Deep Learning distance, torsion and score predictions for de novo structure modelling

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

Deep learning

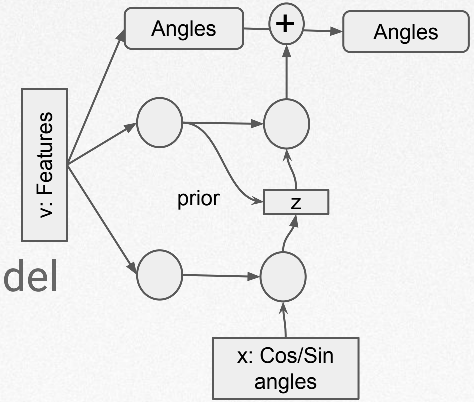
- Neural networks are function approximators trained to optimize an objective
  - Parameters or weights trained by gradient descent
- Hugely successful in recent years, has revolutionized many domains
  - Speech recognition
  - Speech synthesis
  - Machine translation
  - Image recognition / segmentation
  - Agents
    - Playing games: Go, Chess, Atari
    - self-driving cars
- Capable of modelling complex data
  - Long range, subtle patterns, with redundancy, needing generalization
  - Structure of the network gives *inductive bias* to certain kinds of modelling

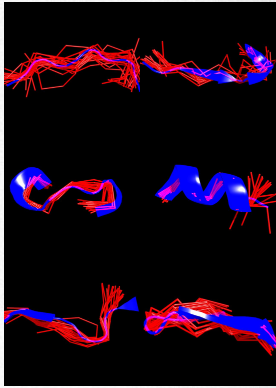
Why machine learning for protein structure modelling

- A complex problem
- Hard to model all the complex interactions in a long molecule
  - Local and long-range dependencies
- There is data thanks to experimental structure techniques
  - 146,000 PDB entries
  - highly redundant, not the scale of many problems
    - 10s of millions of utterances for speech
    - 15 million labelled images in ImageNet
- CASP assessment provides a benchmark with well-defined goals

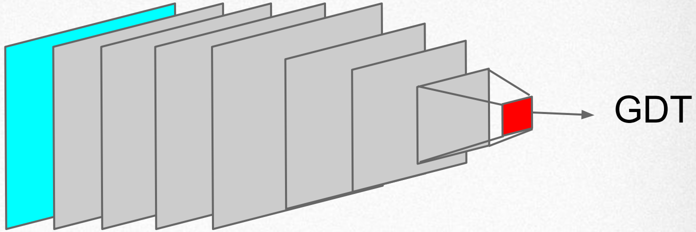
Where have we applied machine learning in CASP13?

- Torsion prediction
  - **End-to-end** training:
    - {Sequence, MSA features} → torsions
  - As a generative model from which we can draw samples
  - Based on DRAW*, a Variational Auto Encoder model
  - Used for fragment generation





- Scoring
  - Score a decoy by predicting the GDT distribution
    - {Distance map, contact prediction, MSA features} → score

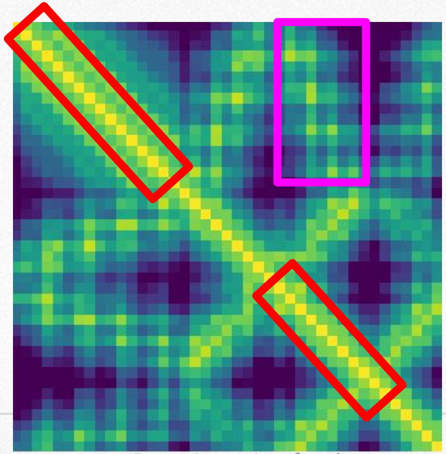


- Residue distance prediction

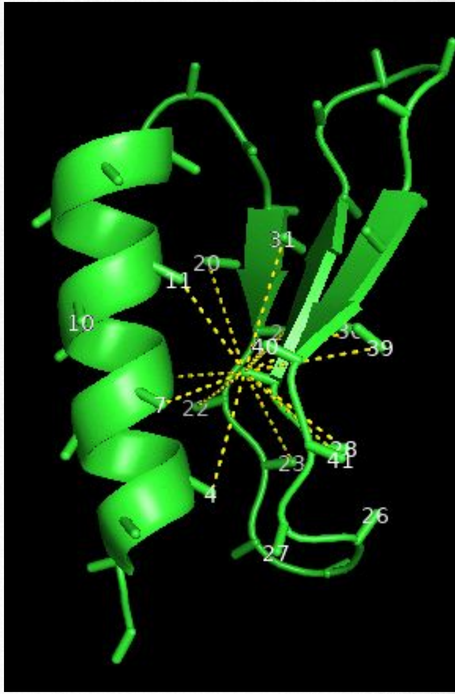
*DRAW: A Recurrent Neural Network For Image Generation K.Gregor, I. Danihelka, A. Graves, D. J. Rezende, D. Wierstra arxiv.org/abs/1502.04623

Predicting inter-residue distances

- Much focus in recent years on predicting residue contacts
  - Contacts provide a strong constraint on non-sequence-local structure
  - DCA, CCMPred, MetaPSICov, Raptor-X, ...
  - Explosion in sequencing expands multiple sequence alignments and coevolution data
- Previous work has predicted distances, or contacts with various thresholds
- Distances are predictable not just from coevolutionary contact information
  - Local propagation of distance constraints
  - Secondary structure interactions

