LLM-powered deep learning models for protein-nucleic acid interactions

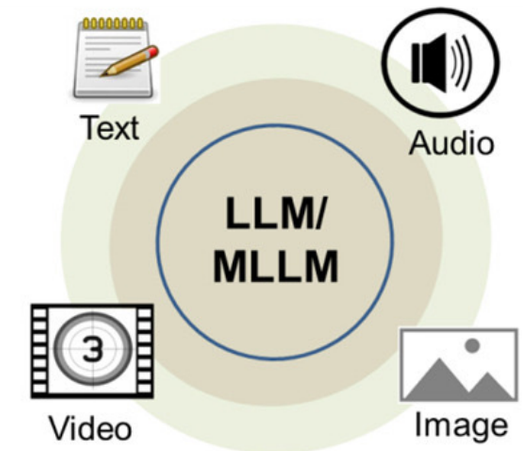
Debswapna Bhattacharya
Virginia Tech

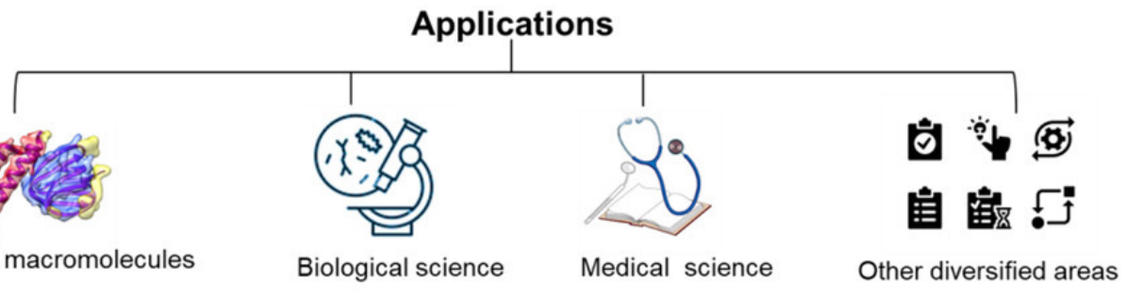
Workshop for AI-Powered Materials Discovery
at Great Plains

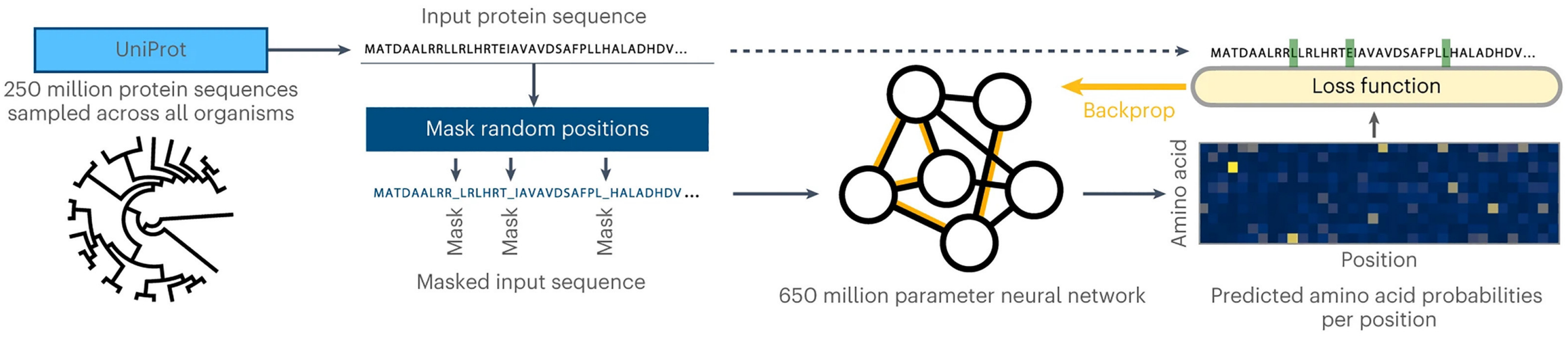
June 24, 2025

Large language models (LLMs) for Bio

ChatGPT

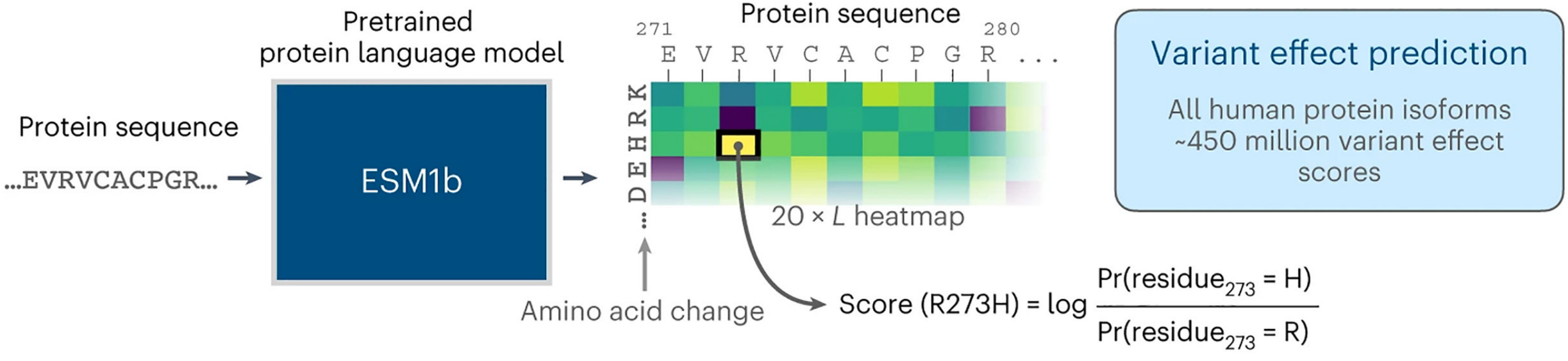


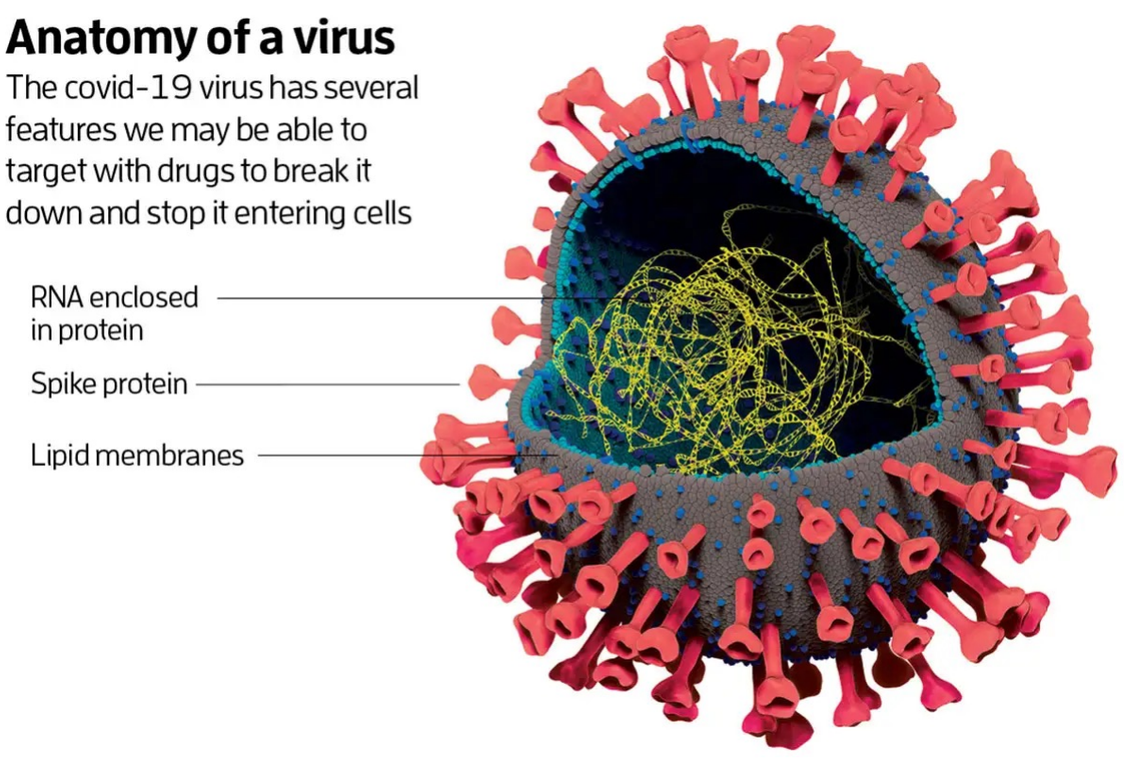




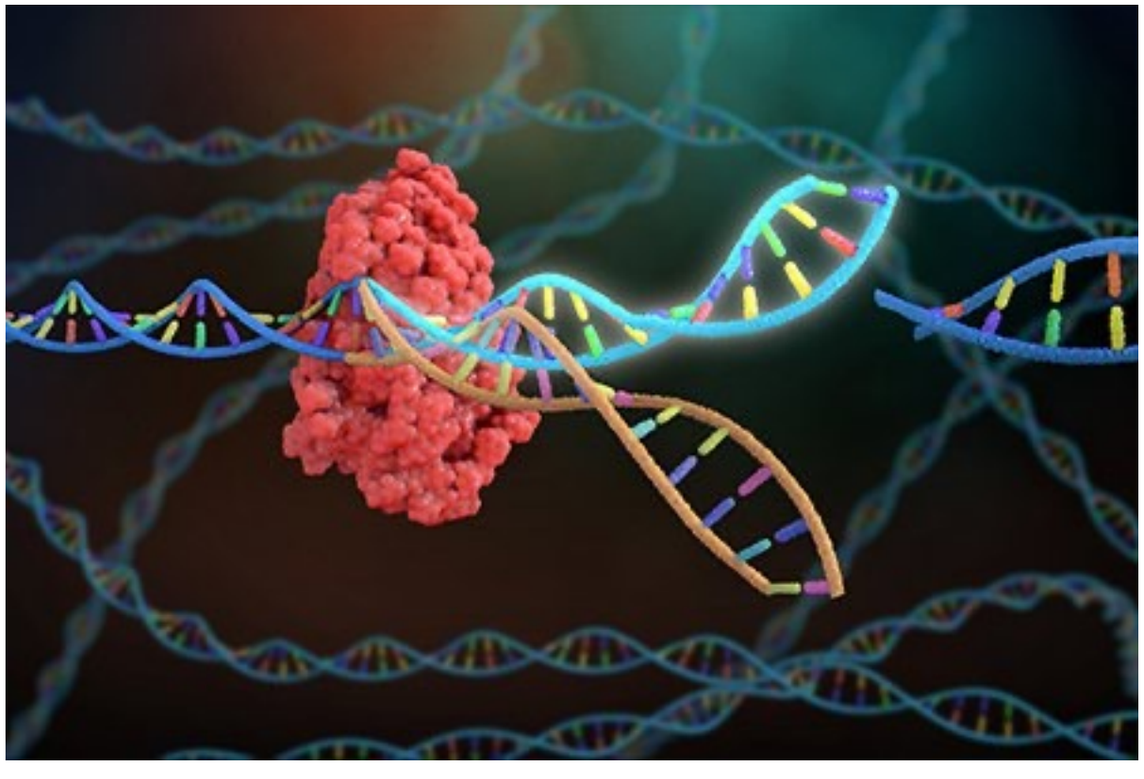
Lin, Z. et. al., 2023

Applications of LLMs for Bio



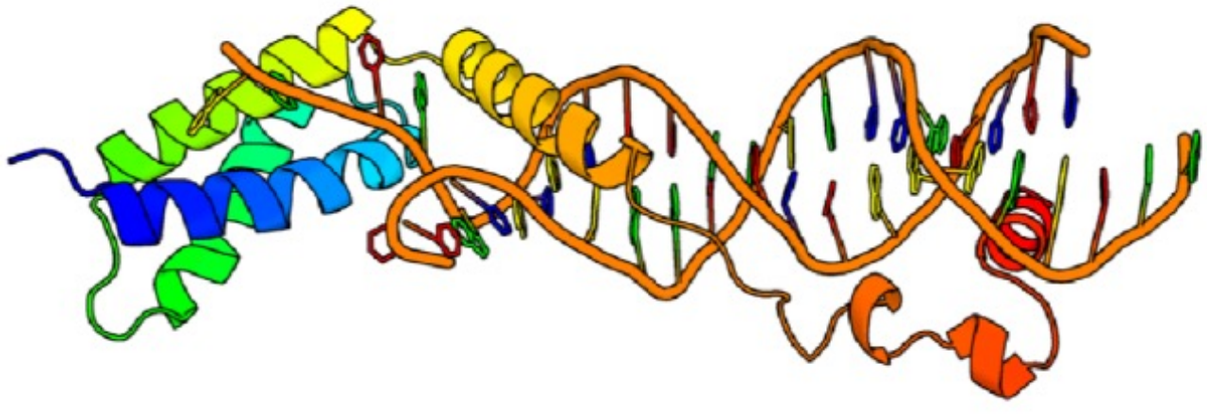


Viral pathogenesis



CRISPR/Cas9-Based Therapeutics

Protein-nucleic acid interactions



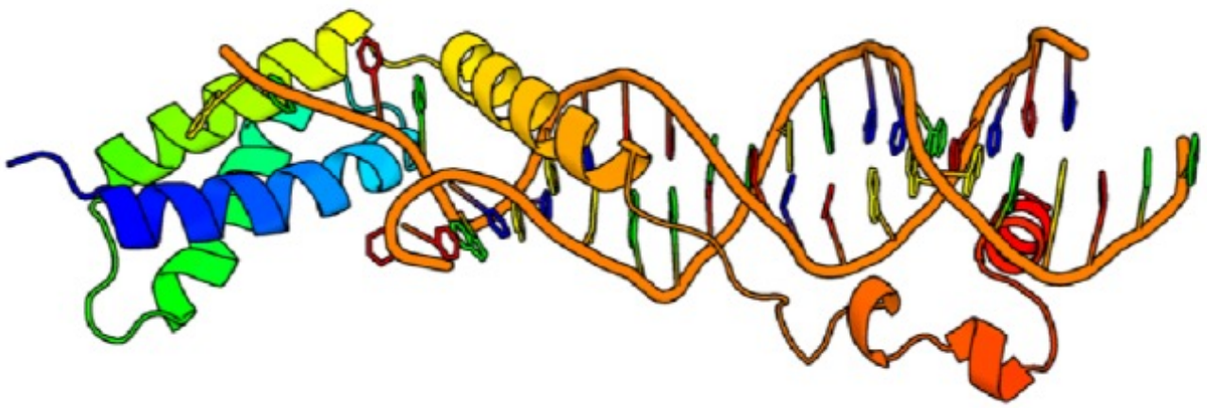
Protein-DNA interaction

Protein-RNA interaction

- Underpins a wide range of cellular processes: from gene replication to regulation to signal transduction to metabolism
- Reliable and accurate characterization of protein-nucleic acid 3D interactions in a large-scale screening manner is highly desirable

