group of machine
gmum
learning research

Update on Protein Folding

Tomasz Danel

Jagiellonian University

Agenda

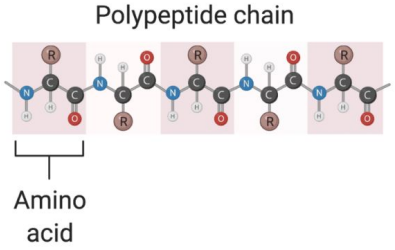
group of machine gmum learning research

1. Protein Structure and Function
2. Introduction to the Protein Folding Problem
3. Classical Approach to Protein Modeling (Homology)
4. Deep Learning Protein Folding Models
    a. AlphaFold
    b. RoseTTAFold
    c. EigenFold
    d. QFold

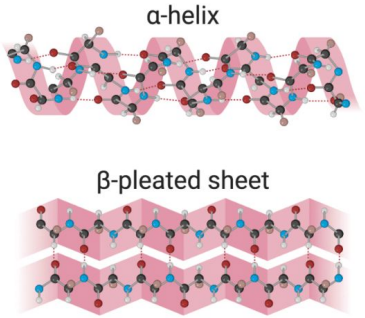
Protein Structure

group of machine learning research

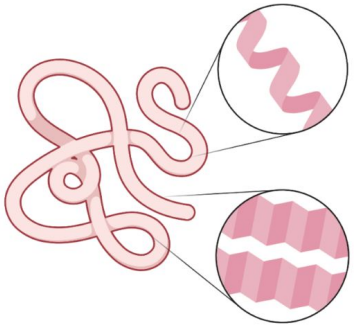
Primary structure



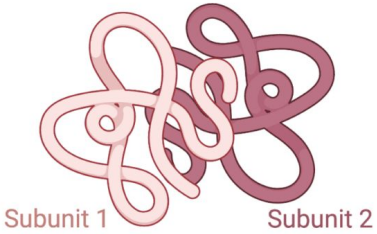
Secondary structure



Tertiary structure

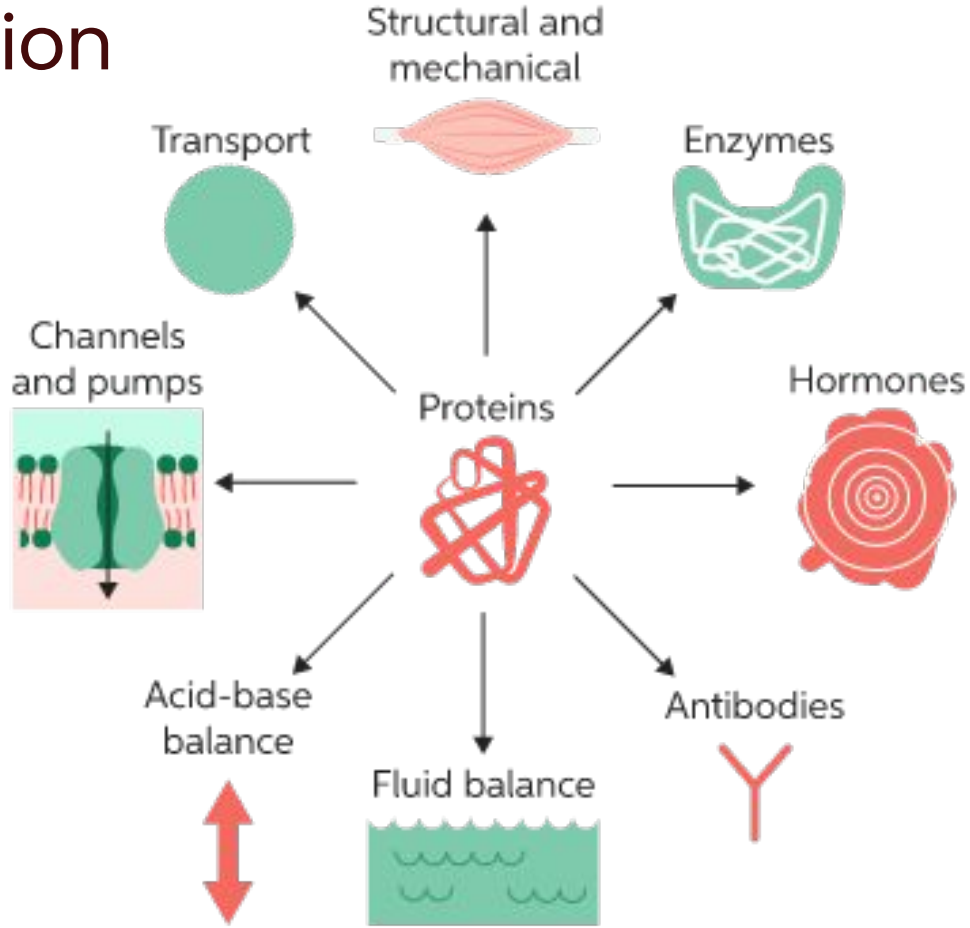


Quaternary structure

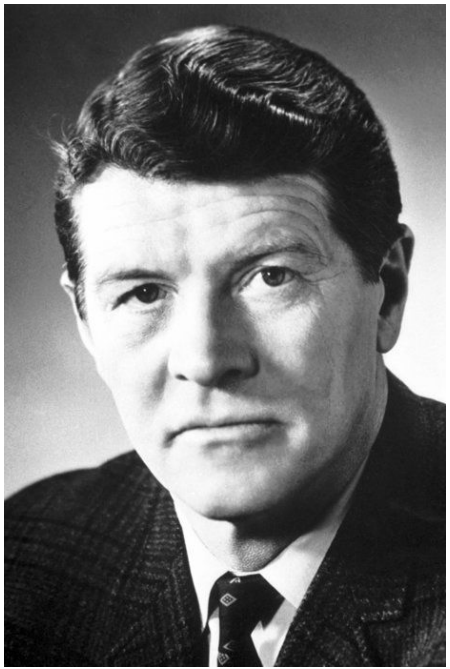


source: https://www.rapidnovor.com/

Protein Function



Genetic Code Defines Protein Structure

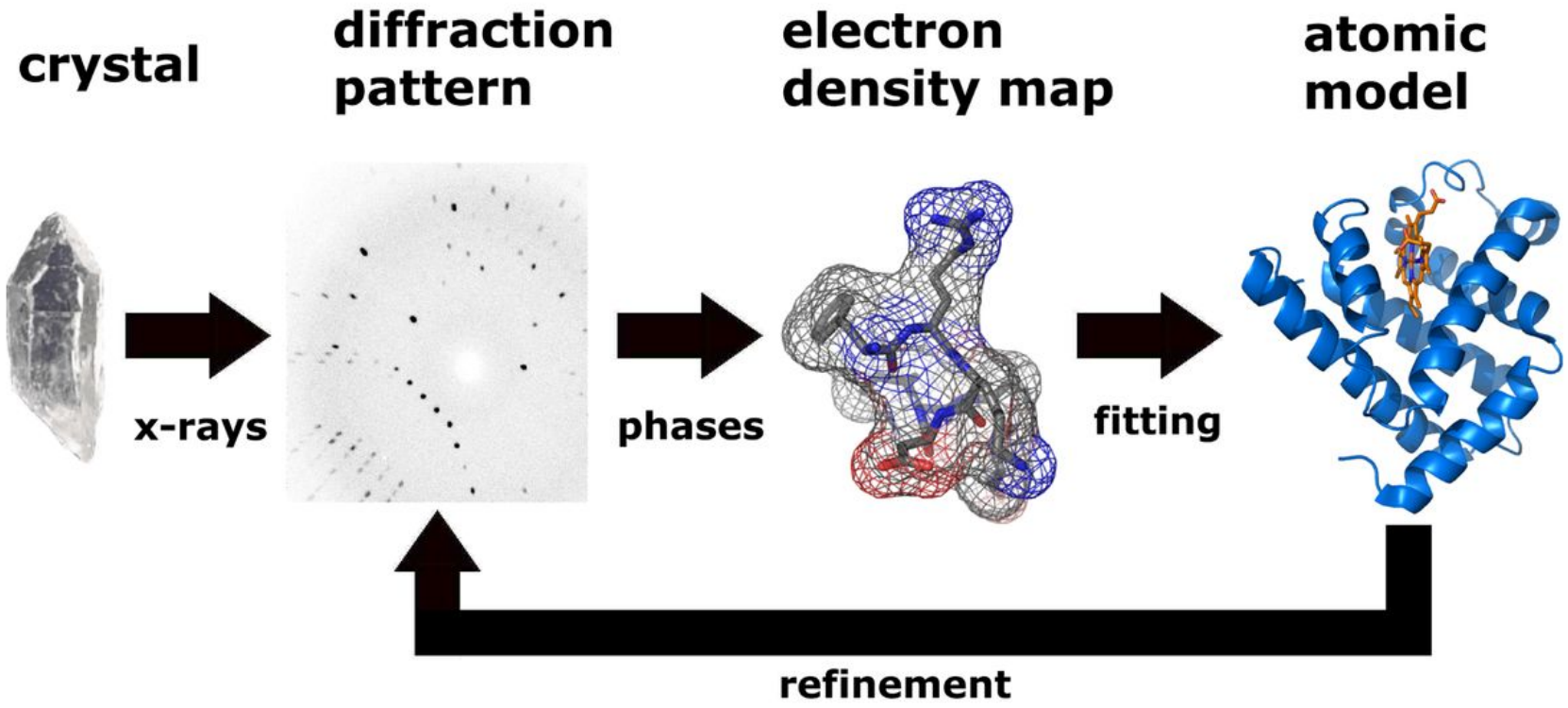


Christian B. Anfinsen

The Nobel Prize in Chemistry 1972

In 1961, Anfinsen proved that the sequence of amino acids, in itself, determines the way the chain folds itself and that no additional genetic information is required in this process.

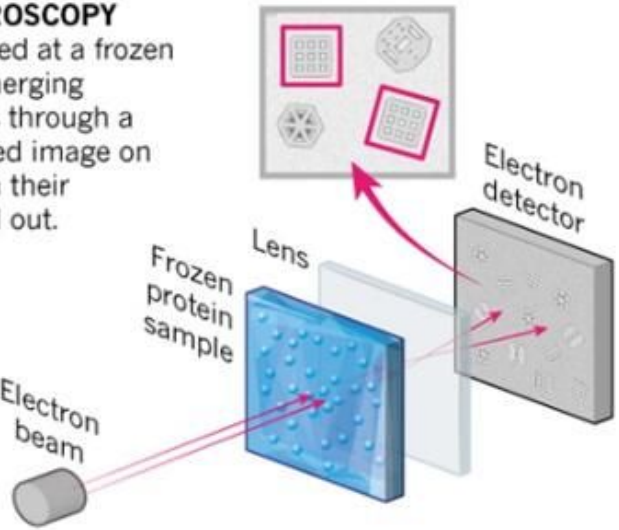
Crystallography Methods: X-Ray

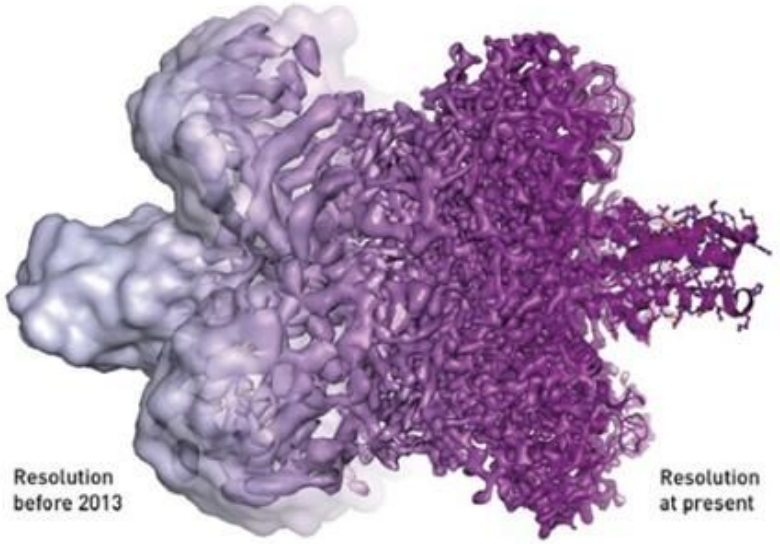


Crystallography Methods: X-Ray

group of machine gmum learning research

CRYO-ELECTRON MICROSCOPY
A beam of electron is fired at a frozen protein solution. The emerging scattered electrons pass through a lens to create a magnified image on the detector, from which their structure can be worked out.

