DeepMetaPSICOV (DMP) in CASP13

Shaun M Kandathil

University College London
&
The Francis Crick Institute

DeepMetaPSICOV (DMP)

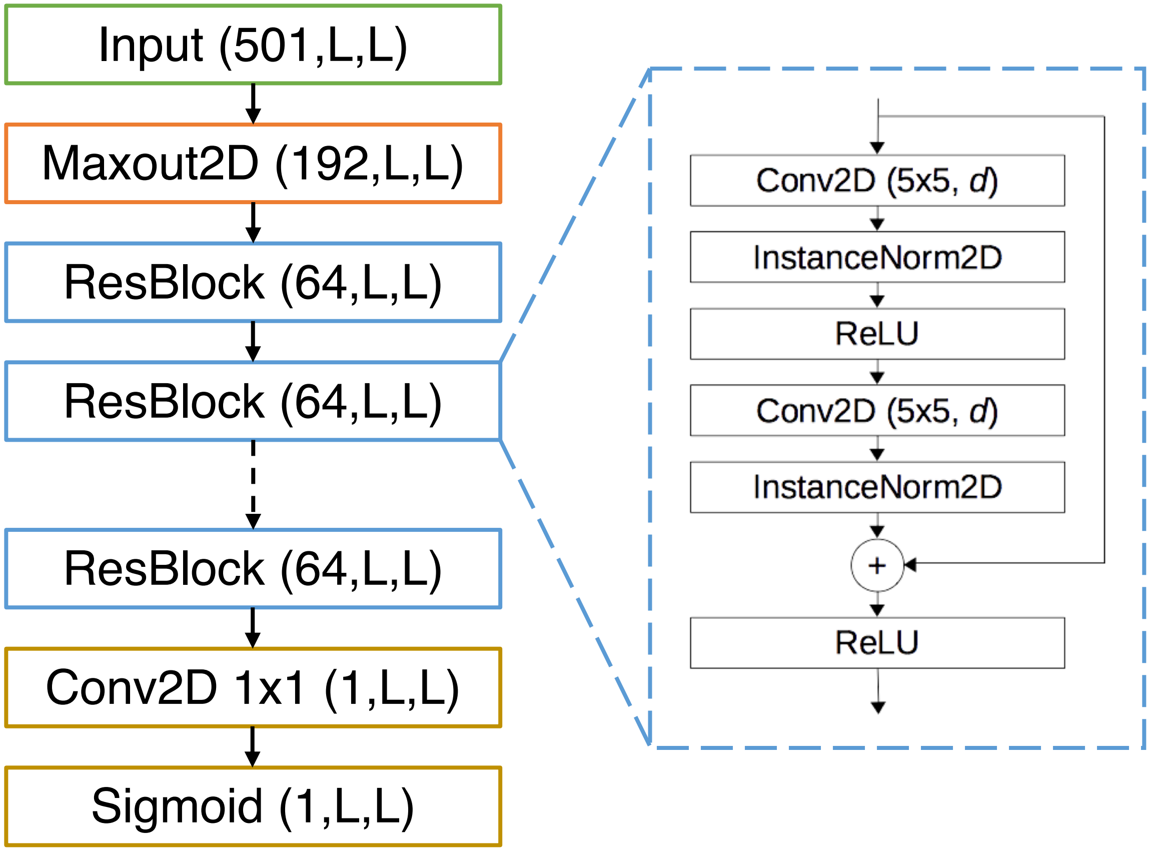
- Combines input features from MetaPSICOV and DeepCov
  - 501 inputs per residue pair
- Fully convolutional deep residual net
- Data augmentation procedures for training
- New alignment generation procedures

DeepMetaPSICOV model architecture

Deep, fully convolutional residual network

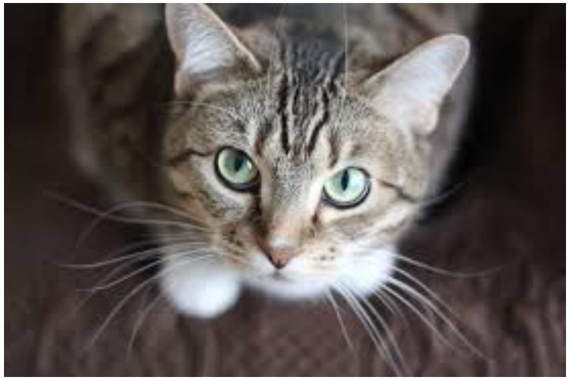
Total of 18 residual blocks plus one Maxout layer

