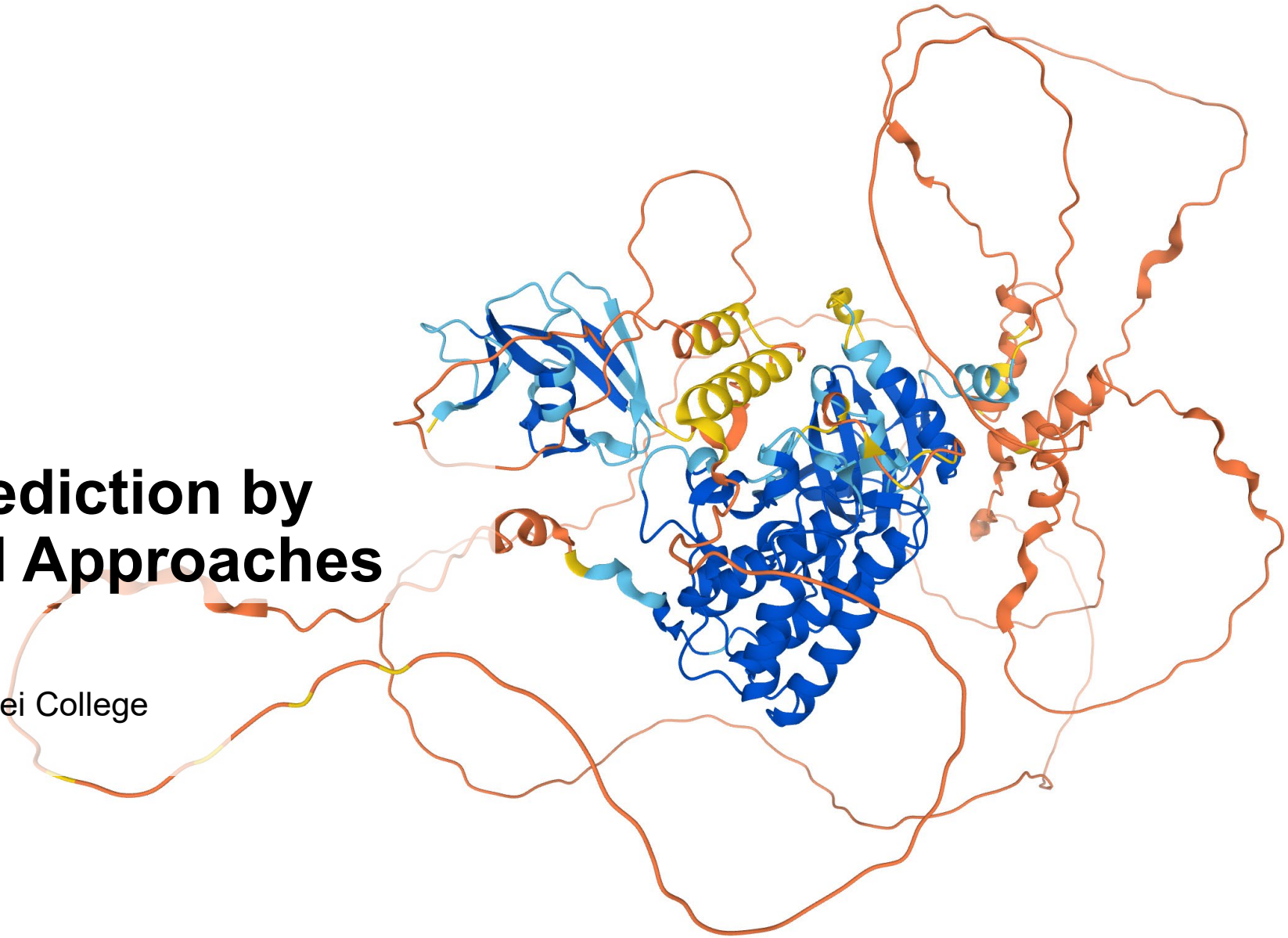
Discussion

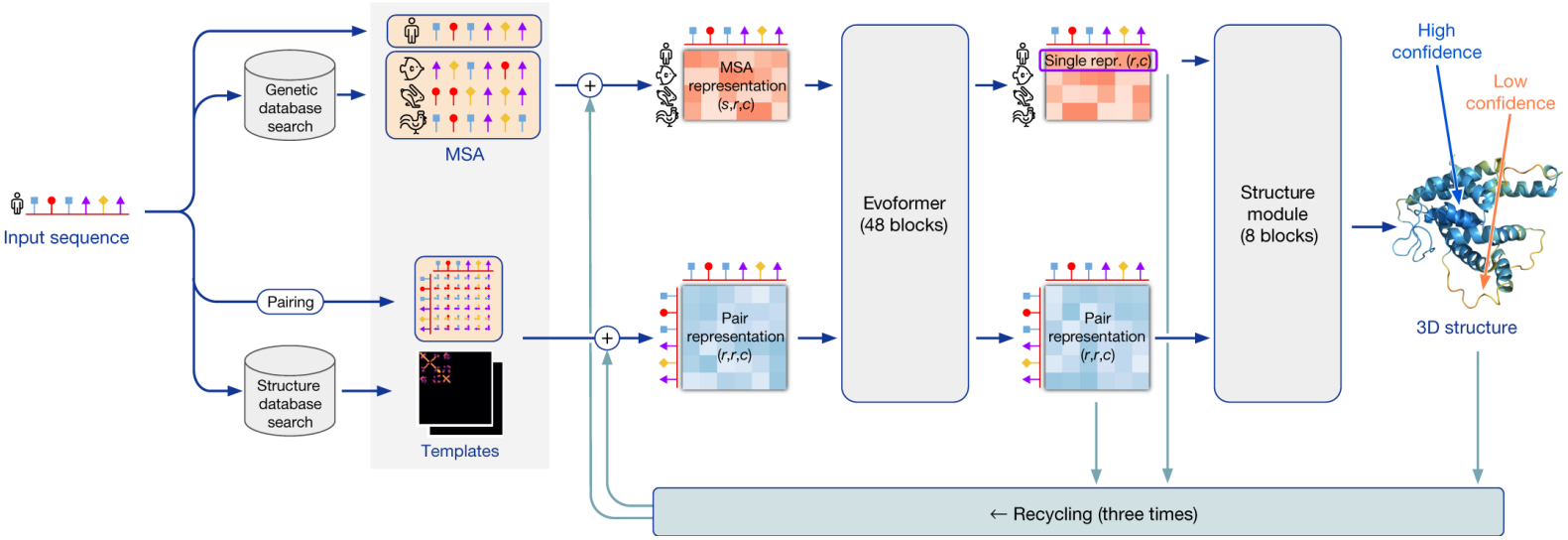
Protein Structure Prediction by Deep-learning-based Approaches

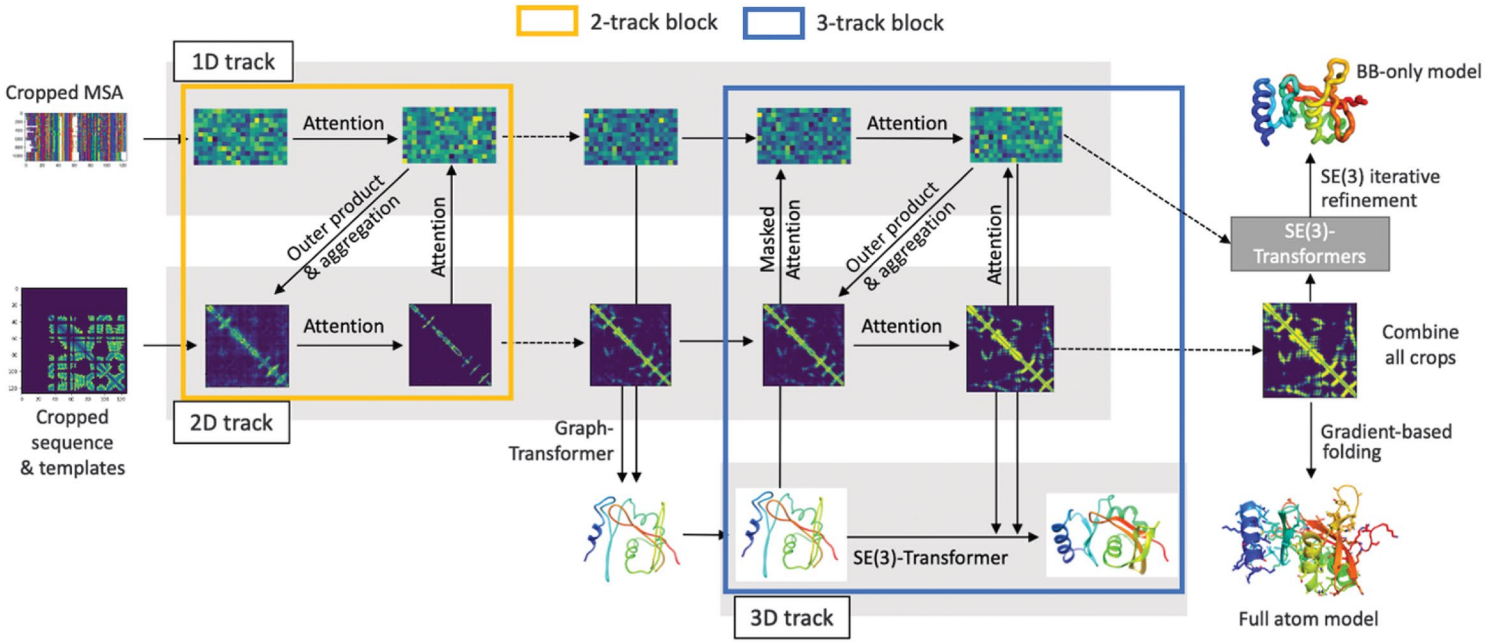
Yongcheng Jiang

2018 Integrated Science Program, Yuanpei College



Review of previous parts





Methods	AF2	RF
Input	MSA, paired rep.	
Neural net.	Attention	
Track num.	2-track	3-track
Speed	Slower	Faster
Database	Yes	No

Jumper, J. *Nature* (2021)

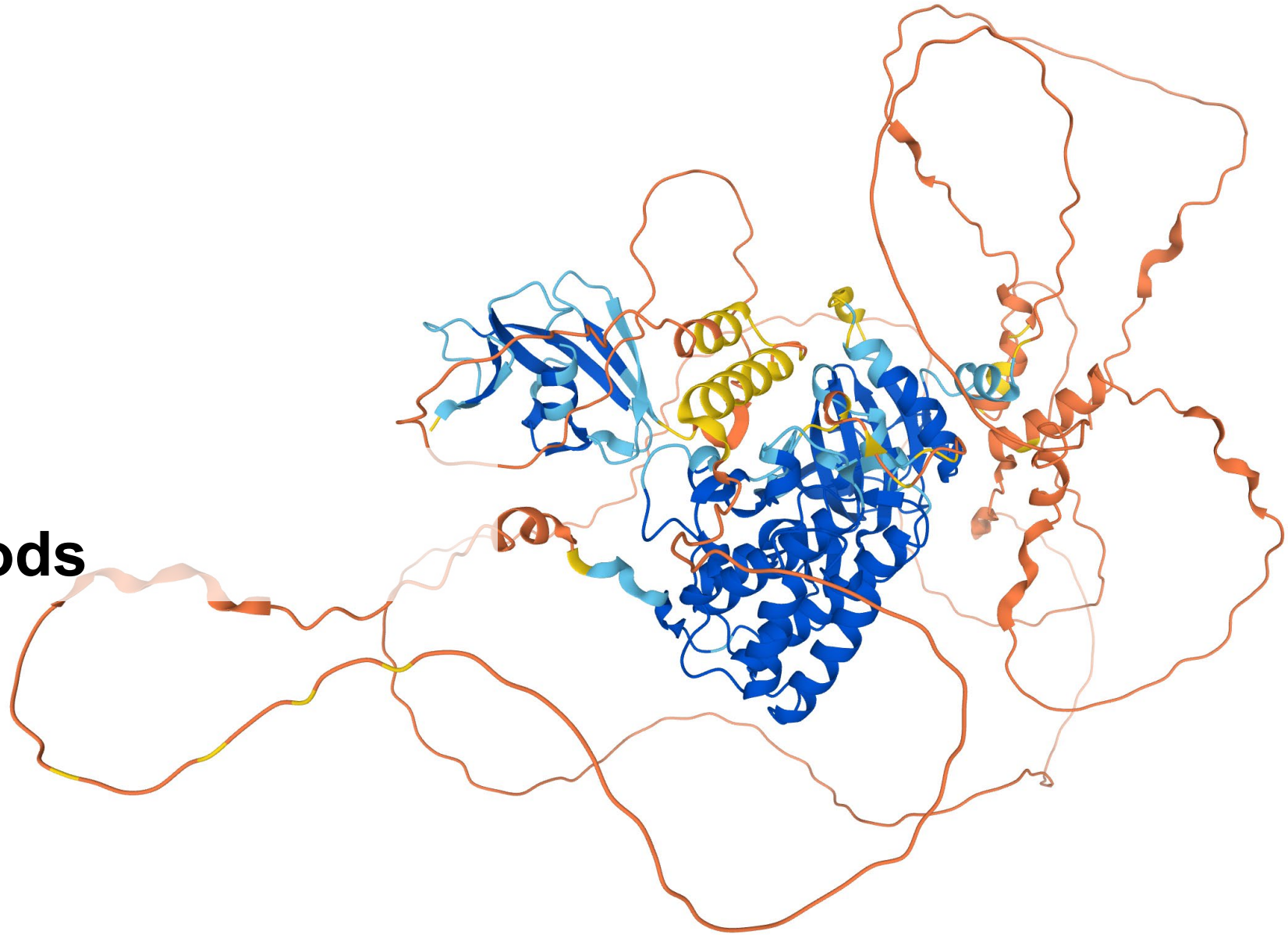
Baek, M. *Science* (2021)

Outline

- Benchmarking deep-learning methods
- Emerging research works involving AlphaFold2 or RoseTTAFold
- Remaining opportunities and challenges for structural biology

