CASP13: Sparse Experimental Data Assisted Prediction in CASP

J. Y. Huang, Y. Ishida, G. T. Montelione, N Denissova,
G. Liu, A. Rosato, D. Sala, D. Snyder, G. V. T. Swapna,
R. Tejero, H. Valafar

G. Hura, J. Tainer, S. Tsutakawa

Fiser, A. Leitner, J. Rappsilber

J. Duarte, C. Seidel

K. Fidelis, A. Kryshtafovych

Dec 2012: Proposal for CASP11
Contact Assisted Prediction

Contacts could be sparse, experimentally accessible distances:

- chemical cross links (Mass Spec)
- backbone NH – NH and or ILV
  Me-Me contacts (< 6.5 Å, ^{2}H proteins)
- Paramagnetic Relaxation Enhancement (PRE) (15 – 30 Å)

Methods will be developed that use realistic types of contacts that can potentially be obtained on larger (20 – 80 kDa) proteins

CASP project will drive the experimental community to generate such contact data and to collaborate with CASP methods developers on specific projects

PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Assessment of CASP11 contact-assisted predictions

Lisa N. Kinch,[1]* Wenlin Li,[2,3] Bohdan Monastyrskyy,[4] Andriy Kryshtafovych,[4] and Nick V. Grishin[1,2,3]

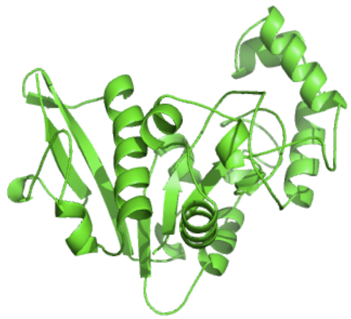
RESEARCH ARTICLE

WILEY PROTEINS

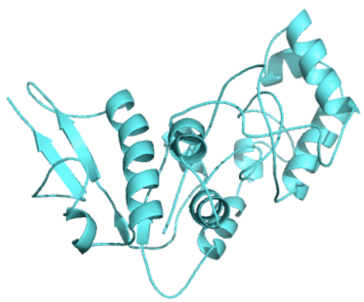
Assessment of data-assisted prediction by inclusion of crosslinking/mass-spectrometry and small angle X-ray scattering data in the 12th Critical Assessment of protein Structure Prediction experiment

Giorgio E. Tamò[1,2] | Luciano A. Abriata[1,2] | Giulia Fonti[1,2] | Matteo Dal Peraro[1,2]

Some CASP 11 ‘Predictors’ did better than standard ASDP NMR Methods



256 residues



EC-NMR
GDT: 0.61
RMSD 2.6 Å

ASDP
GDT: 0.49
RMSD: 3.6 Å

General				LGA Sequ
Model	GR#	GR Name	Charts	GDT_TS
Ts806TS038_1	038 s	nns	A D I G	76.66
Ts806TS044_1	044	LEER	A D I G	76.17
Ts806TS169_1	169	LEE	A D I G	76.17
Ts806TS064_1	064	BAKER	A D I G	71.39
Ts806TS276_1	276	FLOUDAS_A4	A D I G	34.38
Ts806TS065_1	065	Jones-UCL	A D I G	27.93
Ts806TS041_1	041 s	MULTICOM-NOVEL	A D I G	24.61
Ts806TS479_1	479 s	RBO_Aleph	A D I G	19.43
Ts806TS287_1	287	RBO-Human	A D I G	19.43
Ts806TS162_1	162	McGuffin	A D I G	18.85
Ts806TS420_1	420 s	MULTICOM-CLUSTER	A D I G	17.38
Ts806TS345_1	345 s	FUSION	A D I G	16.11
Ts806TS357_1	357	STAP	A D I G	12.79
Ts806TS032_1	032	Legato	A D I G	12.40
Ts806TS080_1	080	MeilerLab	A D I G	11.13
Ts806TS219_1	219	Sternberg	A D I G	9.28

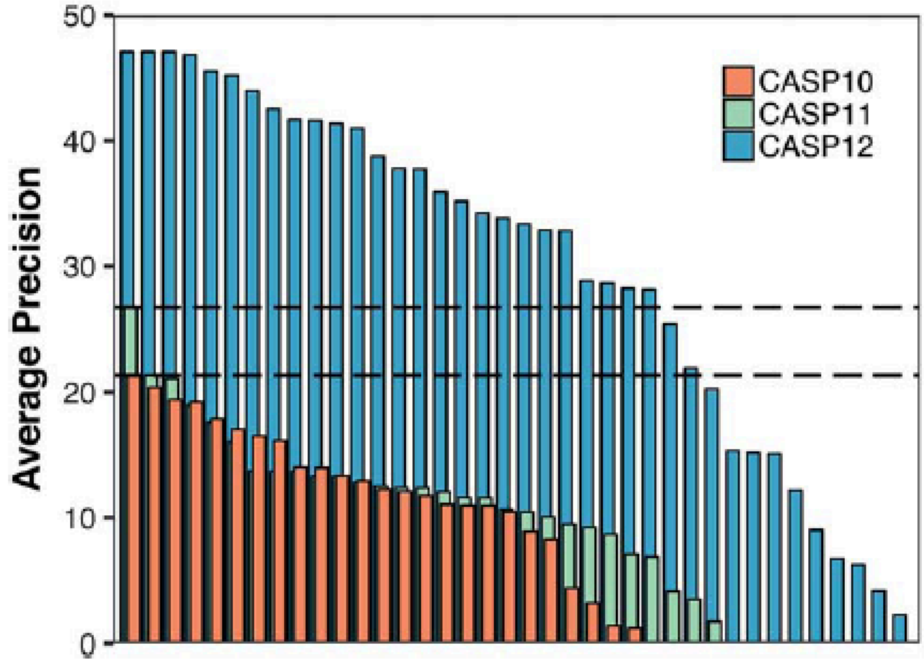


FIGURE 2 Average precision of long range contacts on L/5 lists for free modeling targets in CASP10 (red), CASP11 (green), and CASP12 (blue) sorted by rank. Grey dashed lines indicate the levels of the best performing group in CASP10 and CASP11, respectively. While only one group showed a significantly better average precision than all the others in CASP 11 compared to CASP10, 26 groups showed an improved average precision in CASP12 compared to the best performing group of CASP11

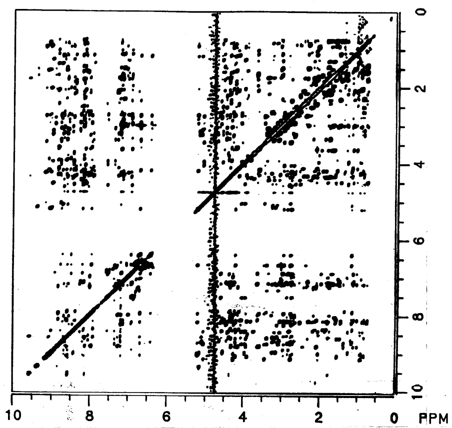
The idea of "more realistic contacts based on what can be obtained by experiments" has been superseded by the advances since 2012 in contact prediction.

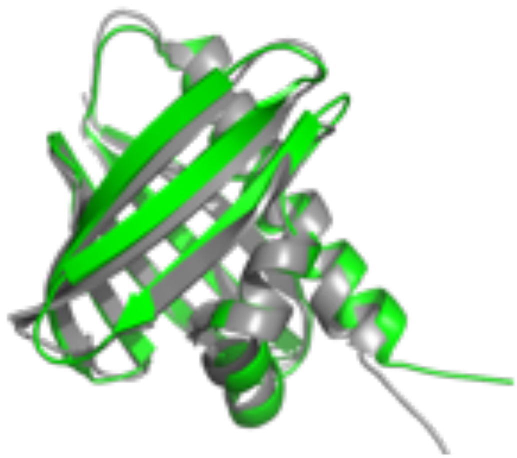
No need to have a CASP category for how well modelers can do with "simulated contacts".

Bonvin et al, PROTEINS 2017

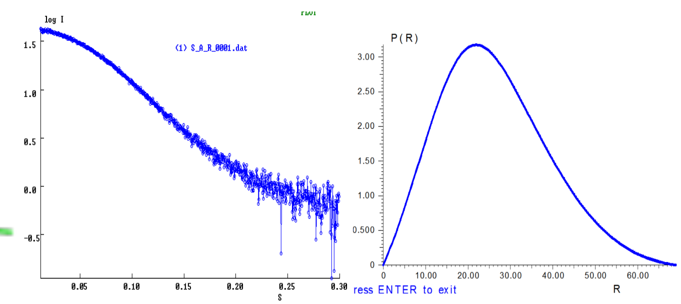
Vision: Combine simple, rapidly obtained experimental data with advance modeling methods to provide accurate 3D structures of proteins

Nuclear Magnetic Resonance (NMR) Data

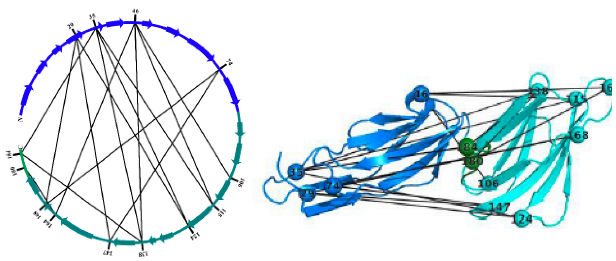


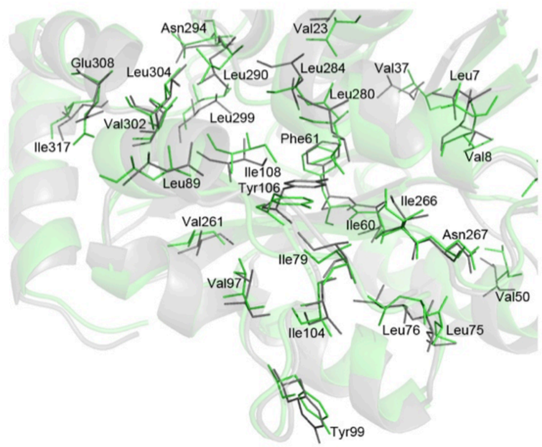


Small Angle X-ray Scattering (SAXS) or SANS

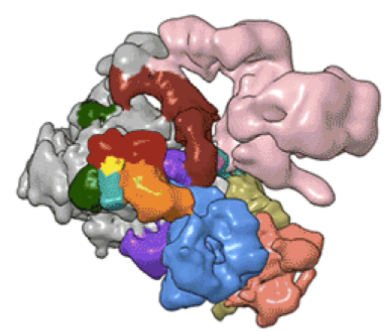


Cross-link or FRET Data



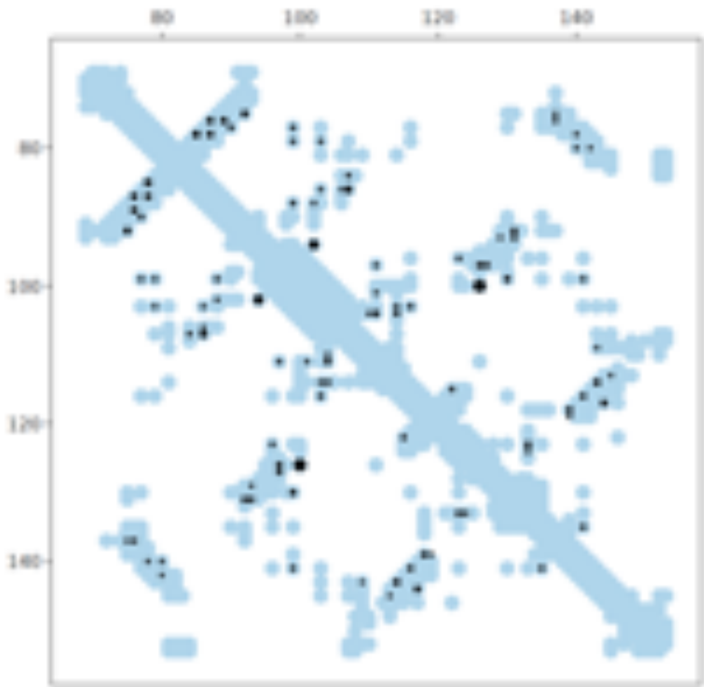


Low Resolution cryoEM

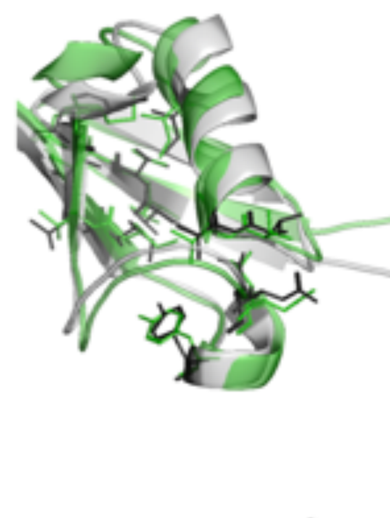


The Sparse Data Problem

48 constraints for 84 residues
HN-HN, HN-Me, Me-Me



Contact Map



Sparse Experimental Data Assisted Prediction in CASP13

How can we combine sparse experimental data with advanced modeling methods for determining accurate structures of proteins and their complexes?

Does the experimental data improve the accuracy of the predicted model?

Do predictors using sparse data provide higher accuracy models than the *best* non-data-assisted predictors?

How is the ranking of data-assisted predictors affected if we assess against data rather than reference structure?

How can we model distributions of conformations?

Sparse Experimental Data Assisted Prediction in CASP13

How can we combine sparse experimental data with advanced modeling methods for determining accurate structures of proteins and their complexes?

Does the experimental data improve the accuracy of the predicted model?

Do predictors using sparse data provide higher accuracy models than the *best* non-data-assisted predictors?

How is the ranking of data-assisted predictors affected if we assess against data rather than reference structure?

How can we model distributions of conformations?

Sparse Experimental Data Assisted Prediction in CASP13

How can we combine sparse experimental data with advanced modeling methods for determining accurate structures of proteins and their complexes?

Does the experimental data improve the accuracy of the predicted model?

Do predictors using sparse data provide higher accuracy models than the *best* non-data-assisted predictors?

How is the ranking of data-assisted predictors affected if we assess against data rather than reference structure?

How can we model distributions of conformations?

Sparse Experimental Data Assisted Prediction in CASP13

How can we combine sparse experimental data with advanced modeling methods for determining accurate structures of proteins and their complexes?

Does the experimental data improve the accuracy of the predicted model?

Do predictors using sparse data provide higher accuracy models than the *best* non-data-assisted predictors?

How is the ranking of data-assisted predictors affected if we assess against data rather than reference structure?

How can we model distributions of conformations?

Sparse Experimental Data Assisted Prediction in CASP13

How can we combine sparse experimental data with advanced modeling methods for determining accurate structures of proteins and their complexes?

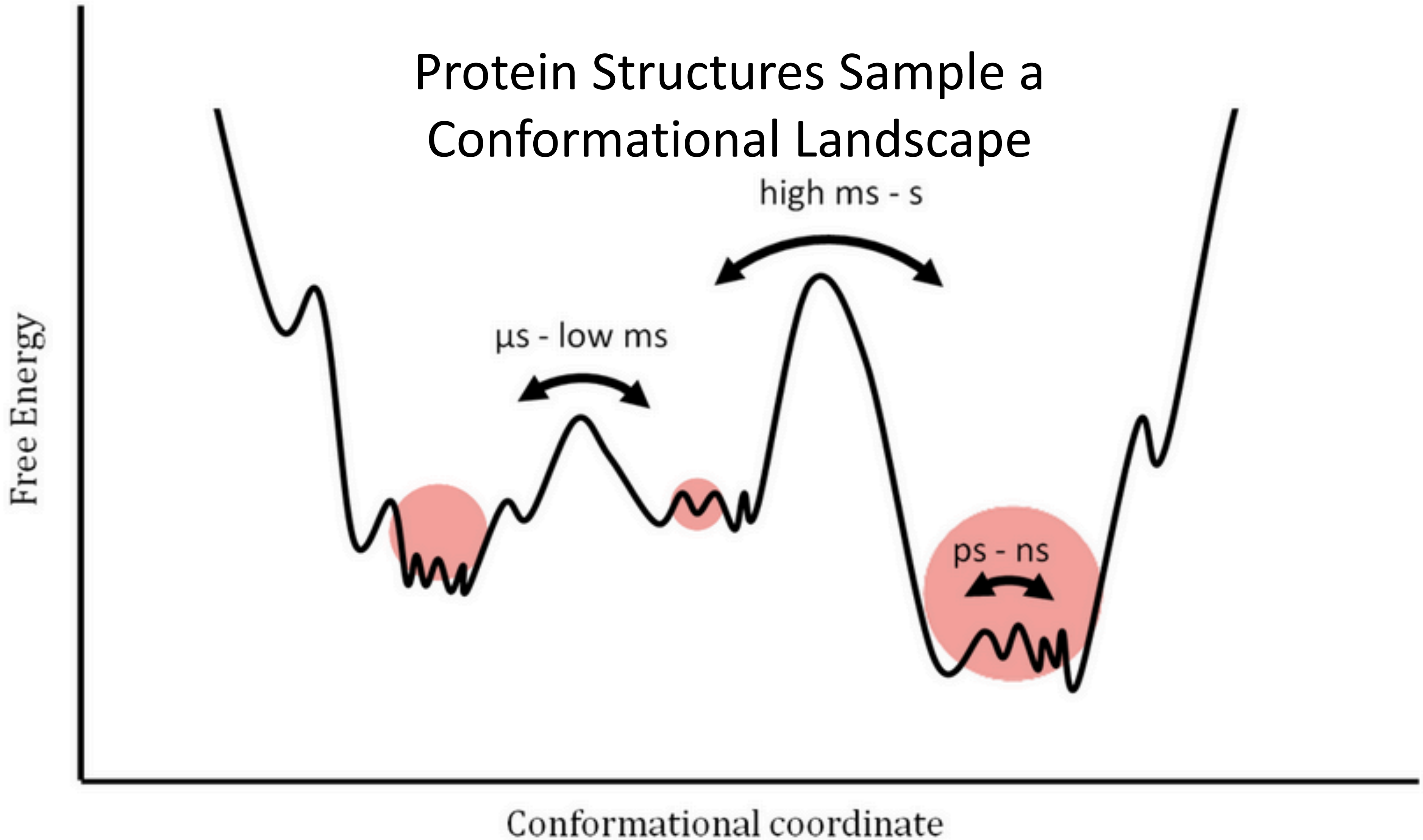
Does the experimental data improve the accuracy of the predicted model?

Do predictors using sparse data provide higher accuracy models than the *best* non-data-assisted predictors?

How is the ranking of data-assisted predictors affected if we assess against data rather than reference structure?

