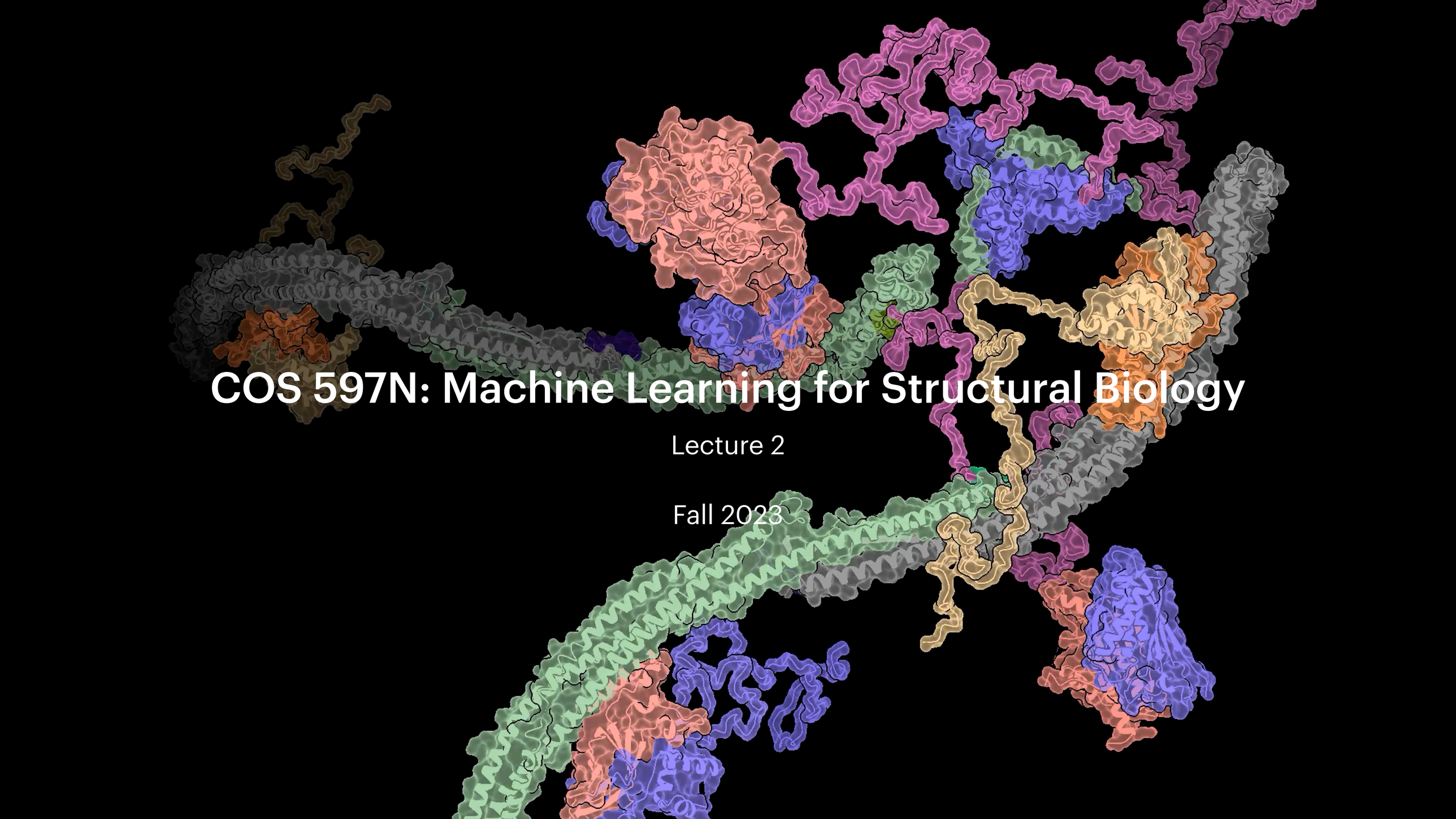


COS 597N: Machine Learning for Structural Biology

Lecture 2

Fall 2023

Course Logistics

- Make sure you are added to slack
- Occasionally I will send emails through Canvas (i.e. make sure you are registered for the course)
- Optional student-only “precept”? I can send out a doodle poll for anyone who is interested in discussing papers/ prepping presentations
- Course website: Draft schedule is almost up... sorry for the delay
- Next week:
  - Viola Chen and Xiaxin Shen will present the AlphaFold2 paper
  - I will present AlphaFold DB results
  - Feedback: Andy Zhang, Brendan Wang

This lecture
AlphaFold1

- Prior work & background
  - “The Protein-Folding Problem, 50 Years on”
- The AlphaFold1 Pipeline
- Results

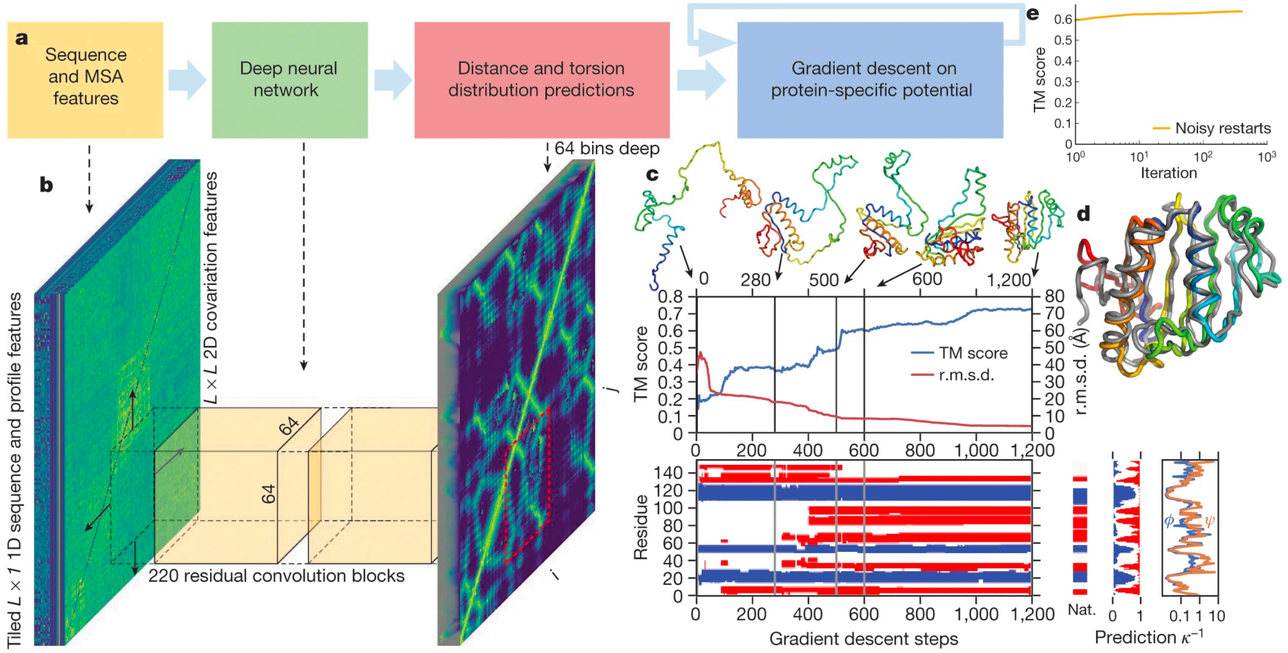
- Question:
  - What did you think of the papers?
  - Which one did you prefer reading?