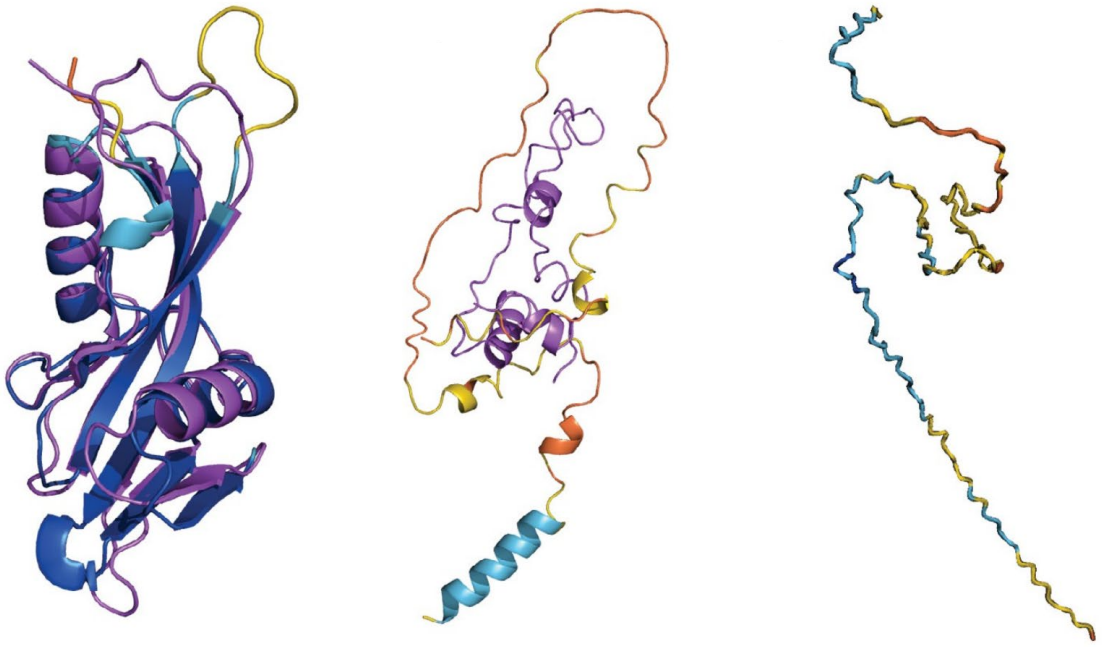
Discussion

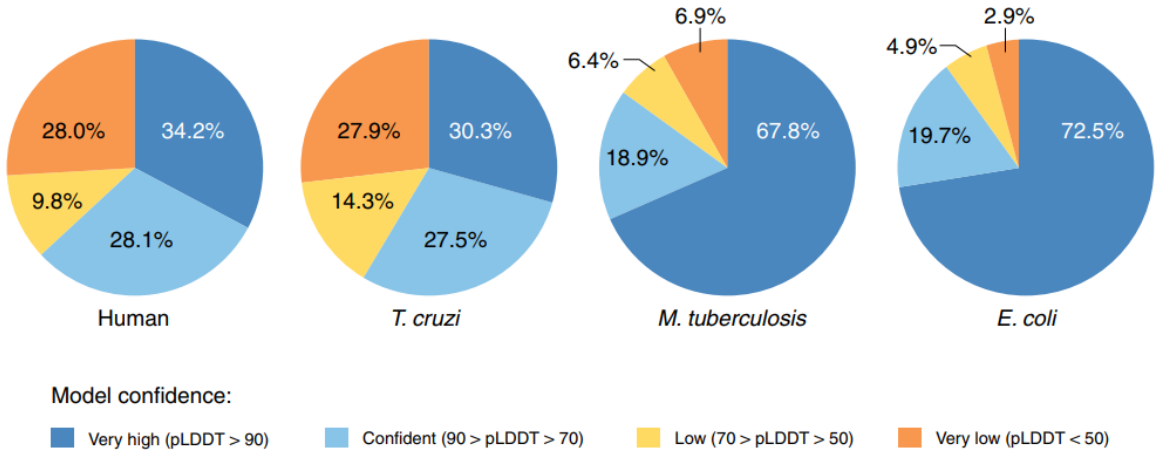
Benchmarking Deep-learning methods

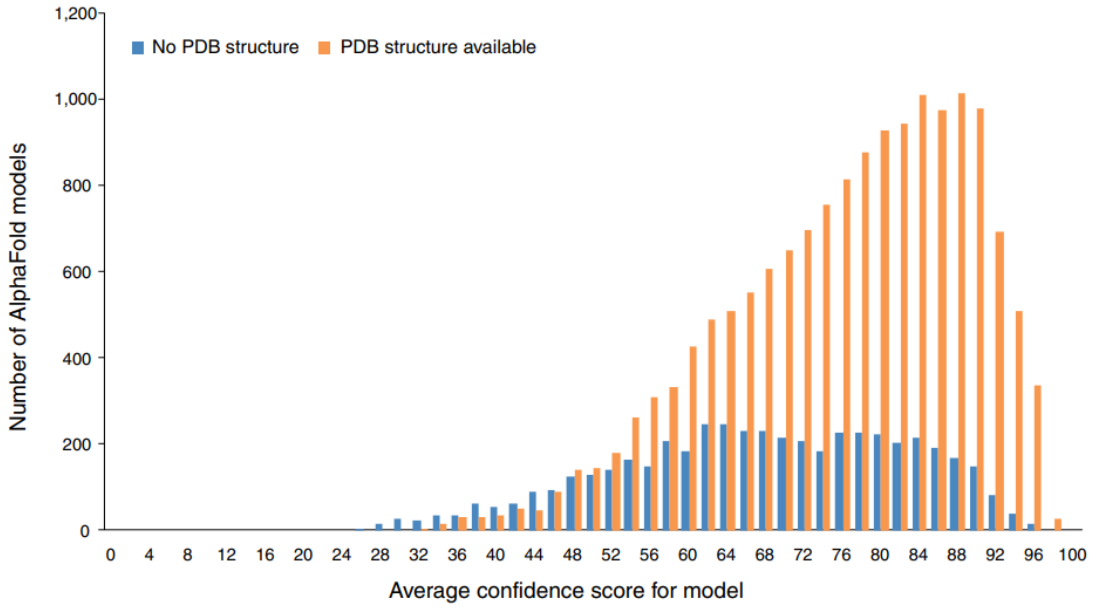


The inevitable doubts deep-learning methods encounter

- Strong MSA-derived bias?
- Over-engineered models?
- No new biological insights?





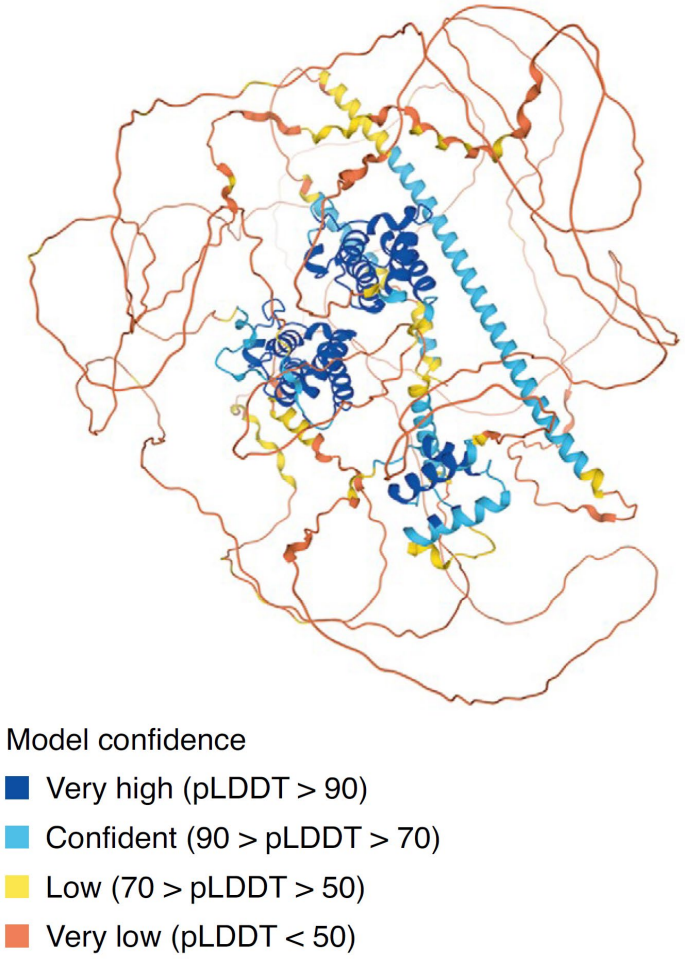


Thornton, J. M. *Nature Medicine* (2021)

Benchmarking AlphaFold2 and RosettaFold requires care and attention

Benchmark = validate a method using various datasets

- What are the strengths and drawbacks?
- Are they immediately applicable for structural biologist?
- Are low-confidence structures completely useless?



Subramaniam, S. *Nature Methods* (2022)

A structural biology community assessment of AlphaFold2

Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

New Results

Follow this preprint

A structural biology community assessment of AlphaFold 2 applications

Mehmet Akdel, Douglas E V Pires, Eduard Porta Pardo, Jürgen Jänes, Arthur O Zalevsky, Bálint Mészáros, Patrick Bryant, Lydia L. Good, Roman A Laskowski, Gabriele Pozzati, Aditi Shenoy, Wensi Zhu, Petras Kundrotas, Victoria Ruiz Serra, Carlos H M Rodrigues, Alistair S Dunham, David Burke, Neera Borkakoti, Sameer Velankar, Adam Frost, Kresten Lindorff-Larsen, Alfonso Valencia, Sergey Ovchinnikov, Janani Durairaj, David B Ascher, Janet M Thornton, Norman E Davey, Amelie Stein, Arne Elofsson, Tristan I Croll, Pedro Beltrao

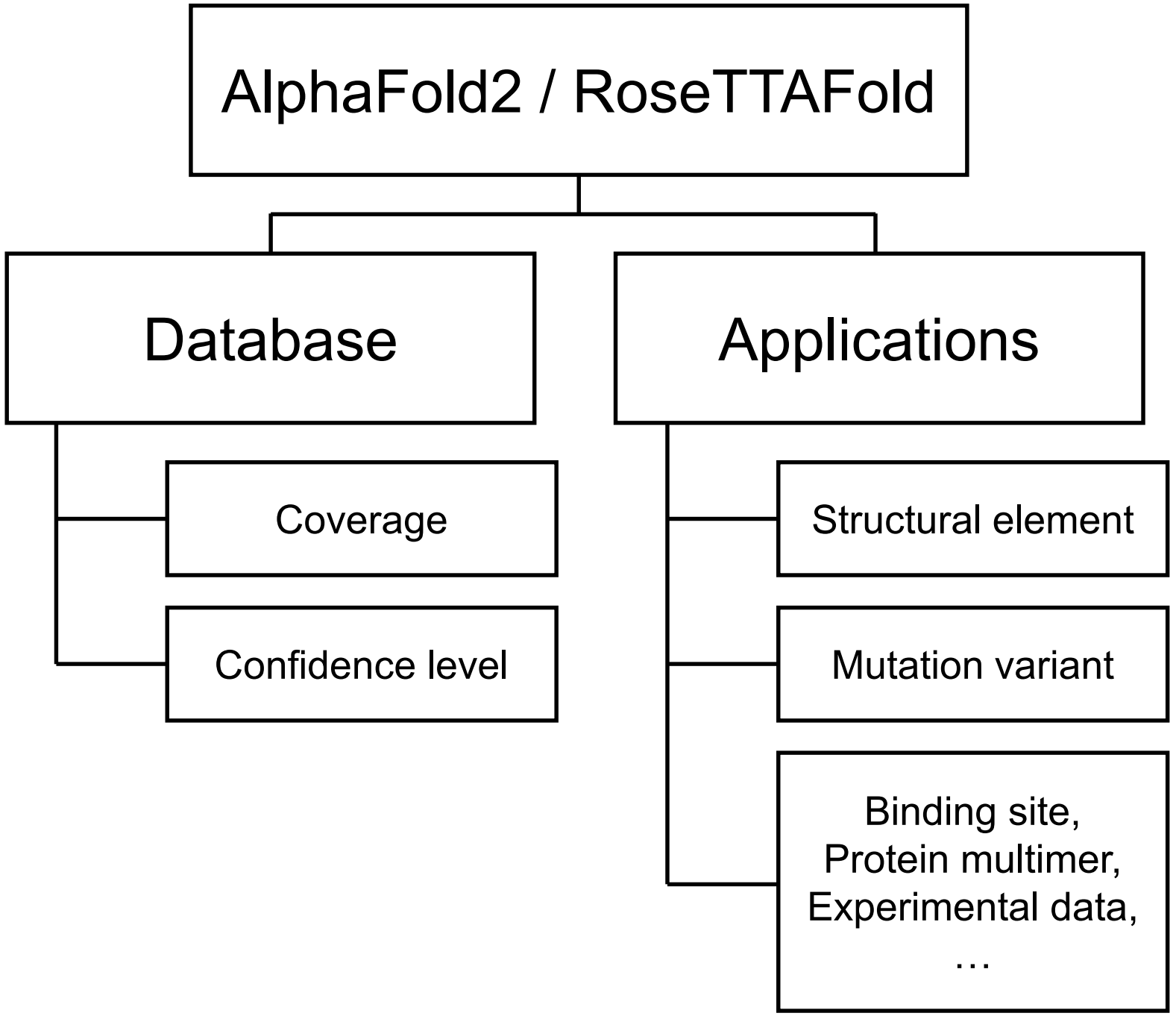
doi: https://doi.org/10.1101/2021.09.26.461876

Pedro Beltrão

2022- ETH Zürich

2013-2021 EMBL-EBI

Akdel, M. *BioRxiv* (2021)


AlphaFold2 / RoseTTAFold
Database
Applications
Coverage
Confidence level
Structural element
Mutation variant
Binding site,
Protein multimer,
Experimental data,
…

Existing databases have already generated hundreds of thousands of protein models

SWISS-MODEL Repository
(homology modeling)

SIB
Swiss Institute of Bioinformatics

BIOZENTRUM
Universität Basel
The Center for
Molecular Life Sciences

Pfam database
(trRosetta)

EMBL-EBI

+

AlphaFold2 database
(AlphaFold2)

