PROTEINS:
Structure, Function, and Bioinformatics

Quantitative use of chemical shifts for the modeling of protein complexes

Journal:	*PROTEINS: Structure, Function, and Bioinformatics*
Manuscript ID:	Prot-00076-2011.R1
Wiley - Manuscript type:	Research Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Stratmann, Dirk; University Pierre et Marie Curie, Paris 6, IMPMC Boelens, Rolf; Utrecht University, Faculty of Science, Department of Chemistry Bijvoet Center for Biomolecular Bonvin, Alexandre; Utrecht University, Faculty of Science, Department of Chemistry
Key Words:	Structural biology, Structural bioinformatics, NMR, Docking, Scoring, HADDOCK

Quantitative use of chemical shifts for the modelling of protein complexes

Dirk Stratmann[1,2], Rolf Boelens[1], Alexandre M.J.J. Bonvin[1,*]

22nd April 2011

Short title: CS-HADDOCK

Keywords: Structural biology, Structural bioinformatics, NMR, Docking, Scoring, HADDOCK

1) Bijvoet Center for Biomolecular Research, Faculty of Science, Utrecht University

Padualaan 8, 3584 CH Utrecht, the Netherlands

2) *current address*:

IMPMC, UMR 7590 CNRS, University Pierre et Marie Curie, Paris 6

Case courrier 115; 4, place Jussieu, 75252 Paris cedex 05, France

∗ *Correspondence to*:

Prof. Alexandre M.J.J. Bonvin

Computational Structural Biology, Bijvoet Center for Biomolecular Research

Faculty of Science, Utrecht University

Padualaan 8, 3584 CH Utrecht, the Netherlands

Phone: +31-30-2533859, Fax: +31-30-2537623

Email: a.m.j.j.bonvin@uu.nl

1 Abstract

Despite recent advances in the modelling of protein-protein complexes by docking, additional information is often required to identify the best solutions. For this purpose, NMR data deliver valuable restraints that can be used in the sampling and/or the scoring stage, like in the data-driven docking approach HADDOCK that can make use of NMR chemical shift perturbation (CSP) data to define the binding site on each protein and drive the docking. We show here that a quantitative use of chemical shifts (CS) in the scoring stage can help to resolve ambiguities. A quantitative CS-RMSD score based on $^{1}H^{\alpha}$,$^{13}C^{\alpha}$ and ^{15}N chemical shifts ranks the best solutions always at the top, as demonstrated on a small benchmark of four complexes. It is implemented in a new docking protocol, CS-HADDOCK, which combines CSP data as ambiguous interaction restraints in the sampling stage with the CS-RMSD score in the final scoring stage. This combination of qualitative and quantitative use of chemical shifts increases the reliability of data-driven docking for the structure determination of complexes from limited NMR data.

2 Introduction

Over the last years, it has been shown that the combination of protein structure prediction programs with experimental NMR chemical shifts can already be sufficient to obtain high-resolution structures of small to medium-sized proteins.[1–3] The approaches developed for this purpose require reasonably accurate predictions of chemical shifts. Thanks to the growing number of protein structures solved by NMR for which chemical shifts have been deposited into the BioMagResBank (BMRB)[4], chemical shifts can be predicted from such databases. Chemical shift predictors are already quite accurate in grasping short-range conformational effects on chemical shifts from such databases and long-range effects, like electrostatics or ring-current effects, from classical equations.[5–10]

Chemical shifts are also used in the context of biomolecular complexes. Measurements of chemical shifts on both the free and complexed forms of a protein yield chemical shift perturbation (CSP) data. Chemical shifts of residues in the interface of the complex are likely to differ between the bound and

the free forms. The perturbation of the chemical shift upon complex formation can be used to map the interaction interface and model protein complexes from the known free form structures.[11–13] For example, the data-driven docking program HADDOCK converts CSP data into ambiguous distance restraints (AIRs) between the two proteins.[14–16]

CSP data are more widely used in a qualitative rather than quantitative manner. They have been used quantitatively mainly for the binding of small molecules to proteins,[17–23] and for the ranking of heme-containing protein-protein complexes obtained with HADDOCK[24], as aromatic rings of small ligands and heme groups generate significant CSP on the protein's protons, and in combination with residual dipolar couplings on the EIN-HPR complex.[25] With the introduction of the CamDock protocol, chemical shift data for various nuclei were used for the first time quantitatively and without any other data to model the E9-Im9 complex.[26]

For the quantitative use of chemical shifts for the modelling of protein complexes, we developed the CS-HADDOCK protocol as an extension of the widely used docking program HADDOCK. We tested the method using the few complete chemical shift data sets of protein complexes currently available from the BMRB resulting in a small benchmark of four protein-protein complexes. Our results on those complexes show that not all chemical shift types are equally useful in defining the complex: $^{1}H^{\alpha}$,$^{13}C^{\alpha}$ and ^{15}N chemical shifts are the most useful, while including $^{1}H^{N}$ and $^{13}C^{\beta}$ shifts, in combination or separately, gives worse results (see supplementary information). Furthermore, we show that the quantitative use of chemical shifts is only robust if the interaction site is approximately known in advance to restrict the search space, for example from a qualitative analysis of CSP data.

3 Materials and Methods

3.1 Input structures and chemical shift data

CS-HADDOCK was tested on four complexes (see Table I): E9 - IM9 (PDB-ID 1EMV), EIN - HPR (PDB-ID 3EZA), Z_{Taq} - anti-Z_{Taq} (PDB-ID 2B87) and ILK ARD - PINCH-1 LIM1 (PDB-ID 3F6Q). The unbound starting structures used for the docking and their distances in terms of C^{α}-RMSD to the

Table I: Reference PDB structures and CS-data of the complexes

name	PDB-ID	Experimental method	CS-data (BMRB-ID[4])
E9-IM9	1EMV	X-ray (1.7Å)	4352 (E9), 4115 (IM9)
EIN-HPR	3EZA	NMR	4264
Z_{Taq} - anti-Z_{Taq}	2B87	NMR	6806
ILK ARD - PINCH-1 LIM1	3F6Q	X-ray (1.6Å)	16063

Table II: Number of CS of the complexes

name	$^{1}H^{\alpha}$	$^{1}H^{N}$	^{15}N	$^{13}C^{\alpha}$	$^{13}C^{\beta}$
E9	119	122	122	131	95
IM9	85	81	81	86	79
EIN	238	248	248	253	238
HPR	84	81	81	85	79
Z_{Taq}	58	53	53	58	56
anti-Z_{Taq}	58	54	54	58	57
ILK ARD	171	164	165	171	157
PINCH-1 LIM1	70	67	67	70	65

reference bound structures are listed in Table III. In general, the higher the C^{α}-RMSD values the more difficult is the docking. The first two complexes, E9-IM9 and EIN-HPR, are in a moderate range of 0.5-2Å. The Z_{Taq} - anti-Z_{Taq} complex is already in a difficult range of 1.5-3.5Å, and the PINCH-1 LIM1 input structures are in a even more difficult range of 4.3-5.4Å C^{α}-RMSD. For each model the interface-RMSD (defined as the backbone RMSD over all residues within 10Å of the partner molecule) to the reference complex structure was also calculated.

3.2 Docking and scoring protocol

Figure 1 shows a flowchart of the CS-HADDOCK protocol which is explained here in detail. The standard HADDOCK 2.1 protocol[15, 16] was used to generate the models of the protein complexes,

Table III: Unbound, free-form PDB structures used for the docking.

name	PDB-ID	Experimental method	C^{α}-RMSD to reference-complex	interface-RMSD to reference-complex	residues
E9	1FSJ, chain B	X-ray (1.8Å)	0.96Å (1EMV:B)	0.39Å	134
Im9	1IMP (all 21 structures)	NMR	1.4-2.0Å (1EMV:A)	1.15-1.49Å	86
EIN	1ZYM, chain A	X-ray (2.5Å)	1.48Å (3EZA:A)	0.98Å	249
HPR	1POH	X-ray (2.0Å)	0.64Å (3EZA:B)	0.64Å	85
Z_{Taq}	2B88 (all 40 structures)	NMR	1.63-3.43Å (2B87:A)	1.18-2.93Å	58
anti-Z_{Taq}	2B89 (all 40 structures)	NMR	2.35-2.67Å (2B87:B)	0.84-0.99Å	58
ILK ARD	3IXE, chain A	X-ray (1.9Å)	0.65Å (3F6Q:A)	0.32Å	171
PINCH-1 LIM1	1G47 (all 25 structures)	NMR	4.34Å-5.39Å (3F6Q:B)	3.58-4.17Å	70

