Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks

Yang Shen and Ad Bax

Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892-0520, U.S.A.

SUPPORTING INFORMATION

Table S1 Architecture of neural networks used by TALOS-N.

	(ϕ,ψ)-ANN	$(\phi,\psi)_{ssNMR}$-ANN	SS-ANN	SS_{seq}-ANN	$(\chi^1)_k$-ANN
level1					
Input	(5-mer) size: 5×32 $[AA_{20}{}^{a},CS_{12}{}^{b}]_{\times5}$	(5-mer) 5×28 $[AA_{20}{}^{a},CS_{8}{}^{b}]_{\times5}$	(5-mer) 5×32 $[AA_{20}{}^{a},CS_{12}{}^{b}]_{\times5}$	(15-mer) 15×20 $[AA_{20}{}^{a}, B_{1}{}^{g}]_{\times15}$	(3-mer) 3×19 $[\phi/\psi/\chi^{1\,h},CS_{12}{}^{b}]_{\times3}$
Output	$[D_{324}]^{c}$	$[D_{324}]^{c}$	$[H,E,C]^{e}$	$[H,E,C]^{e}$	$[\Delta\chi 1]$
level2					
Input	(5-mer) size: 5×324 $[OUT1_{324}]^{d}$	(5-mer) 5×324 $[OUT1_{324}]^{d}$	(5-mer) 5×3 $[OUT1_{3}]^{f}$	(15-mer) 15×4 $[OUT1_{3}, B_{1}{}^{g}]^{f}$	-
Output	$[D_{324}]^{c}$	$[D_{324}]^{c}$	$[H,E,C]^{e}$	$[H,E,C]^{e}$	-
Training/ Validation					
Protocol	3-2-1	3-2-1	3-2-1	3-1-2	3-2-1
Database	chemical shift	chemical shift	chemical shift	structural	chemical shift & structural
Training Performance[h]					
Q_{obs}/Q_3	96.5%	95.8%	88.2% (88.6%)	81.6% (81.2%)	-

[a] Amino acid type represented by its values (unit size n=20 for each amino acid) in the BLOSUM62 sequence homology matrix.

[b] For the (ϕ,ψ)-ANN, $^1H^N/^{15}N/^1H^{\alpha}/^{13}C^{\alpha}/^{13}C^{\beta}/^{13}C'$ secondary chemical shifts (n=6) and Boolean number indicators for missing chemical shifts (n=6); for the $(\phi,\psi)_{ssNMR}$-ANN, $^{15}N/^{13}C^{\alpha}/^{13}C^{\beta}/^{13}C'$ secondary chemical shifts (n=4) and Boolean number indicators for missing chemical shifts (n=4).

[c] 324-state ϕ/ψ distribution of the center residue of query pentapeptides (see Methods).

[d] 324-state ϕ/ψ distribution prediction output (n=5) from the first level ANN.

[e] Three-state secondary structure classification of the center residue of the query pentapeptides (for SS-ANN) and 15-mers (for SS_{seq}-ANN).

[f] Three-state secondary structure classification prediction output (n=5) from the first level ANN.

[g] An additional Boolean value is used to describe the absence of amino acids for positions of the sliding window located at the edge of the chain.

[h] Training performance of the (ϕ,ψ)-ANN, $(\phi,\psi)_{ssNMR}$-ANN and SS-ANN is evaluated for the validation dataset of the chemical shift database; training performance of the SS-ANN on the 34-protein validation dataset is given in parentheses; training performance of SS_{seq}-ANN is evaluated for the protein structure database, and the performance on the 91 target proteins used by CASP9 (Kryshtafovych et al. 2011) is given in parentheses.

Table S2 Performance of TALOS-N for solid-state chemical shift data.

	Consistent		***Ambiguous***	***<sd>***[b] ***(ϕ/ψ)***	***Rmsd***[c] ***(ϕ/ψ)***
	All	**Bad**	**Warn**		
			For chemical shift database		
(ϕ,ψ)-ANN	89.2%[d] *[84.6%/4.6%]*	3.6%[e] *[3.0%/23%]*	10.8%[d]	8.8/8.6	12.2/11.5
$(\phi,\psi)_{\mathrm{ssNMR}}$-ANN	89.6%[d] *[85.5%/4.1%]*	3.6%[e] *[2.9%/26%]*	10.4%[d]	8.7/8.6	12.5/12.1
			For 34-protein validation dataset		
(ϕ,ψ)-ANN	90.3%[d] *[85.5%/4.8%]*	3.5%[e] *[2.8%/26%]*	9.7%[d]	8.5/8.4	12.2/11.6
$(\phi,\psi)_{\mathrm{ssNMR}}$-ANN	90.7%[d] *[86.0%/4.7%]*	3.5%[e] *[2.9%/23%]*	9.3%[d]	8.5/8.3	12.2/12.0

[a] The "solid-state" chemical shift data were prepared by removing all ^{1}H chemical shifts from the two validation databases. TALOS-N runs were performed by using the trained (ϕ,ψ)-ANN and $(\phi,\psi)_{\mathrm{ssNMR}}$-ANN, respectively.