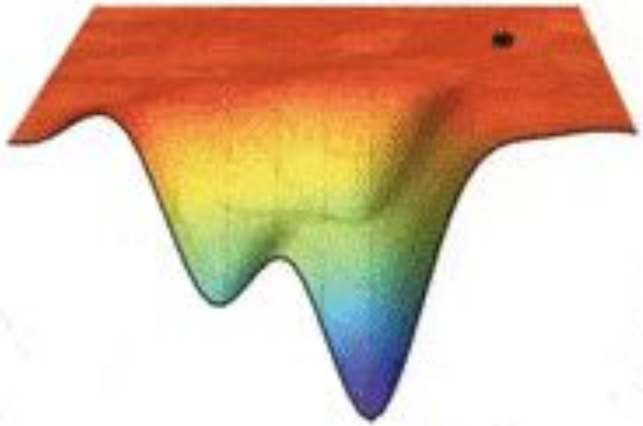




source: https://www.creative-biostructure.com/

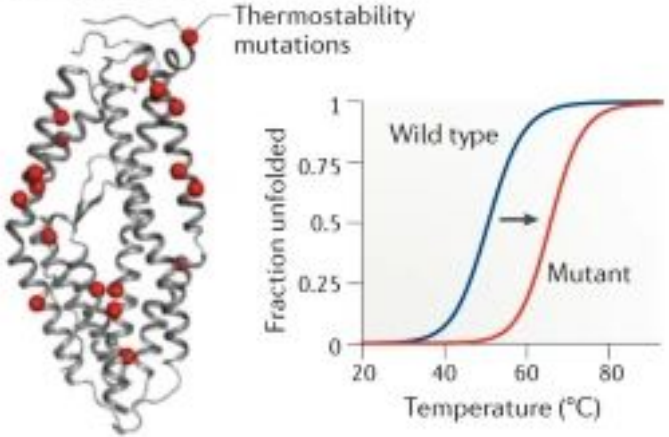
Protein Folding

group of machine
gmum
learning research

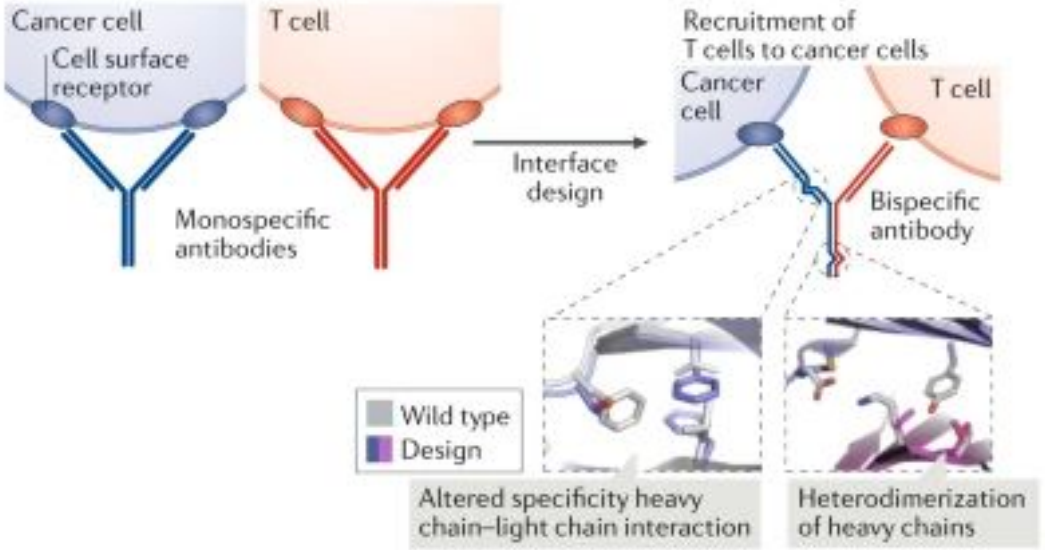


Inverse Protein Folding

a Stabilizing a malaria invasion protein for use as a vaccine antigen

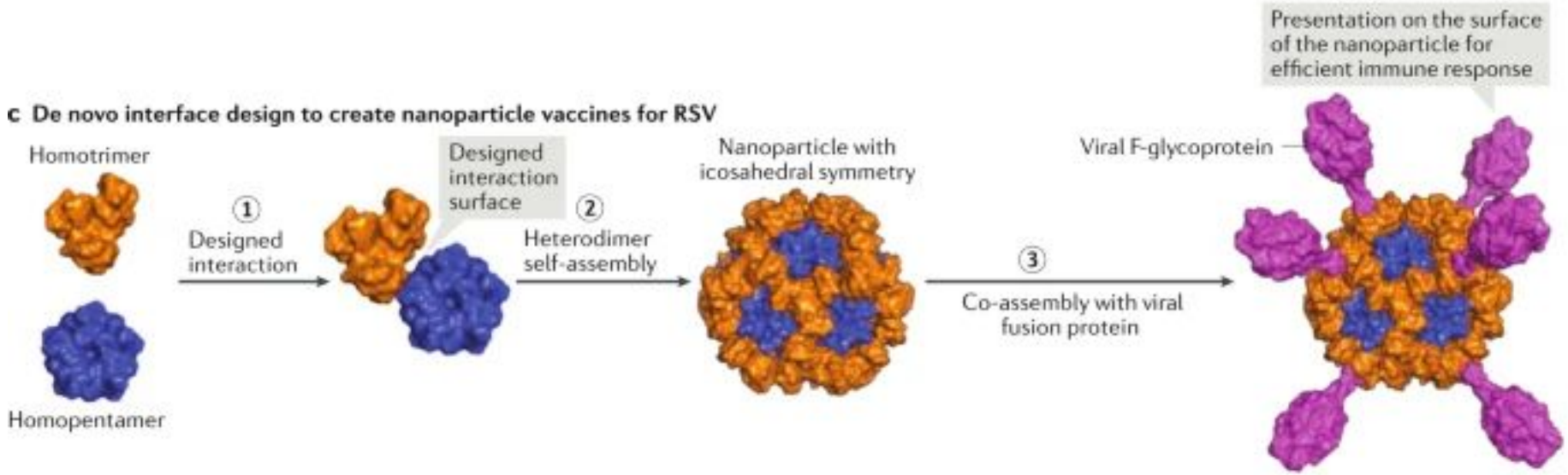


b Orthogonal interface design to generate bispecific antibodies for cancer immunotherapy



Inverse Protein Folding

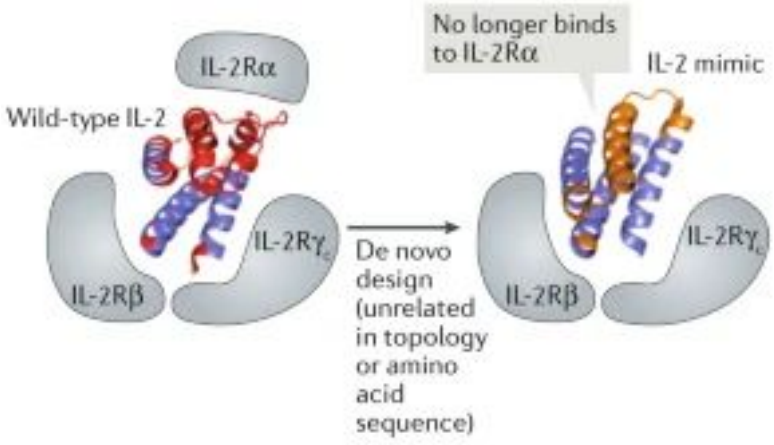
group of machine gmum learning research



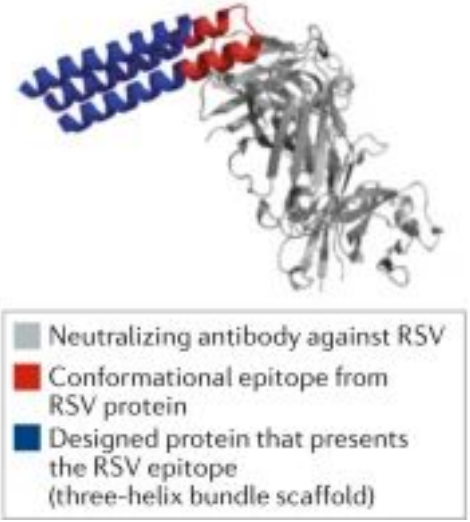
Inverse Protein Folding

group of machine gmum learning research

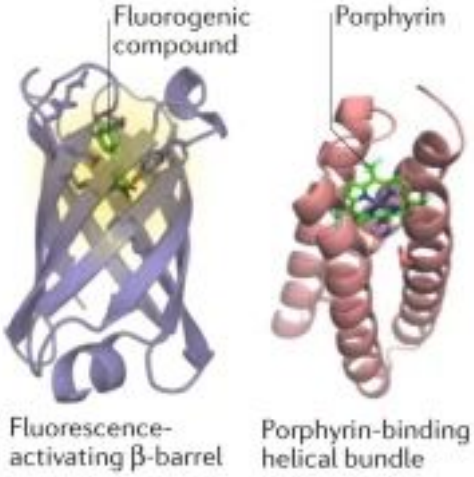
d De novo-designed mimic of IL-2 with receptor specificity altered to reduce toxicity in anti-cancer treatment



e Scaffolding of RSV epitope for vaccine design

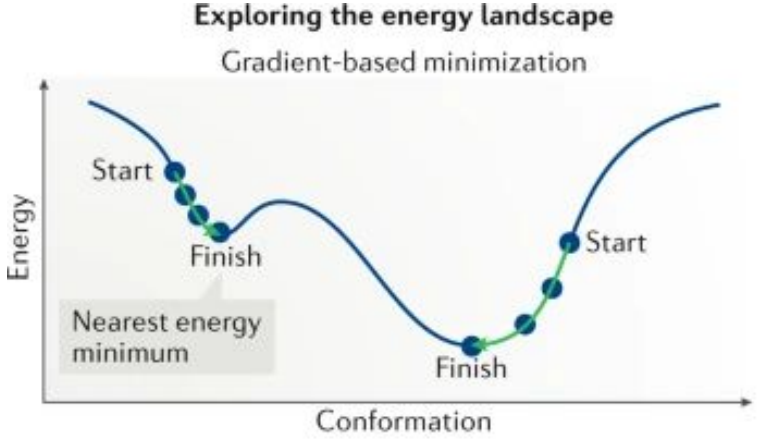


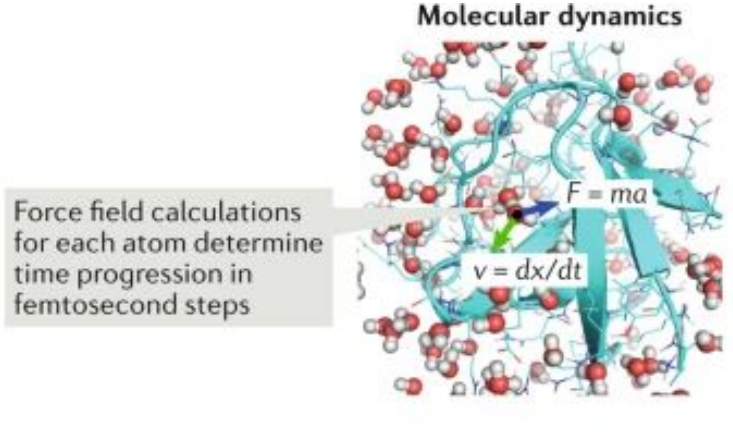
f De novo design of ligand-binding proteins