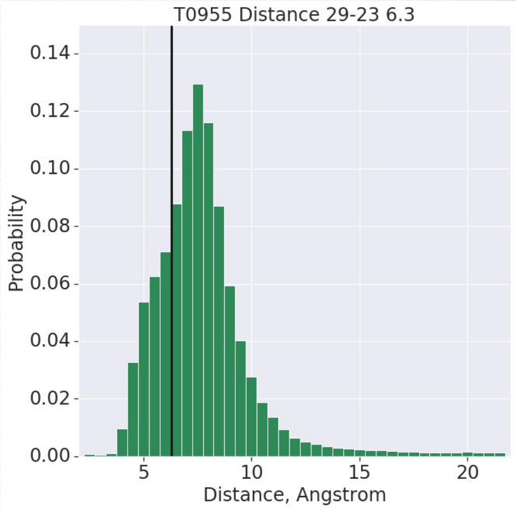


T0955 Native

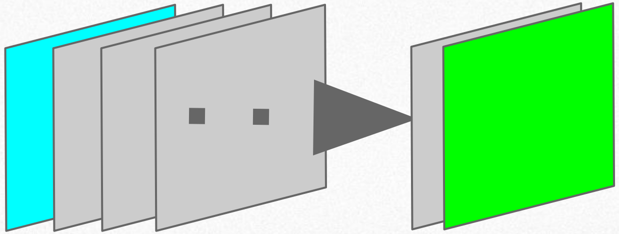


Deep distance distribution network

- Train a large 2-dimensional dilated residual convolutional network to predict CB atom distances
  - For each i, j pair, output is a softmax probability distribution
  - Well-calibrated
  - Train to cross-entropy objective
  - 40 0.5Å bins from 2–22Å (later 64 bins)
  - Distance histograms → “distograms”
  - We predict the highly-correlated distance *marginals,* not a joint distribution
- 2-dimensional throughout



N x N
Input features



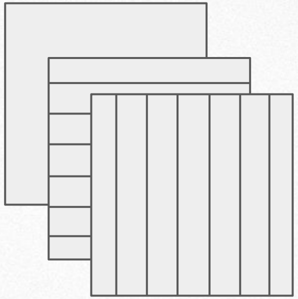
N x N
Distance predictions

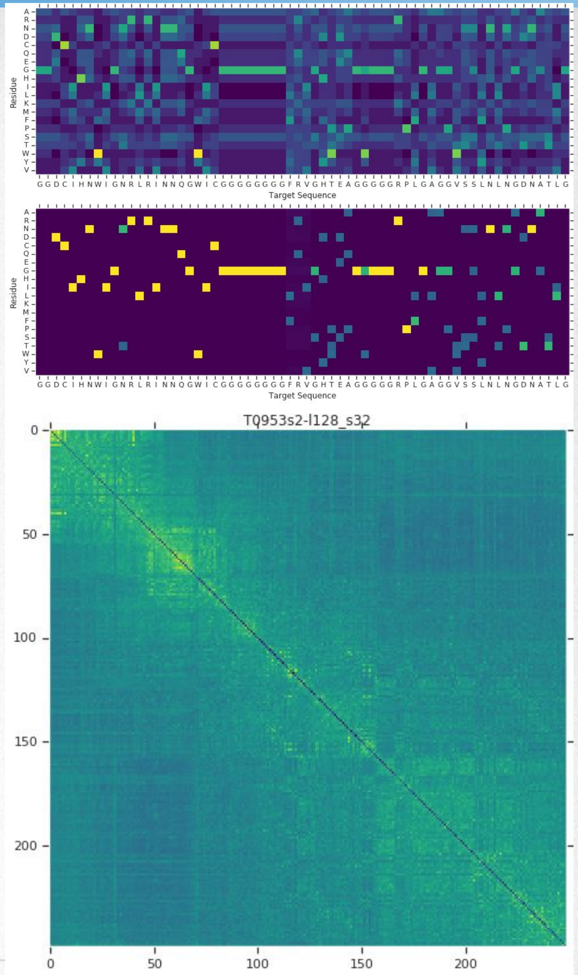
Residual network blocks with NxN representations

Data

- PDB 2018-03-15 / Uniclust30 2017-10
- Train on 29,400 CATH (2018-03-16) s_35 cluster representatives
- MSA features e.g.
  - HHBlits and PSIBLAST profiles
  - 2D features from Potts model fit in TensorFlow
    - Frobenius norm L x L x 1
    - **Raw parameters** L x L x 22 x 22
  - No Mutual Information

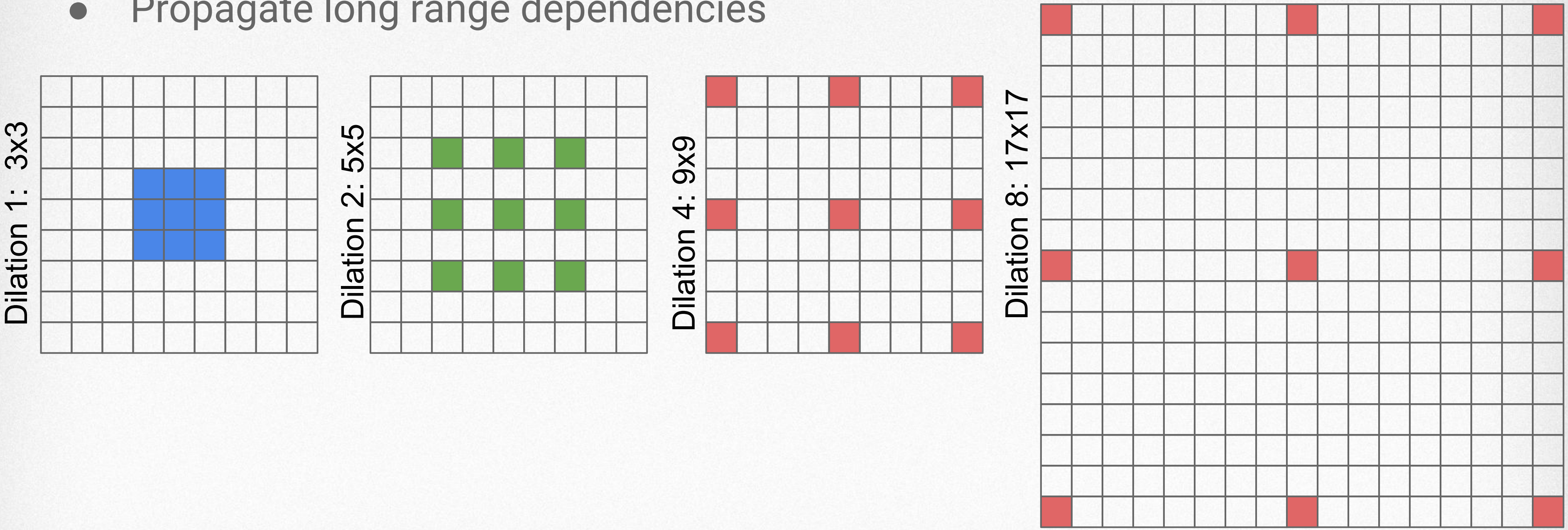
Repeat 1D features, tiling in x and y then concatenate with 2D features





