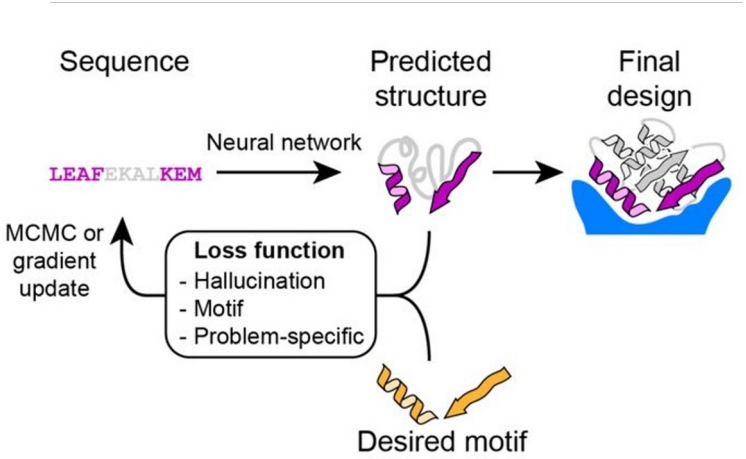


Anishchenko, I., Pellock, S., et al. *Nature* (2021)

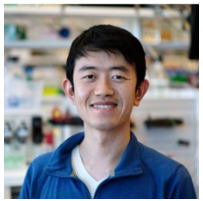
Constrained hallucination:
design a new protein having a given functional motif

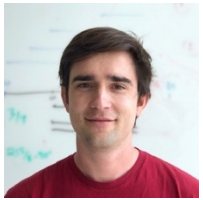


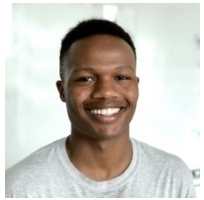
Free hallucination:
generate novel protein folds



Constrained hallucination:
generate scaffolds harboring
pre-specified functional sites


Jue


Doug

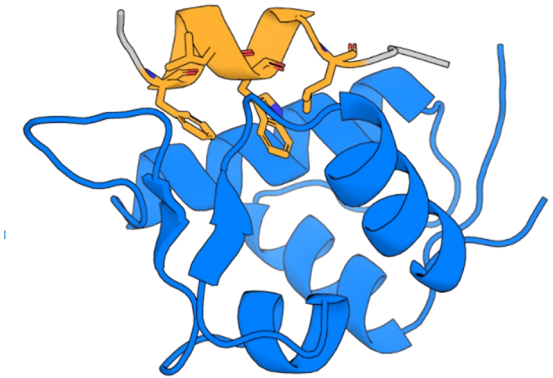

Sidney

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

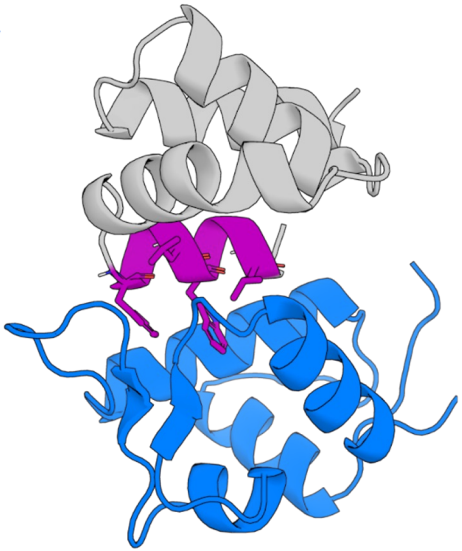
Applications of constrained hallucination

Scaffolding p53 helix to bind cancer-signaling protein mdm2

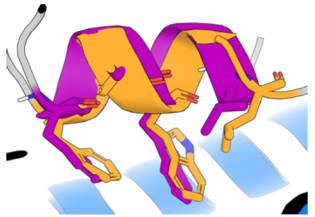
Native motif
Design motif
Binding partner