EMBL-EBI

DeepMind

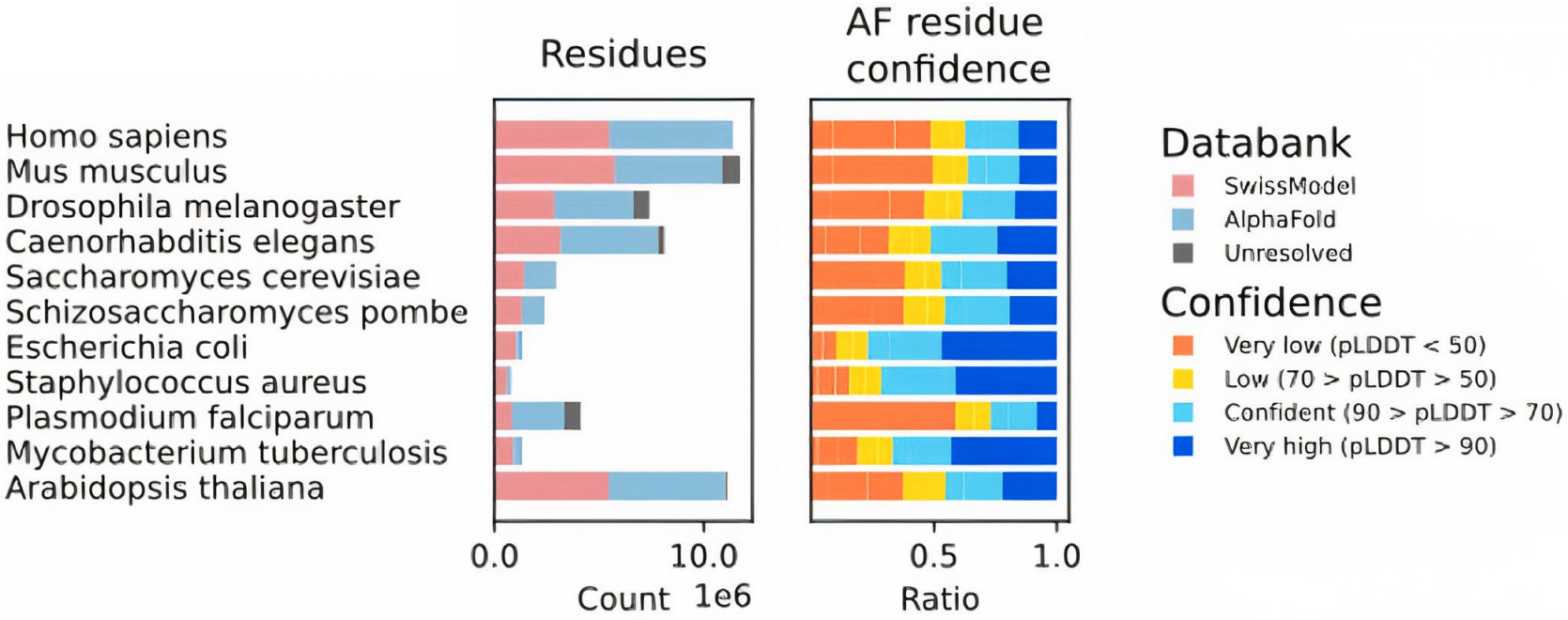
Kiefer, F. *Nucleic Acids Research* (2009)

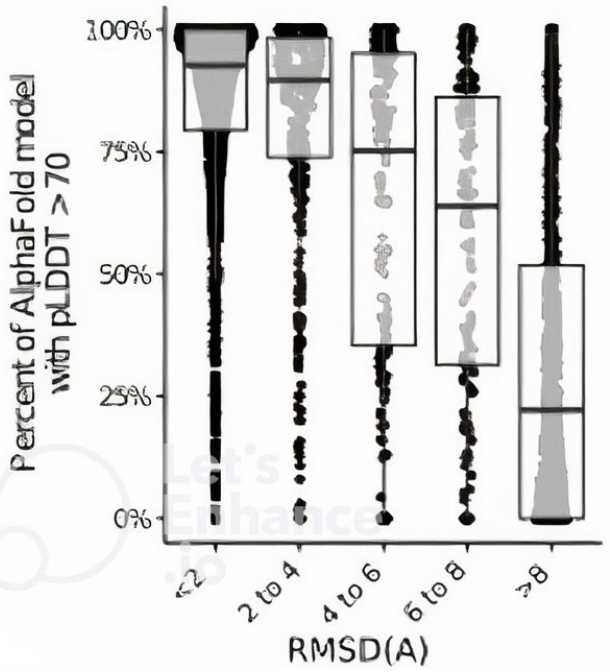
Yang, J. *PNAS* (2020)

Jumper, J. *Nature* (2021)

AlphaFold2 offered additional structures with an applicable confidence metric

- AF2 added 25% residues with novel and confident predictions compared to SMR.
- AF2 confidence score pLDDT correlated with RMSD value from trRosetta model.

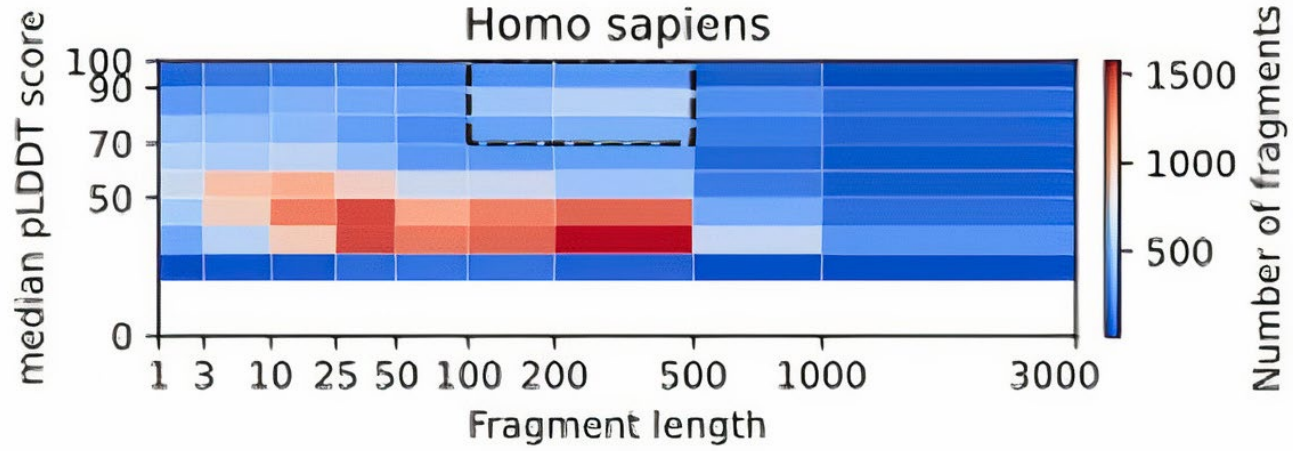




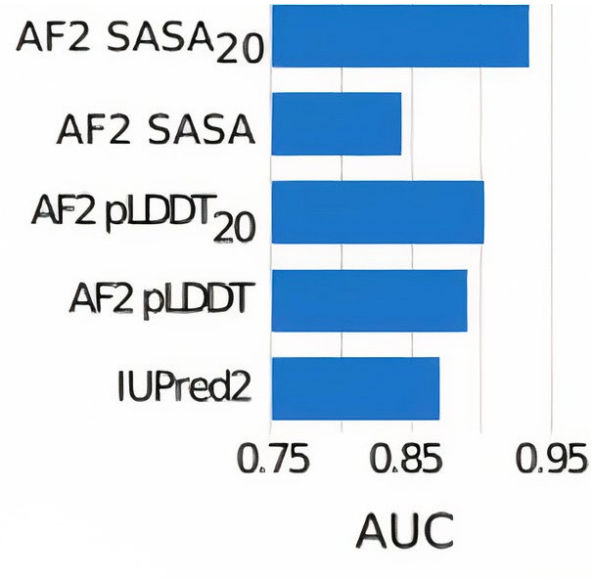
Remarks: novel = not in SMR; confident = pLDDT > 70

Akdel, M. *BioRxiv* (2021)

pLDDT stood as a predictor for novel protein fragments



Wheelan, S. J. *Bioinformatics* (2000)



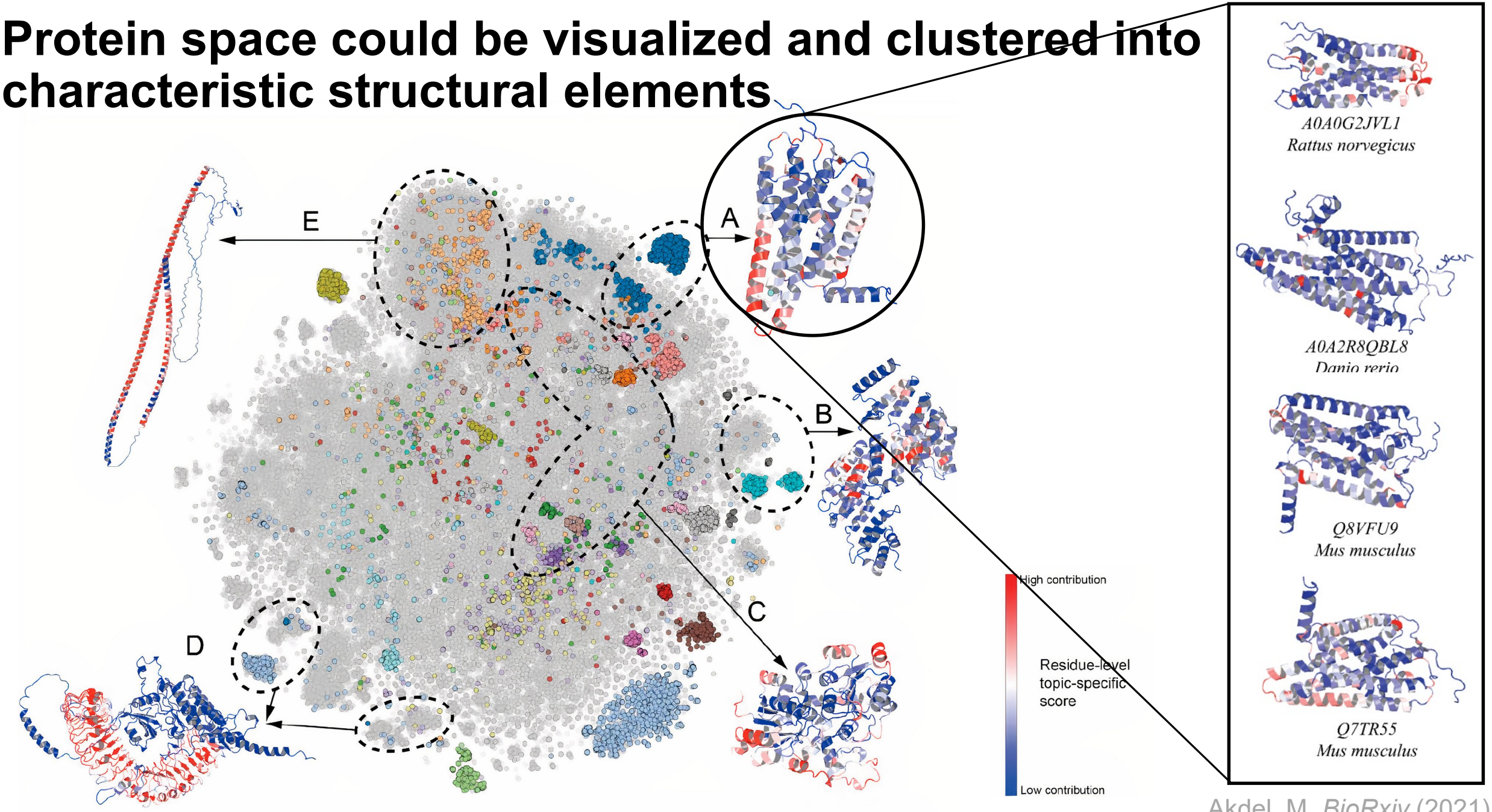
Mészáros, B. NAS (2018)

- Across 11 species, 18429 contiguous regions are “domain-like” with pLDDT > 70.
- Low confidence predictions are significantly enriched for IDRs.

Remarks: SASA = solvent accessible surface area; IUPred2 = a disorder prediction method

Akdel, M. *BioRxiv* (2021)

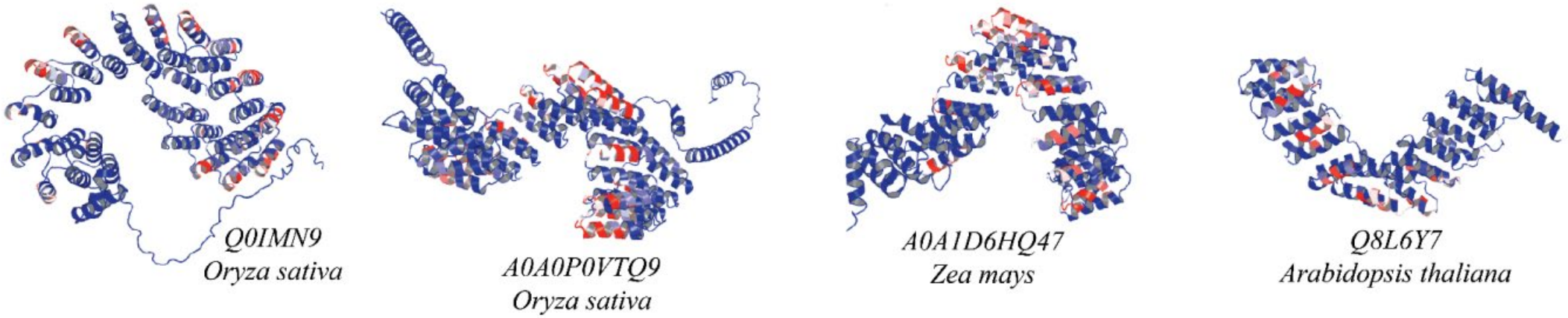
Protein space could be visualized and clustered into characteristic structural elements



Akdel, M. *BioRxiv* (2021)

Protein space could be visualized and clustered into characteristic structural elements

- Clusters exclusively composed of AF2-derived structures

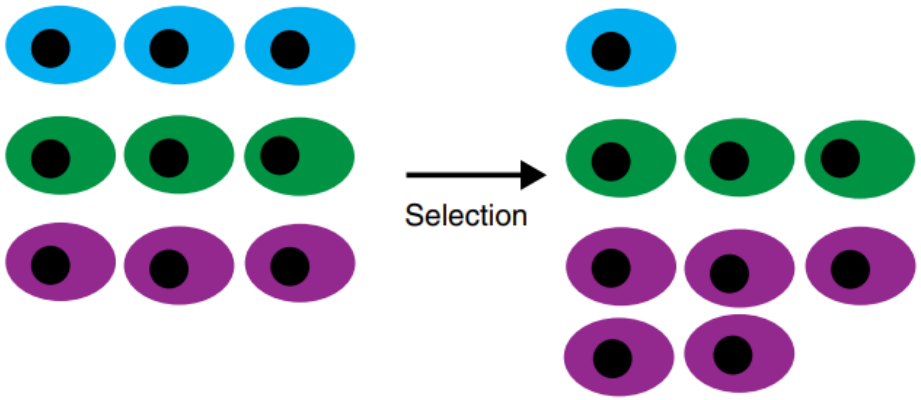


- Clusters exclusively composed of PDB proteins
  - Limited number of species and proteins covered by AF2 database.
  - Structure under intense studies by the academia/industries (i.e, antibodies)

AF2 database indicates rarely studied fields as well as topics of high interest.

