Supporting Information Appendix

Section I

Dependence of the accuracy and speed of DFT calculations of the $^{13}C^{\alpha}$ chemical shifts of proteins on the size of the basis set used.

The purpose of this section is to study the dependence of the accuracy and speed of DFT calculations of the $^{13}C^{\alpha}$ chemical shifts on the size of the basis set used. Six basis sets (see Table S1), viz., five *locally-dense* basis-set approximation 6-31G/3-21G, 6-31G(d)/3-21G, 6-311G(d,p)/3-21G, 6-311+G(d,p)/3-21G, and 6-311+G(2d,p)/3-21G, and the uniform 3-21G/3-21G basis set were initially applied to 10 NMR-derived conformations of the 76-residue α/β protein ubiquitin (Protein Data Bank id 1D3Z[1]). For each of these six basis sets, combined with the OB98 functional,[13] the $^{13}C^{\alpha}$ shielding was computed for 760 amino acid residues by treating each amino acid **X** in the sequence as a terminally-blocked tripeptide with the sequence Ac-G**X**G-NMe in the conformation of the regularized experimental protein structure. Analysis of the results (see Table S1), in terms of the agreement between the computed and observed $^{13}C^{\alpha}$ chemical shifts, shows that the accuracy, with which the observed $^{13}C^{\alpha}$ chemical shifts are reproduced by using either the small basis set (6-31G/3-21G) or the larger basis set [6-311+G(2d,p)/3-21G], is very similar, although use of the small basis set leads to a significant decrease in computational time. An additional analysis was carried out here for: (*a*) two other proteins with different numbers of residues and topology

solved by NMR spectroscopy and X-ray diffraction (PDB id 2JVD[2]; 1NS1[3] and 3HBP[4]; 1AIL[5]), and (*b*) Val and Arg hypersurfaces constructed by calculating a grid of 6,864 and 6,794 points, respectively, corresponding to different combinations of the ϕ, ψ, $\chi 1$ (and $\chi 2$ only for Arg) torsional angles. The results of this analysis, reported in Table S2, and Figures S1 provide evidence that the conclusions derived here apply to proteins of any size or class (see section *Transferability of the results*, below). The results also indicate that the $^{13}C^{\alpha}$ chemical shifts computed with the small basis set (6-31G/3-21G), and extrapolated by an empirically-determined linear regression formula to reproduce the values obtained with a larger basis set [6-311+G(2d,p)/3-21G], constitute an adequate compromise between accuracy (within the average error of ~0.4 ppm) and computational cost (~9 times faster) with which to reproduce the observed $^{13}C^{\alpha}$ chemical shifts of proteins in solution.

Method used to compute the $^{13}C^{\alpha}$ chemical shifts. All the experimentally determined conformations were *regularized*,[6] i.e., all residues were replaced by the standard ECEPP/3[7] residues in which bond lengths and bond angles are fixed (rigid-body geometry approximation) at the standard values,[7] and hydrogen atoms are added, if necessary.

The computations of the $^{13}C^{\alpha}$ chemical shifts involve a series of approximations. For each amino acid residue **X** in the protein sequence: (*a*) it is assumed that the observed $^{13}C^{\alpha}$ chemical shift is a conformational-averaged one (see *Computation of the conformationally-averaged rmsd* section); (*b*) computation of the $^{13}C^{\alpha}$ shielding was carried out on a terminally-blocked tripeptide with the sequence Ac-G**X**G-NMe in the conformation of the regularized experimental protein structure; (*c*) computation of the $^{13}C^{\alpha}$ shielding for each residue **X** was carried out with a Γ *locally-dense* basis set approach,[8] with Γ = 6-31G, 6-31G(d), 6-311G(d,p), 6-311+G(d,p) or 6-311+G(2d,p), while the remaining residues in the

tripeptide were *always* treated with the 3-21G basis set. From here on, this combination of uniform and locally-dense basis sets is referred to as: Set_1, Set_2, Set_3, Set_4, Set_5 and Set_6, respectively; (*d*) all ionizable residues were considered neutral during the gas-phase quantum chemical calculations;[9] (*e*) no geometry optimization is necessary since such optimization by *ab-initio* (HF) or DFT methods has only a small effect on the computed chemical shifts;[10] (*f*) the computed $^{13}C^{\alpha}$ shieldings ($\sigma_{subst,th}$) were converted to $^{13}C^{\alpha}$ chemical shifts (δ) by employing the equation $\delta_{th} = \sigma_{ref} - \sigma_{subst,th}$ where the indices denote a theoretical (*th*) computation, the reference substance (*ref*), and the substance of interest (*subst*), i.e., the $^{13}C^{\alpha}$ shielding of a given amino acid residue **X**. The observed shielding value of tetramethylsilane (TMS) in the gas phase[11], namely 188.1 ppm, was adopted as a reference value.

All the computed $^{13}C^{\alpha}$ shielding ($\sigma_{subst,th}$) values were calculated using the gauge-invariant atomic orbital (GIAO) method at the DFT level of theory as implemented in the GAUSSIAN 03 suite of programs.[12] We have used only one exchange-correlation functional in this section, namely OB98, because it was shown that this functional is, among others, one of the most accurate *and* faster one with which to reproduce the observed $^{13}C^{\alpha}$ chemical shift of proteins in solution.[13]

Determination of an effective TMS shielding value. The determination of an *effective* TMS shielding value follows the procedure introduced recently[13] and, hence, only a brief description will be offered here. By adopting the observed TMS value of 188.1 ppm as a reference, it is possible to find the characteristic mean (x_o) and standard deviation (σ) of the Normal (or Gaussian) fit of the frequency of the error distribution for each of the six basis sets. In other words, for each basis set it is feasible to find an 'effective' TMS shielding value

for which the Normal (or Gaussian) fit shows a zero displacement, i.e., an *effective* TMS value that gives $x_o = 0.0$. The following *effective* TMS values were obtained by applying this procedure and are used throughout this work: 195.4 ppm and 184.5 ppm,[13] for the small (Set_2) and larger (Set_6) basis sets, respectively.

Computation of the conformationally-averaged rmsd. Although the methodology has been published in several previous papers[4,6,9,13] we reproduce it here for the reader's convenience. A protein in solution exists as an ensemble of conformations. As a consequence, we can assume that the observed chemical shifts $^{13}C^{\alpha}_{observed,\mu}$ for a given amino acid μ can be interpreted as a conformational average over different rotational states represented by a discrete number of different conformations, all of which satisfied the NMR constraints from which the conformations were derived.[6] Thus, the following quantity can be computed: $^{13}C^{\alpha}_{computed,\mu} = \sum_{i=1}^{\Omega} \lambda_i \ ^{13}C^{\alpha}_{\mu,i,}$ where $^{13}C^{\alpha}_{\mu,i}$ is the computed chemical shift for amino acid μ in conformation i out of Ω protein conformations, and λ_i is the Boltzmann weight factor for conformation i, with the condition $\sum_{i=1}^{\Omega} \lambda_i \equiv 1$. With existing computational resources, it is not feasible to determine λ_i at the quantum chemical level, and, hence, it is assumed that, under conditions of fast conformational averaging, all Boltzmann weight factors contribute equally and, hence, $\lambda_i \equiv 1/\Omega$. Under this assumptions, the computation of the *ca*-rmsd for a protein containing N amino acids residues, is straightforward:[26] $ca\text{-rmsd}^{\alpha} = (1/N) \sum_{\mu=1}^{N} (^{13}C^{\alpha}_{observed,\mu} - <^{13}C^{\alpha}_{computed,\mu}>)^2]^{½}$ with $<^{13}C^{\alpha}_{computed}>_{\mu} = (1/\Omega) \sum_{i=1}^{\Omega} \ ^{13}C^{\alpha}_{\mu,i}$. Naturally, if $\Omega = 1$, *ca*-rmsd ≡ rmsd, as for any single structure.

In addition, for each amino acid μ, we define an error function $\Delta^{\alpha}{}_{\mu} \cong ({}^{13}C^{\alpha}{}_{\text{observed},\mu} - \langle {}^{13}C^{\alpha}{}_{\text{computed}} \rangle_{\mu})$.