Dilated convolutions



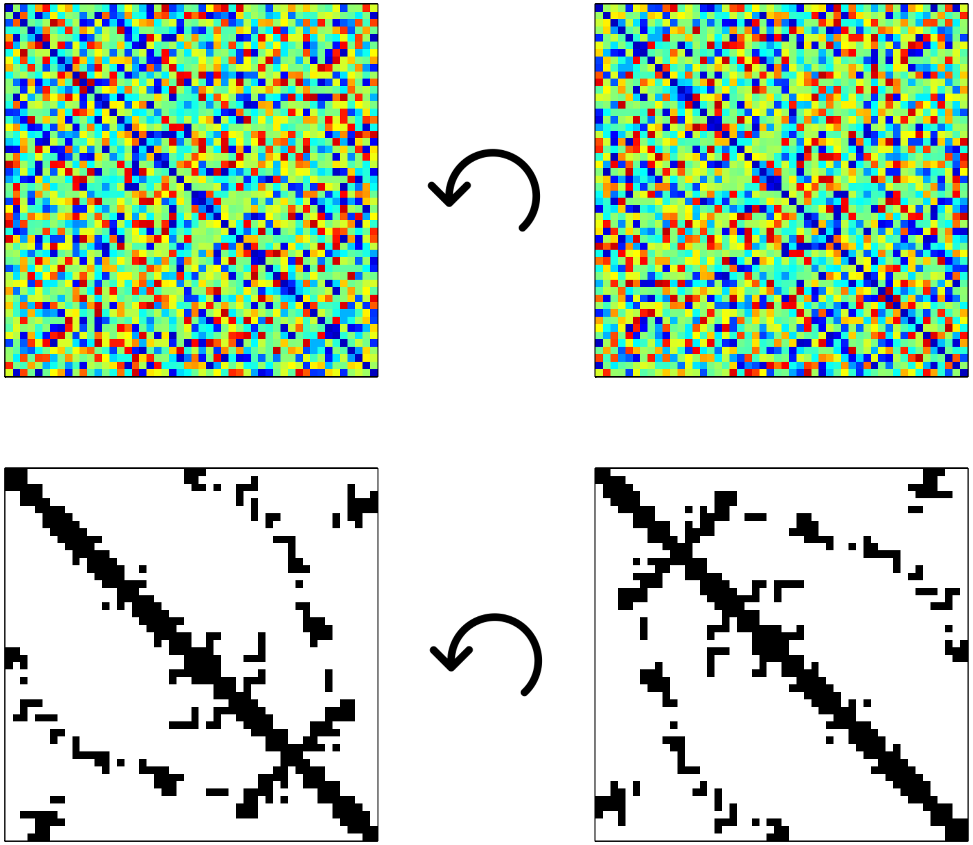
Data augmentations

- More mileage from limited training data
- Generate *plausible* new training examples from existing ones
  - e.g. mirror images, rotated images, pitch-shifted audio
- Discourage memorisation and improve generalisation



