Exercises

What determines protein stability? Electrostatics (ion-ion and hydrogen bonds), van der Walls, hydrophobicity

List two types of secondary structure and what type of interactions mediate those structures? alpha helix, beta sheet, H-bonds

How do we know a protein's structure is determined by amino acid sequence? It refolds to native state in vitro

Solvant (e.g. water) are not important to protein structure/stability [T/F]

What is force plays a dominant role in protein folding? Hydrophobicity

Entropy important in protein folding [T/F]?

Today's objectives

- Protein structure and function
- Protein structure prediction
- BioPhysics, homology modeling, evolution
- Predicting changes in structure

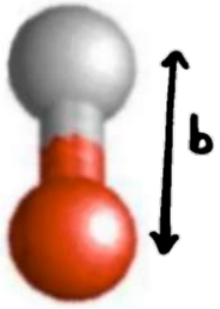
Molecular mechanics force fields

Molecular mechanics force

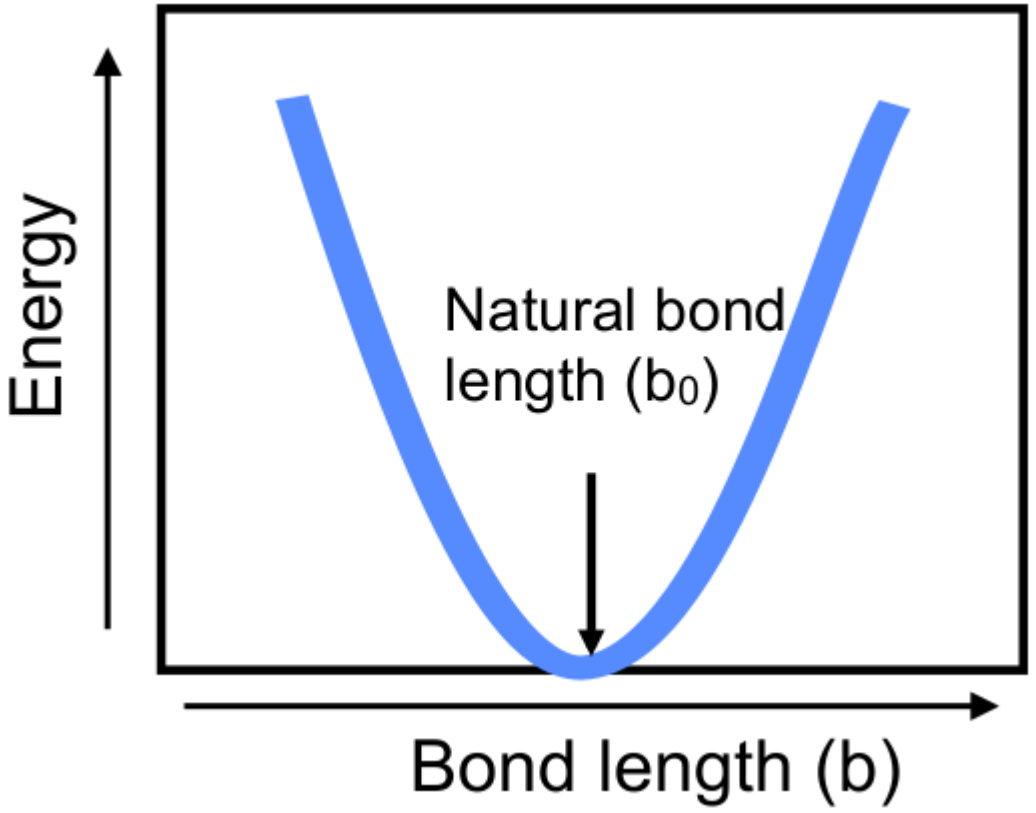
- used for molecular dynamics simulations
- more toward the physics-based, all-atom end (i.e., the more “realistic” force fields)
  - – Represent physical forces explicitly
  - – Typically represent solvent molecules (e.g., water) explicitly
- Forces:
  - Bond length stretching
  - Bond angle bending
  - Torsional angle twisting
  - Electrostatics interaction
  - van der Waals interaction

Bond length stretching

- A bonded pair of atoms is effectively connected by a spring with some preferred (natural) length.
- Stretching or compressing it requires energy.

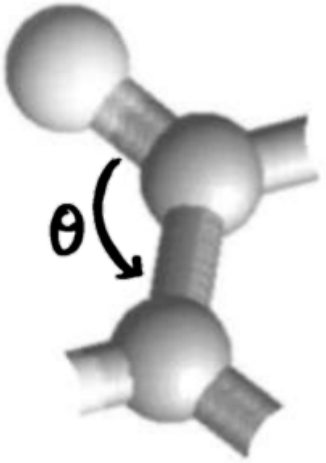


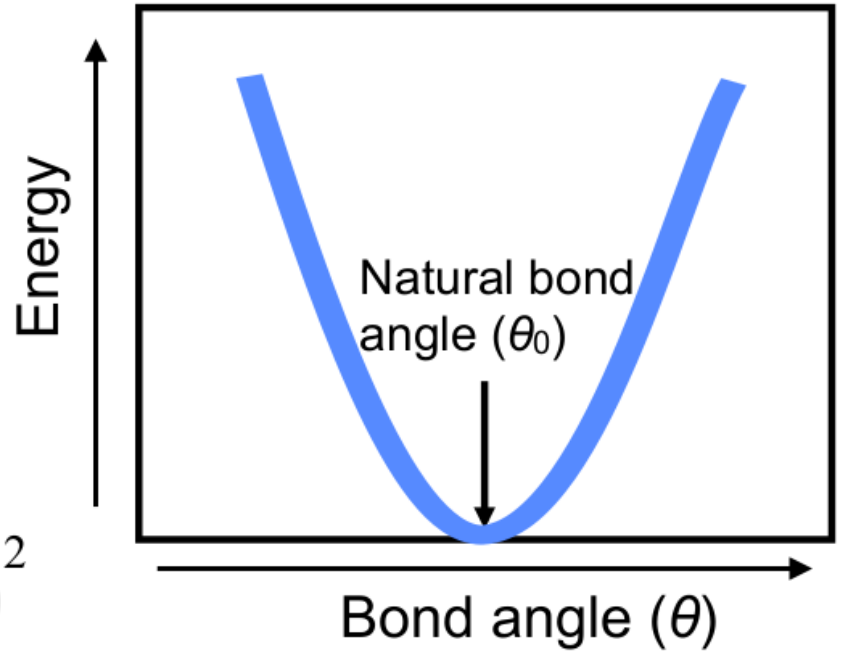
$$U(b) = k_b\left(b - b_0\right)^2$$



Bond angle bending

Each bond angle has some natural value. Increasing or decreasing it requires energy.

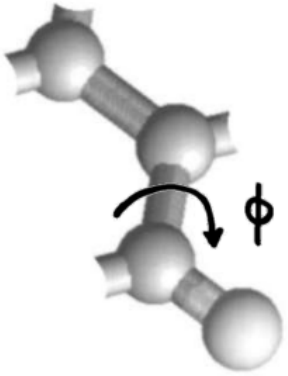


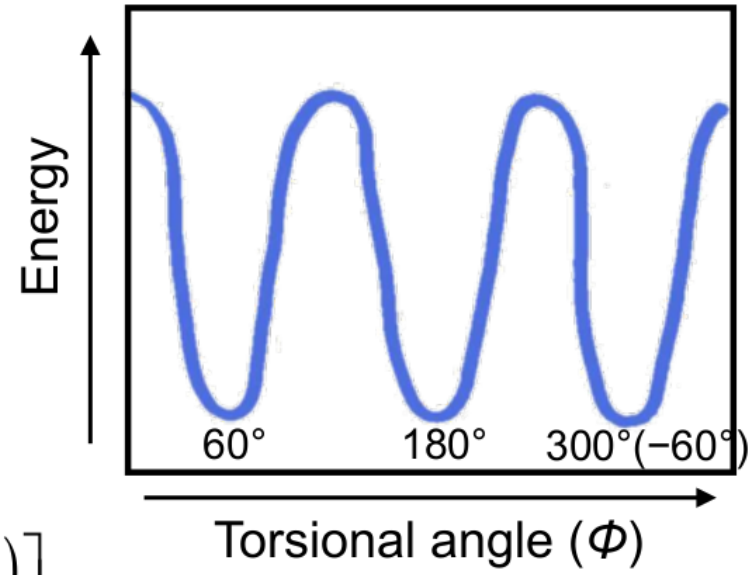


$$U(\theta) = k_\theta \left(\theta - \theta_0\right)^2$$

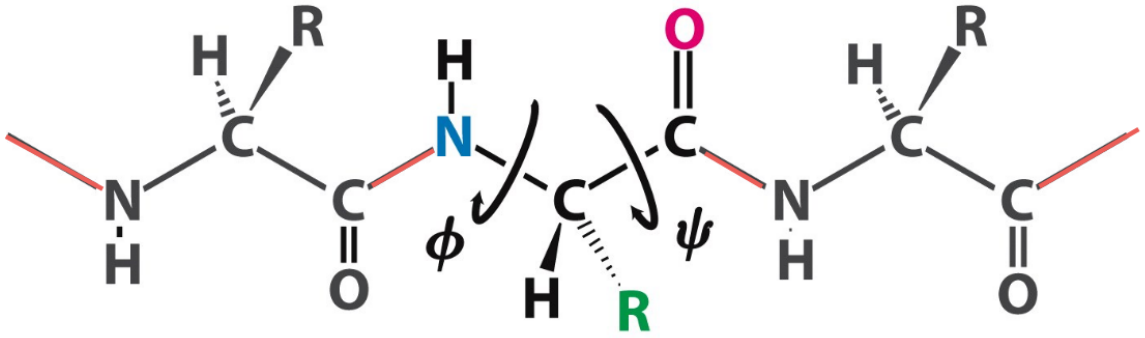
Torsional angle twisting

Certain values of each torsional angle are preferred over others.



