

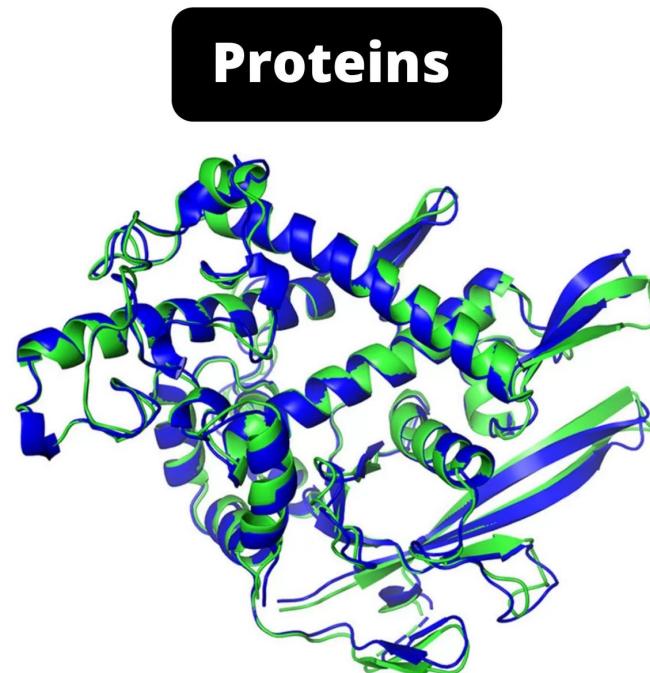
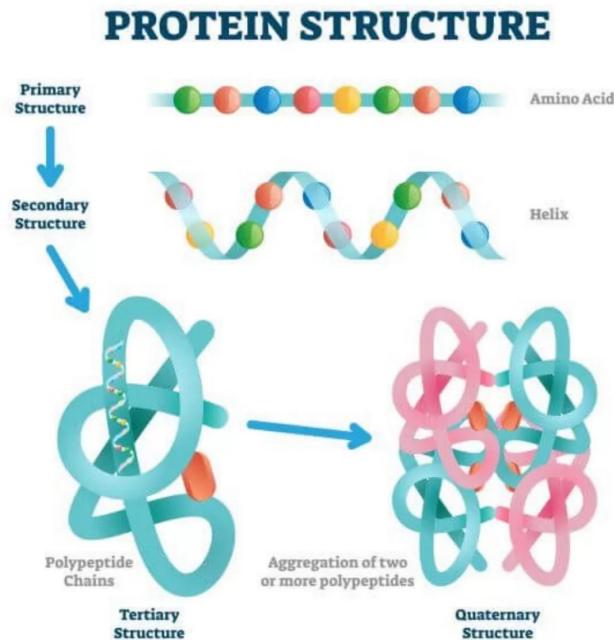
# Transformer-based protein generation with regularized latent space optimization



Xinchun Ran  
Department of Chemistry  
Vanderbilt University

# Background: Protein Sequences

The amino acid sequence of a polypeptide chain determines the final 3D structure of the protein

