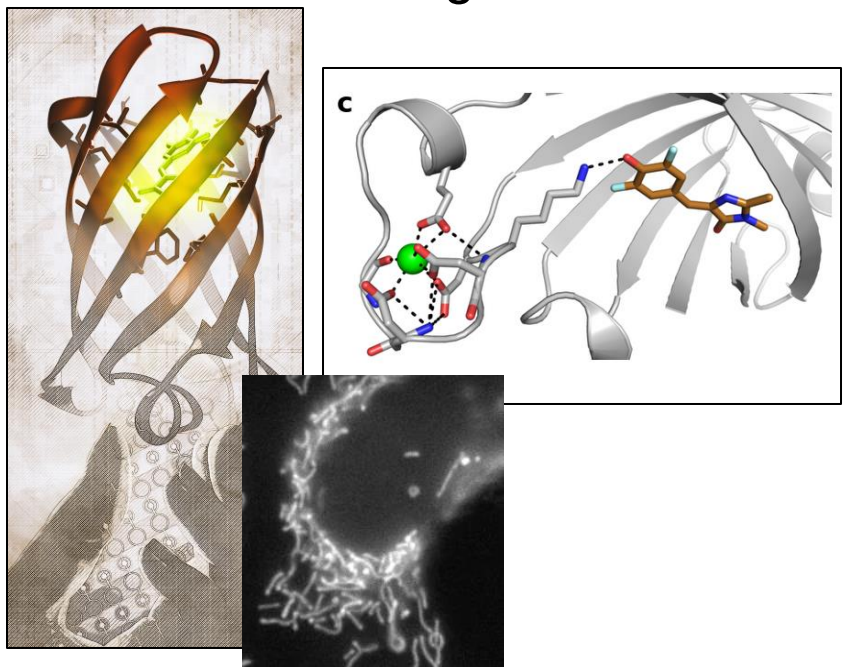
Protein design
Part 1: structure prediction

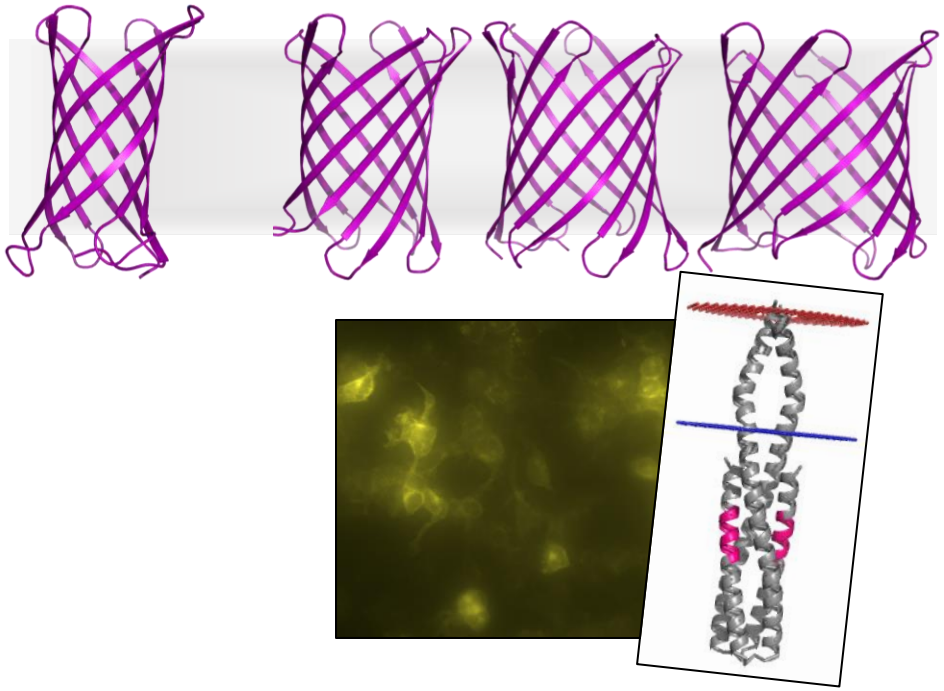
Anastassia Vorobieva

Small-molecule binder/sensor design



Enzyme design

Membrane protein design



2009 — PhD UCLouvain, BE — Directed evolution

2014 — Postdoc UW, Seattle — David Baker

2019 — Postdoc EPFL

2021 — Group leader, VIB

2022 — Assistant professor, VUB

📝 Workshop Schedule

TIME	WEDNESDAY 28 MAY	THURSDAY 29 MAY	FRIDAY 20 MAY
09:00 – 10:30	Protein modelling and structure prediction: Intro to key concepts, from physics-based modelling to AI	De novo protein design: introduction, minimal sequence design, structure-based design principles	De novo design with AI models: RFDiffusion, ProteinMPNN, and ColabFold
10:30 – 11:00	Break and questions	Break and questions	Break and questions
11:00 – 12:00	Introduction to protein design: predicting the effect of mutations on protein stability	Structure-based de novo design: How to generate new structures? The chicken-and-egg problem	Practical session: De novo design of a SARS-CoV-2 RBD binder using RFDiffusion and ProteinMPNN
12:00 – 13:30	Lunch	Lunch	Lunch
13:30 – 15:00	Practical session: AlphaFold hands-on	Practical session: Parametric design of alpha-helical bundles	Practical session: Data analysis and group slide preparation
15:15 – 17:45	Practical session: In silico mutational scanning and ΔΔG calculations	Practical session: Sequence design for parametric bundles with PyRosetta	Practical session: Group presentations and results discussion

Structure prediction - outline

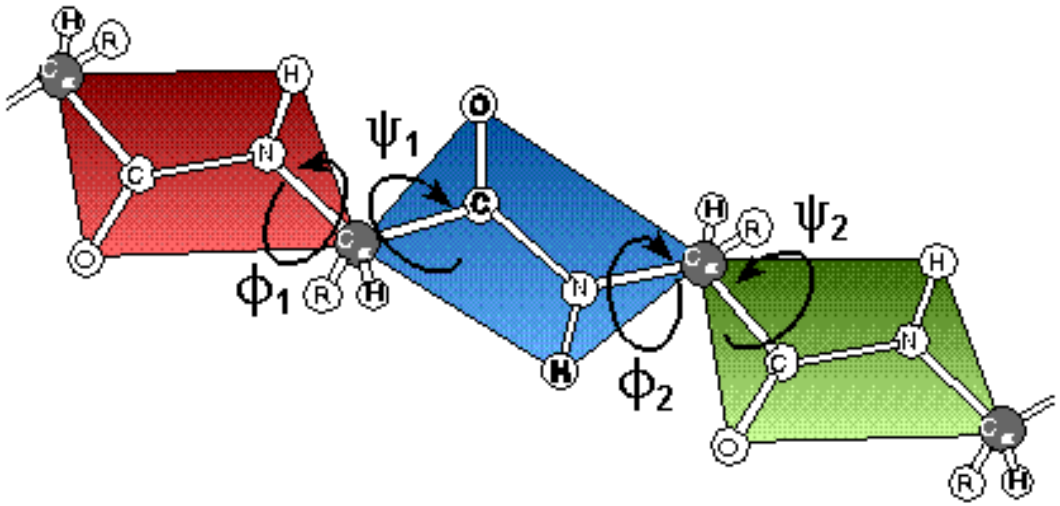
1. Introduction to protein folding and structure:
    - Backbone torsion and secondary structure
    - Levinthal's paradox
    - The thermodynamic hypothesis of protein folding
    - Structure prediction/design *vs* molecular dynamics simulations

2. Predicting the structure of proteins from their sequence
    - Physics-based models:
        - Scoring
        - Sampling
        - Building a prediction pipeline
    - Co-evolution and contacts prediction
    - AI models: AlphaFold 1 and 2 architectures
    - AI models: protein language models

Introduction - protein folding and structure

Accelerated *in vitro* evolution, inspired by natural selection.

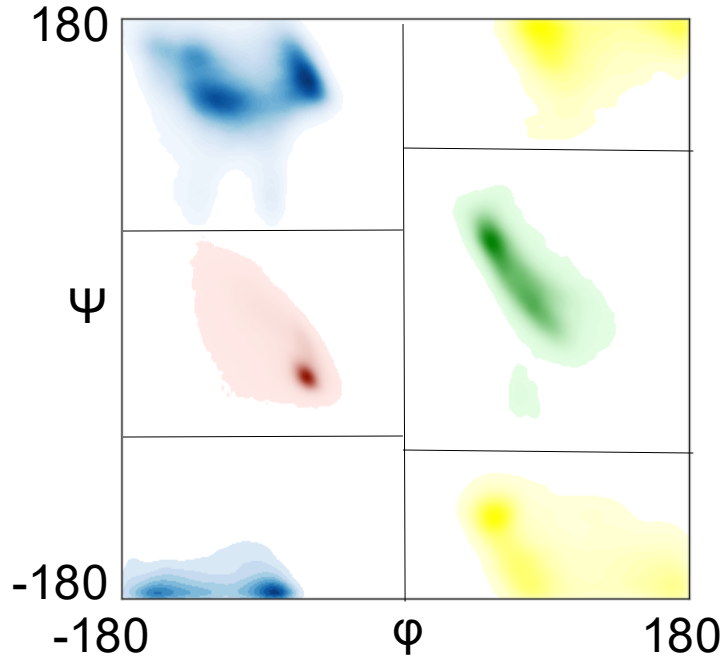
Polypeptide folding: torsion around several degrees of freedom



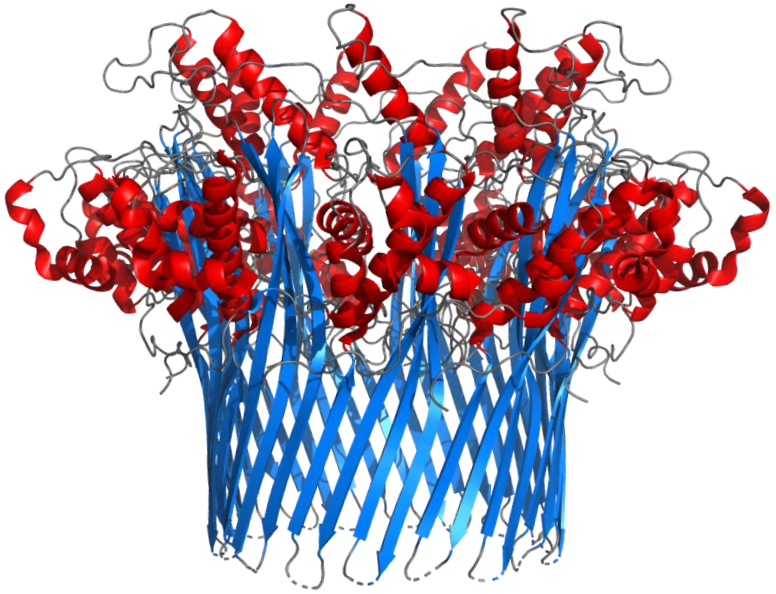
Φ – rotation around the N-C_α bond

Ψ – rotation around the C_α-C bond

Ω – peptide bond

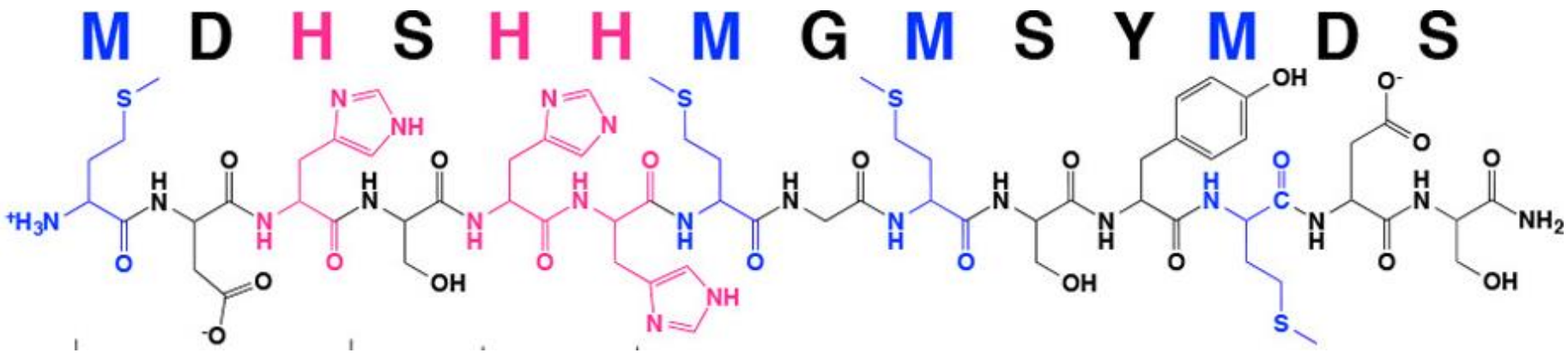


Secondary structure



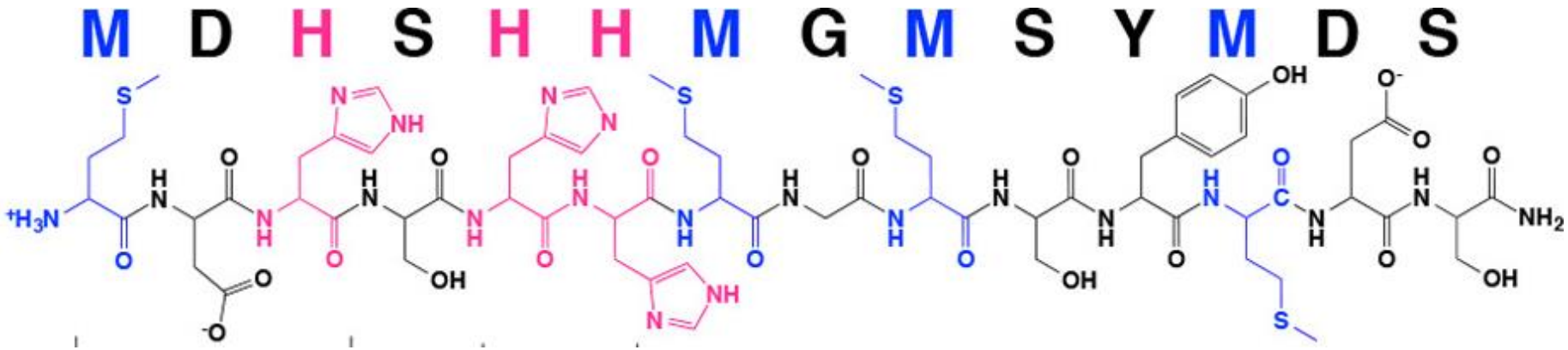
Tertiary structure

The torsion in the backbone is stabilized by the side chains

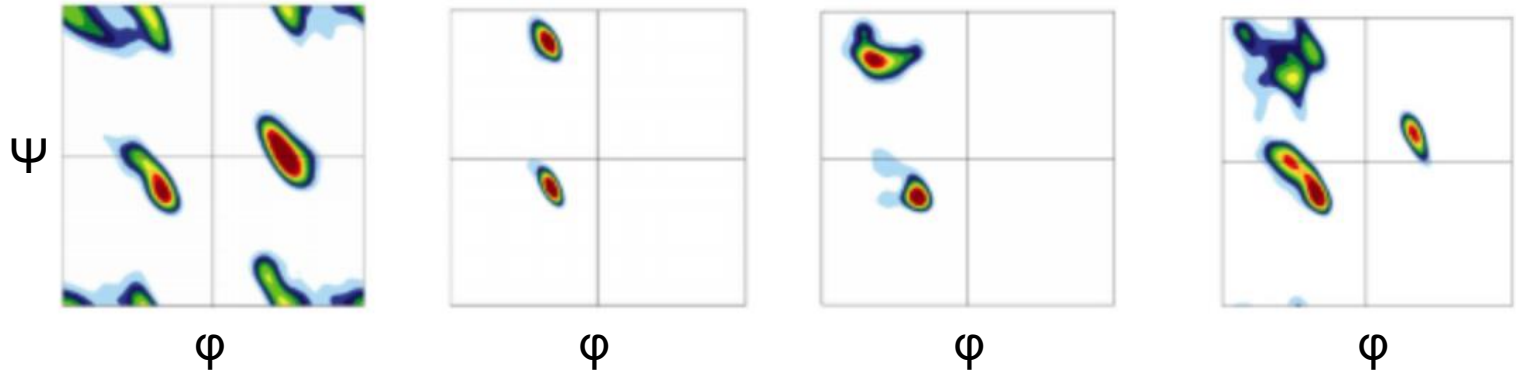


20 letters

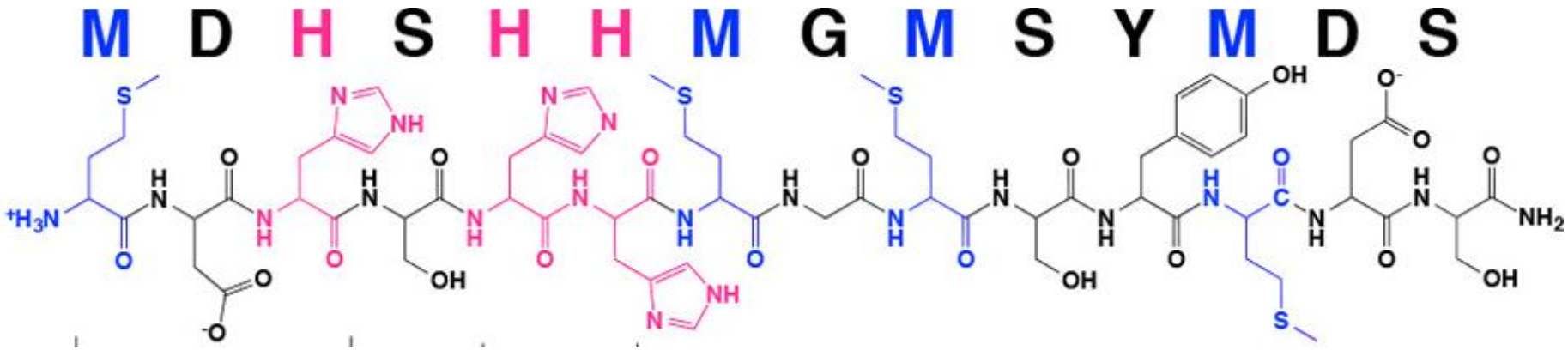
The torsion in the backbone is stabilized by the side chains