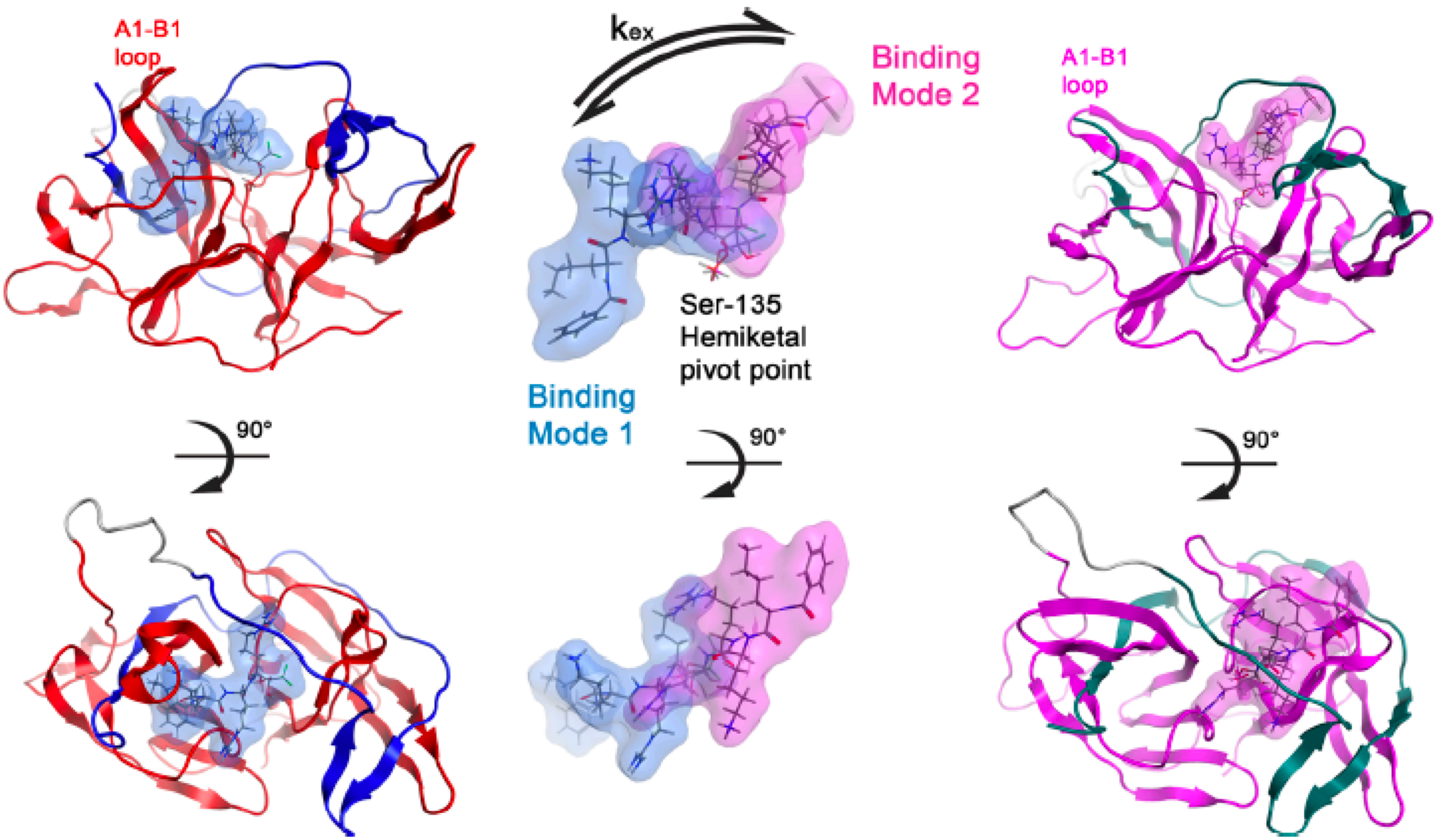
How can we model distributions of conformations?

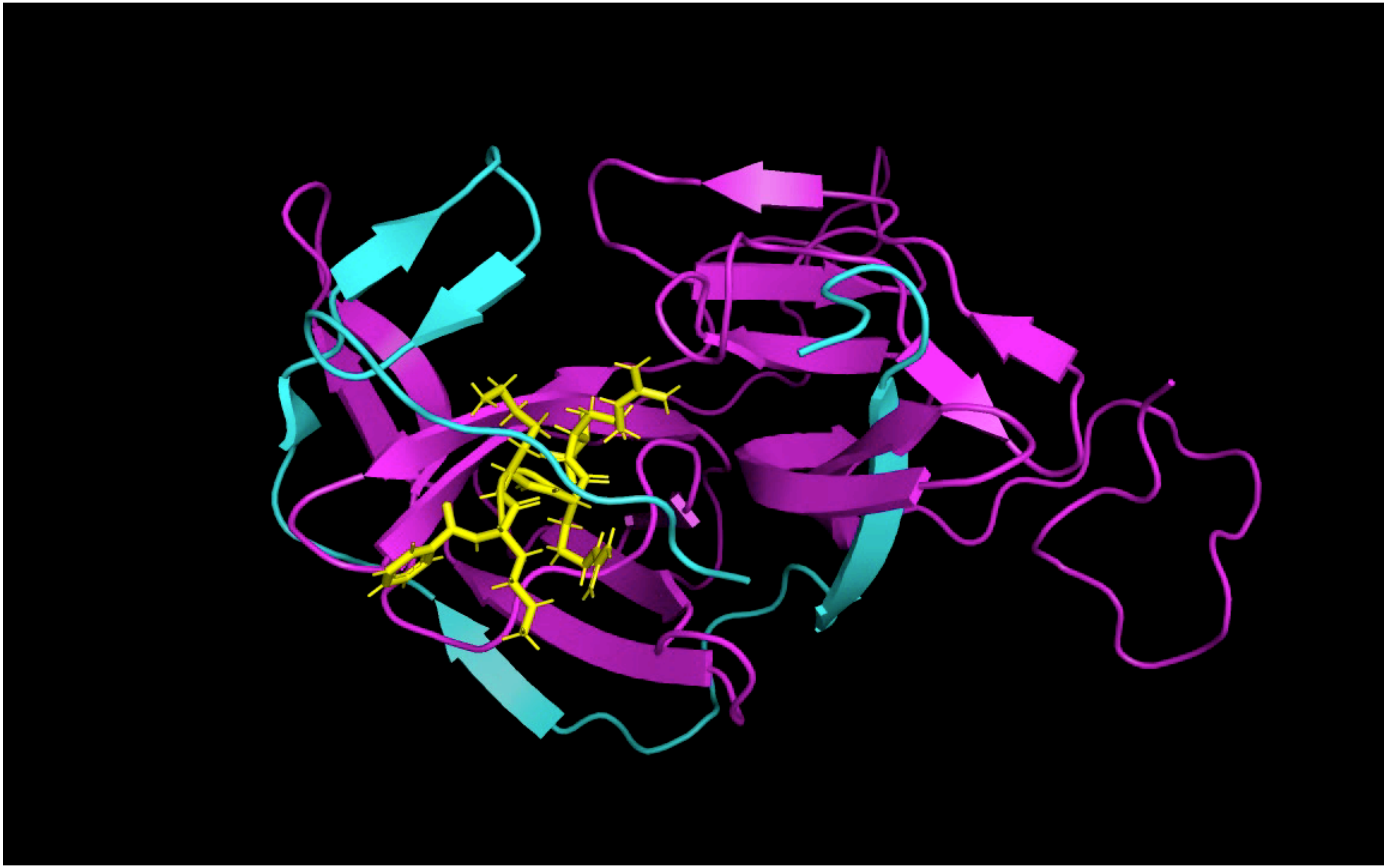
Protein Dynamics



www.mdpi.com

NMR Reveals Two Non-Overlapping Inhibitor Binding Sites in DENV2-NS2B-NS3pro Protease Complex





Sparse Experimental Data Assisted Prediction in CASP13

NMR: J. Duarte, J. Y. Huang, A. Rosato, D. Snyder, G.T. Montelione, H. Valafar

Simulated Sparse NMR data for 11 CASP FM Targets and two real NMR data sets

SAXS and SANS: J. Duarte, G. Hura, J. Tainer, S. Tsutakawa

Real SAXS data for 11 CASP FM Targets

Chemical Cross–Link (X-link): J. Duarte, A. Fiser, A. Leitner, J. Rappsilber

Real X-LInk Data for 29 domains/subunits/full complexes

Fluorescence Resonance Energy Transfer (FRET): C. Seidel

Real FRET data for a multidomain protein

Guided Prediction with Sparse NMR Data

Gaetano T. Montelione, Natalia Denissova, Janet Y. Huang,
Yojiro Ishida, Gaohua Liu, Roberto Tejero, G.V.T. Swapna ,
Rutgers University, New Jersey, USA

Antonio Rosato, Davide Sala
CERM, University of Florence, ITALY

Homay Valafar
University of South Carolina

David Snyder
William Patterson University, New Jersey, USA

NMR-Guided Prediction

13 CASP Targets

17 Assessment Units

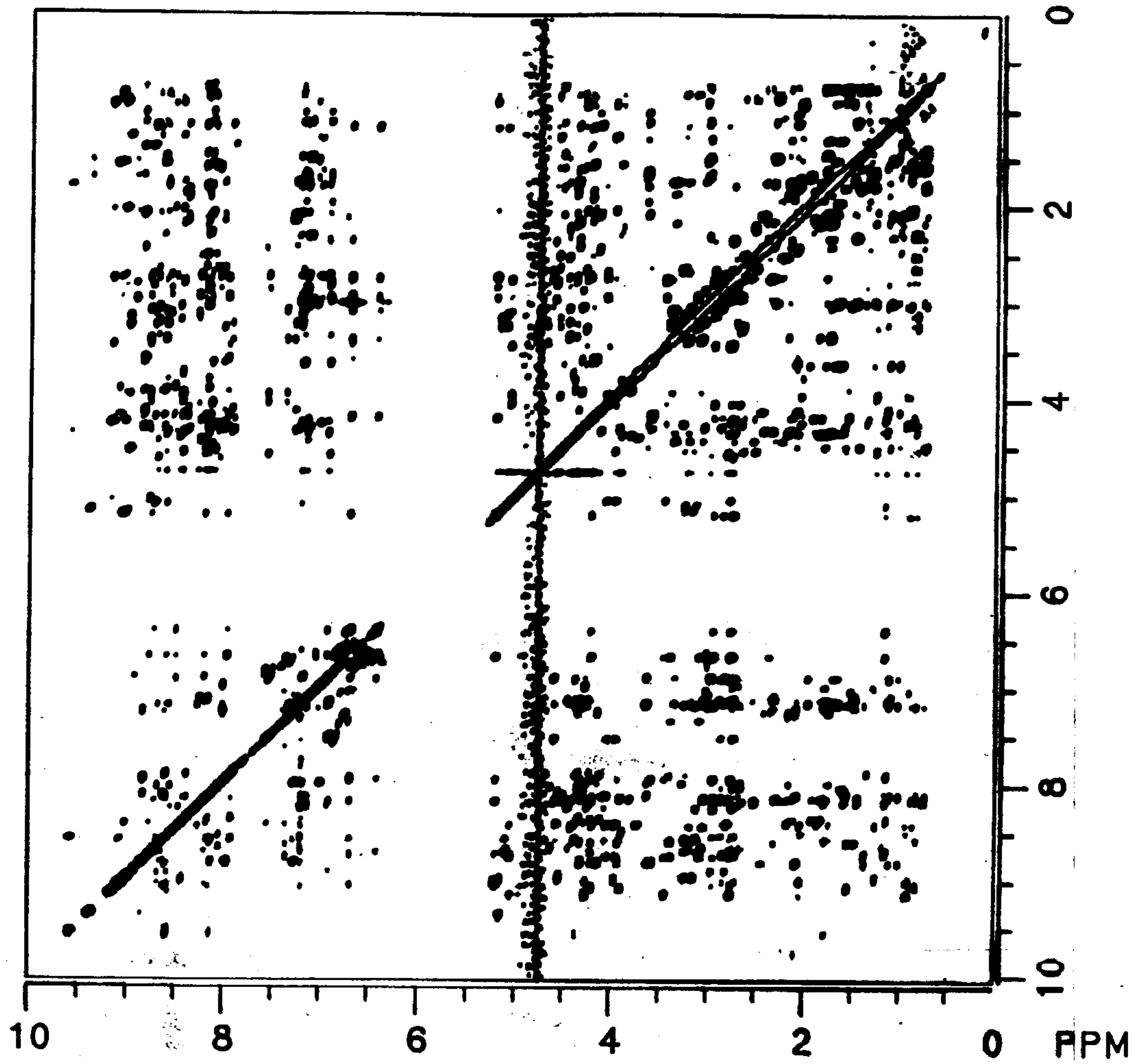
12 Simulated NMR Data Sets (FM Targets)

2 Real NMR Data Sets (Designed Protein)

6 Predictors

3 “Baseline” Groups

2D NOESY Spectrum of a Protein



Acc. Chem. Res., Vol. 22, No. 1, 1989 Wüthrich

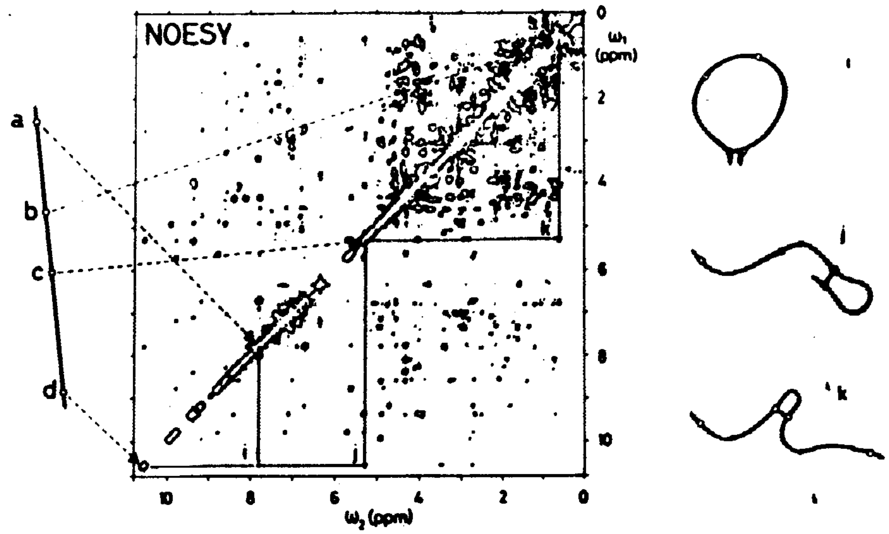


Figure 4. Illustration of the description of the NMR method for protein structure determination in solution. In the center, a contour plot of a 500-MHz ^{1}H NOESY spectrum of the protein basic pancreatic trypsin inhibitor (BPTI) is shown, with the two frequency axes ω_1 and ω_2. Three cross peaks are marked i–k and linked by horizontal and vertical lines with the diagonal positions of the protons connected by the corresponding NOEs. On the left, an extended polypeptide chain is represented by a straight line, and four protons in this chain are identified by circles and the letters a–d. The broken arrows connect these protons with their resonance positions on the diagonal of the NOESY spectrum. On the right, there is a schematic representation of three circular structures formed by the polypeptide chain, which are manifested by the cross peaks i–k.

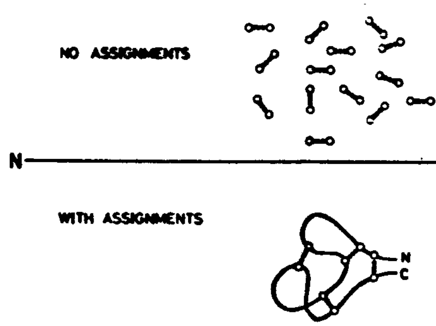
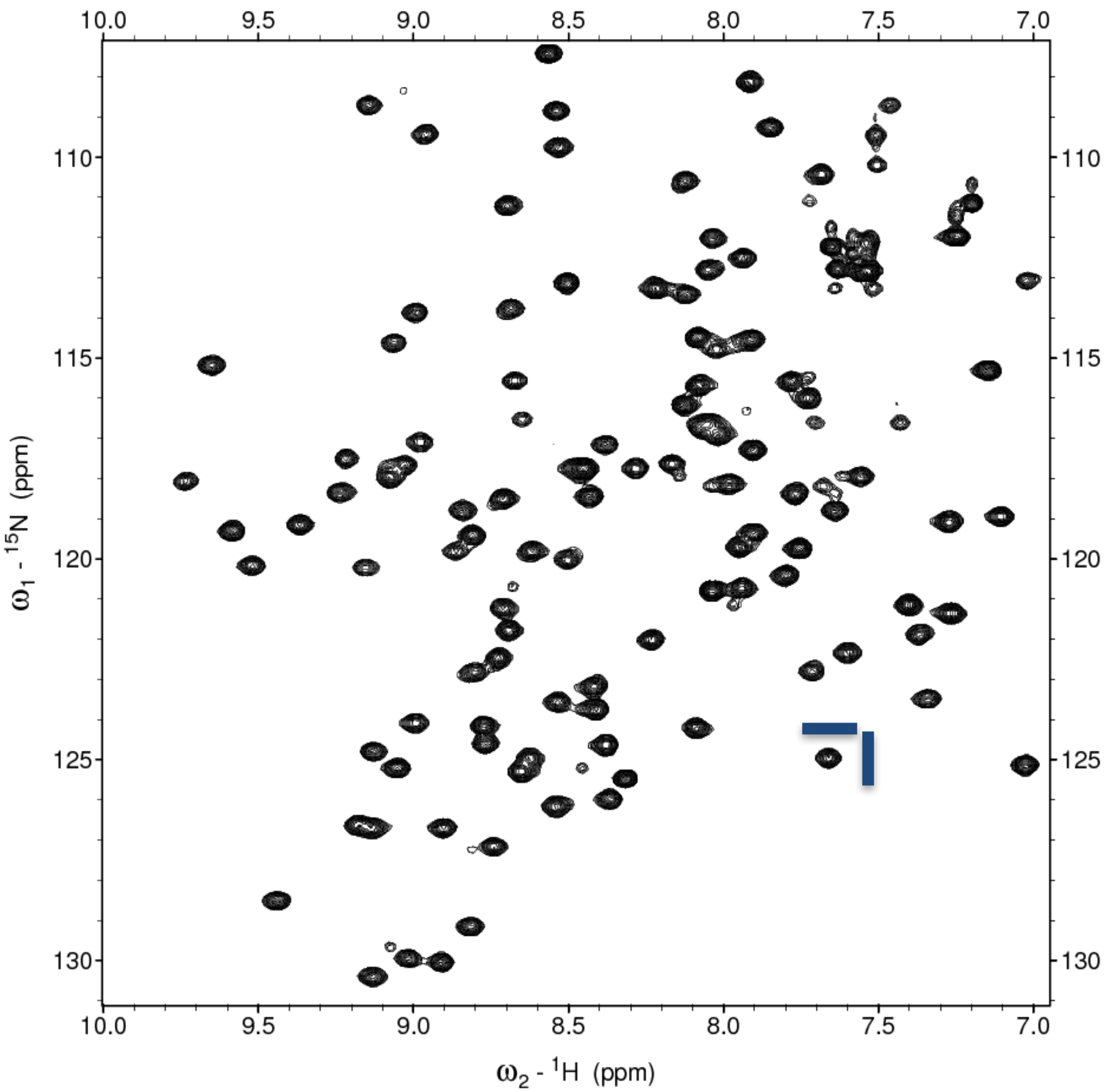


Figure 1.1. Information content of ^{1}H–^{1}H NOE's in a polypeptide chain with and without sequence-specific resonance assignments. Open circles represent hydrogen atoms of the polypeptide. The polypeptide chain is represented by the horizontal line in the center.