Dilated convolutions

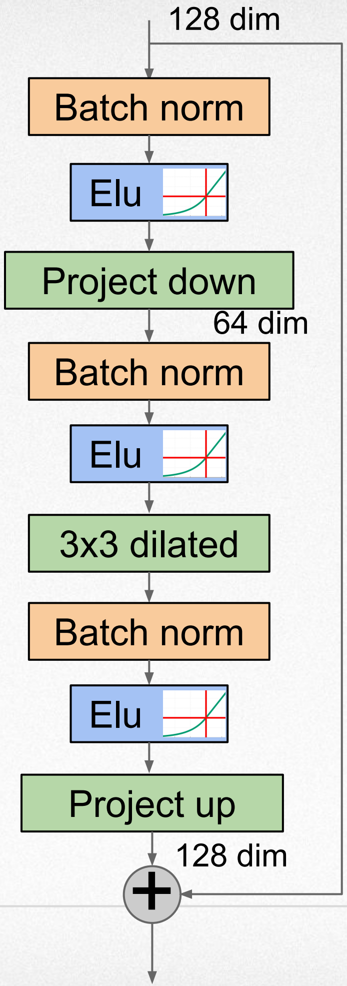
- Dilated convolutions skip pixels
    - Allow wide receptive fields with few parameters and low computation
- Propagate long range dependencies



Residual network

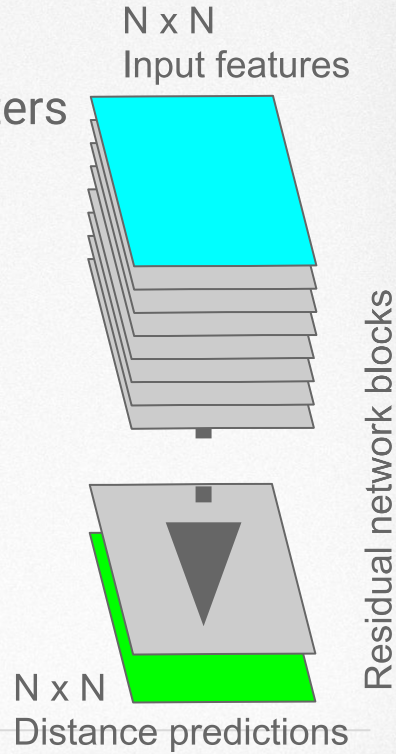
1 residual block

Modifies a 64x64x128 representation from the previous block



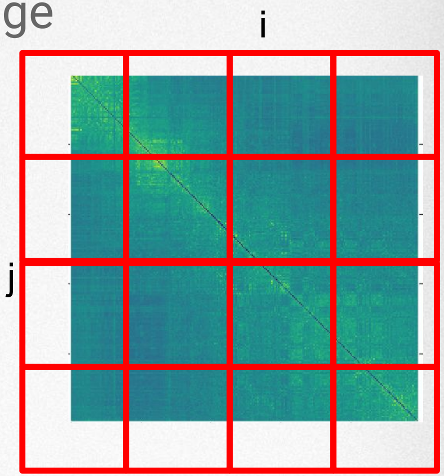
Repeat **220** times, cycling through dilations 1, 2, 4, 8

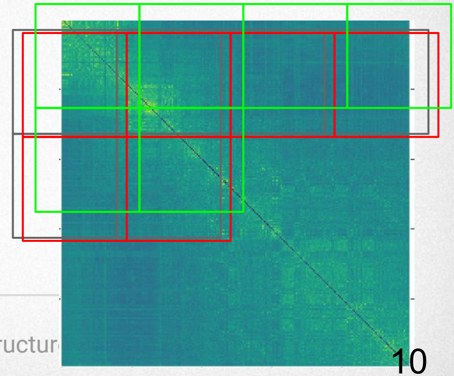
21 million parameters



Cropping

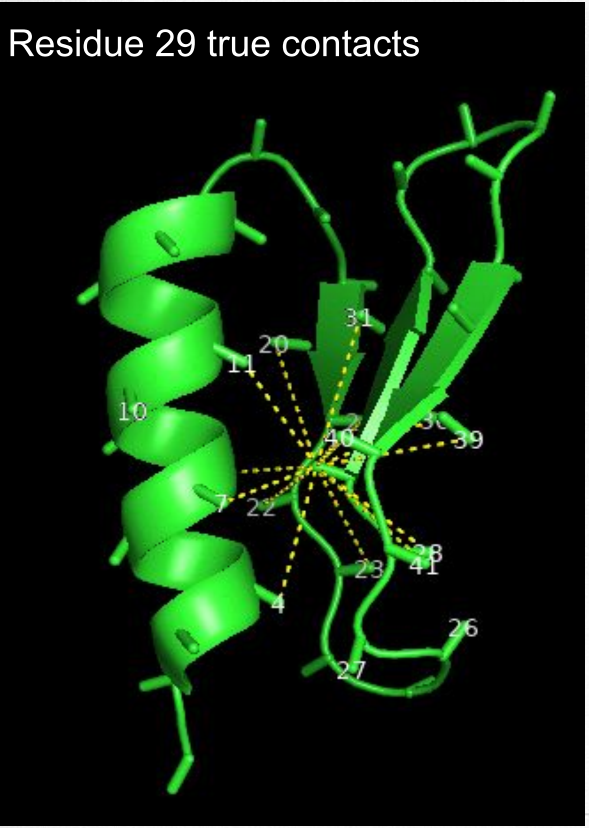
- Handling arbitrary protein length L leads to $O(L^2)$ memory usage
  - Consistent size helps distributed training
- Train on all 64x64 crops from proteins
  - Random offset
  - Including up to 32 residues off-edge
- For a crop (i, i+63)x(j, j+63)
  - Crop corresponding 2D input features
  - Tile corresponding (i, i+63) and (j, j+63) 1D parameters
  - Still allows modelling long range correlations from i to j
- Helps avoid overfitting
  - Data augmentation
  - Each protein leads to many different training examples
- Ensembling:
  - At test time weighted average across alternative offsets
  - Also average across 4 slightly different models

