Recitation 11

Structural biology

Surfaces

Molecular surface

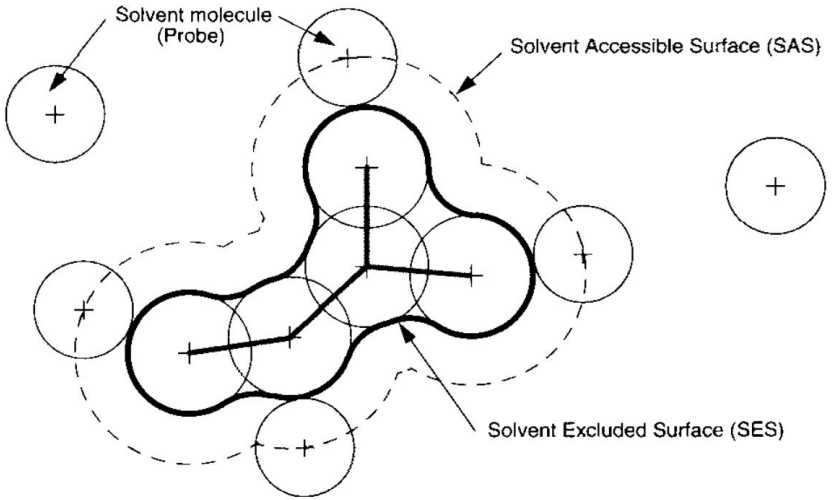
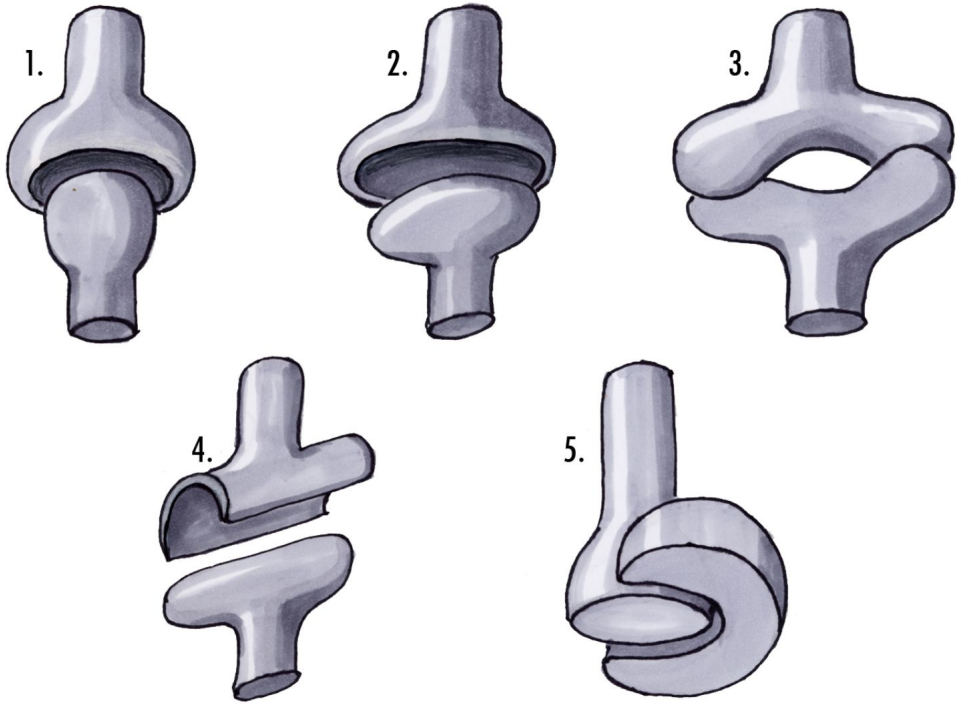


FIGURE 1 The solvent-accessible surface (SAS) is traced out by the center of the probe representing a solvent molecule. The solvent-excluded surface (SES) is the topological boundary of the union of all possible probes that do not overlap with the molecule.

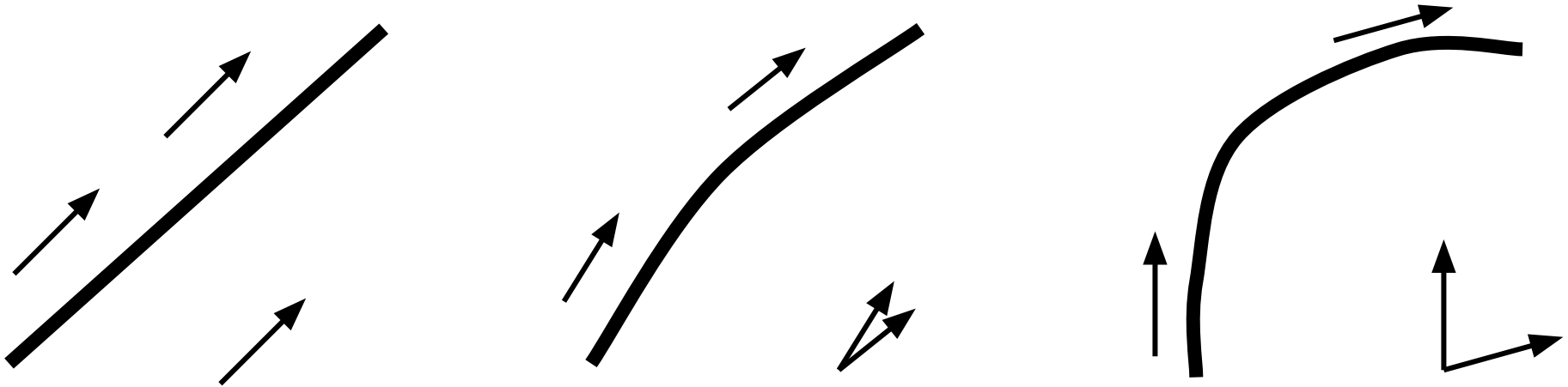
REDUCED SURFACE: an Efficient Way to Compute Molecular Surfaces

Understanding surfaces by understanding curvature

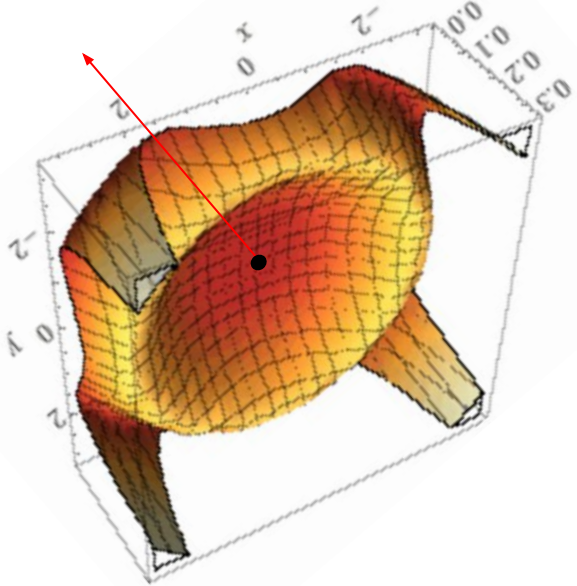


https://upload.wikimedia.org/wikipedia/commons/4/4b/Gelenke_Zeichnung01.jpg

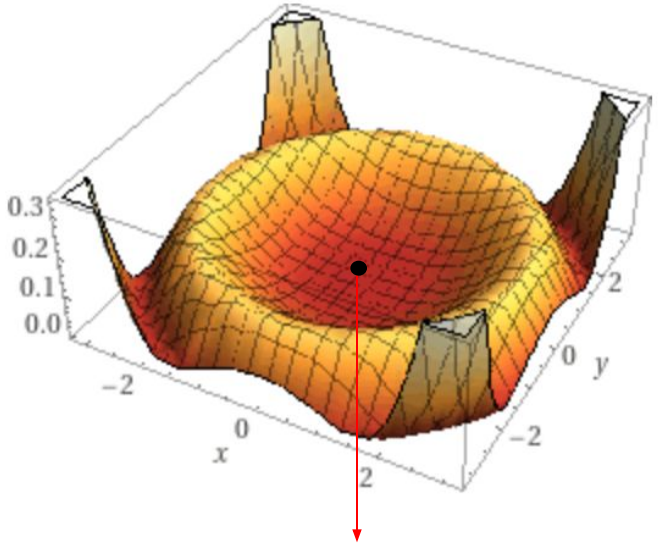
“Curvature is a second derivative”



Curvature of a surface



Rigid transformation



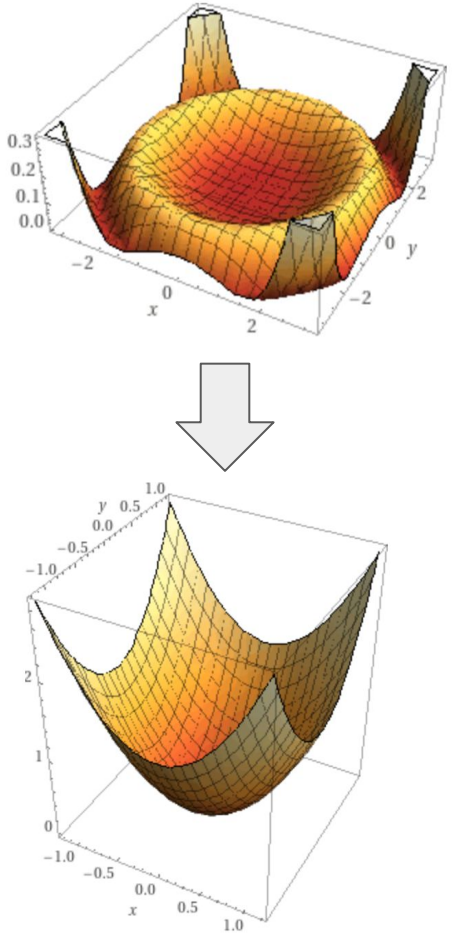
Figures generated from https://www.wolframalpha.com/

Curvature of a surface

- Let v = [x,y]
- Taylor expansion: height = 0.5 vHv^T + $O(|v|^3)$

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_1 x_2} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix}$$

- H is symmetric, so it has 2 real eigenvalues with orthogonal eigenvectors

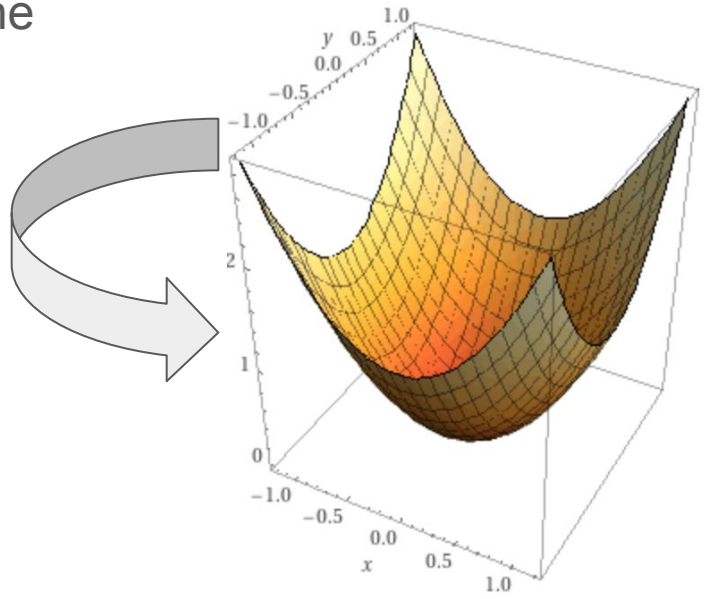


Figures generated from https://www.wolframalpha.com/

Curvature of a surface

- With the appropriate rotation we can make the Hessian diagonal

$$H = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & 0 \\ 0 & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix}$$