The Ambiguity Problem in Analysis in Cross Peak Assignment



In NOESY

For a given cross peak, the Y-axis will, in general, match, within a "match tolerance", to Y possible resonances assignments.

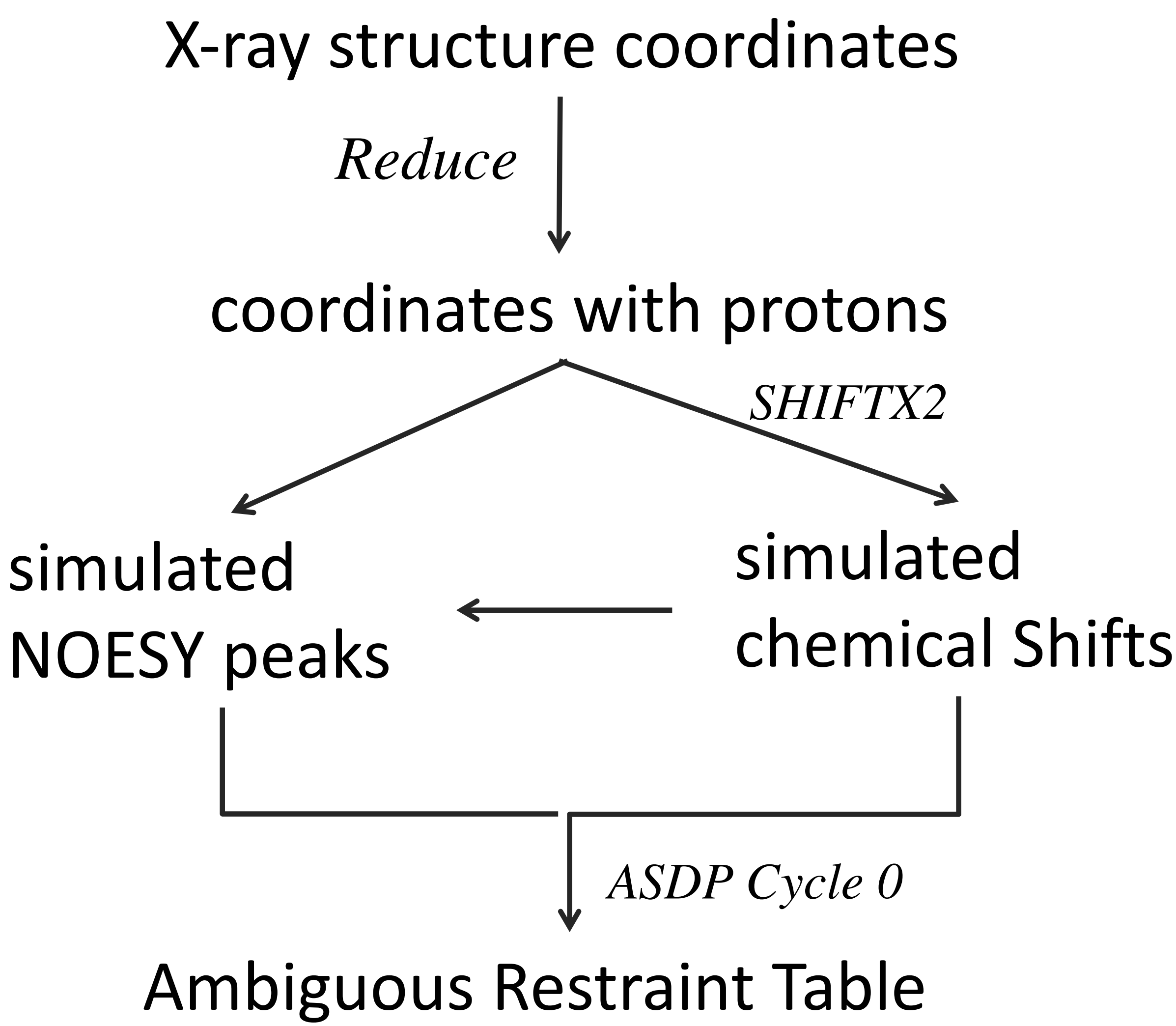
The X-axis will, in general, match, within a "match tolerance", to X possible resonance assignments.

Hence – the NOESY cross peak may arise from any one (or more) of X * Y short (< 5 Å) distance interactions

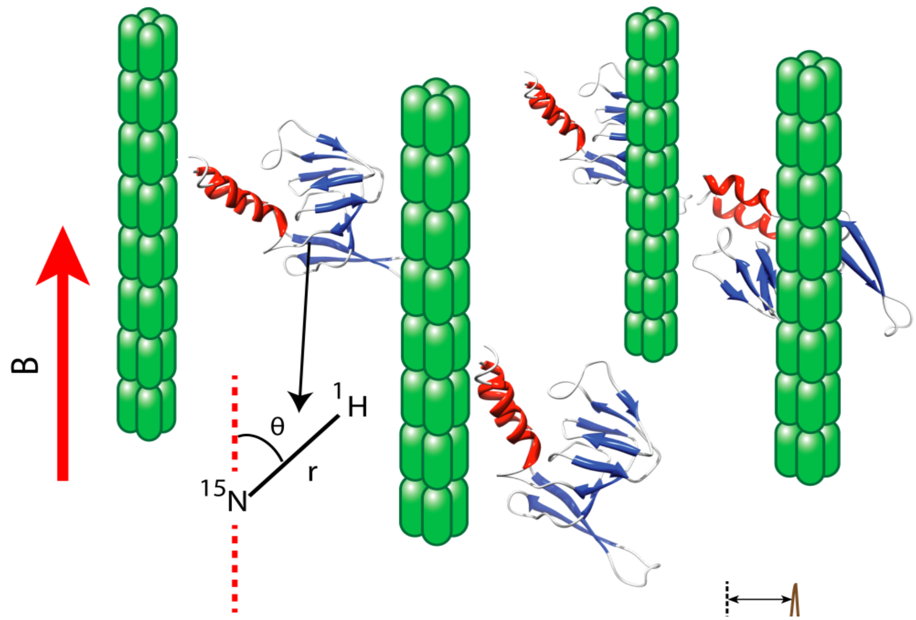
Ambiguous NOE-based Contact List

(H^N-H^N, H^N-Me, Me-Me ^1H-^1H Contacts)

Residue 1	Residue 2	Peak No.	Upper-bound		Atom 1	Atom 2	
R1	**R2**	**P#**	**UPL**	**Confid**	**A1**	**A2**	
79	77	17	5.0	0.95	H	H	Peak 17
79	177	20	6.0	0.67	H	HD2	
79	135	20	6.0	0.97	H	HD1	Peak 20
79	249	20	6.0	0.96	H	HD1	
79	50	20	6.0	0.81	H	HD2	
79	217	23	5.0	0.68	H	H	
79	230	23	5.0	0.75	H	H	
79	232	23	5.0	0.72	H	H	
79	106	23	5.0	0.76	H	H	Peak 23
79	166	23	5.0	0.83	H	H	
79	100	23	5.0	0.83	H	H	
79	82	23	5.0	0.74	H	H	
79	246	23	5.0	0.71	H	H	
79	216	23	5.0	0.67	H	H	
45	37	28	7.5	0.84	HD2	HG1	Peak 28


X-ray structure coordinates
Reduce
coordinates with protons
SHIFTX2
simulated
NOESY peaks
simulated
chemical Shifts
ASDP Cycle 0
Ambiguous Restraint Table

Residual Dipolar Couplings – Measured in Orienting Media



Alignment of a protein in an orienting solution (the molecules of the orienting medium are depicted as green rods).

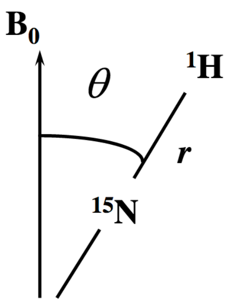
The rods align with the magnetic field due to their large magnetic anisotropy; **the protein interacts weakly with the rods**, yielding a **partial alignment** of the protein molecules.

This allows the measurement of residual dipolar couplings for bond vectors, e.g. the ^{1}H-^{15}N moieties.

N. Tjandra and A. Bax, *Science* 278, 1111 (1997).

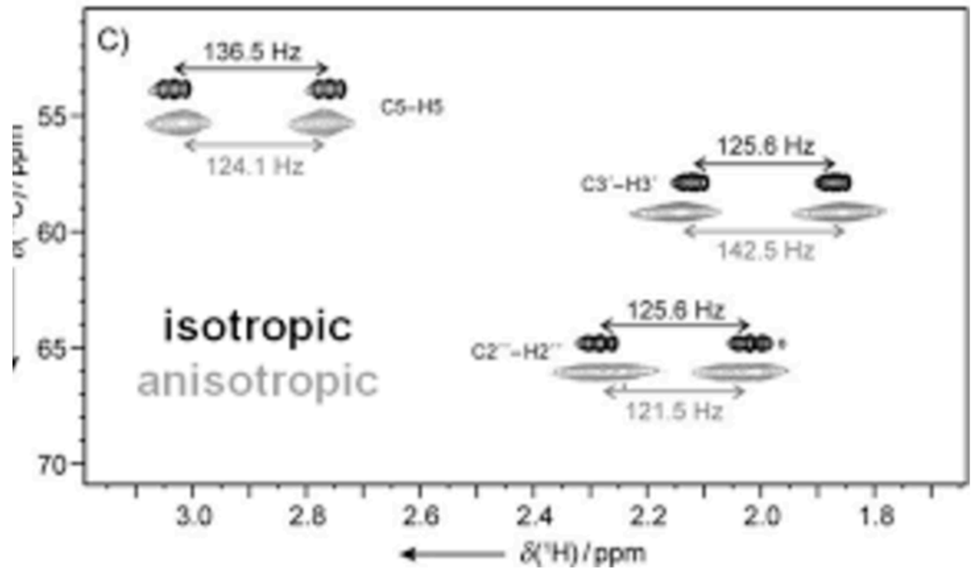
Residual Dipolar Couplings

Provide Information about Bond Vector Orientations



$$D = \frac{C}{r^3}\left\langle \frac{3\cos^2\theta - 1}{2} \right\rangle I_{NZ} I_{HZ}$$

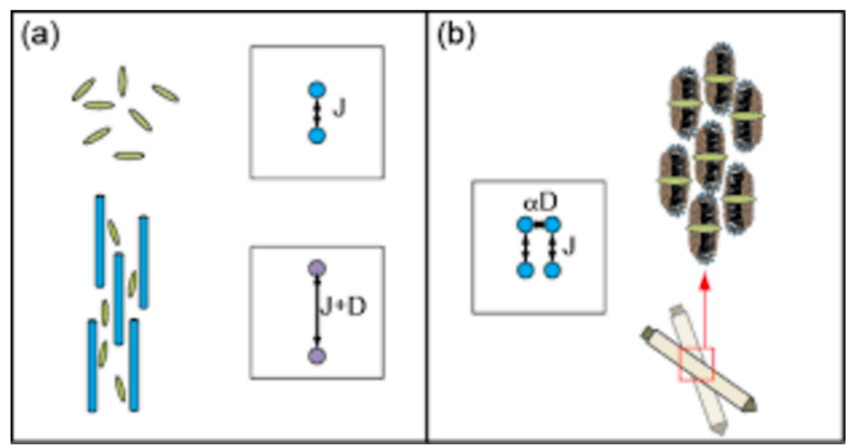
Brackets denote averaging –
goes to zero without partial orientation



Prestegard, A-Hashimi & Tolman, Quart. Reviews Biophys. 33, 371-424 (2000)

Bax, Kontaxis & Tjandra, Methods in Enzymology 339 , 127-174 (2001)

Prestegard, Bougault & Kishore, Chemical Reviews, 104 , 3519-3540 (2004)



Residual Dipolar Couplings – Measured in Orienting Media

$$D_{PQ}(\theta,\phi) = \frac{\mu_0}{4\pi}\frac{\gamma_P\gamma_Q h}{4\pi^2 r_{PQ}^3}\left[A_{ax}(3\cos^2\theta - 1) + \frac{3}{2}A_{rh}(\sin^2\theta\cos 2\phi)\right]$$

$$D^{PQ}(\theta,\phi) = D_{ax}^{PQ}(3\cos^2\theta - 1) + \frac{3}{2}D_{rh}^{PQ}(\sin^2\theta\cos 2\phi)$$

If one is measuring couplings for different atom pairs (P,Q), it is useful to apply a normalization:

$$D^{PQ}(NH) = D^{PQ}\left(\frac{\gamma_N\gamma_H\left\langle r_{NH}^{-3}\right\rangle}{\gamma_P\gamma_Q\left\langle r_{PQ}^{-3}\right\rangle}\right)$$

RDC's are global restraints

Backbone Dihedral Restraints Can be Estimated from Backbone Chemical Shift Values

$^{13}C\alpha$ / $^{13}C\beta$ chemical shifts → backbone dihedral ranges (+/- 30 deg)

Y. Shen, A. Bax. Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. J. Biomol. NMR, 56, 227-241(2013)

https://spin.niddk.nih.gov/bax/software/TALOS-N/

Features of Simulated Sparse NMR Data

Assume ^{13}C,^{15}N-enriched perdeuterated samples, with ILV $^{13}CH_3$ methyl resonances.

NOESY peak frequencies were “wiggled” to simulate inaccuracies in peak picking due to broad line widths.

NOESY Peaks or Resonance Assignments were deleted to account for line broadening due to internal motions and/or incomplete assignments.

Random “noise” peaks were added to the NOESY Peak Lists.

Backbone dihedral angle phi and psi restraints (chosen randomly within +/- 30 deg, with uncertainty +/- 30 deg) were provided. These would normally be available from the backbone chemical shift data.

^{15}N-^{1}H RDC data was provided for 2 alignments, assuming typical precisions.

Features of Simulated NMR Data

Assume ^{13}C,^{15}N-enriched perdeuterated samples, with ILV $^{13}CH_3$ methyl resonances.

NOESY peak frequencies were "wiggled" to simulate inaccuracies in peak picking due to broad line widths.

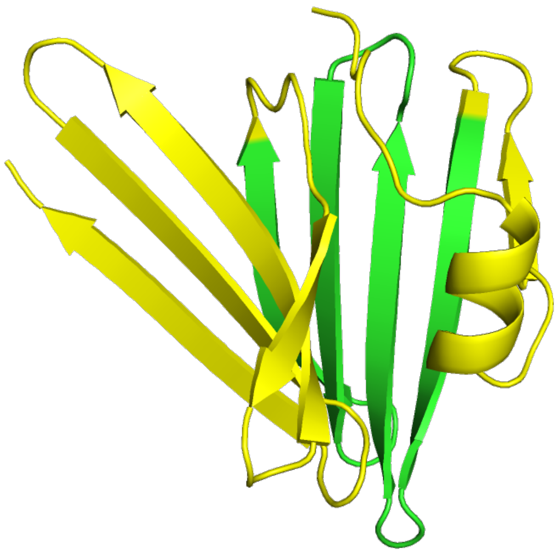
NOESY Peaks or Resonance Assignments were deleted to account for line broadening due to internal motions and/or incomplete assignments.

Random "noise" peaks were added to the NOESY Peak Lists.

Backbone dihedral angle phi and psi restraints (chosen randomly within +/- 30 deg, with uncertainty +/- 30 deg) were provided. These would normally be available from the backbone chemical shift data.

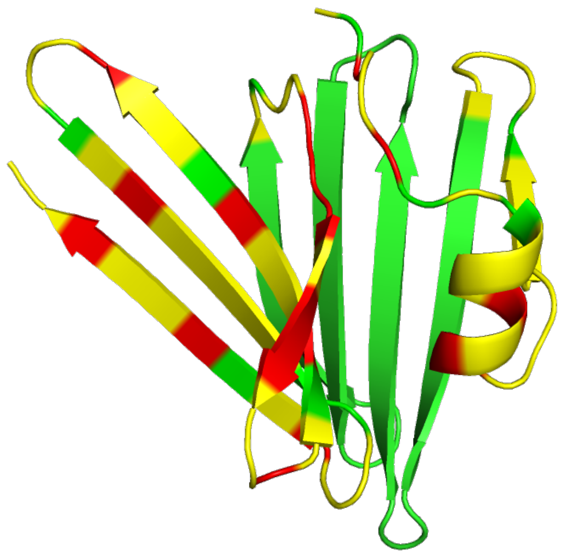
^{15}N-^{1}H RDC data was provided for 2 alignments, assuming typical precisions.