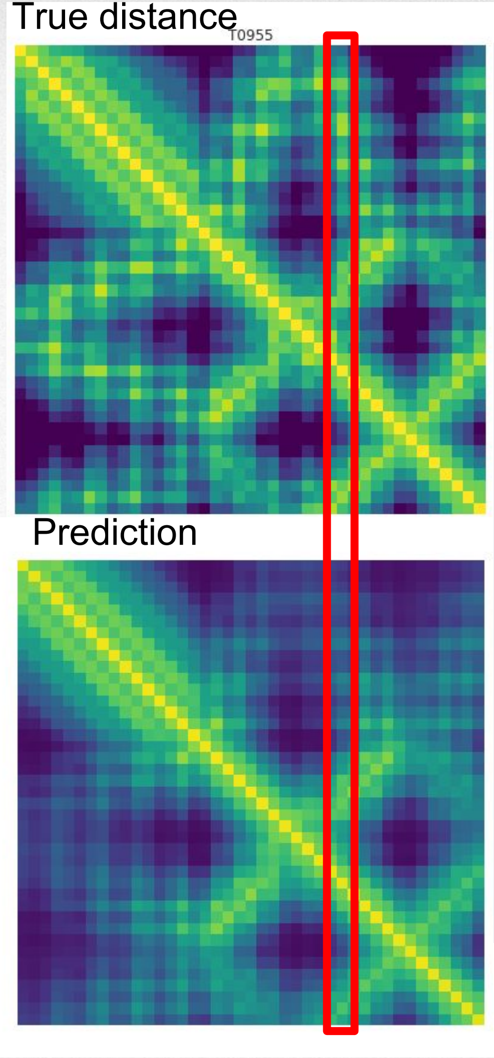




T0955 example

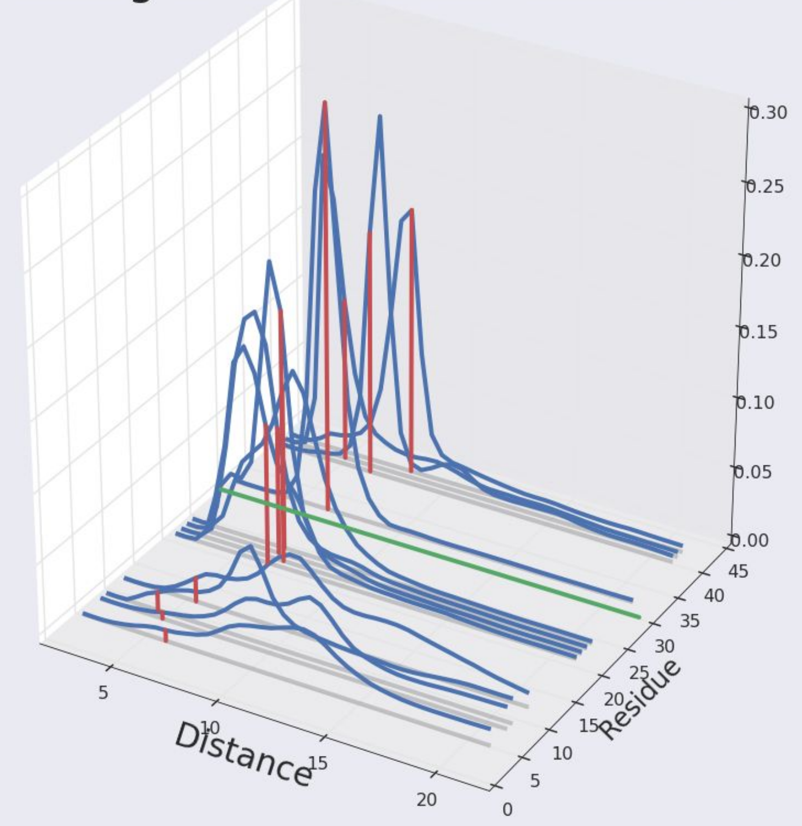
TBM/FM 88.4GDT





True contacts'