Computation of the average CPU time. The average computational time (reported in Table S1) was computed as an average over all 76 residues of conformation 1 out of 10 of 1D3Z:

$$\textit{Averaged-CPU time} = (1/N) \sum_{\mu=1}^{N} \mathrm{T}_{\mu} \ , \ \text{with} \ N = 76$$

where T_{μ} represents the total cpu time (in seconds) for residue μ, as reported by the output file of the GAUSSIAN 03 suite of programs.[12]

Transferability of the results. The current methodology[4,6,9,13] relies on a crucial observation that once the residue conformations are established by their interactions with the rest of the protein, the ${}^{13}C^{\alpha}$ shielding of each residue depends, mainly, on its backbone and its side-chain conformation, with no significant influence of either the amino acid sequence or the position of the given residue in the sequence. This observation allows us to parallelize the ${}^{13}C^{\alpha}$ shielding calculations in proteins and, hence, to make them feasible. In addition, this means that a given set of accurately-determined amino acid residue conformations, representing the accessible conformational space for all the 20 naturally occurring amino acids and showing a good distribution of side-chain conformations, will constitute a reasonable ensemble with which to carry out tests of the current methodology. In other words, the results of such tests will not depend on whether such ensembles of residue conformations belong to a single or many proteins and, therefore, the results should be transferable to proteins of any class or size. For this purpose, three proteins solved by NMR and X-ray were chosen (see Table S2). The analysis of the three proteins includes information for *all* the 20 naturally occurring amino acid residues with their backbone

torsional angles populating the α-helical, β-sheet, turn and extended regions of the Ramachandran map.

Statistical analysis for the six basis sets. For each basis set, we compute the correlation coefficient,[14] *R* (or *Pearson* coefficient), between the average, $<^{13}C^{\alpha}_{computed}>_{\mu}$, chemical shift calculated for the 10 conformations of 1D3Z (as described in *Computation of the conformationally-averaged rmsd* section) and the observed $^{13}C^{\alpha}$ chemical shifts, the standard deviation of the correlation and the average CPU-time (see Table S1). Adopting the *R* value obtained with the largest basis set, i.e., Set_6, as a 'basis set limit result' enables us to conclude that the results obtained with a small basis set, Set_2, appear as the best tradeoff between accuracy and speed of the calculations. In other words, use of the small basis set leads to comparable correlation, in terms of *R*, to that obtained with the larger basis set but at a significantly lower computational cost (3,268/363 ~9 times; see Table S2). Hence, from here on, our work will be focused on a comparison of the results obtained with a small basis set (Set_2), rather than a large (Set_6) basis set.

The previous analysis enabled us to select the smaller basis set that provides similar accuracy as a 'basis set limit', to reproduce the computed shielding rather than chemical shifts, but at a significantly lower CPU time. However, such analysis says nothing about the accuracy of the $^{13}C^{\alpha}$ chemical shifts, not the shielding, obtained with the Set_2. The answer to this important question is provided in the next section.

Comparison of the results obtained with basis Set_2 and Set_6. An analysis of the correlation of the results obtained from Set_2 and Set_6 was carried out for 3 proteins, namely, 1D3Z (10 conformations), 2JVD (20 conformations), 1NS1 (32 conformations) and

for 6,864 and 6,794 points corresponding to different combinations of the ϕ, ψ, $\chi1$, and $\chi2$ torsional angles for Arg and Val, respectively. The results of the analysis are shown in Figure S1a-e. For all the molecules shown in Figure S1, the slopes of the linear regression are very similar and, even more importantly, very close to the ideal value of 1.0. The largest difference appears on the y intercept ($-2.23 < y < -0.620$) of the linear regression. Nevertheless, the $^{13}C^{\alpha}$ chemical shifts computed with Set_6 can be obtained from Set_2 by using the following linear regression: $^{13}C^{\alpha} = -1.597 + 1.040 \times {}^{13}C^{\alpha}_{\mu}$, where $^{13}C^{\alpha}_{\mu}$ represents the $^{13}C^{\alpha}$ chemical shifts computed for a given residue μ with the Set_2 and, -1.597 and 1.040 representing the averaged values over the five linear regressions, namely from each panel of Figure S1, for both the y intercept and the slope of the regression, respectively.

The main goal of this section I is to determine the basis set size that would enable us to compute the $^{13}C^{\alpha}$ chemical shifts in proteins accurately and fast. Once the small basis set has been chosen, namely Set_2, the next step is to determine how accurately the extrapolated values of the $^{13}C^{\alpha}$ chemical shifts computed with such basis set reproduce the 'basis set limit' results, i.e. the results obtained with the larger basis set (Set_6). The results of such an analysis are shown in Table S2. The accuracy of the results is evaluated here in terms of the *ca*-rmsd. As can be seen from Table S2, the quality of the protein structures in terms of these scoring parameters, computed by using either Set_2 or Set_6, is comparable. In other words, extrapolating the $^{13}C^{\alpha}$ chemical shifts computed with Set_2 enables us to reproduce the results obtained with the more computationally expensive basis set (Set_6) with high accuracy.

In Table S2, we also list the average error (Δ) obtained between the $^{13}C^{\alpha}$ chemical shifts computed with the small basis set (after extrapolation by using the linear

relationship: $^{13}C^{\alpha} = -1.597 + 1.040 \times {}^{13}C_{\mu}^{\alpha}$) and the values obtained with the large basis set. Notably, the average error (Δ) among all the proteins listed in Table S2 is quite low, namely ~0.4 ppm.

The results of this Section I indicate that the $^{13}C^{\alpha}$ chemical shifts in proteins, computed at the DFT level of theory with the large (Set_6) basis set, can be reproduced accurately (within an average error of ~0.4 ppm; see Table S2) and ~9 times faster by using the small (Set_2) basis set with an *effective* TMS value of 195.4 ppm and extrapolating it with: $^{13}C^{\alpha} = -1.597 + 1.040 \times {}^{13}C_{\mu}^{\alpha}$.

Table S1

Test of Six Basis Sets[a]

Basis sets [b]	**Correlation Coefficient** R [c]	**Average CPU time** [d] **(sec)**
Set_1: 3-21G/3-21G	0.892 (2.13)	201
Set_2: 6-31G/3-21G	**0.903 (2.04)**	**230 (363; 118)**
Set_3: 6-31G(d)/3-21G	0.894 (2.12)	338
Set_4: 6-311G(d,p)/3-21G	0.900 (2.06)	568
Set_5: 6-311+G(d,p)/3-21G	0.903 (2.03)	1,092
Set_6: 6-311+G(2d,p)/3-21G	**0.908 (1.97)**	**1,535 (3,268; 372)**

a) The entire test was carried out with 10 conformations of the protein Ubiquitin (PDB id 1D3Z).[1] Two basis sets, selected from the results of column 2 and 3 for further test (see results in Table S2), are represented in boldface.

b) As described in the *Method used to compute the* $^{13}C^{\alpha}$ *chemical shifts* section.

c) The correlation coefficient,[14] R (or *Pearson* coefficient), between the average chemical shifts, $<^{13}C^{\alpha}_{computed}>_{\mu}$ computed from the 10 conformations of 1D3Z, and the

observed $^{13}C^{\alpha}$ chemical shifts. The standard deviation of the correlation is in parenthesis.

d) As an average over 76 residues computed from model 1 out of 10 models of 1D3Z,[1] as explained in the *Computation of the average CPU time* section. The maximum and minimum CPU times of the two selected basis set are shown in parentheses in boldface.

Table S2

Test on Proteins Structures with Two Selected Basis Sets[a]

Protein[b]		Basis sets		
		Set_6	Set_2	
		ca-rmsd[c] (ppm)	*ca*-rmsd[d] (ppm)	Δ[e] (ppm)
Ubiquitin (6457)	1UBQ (X-ray) [76] {1;*i*}	2.60	2.57	0.63±0.40
YnzC protein (15476)	2JVD (NMR) [46] {20; *ii*}	1.64	1.59	0.36±0.22
	3HBP (X-ray) [52] {1;*iii*}	2.51	2.55	0.34±0.22
		1.86	1.85	0.37±0.20
		1.88	2.00	0.40±0.21
Non-structural Protein 1 (15117)	1NS1 (NMR) [73] {32; *iv*}	2.48	2.47	0.26±0.17
	1AIL (X-ray) [70] {1; *v*}	2.07	2.09	0.34±0.29

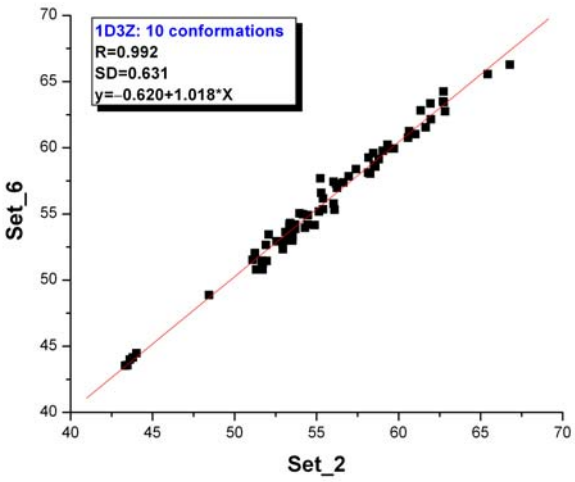
a) Carried out for the two highlighted basis sets in Table S1, namely Set_6 and Set_2.