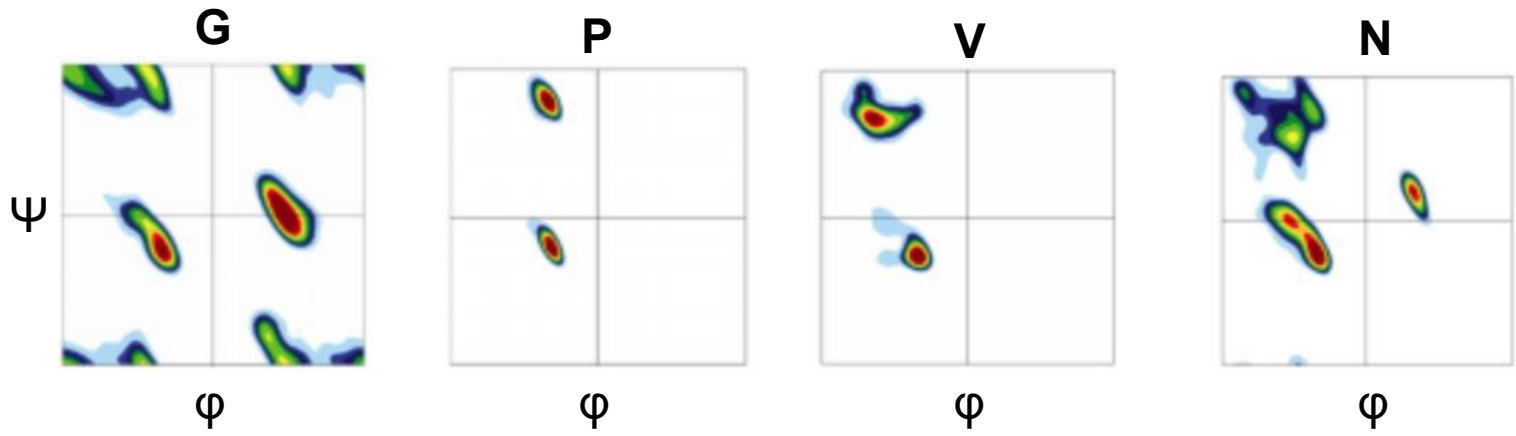
20 letters



The torsion in the backbone is stabilized by the side chains



20 letters



The folded state of a protein is likely the energy minima for its sequence (C. Anfinsen)

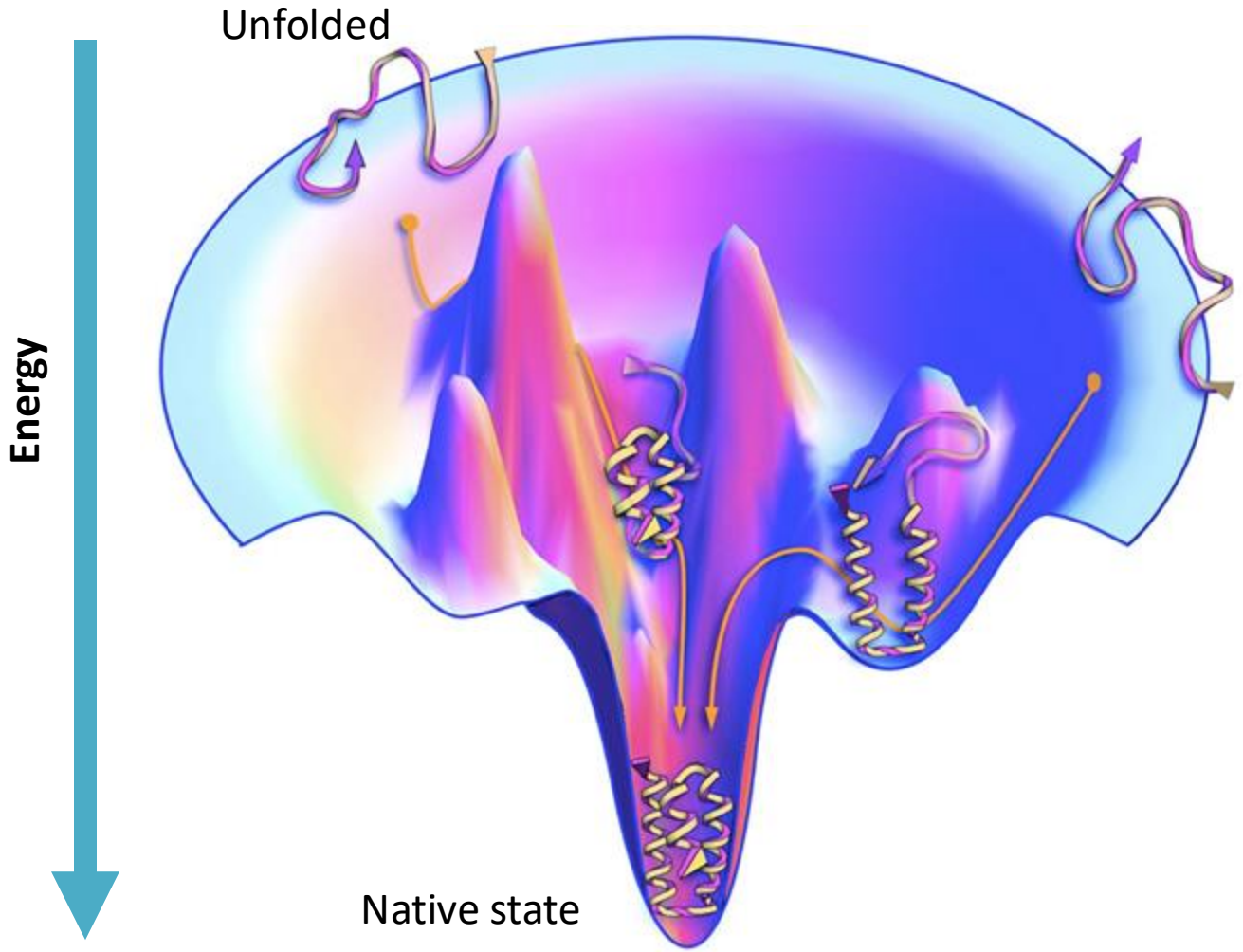


Image from: "The protein-folding problem, 50 years on." *science* 338, no. 6110 (2012): 1042-1046.

The native state of a protein is key to its function

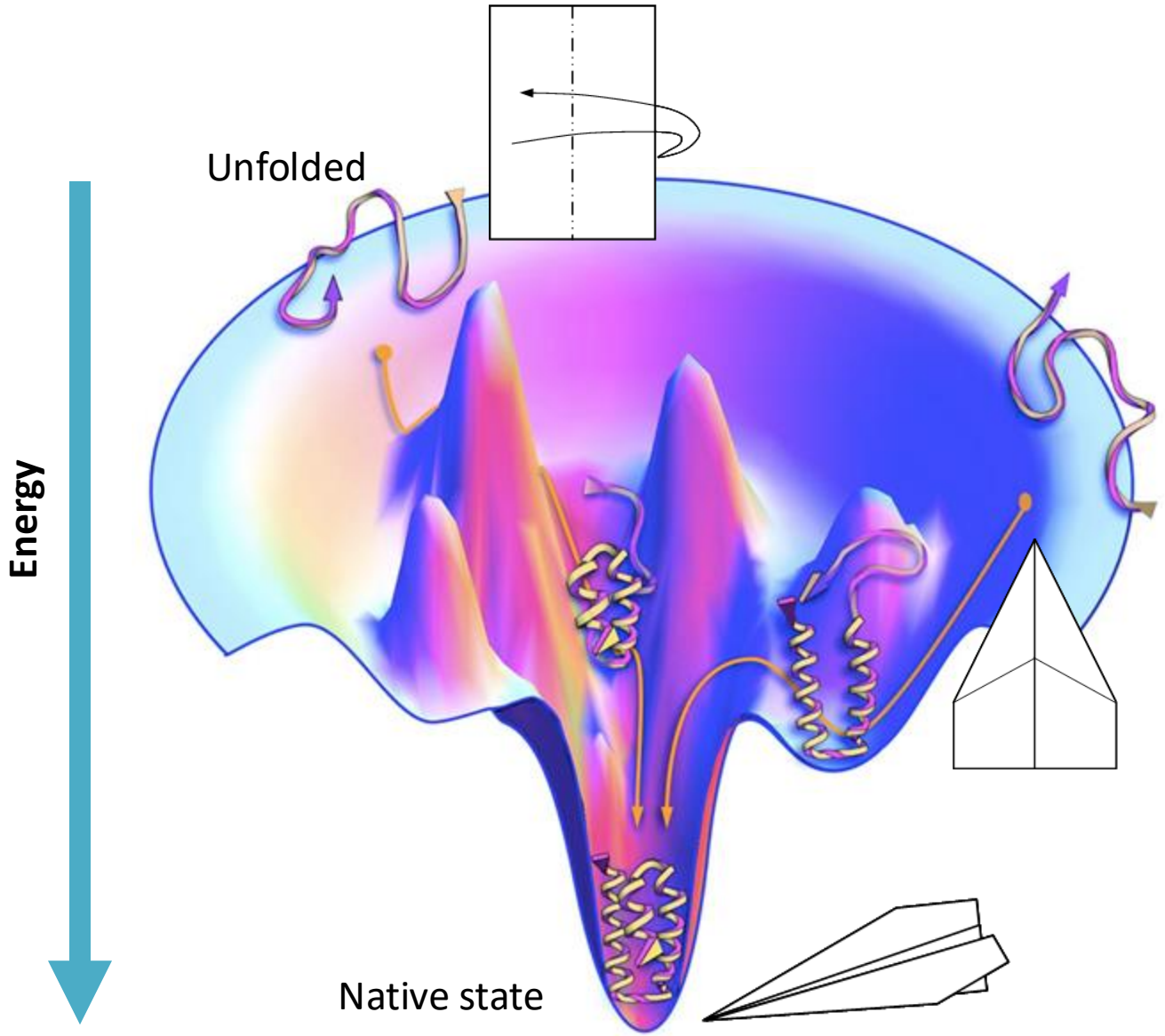


Image from: "The protein-folding problem, 50 years on." *science* 338, no. 6110 (2012): 1042-1046.

Levinthal's paradox

An unfolded protein has an astonishing number of possible conformations.

Yet, most proteins fold on a µs-ms timescale.

➔ Protein folding is driven by a few stable interaction formed early-on (folding pathway).

Molecular Dynamics simulations are used to study the dynamics and the stability of an existing state

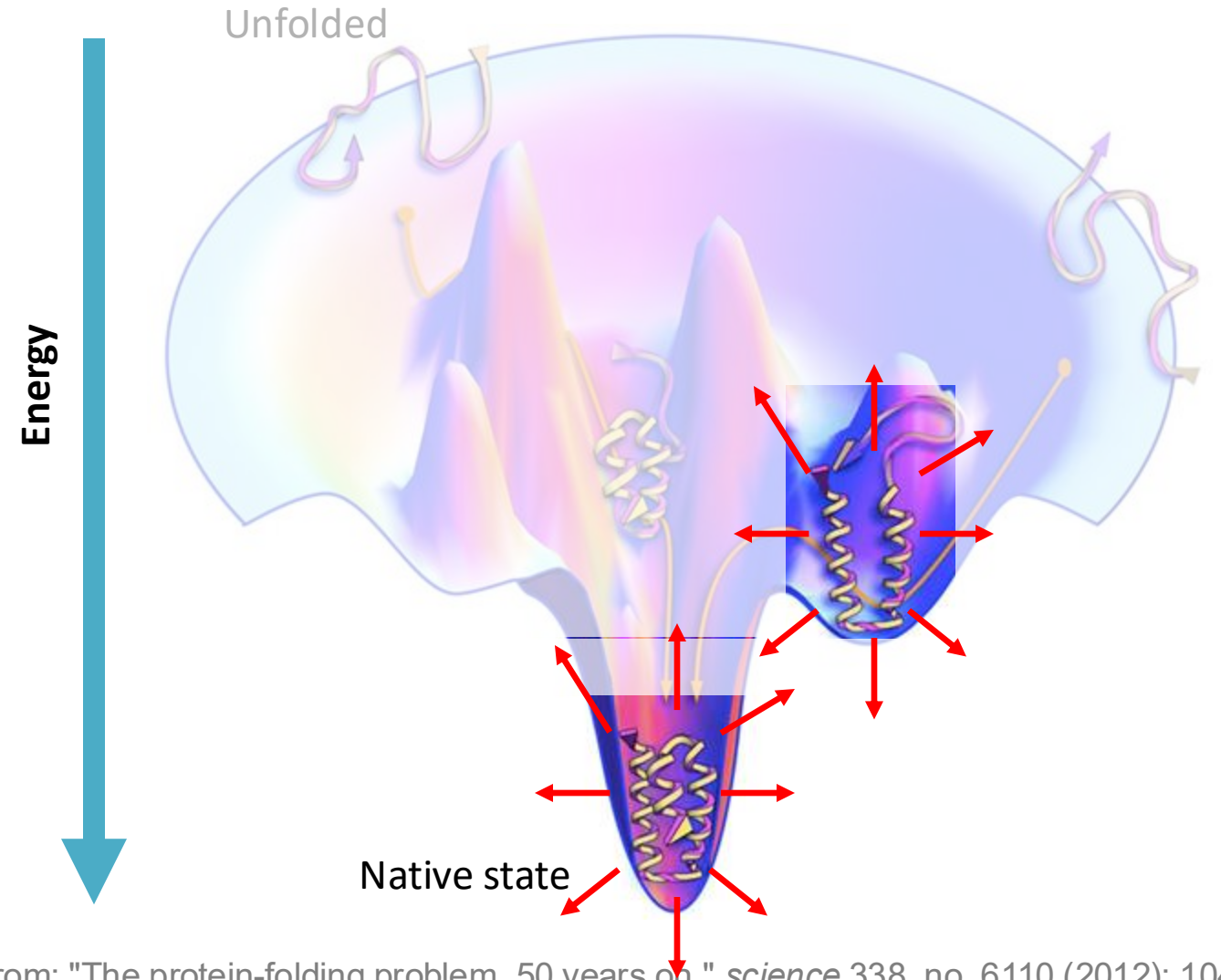
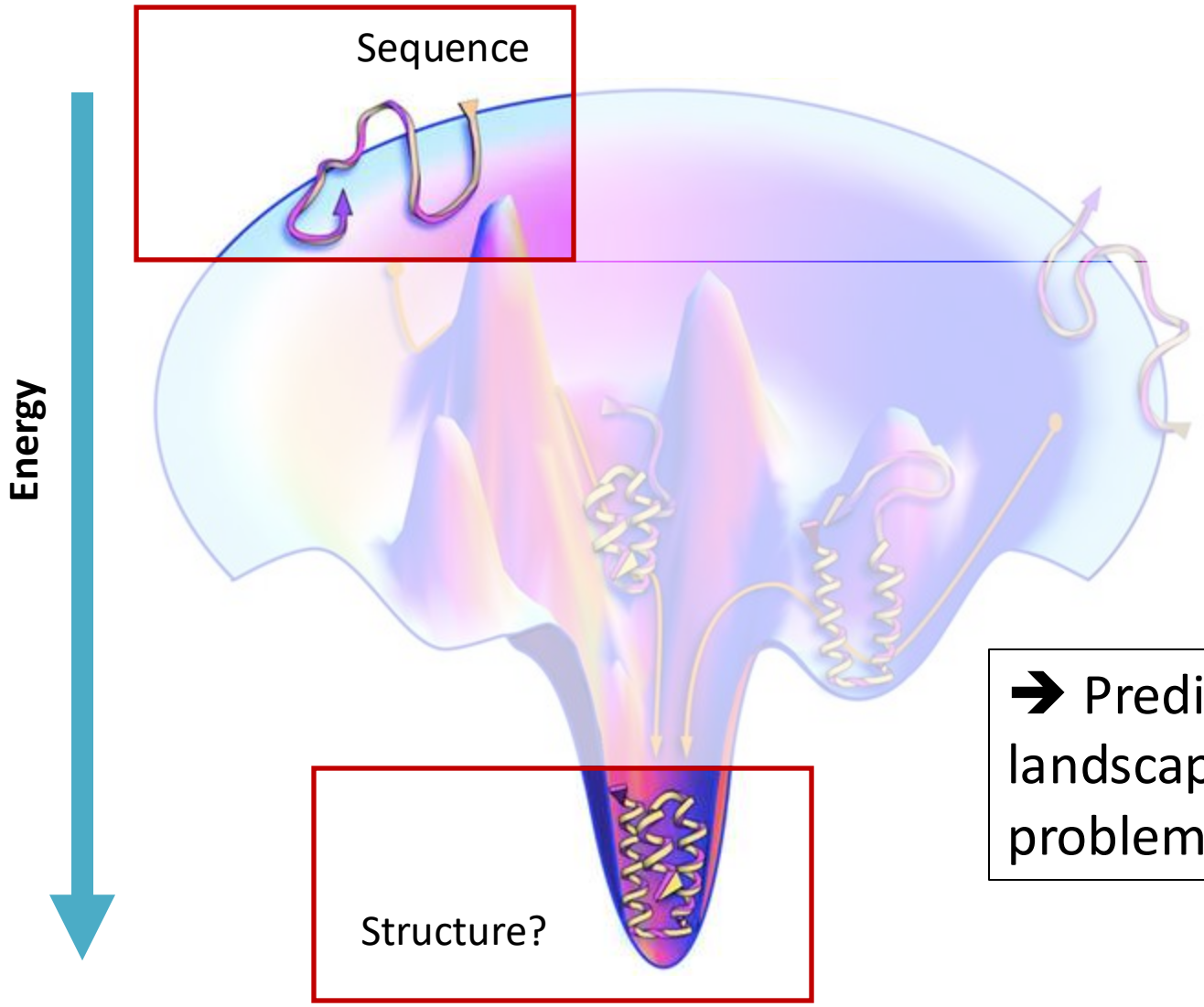


Image from: "The protein-folding problem, 50 years on." *science* 338, no. 6110 (2012): 1042-1046.