Candidates

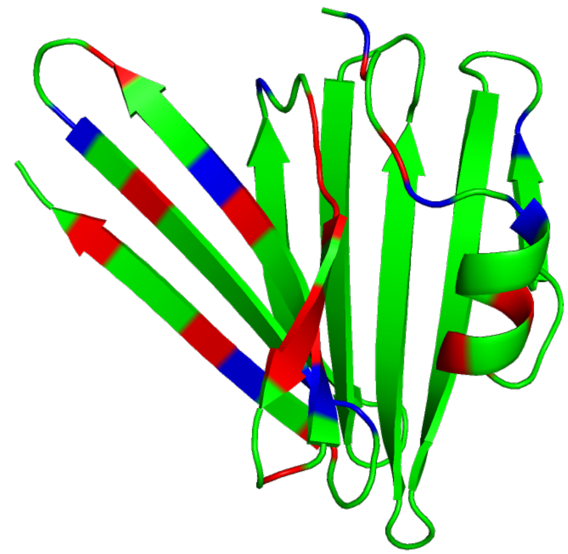


T0968s1

Removed
Did not generate peaks



Removed from
Assignment list



Blue residues are a source of error: their peaks cannot be satisfied

Candidates

Removed
Did not generate peaks

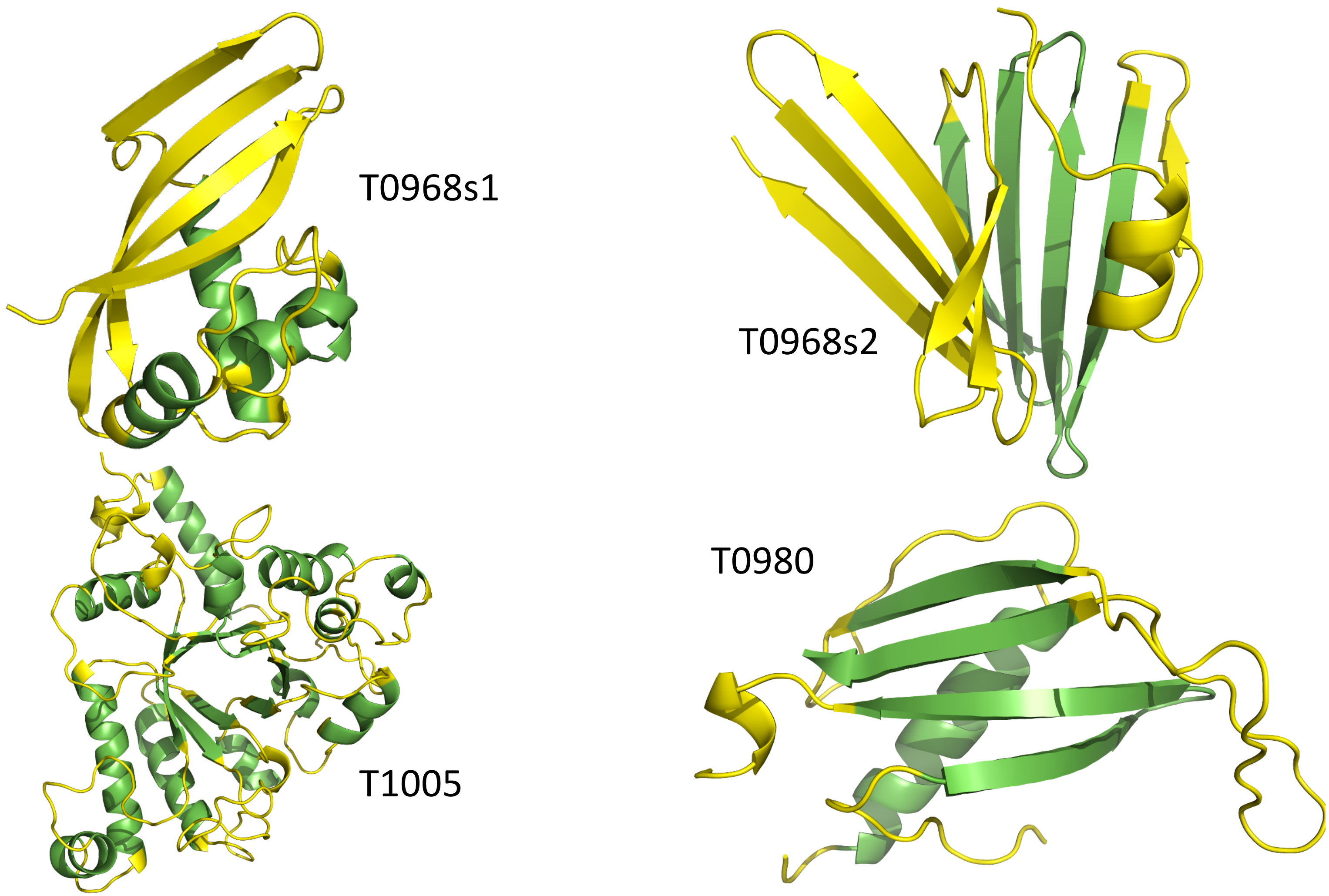
Removed from
Assignment list

90

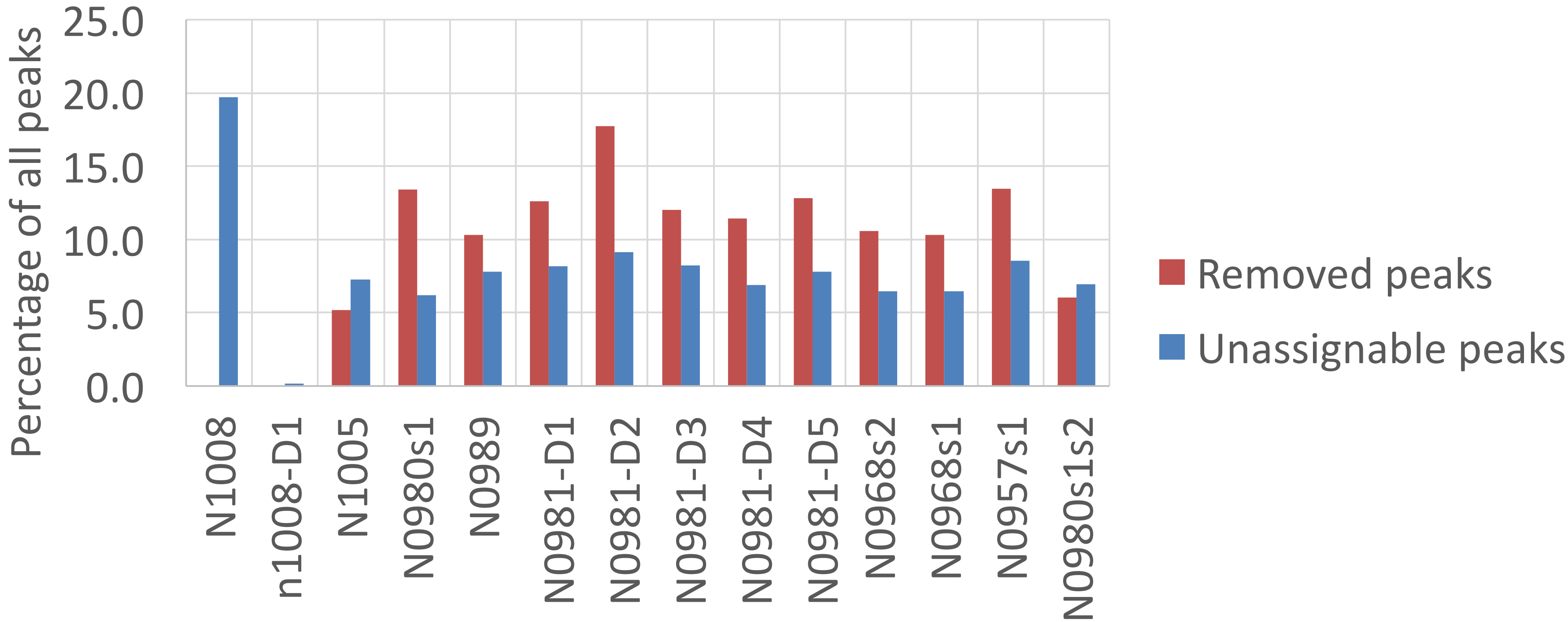
CASP:
Images redacted

Blue residues are a source of error: their peaks cannot be satisfied

CASP:
Images redacted


T0968s1
T0968s2
T0980
T1005

Statistics on NOESY Datasets



We are providing on average 6 peaks/residue. High ambiguity.

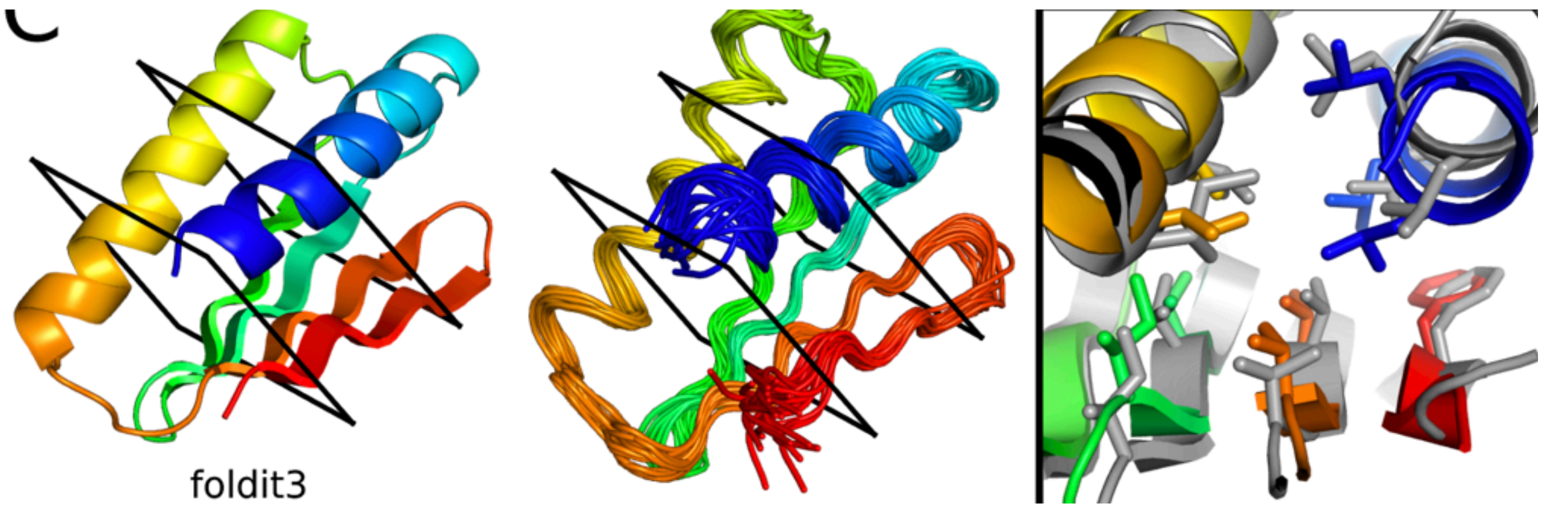
About 40 peaks/residue would be typical for NMR structures.

In the complete real n1008 dataset there are 43 peaks/residue.

Real NOESY Data

De novo protein design by citizen scientists

Koepnick, Liu et al. submitted



N1008
CASP COMMONS Target UW-Eng

Real NMR Data: Targets N1008 and n1008

CASP Commons Target: UW-eng (aka CASP5)

80 Residues, No deuteration, No RDCs, No ECs

^{15}N,^{13}C-enriched sample was produced at Rutgers. Data collection included TR-NMR for assignments, and sim(CN)-NOESY. Data collected at ~ 200 micromolar concentration, at 600 MHz and 800 MHz. Talos_N used to generate dihedral restraints in secondary structure elements.

Reference NMR structure was determined with automated methods using Cyana, and **refined by manual interactive analysis of NOESY** spectra. The resulting structures were then energy refined with CNS in explicit water. Final DP score – 0.78 (OK structure)

N1008: Using only backbone assignments, NOESY peak list was assigned, and used to generate ambiguous contact contact list for CASP predictors.

- This is the strategy we would use for larger (> 15 kDa) ^{2}H-enriched proteins
- However, since the sidechain NOESY peaks are still present in the NOESY spectrum, this data set has a very high number of unassignable / incorrectly assigned NOESY peaks, and is very challenging for automated structure determination.

n1008: Using complete backbone and sidechain assignments, NOESY peak list was assigned, and used to generate ambiguous contact list for CASP predictors.

- This is a standard strategy for small (< 15 kDa) proteins.

Contact Predictions

Provided from the CASP13 Submissions for Jones – Meta PSI COV

Target	M_{eff} / L
N0957s1:	3.3
N0989:	1-130: 107; 120-185: 16.0; 185-246: 612
N0968s1:	15.9
N0968s2 :	3.3
N0980s1 (74 / 111 residues):	2.9
N1005 (residues 72-340):	56.4

The remaining targets had M_{eff}/L < ~ 1

Assessment Units

Target	Data	# residues	Assessment units
N0957	simNMR, dihedrals, 2x rdc's	163	**N0967-D1.D2** **N0957-D1** **N0957-D2**
N0968s1	simNMR, dihedrals, 2x rdc's	123	N0968s1
N0968s2	simNMR, dihedrals, 2x rdc's	116	N0968s2
N0980s1	simNMR, dihedrals, 2x rdc's	105	N0980s1
N0981-D1	simNMR, dihedrals, 2x rdc's	86	N0981-D1
N0981-D2	simNMR, dihedrals, 2x rdc's	80	N0981-D2
N0981-D3	simNMR, dihedrals, 2x rdc's	203	N0981-D3
N0981-D4	simNMR, dihedrals, 2x rdc's	111	N0981-D4
N0981-D5	simNMR, dihedrals, 2x rdc's	127	N0981-D5
N0989	simNMR, dihedrals, 2x rdc's	134	**N0989-D1.D2** **N0989-D1** **N0989-D2**
N1005	simNMR, dihedrals, 2x rdc's	326	N1005
N1008	Limited exp. NMR, dihedrals	80	N1008
n1008	Full exp. NMR, dihedrals	80	n1008

Correlation Coefficients for Z-Scores

Correlations Between Assessment Scores

	GDT_HA	GDT_SC	RPF	SphGrdr	CAD_AA	MolPrbty
GDT_HA		0.959	0.923	0.907	0.929	0.518
GDT_SC	0.952		0.902	0.891	0.937	0.521
RPF	0.918	0.902		0.952	0.969	0.557
SphGrdr	0.901	0.895	0.947		0.927	0.555
CAD_AA	0.915	0.932	0.966	0.920		0.588
MolPrbty	0.546	0.554	0.573	0.562	0.610	

Friedman's Test indicates different scoring techniques do not give significantly different rankings

(upper right: Pearson; lower left: Spearman)

Correlation between LDDT and RPF

Pearson	**0.974**
Spearman	0.977

D. A. Snyder

Baseline Models

Structures Generated by Janet Huang (blind) with ASDP / CYANA -> Restrained Rosetta Refinement

Group 321
Sparse NMR Data, RDCs, no ECs

Janet Huang

Group 459
Sparse NMR Data, RDCs, Meta PSI COV (Jones) ECs

Group 313
Sparse NMR Data, RDCs, “Best” ECs

Best ECs (Jones, Sanders, or none): Picked best 5 from 15 calculated based on DP score.

Generally expect 313_J > 459_J > 321_J

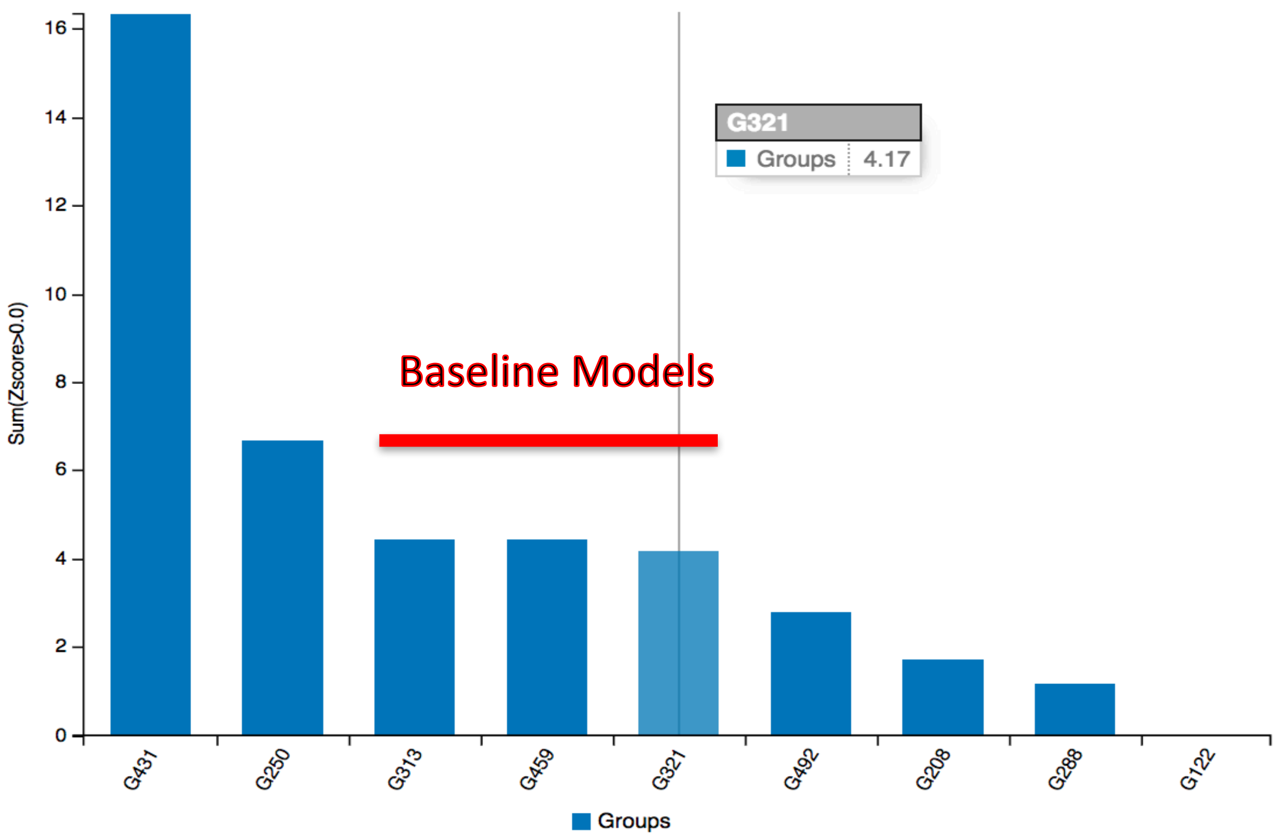
Baseline Models
Structures Generated by Janet Huang (blind) with ASDP / CYANA -> Restrained Rosetta Refinement

Generally expect:	313_J > 459_J > 321_J	
GDT_TS	313_J > 459_J > 321_J	all very similar raw and Z scores
GDT-HA	313_J > 321_J > 459_J	
GDT_ALL	459_J > 313_J > 321_J	
GDT_SC	459_J > 313_J > 321_J	
SphereGrinder	459_J > 321_J > 313_J	
RPF	459_J > 321_J > 313_J	
MolProbity	321_J > 313_J > > > 459_J	

Initial Z-Score Based Ranking
(Z = -2 Cutoff)

<u>Z Score Based Ranking - GDT-TS Score alone</u>

431 > 250 > (313_ > 459_J > 321_J) > 492 > 208 > 288 > 122



Initial Z-Score Based Ranking (Z = -2 Cutoff)

GDT_TS Z_Scores

431 > 250 > (313_J > 459_J > 321_J) > 492 > 208 > 288 > 122

GDT_HA Z_Scores

431 > 250 > (313_J > 321_J > 459_J) > 492 > 208 > 288 > 122

GDT_All Z Scores

431 > 250 > (459_J > 313_J > 321_J) > 492 > 208 > 288 > 122

GDT_SC Z Scores

431 > 250 > (459_J > 313_J > 321_J) > 208 > 492 > 288 > 122

Sphere Grinder Z Scores

431 > 250 > (459_J > 321_J > 313_J) > 492 > 288 > 208 > 122

RPF Z Scores

431 > 250 > (459_J > 321_J > 313_J) > 492 > 288 > 208 > 122

Molprobity Z Scores

250 > 431 > (321_J > 313_J) > 492 > 288 > 459_J > 208 > 122

- note that 459_J drops in ranking – why is this?
- 250 and 431 switch order; 250 does a better job in regularizing the structures?
- the assessment of method 250 is greatly enhanced by including MolProbity score