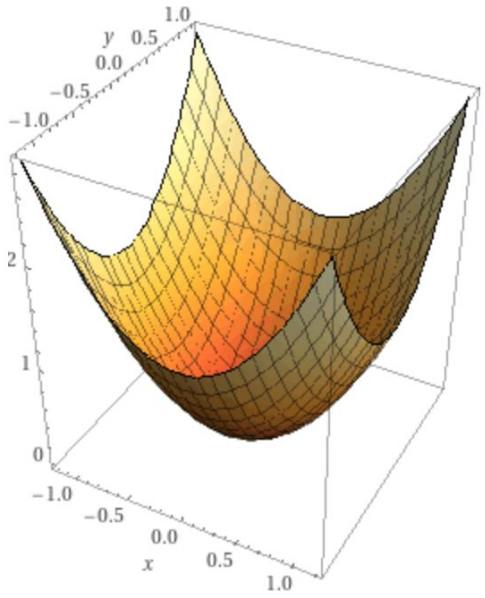


- We still have height = 0.5 vHv^T + $O(|v|^3)$
- = 0.5 $(d^2f/dx^2)x^2$ + 0.5 $(d^2f/dy^2)y^2$ + $O(|v|^3)$
- = 0.5 $\kappa_1 x^2$ + 0.5 $\kappa_2 y^2$ + $O(|v|^3)$

Figures generated from https://www.wolframalpha.com/

Curvature of a surface

- Height = 0.5 $\kappa_1 x^2$ + 0.5 $\kappa_2 y^2$ + O($|v|^3$)
- Choose a direction: (x,y) = (t cos(θ) + t sin(θ))
- Height" = κ_1 cos^2(θ) + κ_2 sin^2(θ)
  - Curvature along the chosen direction



- Mean curvature:

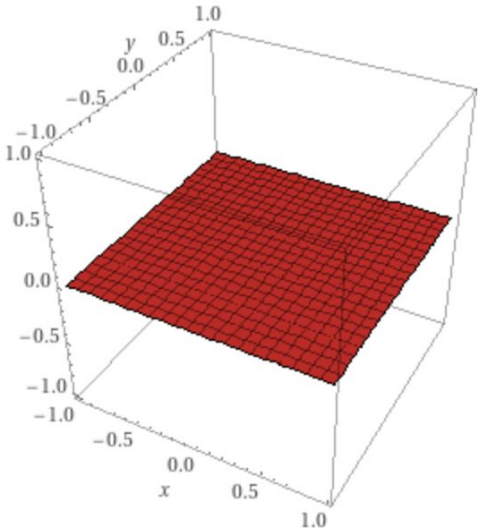
$$\frac{1}{2\pi}\int_0^{2\pi} \kappa_1 cos^2(\theta) + \kappa_2 sin^2(\theta) d\theta = \frac{\kappa_1 + \kappa_2}{2}$$

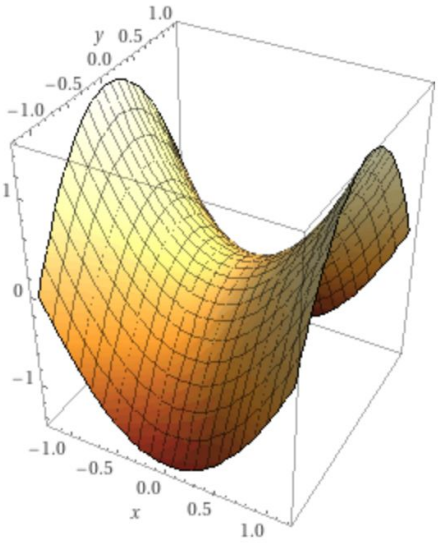
Figures generated from https://www.wolframalpha.com/

Another notion of curvature

Mean curvature: $(\kappa_1 + \kappa_2)/2$

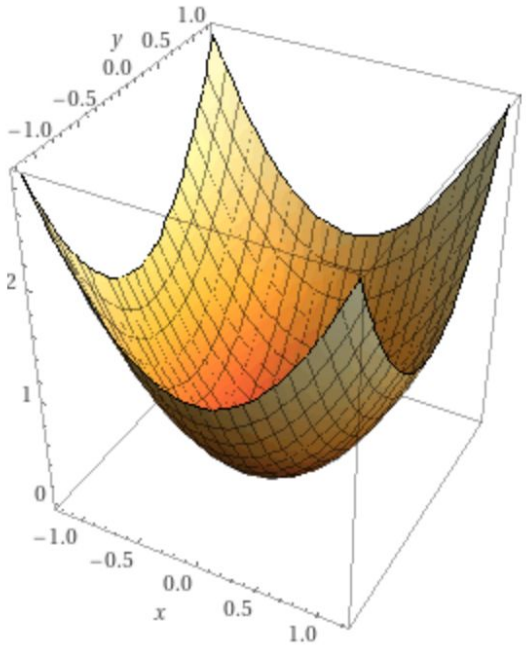
Gaussian curvature: $\kappa_1 * \kappa_2$

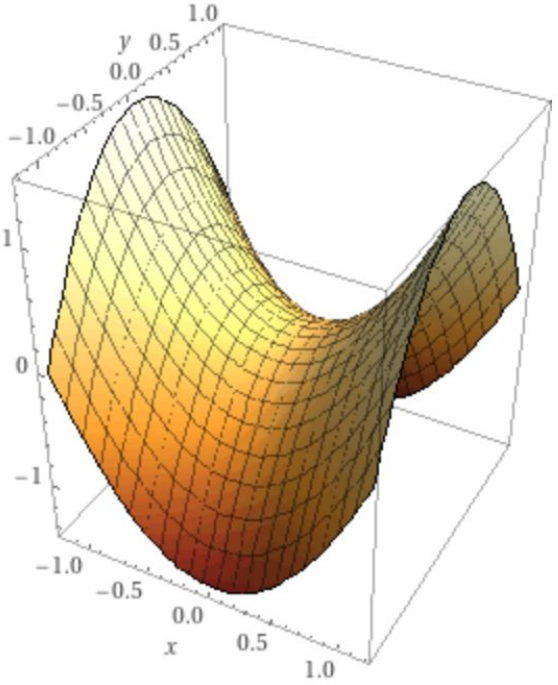




Figures generated from https://www.wolframalpha.com/

Gaussian curvature measures space deficit/excess





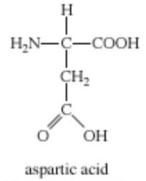
Figures generated from https://www.wolframalpha.com/

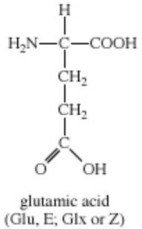
Protein structure

Amino acids are the Building Blocks of Proteins

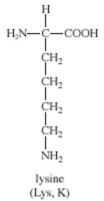
Amino acids have a side chain that endows them with a diverse range of chemical and physical properties, which give proteins their wide range of functions.

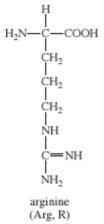
Acidic

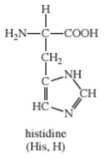


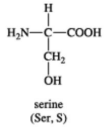


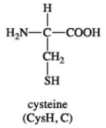
Basic

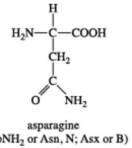


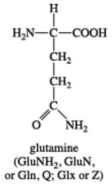


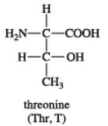


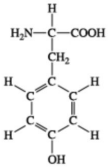




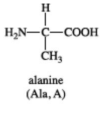


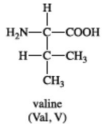


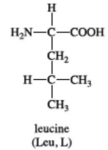


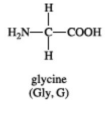


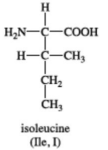
Polar

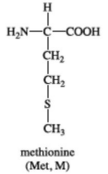


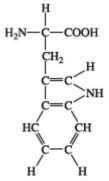


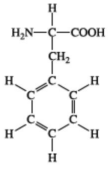


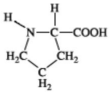








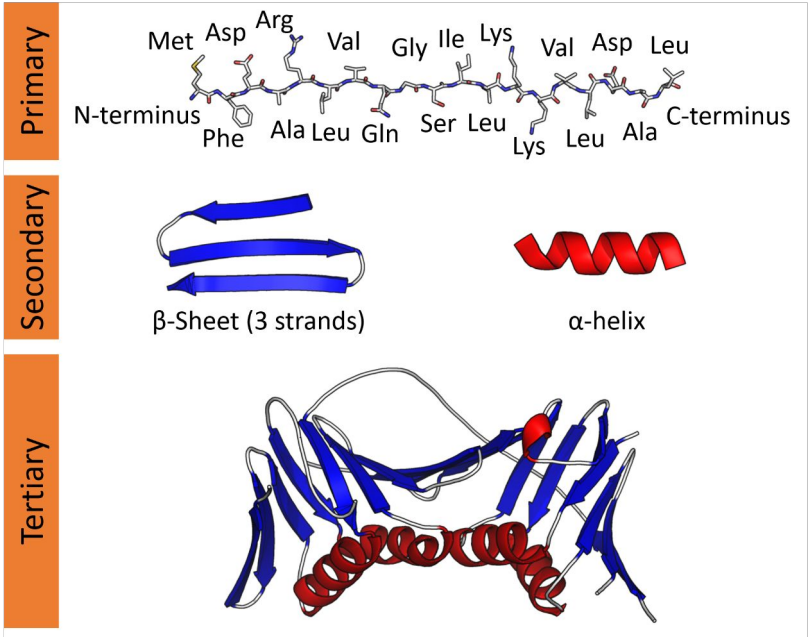


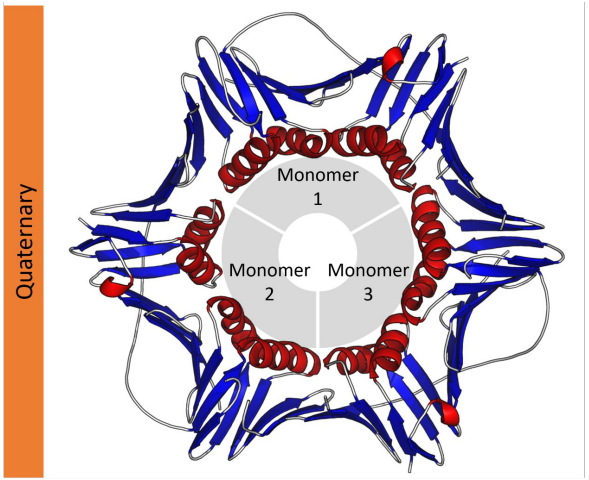


Hydrophobic

Amino acid, Encyclopedia Britannica, Reddy (2020)