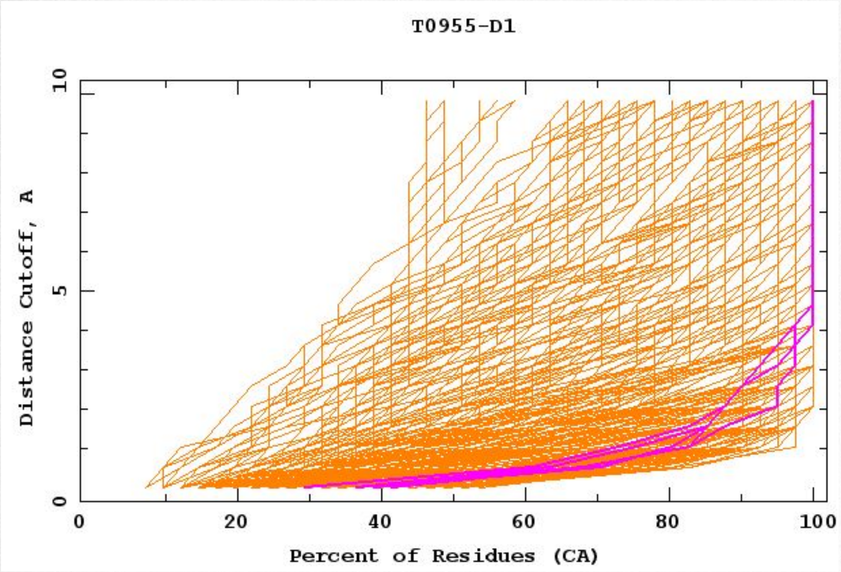
Distograms for T0955 residue 29

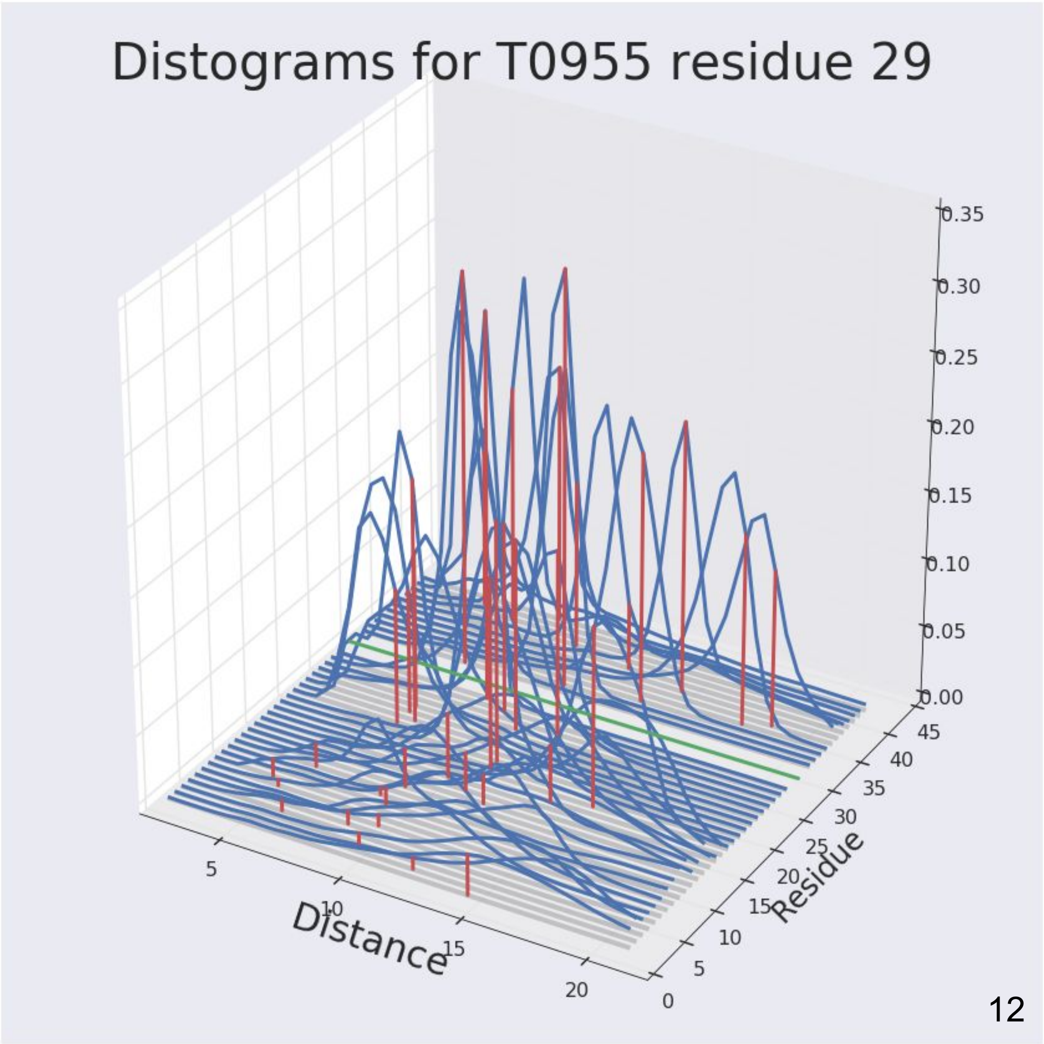


T0955

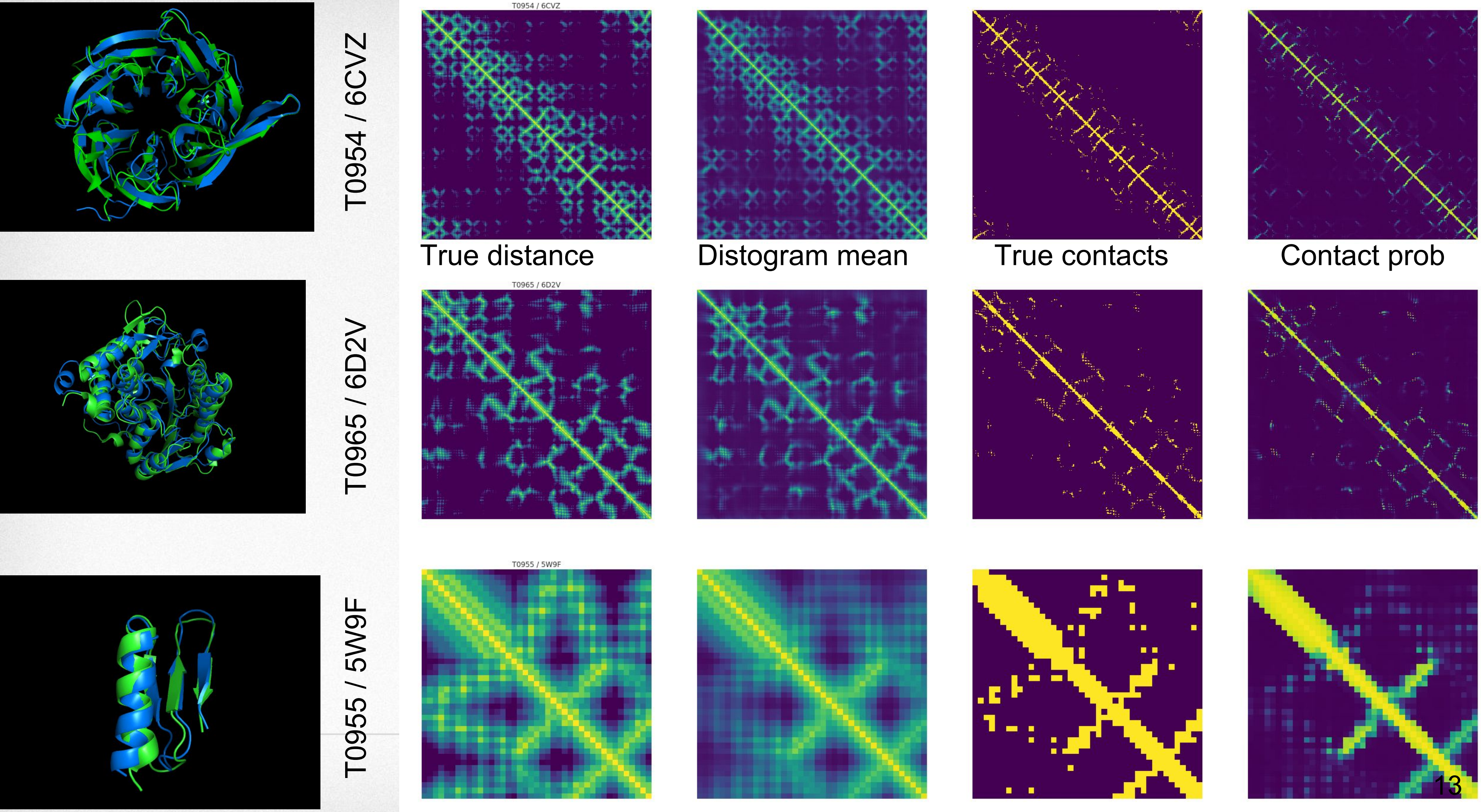
All predicted distributions
for residue 29 to other residues

Red line at true distance





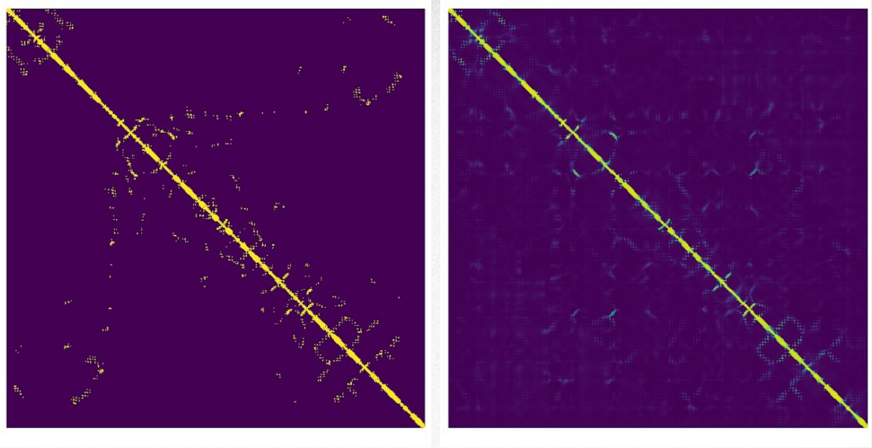
DeepMind


T0954 / 6CVZ
T0954 / 6CVZ
True distance
Distogram mean
True contacts
Contact prob