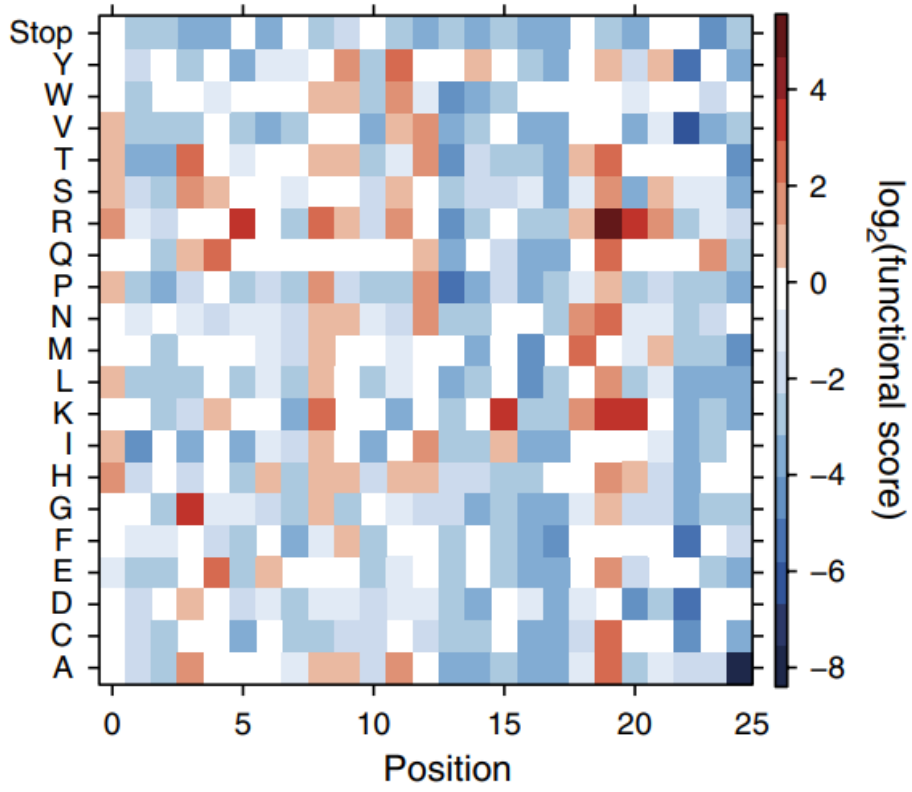
Akdel, M. *BioRxiv* (2021)

Deep mutational scanning revealed phenotypic consequences of genetic variation but lacked structural clues



Variant	Mutation	Counts (input)	Counts (selected)	Functional score
Blue	A60P	3	1	0.33
Green	WT	3	3	1
Purple	S36T	3	5	1.67

Enrichment ratio (ER)



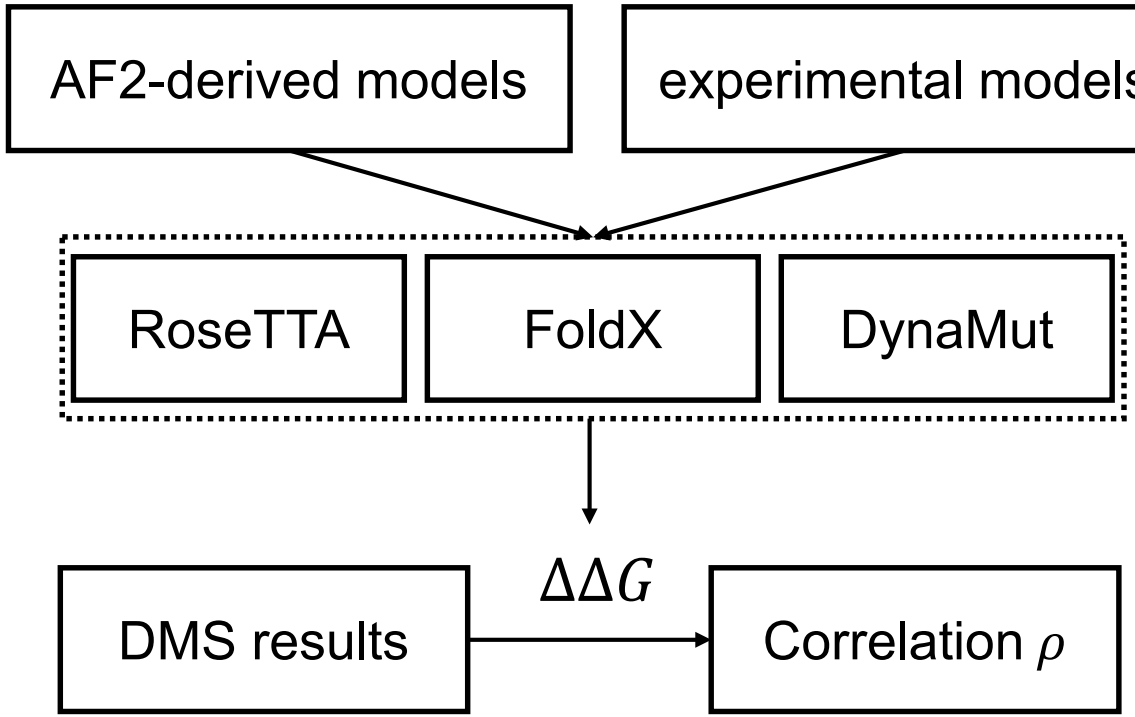
Rosetta Commons

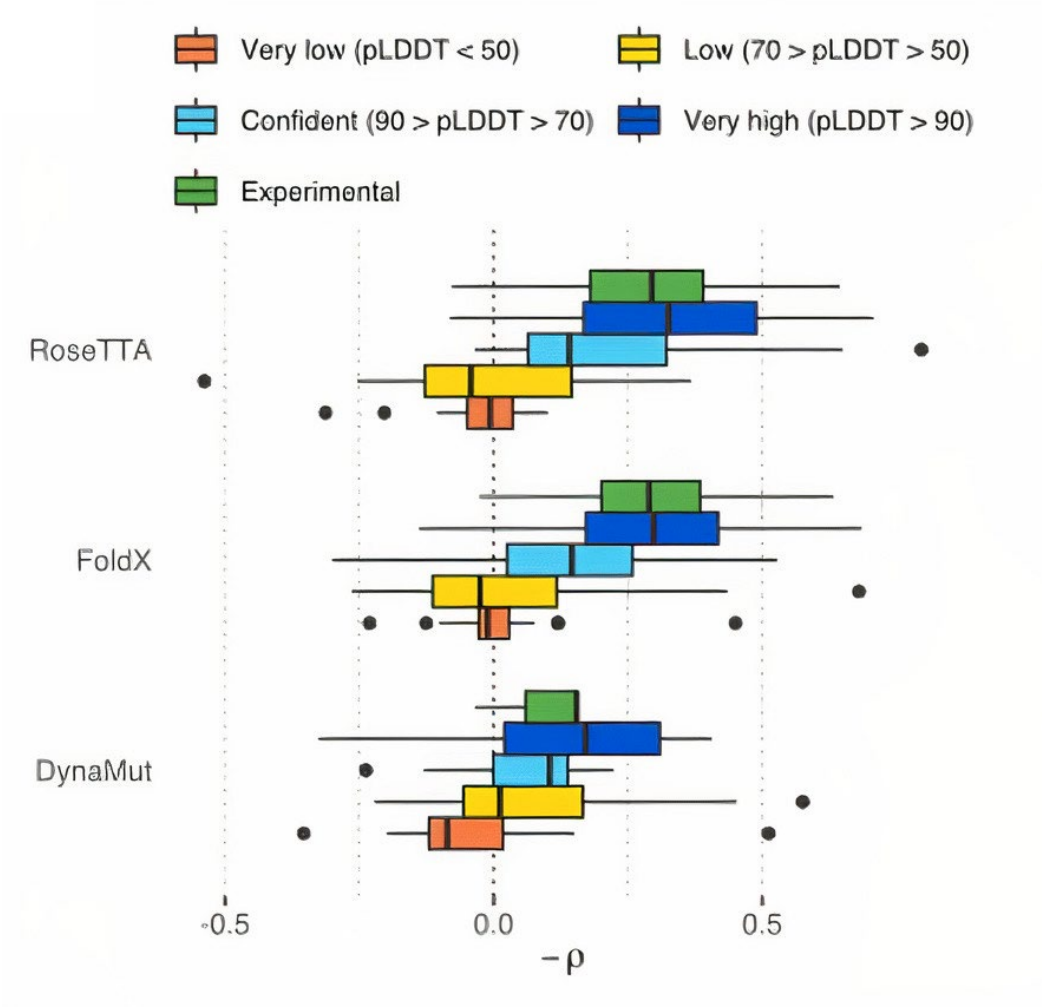
FoldX

DynaMut

Fowler, D. M. *Nature Methods* (2014)

AF2-derived structures could be applied in structural hypotheses about the impact of mutations

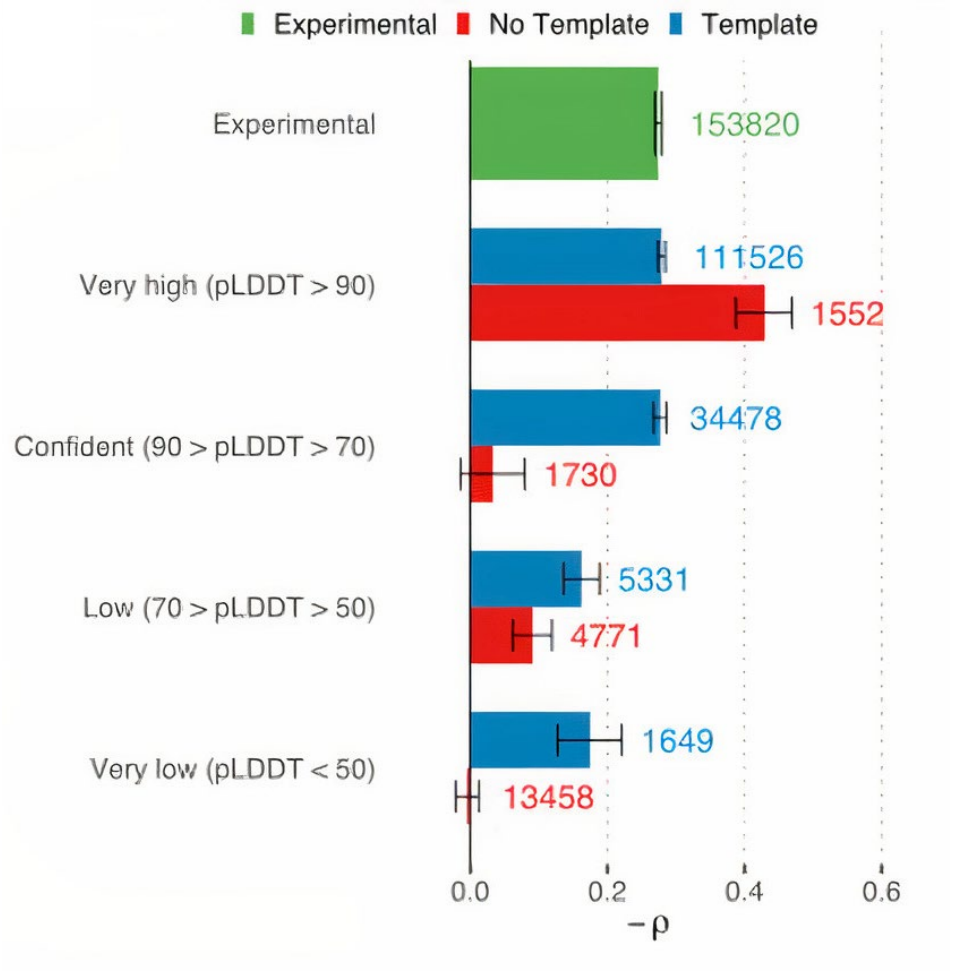


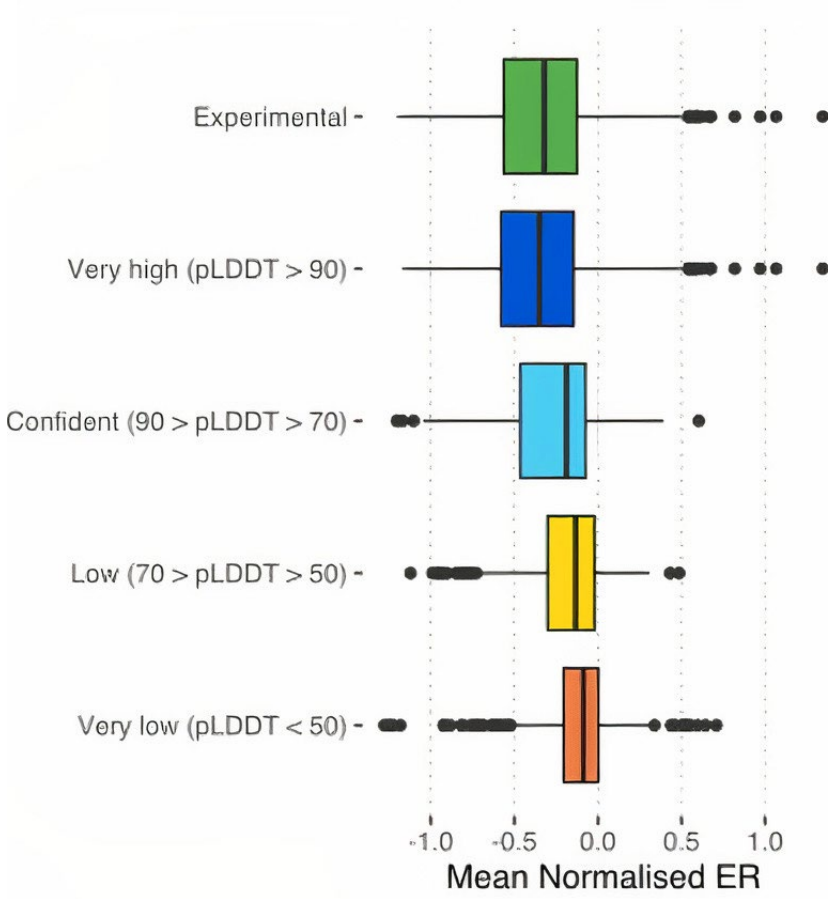


Remarks: DMS = deep mutational scanning

Akdel, M. *BioRxiv* (2021)

High-confidence and low-confidence structures indicate different tolerance to mutations