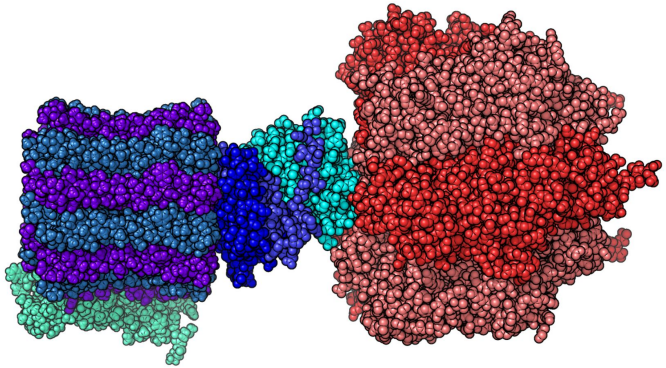
Primary Structure Hierarchy



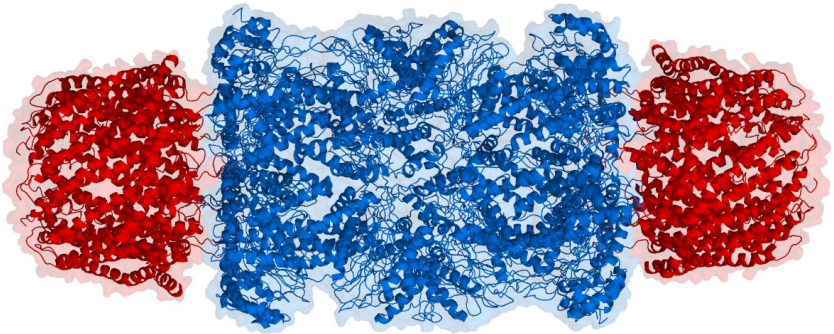


https://upload.wikimedia.org/wikipedia/commons/5/5f/Protein_structure_%28full%29.png

Structure informs function



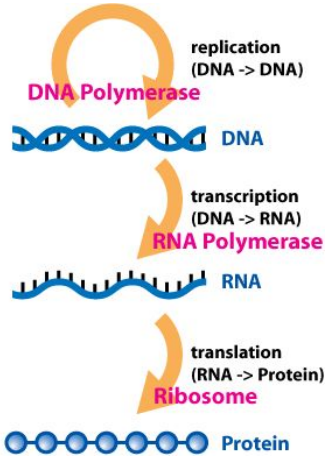
ATP Synthase



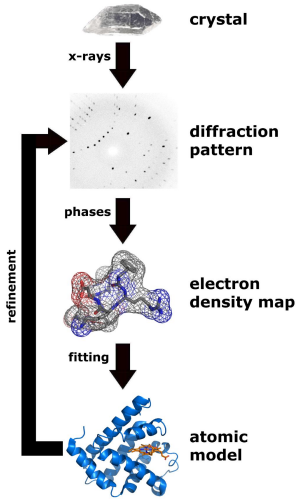
Proteasome

https://upload.wikimedia.org/wikipedia/commons/0/00/Atp_synthase.PNG
https://upload.wikimedia.org/wikipedia/commons/5/54/Proteaosome_1fnt_side.png

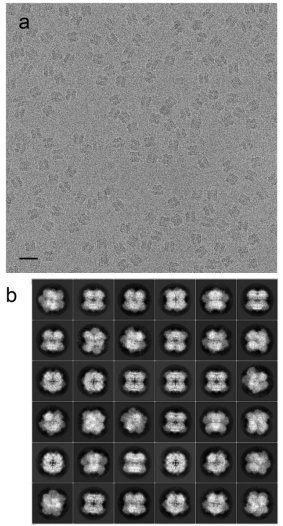
Primary structure is easier to infer than higher order structures



Central dogma of molecular biology



X-ray crystallography



Cryo-Electron microscopy

https://upload.wikimedia.org/wikipedia/commons/6/68/Central_Dogma_of_Molecular_Biochemistry_with_Enzymes.jpg
https://upload.wikimedia.org/wikipedia/commons/7/73/X_ray_diffraction.png
Structure of Alcohol Oxidase from Pichia pastoris by Cryo-Electron Microscopy, Vonck et al. (2016)

Protein Structure Encodes Higher Order Structure

- It was discovered that Bovine Pancreatic Ribonuclease denatured in urea renatures under more optimal conditions
- Thermodynamic hypothesis: protein structure is the solution to an optimization problem encoded by the sequence
    - This is not always true! Counterexamples exist in the form of aggregation, chaperones, misfolding, etc.

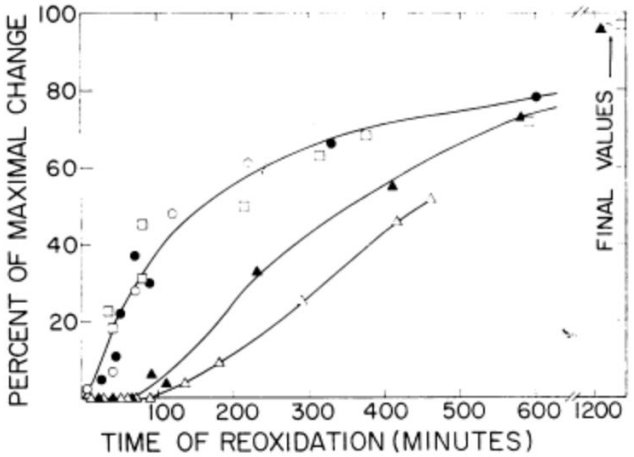


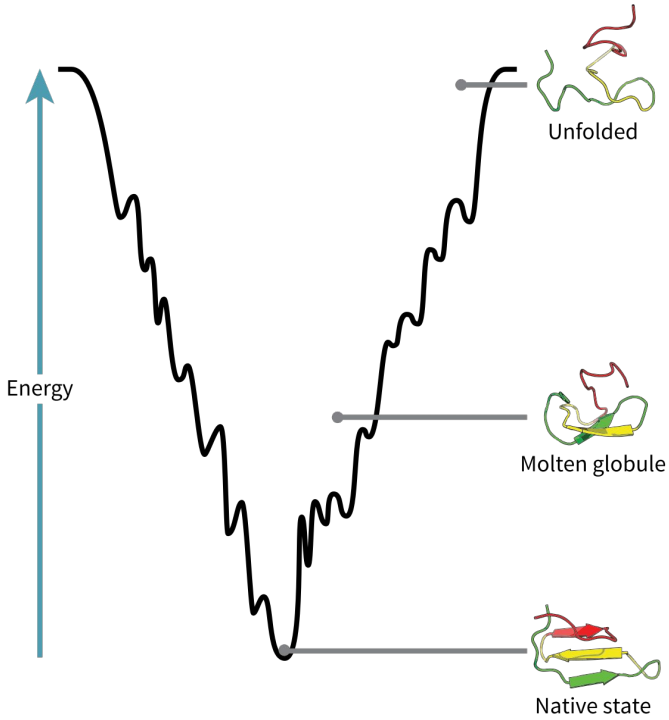
FIG. 1.—Changes, during the oxidation of reduced ribonuclease, in SH groups as followed by titration with *p*-chloromercuribenzoate (●) and by reaction with radioactive iodoacetate (○), in optical rotation (□), and in enzymatic activity as measured against ribonucleic acid (▲) and against uridylic-2′,3′-cyclic phosphate (△).

The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain, Anfinsen et al. (1961)

The thermodynamic hypothesis implies sequence based structure prediction is possible

If the protein structure is the solution to an optimization problem encoded by the sequence, then maybe we can find this solution computationally.

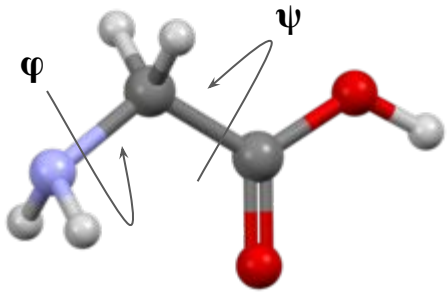
This is the goal of protein structure prediction.



https://upload.wikimedia.org/wikipedia/commons/9/91/Folding_funnel_schematic.svg

3D Protein Structure is Parameterized by Amino Acid Torsion Angles

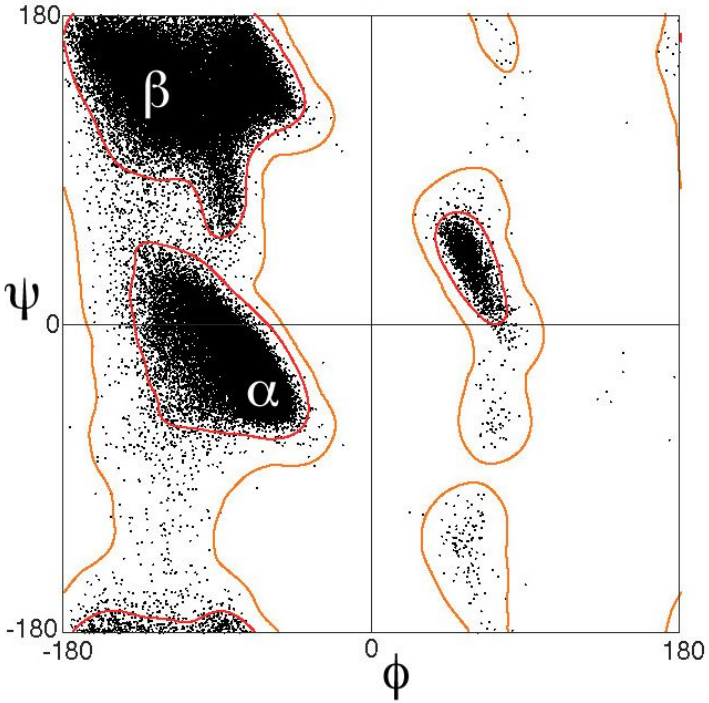
- An amino acid has two degrees of freedom arising from its φ and ψ angles



- The states of the angles parametrize the structure of the protein
- If we discretize each angle to be in one of three states due to torsional strain, we have a massive combinatorial space of possible conformations

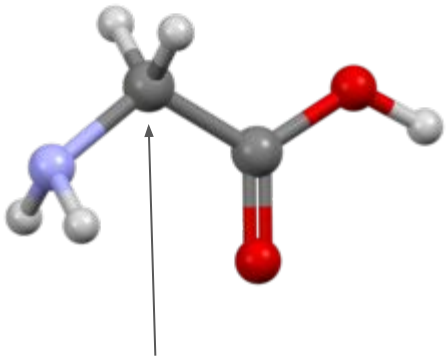
https://upload.wikimedia.org/wikipedia/commons/0/01/Glycine-neutral-lpttt-conformer-3D-bs-17.png

Ramachandran plots



https://upload.wikimedia.org/wikipedia/commons/9/90/Ramachandran_plot_general_100K.jpg

