

CASP9

TBM Assessment

Valerio Mariani
Torsten Schwede

SIB – Swiss Institute of Bioinformatics &
Biozentrum University of Basel
Torsten.Schwede@unibas.ch

Looking back on CASP TBM rounds 6, 7, and 8 ...

Looking back: CASP 6 – 7 – 8 ...

Assessment with increasing focus on molecular details of the final models:

- CASP6: C α measures (GDT-TS / AL0) [1]
- CASP7: HB-Score (Backbone and Side chain H-bond accuracy) and GDT-HA (0.5, 1.0, 2.0, 4.0 Å cut-off) [2]
- CASP8: Molprobity Score, Refinement [3]

[1] Tress et al., Proteins. 2005;61, 27-45.

[2] Kopp et al., Proteins. 2007;69, 38-56; Battey et al., Proteins. 2007;69 Suppl 8:68-82.

[3] Proteins. 2009;77, 29-49.; MacCallum et al., Proteins. 2009; 77, 66-80.

Looking back: CASP 6 – 7 – 8 ...

But, some atomistic scores also have their draw-backs:

- Some methods are complex black boxes with unclear criteria ("Molprobability score").
- Some have free parameters and cut-off definitions ("What exactly is an H-bond?")
- Some are topology dependent (α -helical structures score differently than topologies dominated by β -sheets, etc.)

Increasing emphasis on ranking – rather than on the individual methods to do better predictions.



Looking at scores is ok, but ...
.... we want to watch the match!

9th Critical Assessment of Techniques for Protein Structure Prediction

1. Emphasize interesting techniques –
not ranking.
2. Be critical - but realistic.

It's all about protein structures.
3. Keep it simple.

CASP9 TBM in Numbers

- 102 targets with at least one TBM domain
- 121 TBM assessment units
 - 55 "Human + Server"
 - 66 "Server Only"
- 176 prediction groups
 - 79 servers (2 AL)
 - 97 human groups
- 61'665 TS predictions
 - 14'659 assigned as #1

TBM Assessment Workflow

- Some visual inspection of models for indentifying problems and highlights.
- All TBM assessment was done numerically.
- 121 "Server" TBM targets (all):
 - Predictions **by server groups** are assessed numerically (i.e. human groups are not included in the server ranking).
- 55 "Human+Server" TBM targets (subset):
 - Predictions by all groups are assessed numerically (i.e. both human and server groups are included in the ranking).

Be critical - but realistic:

Physically impossible models

Template based models can be expected to have a correct chemical topology and be physically realistic / plausible.

- We checked for clashes / compressions in predictions ^[1]:
Clash if $d_{A,B} < \text{contact distance observed in PDB} + 0.5 \text{ \AA}$
- We also used WhatIf to identify physically unrealistic predictions ^[2].

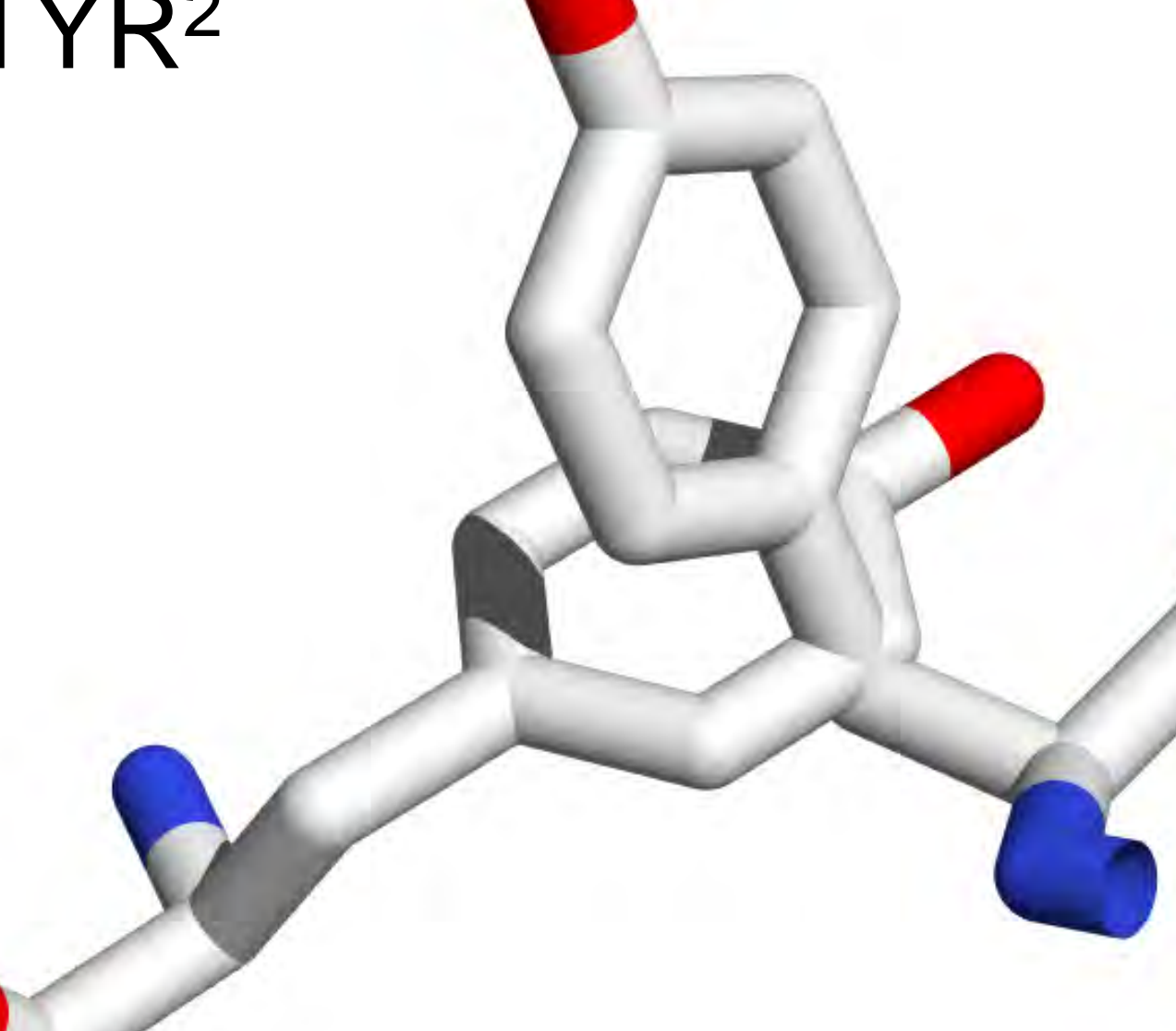
Problem:

Some errors are minor (fusion boundaries, side chain rotamers) - some are fundamental (overlapping backbones).

[1] Biasini et al., Bioinformatics. 2010, 26:2626-2628; <http://OpenStructure.org>

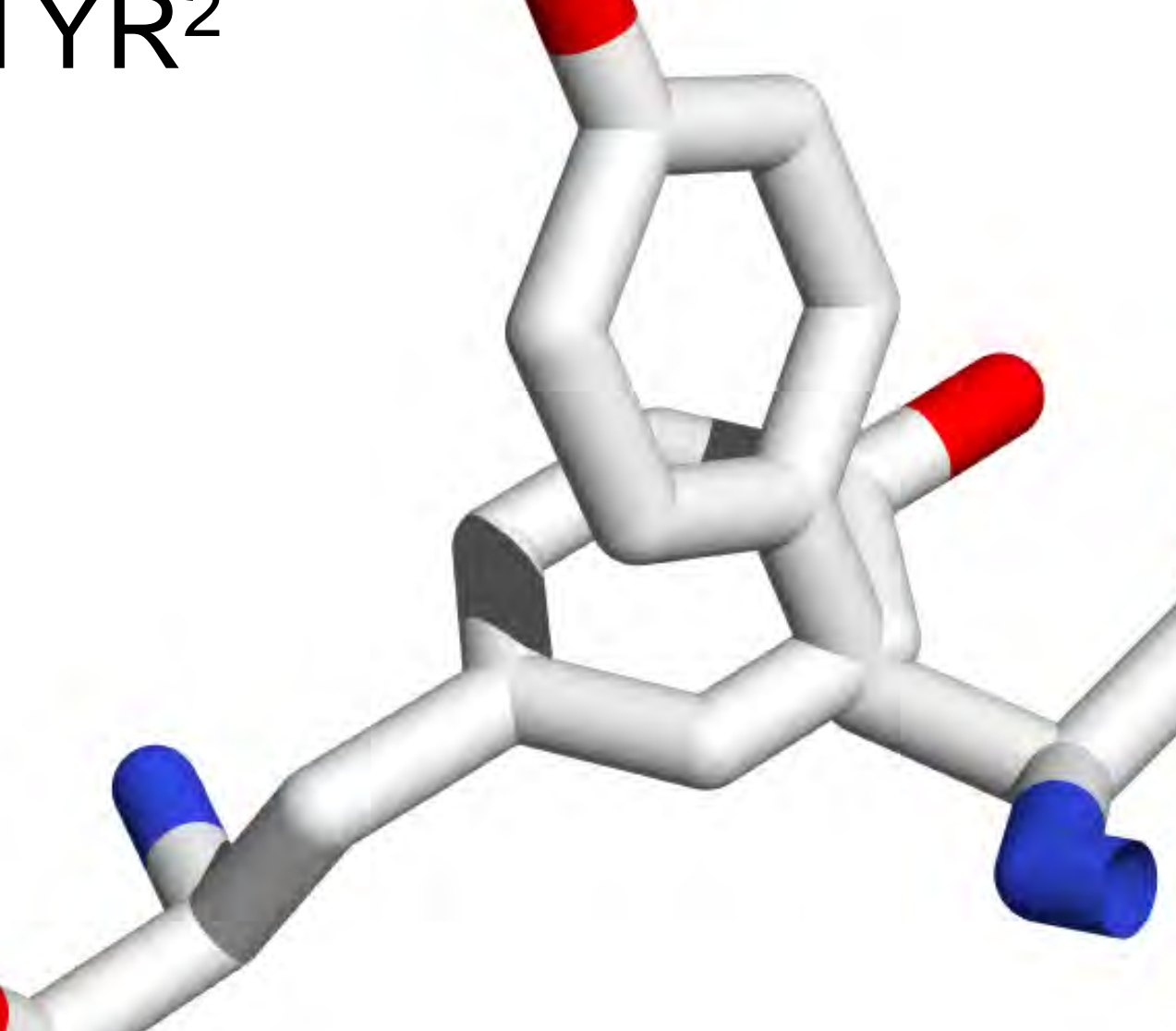
[2] Hooft et al., Nature. 1996, 381:272.

TYR²



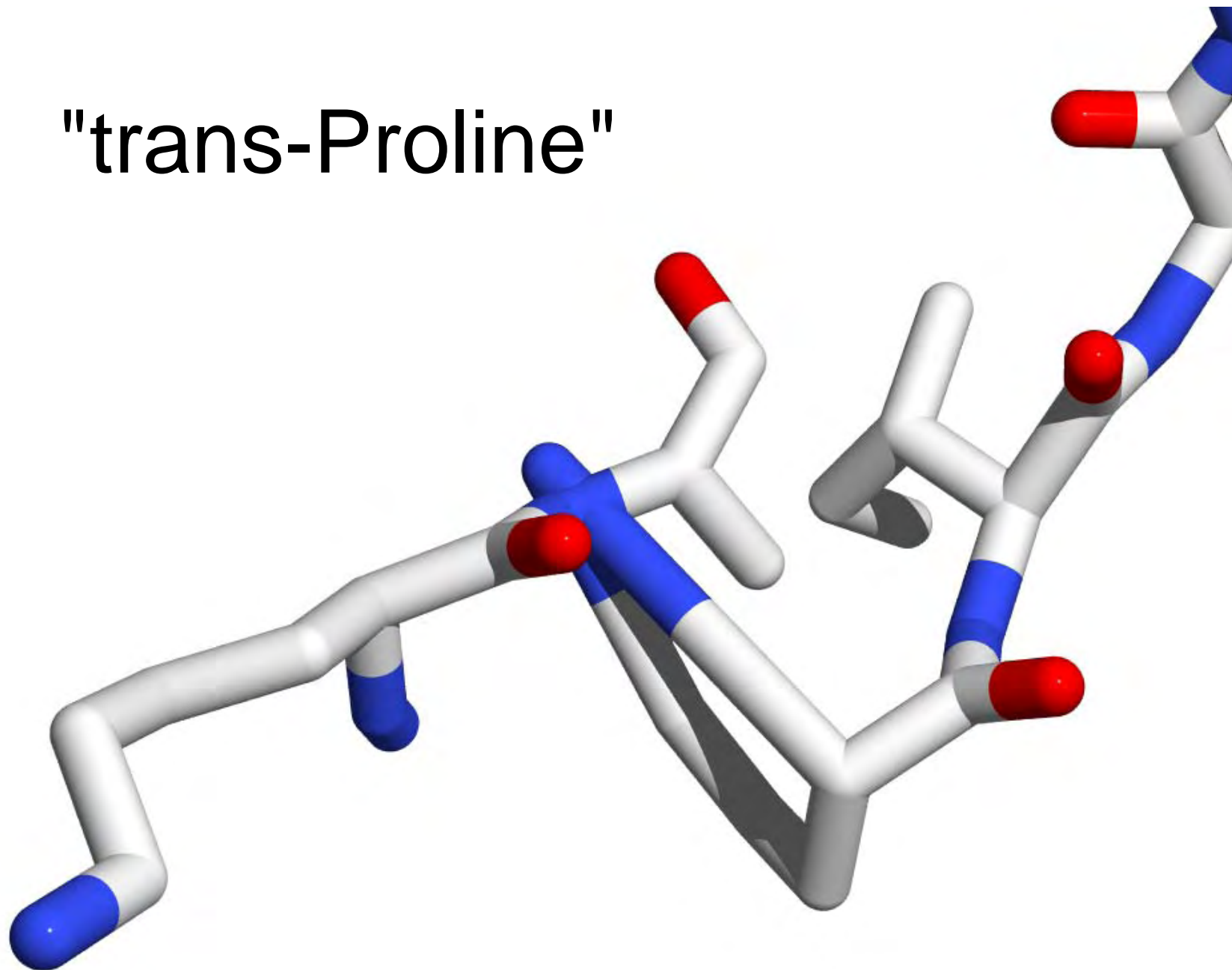
T0550

TYR²



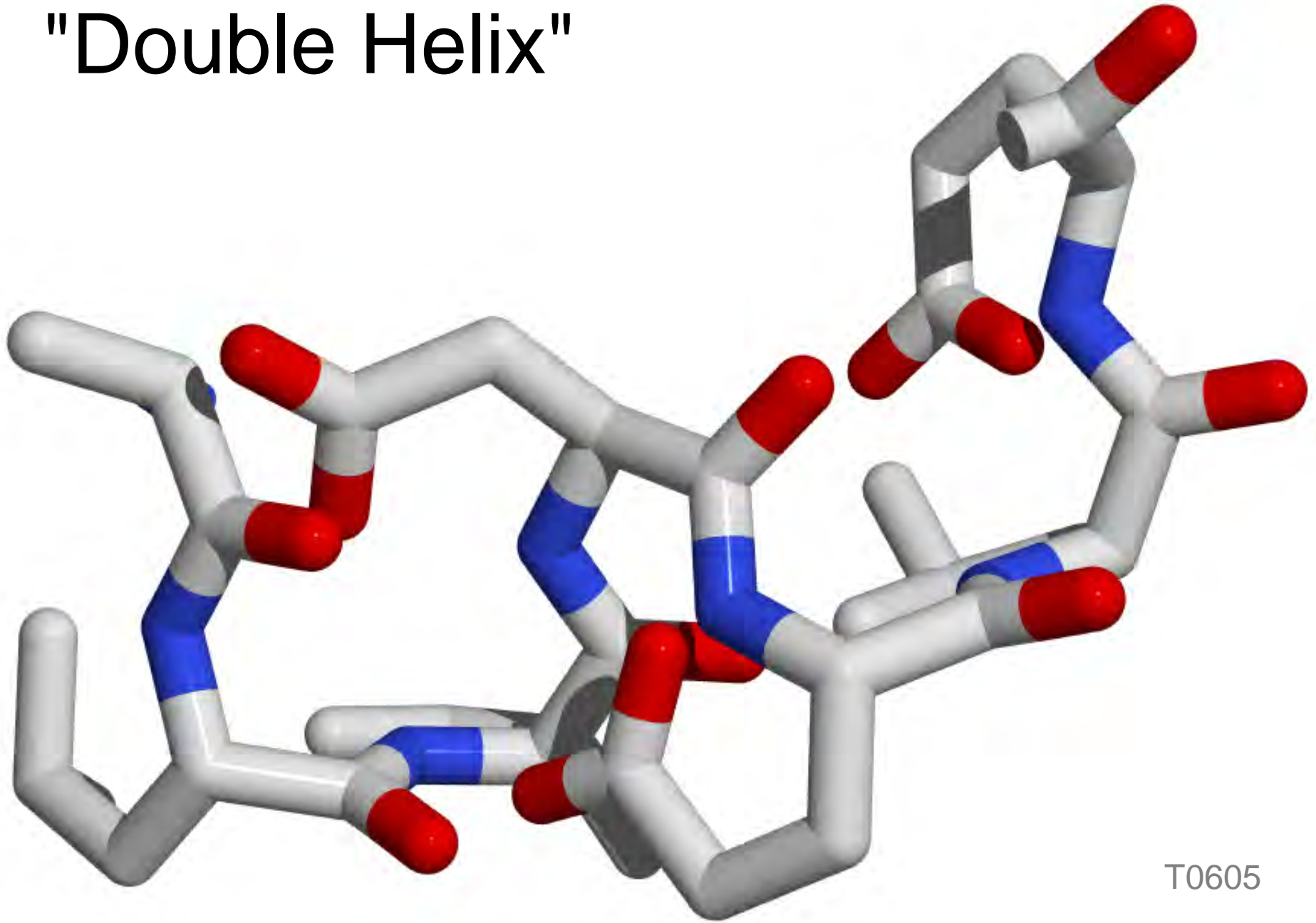
T0550

"trans-Proline"



