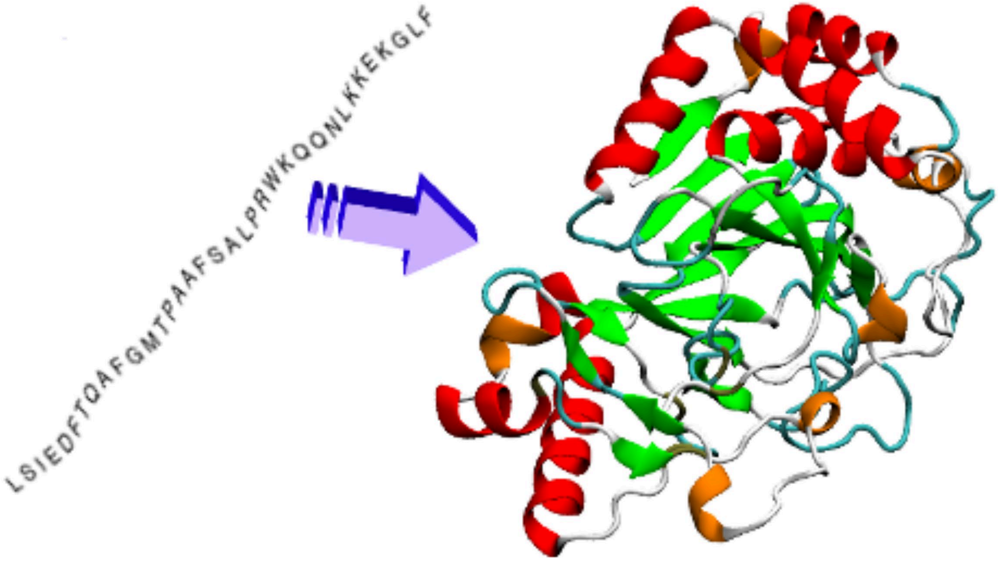
CASP14: Critical Assessment of Structure Prediction

- Rigor
- Transparency
- Collaboration
- Communication



Advancing solutions to the Protein Folding Problem

NIH National Institute of General Medical Sciences

CASP Requirements

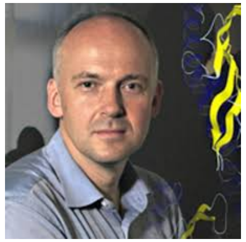
- Many Targets.
- Each Participant makes many predictions.
- Many Participants.
- Gold standard
- Head-to-head methods comparison
- Clear Metrics
- Authoritative evaluation.

Founders and Organizers


Anna Tramontano


Krzysztof Fidelis


Andriy Kryshtafovych


Torsten Schwede


Maya Topf


Jan Pedersen


Tim Hubbard


Steve Bryant


Burkhard Rost

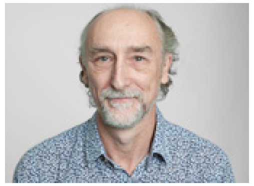
CASP14 ASSESSORS

• Topology	Nick Grishin, Lisa Kinch
• High accuracy modeling	Andrie Lupus, Joana Pereira, Marcus Harman
• Refinement	Dan Rigden
• Contacts	Alfonso Valencia, Rosalba Lepore
• Accuracy estimation	Chaok Seok
• Assemblies	Ezgi Karaca
• CAPRI	Marc Lesink, Shoshana Wodak
• Function	Sandor Vadja, Dima Kozakov











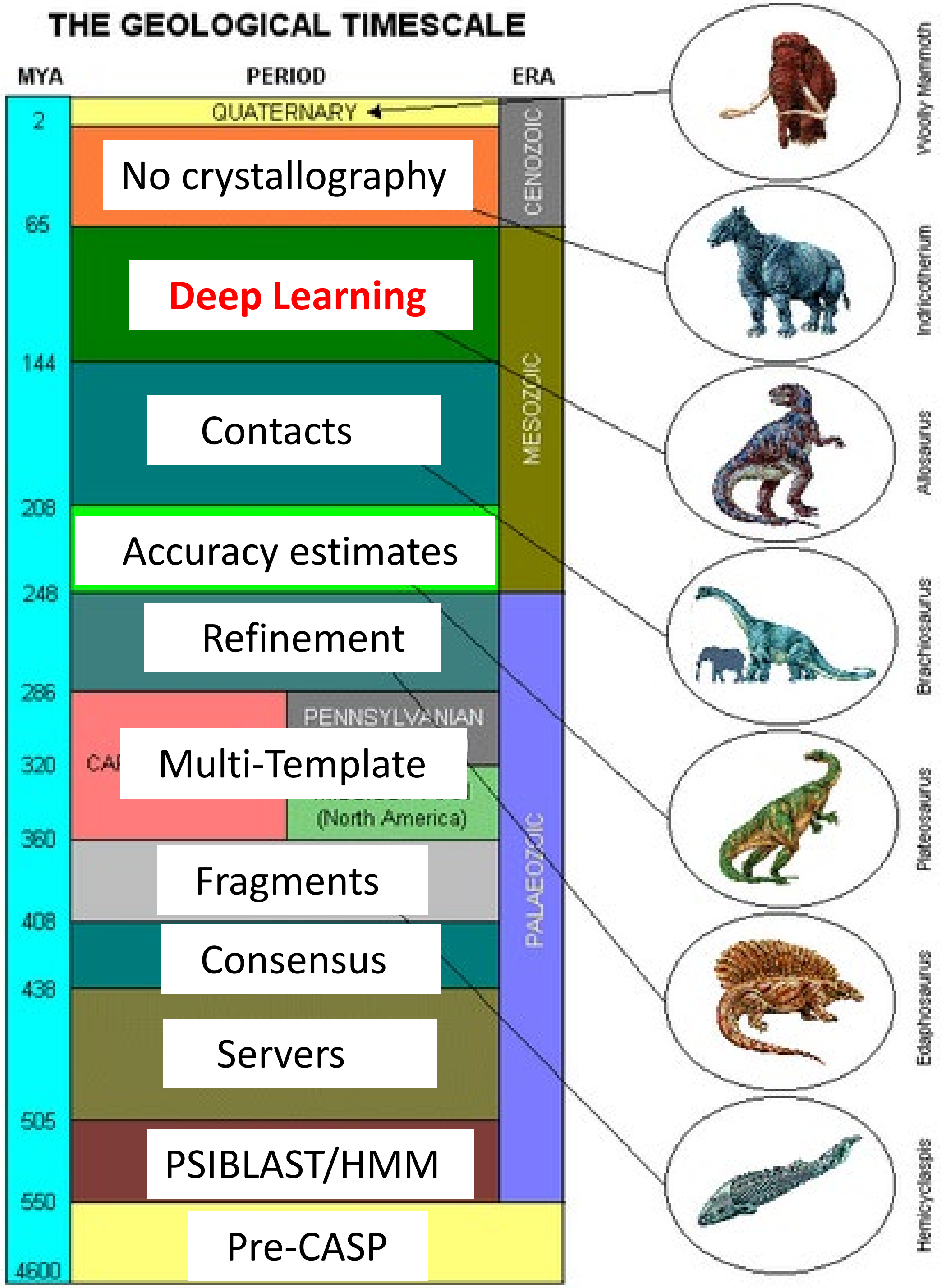



How this meeting (should) work

- Main zoom sessions
- Main interface – posters links, How to, link to Airmeet
- Airmeet lounge – help, informal meetings, poster sessions.
- DISCORD channels
- Extra sessions
- Ongoing interest groups

Protein structure
Refinement
Protein Assemblies
Contacts and Distances
Accuracy Estimation
Deriving function
Covid and CASP
The future of deep learning

CASP Punctuated Equilibrium



No crystallography

Deep Learning

Contacts

Accuracy estimates

Refinement

Multi-Template

Fragments

Consensus

Servers

PSIBLAST/HMM

Pre-CASP

PROTEINS: Structure, Function, and Genetics 18:309–317 (1994)

Correlated Mutations and Residue Contacts in Proteins

Ulrike Göbel, Chris Sander, Reinhard Schneider, and Alfonso Valencia
Protein Design Group, European Molecular Biology Laboratory, D-69012 Heidelberg, Germany.

