Fitting CASP Models to NMR Data

Janet Yuanpeng Huang
Roberto Tejero
Gaohua Liu
Swapna Gurla
Yojiro Ishida
Masayori Inouye
Gaetano Montelione

Andriy Kryshtafovych
Krzysztof Fidelis

Kelly Brock
Chris Sander


Janet Huang


Roberto Tejero


Andriy Kryshtafovych


Gaohua Liu


G.V.T. Swapna

Can We Fit Computational Models into NMR Data?

Inverse Structure Determination

(Make a Model -> Validate/Refine Against Data)

GT Montelione
CASP14 Dec 3, 2020

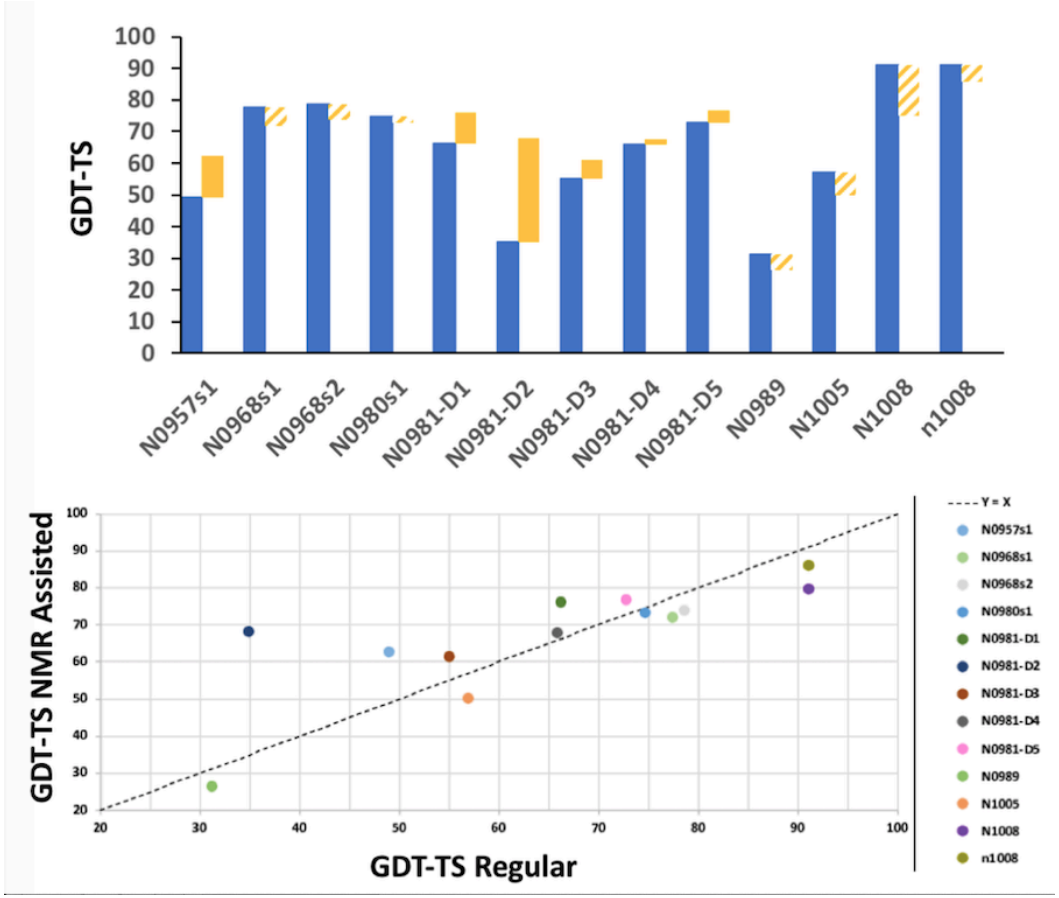
CASP13 – NMR-Guided Modeling
Sparse NMR Data -
Perdeuterated Proteins

11 simulated NMR data sets
2 real NMR data sets

For 6 targets – NMR-guided better than regular prediction

For 7 targets – Best Regular Prediction better than NMR-assisted !!

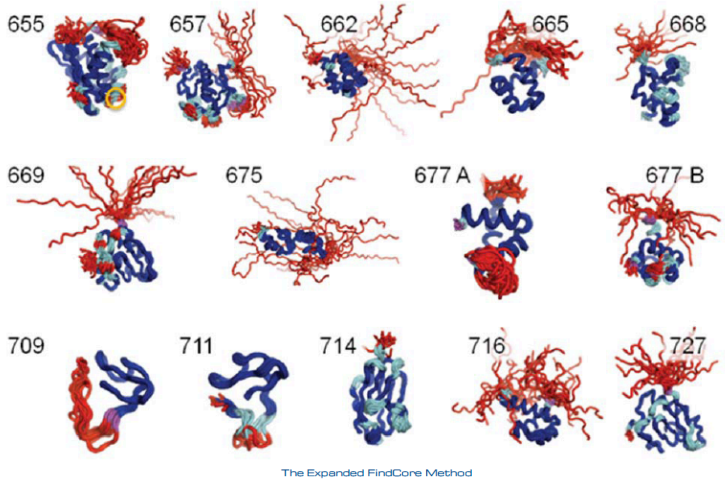
Suggests potential of using Best Regular Prediction to guide NMR structure determination – for difficult targets.



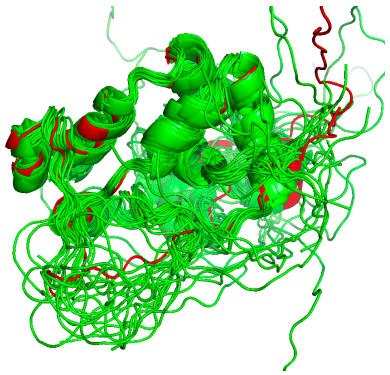
Sala, Huang, et al. PROTEINS 2019

Well- Defined vs Not-well-Defined Regions of NMR Models

Proteins are often more flexible in solution (or in cryoEM ‘glass’) than in the crystal lattice



The Expanded FindCore Method



T1027

FINDCORE2 PDB-Stat (distance variance matrix)

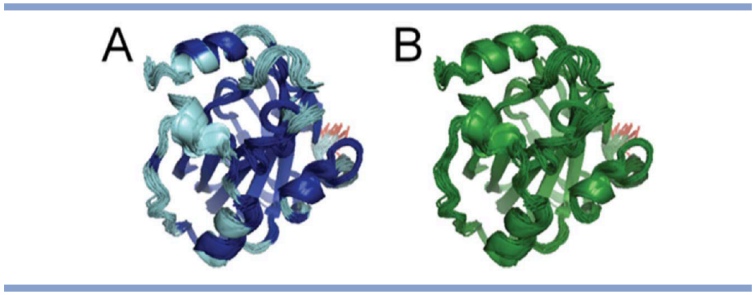


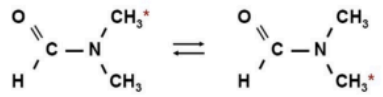
Figure 3

Comparison of FindCore, Expanded FindCore, and CYRANGE methods for identifying core atom sets. (A) FindCore (blue) and Expanded FindCore (cyan) regions of the human Raf-1 kinase inhibitor protein (PDB ID 2L7W). In this case, Expanded FindCore greatly expands upon the original core atom set identified by FindCore, only excluding the poorly converged N-terminal residue (red) from the core. (B) CYRANGE [2] produces a result similar to Expanded FindCore, except that CYRANGE defines its core (green) on a per-residue, rather than per-atom basis.

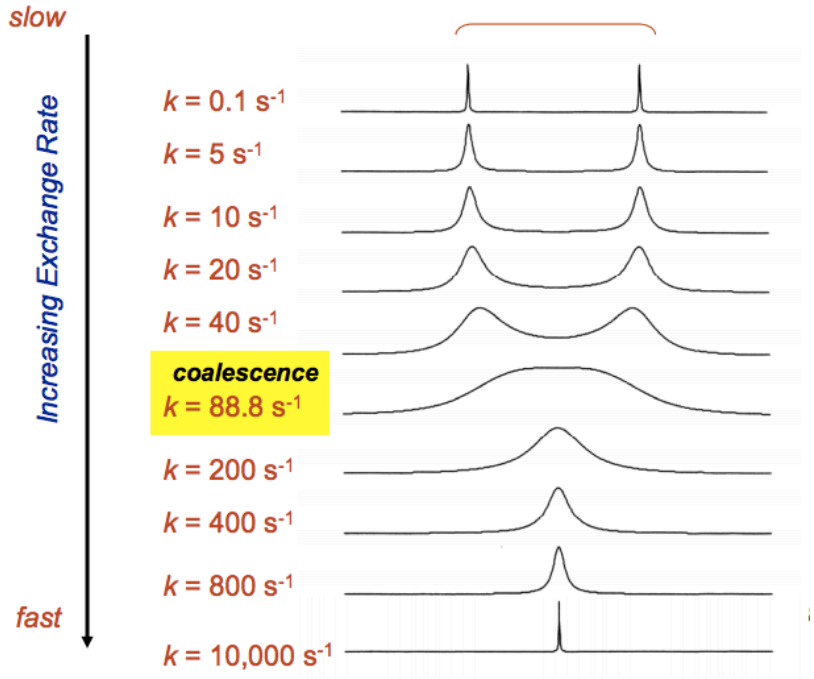
CYRANGE (backbone dihedral angle convergence)

Two-Site Exchange:
Rotation about a partial double bond in dimethylformamide

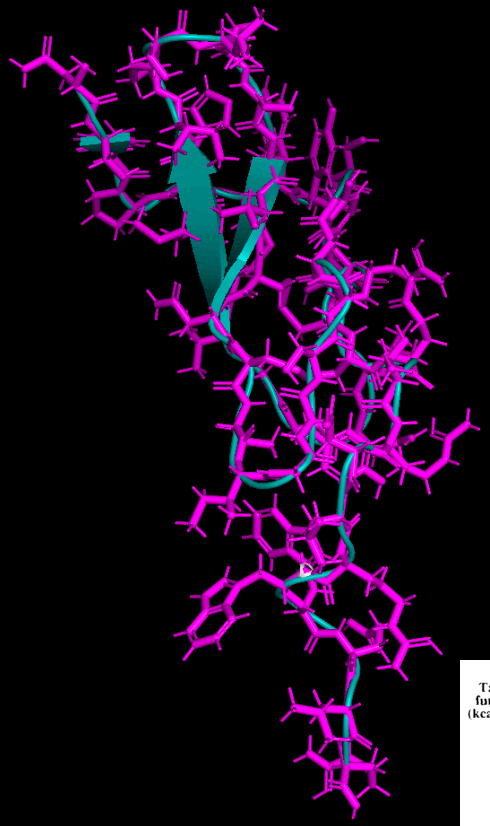
Equal Population of Exchange Sites

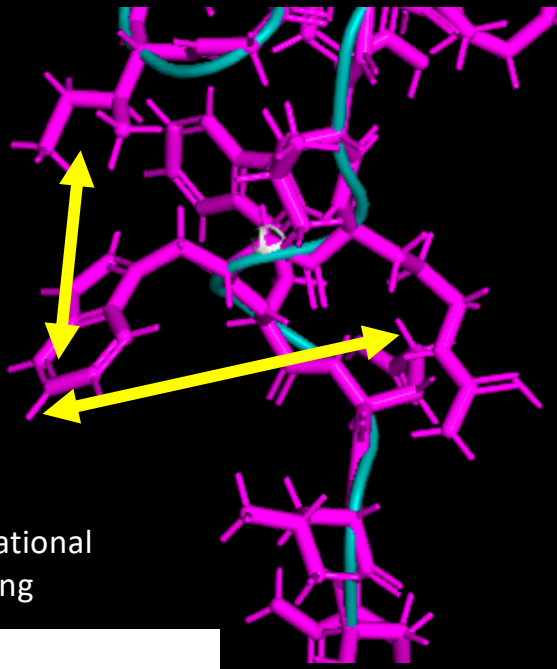


NMR is a powerful tool for studying relationships between protein dynamics and protein function.