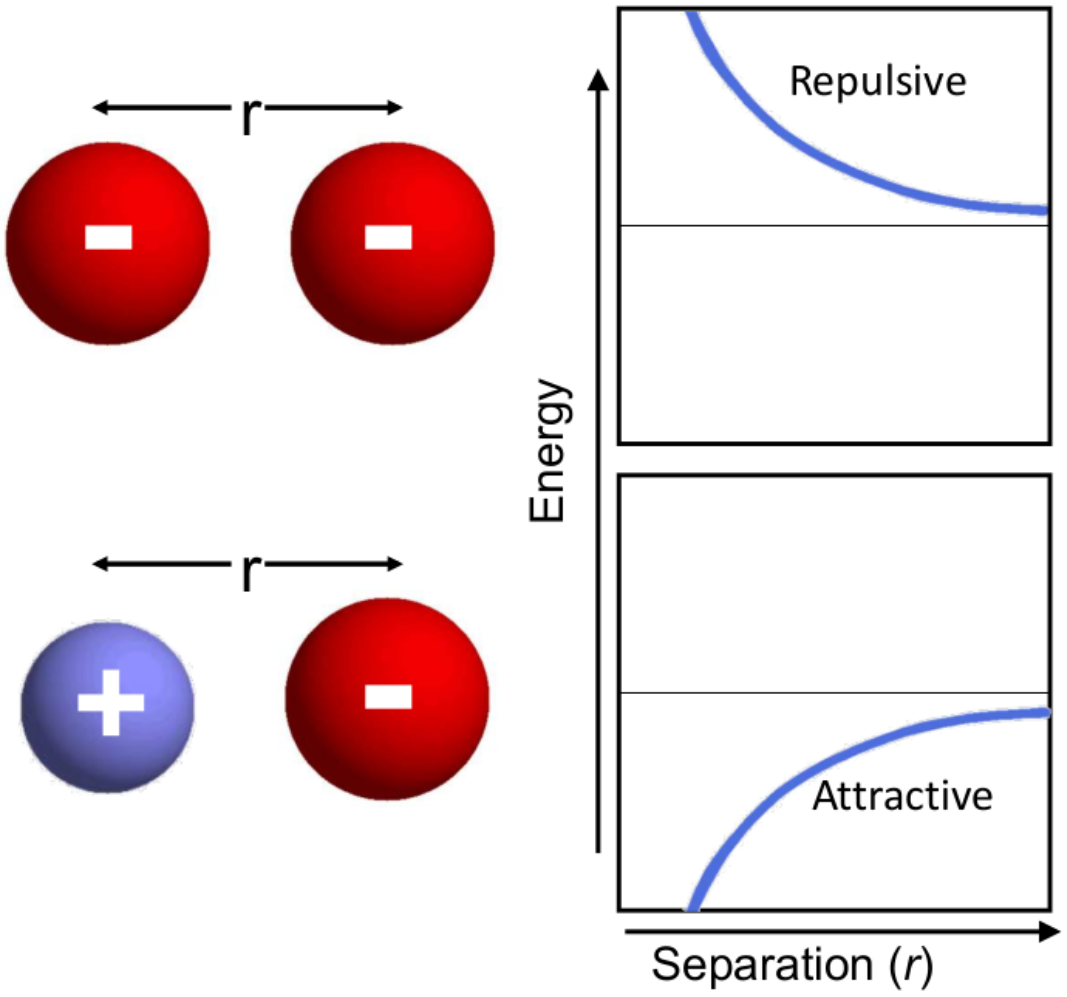


$$U(\phi) = \sum_n k_{\phi,n}\left[1 + \cos\left(n\phi - \phi_n\right)\right]$$



Only two bonds can freely rotate

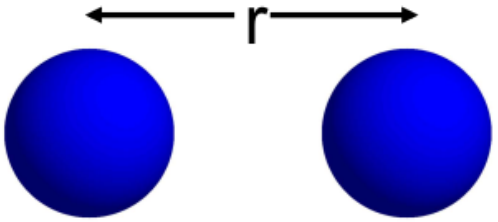
Electrostatics

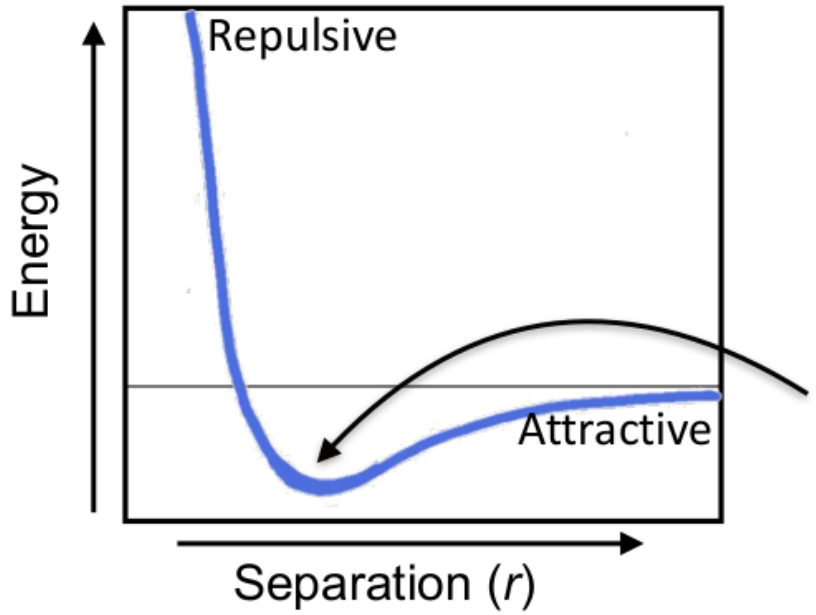


- Acts between all pairs of atoms, including those in different molecules.
- Each atom carries some "partial charge" (may be a fraction of an elementary charge), which depends on which atoms it's q_i q_j, where q_i and q_j are partial charges.

$$U(r) = \frac{q_i q_j}{r}$$

van der Waals interaction





- van der Waals forces act between all pairs of atoms and do not depend on charge.
- When two atoms are too close together, they repel strongly.
- When two atoms are a bit further apart, they attract one another weakly.

$$U(r) = \frac{A_{ij}}{r^{12}} - \frac{B_{ij}}{r^{6}}$$

Energy is minimal when atoms are "just touching" one another

Molecular mechanics force fields

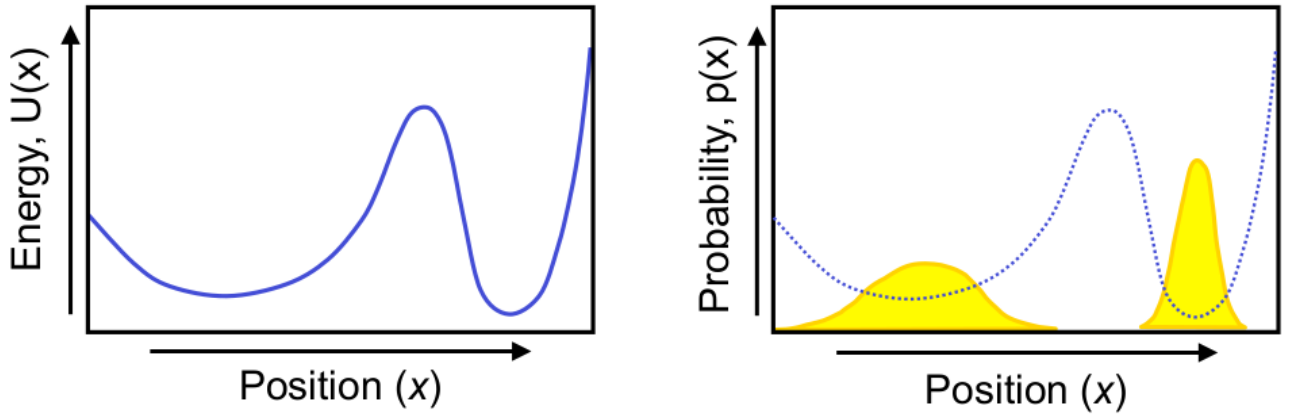
Parameters come from a combination of:
– Quantum mechanical calculations
– Experimental data (knowledge-based potentials from protein structures in the Protein Data Bank)

Given the potential energy associated with a particular arrangement of atoms (set of atom positions), what is the probability that we'll see that arrangement of atoms?

The Boltzmann distribution relates potential energy to probability

$$p(\boldsymbol{x}) \propto \exp\left({-U(\boldsymbol{x})}/{k_B T} \right)$$

k_B is the Boltzmann constant



Application of force fields

Protein structure prediction (constrained application)

- Folding timescales are usually much longer than simulation timescales
- But force fields can be used to 'evaluate' energy of a proposed structure
- Can be used for structure refinement with course grain approximations

Molecular dynamics simulations (many applications)

- Divide time into discrete time steps, no more than a few femtoseconds (10e–15 s) each, or inaccurate
- Require multipl CPU days to simulate nanoseconds (10e−9 s) to microseconds (10e−6 s), $O(n^2)$ n=particles
- optimize potential rather than free energy, doesn't include entropy and hydrophobic
- many approximations, some improvement in structure prediction, but not yet practical
- Computationally intensive (non-bond interactions), GPUs widely used

Two approaches to protein structure prediction

1) Template-based modeling (homology modeling)

– Used when one can identify one or more likely homologs of known structure

2) *Ab initio* structure prediction

– Does not require any homologs

– Even ab initio approaches usually take advantage of available structural data, but in more subtle ways

Template (homology) predictions

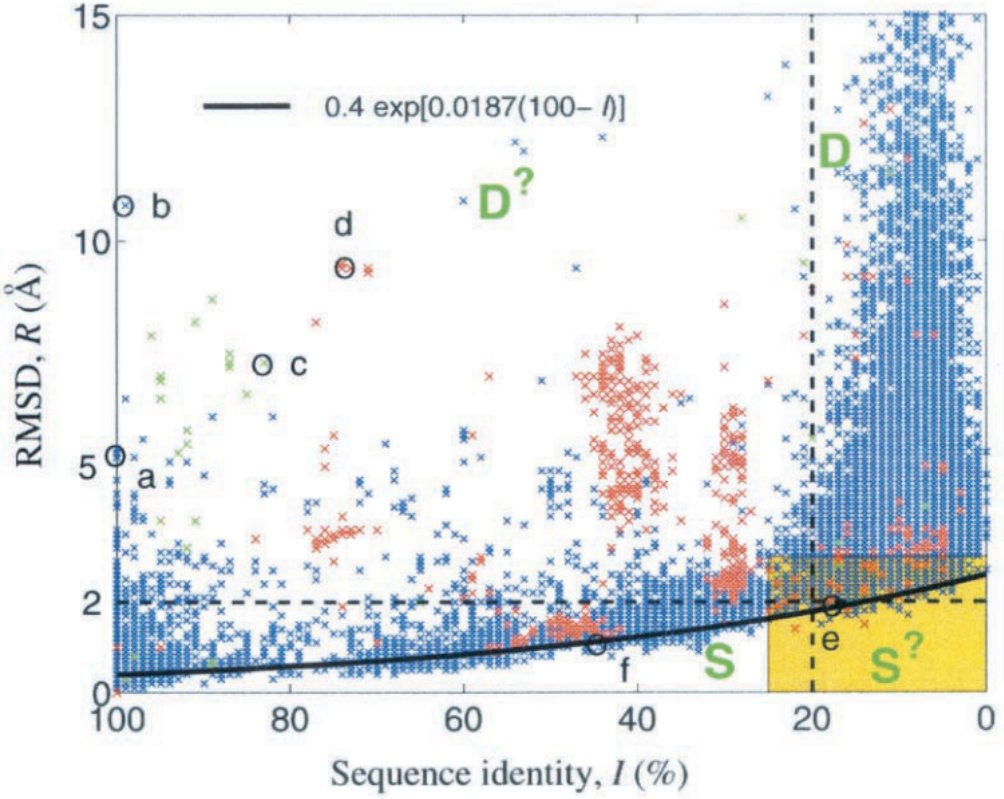
Given a query sequence with unknown structure

- Search the PDB for proteins with similar sequence and known structure. Pick the best match (the template).
- Build a model based on that template

  – One can also build a model based on multiple templates, where different templates are used for different parts of the protein.
- Assessment of homology models without reference to the true target structure using either statistical potentials or physics-based energy calculations

Structures conserved for even distant homologs

- Above 50% sequence identity, models tend to be reliable
- Below 20% identity "twilight zone", serious errors occur
- Error in homology modeling are poor template selection and inaccuracies in target-template sequence alignment
- Best methods use profile models (HMMs, e.g. HHblits), most distant homologs

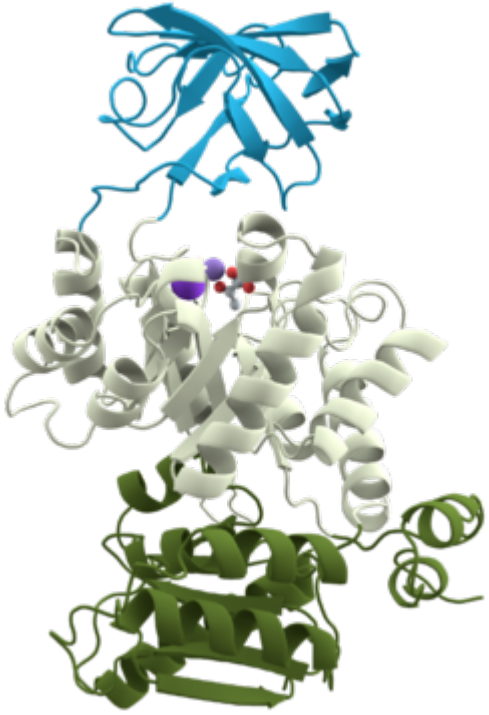


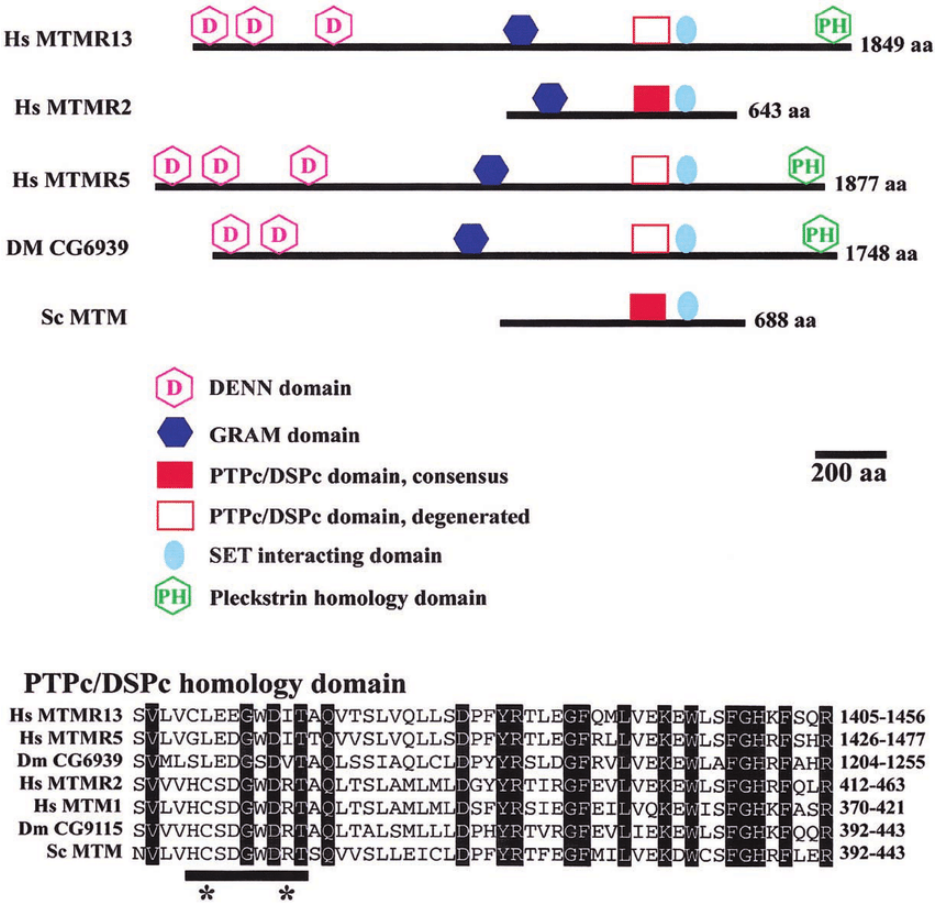
(e) Hemoglobins, S?
1abwA (green)/3sdhA (magenta)
(19%, 1.9Å)