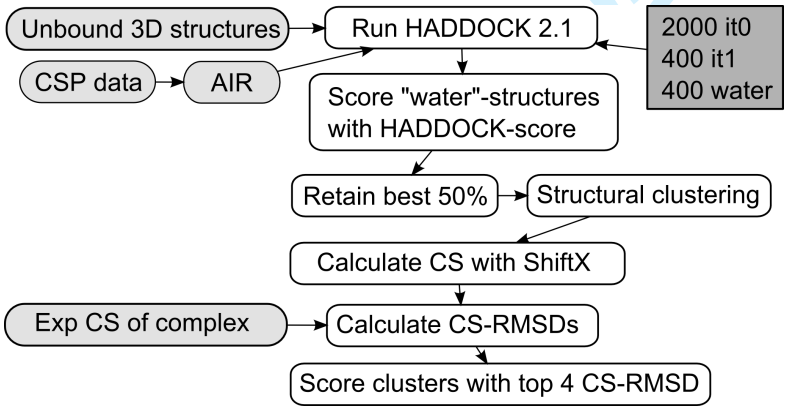


Figure 1: CS-HADDOCK protocol flowchart (see article for explanations)

starting from the unbound input structures listed in Table III. 2000 models were generated in the rigid-body docking step (it0-step), from which the top 400 according to the HADDOCK score were further refined with a flexible interface (it1-step) and with a water-layer around the complex (water-step). $[^1H^N, ^{15}N]$ chemical shift perturbation (CSP) data were used to define the residues potentially involved in binding (active + passive residues) and from these ambiguous distance restraints (AIRs) (see Table S1). The final 400 water-refined complex structures were first ranked according to the standard HADDOCK score[15], which includes the AIR, electrostatic and van der Waals energies and an empirical desolvation[27] term:

$$E_{HADDOCK} = 0.1 * E_{AIR} + 0.2 * E_{elec} + 1.0 * E_{vdW} + 1.0 * E_{desolv} \quad (1)$$

The top 200 structures were clustered according to their pairwise interface ligand RMSD-matrix (RMSD of the backbone interface atoms of the ligand calculated after superimposition on the backbone interface atoms of the receptor). Structures falling into a cluster were further rescored with a new CS-RMSD score, defined as follows for one chemical shifts type (e.g. H^{α}-CS):

$$\text{CS-RMSD}_{\,k} \;=\; \frac{\sqrt{\frac{\sum_{i=1}^{n_A}(\delta_i^{exp}-\delta_{i,k}^{theo})^2}{n_A}} + \sqrt{\frac{\sum_{i=1}^{n_B}(\delta_i^{exp}-\delta_{i,k}^{theo})^2}{n_B}}}{2} \quad (2)$$

Theoretical chemical shifts $\delta_{i,k}^{theo}$ (i = residue number, k = model number) are calculated from the generated complex structures with ShiftX.[7] Experimental chemical shifts δ_i^{exp} are those of the complex (see Table I and II). We tested also other chemical shift predictor programs including SPARTA[8], ShiftS[5] and 4DSPOT[10]. Since we did not find any difference in the performance of the CS-RMSD score we chose ShiftX because of its speed of execution.

The CS-RMSDs are calculated for each binding partner separately and combined to an average CS-RMSD value of the complex. Each binding partner has so the same weight for the CS-RMSD score, regardless of its number of residues or amount of available chemical shifts. Chemical shifts of different nuclei are combined as follows: the CS-RMSD values of all generated models are calculated

for each nucleus separately and then normalized to a scale from 0.0 to 1.0:

$$\text{n-CS-RMSD}_k = \left(\frac{\text{CS-RMSD}_k - \text{CS-RMSD}_{min}}{\text{CS-RMSD}_{max} - \text{CS-RMSD}_{min}} \right) \quad (3)$$

where $\text{CS-RMSD}_{min/max}$ are the minimum, respectively maximum CS-RMSD values among all generated models for a specific nucleus. Finally, the normalized n-CS-RMSD values of each nucleus are summed up to the combined CS-RMSD score, which is also normalized to a scale from 0.0 to 1.0. This approach ensures that each chemical shift type contributes equally to the combined CS-RMSD score. We did not optimize the weights between the different nuclei as in CamDock[26], because this would need a much larger benchmark of protein complexes with chemical shift data. Note that we also investigated a weighting scheme accounting for both the prediction accuracy of ShiftX and the variability of a given nuclei in the BMRB (see supplementary material, Figure S8 and S9). Despite slightly different weights of the various nuclei, the overall scoring performance did not change.

The top 4 structures of each cluster were selected according to the score used (n-CS-RMSD or HADDOCK), and the average score and interface-RMSD were calculated among these 4 structures.

Table IV: Accuracy of the top 4 structures of the best scored cluster. The accuracy is measured in terms of fraction of native contacts f_{nat}[1] and interface-RMSD (i-RMSD[2]) according to the CAPRI standards[28, 29]. The average ± standard deviation, as well as the minimum and maximum, among the values of the top 4 structures is given here.

complex	CS-RMSD score		HADDOCK score	
	f_{nat}	i-RMSD	f_{nat}	i-RMSD
E9-IM9	0.57 ± 0.13 [0.43...0.79]	(1.90 ± 0.30)Å [1.57...2.39Å]	0.06 ± 0.01 [0.05...0.07]	(11.47 ± 0.09)Å [11.35...11.59Å]
EIN-HPR	0.32 ± 0.11 [0.14...0.41]	(2.94 ± 0.89)Å [2.06...4.26Å]	0.53 ± 0.06 [0.44...0.60]	(1.89 ± 0.17)Å [1.68...2.07Å]
Z_{Taq} - anti-Z_{Taq}	0.32 ± 0.06 [0.25...0.38]	(3.04 ± 0.62)Å [2.18...3.93Å]	0.06 ± 0.01 [0.05...0.08]	(8.60 ± 0.39)Å [8.19...9.24Å]
ILK ARD - PINCH-1 LIM1	0.42 ± 0.07 [0.34...0.51]	(4.86 ± 0.88)Å [3.87...5.74Å]	0.06 ± 0.01 [0.03...0.07]	(10.33 ± 0.14)Å [10.11...10.48Å]

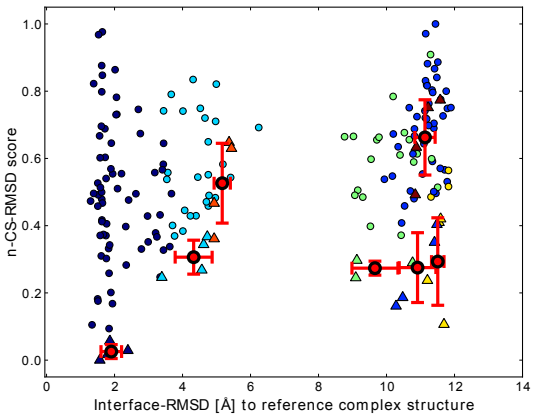
[1] f_{nat} = number of native (correct) residue–residue contacts in the predicted complex divided by the number of contacts in the reference complex. A pair of residues on different sides of the interface was considered to be in contact if any of their atoms were within 5 Å.
[2] i-RMSD = RMSD after optimal superimposition of the backbone atoms of interface residues only in the predicted versus reference complex. Here, a residue belongs to the interface, if it has at least one atom within 10Å of any atom of the partner molecule.

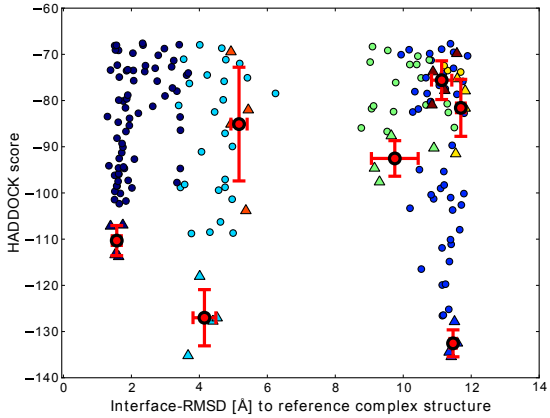
4 Results and Discussion

Figures 2a, 3a, 4 and S1 show the results obtained on the E9-Im9 complex. The left panel of Figure 2a shows the CS-RMSD score of the water-refined models against the interface-RMSD from the reference complex. Although the best generated models (interface-RMSD = 1.3-1.5Å) do not have the best CS-RMSD scores, the best ranked models in terms of CS-RMSD score are still quite close to the reference structure (interface-RMSD = 1.6-2.3Å, see Table IV). Moreover, the models far from the

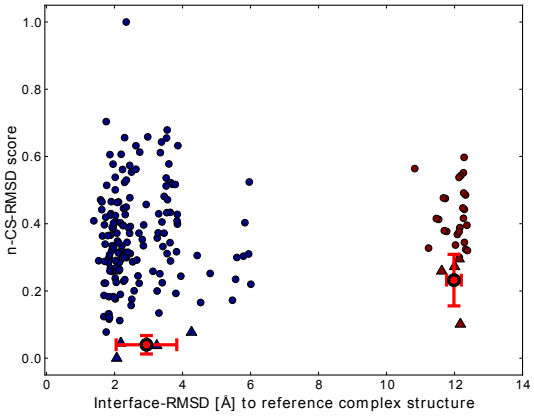
Figure 2 *(following page)*: Comparison of the n-CS-RMSD (left column) and HADDOCK (right column) scores versus interface RMSD from the reference structure of the 200 clustered water-refined models. Each structural cluster has a different color. The triangles indicate the top 4 structures of each cluster. The red crosses indicate the average and the standard deviation of the scores and the interface-RMSDs of the top 4 structures of each cluster. The clusters are ranked according to the average score of the top 4 structures of each cluster. The n-CS-RMSD score is the normalized combined score of the $^{13}C^{\alpha}$,[1] H^{α} and ^{15}N nuclei (see Material and Methods). **(a)** E9-IM9, **(b)** EIN-HPR, **(c)** Z_{Taq} - anti-Z_{Taq} and **(d)** ILK-PINCH