Data augmentations

Rotate inputs and contact maps by 180°

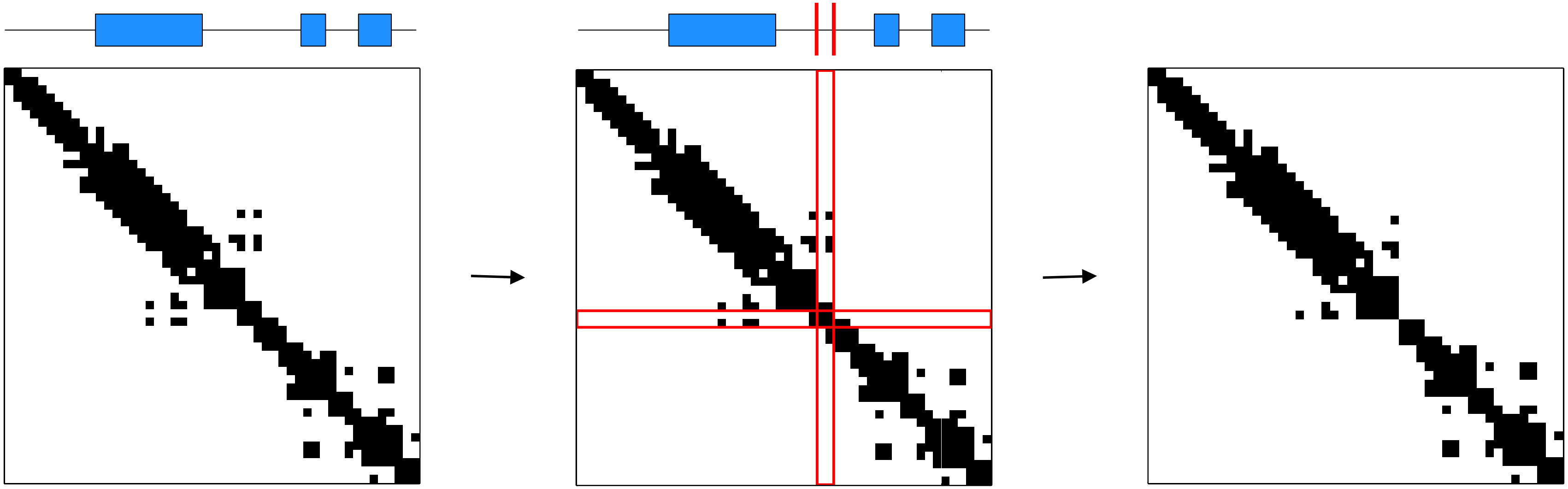


Original example

New example

Data augmentations

Simulate deletions in loops



Original example

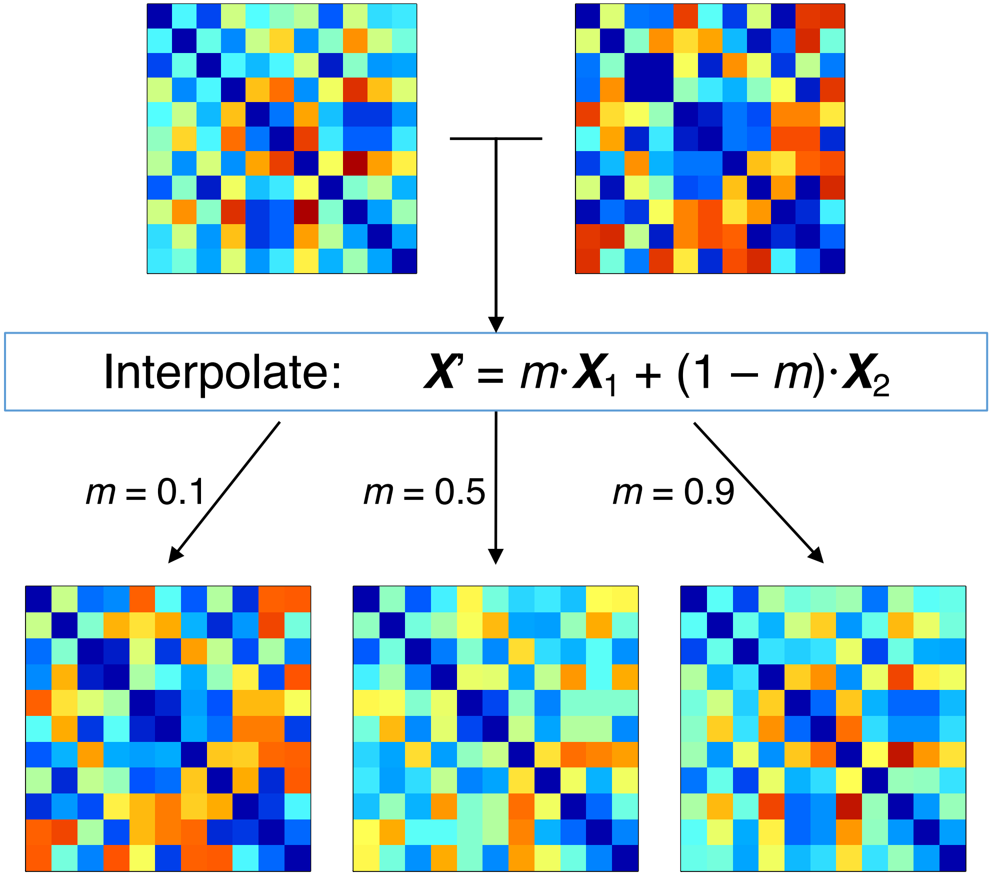
Deletion in loop

New example

Data augmentations

Input feature interpolation

Original examples:
Inputs generated using PSI-BLAST/SwissProt or HHblits/UniClust30 alignments