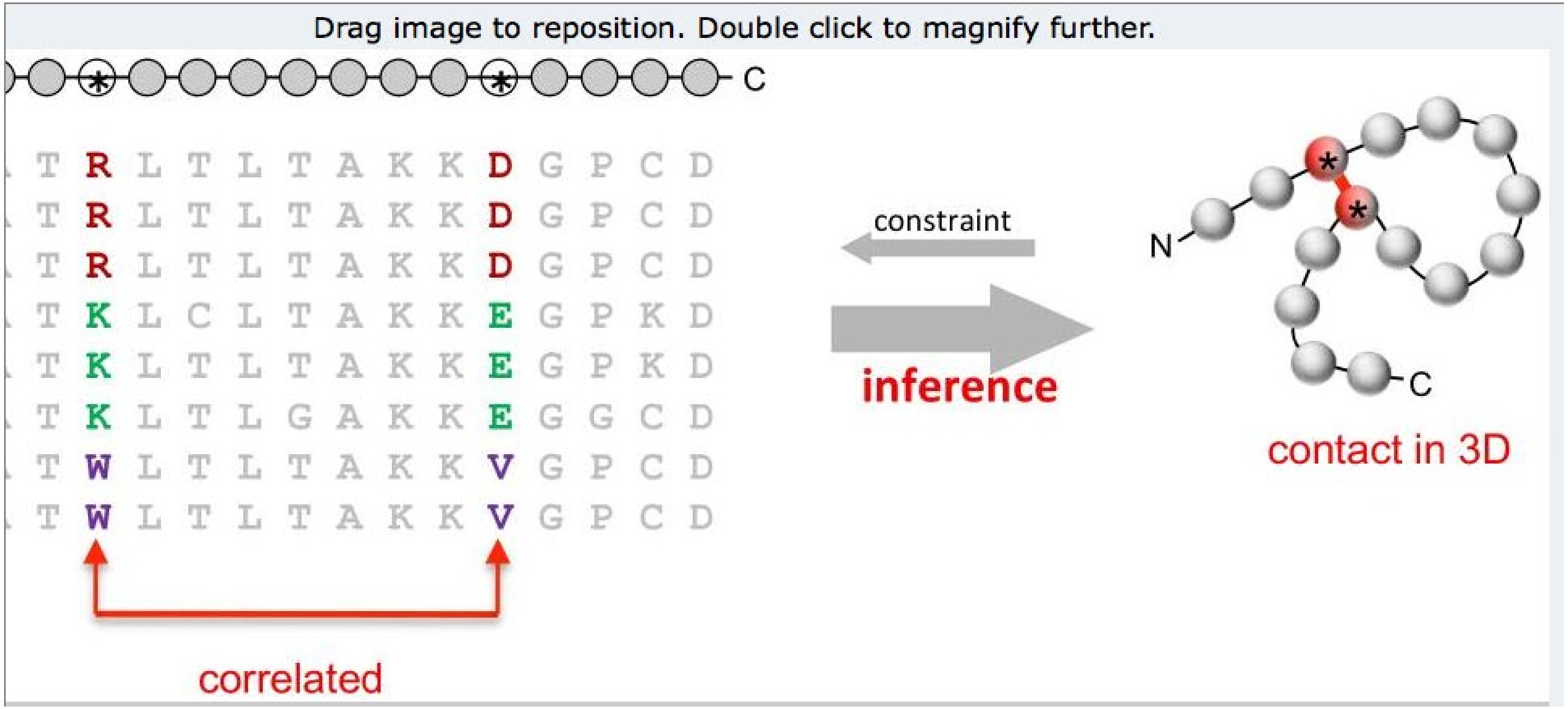
ABSTRACT **The maintenance of protein function and structure constrains the evolution of amino acid sequences. This fact can be exploited to interpret correlated mutations observed in a sequence family as an indication of probable physical contact in three dimensions. Here we present a simple and general method to analyze correlations in mutational behavior between different positions in a multiple sequence alignment. We then use these correlations to predict contact maps for each of 11 protein families and compare the result with the contacts determined by crystallography. For the most strongly correlated residue pairs predicted to be in contact, the prediction accuracy ranges from 37 to 68% and the improvement ratio relative to a random prediction from 1.4 to 5.1. Predicted contact maps can be used as input for the calculation of protein tertiary structure, either from sequence information alone or in combination with experimental information.**

Key words: protein structure prediction, predicted contact maps, correlated mutations

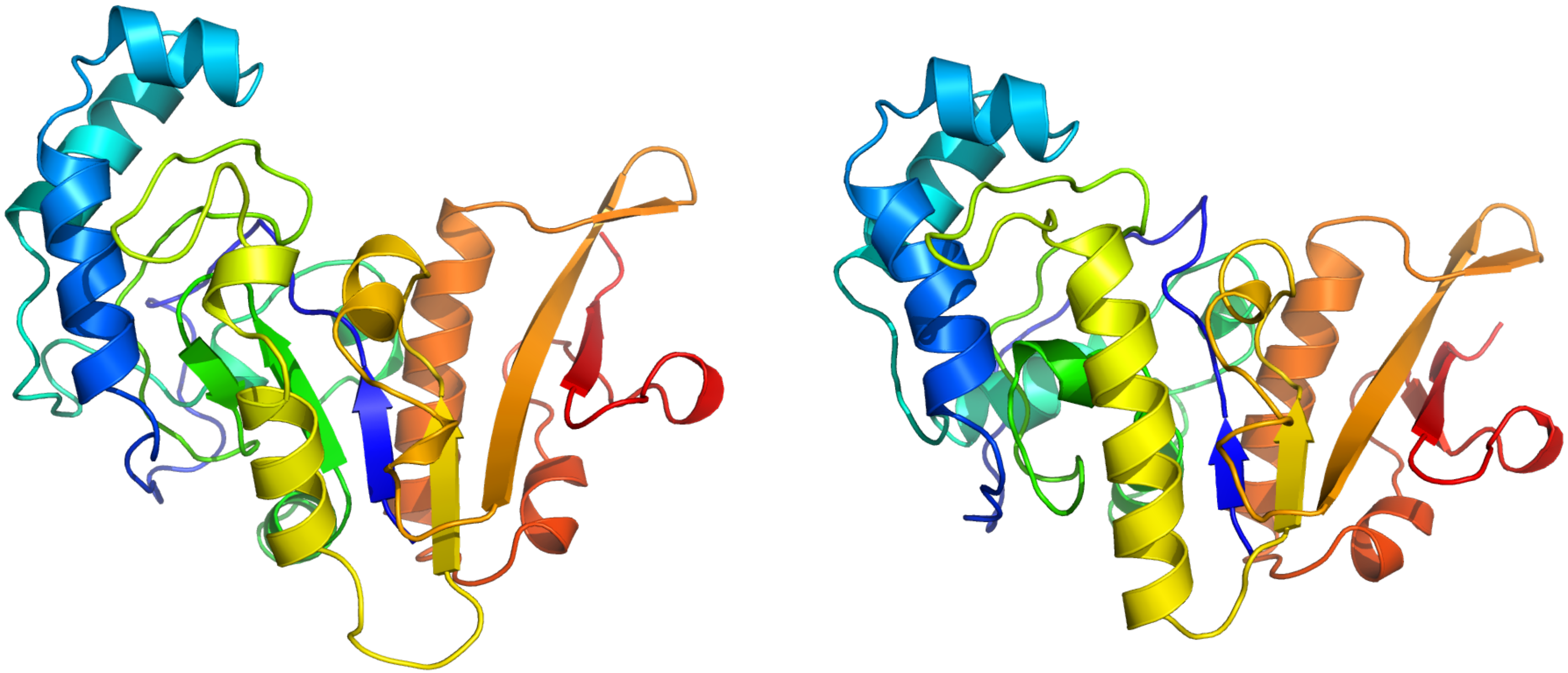
tions, differential replication of genetic information depending on genotype and phenotype, and elimination of cells containing dysfunctional or noncompetitive sequences. When functionally negative point mutations are compensated for by other mutations, the cells may survive and with it the protein sequence information. We need to learn how to extract information about the various types of evolutionary constraints from multiple sequence data and then exploit this information for the prediction of three-dimensional structure, via distance constraints.