b) First column list the set of proteins used and, in parenthesis, the BMRB[24] accession number under which the observed $^{13}C^{\alpha}$ Chemical shifts can be found. Second column, in parentheses, the experimental method used; in brackets, the numbers of residues for each protein; and, in braces, the number of conformations and the reference for the protein, namely {*i*} Vijay-Kumar *et al.*[17]; {*ii*} Aramini *et al.*[2]; {*iii*} Vila *et al.*[4]; {*iv*} Chien *et al.*[3]; {*v*} Liu *et al.*[5]

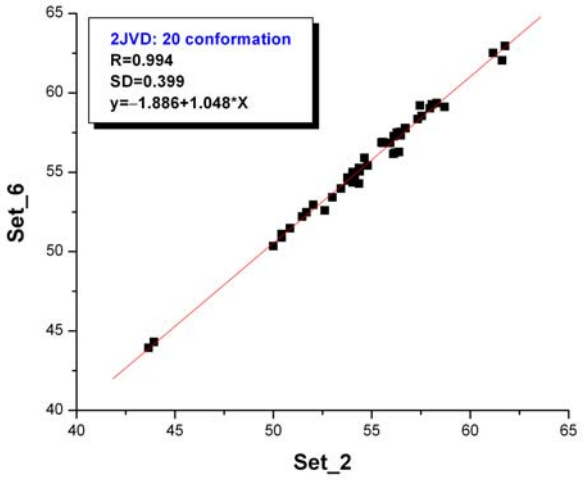
c) The *ca*-rmsd, computed as described in *Computation of the conformationally-averaged rmsd* section, with Set_6 using an *effective* TMS value of 184.5 ppm. Note that the *ca*-rmsd ≡ rmsd for the single X-ray conformations.

d) Same as item (c) but with *all* the $^{13}C^{\alpha}$ chemical shifts given by: $^{13}C^{\alpha} = -1.597 + 1.040 \times {}^{13}C^{\alpha}_{\mu}$ where $^{13}C^{\alpha}_{\mu}$ represent the $^{13}C^{\alpha}$ chemical shifts computed for the residue μ with a small (Set_2) basis set, and using an *effective* TMS value of 195.4 ppm.

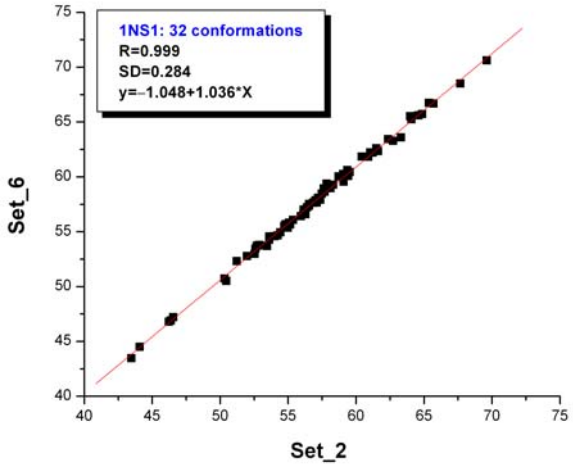
e) The value of the absolute averaged error per-residue, Δ, between the $^{13}C^{\alpha}$ chemical shifts computed with the small basis set (after extrapolation by using the linear relationship) and the values obtained with the large basis set.


1D3Z: 10 conformations
R=0.992
SD=0.631
y=−0.620+1.018*X
Set_6
Set_2

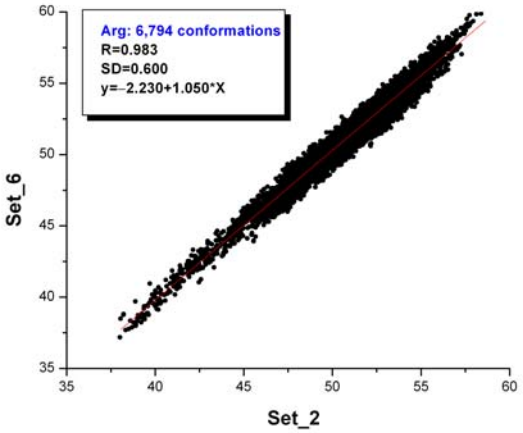
(a)


2JVD: 20 conformation
R=0.994
SD=0.399
y=−1.886+1.048*X
Set_6
Set_2

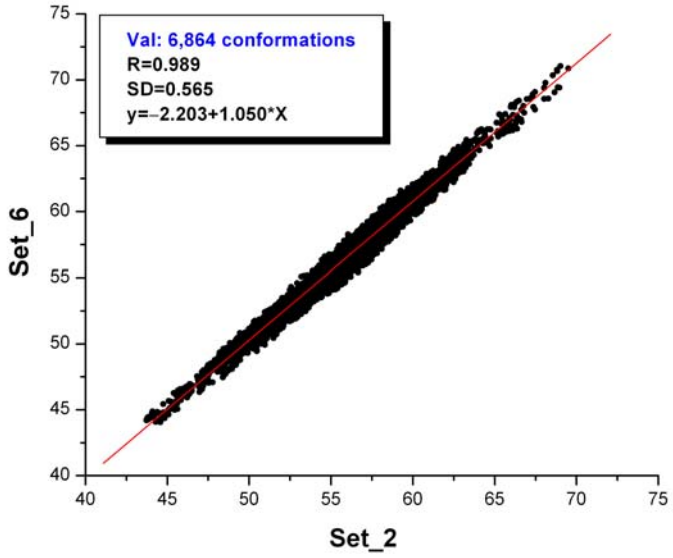
(b)


1NS1: 32 conformations
R=0.999
SD=0.284
y=−1.048+1.036*X
Set_6
Set_2

(c)


Arg: 6,794 conformations
R=0.983
SD=0.600
y=−2.230+1.050*X
Set_6
Set_2

(d)



(e)

Figure S1. **(a)** Correlation between average chemical shifts, $<^{13}C^{\alpha}_{computed}>_{\mu}$, computed from the 10 conformations of 1D3Z[1] with Set_6 versus Set_2. The red line represents the linear regression. Values for the correlation coefficient[14] (*R*), the standard deviation from the linear regression (SD) and the slope and the *y*-intercepts of the linear regression are inserted in the panel; **(b)** same as (**a**) for 20 conformations of 2JVD[2]; **(c)** same as (**a**) for 32 conformations of 1NS1[3]; **(d)** same as (**a**) for 6,864 points of Arg obtained by sampling the ϕ, ψ, $\chi 1$, and $\chi 2$ Ramachandran space; and **(e)** same as (**d**) for 6,794 points for Val computed obtained by sampling the ϕ, ψ, $\chi 1$ Ramachandran space.

Section II

Analysis of X-ray and X-ray-NMR pairs of structures

In this section, we provide information about the quality of the prediction, in terms of the correlation coefficient[14] (*R*), between observed and predicted $^{13}C^{\alpha}$ chemical shifts from different databases, for a set of X-ray derived structures (listed in Table S3), and for a set of pairs of X-ray and NMR-determined structures (listed in Table S4). The analysis of the set listed in Table S3 was carried out with five databases, namely SHIFTS,[18,19] SHIFTX,[20] PROSHIFT,[21] SPARTA[22] and *Che*Shift. PROSHIFT predictions were not carried out for NMR-derived ensembles, because this database web server provides predictions for *only* one structure at a time, making the analysis of a large number of structures very tedious, e.g., as for PDB id 1XQQ[23] containing 128 conformers.