Protein-nucleic acid interactions

in atomic detail

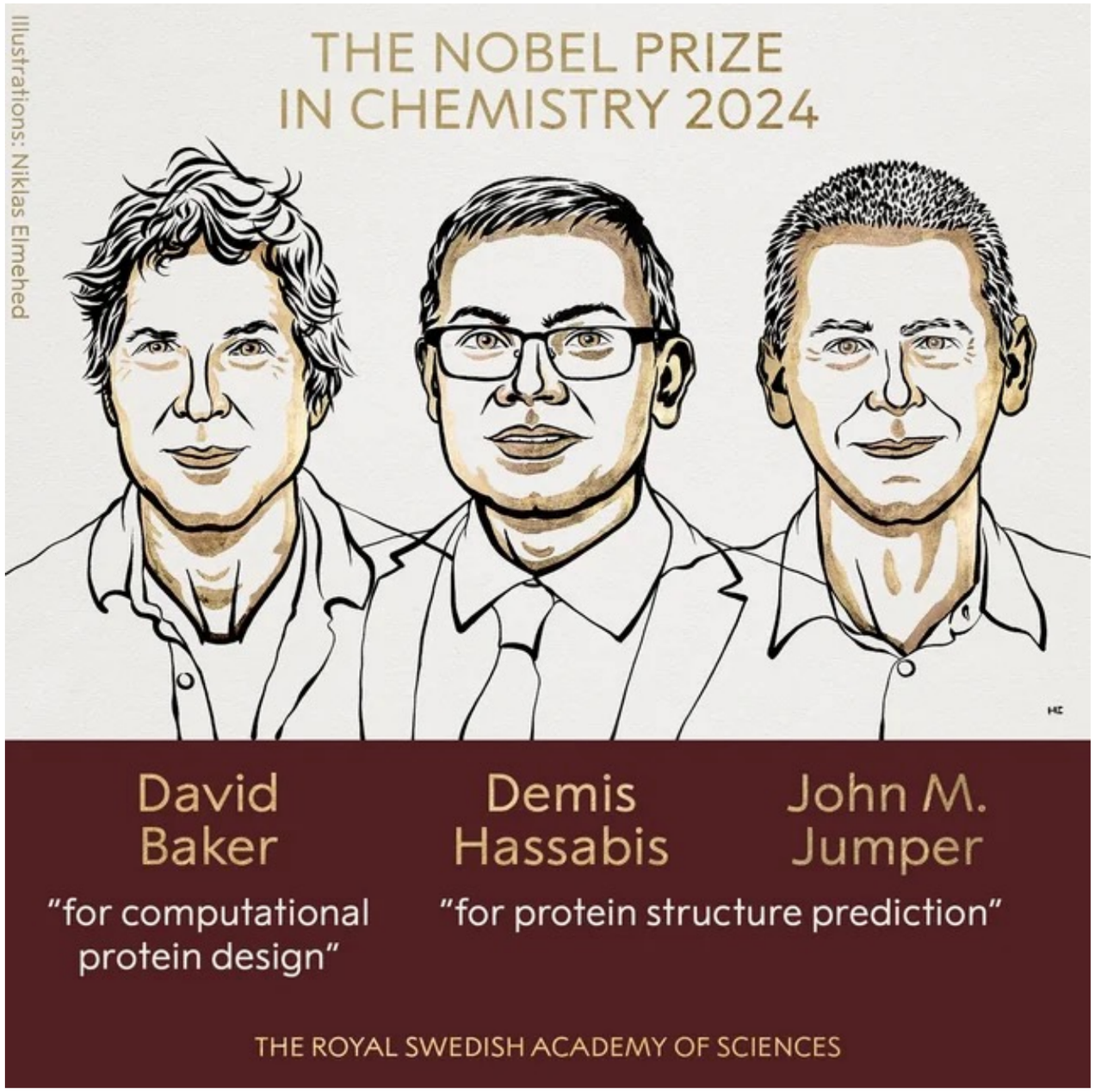


Protein-DNA interaction

Protein-RNA interaction

- Experimental structure determination is not always feasible or practical
- Can we use computational modeling to address this gap?

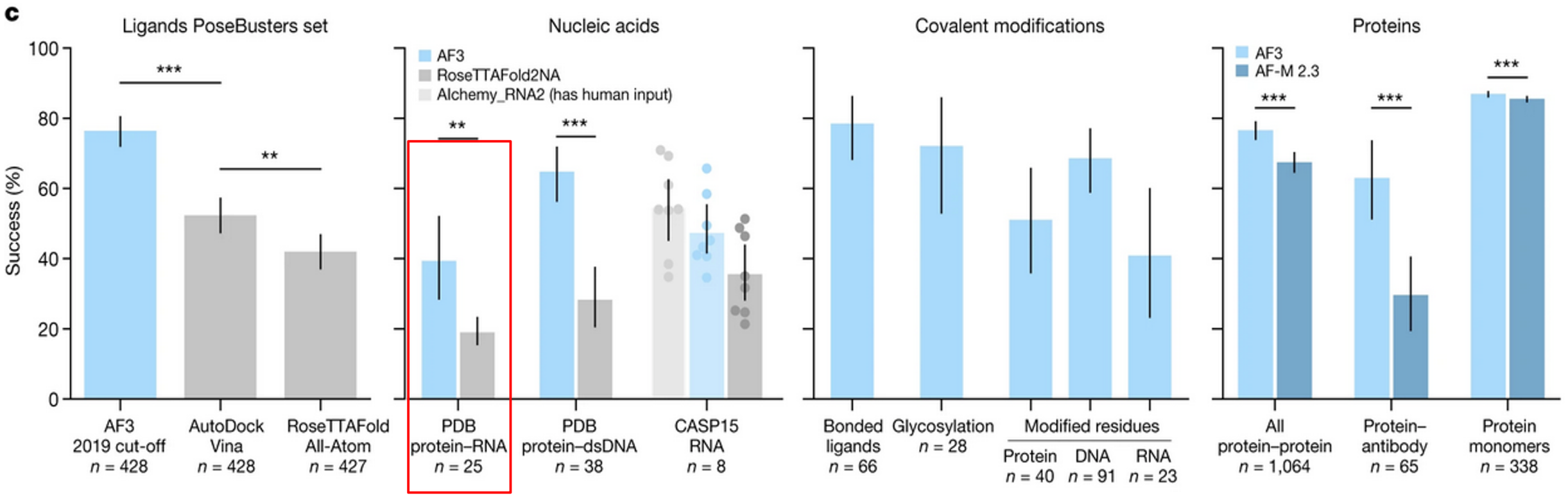
Harness AI to predict structure from sequence using AlphaFold



Can AlphaFold address this gap?

www.nobelprize.org

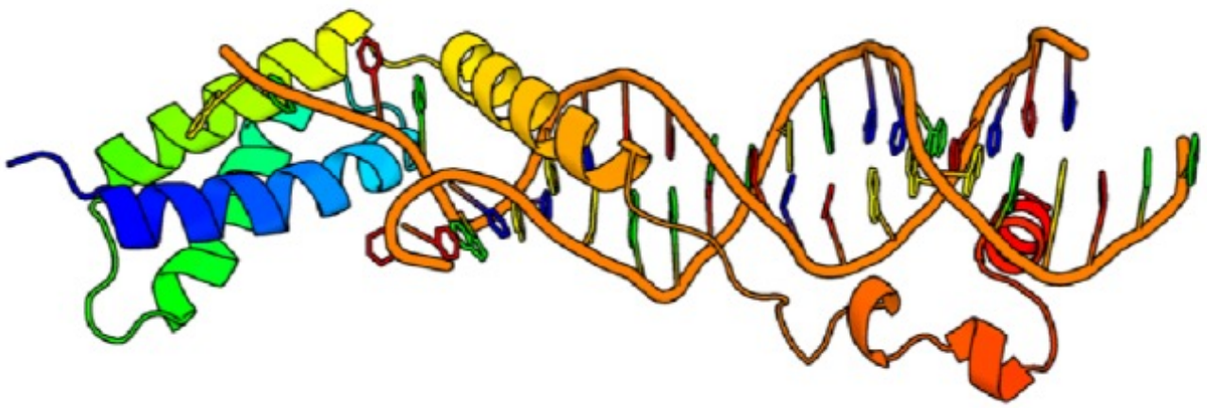
Protein-RNA complex structure prediction is not highly accurate even with AlphaFold3



Computational methods for predicting the structures are not highly accurate

Abramson, J. et. al., 2024

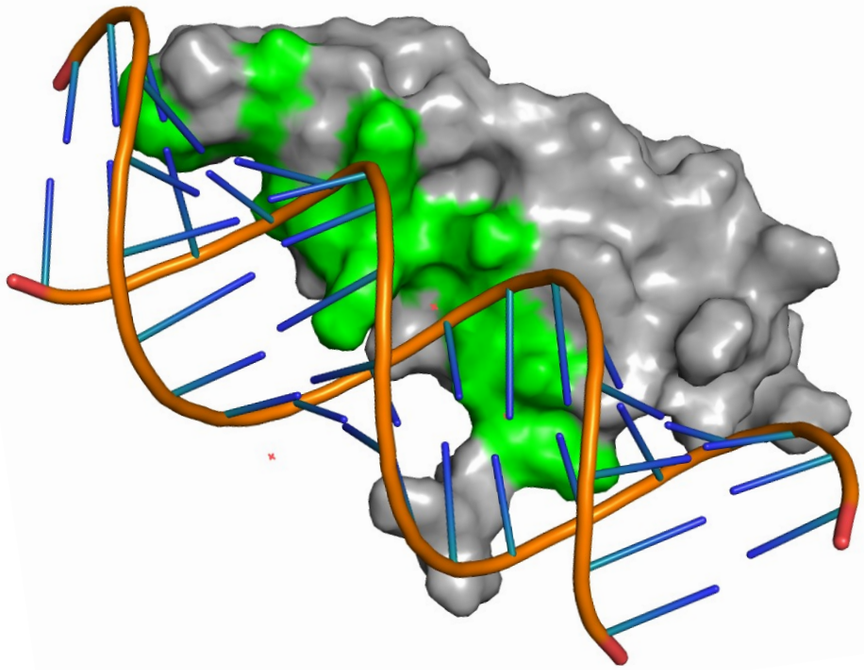
Protein-nucleic acid interactions:
from structures to binding sites