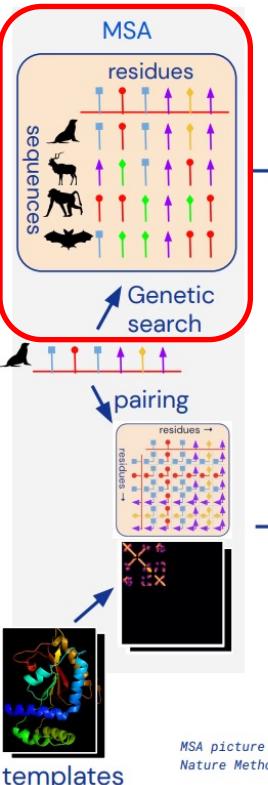


# Background: Evolutionary information (MSA)

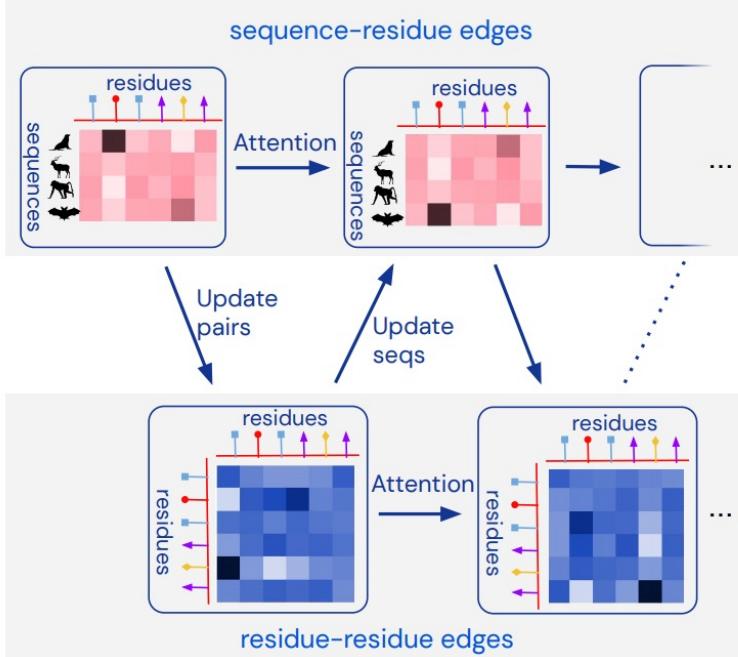
	*	:	:	*	:	:	*	:	:	*	:	:
Q5E940_BOVIN	-	-	-	-	-	-	-	-	-	-	-	76
RLAO_HUMAN	-	-	-	-	-	-	-	-	-	-	-	76
RLAO_MOUSE	-	-	-	-	-	-	-	-	-	-	-	76
RLAO_RAT	-	-	-	-	-	-	-	-	-	-	-	76
RLAO_CHICK	-	-	-	-	-	-	-	-	-	-	-	76
RLAO_RANSY	-	-	-	-	-	-	-	-	-	-	-	76
Q7ZUG3_BRARE	-	-	-	-	-	-	-	-	-	-	-	76
RLAO_ICTPU	-	-	-	-	-	-	-	-	-	-	-	76
RLAO_DROME	-	-	-	-	-	-	-	-	-	-	-	76
RLAO_DICDI	-	-	-	-	-	-	-	-	-	-	-	75
Q54LP0_DICDI	-	-	-	-	-	-	-	-	-	-	-	75
RLAO_PLAF8	-	-	-	-	-	-	-	-	-	-	-	76
RLAO_SULAC	-	-	-	-	-	-	-	-	-	-	-	79
RLAO_SULTO	-	-	-	-	-	-	-	-	-	-	-	80
RLAO_SULSO	-	-	-	-	-	-	-	-	-	-	-	80
RLAO_AERPE	-	-	-	-	-	-	-	-	-	-	-	86
RLAO_PYRAE	-	-	-	-	-	-	-	-	-	-	-	85
RLAO_METAC	-	-	-	-	-	-	-	-	-	-	-	78
RLAO_METMA	-	-	-	-	-	-	-	-	-	-	-	78
RLAO_ARCFU	-	-	-	-	-	-	-	-	-	-	-	75
RLAO_METKA	-	-	-	-	-	-	-	-	-	-	-	88
RLAO_METTH	-	-	-	-	-	-	-	-	-	-	-	74
RLAO_METTL	-	-	-	-	-	-	-	-	-	-	-	82
RLAO_METVA	-	-	-	-	-	-	-	-	-	-	-	82
RLAO_METJA	-	-	-	-	-	-	-	-	-	-	-	81
RLAO_PYRAB	-	-	-	-	-	-	-	-	-	-	-	77
RLAO_PYRHO	-	-	-	-	-	-	-	-	-	-	-	77
RLAO_PYRFU	-	-	-	-	-	-	-	-	-	-	-	77
RLAO_PYRKO	-	-	-	-	-	-	-	-	-	-	-	76
RLAO_HALMA	-	-	-	-	-	-	-	-	-	-	-	79
RLAO_HALVO	-	-	-	-	-	-	-	-	-	-	-	79
RLAO_HALSA	-	-	-	-	-	-	-	-	-	-	-	79
RLAO_THEAC	-	-	-	-	-	-	-	-	-	-	-	72
RLAO_THEVO	-	-	-	-	-	-	-	-	-	-	-	72
RLAO_PICTO	-	-	-	-	-	-	-	-	-	-	-	72
ruler	1.....	10.....	20.....	30.....	40.....	50.....	60.....	70.....	80.....	90		