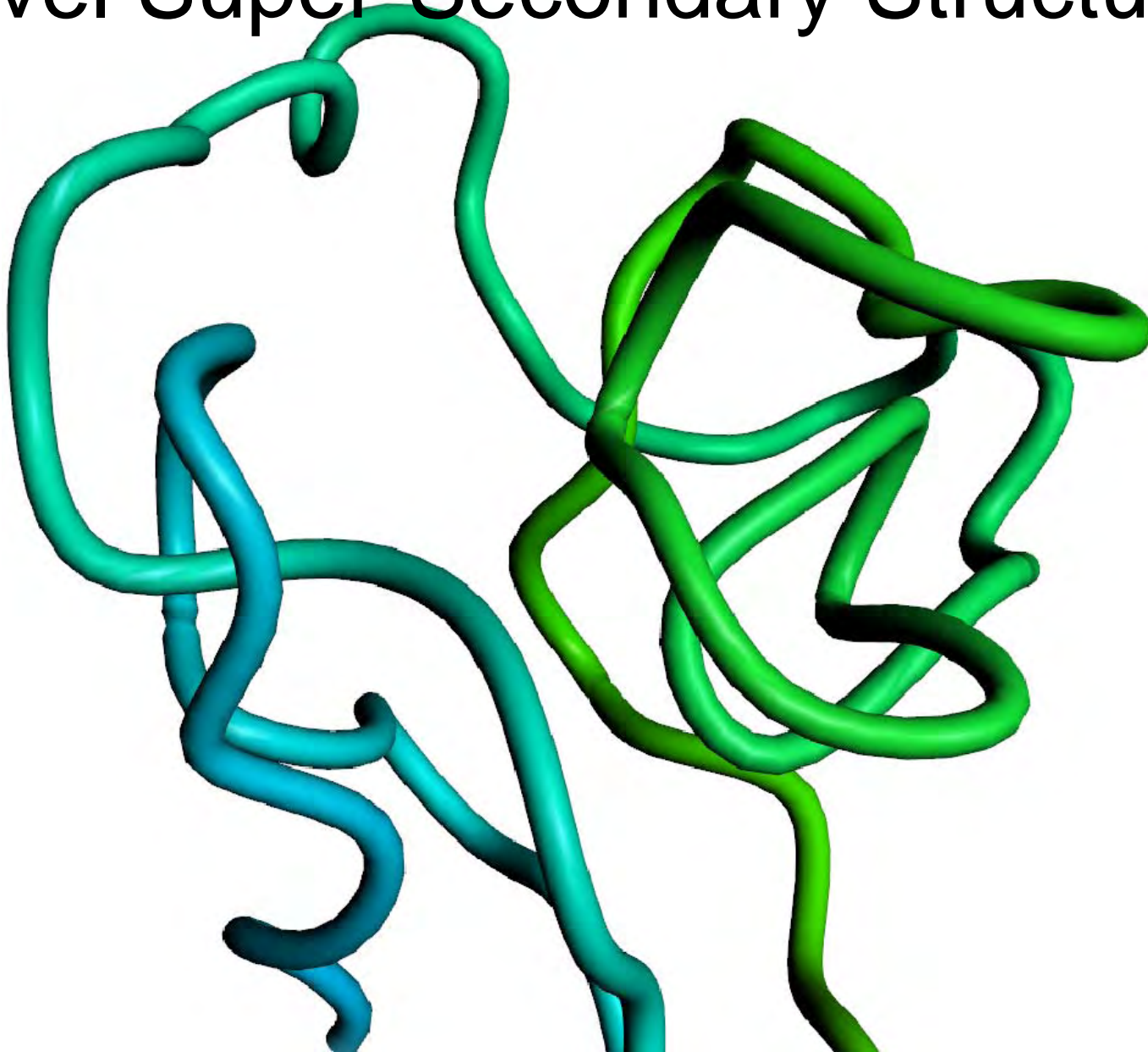
T0550

"Double Helix"



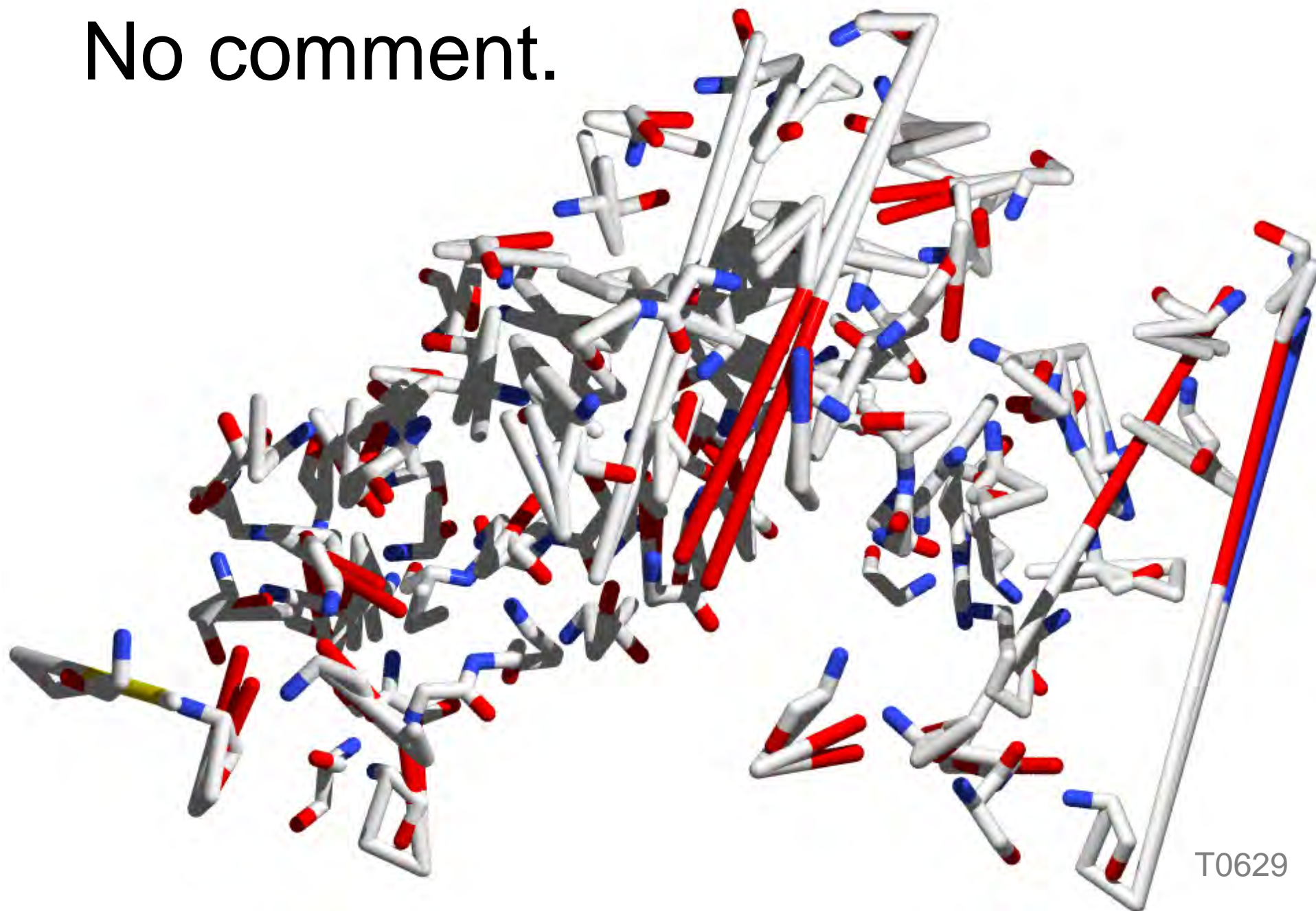
T0605

Novel Super-Secondary Structures

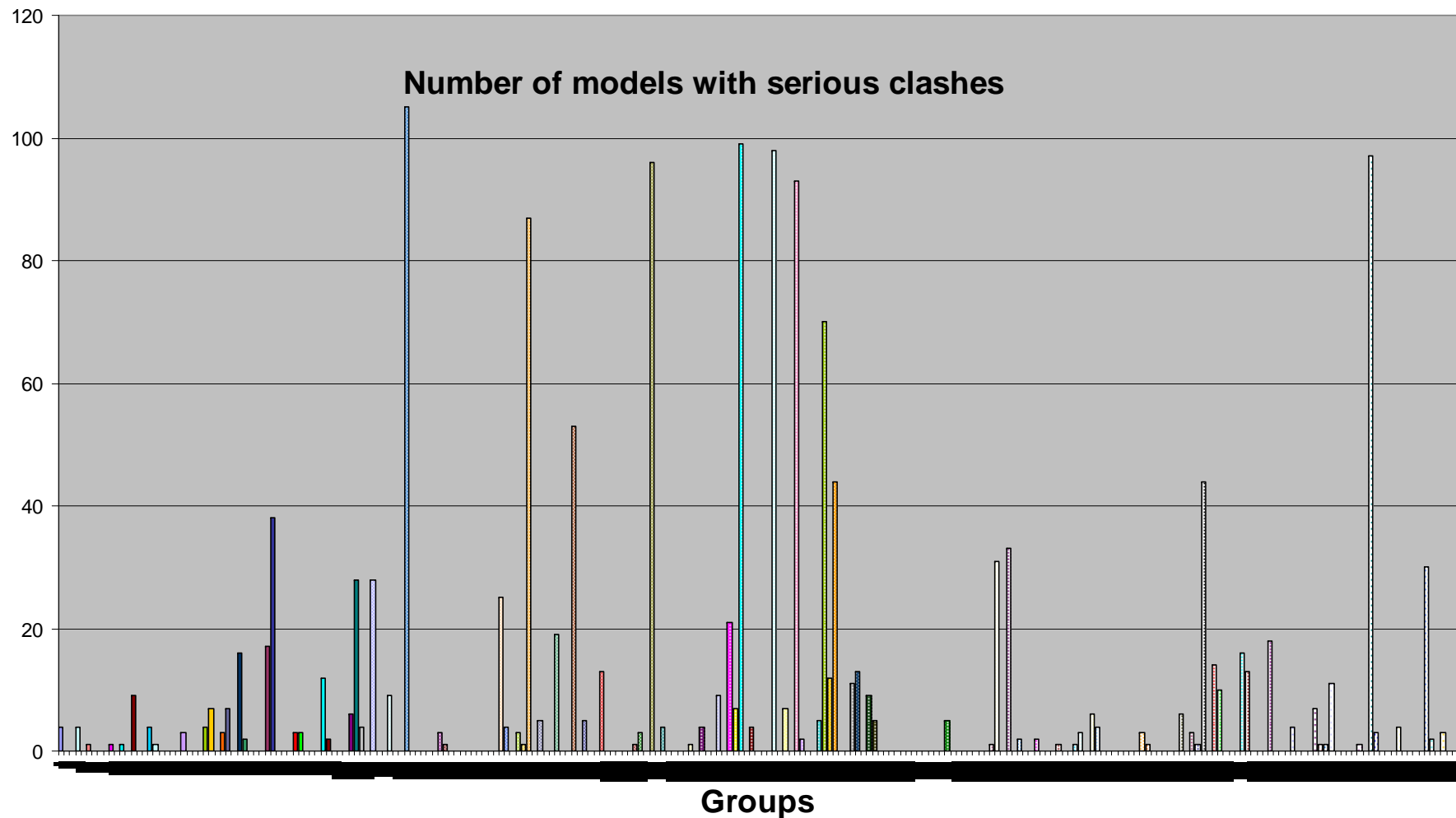


T0629

No comment.



Most models with serious clashes are produced by a small number of methods:

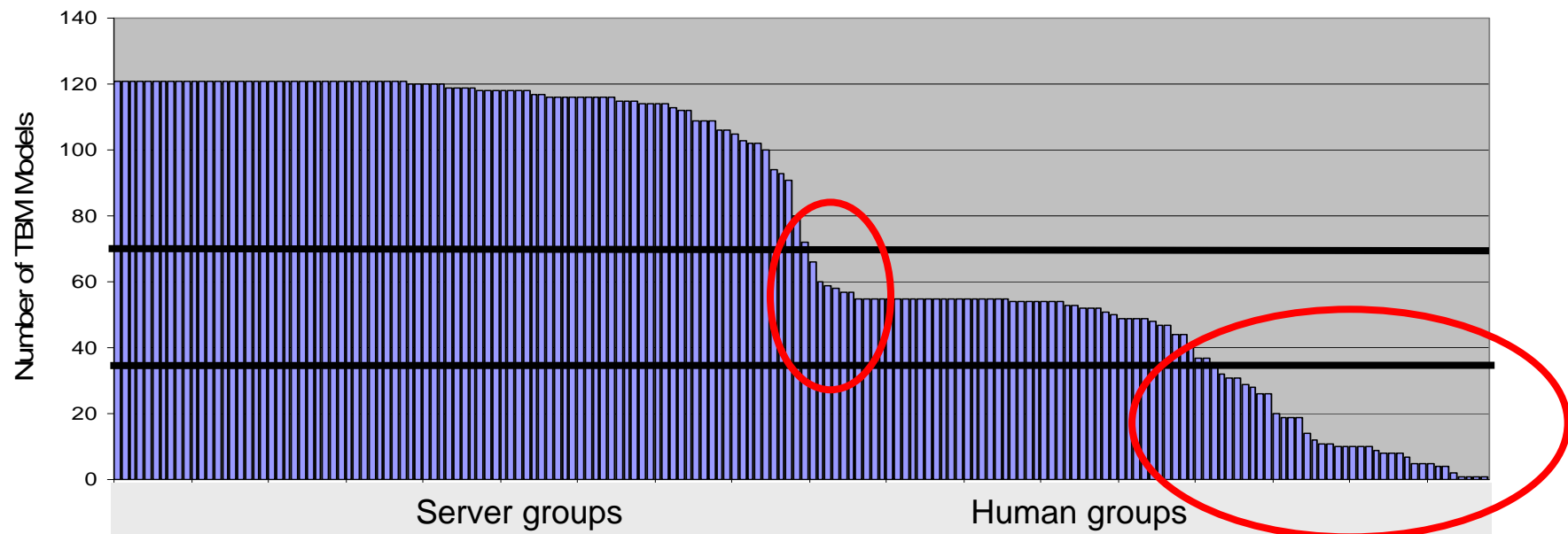


Physically impossible models

- Models with more than 5% of all residues involved in physically impossible interactions were weighted down to average (i.e. assigned a Z-score of 0).
- Chemically implausible residues, i.e. residues involved in unphysical interactions, were excluded in local all-atom scores resulting in lower scores for these models.

CASP "Light"

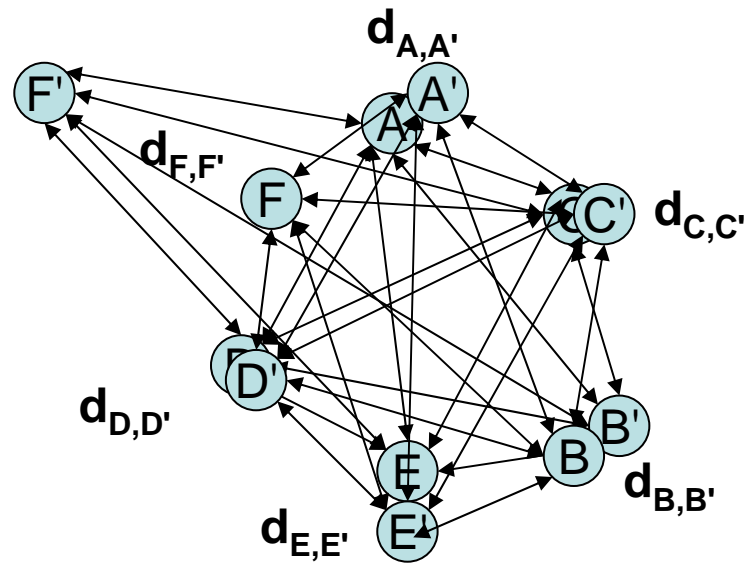
- Some groups only predicted a small number of easy targets – not surprisingly with quite good results. We need a direct comparison on common targets to draw significant conclusions on the ranking.
- We require **at least 80** predictions for server targets and **40** predictions for human targets to be included in the final ranking.