Functional or Structural Constraints?

Given many sequences in a protein family, the first tasks in decoding the observed mutational patterns are to identify constrained sequence changes on a background of neutral mutational drift, i.e., to separate signal from noise, and to separate the effects of structural constraints from those of functional constraints that have no structural implications. The task is difficult for several reasons. First, evolutionary constraints are not directly observable, as they specify which attempted point mutations would lead to the elimination of a particular genotype from a population. Second, the distinction between functional and structural constraints is not clear-cut: e.g., mutations in some residues may af-

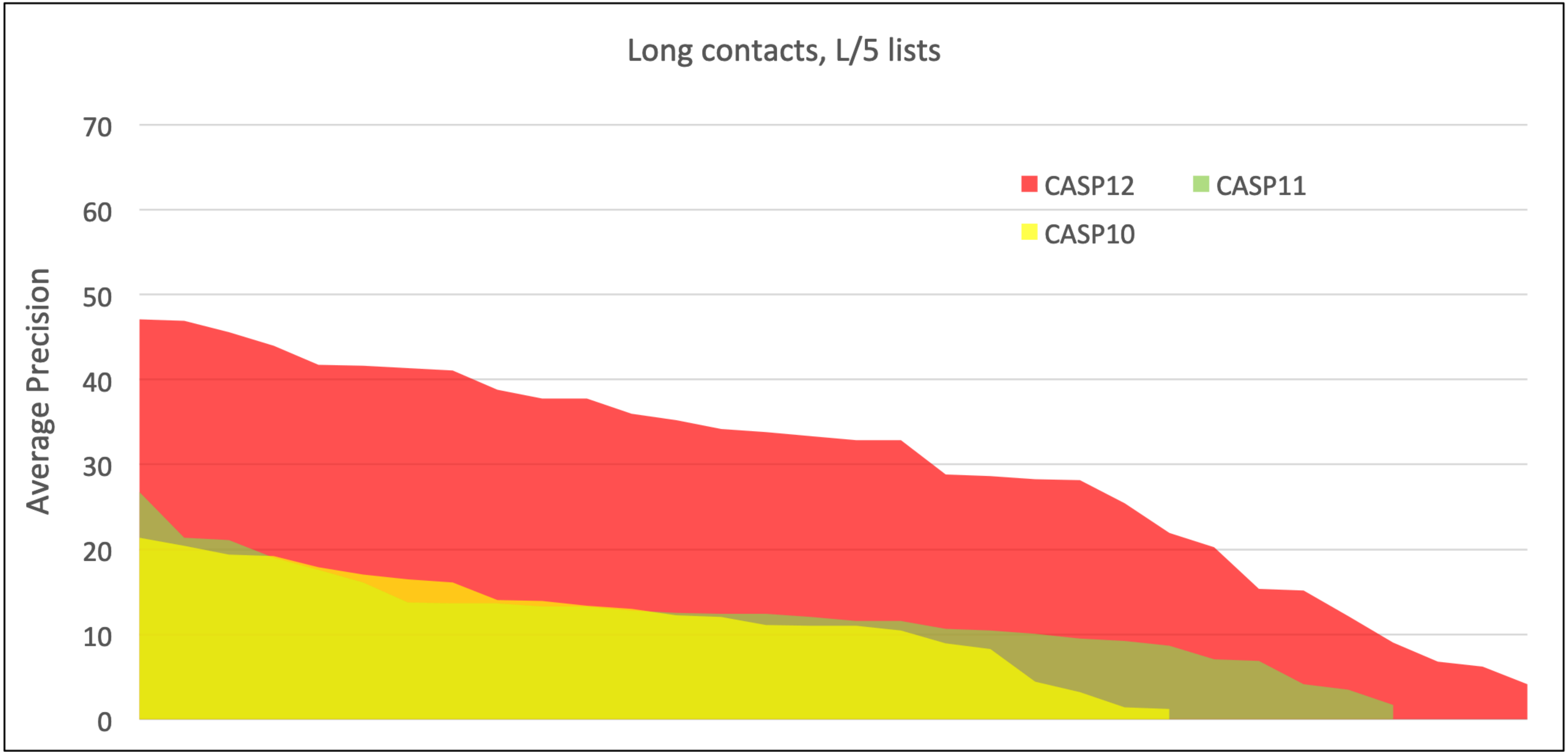


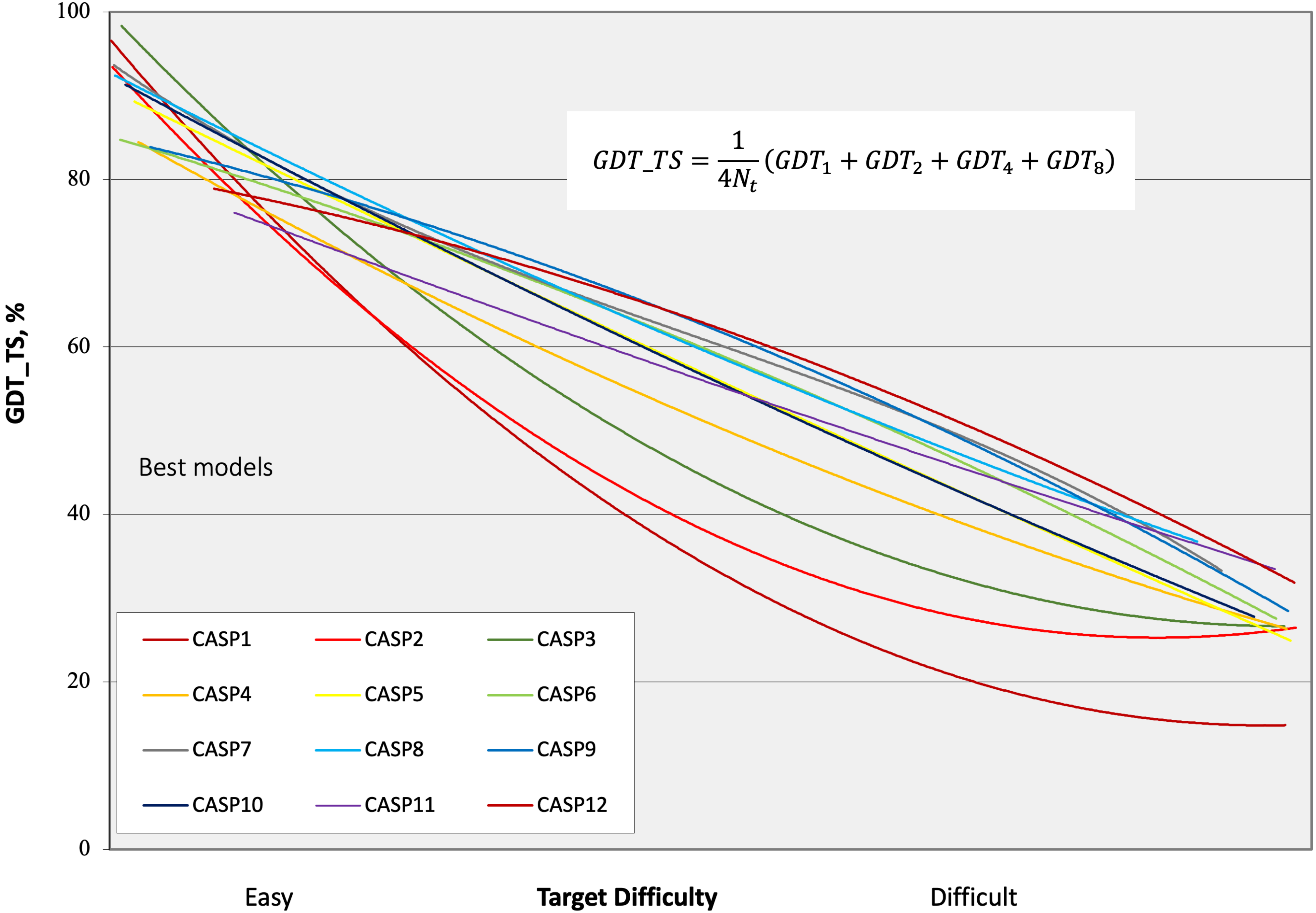
```
T R L T L T A K K D G P C D
T R L T L T A K K D G P C D
T R L T L T A K K D G P C D
T K L C L T A K K E G P K D
T K L T L T A K K E G P K D
T K L T L G A K K E G G C D
T W L T L T A K K V G P C D
T W L T L T A K K V G P C D
```