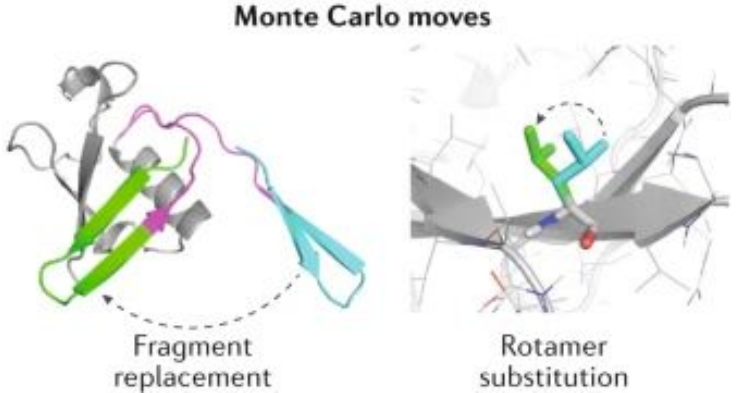


Energy Minimization

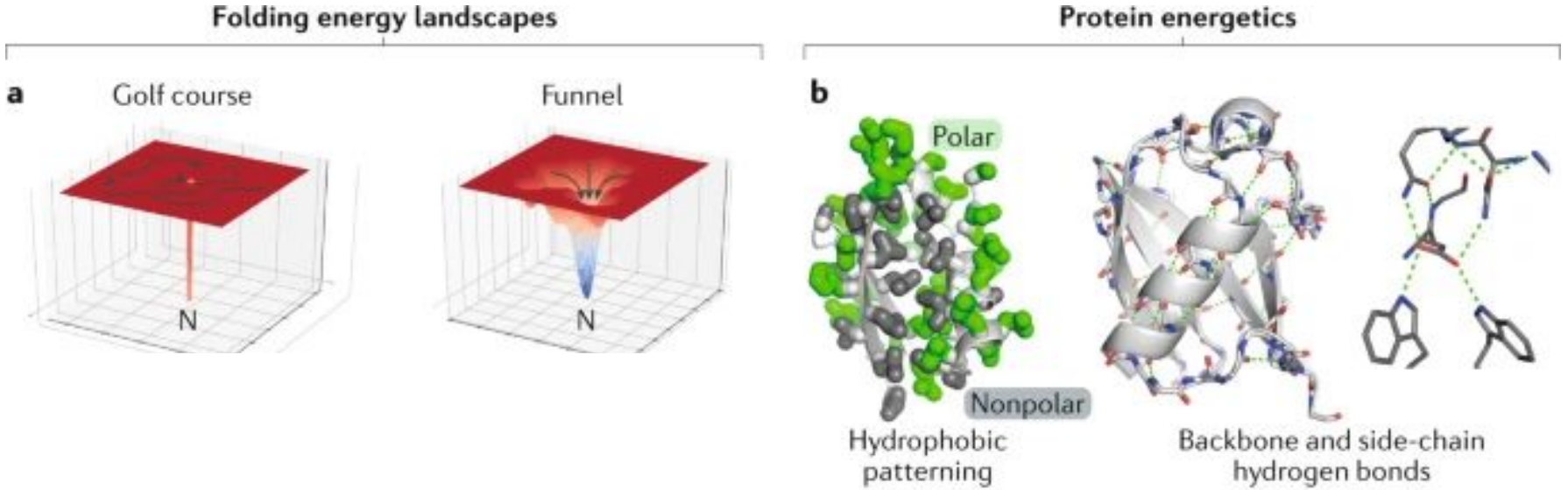
group of machine gmum learning research





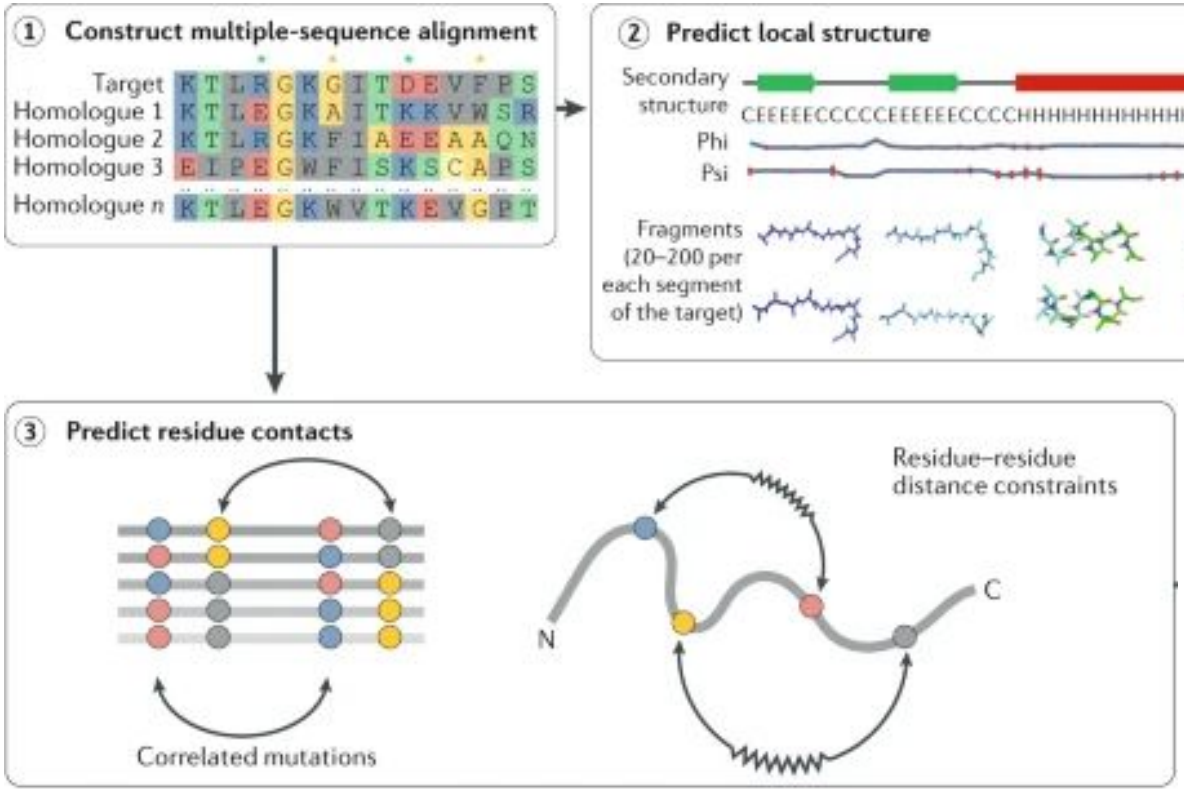


Energy Minimization



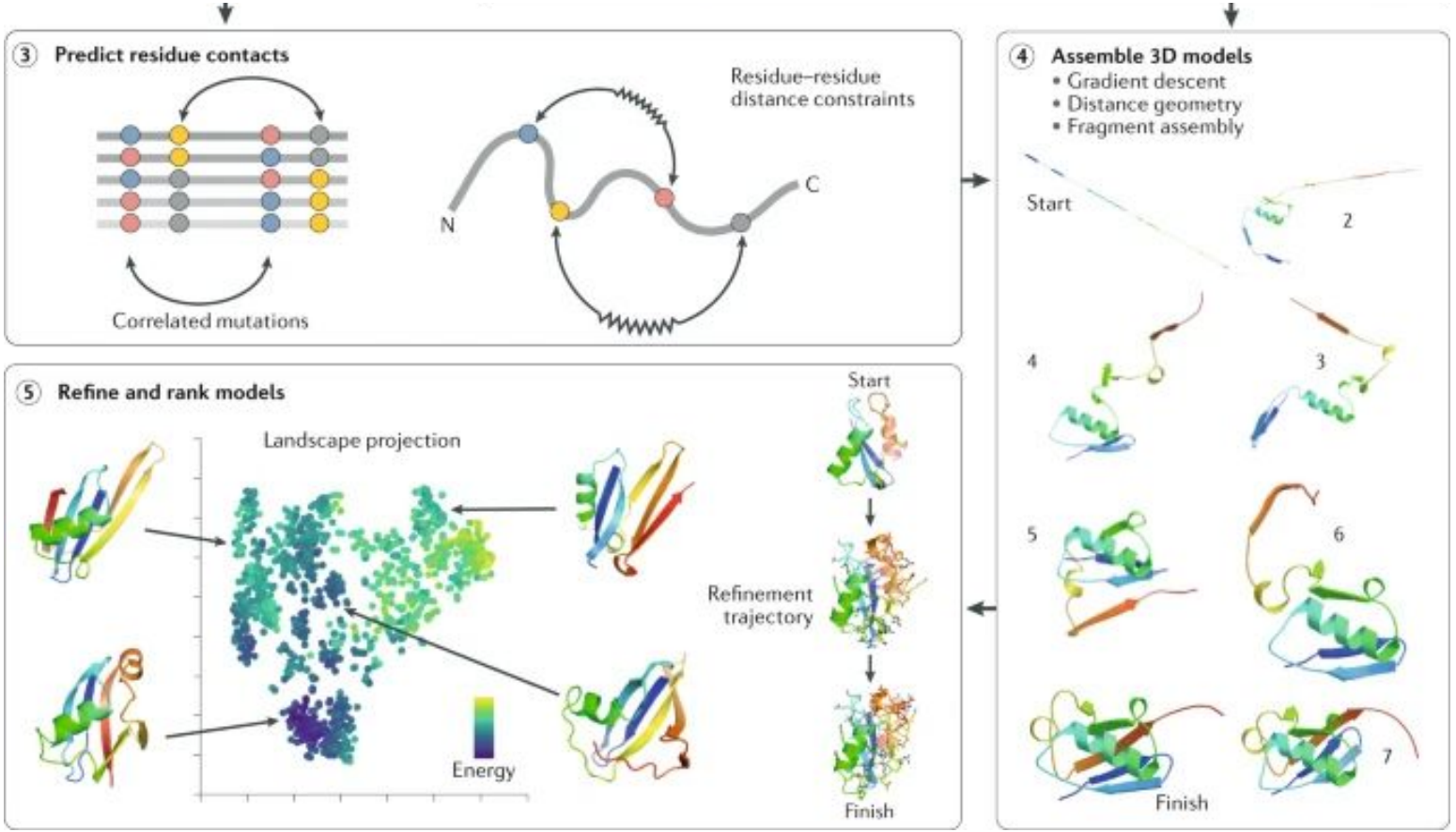
Homology Modeling

group of machine gmum learning research



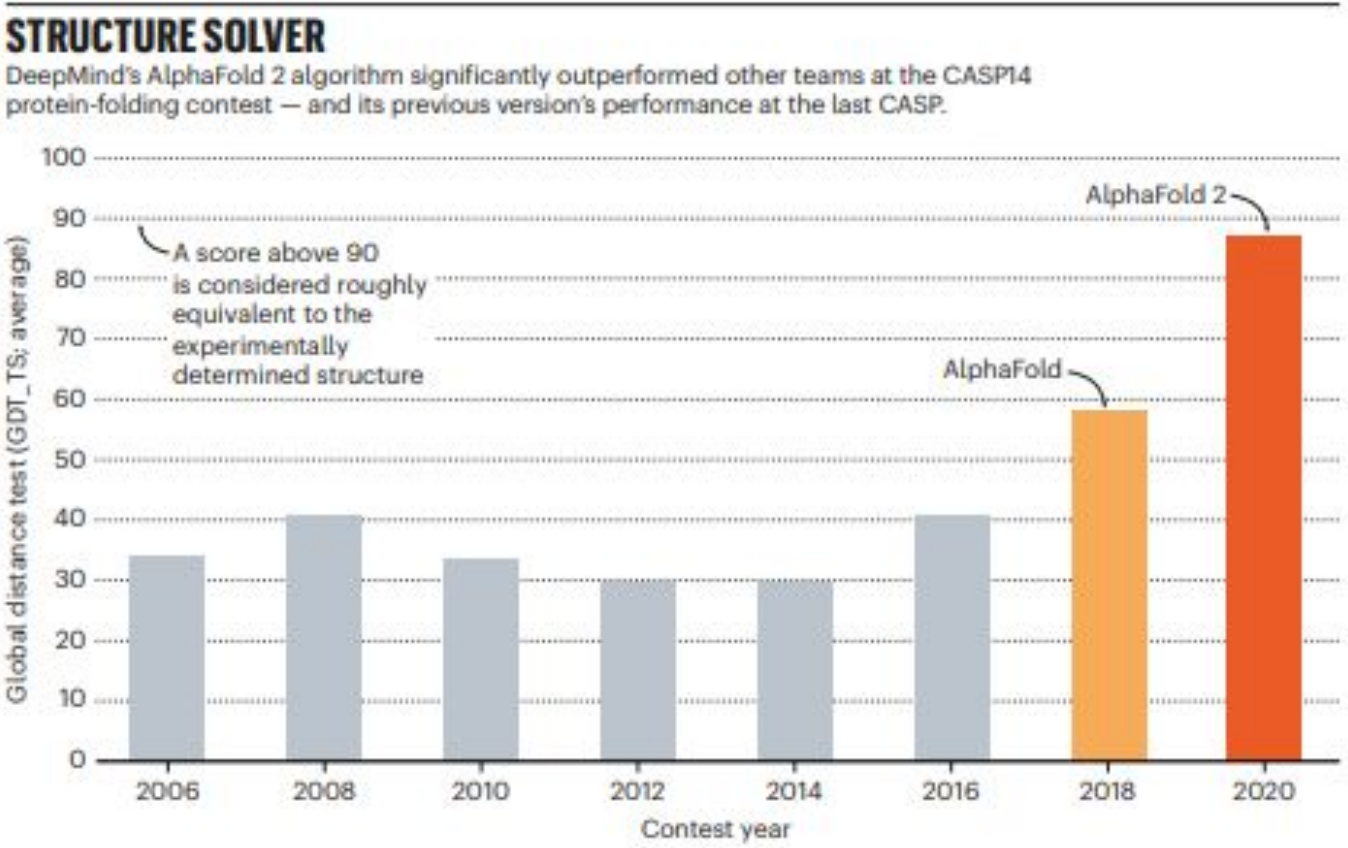
Homology Modeling

group of machine gmum learning research



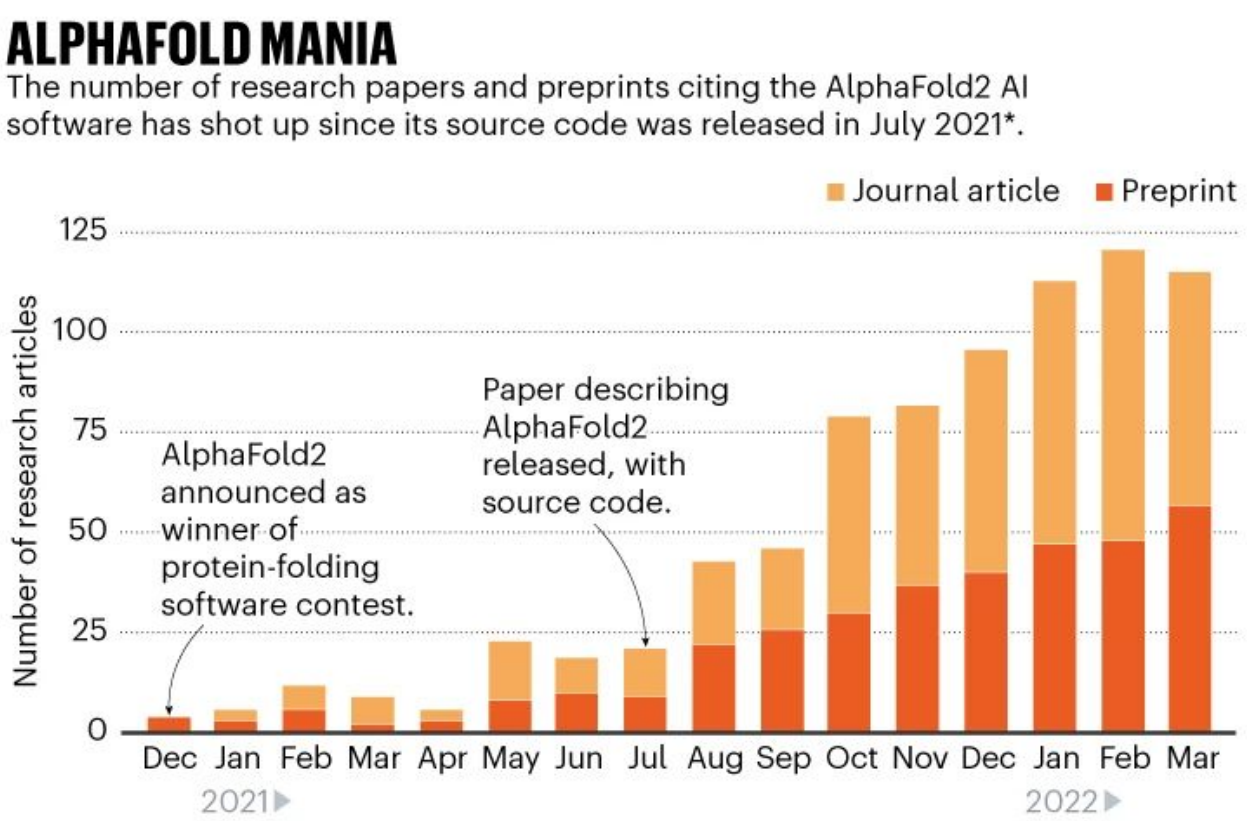
