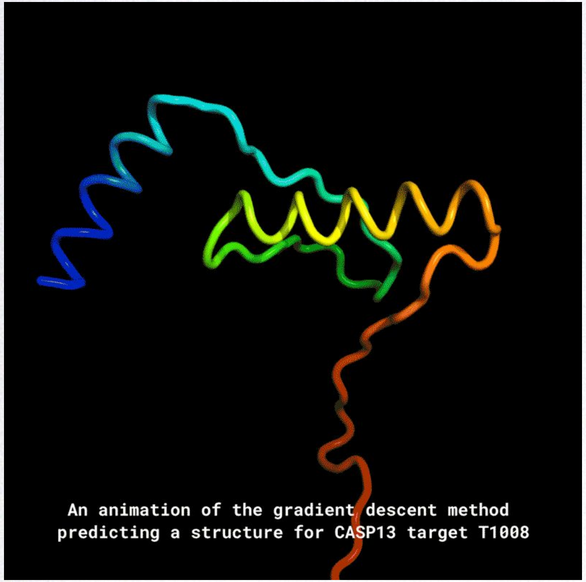
De novo protein folding using statistical potentials from deep learning

R.Evans, J.Jumper, J.Kirkpatrick, L.Sifre, T.F.G.Green, C.Qin, A.Zidek, A.Nelson, A.Bridgland, H.Penedones, S.Petersen, K.Simonyan, D.T.Jones [UCL] , K.Kavukcuoglu, D.Hassabis, A.W.Senior

DeepMind

Group 043 / A7D / AlphaFold

Protein folding at DeepMind



DeepMind's core mission is to develop advanced artificial intelligence and use it to solve important problems

Protein folding allows us to work on a central problem in biology that has clear goals and rich data

Major progress in protein folding should allow rapid advances in understanding protein interactions

Free modeling

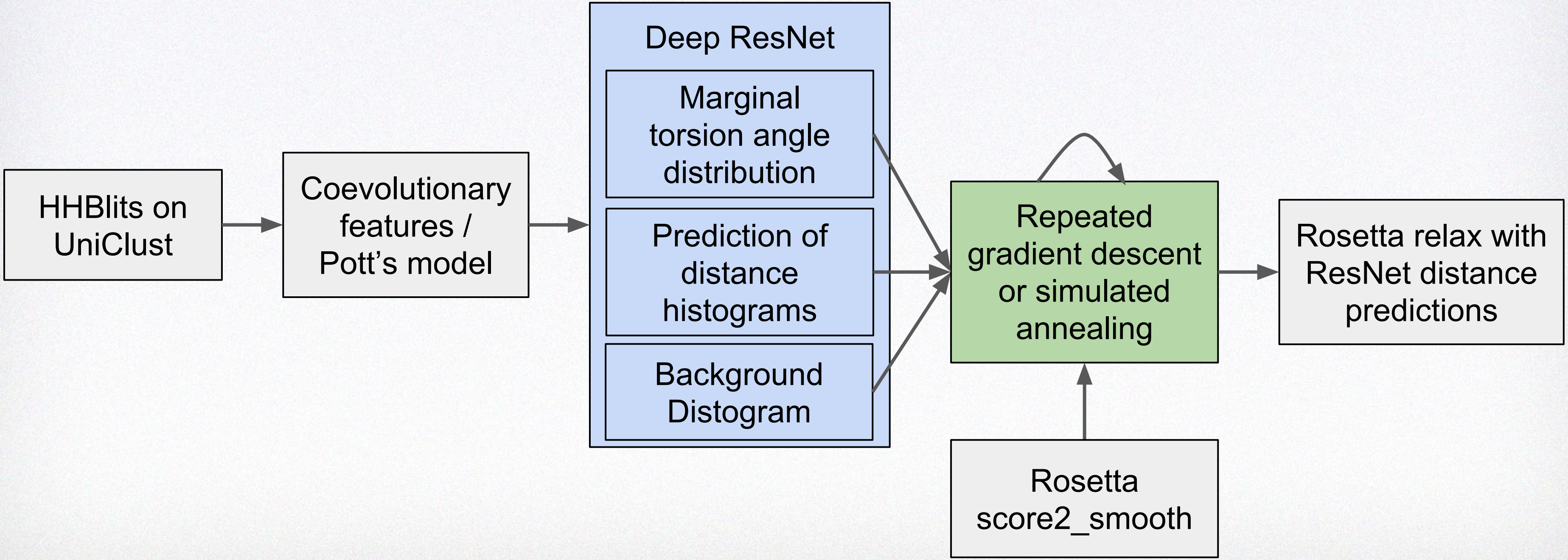
- Our system is exclusively free-modelling system and does not use templates*
    - (*) except we adjusted a pair of domain segmentations by hand on strong templates
    - We ran TBM targets identically to FM targets
    - No use of stoichiometry information or server models