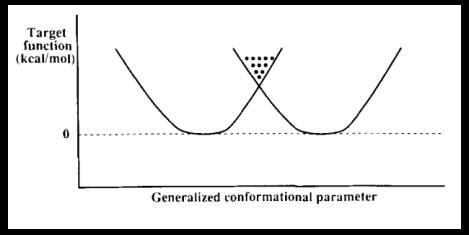


http://www.bioc.aecom.yu.edu/labs/girvlab/nmr/course





Conformational Pinning



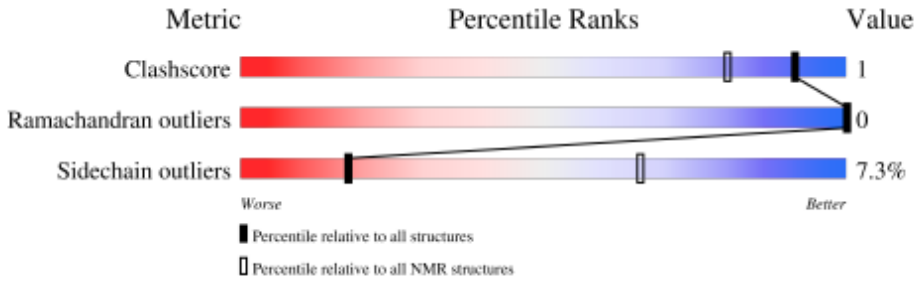
3EGF Montelione, G.T.; … Wüthrich, K.; Scheraga, H.A. Proc Natl Acad Sci 1987

Tejero, R.; Bassolino-Klimas, D.; Bruccoleri, R.E.; Montelione, G.T. **Protein Science** 1996

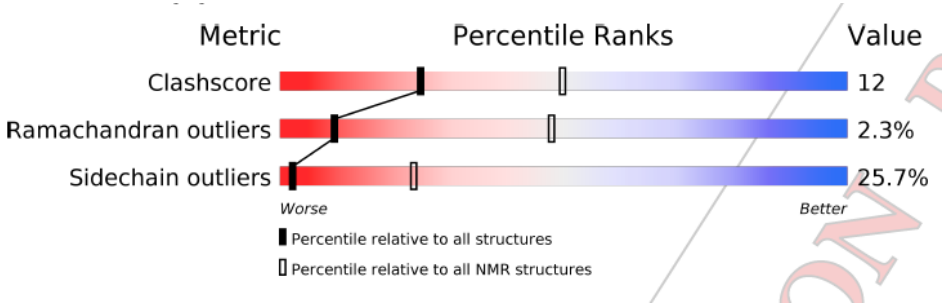
Knowledge-Based Validation for 3 CASP NMR Targets

PDB Structure Validation Server – Slider Bars

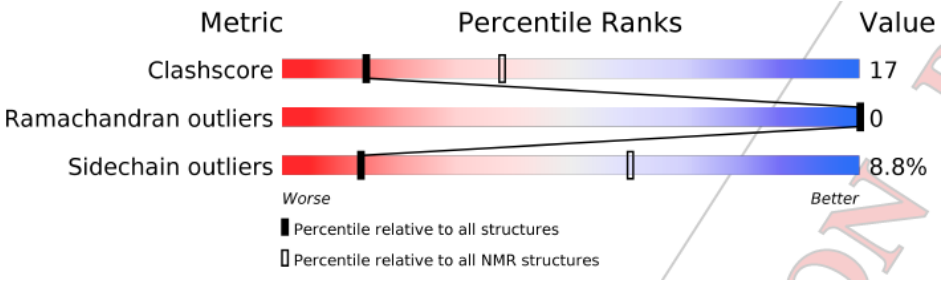
Great thanks to the labs who determined these structures and shared their NMR data !



T1029
N. Zhang
A. LiWang

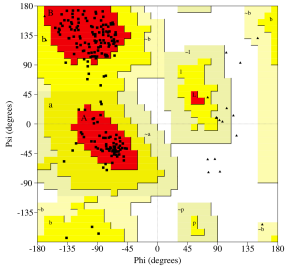


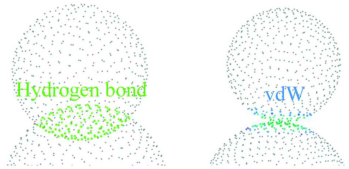
T1027
N. Kobayashi
Y. Kuroda

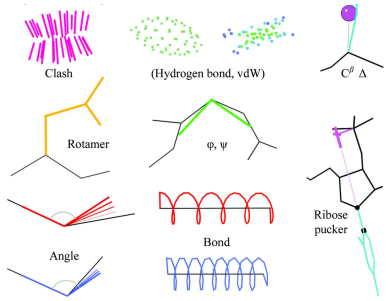


T1055
B. Bersch

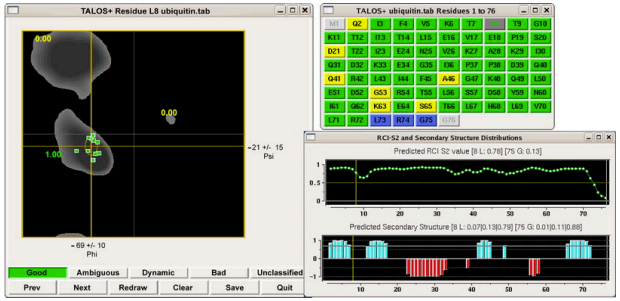
Knowledge-Based - Have I made something that looks like “a protein structure”?

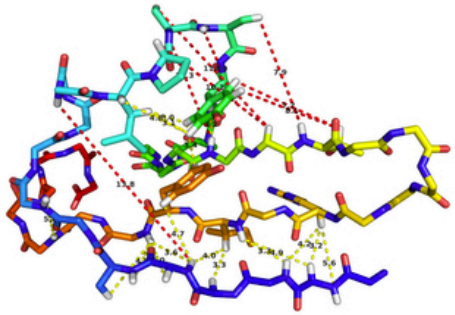






Model vs Data - Does the structure explain the experimental data?





Model vs Data Validation

NMR Restraint Analysis

Back Calculation of NOESY Peaks
(Relaxation Matrix, RPF-DP scores)

Back Calculation of Residual Dipolar Couplings

Back Calculation of Chemical Shifts

GT Montelione
CASP14 Dec 3, 2020

Model vs Data Validation

NMR Restraint Analysis

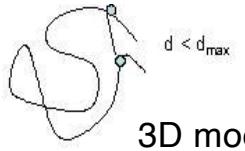
Back Calculation of NOESY Peaks
(Relaxation Matrix, RPF-DP scores)

Back Calculation of Residual Dipolar Couplings

Back Calculation of Chemical Shifts

GT Montelione
CASP14 Dec 3, 2020

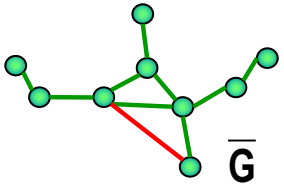
RPF-DP Scores



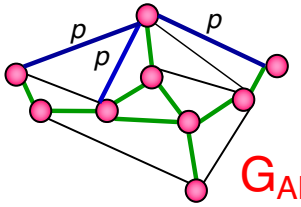
3D model

$$\text{Recall} = \frac{TP}{TP + FN} \qquad \text{Precision} = \frac{TP}{TP + FP}$$