Interpolate: $\boldsymbol{X}' = m \cdot \boldsymbol{X}_1 + (1 - m) \cdot \boldsymbol{X}_2$

$m = 0.1$

$m = 0.5$

$m = 0.9$

New examples:

Alignment generation

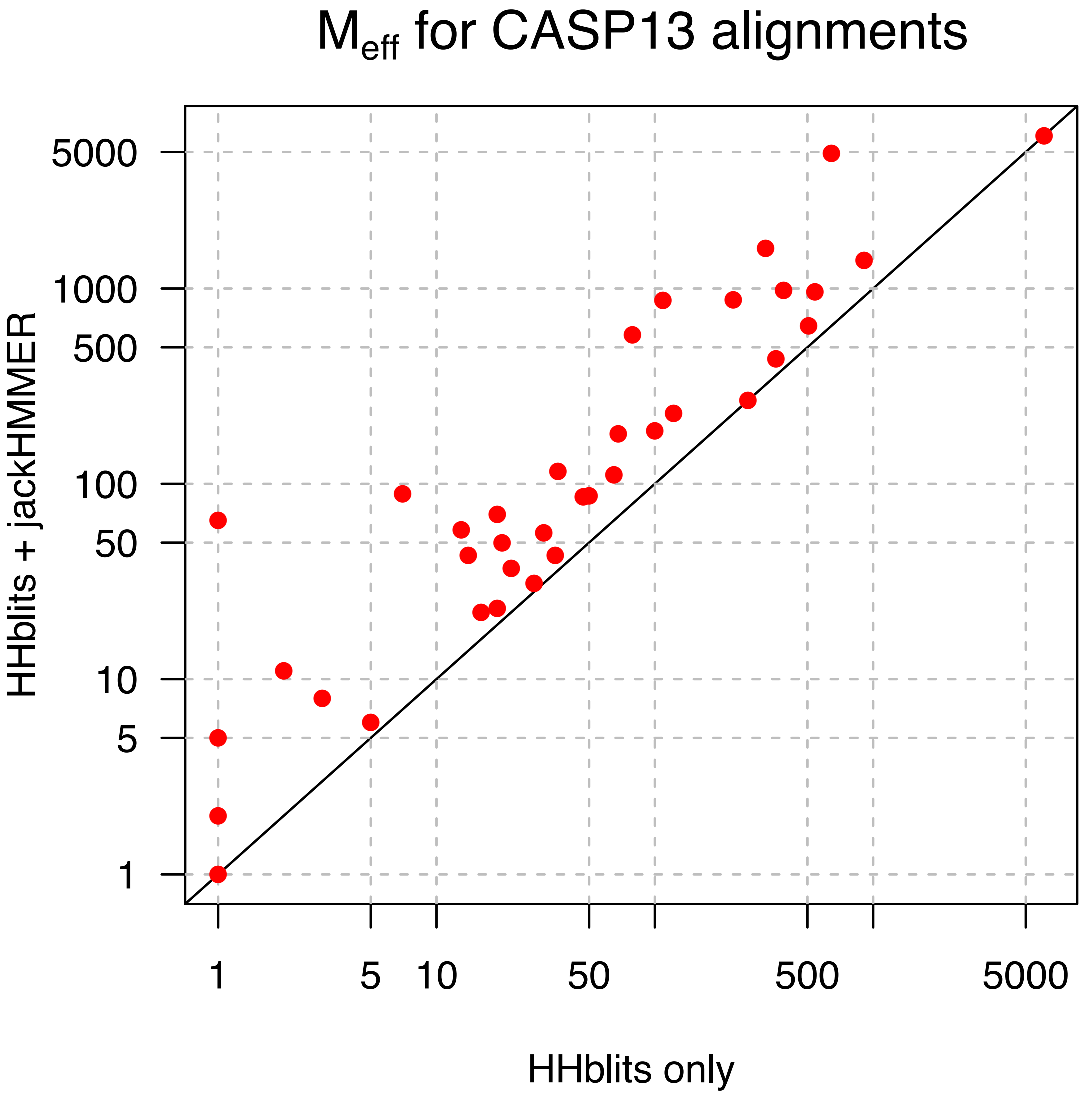
- Initial search using HHblits and UniClust30
- If fewer than 10L raw sequences, search larger database using jackHMMER + HHblits (jack_hhblits)
    - UniRef100 + EBI MGnify peptides