- Very minor human intervention
    - Two domain segmentations (especially on T0999)
    - We used human judgement when deciding whether two models were too similar to include both
    - We reordered predictions in one case because the second prediction had strictly more structure

- We use standard PDB (training) and UniClust (MSA generation) databases

System overview

Details of the distance and torsion prediction network will be in Andrew Senior's talk in the AI session



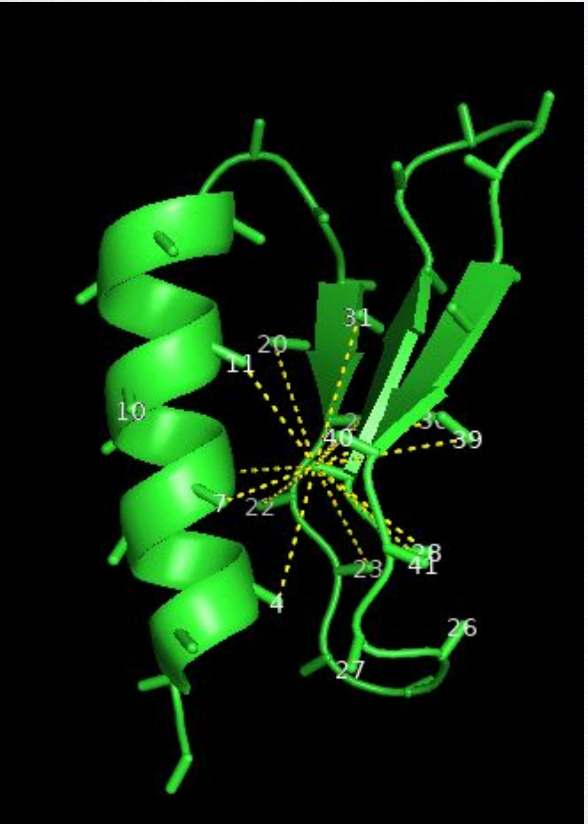
Key aspects of our system

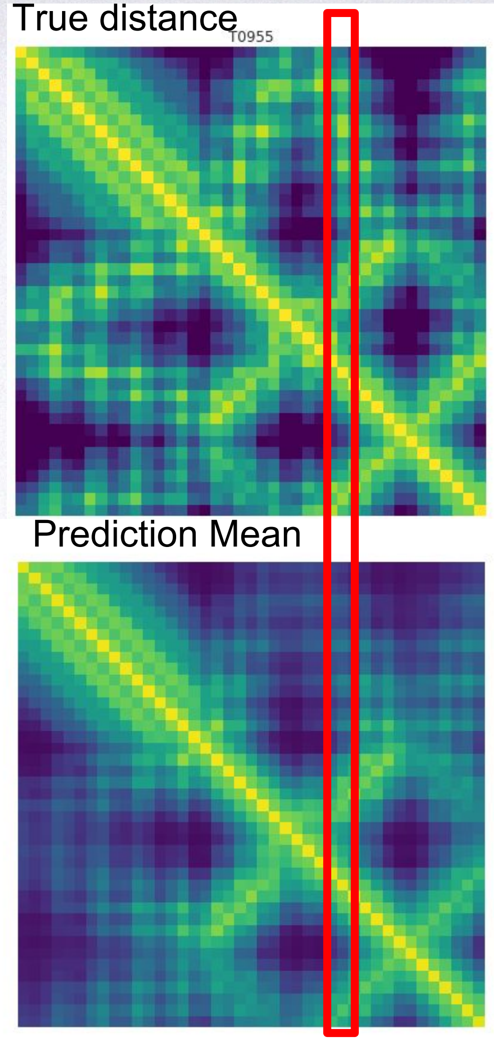
- Use a very large number of distributional predictions from a neural network
  - P_{ij}(distance) for all pairs
  - $P_i(\phi, \psi)$ for each residue