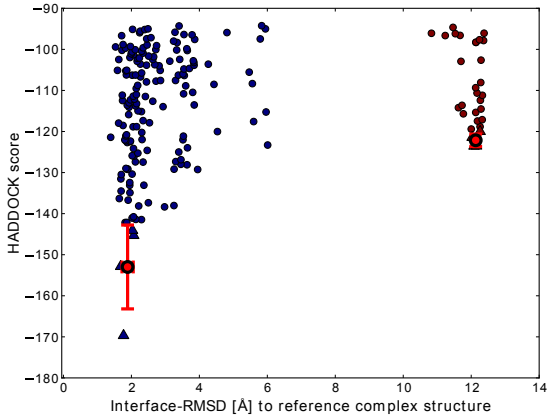
(a) E9-IM9



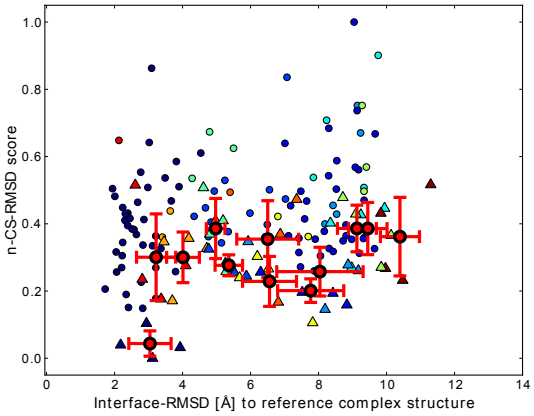


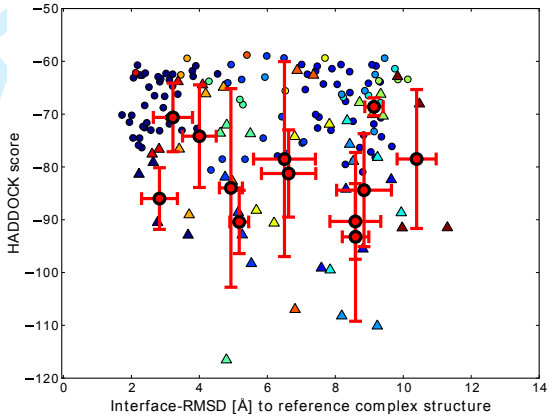
(b) EIN-HPR



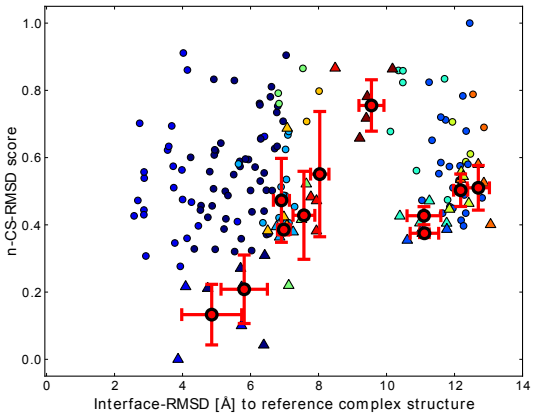


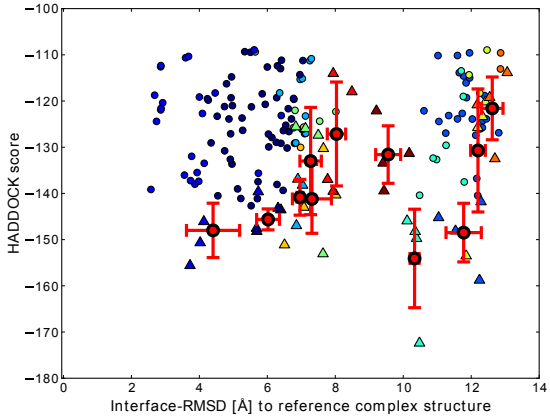
(c) Z_{Taq}-anti-Z_{Taq}





(d) ILK-PINCH





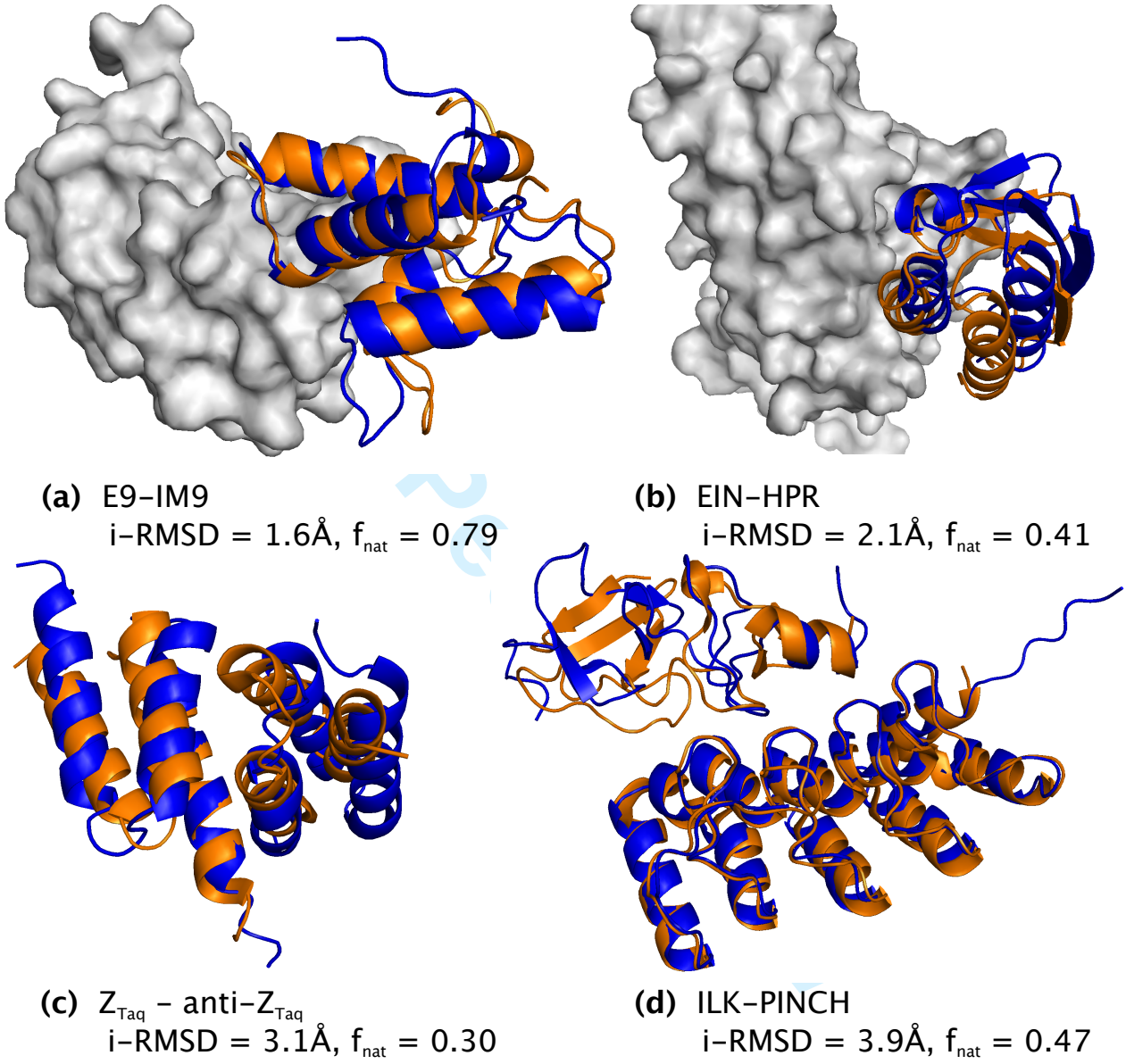
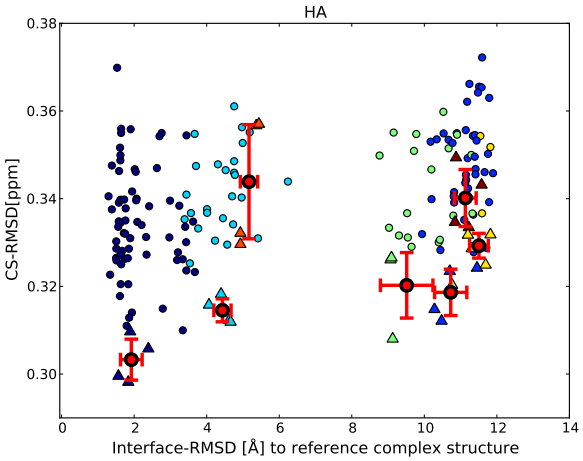
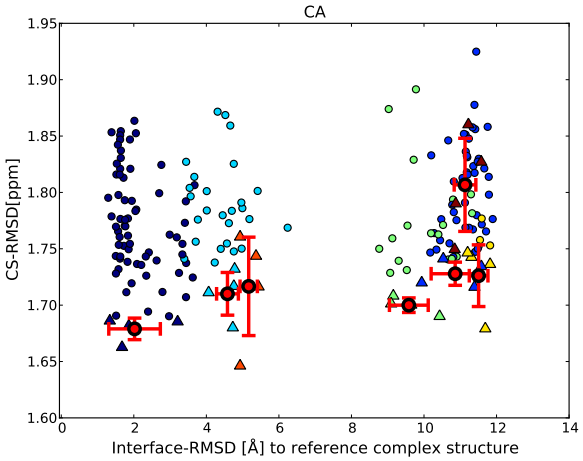


