CASP14 data and statistics

Andriy Kryshtafovych

Protein Structure Prediction Center
Genome Center
University of California, Davis

Website traffic
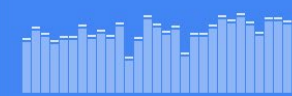
Predictors

Experimentalists
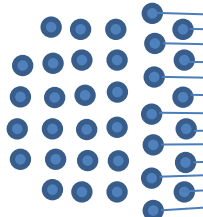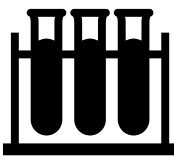
Center for CASP

Organizers

Assessors

CASP14 data at a glance
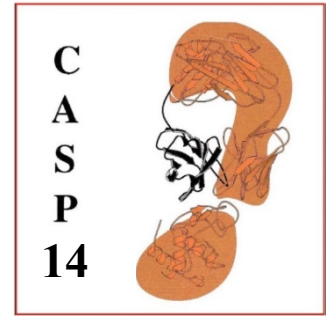
- >200 prediction methods

- ~100 research centers

- ~350 predictors

- >80 targets

- >67,000 models

- >5,000,000 scores

- ~430 GB of data

- >30 different software tools

- >20 visualization tools
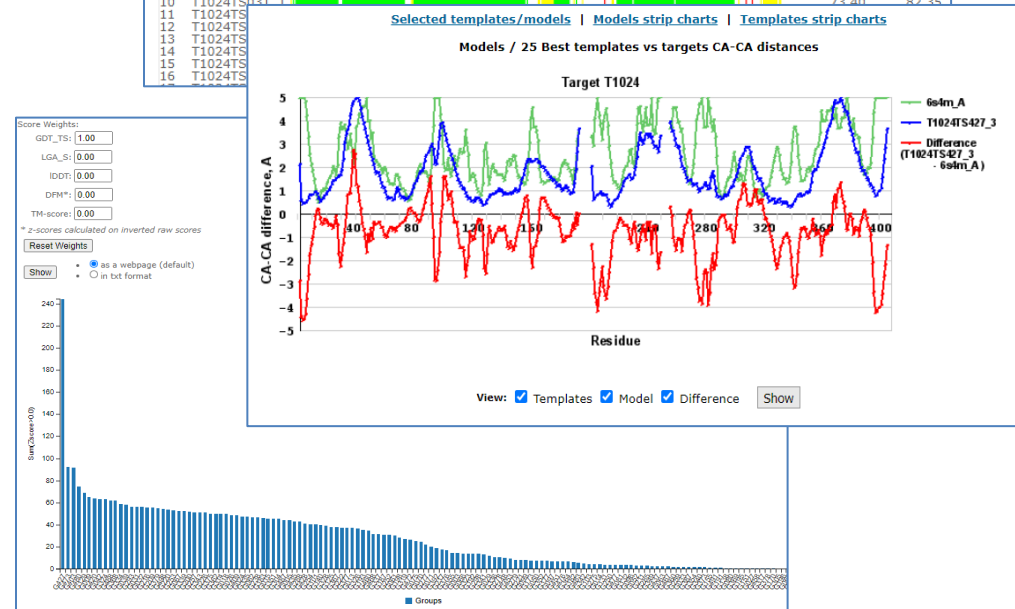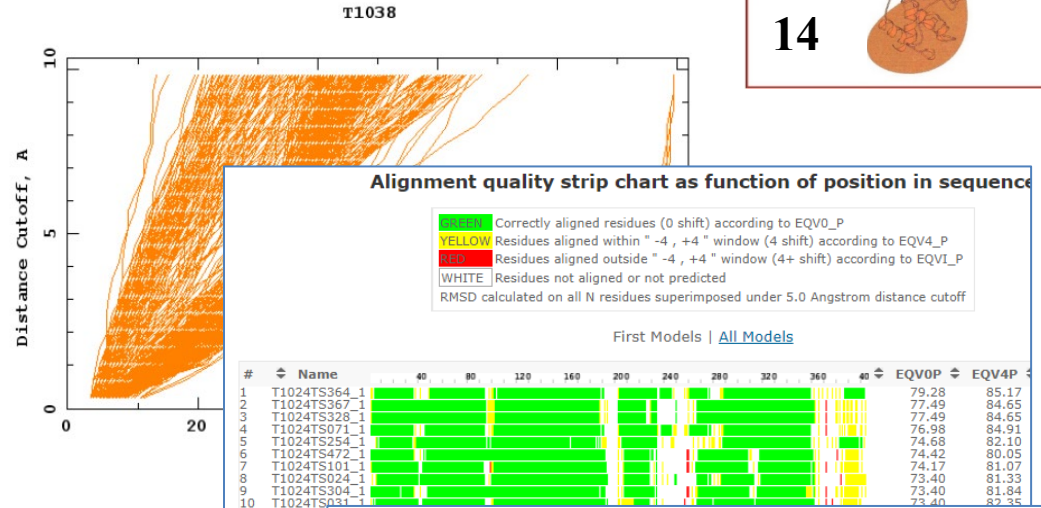
New elements of CASP14 system at a glance
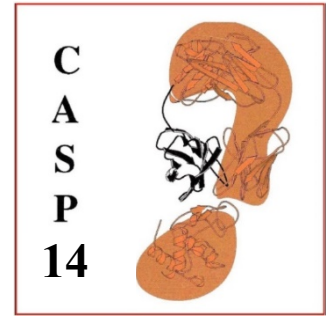
- new CASP-covid infrastructure
- new domain interactions system
- multimeric and cryo-EM analyses
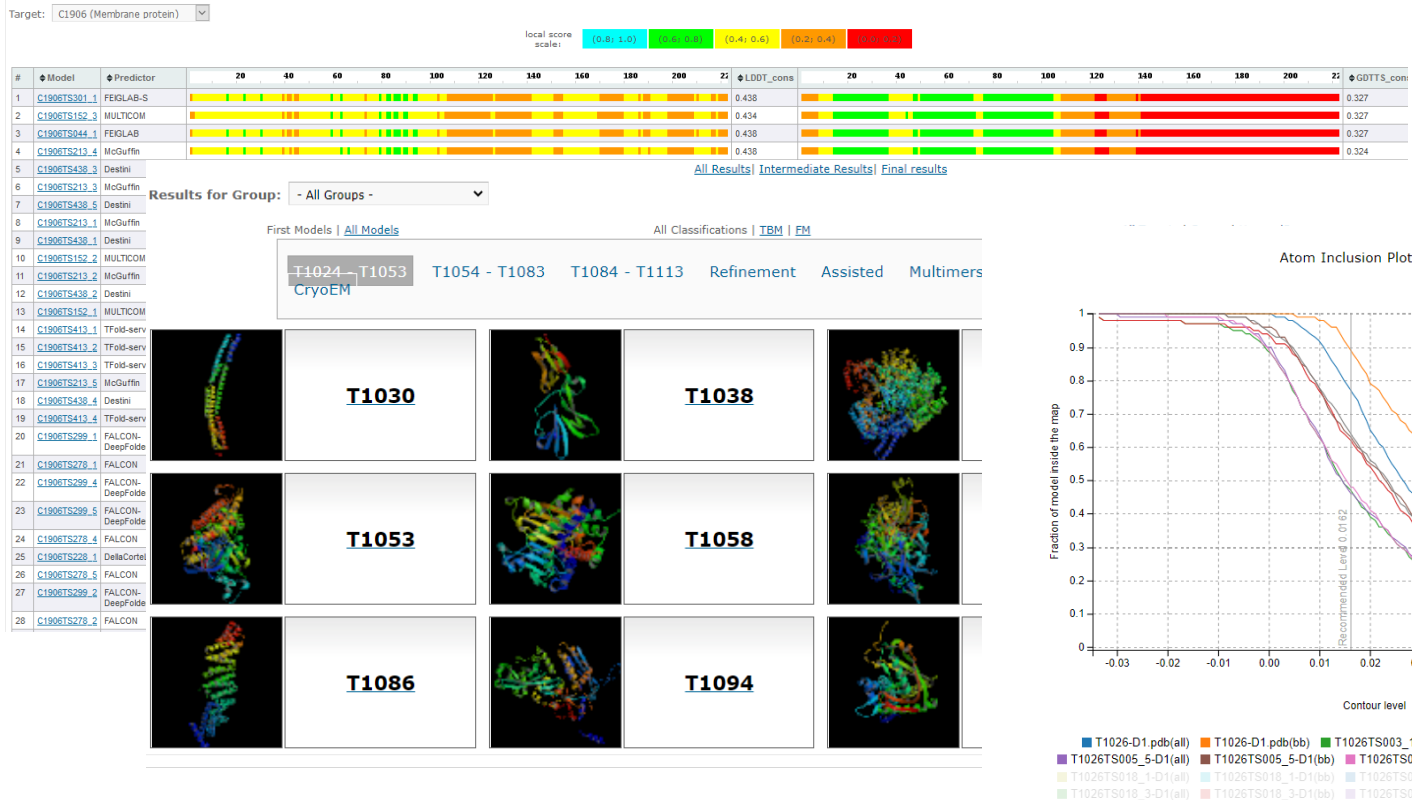
How would I remember CASP14
from the operational point of view?