$$F\text{-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

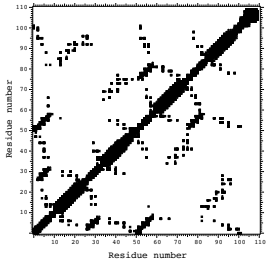


— TP
— FP
— FN
TN

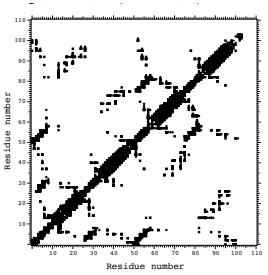


G_{ANOE}

← R and NOE

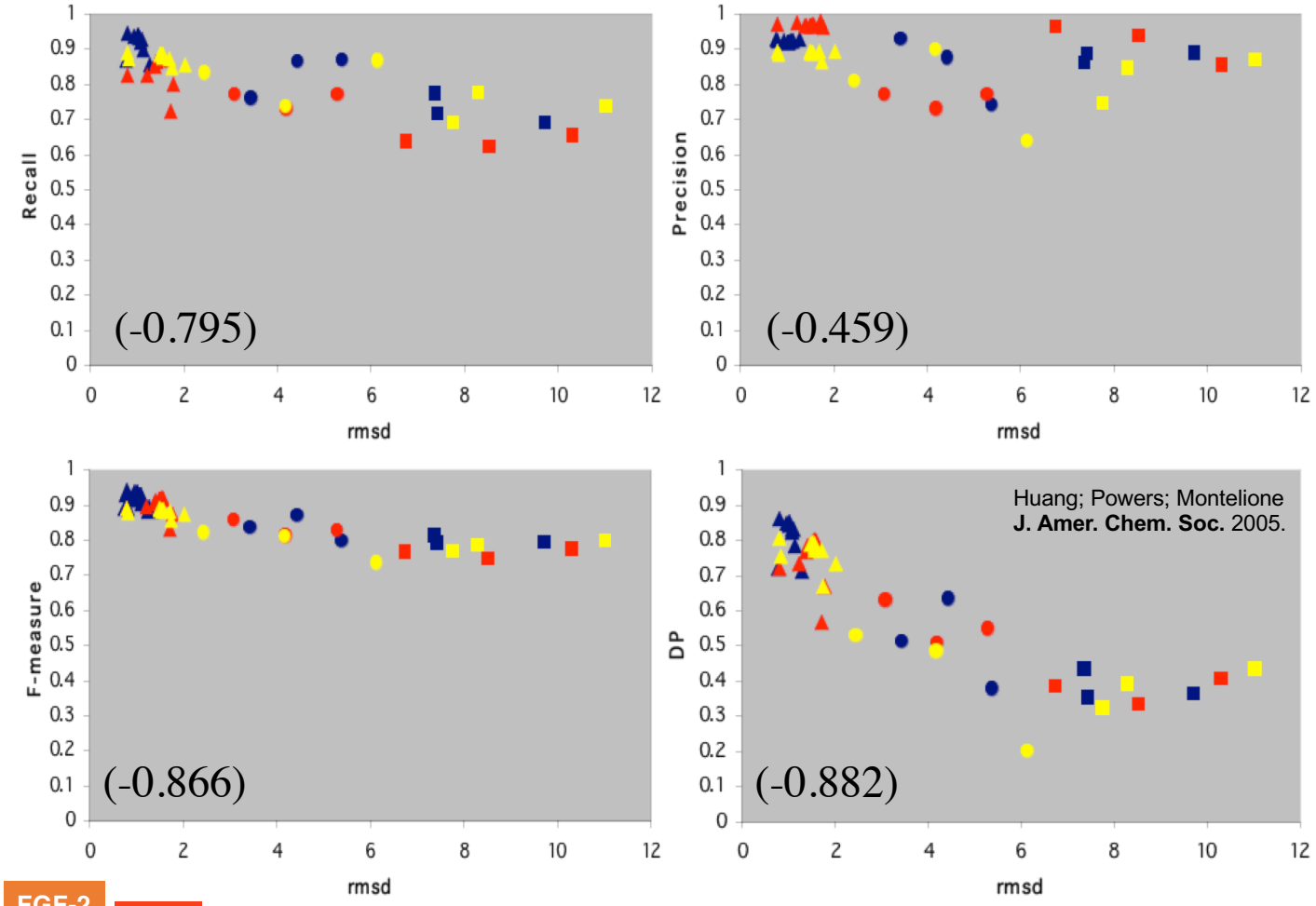


By comparing the differences between the two graphs $\overline{G}$ (derived from the structure) and G_{ANOE} (derived from the peaklists), a global measure of the goodness-of-fit of the query structures with the original peaklists can be formulated.



Similar to lDDT developed later by Schwede and coworkers

Huang, Y J ; Powers, R ; Montelione, G T **J. Amer. Chem. Soc.** 2005, 127: 1665.
Huang, Y J ; Rosato, A ; Singh, G ; Montelione, G T **Nucleic Acids Research** 2012, 40:542



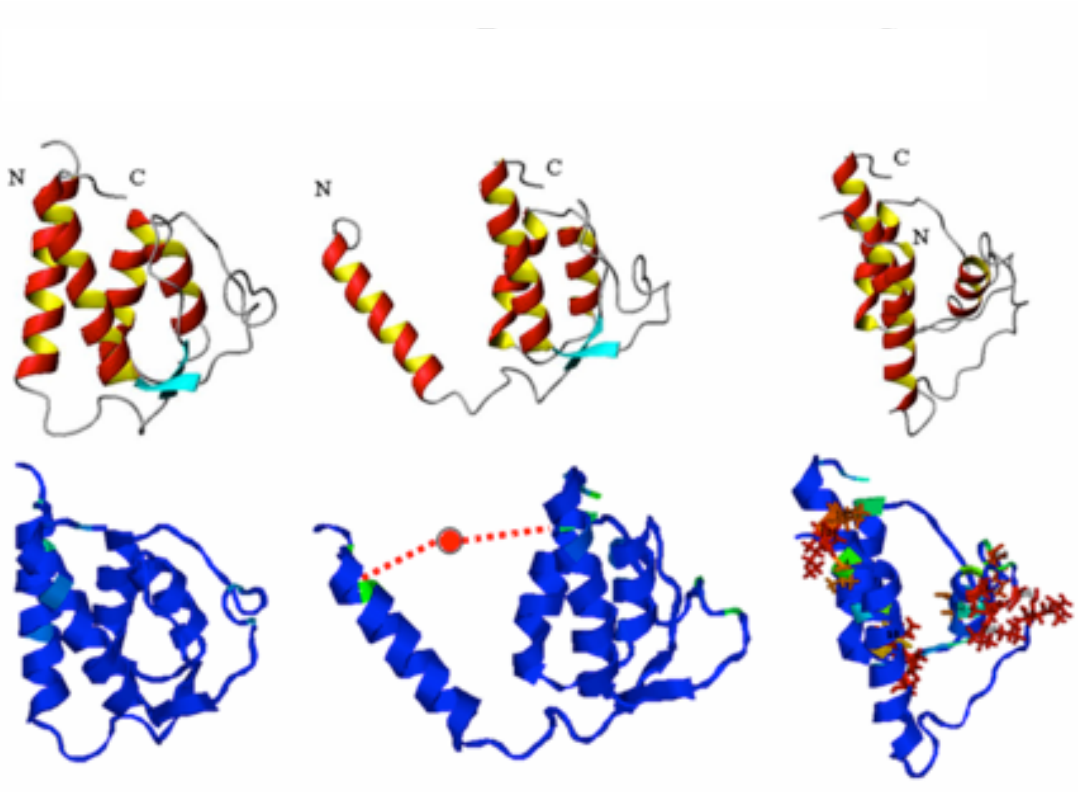
FGF-2

IL-13

MMP-1

△ < 2 Å ○ Partially correct □ Different fold

Recall and Precision Violations



Huang et al.

JACS 2005

Recall = 0.825, Precision = 0.971
F = 0.892 and DP = 0.723

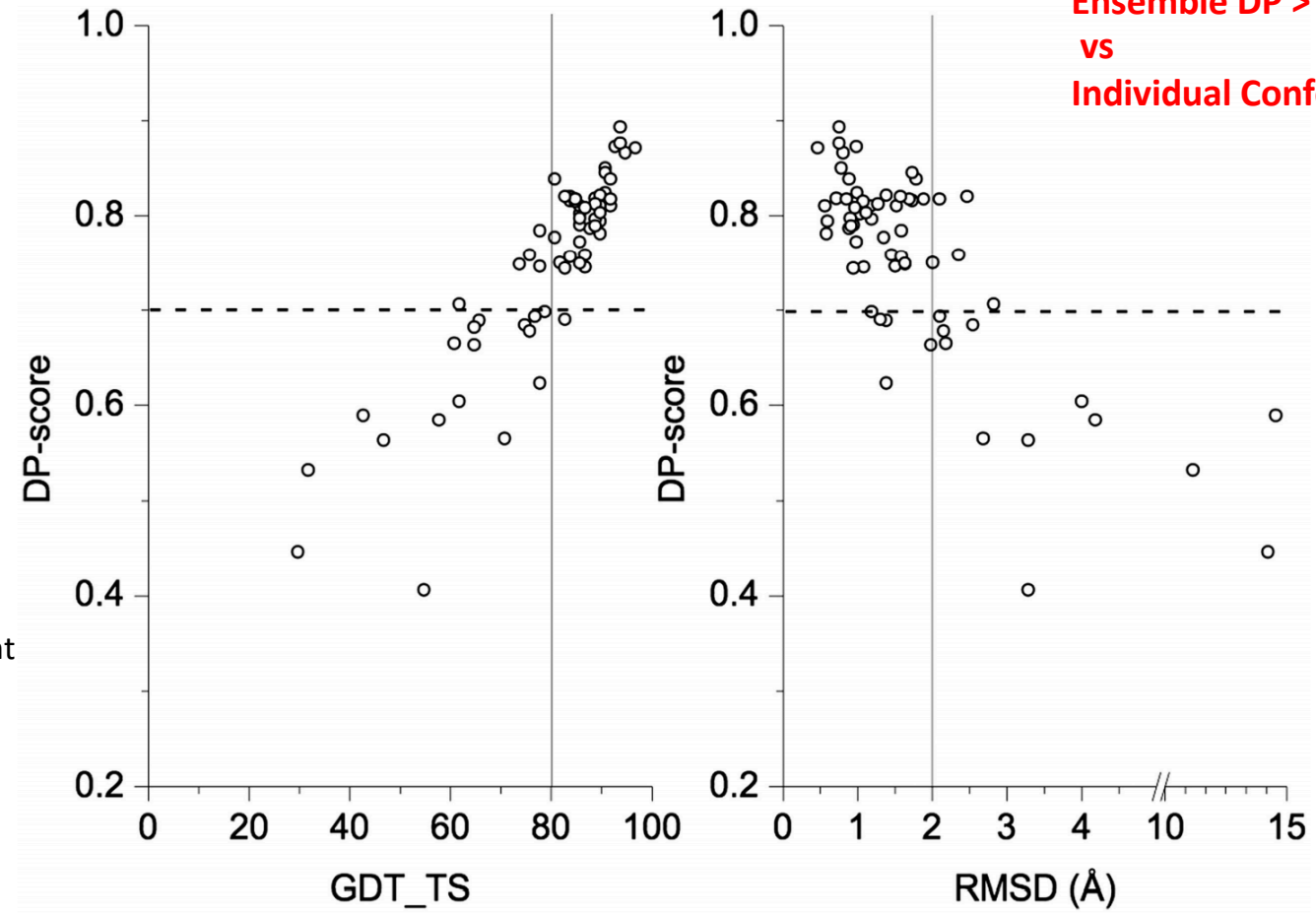
Recall Violations

Recall = 0.769, Precision = 0.969
F = 0.857 and DP = 0.629

Precision Violations

Recall = 0.729, Precision = 0.917
F = 0.812 and DP = 0.508

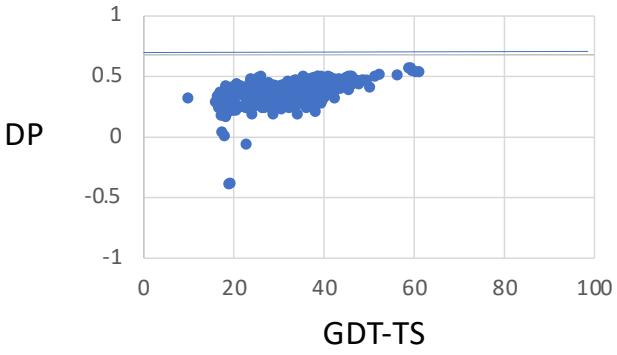
Critical Assessment of Automated Protein Structure Determination by NMR (CASD-NMR)