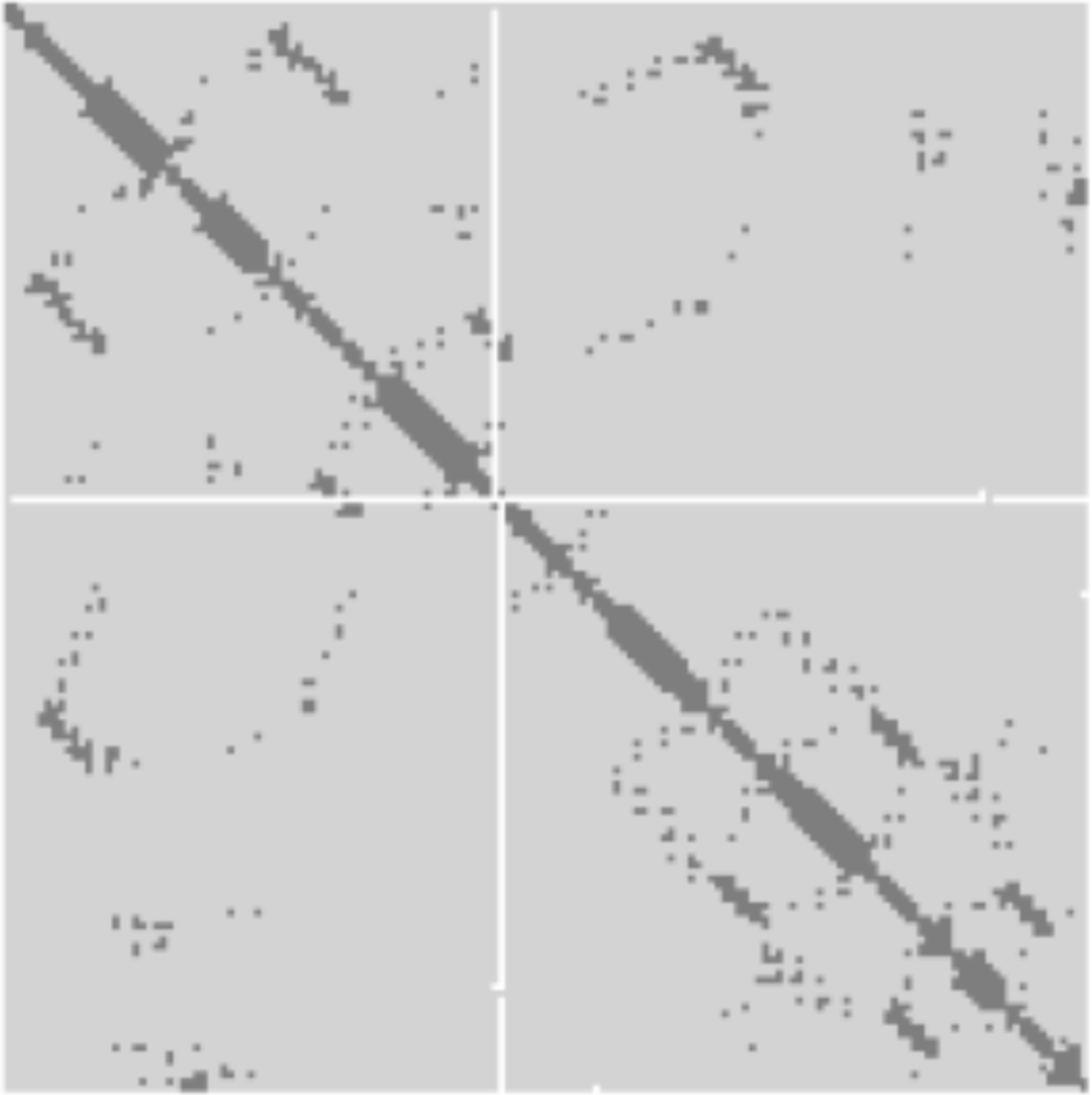
‘So, either this group is close to solving the folding problem or they cheated somehow.’

Nick Grishin

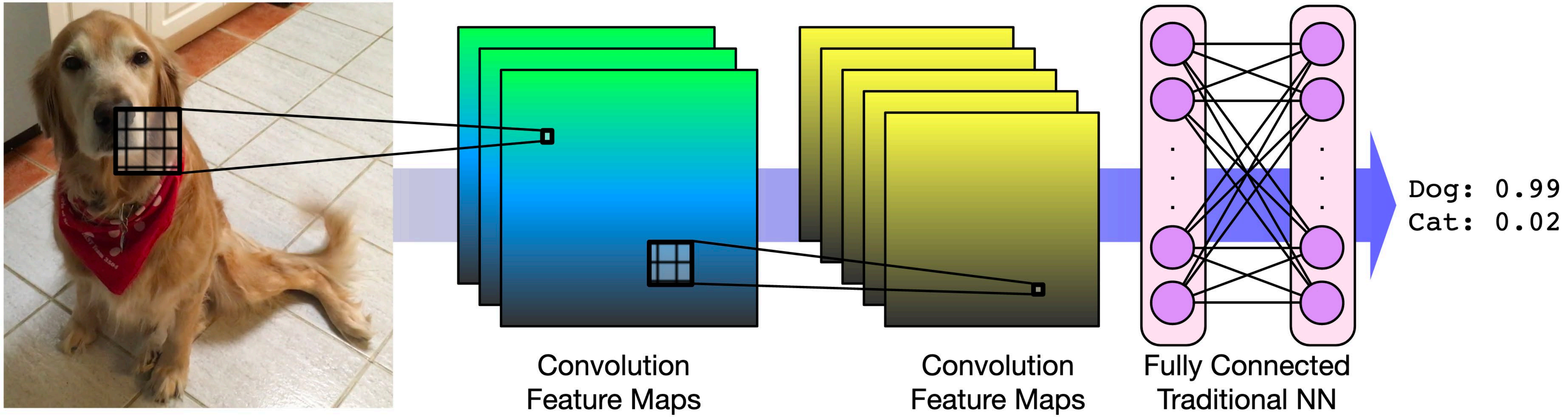

Long contacts, L/5 lists
Average Precision
70
60
50
40
30
20
10
0
CASP12
CASP11
CASP10