Homology models: Domains

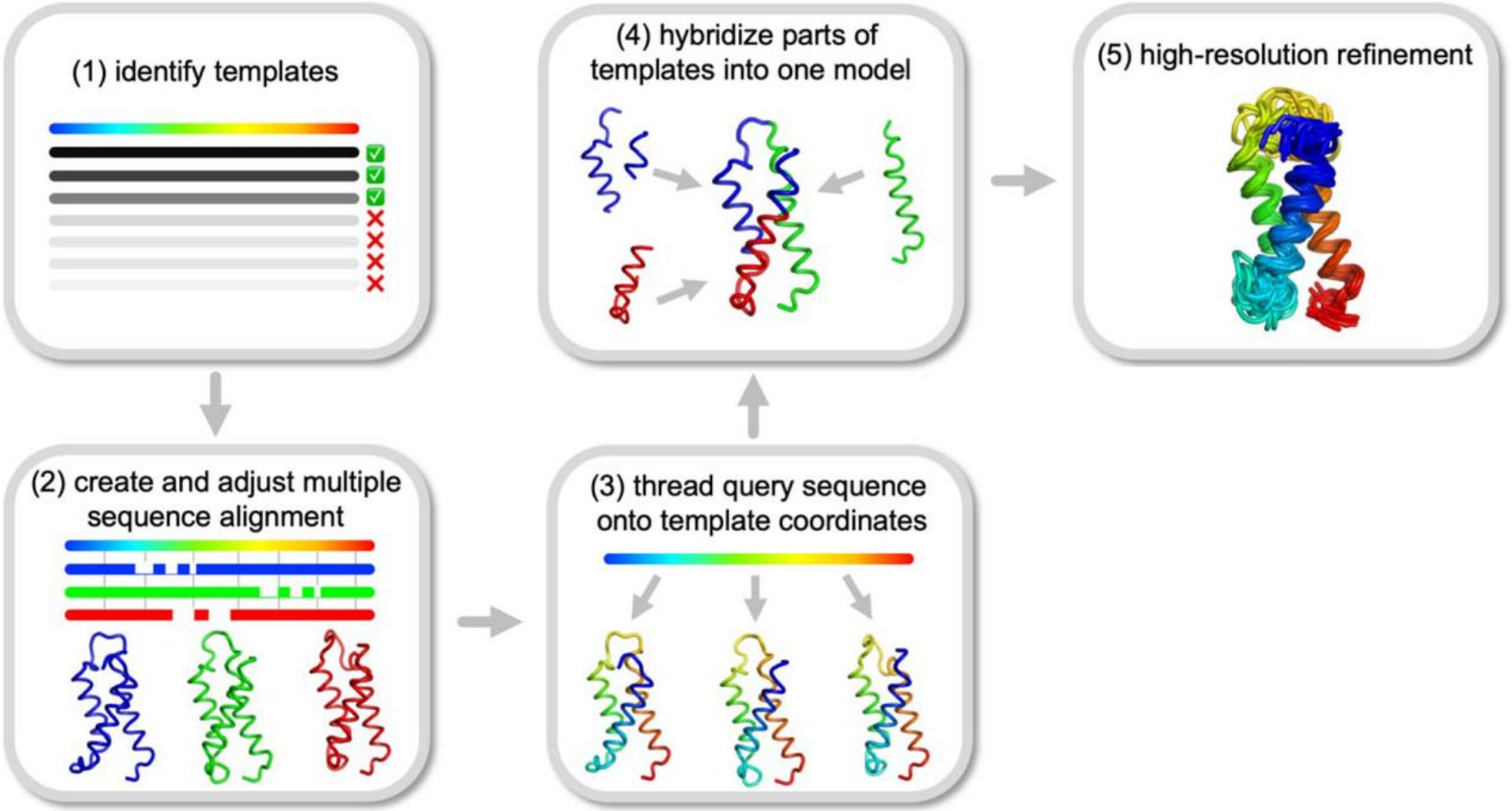
- A protein domain is a region of the protein's polypeptide chain that is self-stabilizing and that folds independently from the rest
- Can have their own function (e.g. regulatory, binding), and be swapped and shared with other proteins
- Homologous protein domains useful for homology modeling

Pyruvate kinase (3 domains)

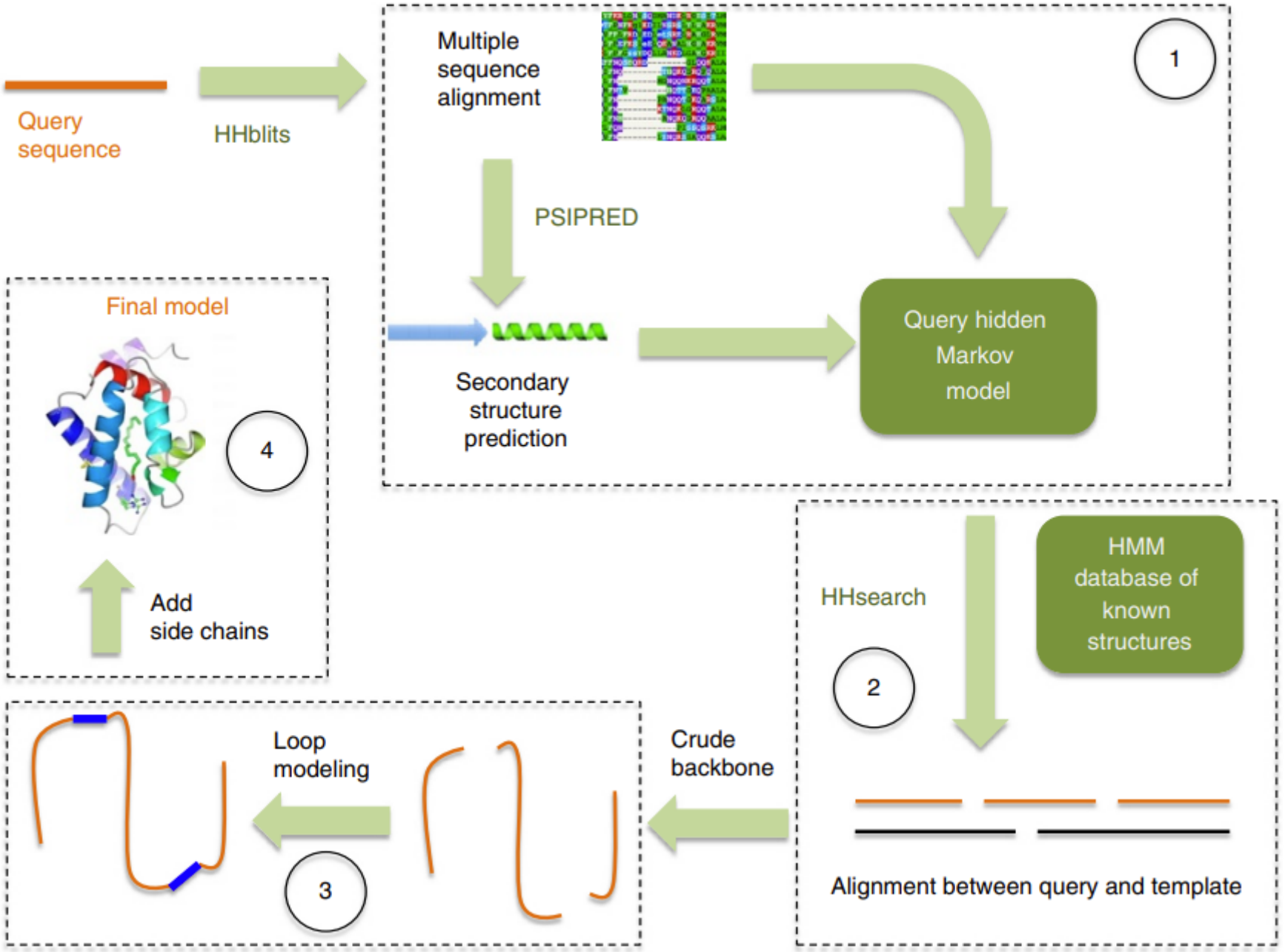




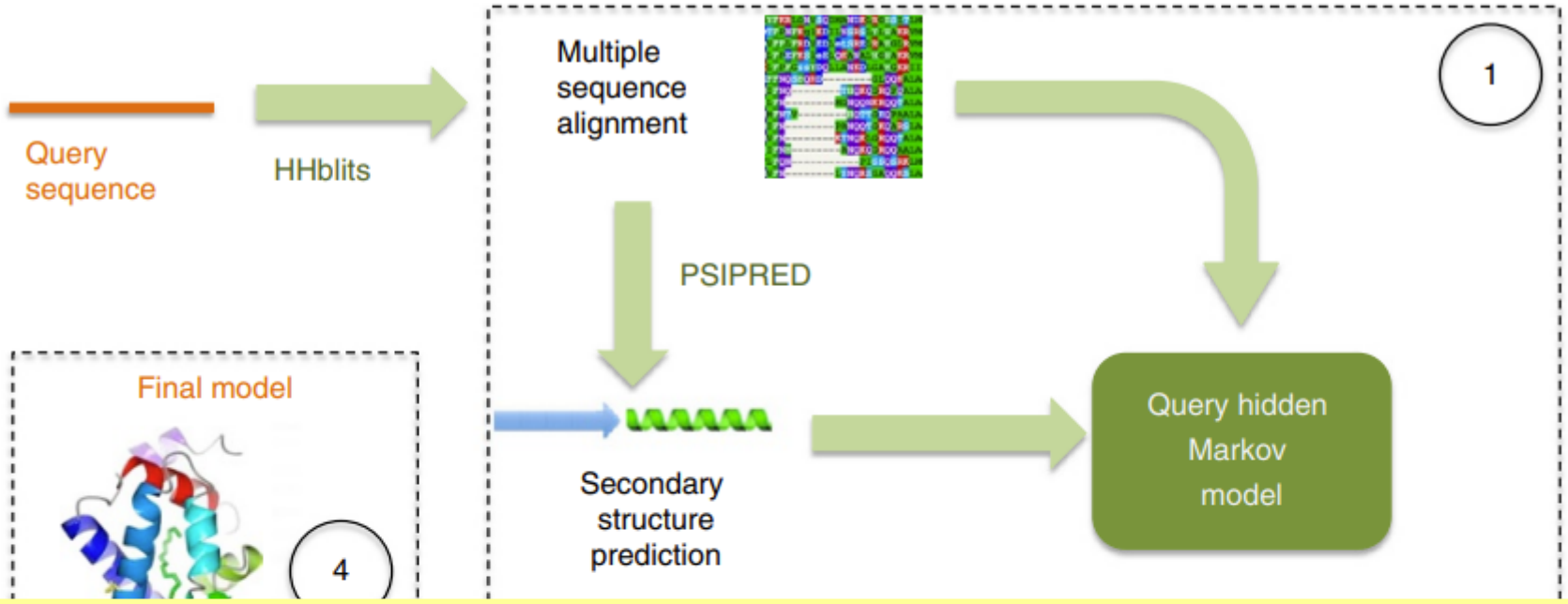
Multi-template homology models



Phyre2: algorithm pipeline



Phyre2: algorithm pipeline



Step 1: Identify similar sequences in protein sequence database

Choose a template structure by:

(1) comparing sequence profiles and

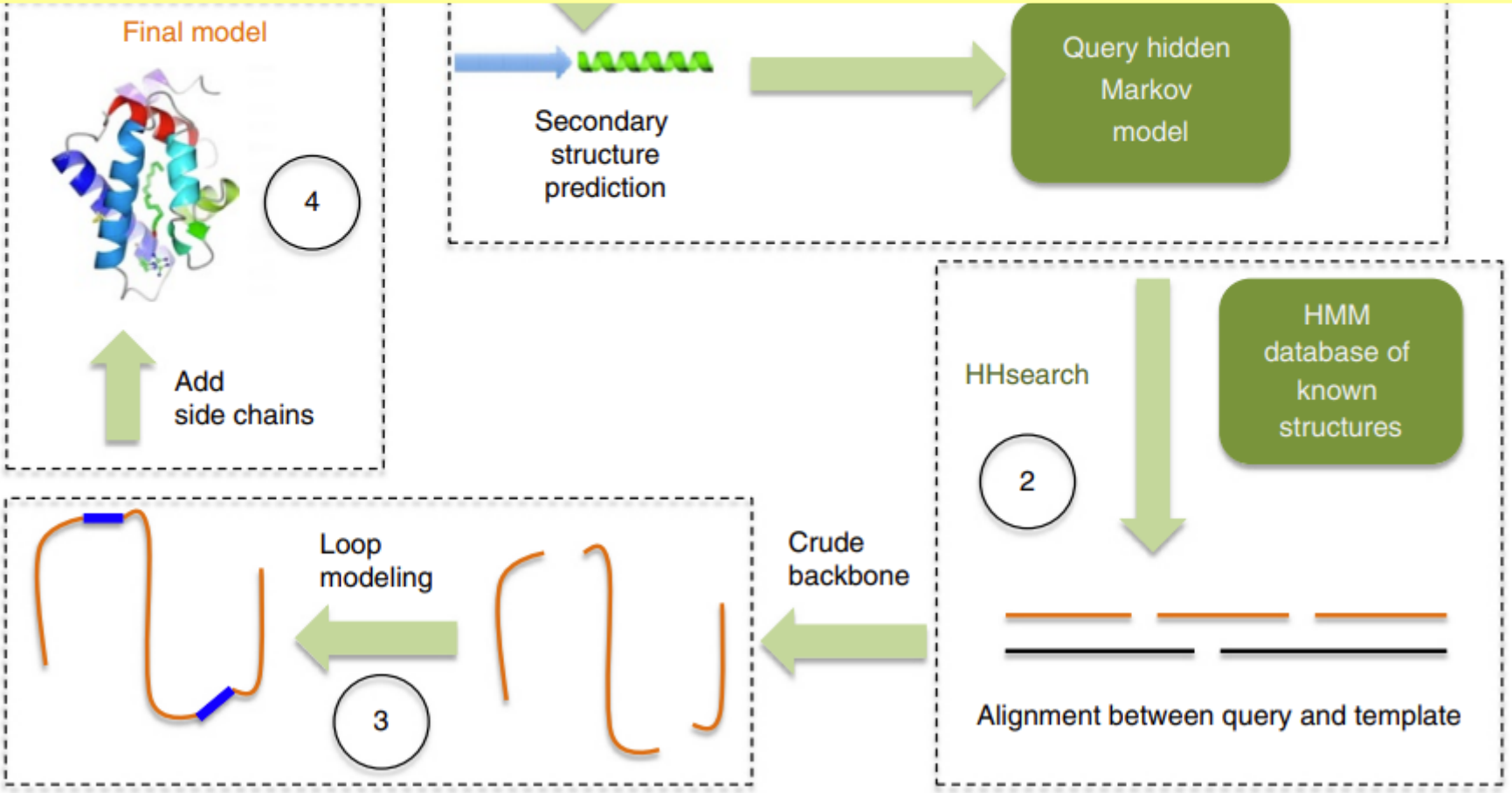
(2) predicting secondary structure for each residue in the query sequence and comparing to candidate template structures.

*Secondary structure (alpha helix, beta sheet, or neither) is predicted for segments of query sequence using a neural network trained on known structures.

Phyre2: algorithm pipeline

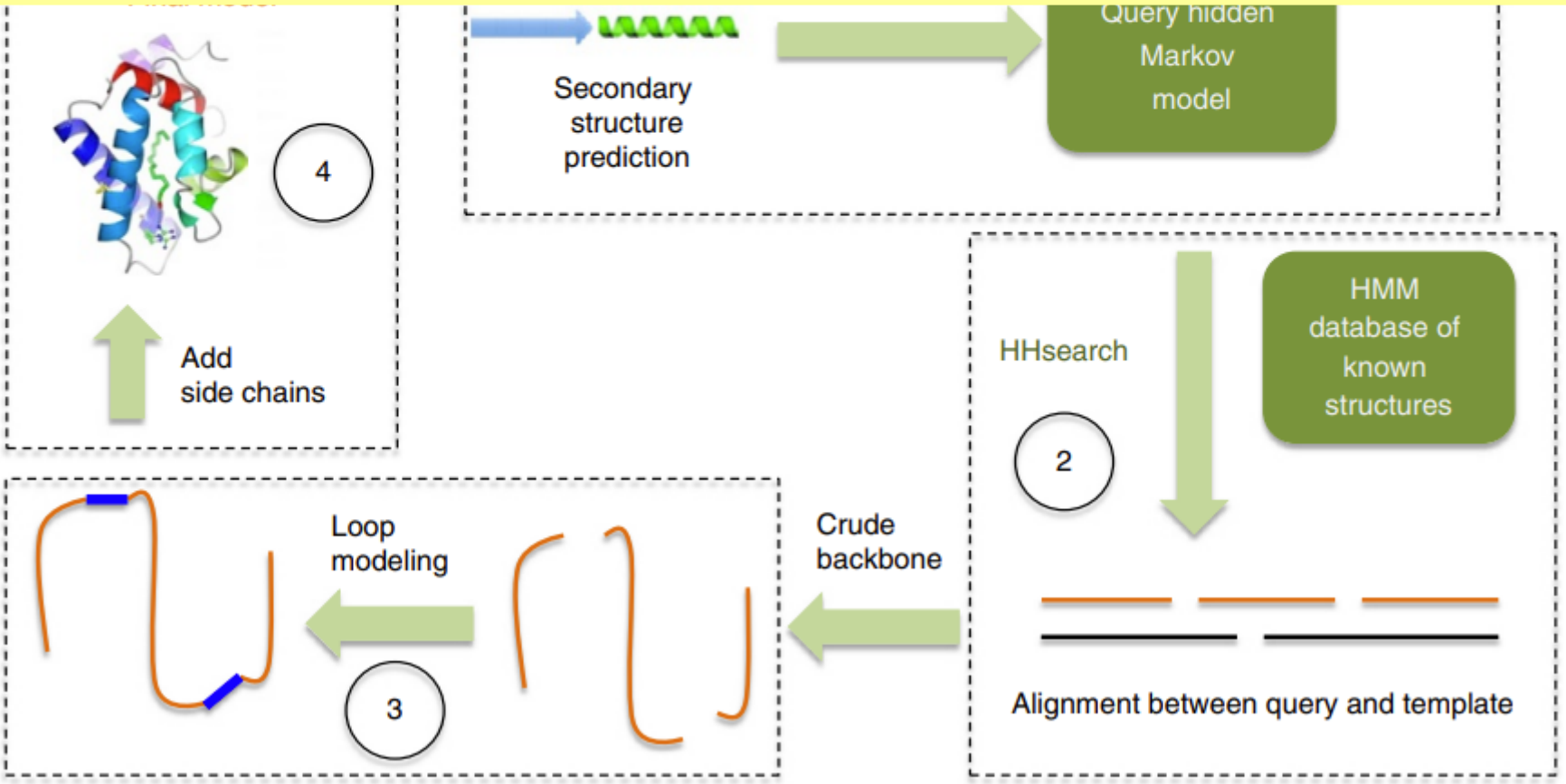
Step 2: Compute optimal alignment of query sequence to template structure

Build a crude backbone model (no side chains) by superimposing corresponding amino acids. Some of the query residues will not be modeled (insertions), there will be some physical gaps (deletions).



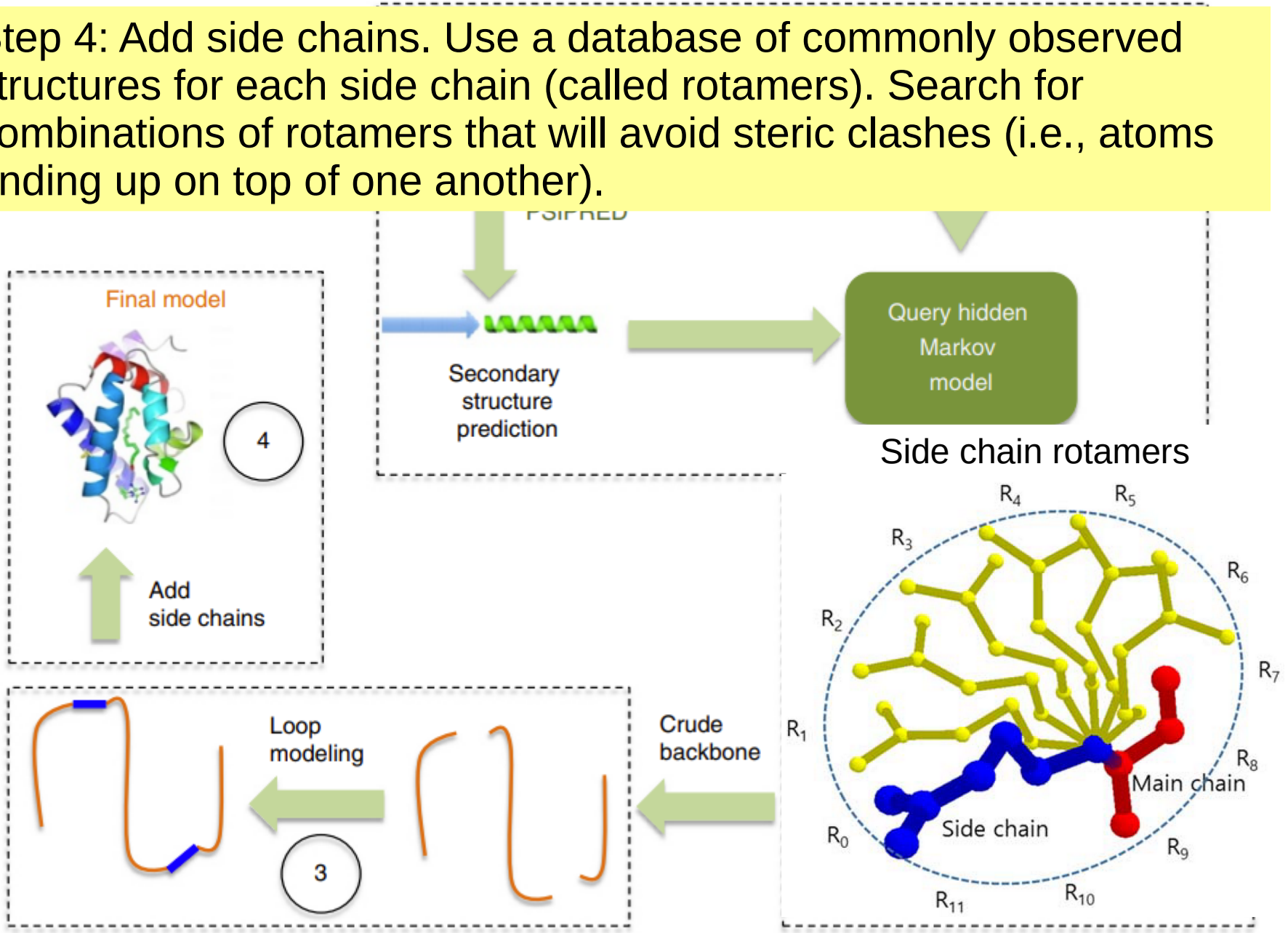
Phyre2: algorithm pipeline

Step 3: Use loop modeling to patch up defects in the crude model due to insertions and deletions. For each InDel, search a large library of loop fragments (2-15 residues) of PDB structures for ones that match local sequence and fit the geometry best. Tweak backbone dihedrals within these fragments to make them fit better.



Phyre2: algorithm pipeline

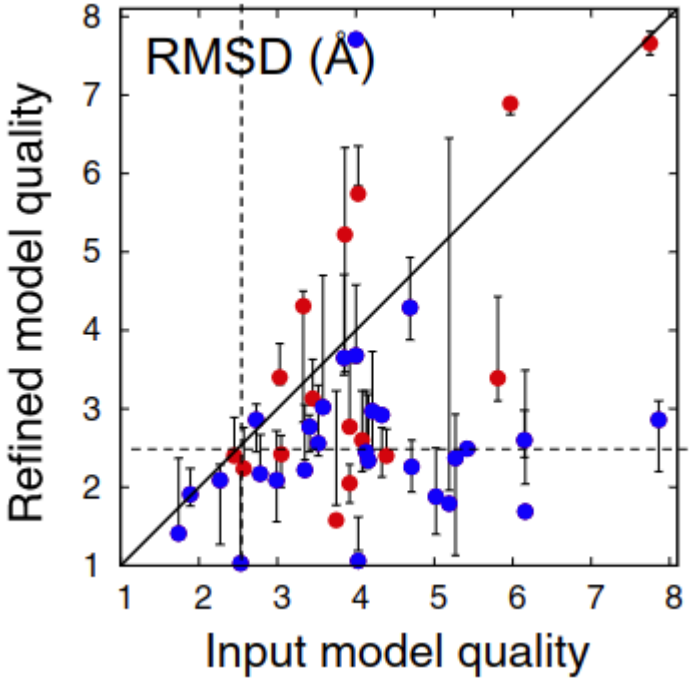
Step 4: Add side chains. Use a database of commonly observed structures for each side chain (called rotamers). Search for combinations of rotamers that will avoid steric clashes (i.e., atoms ending up on top of one another).