- **Conclusion: Get pretty much the same ranking regardless of the score used.**

PCA Results: Thresholding at Z = -1

Component	GDT_HA	GDT_SC	RPF	SphGrdr	CAD_AA	MolPrbty		% Variance Explained
1	0.442	0.449	0.425	0.428	0.433	0.227		86.702
2	-0.146	-0.188	-0.067	-0.056	-0.040	0.966		8.351
3	-0.388	-0.562	0.389	0.608	0.050	-0.104		2.511
4	-0.371	-0.034	0.373	-0.567	0.632	-0.044		1.331
5	0.655	-0.548	0.380	-0.319	-0.156	-0.007		0.800
6	0.256	-0.383	-0.616	0.145	0.621	-0.044		0.306

GDT-like: Z-Score Based Ranking (Z = 0 Threshold, Model 1)

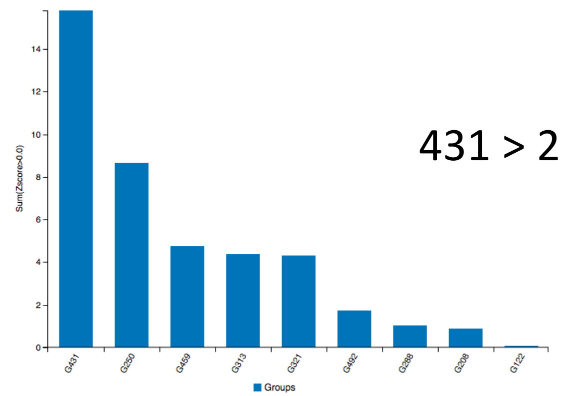
GDT_TS: 0.0
GDT_HA: 0.44
GDT_ALL: 0.00
GDT_SC: 0.45
RMSD*: 0.00
LGA_S: 0.00
ALO: 0.00
Dali(raw): 0.00
Mammoth: 0.00
Molprb*: 0.23
lDDT: 0.00
CAD(aa): 0.43
CAD(ss): 0.00
RPF: 0.42
CODM: 0.00
DFM*: 0.00
Hand.: 0.00
SOV: 0.00
QCS: 0.00
ContS: 0.00
TM-score: 0.00
SG: 0.43
RDC1*: 0.00
RDC2*: 0.00
DP: 0.00

* *z-scores calculated on inverted raw scores*

Reset Weights

Equal Weights

Show
- as a webpage (default)
- in txt format



431 > 250 > Baseline

#	GR code	GR name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)	AVG Zscore (>0.0)	Rank AV (>0.0)
1	431	-	14	14.3277	1	1.0328	1	15.7915	1	1.1280	1
2	250	-	14	5.4434	3	0.1443	5	8.6442	2	0.6174	2
3	459	-	13	3.5906	4	0.1591	4	4.7489	3	0.3653	3
4	313	-	12	2.0121	5	0.2515	2	4.3726	4	0.3644	4
5	321	-	14	5.9091	2	0.1909	3	4.3101	5	0.3079	5
6	492	-	13	-2.6627	7	-0.5181	7	1.7315	6	0.1332	6
7	288	-	14	-1.9207	6	-0.5921	8	1.0108	7	0.0722	8
8	208	-	7	-12.9425	9	-0.4904	6	0.8941	8	0.1277	7
9	122	-	8	-12.1411	8	-1.0353	9	0.0468	9	0.0058	9

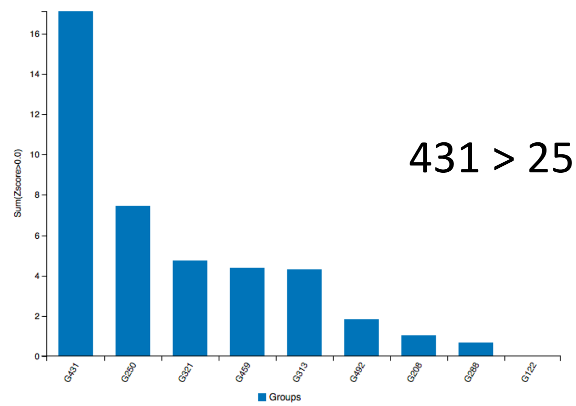
Z-Score Based Ranking
(Z = 0 Threshold, Best Model)

GDT_TS: 0.0	GDT_HA: 0.44	GDT_ALL: 0.00	GDT_SC: 0.45	RMSD*: 0.00
LGA_S: 0.00	AL0: 0.00	Dali(raw): 0.00	Mammoth: 0.00	Molprb*: 0.23
lDDT: 0.00	CAD(aa): 0.43	CAD(ss): 0.00	RPF: 0.42	CODM: 0.00
DFM*: 0.00	Hand.: 0.00	SOV: 0.00	QCS: 0.00	ContS: 0.00
TM-score: 0.00	SG: 0.43	RDC1*: 0.00	RDC2*: 0.00	DP: 0.00

* z-scores calculated on inverted raw scores

Reset Weights   Equal Weights

Show
- as a webpage (default)
- in txt format



431 > 250 > Baseline

#	GR code	GR name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)	AVG Zscore (>-2.0)	Rank AVG Zscore (>-2.0)	SUM Zscore (>0.0)	Rank SUM Zscore (>0.0)	AVG Zscore (>0.0)	Rank (>0.0)
1	431	-	14	15.2599	1	1.1260	1	17.1015	1	1.2215	1
2	250	-	14	4.7298	3	0.0730	5	7.4357	2	0.5311	2
3	321	-	14	5.6718	2	0.1672	4	4.7371	3	0.3384	4
4	459	-	13	3.8357	4	0.1836	3	4.3806	4	0.3370	5
5	313	-	12	2.4560	5	0.3070	2	4.3171	5	0.3598	3
6	492	-	13	-2.2360	6	-0.4707	6	1.8384	6	0.1414	7
7	208	-	7	-13.2321	9	-0.5387	7	1.0315	7	0.1474	6
8	288	-	14	-2.4610	7	-0.6461	8	0.6623	8	0.0473	8
9	122	-	8	-12.6000	8	-1.1500	9	0.0000	9	0.0000	9

Real Sparse NMR Data

G.V.T. Swapna

Target N1008 – real NMR data
(bb assignments only)

Note that no EC contact predictions or RDCs were available, as this is a FoldIt designed protein from the David Baker group (Crowd source).

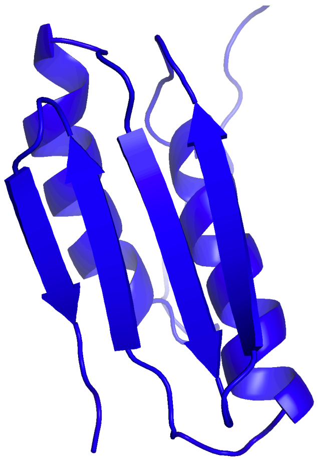
General					LGA Sequence Dependent (4Å) Full			LGA Sequence Independent (4Å) Full		MAMMOTH	Dali Full	Molprobity Full	IDDT	SphGr
#	◆ Model	◆ GR#	◆ GR Name	Charts	◆ GDT_TS	◆ NP_P	◆ Z-M1-GDT	◆ AL0_P	◆ AL4_P	◆ Z-score	◆ Z-Score	◆ MP-Score	◆ Global score	◆ SG
1.	N1008TS250_1-D1	250		A D I G	75.00	100.00	1.42	81.82	96.10	6.75	11.2	0.50	0.69	93.51
2.	N1008TS208_1-D1	208		A D I G	73.05	100.00	1.28	83.12	97.40	6.61	10.1	1.34	0.60	82.47
3.	N1008TS431_1-D1	431		A D I G	68.18	100.00	0.92	58.44	93.51	6.18	9.2	0.50	0.59	81.82
4.	N1008TS313_1-D1	313		A D I G	52.92	98.70	-0.21	46.75	75.32	3.41	5.1	0.89	0.49	59.95
5.	N1008TS321_1-D1	321		A D I G	52.92	98.70	-0.21	46.75	75.32	3.41	5.1	0.89	0.49	59.95
6.	N1008TS122_1-D1	122		A D I G	42.86	100.00	-0.95	33.77	48.05	1.09	2.3	2.78	0.43	57.79
7.	N1008TS492_1-D1	492		A D I G	40.58	100.00	-1.11	27.27	42.86	0.95	1.5	1.18	0.41	47.40
8.	N1008TS288_1-D1	288		A D I G	40.26	100.00	-1.14	22.08	45.45	0.53	1.4	1.78	0.41	44.80

Interestingly -- for real data (bb only) the GDT-TS performance order is:
Group 250 > Group 208 > Group 431 > Janet 313

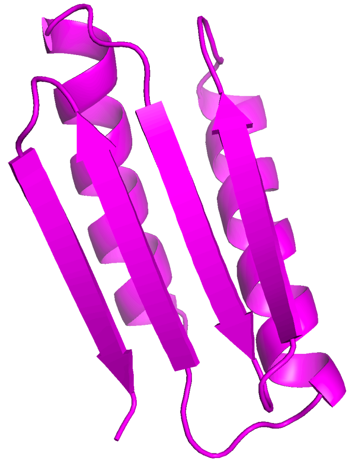
Group 208 did relatively better with this real NMR data set than with most simulated data, and Group 431 did relatively less well that they did for other targets.

GDT gain over 313_J: 431: 15 pts 208: 20 pts 250: 22 pts

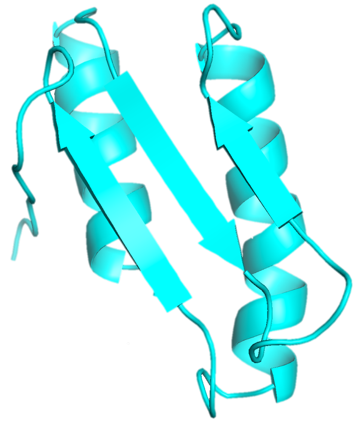
Target N1008 - real data; bb only



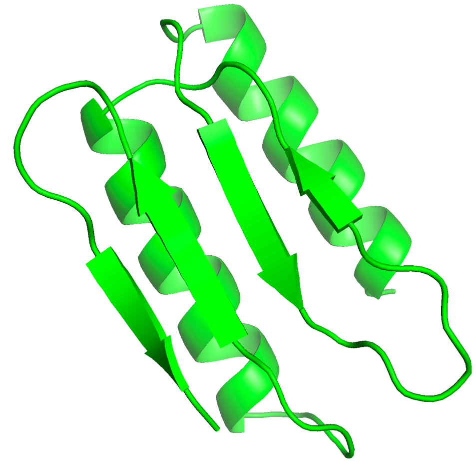
PDB



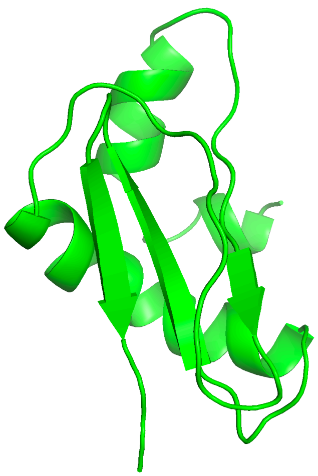
Best Regular
Prediction – SHORTLE
91.23



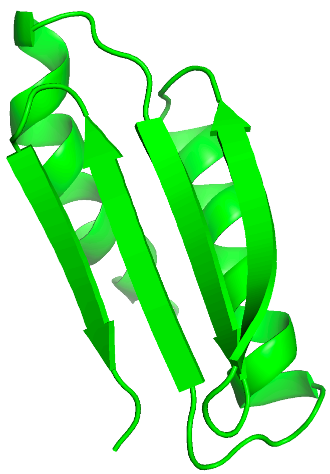
ASDP no EC
52.92



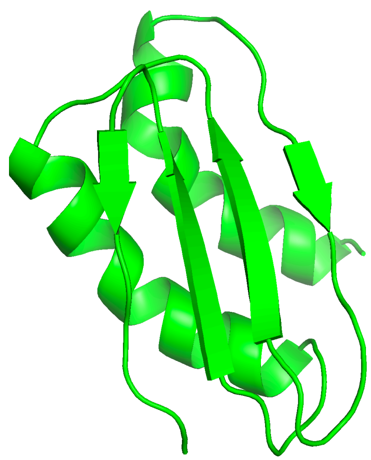
122-Forbidden
42.86



492-wfBakerUNRES
40.58

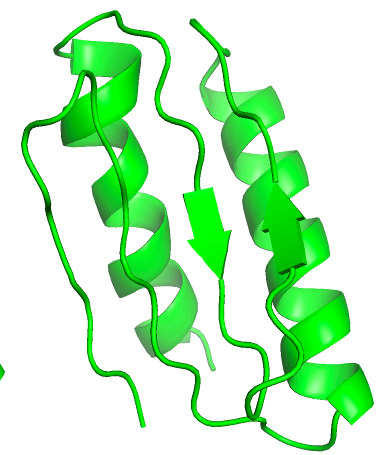


431-Laufer
68.18



288-UNRES
40.26

250-Meilerlab
75.00

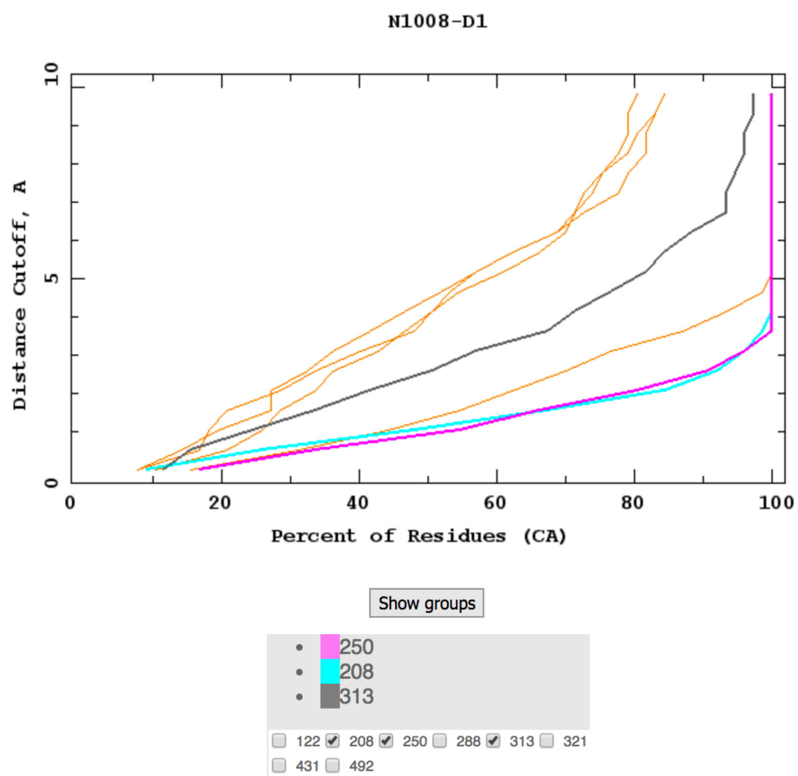


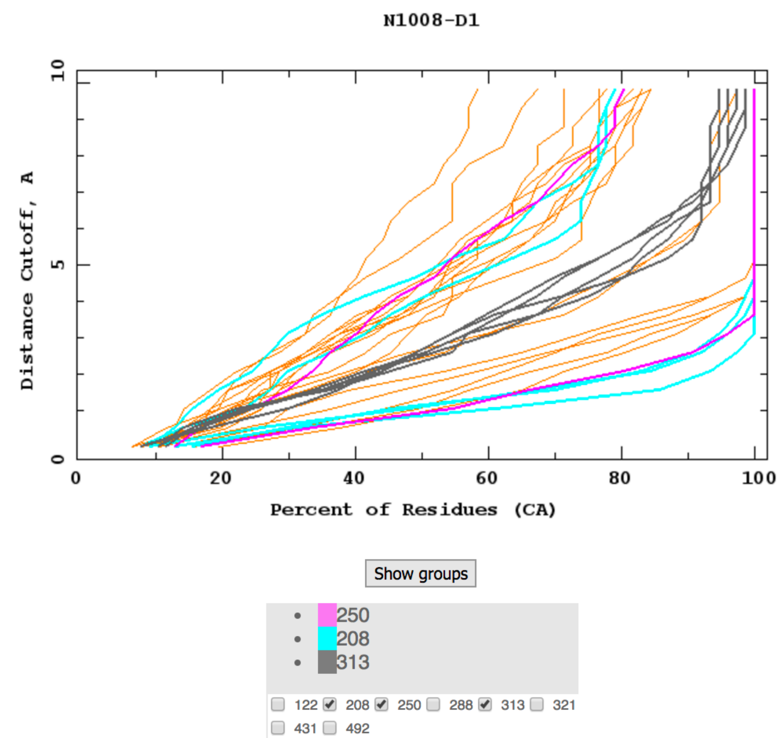
208-KIAS
73.05

Real NMR Data

Target N1008 – real NMR data (bb only)

The "top 1" target of Groups 250 and 208 are significantly better than baseline (Group 313_J), but the variability across their submissions is high - they do a good job of selecting their best model of the 5 submitted.





Real NMR Data

Target n1008 – real NMR data (bb + sc assignments)

Note that no EC contact predictions or RDCs were available, as this is a FoldIt designed protein from the David Baker group (Crowd source).

General					LGA Sequence Dependent (4Å) Full			LGA Sequence Independent (4Å) Full		MAMMOTH	Dali Full	Molprobity Full	lDDT	SphGr
#	Model	GR#	GR Name	Charts	GDT_TS	NP_P	Z-M1-GDT	AL0_P	AL4_P	Z-score	Z-Score	MP-Score	Global score	SG
1.	n1008TS321_1-D1	321		A D I G	82.79	100.00	1.73	88.31	97.40	7.03	12.5	1.10	0.73	95.45
2.	n1008TS431_1-D1	431		A D I G	57.47	100.00	0.40	41.56	83.12	5.34	6.1	2.40	0.54	76.62
3.	n1008TS288_1-D1	288		A D I G	41.56	100.00	-0.44	23.38	64.94	1.38	3.1	1.27	0.41	44.80
4.	n1008TS122_1-D1	122		A D I G	40.26	100.00	-0.50	0.00	41.56	0.39	1.2	3.71	0.41	44.16
5.	n1008TS492_1-D1	492		A D I G	27.27	100.00	-1.19	19.48	27.27	-0.60	0	2.27	0.34	24.68

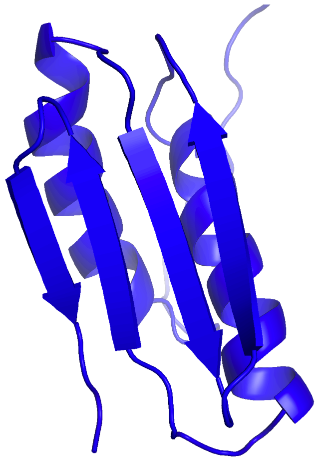
For real data (bb + sc assignments) the GDT-TS performance order is:

Janet 313_J > Group 431 > Group 288 > Group 122 etc

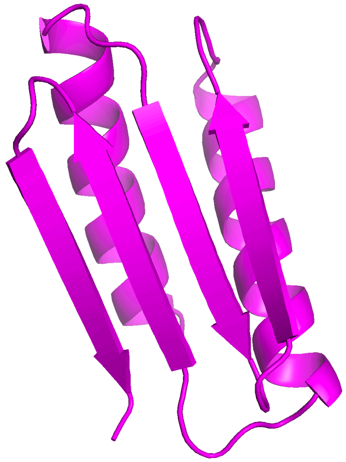
Groups 208 and 250 did not submit .