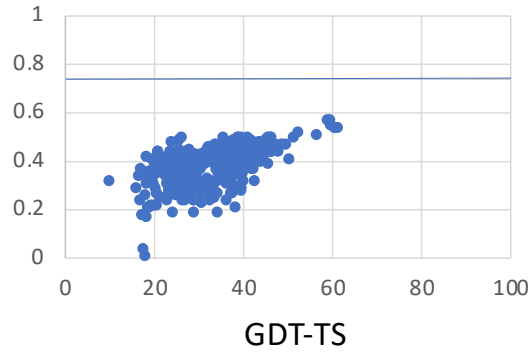


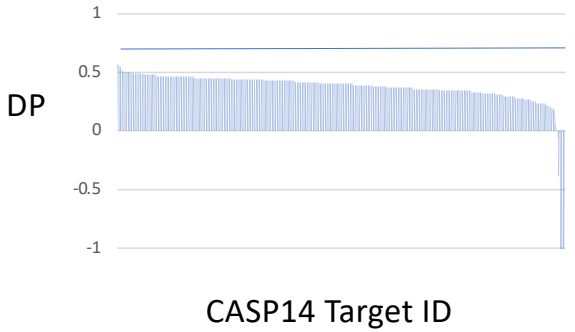
Ensemble DP > 0.7
vs
Individual Conformer DP > 0.6

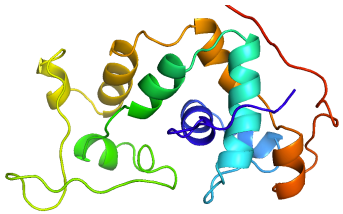
GT Montelione
CASP14 Dec 3, 2020

CASP14 Target T1027



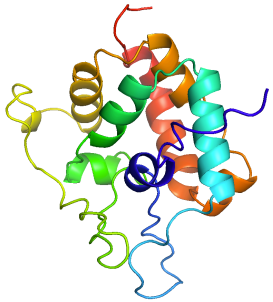




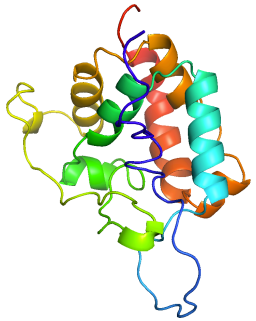


NMR_19 (DP best)

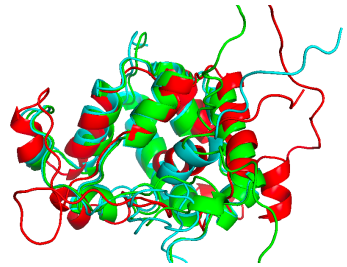
DP: 0.68 (R 0.89, P 0.80)



CASP14 427_4 (DP best)
GDT_TS: 59.3
DP: 0.57 (R 0.85, P 0.78)

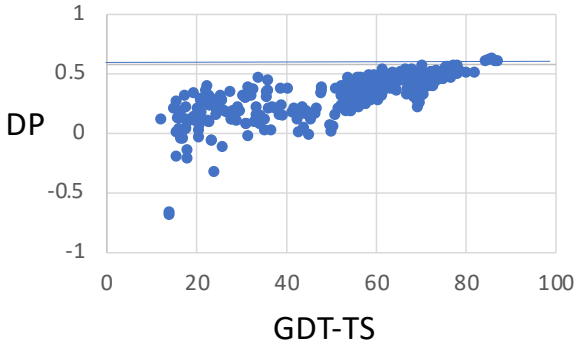


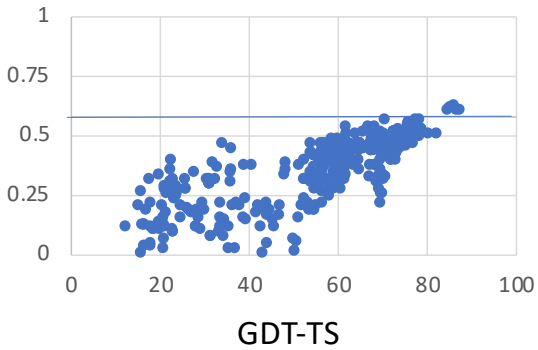
CASP14 427_1 (GDT_TS best)
GDT_TS: 61.1
DP: 0.54 (R 0.84, P 0.76)

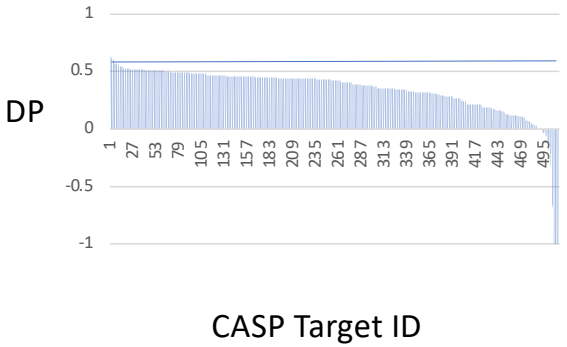


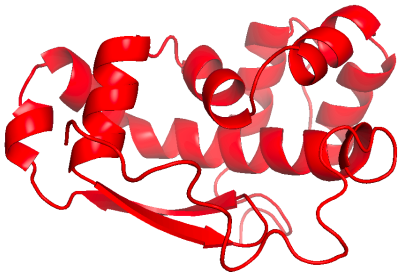
GT Montelione
CASP14 Dec 3, 2020

CASP14 Target T1055



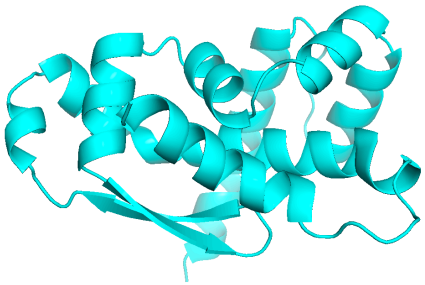




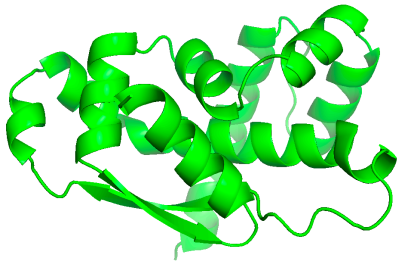


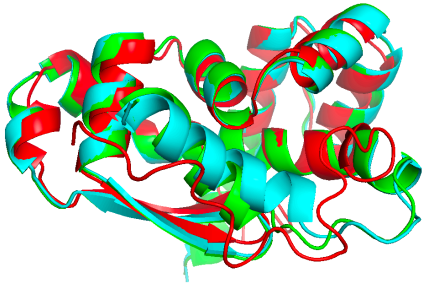
NMR (DP best)

DP: 0.58 (R 0.97, P 0.76)



CASP14 427_2 (DP best)
GDT_TS: 85.8
DP: 0.63 (R 0.96, P 0.78)

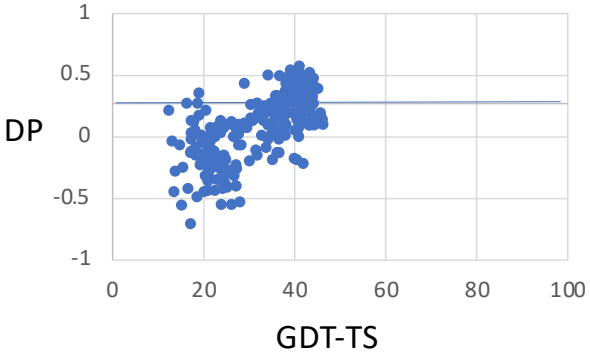


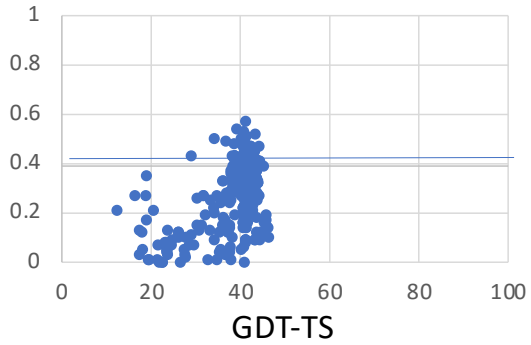


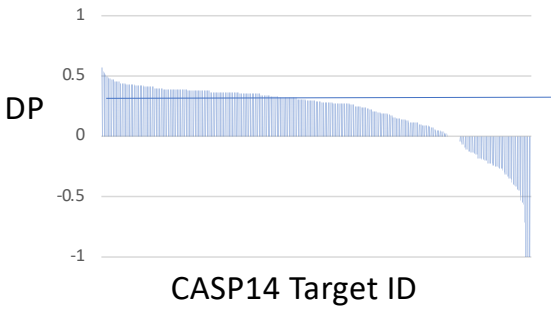
CASP14 427_3 (GDT_TS best)
GDT_TS: 87.1
DP: 0.61 (R 0.95, P 0.78)

GT Montelione
CASP14 Dec 3, 2020

CASP14 Target T1029

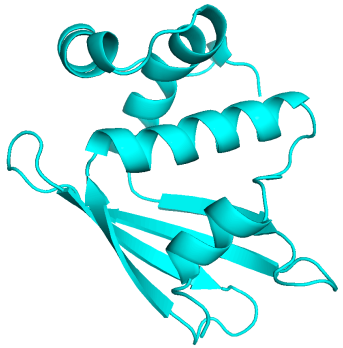




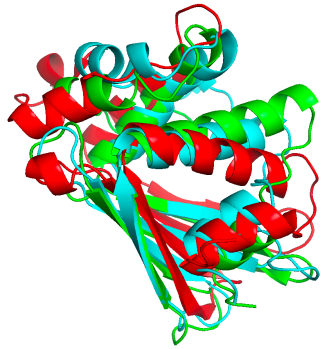


NMR (DP best)

DP: 0.27 (R 0.89, P 0.57)