Table S3

Set of X-ray Structures and Corresponding $^{13}C^{\alpha}$ Chemical Shifts

PDB id[a] [Resolution (Å); BMRB accession]		SERVERS[b]				
		*Che*Shift	SHIFTX	SPARTA	SHIFTS	PROSHIFT
1A6K [1.10; 4061]		0.94	0.97	0.97	0.96	0.97
1BKF [1.60; 4077]		0.93	0.98	0.99	0.96	0.96
1CEX [1.00; 4101]		0.95	0.98	0.99	0.98	0.96
1CLL [1.70; 547]		0.91	0.97	0.99	0.97	0.98
1DMB [1.80; 4354]		0.93	0.98	1.00	0.96	0.97
1RGE[c] [1.15; 4259]		0.95	0.98	0.99	0.97	0.97
		0.96	**1.00**	**1.00**	**0.98**	**1.00**
1HFC [1.56; 4064]		0.94	0.97	0.99	0.95	0.97
1HKA [1.50; 4299]		0.94	0.97	0.99	0.95	0.97
1ONC [1.7 ;4371]		0.90	0.96	0.99	0.95	0.95
1HCB [1.60; 4022]		0.92	0.97	0.98	0.96	0.96
1RUV [1.30; 4031]		0.91	0.96	0.99	0.95	0.95
1TOP [1.78; 4401]		0.95	0.97	(0.97	0.95	0.96
3LZT [0.92; 4562]		0.94	0.96	0.97	0.95	0.96
4FGF[d] [1.60; 4091]		0.93 / 0.95	0.98	0.99	0.98	0.97
Interleukin 1β[e] (human)	4I1B [2.00;1061]	**0.87**	0.95	0.97	0.95	0.96
	2I1B [2.00;1061]	0.91	0.96	0.98	0.94	0.96
5PTI [1.00; 46]		0.95	0.97	1.00	0.98	0.97

PDB id[a] [Resolution (Å); BMRB accession]	SERVERS[b]				
	*Che*Shift	SHIFTX	SPARTA	SHIFTS	PROSHIFT
1AIL [1.90; 4317]	0.96	0.99	1.00	0.98	0.98
1BSY [2.24; 10010]	0.90	0.95	0.97	0.93	0.95
1BV1 [2.00; 4417]	0.94	0.97	0.98	0.97	0.97
1CHN [1.76; 4083]	0.95	0.96	0.97	0.95	0.96
1EKG [1.80; 4342]	0.94	0.97	1.00	0.96	0.96
1FIL [2.00; 4082]	0.92	0.97	0.99	0.97	0.97
1GSV [1.75; 6321]	0.94	0.98	0.98	0.96	0.97
1HB6 [2.00; 5351]	0.92	0.95	0.95	0.94	0.96
1HOE [2.00; 1642]	0.94	0.97	0.97	0.95	0.97
1I27 [1.02; 5685]	0.92	0.98	0.98	0.96	0.97
1JF4 [1.40;4083]	0.95	0.97	0.97	0.97	0.98
1RBV [1.80; 4012]	0.94	0.98	0.98	0.97	0.98
1RNB [1.90; 7126]	0.93	0.96	0.98	0.96	0.95
1UBI [1.80; 5387]	0.91	0.97	0.99	0.97	0.97
1ZON [2.00; 4553]	0.92	0.96	0.98	0.95	0.96
2CI2 [2.00; 4974]	0.92	0.98	0.99	0.96	0.97
2CPL [1.63; 2208]	0.97	0.98	1.00	0.97	0.97
2OVO [1.50; 5473]	0.93	0.97	0.97	0.97	0.97
3ICB [2.30; 6699]	0.93	0.97	0.98	0.96	0.97

a) PDB ID code identifying the X-ray structure with which the correlations, *R*, between observed and predicted $^{13}C^{\alpha}$ chemical shifts (otherwise noted) were used for each of the servers. In brackets the resolution at which the protein was solved, and the BMRB[24] accession number under which the observed $^{13}C^{\alpha}$ Chemical shifts can be found. Details for each of the mentioned structures can be found in the cited manuscript in the PDB web site.

b) Server name used for automatic prediction of the $^{13}C^{\alpha}$ chemical shifts. The agreement is analyzed here in terms of the correlation coefficient *R*. The *R* values are reported with only two digits to facilitate a fast comparison of the results among servers. All the *R* values from the servers *Che*Shift, SHIFTX, SPARTA and SHIFTS include all residues, except the first and last one for *Che*Shift, SPARTA and SHIFTS. All correlations computed with SHIFTS do not include predictions for the first and last one as well as for cysteines. If the inclusion/exclusion of cysteines leads to a significant difference in the *R* value computed with the *Che*Shift server, as with protein 4FGF, two correlation coefficients are reported (see footnote *d*, below). All *R* values lower than an arbitrary selected cut off value of 0.90 are highlighted in red because they indicate that more than 20% of the observed $^{13}C^{\alpha}$ chemical shifts cannot be explained by a given protein model and, hence, further analysis may be required.

c) See discussion about the results for this protein in section: *Protein 1RGE (Ribonuclease Sa)*, of the main text. First row shows the correlation coefficient, *R*, between observed and predicted $^{13}C^{\alpha}$ chemical shifts. Second row shows, in bold

face and green color, the correlation coefficient, *R*, between $^{13}C^{\alpha}$ chemical shift predictions for molecule A and B, respectively, of 1RGE, for each of the servers.

d) There are two values of *R* listed for prediction by the *Che*Shift server. In the first one, all residues were included and, in the second one, *all* cysteines were omitted from the calculations of *R* (as with the SHIFTS calculations).

e) The first and second rows of this entry list the *R* values, obtained from each server, for protein 4I1B and 2I1B, respectively. For a discussion of the results obtained for this protein, see section: *Protein interleukin 1β (human)*, of the main text.

Table S4

Pairs of X-ray Structures and sets of NMR Structures

PDB id[a]	BMRB[24] accession	SERVERS[b]			
		*Che*Shift	SHIFTX	SPARTA	SHIFTS
1A6K 1MYF(12)	4061	0.94 0.95	0.97 0.96	0.97 0.96	0.96 0.96
1BKF 1FKR(20)	4077	0.93 0.95	0.98 0.97	0.99 0.97	0.96 0.96
1CLL[c] 2BBN(21)	1634	0.92 0.94	0.97 0.98	0.98 0.98	0.96 0.97
1DMB 1EZP(10)	4354	0.93 0.91	0.98 0.96	1.00 0.97	0.96 0.95
1RGE 1C54(20)	4259	0.95 **0.89**	0.98 0.95	0.99 0.95	0.97 0.94
1HFC 4AYK(30)	4064	0.94 0.91	0.97 0.96	0.99 0.98	0.95 0.95
1HKA 1EQ0(20)	4299	0.94 0.91	0.97 0.96	0.99 0.97	0.95 0.95
1ONC 1PU3(20)	4371	0.90 **0.88**	0.96 0.95	0.99 0.97	0.95 0.95
1RUV 2AAS(32)[d]	4031	0.91 / 0.92 **0.89** / 0.91	0.96 / 0.96 0.94 / 0.96	0.99 / 0.99 0.96 / 0.97	0.95 0.94
1TOP 1SKT(40)	4401	0.95 0.92	0.97 0.97	0.97 0.97	0.96 0.96
3LZT **1E8L(50)**[e]	4562	0.94 / 0.95 **0.89** / 0.91	0.96 / 0.97 0.95 / 0.97	0.97 / 0.98 0.96 / 0.97	0.95 0.95
4I1B / 2I1B 7I1B (32)[f]	1061	**0.87** / 0.91 0.90	0.95 / 0.96 0.96	0.97 / 0.98 0.97	0.95 / 0.94 0.95
5PTI 1UUA(20)	46	0.95 0.93	0.97 0.96	1.00 0.97	0.98 0.96
2CI2 3CI2 (20)	4974	0.92 0.93	0.98 0.97	0.99 0.98	0.96 0.96
1UBQ[g] **1XQQ (128)**	6457	0.91 0.95	0.98 0.98	0.99 0.98	0.97 0.98
2B95(20)[h] 1TGQ (40)	6210	0.93 **0.87**	0.95 0.91	0.97 0.95	0.97 0.95

a) PDB ID code identifying the X-ray- and NMR-derived conformations, in the first and second lines, respectively, for each protein. The total number of structures in the NMR ensemble for which the conformational-average was computed for each database is given in parentheses.

b) Server name used for automatic prediction of the $^{13}C^{\alpha}$ chemical shifts. The agreement is analyzed here in terms of the correlation coefficient, *R*,[14] between observed and predicted $^{13}C^{\alpha}$ chemical shifts. As with Table S3, only two significant digits are reported for each R value to facilitate the comparison among the servers. All *R* values lower than an arbitrary selected cut off, namely 0.90, are highlighted in red because they indicate that more than 20% of the observed $^{13}C^{\alpha}$ chemical shifts cannot be explained by a given protein model and, hence, further analysis may be required.

c) The observed $^{13}C^{\alpha}$ chemical shifts for the protein Calmodulin (BMRB[24] accession 1634) were obtained by NMR for Calmodulin complexed with a 26-residue synthetic peptide, while the X-ray structure was obtained for a peptide-free Calmodulin at 1.7 Å resolution. There are conformational differences between the X-ray (1CLL) and NMR-derived structures (2BBN) along the whole sequence but the larger differences occur for the long central helix (residues 65-93, in the X-ray structure) which is disrupted into two helices connected by a long flexible loop in the NMR-determined conformations (2BBN).