Representation of a protein structure as a point cloud in 3D space

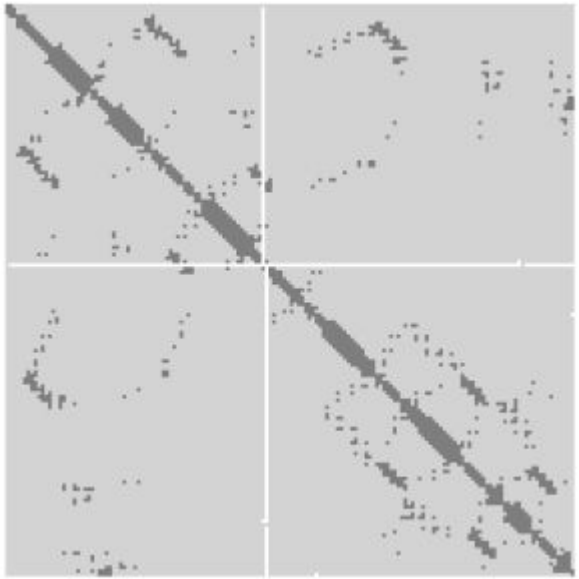


The alpha carbon (Cα) is often used as the “center” of the amino acid

https://upload.wikimedia.org/wikipedia/commons/0/01/Glycine-neutral-lpttt-conformer-3D-bs-17.png

https://www.rcsb.org/3d-view/4hhb/1

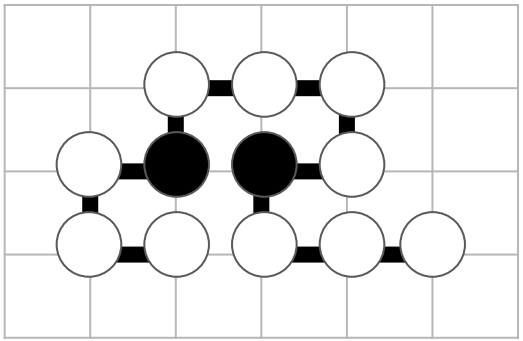
Alternative representations/simplifications of protein structure



Contact map

EEEEEEECCCCCCHHHHHHHHCC

Secondary structure



Lattice walk

https://upload.wikimedia.org/wikipedia/commons/8/8d/Protein_Contact_Map%2C_2-Color%2C_2QIP-A.png

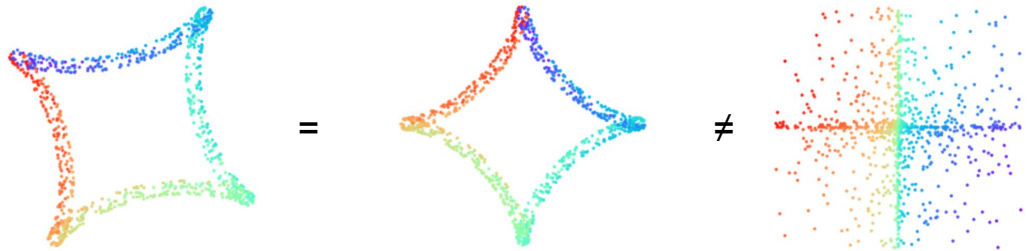
Structure assessment

Critical Assessment of protein Structure Prediction (CASP)

- Since 1994, every two years a contest is held to see who can best predict protein structures from peptide sequences
- Targets structures are held from publication until results are in

Aligning two point clouds

We want to factor out rigid motion when comparing a pair of point clouds



A popular approach is to minimize the root mean square distance (RMSD)

$$\mathrm{RMSD(P,Q)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\|p_i - q_i\|_2^2\right)}$$

Root mean squared distance is a computable but flawed metric

- Given a pair of structures, minimize the sum of the squared distances between Cα
  - Kabsch algorithm
- Taking the square root gives the RMSD
- Small imperfections (outliers) can ruin a pretty good alignment since distances are squared

Finding the optimal rigid transformation

- Define data matrices P and Q

$$P = \begin{bmatrix} p_{1x} & p_{1y} & p_{1z} \\ & \vdots & \\ p_{nx} & p_{ny} & p_{nz} \end{bmatrix} \qquad Q = \begin{bmatrix} q_{1x} & q_{1y} & q_{1z} \\ & \vdots & \\ q_{nx} & q_{ny} & q_{nz} \end{bmatrix}$$

- Translate each point cloud so that they are centered at the origin
- Want to find optimal rotation R such that RMSD between PR and Q is minimized

Finding the optimal rigid transformation

$$\begin{aligned}
\underset{R}{\operatorname{argmin}}(\mathrm{RMSD}(PR, Q)) &= \underset{R}{\operatorname{argmin}} \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\|R(p_i) - q_i\|_2^2\right)} \\
&= \underset{R}{\operatorname{argmin}} \sum_{i=1}^{n}\sum_{j\in\{x,y,z\}}\left((R(p_i)_j - q_{ij})^2\right) \\
&= \underset{R}{\operatorname{argmax}} \sum_{i=1}^{n}\sum_{j\in\{x,y,z\}} R(p_i)_j q_{ij} \\
&= \underset{R}{\operatorname{argmax}} \Big(tr(R^T P^T Q)\Big)
\end{aligned}$$

R is a rotation so it cannot affect the singular values that upper bound eigenvalues

$$tr(R^T U\Sigma V^T) \leq tr(\Sigma)$$

Choose R such that eigenvalues coincide with singular values

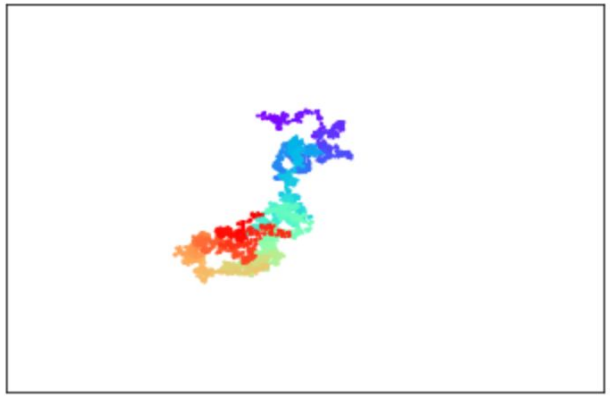
$$\begin{aligned}
R &= UV^T \\
\implies tr(R^T U\Sigma V^T) &= tr(V\Sigma V^T) \\
&= tr(\Sigma)
\end{aligned}$$

Problems with RMSD

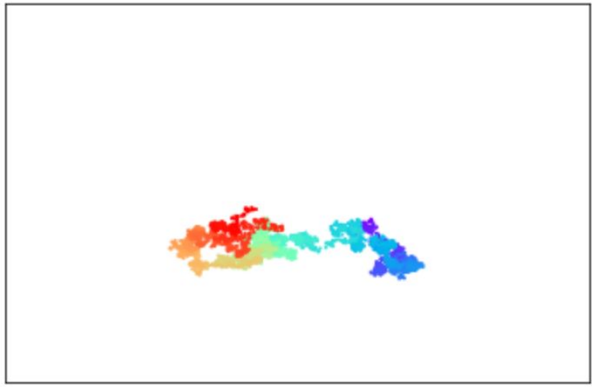
A large error in a small part of a prediction can tank a structure's RMSD due to squaring

$$\mathrm{RMSD(P,Q)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\|p_i - q_i\|_2^2\right)}$$

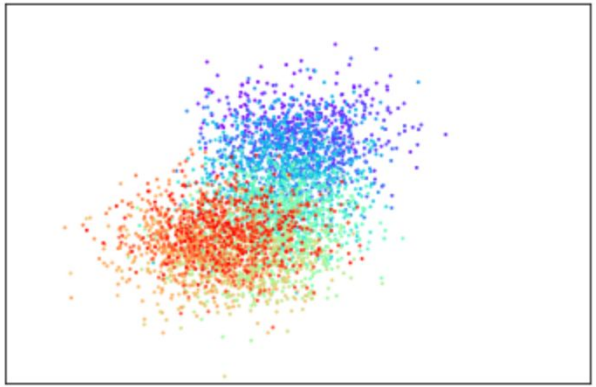
The Global Distance Test (GDT) instead measures the fraction of points that can be matched up between two point clouds up to a given distance threshold.



RMSD: 0.67
GDT: ~0.5



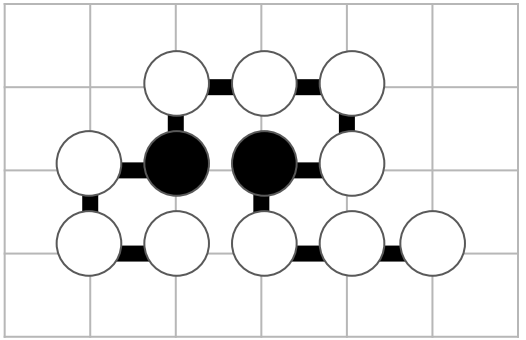
RMSD: 0.35
GDT: ~0



Physical modelling

Simplified models of protein folding

CCCCCCCCCHHHHCCCCCCCC
CCCCCCCCHHHHHHCCCCCCC
CCCCCCHHHHHHHHHCCCCCC

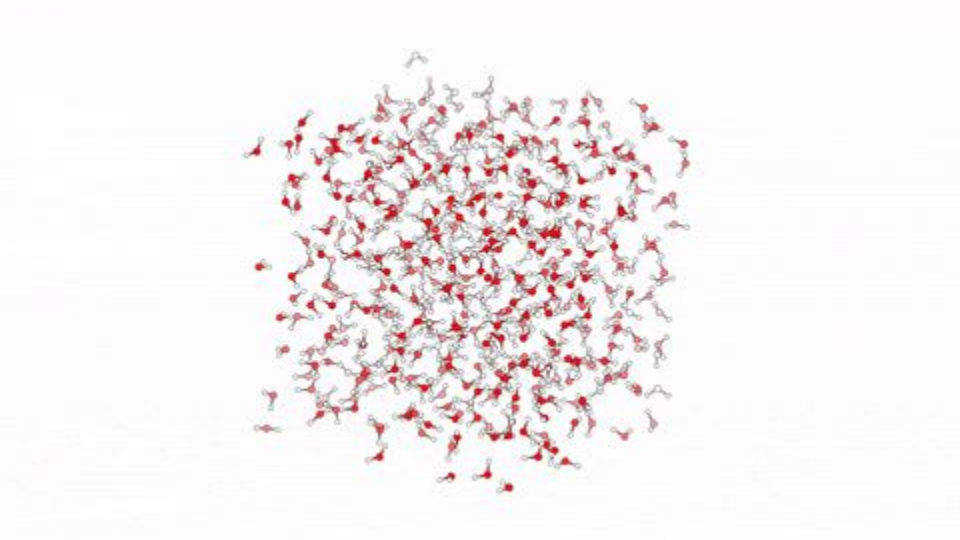


Ising models allow us to derive the average helicity of a canonical ensemble and model its behaviour (e.g. helix propagation)

Lattice based folding models allow us to understand the combinatorial complexity of protein folding (protein folding is NP-hard)