Structure prediction(*ab initio*) and design



Structure prediction

Design

➔ Predicting the entire energy landscape is an un-tracktable problem

Image from: "The protein-folding problem, 50 years on." *science* 338, no. 6110 (2012): 1042-1046.

Modelling and predicting protein structures

From physics-based modelling to deep learning

“Classic” physics-based models

- Both MD simulations and protein structure prediction/design
- Explicit modeling of possible conformations
- **Sampling** conformations from the energy landscape
  Proteins can fold into an astronomically large number of conformations, and sampling methods aim to explore these efficiently.
- **Scoring** the sampled conformations with an energy function or force field
  Estimate the free energy of a state or a conformation as accurately as possible.

Scoring – energy functions

- Mathematical models used to evaluate the stability or quality of a protein model
  ➔ Models ranking and comparison tools

- Trade-off between speed (e.g. energy functions for design) and accuracy (e.g. MD simulations force fields)

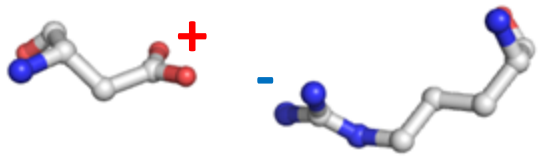
Scoring – energy functions

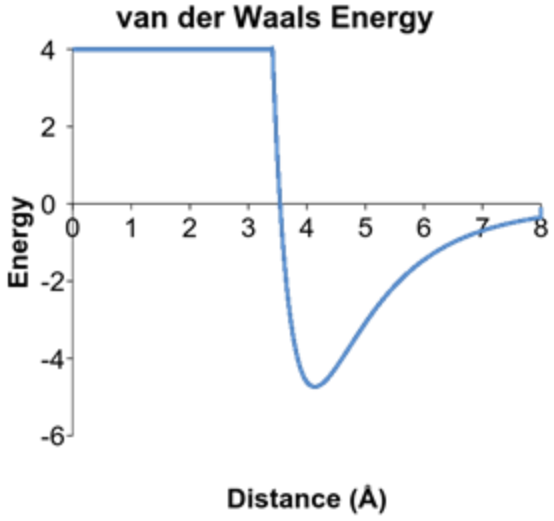
- Mathematical models used to evaluate the stability or quality of a protein model
  ➔ <u>Models ranking and comparison tools</u>

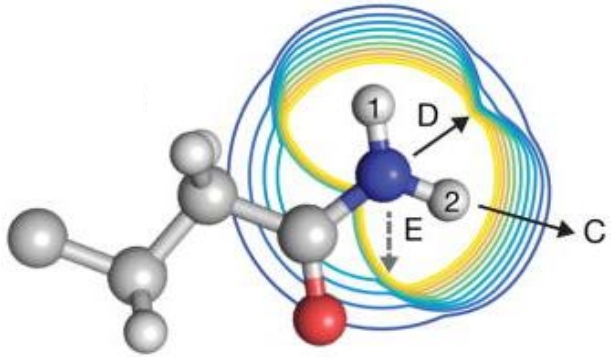
- Trade-off between speed (e.g. energy functions for design) and accuracy (e.g. MD simulations force fields)

- Physics-based parameters

Electrostatic interactions







Lazaridis-Karplus solvation

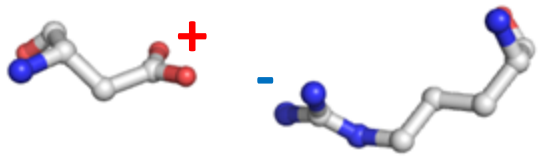
Scoring – energy functions

- Mathematical models used to evaluate the stability or quality of a protein model
  ➔ <u>Models ranking and comparison tools</u>

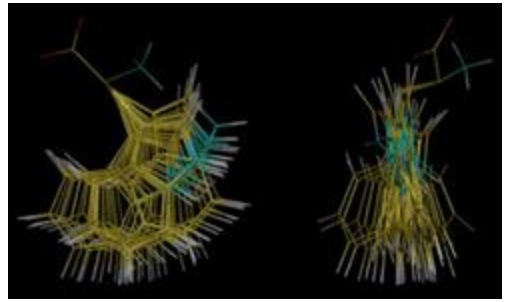
- Trade-off between speed (e.g. energy functions for design) and accuracy (e.g. MD simulations force fields)

- Physics-based parameters
- Statistical potentials (from PDB)

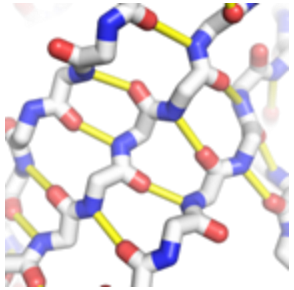
Electrostatic interactions

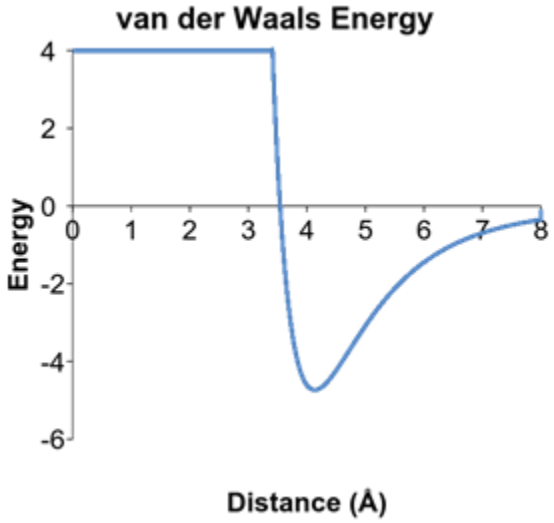


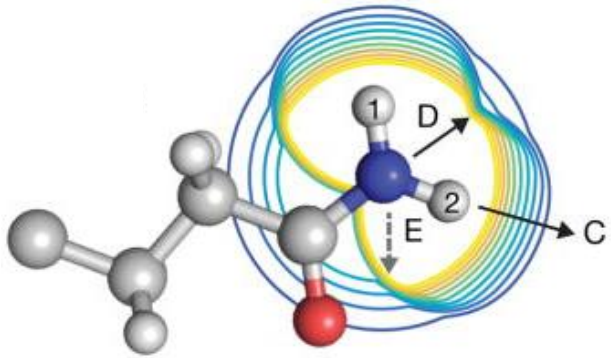
Sidechain conformations



Hydrogen bond

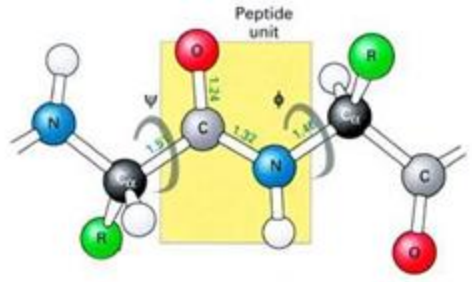






Lazaridis-Karplus solvation

Backbone conformations



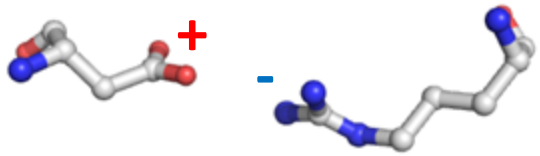
Scoring – energy functions

- Mathematical models used to evaluate the stability or quality of a protein model
  ➔ <u>Models ranking and comparison tools</u>

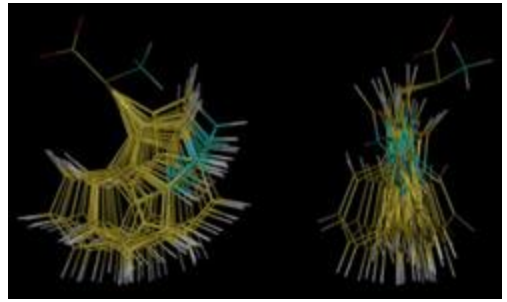
- Trade-off between speed (e.g. energy functions for design) and accuracy (e.g. MD simulations force fields)

- Physics-based parameters
- Statistical potentials (from PDB)
- Hybrid potentials: **e.g. Rosetta**

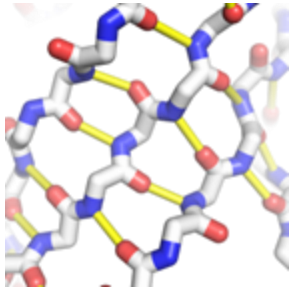
Electrostatic interactions

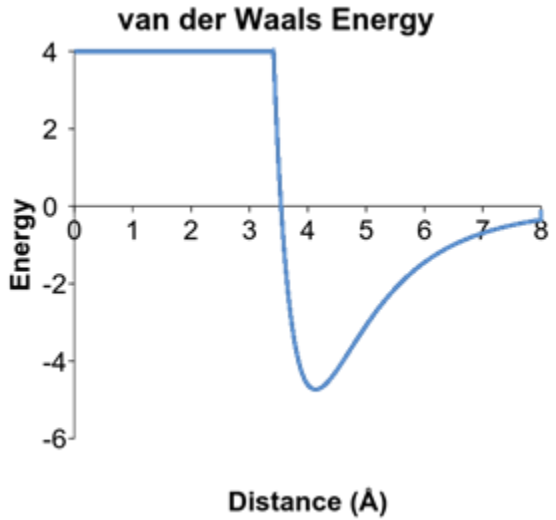


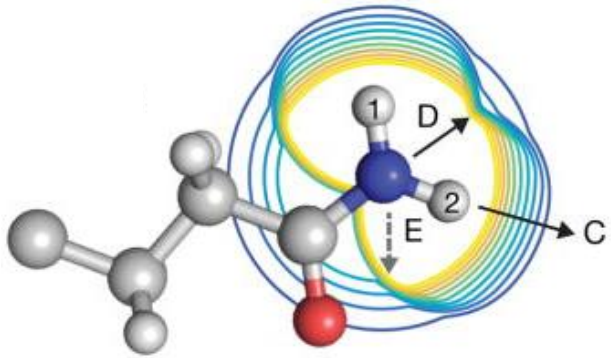
Sidechain conformations



Hydrogen bond

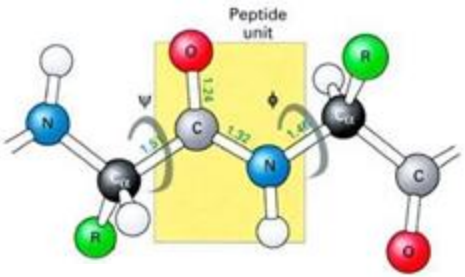






Lazaridis-Karplus solvation

Backbone conformations



weight

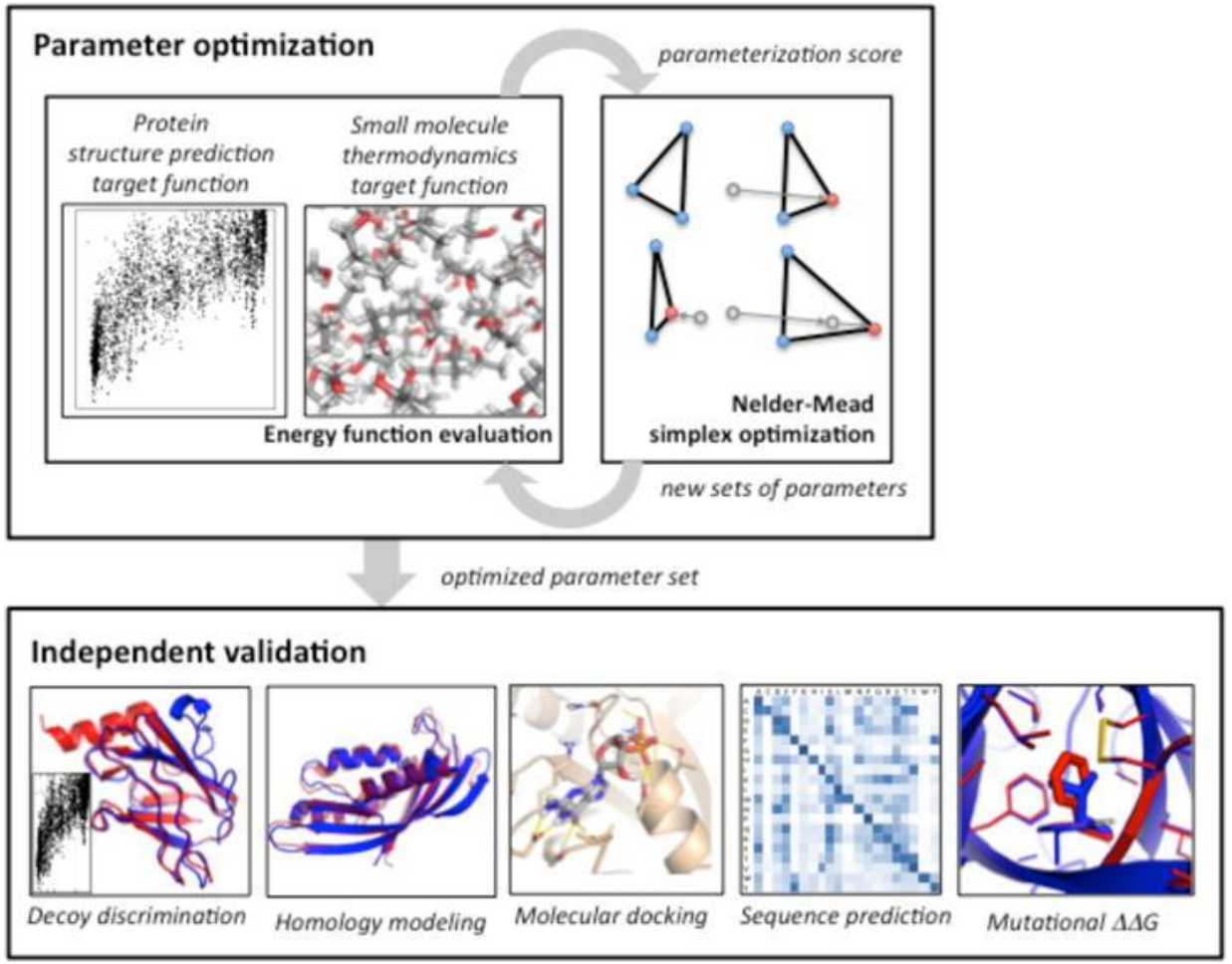
$$\Delta E_{\text{total}} = \sum_i w_i E_i(\Theta_i, \text{aa}_i)$$

https://doi.org/10.1021/acs.jctc.7b00125

Training an energy function

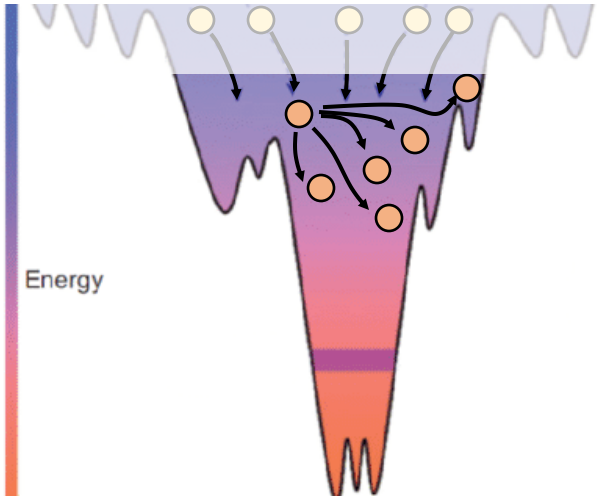
- Adjusting the weights of the parameters in the energy function to fit experimental data