GDT gain by Janet: 431: 22 pts 288: 41 pts 122: 43 pts 492: 55 pts

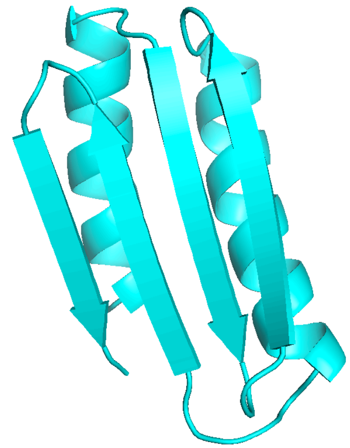
Target n1008 – Real Data; bb + sc



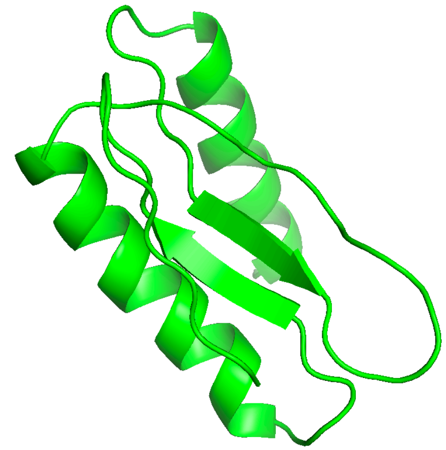
PDB



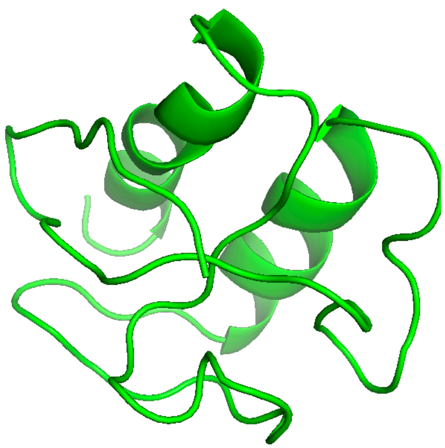
Best Regular
Prediction – SHORTLE
91.23



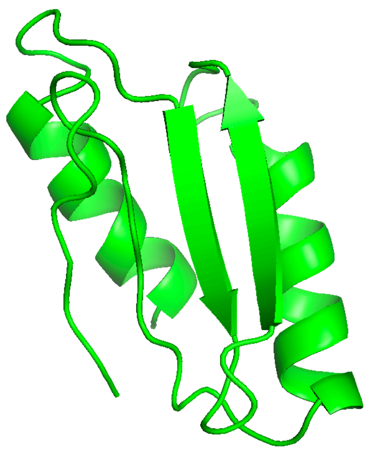
ASDP no EC
82.79



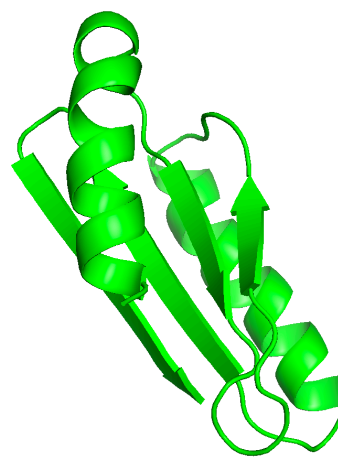
122-Forbidden
40.26



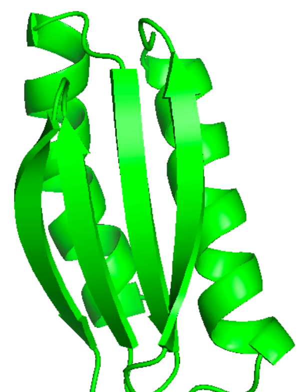
492-wfBakerUNRES
27.27



431-Laufer
57.47



288-UNRES
41.56



250-Meilerlab
62.11

208-KIAS
??

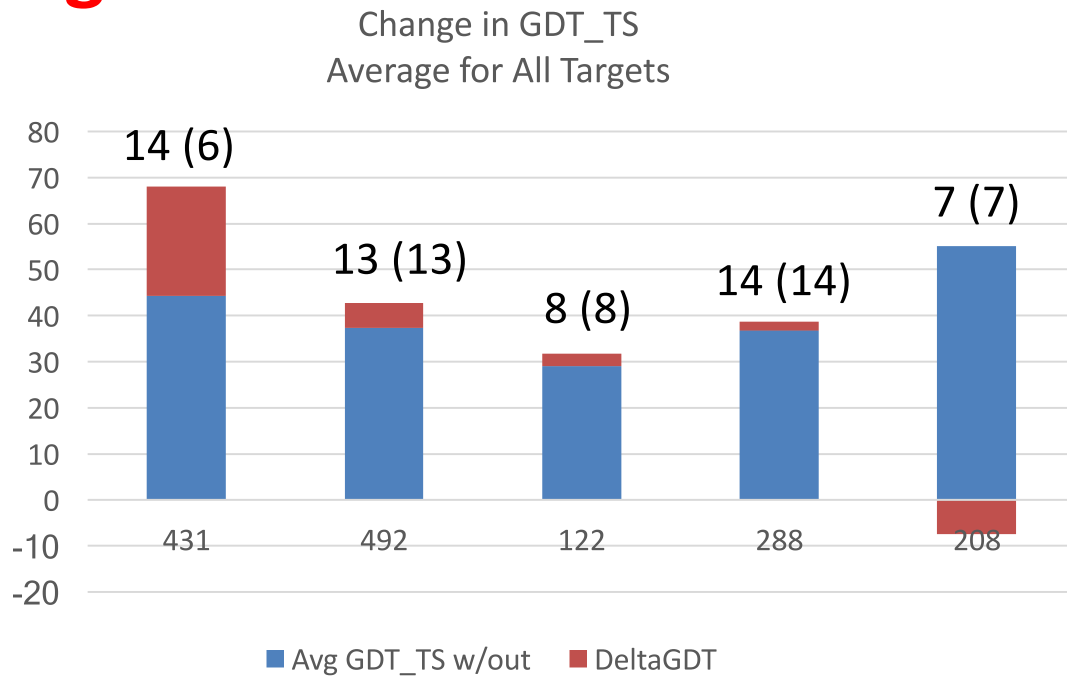
Overall Performance
Per Target Per Group
GDT Scores, First Model

Best Regular Prediction	492	431	288	250	208	122	GDT-NO EC	GDT-JonesEC
45.22	31.94	52.93	28.86	56.02		20.52	32.25	30.25
31.3	12.7	23.58	13.92	17.58		10.77	15.45	16.67
71.4	64.62	59.53	45.34	69.07		31.57	59.75	54.66
78.7	60	73.7	55.44	43.26		30.43	33.7	49.56
54.81	29.81	67.79	25	59.86		28.61	62.02	72.11
66.28	49.42	58.43	53.78	55.52	61.05		70.35	69.77
34.06		40	33.75	34.06	42.19		64.38	67.5
55.17	37.93	41.01	39.04	17.49	49.38		55.79	55.17
65.99	50.68	65.77	47.75	61.71	59.69		60.13	58.11
							40.55	24.41
72.83	53.15	76.58	39.76	59.84	40.95		38.98	25.98
56.37	28.99	49.85	26.46	36.27	29.29		33.97	29.45
91.23	40.58	68.18	40.26	75	73.05	42.86	52.92	NA
91.23	27.27	57.47	41.56		40.26		82.79	NA

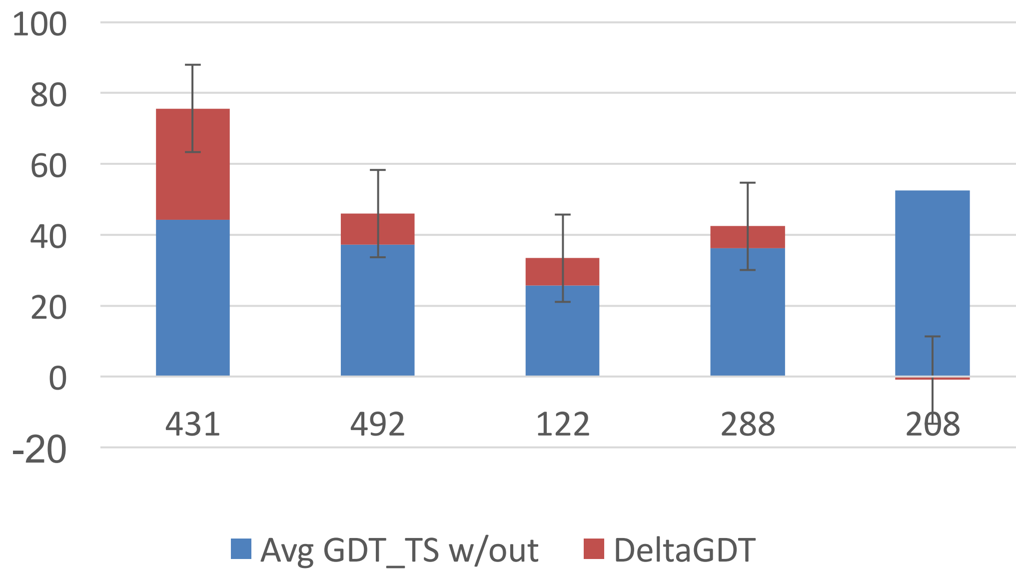
Could predictors use sparse NMR data to improve the accuracy of their models.

Regular vs NMR Assisted

The top performing group (431) was greatly enhanced (25 pts) by NMR data



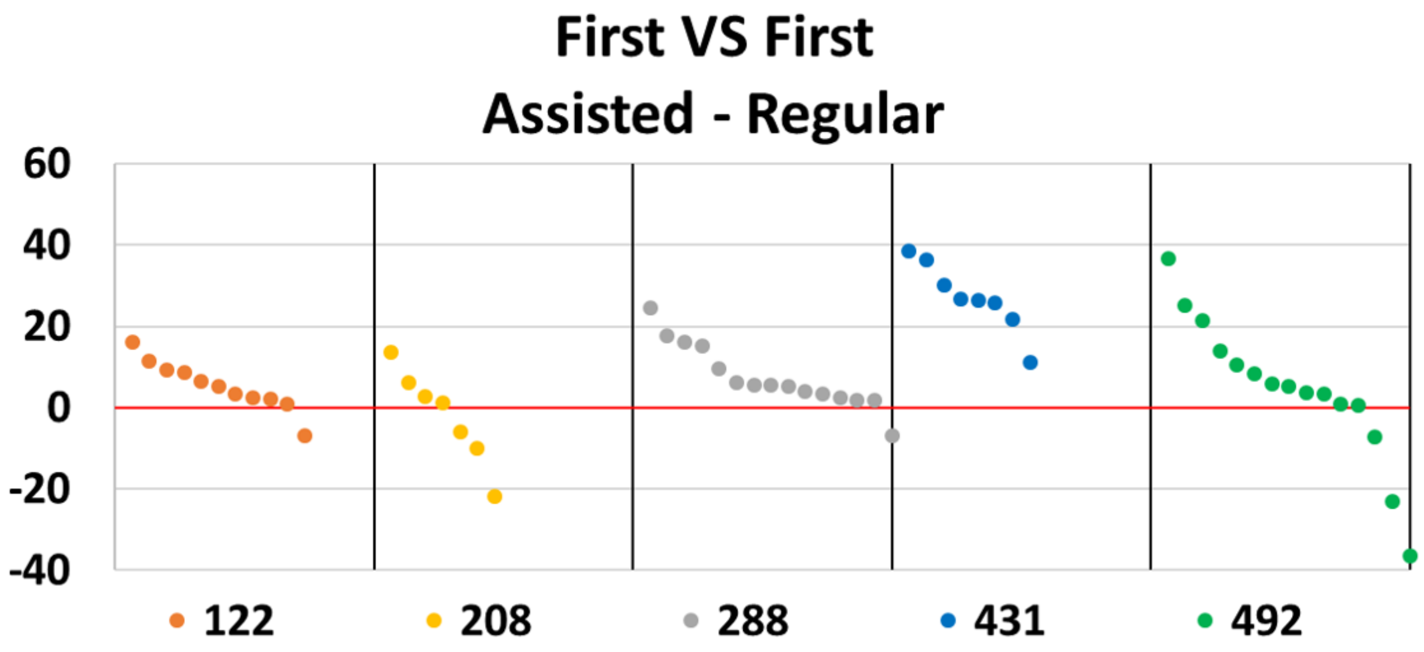
First vs First

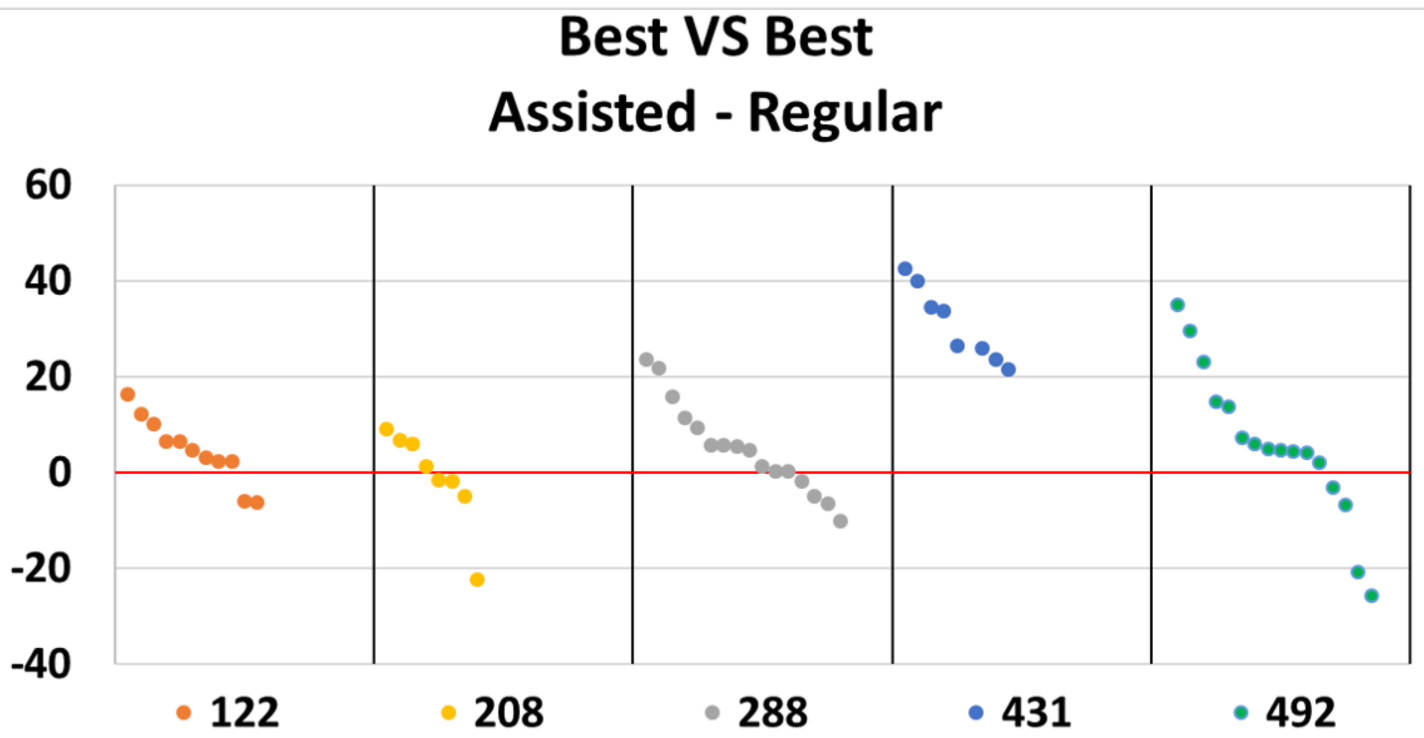


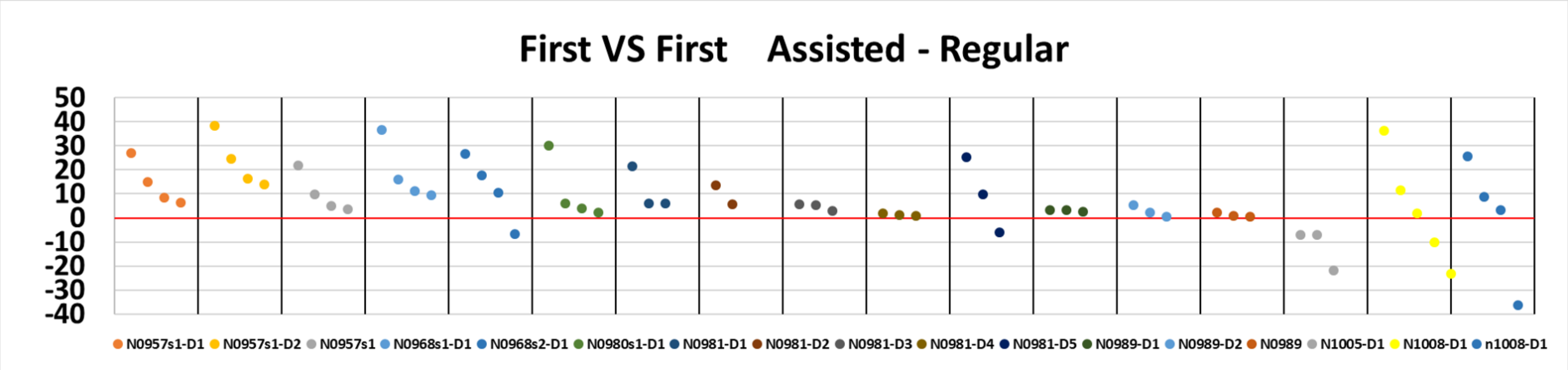
Best vs Best

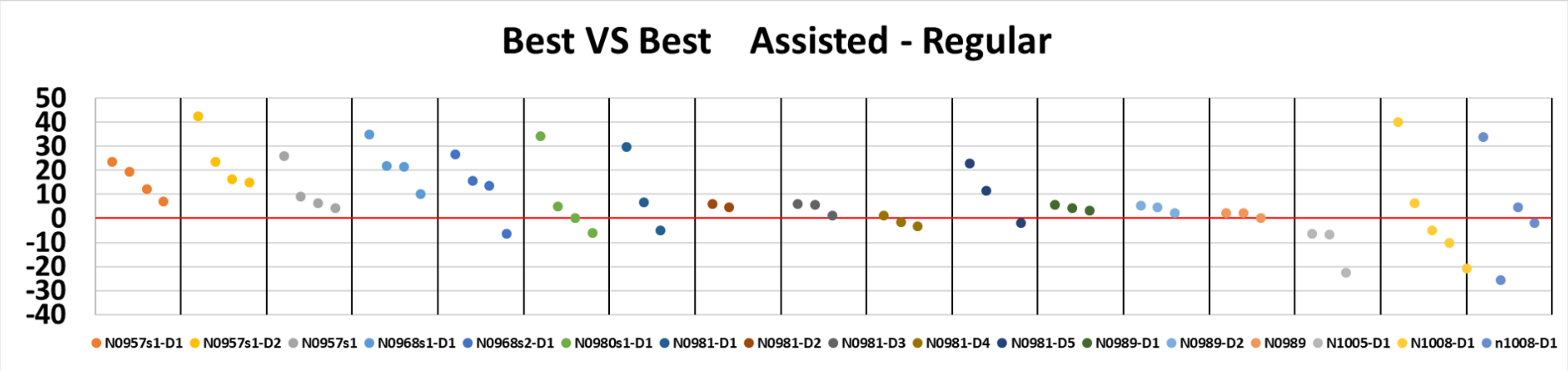
Group 250 and Janet groups provided no "regular" predictions