- Also experimented with a version of this procedure that uses hmmbuild and hmmsearch instead of jackHMMER
    - Can iteratively search the custom database using HMMs

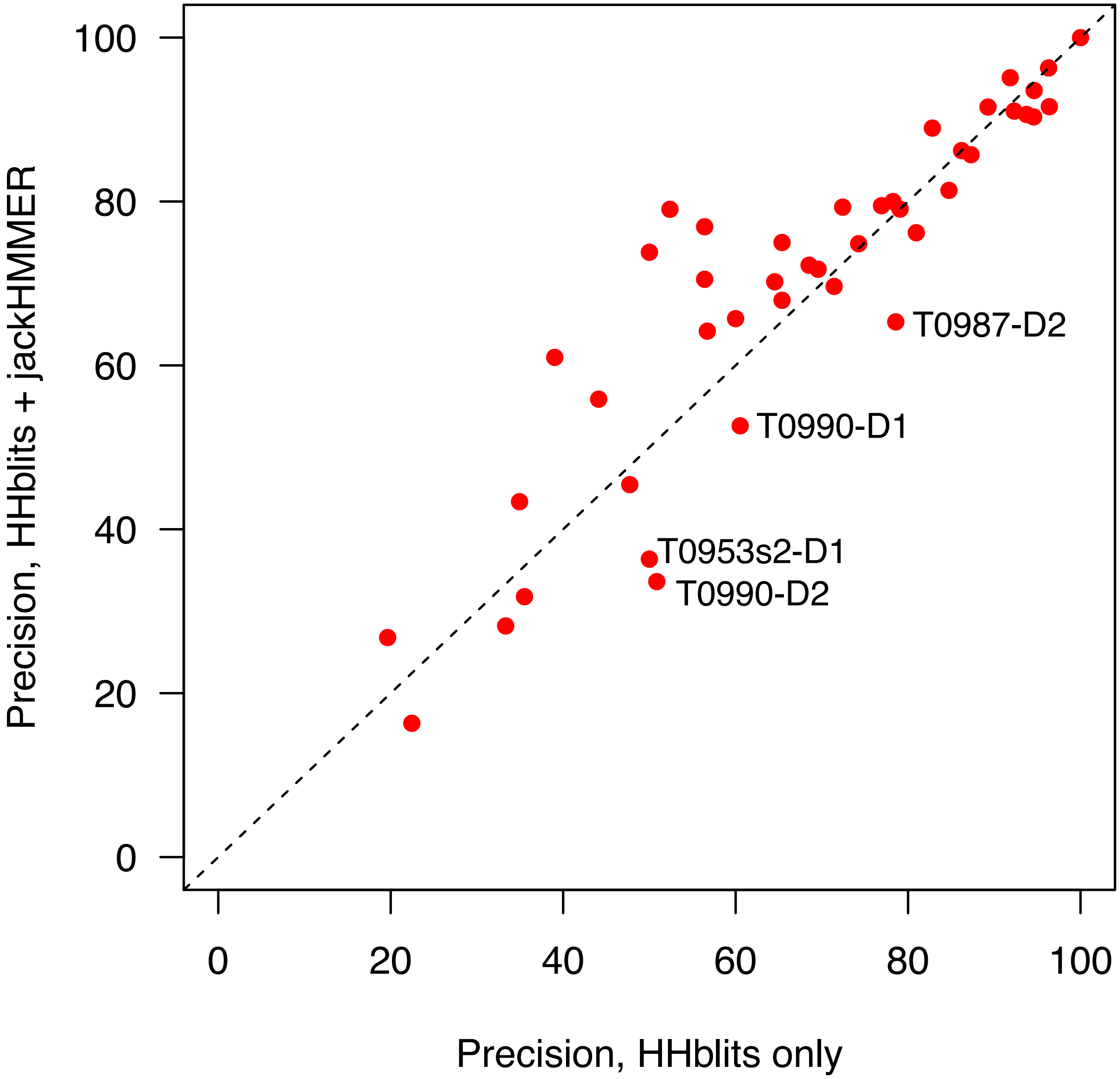
CASP13 results

What went right,
what could have gone better, and
what went wrong

Alignment generation



Was jack_hhblits always better than HHblits only?
Medium + long-range precision, Top-L/2 contacts