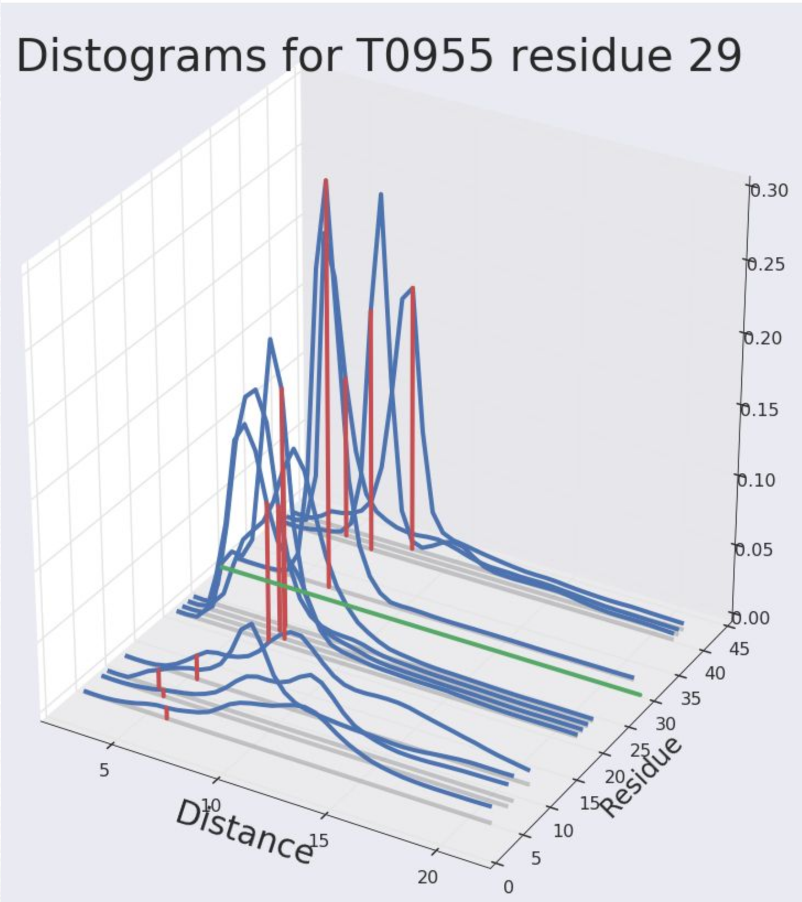
- Individual predictions are *detailed*, *calibrated*, and *smooth*

- Averaging the agreement scores over large numbers of distributional predictions (e.g. all distances) gives an accurate and smooth scoring function

Deep distance distribution Network (D^3N)







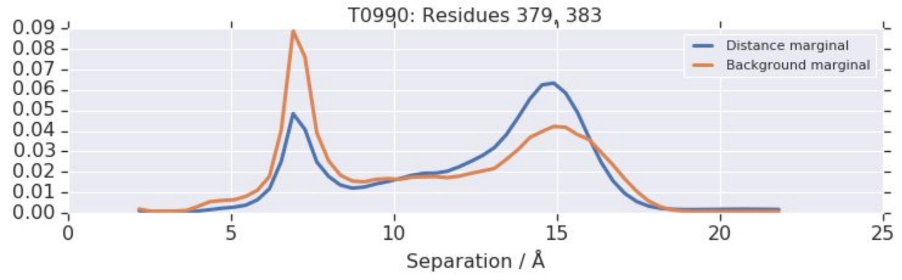
Using deep learning to construct a reference state

The outputs of the distance prediction network are analogous to raw counts in a tabular knowledge-based potential

To obtain a potential, we must apply a reference state correction

We train a neural network to produce reference state distance distributions

- Only input features are i, j, N, and is_glycine
- No other sequence or MSA information



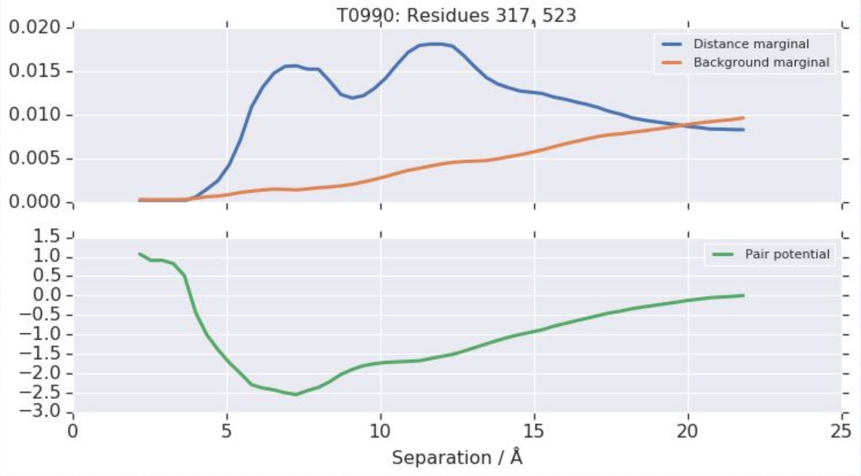
Potential construction

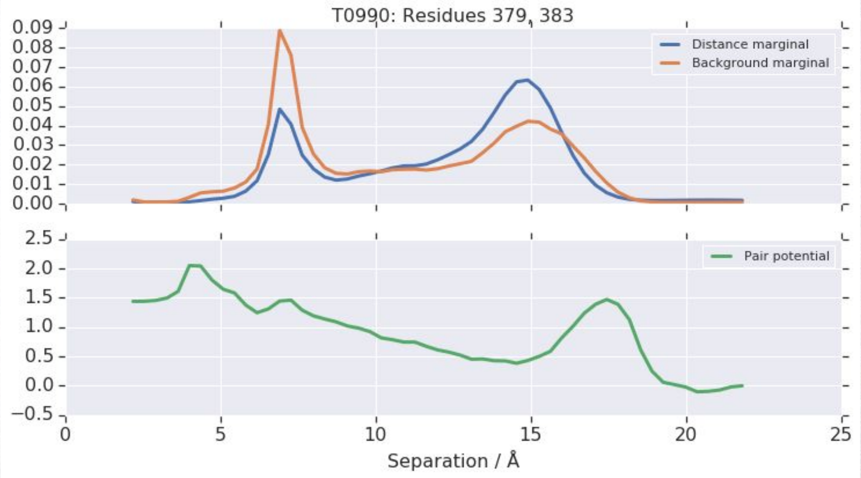
The log ratio tends to be more convex than the distance predictions

$$V_{ij}(d_{ij}) = -\log\left(\frac{\Pr(d_{ij}|i,j,N,\text{sequence},\text{co-evolution})}{\Pr(d_{ij}|i,j,N,\text{is\_glycine})}\right)$$