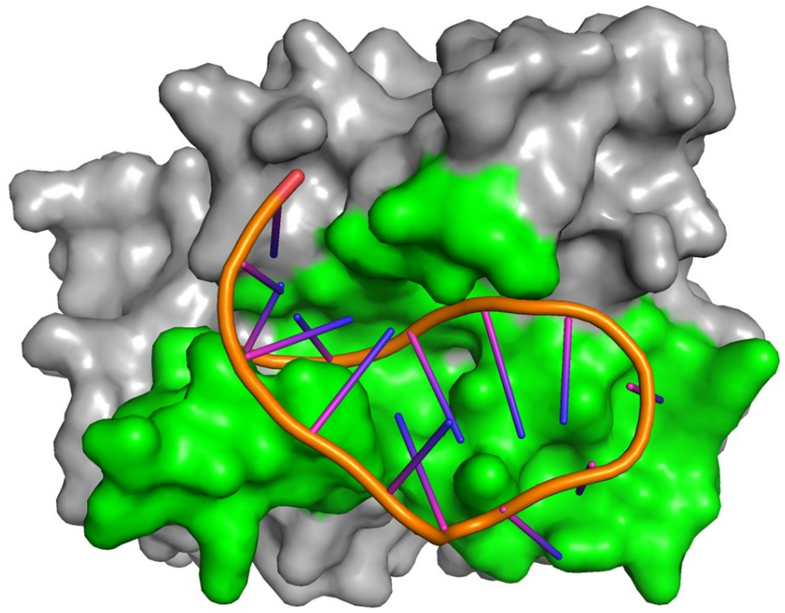
Protein-DNA interaction

atomic level
structure

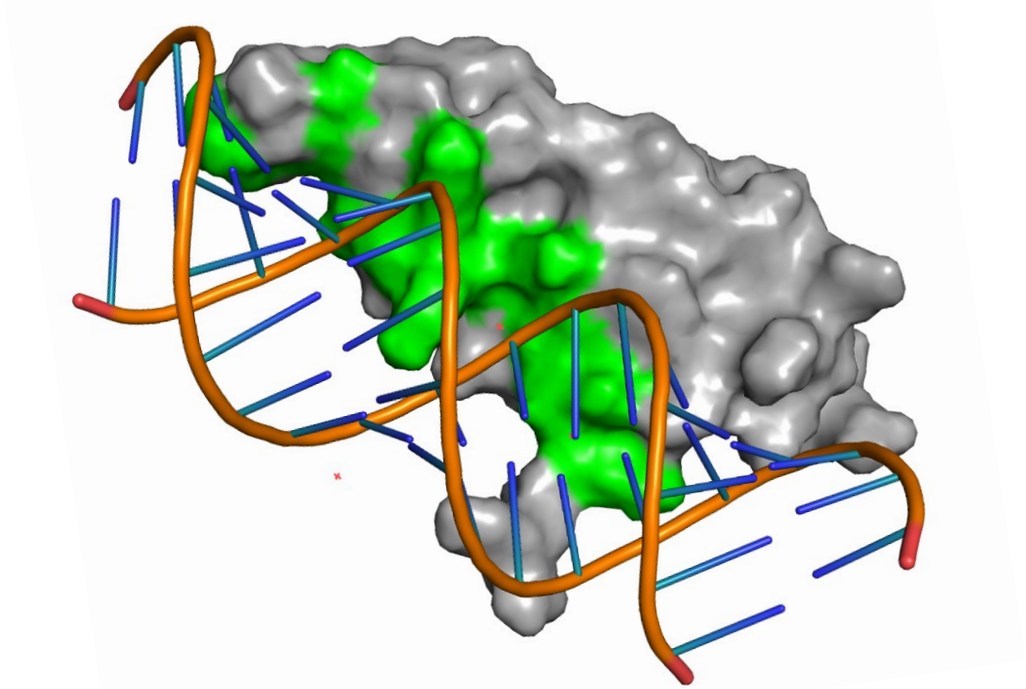
Protein-RNA interaction



binding sites



Protein-nucleic acid binding sites



Protein-DNA interaction

Protein-RNA interaction

- Binding sites = atom pairs within sum of van der Waal's radius+0.5Å
- Unfortunately, experimental characterization is time-consuming and expensive

This talk…

I. **Protein-nucleic acid binding site prediction**
powered by LLMs & deep graph learning

II. **Single-sequence protein-nucleic acid 3D structure prediction**
using geometric attention-enabled pairing of bio LLMs

III. **Future directions**
AI-powered biomolecular modeling

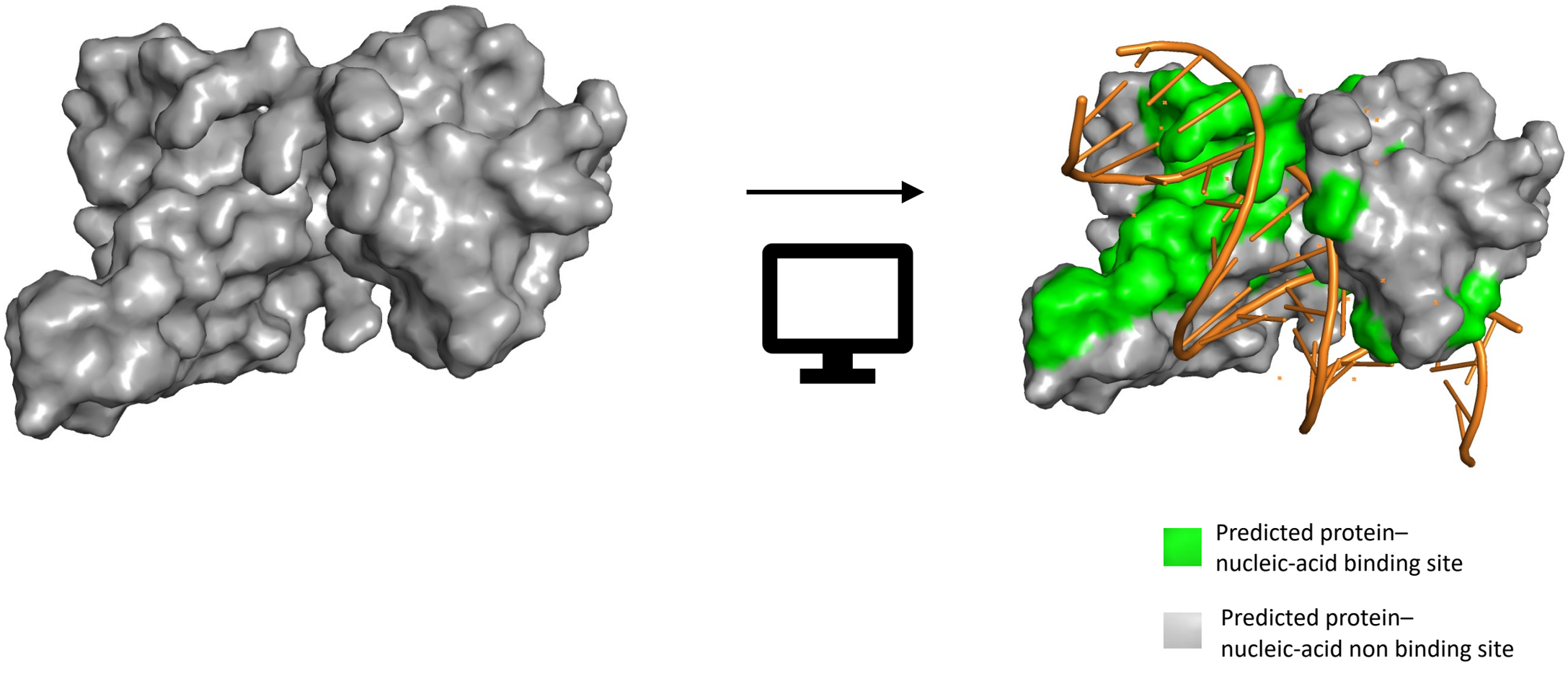
This talk…

I. **Protein-nucleic acid binding site prediction**
powered by LLMs & deep graph learning

II. **Single-sequence protein-nucleic acid 3D structure prediction**
using geometric attention-enabled pairing of bio LLMs

III. **Future directions**
AI-powered biomolecular modeling

Protein-nucleic acid binding site prediction:
a protein-centric view



Partner-independent protein-nucleic acid binding site prediction

Protein-nucleic acid binding site prediction: existing approaches

Sequence-based

Utilizes protein sequence + evolutionary information

(e.g., SVMnuc, NCBRPred,…)

+ Sequence readily available

- Tends to be less accurate

Structure-aware

Utilizes protein sequence + evolutionary + structural information

(e.g., GraphBind, GraphSite)

+ Generally more accurate

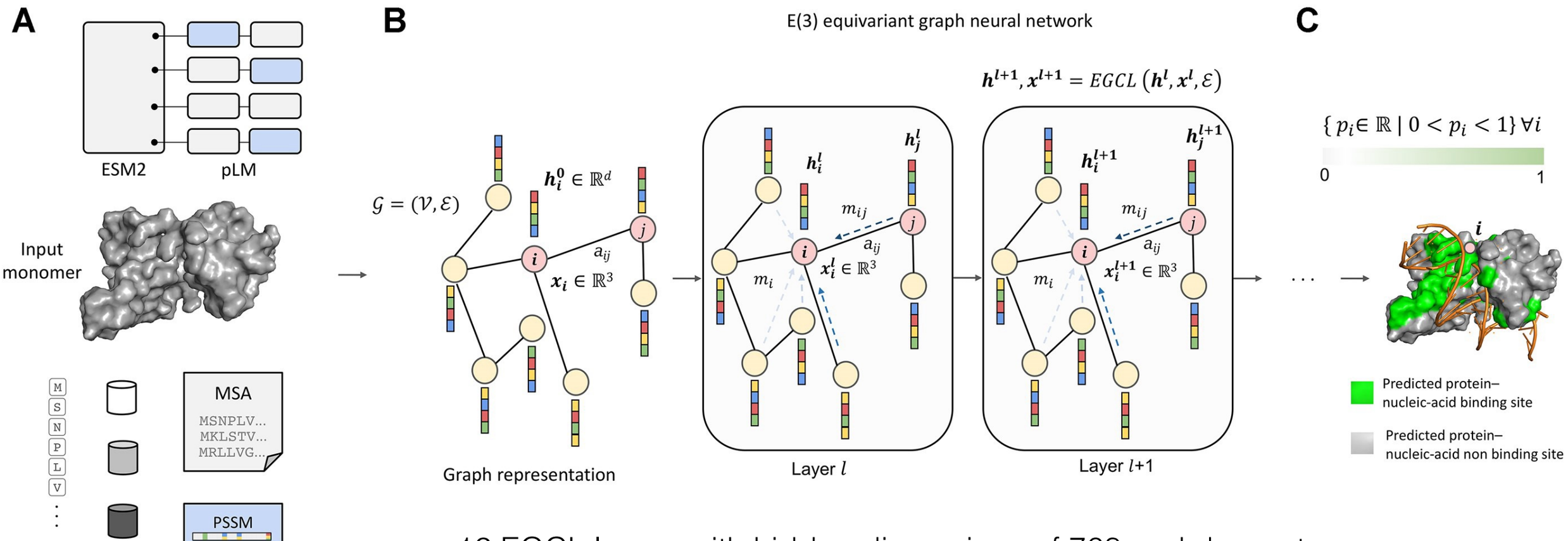
- Needs structural data, but AlphaFold can be used

What if we integrated
protein language model (pLM) embeddings?

Su, H. et. al., 2019
Zhang, J. et. al., 2021
Yuan, Q. et. al., 2022
Xia, Y. et. al., 2021

EquiPNAS:

pLM-informed equivariant deep graph learning



12 EGCL layers with hidden dimensions of 768,and dropout rate of 0.1 determined through 5-fold cross-validation

Graph node classification for residue-level binding site prediction

Roche, R. et. al., 2024

EquiPNAS features:
sequence- and pLM-based features

Features *[shape]*	**Description**
aa *[L, 20]*	One-hot encodings of 20 amino acid residue types.
PSSM *[L, 20]*	Normalized position specific scoring matrix (PSSM).
MSA *[L, 256]*	Multiple sequence alignment (MSA) representation distilled through ColabFold's EvoFormer blocks.
pLM *[L, 5120]*	pLM embeddings from ESM-2 with 15B parameters.

Lin, Z. et. al., 2023

EquiPNAS features:
structures-based features

Features *[shape]*	**Description**
SS *[L, 11]*	One-hot encodings of 3- and 8-state secondary structure.
RSA *[L, 10]*	One-hot encodings of 2- and 8-state relevant solvent accessibility.
Local geometry *[L, 11]*	Cosine angle between the C=O of consecutive residues, normalized values of virtual bond and torsion angles, and normalized peptide backbone torsion angles.
Residue orientation *[L, 9]*	Unit vectors pointing towards the directions of $C_\alpha^{(i+1)} - C_\alpha^i$, $C_\alpha^{(i-1)} - C_\alpha^i$ and $C_\beta^i - C_\alpha^i$.
Relative residue positioning *[L, 2]*	Two types of relative positional features for the i^{th} residue: (1) inverse of i representing the relative sequence position, and (2) inverse of the Euclidean distance of C_α atom from the centroid representing the relative spatial positioning.
Residue virtual surface area *[L, 1]*	Virtual surface area of the conceptual convex hull constructed by the atoms in a residue.
Contact count *[L, 1]*	The number of spatial neighbors of each residue.

EquiPNAS features:
coordinate and edge features

- Coordinate feature:

  C_α (x, y, z) coordinates from input protein structure

- Edge feature:

  Ratio of logarithmic sequence separation & Euclidian distance

$$a_{ij} = \frac{log(abs(i-j))}{||x_i - x_j||}$$

EquiPNAS datasets:
standard sets from the BioLiP database

- Protein-DNA
    - Train: 573 targets; 14,479 binding & 145,404 non-binding sites
    - Test:181 targets; 3,208 binding & 72,050 and non-binding sites
- Protein-RNA
    - Train: 495 targets;14,609 binding & 122,290 non-binding sites
    - Test: 117 targets; 2,031 binding & 35,314 non-binding sites

NOTE: Pre-processed to filter out protein chains with >30% sequence similarity between the train and test sets using CD-HIT

Yang , J. et. al., 2012
Yuan, Q. et. al., 2022
Xia, Y. et. al., 2021

EquiPNAS results:
protein-DNA binding site prediction

- Use AlphaFold2 predicted structural models as input

- Randomly sample 70% of the targets for each of the test sets, repeating it 10 times, means and standard deviations are reported

Datasets		Methods	ROC-AUC	PR-AUC
Protein-DNA	Test_181	GraphBind	0.8916 ± 0.006003703	0.3102 ± 0.017706245
		p-value	8.63327E-08	7.16361E-09
		GraphSite	0.8964 ± 0.006292853	0.3286 ± 0.018124262
		p-value	2.25585E-07	7.9832E-07
		EquiPNAS	**0.9159** ± 0.00395671	**0.3717** ± 0.018372987

EquiPNAS results:
protein-RNA binding site prediction

- Use AlphaFold2 predicted structural models as input
- Randomly sample 70% of the targets for each of the test sets, repeating it 10 times, means and standard deviations are reported

	Datasets	Methods	ROC-AUC	PR-AUC
Protein-RNA	Test_117	GraphBind	0.7942 ± 0.006250333	0.2019 ± 0.009573691
		p-value	2.3402E-11	1.44E-10
		EquiPNAS	**0.8856** ± 0.006221825	**0.3118** ± 0.013003