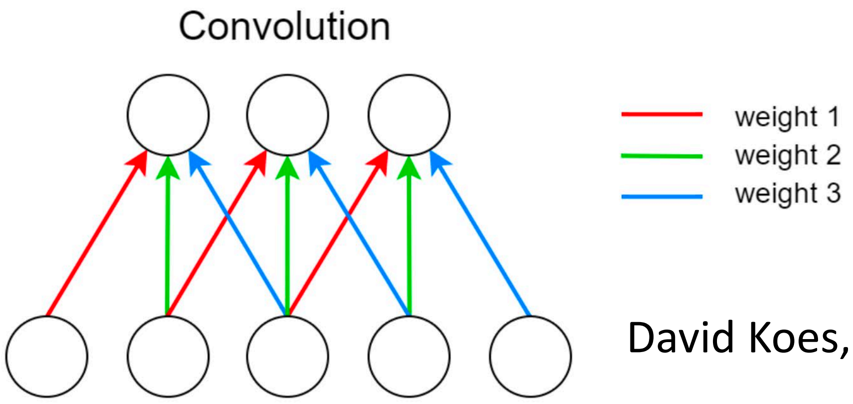


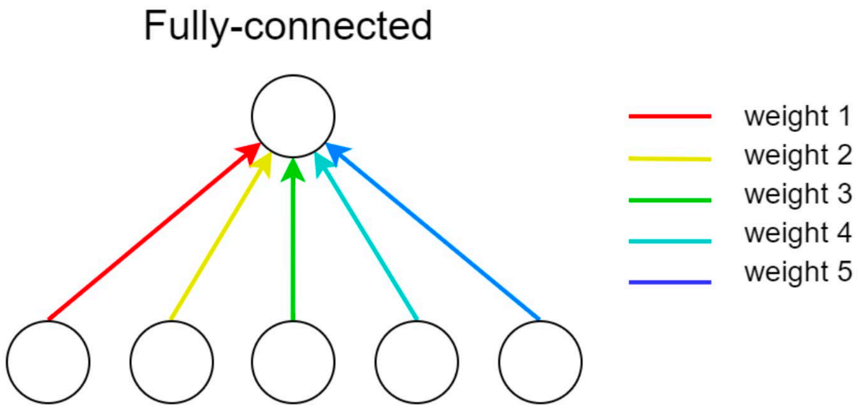
$$GDT\_TS = \frac{1}{4N_t}(GDT_1 + GDT_2 + GDT_4 + GDT_8)$$