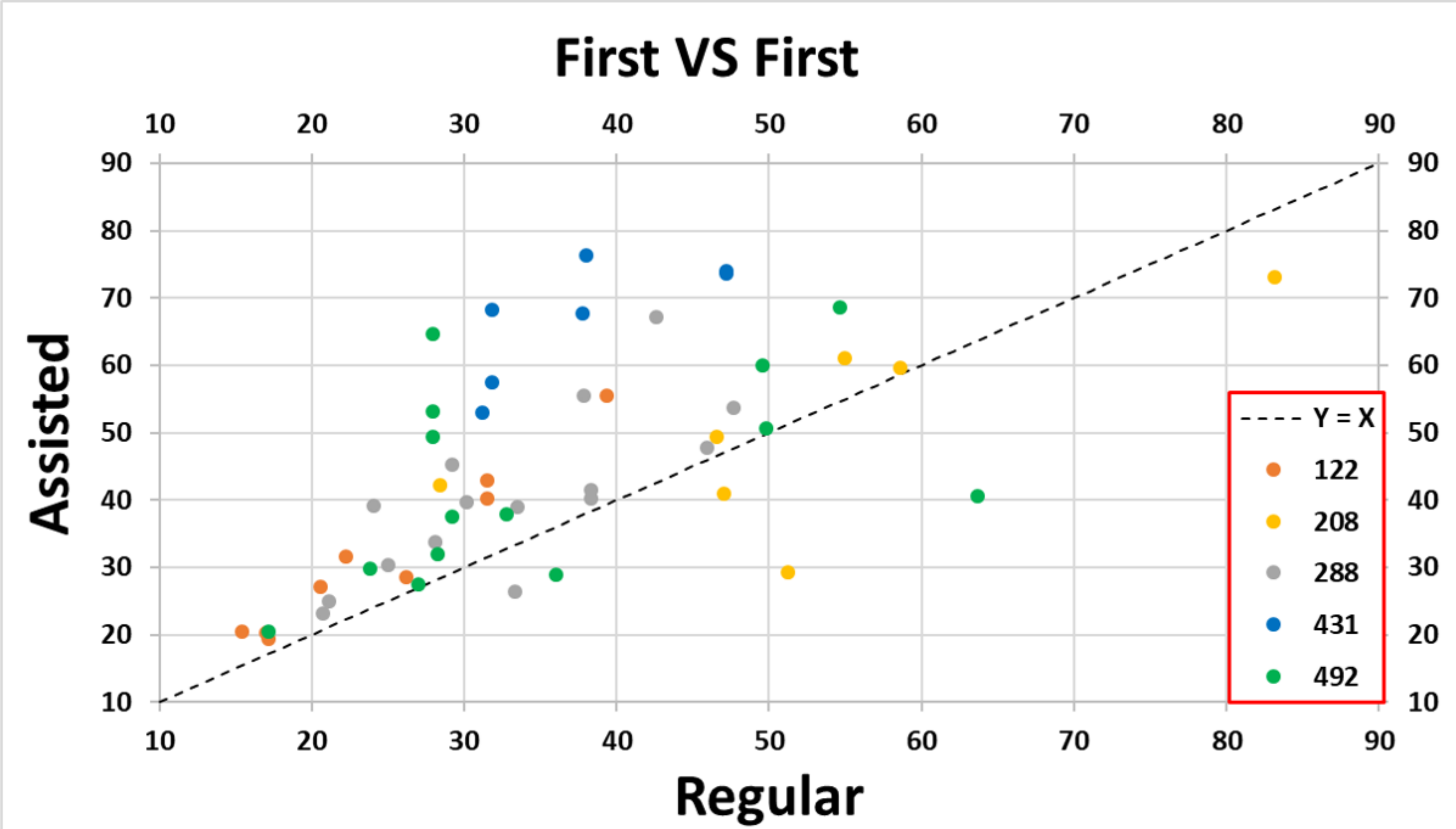
A. Rosato


First VS First
Assisted - Regular
60
40
20
0
-20
-40
122
208
288
431
492


Best VS Best
Assisted - Regular
60
40
20
0
-20
-40
122
208
288
431
492

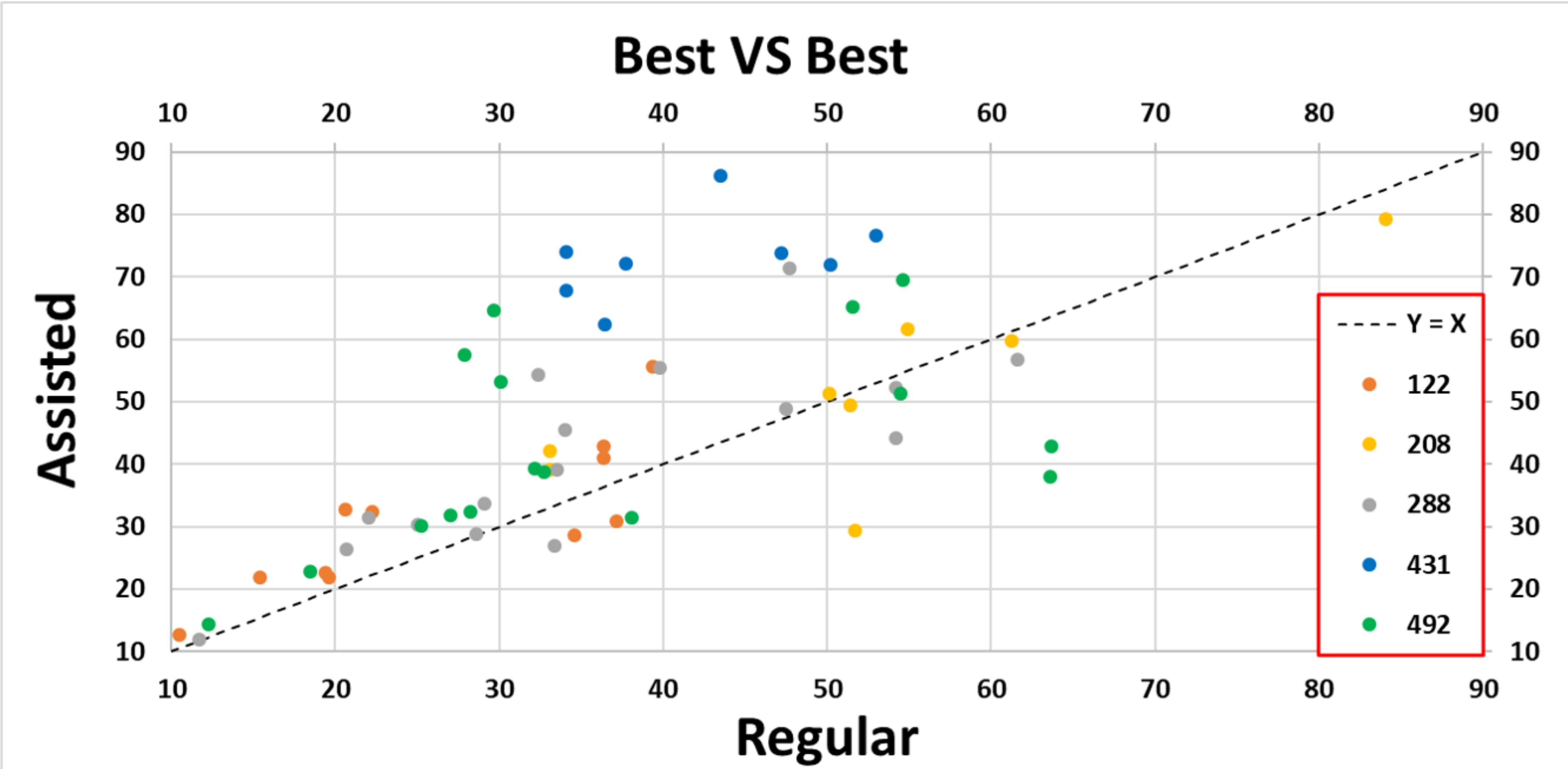

First VS First Assisted - Regular
50
40
30
20
10
0
-10
-20
-30
-40
N0957s1-D1 N0957s1-D2 N0957s1 N0968s1-D1 N0968s2-D1 N0980s1-D1 N0981-D1 N0981-D2 N0981-D3 N0981-D4 N0981-D5 N0989-D1 N0989-D2 N0989 N1005-D1 N1008-D1 n1008-D1


Best VS Best Assisted - Regular
50
40
30
20
10
0
-10
-20
-30
-40
N0957s1-D1 N0957s1-D2 N0957s1 N0968s1-D1 N0968s2-D1 N0980s1-D1 N0981-D1 N0981-D2 N0981-D3 N0981-D4 N0981-D5 N0989-D1 N0989-D2 N0989 N1005-D1 N1008-D1 n1008-D1



122 - Forbidden
208 - KIAS
250 - Meilerlab
288 - UNRES
431 - Laufer
492 - wfBakerUNRES

Baseline Predictions
321 - Janet_ASDP No ECs
459 - Janet_ASDP MetaPSICOV ECs
313 - Janet_ASDP Best Method



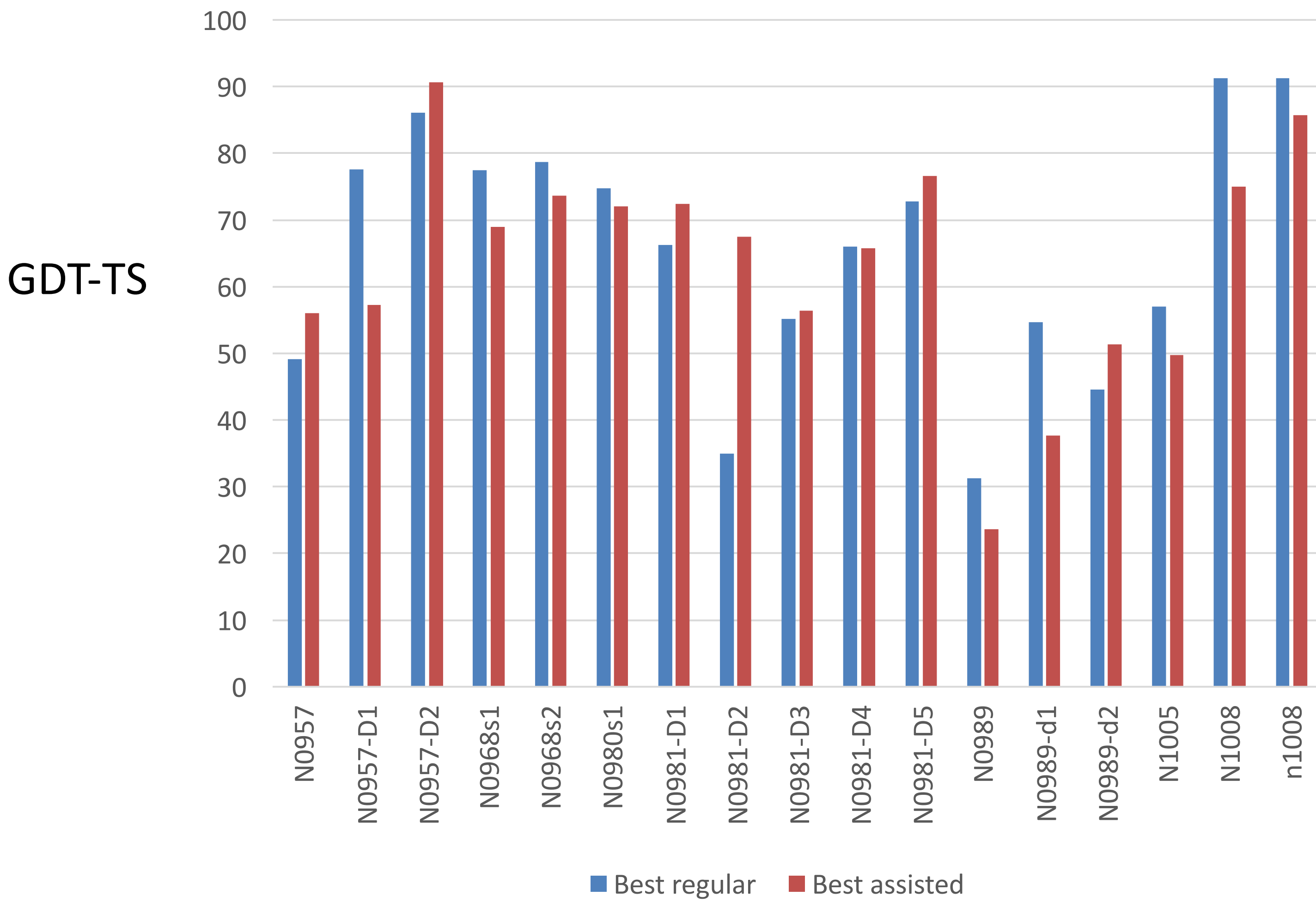
122 - Forbidden
208 - KIAS
250 - Meilerlab
288 - UNRES
431 - Laufer
492 - wfBakerUNRES

Baseline Predictions
321 - Janet_ASDP No ECs
459 - Janet_ASDP MetaPSICOV ECs
313 - Janet_ASDP Best Method

Do predictors using sparse NMR data have higher accuracy than the *best* non-data-assisted predictors

Regular vs NMR Assisted

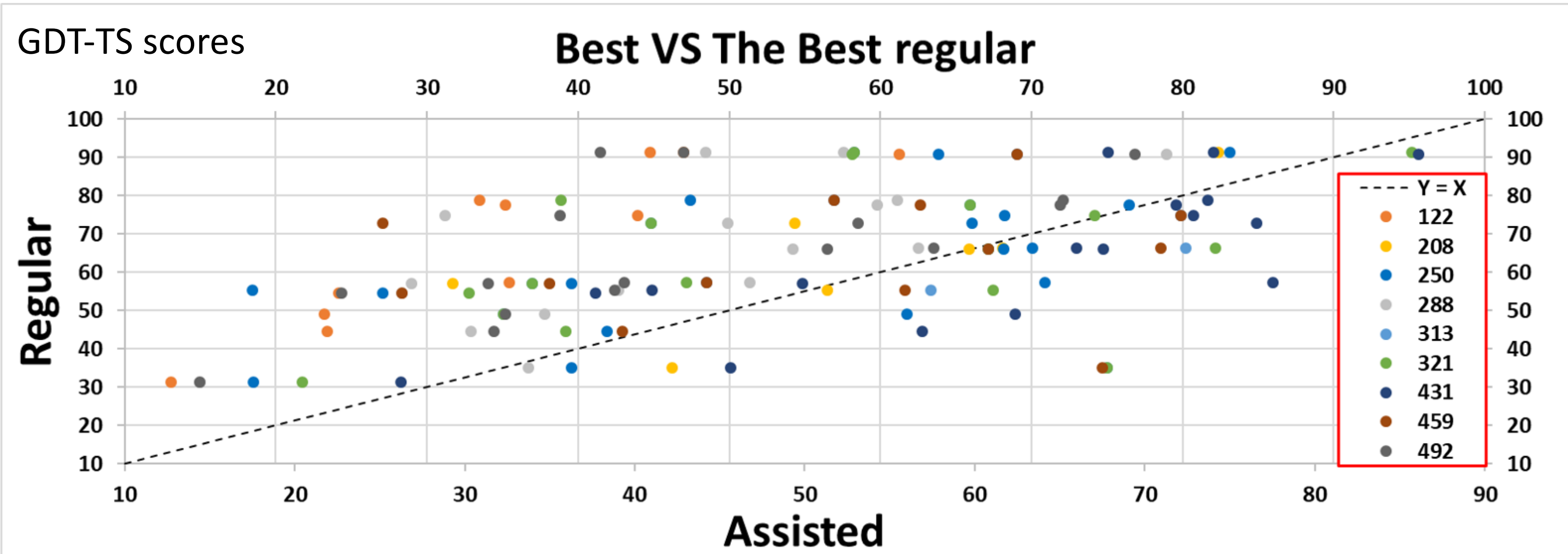
REMARKABLE RESULT!!



Best among all regular predictions vs
best NMR-assisted prediction for each target

A. Rosato

In many cases the best "regular" prediction for a target was more accurate than the best "data assisted" prediction.



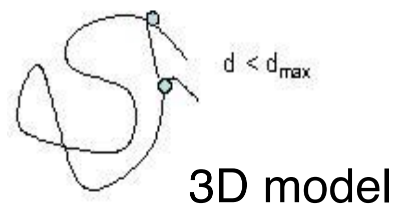
GROUP	RANK 1	RANK 1 or 2	RANK 1 or 2 or 3
043 - A7D	10 / 16	18 / 32	27 / 48
322- Zhang	2 / 16	4 / 32	4 / 48
366-Venclovas	1 / 16	2 / 32	2 / 48
266s – slbio_serve	1 / 16	1 / 32	1 / 48
281 – SHORTLE	1 / 16	1 / 32	1 / 48
089 – MULTICOM	1 / 16	1 / 32	1 / 48

How is the ranking of NMR-Assisted predictors affected if we assess against data rather than reference structure?