Potential is score2 + distance potential

Alternatively, can train a scoring network to predict GDT





Optimizing the statistical potential

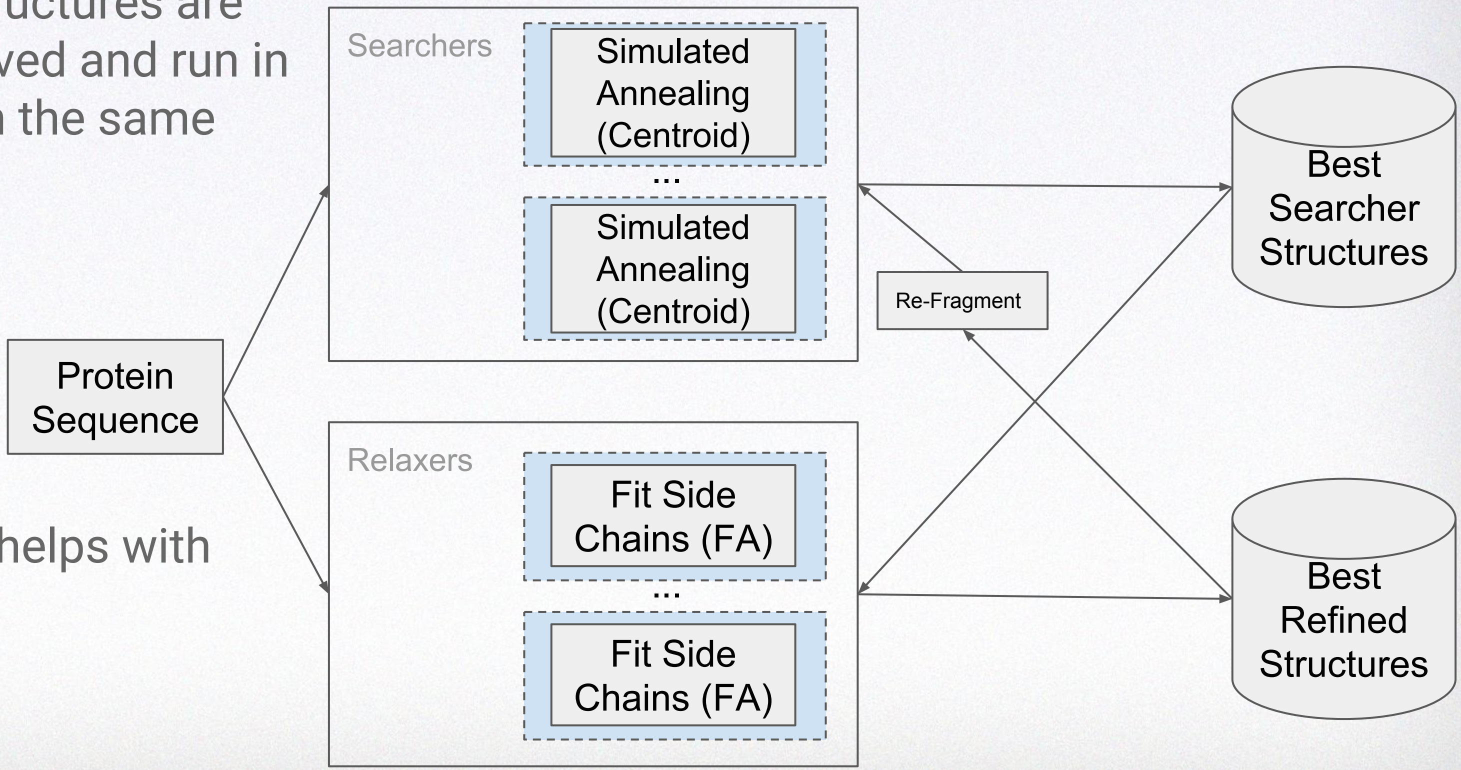
Two methods

- Simulated annealing with fragment insertion
  - Domain segmented
  - Generative model of protein fragments
  - Higher diversity
- Repeated gradient descent
  - Full chains
  - Lower diversity