Native



Hallucination

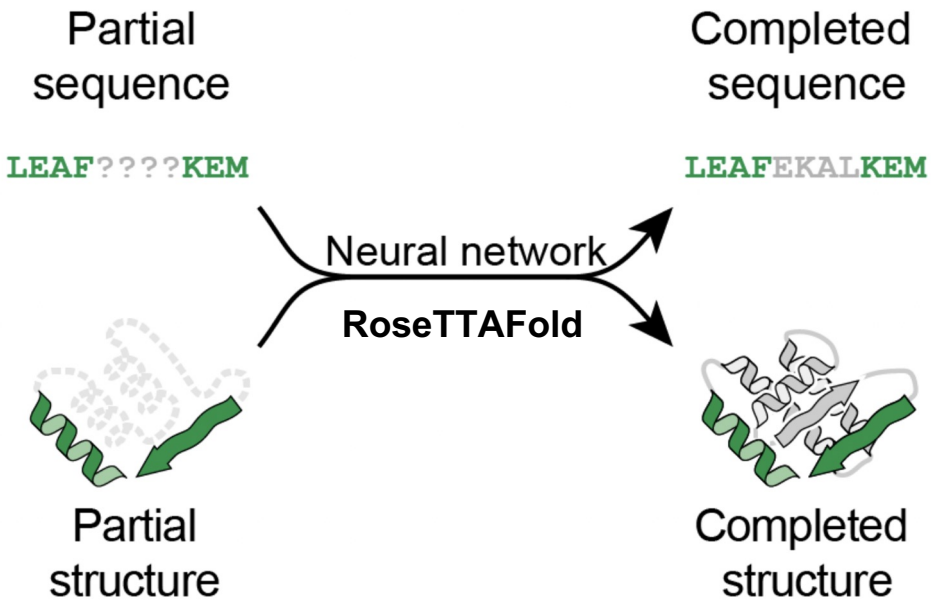


Native vs designed motif

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

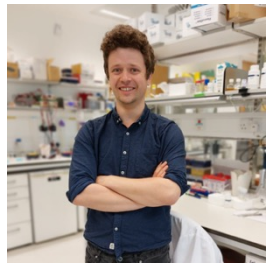
Protein Design via Inpainting

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