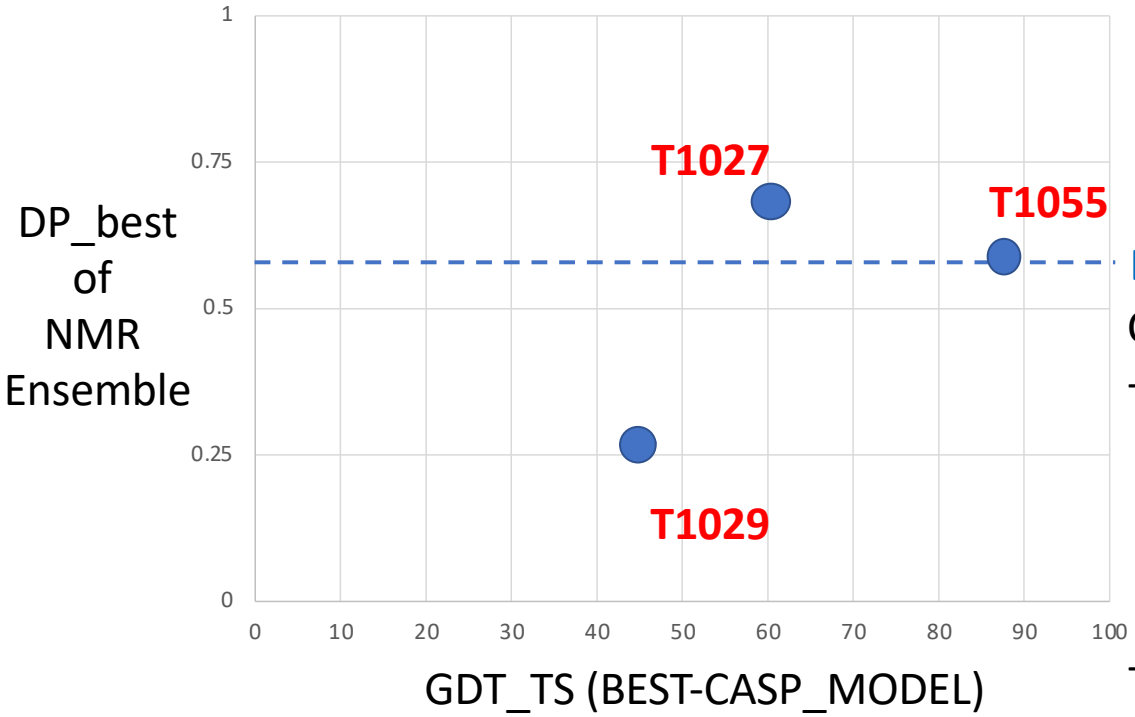
CASP14 323_4 (DP best)
GDT_TS: 41.4
DP: 0.57 (R 0.90, P 0.62)



CASP14 071_2 (GDT_TS best)
GDT_TS: 46.4
DP: 0.10 (R 0.91, P 0.54)

GT Montelione
CASP14 Dec 3, 2020

DP (NMR) vs GDT_TS (BEST_CASP_MODEL)



DP threshold

Conclusions

- Best regular prediction models can sometimes fit NMR data better than the "NMR structure" reported (as observed in CASP13)

- Highly dynamic proteins are difficult to model correctly

- CASP is advised to institute more rigorous quality control on NMR targets

T1027 168 residues
T1029 125 residues
T1055 148 residues

GT Montelione
CASP14 Dec 3, 2020

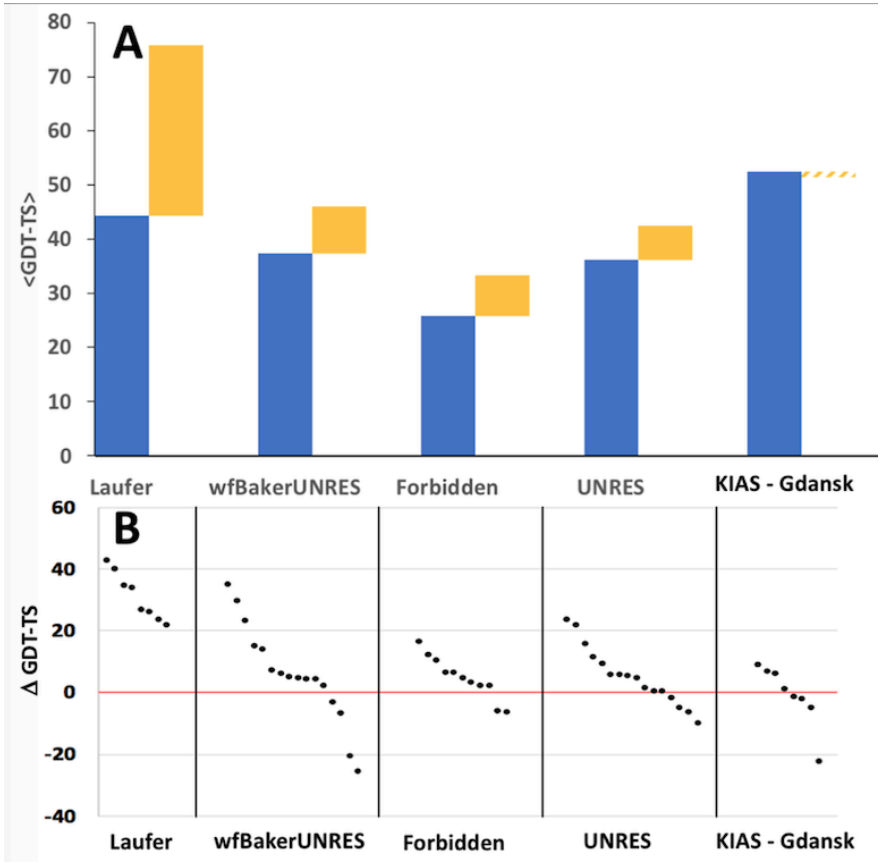
Can Sparse NMR Data Guide Structure Prediction?

GT Montelione
CASP14 Dec 3, 2020

CASP13 – NMR-Guided Modeling – Perdeuterated Proteins

11 simulated data sets
2 real data sets

In most cases where predictor submitted both regular and NMR-guided predictions, the NMR-guided models are more accurate



Sala, Huang, et al. PROTEINS 2019

CASP14 NMR Guided Prediction

No Simulated NMR Data Sets

Two Targets – Perdeuterated Integral Membrane Proteins in Micelles

- T1077 YfaZ 170 Residues Unknown Function
- T1088 MipA 238 Residues Drug Transporter

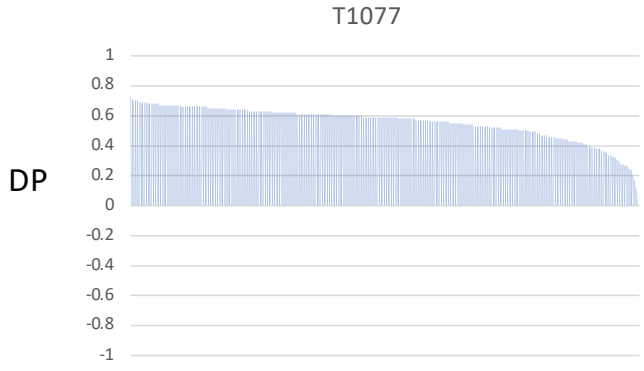
No structures available for any member of these two large pfam families

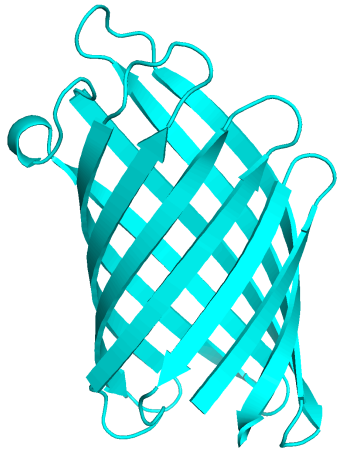
Samples prepared with ^{2}H,^{15}N,^{13}C-enrichment and ILV ^{13}C Methyl labeling in perdeuterated micelles.

Backbone and sidechain methyl resonance assignments
HN/HN and HN/Me NOEs -> Ambiguous Contacts -> 9 CASP Predictor Groups
Chemical Shifts -> Backbone dihedral angle ranges -> 9 CASP Predictor Groups
EC contacts (C. Sander & K. Brock) -> EC-NMR structures

GT Montelione
CASP14 Dec 3, 2020

Target T1077
YfaZ in detergent micelles
EC-NMR
Perdeuterated Sparse Restraints

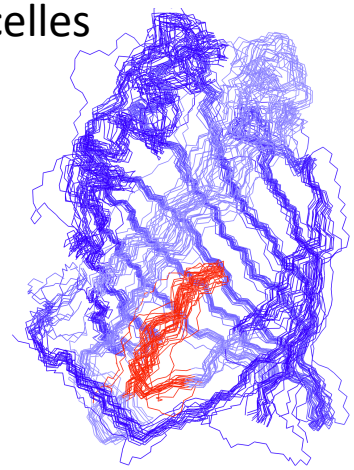




DP = 0.71

GT Montelione
CASP14 Dec 3, 2020

Target 1088 MipA in detergent micelles
EC-NMR
Perdeuterated Sparse Restraints
~ 1000 Conformational Restraints
~ 4.2 restraints / residue

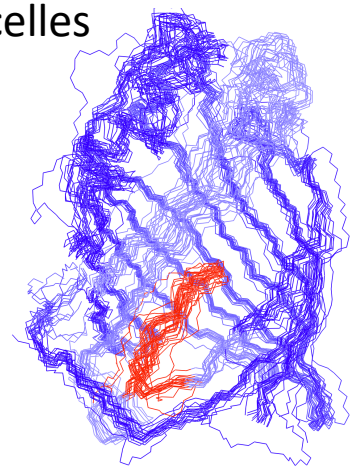


Extensive exchange broadening in red hairpin; do not see many HN-HN NOEs

Chemical shift data indicate hairpin is not beta strand -> rather suggests dynamic structure