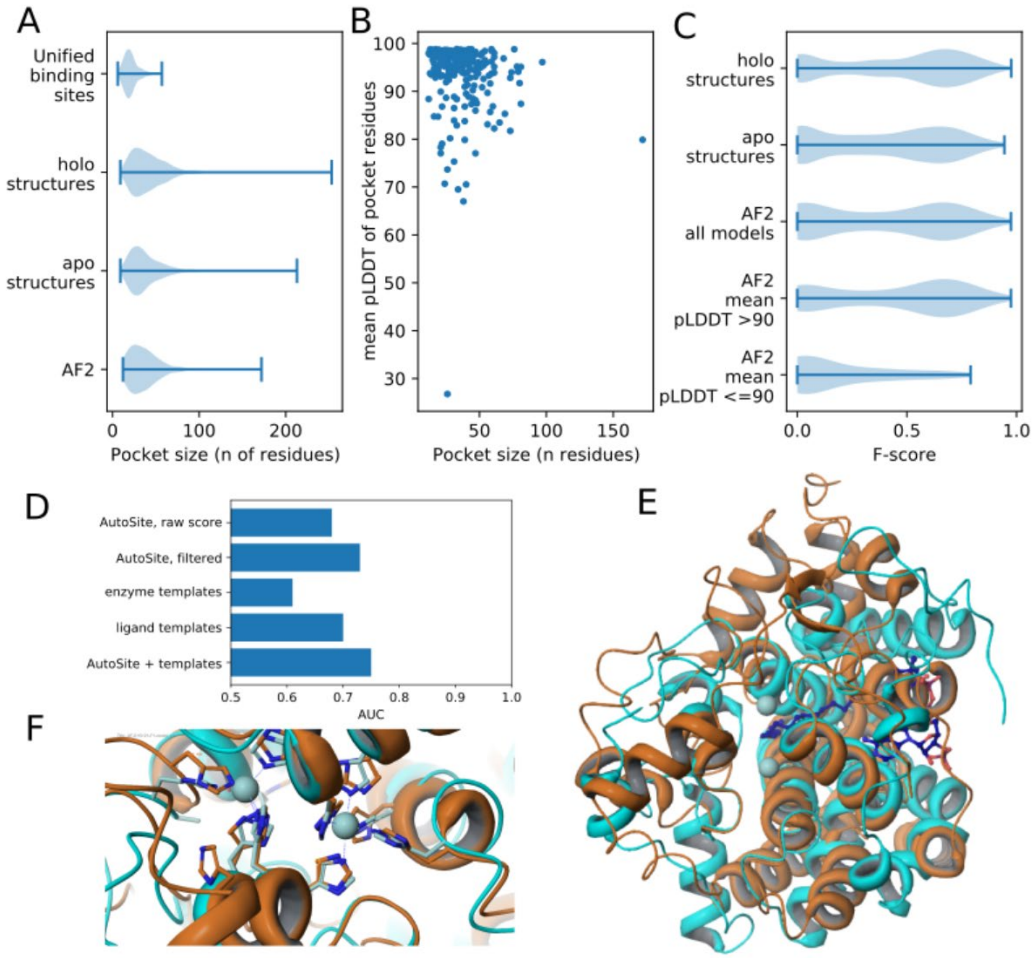


Remarks: DMS = deep mutational scanning

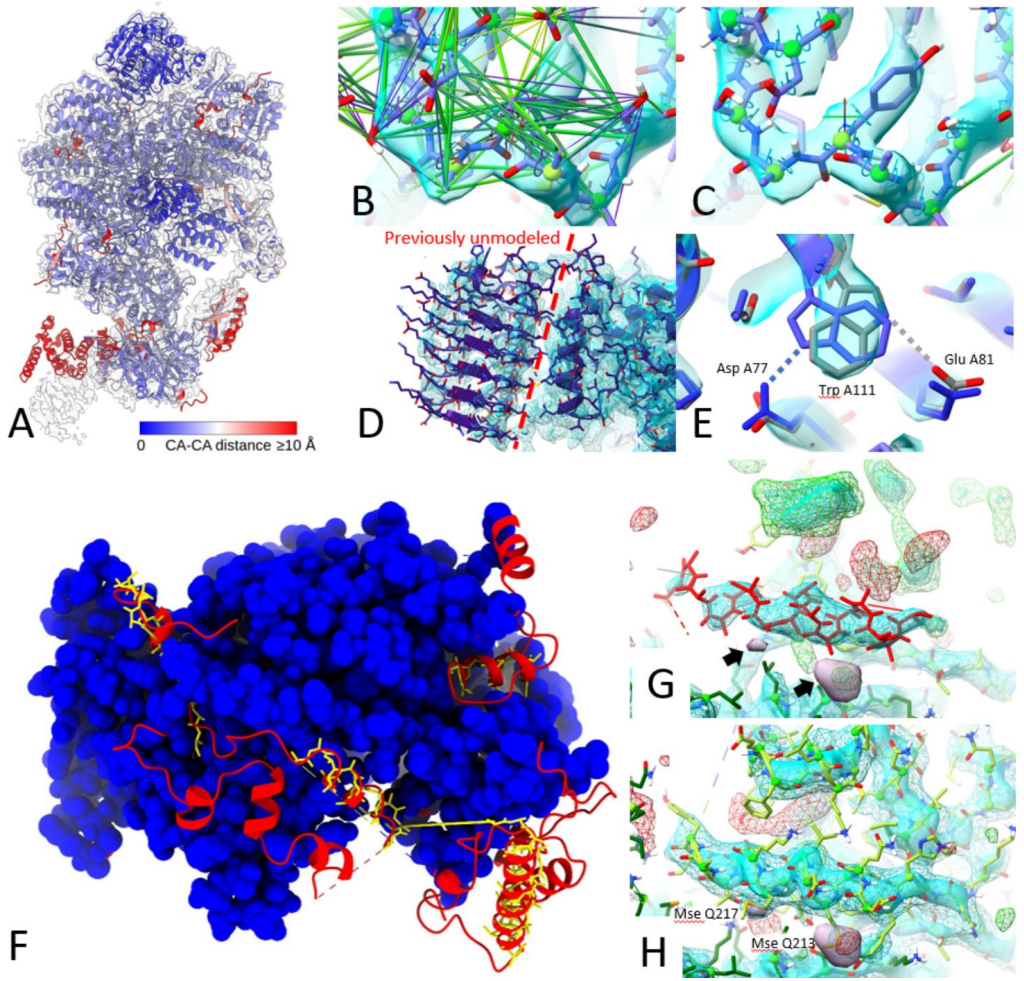
Akdel, M. *BioRxiv* (2021)

Other aspects worthy of paying attention…

Pocket detection and function prediction



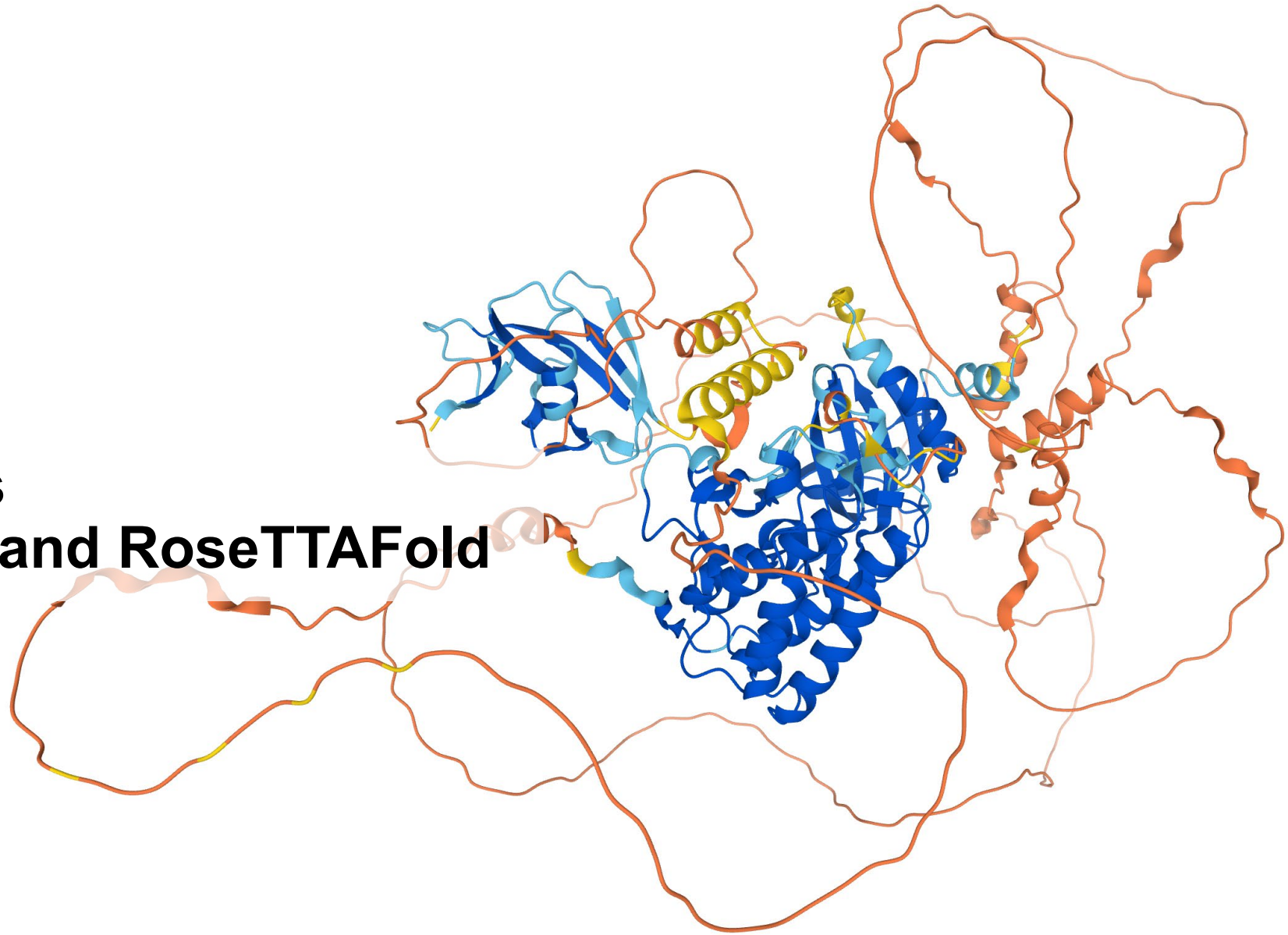
Modelling into cryo-EM/crystallographic data



Akdel, M. *BioRxiv* (2021)

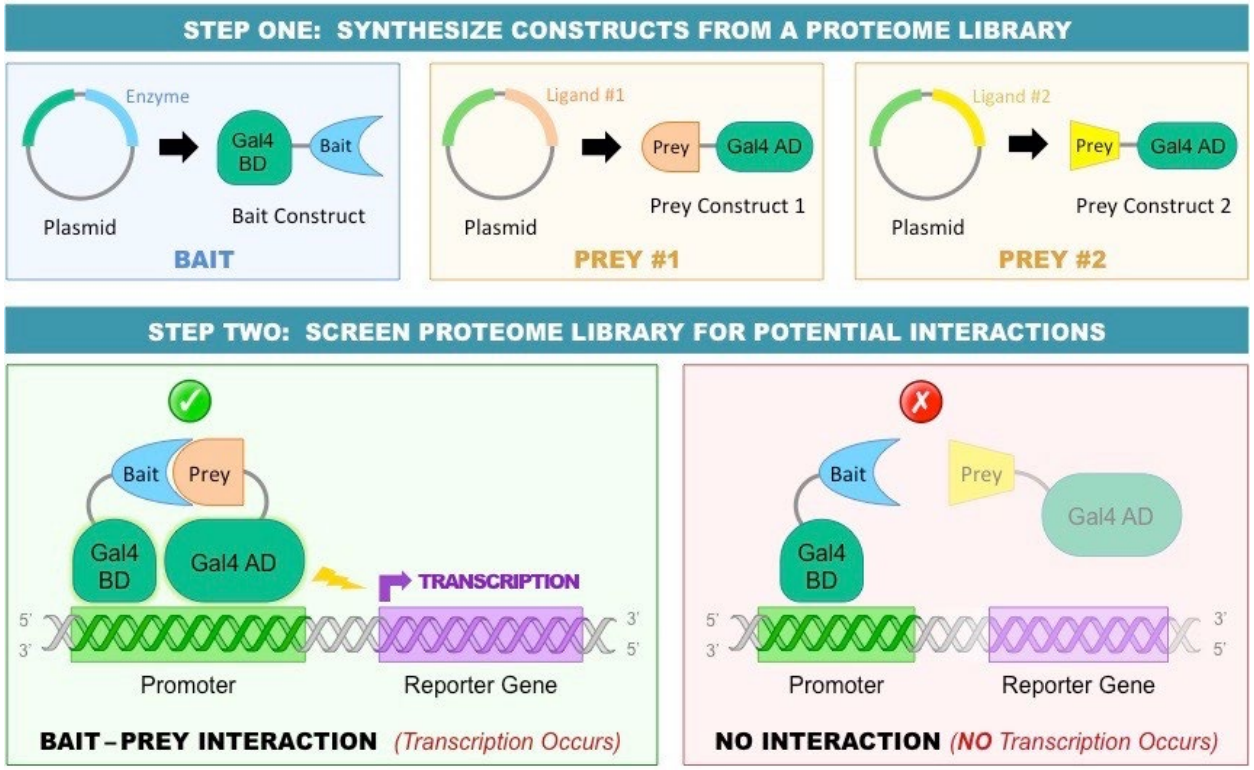
Discussion

Emerging Researches Involving AlphaFold2 and RoseTTAFold

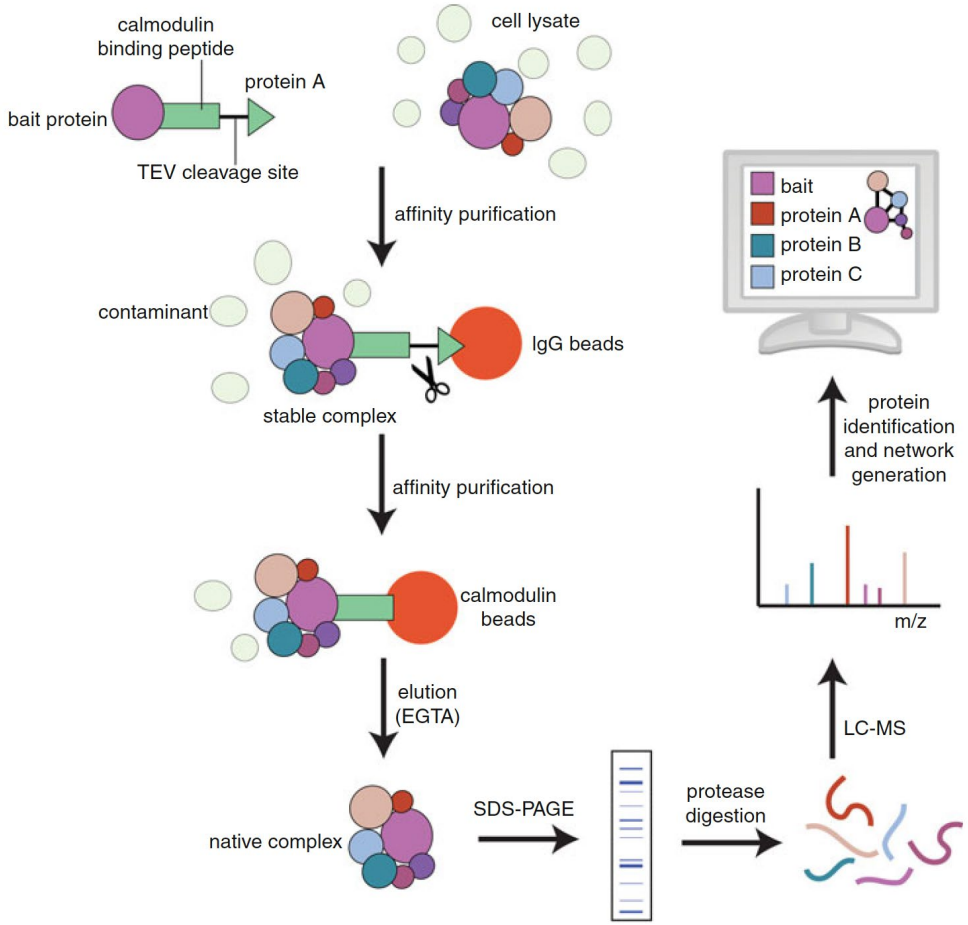


Experimental methods inspecting protein-protein interaction (PPI) lose high-resolution structure information