Article

Improved protein structure prediction using potentials from deep learning

https://doi.org/10.1038/s41586-019-1923-7
Received: 2 April 2019
Accepted: 10 December 2019
Published online: 15 January 2020

Andrew W. Senior[1,4*], Richard Evans[1,4], John Jumper[1,4], James Kirkpatrick[1,4], Laurent Sifre[1,4], Tim Green[1], Chongli Qin[1], Augustin Žídek[1], Alexander W. R. Nelson[1], Alex Bridgland[1], Hugo Penedones[1], Stig Petersen[1], Karen Simonyan[1], Steve Crossan[1], Pushmeet Kohli[1], David T. Jones[2,3], David Silver[1], Koray Kavukcuoglu[1] & Demis Hassabis[1]



CASP13: Competition throughout 2018
CASP13: Results announced December 2018

Recap: “The Protein-Folding Problem, 50 Years on”
Dill & Maccallum, Science 2012

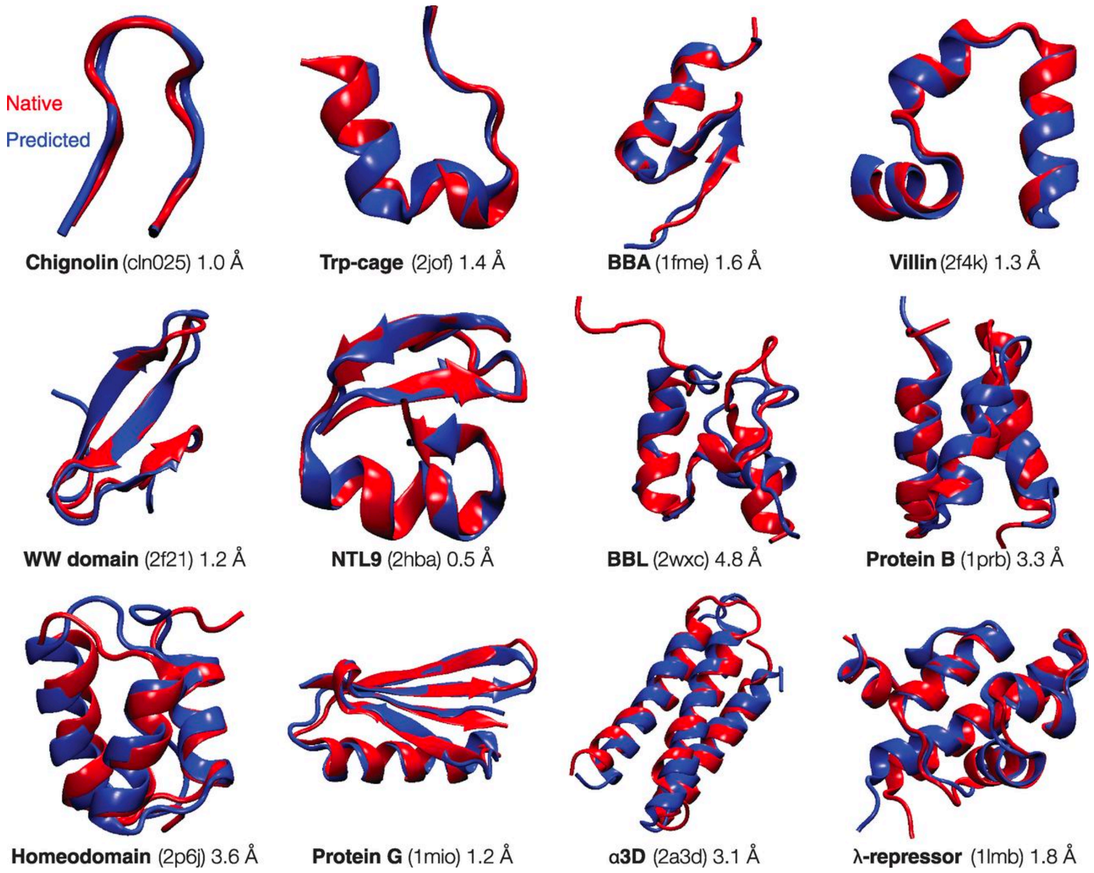
Three broad questions:

1. What is the physical code by which an amino acid sequence dictates a protein’s native structure? (Themodynamics)
2. How can proteins fold so fast? (Kinetics)
3. Can we devise a computer algorithm to predict protein structures from their sequences?

> “Perhaps the most remarkable features of the molecule are its complexity and its lack of symmetry. The arrangement seems to be almost totally lacking in the kind of regularities which one instinctively anticipates, and it is more complicated than has been predicated by any theory of protein structure.”

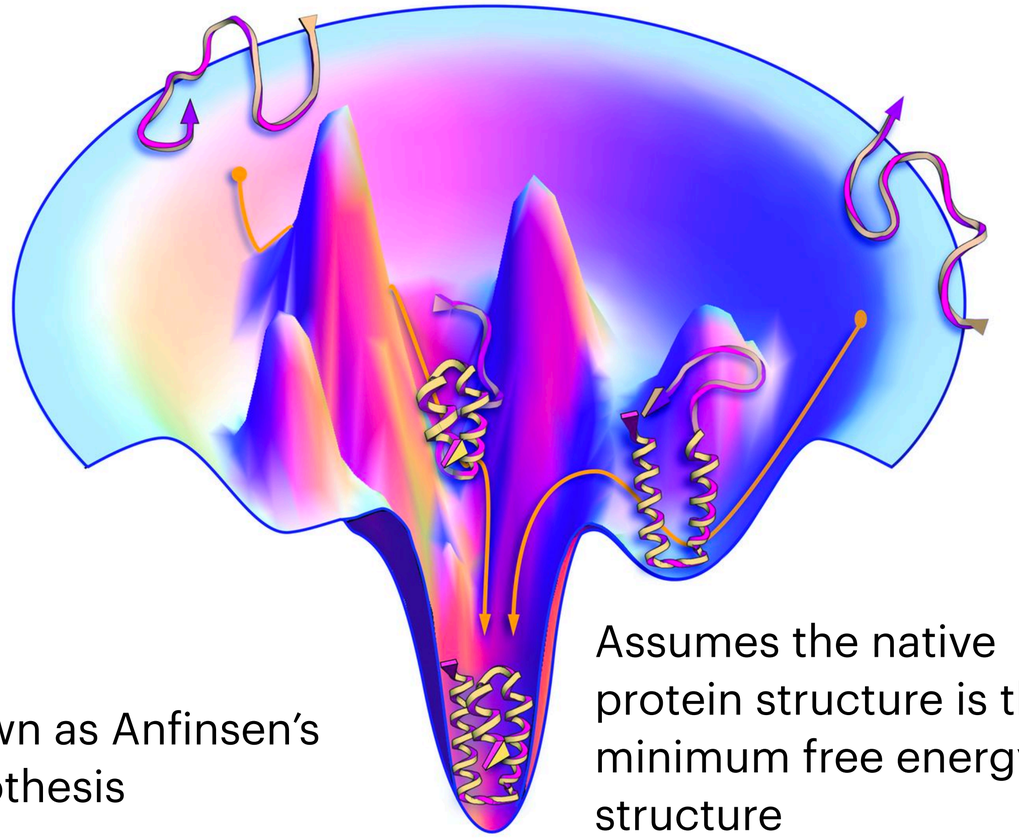
Fig 1: In 1958, Kendrew and coworkers published the first structure of a globular protein, myoglobin at 6A resolution

Protein folding is driven by physics



Lindorff-Larsen et al, Science 2011

Energy landscape theory of protein folding



Known as Anfinsen's hypothesis

Assumes the native protein structure is the minimum free energy structure

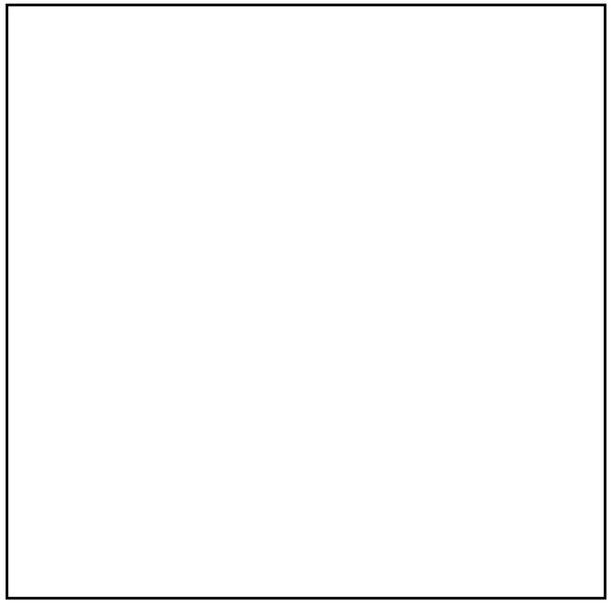
Figure from Dill & Maccallum, Science 2012

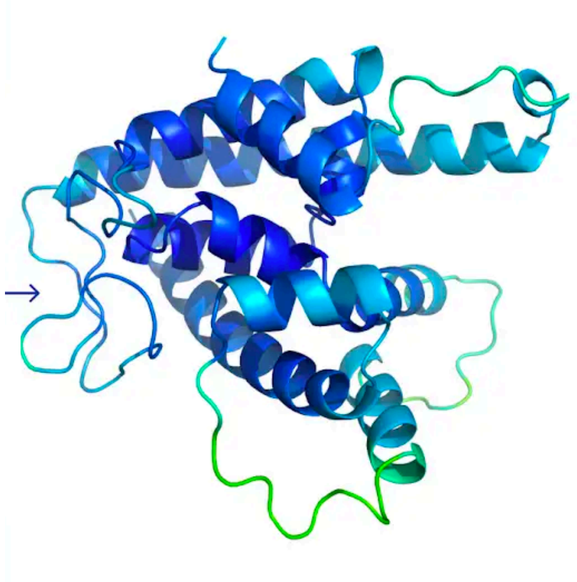
Protein structure prediction

Task: Predict the 3D coordinates of all atoms of a given input protein sequence

- How would you approach it?
  - In 2012?
  - If you had infinite compute?