d) The solution NMR-derived conformations for RNase A (2AAS) and the observed $^{13}C^{\alpha}$ chemical shifts (BRMB accession 4031) are for the wild-type protein. However, the X-ray structure (1RUV) was solved at 1.30 Å resolution for the

ribonuclease A-Uridine Vandate (UV) complex. Analysis of the correlation between the observed $^{13}C^{\alpha}$ chemical shifts (from the wild-type protein) and the phosphate-free ribonuclease A (7RNS) solved at 1.26 Å, shows, not surprisingly, a slightly better correlation coefficient than 1RUV, namely R = 0.92. The close agreement, in terms of R, between 1RUV (0.91) and 7RNS (0.92) indicates that both structures are very similar. In fact, the overall rmsd between the C^{α} positions of these two structures is only 0.2 Å, indicating no major conformational change upon UV binding.[25] In other words, the X-ray structure for Ribonuclease A solved with (1RUV) or without (7RNS) ligand (0.91 and 0.92, respectively) are better representations of the observed $^{13}C^{\alpha}$ chemical shifts in solution than the NMR-derived ensemble (2AAS, R = 0.89). For some residues of the ensemble of conformations of 2AAS the predictions with *Che*Shift fail because those residues are in high energy regions of the Ramachandran map, for which the DFT method fails to converge, e.g., Asn34 of conformation N^{o} 12 shows a backbone $\phi = -10.9^{o}$ and $\psi = 20.6^{o}$. In such cases, the whole conformation was removed from the analysis using *Che*Shift. Despite this, all of the other servers, namely SHIFTS, SHIFTX or SPARTA, do predict $^{13}C^{\alpha}$ chemical shifts for residues populating high-energy regions of the Ramachandran map. The second value in each row, for the X-ray and NMR-derived conformations, denotes the R value without 8 cysteines in the sequence (for a straightforward comparison with the SHIFTS predictions which do not include values for cysteines).

e) For a discussion of the results, see section: *A comparative validation analysis of proteins 1E8L and 3LZT*, of the main text. For each server, except SHIFTS, there

are two *R* values for both the X-ray and the NMR-derived models. The first one was computed by using all residues in the sequence, and the second one without the cysteines (for a straightforward comparison with the SHIFTS predictions which do not include values for cysteines).

f) The two values in the first row are the *R* values obtained for the X-ray structures of 4I1B and 2I1B, respectively.

g) For a discussion of the results see section: *A comparative validation analysis of proteins 1UBQ and 1XQQ*, of the main text.

h) There is no X-ray-derived protein structure here, i.e., both ensembles of conformations were obtained by NMR-spectroscopy, and the main differences between these two sets of conformations and the relevance of their analysis was recently discussed.[26] The reported *R* values belong to a segment of 27 residues, from Asp 45 to Asp 71 of protein 1TGQ (now obsolete) and the corresponding segment of protein 2B95. In very good agreement with previous calculations, using the 'internal standard reference',[26] only the results of *Che*Shift, among all the servers, indicates that careful attention should be paid to the fold of this segment in the protein 1TGQ. In other words, this segment of protein 2B95 (R = 0.93) is a significantly better representation of the observed $^{13}C^{\alpha}$ chemical shifts in solution than 1TGQ (R = 0.87).

Section III

Approximations used to interpolate computed $^{13}C^{\alpha}$ chemical shift values

A Gaussian[27] and a linear interpolation function, described in detail for a one-dimensional case by Eq. (1) and (2) below, were used to reproduce the DFT results obtained on a fine grid (see section *Approximations used to interpolate computed $^{13}C^{\alpha}$ chemical shift values*, in the main text). The fill-red circles in each panel of Figure S5 correspond to the results computed using the Gaussian [panels (**a**)-(**b**)] and linear [panels (**c**)-(**d**)] interpolations, respectively. Generalization of these equations to higher dimension of the torsional angle space (i.e. for ϕ, ψ, $\chi1$, and $\chi2$), as used here, is straightforward. The frequencies of the error distribution obtained using three-dimensional interpolations are shown in Figure S6.

I) Gaussian[27] interpolation

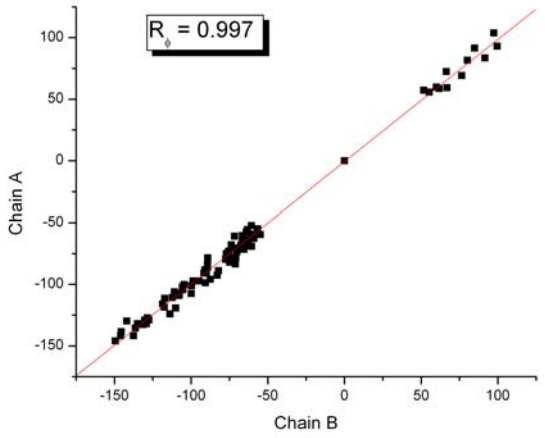
$$ {}^{13}C^{\alpha}_{computed}(\chi_1^n)\,|_{\phi,\psi,\chi^2} = \frac{\sum_{k=1}^{2} {}^{13}C_k^{\alpha}\,\exp[-(\chi_1^n - \chi_{1,k}^o)^2 / S]}{\sum_{k=1}^{2}\exp[-(\chi_1^n - \chi_{1,k}^o)^2 / S]} \qquad (1) $$

with n = 1, …,16; $\chi_1^1 = \chi_{1,1}^0$ and $\chi_1^{16} = \chi_{1,2}^0$; and S a Gaussian scale factor.

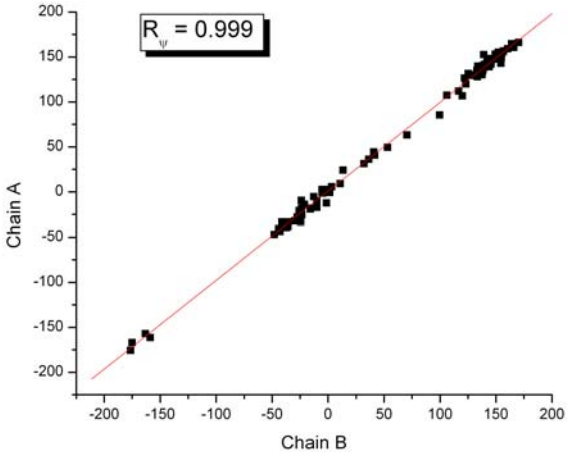
II) Linear interpolation

$$ {}^{13}C^{\alpha}_{computed}(\chi_1^n)\,|_{\phi,\psi,\chi^2} = \sum_{k=1}^{2} {}^{13}C_k^{\alpha}\left[\frac{(\chi_{1,k+1}^o - \chi_1^n)}{\Delta\chi_1}\right]^{2-k}\left[\frac{(\chi_1^n - \chi_{1,k-1}^o)}{\Delta\chi_1}\right]^{k-1} \textbf{ with } n = 1,....,16; \qquad (2) $$

$(\chi_{1,2}^o - \chi_1^n)$ and $(\chi_1^n - \chi_{1,1}^o)$ equal to $\Delta\chi_1$ (= 30°) for n = 1 and 16, respectively.

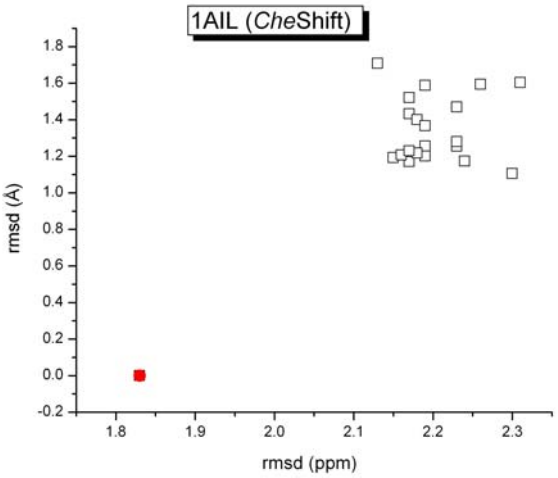


(**a**)

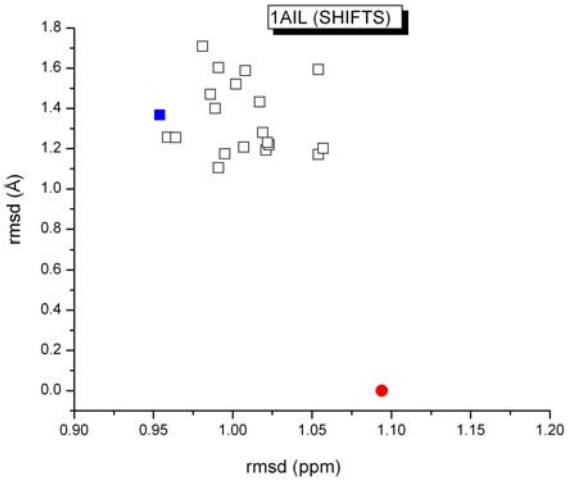


(**b**)

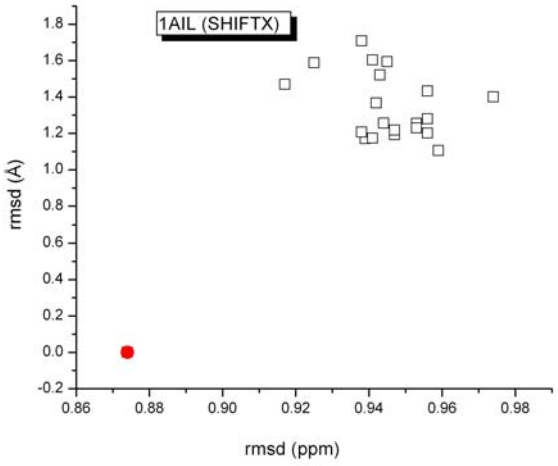
Figure S2. Correlation between the identical backbone torsional angles derived from the molecules A and B of protein PDB id 1RGE, from Ribonuclease Sa. Panel (**a**) for the ϕ torsional angle and (**b**) for the ψ torsional angle.