Yeast two-hybrid



Affinity purification mass spectrometry



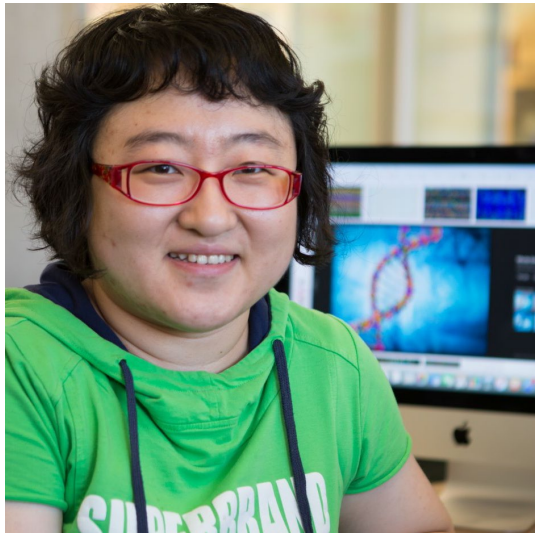
Building accurate models of core eukaryotic protein complexes with combination of RoseTTAFold and AlphaFold2

RESEARCH ARTICLE

STRUCTURE PREDICTION

Computed structures of core eukaryotic protein complexes

Ian R. Humphreys[1,2]†, Jimin Pei[3,4]†, Minkyung Baek[1,2]†, Aditya Krishnakumar[1,2]†, Ivan Anishchenko[1,2], Sergey Ovchinnikov[5,6], Jing Zhang[3,4], Travis J. Ness[7]‡, Sudeep Banjade[8], Saket R. Bagde[8], Viktoriya G. Stancheva[9], Xiao-Han Li[9], Kaixian Liu[10], Zhi Zheng[10,11], Daniel J. Barrero[12], Upasana Roy[13], Jochen Kuper[14], Israel S. Fernández[15], Barnabas Szakal[16], Dana Branzei[16,17], Josep Rizo[4,18,19], Caroline Kisker[14], Eric C. Greene[13], Sue Biggins[12], Scott Keeney[10,11,20], Elizabeth A. Miller[9], J. Christopher Fromme[8], Tamara L. Hendrickson[7], Qian Cong[3,4]*§, David Baker[1,2,21]*§

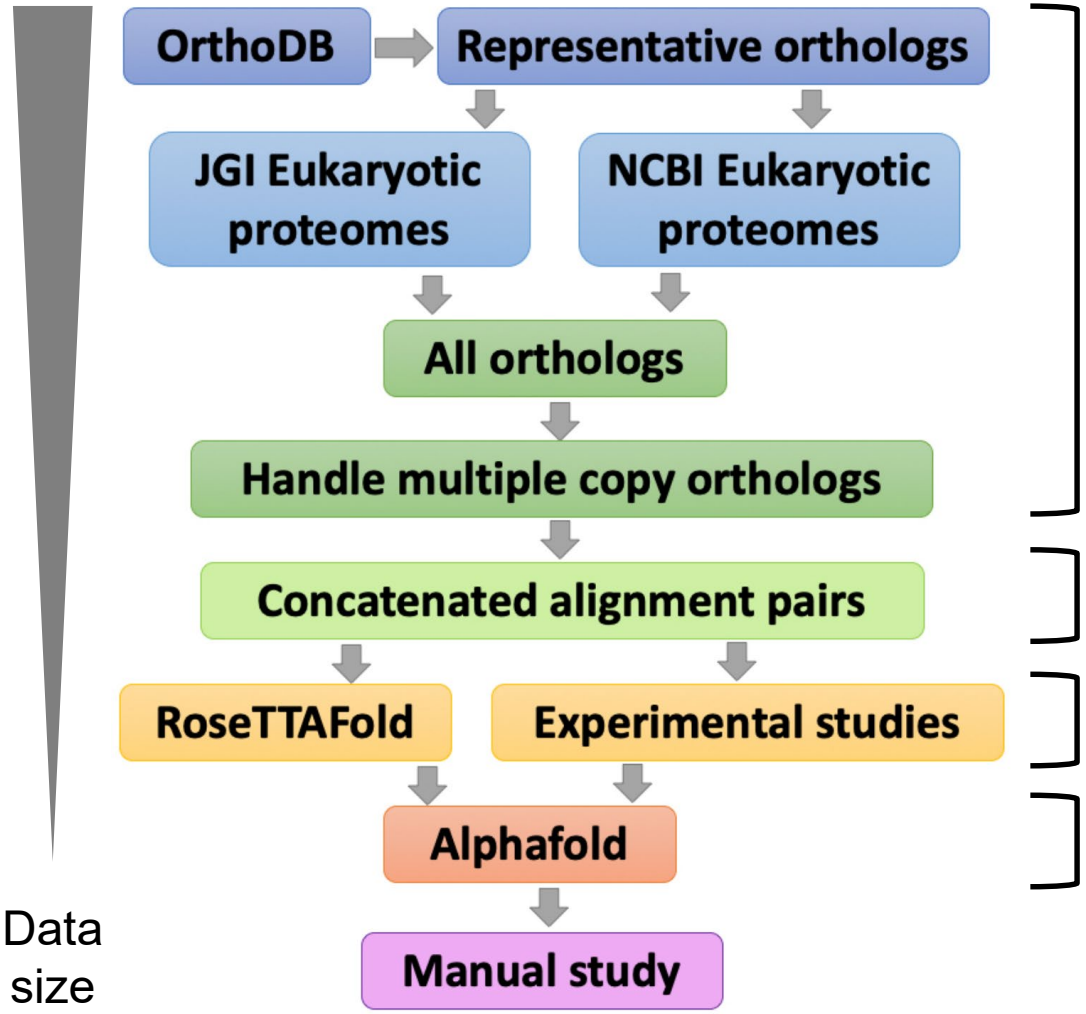


Qian Cong

2020- UT Southwestern
2017-2020 UWashington

Key idea: residues in interprotein contacts coevolve!

PPI screen using RoseTTAFold + AlphaFold2 with paired multiple sequence alignments (pMSAs)



Selected 4090 yeast proteins and their orthologs

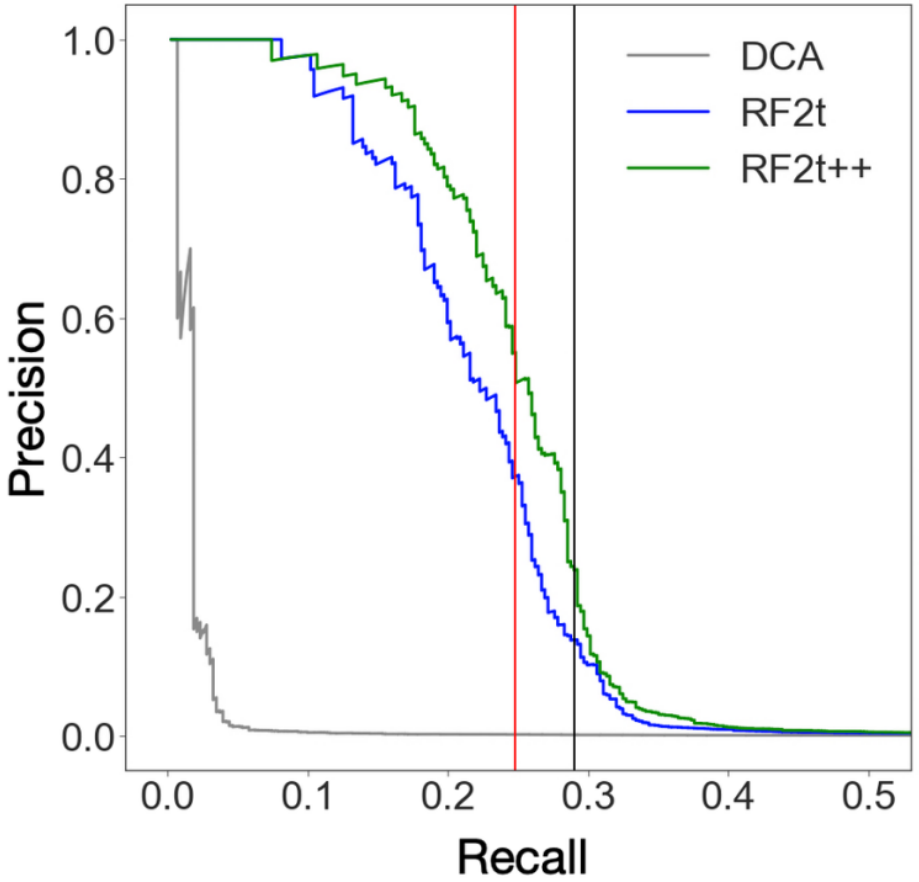
Built 4286433 paired alignments

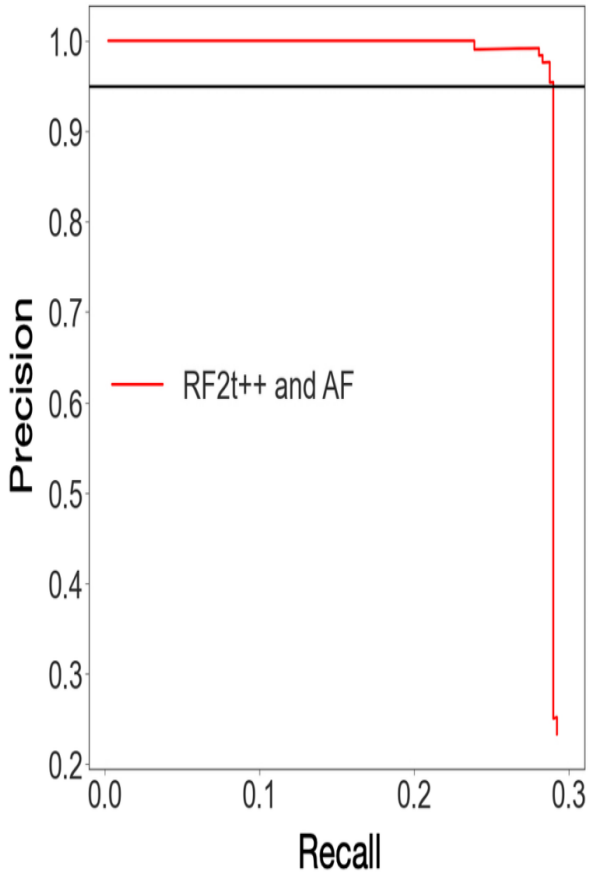
Got 5495 PPIs with RoseTTA or skip

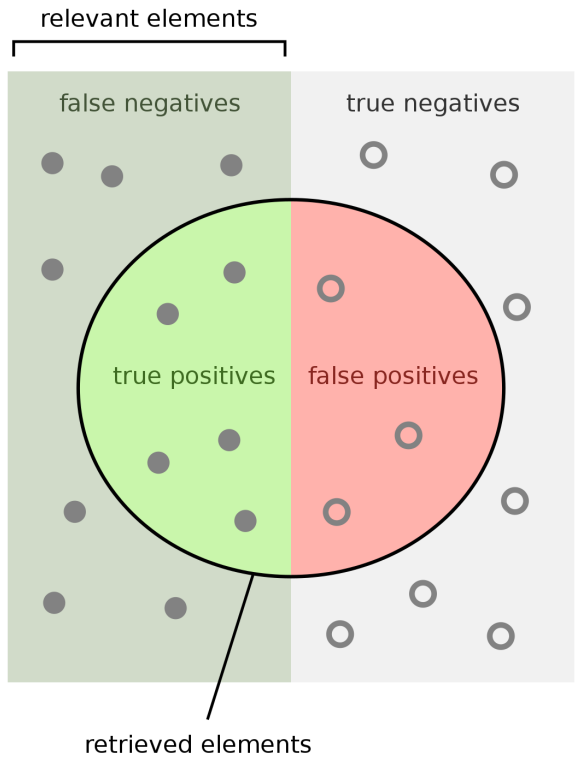
Got 715 PPIs with modified AlphaFold

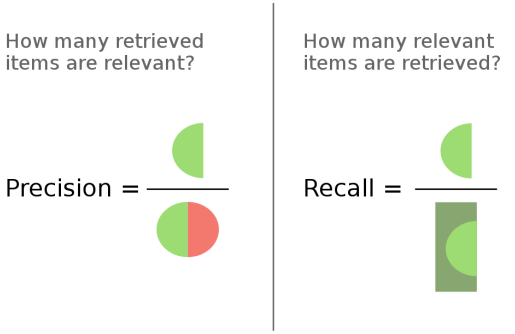
Humphreys, I. R. *Science* (2021)