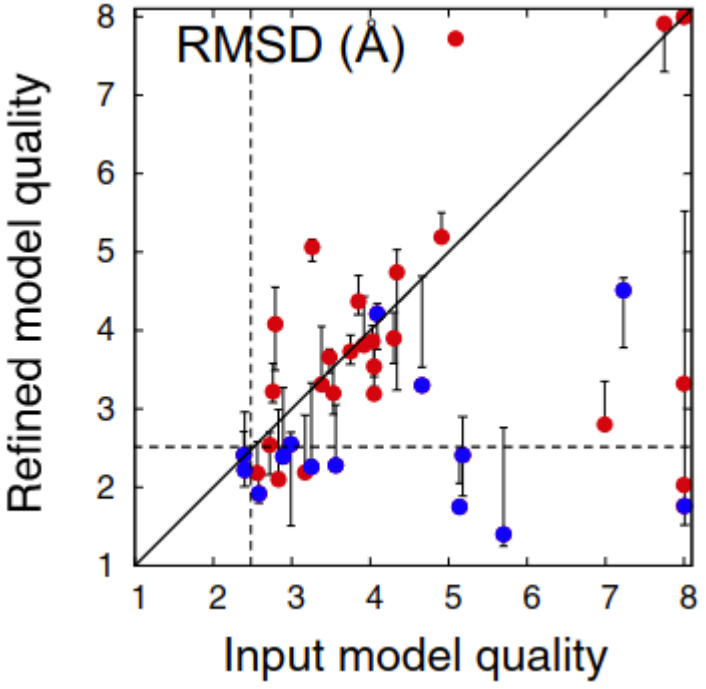


Optimization

Protein homology models can be refined by large-scale energy optimization

- Molecular dynamics works when starting models close to native structure
- Coarse-grained conformational search and unrestrained simulations can sample more extensively but suffer from inaccuracy in energy functions
- Rosetta modeling, two stage:
  - generate a population of diverse structures in different low-energy minima
  - utilize an evolutionary algorithm (Monte Carlo) to guide this model population toward the lowest all-atom energy





RMSD = root mean squared deviation
Red > 120 residues
Blue < 120 residues

Park et al (2018)

Two approaches to protein structure prediction

1) Template-based modeling (homology modeling)

– Used when one can identify one or more likely homologs of known structure

2) *Ab initio* (*de novo*) structure prediction

– Does not require any homologs

– Even *ab initio* approaches usually take advantage of available structural data, but in more subtle ways

Rosetta *ab initio* predictions

- Software developed over the last 20–25 years by David Baker, well known, and similar to other methods
- Knowledge-based energy function (PDB structures)
  - The “Rosetta energy function,” which is coarse-grained (i.e., does not represent all atoms in the protein), is used in early stages of protein structure prediction
  - The “Rosetta all-atom energy function,” which depends on the position of every atom, is used in late stages

Rosetta energy function
– Do not explicitly represent solvent (e.g., water)
– Assume all bond lengths and bond angles are fixed
– Represent the protein backbone using torsion angles (three per amino acid: Φ, Ψ, ω)
– Represent side chain position using a single “centroid,” located at the side chain’s center of mass.
Centroid position determined by averaging over all structures of that side chain in the PDB

Conformational search

Two steps:
– Coarse search: fragment assembly
– Refinement
Perform coarse search many times, and then perform refinement on each result

Coarse search
- Uses a large database of 3-residue and 9-residue fragments, taken from structures in the PDB (e.g. torsion angles)
- Monte Carlo sampling algorithm proceeds as follows:

1. Start with the protein in an extended conformation
2. Randomly select a 3-residue or 9-residue section
3. Find a fragment in the library whose sequence resembles it
4. Consider a move in which the backbone dihedrals of the selected section are replaced by those of the fragment. Calculate the effect on the entire protein structure.
5. Evaluate the Rosetta energy function before and after the move.
6. Use the Metropolis criterion to accept or reject the move.
7. Return to step 2

Conformational search

Two steps:
– Coarse search: fragment assembly
– Refinement
Perform coarse search many times, and then perform refinement on each result

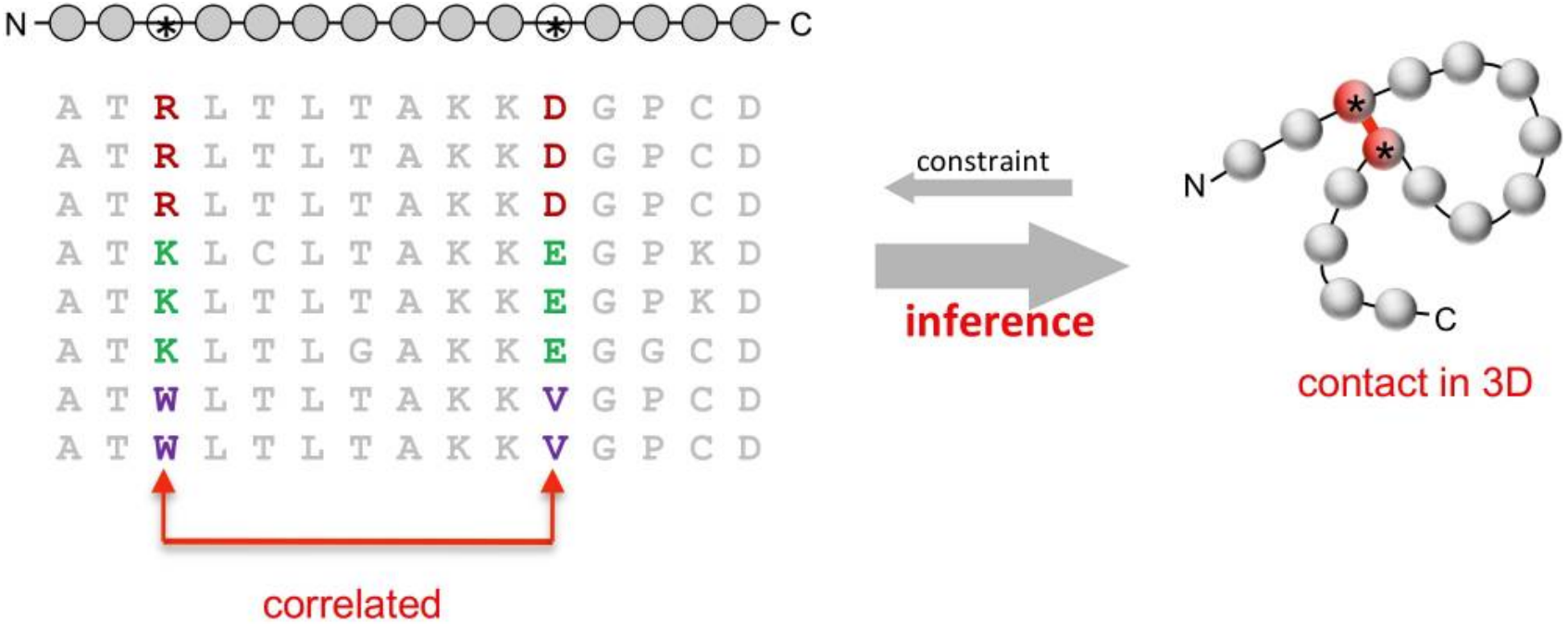
Refinement
- Refinement is performed using the Rosetta all-atom energy function, after building in side chains
- Refinement involves a combination of Monte Carlo moves and energy minimization
- The Monte Carlo moves are designed to perturb the structure much more gently than those used in the coarse search

Alphafold2

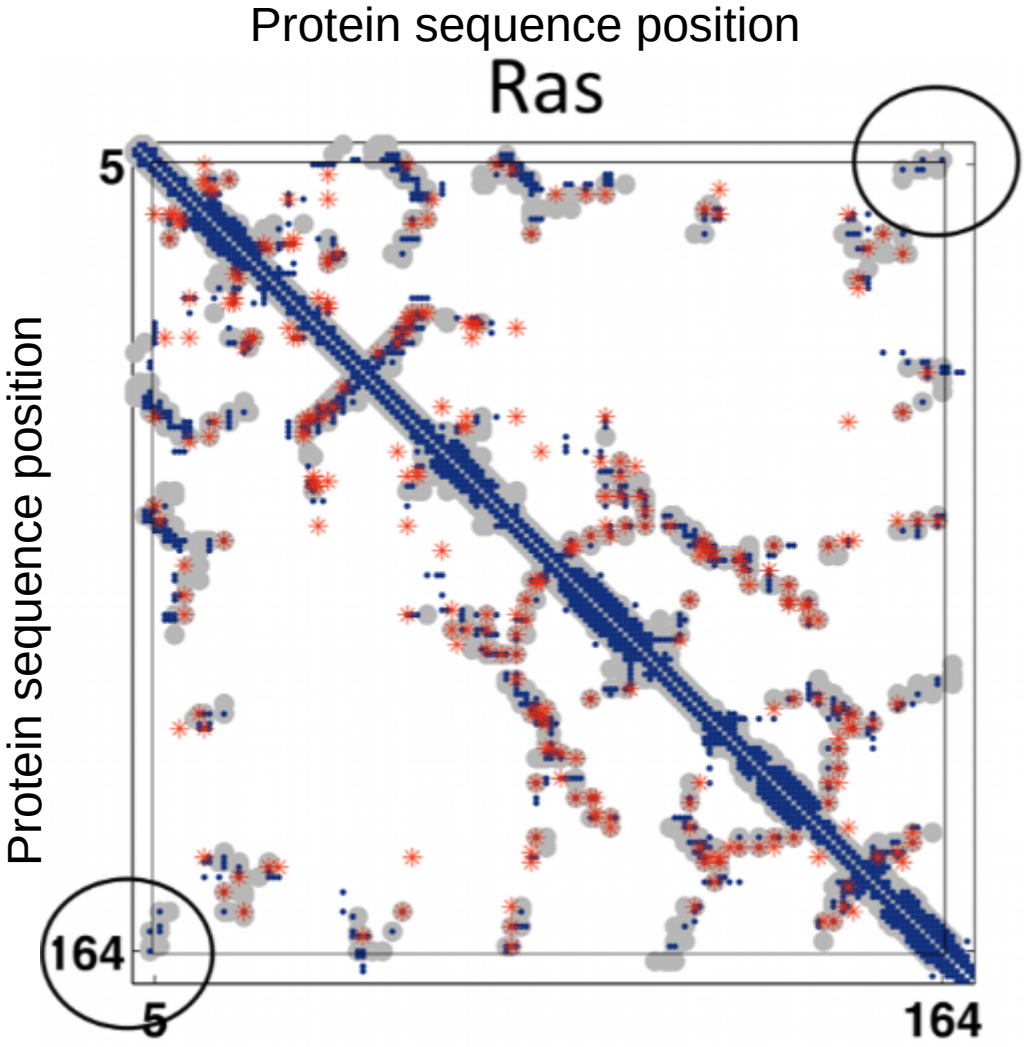
A Deep neural network:
- physical/statistical interactions
- evolution: constraints, homology models, pairwise correlations
- growth of PDB structure database

How does one predict 3D structure?

Observation: residues that physically contact each other show correlated patterns of evolution



Correlations vs Contacts



3D contacts for Ras protein

The contacts derived from the predicted 3D structure (dark blue) are in good general agreement with those from the observed structure (grey).