Protein Folding Challenge

group of machine gmum learning research



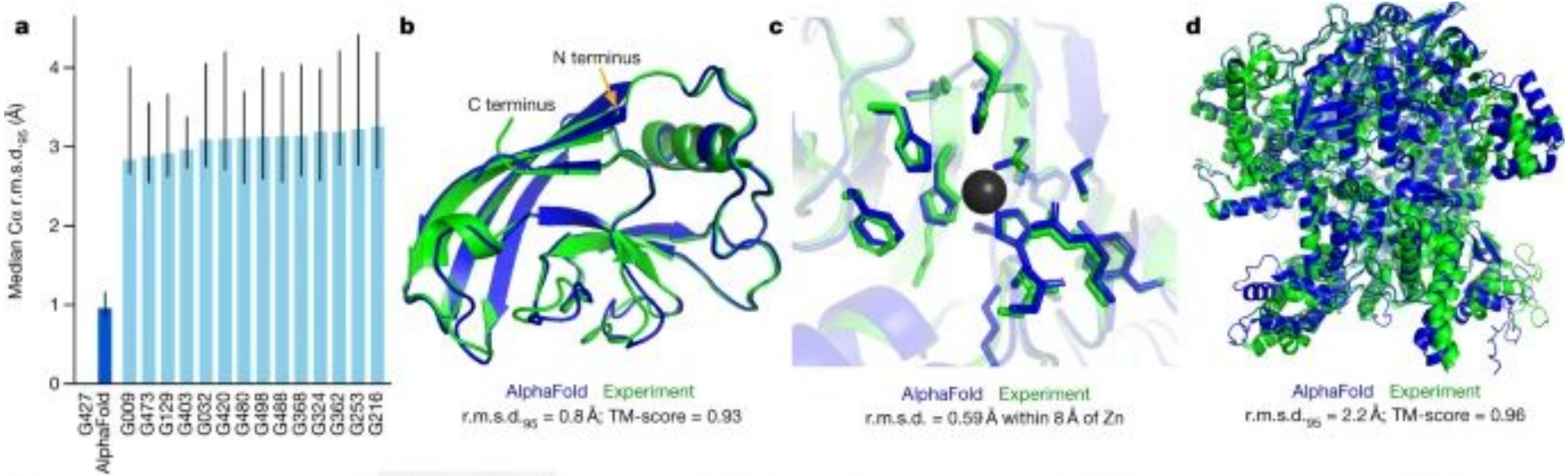
AlphaFold – Implications

group of machine gmum learning research



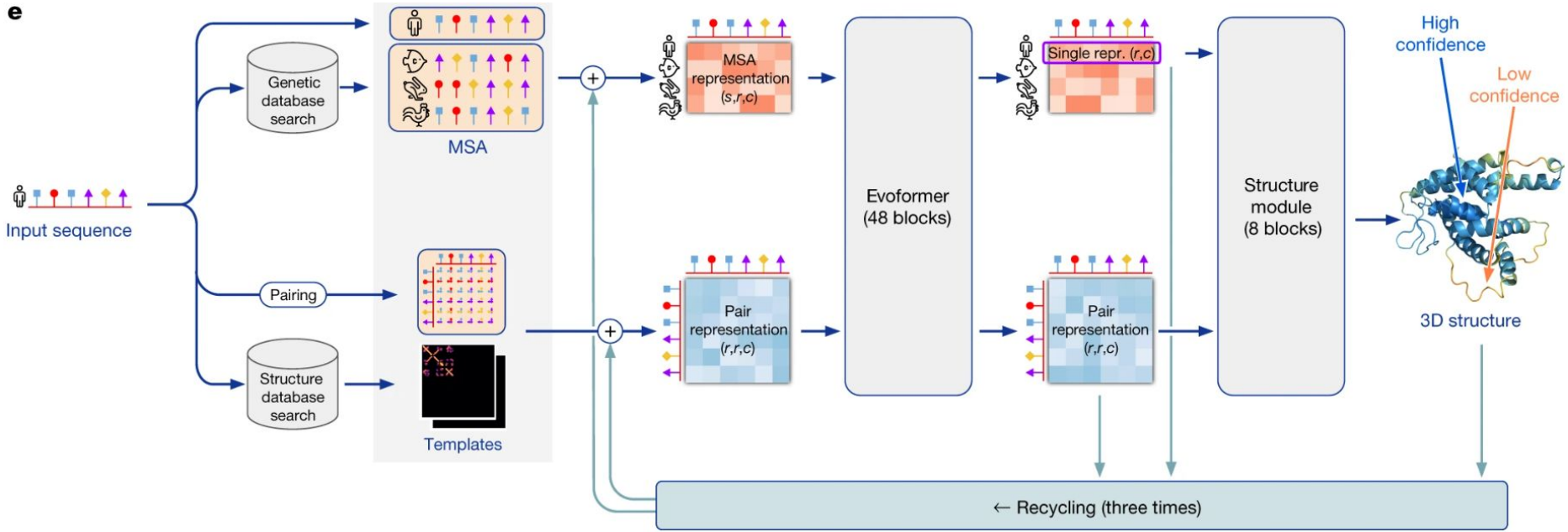
AlphaFold: CASP14 Winner

group of machine gmum learning research



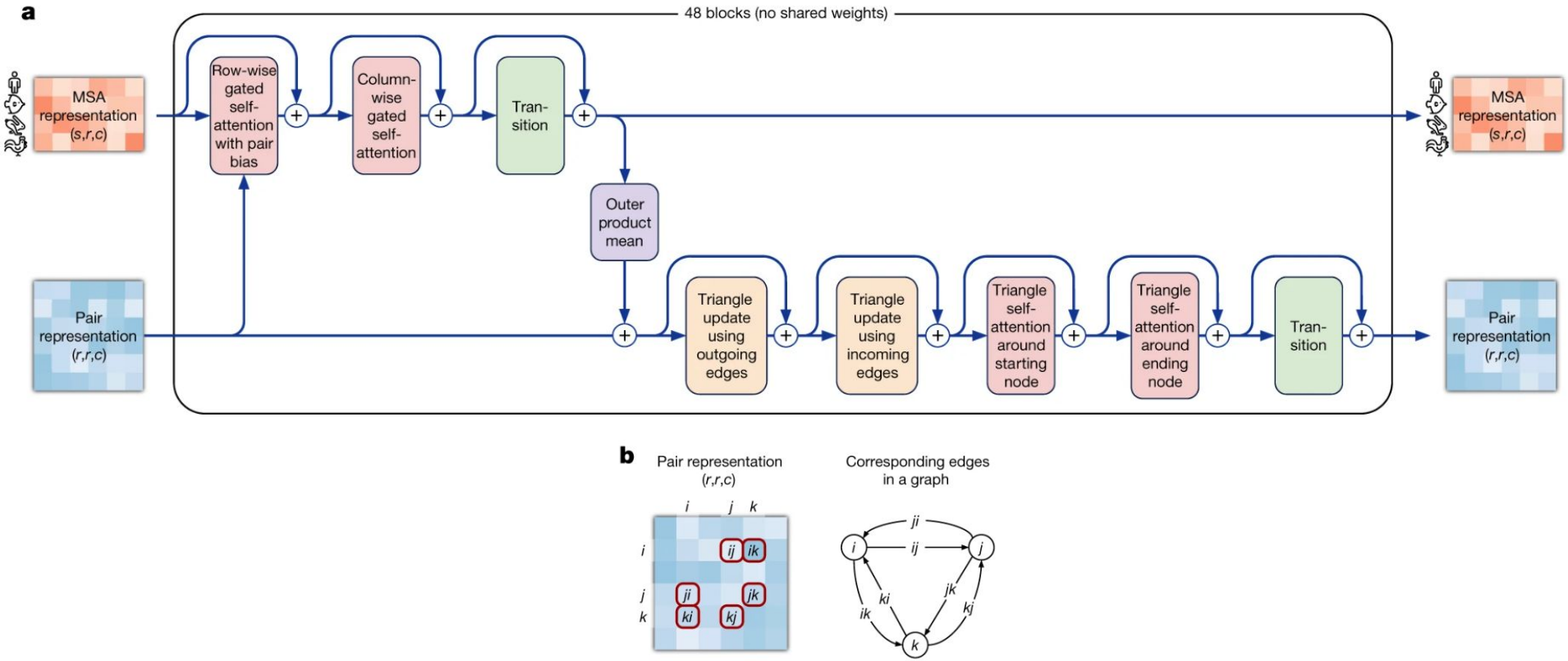
AlphaFold: Architecture

group of machine gmum learning research



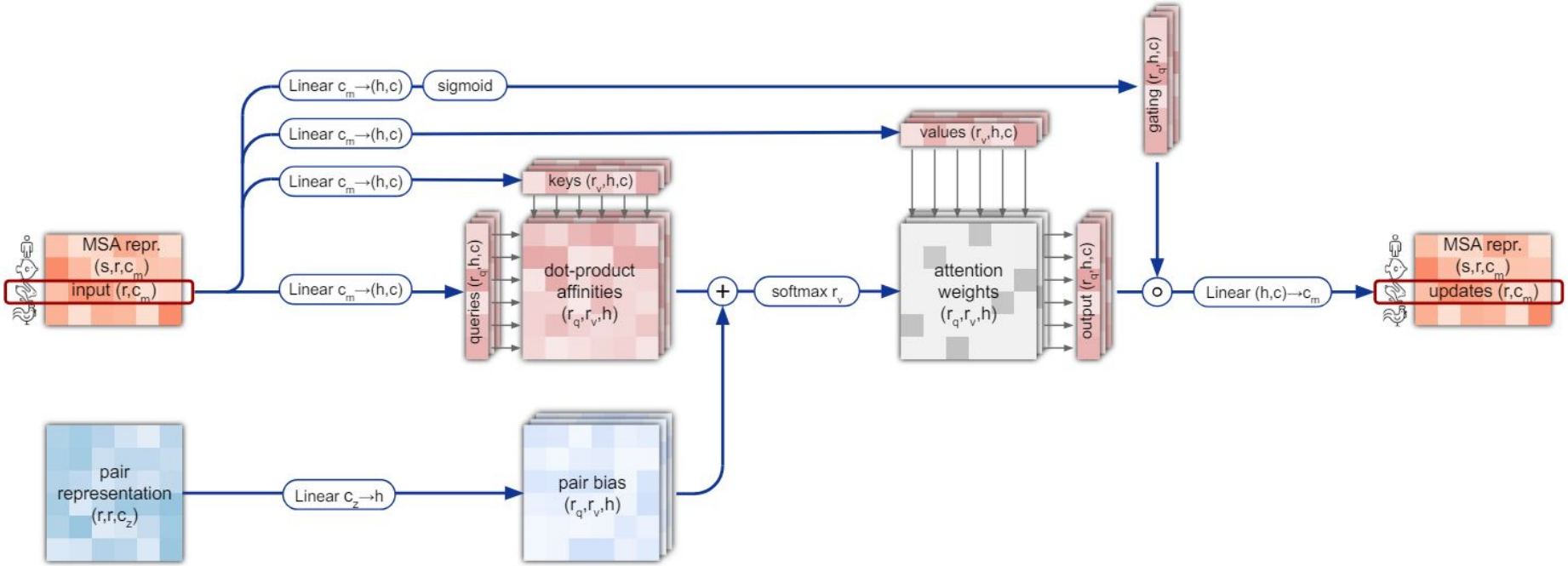