NOESY data

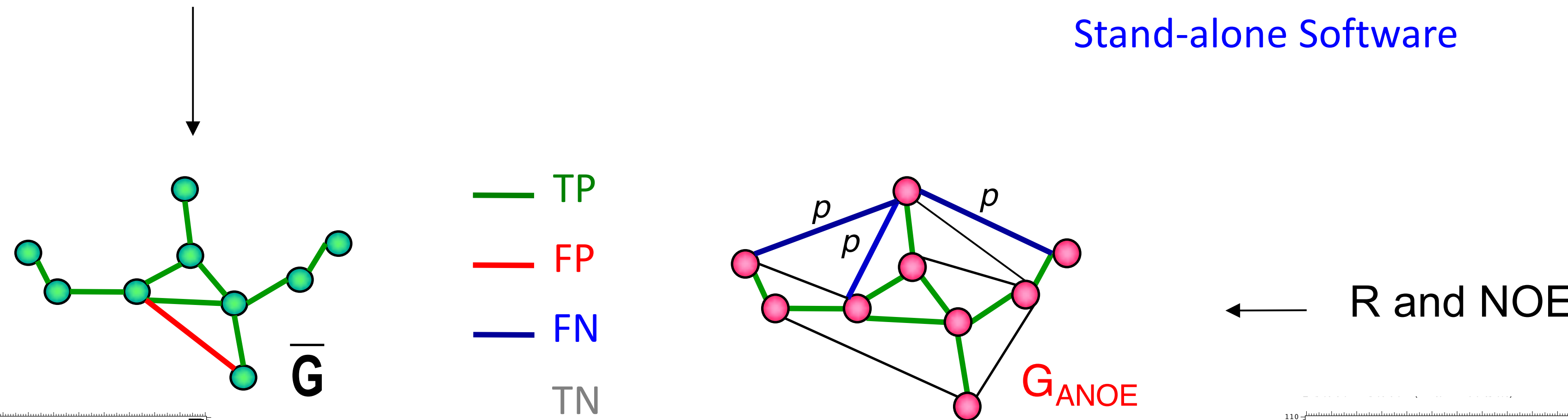
RDC data

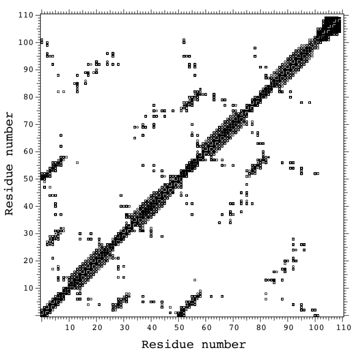
RPF-DP Scores



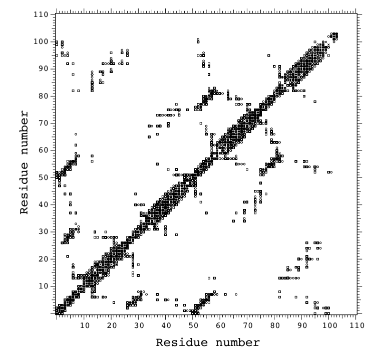
RPF An NMR Protein Structure Quality Assessment Tool

RPF Server and
Stand-alone Software



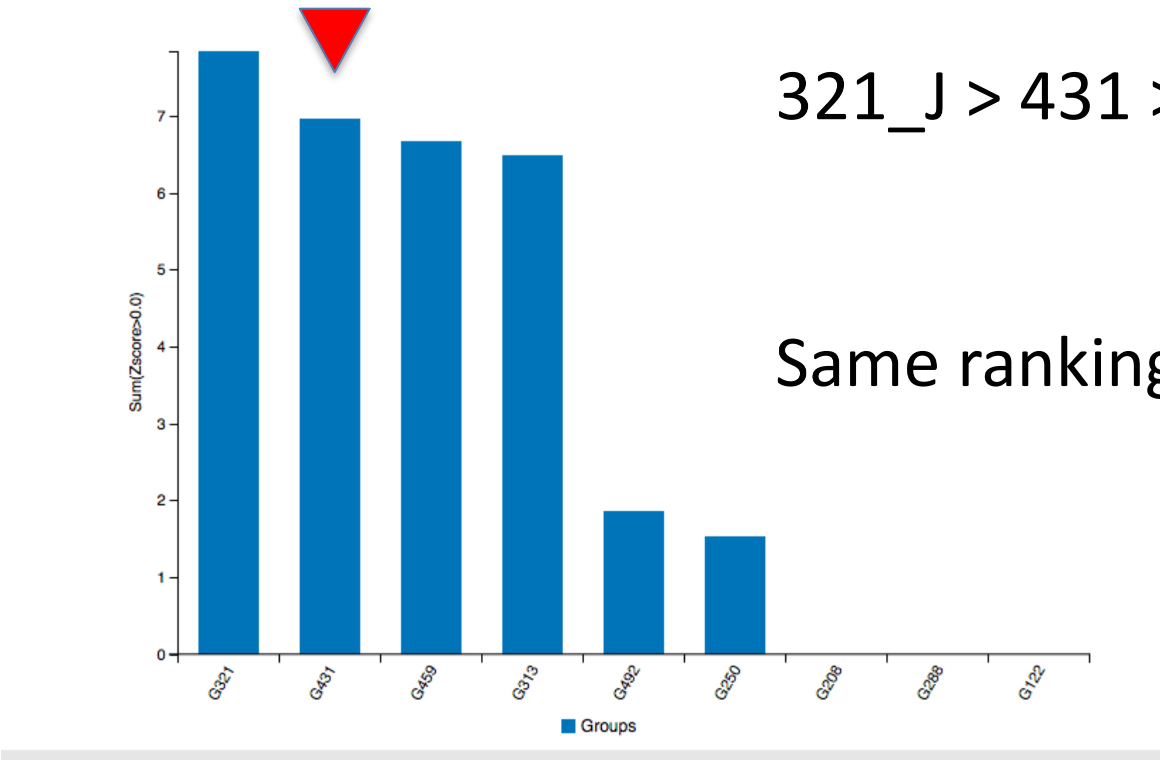


By comparing the differences between the two graphs $\overline{G}$ (derived from the structure) and G_{ANOE} (derived from the peaklists), a global measure of the goodness-of-fit of the query structures with the original peaklists can be formulated.



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

DP Score : Z-Score Based Ranking (Z = 0 Threshold, Model 1)

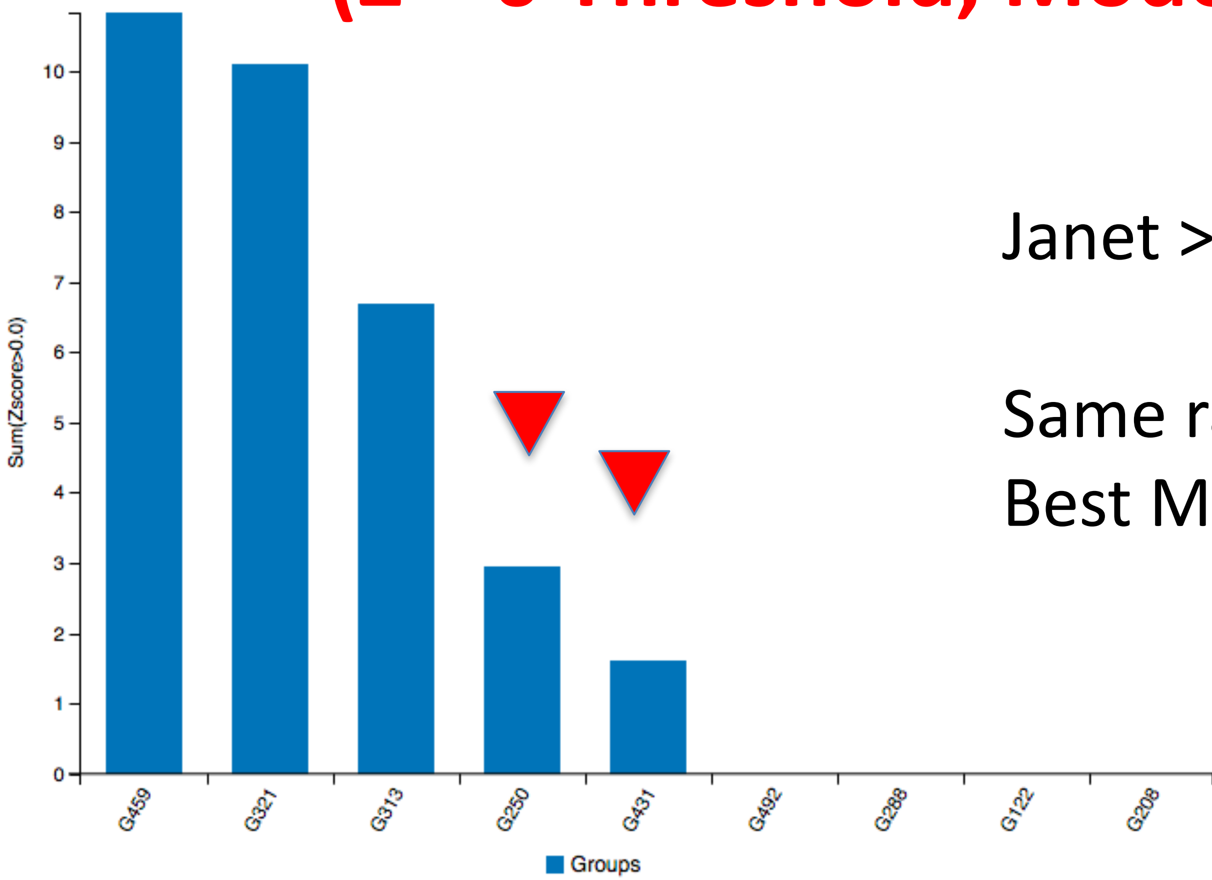


321_J > 431 > 459_J > 313_J

Same ranking using Best Model

Group 431 also does very well with DP score!

RDC1: Z-Score Based Ranking (Z = 0 Threshold, Model 1)

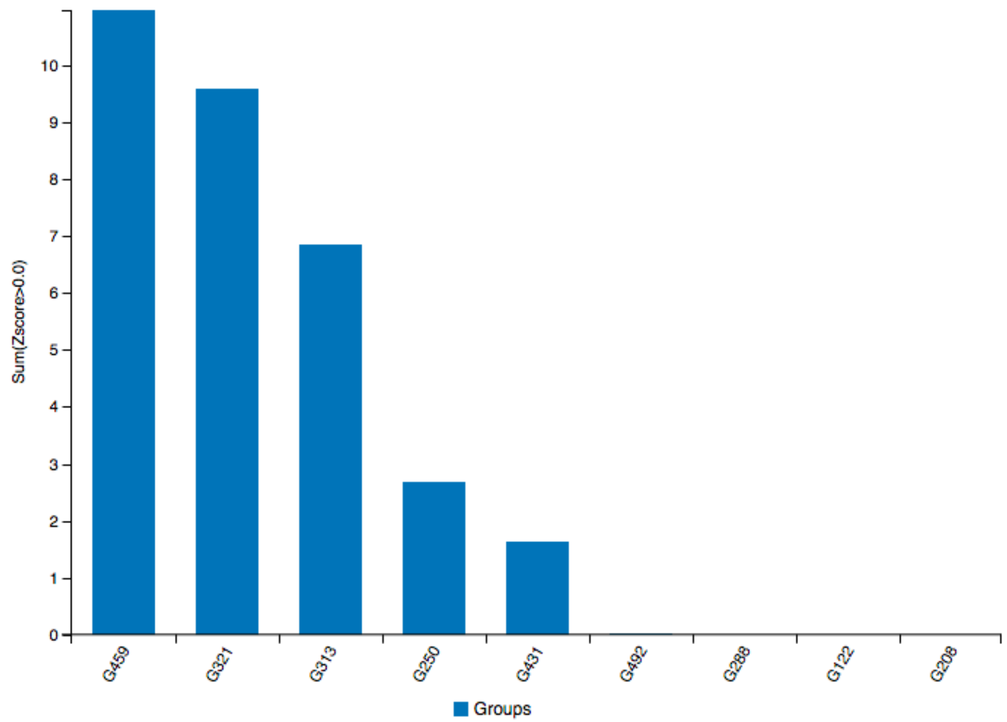


Janet > 250 > 431

Same ranking using Best Model

Groups 250 and 431 do well on RDC scoring - probably used RDC data.

RDC1: Z-Score Based Ranking (Z = 0 Threshold, Model 1)



Janet > 250 > 431

Same ranking using Best Model

Group 250 does well -
probably used RDC data.

Sidechain Rotamer comparisons between predicted and reference structures.

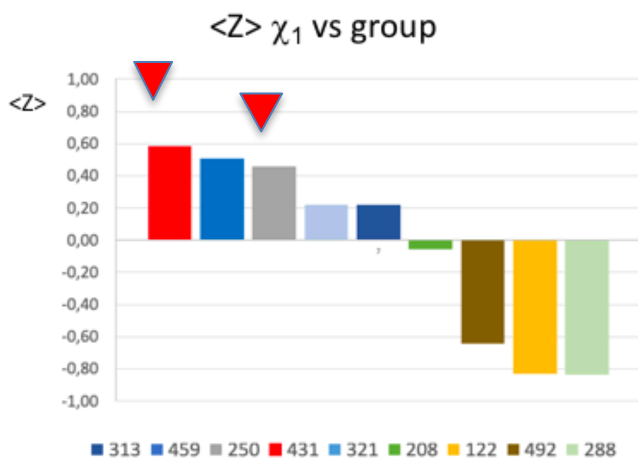
Rotamer states for residues with both buried and converged side chains were compared between the predicted models and the corresponding reference structure.

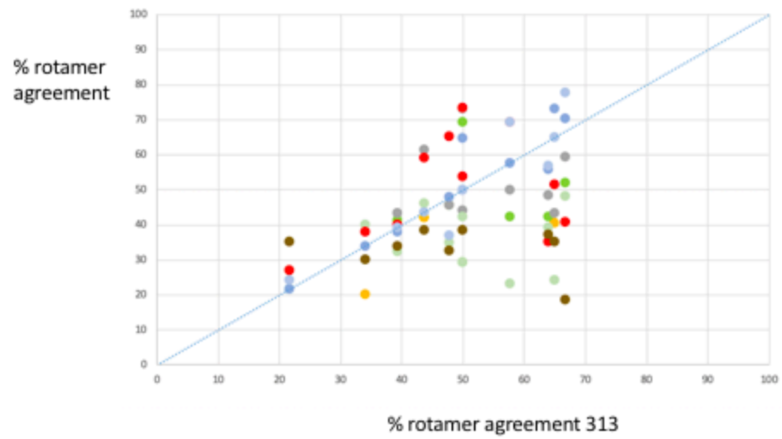
The χ_1 and χ_2 rotamers for all residues in each reference structure were assigned to the nearest g^+, t, or g^- conformational state.

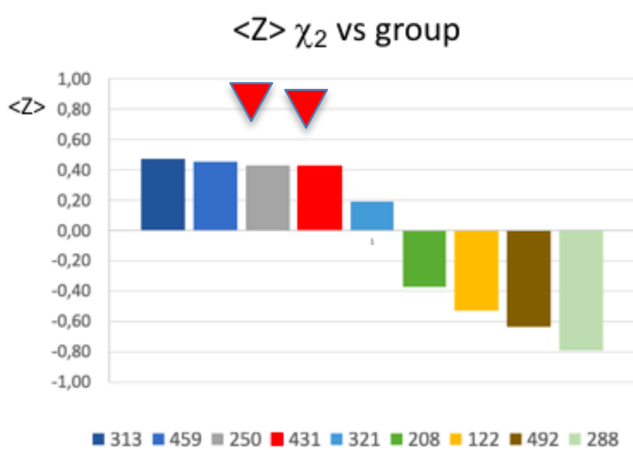
Side chains with solvent accessible surface area less than 40 $Å^2$ in the reference structure (calculated using the program Molmol) were considered as buried side chains.

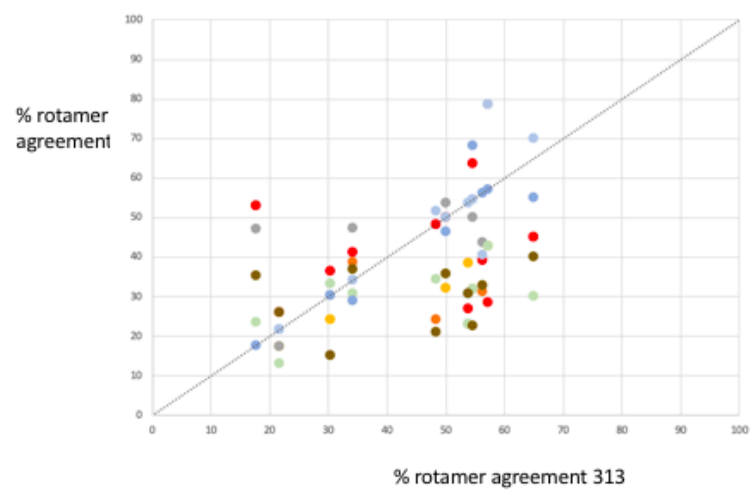
For NMR-derived reference structures, the medoid conformer of the ensemble was selected as the representative structure.

Chi-1 and Chi-2 Rotamer Agreement









PDBStat
Roberto Tejero

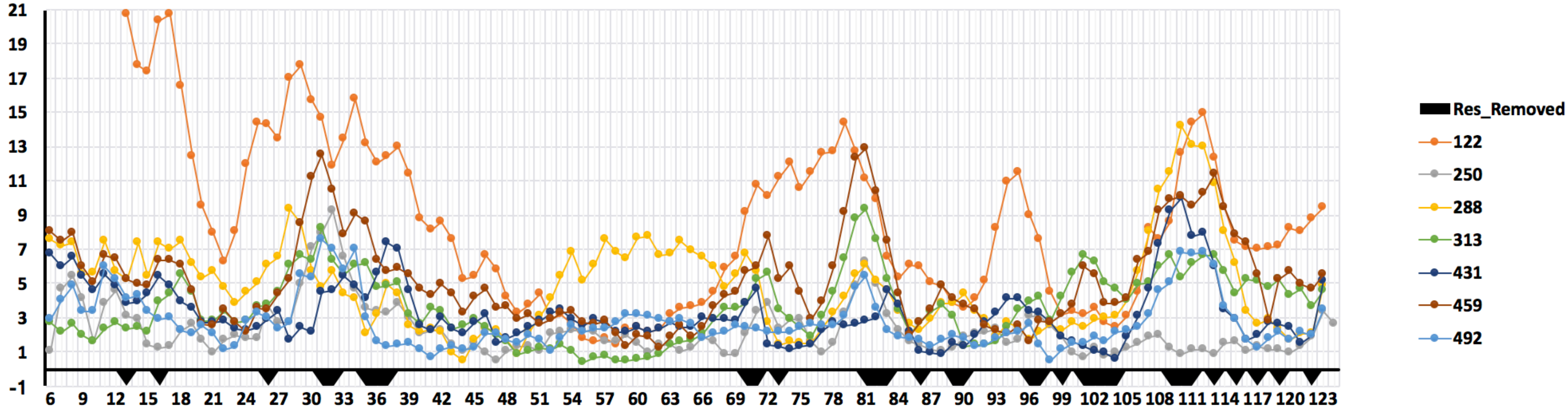
Why did data-guided groups 431 and 250 provide more accurate structures than Janet baseline.

Was this difference largely due to regions with missing?

Local Backbone RMSD vs Sequence

Group 313 – ASDP Baseline

Target N0968s1



There is some tendency to have higher local rmsd for ASDP method in regions where data is missing, which can be overcome to some extent by prediction methods.

Future CASP Challenges

Ongoing process of generating CASP Commons Targets, Data (NMR, SAXS, X-Link, FRET), and Structure

Modeling Multiple Conformational States

Modeling Using Unassigned NOESY spectra

Modeling Using Unassigned RDC data

Combining SAXS, NMR, X-Link, FRET, CryoEM with advance modeling / prediction methods.

J. Y. Huang

G. Liu

A. Rosato

D. Sala

D. Snyder

R. Tejero

H. Valafar

A. Kryshtafovych