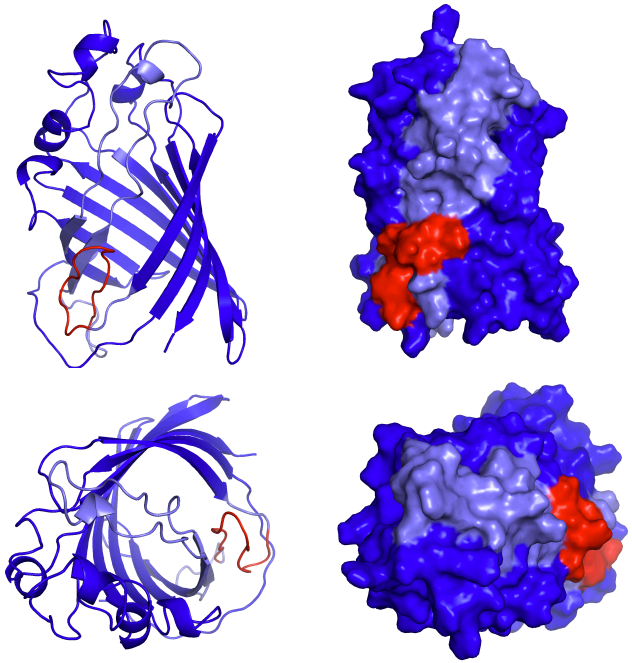
GT Montelione
CASP14 Dec 3, 2020

Target 1088 MipA in detergent micelles
EC-NMR
Perdeuterated Sparse Restraints
~ 1000 Conformational Restraints
~ 4.2 restraints / residue

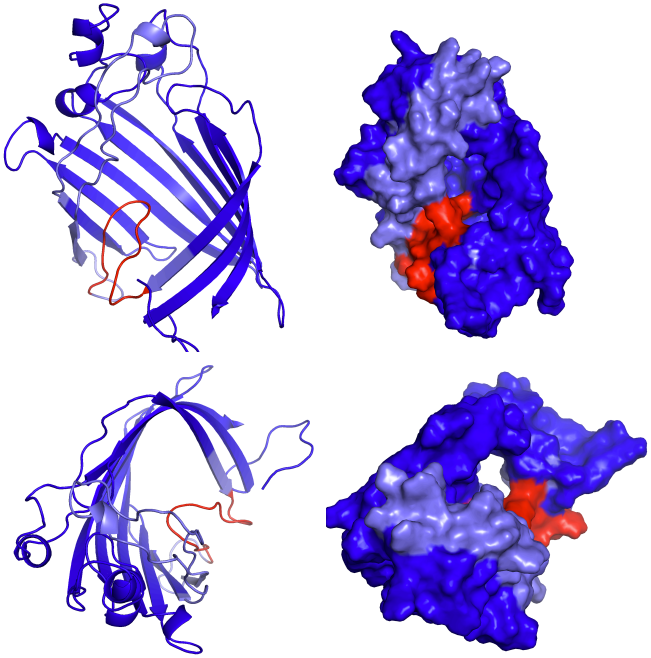


Extensive exchange broadening in red hairpin; do not see many HN-HN NOEs

Chemical shift data indicate hairpin is not beta strand -> rather indicate a dynamic local structure

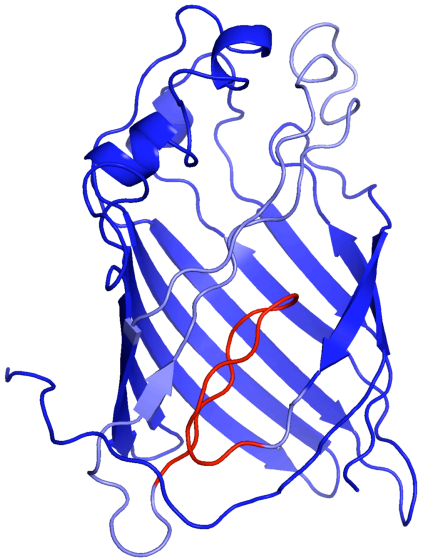


NMR_10 Closed Conformer
DP = 0.53



NMR_13 Open Conformer
DP = 0.49

GT Montelione
CASP14 Dec 3, 2020

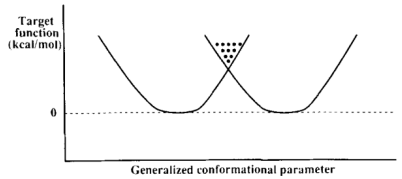


T1088

MipA – Integral Membrane Protein – Multi Drug Transporter

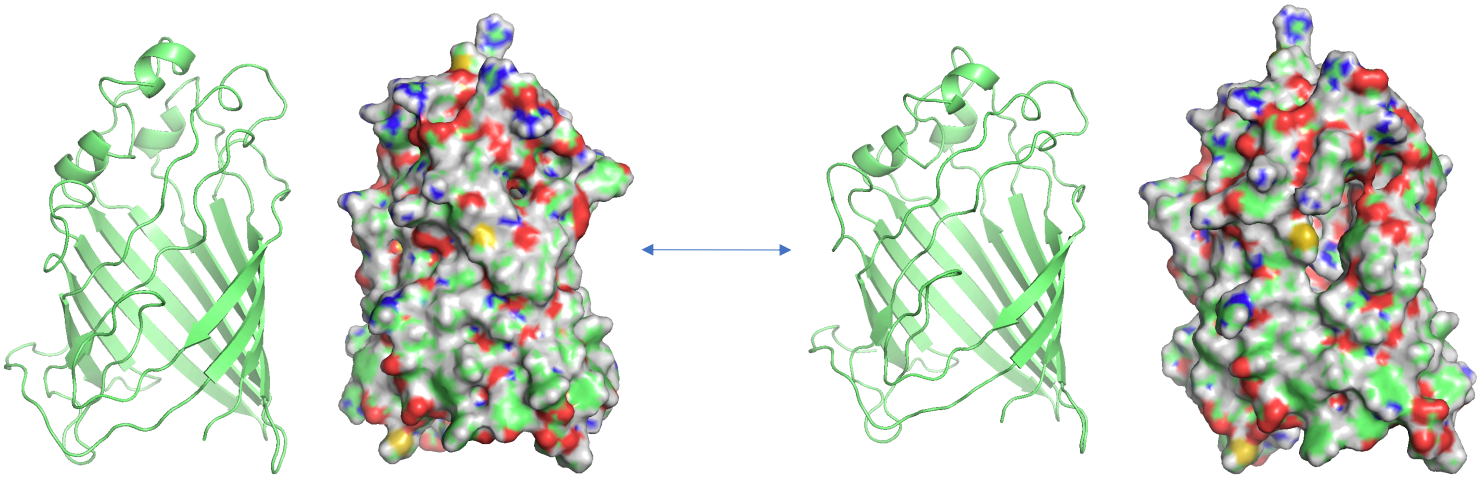
NMR data indicate exchange broadening in red hairpin; 13C chemical shifts inconsistent with static beta-strands

Experimental NMR structures – each conformer is best fit to all of the data.



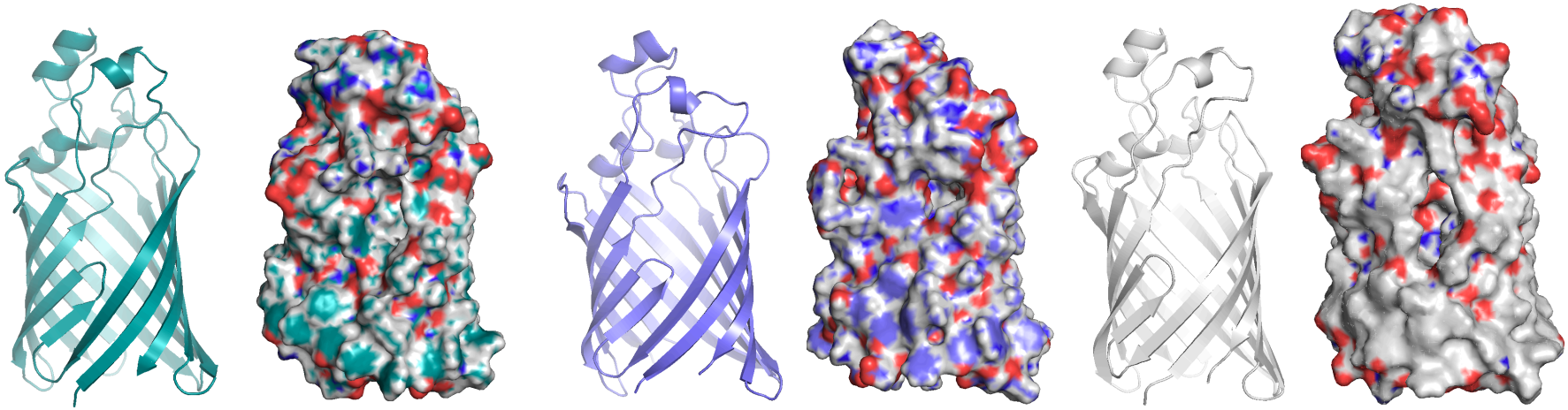
Need to instead model the distribution that is best fit to the ensemble- averaged NMR data

GT Montelione
CASP14 Dec 3, 2020



EC-NMR closed form

EC-NMR open form

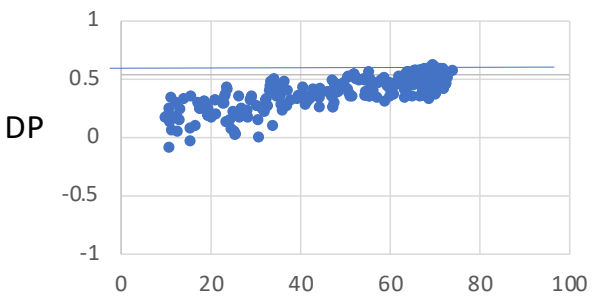


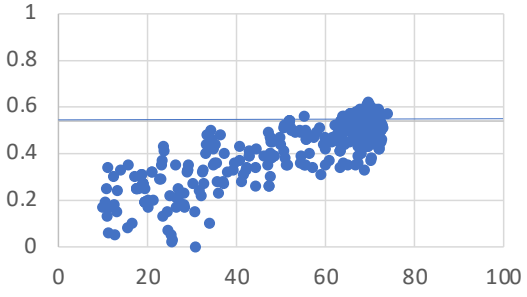
T1088 362_2 closed form

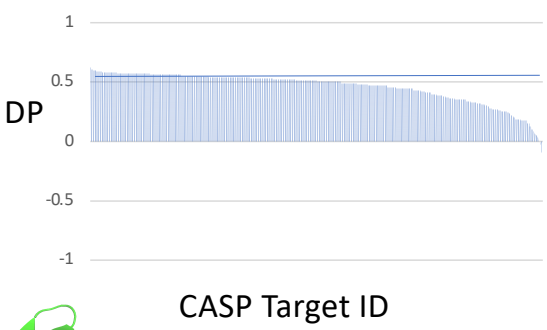
T1088 314_5 smaller cavity

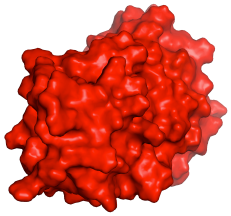
T1088 226_5 open form

GT Montelione
CASP14 Dec 3, 2020



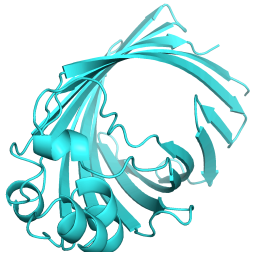


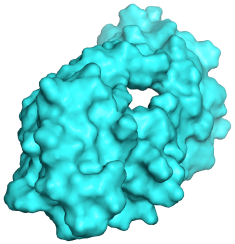




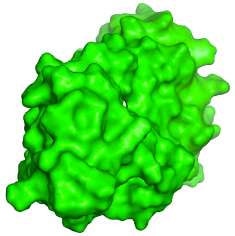
NMR_8 Closed

DP: 0.53 (R 0.73, P 0.63)