715 candidate PPIs were selected by *de novo* RF → AF pipeline







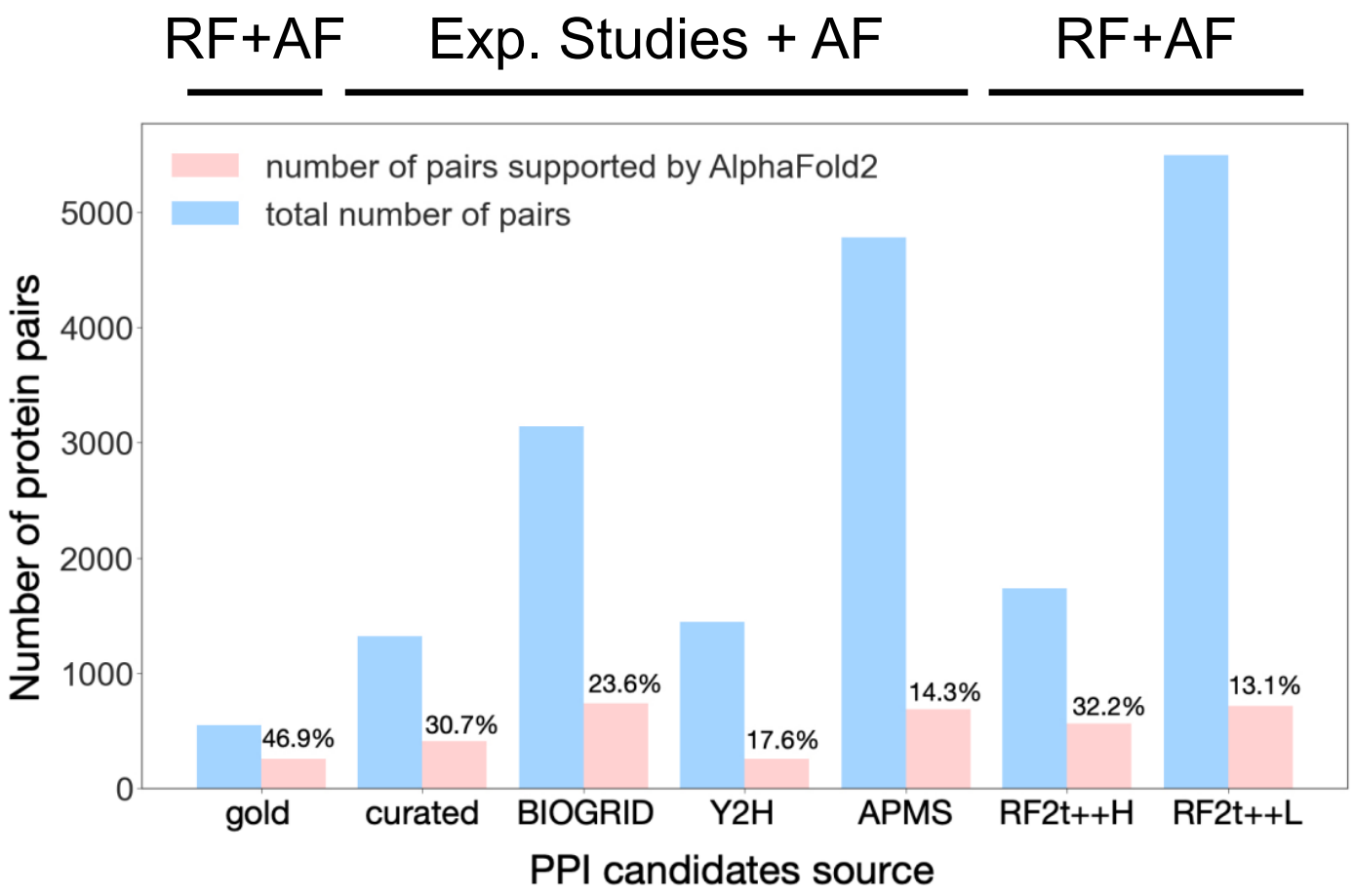


Remarks: DCA = direct coupling analysis

Humphreys, I. R. *Science* (2021)

https://en.wikipedia.org/wiki/Precision_and_recall

De novo PPI screen procedure identified much fewer PPIs than experimental methods



Gold = gold standard (ground truth)

Curated = literature dataset

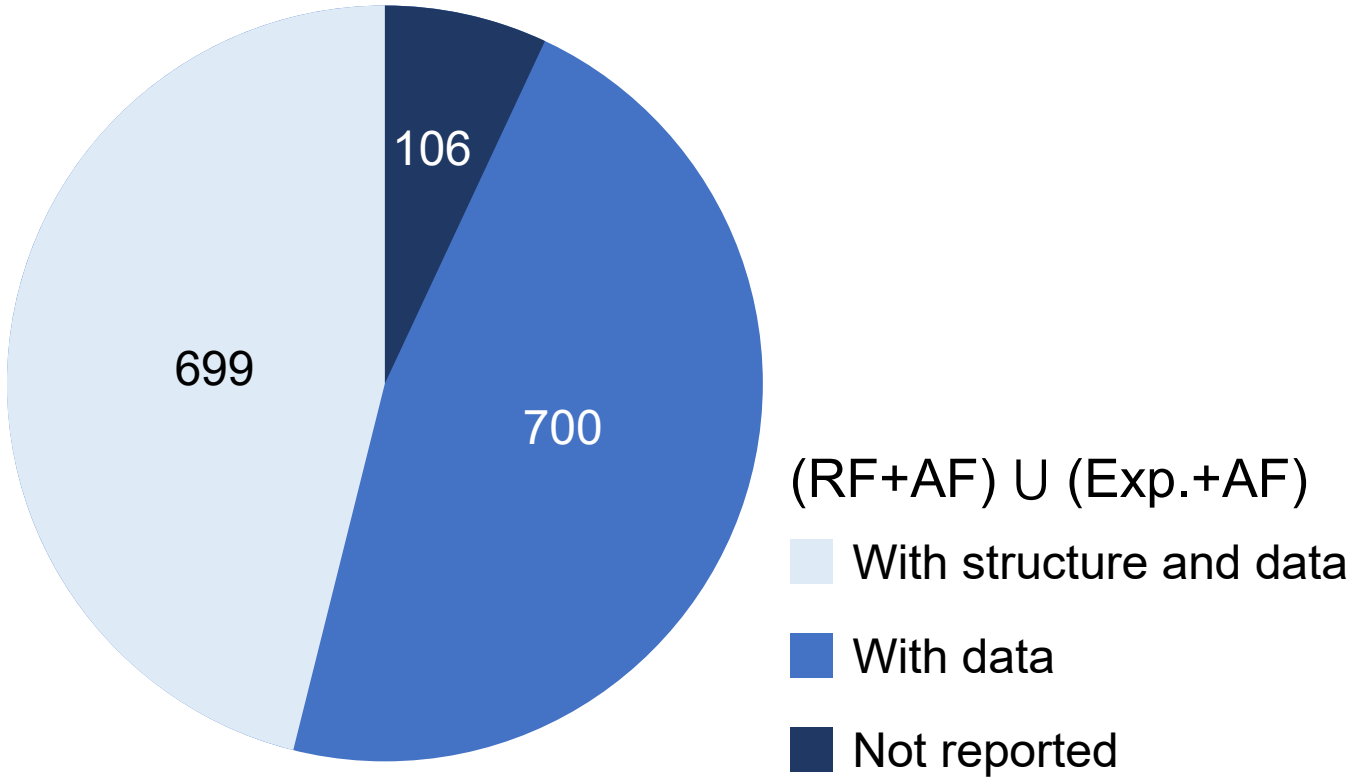
BIOGRID = curated PPI database

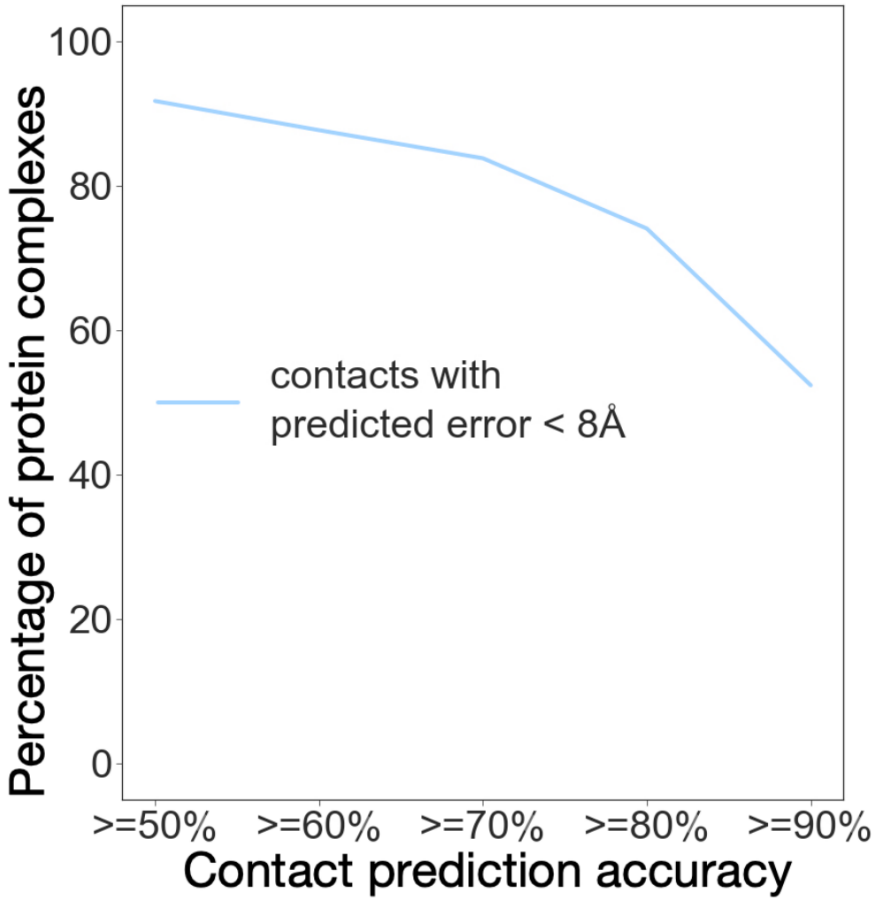
- Higher ratio = more true positive
- Lower ratio = more false positive

AF helps filtering out false positives

Humphreys, I. R. *Science* (2021)

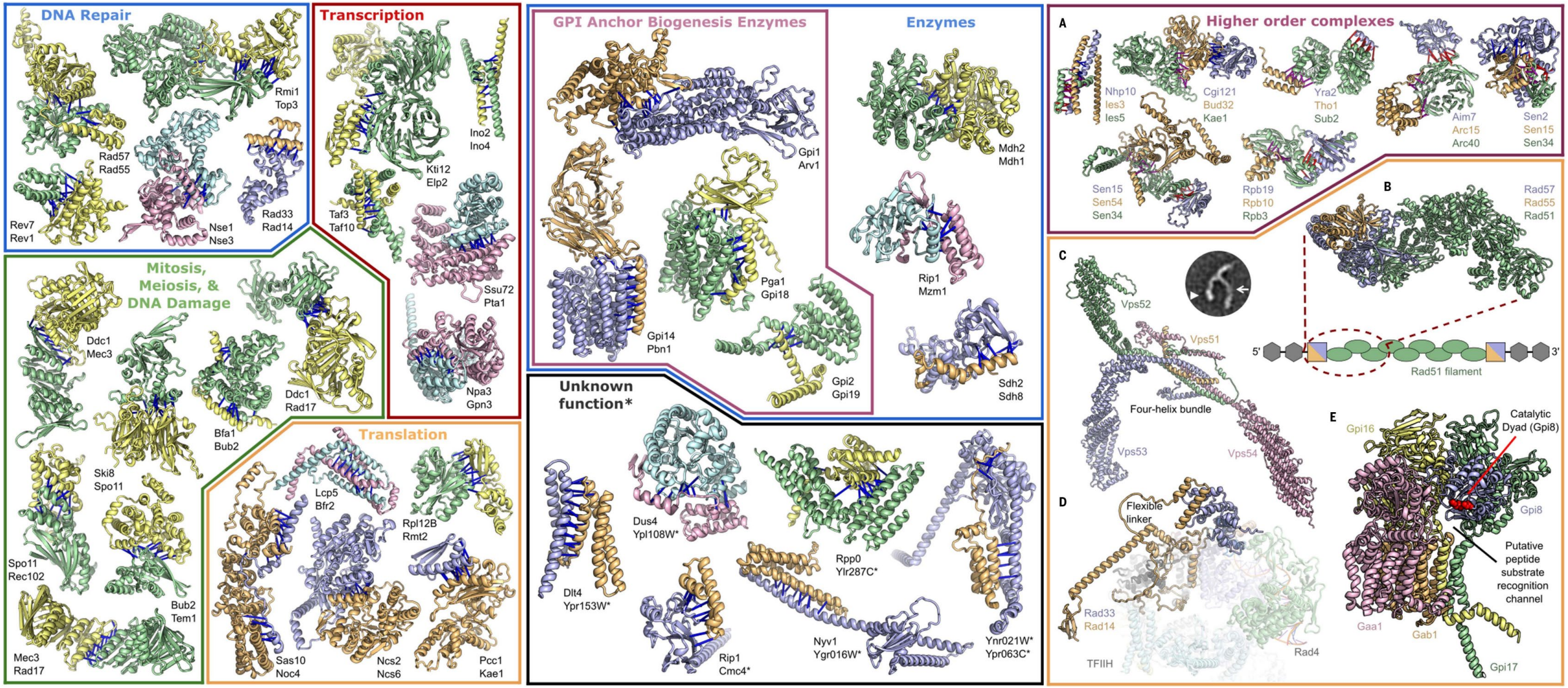
AF predicted interprotein contacts with high accuracy





Humphreys, I. R. *Science* (2021)

The protein-protein interaction gallery



Humphreys, I. R. *Science* (2021)

Limitation of the *de novo* RF → AF pipeline

General limitations

- Available pMSAs are limited for specific organism.
- PPIs with stronger coevolutionary signals are easier to be identified.
- PPIs with stronger interactions between ordered elements are easier to be found.

Specific limitations

- Single hydrophobic/amphipathic helices interactions may be overpredicted.
- High-order obligate protein complexes may be quite inaccurate.