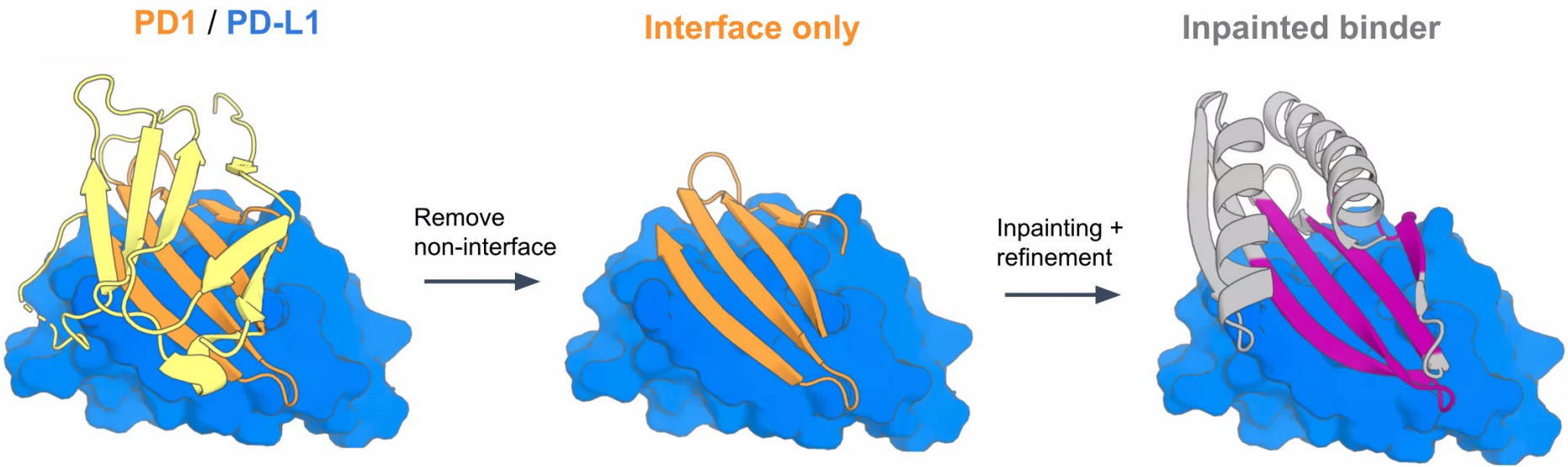
David Juergens

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



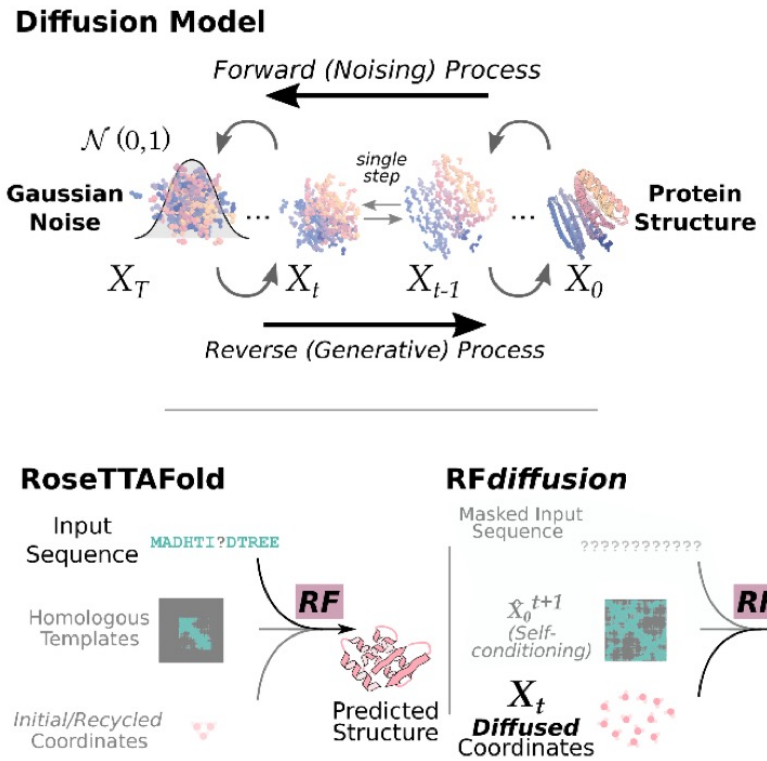
Joseph Watson

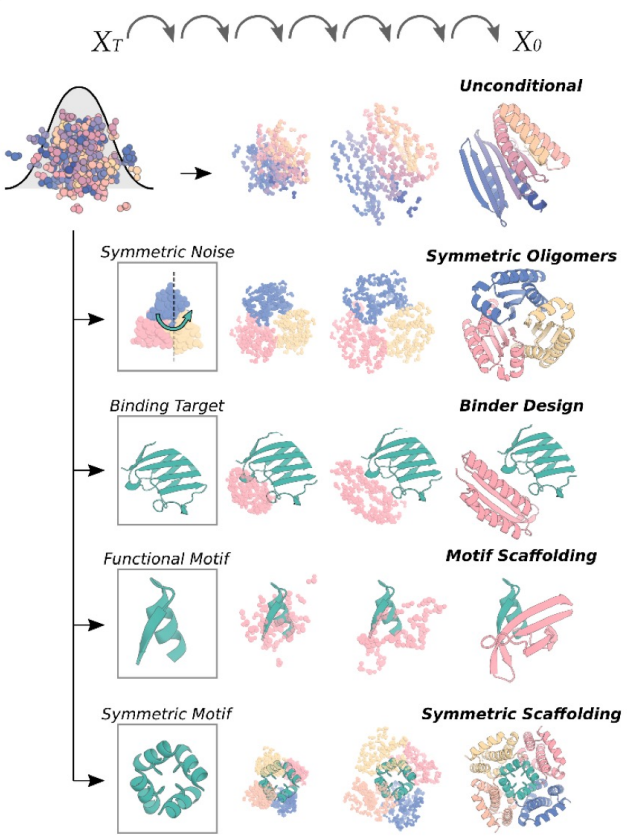
Design PD-L1/PD-1 binding inhibitor via inpainting