AlphaFold: Evoformer Block

group of machine gmum learning research

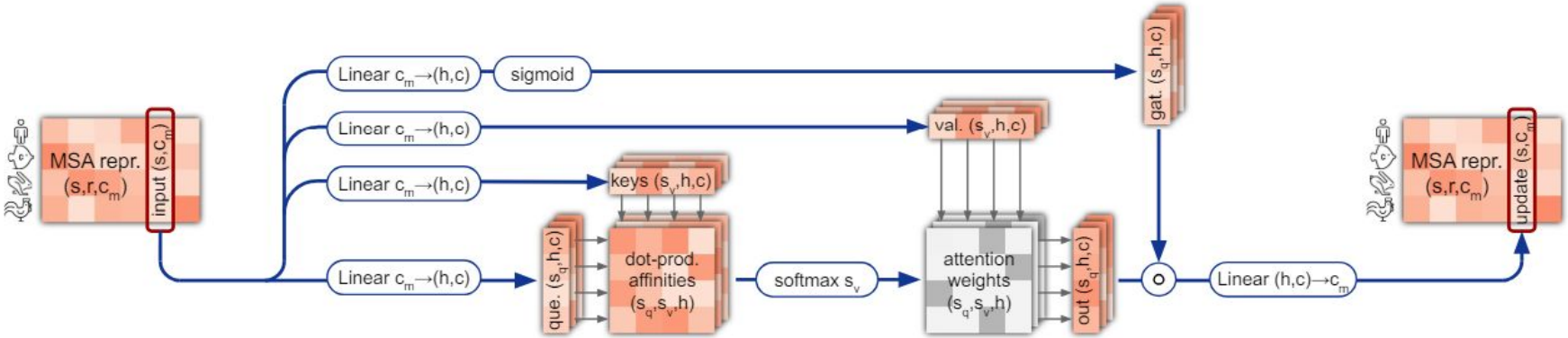


AlphaFold: Row-Wise Gated SA

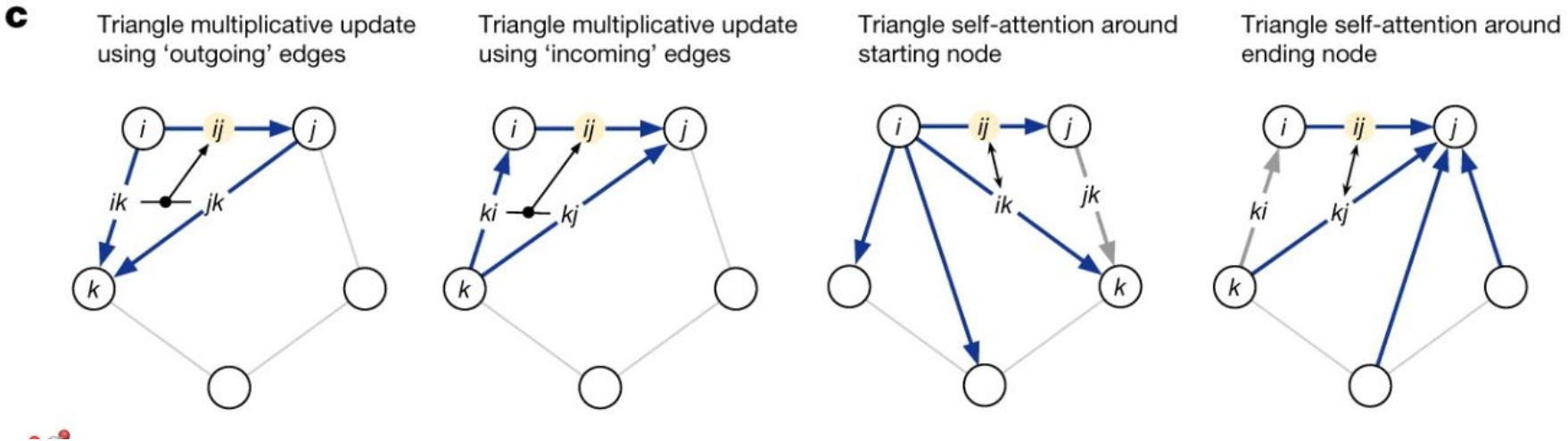
group of machine gmum learning research



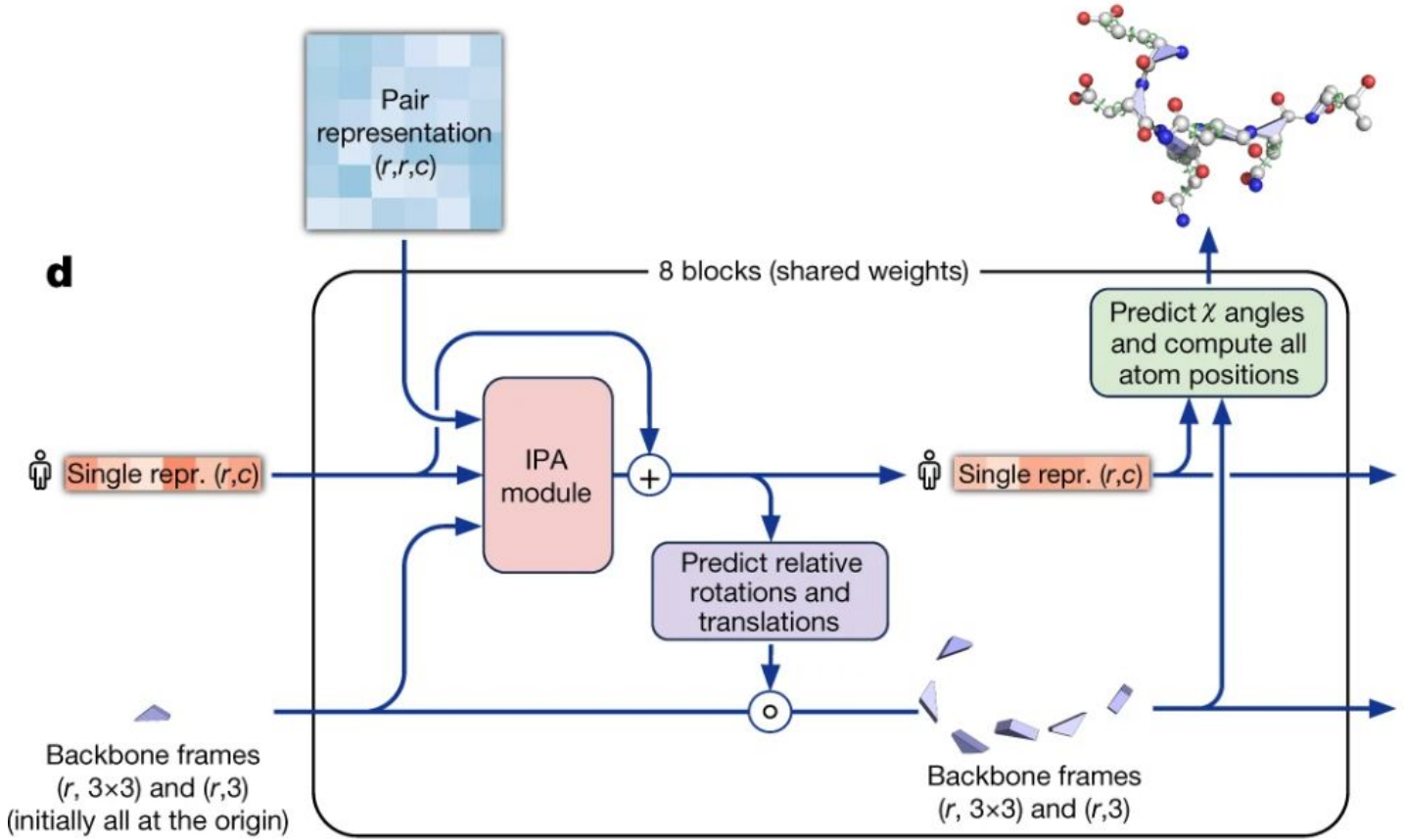
AlphaFold: Column-Wise Gated SA



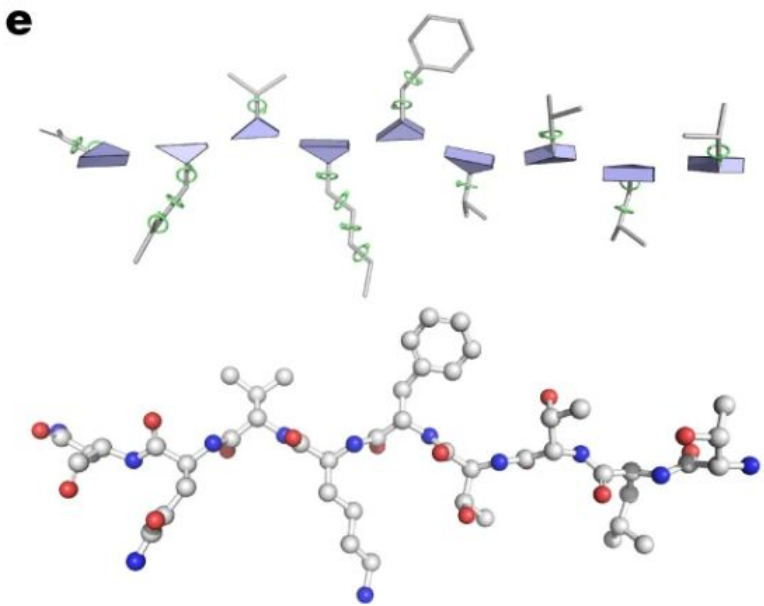
AlphaFold: Triangle Layers

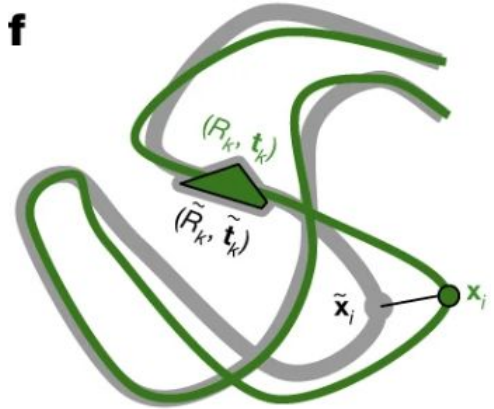


AlphaFold: Structure Module



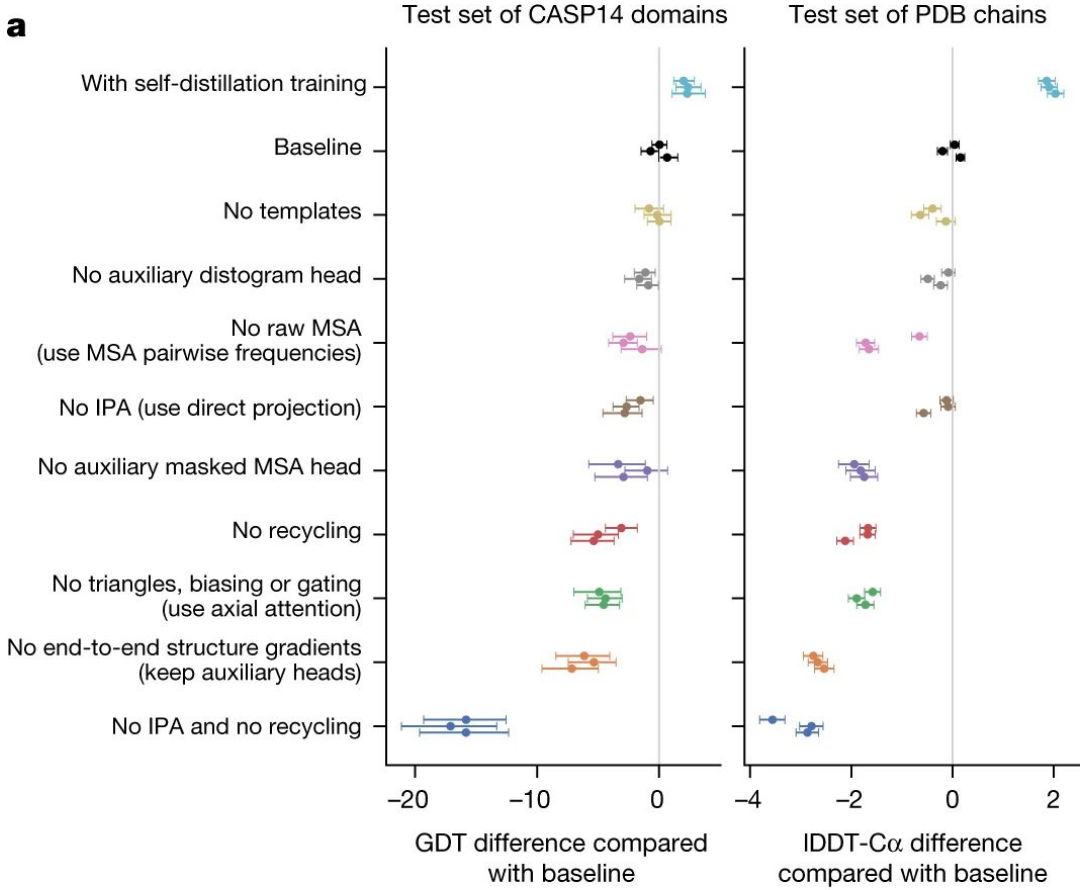