# Background: AlphaFold2 solutions?

## Embedding

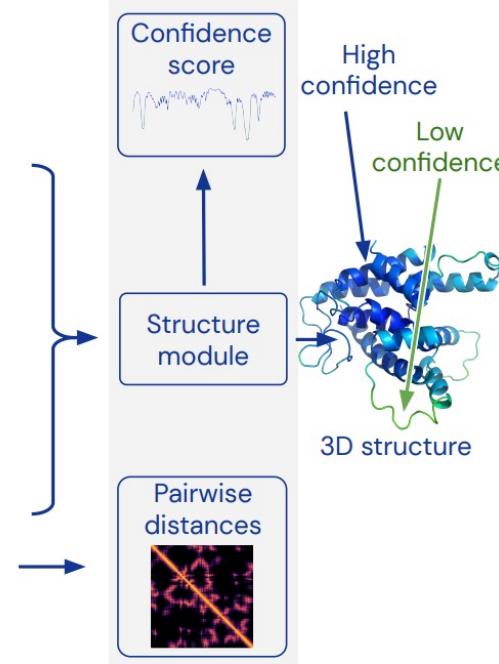


## Trunk



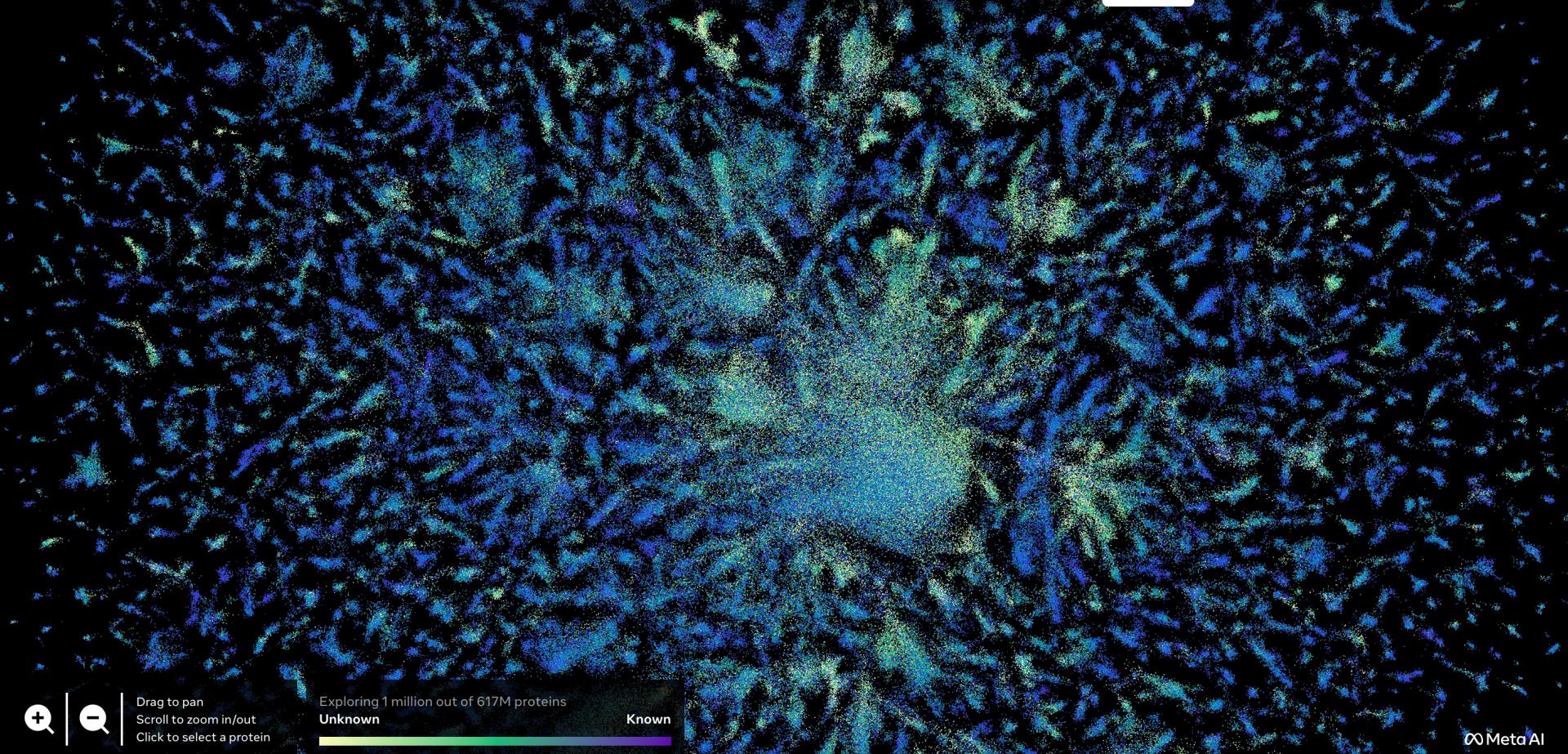
## Heads

© 2020 DeepMind Technologies Limited