DI constraints

Predicted structure contacts

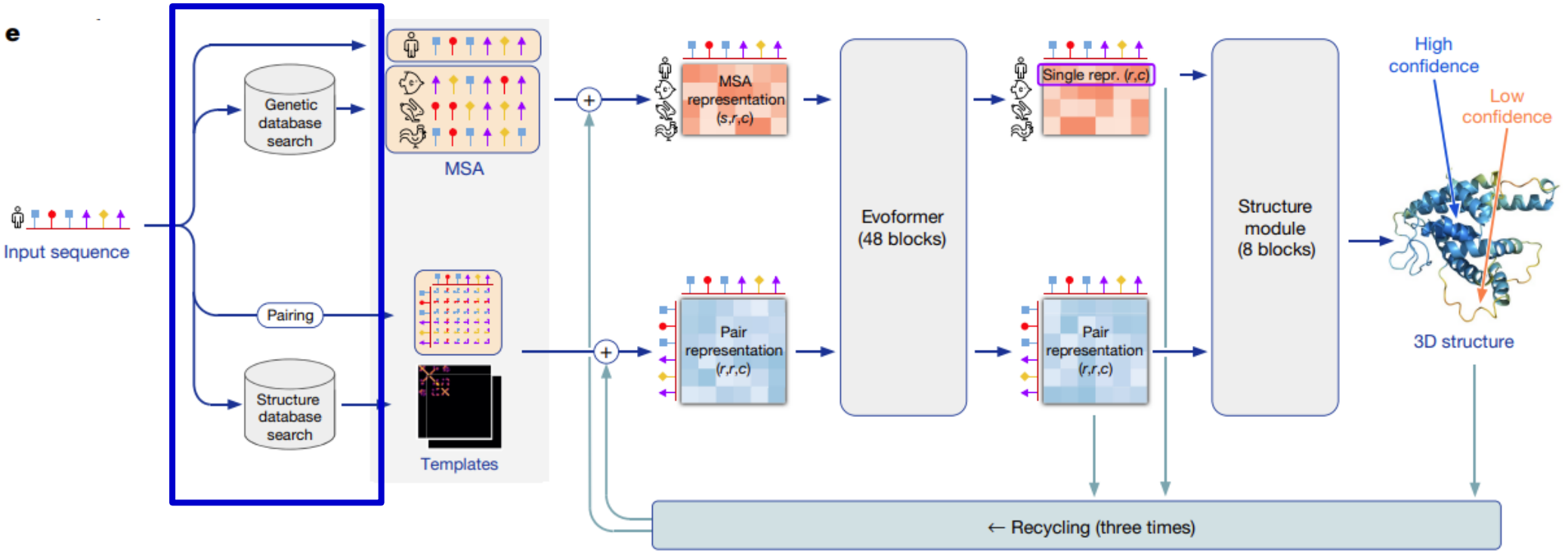
Experimental structure contacts

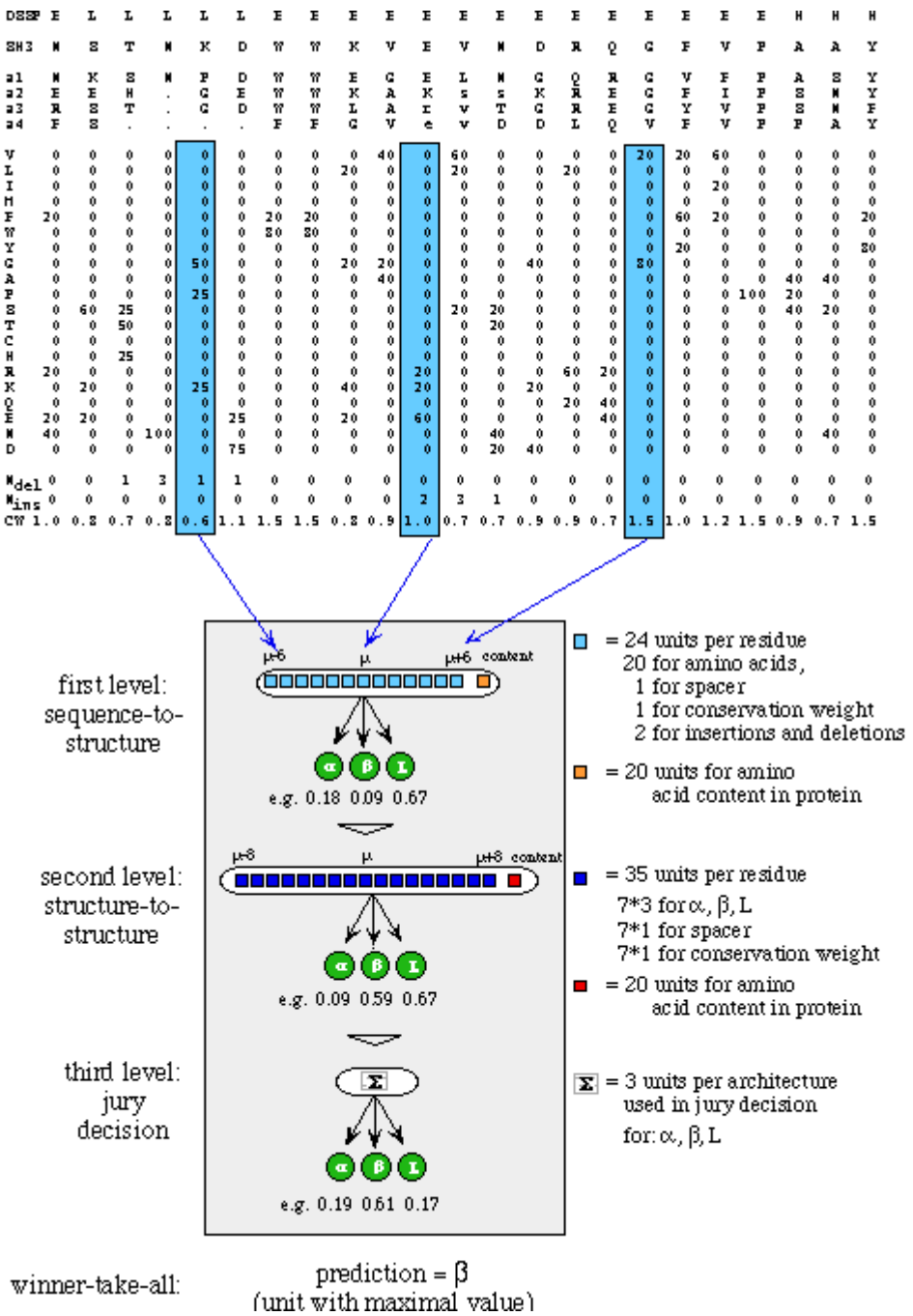
AlphaFold2 Inputs

Inputs

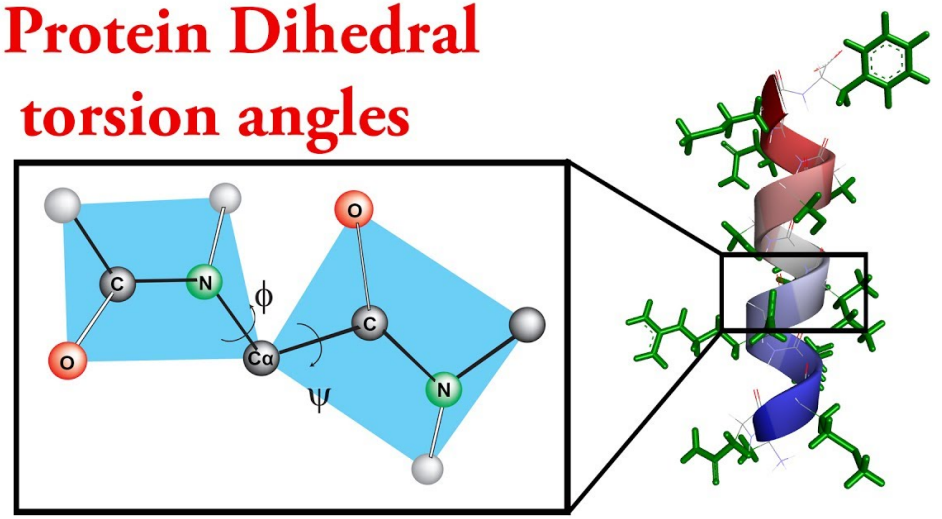
- Multiple sequence alignment (MSA)
- Pair representations (correlated evolution)
- PDB (homologs)

MSA provides: profile of amino acids that have same structure





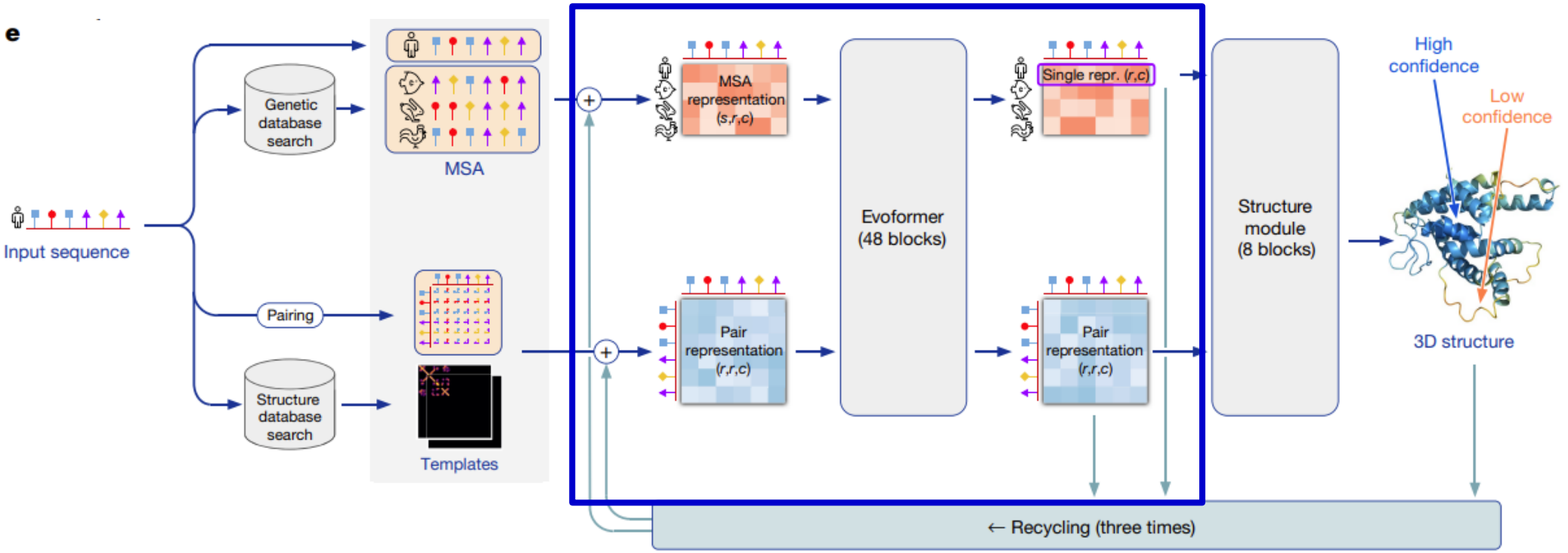
Protein Dihedral torsion angles



AlphaFold2 Transformer

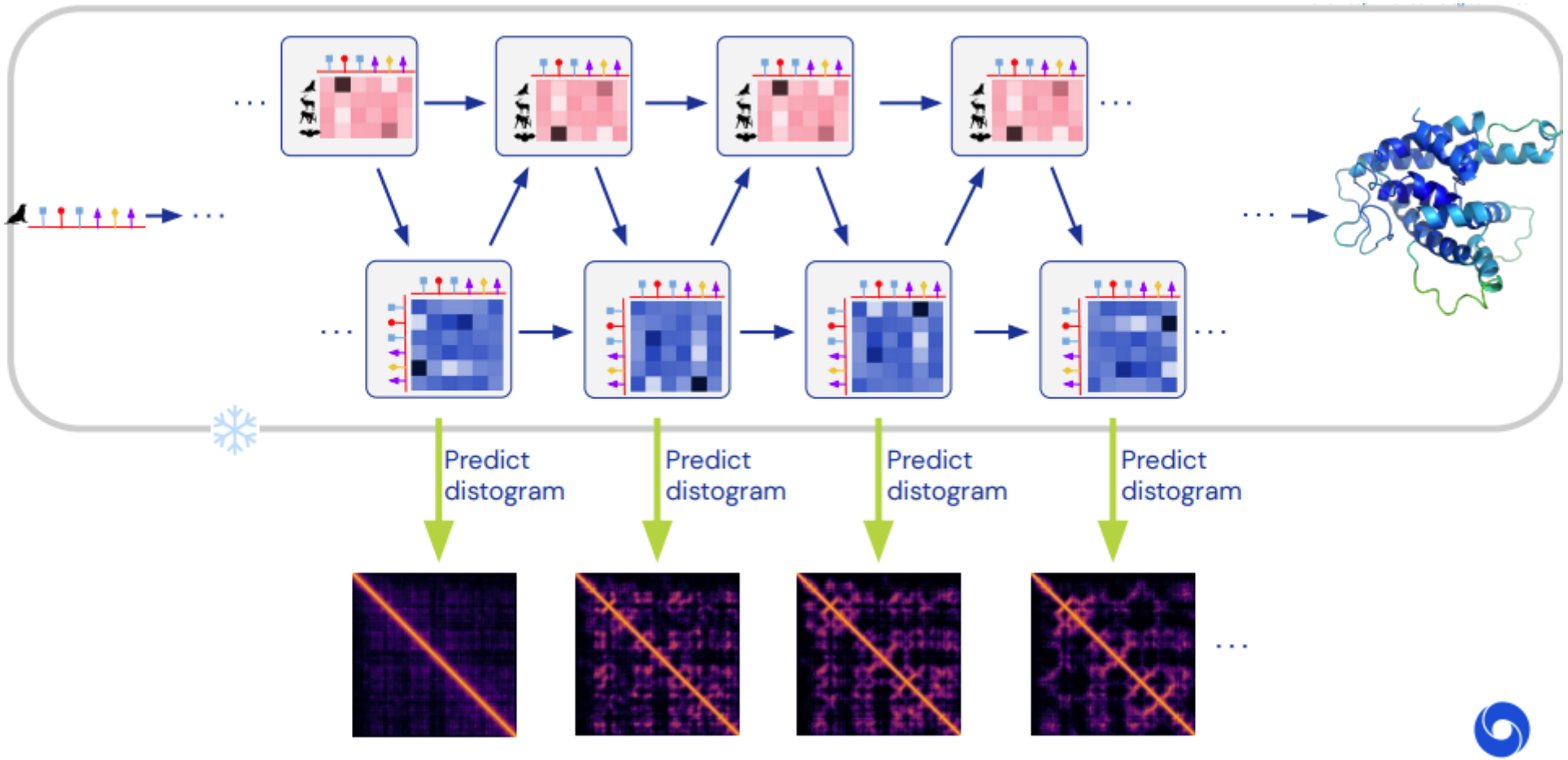
Transformer (evoformer 48 blocks)

- identify which pieces of information are more informative
- refine the representations for both the MSA and the pair interactions (both a product and an intermediate)
- iteratively for three cycles (iteratively exchange information)