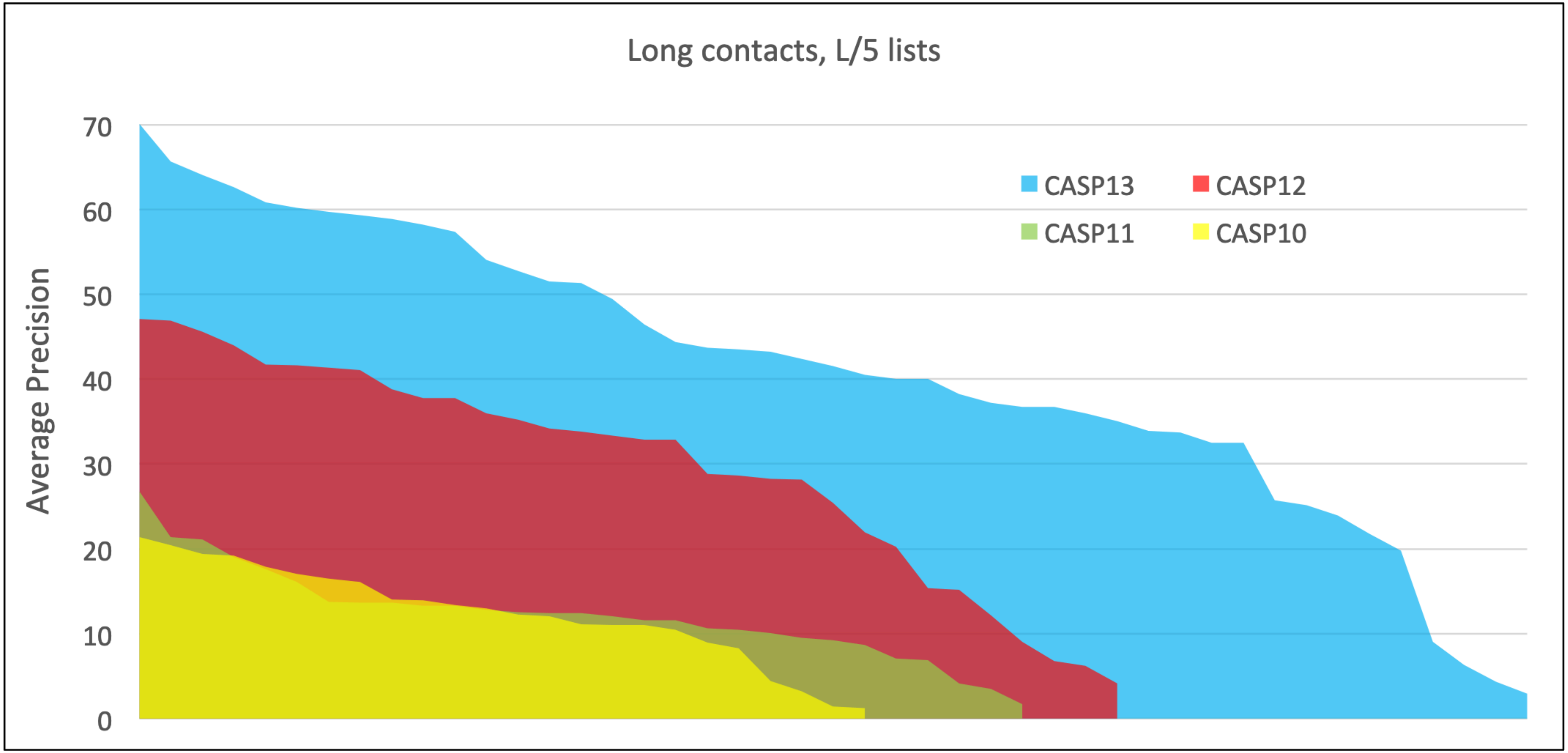
Convolutional Neural Networks

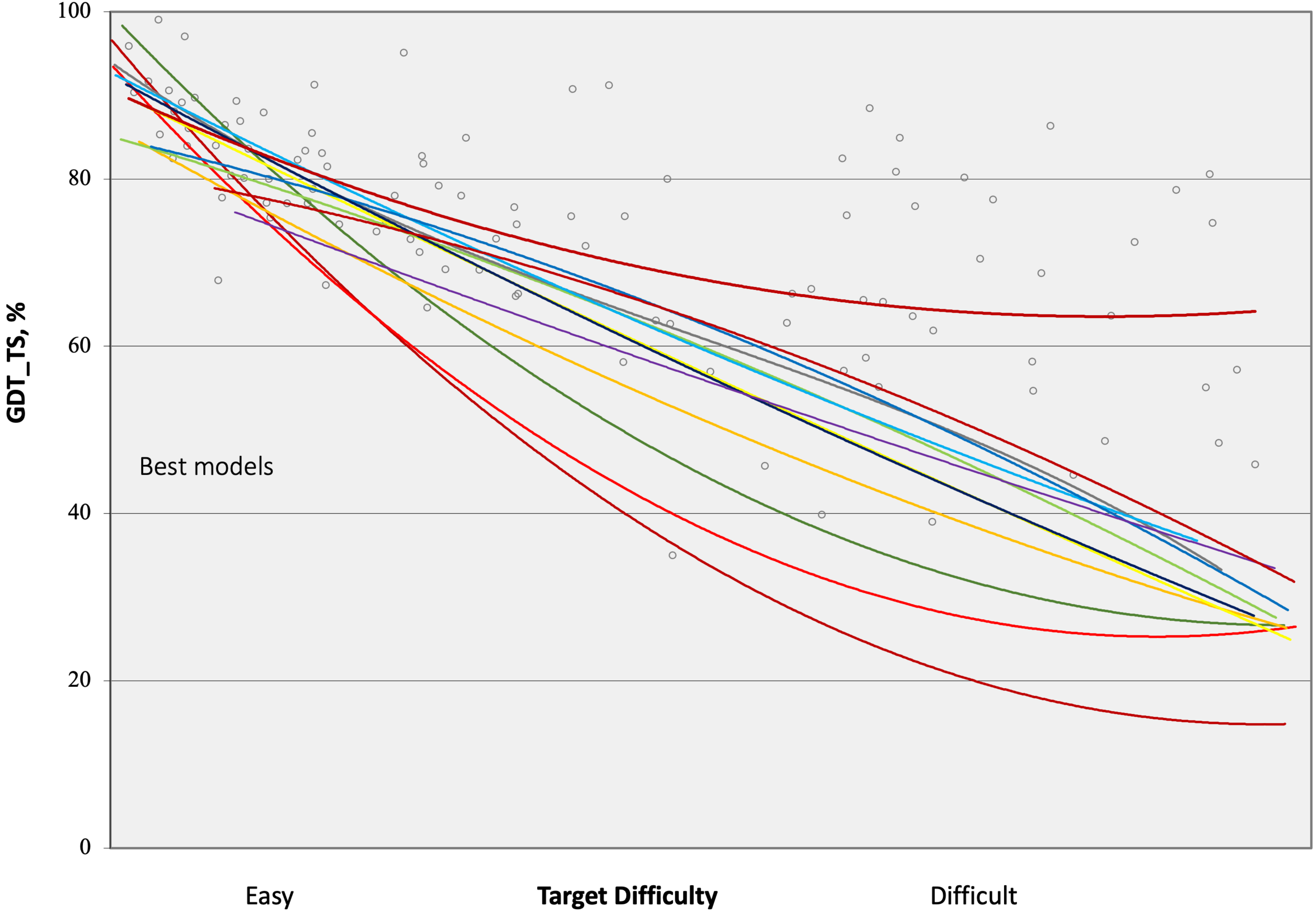


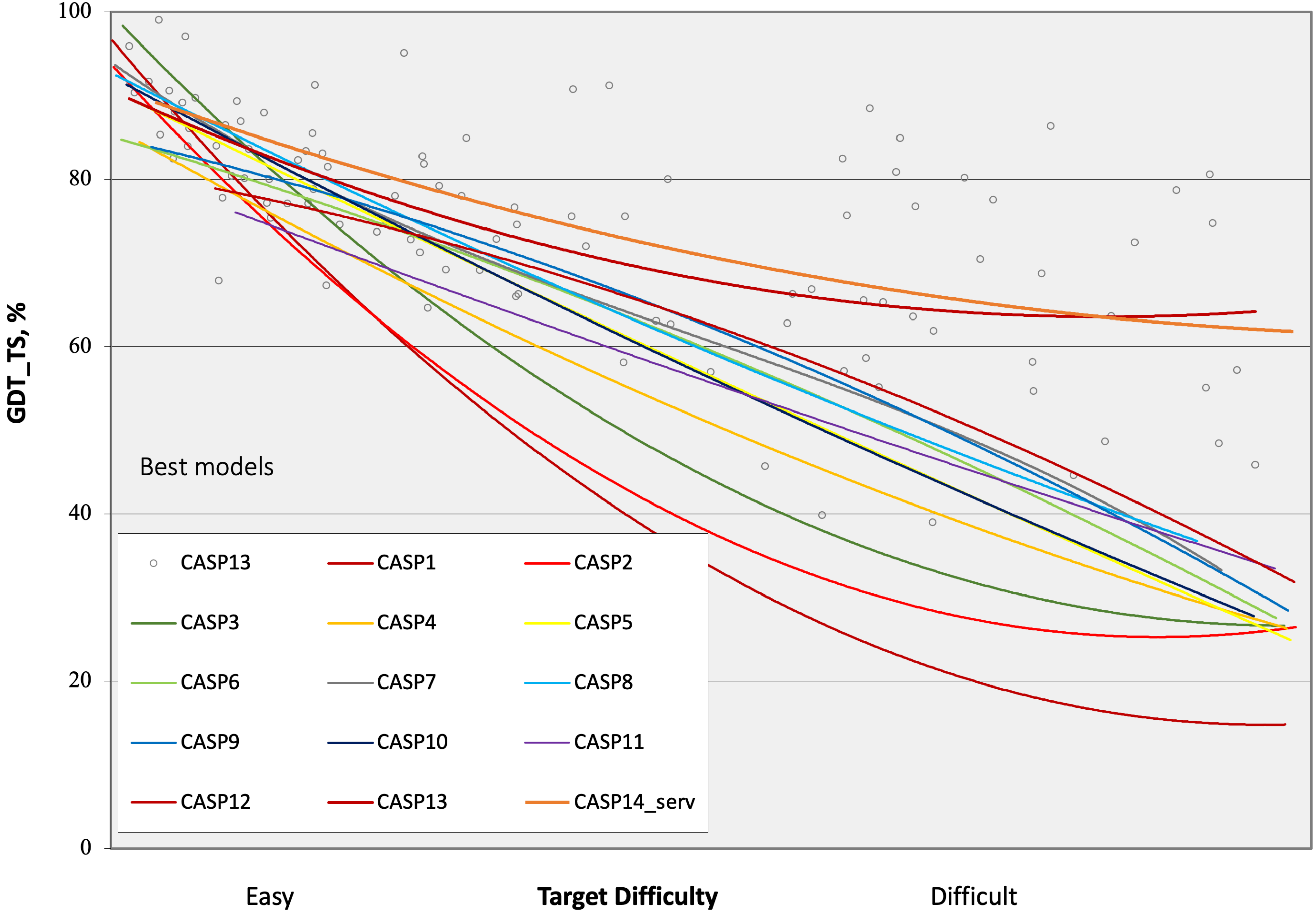