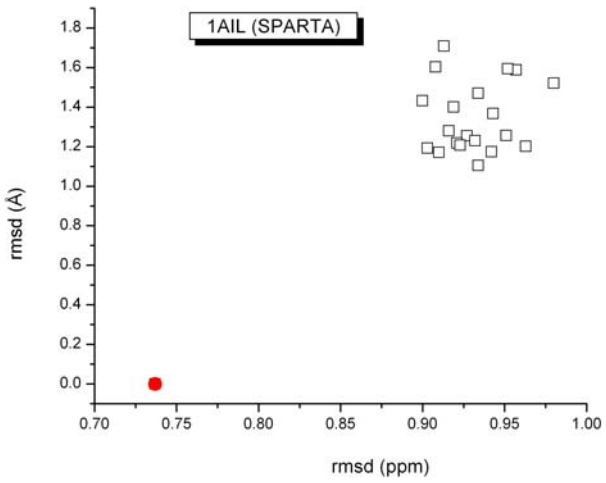
(a)



(b)



(c)



(d)

Figure S3. Plot showing rmsd of the protein decoys (open squares), lower than 3 Å from the 'native' structure (PDB id 1AIL), versus the rmsd (ppm) between observed and predicted $^{13}C^{\alpha}$ chemical shifts, computed with four different servers, namely (**a**) *Che*Shift; (**b**) SHIFTS;

(**c**) SHIFTX; and (**d**) SPARTA. In each panel, the red-filled circle denotes the 'native' structure (1AIL) from which the decoys were obtained; if one decoy shows a lower rmsd (ppm) than that of the 'native' structure, it is denoted by a blue-filled square.

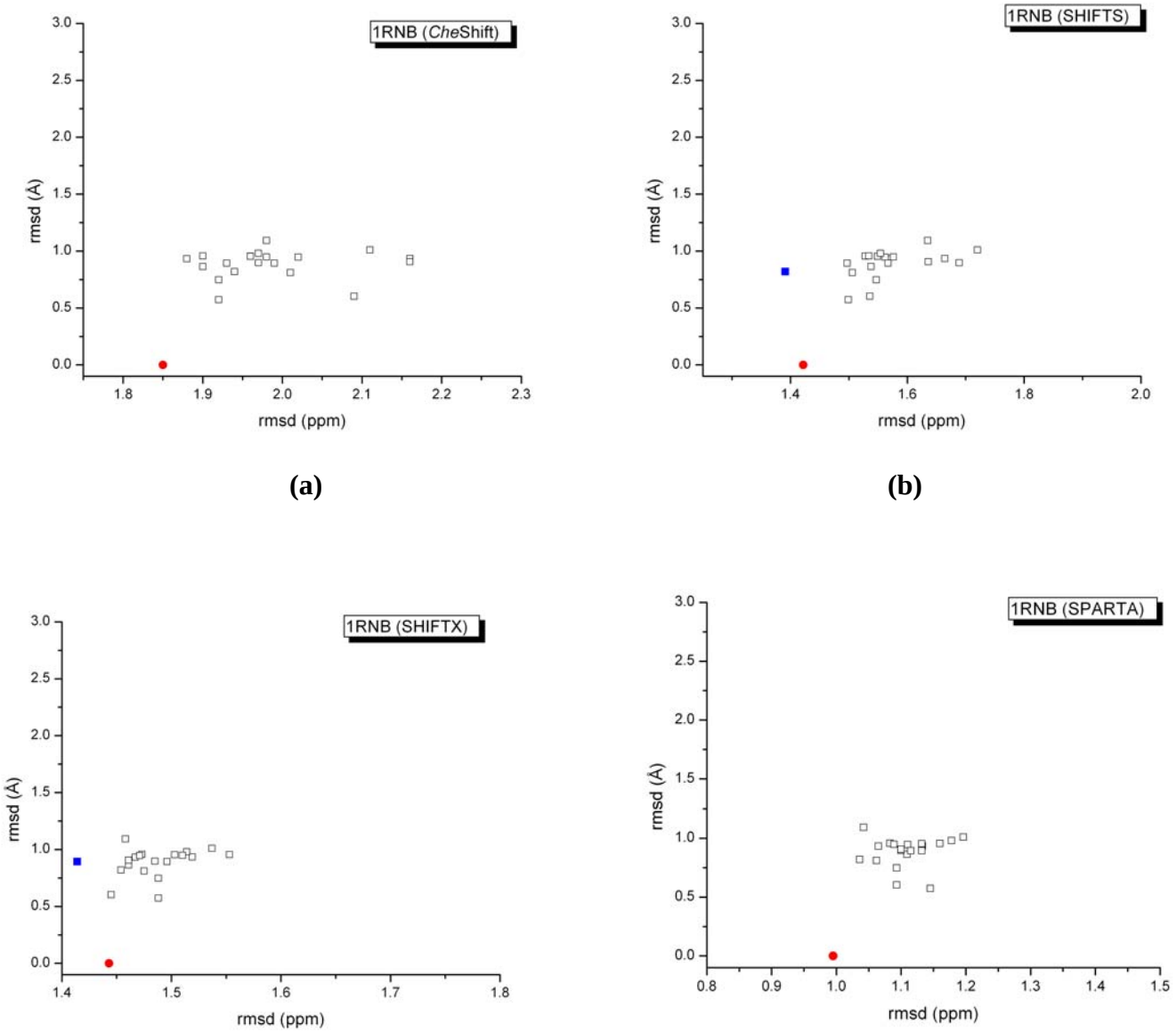


Figure S4. Plot showing rmsd of the protein decoys (open squares), lower than 3 Å from the 'native' structure (PDB id 1RNB), versus the rmsd (ppm) between observed and predicted

$^{13}C^{\alpha}$ chemical shifts, computed with four different servers, namely (**a**) *Che*Shift; (**b**) SHIFTS; (**c**) SHIFTX; and (**d**) SPARTA. In each panel, the red-filled circle denotes the 'native' structure (1RNB) from which the decoys were obtained; if one decoy shows a lower rmsd (ppm) than that of the 'native' structure, it is denoted by a blue-filled square.