AlphaFold: Frame Aligned Point Error





AlphaFold: Self-Distillation

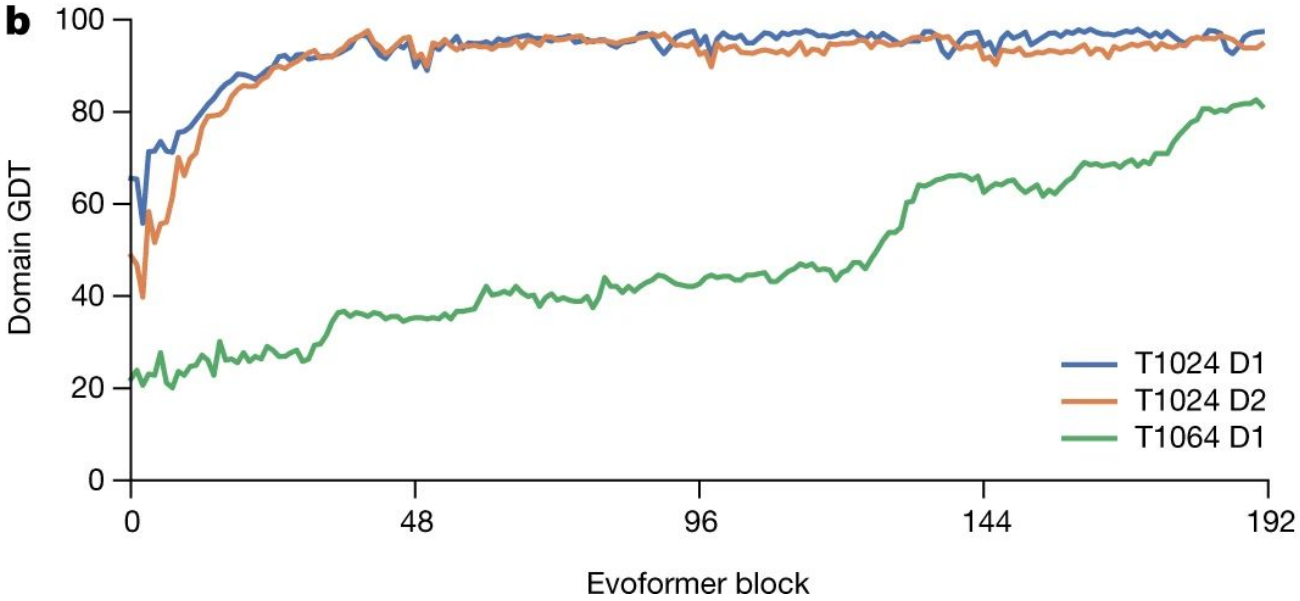
group of machine gmum learning research

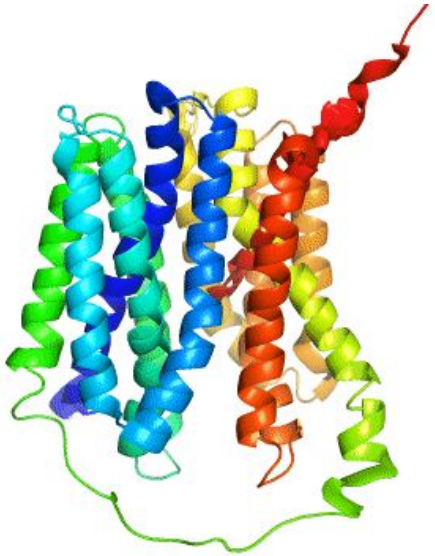


- 350,000 diverse sequences from Uniclust30
- predicted structures filtered to a high-confidence subset
- various training data augmentations such as cropping and MSA subsampling
- randomly mask out or mutate individual residues within the MSA