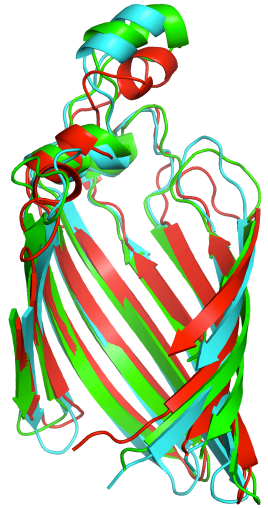




CASP14 226_5 (DP best)
GDT_TS: 69.5
DP: 0.62 (R 0.75, P 0.73)



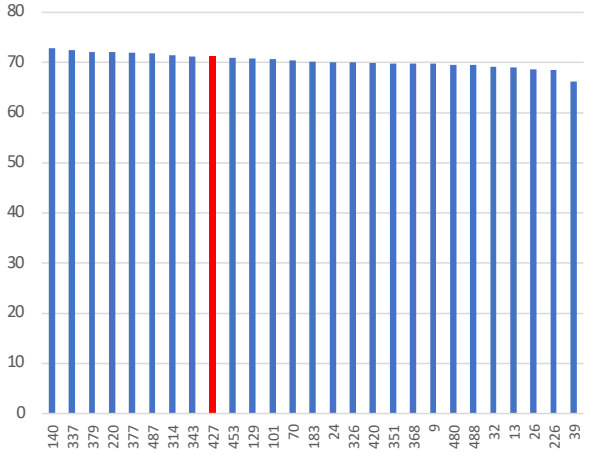
CASP14 314_5 (GDT_TS best)
GDT_TS: 73.9
DP: 0.57 (R 0.73, P 0.68)



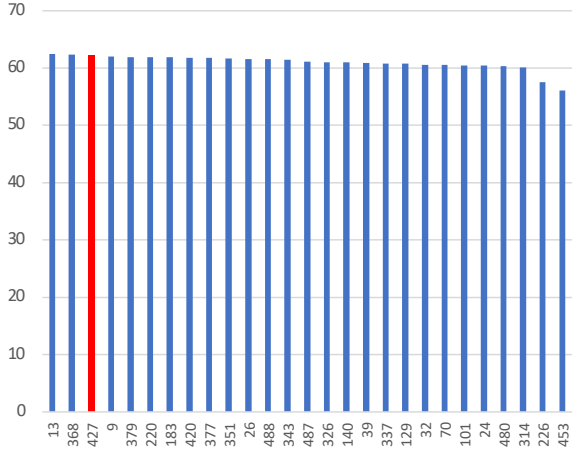
GT Montelione
CASP14 Dec 3, 2020

T1088 Ranking – Regular Prediction (top 27 groups)

GDT_TS_First

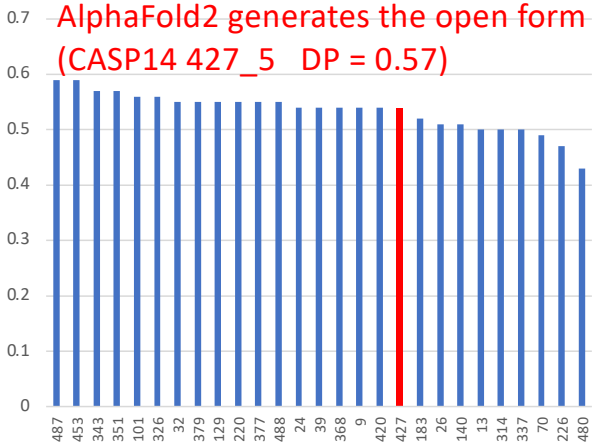


Combined_First



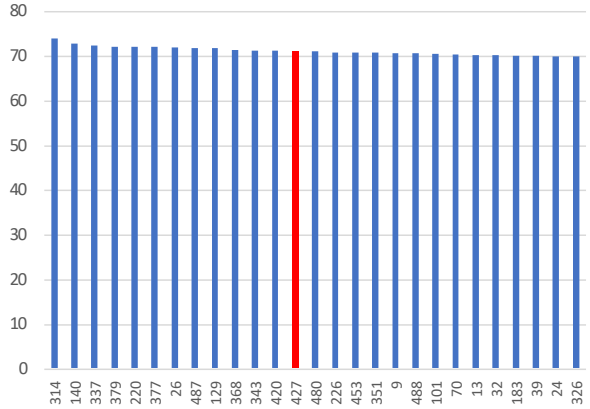
DP_First

AlphaFold2 generates the open form (CASP14 427_5 DP = 0.57)

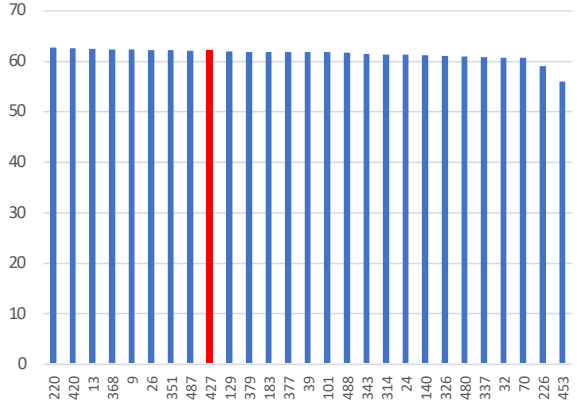


GDT-HA*0.4 + GDT-SC*0.4 + RPF * 0.4 + SphereGrinder * 0.4 + CAD-AA * 0.4 + MolProbity * 0.2)/2.2

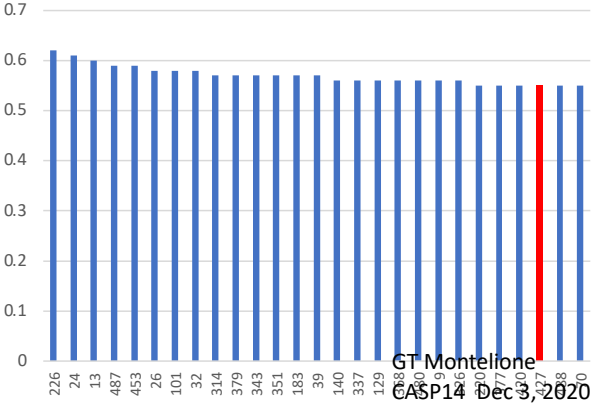
GDT_TS_BEST



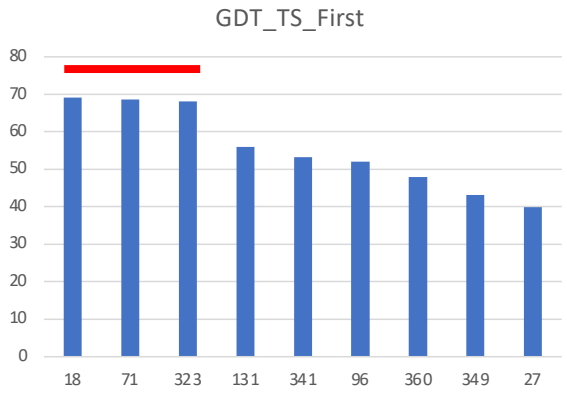
Combined_Best

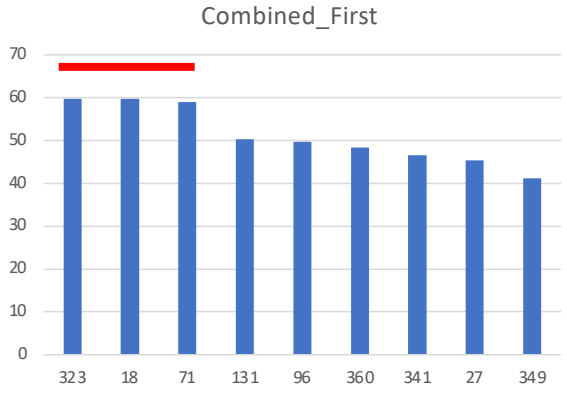


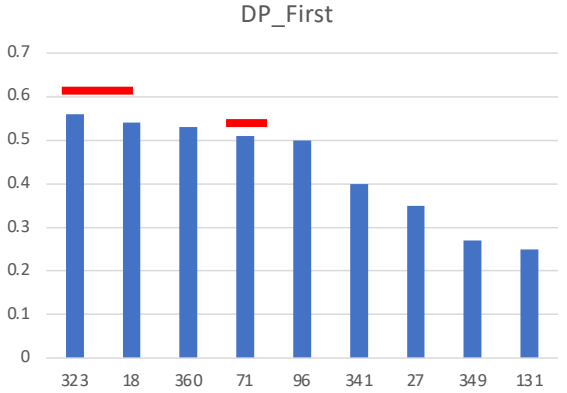
DP_Best



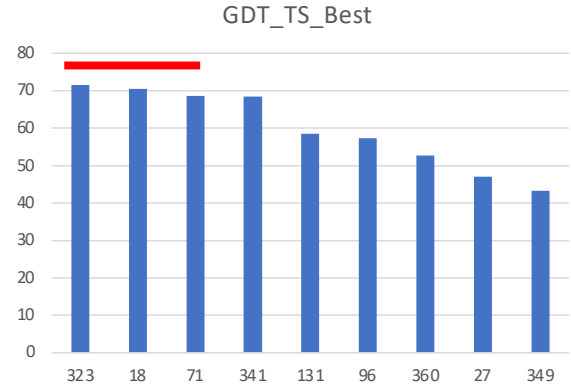
NMR Guided Predictions
N1088 Ranking - Nine Prediction Groups