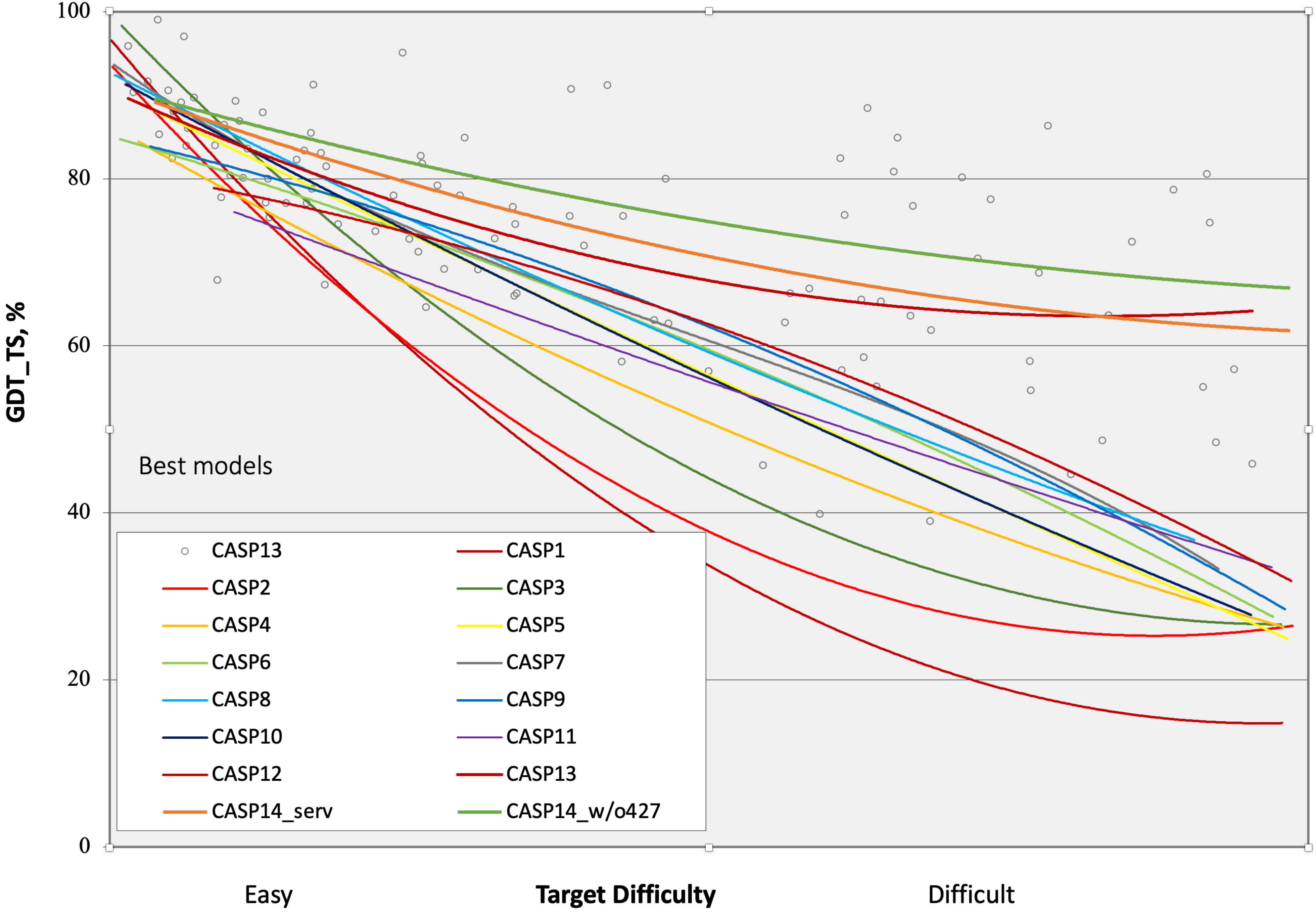
David Koes, CASP13

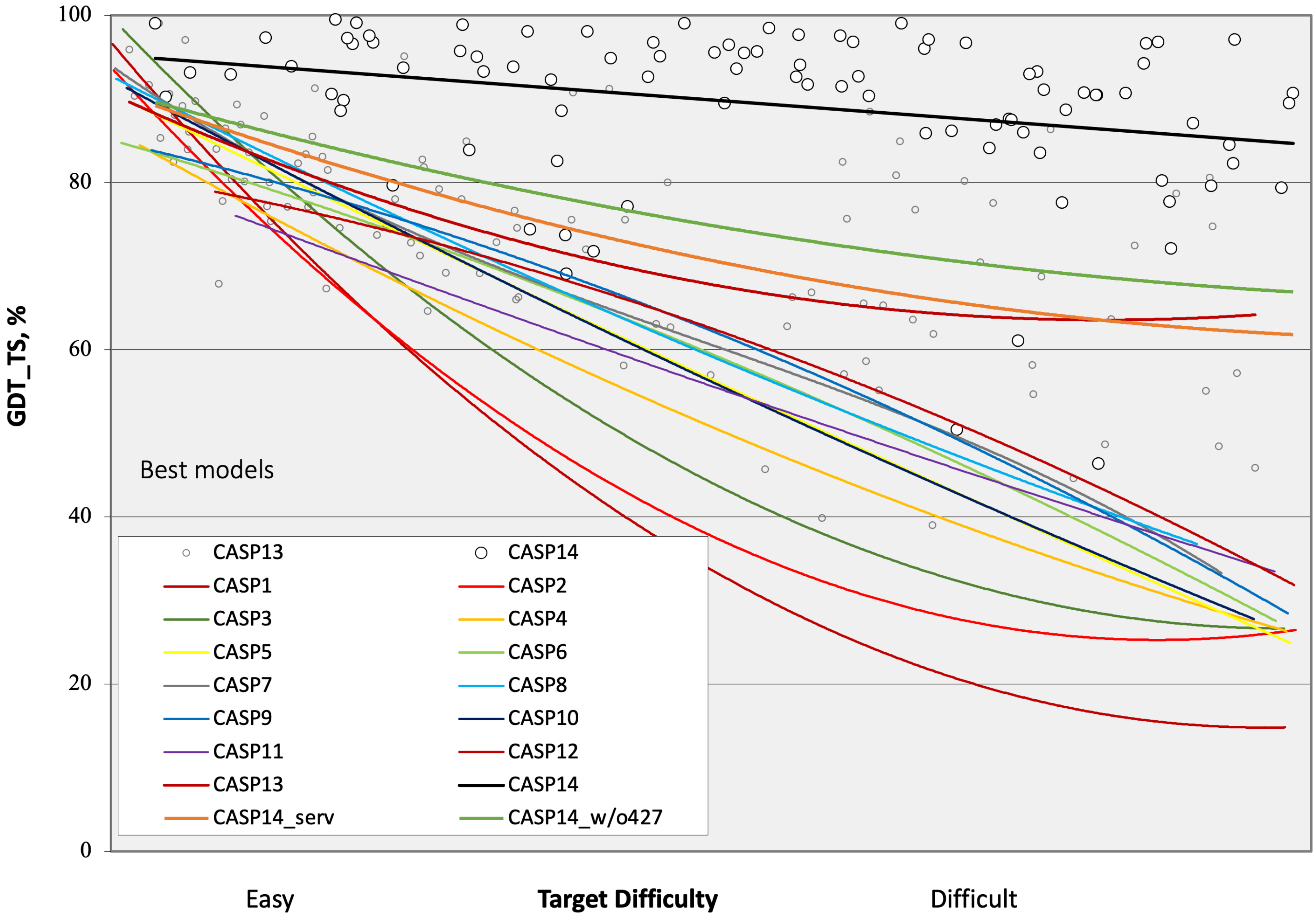

Long contacts, L/5 lists
Average Precision
70
60
50
40
30
20
10
0
CASP13
CASP12
CASP11
CASP10