- Example: training the Rosetta energy function

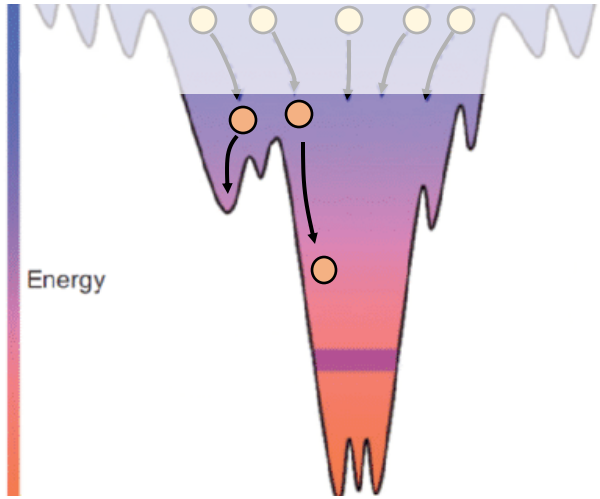


https://doi.org/10.1021/acs.jctc.6b00819

Sampling *vs* minimization



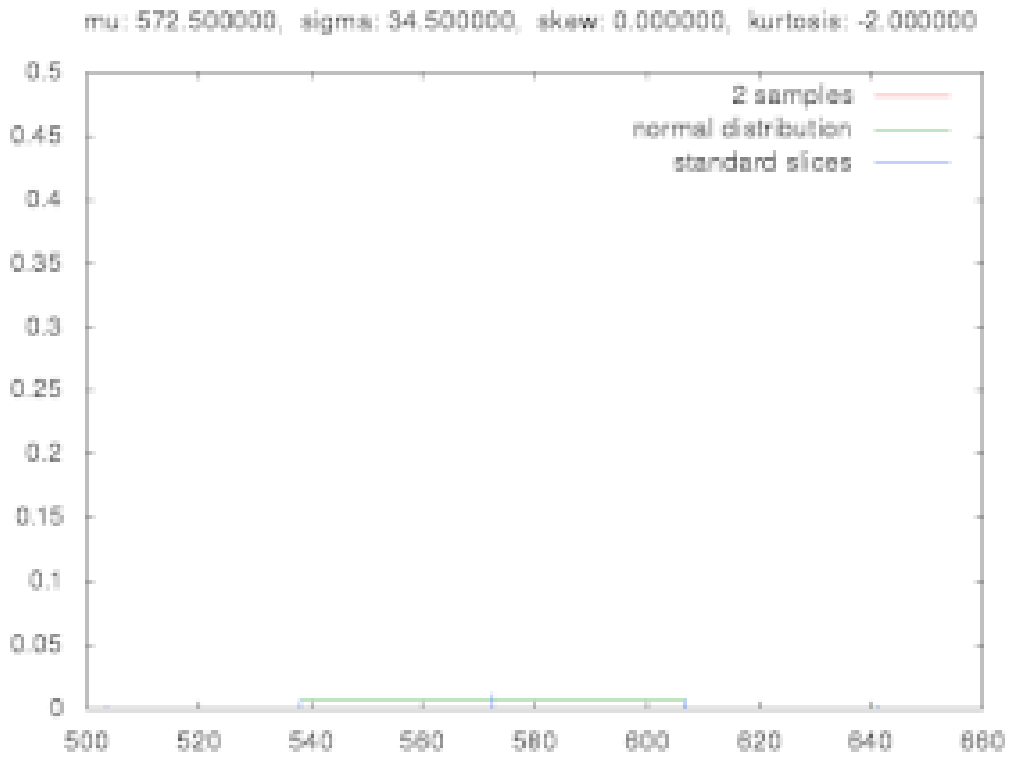
Sampling: random exploration



minimization: rolling down the energy gradient

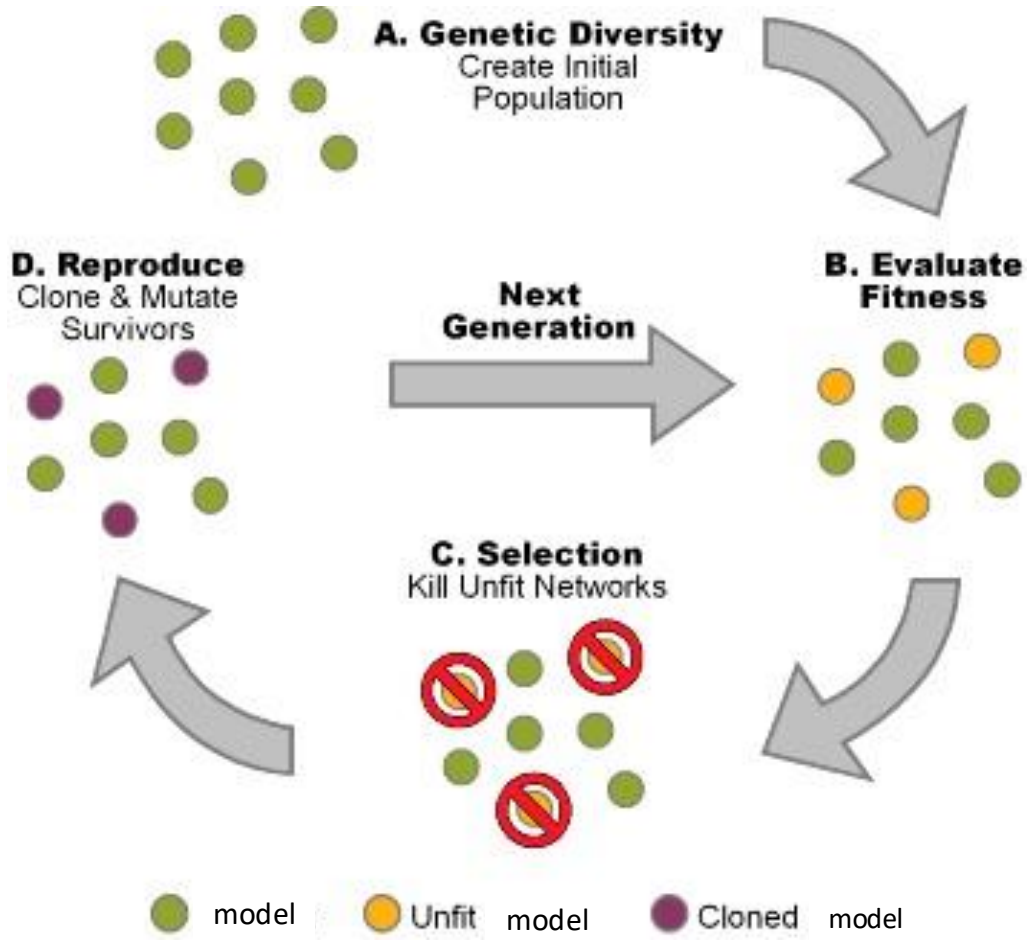
Sampling methods

- Approaches and algorithms to efficiently navigating the conformational landscape
- Monte Carlo algorithm to simulate random sampling of a normal distribution

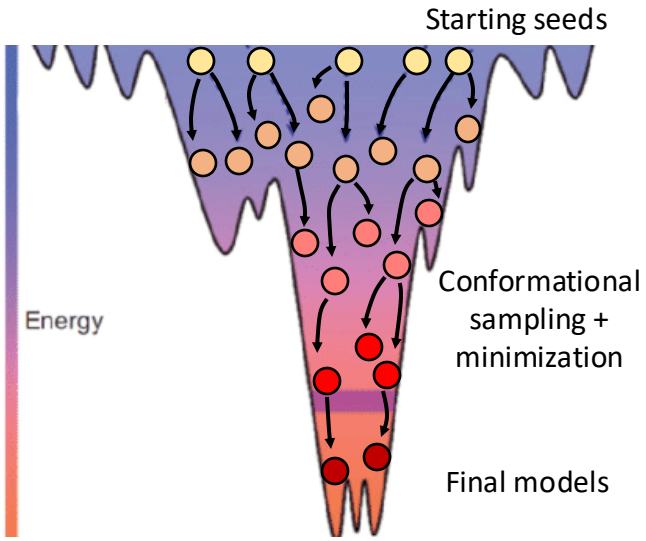


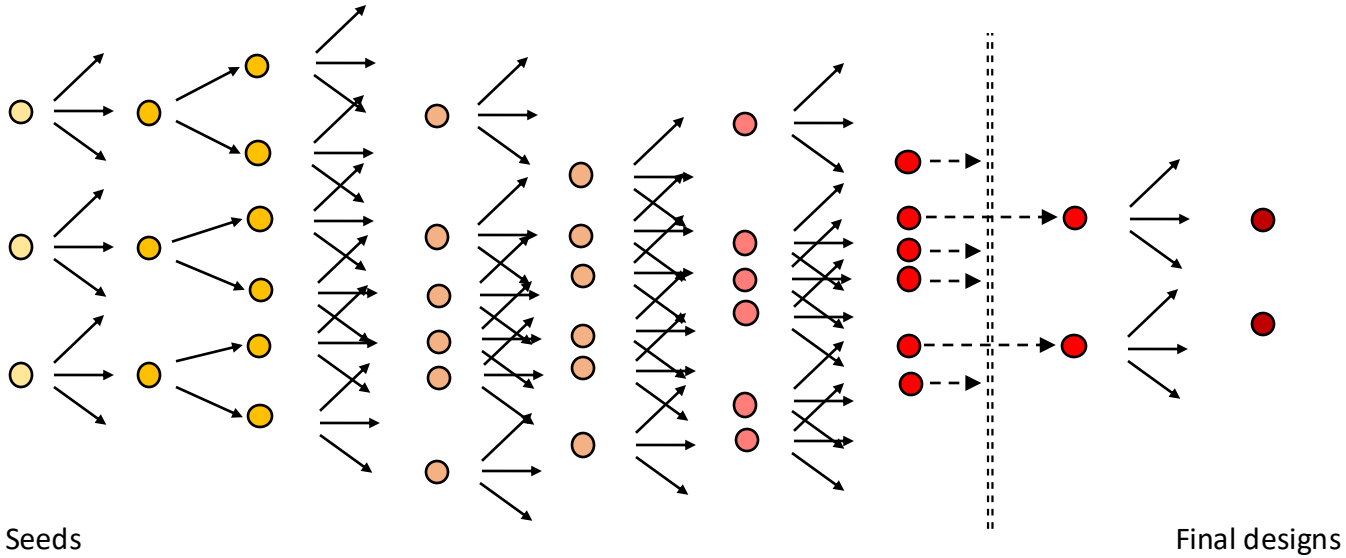
Sampling methods

- Approaches and algorithms to efficiently navigating the conformational landscape
- Monte Carlo algorithm to simulate random sampling of a normal distribution
- Genetic algorithm

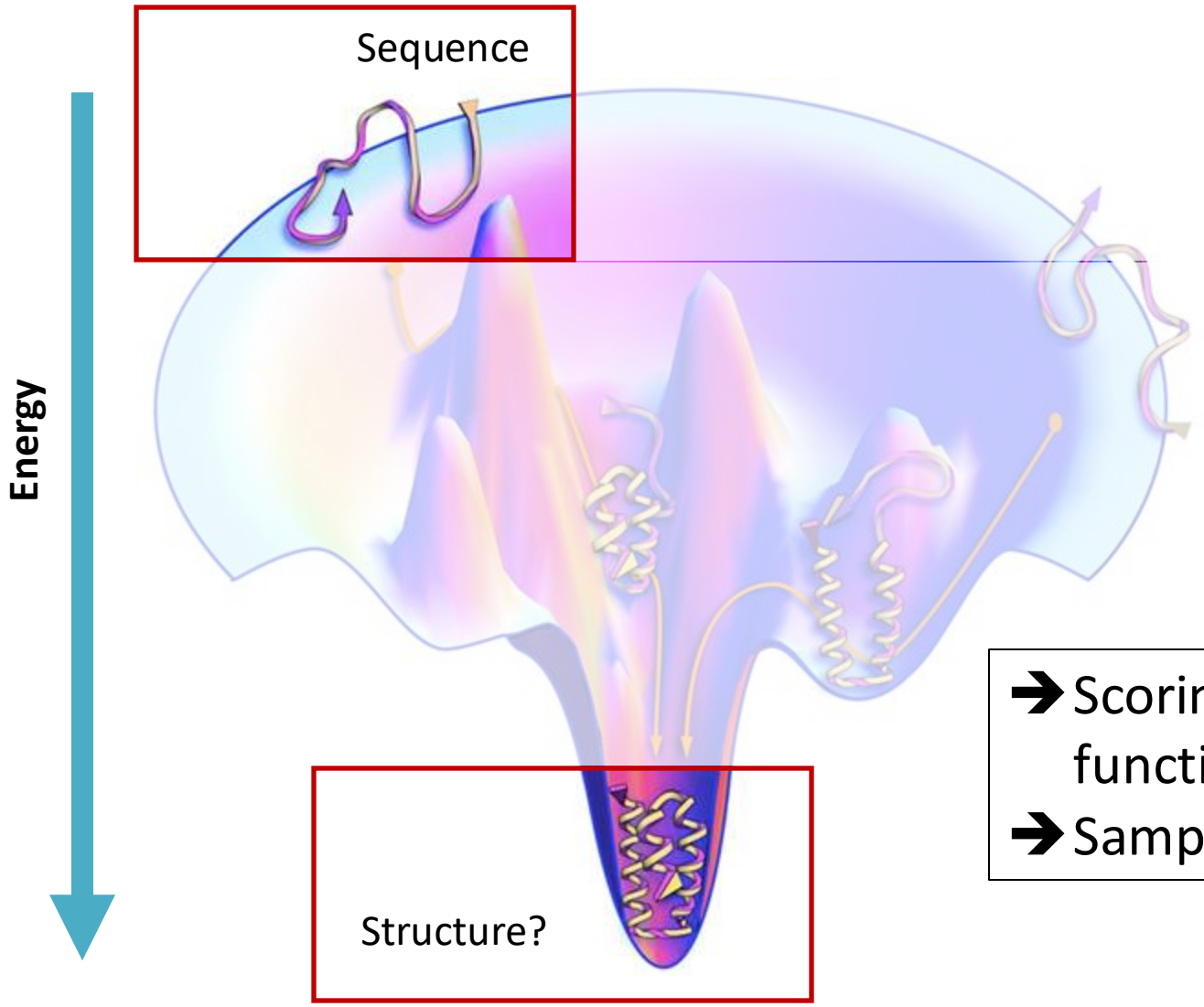


Example modelling trajectory





Predicting the structure of a protein based on its sequence (*ab initio*)



➔ Scoring: Rosetta energy function

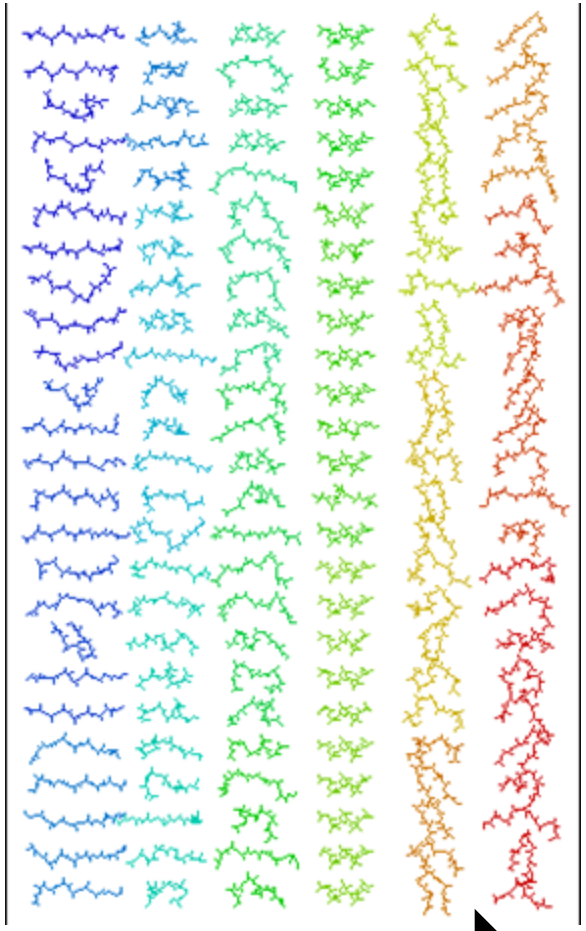
➔ Sampling structures ?

Image from: "The protein-folding problem, 50 years on." *science* 338, no. 6110 (2012): 1042-1046.

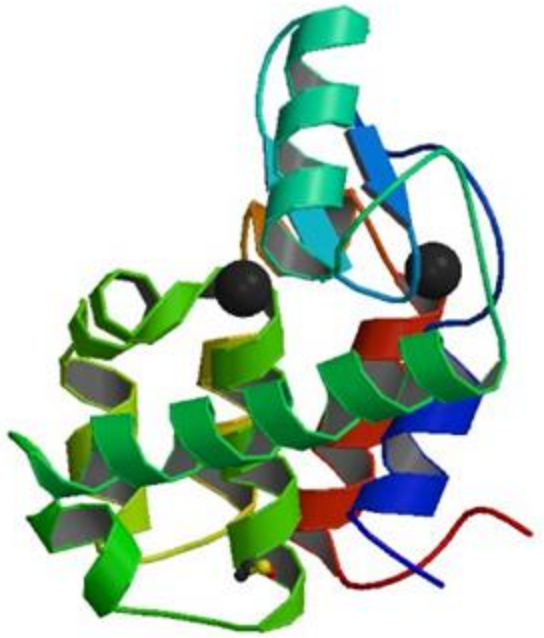
Classic structure prediction: **Monte Carlo sampling** of existent conformations

MNIFEMLRIDEGLRLKI
YKDTEGYYTIGIGHLLT
KSPSLNASKSELDKAIG
RNTNGVITKDEAEKLFN
QDVDAAVRGILRNAKLK
PVYDSLDAVRRAALINM
VFQMGETGVAGFTNSLR
MLQQKRWDEAAVNLAKS
RWYNQTPNRAKRVITTF
RTGTWDAYKNL

Primary Sequence

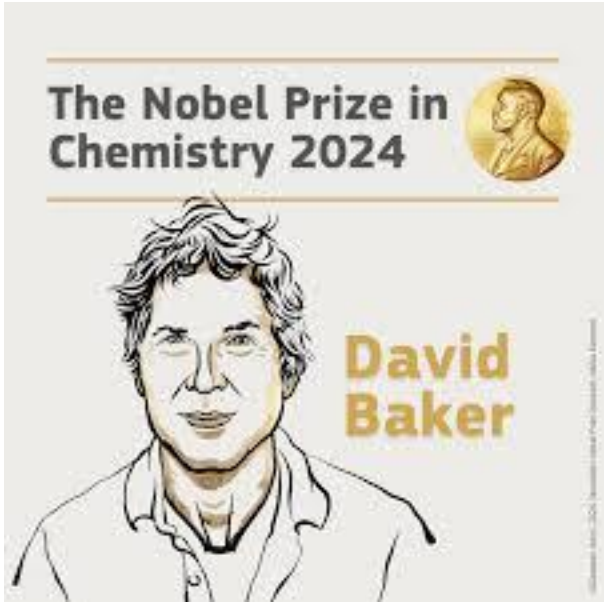


***Ab initio* folding**



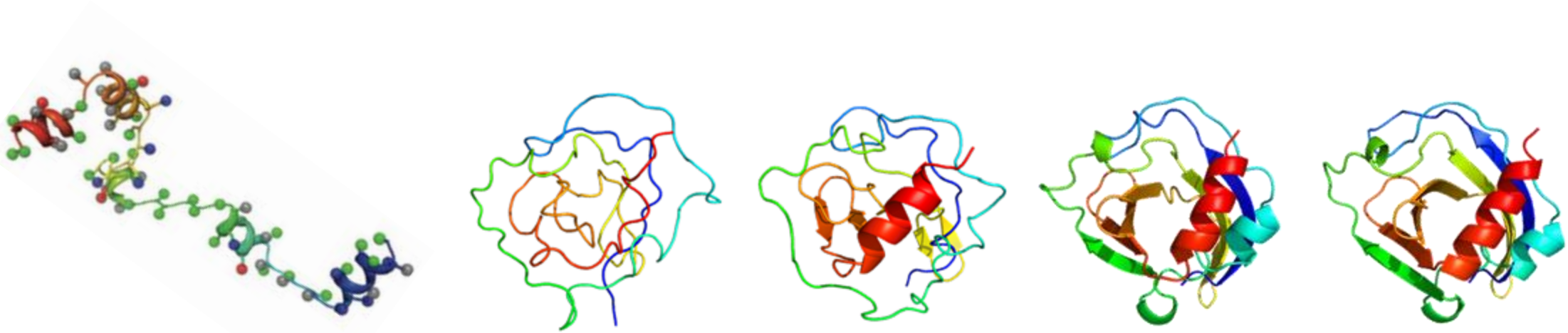
Tertiary Structure

Rosetta structure prediction (state-of-the-art years ago)



This is not a folding pathway simulation, but one successful trajectory of random exploration of the conformational space.