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

Generative model for protein design (RFdiffusion)

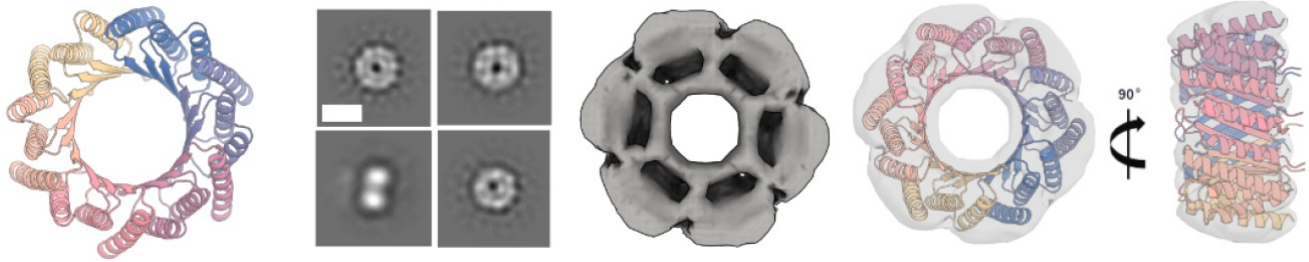


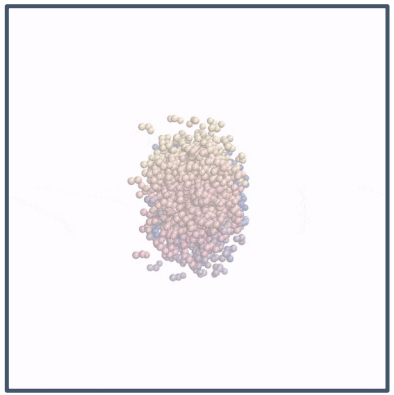


Watson, J., Juergens, D., Bennett, N., Trippe, B., Yim, J., Eisenach, H., Ahern, W., *Biorxiv* (2022)