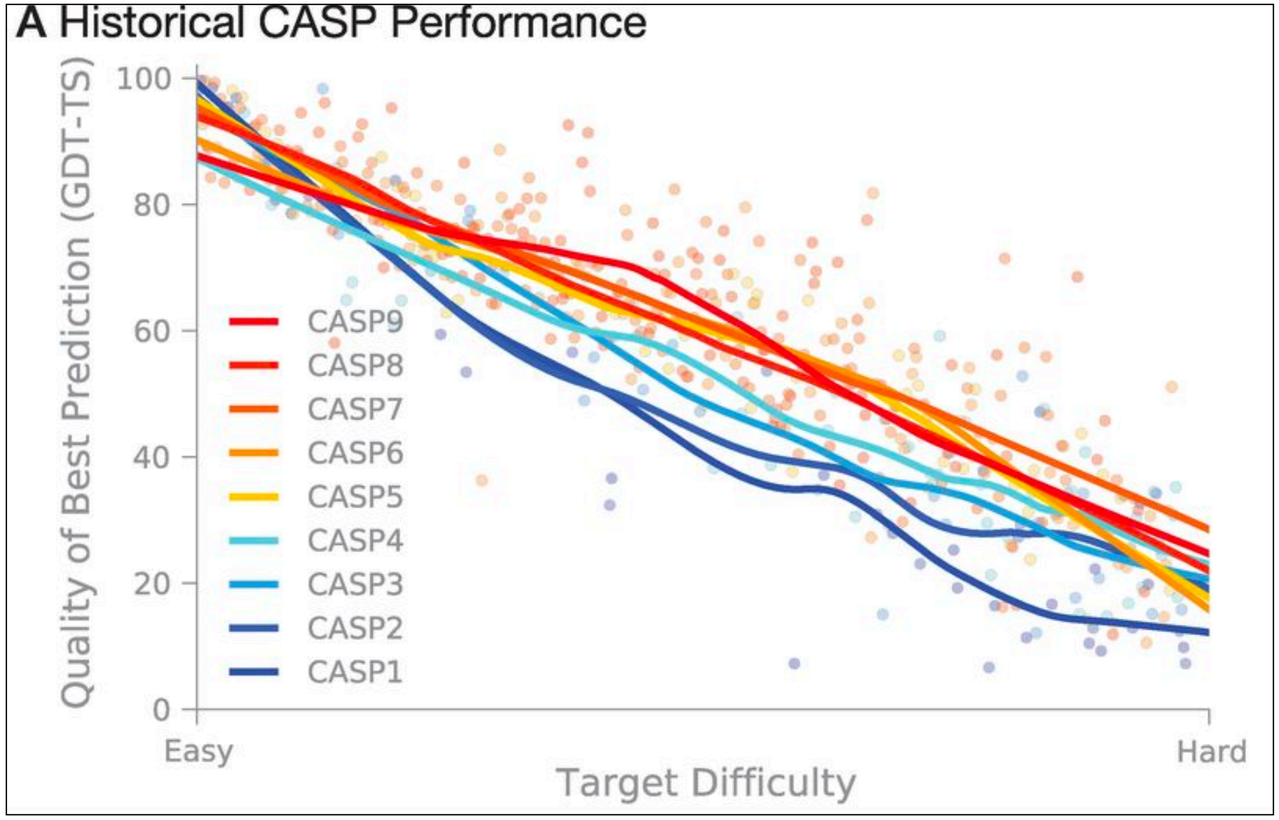
Input sequence
MRKPRTPFTT...

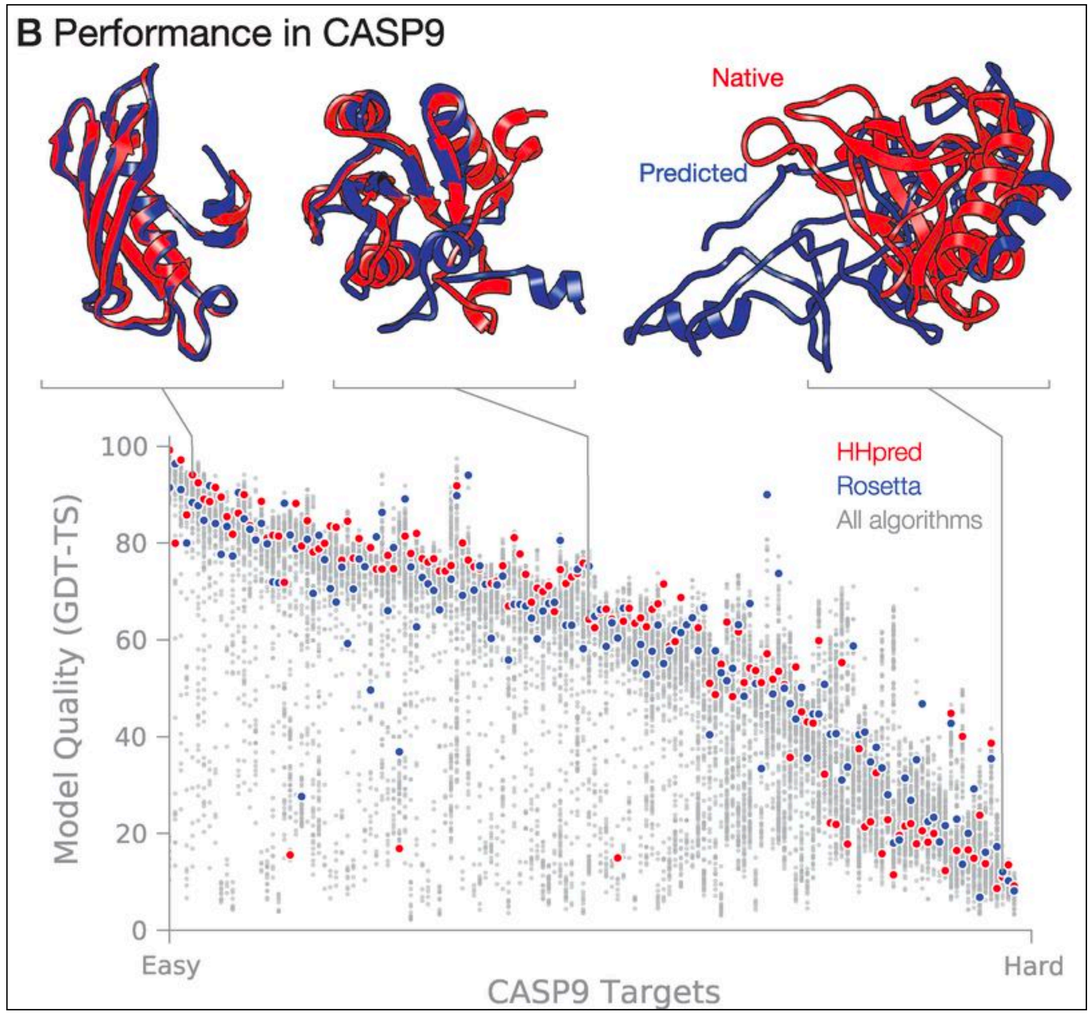




CASP: Critical Assessment of protein Structure Prediction

- Biennial, community-wide blind competition initiated by John Moult and colleagues in 1994
- ~100 target sequences, ~150 research groups
- Release ~1 protein per day, May-August. 3 weeks to return 5 predictions
- What makes a target sequence “easy” vs “hard”?