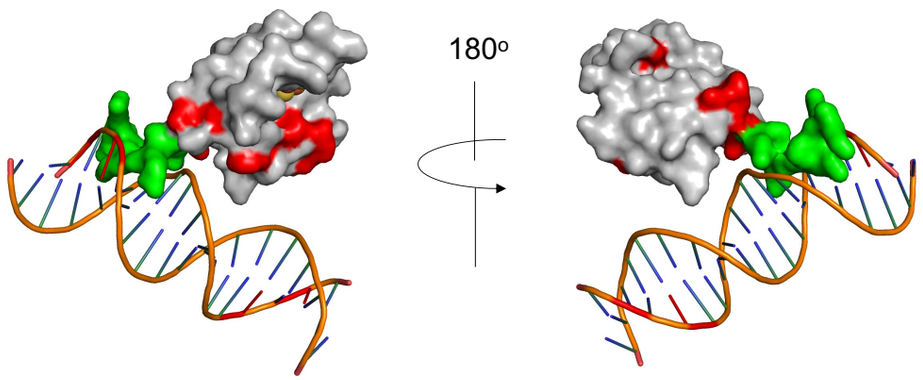
Case study

Green: TP
Red: FP
Yellow: FN

Protein-DNA

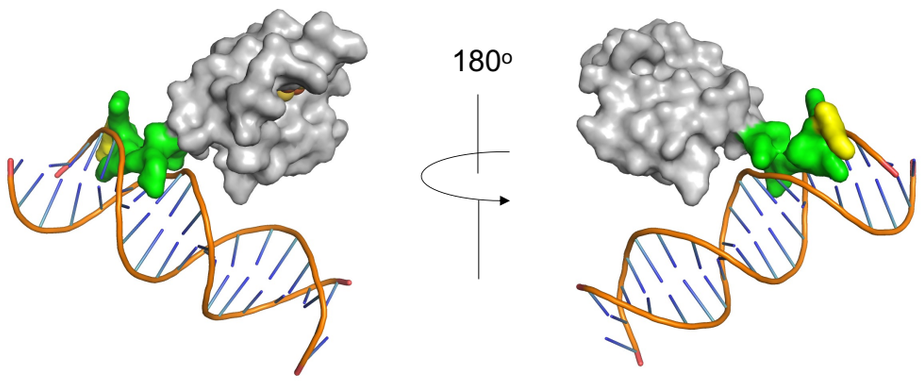
7kuf_A

GraphSite



F1-score = 0.64, MCC = 0.637
ROC-AUC = 0.985, PR-AUC = 0.837

EquiPNAS

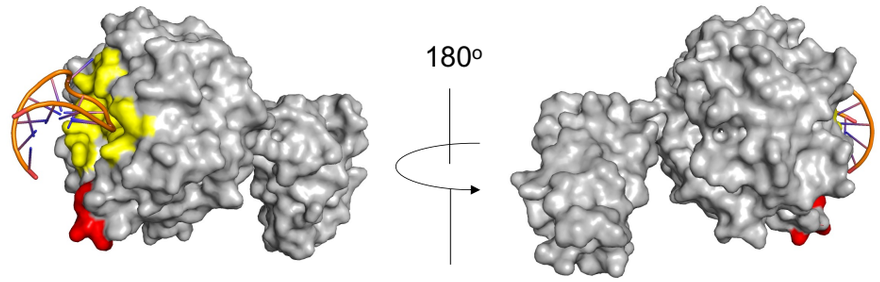


F1-score = 0.933, MCC = 0.928
ROC-AUC = 1.0, PR-AUC = 1.0

Protein-RNA

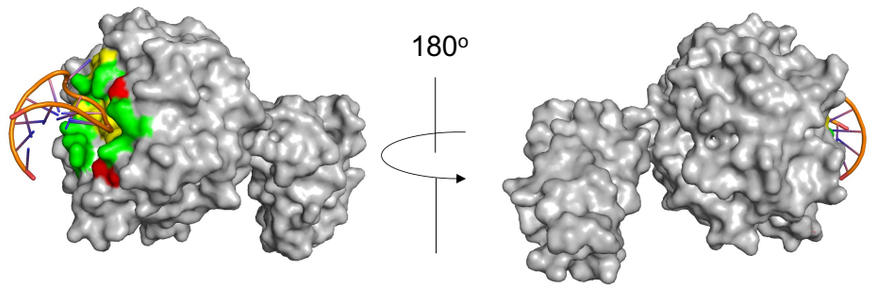
6fq3_A

GraphBind



F1-score = 0, MCC = -017
ROC-AUC = 0.447, PR-AUC = 0.024

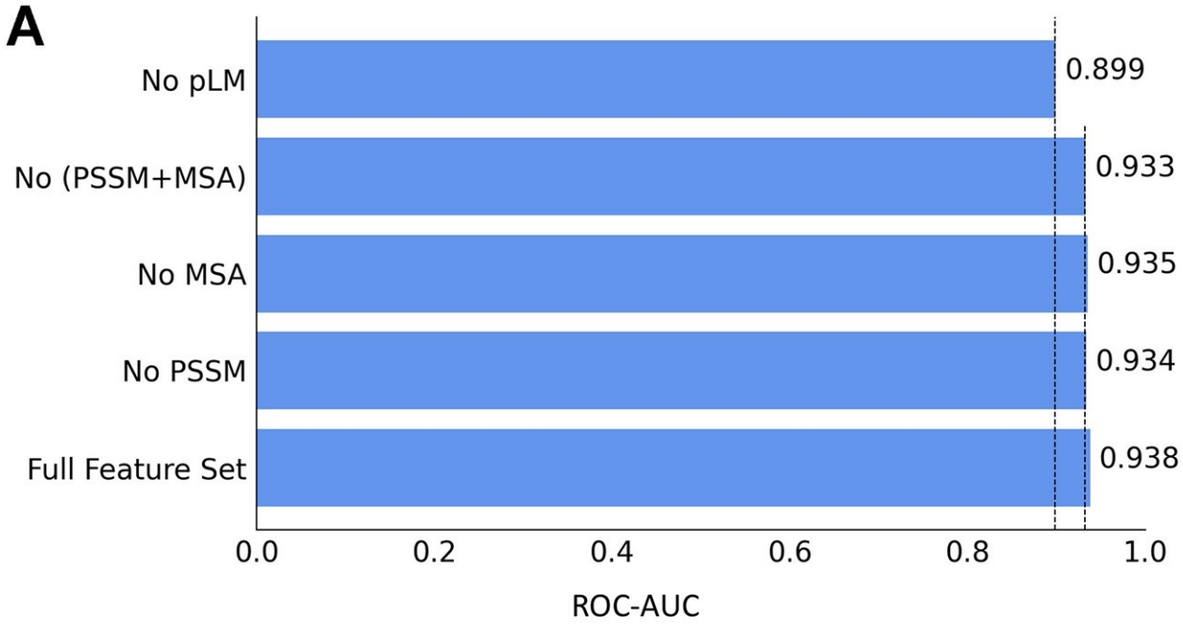
EquiPNAS

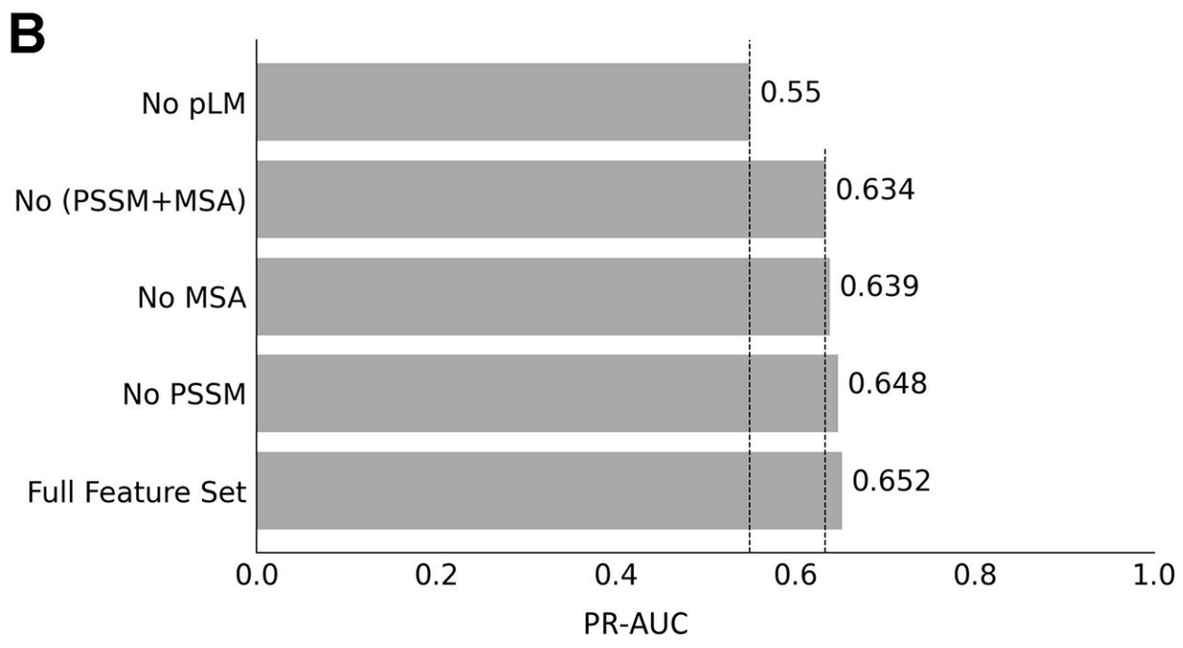


F1-score = 0.545, MCC = 0.555
ROC-AUC = 0.988, PR-AUC = 0.732

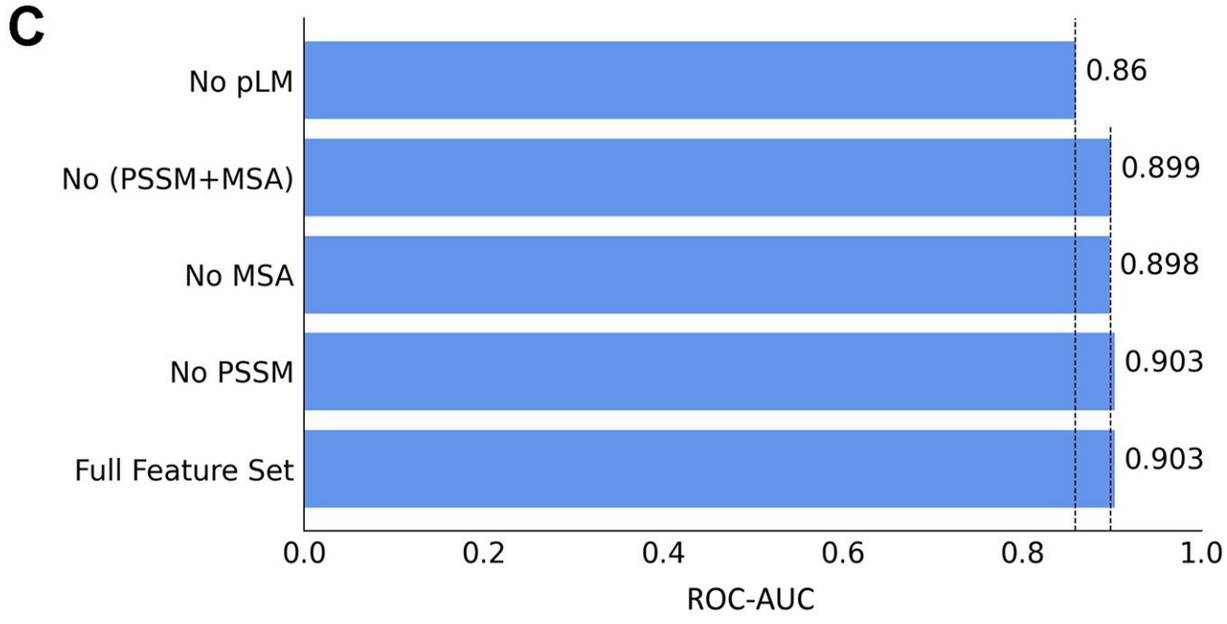
Ablation study using 5-fold cross-validation: contribution of pLM embeddings

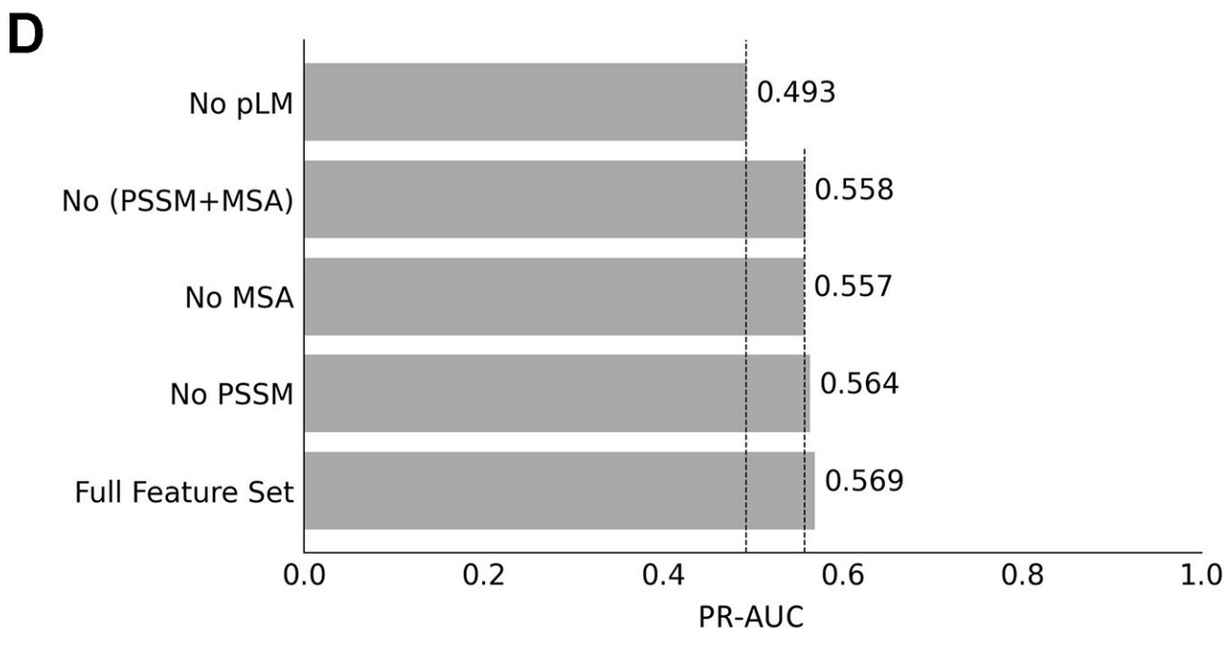
Protein-DNA





Protein-RNA





Ablation study:

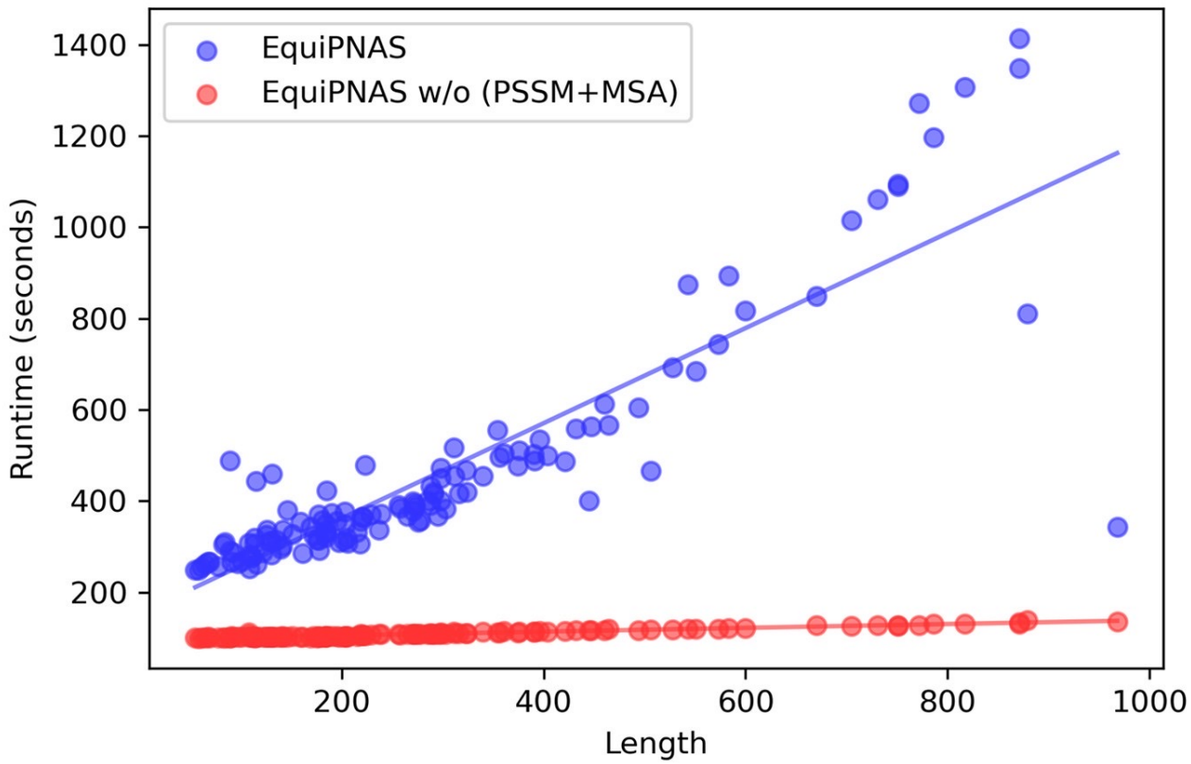
EquiPNAS results w/o (MSA +PSSM)

	Datasets	Methods	ROC-AUC	PR-AUC
		GraphBind	0.793	0.204
Protein-RNA	Test_117	EquiPNAS w/o (MSA+PSSM)	0.877	0.299
		EquiPNAS	**0.886**	**0.320**

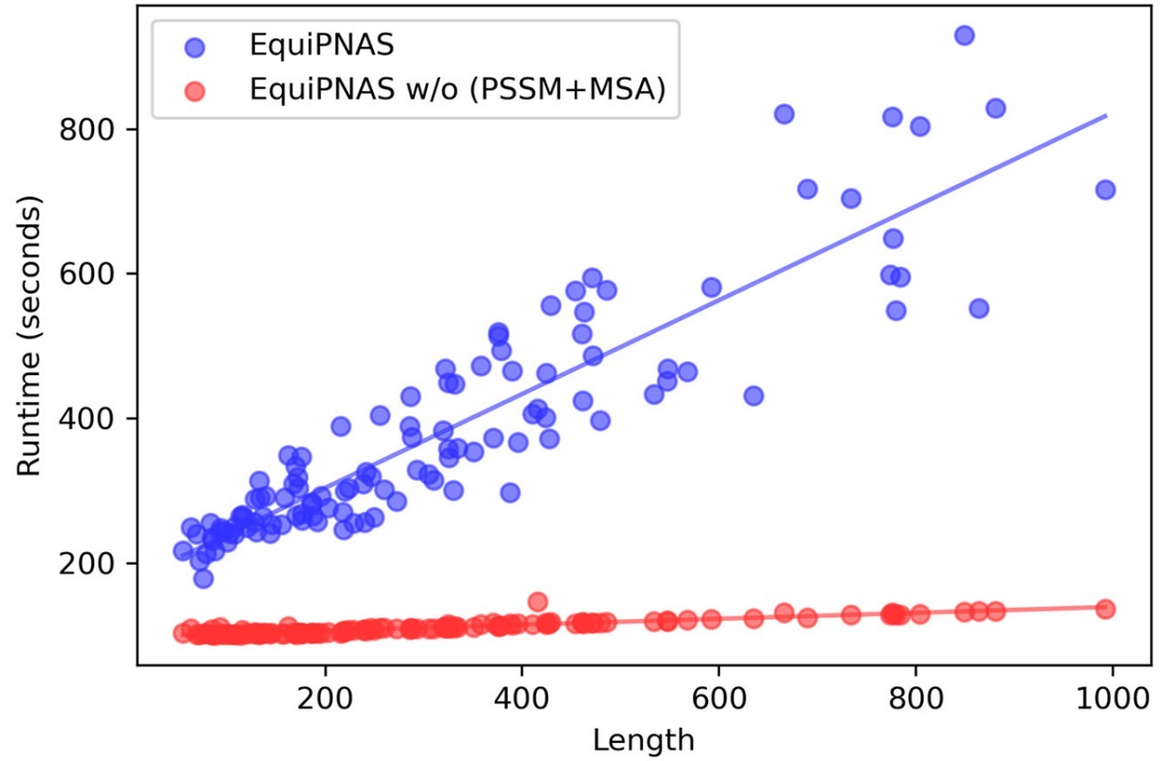
pLM embeddings reduce the dependence on the availability of explicit evolutionary information without a drastic drop in accuracy

Ablation study:
running time

Protein-DNA

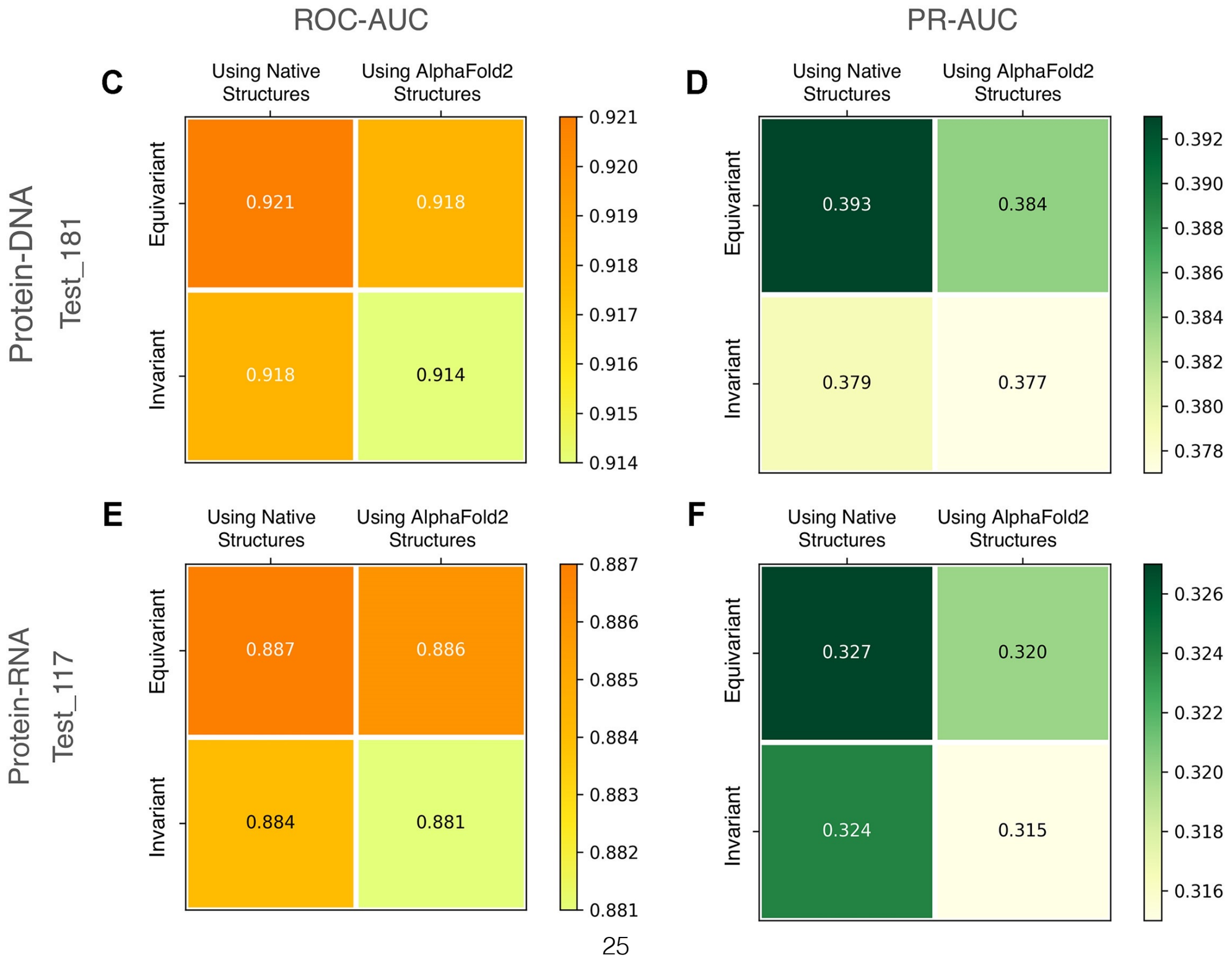


Protein-RNA



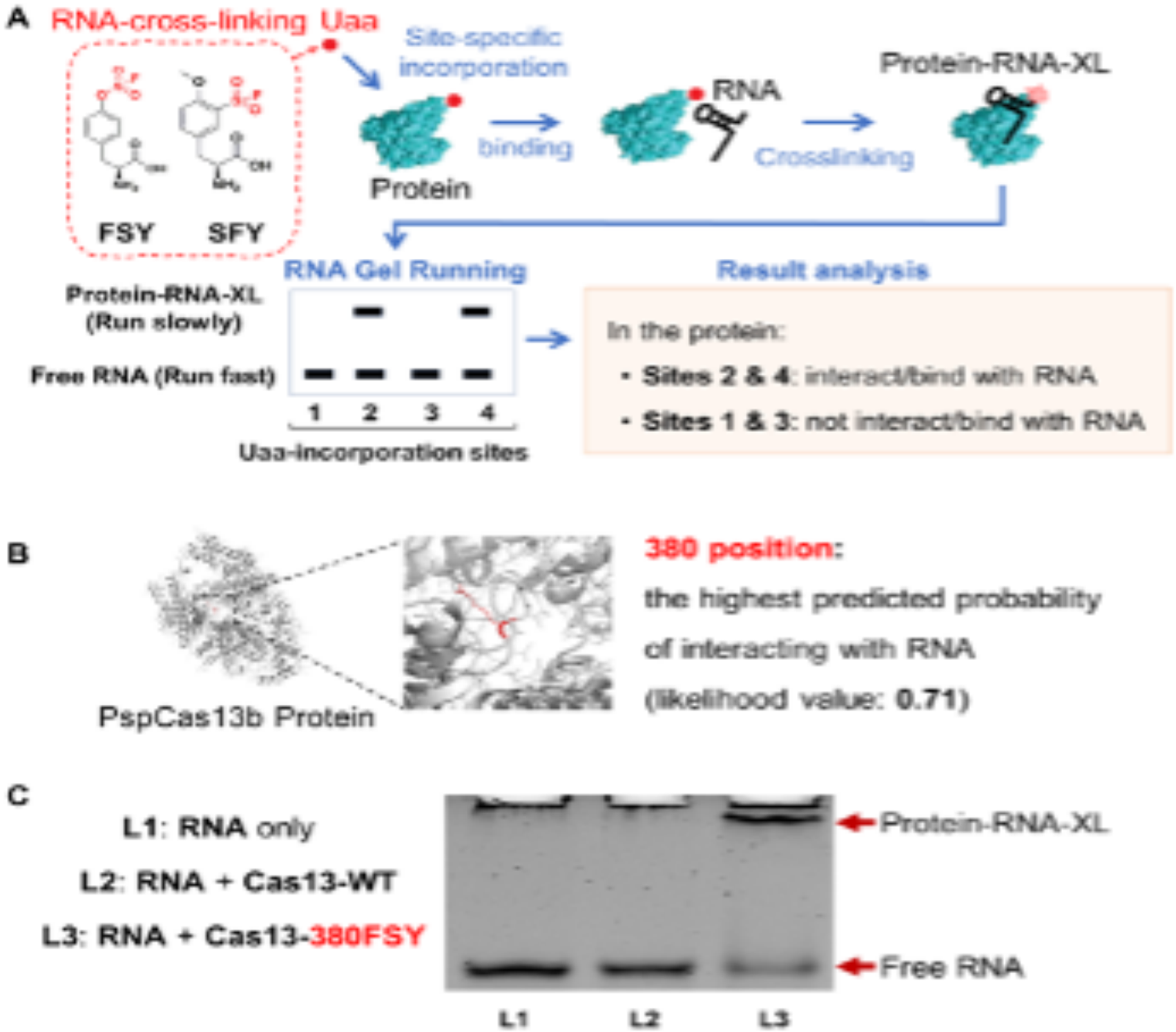
Bypassing the search for explicit evolutionary information leads to orders of magnitude speedup