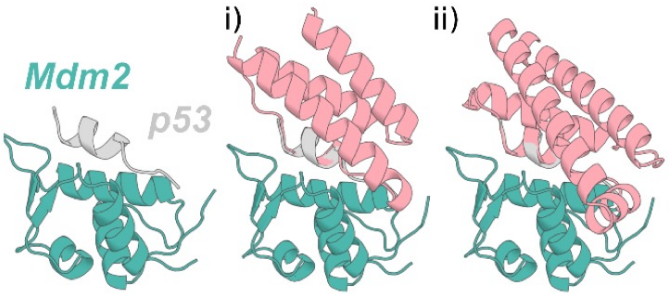
RFdiffusion can do various design tasks

High-order symmetric assemblies

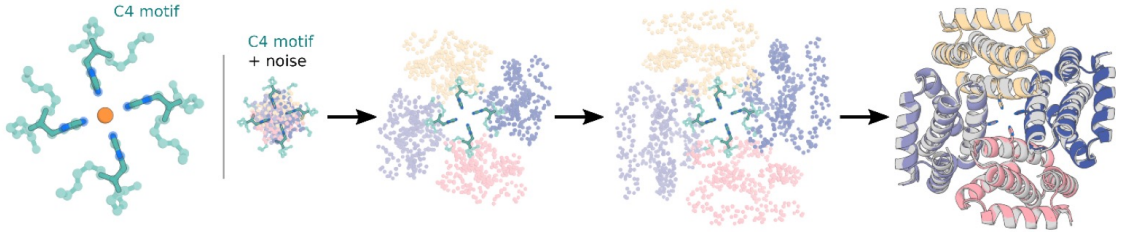


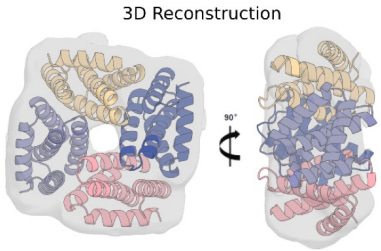


Motif scaffolding



K_D=0.7nM K_D=0.5nM

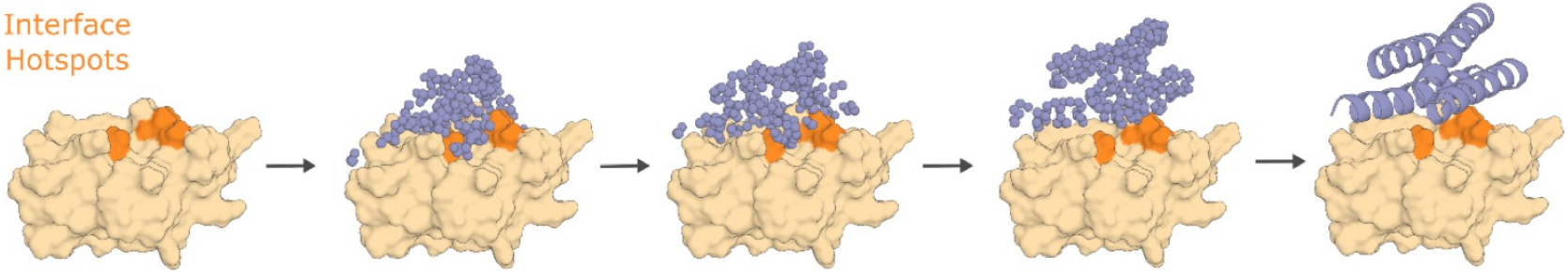


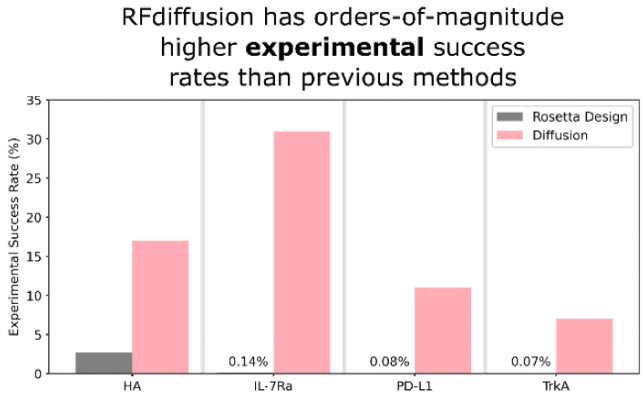


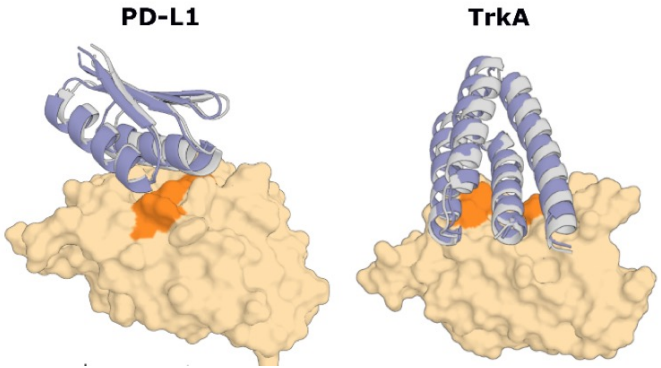
Watson, J., Juergens, D., Bennett, N., Trippe, B., Yim, J., Eisenach, H., Ahern, W., *Biorxiv* (2022)