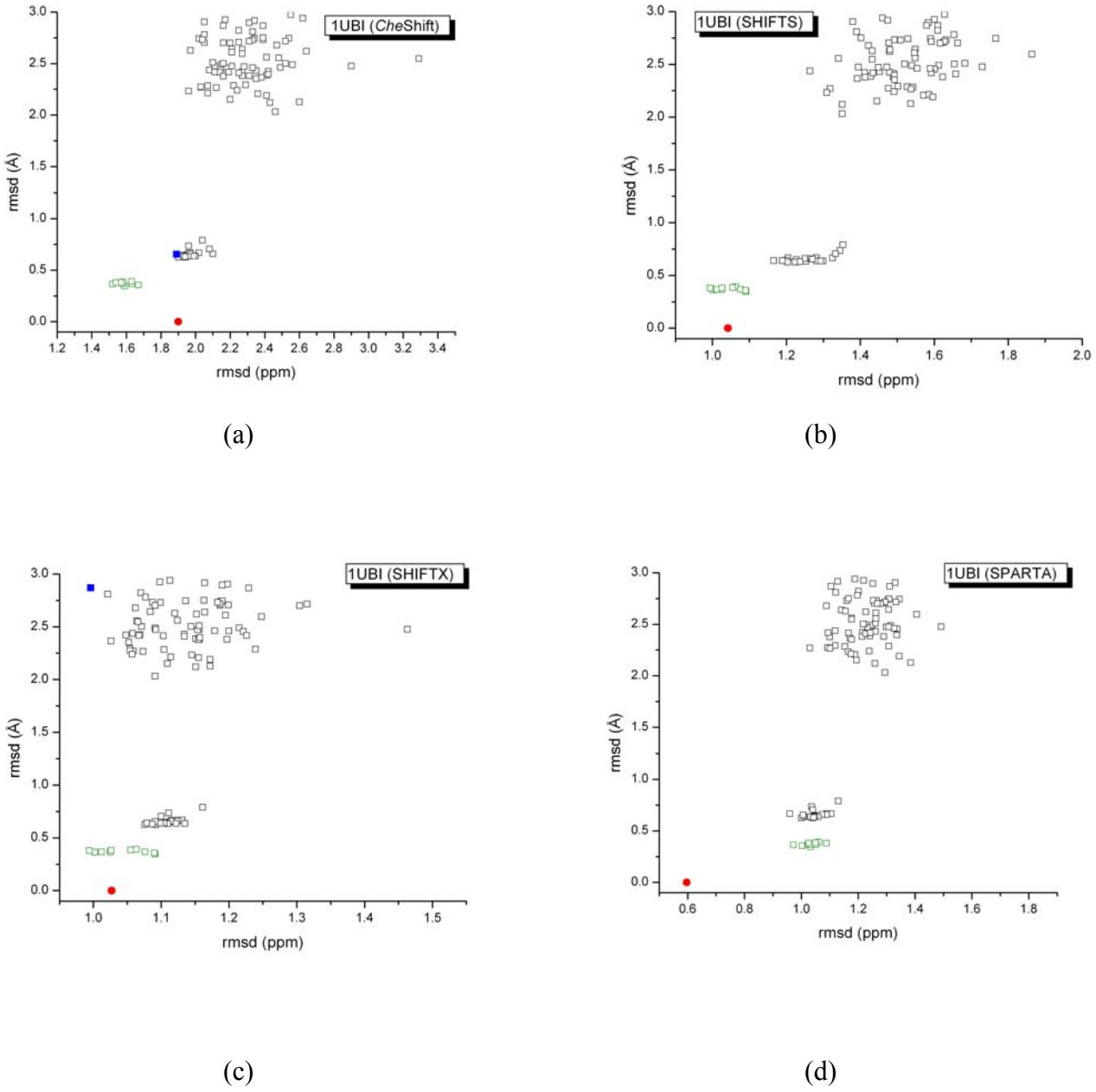


Figure S5.- Plot showing rmsd of the protein decoys (open squares), lower than 3 Å from the 'native' structure (PDB id 1UBI), versus the rmsd (ppm) between observed and predicted $^{13}C^{\alpha}$ chemical shifts, computed with four different servers, namely (**a**) *Che*Shift; (**b**) SHIFTS; (**c**) SHIFTX; and (**d**) SPARTA. In each panel, the red-filled circle denotes the 'native'

structure (1UBI) from which the decoys were obtained; if one decoy shows a lower rmsd (ppm) than that of the 'native' structure, it is denoted by a blue-filled square. Each of the open green squares in each panel, were computed from 10 conformations of protein PDB id 1D3Z solved by NMR spectroscopy at very-high resolution.[1]

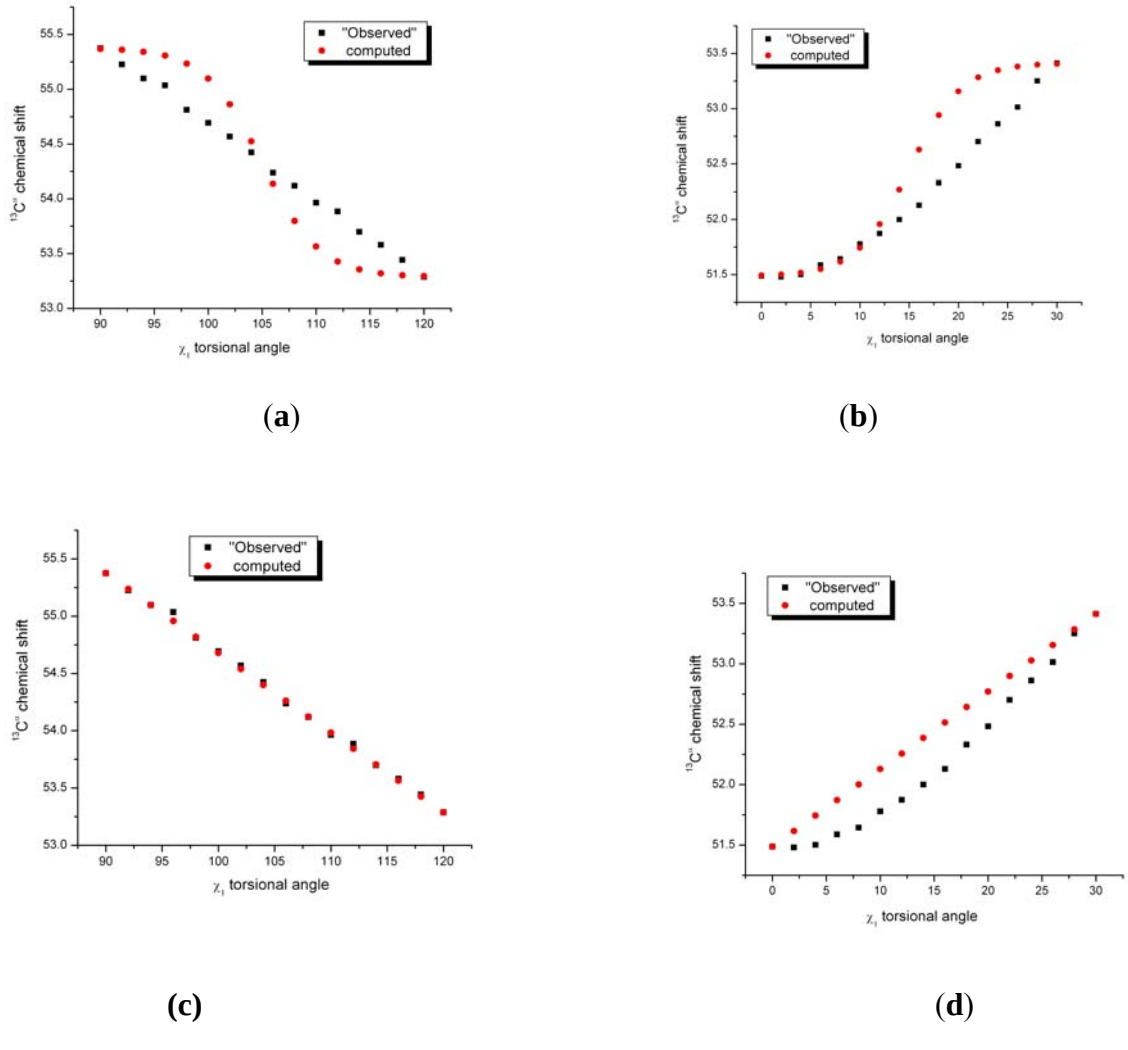
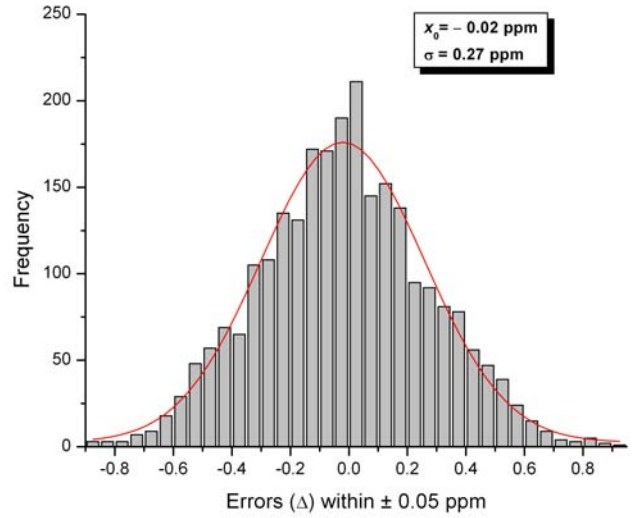
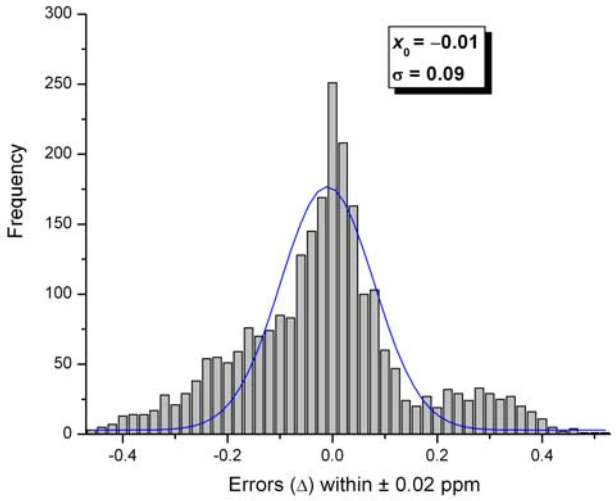