AlphaFold: Interpretation

group of machine gmum learning research

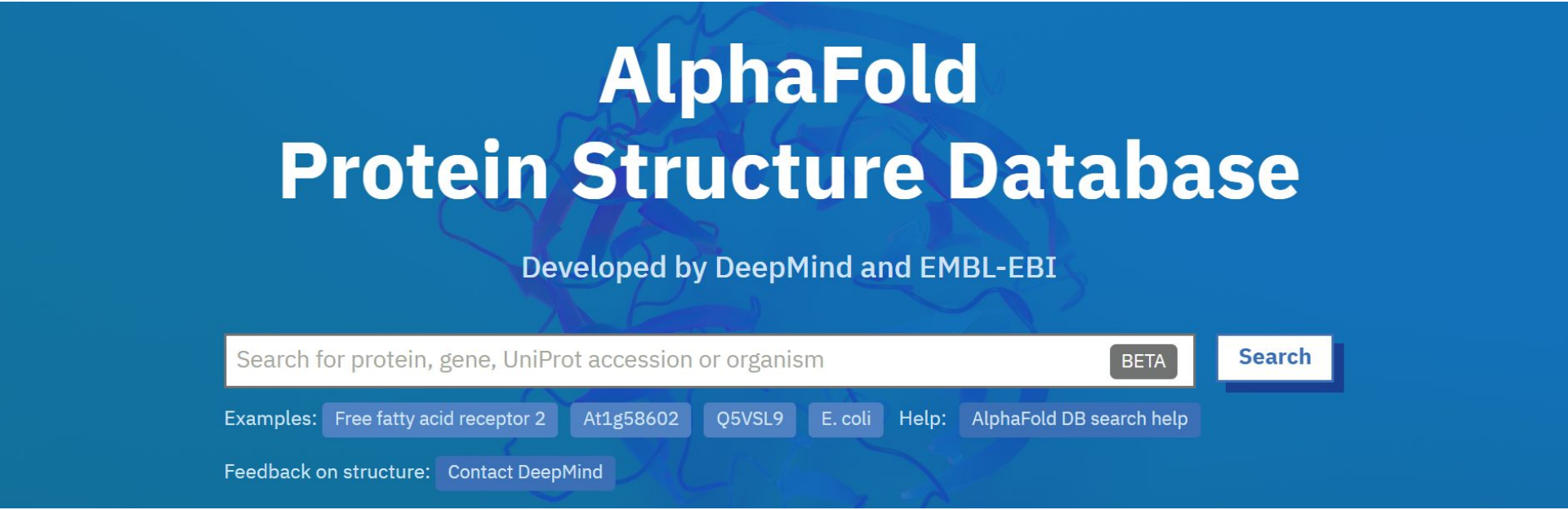




1

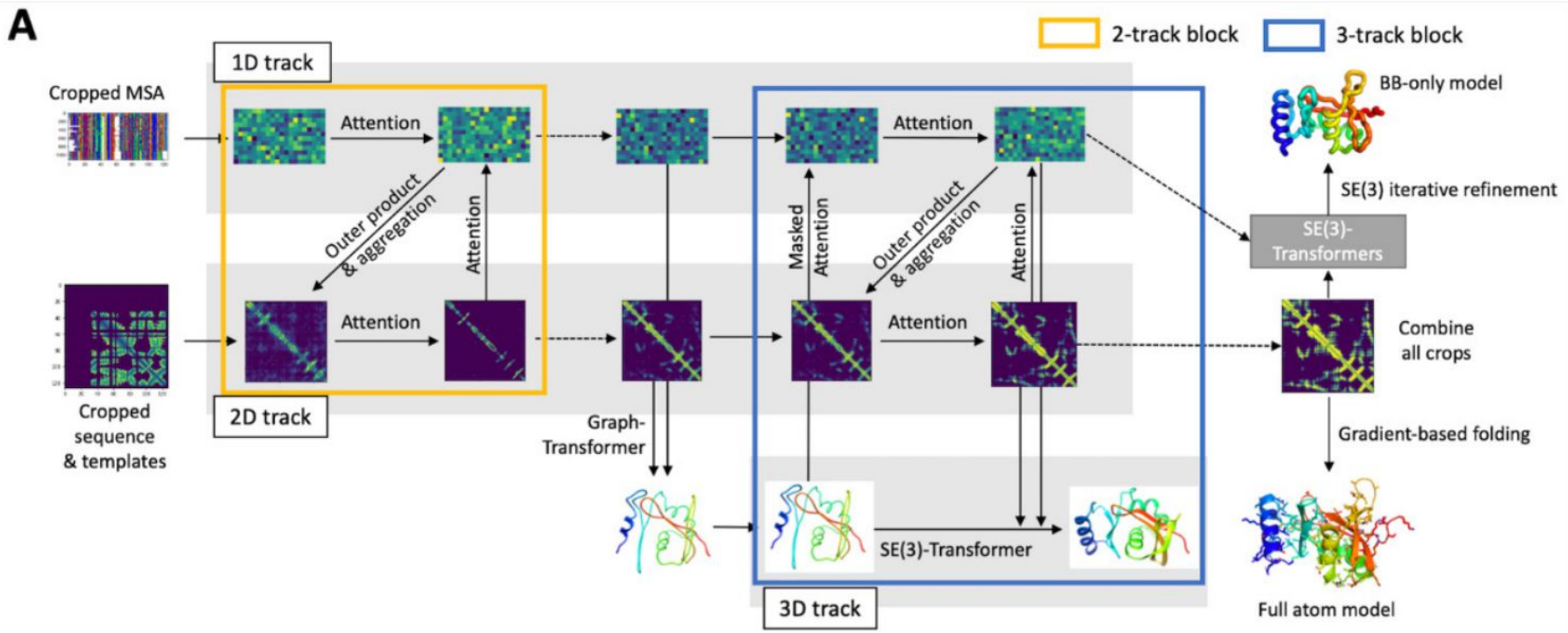
AlphaFold Protein Structure Database

group of machine gmum learning research

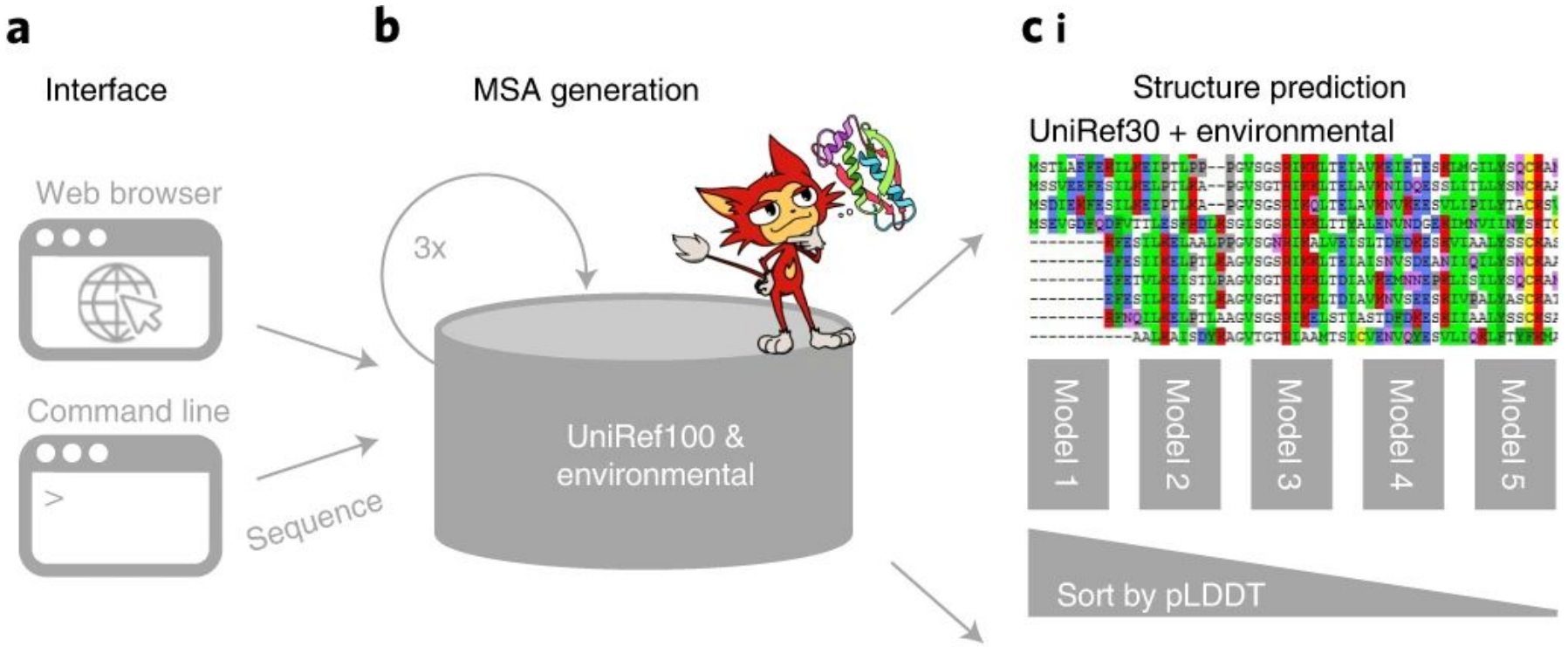


RoseTTAFold

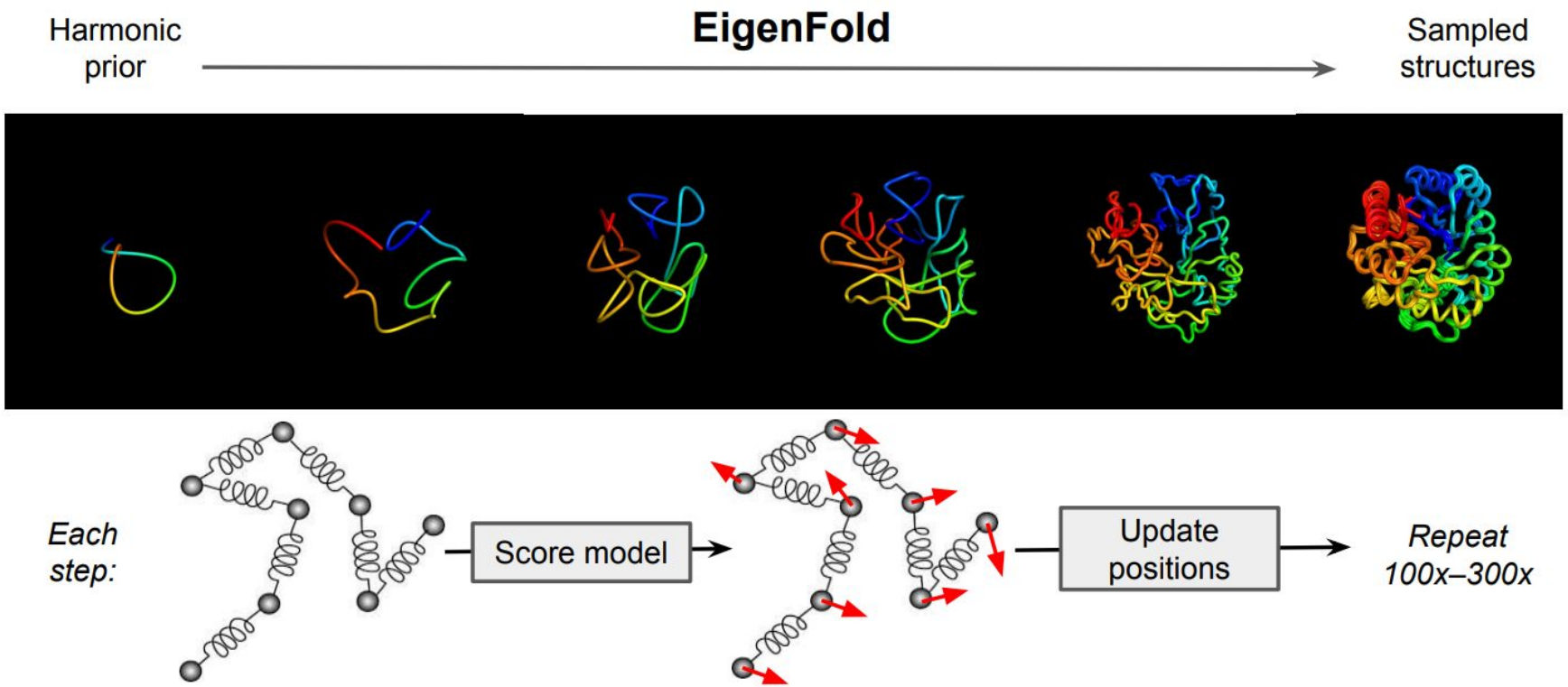
group of machine
gmum
learning research



ColabFold

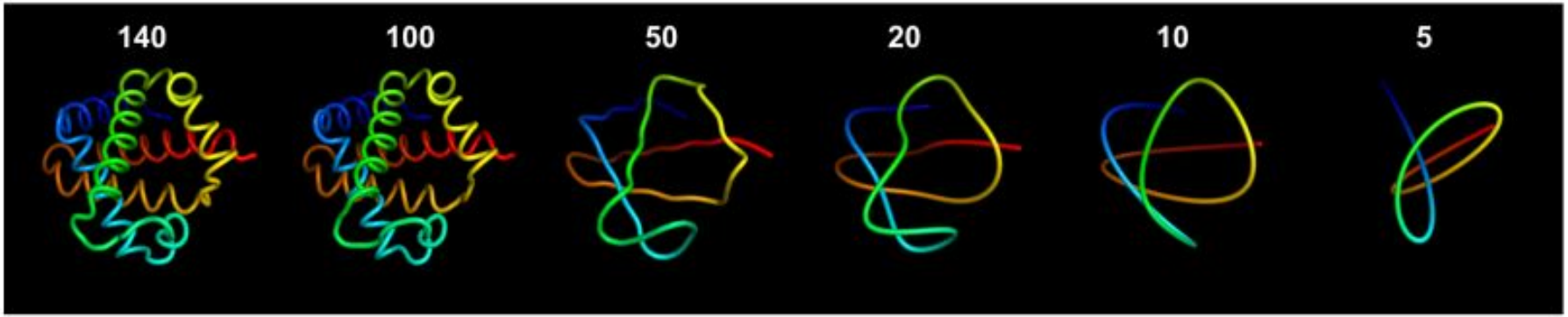

a
Interface
Web browser
Command line
>
Sequence
b
MSA generation
3x
UniRef100 &
environmental
c i
Structure prediction
UniRef30 + environmental
Model 1
Model 2
Model 3
Model 4
Model 5
Sort by pLDDT

EigenFold: Diffusion Protein Folding



EigenFold: Harmonic Diffusion

group of machine gmum learning research

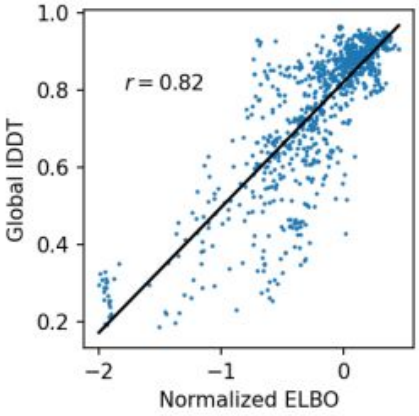


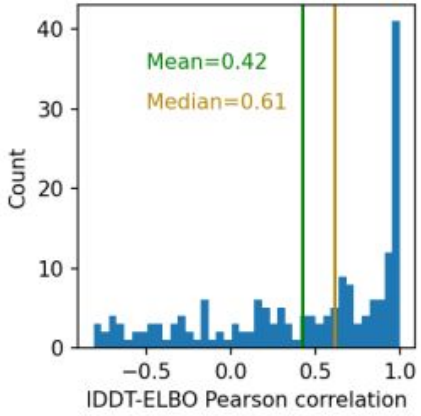
EigenFold: Results

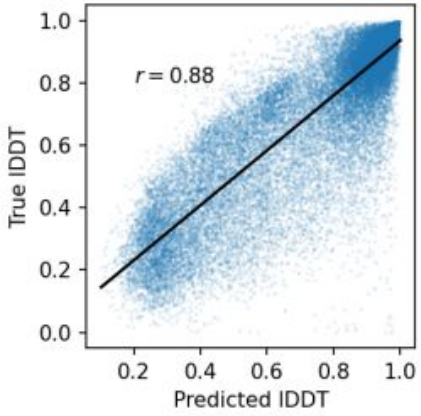
group of machine gmum learning research

Table 1: Single-structure prediction accuracy of EIGENFOLD and baseline methods on CAMEO targets under 750 residues from Aug 1–Oct 31, 2022. All metrics are reported as mean / median.

	$RMSD_{C\alpha}$ ↓	TMScore ↑	GDT-TS ↑	$lDDT_{C\alpha}$ ↑
ALPHAFOLD2	3.30 / 1.64	0.87 / 0.95	0.86 / 0.91	0.90 / 0.93
ESMFOLD	3.99 / 2.03	0.85 / 0.93	0.83 / 0.88	0.87 / 0.90
OMEGAFOLD	5.26 / 2.62	0.80 / 0.89	0.77 / 0.84	0.83 / 0.89
ROSETTAFOLD	5.72 / 3.17	0.77 / 0.84	0.71 / 0.75	0.79 / 0.82
EIGENFOLD	7.37 / 3.50	0.75 / 0.84	0.71 / 0.79	0.78 / 0.85