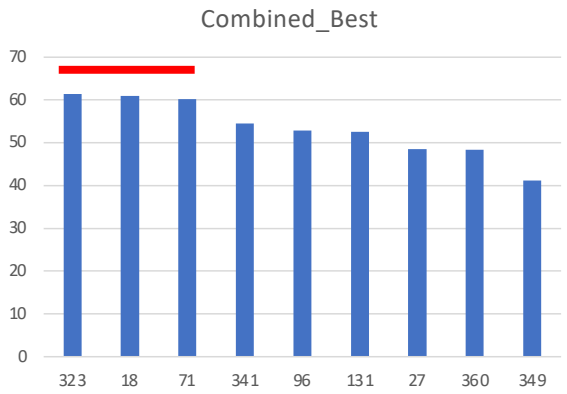


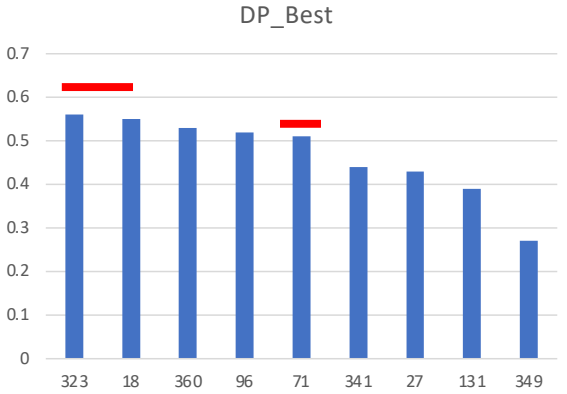




GDT-HA*0.4 + GDT-SC*0.4 + RPF * 0.4 + SphereGrinder * 0.4 + CAD-AA * 0.4 + MolProbity * 0.2)/2.2

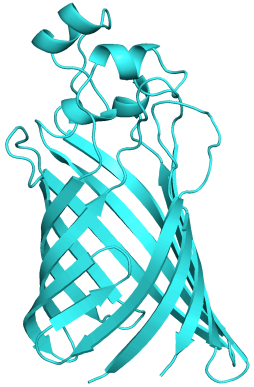




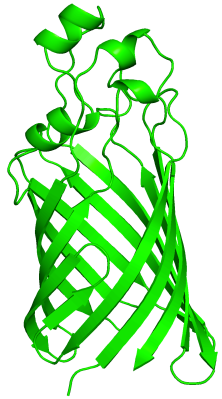


GT Montelione
CASP14 Dec 3, 2020

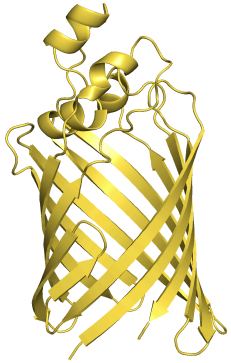
CASP14 NMR-Guided Prediction Results



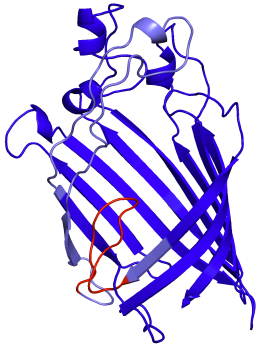
N1088_323_1
DP = 0.56
GDT_TS = 68.0



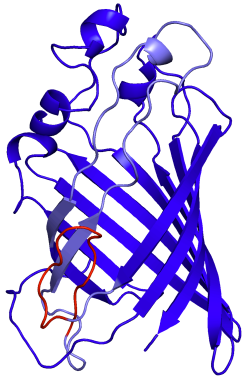
N1088_018_2
DP = 0.55
GDT_TS = 70.4



N1088_071_1
DP = 0.51
GDT_TS = 68.6



EC-NMR – open
DP = 0.49



EC-NMR – closed
DP = 0.53

Note that the best DP score for all CASP models is regular prediction result
DP: 0.62 CASP14 226_5

For NMR-Assisted Target N1088 - the 3 top groups are:
323 DellaCorteLab
18 UNRES-template
71 Kihara Lab

Experimental NMR structures – each conformer is best fit to all of the data. Can we use CASP models to segregate the data to more accurately determine the structures of each state?

GT Montelione
CASP14 Dec 3, 2020

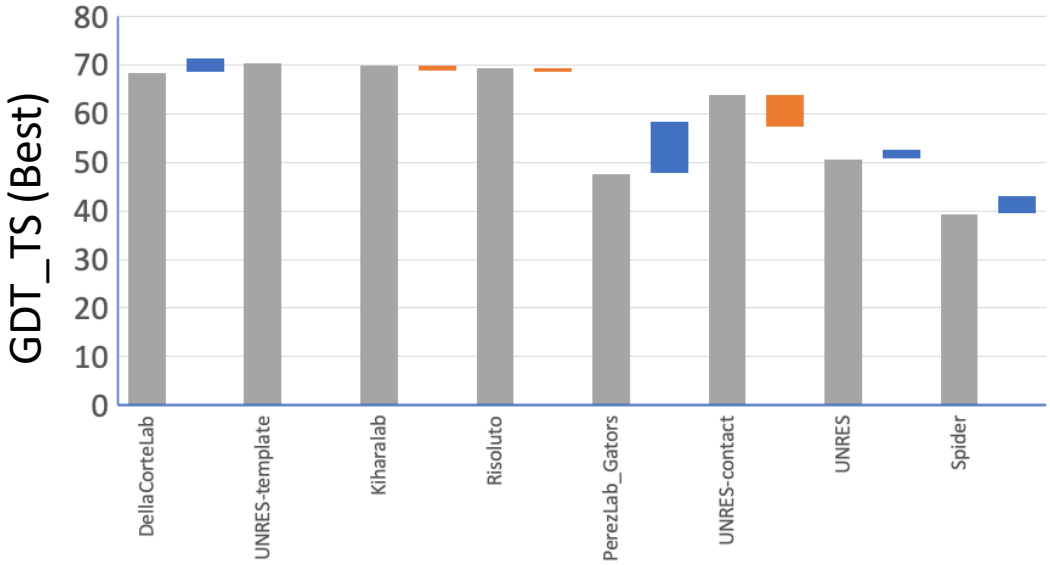
CASP14 – NMR-Guided Modeling – Perdeuterated Integral Membrane Protein

N1088 MipA 238 residues

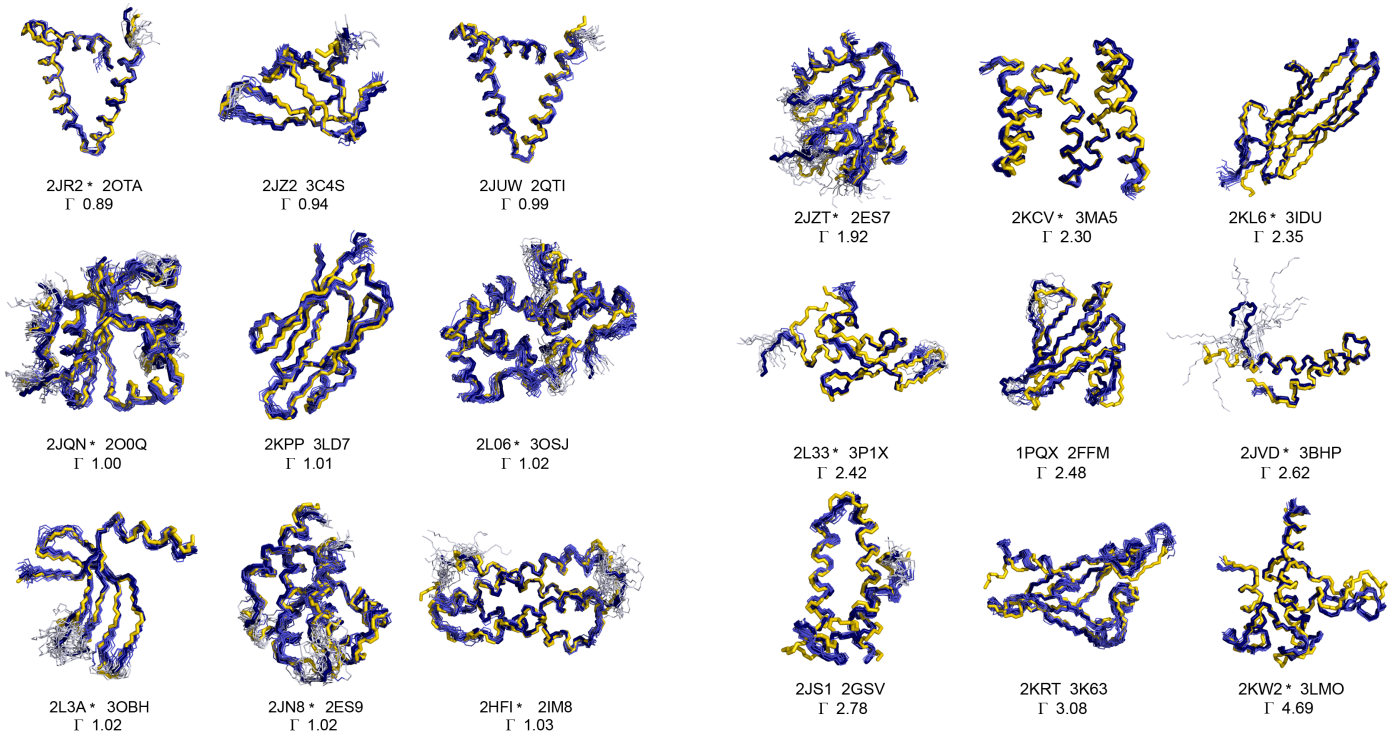
4 groups – NMR data improved prediction accuracy, as much as 10 GDT points

1 group – NMR data degraded prediction accuracy

3 groups – NMR data did not much impact accuracy



41 NMR X-ray Pairs



Structure

Ways & Means

Improved Technologies Now Routinely Provide Protein NMR Structures Useful for Molecular Replacement

Binchen Mao,[1] Rongjin Guan,[1] and Gaetano T. Montelione[1,2,*]

Structure 2011

Conclusions

Integral membrane protein N1088 MipA functions as a drug transporter to confer antibiotic resistance.

NMR data indicate dynamic equilibrium with fast / intermediate exchange between closed and open forms. Regular CASP Prediction models include both closed and open forms.

Some predictors could use NMR data to improve their models (PerezLab)

Some NMR-Guided Prediction models fit to NMR NOESY data better than sparse EC-NMR structure (DellaCorte, UNRES-template, UNRES)

Some Regular Prediction models fit to NMR NOESY data better than any NMR-guided models → could be used to guide data analysis