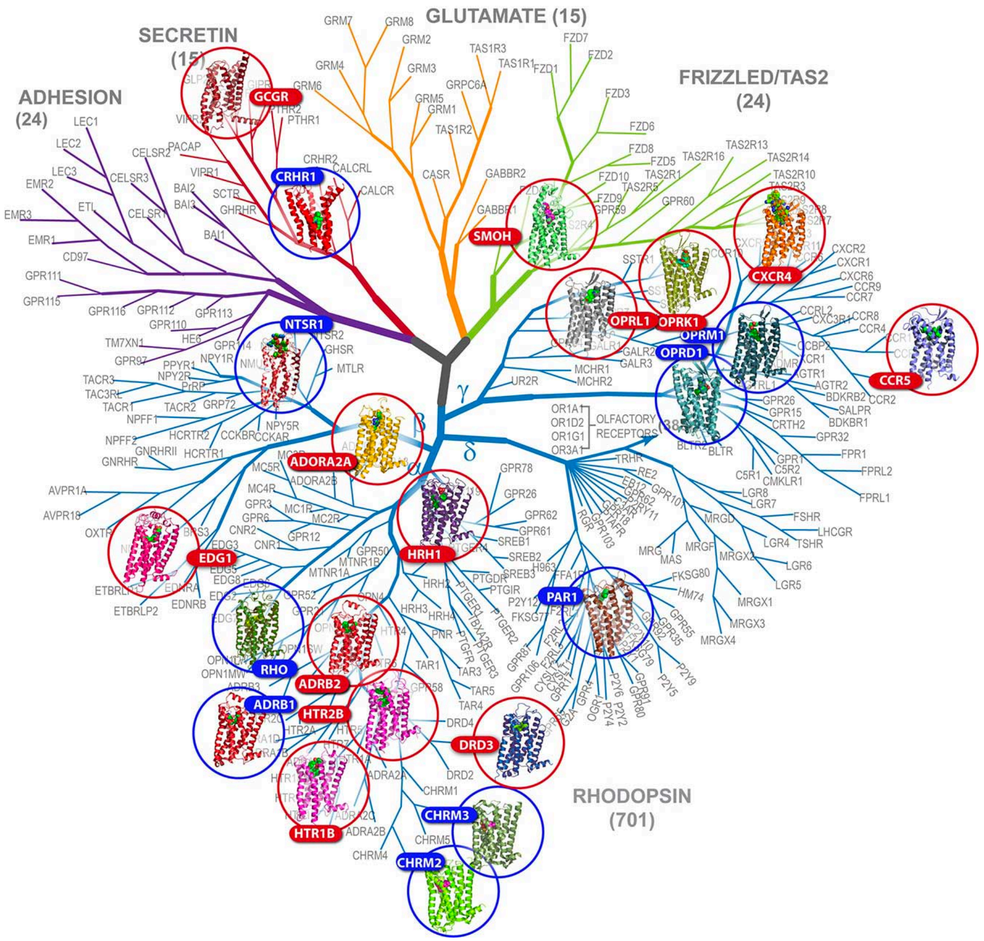


Figures from Dill & Maccallum, 2012

Template-based modeling (TBM) vs. Free modeling (FM)

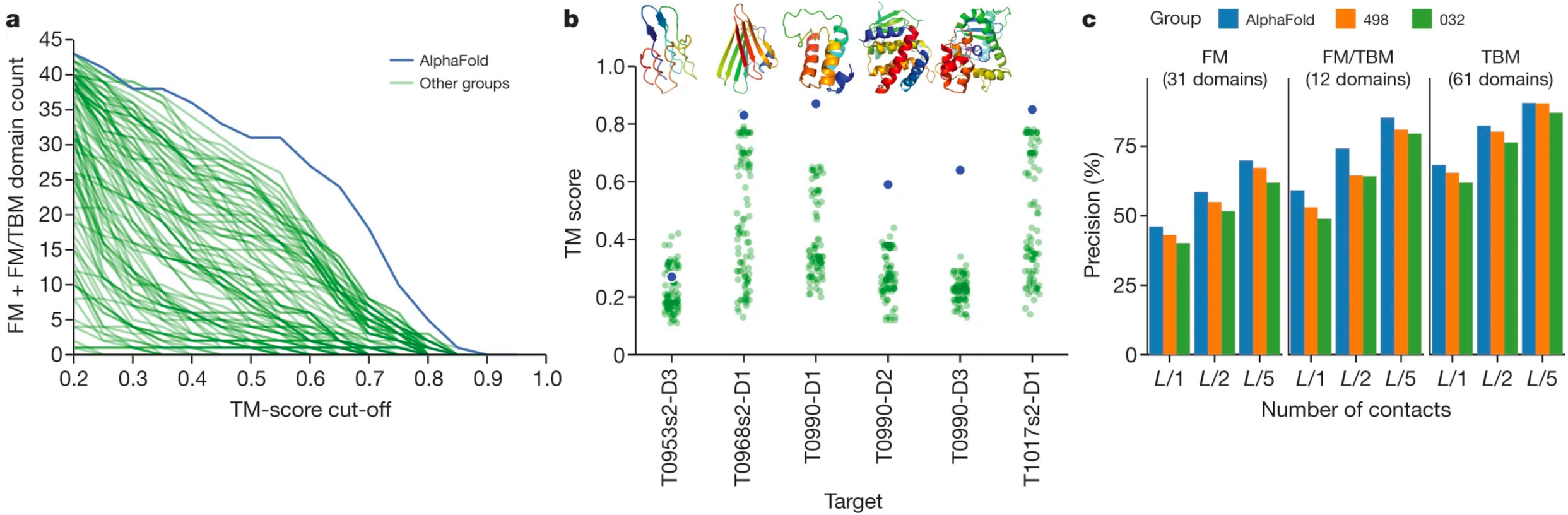
- Template-based modeling (TBM)
  - Availability of a *homologous* structure (i.e. the template)
- Free modeling (FM)
  - No similar sequence with a known structure
  - Methods typically rely on fragment assembly approaches and/or predict residue-residue correlations from sequence alone
- What data or information do methods use?
  - Assume that similar sequences lead to similar structures.
  - 2.4B protein sequences
  - In 2012: 80,000 structures in the PDB (4000 structural families, 1200 folds)

A *homologous* gene (or homolog) is a gene inherited in two species from a common ancestor.



https://www.cusabio.com/Transmembrane/G-protein-Coupled-Receptor-Targets-for-Drug-Therapy.html