New researches on the way…

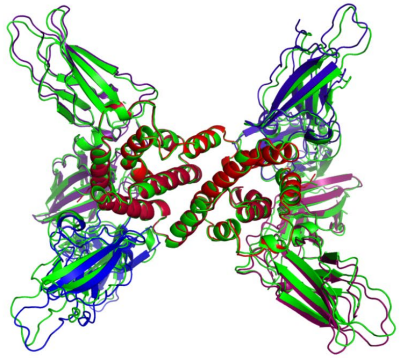
New Results

Follow this preprint

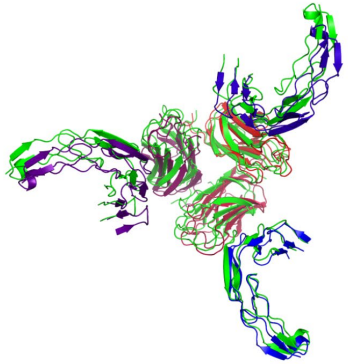
Protein complex prediction with AlphaFold-Multimer

Richard Evans, Michael O'Neill, Alexander Pritzel, Natasha Antropova, Andrew Senior, Tim Green, Augustin Žídek, Russ Bates, Sam Blackwell, Jason Yim, Olaf Ronneberger, Sebastian Bodenstein, Michal Zielinski, Alex Bridgland, Anna Potapenko, Andrew Cowie, Kathryn Tunyasuvunakool, Rishub Jain, Ellen Clancy, Pushmeet Kohli, John Jumper, Demis Hassabis

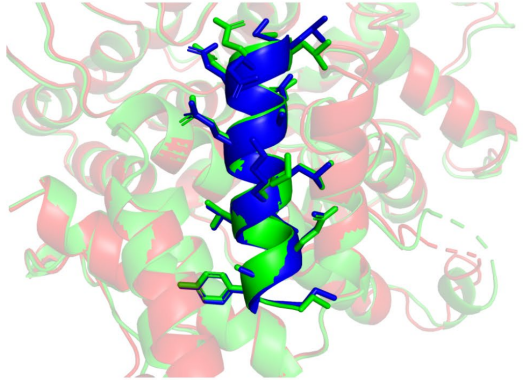
doi: https://doi.org/10.1101/2021.10.04.463034



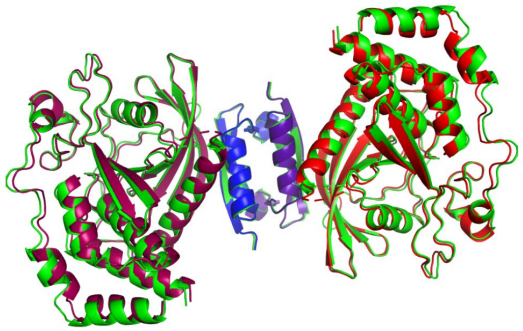
(a) A2B2C2 heteromer
TM-score = 97.4, N_{res} = 1,246, PDB ID = 6E3K



(b) A3B3 heteromer
TM-score = 85.4, N_{res} = 795, PDB ID = 7KHD



(c) Protein-peptide complex
TM-score = 96.6, DockQ = 0.954,
N_{res} = 385, PDB ID = 6JMT

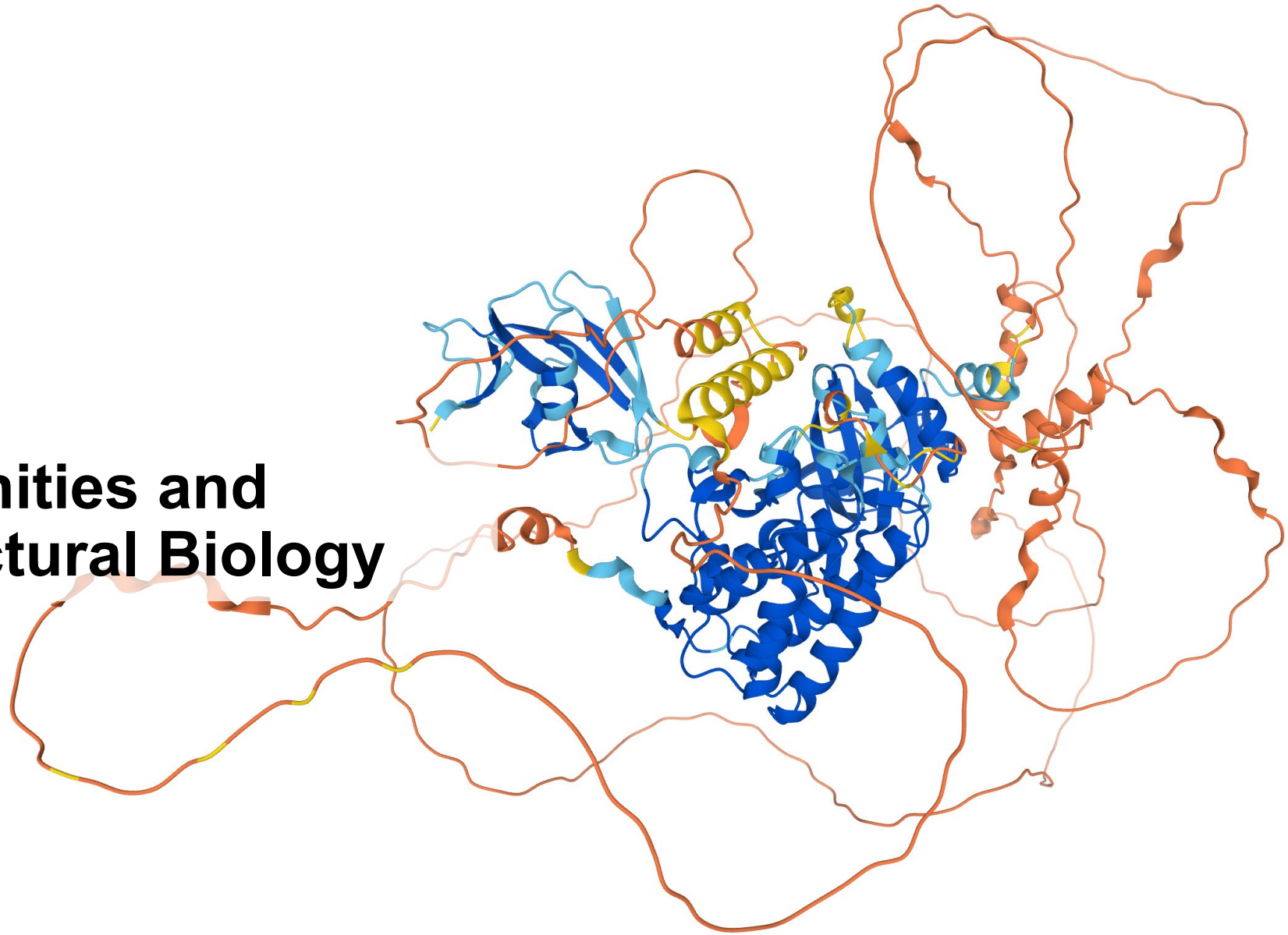


(d) A2B2 heteromer
TM-score = 98.5, N_{res} = 716, PDB ID = 6IWD

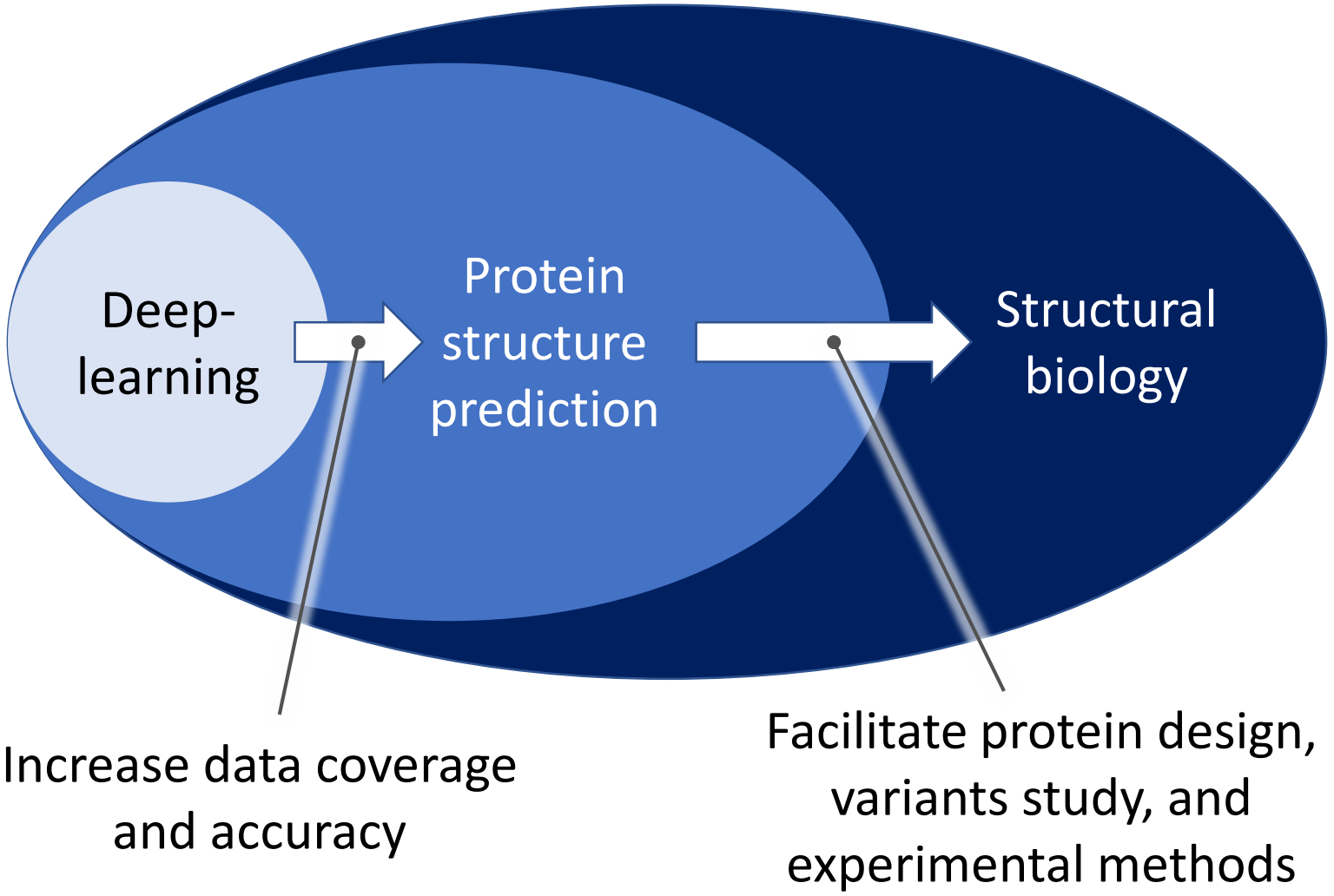
Evans, R. *BioRxiv* (2022)

Discussion

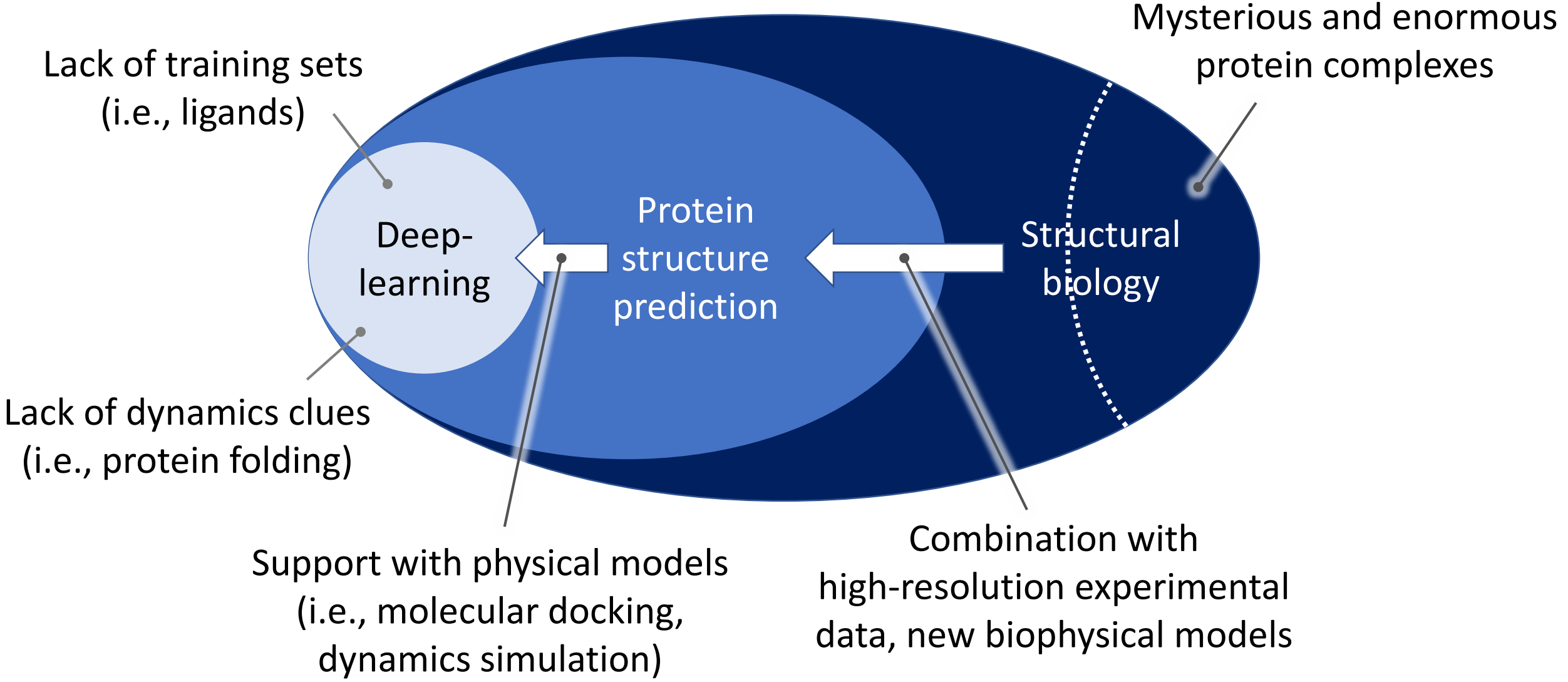
Remaining Opportunities and Challenges for Structural Biology



Deep-learning-based methods facilitate biomedical researches



Deep-learning also gains support from existing methods



Summary

- AlphaFold2 and RoseTTAFold are deep-learning-based methods that apply attention algorithms on MSA and paired distance matrices to iterate accurate protein structures.
- Both high- and low-confidence predicted structures have biological implications.
- Predicted models have potentials in studying mutational variants, enzymatic domains, ligand-biding sites, protein design, etc.

RaptorX

I-TASSER
Protein Structure & Function Predictions

DP Technology
深势科技

It is the prelude to solving protein mechanism and function.

Thank you for your attention
Questions are welcomed