Ab initio Relax



Sequence

uniform: 9 (25) → **Stage1** → **score0**: vdw

uniform: 9 (25) → **Stage2** → **score1**: vdw, pair, env, hs_pair, ss_pair, sheet

uniform: 9 (25) → **Stage 3a** → **score2**: vdw, cenpack, pair, env, cbeta, hs_pair, ss_pair, sheet

uniform: 9 (25) → **Stage 3b** → **score5**: vdw, cenpack, pair, env, cbeta, hs_pair, ss_pair, sheet

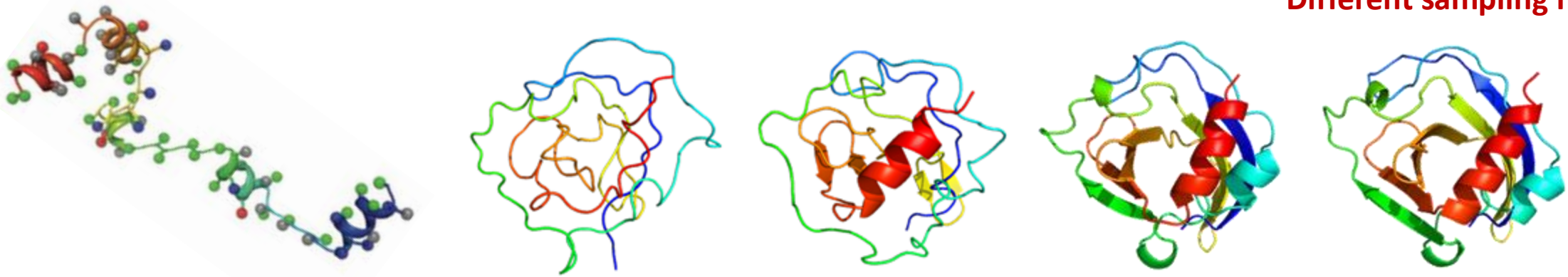
uniform: 3 (200)
smooth: 3 (200) → **Stage4** → **score3**: vdw, cenpack, pair, env, cbeta, **rg**, hs_pair, ss_pair, rsigma, sheet

pack
minimize → **Relax** → **ref2015**

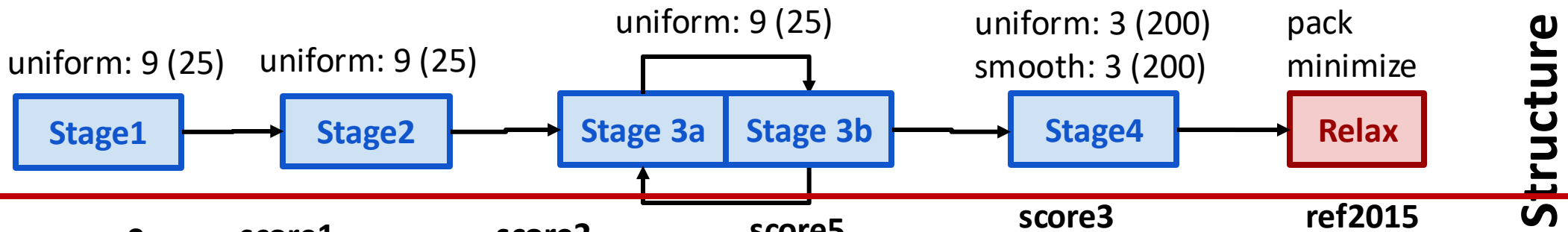
Structure

Ab initio Relax

Different sampling methods



Sequence



Structure

score0
vdw

score1
vdw, pair
env, hs_pair
ss_pair, sheet

score2
vdw, cenpack
pair, env,
cbeta, hs_pair,
ss_pair, sheet

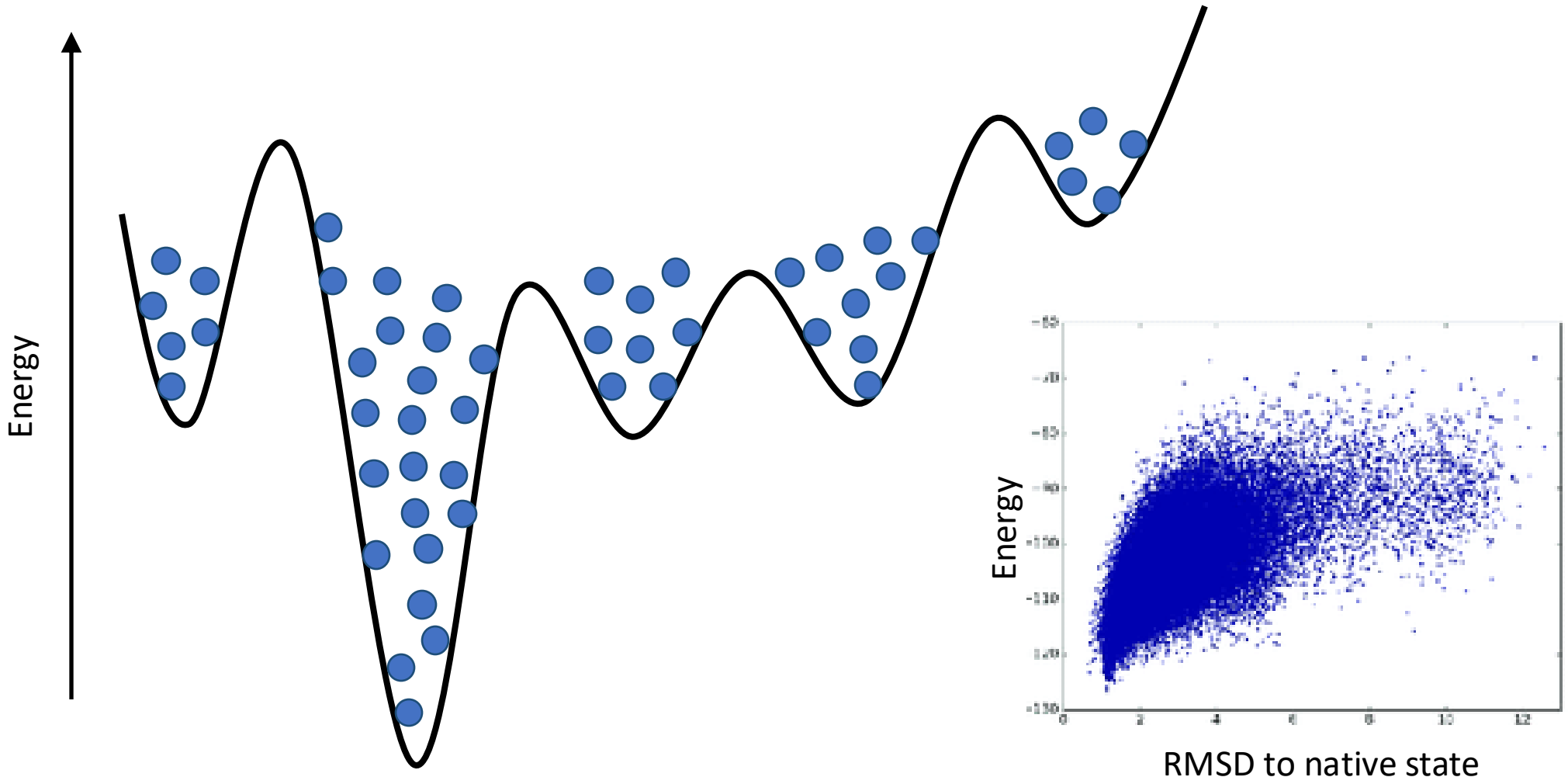
score5
vdw, cenpack
pair, env, cbeta,
hs_pair, ss_pair,
sheet

score3
vdw, cenpack
pair, env, cbeta,
rg, hs_pair, ss_pair
rsigma, sheet

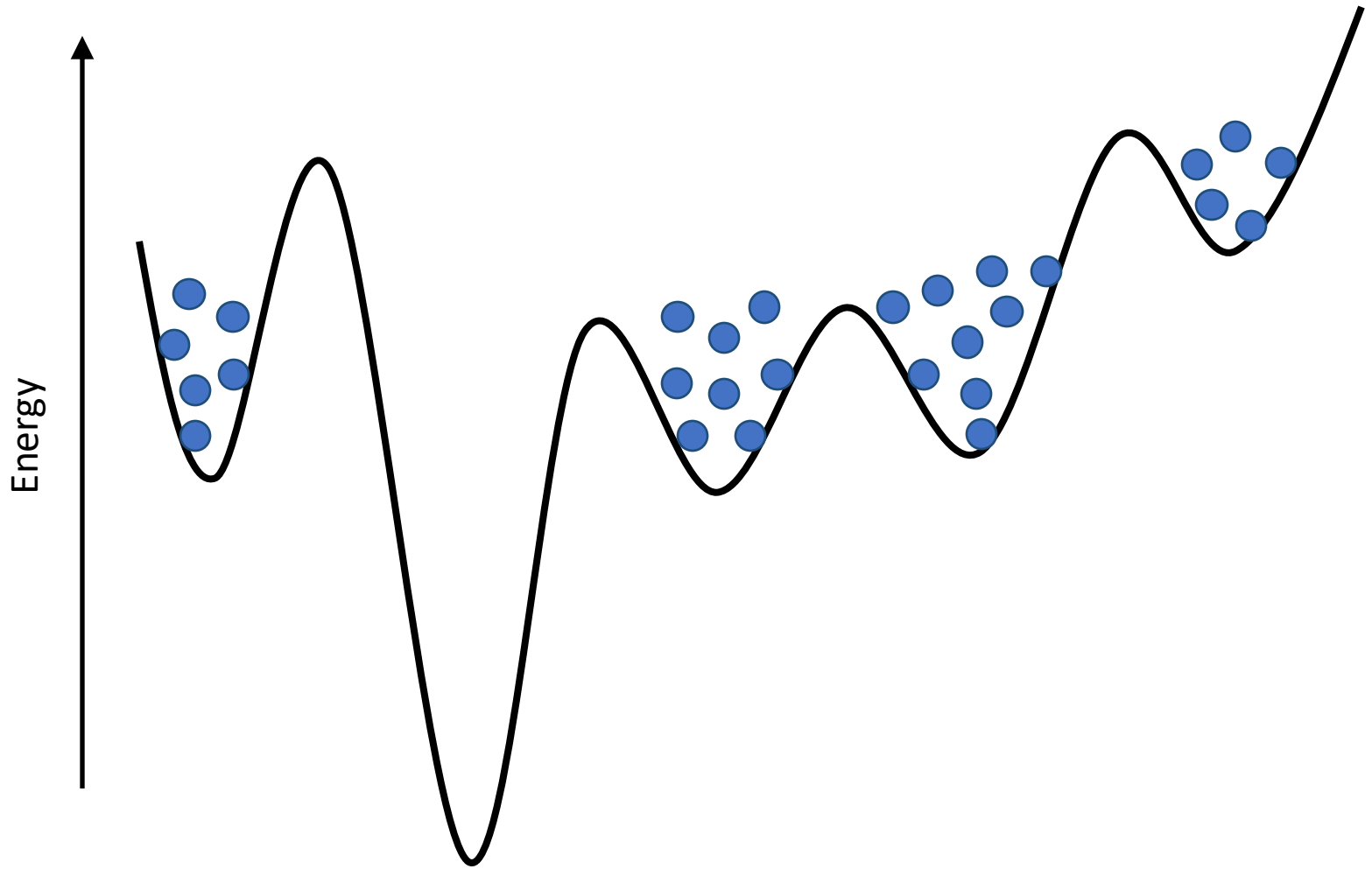
ref2015

Custom energy function

Lowest energy structures converge to same solution



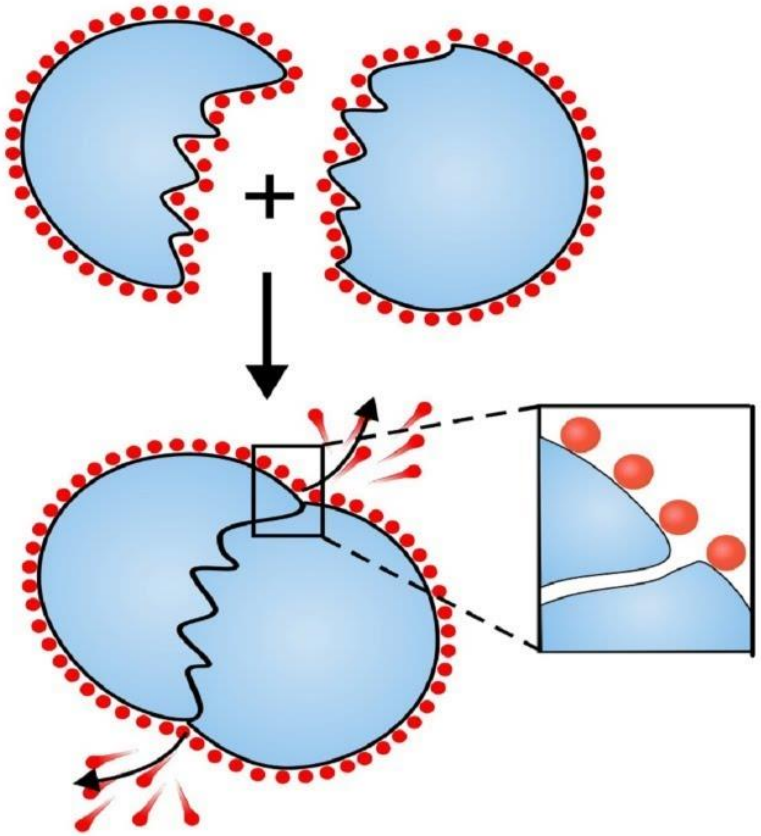
BUT… for many proteins **no** convergence



Why would you see no convergence?

What is the major driving force of protein folding?