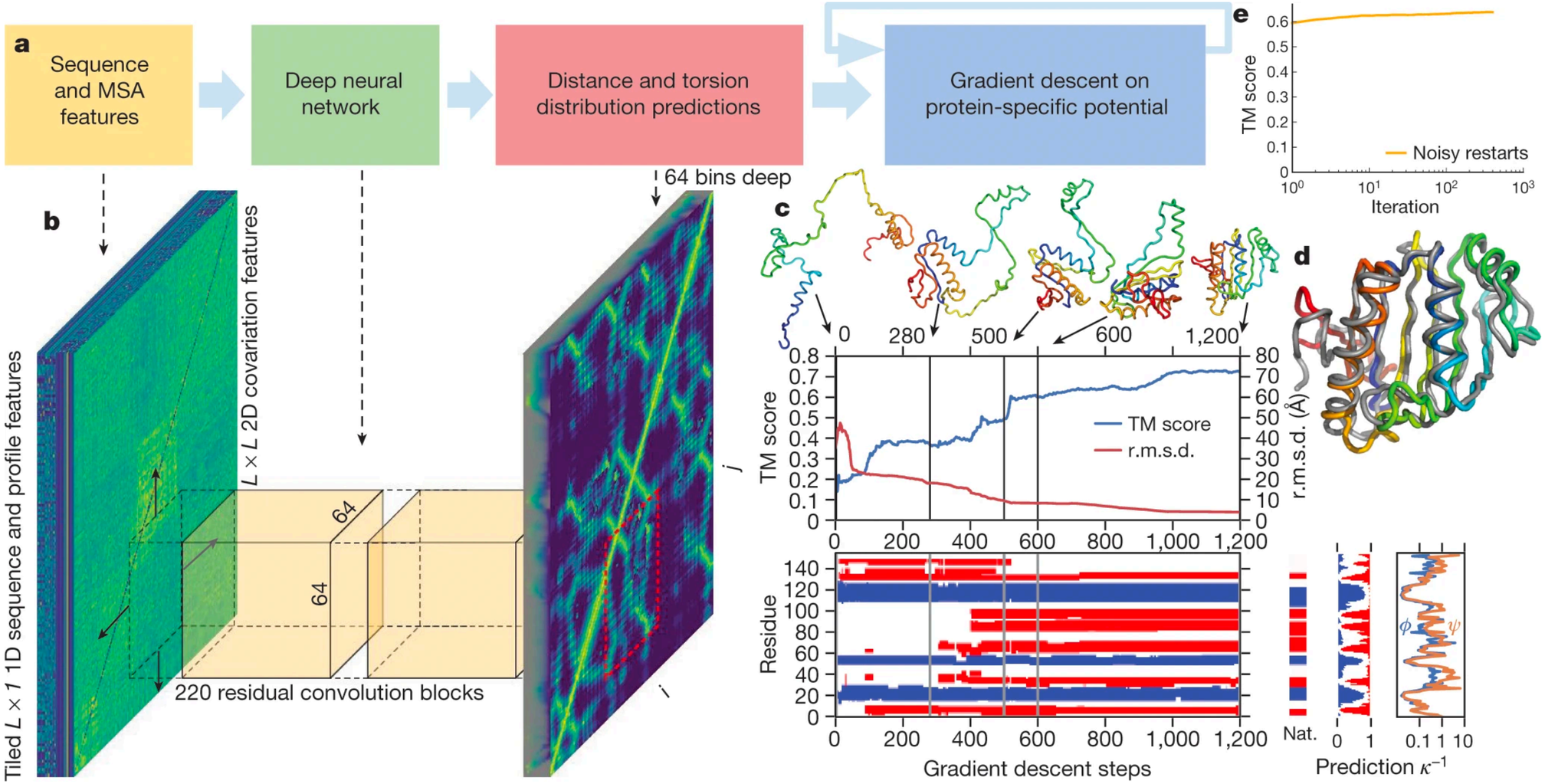
Alphafold1 at CASP13



a, Number of FM (FM + FM/TBM) domains predicted for a given TM-score threshold for AlphaFold and the other 97 groups. **b**, For the six new folds identified by the CASP13 assessors, the TM score of AlphaFold was compared with the other groups, together with the native structures. The structure of T1017s2-D1 is not available for publication. **c**, Precisions for long-range contact prediction in CASP13 for the most probable L, $L/2$ or $L/5$ contacts, where L is the length of the domain. The distance distributions used by AlphaFold in CASP13, thresholded to contact predictions, are compared with the submissions by the two best-ranked contact prediction methods in CASP13: 498 (RaptorX-Contact[26]) and 032 (TripletRes[32]) on 'all groups' targets, with updated domain definitions for T0953s2.

How does Alphafold1 work?

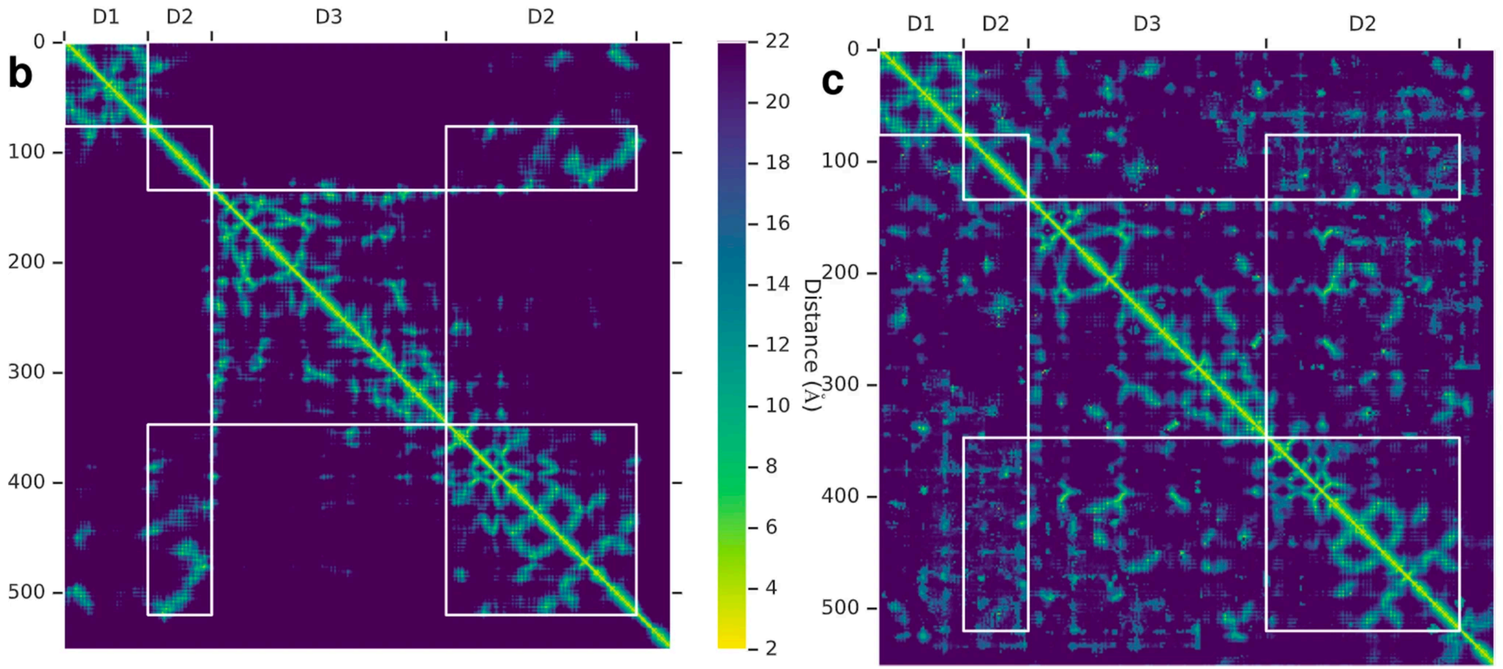
Fig. 2: The folding process illustrated for CASP13 target T0986s2.



What is a distogram?