Ablation study:
contribution of equivariance



Experimental validation:
using GECX-RNA

Slide credit: Sun, W. et. al



Sun, W. et. al., 2023

This talk…

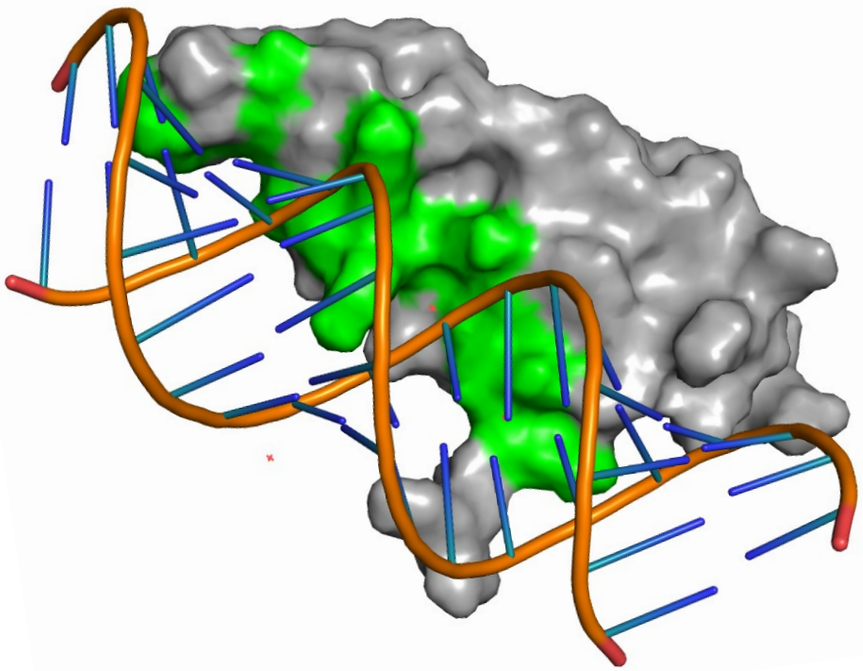
I. **Protein-nucleic acid binding site prediction**
powered by LLMs & deep graph learning

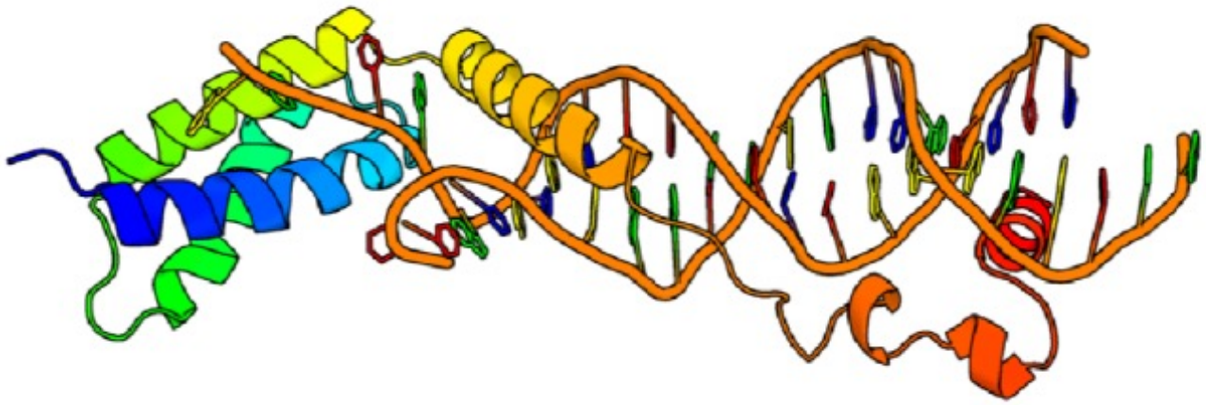
II. **Single-sequence protein-nucleic acid 3D structure prediction**
using geometric attention-enabled pairing of bio LLMs

III. **Future directions**
AI-powered biomolecular modeling

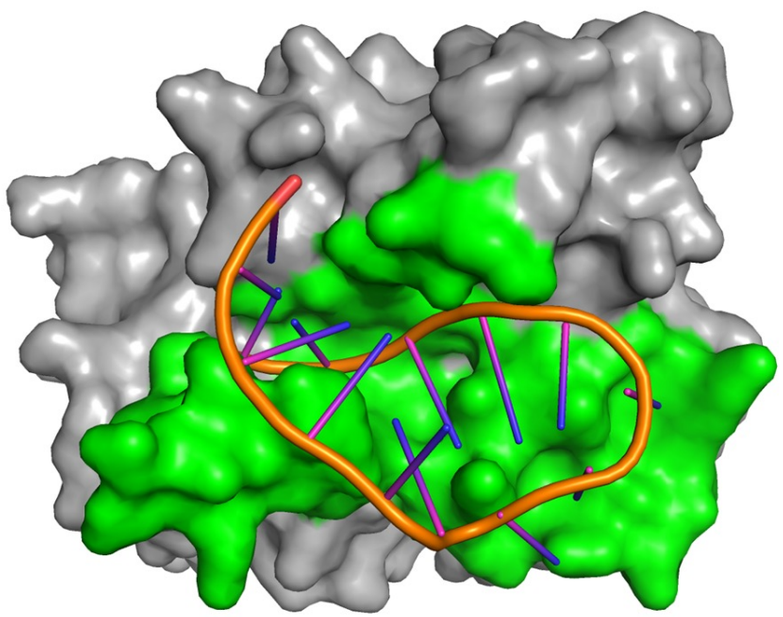
Beyond binding sites:
protein-nucleic acid complex structures

Protein-DNA interaction





Protein-RNA interaction



binding sites

atomic level structure

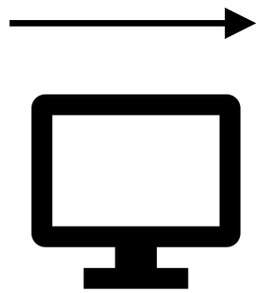
Protein-nucleic acid 3D structure prediction:
from sequence to all-atom coordinates

Protein sequence

A K A P A S P ...

RNA sequence

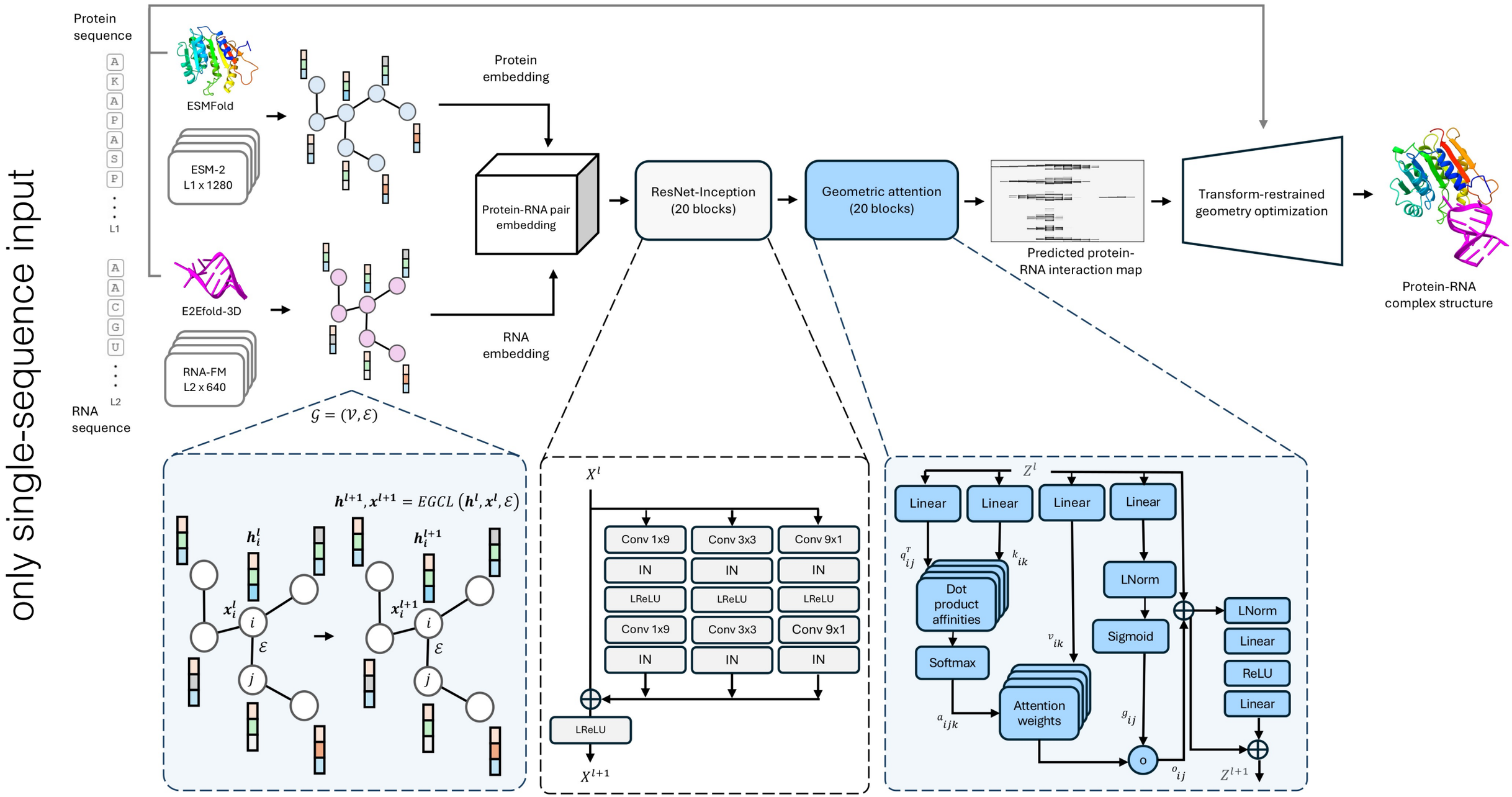
A A C G U ...



Protein-RNA complex
3D structure

ProRNA3D-single :

protein-nucleic acid 3D structure prediction



Geometric attention-enabled pairing of biological LLMs

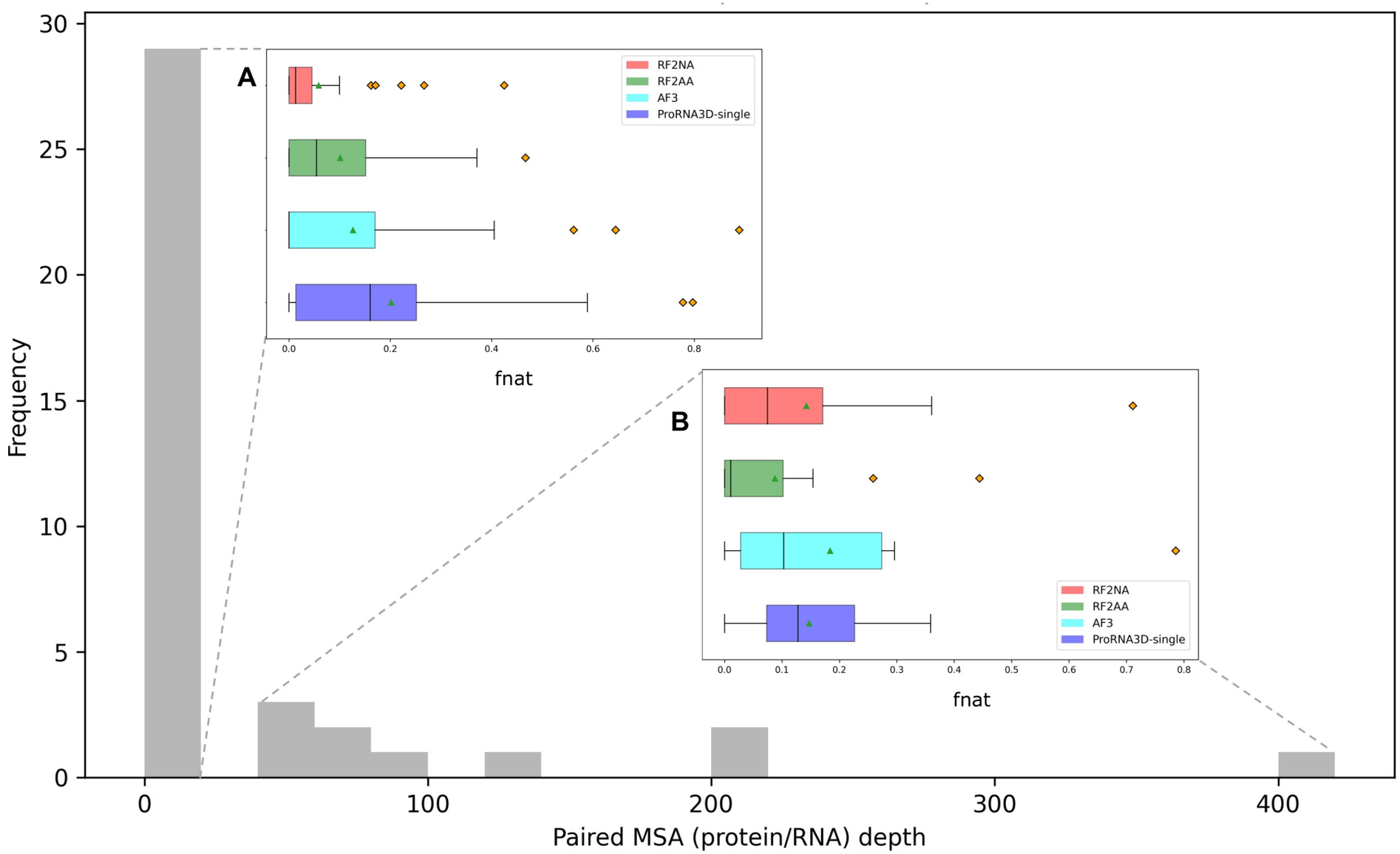
Roche, R. et. al., 2025

ProRNA3D-single datasets:
X-ray crystal structures from the PDB

- Filtering
    - resolution < 3.5 Å
    - deposited to the PDB on and before October 2022
- Datasets
    - Train: 750 targets
    - Validation: 48 targets