T0965 / 6D2V
T0965 / 6D2V
T0955 / 5W9F
T0955 / 5W9F

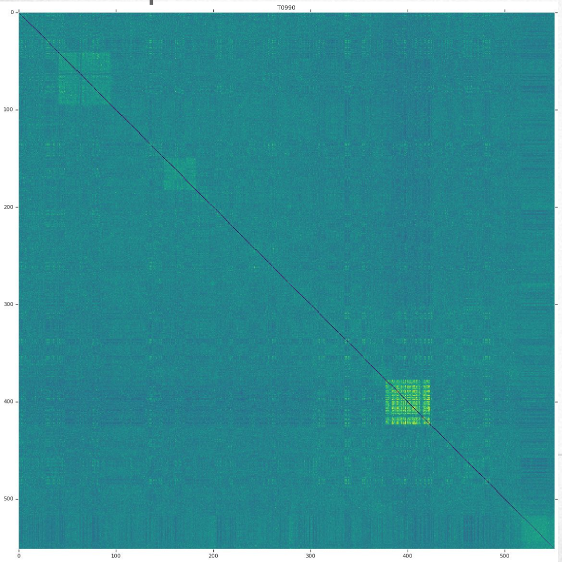
T0990

Precisions at L/k

	L/1 long +Δ	L/2 long +Δ	L/1 medium	L/2 medium	L/1 short	L/2 short	Top 1 GDT+Δ
T0990-D1	51.3 +14.5	68.4 +13.1	30.3	55.3	21.1	39.5	85.2 +17.1
T0990-D2	41.6 +8.3	55.7 +10.9	22.1	39.1	18.2	33.0	45.9 +16.1
T0990-D3	45.5 +15.0	67.9 +23.3	21.6	37.7	27.7	49.1	48.7 +29.5