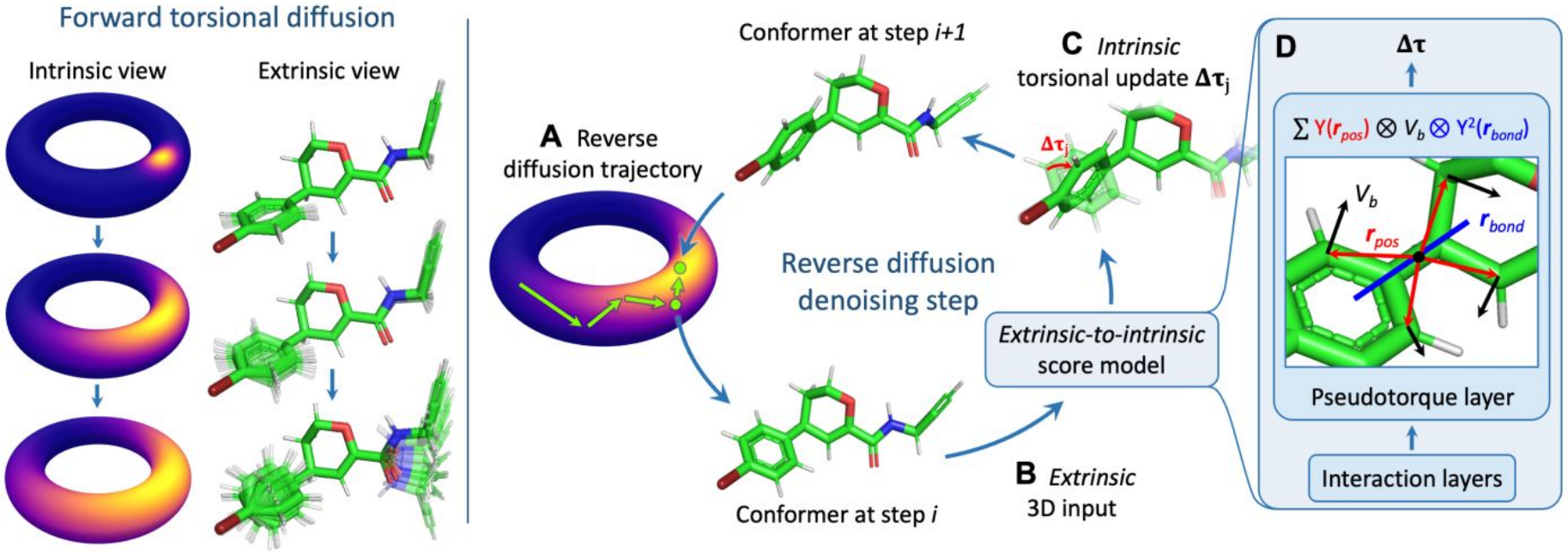






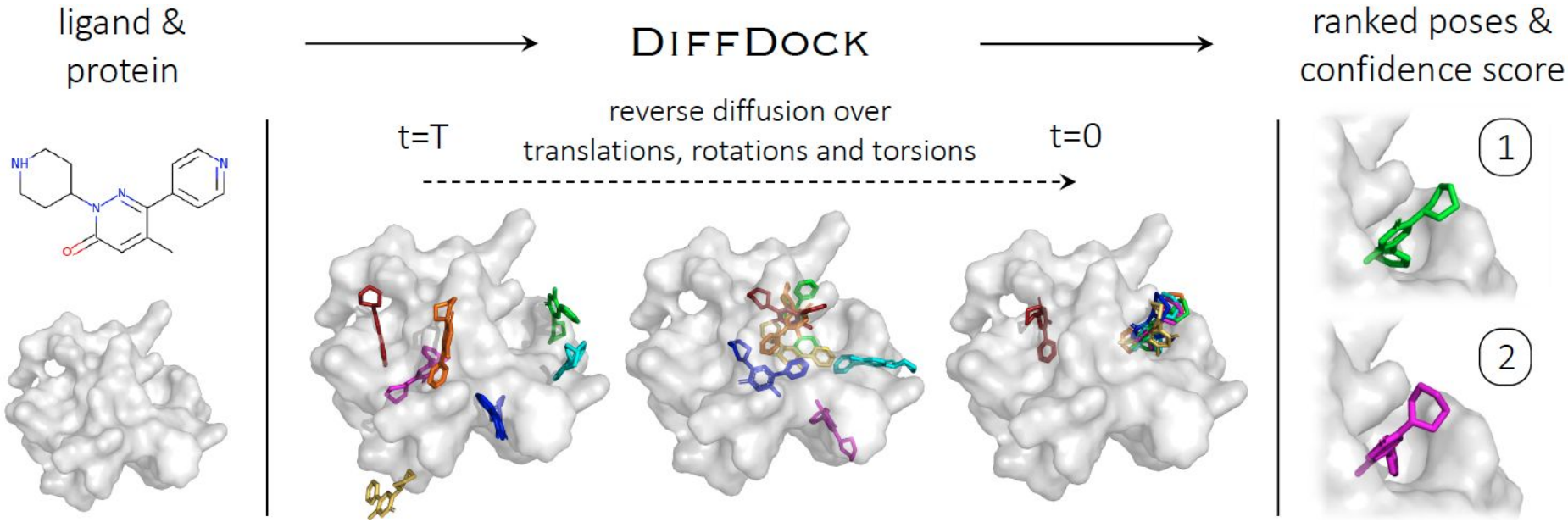
Torsional Diffusion for Small Molecules

group of machine gmum learning research



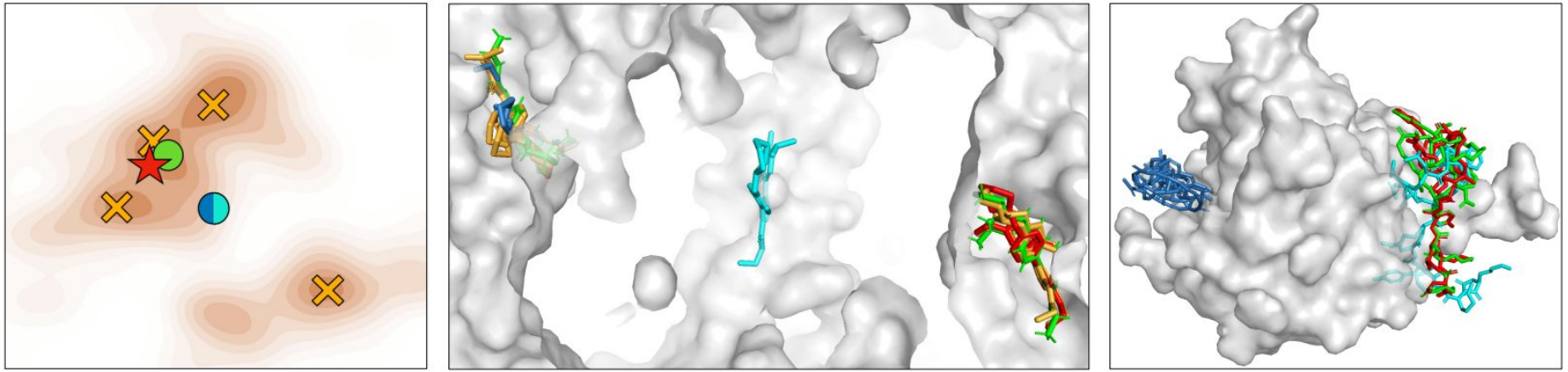
DiffDock: Docking with Deep Learning

group of machine gmum learning research



DiffDock: Motivation

group of machine gmum learning research



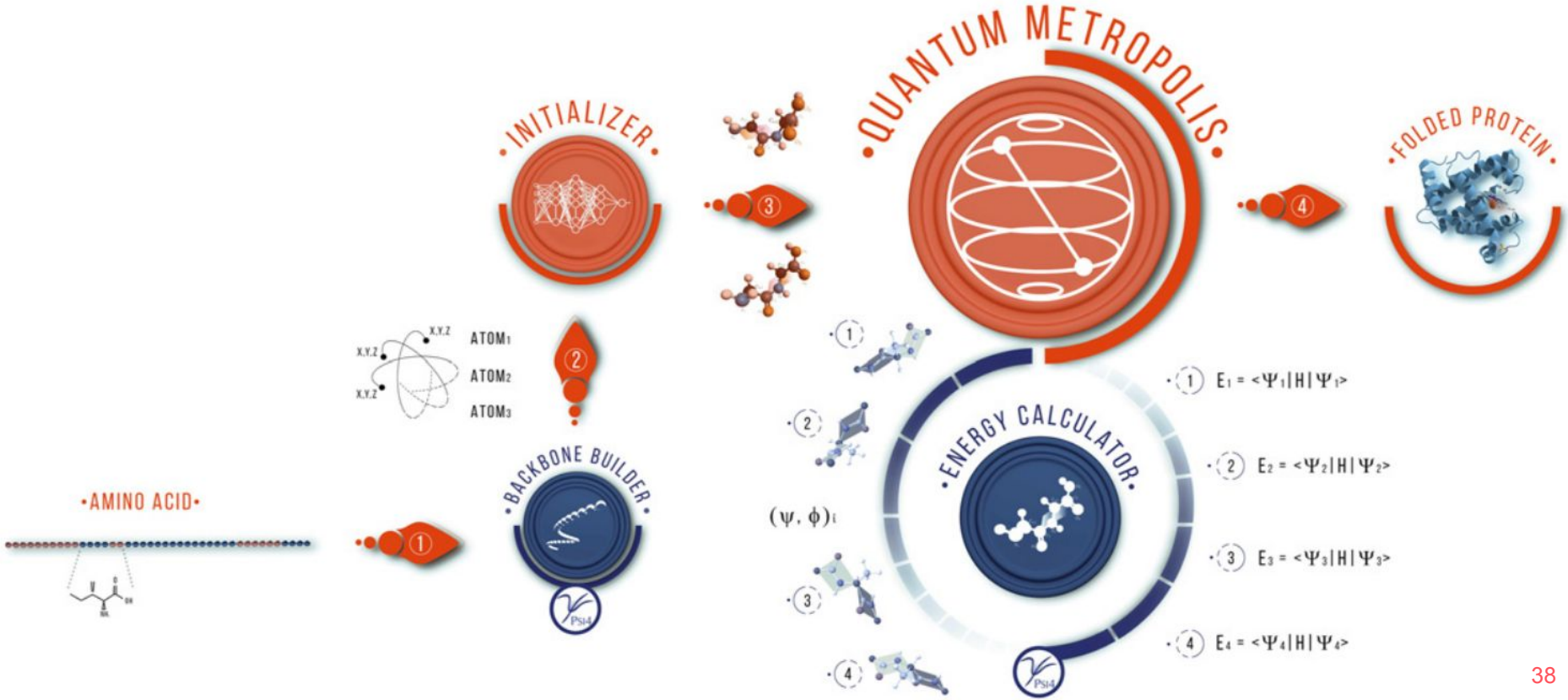
DiffDock: Results

group of machine gmum learning research

Method	Holo crystal proteins Top-1 RMSD %<2	Holo crystal proteins Top-1 RMSD Med.	Holo crystal proteins Top-5 RMSD %<2	Holo crystal proteins Top-5 RMSD Med.	Apo ESMFold proteins Top-1 RMSD %<2	Apo ESMFold proteins Top-1 RMSD Med.	Apo ESMFold proteins Top-5 RMSD %<2	Apo ESMFold proteins Top-5 RMSD Med.	Average Runtime (s)
GNINA	22.9	7.7	32.9	4.5	2.0	22.3	4.0	14.22	127
SMINA	18.7	7.1	29.3	4.6	3.4	15.4	6.9	10.0	126*
GLIDE	21.8	9.3							1405*
EQUIBIND	5.5	6.2	-	-	1.7	7.1	-	-	0.04
TANKBIND	20.4	4.0	24.5	3.4	10.4	5.4	14.7	4.3	0.7/2.5
P2RANK+SMINA	20.4	6.9	33.2	4.4	4.6	10.0	10.3	7.0	126*
P2RANK+GNINA	28.8	5.5	38.3	3.4	8.6	11.2	12.8	7.2	127
EQUIBIND+SMINA	23.2	6.5	38.6	3.4	4.3	8.3	11.7	5.8	126*
EQUIBIND+GNINA	28.8	4.9	39.1	3.1	10.2	8.8	18.6	5.6	127
DIFFDOCK (10)	35.0	3.6	40.7	2.65	**21.7**	**5.0**	**31.9**	**3.3**	10
DIFFDOCK (40)	**38.2**	**3.3**	**44.7**	**2.40**	20.3	5.1	31.3	**3.3**	40

QFold: Quantum Walks and DL

group of machine gmum learning research



Literature

group of machine gmum learning research

1. Kuhlman, Brian, and Philip Bradley. "Advances in protein structure prediction and design." Nature Reviews Molecular Cell Biology 20.11 (2019): 681-697.
2. Jumper, John, et al. "Highly accurate protein structure prediction with AlphaFold." Nature 596.7873 (2021): 583-589.
3. Baek, Minkyung, et al. "Accurate prediction of protein structures and interactions using a three-track neural network." Science 373.6557 (2021): 871-876.
4. Mirdita, Milot, et al. "ColabFold: making protein folding accessible to all." Nature methods 19.6 (2022): 679-682.
5. Jing, Bowen, et al. "EigenFold: Generative Protein Structure Prediction with Diffusion Models." arXiv preprint arXiv:2304.02198 (2023).
6. Jing, Bowen, et al. "Torsional diffusion for molecular conformer generation." arXiv preprint arXiv:2206.01729 (2022).
7. Corso, Gabriele, et al. "Diffdock: Diffusion steps, twists, and turns for molecular docking." arXiv preprint arXiv:2210.01776 (2022).
8. Casares, Pablo Antonio Moreno, Roberto Campos, and Miguel Angel Martin-Delgado. "QFold: quantum walks and deep learning to solve protein folding." Quantum Science and Technology 7.2 (2022): 025013.