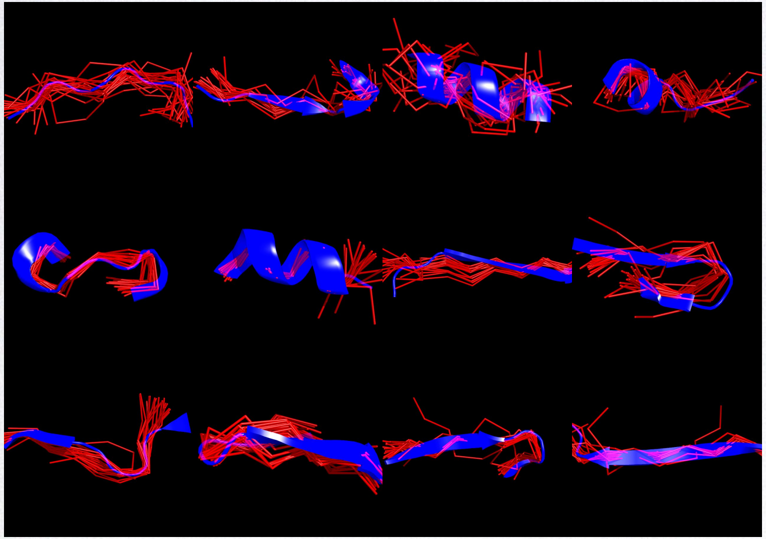
Simulated annealing with fragment insertion

Lowest energy structures are periodically removed and run in Rosetta relax with the same pairwise energy.

Refragmentation helps with accuracy



Generative model of fragments

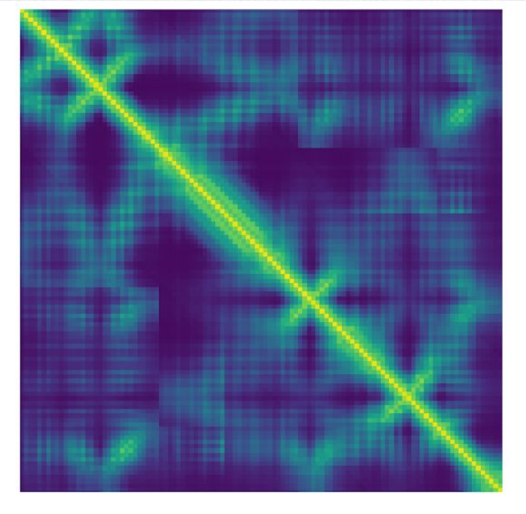


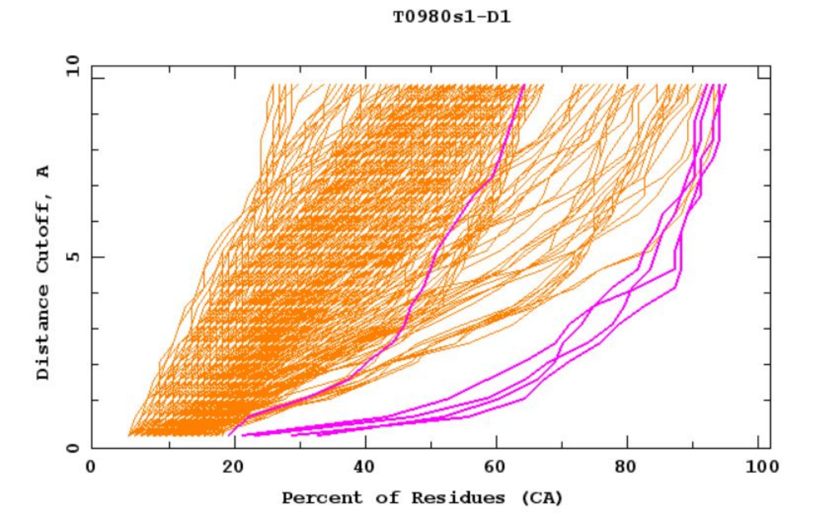
End-to-end trained model of 32-residue fragments

Based on VAE (variational auto-encoder) with recurrent "canvas"