Molecular Dynamics Simulations

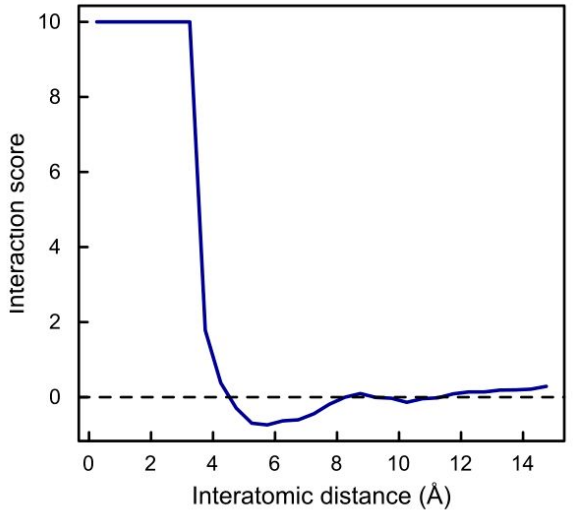
- Simulate the folding of the protein by simulating the physics directly
- Very computationally expensive



https://upload.wikimedia.org/wikipedia/commons/f/f4/MD_water.gif

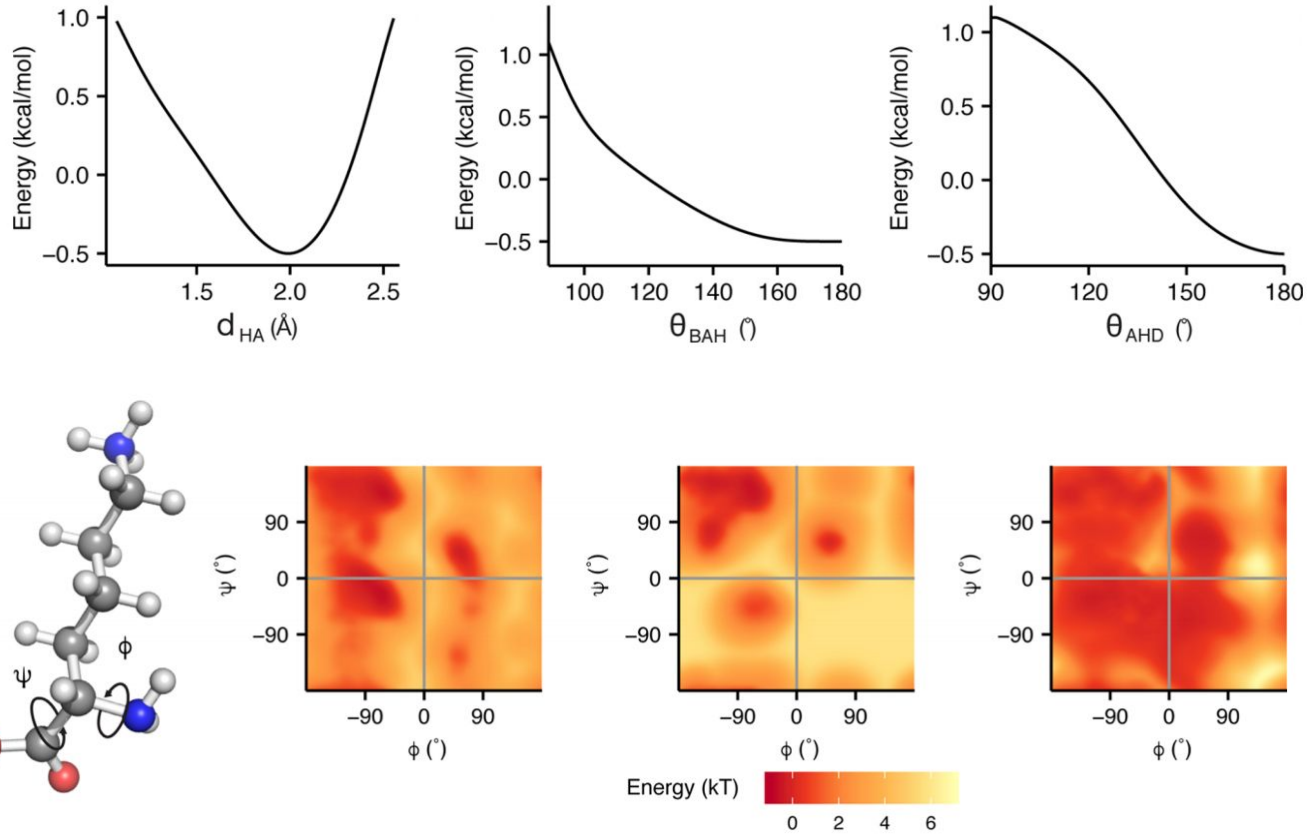
Statistical Potentials

- Given a triplet (amino acid 1, amino acid 2, distance between them), we can look at how common that configuration occurs
  - This gives us an indication of how stable the configuration is (potential energy)
- Initially these potentials are interpreted to be physical energy values
- Later shown that they are more reflective of evolutionary history than physical energy



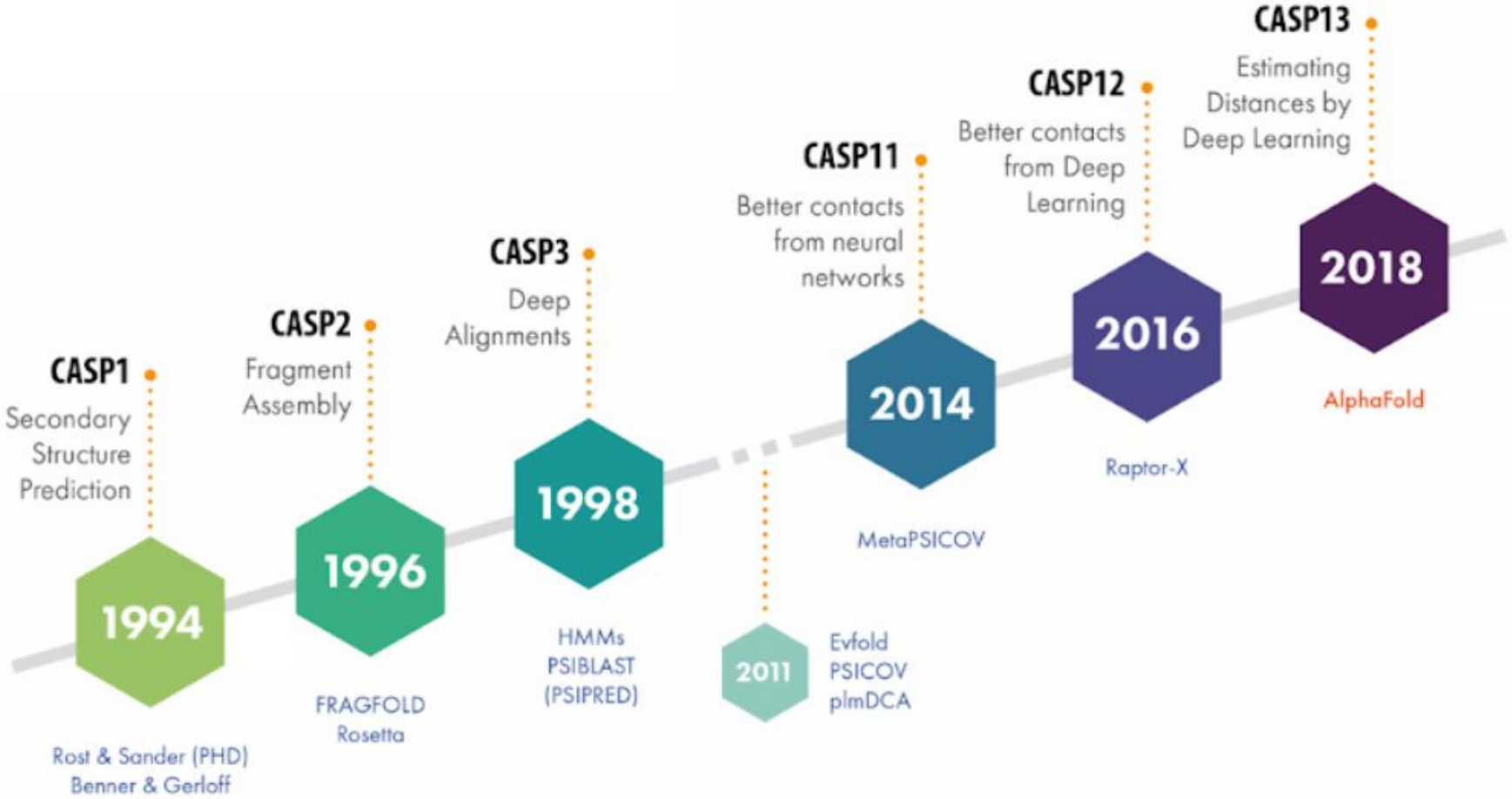
https://upload.wikimedia.org/wikipedia/commons/2/2b/ICBVCB.svg