Exceptions

- "Server groups" with less than 80 predictions, or "human groups" with less than 40 predictions for human targets were assessed, but not included in the final ranking.
- We excluded groups 453 (HHpredA) and 346 (HHpredC) from the assessment since predictions were too similar to 449 HHpredB.
- We excluded target T0562 from the assessment for the Asilomar meeting due to a technical problem with data handling at PredictionCenter.

Global-Distance-Tests



Using many different global super-positions, we compute the fraction of residues in the model deviating from the target by not more than a specified distance cutoff .

- GDT: $C\alpha$ based measure
- GDC: all-atom measure

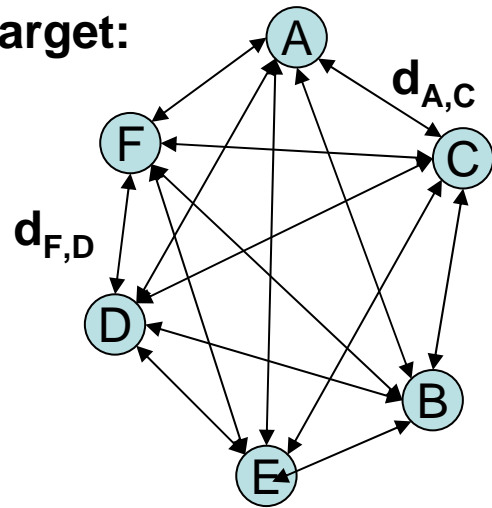
Which scores to use in CASP9 TBM assessment?

Global vs. Local Scores

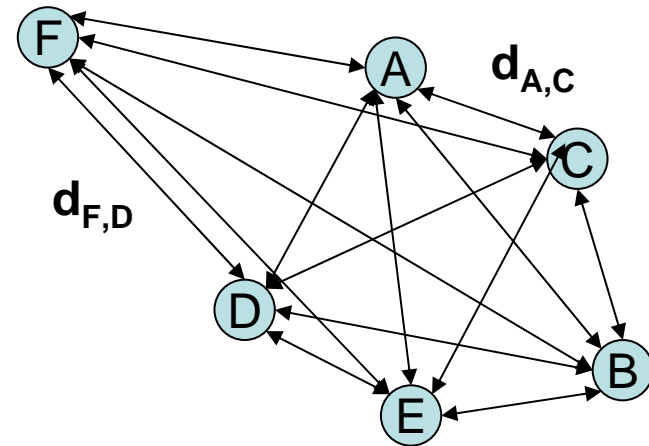
- Global scores (RMSD, GDT, etc.) capture the overall spatial arrangement of a structure well.
- Local scores capture the correctness of packing, e.g. side chain packing, active sites, packing of secondary structures and β -sheet topology well.

Distance-Difference-Test (DDT)

Target:

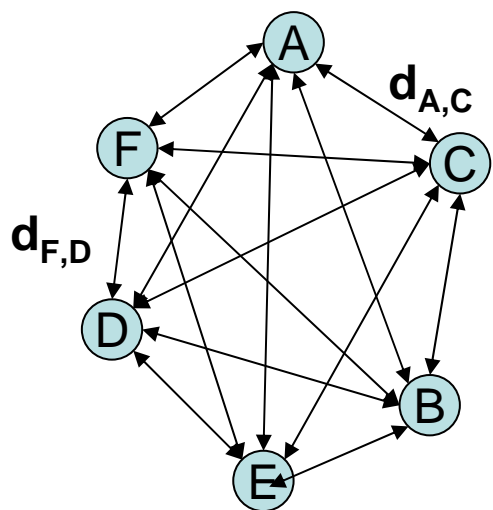


Prediction:



Compute the ***fraction of correctly predicted inter-atomic distances*** for each atom to its neighbors within certain error thresholds. "dRMSD-version" of GDT ...

Distance-Difference-Test (DDT)



$$DDT = \frac{\sum_{cutoff}^{0.5,1,2,4} \left(\begin{cases} 1 & \text{if } |d_{i,j}^{Target} - d_{i,j}^{Prediction}| < cutoff \\ 0 & \text{otherwise} \end{cases} \right)}{4 * \text{number of distances } d_{i,j}^{Target}}$$

- Can be used **globally** (gDDT) for all distances in the target structure, or **locally** (lDDT) for distances below a certain threshold.
- Can be applied to all atoms, or subsets (e.g. back-bone).
- Accounts for coverage (like GDT)
- Note: Like any all-atom method, we must correct for swaps of chemically equivalent atoms, e.g. in Phe, Tyr, Arg, Asp, Glu, Val, Leu ^[1,2].

[1] Biasini et al., Bioinformatics. 2010, 26:2626-268; <http://OpenStructure.org>

[2] Kopp et al., Proteins. 2007;69, 38-56;

Scores used in CASP9 TBM

1. **C α global distance test (GDT_HA)**
calculated with LGA ^[1] with 0.5, 1.0, 2.0, 4.0 Å cutoffs
2. **All-atom global distance test (GDC_all)**
calculated with LGA for 20 bins from 0.5 to 10.0 Å
3. **All atom local distance difference test (IDDT_all)**
within a 5 Å radius of each atom and 0.5, 1.0, 2.0, 4.0 Å cutoff values for distance errors calculated with OpenStructure ^[2].

[1] Zemla, A. Nucleic Acids Res. 2003, 31:3370-3374.

[2] Biasini, Bioinformatics 2010, OpenStructure.org

TBM Assessment Workflow

For all TBM targets:

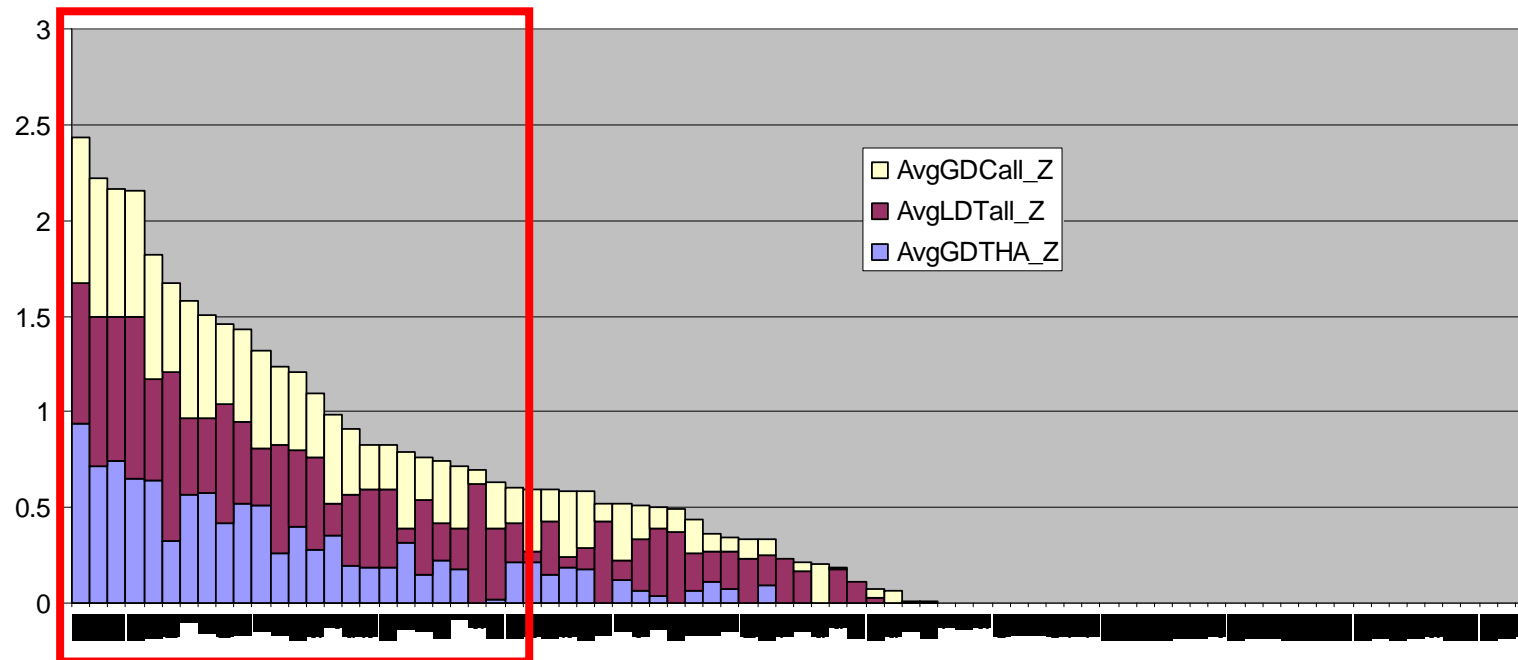
- we assess the prediction labeled #1; if there is no #1, take the one with the lowest number. For fragmented predictions, we assess the longest overlapping fragment.
- for each measure (e.g. GDT), we compute Z-scores for each target (excluding bad outliers worse than 2 sigma when calculating mean / standard deviation)
- we assign predictions with negative Z-scores to 0 (= average score over all groups for this target).
- we assign physically impossible predictions to $Z=0$.

TBM Assessment Workflow

- For each prediction group, compute median Z-scores for each group for each of the scores (GDT-HA, GDC, IDDT). Select the top 25 groups based on sum of median Z-scores.
 - Compute all pair wise differences (***paired t-test on common targets***) between top 25 groups on raw scores.
- ➔ Direct head-to-head comparison: count number of significant wins against all other groups for final ranking.

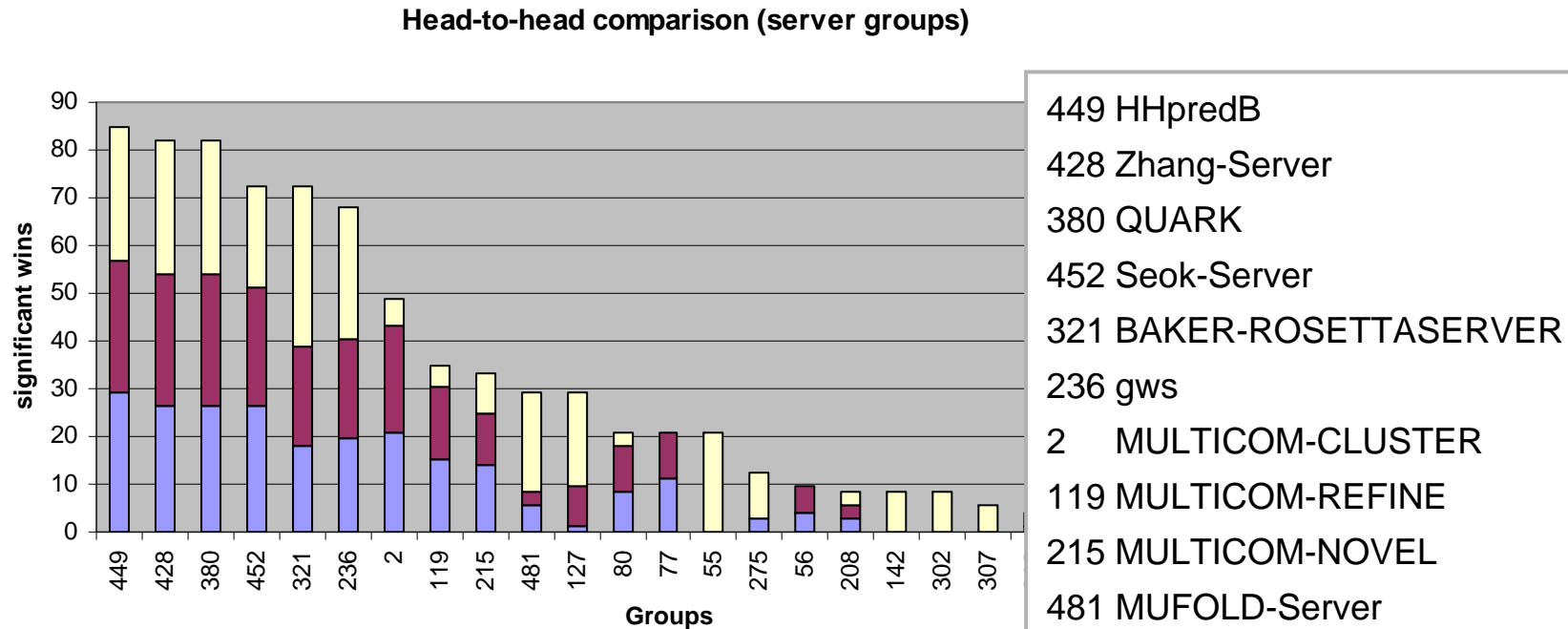
TBM – Server Groups

- Top 25 groups selected based on sum of median Z-scores for GDT-HA, GDC-all, and IDDT-all for all prediction targets (human + server):



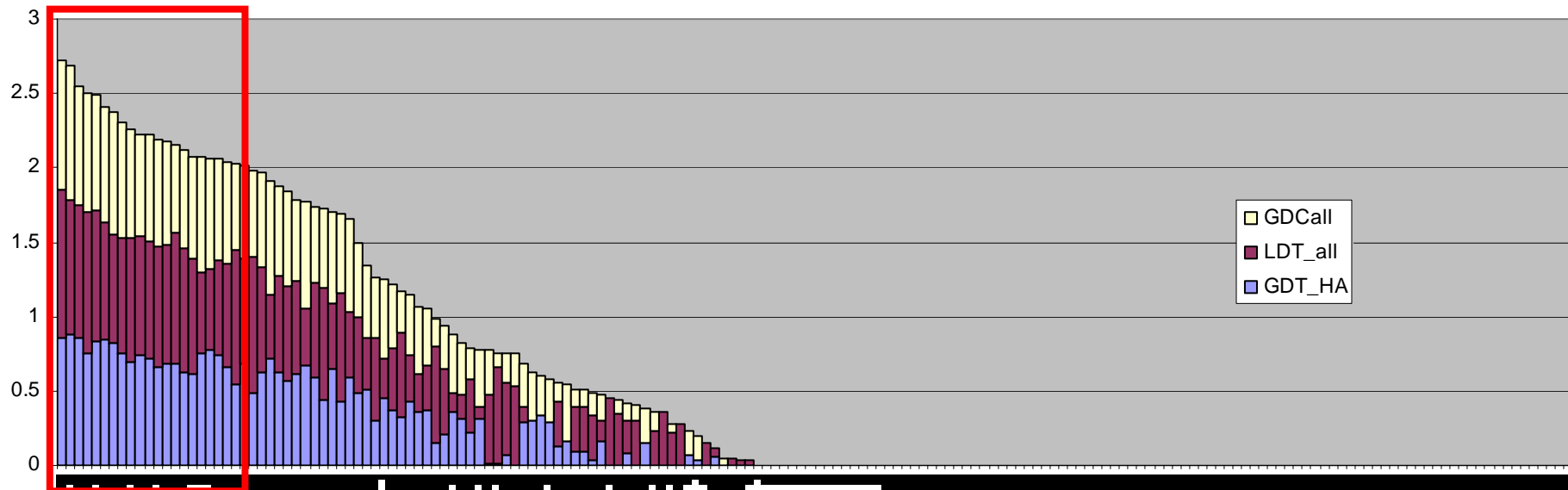
TBM – Server Groups

- Direct head-to-head comparison (number of significant wins within the top 25 groups on paired t-Tests on common targets on all three scores.)
- ***All atom model accuracy and weighting of unphysical interactions has significant influence on the final ranking.***



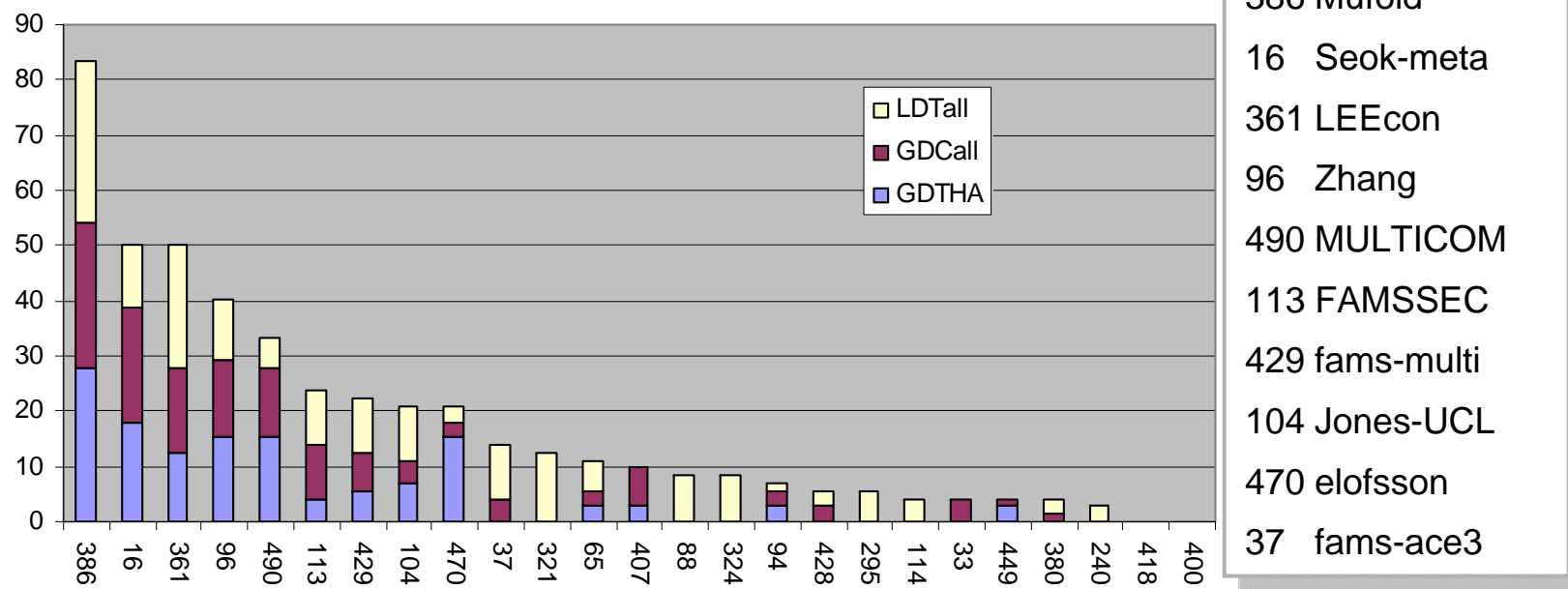
TBM – Human Targets

- Top 25 groups selected based on sum of median Z-scores for GDT-HA, GDC-all, and IDDT-all for human prediction targets:



TBM – Human Targets

- Direct head-to-head comparison (number of significant wins within the top 25 groups on paired t-Tests on common targets on all three scores.)
- All atom model accuracy and weighting of "unphysical models" has significant influence on final ranking of top groups.