Not single sequence structure prediction





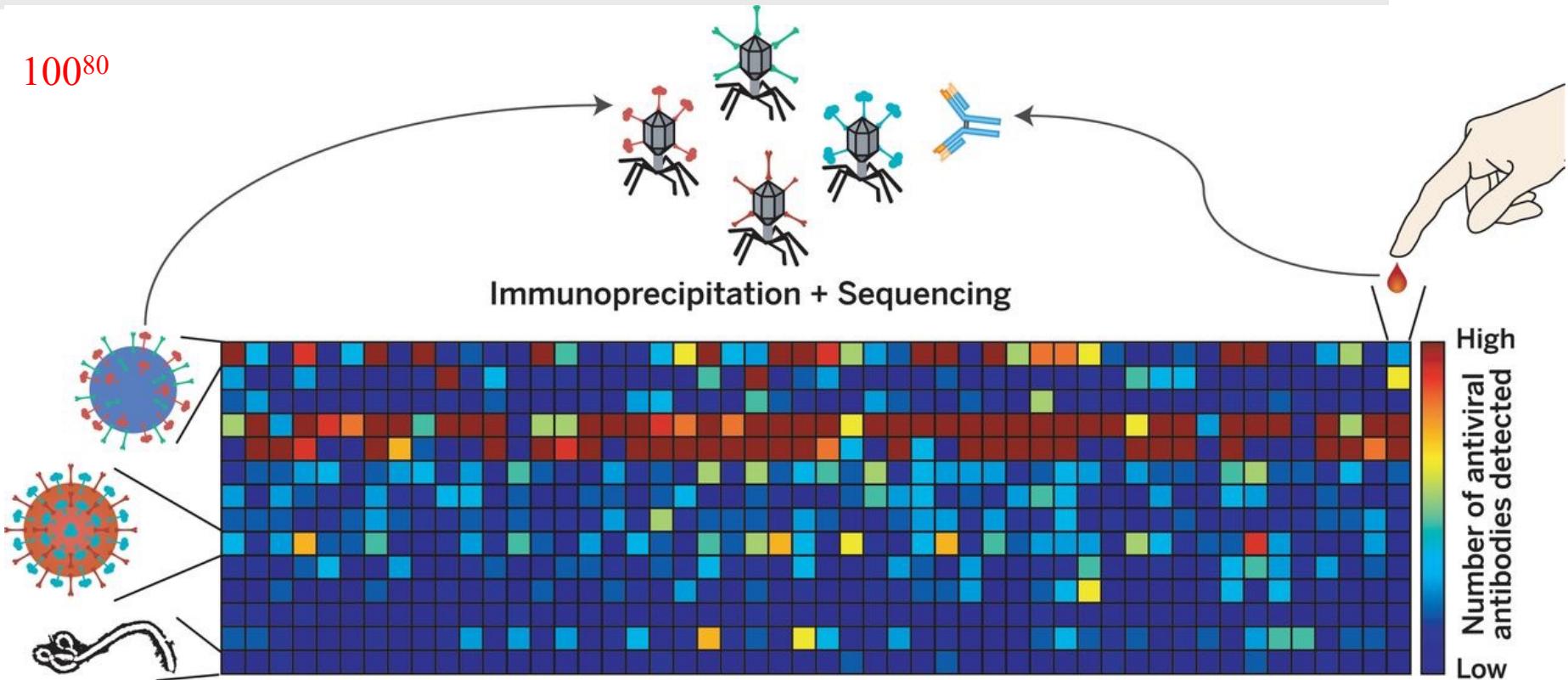
Drag to pan  
Scroll to zoom in/out  
Click to select a protein