- No hiccups with hardware – good for a change

- Moved to a secure data transfer protocol (https)

- Staffing changes: Bohdan Monastyrskyy moved on
  in his career, and his experience and
  dedication were hard to replenish.
  Thank you Богдан for all your help!

- It was Y2020: proceed or postpone?
  - target availability
  - willingness /ability of predictors to participate

	Go as scheduled	Slight delay	Significant Delay
David Jones	1		
David Baker			
Chaok Seok			
Michael Sternberg			
Shoshana Wodak			
David Shortle			
Ceslovas Venclovas			
Arne Elofsson			
Daisuke Kihara			
Alberto Perez			
Liam McGuffin			
John Jumper			
Konstantin Weibenow			
Russ Altman			
Jinfeng Zhang			
B. Jarayam (India)			
Mohammed AlQuraishi			
Roy Nassar (Laufer)			
Marcin Skwark		1	
Gianni De Fabritiis (UPF Barcelona)			
Hyung-Rae Kim (S. Korea)			
Georgy Derevyanko, Guillaume Lamoureux			
Firas Khatib			
Mayuko Takeda-Shitaka			
Paul Bates			
Chen Ceasar			
Jinbo Xu			
Yang Zhang			
Michael Feig			1
Total	13	17	13

In this crazy time having some things go on as before is good. (David Baker)

All of us in self-isolation need some successes to hope for. For me it is CASP. (David Shortle)

Can't see any reason to cancel. The experiment can and should go as planned. (David Jones)

It would be a shame to completely cancel CASP this year. (Arne Elofsson)

It just doesn't seem the right time and the right priority at this exact moment to pursue CASP. (Michael Feig)

I can't even claim that I'm again trying to delay CASP due to soccer. If CASP 14 were to go as scheduled, without as many targets or participation as in previous CASPs, then I fear it would just have an * next to it compared to previous years. (Firas Khatib)