"Server" vs. "human" methods in CASP9

- We see different types of methods on top the lists:

449 HHpredB	"Server"	386 Mufold	"Human"
428 Zhang-Server		16 Seok-meta	
380 QUARK		361 LEEcon	
452 Seok-Server		96 Zhang	
321 BAKER-ROSETTASERVER		490 MULTICOM	
236 gws		113 FAMSSEC	
2 MULTICOM-CLUSTER		429 fams-multi	
119 MULTICOM-REFINE		104 Jones-UCL	
215 MULTICOM-NOVEL		470 elofsson	
481 MUFOLD-Server		37 fams-ace3	

- Consensus or Human Creativity? Obviously, Meta-methods dominate the "human" category.
- Most "humans" were fully automated. Role of human interventions was not creative and not essential, but mainly "debugging".

"We consider the method not very interesting but just wanted to see how a simple consensus method would work in CASP. We did not even submit an abstract [...] If we are invited to the panel based on the results of [...] only, I think that we may not be eligible to be in the panel."

One of the top CASP9 predictors in reply to my email on relative success of their different methodologies in CASP9.

What would BLAST be
without an E-Value?

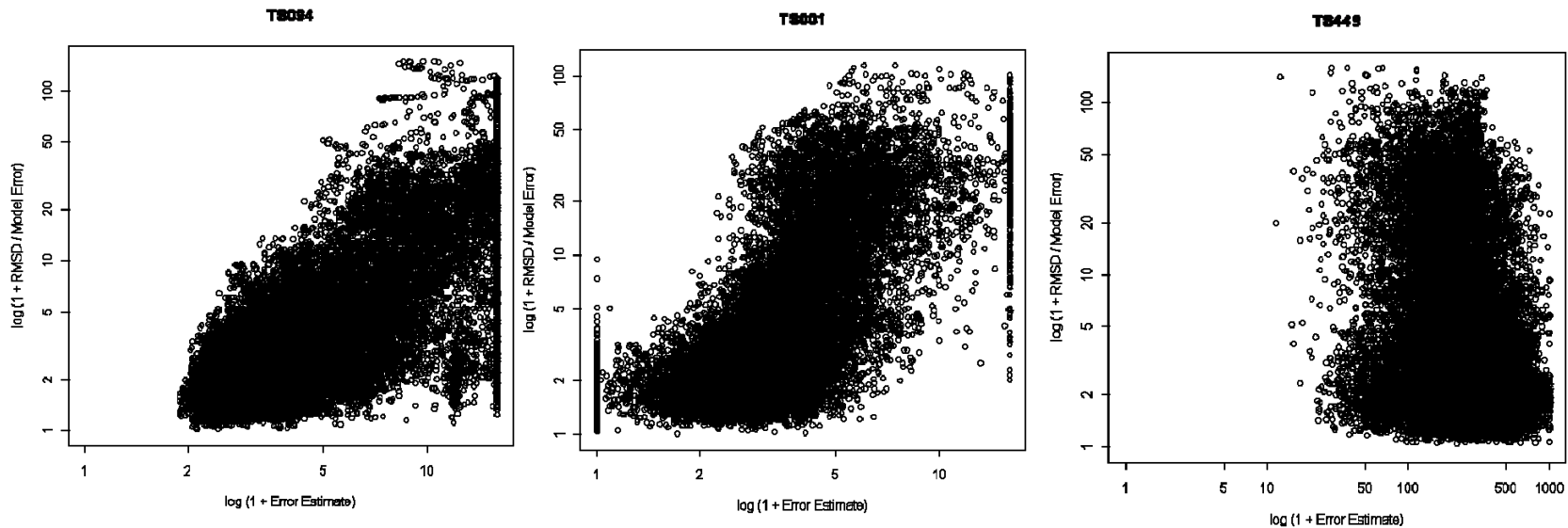
Assessment of model "B-Factors"

What would BLAST be without an E-Value?

- Accuracy differences between different protein targets are much smaller than the differences between the (best) methods.
- Therefore, in order to be practically useful, any structure prediction should come with a confidence measure.

Assessment of model "B-Factors"

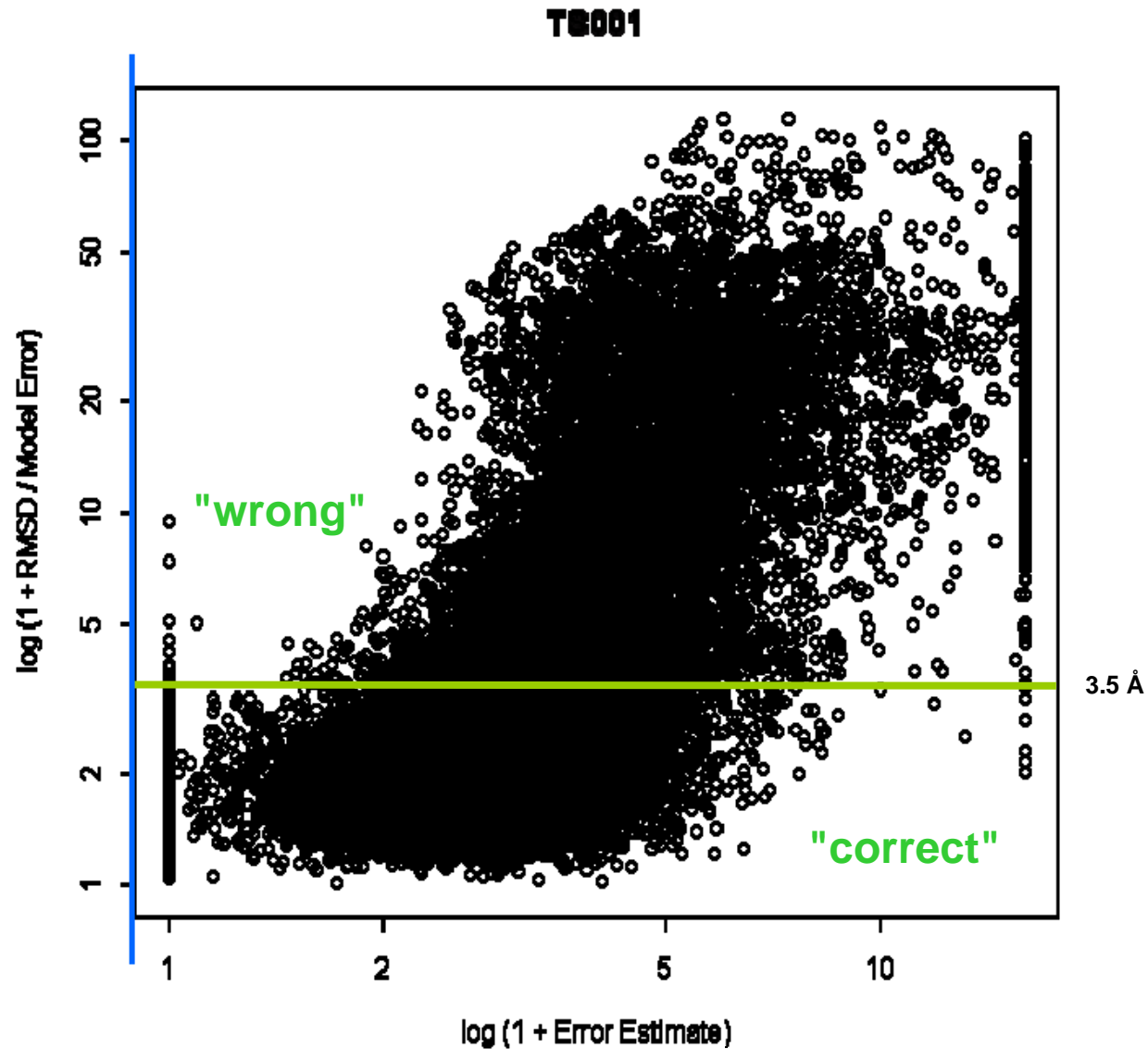
- Some groups did great effort to assign confidence values.
- However, others didn't really bother much:



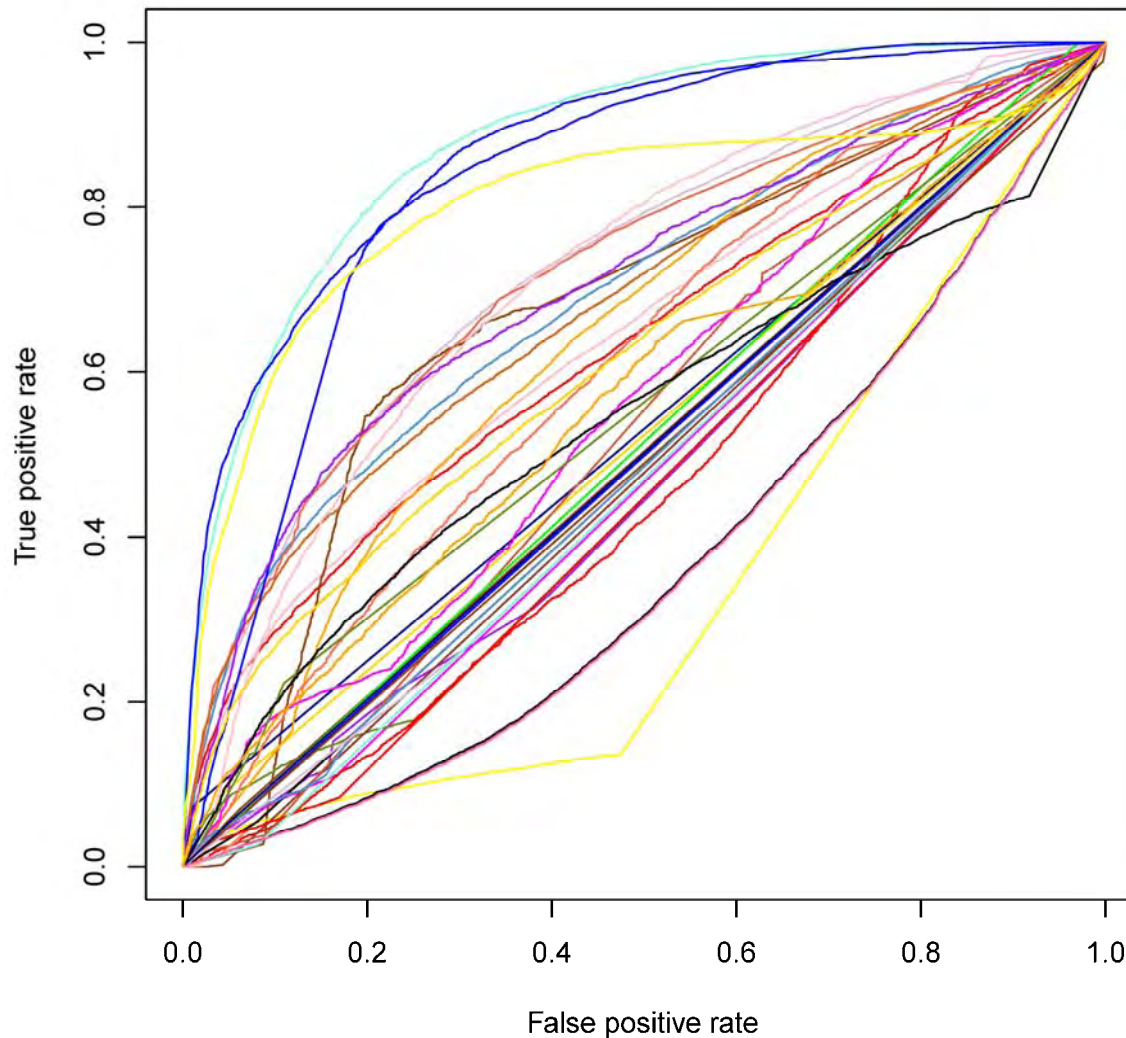
Assessment of model "B-Factors"

- Model confidence values ("B-Factors") were evaluated vs. model $C\alpha$ errors calculated on a 4 Å superposition with LGA.
- **Correlation:** global log-linear correlation of all residues in all models of a group: $\log(1+Bfactor) \sim \log(1+error)$
- **ROC:** Classification ("correct" / "incorrect") of all residues in all models of a group using a **3.5 Å cutoff**. ROC AUC is used as measure to assess the "B-Factors" as global residue accuracy estimate.

Assessment of model "B-Factors"



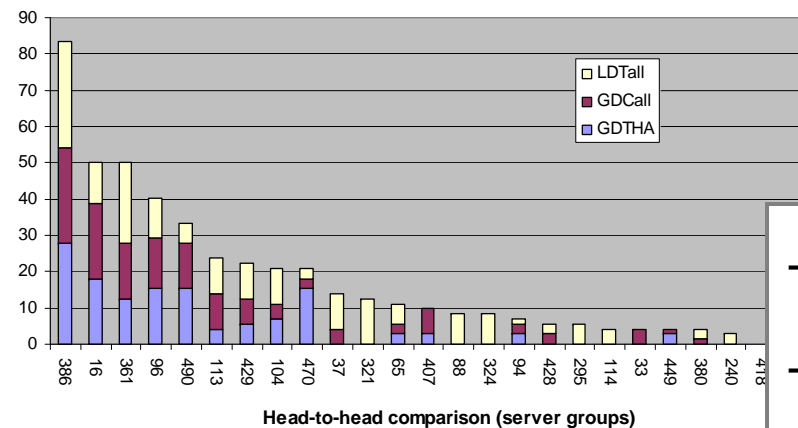
Assessment of model "B-Factors"



TS094	0.888
TS001	0.880
TS102	0.872
<u>TS275</u>	<u>0.866</u>
TS297	0.833
TS423	0.811
TS470	0.804
TS033	0.778
TS018	0.754
TS236	0.730
TS436	0.722
TS147	0.718

Assessment of model "B-Factors"

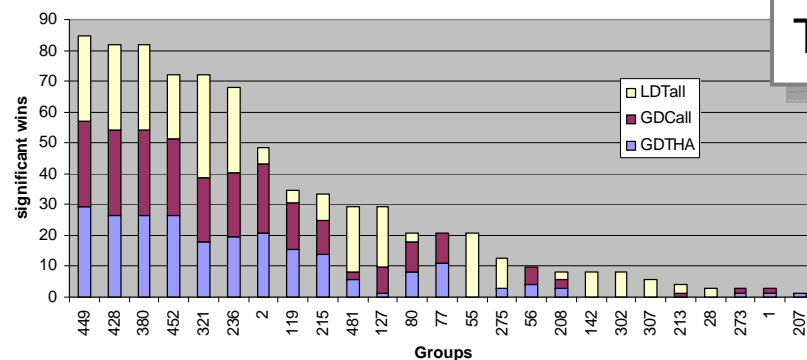
- Good groups should build good models (top 25) and correctly assign reliable and unreliable parts.