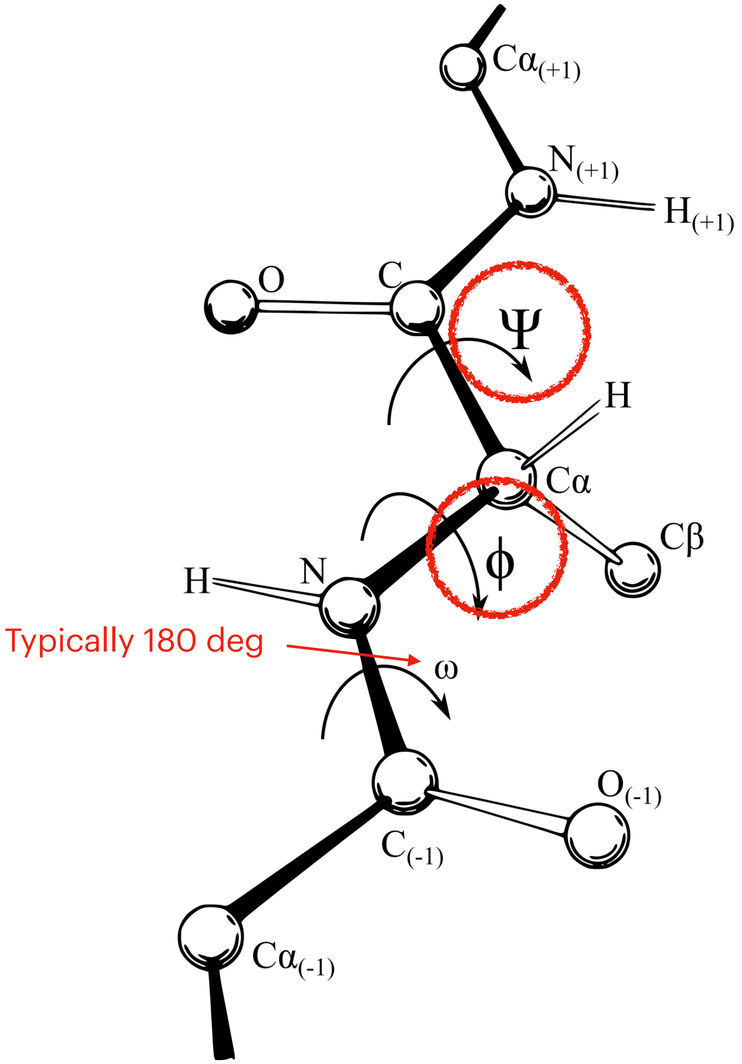
a

Contact precisions		L long			L/2 long			L/5 long		
Set	N	AF	498	032	AF	498	032	AF	498	032
FM	31	**46.1**	43.1	40.1	**58.5**	54.9	51.6	**69.9**	67.3	61.9
FM/TBM	12	**59.1**	53.0	48.9	**74.2**	64.5	64.2	**85.3**	81.0	79.6
TBM	61	**68.3**	65.5	61.9	**82.4**	80.3	76.4	**90.6**	90.5	87.1



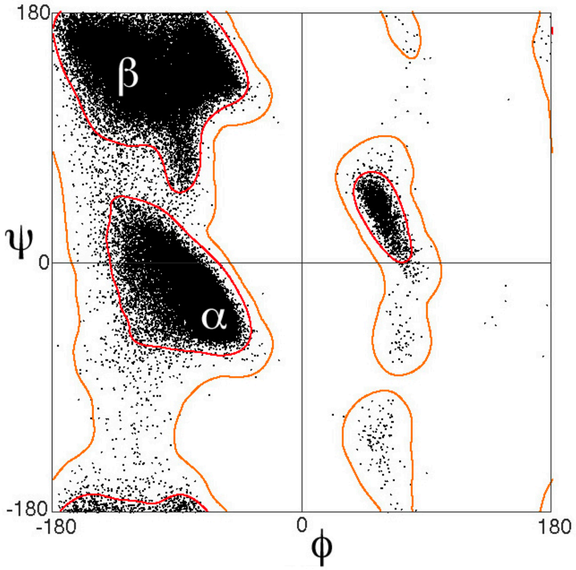
a, Precisions (as shown in Fig. 1c) for long-range contact prediction in CASP13 for the most probable L, $L/2$ or $L/5$ contacts, where L is the length of the domain. The distance distributions used by AlphaFold (AF) in CASP13, thresholded to contact predictions, are compared with submissions by the two best-ranked contact prediction methods in CASP13: 498 (RaptorX-Contact[26]) and 032 (TripletRes[32]), on 'all groups' targets, with updated domain definitions for T0953s2. **b**, **c**, True distances (**b**) and modes of the predicted distogram (**c**) for CASP13 target T0990. CASP divides this chain into three domains as shown (D3 is inserted in D2) for which there are 39, 36 and 42 HHblits alignments, respectively (from the CASP website).

Backbone torsions and Ramachandran plots

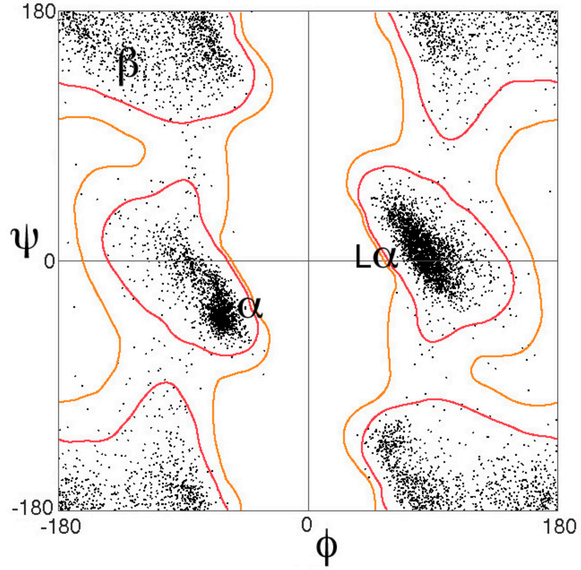


- A Ramachandran plot, or [phi, psi] plot, is a way to visualize backbone torsion statistics in protein structures
- Specifying the torsion angles gives a complete description of the backbone structure and geometry (and constrains side-chain locations)