Rosetta

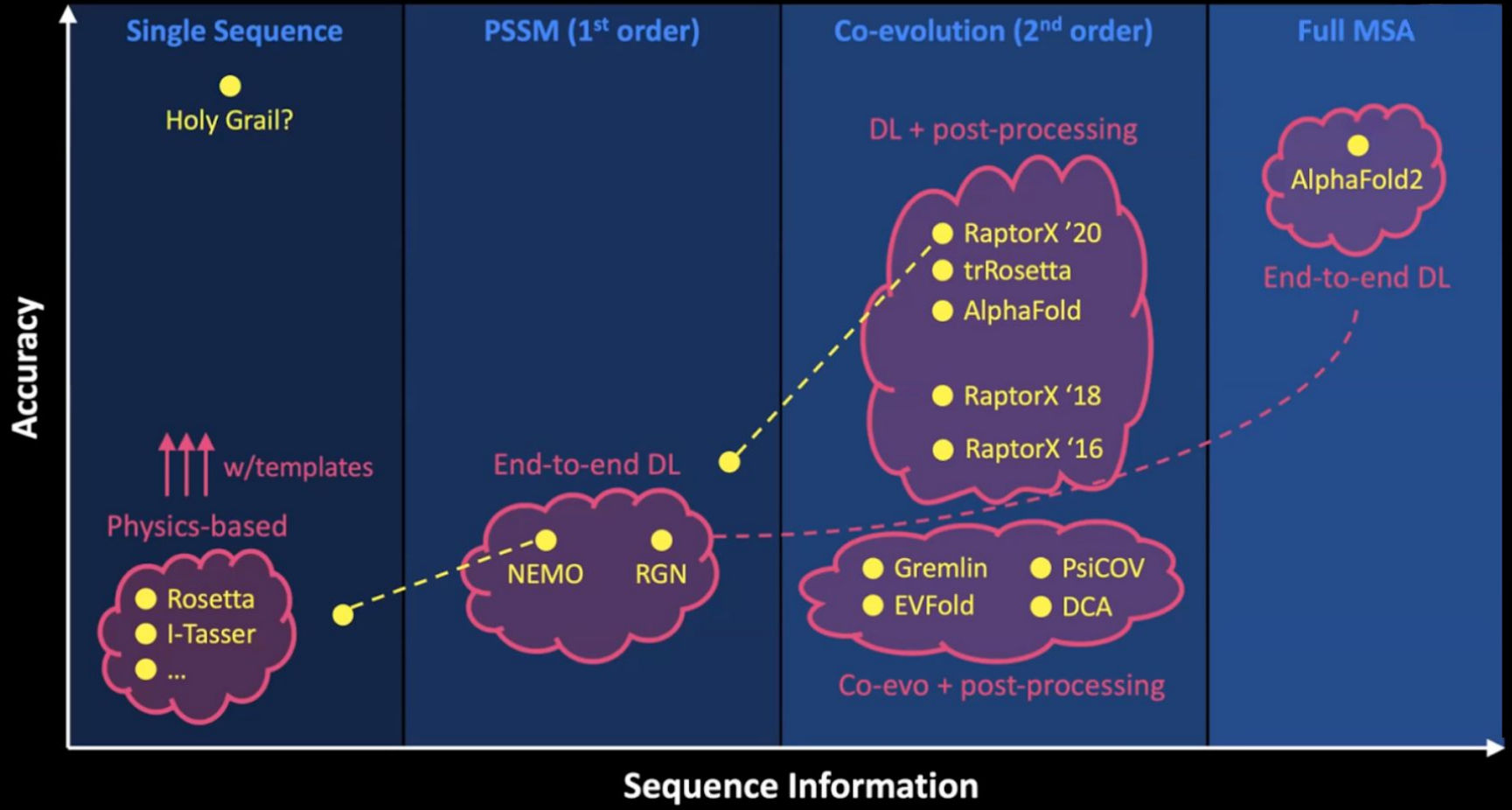


The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design

Evolution of structure prediction



Slide from David Jones



Slide from Mohammed AlQuraishi

AlphaFold

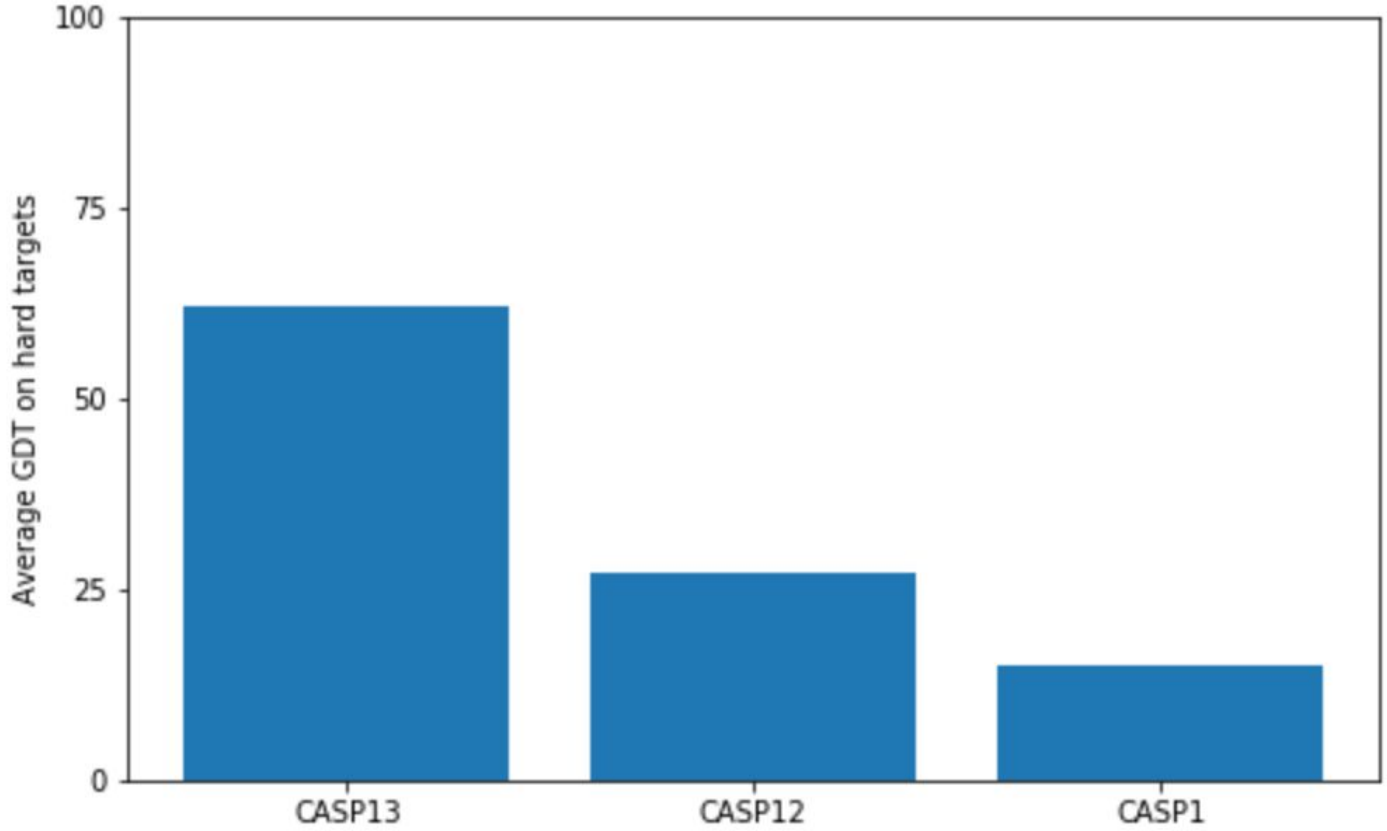
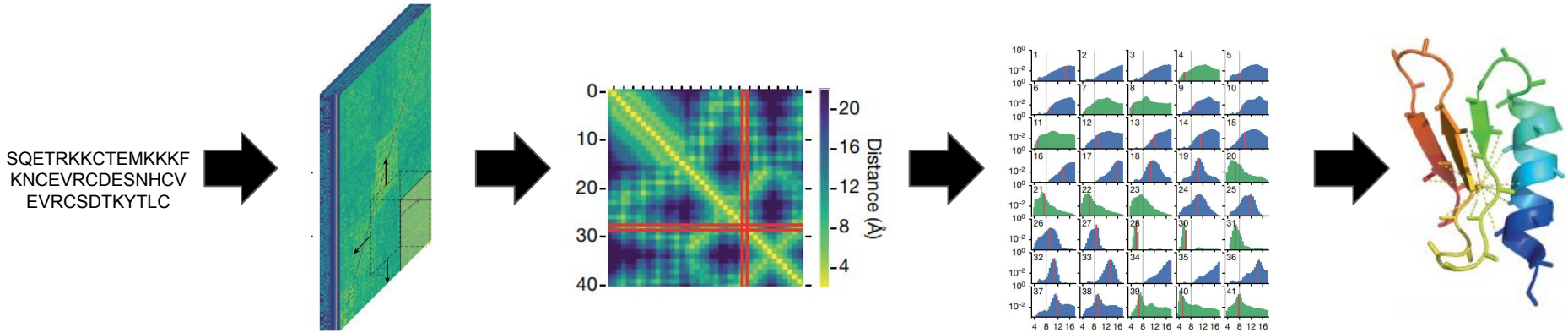


Figure loosely adapted from David Jones

Breaking down the AlphaFold pipeline

1. Raw sequence data is converted into a feature tensor
2. CNNs produces distance predictions for each pair of c_α-atoms
3. Distance predictions are used to derive distance based potentials
4. The φ and ψ angles are derived by optimization using the potentials



Improved protein structure prediction using potentials from deep learning

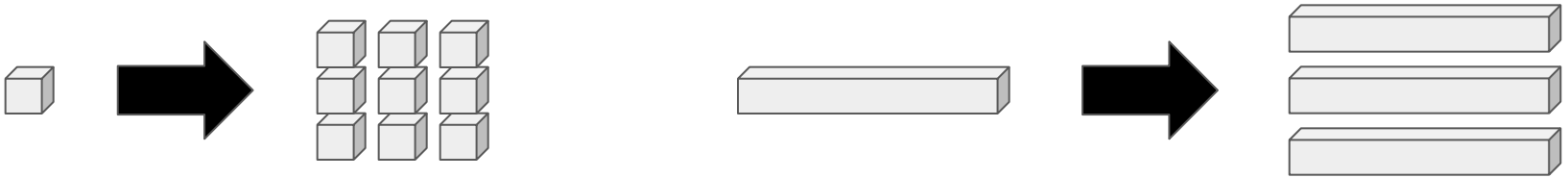
Step 1: Feature engineering

- **Scalar** (1 feature): Number of similar proteins
- **Vector** (139 features): Sequence, alignment, position
- **Matrix** (486 features): Coevolution

- Number of HHblits alignments (scalar).
- Sequence-length features: 1-hot amino acid type (21 features); profiles: PSI-BLAST (21 features), HHblits profile (22 features), non-gapped profile (21 features), HHblits bias, HMM profile (30 features), Potts model bias (22 features); deletion probability (1 feature); residue index (integer index of residue number, consecutive except for multi-segment domains, encoded as 5 least-significant bits and a scalar).
- Sequence-length-squared features: Potts model parameters (484 features, fitted with 500 iterations of gradient descent using Nesterov momentum 0.99, without sequence reweighting); Frobenius norm (1 feature); gap matrix (1 feature).

Improved protein structure prediction using potentials from deep learning

All features are transformed into 2d tensors

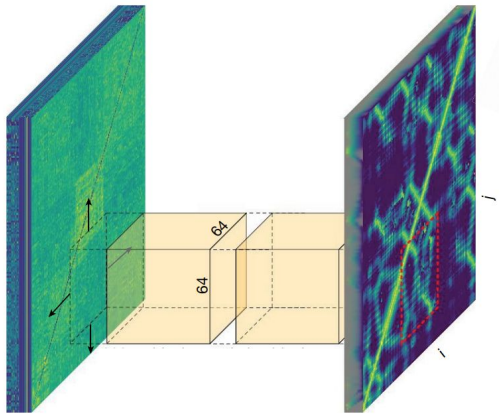


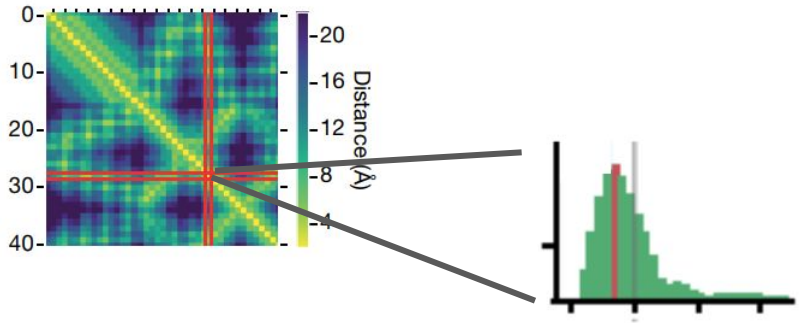
- Subsequent stacking yields a square tensor with 600+ channels

https://upload.wikimedia.org/wikipedia/commons/6/68/Messy_kitchen_sink.jpg

Step 2: Distance map prediction

- The output space consists of 64 discrete equally sized “distance classes”
- Given a 64x64 piece of the square tensor, the task is to output a 64x64x64 tensor that gives the distance distribution for each position