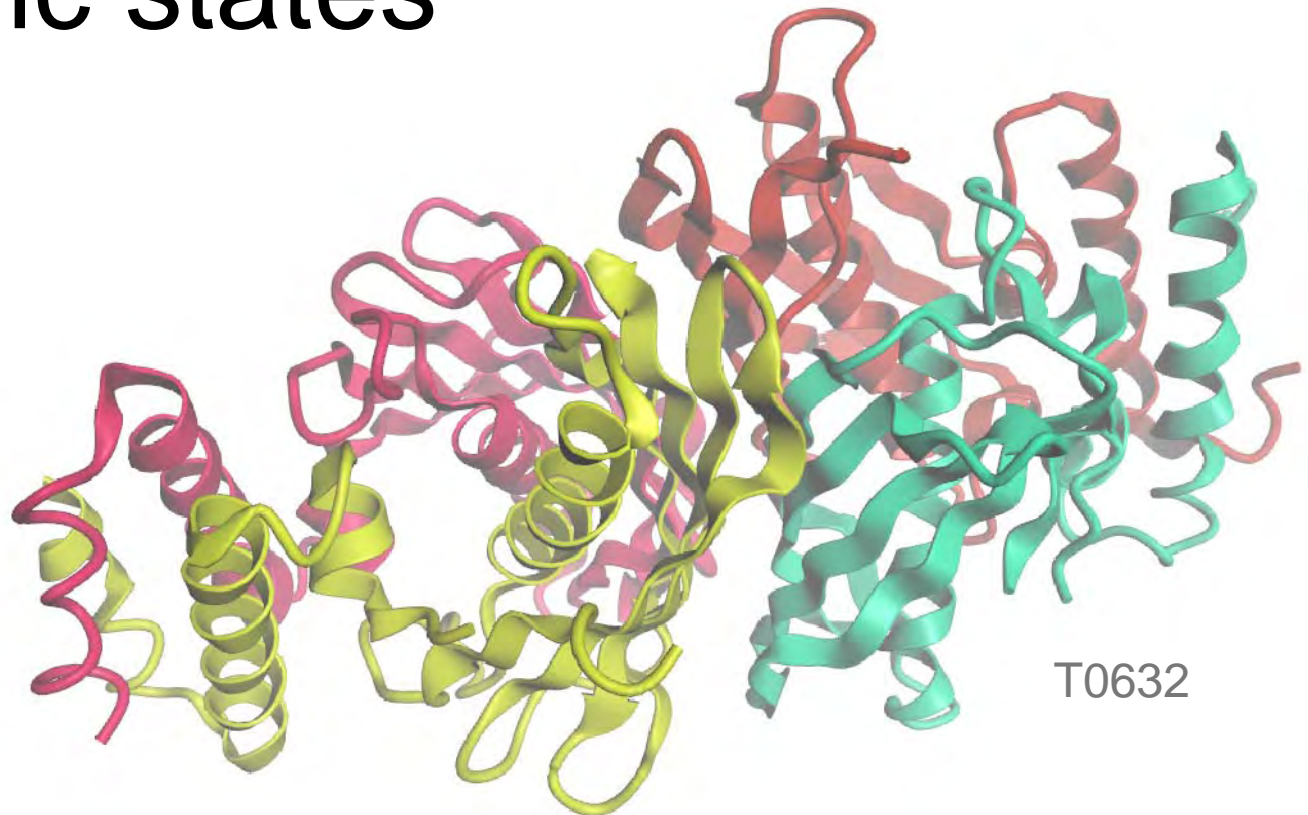
TS094 McGuffin (Liam Mc Guffin)

TS275 (s) IntFOLD-TS (Liam Mc Guffin)

TS001 (s) ProQ (Bjorn Wallner)



Assessing the prediction of oligomeric states



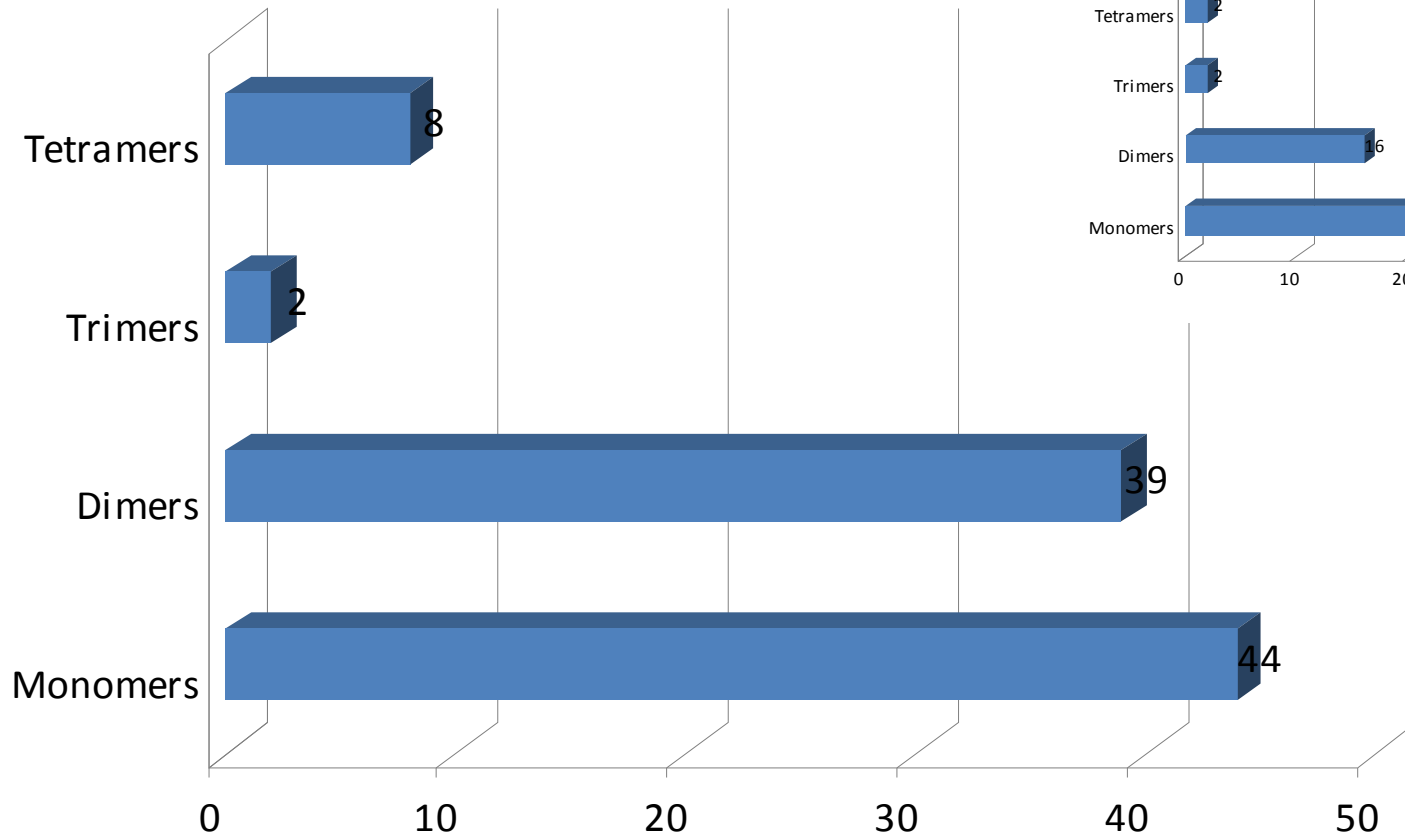
T0632

Oligomer Assessment - Groups

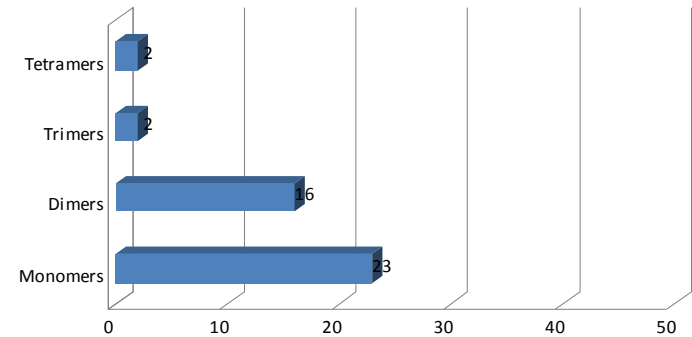
- Oligomer predictions are not a separate category, but part of the "normal" TS prediction process.
- **23 groups** with at least one oligomeric “#1” prediction were considered as "Oligomer Predictor Groups".
 - 16 Human Groups
 - 7 Server Groups

Oligomeric state as assigned by authors:

All Targets:



Only Human/Server Targets:



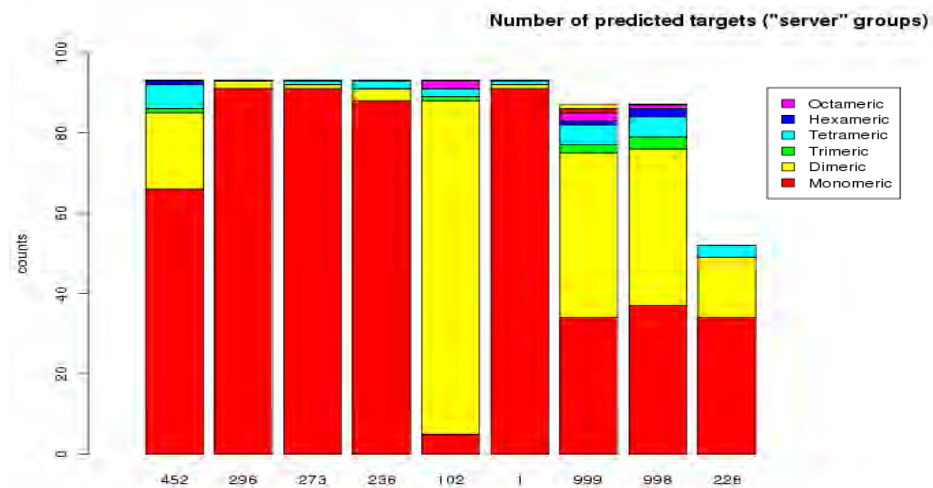
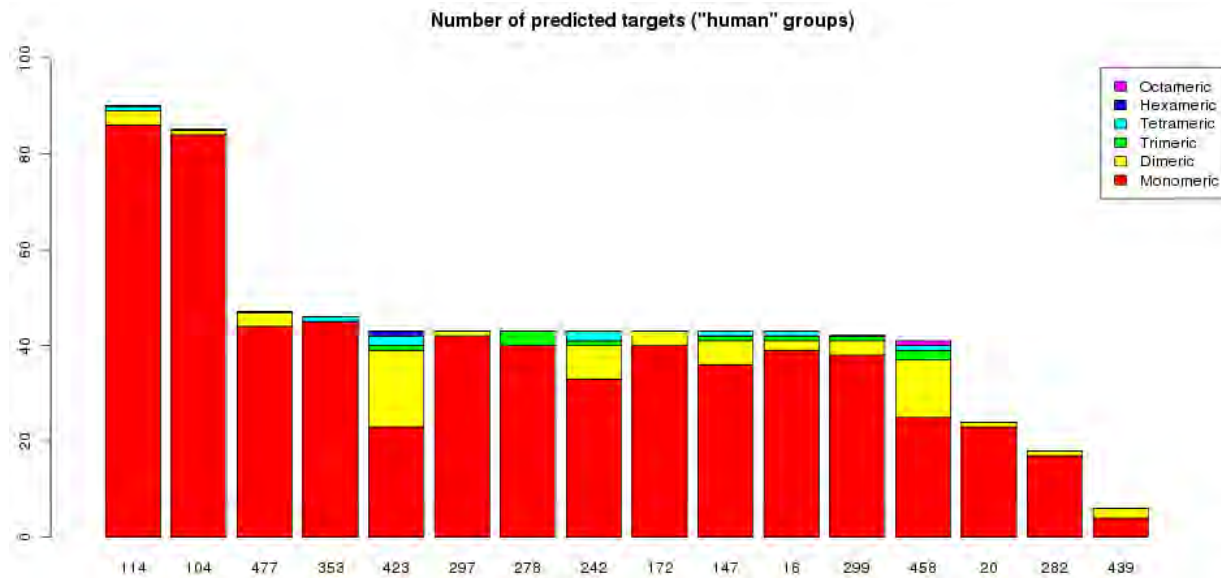
Oligomer assessment – Naive predictors

- Naive predictors were introduced as baseline to estimate the difficulty of the prediction task:
 - "Group 999": Predicts the oligomeric state as the PISA^[1] assembly of the best template structure identified by HHSearch^[2] with highest **sequence coverage** (min 15% seqid, 15% coverage).
 - "Group 998": Predicts the oligomeric state as the PISA assembly of the best template structure identified by HHSearch with highest **sequence identity** (min 15% seqid, 15% coverage).

[1] Krissinel and Henrick, J Mol Biol. 2007, 37:774-797.

[2] Söding, Bioinformatics. 2005, 21:951-960.

Oligomer assessment – Prediction Statistics



Were predictors able to predict structures of oligomeric complexes?

Contact agreement score defined as fraction of correctly predicted contacts. Residue i and j are defined as "in contact" if the **inter-chain distance** between their C β -atoms $\leq 12\text{\AA}$.

- $c(i,j)$ = Number of total contacts in complex between residue i and residue j .

$$S_{\text{agrees}} = \frac{\sum_{i,j} f(x_{ij}, y_{ij})}{\sum_{i,j} g(x_{ij}, y_{ij})}$$

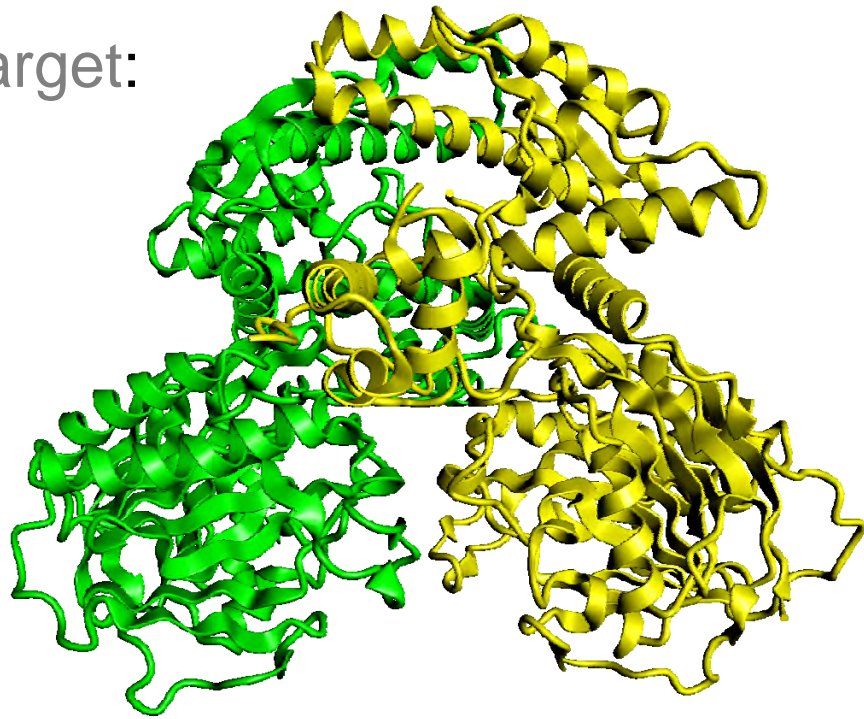
$$f(x_{ij}, y_{ij}) = \begin{cases} 1 - \frac{|x_{ij} - y_{ij}|}{\max(x_{ij}, y_{ij})}, & \max(x_{ij}, y_{ij}) > 0 \\ 0, & \max(x_{ij}, y_{ij}) = 0 \end{cases}$$

$$g(x_{ij}, y_{ij}) = \begin{cases} 1, & \max(x_{ij}, y_{ij}) > 0 \\ 0, & \max(x_{ij}, y_{ij}) = 0 \end{cases}$$

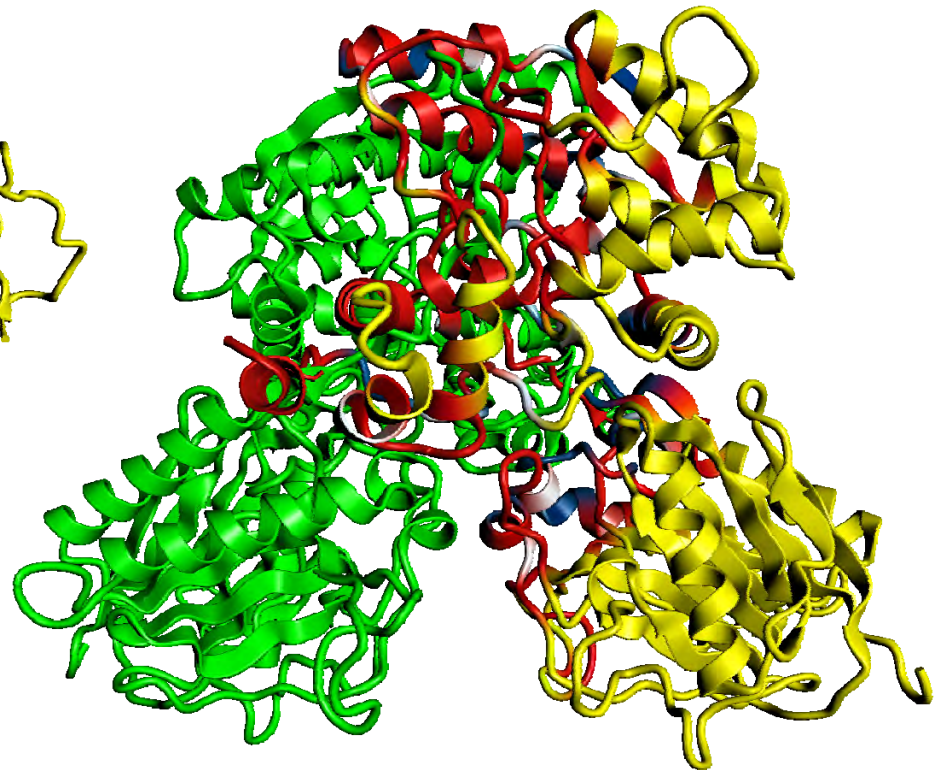
- $f(x_{ij}, y_{ij})$ = fraction of correctly predicted contacts
- $g(x_{ij}, y_{ij})$ = union of number of residue pairs with at least one contact