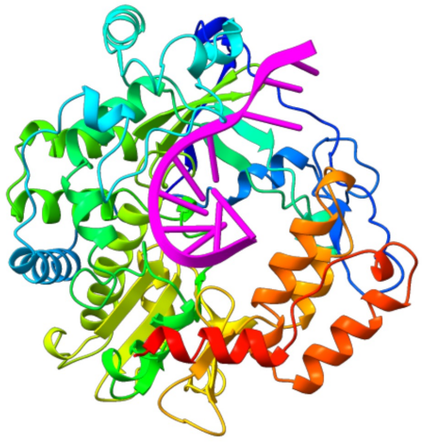
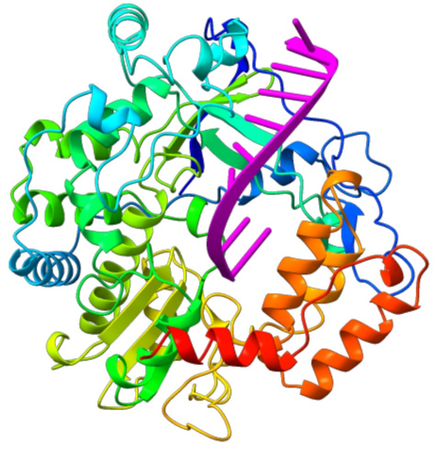
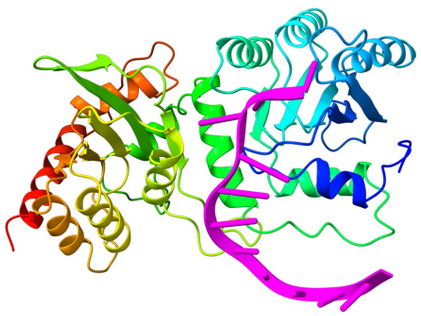
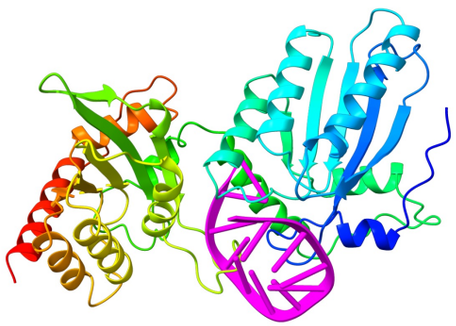
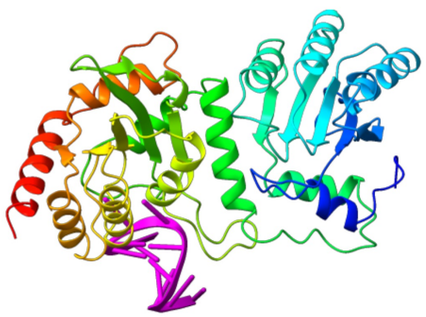
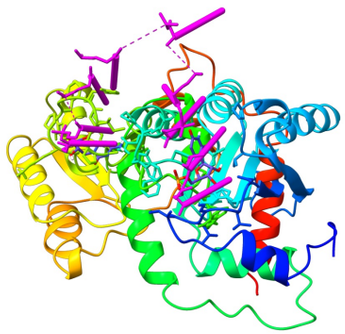
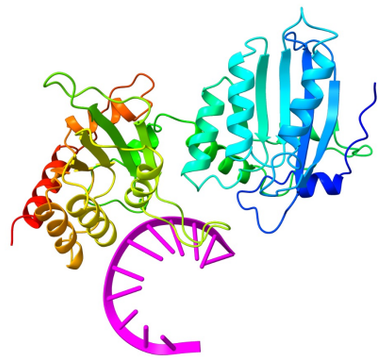
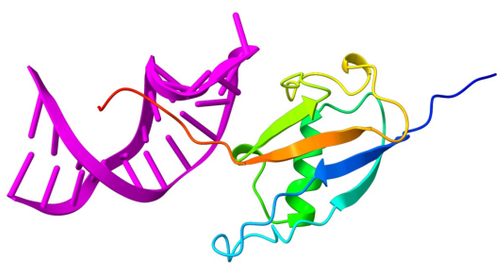
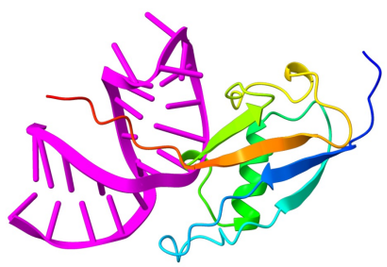
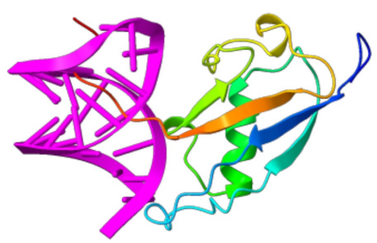
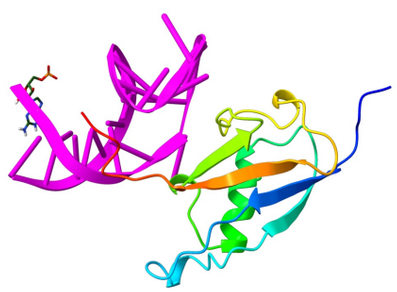
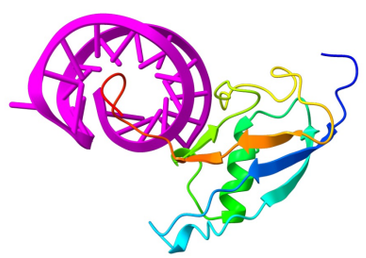
NOTE: Test set consists of targets released on and after November 2022 till November 2023

ProRNA3D-single results:
protein-RNA complex structure prediction

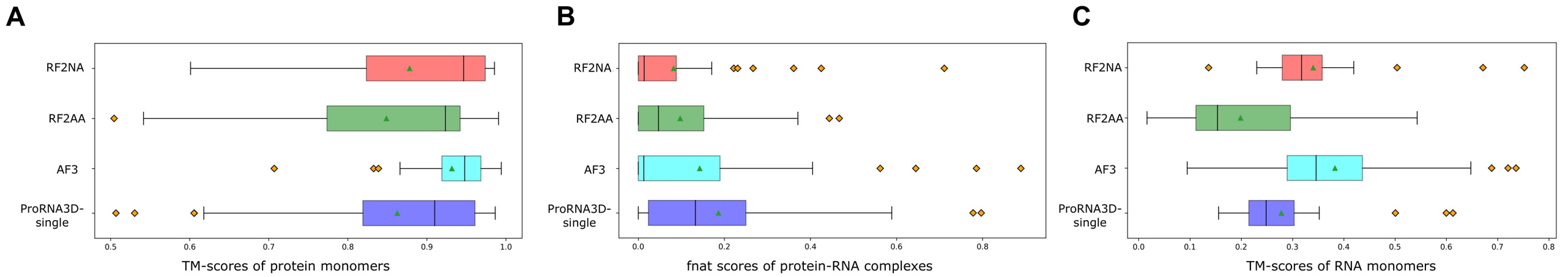


ProRNA3D-single outperforms SOTA methods including AlphaFold3, particularly when evolutionary information is limited

Case study

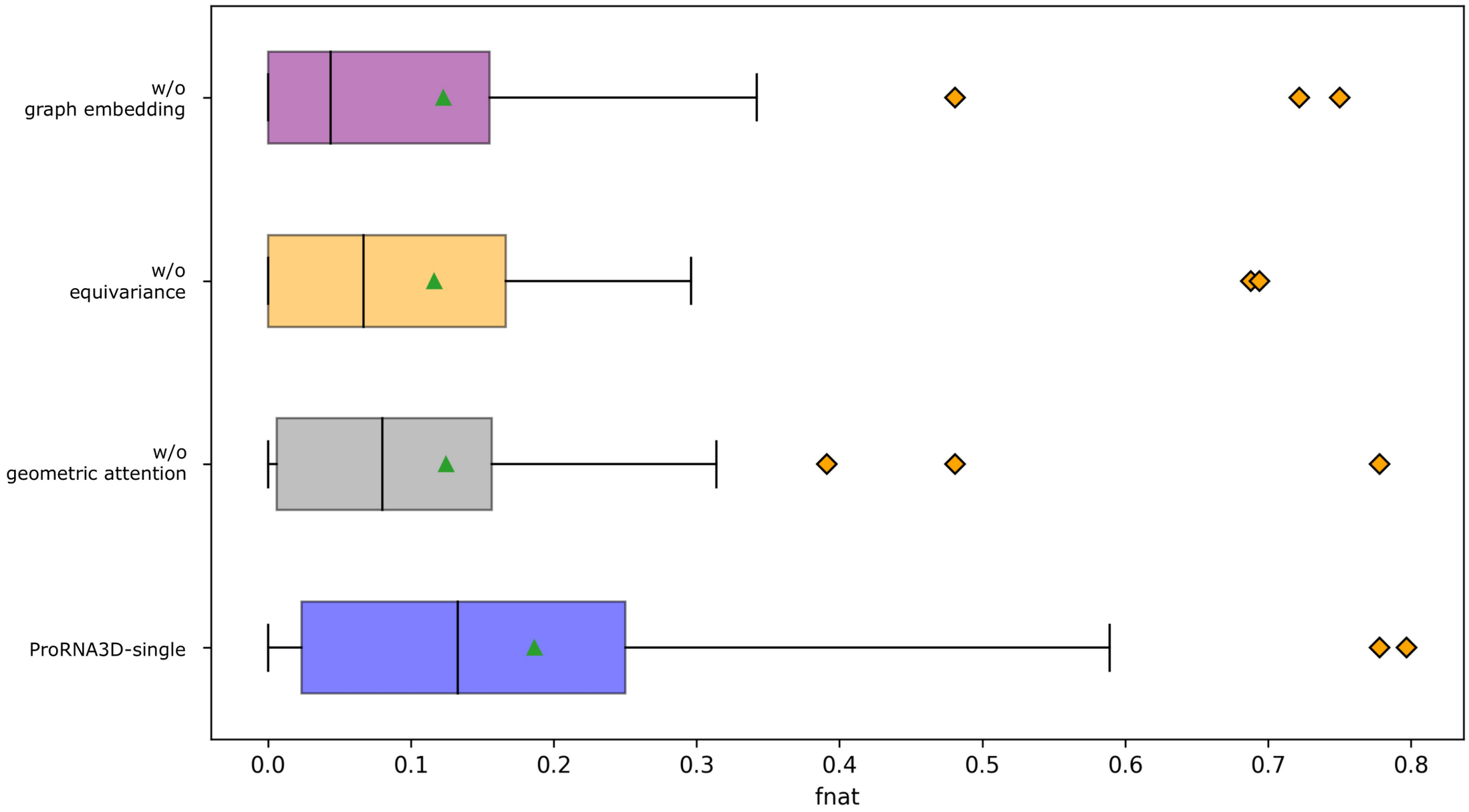
	Experimental	ProRNA3D-single	AF3	RF2AA	RF2NA
7W9S_AC		 fnat = 0.778, iRMS = 1.51, lRMS = 4.074	 fnat = 0, iRMS = 6.355, lRMS = 20.814	 fnat = 0.028, iRMS = 7.488, lRMS = 24.104	 fnat = 0.222, iRMS = 6.661, lRMS = 22.518
7UU3_BD No sequence in Paired MSA		 fnat = 0.457, iRMS = 10.194, lRMS = 18.67	 fnat = 0, iRMS = 14.845, lRMS = 45.129	 fnat = 0.371, iRMS = 12.215, lRMS = 23.528	 fnat = 0, iRMS = 12.375, lRMS = 29.742
8ID2_AC 43 sequences in Paired MSA		 fnat = 0.359, iRMS = 4.341, lRMS = 8.304	 fnat = 0.282, iRMS = 10.591, lRMS = 17.646	 fnat = 0.154, iRMS = 5.095, lRMS = 11.602	 fnat = 0.077, iRMS = 11.944, lRMS = 22.694