The hydrophobic effect



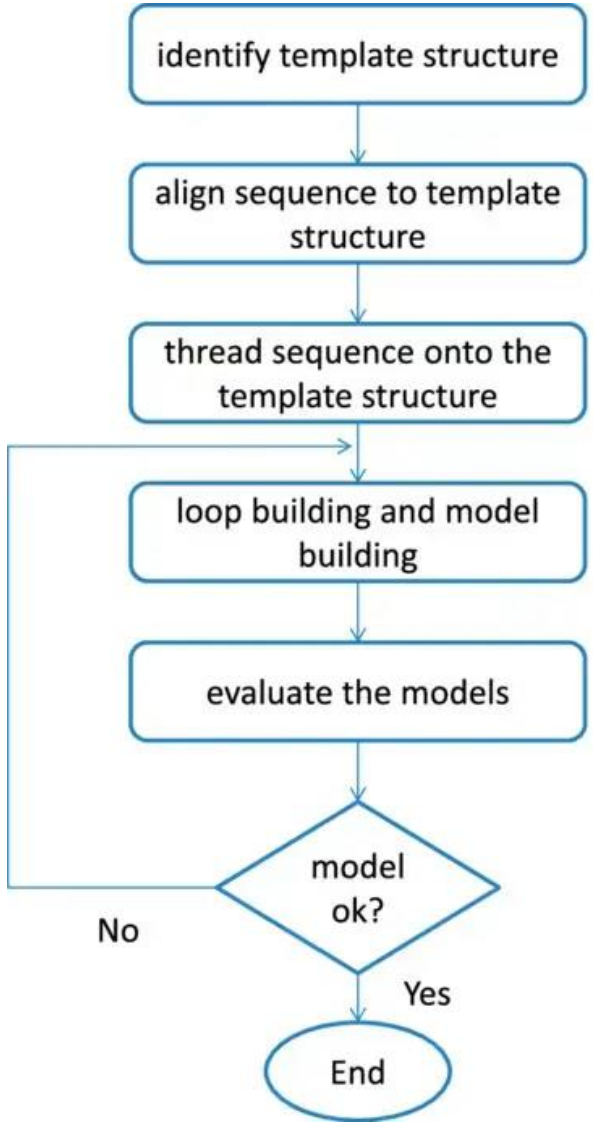
- Entropic effect

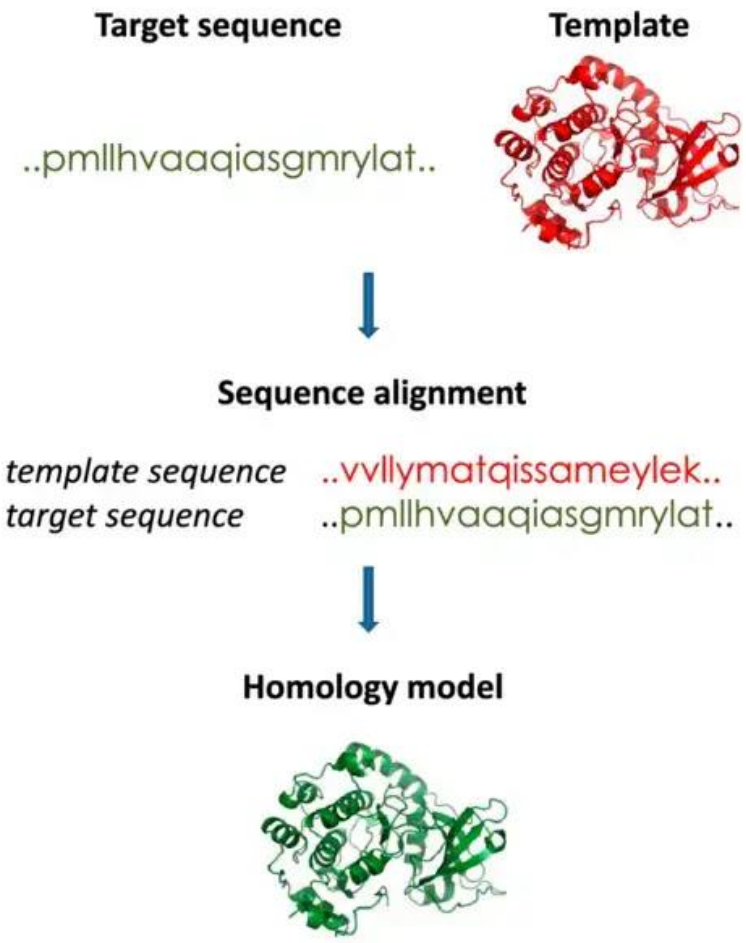
- Water-mediated (water molecules are not explicitely represented to save computing)

- long-range interactions in the protein

➔ The hydrophobic effect is missing from the energy function

Solution? Use a homologous protein with a known structure as a template

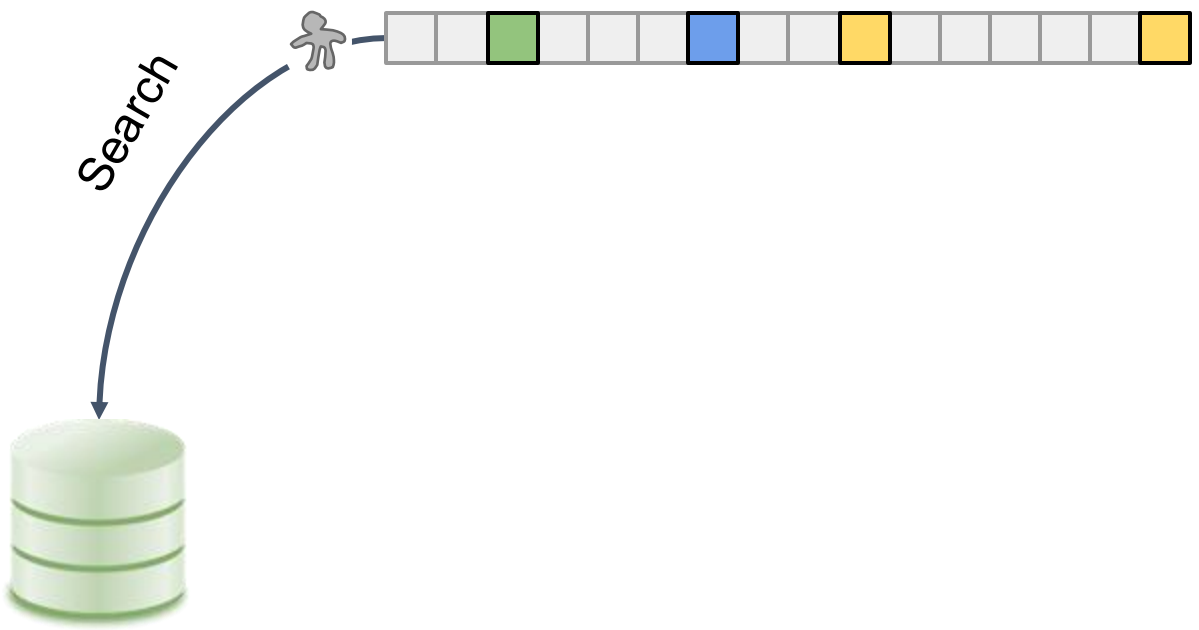




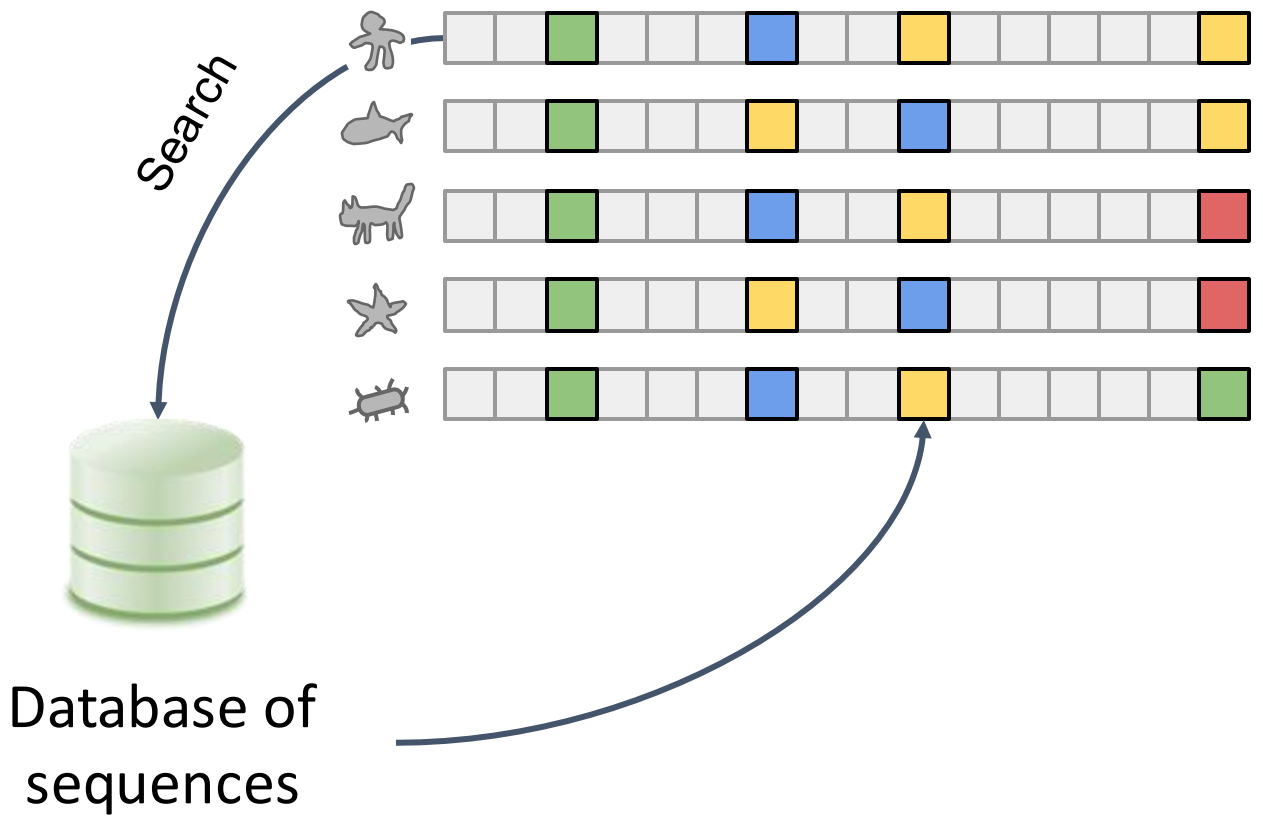
What happens if there is no homolog with a known structure?

We can **CHEAT** by using evolutionary information!

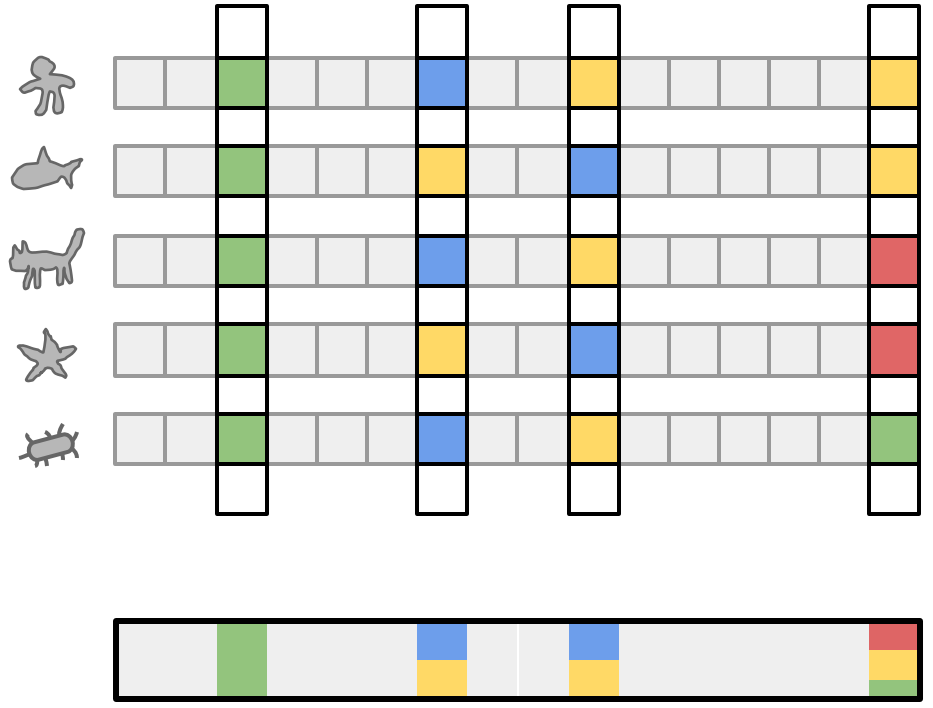
Search against a database of sequences



Generate a multiple sequence alignment

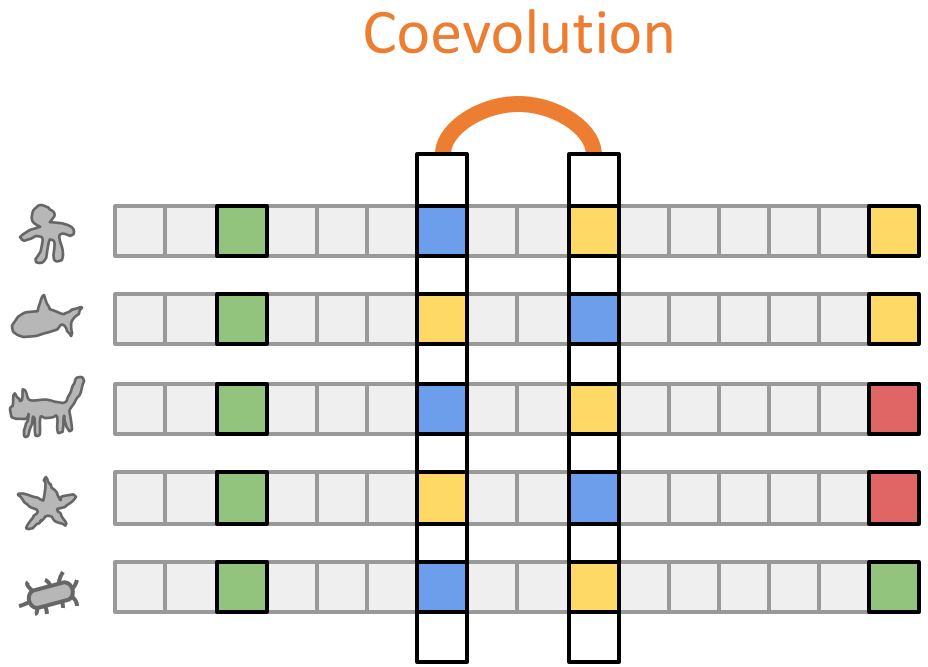


Analyze the MSA for conservation

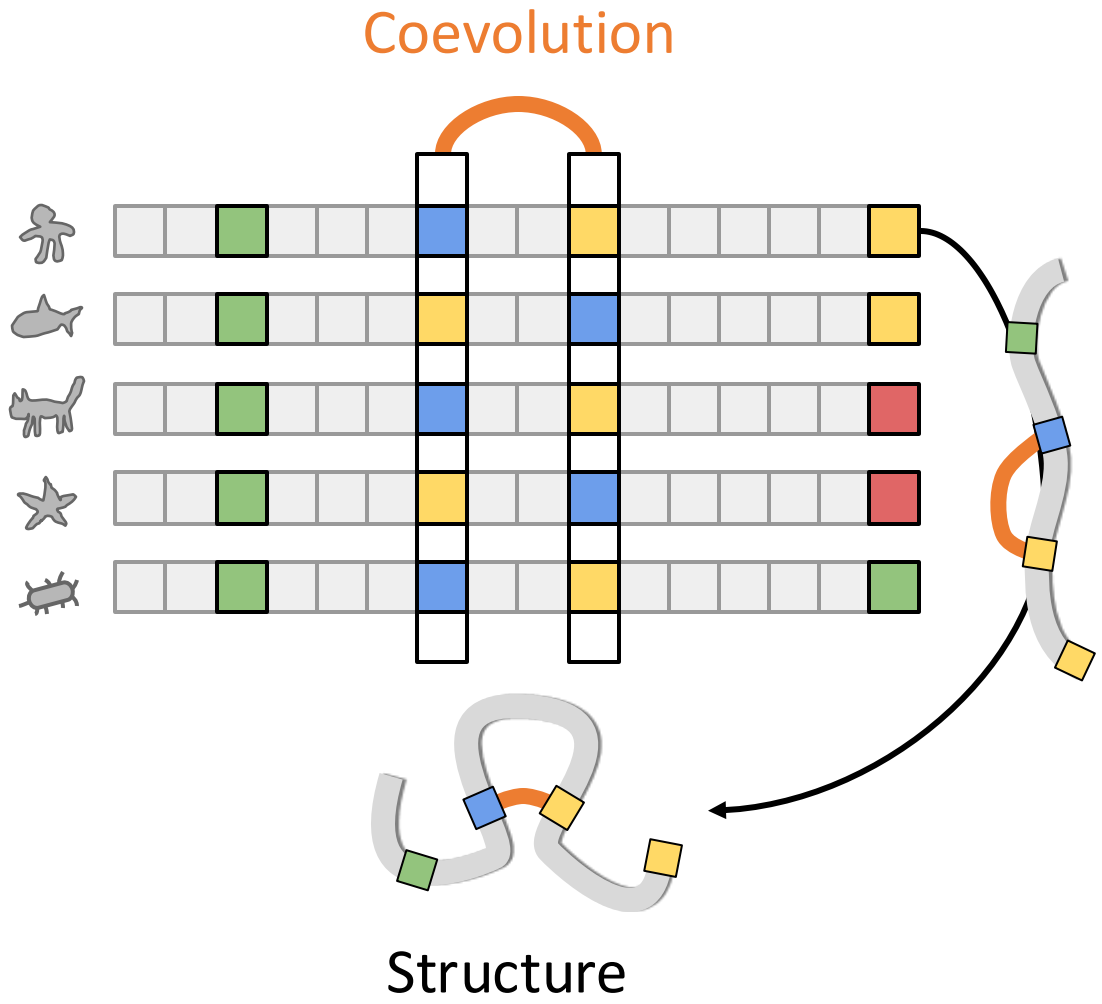


PSSM = Position-specific-scoring matrix

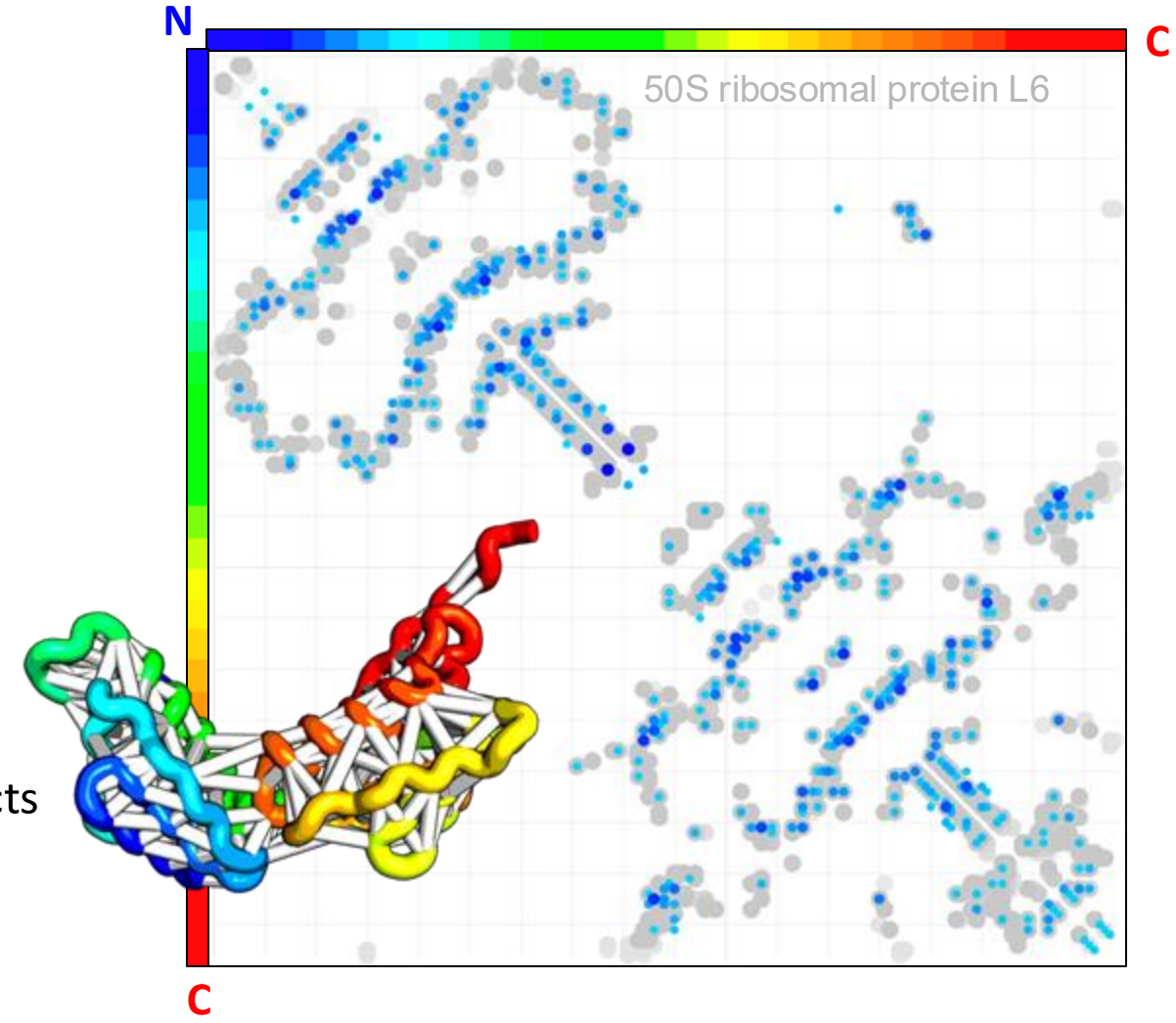
Analyze the MSA for coevolution



Use the as restraints in folding simulations!