General case

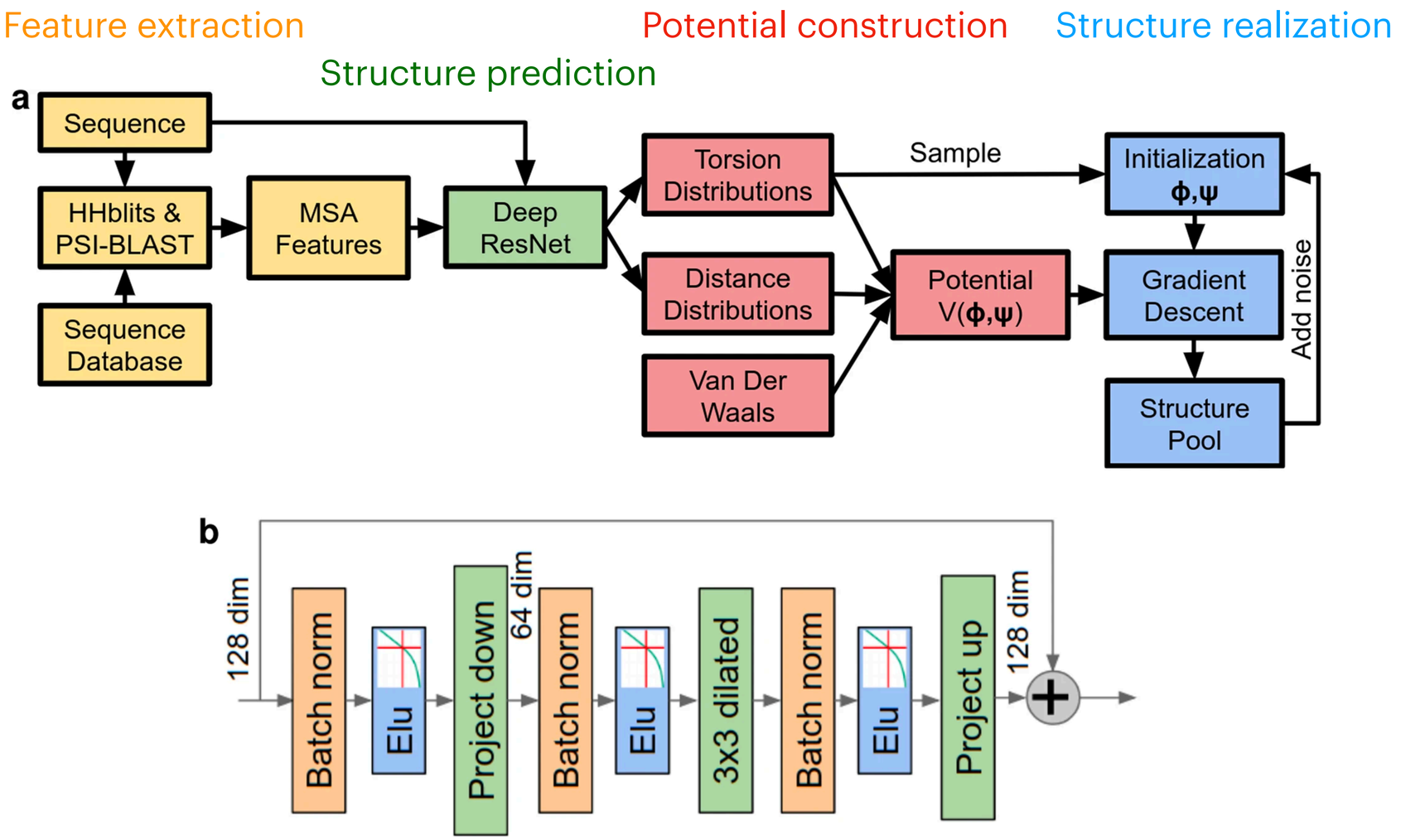


For glycines



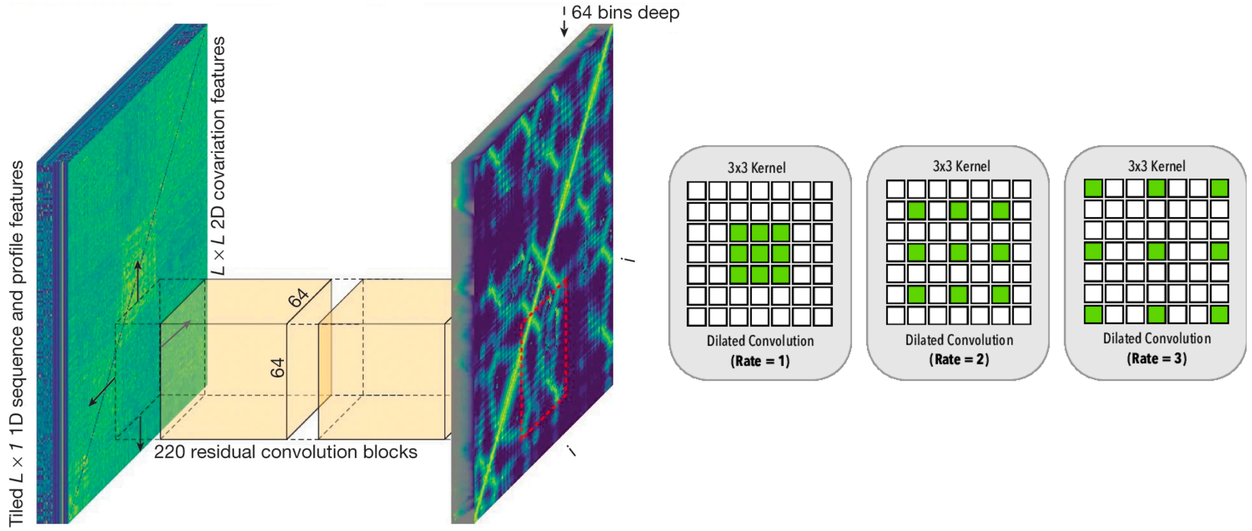
Data from Lovell 2003

The overall folding system



Structure prediction: residual 2D convolutional networks

- A 220 layer residual 2D convolutional network, 21M parameters
- Dilated convolutions for efficient long-range interaction
- Predicts pairwise *distance* distribution instead of contacts
- Mean and max pooling of the 2D activations in the second to last layer to get an 1-D output head:
  - 8-class secondary structure labels (computed by DSSP)
  - Relative accessible surface area
  - Backbone torsions as discretized marginal Ramachandran distributions for each residue
- At test time:
  - Tile crops to cover the whole histogram
  - Average across 4 models



More architecture and training details...

- 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.
- Data augmentation: MSA subsampling, coordinate noise, distogram cropping

How did they design the system?

Building a protein-specific potential

- To construct a differentiable potential, the predicted distance distribution is interpolated with a cubic spline
- A distance potential is created from the negative log likelihood of the distances, summed over all pairs of residues

$$V_{\text{distance}}(\mathbf{x}) = -\sum_{i,j,\ i\neq j} \log P(d_{ij} \mid \mathcal{S}, \text{MSA}(\mathcal{S}))$$

- The distance potential is adjusted based on a predicted reference distribution

$$V_{\text{distance}}(\mathbf{x}) = -\sum_{i,j,\ i\neq j} \log P(d_{ij} \mid \mathcal{S}, \text{MSA}(\mathcal{S})) - \log P(d_{ij} \mid \text{length}, \delta_{\alpha\beta})$$

- The final potential consists of three separate terms

$$V_{\text{total}}(\boldsymbol{\phi}, \boldsymbol{\psi}) = V_{\text{distance}}(G(\boldsymbol{\phi}, \boldsymbol{\psi})) + V_{\text{torsion}}(\boldsymbol{\phi}, \boldsymbol{\psi}) + V_{\texttt{score2\_smooth}}(G(\boldsymbol{\phi}, \boldsymbol{\psi}))$$

1D torsional distributions

Steric

Building a protein-specific potential

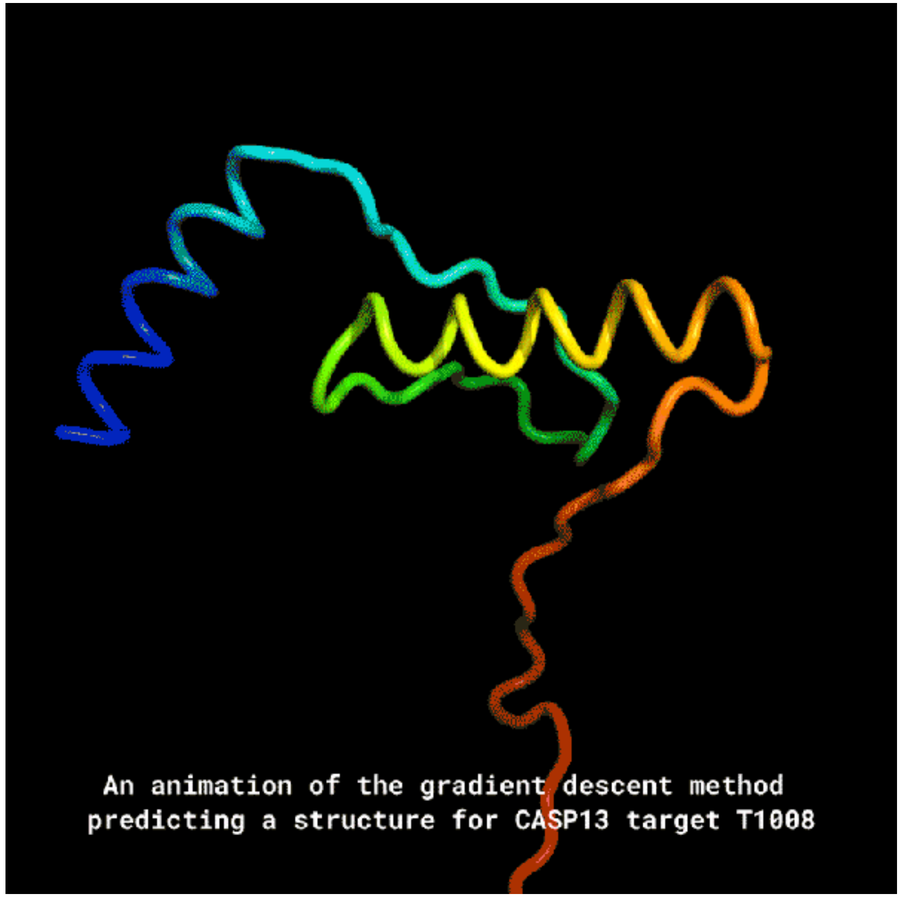
Key aspects

- Use a very large number of distributional predictions from a neural network
  - P_{ij}(distance) for all pairs
  - $P_i(\phi, \psi)$ for each residue
- Individual predictions are detailed, calibrated, and smooth
- Averaging the agreement scores over large numbers of distributions predictions (e.g. all distances) gives an accurate and smooth scoring function
- How is this related to the folding energy funnel?