T0542

Target:



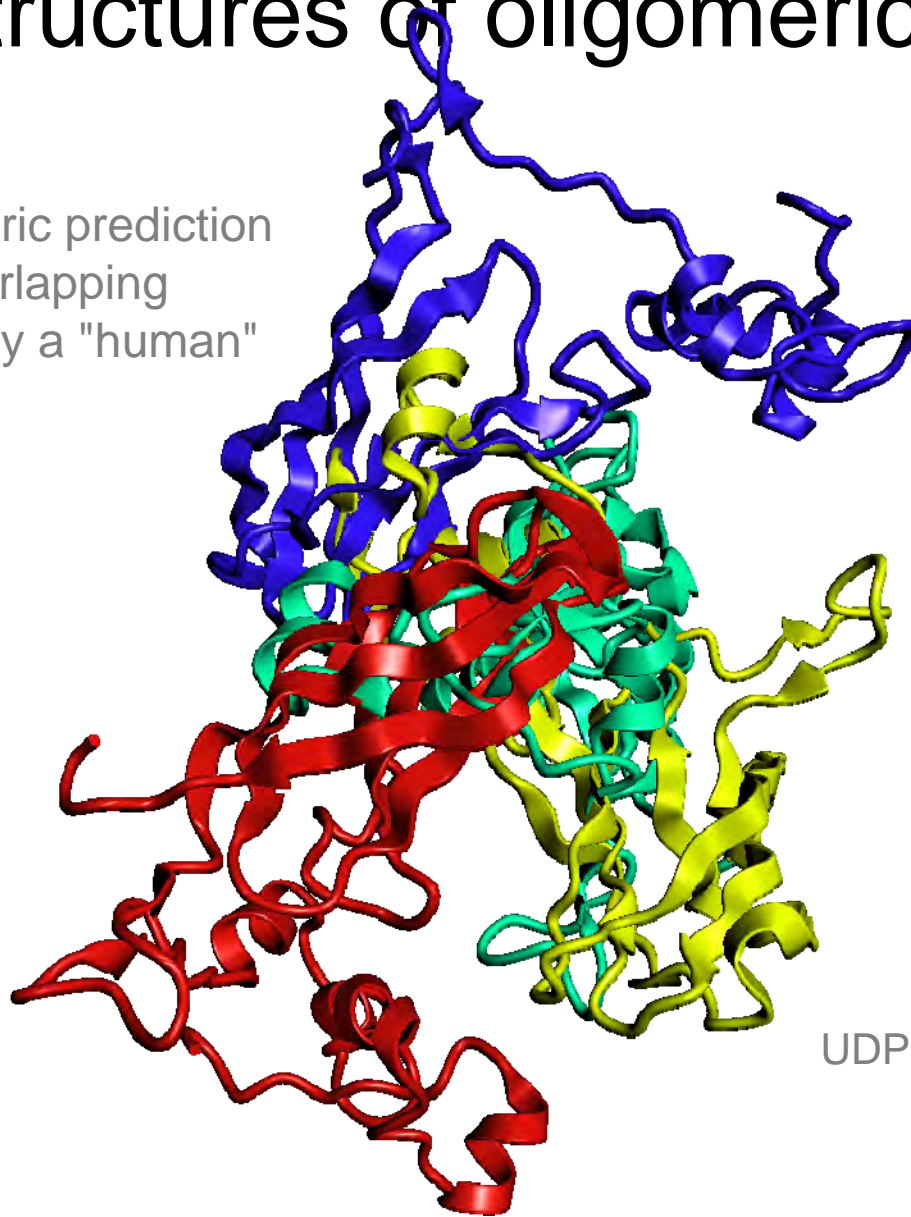
NH₃-Dependent NAD⁺ Synthetase



Group 102

Were predictors able to predict the structures of oligomeric complexes?

Tetrameric prediction with overlapping chains by a "human" group.



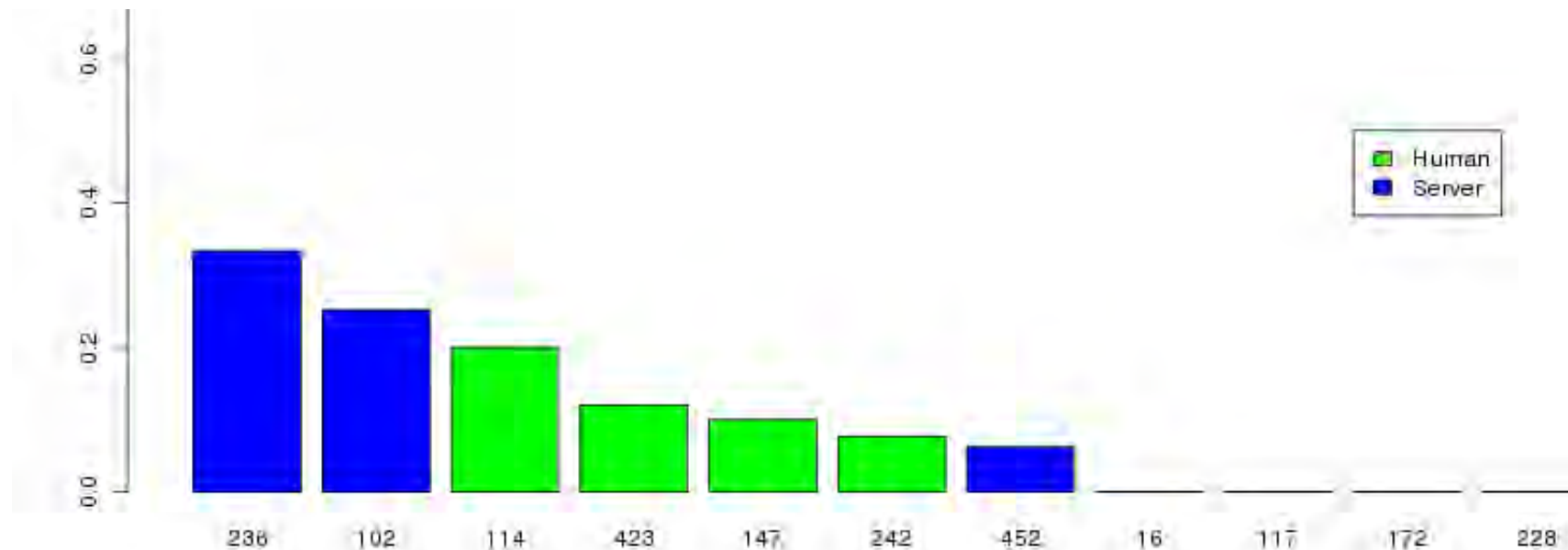
Target T0622



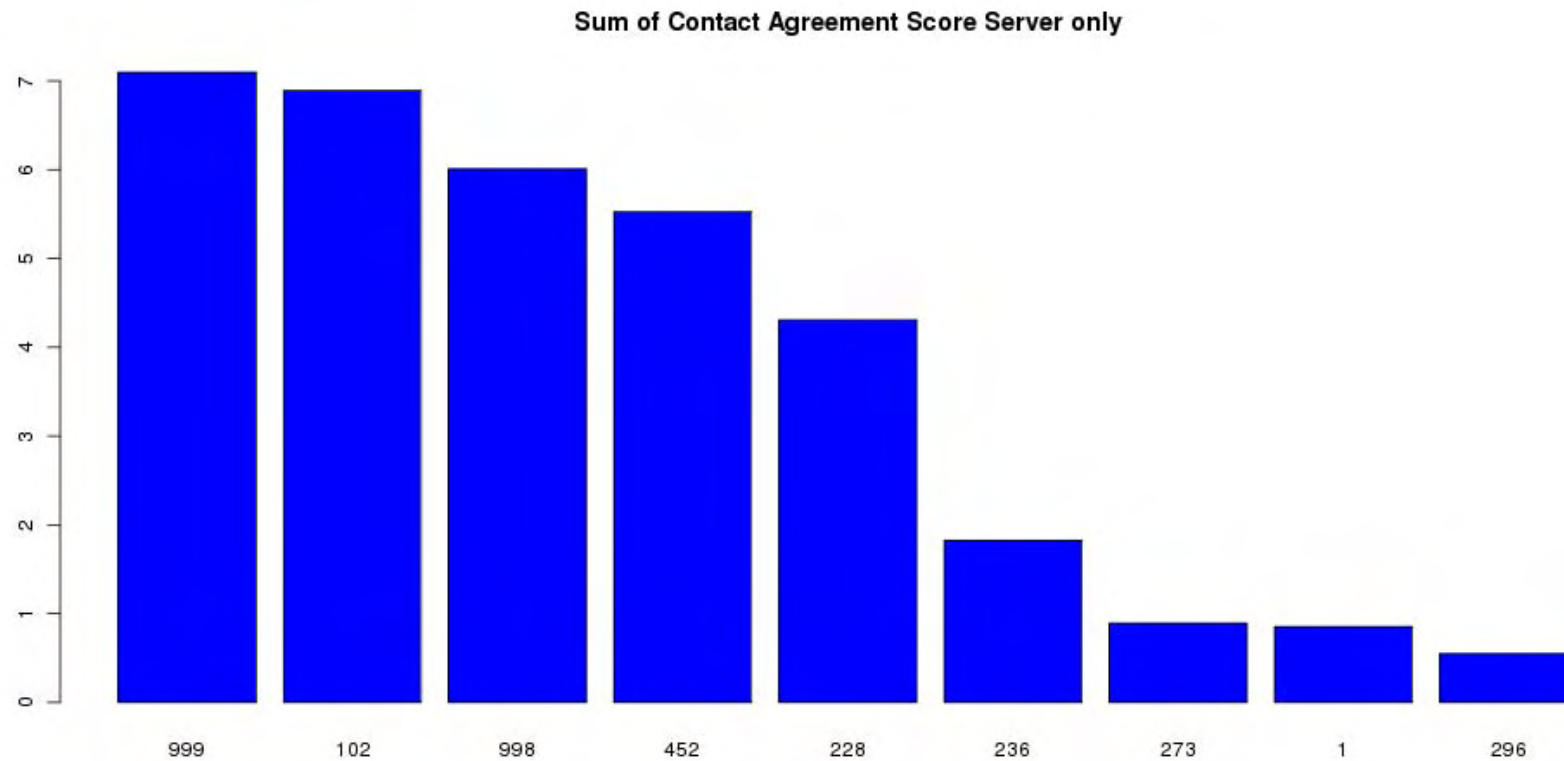
UDP-D-Quinovosamine 4-Dehydrogenase

Were predictors able to predict structures of oligomeric complexes?

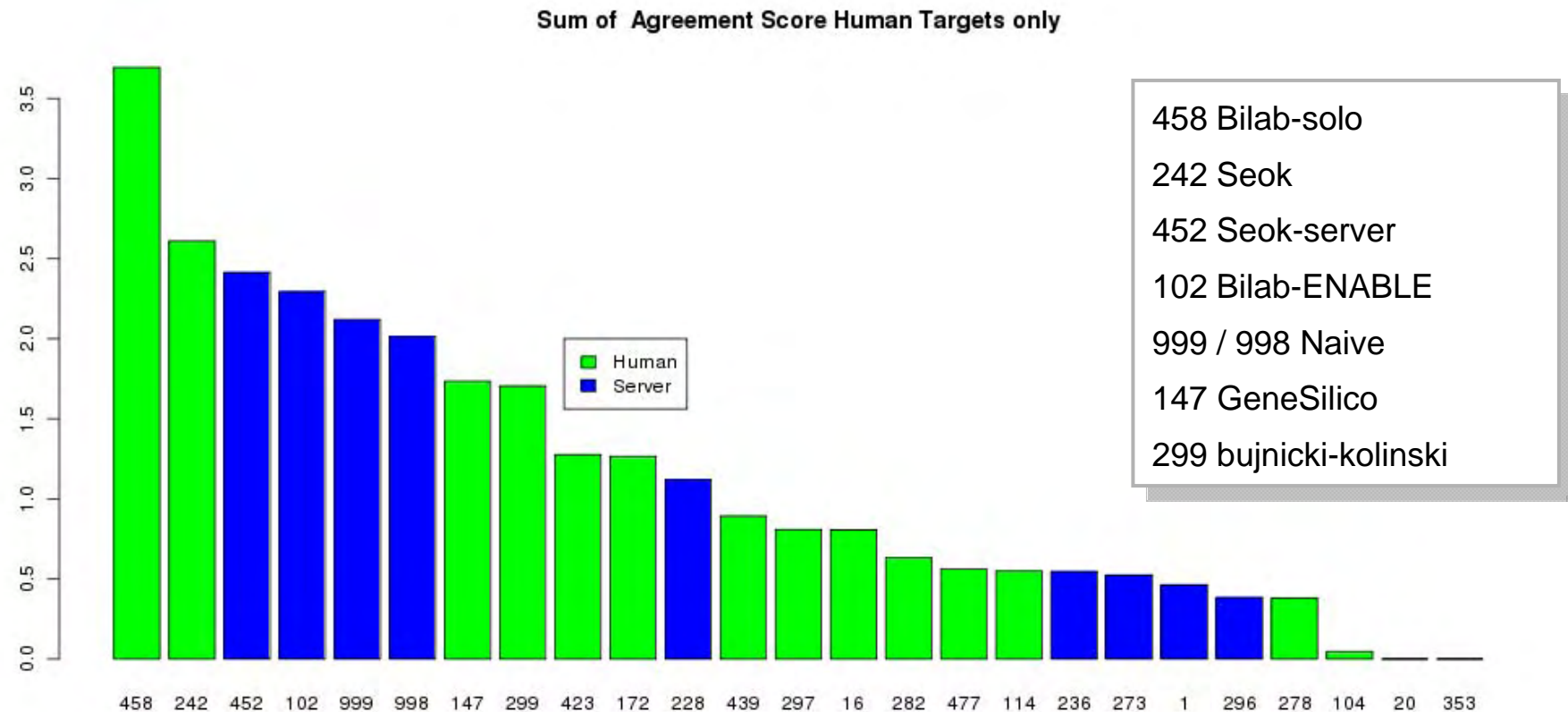
Fraction of models with more than 10 backbone clashes between inter-chain residues (for groups with more than 5 oligomeric predictions):



"Server" oligomer predictions



"Human" oligomer predictions



Assessing the prediction of oligomeric states

Conclusion

- Some promising developments in CASP9. However, there is considerable room for improvement.
- Obviously, there are substantial technical difficulties in modeling – as observed by the overall limited quality of the oligomeric predictions.

The infamous progress question:

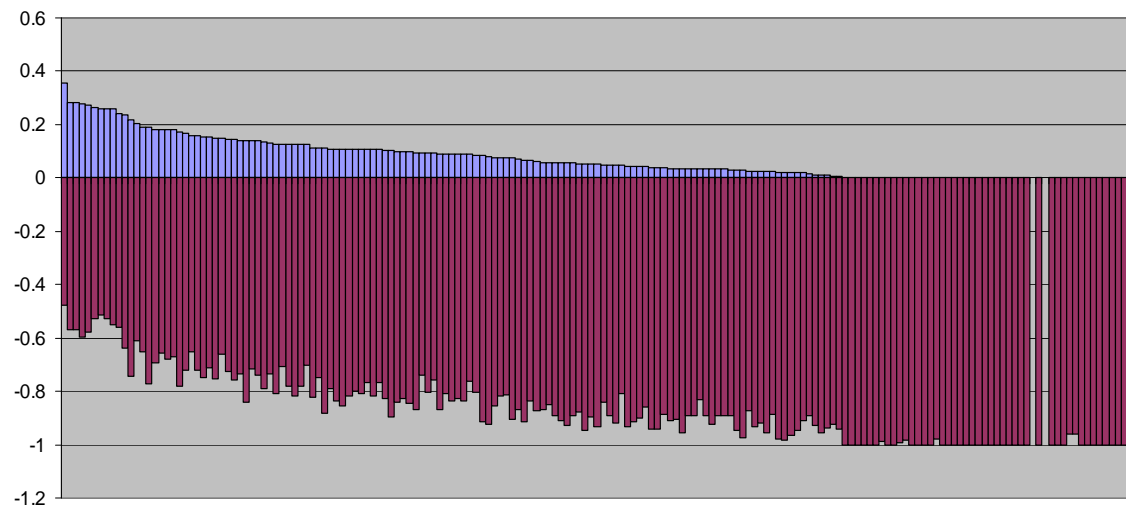
Has there been progress since CASP8?

Has there been progress since CASP8?