The Contact map is correlated to physical contacts

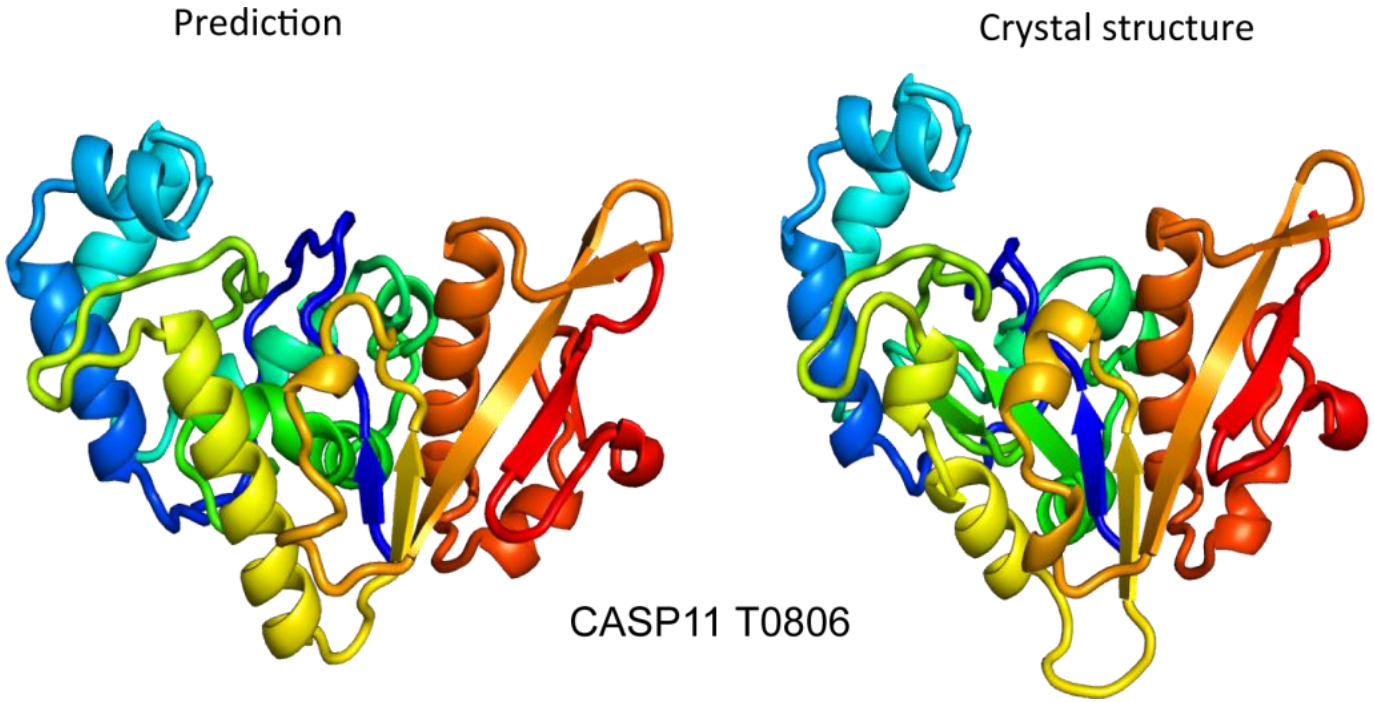


XRAY Contacts

Predicted Contacts

CASP-11 (2014)

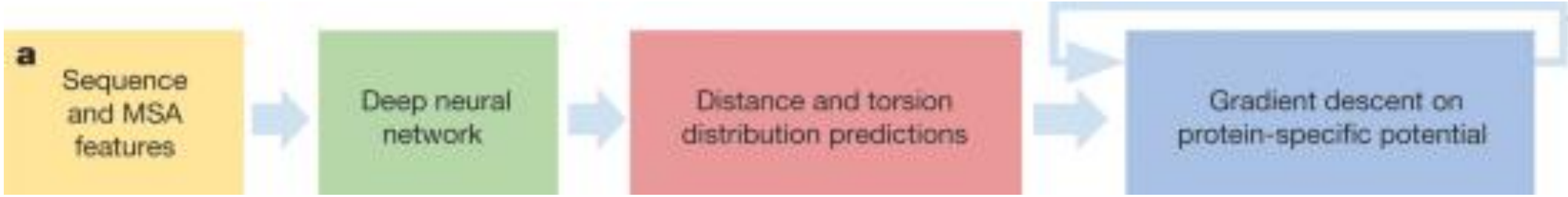
Sergey Ovchinnikov, David Baker

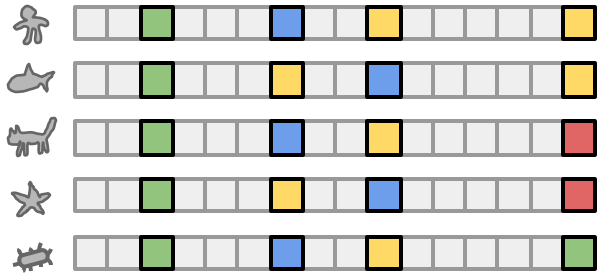


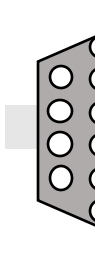
https://doi.org/10.7554/eLife.02030

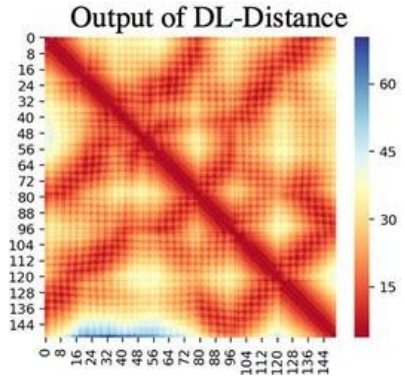
*S*tructure prediction:
Introducing AI

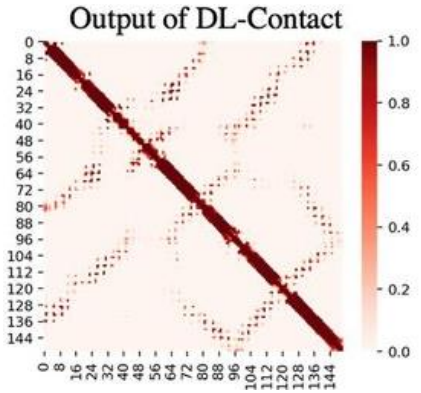
AlphaFold1: coevolution contacts from MSA



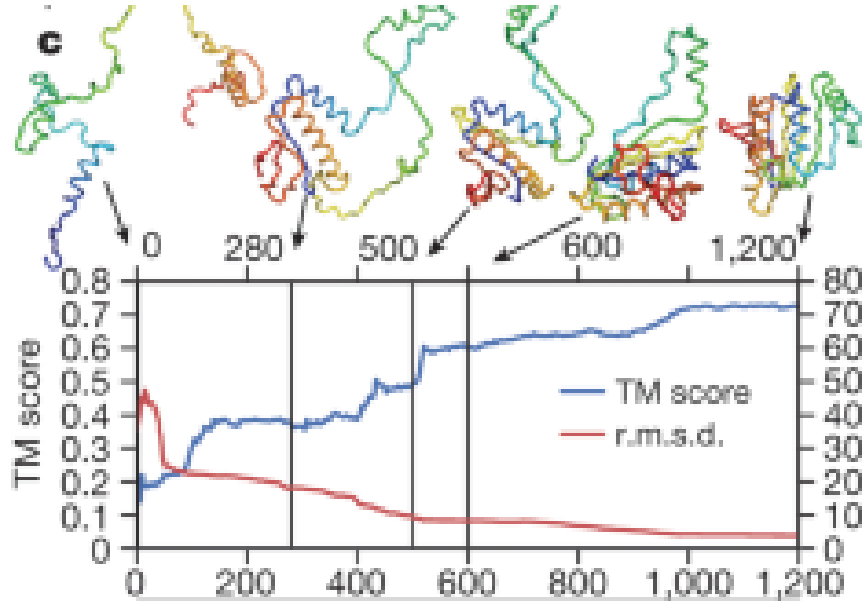








No physical parameters !!



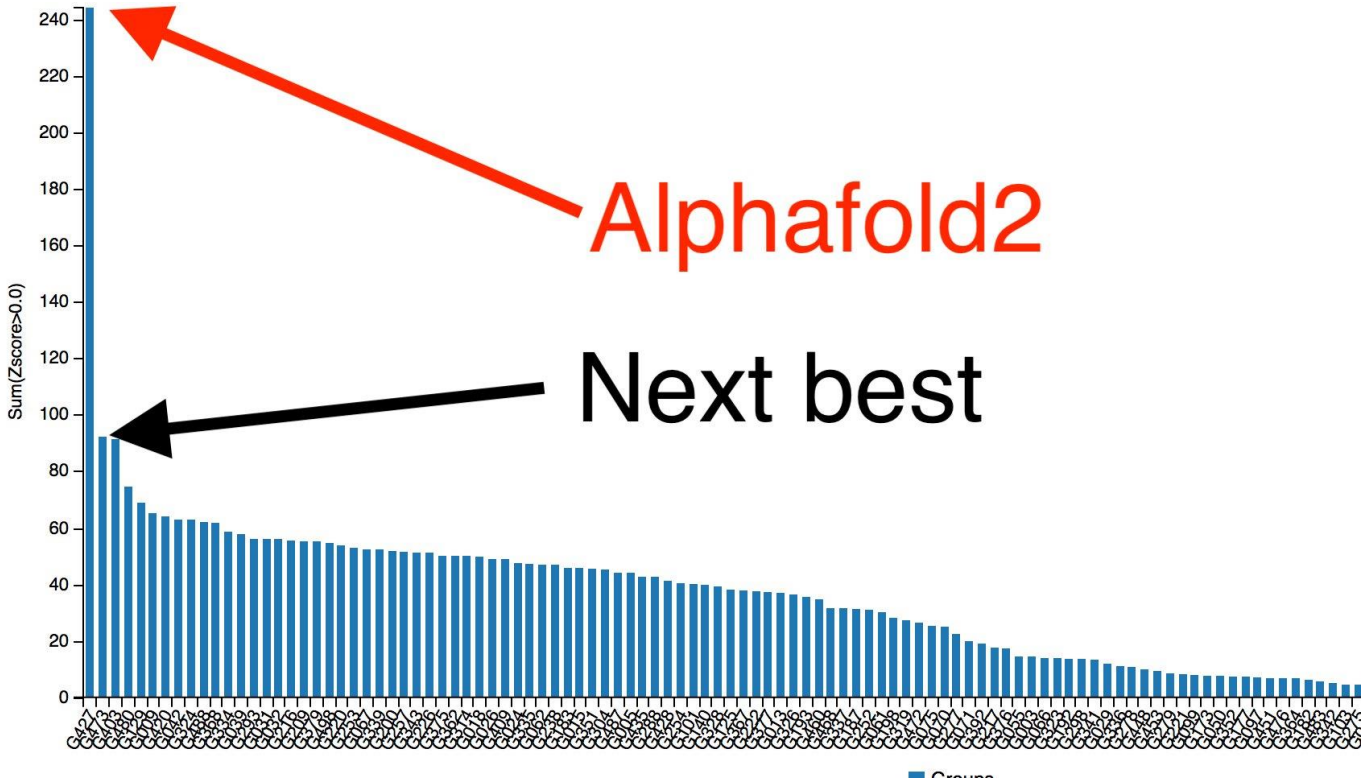
Minimize the restraints

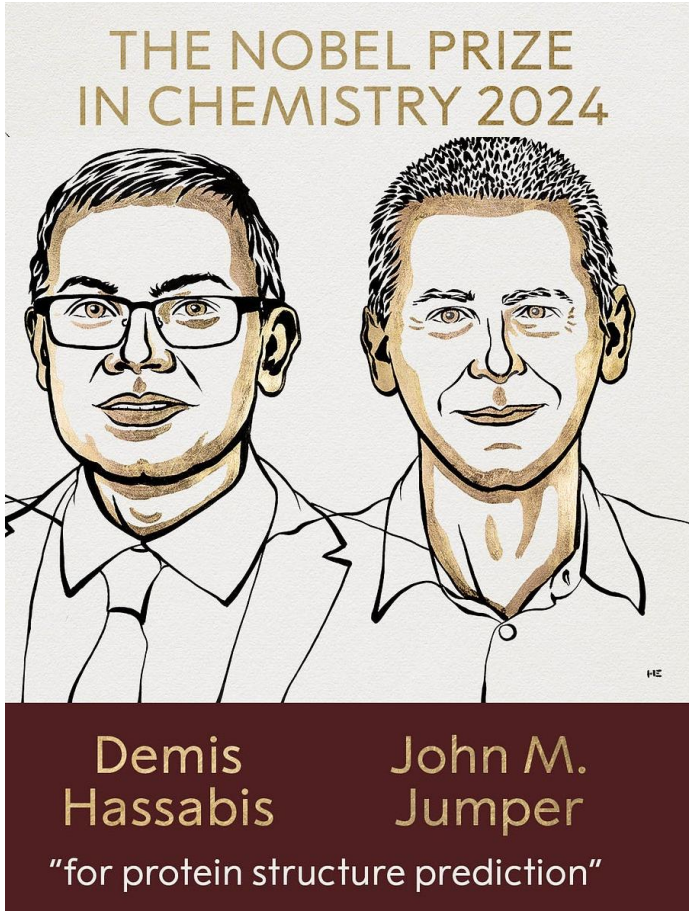
Contacts prediction module

Structure module

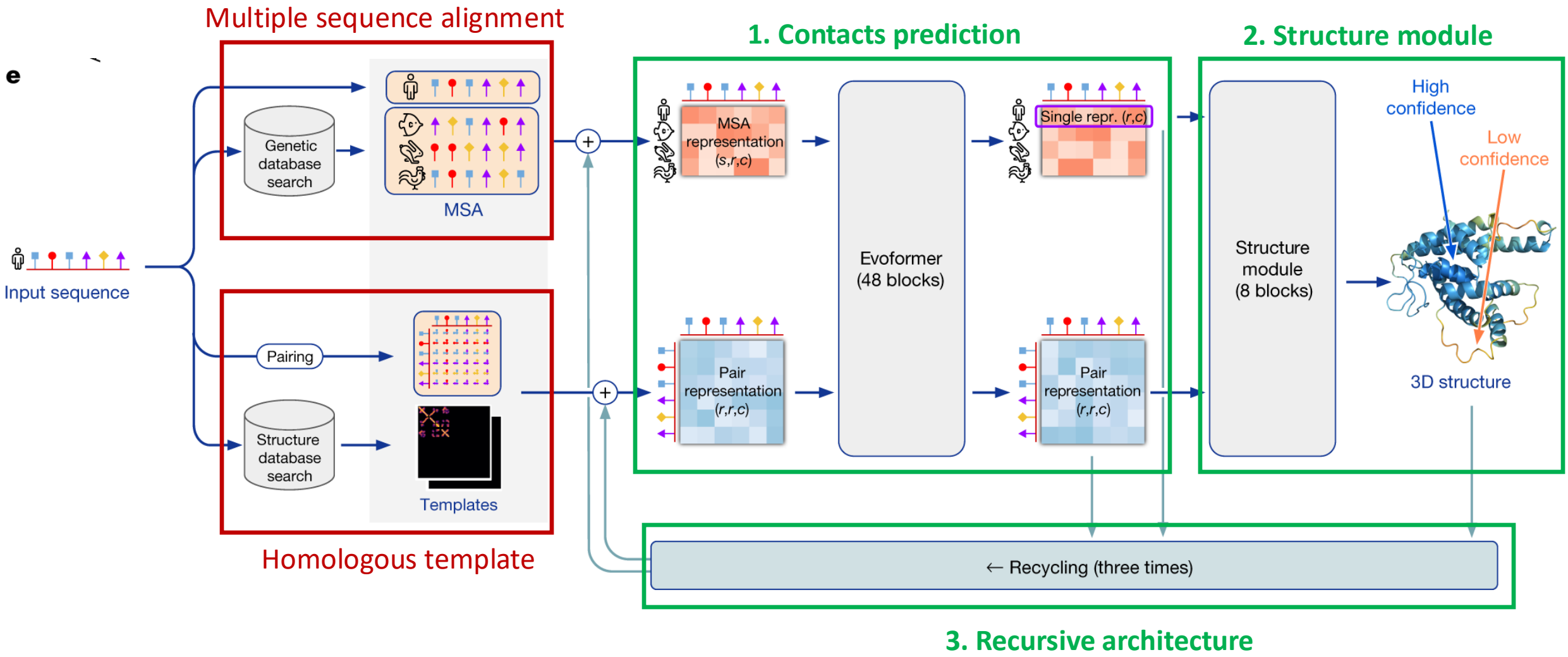
https://doi.org/10.1038/s41586-019-1923-7