Iterations

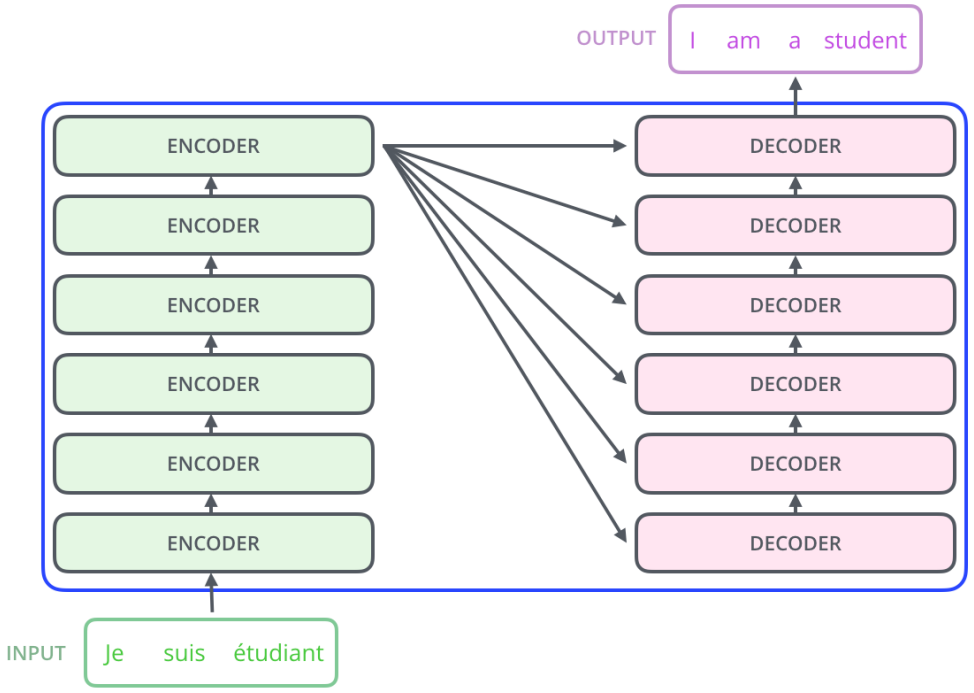
- Predicted contacts inform structure, structure informs MSA
- Contacts inform structure, which informs which sequences are informative for contacts (too close or too distant)

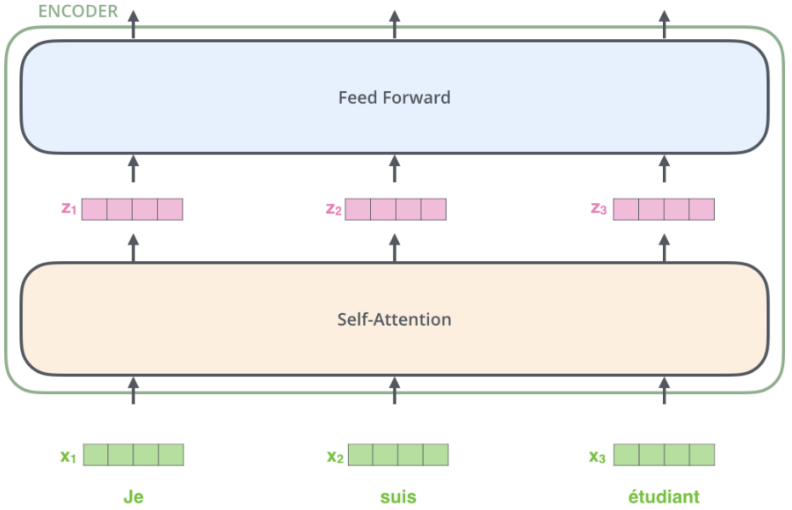


Transformers

A transformer is a deep learning model that adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data.

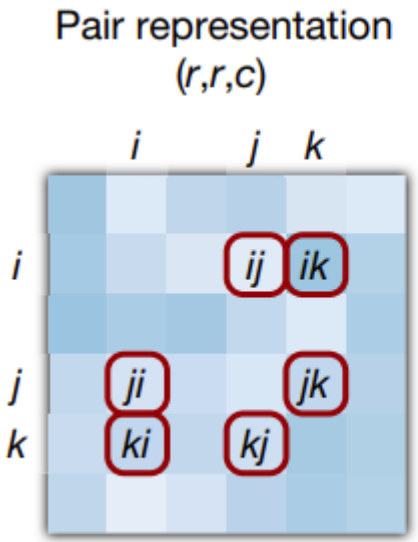
- Designed to handle sequential input data
- Attention mechanism provides context, so doesn't need to be in sequential order
- Transformers were introduced in 2017 by Google Brain and are increasingly the model of choice, replacing Recurrent Neural Networks such as long short-term memory (LSTM).

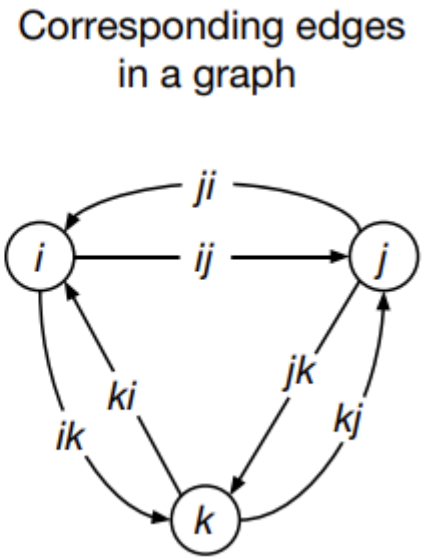


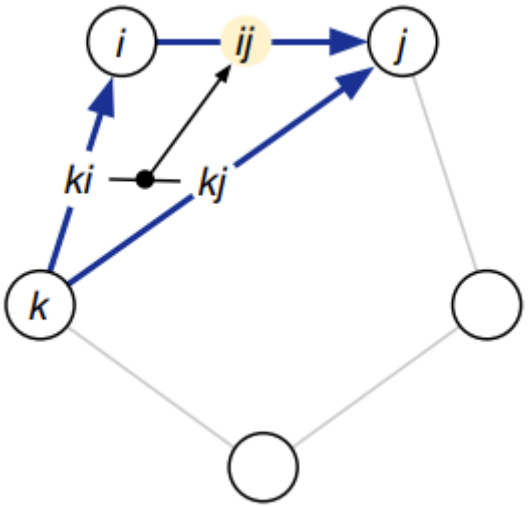


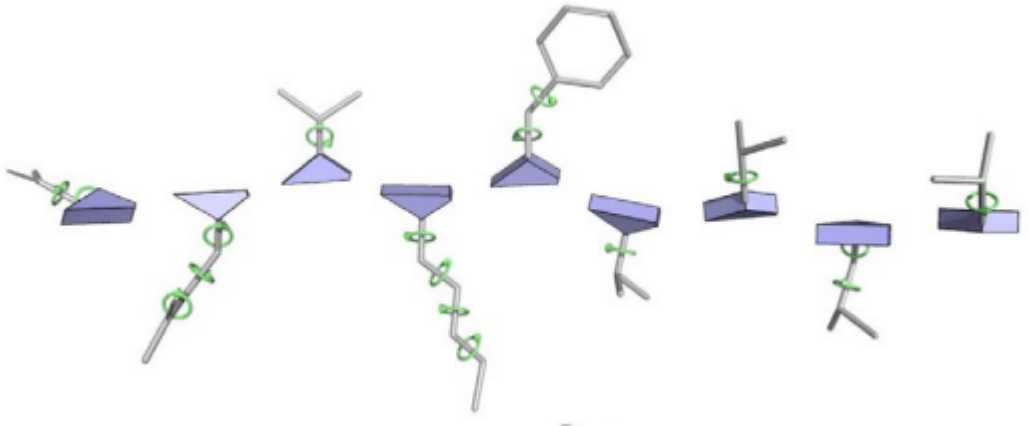
Pair transformer

Attention is arranged in terms of triangles of residues, to enforce the triangle inequality





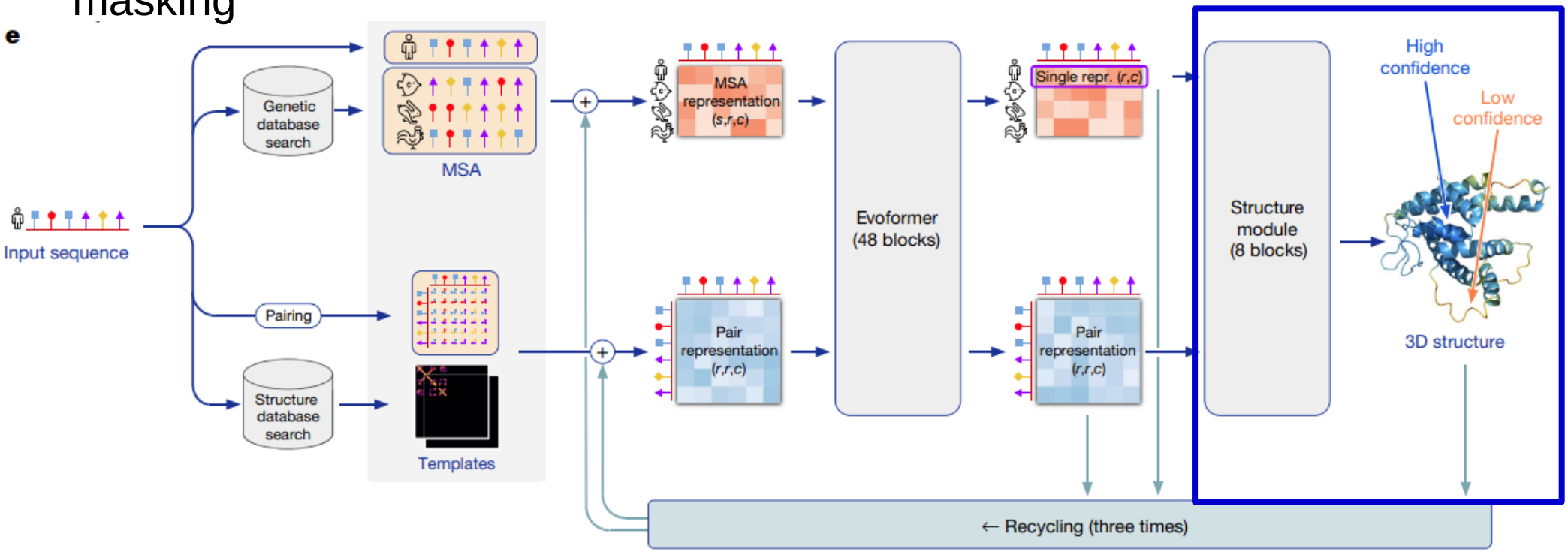




AlphaFold2 Structure

Structure modeule (8 blocks)

- no optimization, outputs 3D structure from pair and MSA
- invariance to translations and rotations
- generates a model of the side chains, their positions are parameterized by a list of torsion angles
- structural loss function FAPE (Frame Aligned Point Error), similar to RMSD
- loss function is a weighted sum of multiple “auxiliary losses”, e.g. MSA masking



AlphaFold2 Training

Training

- supervised learning on PDB data
- predicted structures as the training data, in which the various training data augmentations such as cropping and MSA subsampling make it challenging for the network to recapitulate the previously predicted structures
- encourage the network to learn to interpret phylogenetic and covariation relationships without hardcoding: randomly mask out or mutate individual residues within the MSA and use Bidirectional Encoder Representations from Transformers (BERT) objective
- auxiliary side-chain loss during training, and an auxiliary structure violation loss during fine-tuning

Limitations (poor performance)

- fewer than 30 sequences in the MSA
- few intra-chain or homotypic contacts

AlphaFold2 Summary

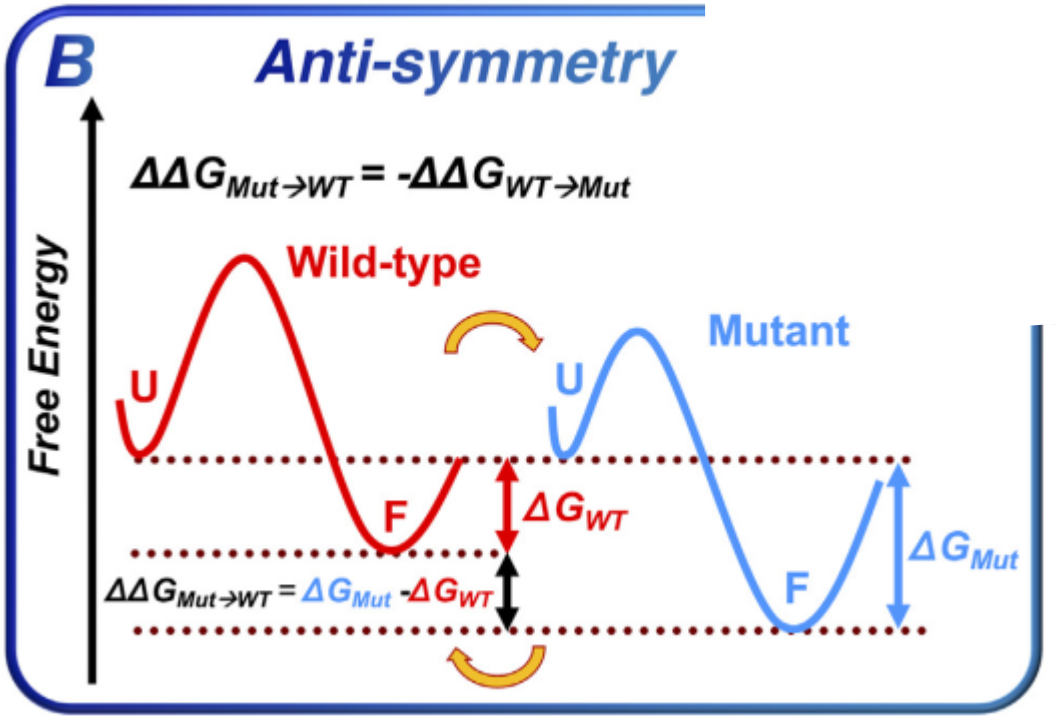
Summary

- No new biological insights; same information used by everyone else
- performance boils down to DeepMind's superb engineering [not black box]
- architecture is experimentally derived: what works?
  - MSA transformer use gated attention
  - MSA representation learn from the pair representation as an input