Cut into 9-residue fragments for fragment insertion

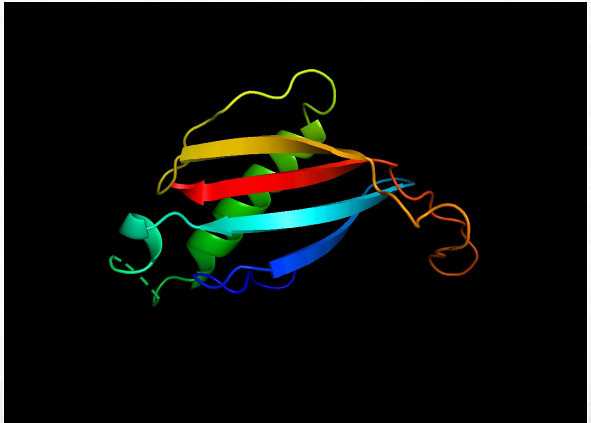
T0980s1-D1





Fragment insertion model

Experimental model



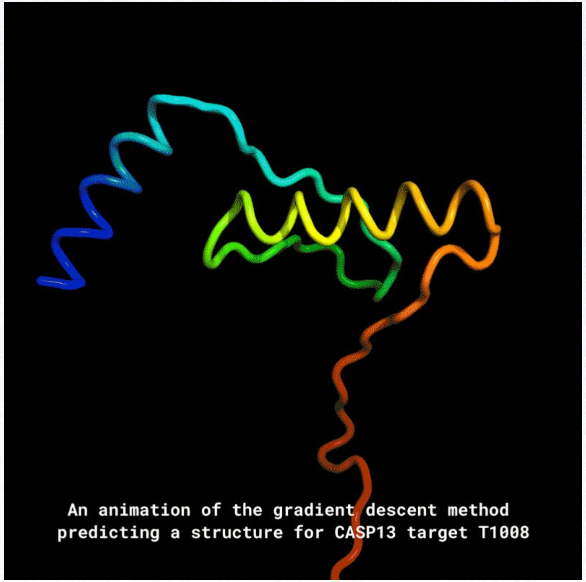
Repeated gradient descent

Repeated gradient descent

With a smooth Rama, the potential minimizes using repeated gradient descent (initialize from corruptions of best results)

Instead of using fragments, we will use a Rama energy term smoothed to a single von Mises

No domain segmentation (except T0999)