AlphaFold2 on CASP-13 competition (2020)





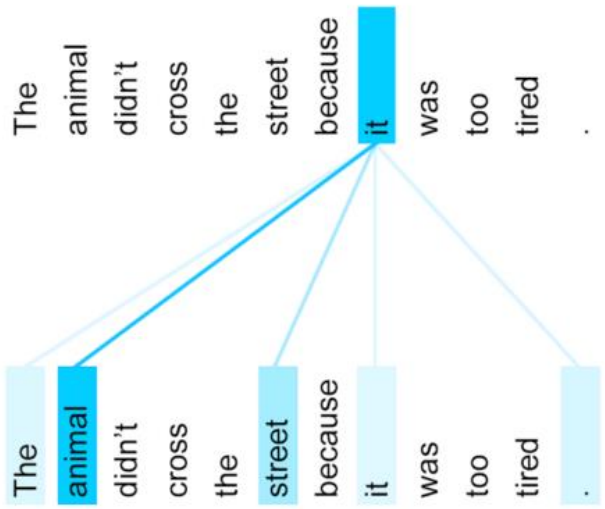
AlphaFold 2 - Architecture

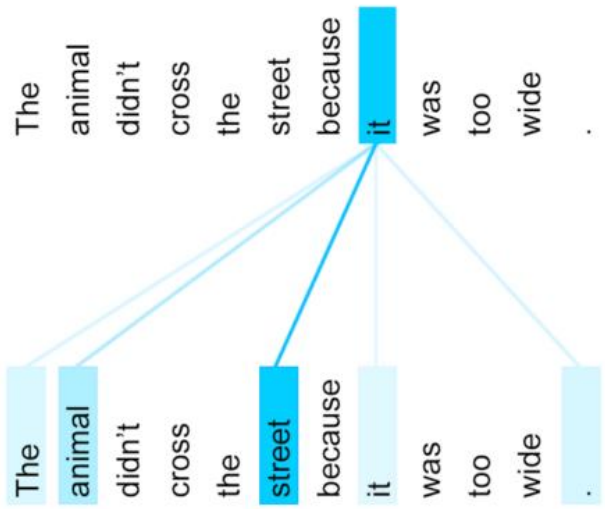


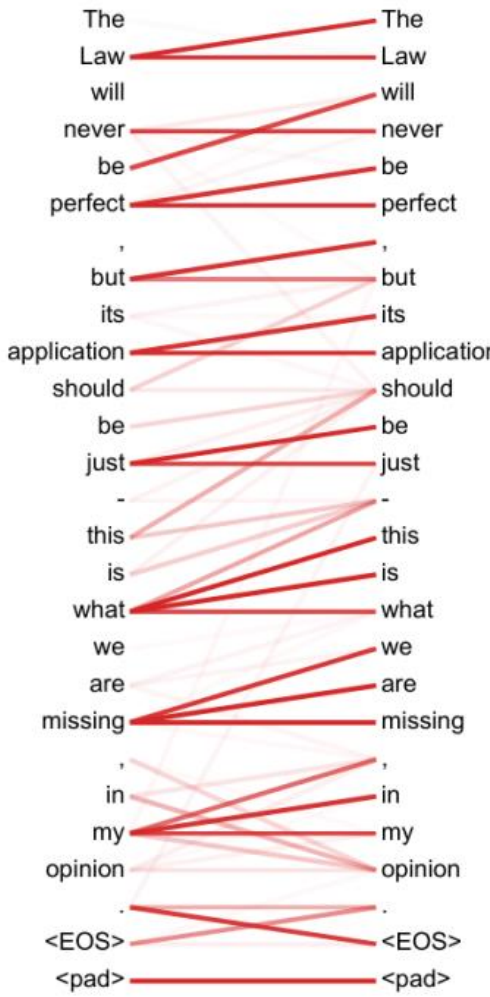
https://doi.org/10.1038/s41586-021-03819-2

1. Contacts prediction module

A **TRANSFORMER** with an **ATTENTION** mechanism



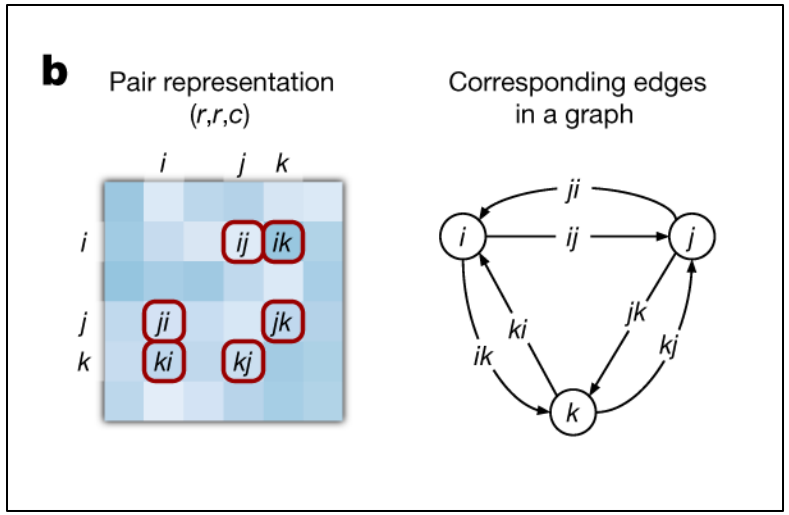


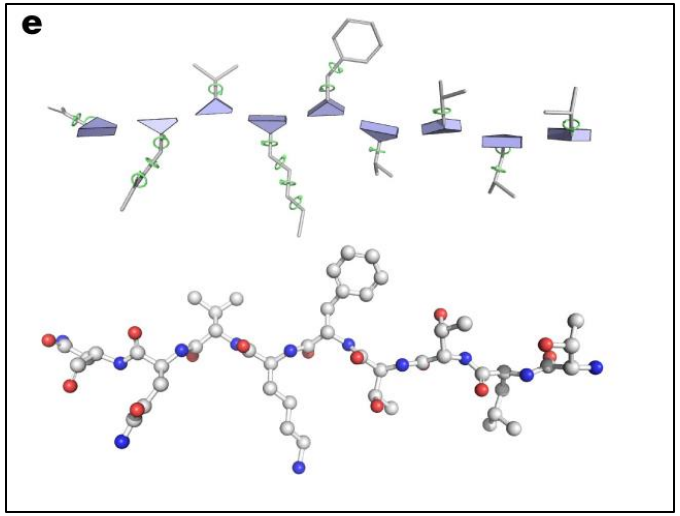


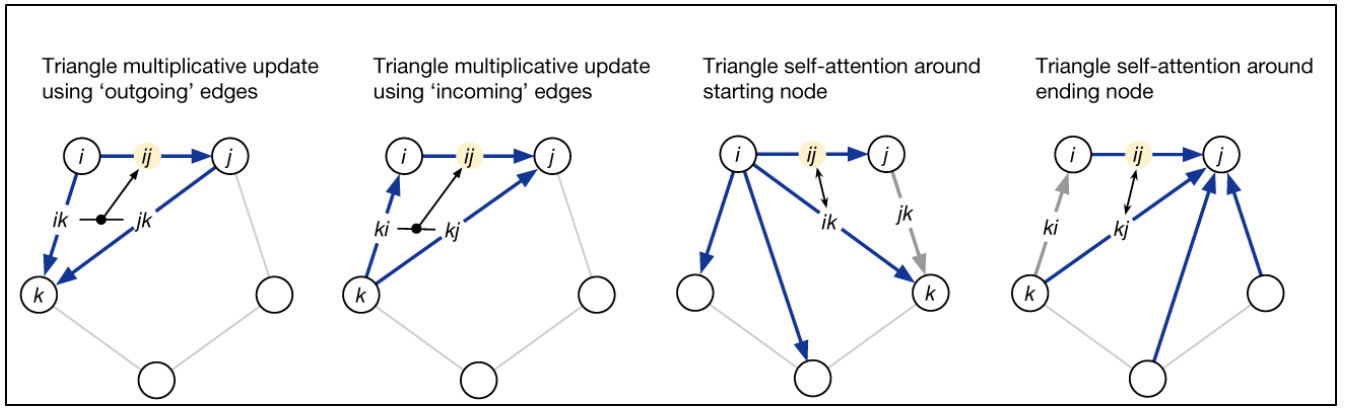
The **TRANSFORMER** that predicts contacts from evolution relationships ➔ **EVOFORMER**

2. Structure prediction module

All residues move independently in the space to satisfy the predicted contacts

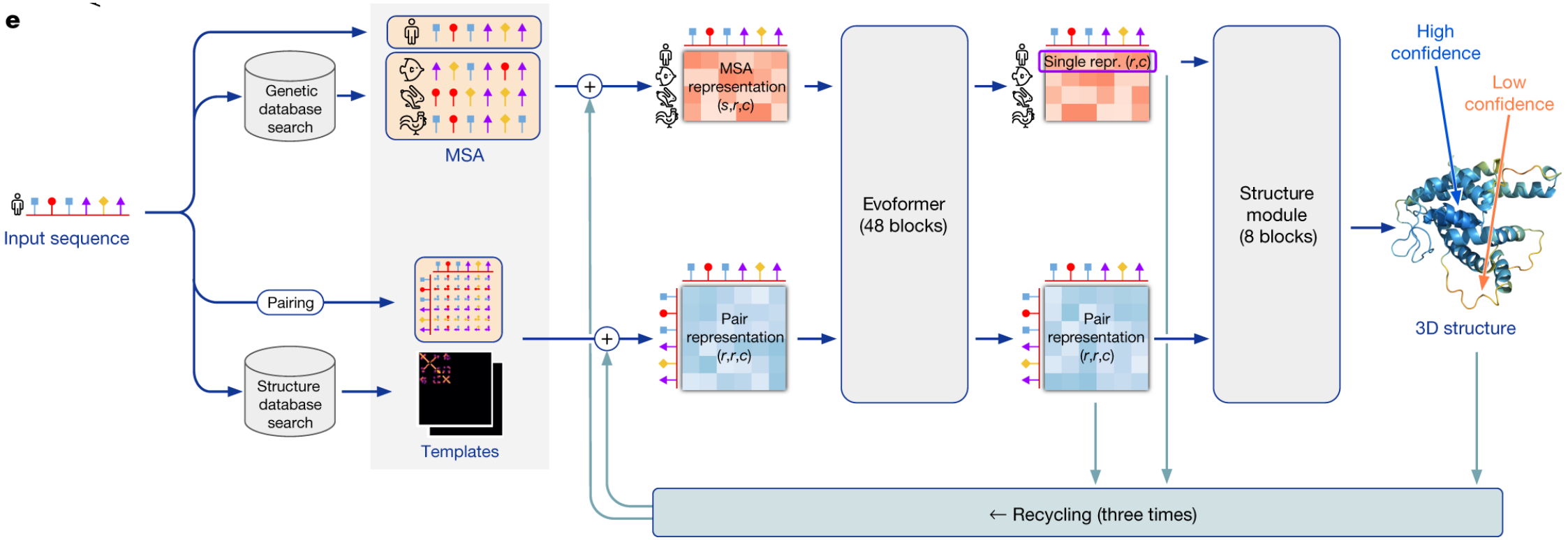






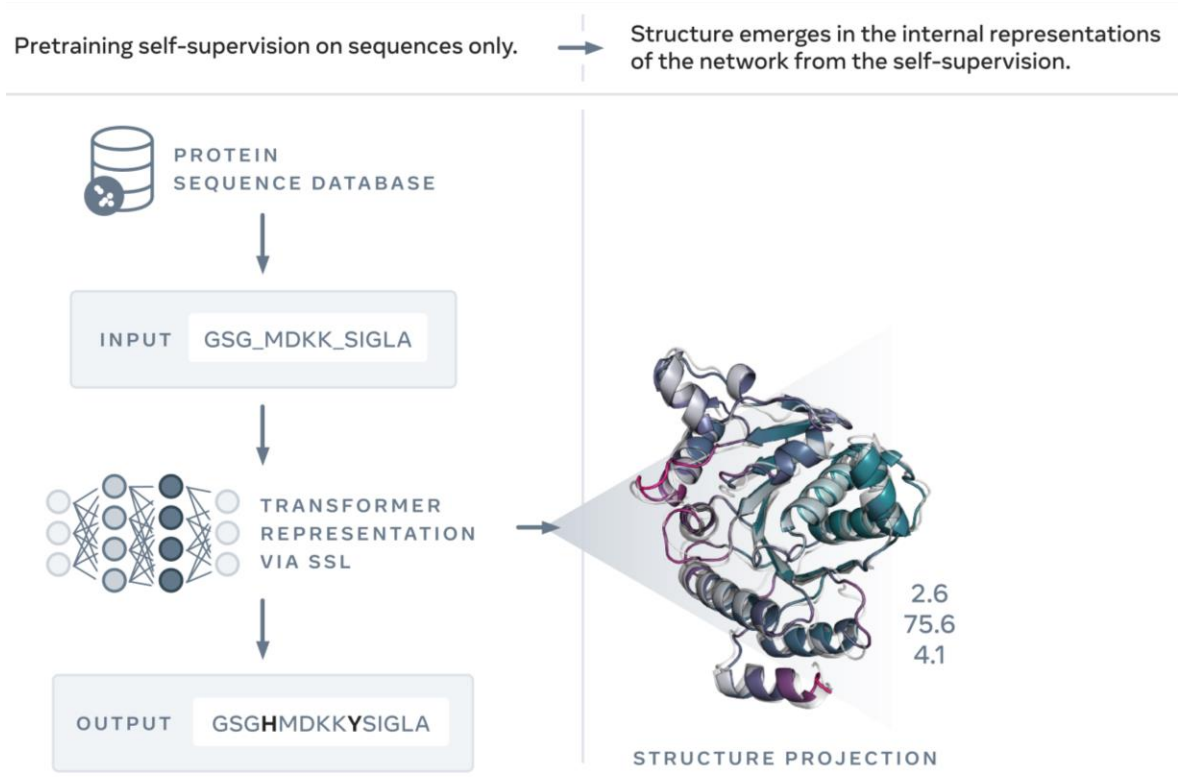
3. Recursive architecture to refine the solution

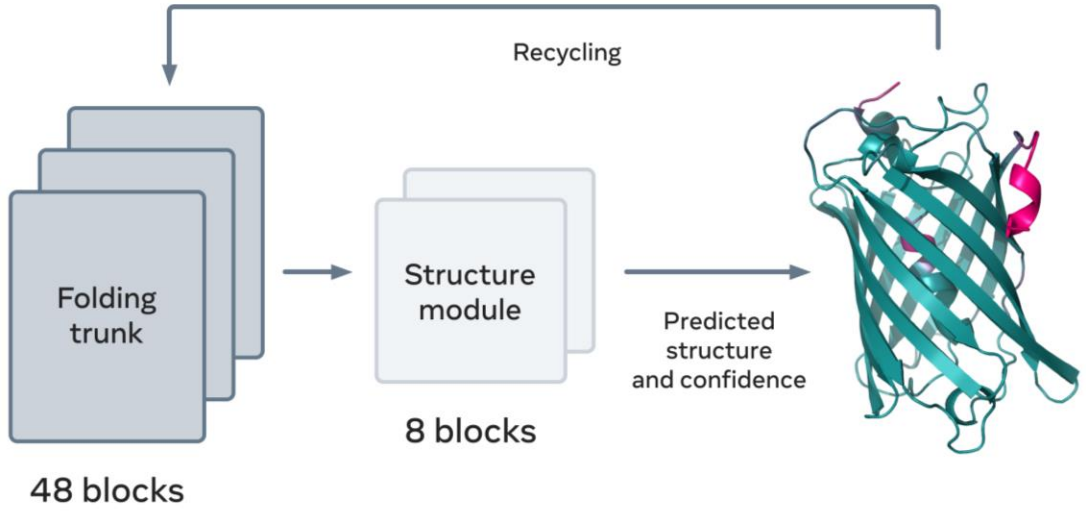
A multiple sequence alignment is a **HYPOTHESIS**



RECYCLING to refine the predicted structure based on local confidence assignments

Other AI models: Structure prediction with a protein language model (e.g. ESMFold from Meta AI)





Less accrate than AlphaFold2 but predicts structure from a single sequence (rather than from MSA)

https://doi.org/10.1126/science.ade2574