RFdiffusion can do various design tasks

Binder design based on a given interface

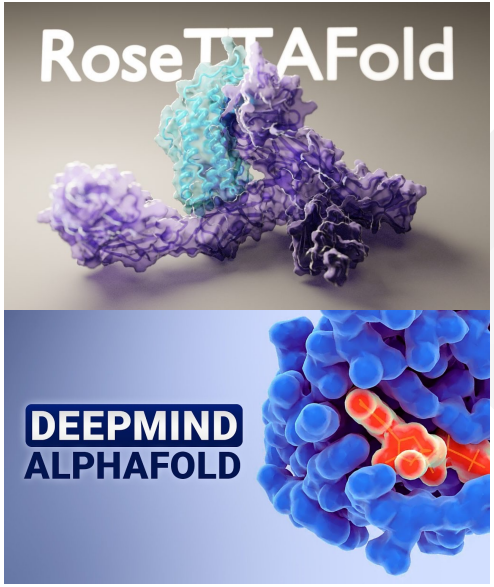




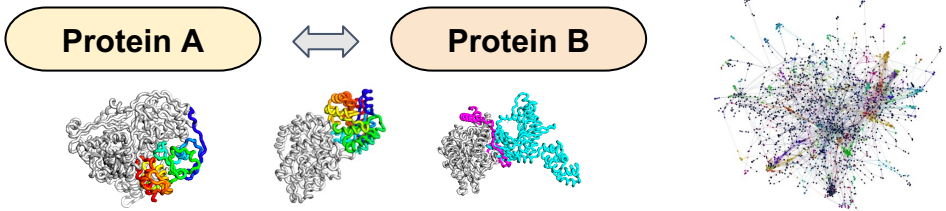


Watson, J., Juergens, D., Bennett, N., Trippe, B., Yim, J., Eisenach, H., Ahern, W., *Biorxiv* (2022)

AI-based protein modeling

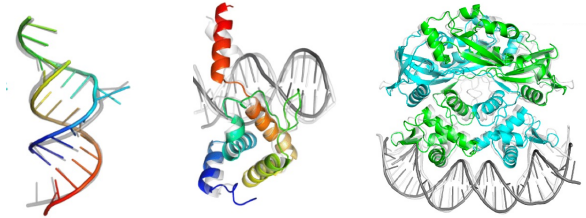


Large-scale *in silico* PPI screening



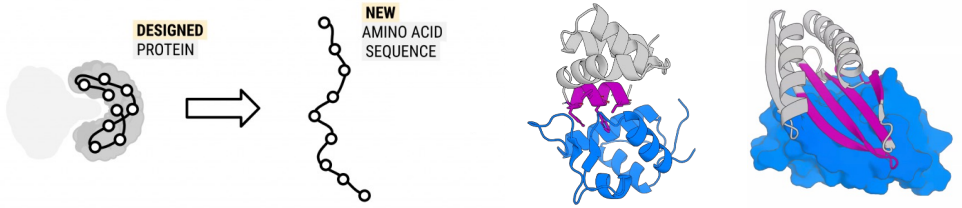
Humphreys, I., Pei, J., Baek, M., Krishnakumar, A., et al, *Science* (2021)

Nucleic acid structure & interaction prediction



Baek, M., McHugh, R., Anishchenko, I., Baker, D., DiMaio, F., et al, *biorxiv* (2022)

De novo functional protein design



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

Acknowledgements

RoseTTAFold
Frank DiMaio
Ivan Anishchenko
Justas Dauparas
Sergey Ovchinnikov
Jue Wang

PPI screening
Qian Cong
Ian Humphreys
Aditya Krishnakumar
Jimin Pei

Hallucination
David Juergens
Joseph Watson

Inpainting & Diffusion
Joseph L. Watson
David Juergens
Nathaniel R Bennett
Brian L Trippe
Jason Yim
Helen E. Eisenach
Woody Ahern

Diffusion
Ivan Anishchenko
Doug Tischer
Jue Wang
Sidney Lisanza
Sam Pellock
Tamuka Chidyausiku
Sergey Ovchinnikov
Chris Norn

IT support
Luki Goldschmidt
David Kim
Microsoft
AWS

Discussion
David Baker
Lance Stewart
Eric Horvitz

Looking for graduate students & postdoc!!
If you're interested in, please e-mail me (minkbaek@snu.ac.kr)