- Thanks, John. Very good question.
However, I'm not sure that I can answer this satisfactorily by only looking at the final models. Difficulty of targets is hard to compare; databases have changed; data from previous CASPs is a mess.
- But: The whole community has been working for 2 years to improve methods, and different groups are now on top. So obviously there is progress.
- Let's try to answer a more interesting question: **In which aspects have we seen progress in CASP9?**

Improvement over best single template model

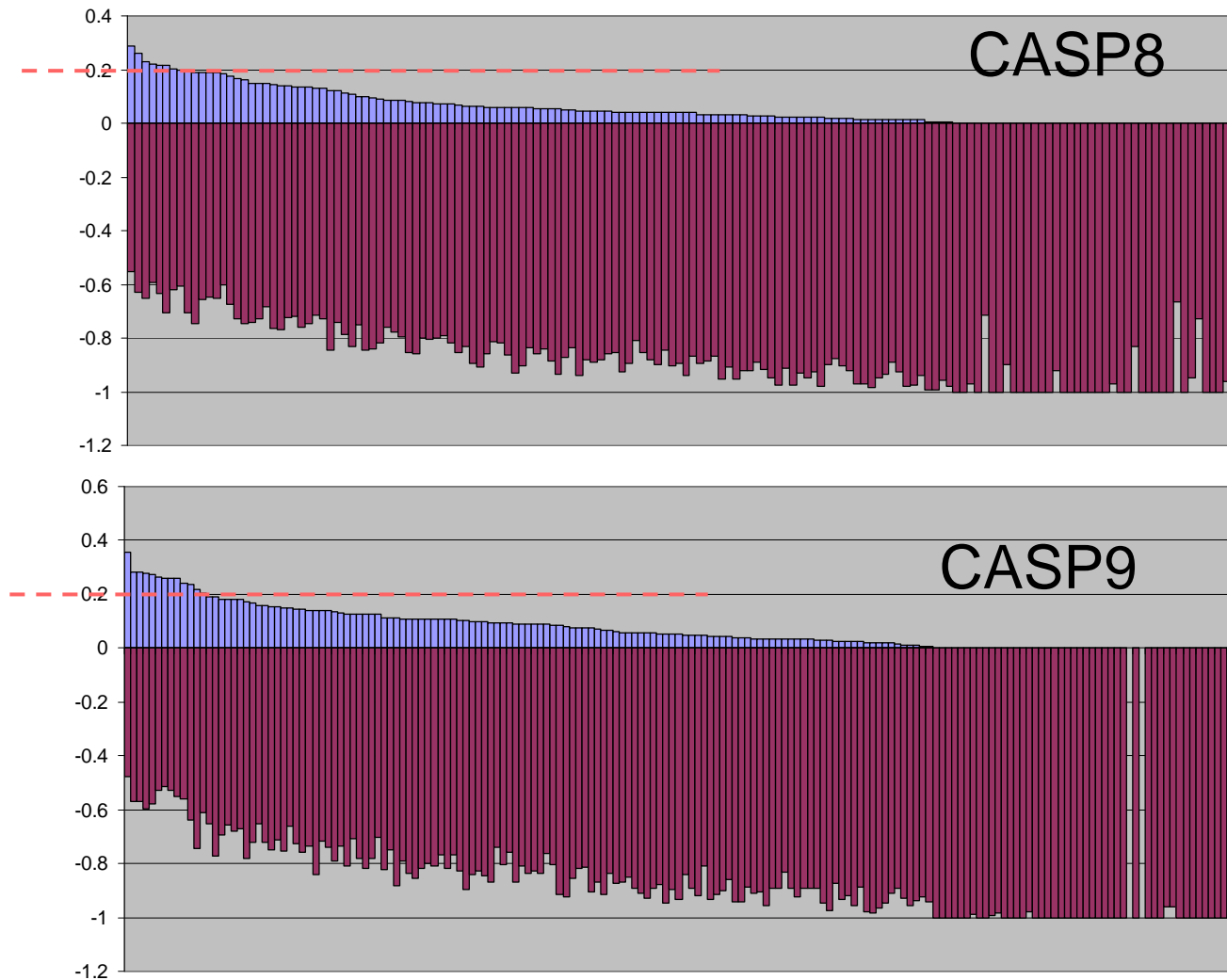
- We scanned the PDB for similar structures using SALAMI [1] and identified the structurally best available template at the CASP target deadline using LGA [2].
- Many predictor groups sometimes improve over the best single template by more than 1 GDT unit:



[1] Margraf et al. Nucleic Acids Res. 2009, 37, W480-4.

[2] Zemla, A. Nucleic Acids Res. 2003, 31:3370-3374.

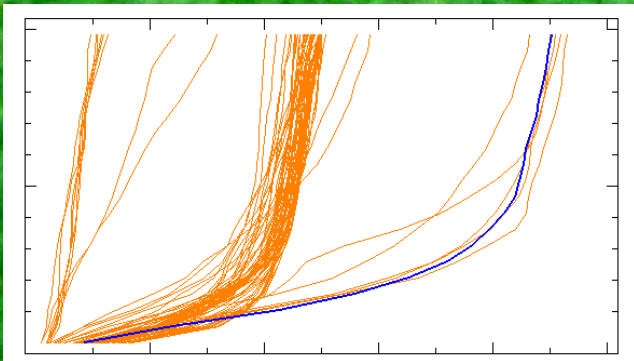
Improvement over best single template model between CASP8 and 9?



It seems that in CASP9, the fraction of models better than a single target structure is higher than in CASP.

(Preliminary analysis; needs to be redone more accurately.)

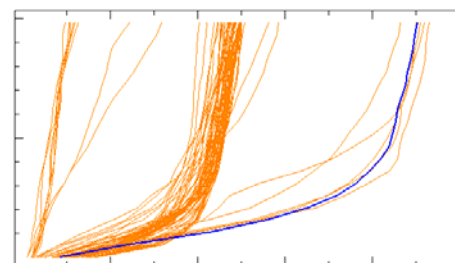
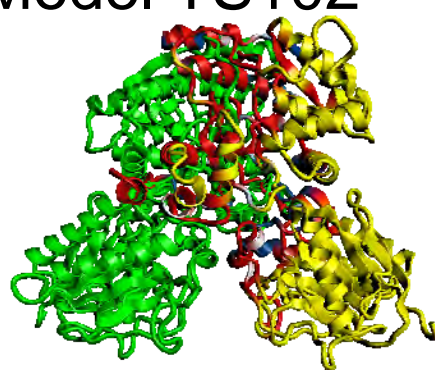
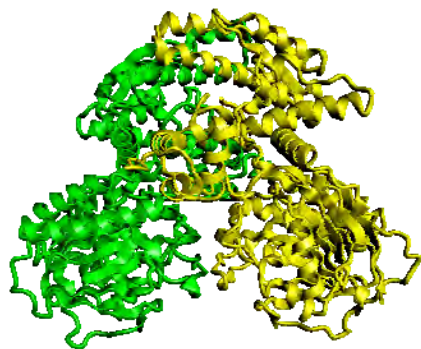
Looking for a "hole in one" ...



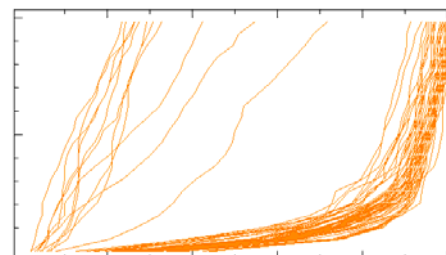
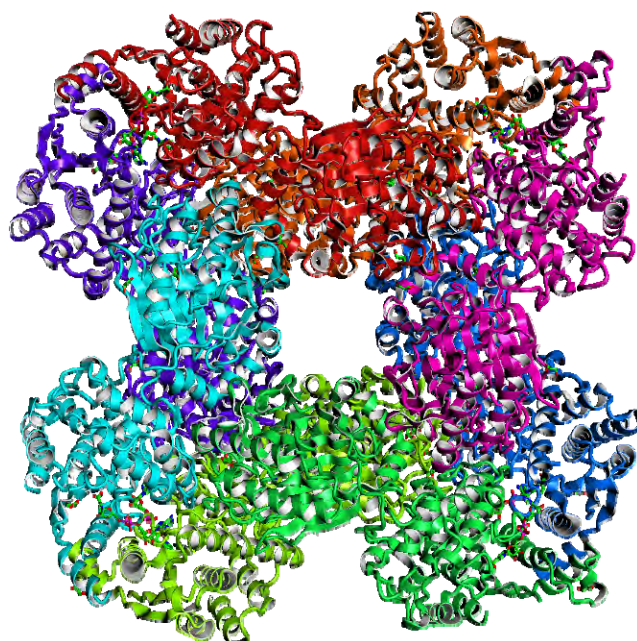
T0542

Target:

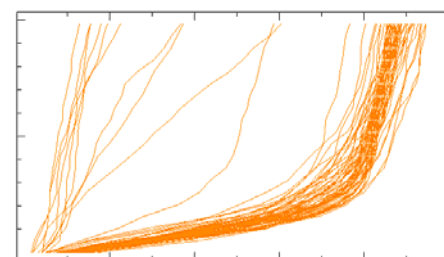
Model TS102



full length

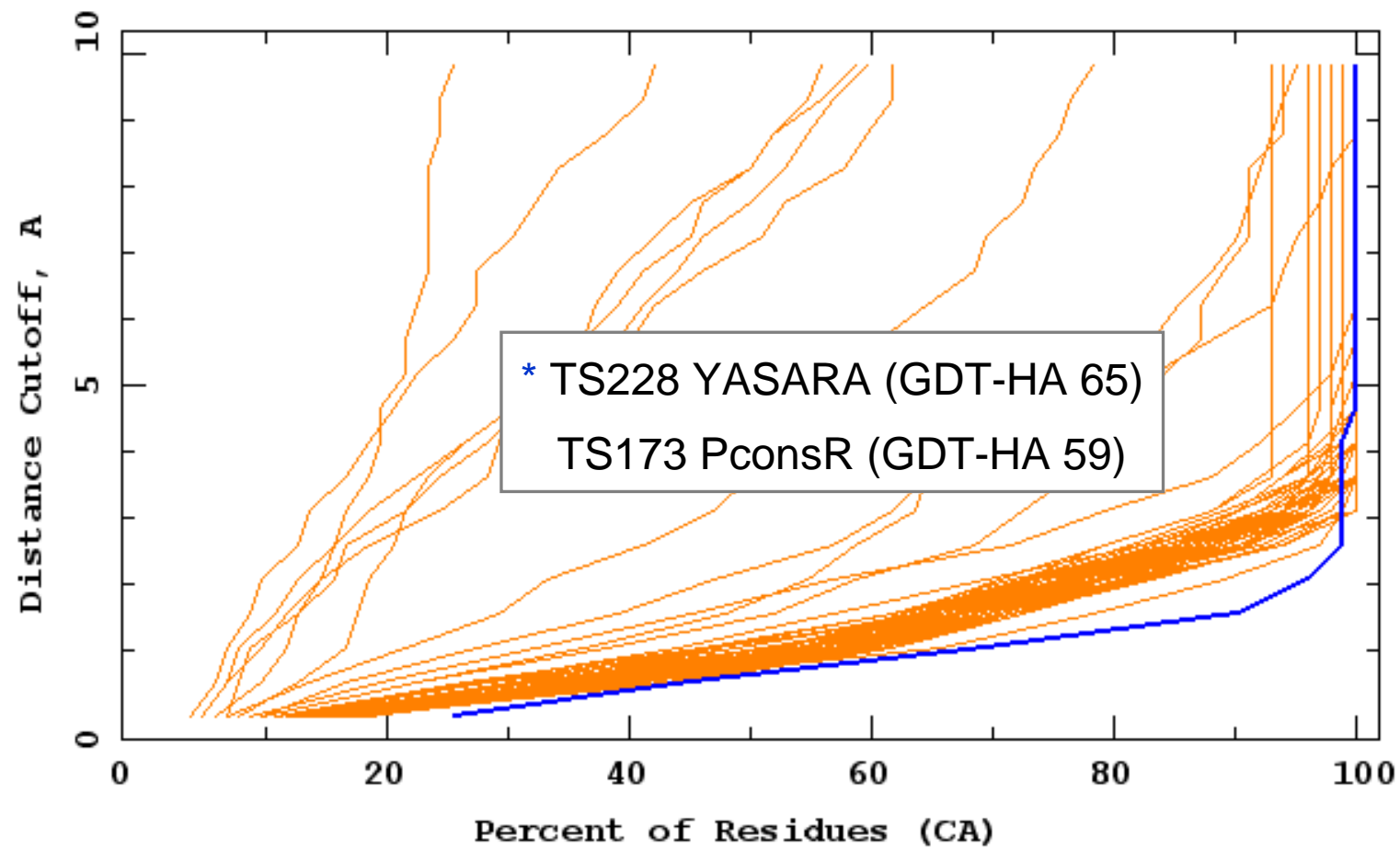


by domains

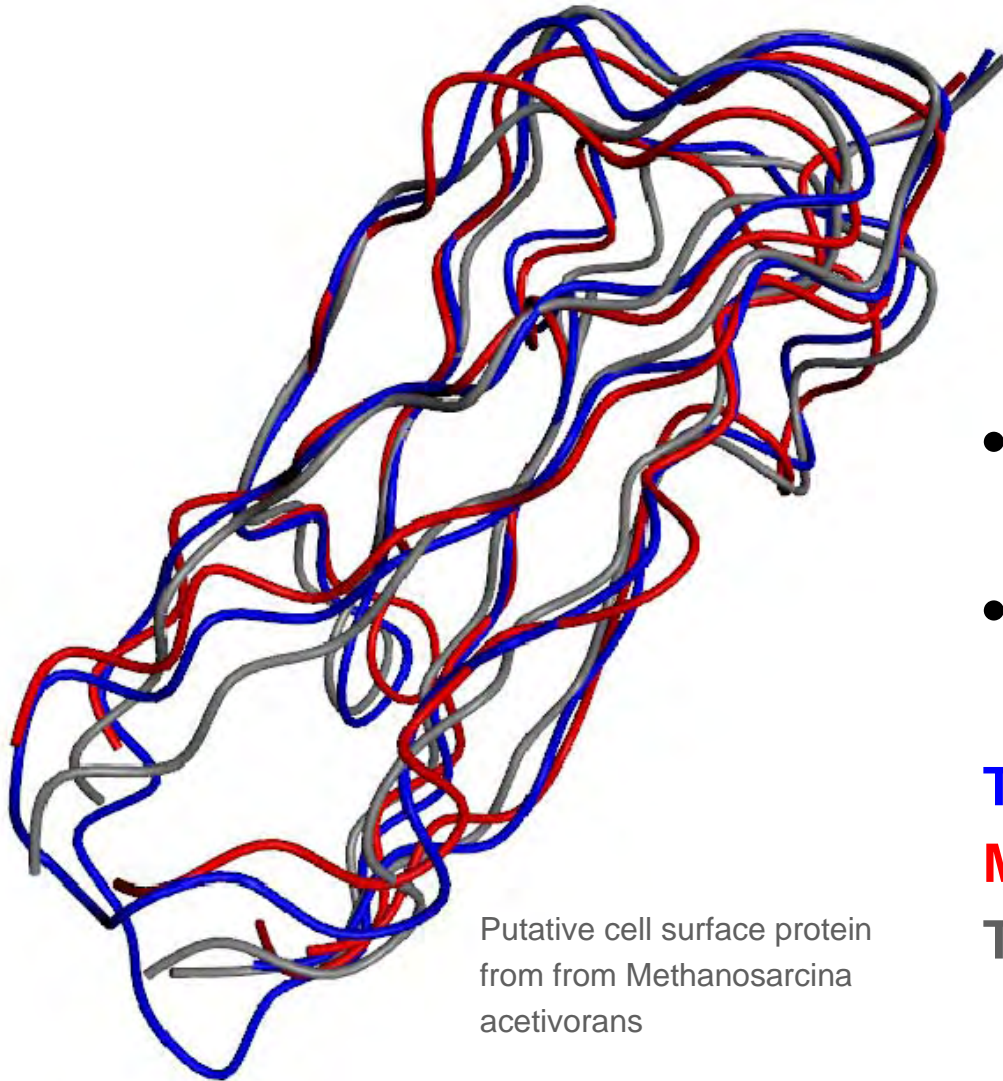


Alternative template with different domain orientation.

T0541-D1



T0541-D1



Putative cell surface protein
from from Methanosarcina
acetivorans

Some outstanding
prediction examples - not
always for obvious
reasons.