Groups
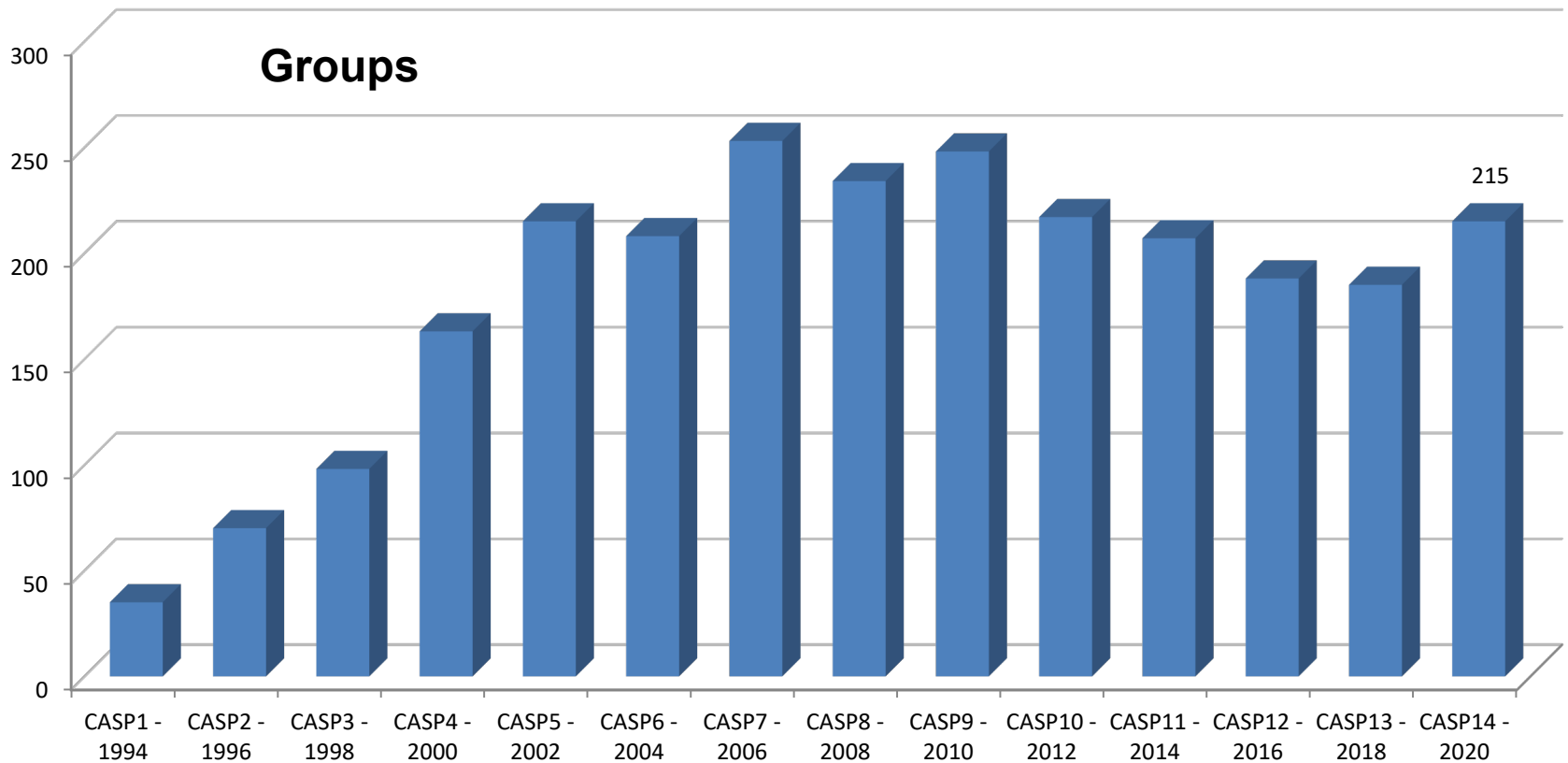
RESEARCH ARTICLE

Critical assessment of methods of protein structure prediction (CASP)—Round XII

John Moult[1] | Krzysztof Fidelis[2] | Andriy Kryshtafovych[2] |
Torsten Schwede[3] | Anna Tramontano[4]

A key aspect of the experiments is that independent assessors are asked to interpret the results. Assessors are encouraged to base their analysis on the established CASP measures and also to develop additional measures they consider appropriate.

The CASP12 prediction period was from May till August 2016. A planning meeting was held in October, at which the assessors presented their findings to each other and to the organizers. After the assessors had reported their conclusions, group identities were revealed and the most successful groups as well as those with the most promising novel methods were invited to talk at the CASP conference. The conference was held in Gaeta, Italy, in December 2016. The program of the CASP12 meeting can be found at http://predictioncenter.org/casp12/doc/CASP12_Meeting_Program.html. Many of the conference presentations as well as all results are also available on the web site.

1.2 | CASP12 statistics: precipitating groups, targets, and submissions

CASP12 maintained the high participation level of recent CASPs with 188 methods from 96 research groups in 19 countries taking part. The number of methods decreased slightly from the 207 of CASP11, primarily as a result of the elimination of the disorder prediction category and limiting the number of methods from the same research group to five.

In between CASP rounds, the CAMEO project complements the experiment by providing an automated continuous benchmarking platform for developers of server methods, using the weekly PDB prerelease information to identity targets. Several of the leading groups tested and benchmarked their new methods in preparation for CASP12. New CAMEO categories currently in implementation are continuous assessment of complexes (homo- and hetero oligomeric), residue-residue contact prediction, and ligand conformation in 3D structure modelling[5] (cameo3d.org).

Almost 55 thousand models were submitted in CASP12, of which 37,672 were three-dimensional coordinate sets. The remaining submissions were for refinement (6,227), estimation of model accuracy (7,400), residue–residue contacts (3,077), and data-assisted predictions (528).