Exploring 1 million out of 617M proteins  
Unknown

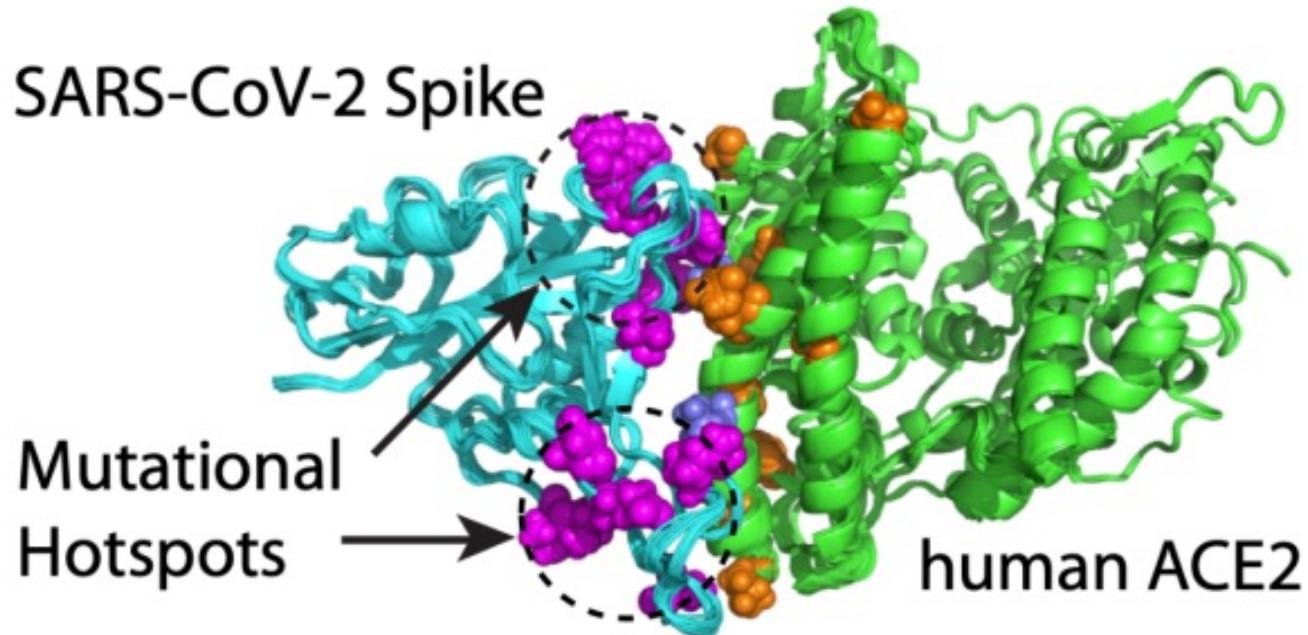
Known

MetaAI

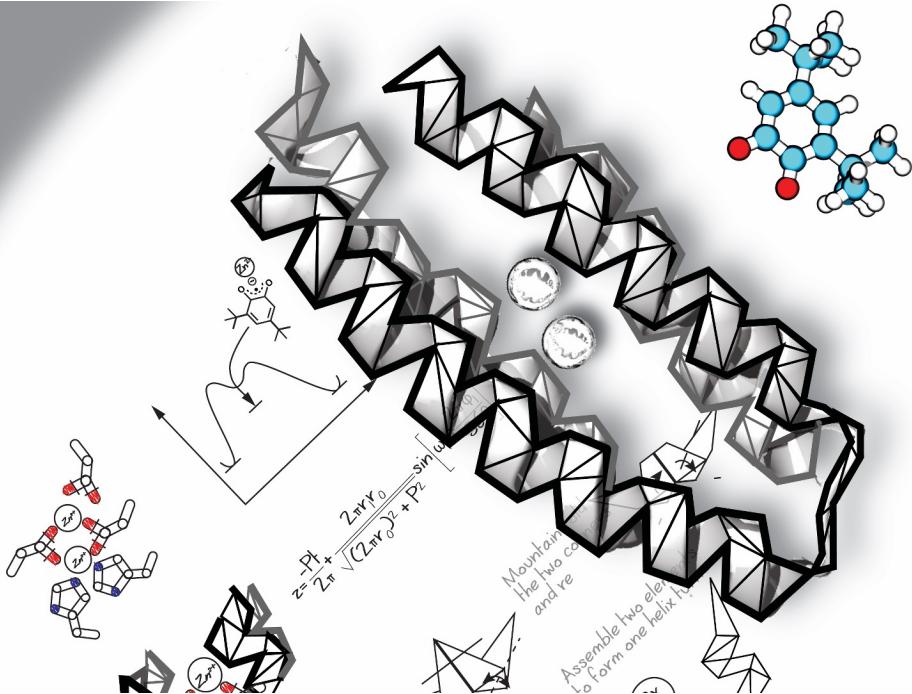
# Real proteins sequence space



# Traverse entire protein sequences space



# Solutions: Protein Language model (PLM)



How to search the entire sequence space combination?

How can we guide the model goes to our desired folds/functions?