Figure S6. In each panel, the black-filled squares indicate the "observed" values, i.e., the $^{13}C^{\alpha}$ chemical shifts computed on a fine grid, namely with 5 degree steps for $\chi 1$, using a small basis set and linearly extrapolated to a large basis set (see Section I, Supporting Information). The red-filled circles indicated the interpolated $^{13}C^{\alpha}$ chemical shifts, between the two known values, namely the first and last one of the 30° interval, by using a Gaussian [panels (**a**) and (**b**)] or linear [panels (**c**) and (**d**)] interpolation, respectively. For a given set of (ϕ, ψ, $\chi 2$) torsional angles, there is a total of 9 panels, using an interval of 30° as in the *Che*Shift database, to fully describe the $\chi 1$ torsional angle variations (–180°, 180°). Among all these 9 panels we selected the best and worst results for each of the two interpolation

methods, i.e., shown in FigureS5 (**a**) and (**b**) for the Gaussian and (**c**) and (**d**) the linear interpolation, respectively. Five out of 9 panels for the linear interpolations are 'similar' to the results shown in Figure S5 (**c**); the good performance of the linear interpolation is reflected in the standard deviation, σ, of the frequency of the error distribution [see Figure S7 (**b**)].



(**a**)



(**b**)

Figure S7. Each panel denotes the frequency of the error distribution, between the "observed" $^{13}C^{\alpha}$ chemical shifts and those computed by using (**a**) Gaussian or (**b**) linear interpolation. The frequency of the error distribution can be fitted by a Gaussian curve with the mean value, x_o, and standard deviation, σ shown in the inserted panels of each Figure. A total of 2,500 conformations of Ser were evaluated, by sampling the ϕ and ψ every 2^o and χ1

every 5^{o}, in the following torsional angle ranges (-150^{o}; -160^{o}), (160^{o}, 170^{o}) and (-180^{o}, 180^{o}), respectively. The computed interpolation values were obtained by using a three-dimensional generalization of Equations (1) and (2) of Section III.

References

1. Cornilescu G, Marquardt JL, Ottiger M, Bax A. Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase, 1998**,** *J. Am. Chem. Soc.***120**, 6836-6837.
2. Aramini JM, Sharma S, Huang YJ, Swapna GVT, Ho CK, Shetty K, Cunningham K, Ma L-C, Zhao L, Owens LA, Jiang M, Xiao R, Liu J, Baran MC, Acton TB, Rost B and Montelione GT. Solution NMR structure of the SOS response protein YnzC from *Bacillus subtilis* (2008) *Proteins* 72:526-530.
3. Chien C.-y., Tejero R., Huang Y., Zimmerman D.E., Rios C.B., Krug R.M., Montelione G.T. A novel RNA-binding motif in influenza A virus non-structural protein 1, 1997, Nature Structural Biology, **4**, 891-895.
4. Vila, J. A.; Aramini J. M.; Rossi P.; Kuzin A.; Su M.; Seetharaman J.; Xiao R.; Tong L.; Montelione G. T.; H. A. Scheraga. Quantum Chemical $^{13}C^{\alpha}$ Chemical Shift Calculations for Protein NMR Structure Determination, Refinement, and Validation. *Proc. Natl. Acad. Sci. USA*. 2008, **105**, 14389-14394.
5. Liu J., Lynch P.A., Chien C.-y., Montelione G.T., Krug R.M., Berman H.M. (1997) Crystal structure of the unique RNA-binding domain of the influenza virus NS1 protein. *Nature Structural Biology* **4**, 896-899.
6. Vila JA, Villegas ME, Baldoni HA and Scheraga HA. Predicting $^{13}C^{\alpha}$ chemical shifts for validation of protein structures (2007) *J Biomol NMR* 38:221-235.
7. Némethy G, Gibson KD, Palmer KA, Yoon CN, Paterlini G, Zagari A, Rumsey S, and Scheraga HA. Energy parameters in polypeptides. 10. Improved geometrical

parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to praline-containing peptides. (1992) *J Phys Chem* 96:6472-6484.

8. Chesnut DB and Moore KD. Locally dense basis-sets for chemical-shift calculations (1989) *J Comp Chem* 10:648-659.
9. Vila JA and Scheraga HA. Factors affecting the use of $^{13}C^{\alpha}$ chemical shifts to determine, refine, and validate protein structures (2008) *Proteins*, 71:641-654.
10. Pearson JG, Le H, Sanders LK, Godbout N, Havlin RH and Oldfield EJ. Predicting chemical shifts in proteins: Structure refinement of valine residues by using *ab initio* and empirical geometry optimizations (1997) *J Am Chem Soc* 119:11941-11950.
11. Jameson A. K.; Jameson C.J. J Chem Phys Lett 1997, 134, 461
12. M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong,

C. Gonzalez, and J. A. Pople, Gaussian 03, Revision E.01, Gaussian, Inc., Wallingford CT, 2004.

13. Vila J.A., Baldoni H.A. and Scheraga H.A. Performance of Density Functional Models to Reproduce Observed $^{13}C^{\alpha}$ Chemical Shifts of Proteins in Solution. *J. Comp. Chem.*, **30**, 884-892 (2009).
14. Press H.W.; Teukolsky S.A.; Vetterling W.T.; Flannery BP, in Numerical Recipes in FORTRAN 77. The Art of Scientific Computing, Second Edition, Cambridge University Press **1992**, Chapter 14, page 630-633.
15. Wang Y.J., Jardetzky O. Probability-based protein secondary structure identification using combined NMR chemical-shift data. *Protein Sci.* 2002, **11**, 852-861.
16. Lovell S.C., Word J.M, Richardson J.S. and Richardson D.C. The penultimate rotamer library. Proteins, 2000, **40**, 389-408.
17. Vijay-Kumar S, Bugg CE, Cook WJ. Structure of ubiquitin refined at 1.8 Å resolution (1987) *J Mol Biol* 194:531-544.
18. Xu X.-P. and Case D.A. Automated prediction of ^{15}N, $^{13}C^{a}$, $^{13}C^{b}$ and $^{13}C'$ chemical shifts in proteins using a density functional database. *J. Biomol. NMR*, 2001, **21**, 321-333.
19. Xu X.-P. and Case D.A. Probing multiple effects on ^{15}N, $^{13}C^{a}$, $^{13}C^{b}$ and $^{13}C'$ chemical shifts in peptides using density functional theory. *Biopolymers*, 2002, **65**, 408-423.
20. Neal S., Nip A.M., Zhang H. and Wishart D.S. Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *J. Biomol. NMR*, 2003, **26**, 215-240.
21. Meiler J. PROSHIFT: Protein chemical shift prediction using artificial neural networks. *J. Biomol. NMR*, 2003, **26**, 25-37.

22. Shen Y. and Bax Ad. Protein backbone chemical shifts predicted from searching a database for torsional angle and sequence homology. J. Biomol. NMR, 2007, **38**, 289-302.

23. Lindorff-Larsen K., Best R.B., Depristo M.A., Dobson C.M. and Vendruscolo M. Simultaneous determination of protein structure and dynamics, 2005, Nature, **433**, 128-132.

24. Ulrich E.L., Akutsu H., Doreleijers F.J., Harano Y., Loannidis E.Y., Lin J., Livny M., Mading S., Maziuk D., Miller Z., Nakatani E., Schulte C.F., Tolmie D.E., Wenger R.K., Yao H. and Markley J.L. BioMagResBank, Nucleic Acids Res., 2007, **36**, D402-D408.

25. Ladner J.E., Wladkowski B.D., Svensson L.A., Sjölin L. and Gilliland G. X-ray structure of Ribonuclease A-uridine vanadate complex at 1.3 Å resolution. Acta Cryst., 1997, D**53**, 290-301.

26. Vila J.A. and Scheraga H.A. Assessing the accuracy of protein structures by quantum mechanical computations of $^{13}C^{\alpha}$ chemical shifts. *Accounts of Chemical Research*, 2009, in press.

27. Kuszewski J., Qin J., Gronenborn A.M., Clore M.G. The impact of direct refinement against $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts on protein structure determination by NMR. J. Mag. Res., 1995, B**106**, 92-96.