1.3 | Management and organization

The CASP12 organizers were unchanged from CASP11 and are the authors of this article. They are responsible for all aspects of the experiment. There is an advisory board composed of senior members of the modeling community. A participants' meeting during each CASP conference allows for more direct interaction, including votes on issues of CASP policy. The Protein Structure Prediction Center is responsible for the experiment data management, including the distribution of target information, collection of predictions, generation of numerical evaluation data, developing tools for data analysis, data security, and maintenance of
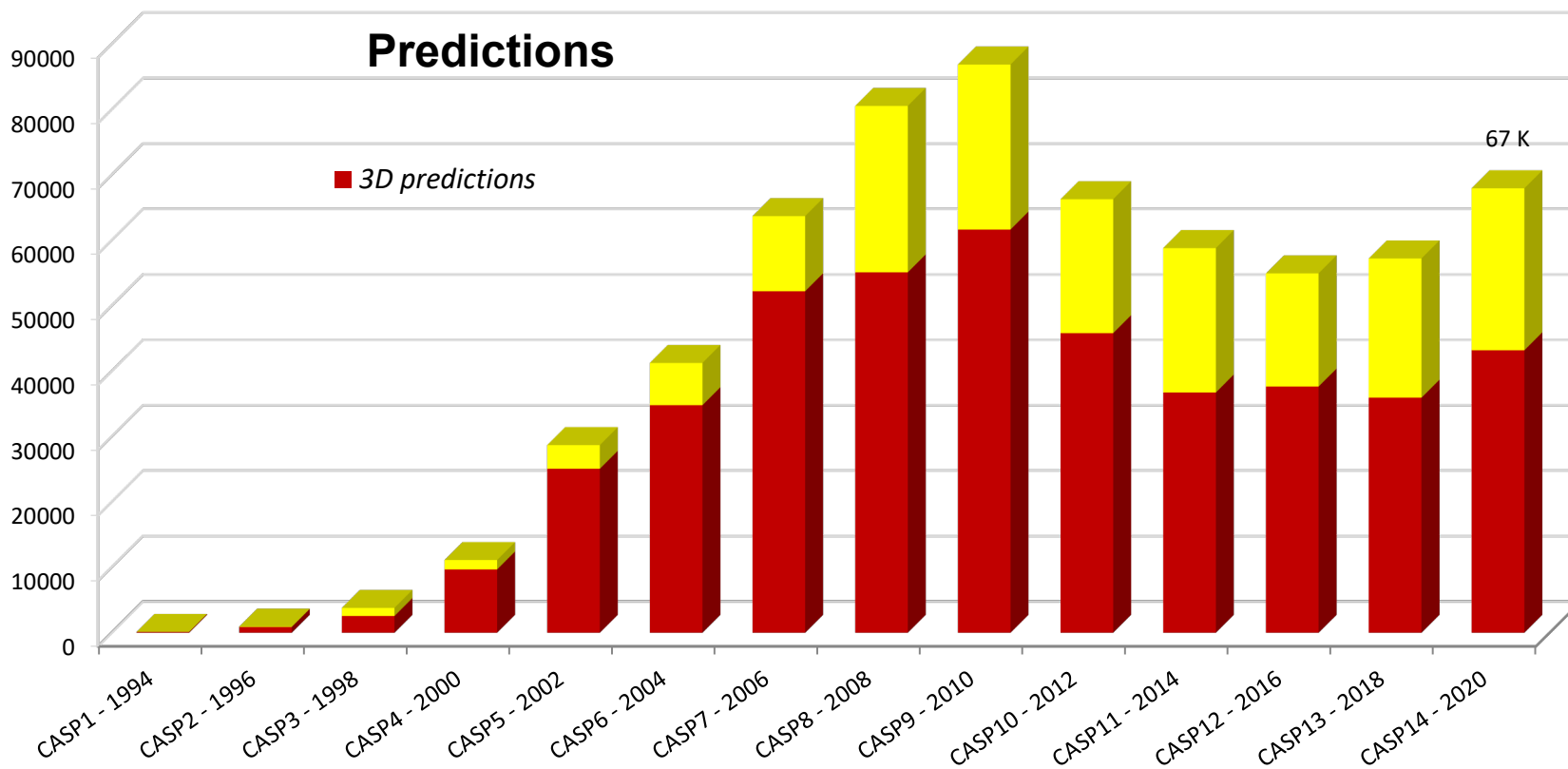