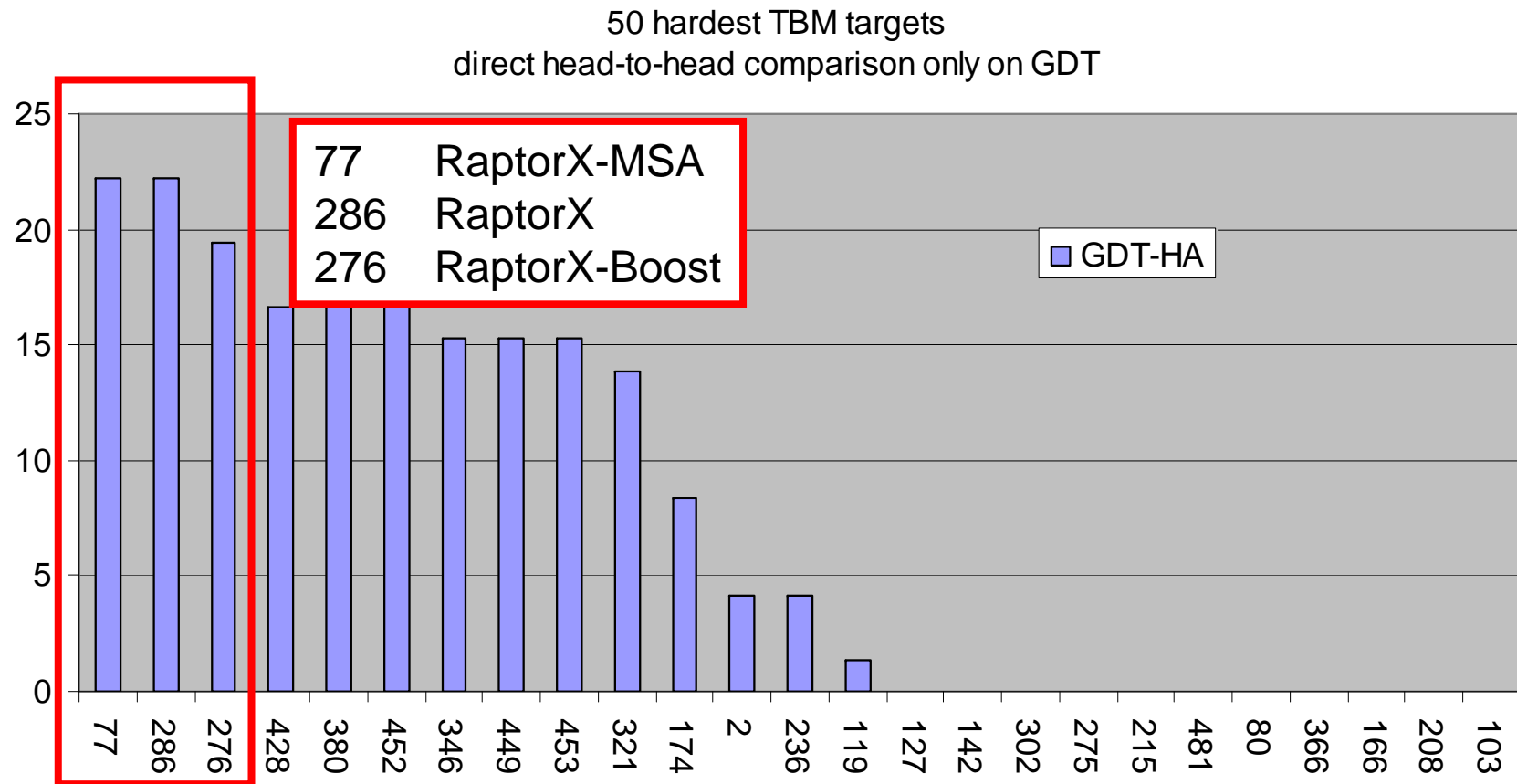
- Multiple template modeling?
- Successful refinement?

Target T0541-D1

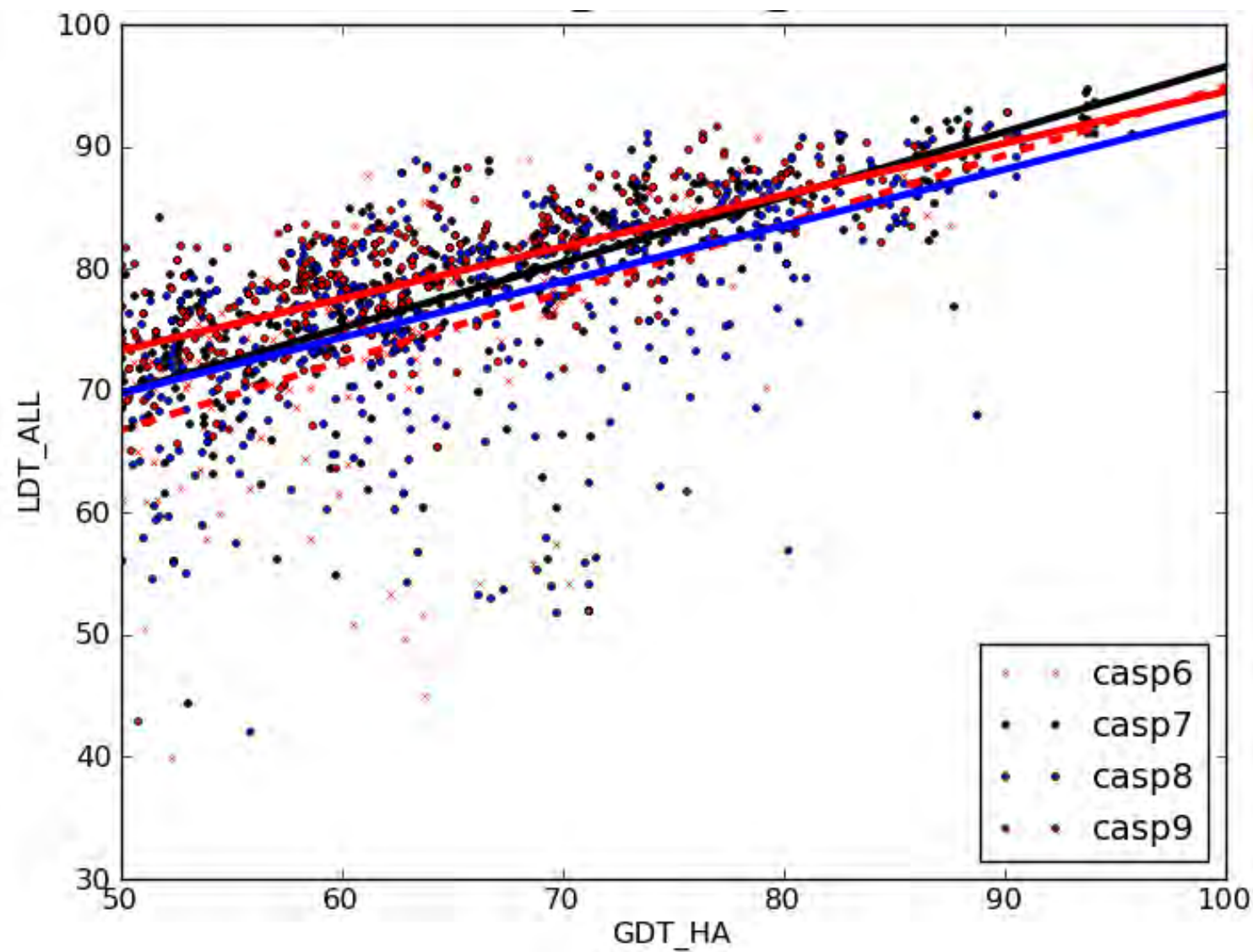
Model TS228

Template 3idu

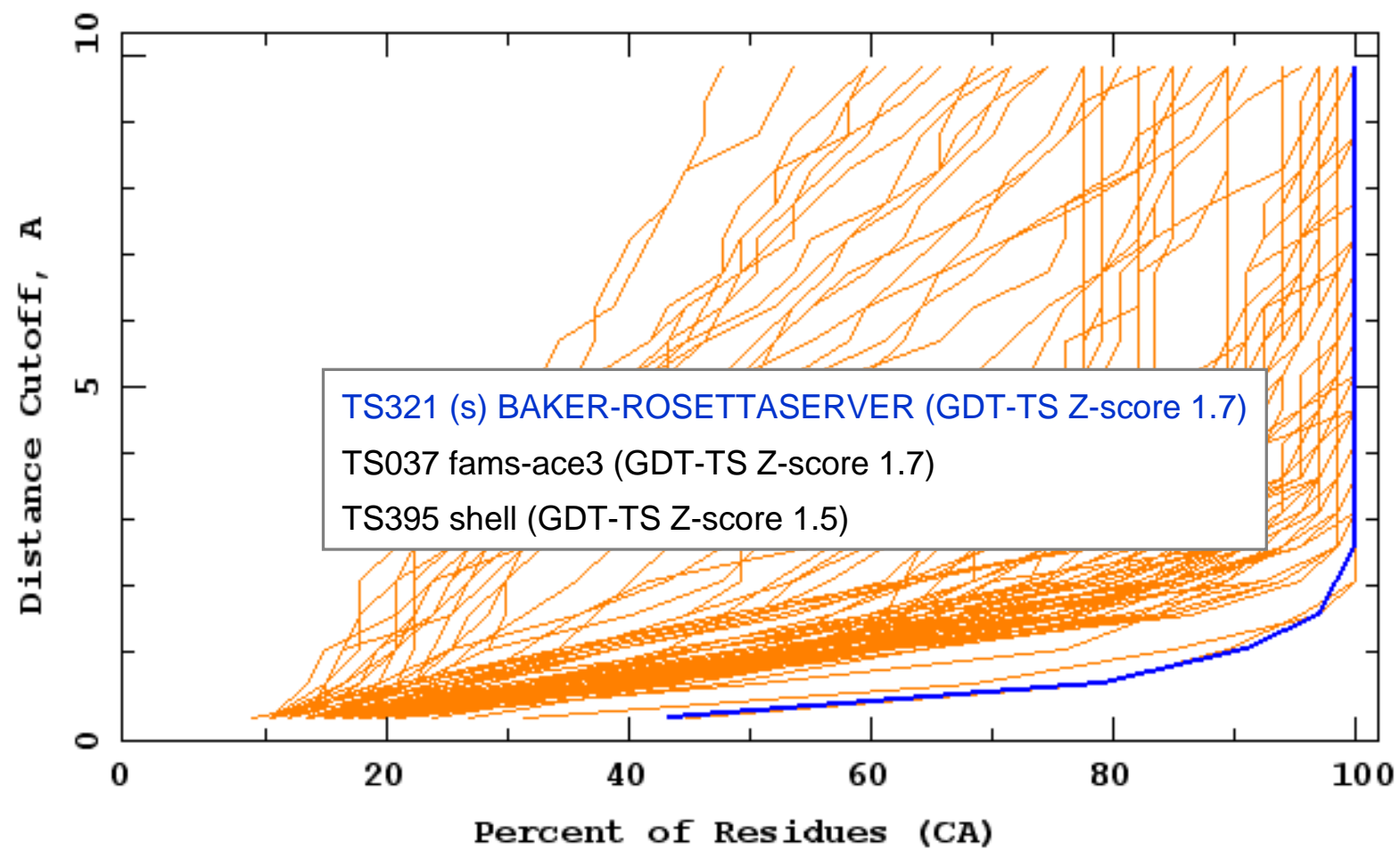
Any improvement in target-template alignment accuracy?

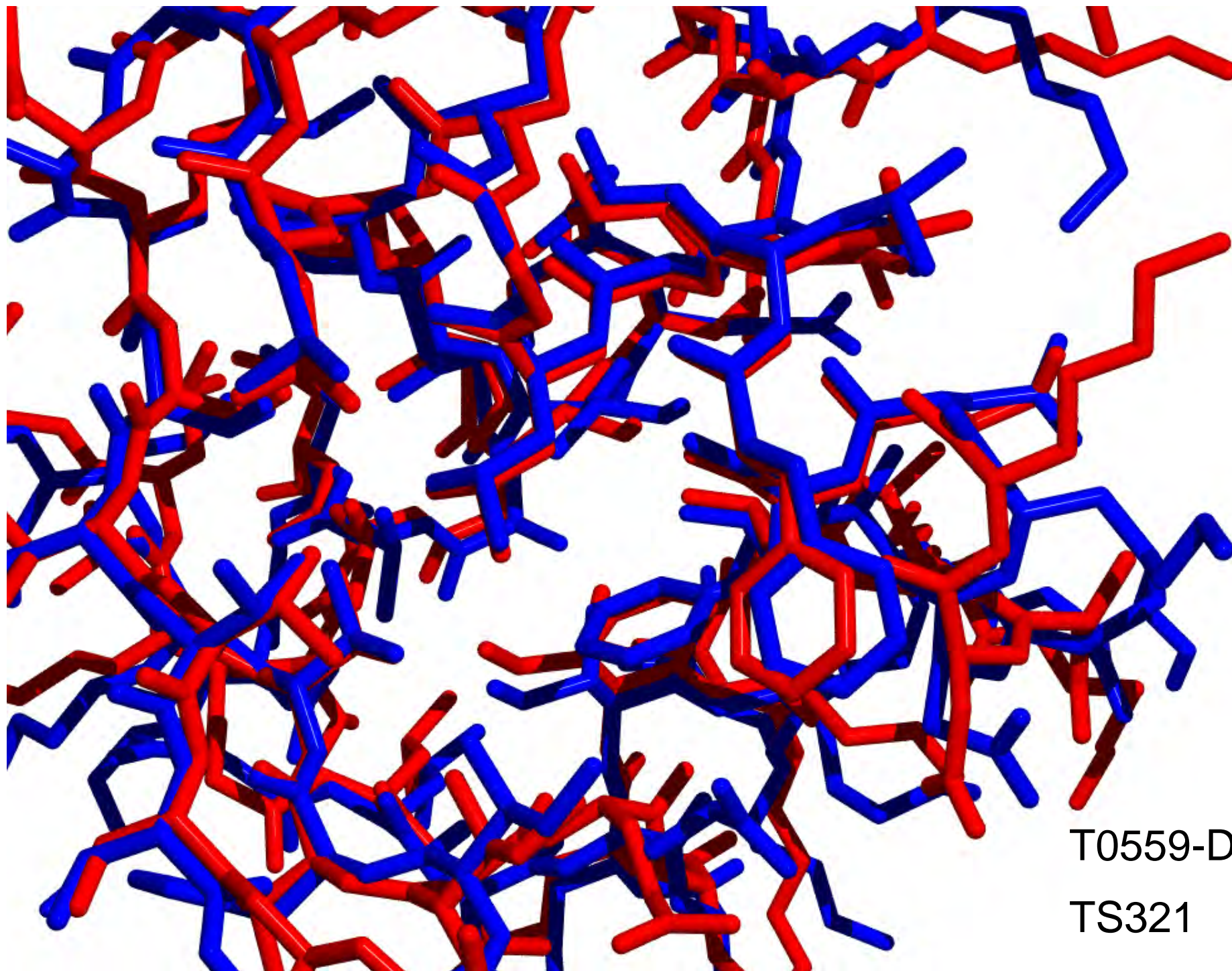


Progress in modeling atomic details?



T0559-D1





CASP9- TBM Round Table

- We tried to identify interesting new methods development during the assessment process.
- We sent questions to top predictor groups to figure out, where innovative methods development has been done for CASP9.
- ***Panelists are asked to discuss about specific progress in methods, challenges and ideas to overcome them.***

TBM Round Table: **New approaches to old challenges in template based protein structure modeling**

*Round table co-chair: **David Jones** (UCL)*

1. Integrating information from multiple templates.
2. How to build accurate atomic models?
3. Alignments and Model Quality Estimation.
4. Oligomer Modeling.
5. Unsolved challenges in template based modeling.

TBM Round Table: New approaches to old challenges in template based protein structure modeling

Round table co-chair: **David Jones** (UCL)

1. *Integrating information from multiple templates.*
 - **Johannes Söding** (449, 453, 346 HHpred)
 - **Dong Xu** (386 Mufold)
2. *How to build accurate atomic models?*
 - **Yang Zhang** (428 Zhang-Server, 380 QUARK)
 - **Jooyoung Lee** (361 LEEcon)
3. *Alignments and Model Quality Estimation.*
 - **Jinbo Xu** (77 RaptorX-MSA, 286 RaptorX)
 - **Liam Mc Guffin** (94 McGuffin)
4. *Oligomer Modeling and other unsolved challenges in template based modeling.*
 - **All previous panelists plus:**
 - **David Baker** (172 BAKER, 321 BAKER-ROSETTA)
 - **Shugo Nakamura** (458 Bilab-solo, 102 Bilab-ENABLE)
 - **Hahnbeom Park** (16 Soek, 242 Seok, 452 Seok-server)
 - **Janusz Bujnicki** (147 Genesilico, 299 bujnicki-kolinski)

Acknowledgments

CASP9 predictors

Valerio Mariani (TBM Assessment)

Florian Kiefer (Oligomer Assessment)

Marco Biasini (OpenStructure, *see poster*)

Lorenza Bordoli ("B-factor" assessment scripts)

Konstantin Arnold (IT System Administration)

Jürgen Haas (WhatIf, NMR constraints evaluation)

Gert Vriend / Jurgen F. Doreleijers (WhatIf, NMR constraints evaluation)

Thomas Margraf / Andrew Torda (Salami)

CASP9 co-assessors: **Nick Grishin, Lisa Kinch, & Justin MacCallum**

CASP Organizers & Prediction Center: **John Moult, Anna Tramontano, Andriy Kryshatovych, Krzysztof Fidelis**

Funding: SIB – Swiss Institute of Bioinformatics, SNF Swiss National Science Foundation, Biozentrum University of Basel, NIH National Institutes of Health