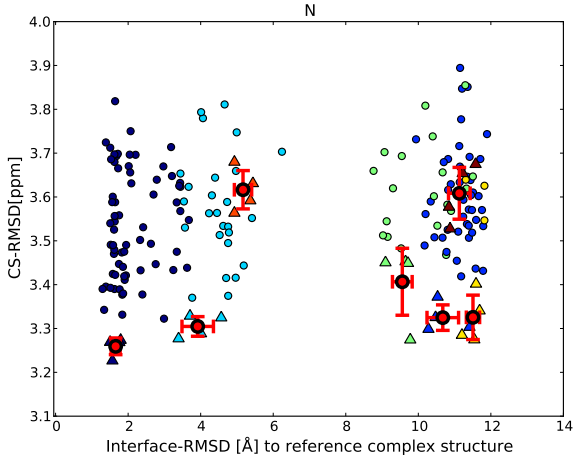
Figure 3: Comparison of the CS-RMSD best-scored model with the reference structure. The structures were fitted on the interface backbone atoms and the interface-RMSD (i-RMSD) and fraction of native contacts f_{nat} values are given. **(a)** E9-IM9 model (IM9 in blue) versus reference structure 1EMV (IM9 in gold and E9 as gray surface). **(b)** EIN-HPR model (HPR in blue) versus reference structure 3EZA (HPR in gold and EIN as gray surface). **(c)** Z_{Taq} - anti-Z_{Taq} model (in blue) versus reference structure 2B87 (in gold). **(d)** ILK-PINCH model (in blue) versus reference structure 3F6Q (in gold).



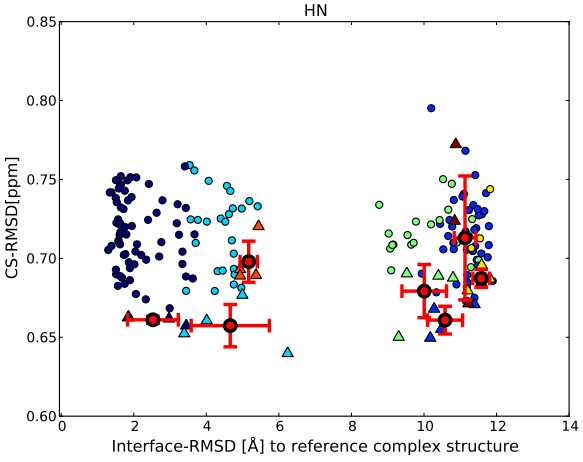
(a) CS-RMSD of $^{1}H^{\alpha}$



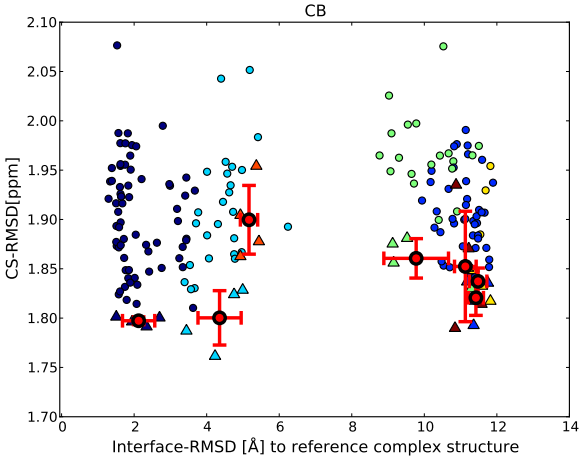
(b) CS-RMSD of $^{13}C^{\alpha}$



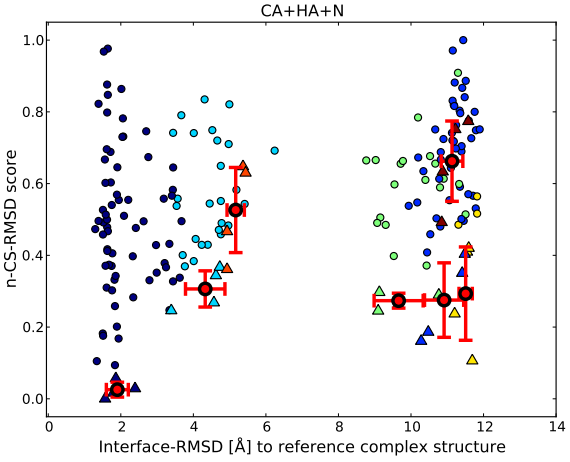
(c) CS-RMSD of ^{15}N



(d) CS-RMSD of $^{1}H^{N}$



(e) CS-RMSD of $^{13}C^{\beta}$



(f) Combined CS-RMSD of $^{13}C^{\alpha} + {}^{1}H^{\alpha} + {}^{15}N$

Figure 4: E9-Im9 complex: (a-e) CS-RMSD scores for single nucleus. (f) Combined n-CS-RMSD score

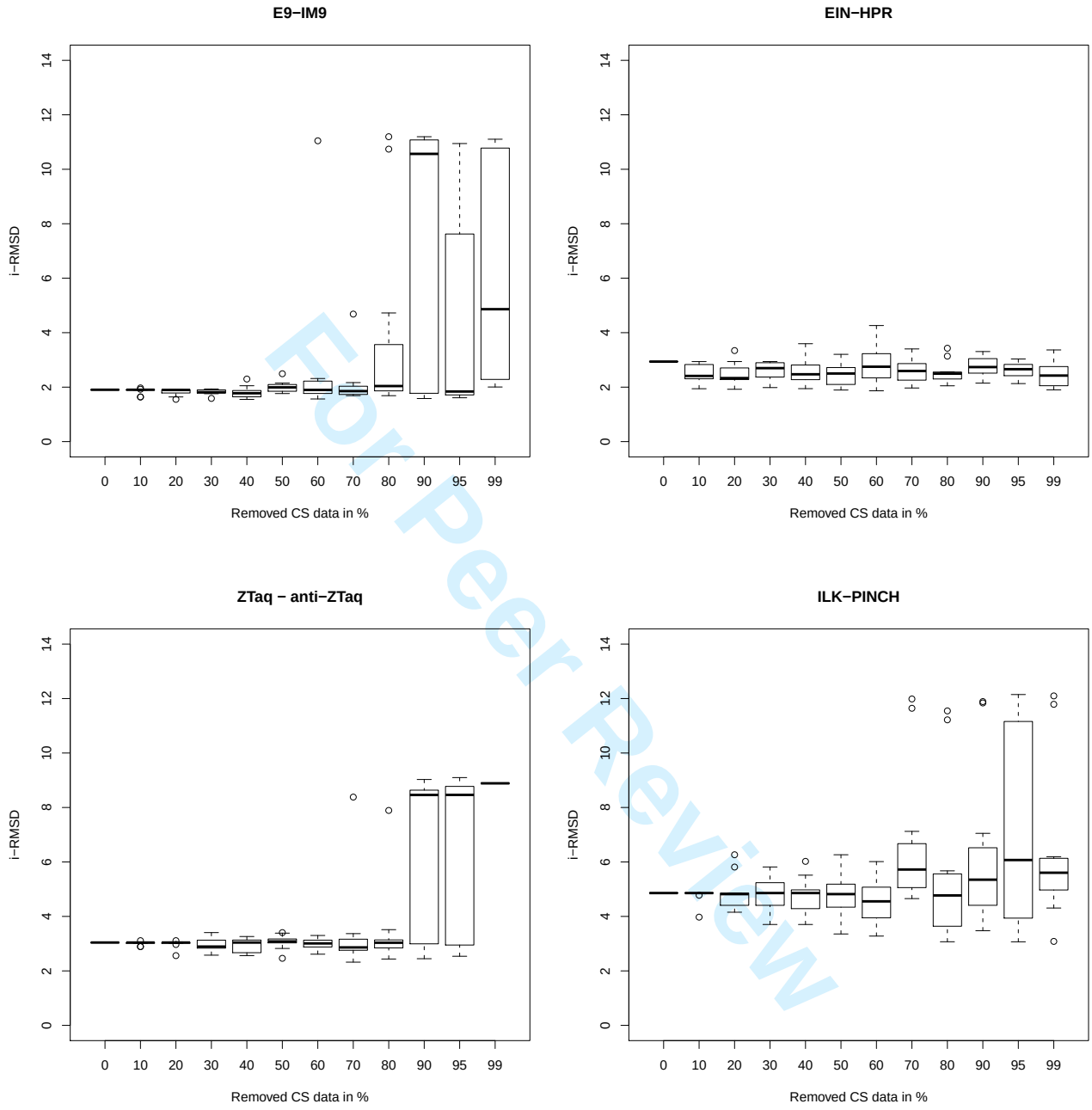


Figure 5: Robustness versus missing CS data: average i-RMSD of the best scored cluster as a function of the fraction of CS data randomly removed. The CS data were randomly removed, separately for each binding partner. This was repeated 50 times. The resulting distribution of i-RMSD values is shown as a boxplot. The black horizontal bar indicate the median, the surrounding box the lower and upper quartile. The dashed lines indicate the smallest and largest values, excluding the outliers which are indicated by circles.

reference structure (interface-RMSD > 8Å) have a significantly worse CS-RMSD score than the best ranked models. Clustering the solutions yields an even better discrimination (see Figure 2a, left panel). Without the quantitative use of chemical shifts, the standard HADDOCK score can, in this particular case, not discriminate between correct and wrong solutions (see Figure 2a, right panel). A possible explanation can be the presence of multiple charged patches on Im9. The ambiguous distance restraints (AIRs) define only the binding site, but not the relative orientation of the binding partners.