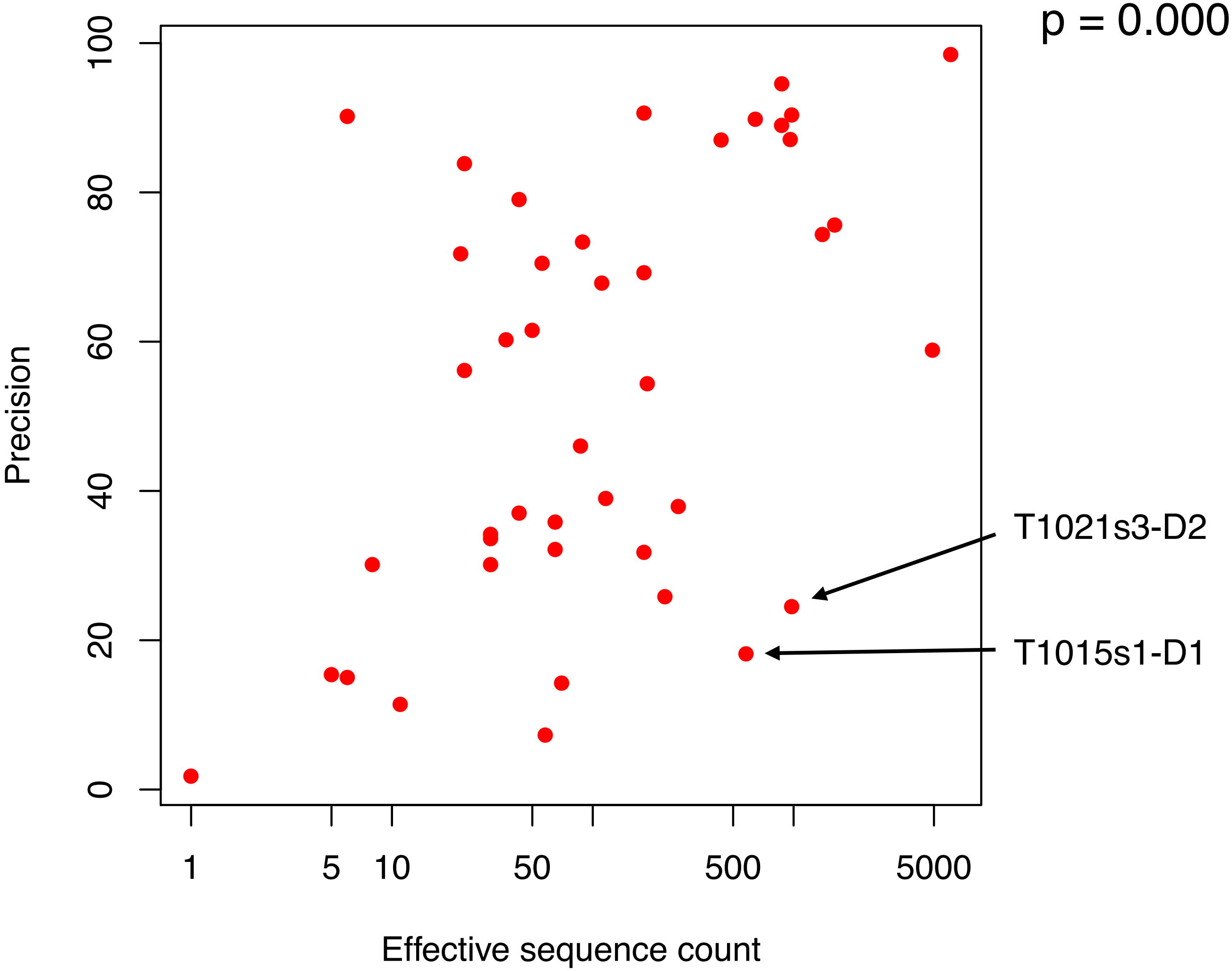


Predictions at $M_{eff} \leq 50$

- 16 domains
- Mean precision:
  - Top-L/2 long-range: 43.61%