Input



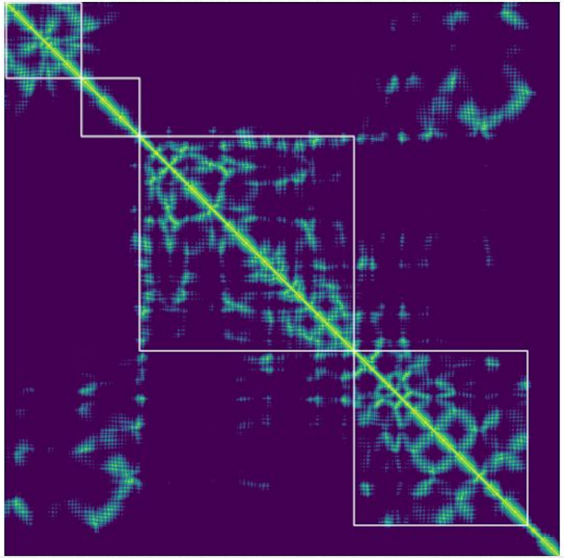
D1

D2

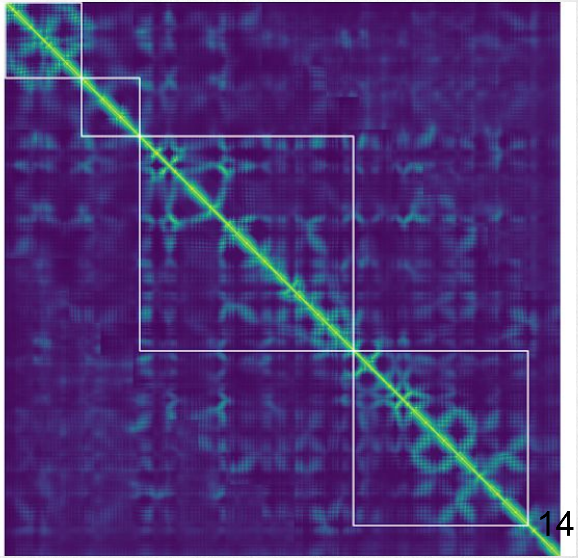
D3

D2

True distance

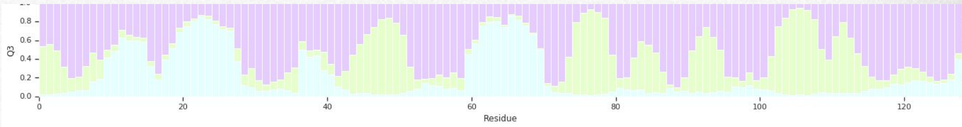


Mean prediction



Auxiliary losses

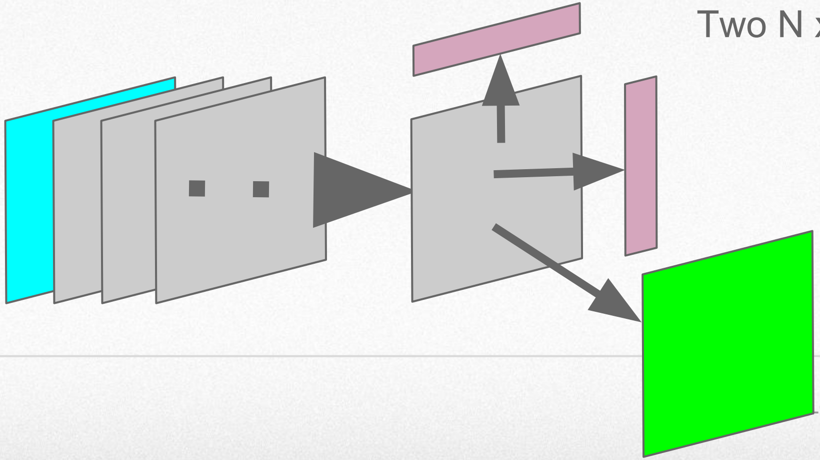
- We know the contact map encodes secondary structure
  - A distance network should be good at predicting it
- *Auxiliary loss* of secondary structure from 1D reductions for **both** (i, i+63) and (j, j+63)
  - Ensembled across all 2D crops



- Q3 Accuracy on CASP11 ~84%
- Predicting secondary structure **improves** contact prediction

Two N x 8 secondary structure predictions

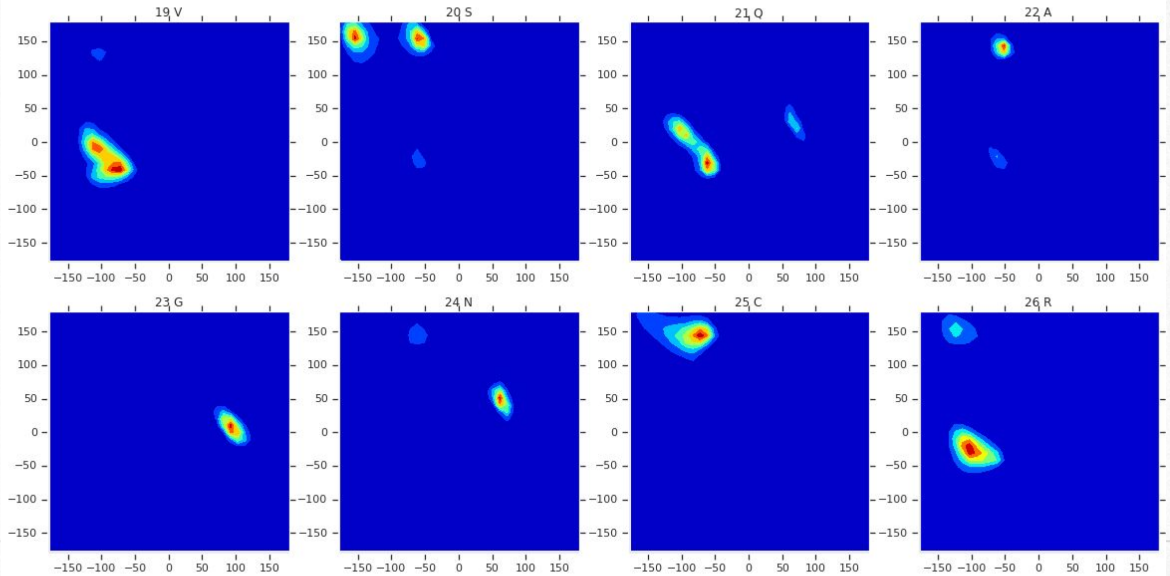
N x N
Input features



N x N x 40
Distance predictions

Auxiliary losses: torsions

- For repeated gradient descent, we need torsion predictions
  - From 1D reduction also predict a joint (phi, psi) Ramachandran probability distribution for each residue (10 degree bins)
  - Again marginal distributions