The best CS-scored model obtained, using the combined CS-RMSD score of $^{1}H^{\alpha}$, $^{13}C^{\alpha}$ and ^{15}N chemical shifts, is compared to the reference complex in Figure 3a. The predicted interaction site superimposes rather well to the reference complex with an interface-RMSD of 1.6Å and a high fraction of native contacts f_{nat} = 78.6%. We can compare these results with CamDock[26] which has also been applied to model the E9-Im9 complex. The authors obtained a very good interface-RMSD of 0.93Å (C^{α}-RMSD = 1.18Å) using a combination of $^{1}H^{\alpha}$, $^{13}C^{\alpha}$, $^{13}C^{\beta}$ and ^{15}N chemical shifts. However, the performance of CamDock seems to depend highly on the completeness of the chemical shift data, as the use of a reduced set of $^{1}H^{\alpha}$, $^{13}C^{\beta}$ and ^{15}N chemical shifts yielded a much higher C^{α}-RMSD of 6.25Å.[26]

CS-HADDOCK seems thus more robust against missing input data, as the use of only one of the three nuclei $^{1}H^{\alpha}$,$^{13}C^{\alpha}$,^{15}N already gives good results in most cases (Figure 4a-c). While $^{1}H^{\alpha}$ had the best scoring properties for the E9-Im9 case, $^{13}C^{\alpha}$ and ^{15}N also scored the best cluster at the first position. In contrast, $^{13}C^{\beta}$ and especially $^{1}H^{N}$ were not as discriminative as the other nuclei (Figure 4d-e). This can be explained as $^{13}C^{\beta}$ chemical shifts depend mainly on the amino acid type and $^{1}H^{N}$ chemical shifts are in general poorly predicted. The prediction error for the latter is twice as large as for $^{1}H^{\alpha}$ chemical shifts, mainly because of the difficulty to correctly predict hydrogen bonding networks. A similar tendency can be observed for the other three complexes (see supplementary material, Figure S2-S7). A scoring based on only one nucleus will not be perfect in all cases. Therefore, the combined use of the three most appropriate nuclei $^{1}H^{\alpha}$, $^{13}C^{\alpha}$ and ^{15}N gives a robust scoring.

The robustness of CS-HADDOCK against missing experimental CS data is also demonstrated in Figure 5. For all four complexes tested, random removal of up to 60-70% of the experimental CS data did not changed the scoring results. Similar observations were made when restricting the random

removal to only interface residues (results not shown). For proper scoring, refinement of the models seems more important than the completeness of the CS data as scoring of rigid body docking solutions only does not allow to identify the native-like solutions.

The second complex on which the CS-RMSD score was tested is the EIN-HPR complex, which was used in the original HADDOCK publication.[14] The right panel of Figure 2b demonstrates that the HADDOCK score is already sufficient to discriminate the best solution cluster from the others. In this case, the interface on HPR contains a single well-defined charged patch that allows HADDOCK to correctly rank the best models. The CS-RMSD score also ranks the best cluster at the top (Figure 2b, left panel and Figure 3b).

As a third test, CS-HADDOCK was applied to the Z_{Taq} and anti-Z_{Taq} affibodies complex. As for E9-Im9, only the combination of the qualitative use of CSP data and the quantitative use of CS made it possible to score the best cluster at the first position (Figure 2c and 3c). The HADDOCK score alone could not rank the best solutions at the top (Figure 2c, right panel) due to the lack of a clear electrostatic signature on the interface.

For the last complex of our small benchmark, the ILK ARD - PINCH-1 LIM1 complex, the CS-RMSD score ranked again the best cluster at the first position (Figure 2d and 3d). The ILK-PINCH complex is a particularly difficult target for a docking method, as the unbound, free-form structures of the PINCH-1 LIM1 protein have high RMSD values compared to the reference complex structure (see Table III, interface-RMSD = 3.6-4.2Å that would classify it as challenging for docking). It is therefore not surprising that the best cluster of ILK-PINCH models, generated by HADDOCK, has quite high interface-RMSD values (between 2.5-6.2Å). The fraction of native contacts recovered is however quite high (between 0.34 and 0.51 for the top 4 structures, see Table IV), which would qualify it as acceptable to medium quality prediction according to CAPRI criteria[28, 29]. Despite the rather large conformational change between the free structures and the reference complex structure, CS-HADDOCK performed very well in selecting the best cluster from the docking results.

Beside assessing the performance of various combinations of nuclei for the calculation of the n-CS-RMSD score (see Figure S1-S7), we also investigated if a combined CS-RMSD HADDOCK score would perform better. This combined score would measure for a model both its fit to CS data and

its interaction energy as given by the force-field. Analysis of our data reveals that the CS-RMSD and HADDOCK score are almost uncorrelated (data not shown). This may not be surprising, as both scores show a very large range of values inside a structural cluster (see Figure 2), i.e. a small structural rearrangement can change each score quite dramatically, but not necessarily in the same direction. Combining the two scores after normalization of each individual score does not lead to improvement and more work will be needed to optimize the various weights of the scoring function. This would, however, require a much larger benchmark set to be of any significance, something difficult to achieve at this time considering the very limited number of complete entries for complexes in the BMRB.

The robustness of our new CS-HADDOCK protocol, as demonstrated here on four protein-protein complexes, comes from the combined use of CSP and CS data. Without restricting the search space to the binding site (obtained here through the qualitative use of CSP data as ambiguous interface restraints), the quantitative use of chemical shifts does not give a robust scoring function, as remote binding sites may result in smaller CS-RMSD values than for the true binding site. We tested on E9-IM9, whether the CS-RMSD score would be able to pick the best solutions among an ensemble of models that sample the whole 6D interaction space, i.e. the models that were generated by ab-initio docking without any information about the binding site. We used for this the FFT docking program ZDOCK[30]. As FFT "soft-docking" models, like the ones from ZDOCK, may contain steric clashes, all models were subjected to the water-refinement step of HADDOCK. From the resulting 3600 models of E9-IM9, neither the CS-RMSD nor the HADDOCK score were able to select the best solutions, irrespective whether the ZDOCK or water-refined models were considered (see Figure S10). These results indicate that the use of CSP data in CS-HADDOCK to concentrate the initial search around putative interface regions, rather than performing a full search of the interaction space as in ab-initio docking, is beneficial to obtain robust results.

Finally, a few cautionary remarks are in place. First, experimental chemical shifts should be properly referenced prior to running CS-HADDOCK, as badly referenced shifts might degrade the performance of the CS-RMSD score. Several automated solutions exist for this purpose [31–34]. Furthermore, CS-HADDOCK is expected to work best on tightly bound complexes. Caution should to be applied to less tightly bound complexes as the observed chemical shifts might represent an average of

the free and the bound form. Among the four tested complexes one is very tightly bound, E9-IM9: $K_d = 10^{-16}M$ [35], two tightly bound, Z_{Taq} - anti-Z_{Taq}: $K_d = 100nM$ [36], ILK-PINCH: $K_d = 68nM$ [37] and one less tightly bound, EIN-HPR: $K_d = 6.7\mu M$ [38]. Even though EIN-HPR is a less tightly bound complex, CS-HADDOCK scores still the best cluster at the first position, showing that it can already be applied in its current version to this class of complexes. In principle, if the dissociation constant is known, the average chemical shifts might be calculated from the mixture of free and bound forms.

5 Conclusion

We have shown on a small benchmark set that the combination of qualitative and quantitative use of chemical shifts increases the reliability of data-driven docking for the structure determination of complexes from limited NMR data. In particular, the combined use of $^{1}H^{\alpha}$, $^{13}C^{\alpha}$ and ^{15}N chemical shifts gives the best discrimination. Furthermore, robust results are only obtained when restricting the search space to the interaction site, as is done for example by the qualitative introduction of CSP data into AIRs. As, hopefully, the number of entries of biomolecular complexes for which chemical shifts are available in the BMRB database will increase in the future, further optimization of the protocol and scoring function will become possible.