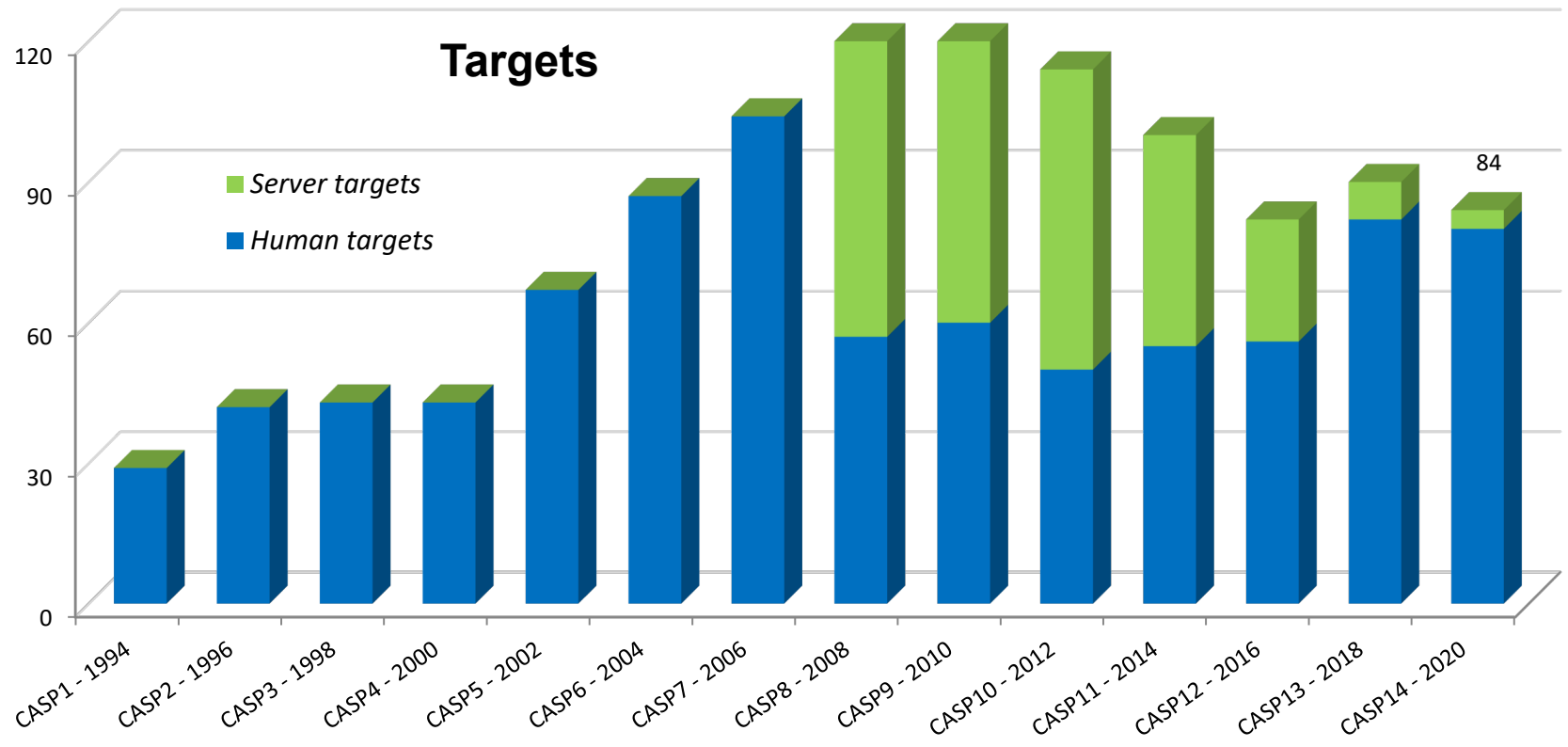Number of 'precipitating' groups
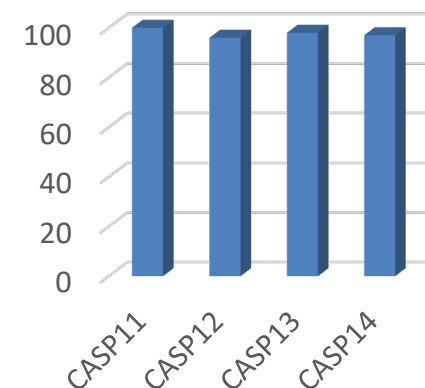
CASP14 predictors geography

19 countries

Predictions

Regular targets

THANKS:
39 structure determination groups from 15 countries

Basle, Arnaud
Davies, Owen
Lea, Susan
Lovering, Andrew

Herzberg, Osnat
Michalska, Karolina
Joachimiak Andrzej
Hunt, John
Montelione, Gaetano