Adapted from https://www.youtube.com/watch?v=uQ1uVbrIv-Q

Realizing a protein structure with gradient descent

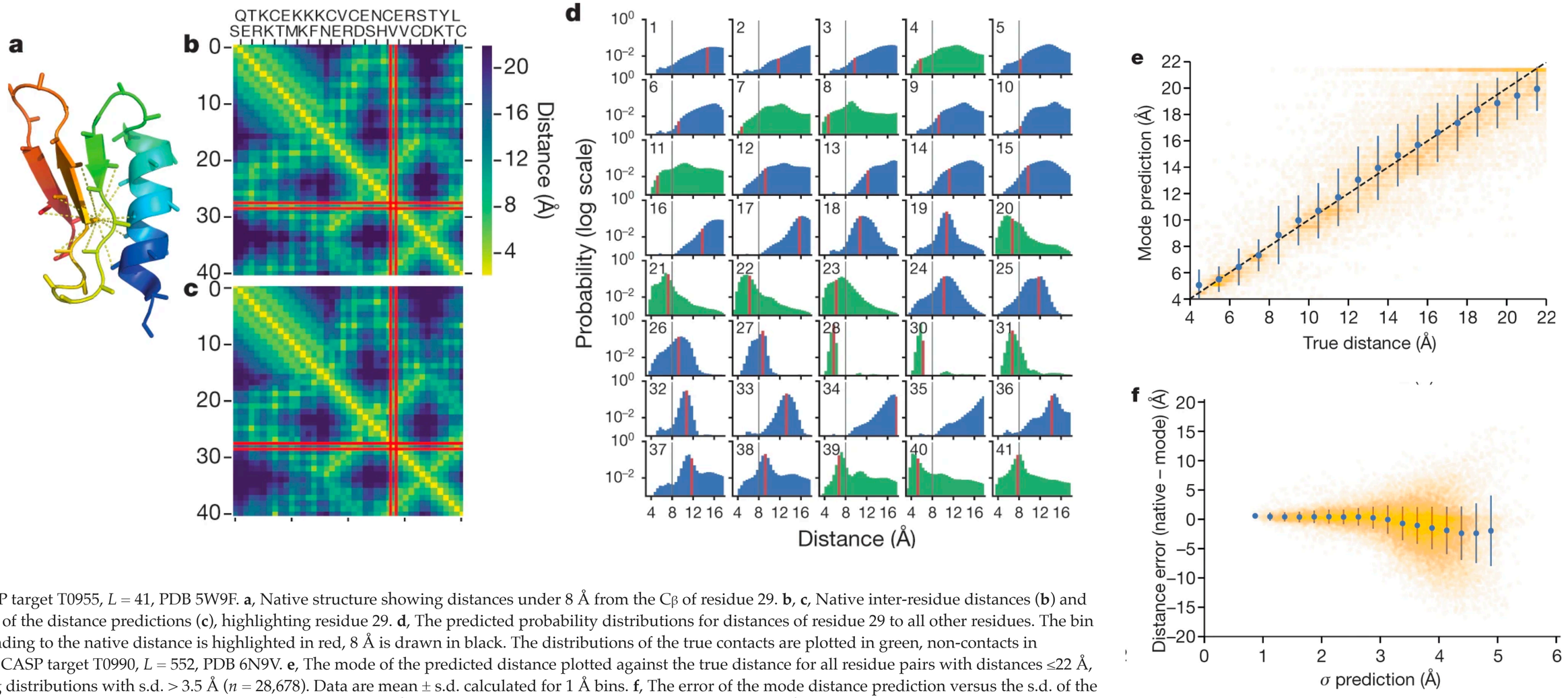
- The potential is differentiable with respect to backbone torsions angles
- Given an initial set of torsions, V_total is minimized by optimization with L-BFGS
- Repeat with multiple initializations and add noisy restarts (addition of 30 deg noise to the backbone torsions)
- Folding is performed on full chains, which avoids error-prone domain segmentation



Data: Input features, training and test sets

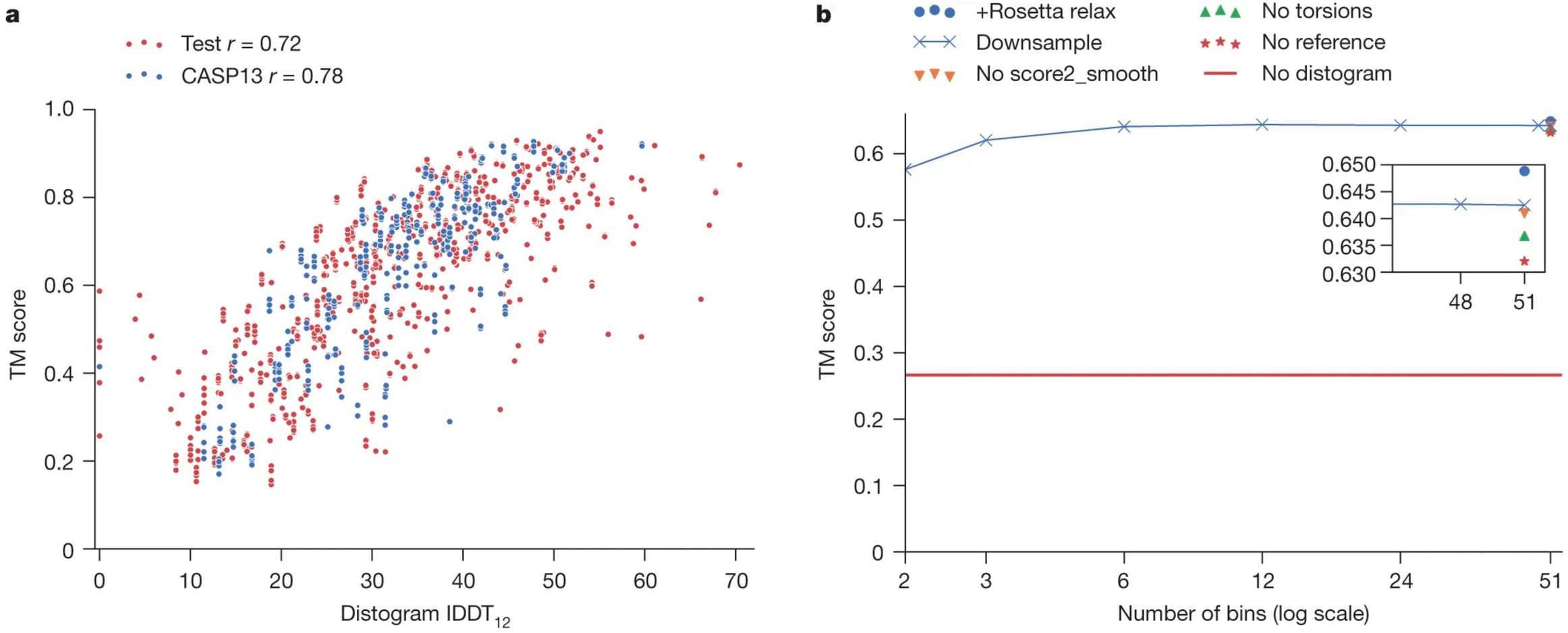
- PDB structures as of March 2018:
  - Extract non-redundant domains by using the CATH 35% sequence similarity cluster representatives
  - 31,247 domains, split into 29,427 train and 1,820 test
  - Keeping all domains from the same homologous superfamilies (H-level) in the same partition
  - CATH superfamilies of FM in CASP11 and CASP12 were excluded from training
- Sequences from Uniclust30 as of Oct 2017:
  - HHblits to generate the MSA and other covariation features
  - Additional features from PSI-BLAST

Predicted distance distributions compared with true distances



a–**d**, CASP target T0955, $L = 41$, PDB 5W9F. **a**, Native structure showing distances under 8 Å from the Cβ of residue 29. **b**, **c**, Native inter-residue distances (**b**) and the mode of the distance predictions (**c**), highlighting residue 29. **d**, The predicted probability distributions for distances of residue 29 to all other residues. The bin corresponding to the native distance is highlighted in red, 8 Å is drawn in black. The distributions of the true contacts are plotted in green, non-contacts in blue. **e**, **f**, CASP target T0990, $L = 552$, PDB 6N9V. **e**, The mode of the predicted distance plotted against the true distance for all residue pairs with distances ≤22 Å, excluding distributions with s.d. > 3.5 Å ($n = 28{,}678$). Data are mean ± s.d. calculated for 1 Å bins. **f**, The error of the mode distance prediction versus the s.d. of the distance distributions, excluding pairs with native distances >22 Å ($n = 61{,}872$). Data are mean ± s.d. are shown for 0.25 Å bins. The true distance matrix and distogram for T0990 are shown in Extended Data Fig. 2b, c.

Method ablations: TM scores vs accuracy of the distogram, components of the potential



a, TM score versus distogram lDDT12 with Pearson's correlation coefficients, for both CASP13 (n = 500: 5 decoys for all domains, excluding T0999) and test (n = 377) datasets. **b**, Average TM score over the test set (n = 377) versus the number of histogram bins used when downsampling the distogram, compared with removing different components of the potential, or adding Rosetta relaxation.