T0954

Distogram performance on contact metrics

- Sum probability mass below 8 Ångstrom
- Roughly a 4% gain when data was refreshed from pre-CASP12 to latest

	CASP12 FM (27 domains) L long
Single model	50.7%
4-model ensemble	52.3%
Without MSA features	13.6%
Reference model (no AA-type, is_glycine only)	3.8%

CASP13 contact accuracies

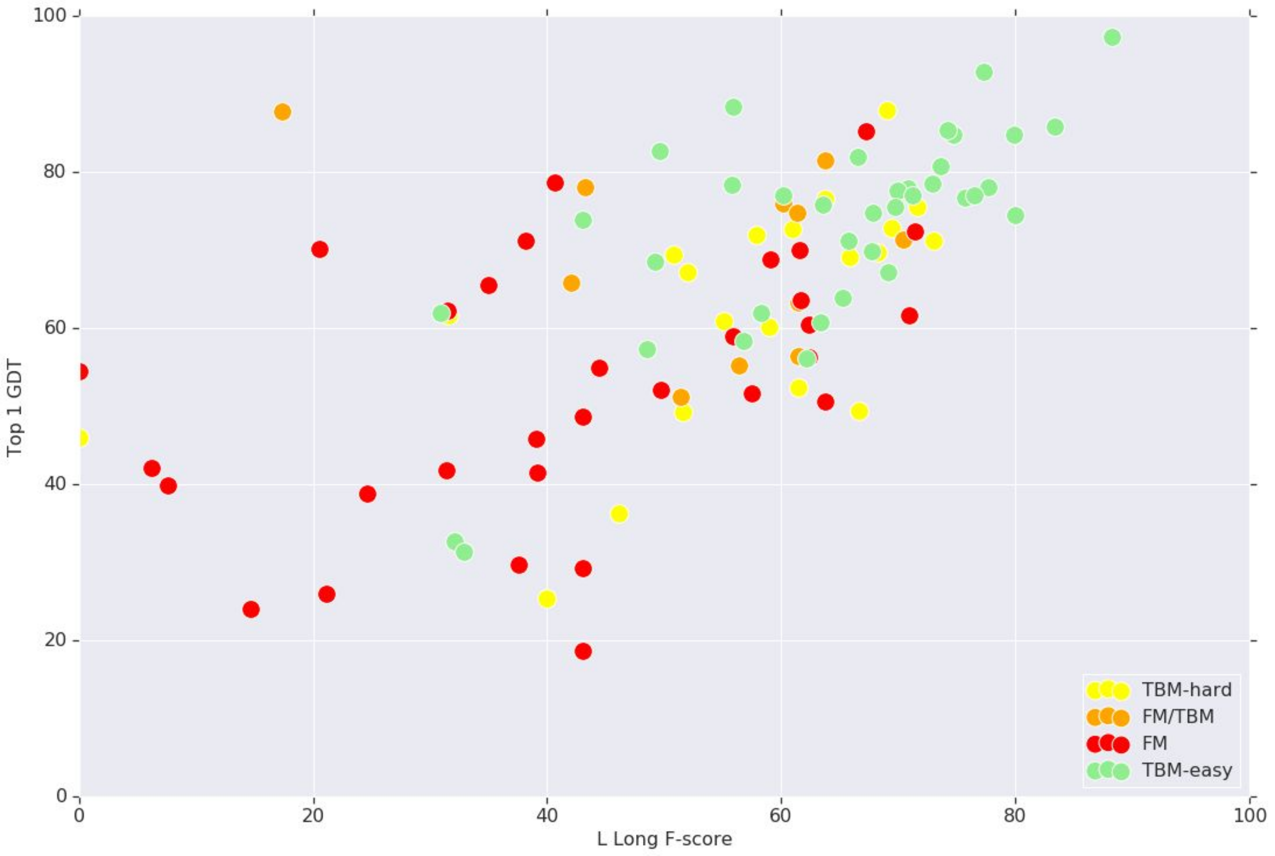
Precisions

Set	Domains	L/1 long +Δ	L/2 long +Δ	L/1 medium	L/2 medium	L/1 short	L/2 short
FM	31	44.7 +0.0	57.9 +0.1	39.6	58.8	32.3	52.2
TBM/FM	12	58.1 -1.8	72.8 -0.4	44.1	65.5	41.9	63.7
Both	43	48.5	62.0	40.8	60.7	35.0	55.4

F scores

Set	Domains	L/1 long +Δ	L/2 long +Δ	L/5 long	L/1 medium	L/2 medium	L/5 medium
FM	31	41.9 +0.8	36.9 +0.7	22.7	49.4	56.5	47.3
TBM/FM	12	55.1 +3.4	48.7 +3.4	31.4	56.4	62.4	47.0
Both	43	45.6	40.2	25.1	51.4	58.1	47.2

GDT vs Long range contact accuracy



Conclusions

What worked well?

- Deep learning!
- Distance prediction
  - Gives greater contact prediction accuracy
  - Is a richer source of information than contact prediction
  - Constructing a potential, with a reference that uses the whole distribution is very valuable
- Crops are effective for modelling even long-range contacts
- Avoiding domain segmentation

What doesn't work well?

- With few or no alignments accuracy is much worse
- T0961-D1 (-35 GDT, TBM Easy), T0966-D1 (-37.8, TBM Hard).....