  - Top-L/5 long-range: 58.0%

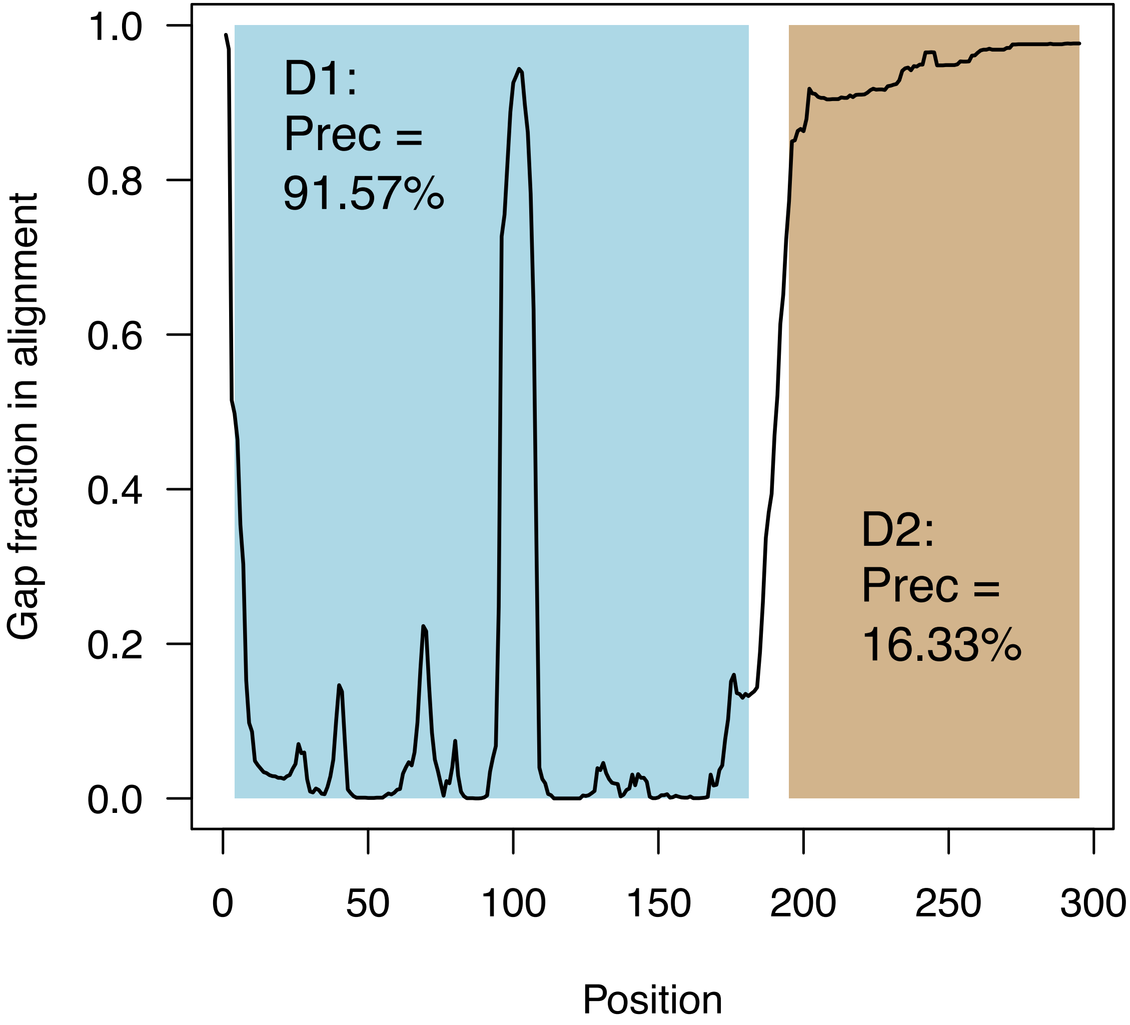
  - Top-L/2 medium+long-range: 60.31%
  - Top-L/5 medium+long-range: 89.03%