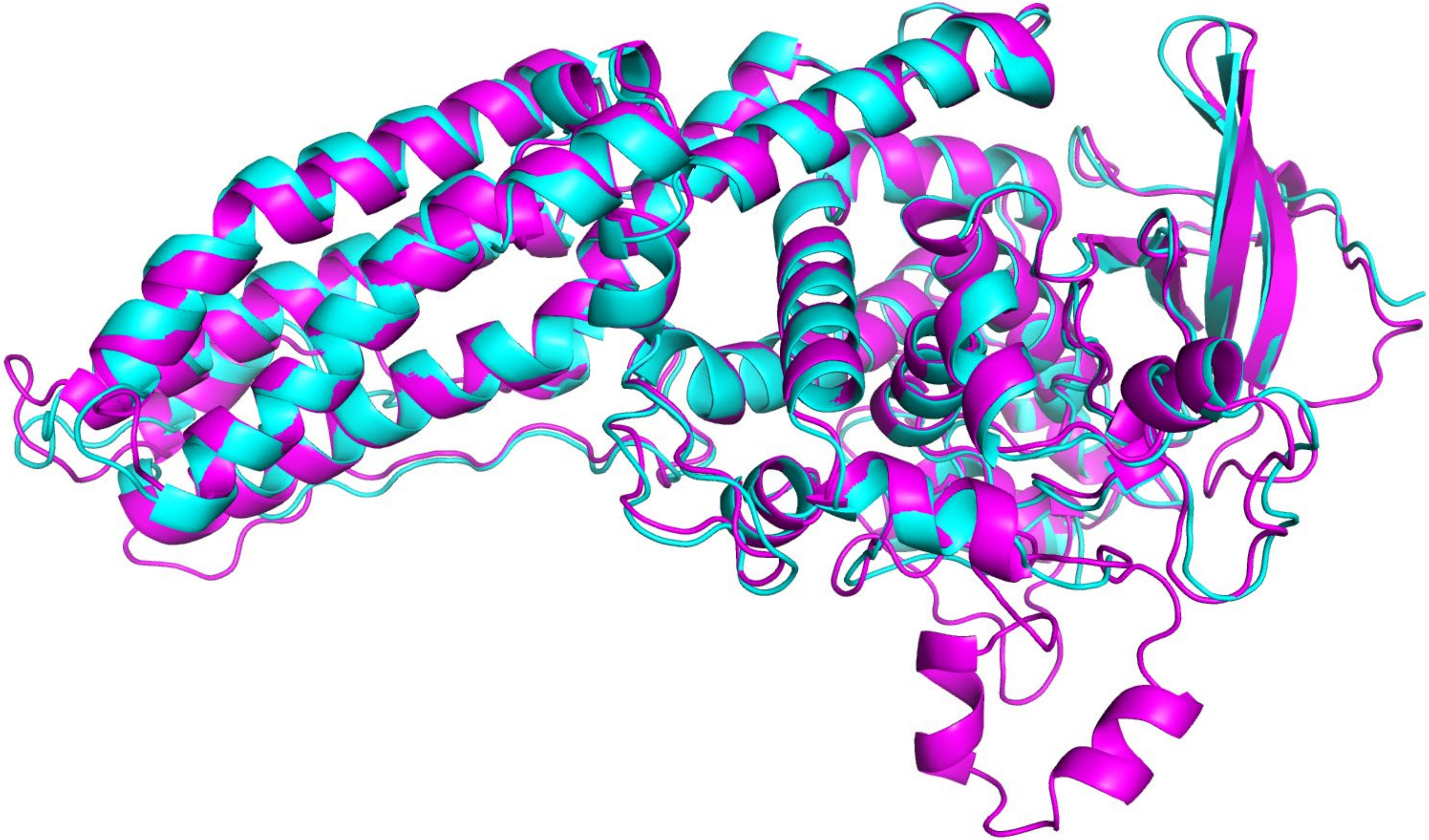

100
80
60
40
20
0
GDT_TS, %
Best models
Easy
Target Difficulty
Difficult

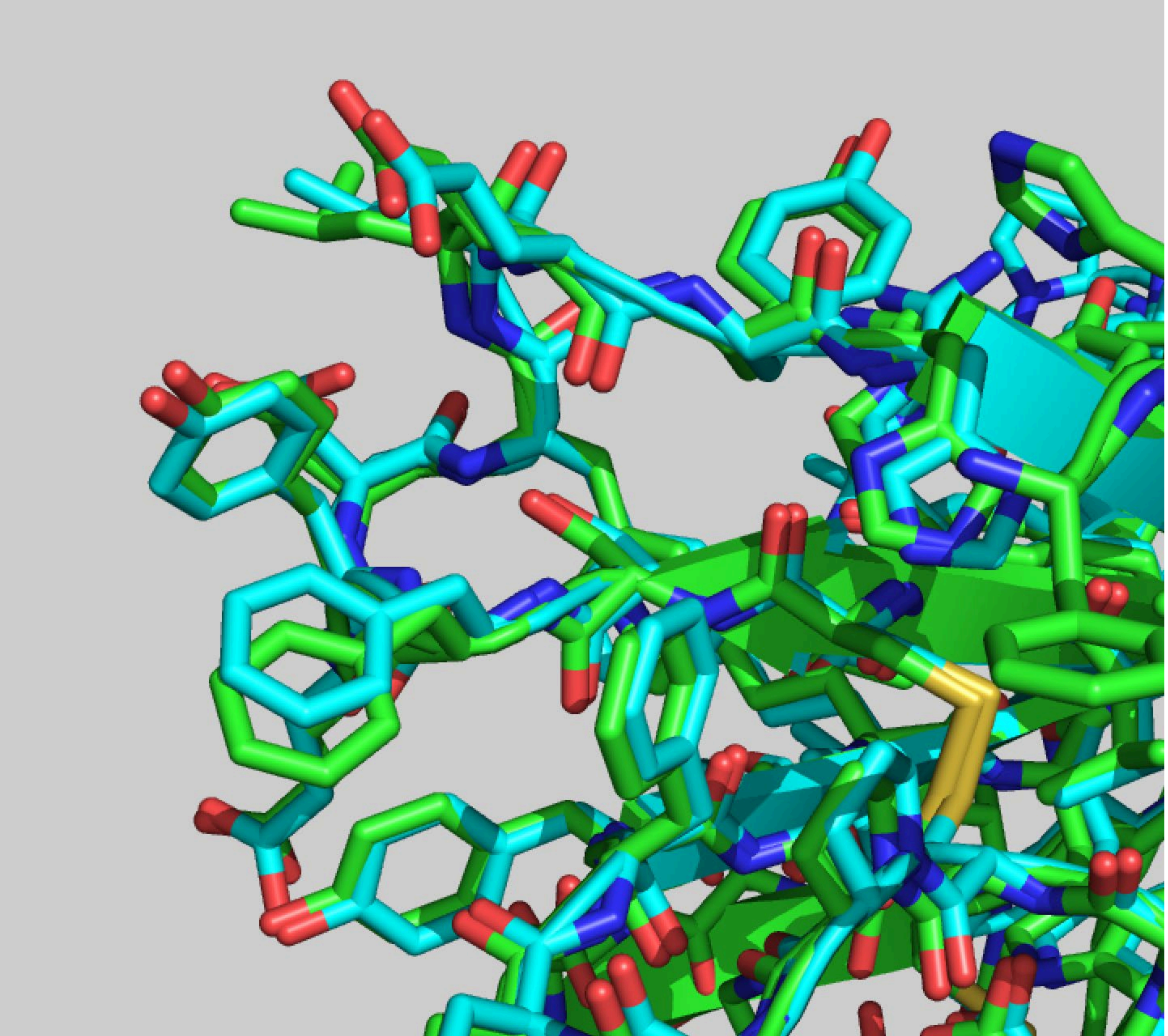


100
80
60
40
20
0
GDT_TS, %
Best models
CASP13
CASP1
CASP2
CASP3
CASP4
CASP5
CASP6
CASP7
CASP8
CASP9
CASP10
CASP11
CASP12
CASP13
CASP14_serv
Easy
Target Difficulty
Difficult


100
80
60
40
20
0
GDT_TS, %
Best models
CASP13
CASP1
CASP2
CASP3
CASP4
CASP5
CASP6
CASP7
CASP8
CASP9
CASP10
CASP11
CASP12
CASP13
CASP14_serv
CASP14_w/o427
Easy
Target Difficulty
Difficult


GDT_TS, %
100
80
60
40
20
0
Best models
CASP13
CASP14
CASP1
CASP2
CASP3
CASP4
CASP5
CASP6
CASP7
CASP8
CASP9
CASP10
CASP11
CASP12
CASP13
CASP14
CASP14_serv
CASP14_w/o427
Easy
Target Difficulty
Difficult





Not an end, but a Beginning:
The door is open to?:

- Protein complexes
- Accuracy estimation
- Protein design
- Protein dynamics
- Protein conformational change
- Preferred conformations of disordered proteins
- Mutation interpretation
- Ligand docking