6 Availability

The python script for CS-RMSD calculations is available from the authors upon request. It will be included in a future release of HADDOCK.

7 Acknowledgement

This work was supported financially by The Netherlands Organization for Scientific Research (VICI grant no. 700.56.442 to A.B.) and by the European FP7 e-Infrastructure ‘e-NMR’ I3 project, grant

number 213010.

References

[1] Andrea Cavalli, Xavier Salvatella, Christopher M. Dobson, and Michele Vendruscolo. Protein structure determination from NMR chemical shifts. *Proc Natl Acad Sci USA*, 104(23):9615–9620, 2007.

[2] Yang Shen, Oliver Lange, Frank Delaglio, Paolo Rossi, James M. Aramini, Gaohua Liu, Alexander Eletsky, Yibing Wu, Kiran K. Singarapu, Alexander Lemak, Alexandr Ignatchenko, Cheryl H. Arrowsmith, Thomas Szyperski, Gaetano T. Montelione, David Baker, and Ad Bax. Consistent blind protein structure generation from NMR chemical shift data. *Proc Natl Acad Sci USA*, 105(12):4685–4690, March 2008.

[3] David S. Wishart, David Arndt, Mark Berjanskii, Peter Tang, Jianjun Zhou, and Guohui Lin. CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. *Nucl. Acids Res.*, 36(suppl_2):W496–502, 2008.

[4] Eldon L Ulrich, Hideo Akutsu, Jurgen F Doreleijers, Yoko Harano, Yannis E Ioannidis, Jundong Lin, Miron Livny, Steve Mading, Dimitri Maziuk, Zachary Miller, Eiichi Nakatani, Christopher F Schulte, David E Tolmie, R. Kent Wenger, Hongyang Yao, and John L Markley. BioMagResBank. *Nucleic Acids Res*, 36(Database issue):D402–D408, Jan 2008.

[5] Klara Osapay and David A. Case. A new analysis of proton chemical shifts in proteins. *J Am Chem Soc*, 113(25):9436–9444, December 1991.

[6] M. Iwadate, T. Asakura, and M. P. Williamson. C alpha and C beta carbon-13 chemical shifts in proteins from an empirical database. *J Biomol NMR*, 13(3):199–211, Mar 1999.

[7] Stephen Neal, Alex M. Nip, Haiyan Zhang, and David S. Wishart. Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *J Biomol NMR*, 26(3):215–240, 2003.

[8] Yang Shen and Ad Bax. Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J Biomol NMR*, 38(4):289–302, 2007.

[9] Kai J Kohlhoff, Paul Robustelli, Andrea Cavalli, Xavier Salvatella, and Michele Vendruscolo. Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. *J Am Chem Soc*, 131(39):13894–13895, October 2009.

[10] Juuso Lehtivarjo, Tommi Hassinen, Samuli-Petrus Korhonen, Mikael Peräkylä, and Reino Laatikainen. 4D prediction of protein 1H chemical shifts. *J Biomol NMR*, 45(4):413–426, December 2009.

[11] Erik R P Zuiderweg. Mapping protein-protein interactions in solution by NMR spectroscopy. *Biochemistry*, 41(1):1–7, Jan 2002.

[12] Frank Schumann, Hubert Riepl, Till Maurer, Wolfram Gronwald, Klaus-Peter Neidig, and Hans Kalbitzer. Combined chemical shift changes and amino acid specific chemical shift mapping of protein–protein interactions. *J Biomol NMR*, 39(4):275–289, December 2007.

[13] Mickael Krzeminski, Karine Loth, Rolf Boelens, and Alexandre Bonvin. SAMPLEX: Automatic mapping of perturbed and unperturbed regions of proteins and complexes. *BMC Bioinformatics*, 11(1):51, 2010.

[14] Cyril Dominguez, Rolf Boelens, and Alexandre M. J. J. Bonvin. HADDOCK: A Protein−Protein Docking Approach Based on Biochemical or Biophysical Information. *J Am Chem Soc*, 125(7):1731–1737, February 2003.

[15] Sjoerd J de Vries, Aalt D J van Dijk, Mickaël Krzeminski, Mark van Dijk, Aurelien Thureau, Victor Hsu, Tsjerk Wassenaar, and Alexandre M J J Bonvin. HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins*, 69(4):726–733, December 2007.

[16] Sjoerd J de Vries, Marc van Dijk, and Alexandre M J J Bonvin. The HADDOCK web server for data-driven biomolecular docking. *Nat. Protocols*, 5(5):883–897, May 2010.

[17] M. A. McCoy and D. F. Wyss. Alignment of weakly interacting molecules to protein surfaces using simulations of chemical shift perturbations. *J Biomol NMR*, 18(3):189–198, Nov 2000.

[18] Mark A. McCoy and Daniel F. Wyss. Spatial Localization of Ligand Binding Sites from Electron Current Density Surfaces Calculated from NMR Chemical Shift Perturbations. *J Am Chem Soc*, 124(39):11758–11763, October 2002.

[19] Bing Wang, Edward N. Brothers, Arjan van der Vaart, and Jr. Merz. Fast semiempirical calculations for nuclear magnetic resonance chemical shifts: A divide-and-conquer approach. *The Journal of Chemical Physics*, 120(24):11392–11400, 2004.

[20] Jaime Stark and Robert Powers. Rapid Protein−Ligand Costructures Using Chemical Shift Perturbations. *J Am Chem Soc*, 130(2):535–545, 2008.

[21] Marina Cioffi, Christopher A Hunter, Martin J Packer, and Andrea Spitaleri. Determination of protein-ligand binding modes using complexation-induced changes in (1)h NMR chemical shift. *J Med Chem*, 51(8):2512–2517, Apr 2008.

[22] Marina Cioffi, Christopher A Hunter, Martin J Packer, Maya J Pandya, and Mike P Williamson. Use of quantitative (1)H NMR chemical shift changes for ligand docking into barnase. *J Biomol NMR*, 43(1):11–19, Jan 2009.

[23] Domingo González-Ruiz and Holger Gohlke. Steering protein-ligand docking with quantitative NMR chemical shift perturbations. *J Chem Inf Model*, 49(10):2260–2271, Oct 2009.

[24] Shashank Deep, Sang-Choul Im, Erik R. P. Zuiderweg, and Lucy Waskell. Characterization and Calculation of a Cytochrome c−Cytochrome b5 Complex Using NMR Data†. *Biochemistry*, 44(31):10654–10668, 2005.

[25] Mark A. McCoy and Daniel F. Wyss. Structures of Protein−Protein Complexes Are Docked Using Only NMR Restraints from Residual Dipolar Coupling and Chemical Shift Perturbations. *J Am Chem Soc*, 124(10):2104–2105, March 2002.

[26] Rinaldo W. Montalvao, Andrea Cavalli, Xavier Salvatella, Tom L. Blundell, and Michele Vendruscolo. Structure Determination of Protein−Protein Complexes Using NMR Chemical Shifts:

Case of an Endonuclease Colicin−Immunity Protein Complex. *J Am Chem Soc*, 130(47):15990–15996, November 2008.

[27] Juan Fernandez-Recio, Ruben Abagyan, and Maxim Totrov. Improving CAPRI predictions: optimized desolvation for rigid-body docking. *Proteins*, 60(2):308–313, August 2005.

[28] Raúl Méndez, Raphaël Leplae, Marc F Lensink, and Shoshana J Wodak. Assessment of CAPRI predictions in rounds 3-5 shows progress in docking procedures. *Proteins*, 60(2):150–169, Aug 2005.

[29] Marc F Lensink and Shoshana J Wodak. Docking and scoring protein interactions: CAPRI 2009. *Proteins*, 78(15):3073–3084, Nov 2010.

[30] Rong Chen and Zhiping Weng. Docking unbound proteins using shape complementarity, desolvation, and electrostatics. *Proteins*, 47(3):281–294, May 2002.

[31] Haiyan Zhang, Stephen Neal, and David S. Wishart. RefDB: A database of uniformly referenced protein chemical shifts. *Journal of Biomolecular NMR*, 25(3):173–195, March 2003.

[32] Yunjun Wang and David S. Wishart. A simple method to adjust inconsistently referenced 13C and 15N chemical shift assignments of proteins. *Journal of Biomolecular NMR*, 31(2):143–148, February 2005.

[33] Liya Wang and John L Markley. Empirical correlation between protein backbone 15N and 13C secondary chemical shifts and its application to nitrogen chemical shift re-referencing. *J Biomol NMR*, 44(2):95–99, Jun 2009.