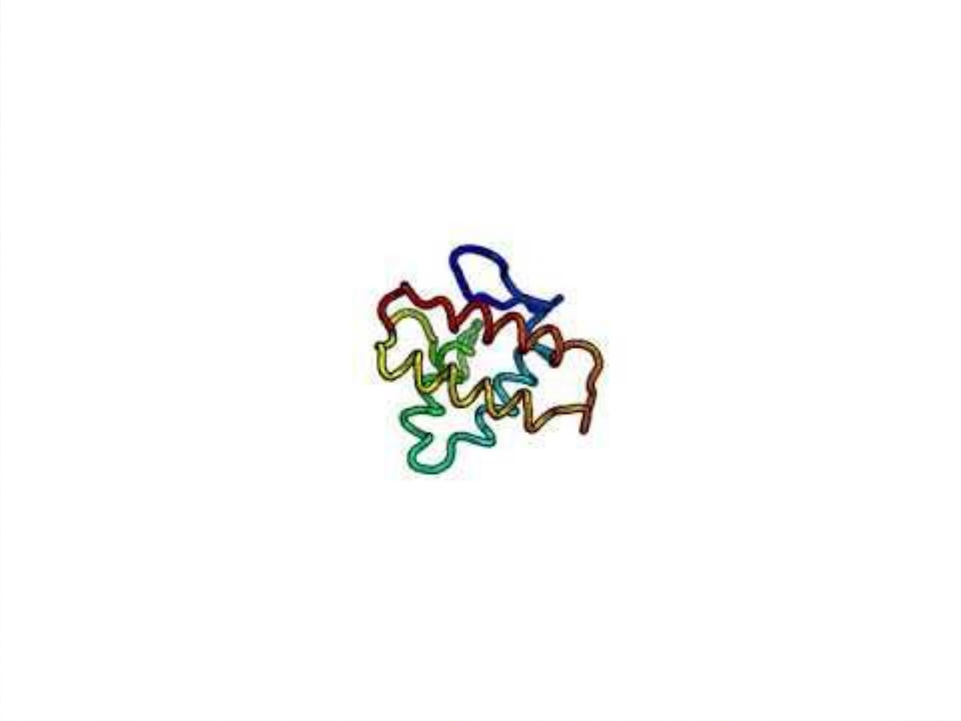


Repeated gradient descent animation

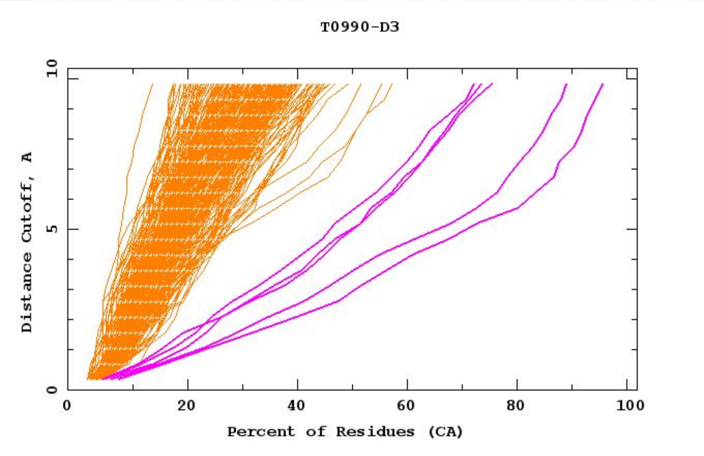
T0869-D1 from Casp12

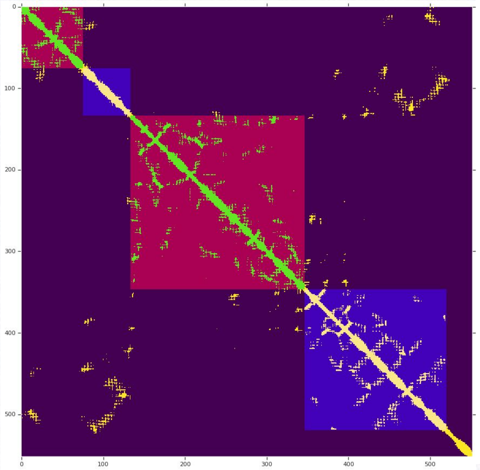
Approximately real time

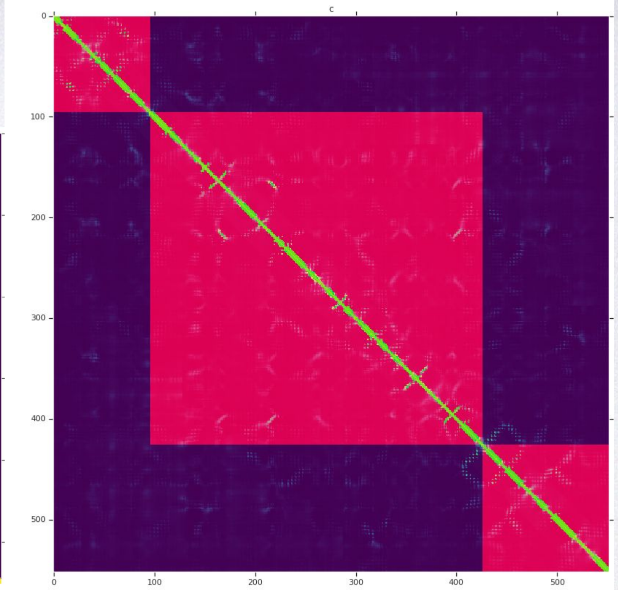
Cherry-picked

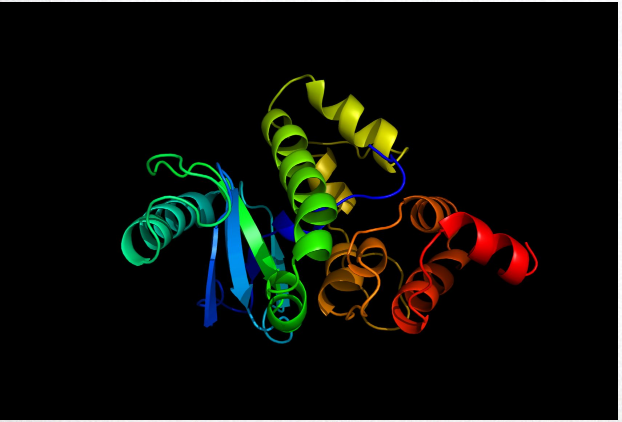


T0990-D3


T0990-D3
Distance Cutoff, A
Percent of Residues (CA)
0
5
10
0
20
40
60
80
100


0
100
200
300
400
500

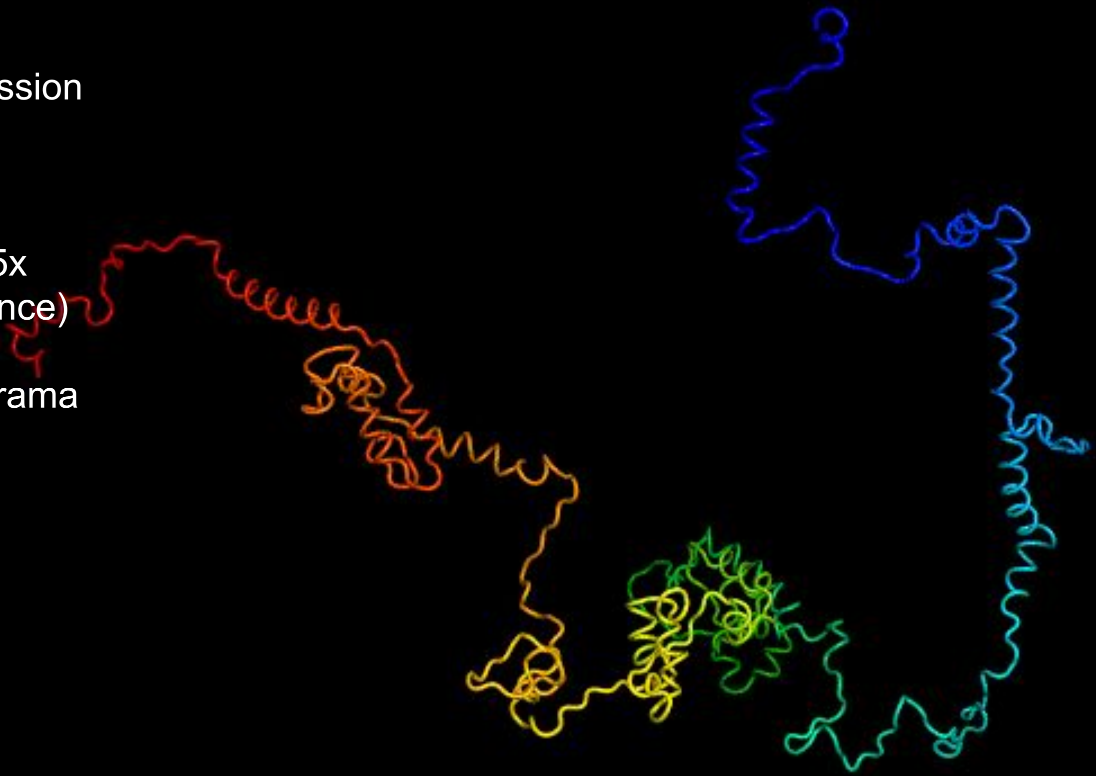

c
0
100
200
300
400
500



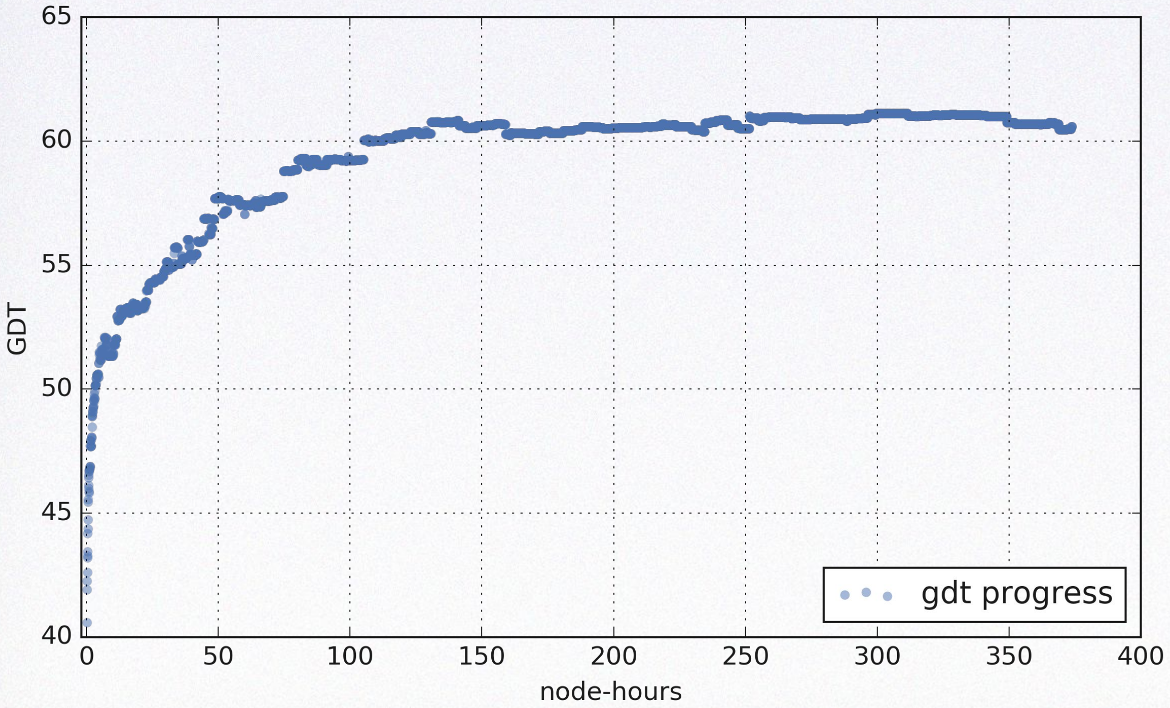
Not our submission model

639 frames (subsampled 5x original sequence)

Initialize from rama predictions



Accuracy vs computational cost



(for a subset of targets, on CPU nodes)

Repeated gradient descent

Using simple vdW instead of score2

Highly parallelizable

What went wrong

- T0999 broke our pipeline
- Repeated gradient descent was found to work midway through CASP
- False confidence (low diversity even when wrong)
- Currently cannot produce a cis proline

Conclusions

- Distance distributions are rich scoring distributions, and calibration of predictions makes combinations efficient
- Reference state correction matters
- Sampling is declining in importance relative to the initial network predictions
- Full chain folding can reduce errors

What’s next

- Still focused on fundamental improvements to structure prediction accuracy
- Will publish detailed methods in a paper
- Open to collaboration on applications
- No current plans to open source or put up a public server

Team

Team lead:

Andrew Senior

Contributors:

Richard Evans, **John Jumper**, **James Kirkpatrick**, Laurent Sifre, **Tim Green**, Chongli Qin**, Augustin Zidek**, Sandy Nelson, **Alex Bridgland**, Hugo Penedones, Stig Petersen, Karen Simonyan, Koray Kavukcuoglu, Demis Hassabis

David Jones consulted

Bold indicates present at CASP13