Predicting destabilizing mutations ΔΔG

- Most deleterious/disease mutations disrupt protein stability
- Protein structure can be used to predict which amino acid mutations destabilize proteins
- MSA (conservation) can also be used, even when no structure (SIFT, PolyPhen2)

ΔG = Gibbs free energy (native vs unfolded)
ΔΔG = ΔG (wildtype) vs ΔG (mutant), measure of destabilization



ΔΔG Prediction Algorithms

FoldX (2002), first method, energy calculated from simplified force field

Three types of methods:

- Classical (slower): free energy differences by classical equations and geometrical features, chemico-physical parameters or potential energy evaluations.
  SRide, CUPSAT, Eris, SDM2, TKSA-MC, pSTAB, PoPMuSiCsym [FoldX]
- AI-based (fast but tend to overfit): AI-based methods based on descriptors of the sequence and/or structural features
  I-Mutant-2 and 3, MUpro, mCSM, NeEMO, AUTO-MUTE 2, INPS-MD, EASE-MM, STRUM, PON-tstab, DeepDDG
- Meta: combinations, use output from other software
  iStable, DUET, ELASPIC, MAESTROweb, DynaMut

Free and available

Predicting ΔΔG

Challenges and areas for improvement

Symmetry: reverse hypothetical mutation (Li et al 2012; PROTS). AI-based methods are biased because database is mostly destabilizing

Database: ProTherm accessed publicly through VariBench: remove errors (VariBench-stable): 99 proteins 1564 entries, 234 stabilize, 864 destabilize, 467 neutral. 29% of Protherm.

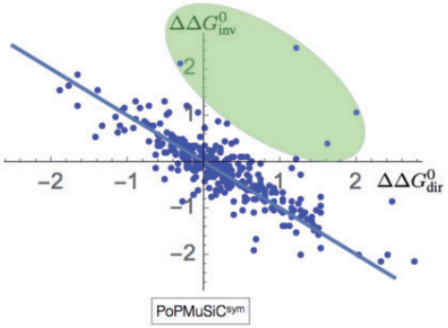
Empirical errors: There is a natural upper bound to the accuracy of predicting protein stability changes upon mutations: 0.7-0.8 correlation, empirical error in ΔΔG = 0.48

Predicting ΔΔG

PoPMuSiC-Sym (2015): incorporates both symmetric and transitive properties of ΔΔG into the architecture, using neural network.
- based on standard statistical potentials, combined with sigmoidal weights that depend on the solvent accessibility of the mutated residues.
- artificial neural network of terms
- 16 terms (distance, local, volume), 2 require forced symmetry
- S^{sym} uses 684 mutations with both WT and mutant structures, 357 proteins

Free energy terms	
Distance Potentials	$\Delta W_1(s,d)$, $\Delta W_2(s,s,d)$ $\Delta W_3(s,a,d)$, $\Delta W_4(s,a,s,a,d)$ $\Delta W_5(s,t,d)$, $\Delta W_6(s,t,s,t,d)$
Local Potentials	$\Delta W_7(s,t)$, $\Delta W_8(s,t,t)$ $\Delta W_9(s,s,t)$, $\Delta W_{10}(s,a)$ $\Delta W_{11}(s,a,a)$, $\Delta W_{12}(s,s,a)$ $\Delta W_{13}(s,t,a)$
Volume terms	ΔV_+, ΔV_-
Independent term	1

Method	σ_{dir}	r_{dir}	σ_{inv}	r_{inv}	$r_{dir-inv}$	$\langle\delta\rangle$
PoPMuSiC^{sym}	1.58	0.48	1.62	0.48	−0.77	0.03
MAESTRO	1.36	0.52	2.09	0.32	−0.34	−0.58
FoldX	1.56	0.63	2.13	0.39	−0.38	−0.47
PoPMuSiC v2.1	1.21	0.63	2.18	0.25	−0.29	−0.71
SDM	1.74	0.51	2.28	0.32	−0.75	−0.32
iSTABLE	1.10	0.72	2.28	−0.08	−0.05	−0.60
I-Mutant v3.0	1.23	0.62	2.32	−0.04	0.02	−0.68
NeEMO	1.08	0.72	2.35	0.02	0.09	−0.60
DUET	1.20	0.63	2.38	0.13	−0.21	−0.84
mCSM	1.23	0.61	2.43	0.14	−0.26	−0.91
MUPRO	0.94	0.79	2.51	0.07	−0.02	−0.97
STRUM	1.05	0.75	2.51	−0.15	0.34	−0.87
Rosetta	2.31	0.69	2.61	0.43	−0.41	−0.69
AUTOMUTE	1.07	0.73	2.61	−0.01	−0.06	−0.99
CUPSAT	1.71	0.39	2.88	0.05	−0.54	−0.72



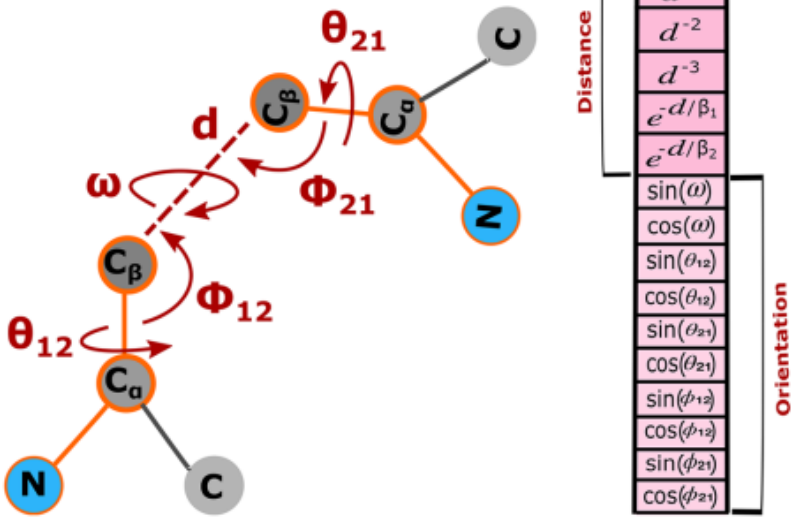
Example: SCONES

SCONES (2021): incorporates both symmetric and transitive properties of ΔΔG into the architecture, using neural network.
- Symmetric (ΔΔG-forward = -ΔΔG-reverse)
- transitive (X to Z = X to Y + Y to Z) A51T+G88S = AG to TS
- assume structure is same for WT and mutant
- no long range interactions, far from the site
- nodes (AA properties) and edges (residue interaction)

Table 1. List of Node Features

feature	description or AAindex[a]
formal charge[b]	Table S1 in Supporting Information
normalized van der Waals volume[b]	FAUJ880103
hydropathy index[b]	KYTJ820101
steric parameter[b]	CHAM810101
polarity[b]	GRAR740102
residue-accessible surface area in tripeptide[b]	CHOC760101
solvent-accessible surface area	calculated using DSSP[64,65]

[a] AAindex[66] is a database of numerical indices for physicochemical and biochemical properties of amino acids and pairs of amino acids. The code in the column is the AAindex database ID of the property. [b] These features are constants for a given amino acid. They are used to initialize the amino acid embedding layer.

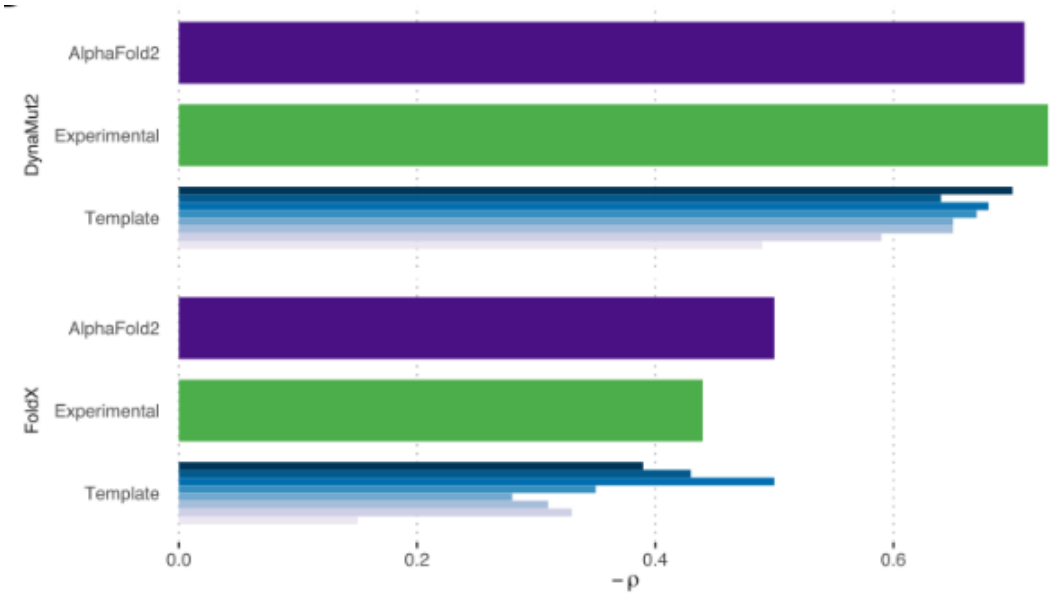


Does AlphaFold help predict ΔΔG?

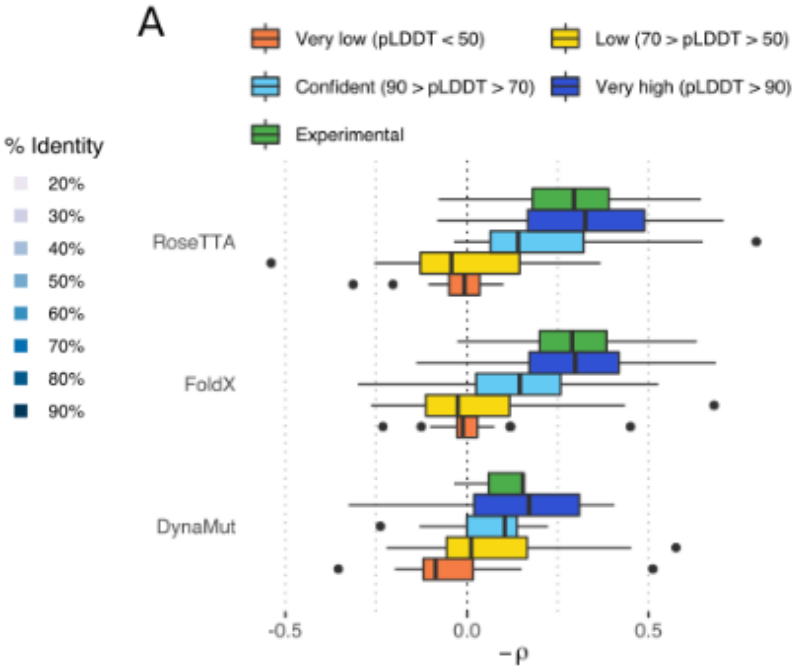
Akdel (bioRxiv 2021):

- AlphaFold structures as good as experimental for ΔΔG predictions

2,648 single-point missense mutations over 121 distinct proteins



33 proteins with 117,135 mutations from deep mutational scanning



Exercises

1) Structures can be conserved with little or no sequence similarity [T/F]
2) Homology modeling is facilitated by using homologous protein domains, even if the entire protein is not homologous [T/F]
3) Molecular mechanics force field parameters can be obtained from what two sources? Quantum mechanical calculations, Experimental data (knowledge-based potentials)
4) Molecular dynamics simulations are: too slow to predicting structure from sequence [T/F], include bond angles and electrostatics [T/F], can be used to refine predicted structures [T/F]
5) Ab initio structure prediction does not use existing structures in PDB [T/F]
6) Why are protein domains useful for homology modeling? Domains often fold independently, use of multiple templates, more homologs
7) How can PDB be used to predict a structure if there are no homologs for a sequence? Alpha helix, beta sheet, torsion angles, side chains
8) What do pair representations (correlated evolution of positions in proteins) provide information about? 3D contacts

Exercises

9) What is AlphaFold2 structure predictions based on [input]? MSA, PDB, pair representations

10) How do protein structures help identify disease mutations? By predicting changes in protein stability, ΔΔG