Alahuhta, Marcus

Chiu, Wah
Oliver, Stefan
Flower, TG
Tsutakawa, Susan
LiWang, Andy
Rees, Steven

Krieger, Inna
Leiman, Petr
Sthanam, Narayana
Tagliabracci, Vincent
Miller, Mitch
George Phillips

Chiarini, Valerio
Knight, Stefan

Boettcher, Bettina
Lupas, Andrei;
Hartmann, Marcus
Tidow, Henning
Perrakis, Anastassis
Ilari, Andrea

Kuroda, Yutaka

Govaerts, Cedric
Bersch, Beate
Desfosses, Ambroise
Linares, Romain
Van Raaij, Mark
Dunne, Matthew

Dessau, Moshe

Adav, Rajnesh Kumari
Gupta, Gagan

Birkinshaw, Richard
Grinter, Rhys
Peat, Tom

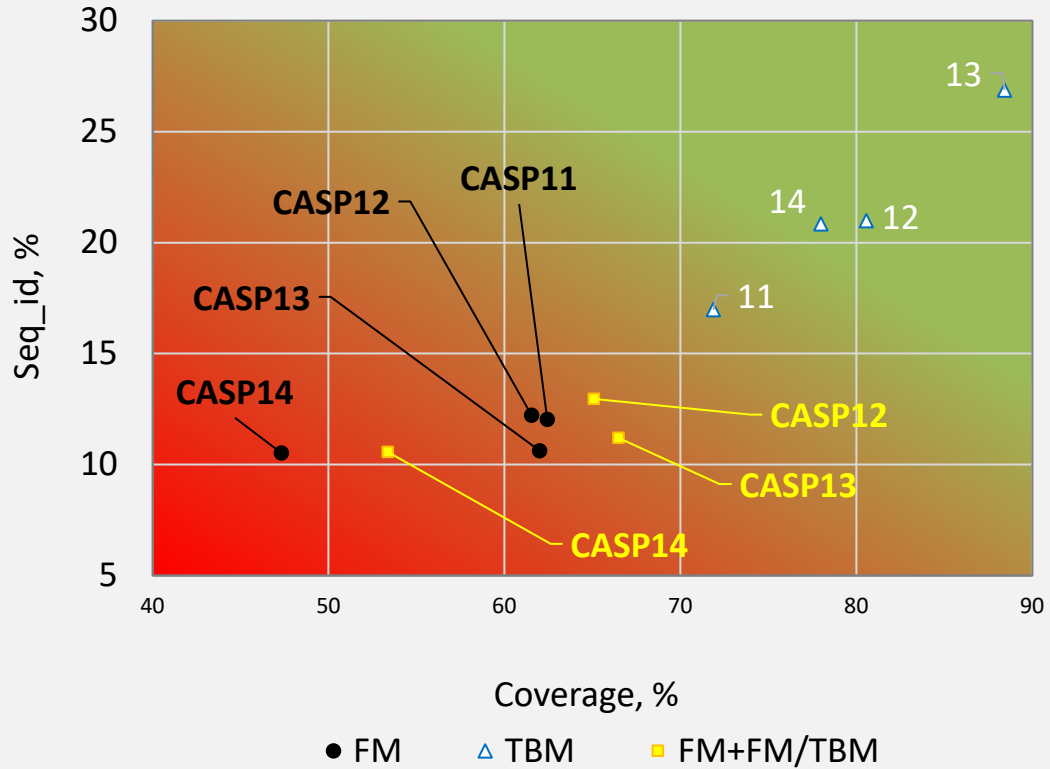| Target type	# Targets CASP14	# Targets CASP13
Regular	84 (68 evaluated, 96 EU + 11 DD) 5: canceled by assessors 9: no structure 2: separate evaluation (NMR)	90 (80 evaluated, 111 EU)
Multimeric	30 (25 SU assessed) = (10 hetero) + (20 homo)	42 = (12 hetero) + (30 homo)
Refinement	49 = 28 one-start + 7 double-barreled (14 total) + 7 extended	29
Assisted	3 = 2 NMR (separate eval.) + 1 SAXS (no structure)	40 = 11 SAXS + 12 Xlink + 15 NMR + 1 SANS + 1 FRET
Cryo-EM	7 (+ 5 domains)	7