Ablation study:
effect of modeling accuracy of the individual components



Performance gain is not because of individual protein or RNA component modeling, but due to improved inter-protein-RNA interaction prediction

Ablation study:
contribution of the neural architecture



Both symmetry-aware graph convolutions and geometric attention module are the key modules of the neural architecture of ProRNA3D-single

This talk…

I. **Protein-nucleic acid binding site prediction**
powered by LLMs & deep graph learning

II. **Single-sequence protein-nucleic acid 3D structure prediction**
using geometric attention-enabled pairing of bio LLMs

III. Future directions
AI-powered biomolecular modeling

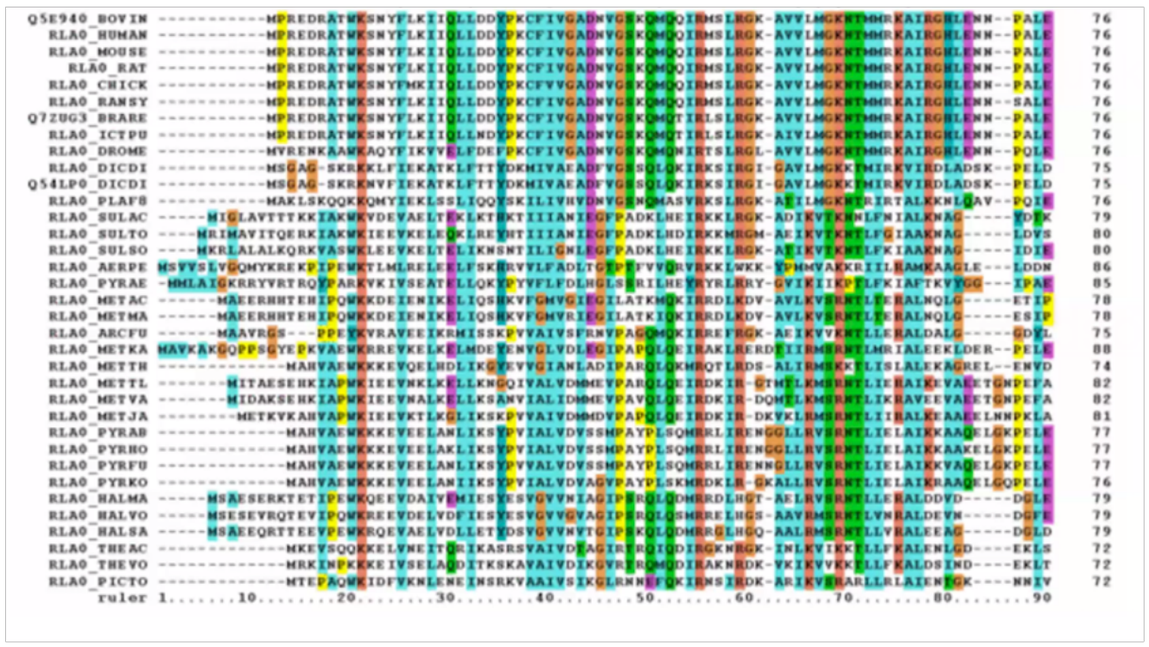
Beyond “Darwinian” biomolecular modeling

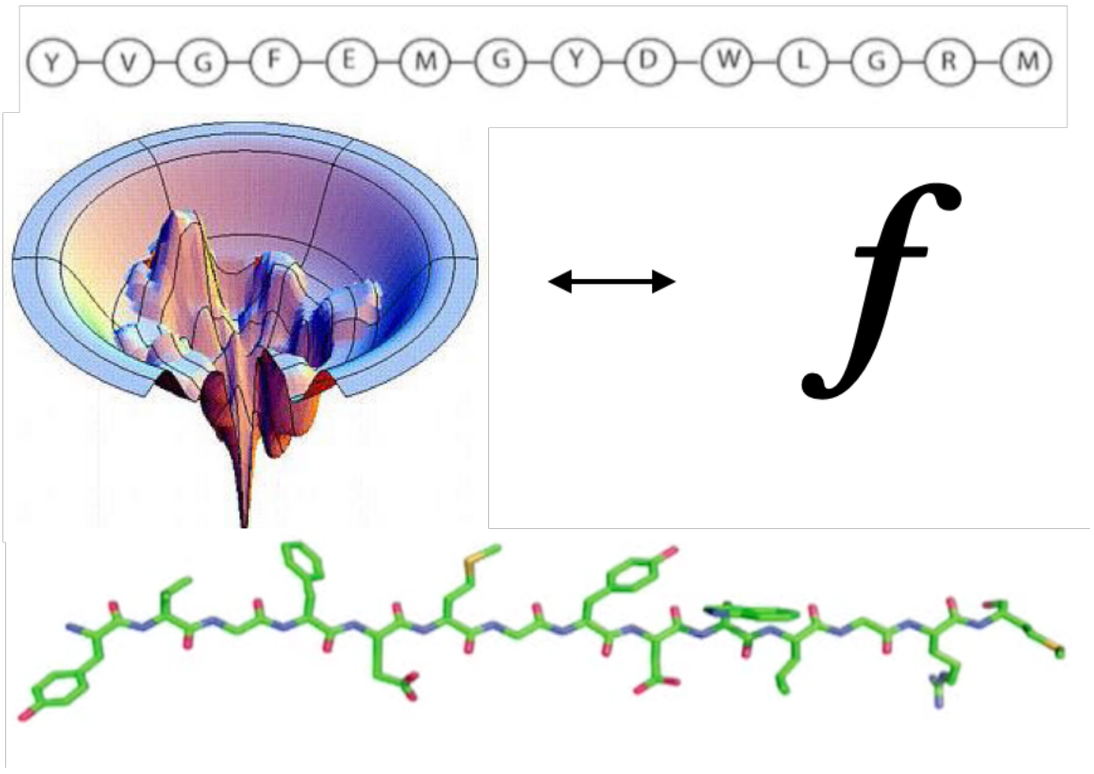
Problem settings

- **Input**: just single sequences
- **Output**: folded and functional 3D structures

Possible solutions

- Representation learning of sequence spaces informed by embeddings from bio LLMs
- Combine biophysics (force fields) with machine learning for first-principles folding





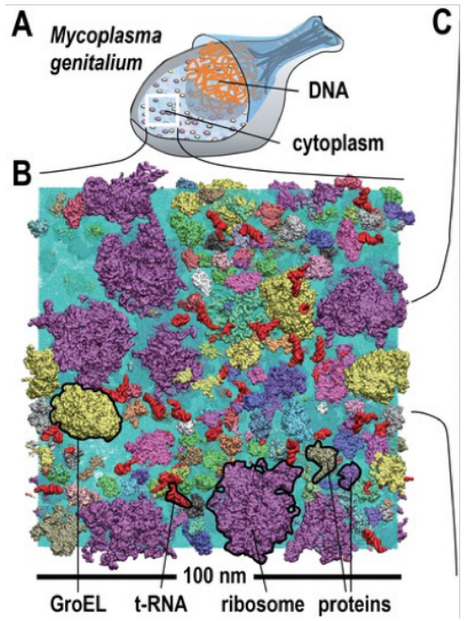
From molecular to cellular scale

Connecting

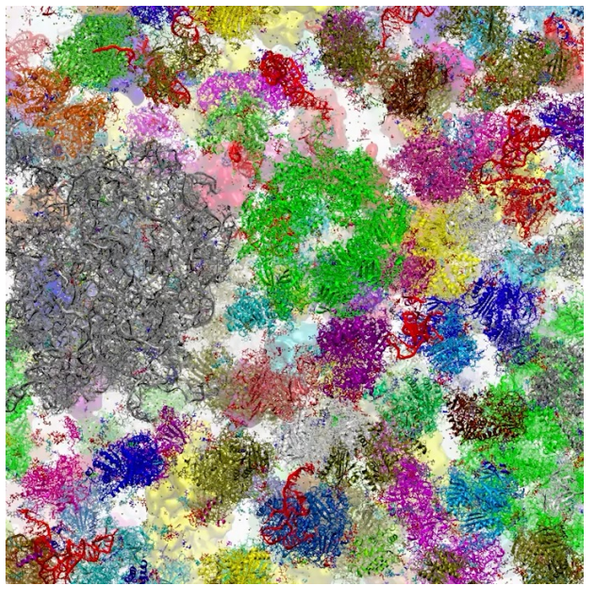
- **Molecular** scale, where fundamental biochemistry takes place
- **Cellular** scales, where biological function (or dysfunction) is realized

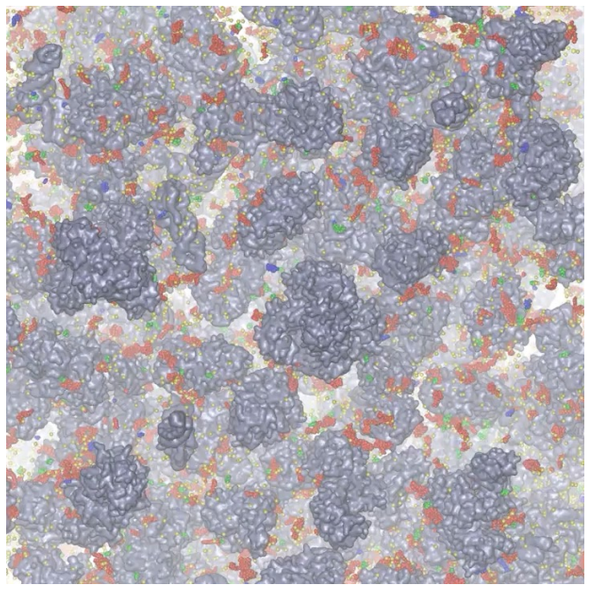
Multiscale modeling

- Computational modeling of large macromolecular complexes and assemblies
- Machine learning for modeling intra- and inter-molecular interactions



Bacterial cytoplasm


Nanosecond dynamics


Diffusive motion of metabolites

Yu, I. et. al., 2016

Thank you!

Students & Collaborators

Rahmatullah Roche

Bernard Moussad

Md Hossain Shuvo

Sumit Tarafder

Trevor Norton

Xinyu Wang

Wei Sun

Public databases

PDB

BioLiP

Organizers

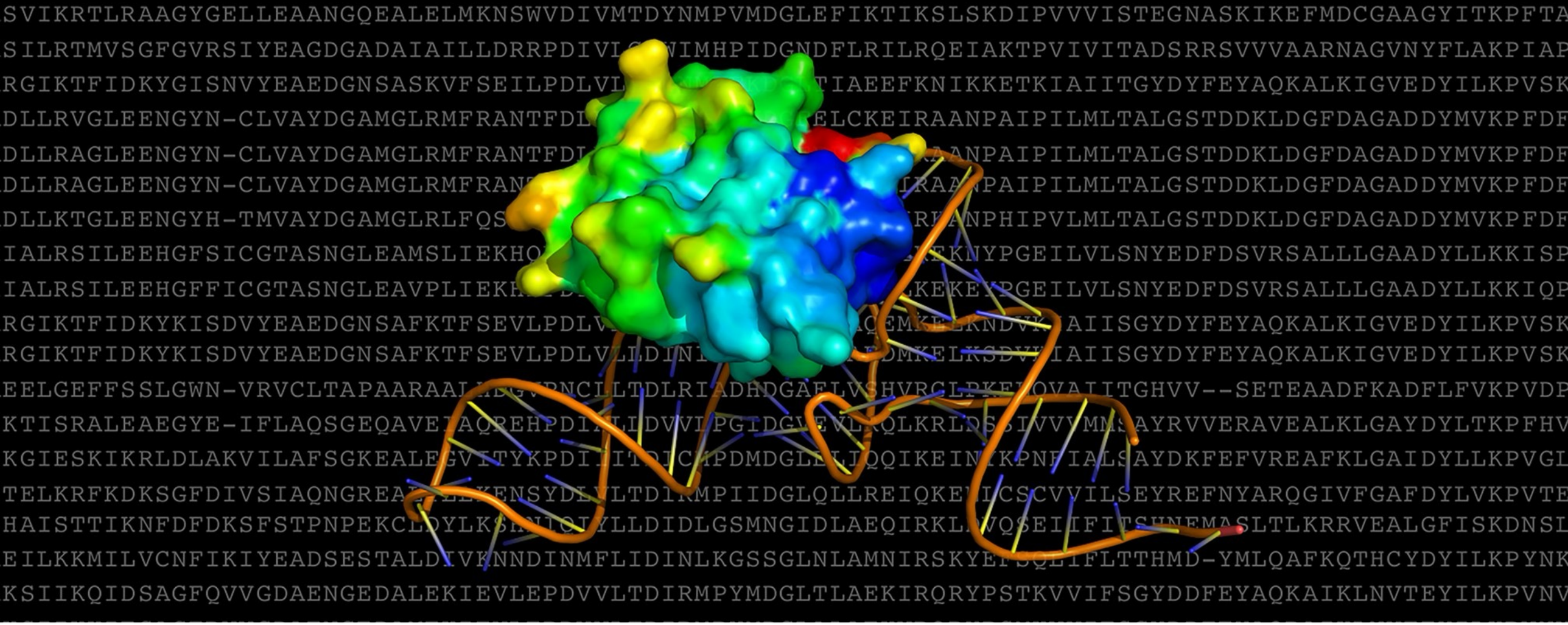
Workshop for AI-Powered Materials Discovery in the Great Plains

University of South Dakota

Funding sources

NIGMS R35 GM138146

NSF CAREER DBI-1942692
NSF CISE IIS-2030722



Open-source code available at:

https://github.com/Bhattacharya-Lab/