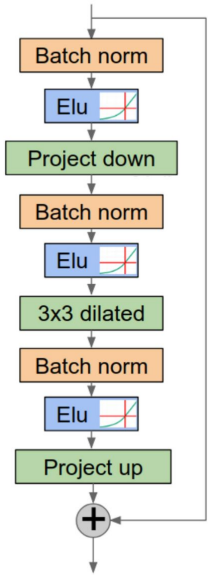




Improved protein structure prediction using potentials from deep learning

The deep learning component of AlphaFold

A convolutional residual network consisting of 220 residual blocks

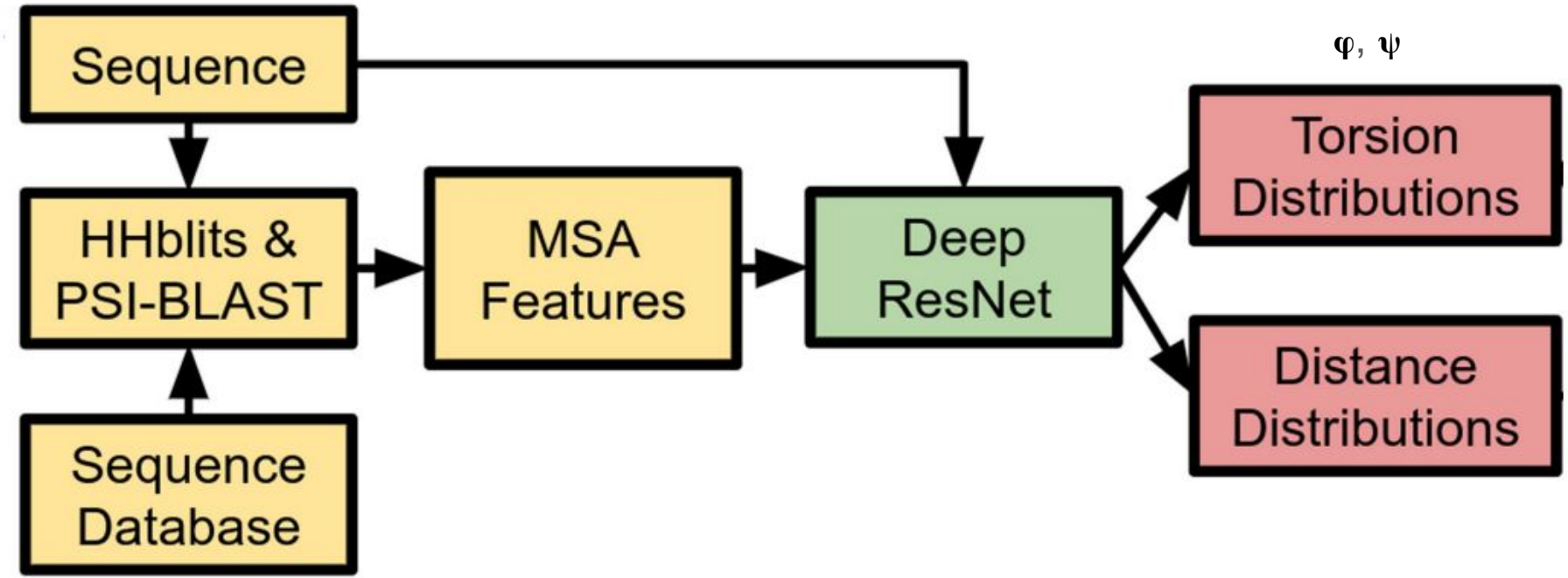


Neural network hyperparameters

- 7 groups of 4 blocks with 256 channels, cycling through dilations 1, 2, 4, 8.
- 48 groups of 4 blocks with 128 channels, cycling through dilations 1, 2, 4, 8.
- Optimization: synchronized stochastic gradient descent
- Batch size: batch of 4 crops on each of 8 GPU workers.
- 0.85 dropout keep probability.
- Nonlinearity: ELU.
- Learning rate: 0.06.
- Auxiliary loss weights: secondary structure: 0.005; accessible surface area: 0.001. These auxiliary losses were cut by a factor 10 after 100 000 steps.
- Learning rate decayed by 50% at 150,000, 200,000, 250,000 and 350,000 steps.
- Training time: about 5 days for 600,000 steps.

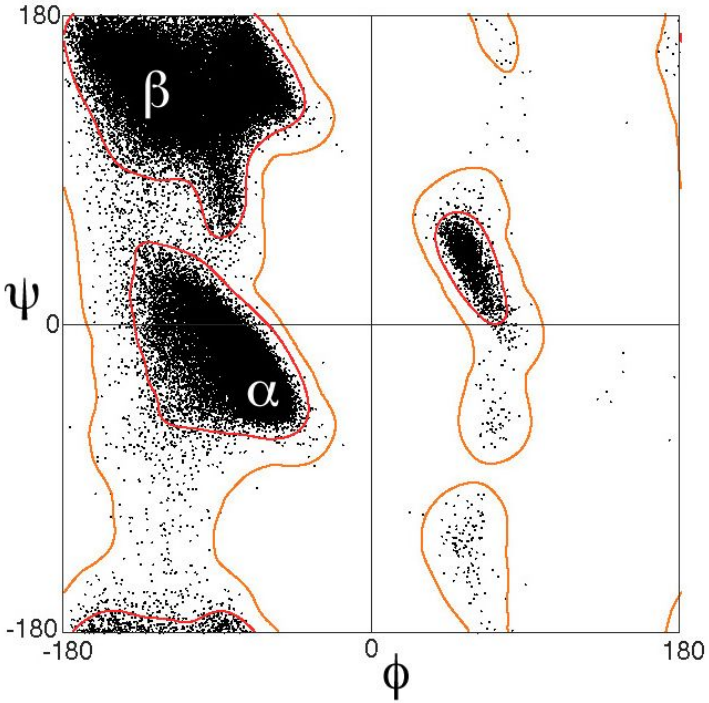
Improved protein structure prediction using potentials from deep learning

Recap



Improved protein structure prediction using potentials from deep learning

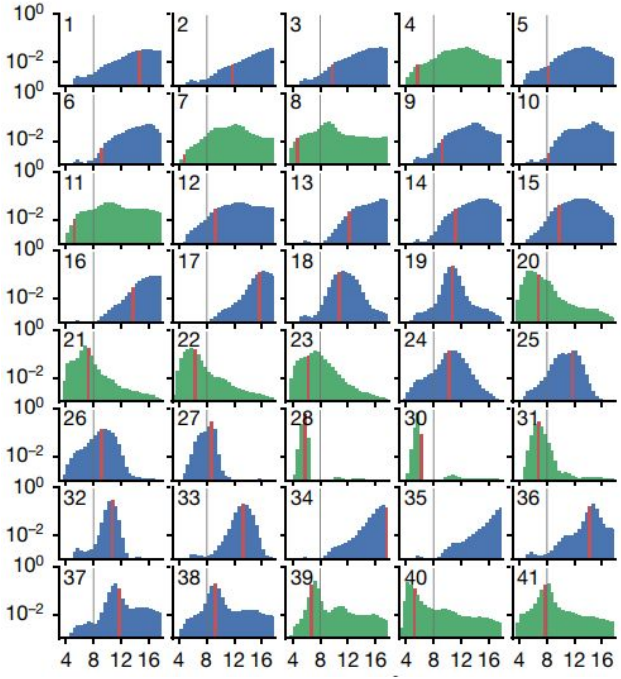
Ramachandran plots



https://upload.wikimedia.org/wikipedia/commons/9/90/Ramachandran_plot_general_100K.jpg

Step 3: Obtaining distance potentials

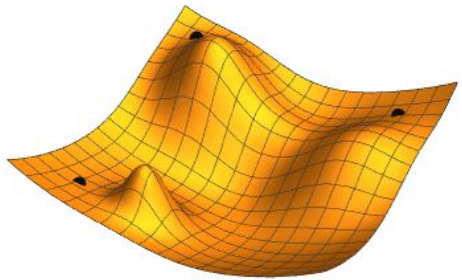
- Model outputs distribution over 64 discrete equally sized “distance classes” between 2-22Å
- Ramachandran plots are discretized to 10° by 10° chunks
- To optimize, these must be made differentiable
  - Splines
  - Fitting “Gaussian distributions”



Improved protein structure prediction using potentials from deep learning

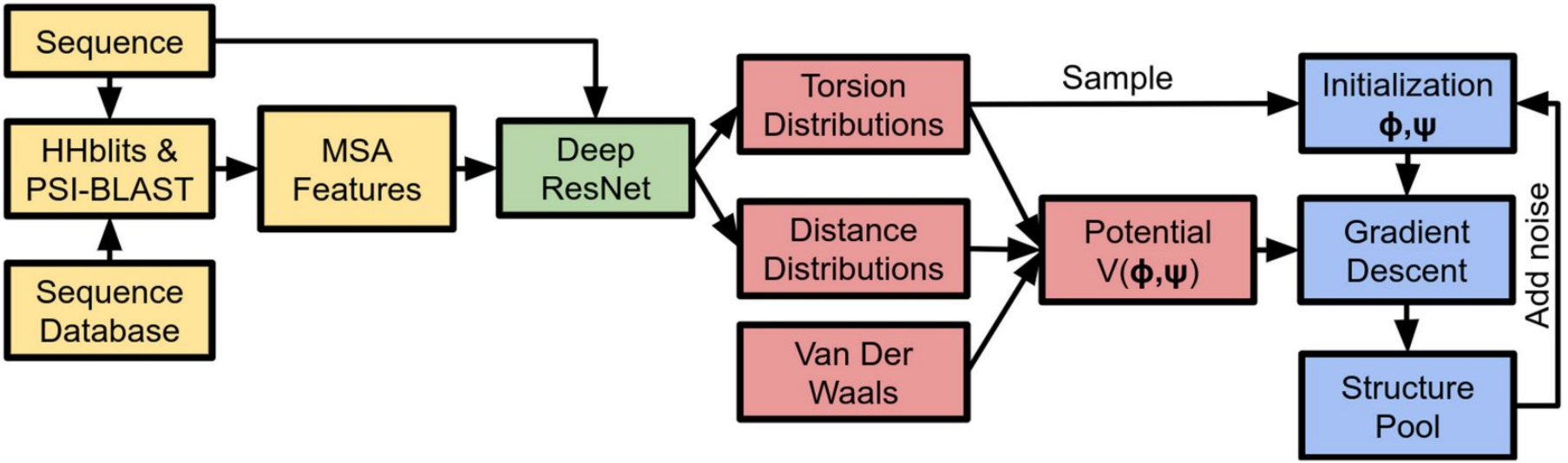
Step 4: Optimize the φ and ψ angles

- So far the distance based potentials and torsion angle potentials have been computed
- An additional van der Waals term from Rosetta is added to prevent physically impossible structures
- Optimization is done by minimizing the sum of these three terms
- Initial state is drawn from Ramachandran plots