Domain definition and classification

Andriy Kryshtafovych

- Pre-processed targets as soon as structures become available.
- Run domain boundary definition programs (DDomain and DomainParser2).
- Compare results of homology search programs (PSIBLAST, HHsearch) with #2.
- Suggested preliminary domain definition based on #2, #3 and visual inspection.
- Run evaluation of models and template search for the suggested domains.
- Suggested composition of evaluation units (EUs) based on the domain-based evaluation results (Grishin plots) and, if needed, rerun evaluation on the adjusted EUs
- Preliminary classified domains in 3 difficulty categories (TBM, TBM/FM, FM) based on the scores of top 20 server models (#1) and homology searches

Lisa Kinch

- Analyzed sequence and structural templates (PSIBLAST, HHsearch, LGA, ECOD).
- Suggested alternative domain definitions, if needed.
- Suggested target categorization based on template quality, server performance and clustering of domains.
- Looked into all borderline cases (TBM/FM).
- Prepared visual material (graphs, ribbon diagrams, structure superpositions) for the discussion.
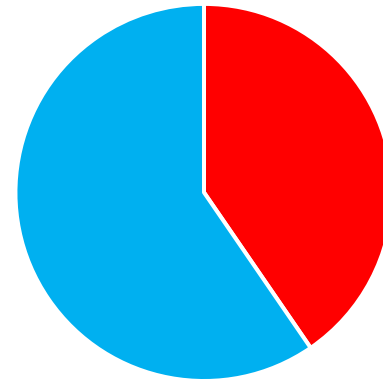
Average target difficulty
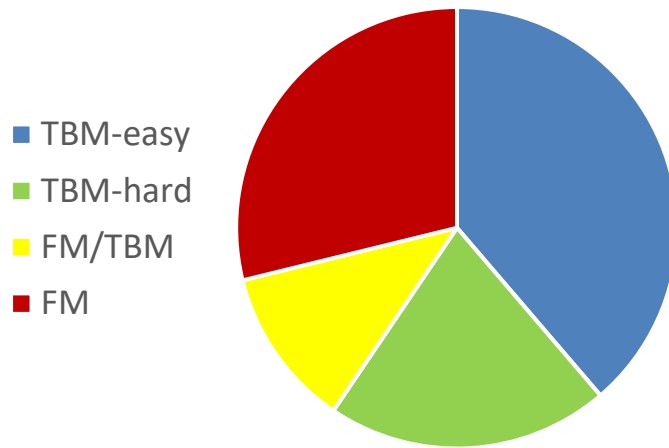
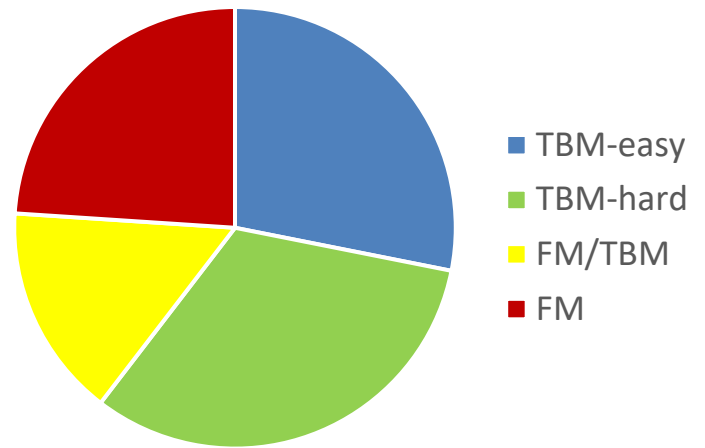# Availability of sequence relatives (Neff)

CASP13
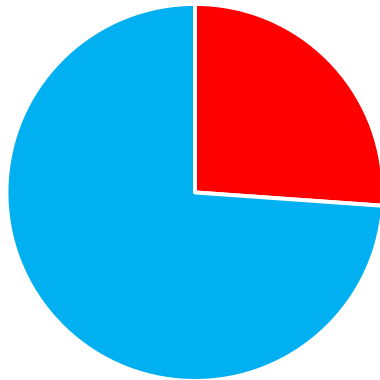
CASP14

■ Neff<1   ■ Neff>1

■ Neff<1   ■ Neff>1

Domain definition and classification

<u>Take home messages:</u>

- CASP14 targets were harder than those from previous CASPs

- We will not have an * next to CASP14 due to lower levels of participation or poorer target set, but due to other reasons