Precision (long-range L/2) versus M_{eff}



Pearson ρ = 0.4957,
p = 0.0007

What went wrong: Domain parsing on T1021s3



M_{eff} = 979 but poor coverage at C-terminus

No domains were detected

What went wrong: T1015s1

- Alignment had M_{eff} of 580 (HHblits M_{eff} was 79)
- Top-L/2 long-range precision of 18.18%

- Highly conserved CXC and CXXC motifs; metal binding site
- Most jack_hhblits hits had these motifs, but many were clearly unrelated

- Using HHblits alignment gives 47.72% precision
  - Iterated search with min query id of 20% gives 61.36%

What went wrong: other issues

- Bugs in our code
  - Incorrect calculation of mutual information affected all predictions (but not training)
  - Loss of around 3-7% mean long-range precision!

- Bugs in other people's code
  - Large size of T0999 revealed issue with dilated convolutions in PyTorch v0.3.1 (fixed in v0.4.1+)
  - Errors in HHblits PDB70 database affected domain parsing for several targets (now fixed)

What could have gone better

- Iterated sequence search deemed too unstable for use during the prediction season
- Prone to profile drift; pulls in unrelated sequences
- However, it did give better results in some cases, e.g. T1010:

	HHblits only	HHblits + jack_hhblits	HHblits + 3 iterations of (hmmsearch + HHblits)
M_{eff}	7	89	200
Medium + long-range precision (L/2)	52.38%	79.05%	91.43%

Conclusions

- Using expanded sequence databanks is valuable
    - Integrating more sources may give even better results

- Alignment generation can be improved
    - Iterated sequence searching shows some promise, but needs care
    - Need effective, objective measures of alignment quality
    - Gap fraction in alignment columns?
    - Interactions with domain parsing

- Don’t have bugs in your code.

Acknowledgements

BBSRC
bioscience for the future

erc

Thank you!

- All methods will be made available on our GitHub: https://github.com/psipred

- Read about our TS method (DMPfold):
    - https://arxiv.org/abs/1811.12355

- New PSIPRED server is coming
    - http://bioinf.cs.ucl.ac.uk/psipred_beta/ **(include the last '/')**

s.kandathil@ucl.ac.uk

d.t.jones@ucl.ac.uk

501 input features per residue pair

- Raw AA covariances (as in DeepCov) : 441
- Sequence profiles: 42
- Secondary structure (PSIPRED v4): 6
- Solvent accessibility (SOLVPRED): 2
- Shannon entropy: 2
- PSICOV: 1
- plmDCA (CCMpred): 1
- mfDCA (FreeContact): 1
- Mutual Information: 2
- Mean contact potential: 1
- Sequence separation: 1
- Channel of 1s: 1

Automatic domain parsing

Same procedure as in CASP12:

- First run DMP on whole target sequence
- HHblits search against PDB70 (Söding group)
- Re-run DMP on any region of sequence that did not match detected domains
- Copy predicted scores for this region into final contact map

Alignment generation

Initial HHblits search against UniClust30 (Söding group)

If fewer than 10*L* raw sequences, jack_hhblits:

- jackHMMER search against custom database comprising UniRef100 + EBI MGnify peptides
- Cluster hits using kClust, align clusters using MAFFT
- Make HHblits database from cluster alignments
  - Include initial HHblits alignment!
- Final HHblits search against this database

- Also experimented with a version of this procedure that uses hmmbuild and hmmsearch instead of jackHMMER
  - Can iteratively search the custom database using HMMs