https://upload.wikimedia.org/wikipedia/commons/a/a3/Gradient_descent.gif

Overview



Improved protein structure prediction using potentials from deep learning

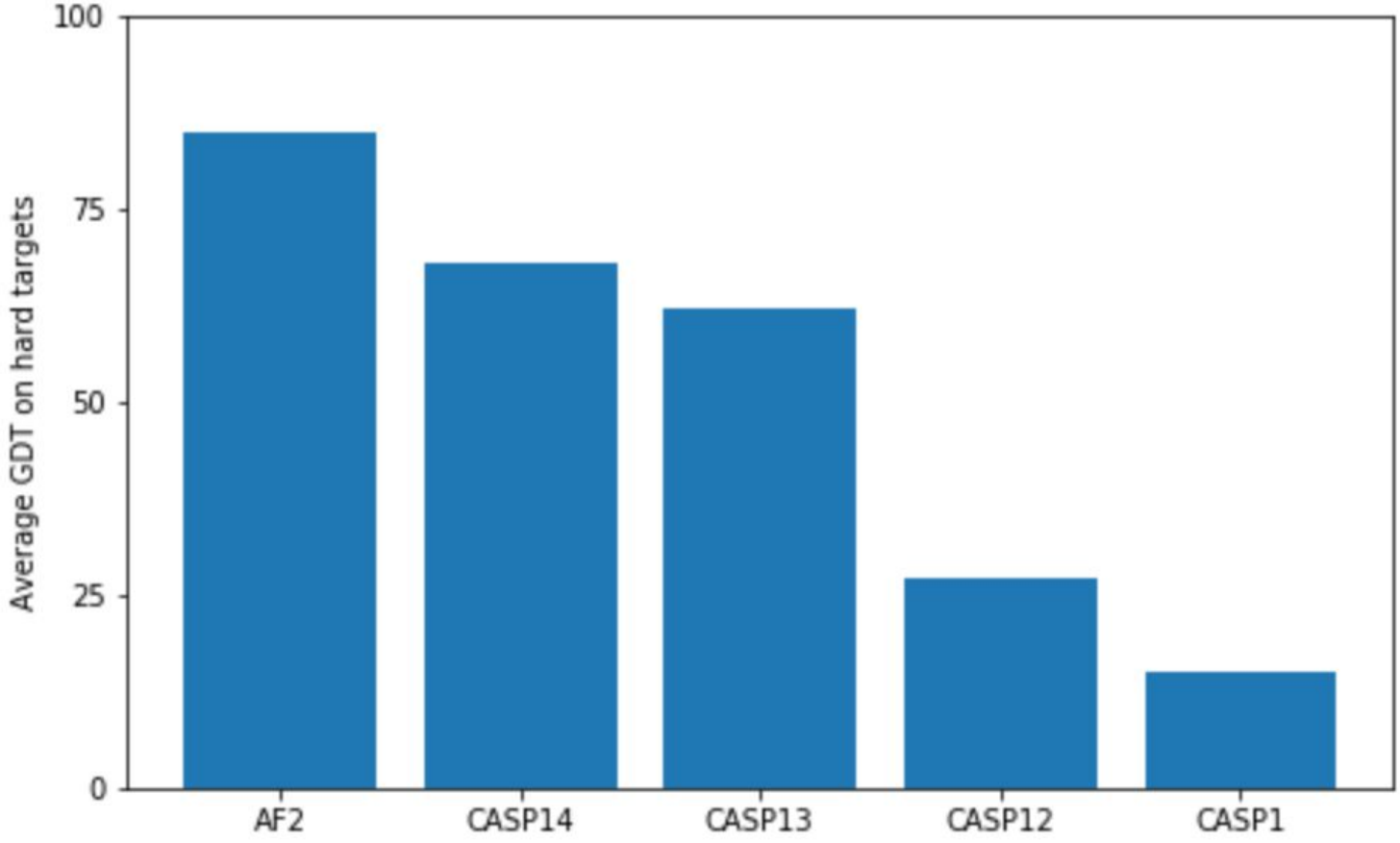


Figure loosely adapted from David Jones

What is AlphaFold2?

A folded protein can be thought of as a “spatial graph”, where residues are the nodes and edges connect the residues in close proximity. This graph is important for understanding the physical interactions within proteins, as well as their evolutionary history. For the latest version of AlphaFold, used at CASP14, we created an attention-based neural network system, trained end-to-end, that attempts to interpret the structure of this graph, while reasoning over the implicit graph that it’s building. It uses evolutionarily related sequences, multiple sequence alignment (MSA), and a representation of amino acid residue pairs to refine this graph.

By iterating this process, the system develops strong predictions of the underlying physical structure of the protein and is able to determine highly-accurate structures in a matter of days. Additionally, AlphaFold can predict which parts of each predicted protein structure are reliable using an internal confidence measure.

DeepMind blog

???

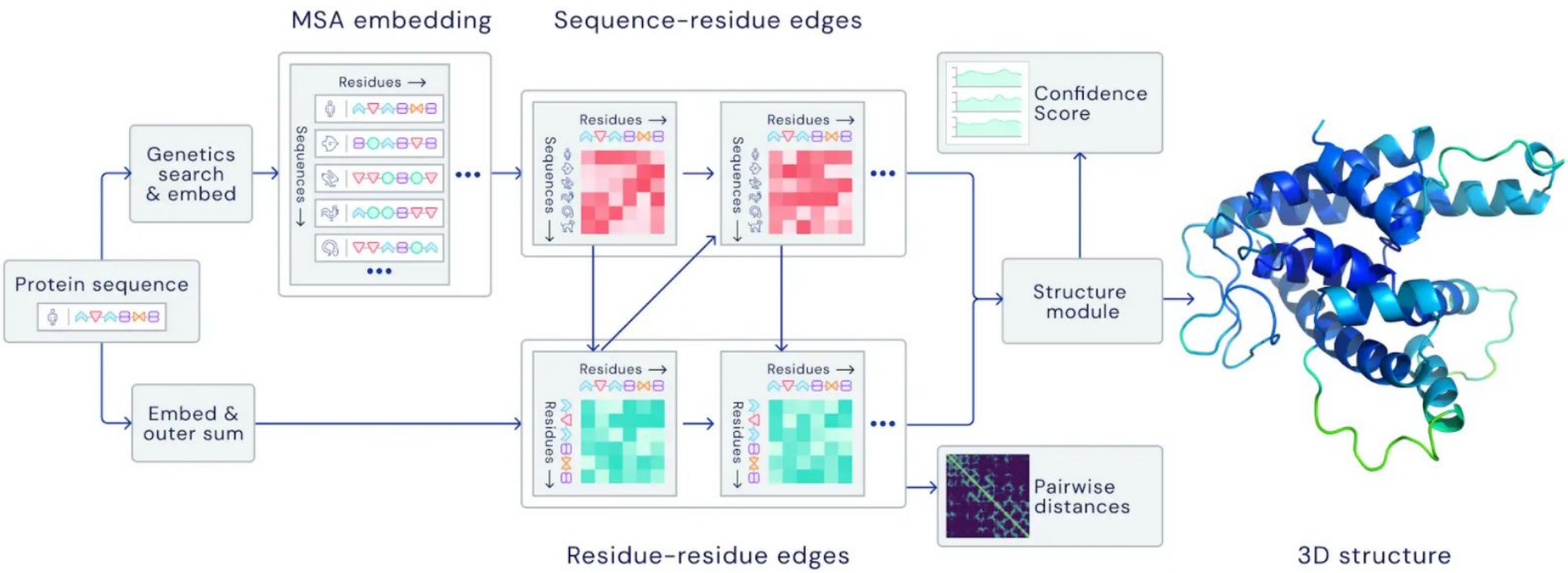
A folded protein can be thought of as a “spatial graph”, where residues are the nodes and edges connect the residues in close proximity. This graph is important for understanding the physical interactions within proteins, as well as their evolutionary history. For the latest version of AlphaFold, used at CASP14, we created an attention-based neural network system, trained end-to-end, that attempts to interpret the structure of this graph, while reasoning over the implicit graph that it’s building. It uses evolutionarily related sequences, multiple sequence alignment (MSA), and a representation of amino acid residue pairs to refine this graph.

By iterating this process, the system develops strong predictions of the underlying physical structure of the protein and is able to determine highly-accurate structures in a matter of days. Additionally, AlphaFold can predict which parts of each predicted protein structure are reliable using an internal confidence measure.

??? ***???*** ***???***

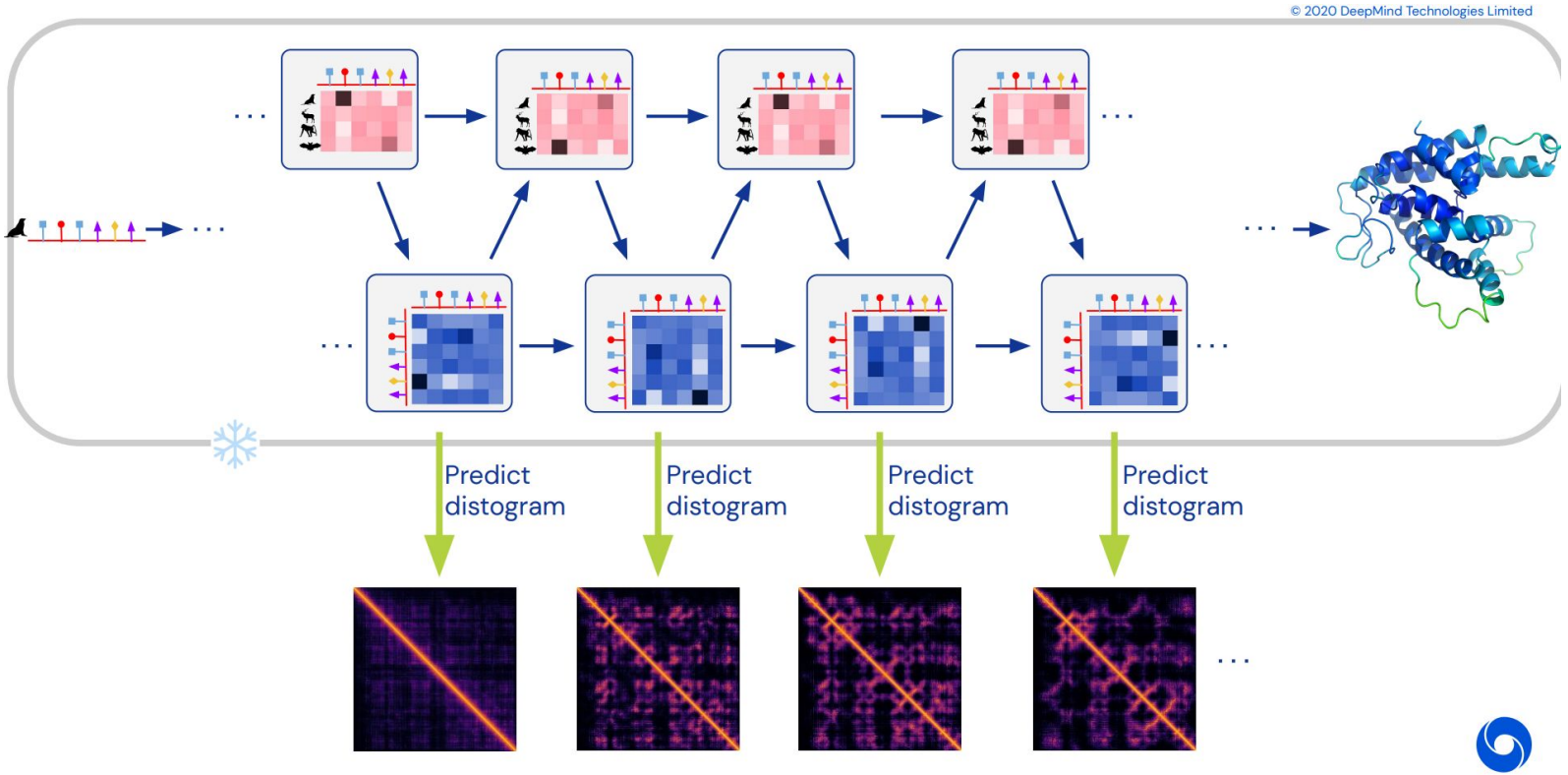
DeepMind blog

Schematic



DeepMind blog

Iterative attention based neural network system?



DeepMind blog