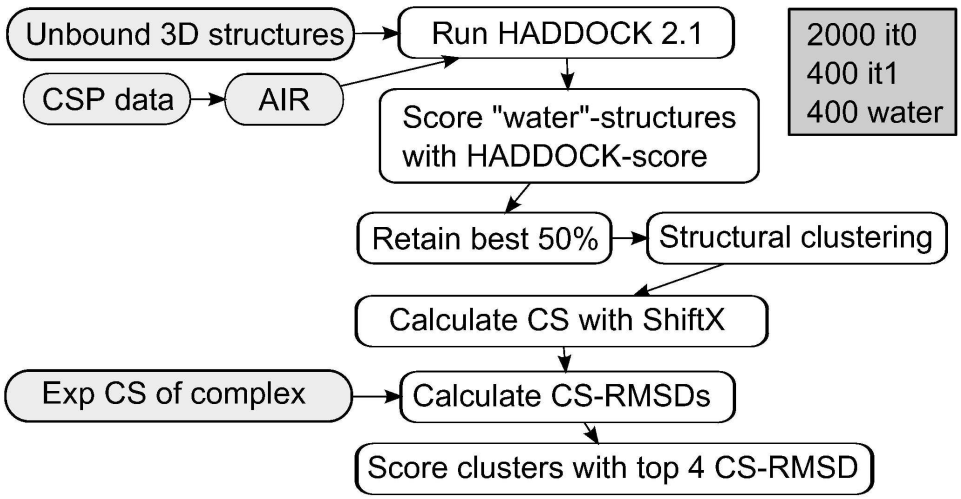
[34] Simon Ginzinger, Marko Skočibušić, and Volker Heun. CheckShift improved: fast chemical shift reference correction with high accuracy. *Journal of Biomolecular NMR*, 44(4):207–211, 2009.

[35] Russell Wallis, Geoffrey R. Moore, Richard James, and Colin Kleanthous. Protein-Protein interactions in colicin e9 DNase-Immunity protein complexes. 1. Diffusion-Controlled association and femtomolar binding for the cognate complex. *Biochemistry*, 34(42):13743–13750, October 1995.

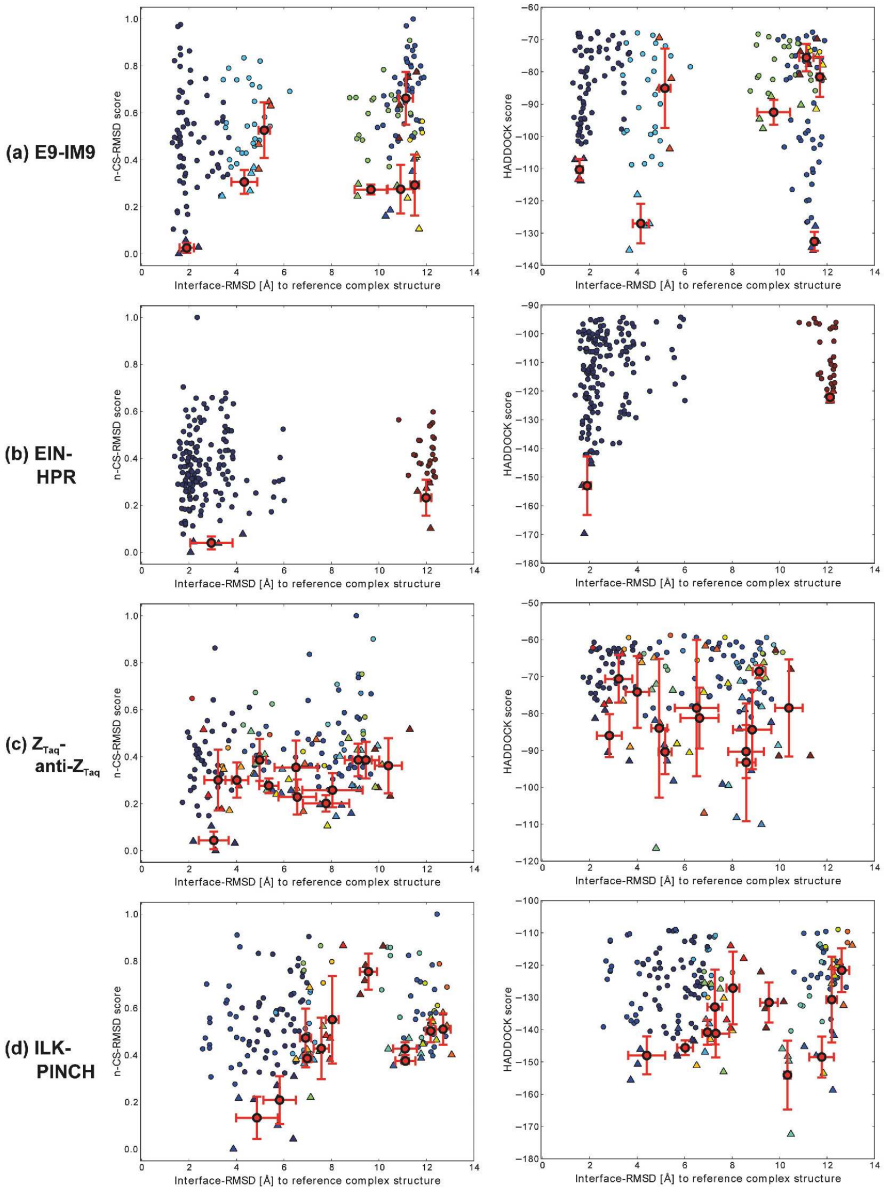
[36] Jakob Dogan, Christofer Lendel, and Torleif Härd. Thermodynamics of folding and binding in an affibody:affibody complex. *Journal of Molecular Biology*, 359(5):1305–1315, June 2006. PMID: 16701696.

[37] Y. Yang, X. Wang, C. A. Hawkins, K. Chen, J. Vaynberg, X. Mao, Y. Tu, X. Zuo, J. Wang, Y. x. Wang, C. Wu, N. Tjandra, and J. Qin. Structural basis of focal adhesion localization of LIM-only adaptor PINCH by integrin-linked kinase. *Journal of Biological Chemistry*, 284(9):5836–5844, 2008.

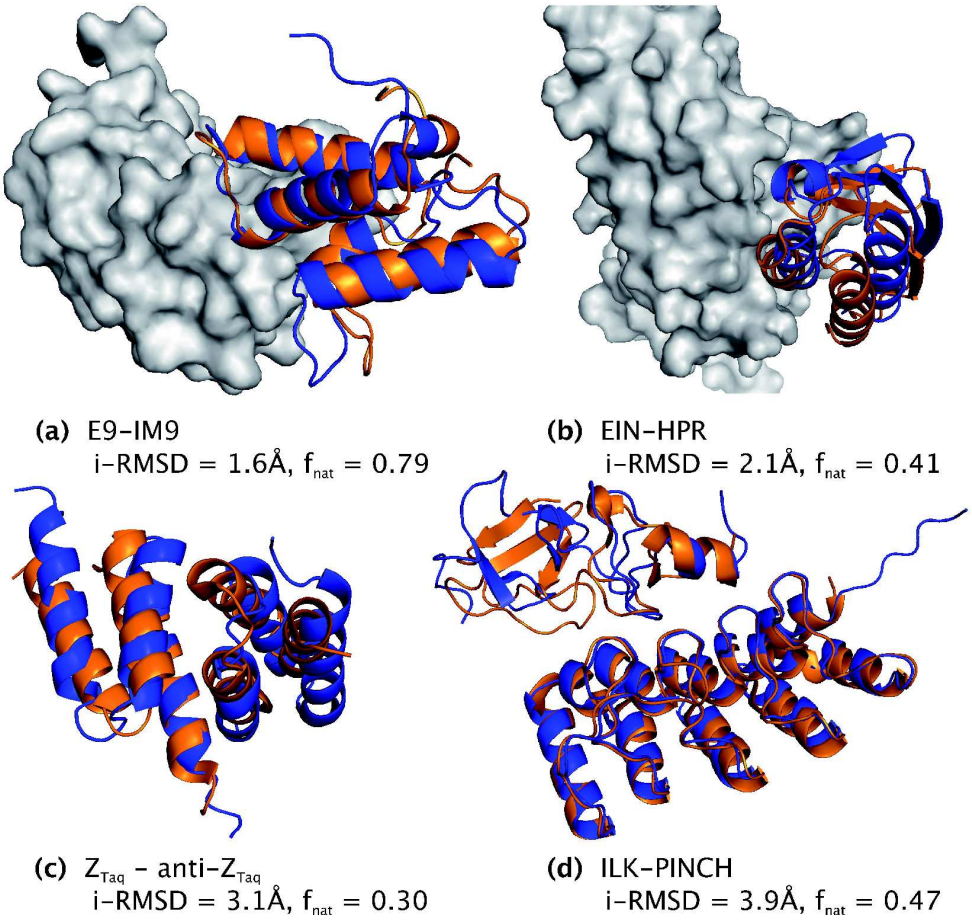
[38] D S Garrett, Y J Seok, A Peterkofsky, G M Clore, and A M Gronenborn. Identification by NMR of the binding surface for the histidine-containing phosphocarrier protein HPr on the n-terminal domain of enzyme i of the escherichia coli phosphotransferase system. *Biochemistry*, 36(15):4393–4398, April 1997. PMID: 9109646.



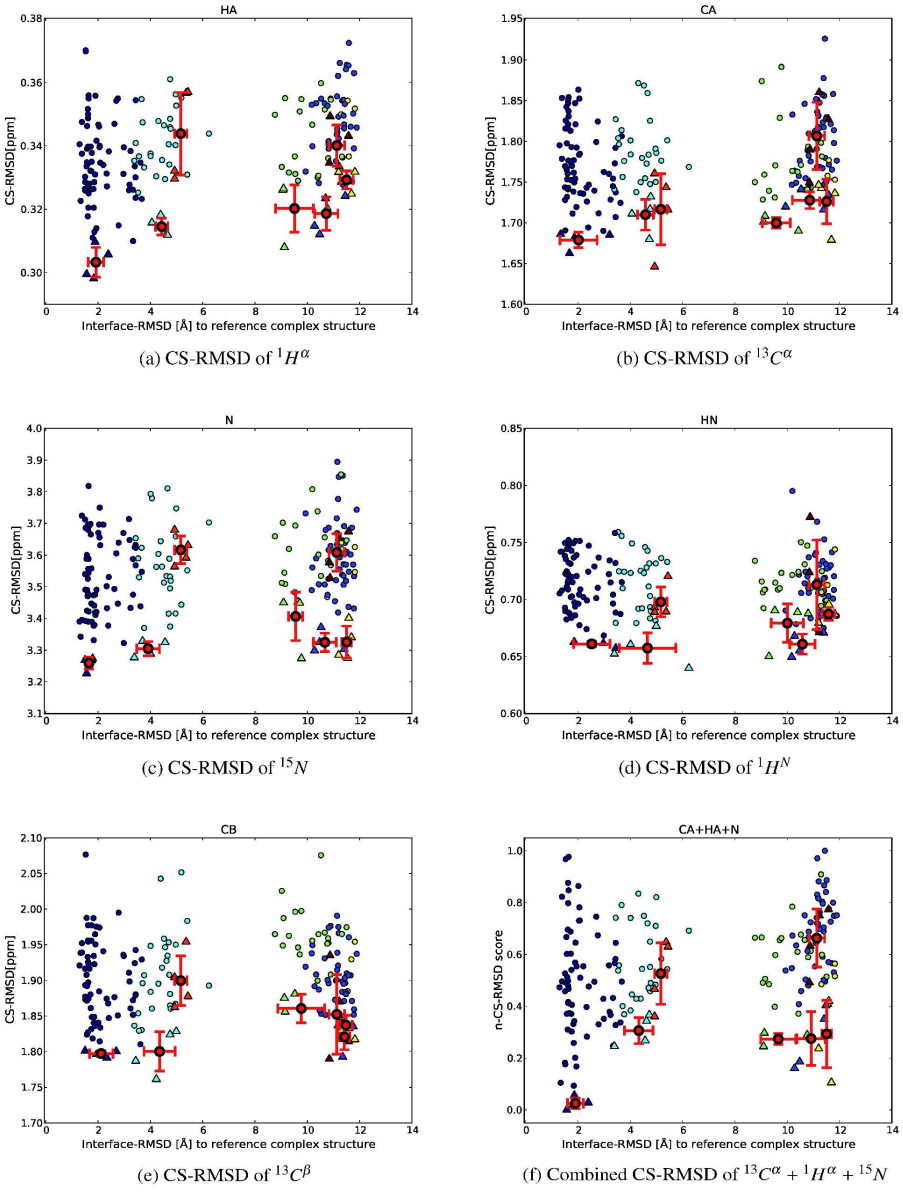
CS-HADDOCK protocol flowchart (see main text for explanations)
181x96mm (600 x 600 DPI)



Comparison of the n-CS-RMSD (left column) and HADDOCK (right col- umn) scores versus interface RMSD from the reference structure of the 200 clustered water-refined models. Each structural cluster has a different color. The triangles indicate the top 4 structures of each cluster. The red crosses indicate the average and the standard deviation of the scores and the interface- RMSDs of the top 4 structures of each cluster. The clusters are ranked according to the average score of the top 4 structures of each cluster. The n-CS-RMSD score is the normalized combined score of the 13Cα ,1 Hα and 15N nuclei (see Material and Methods). (a) E9-IM9, (b) EIN-HPR, (c) Z - anti-Z and (d) ILK-PINCH
215x285mm (600 x 600 DPI)



Comparison of the CS-RMSD best-scored model with the reference structure. The structures were fitted on the interface backbone atoms and the interface-RMSD (i-RMSD) and fraction of native contacts fnat values are given. (a) E9-IM9 model (IM9 in blue) versus reference structure 1EMV (IM9 in gold and E9 as gray surface). (b) EIN-HPR model (HPR in blue) versus reference structure 3EZA (HPR in gold and EIN as gray surface). (c) ZTaq - anti-ZTaq model (in blue) versus reference structure 2B87 (in gold). (d) ILK-PINCH model (in blue) versus reference structure 3F6Q (in gold). 214x199mm (600 x 600 DPI)

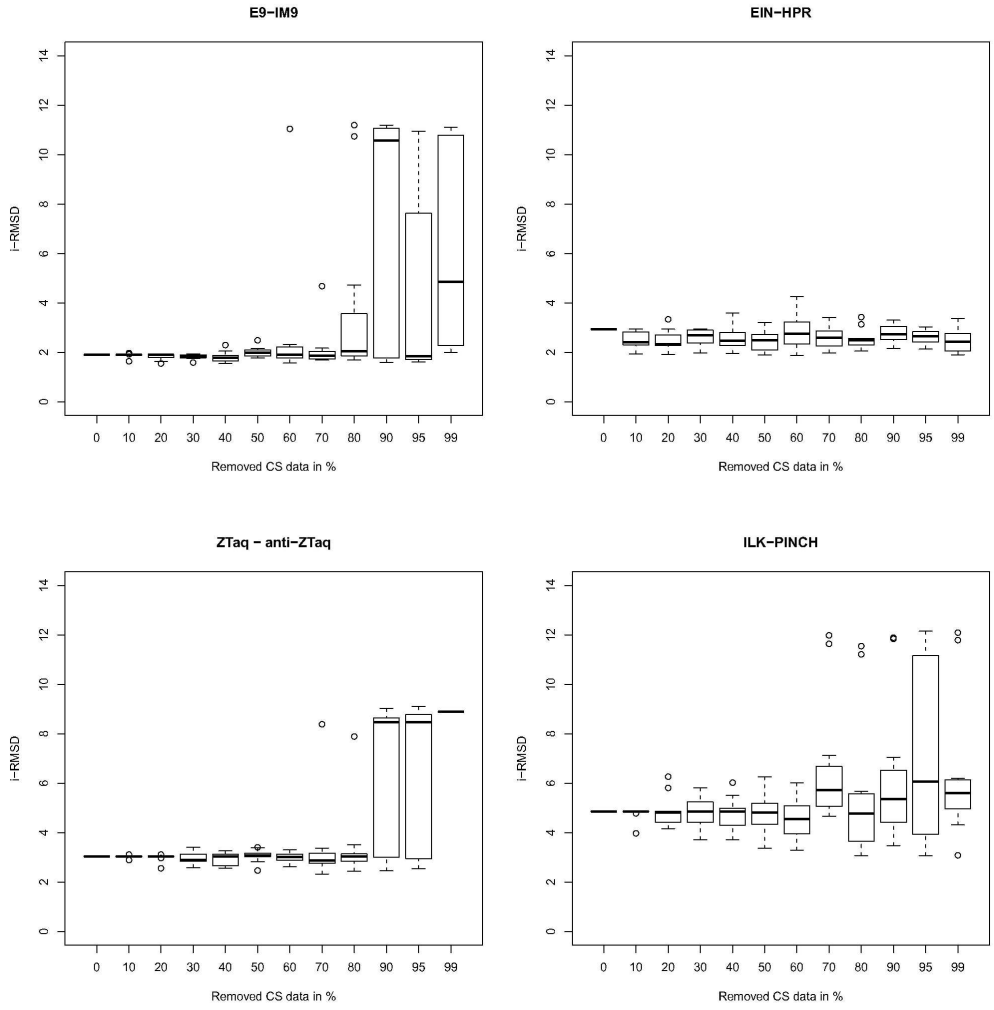


(a) CS-RMSD of $^{1}H^{\alpha}$ (b) CS-RMSD of $^{13}C^{\alpha}$

(c) CS-RMSD of ^{15}N (d) CS-RMSD of $^{1}H^{N}$

(e) CS-RMSD of $^{13}C^{\beta}$ (f) Combined CS-RMSD of $^{13}C^{\alpha} + {}^{1}H^{\alpha} + {}^{15}N$

E9-Im9 complex: (a-e) CS-RMSD scores for single nuclei. (f) Combined n-CS-RMSD score
169x217mm (600 x 600 DPI)



Robustness versus missing CS data: average i-RMSD of the best scored cluster as a function of the fraction of CS data randomly removed. The CS data were randomly removed, separately for each binding partner. This was repeated 50 times. The resulting distribution of i-RMSD values is shown as a boxplot. The black horizontal bar indicate the median, the surrounding box the lower and upper quartile. The dashed lines indicate the smallest and largest values, excluding the outliers which are indicated by circles.
169x173mm (600 x 600 DPI)