[b] Average standard deviation of ϕ/ψ torsion angles among the 25 (or 10) best matched tripeptides for "Good" (or "Generous") TALOS-N predictions, representing the average precision of the predictions. See footnote *e* for the definition of a good/bad prediction.

[c] Rmsd values between TALOS-N predicted ϕ/ψ angles ("Good" predictions only) and observed ϕ/ψ angles in the reference structures, representing the average accuracy of the predictions.

[d] Percentage relative to the total number of residues for which predictions are calculated, excluding those residues that are dynamically disordered on the basis of their RCI value.

[e] Percentage relative to the number of total consistently predicted residues ("Good"+"Bad"). A good prediction is defined based on the criterion sqrt[$(\phi_{\mathrm{obs}} - \phi_{\mathrm{pred}})^2 + (\psi_{\mathrm{obs}} - \psi_{\mathrm{pred}})^2$] ≤ 60°, and as bad otherwise.

Table S3 Proteins selected for the second validation dataset.

BMRB id	*PDB id*	*Resolution (Å)*
15952	3F04	1.35
15962	3MOK	1.55
16042	1FDQ	2.10
16059	2Q2T	2.30
16135	3JSC	1.50
16209	3GL6	1.90
16260	3IHZ	1.67
16265	1ZYN	2.30
16327	1FVK	1.70
16329	3BCI	1.81
16330	3BD2	1.80
16408	3DSO	1.55
16409	2WF7	1.05
16430	3BBE	2.20
16465	2O5G	1.08
16481	2VY0	2.16
16514	1SFC	2.40
16599	3KPX	1.90
16656	3IPF	1.99
16717	2QYJ	2.05
16739	3E7R	1.00
16770	2X9C	2.45
16805	3LMO	2.00
16806	3MER	2.20
16813	3IQL	1.40
16822	3DAT	2.30
16891	1IRD	1.25
16925	2PPN	0.92
16982	1GUD	1.71
17010	3WRP	1.80
17130	1AAY	1.60
17160	2BEM	1.55
17162	2Z1O	1.75
17264	2O5G	1.08

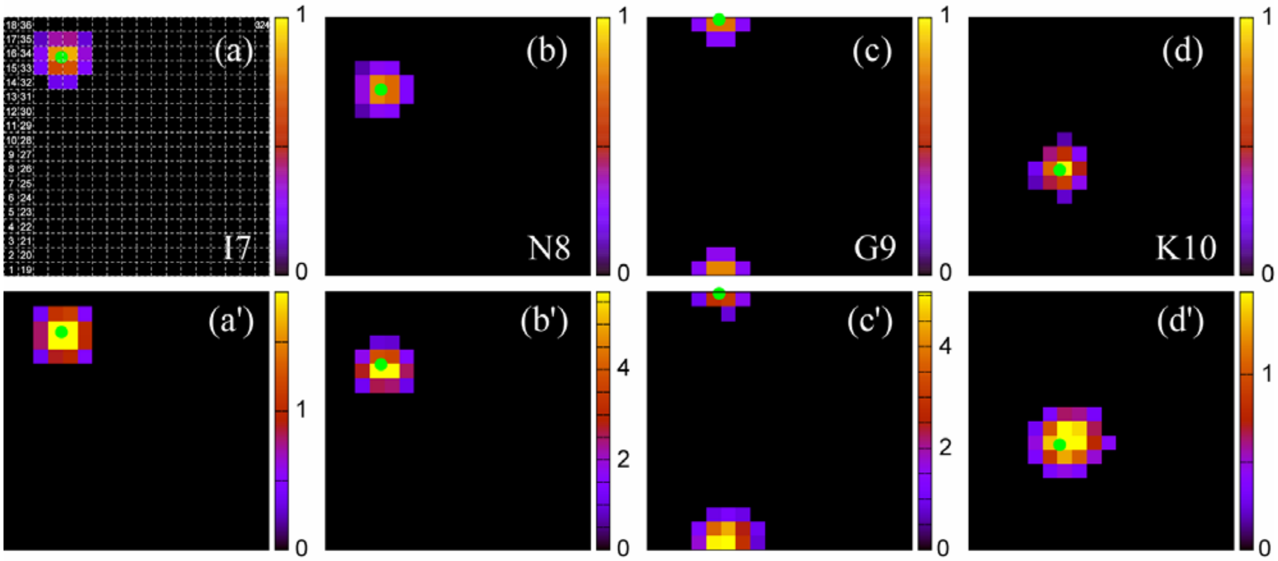

Fig. S1 Residue density distributions predicted for I7-K10 of protein GB3. Each panel presents a (ϕ,ψ) map ranging from -180 to +180º for both ϕ (horizontal axis) and ψ, with each map divided into 18×18 voxels (a). Panels (a-d) represent raw residue density distributions generated using Eq. 1. Panels (a'-d') show the residue density distribution after normalization according to $D(\phi_i,\psi_i)_k/\sqrt{\langle D_k \rangle}$. The indexing numbers of the 18×18 (ϕ,ψ) voxels are included for the left two columns of panel (a).

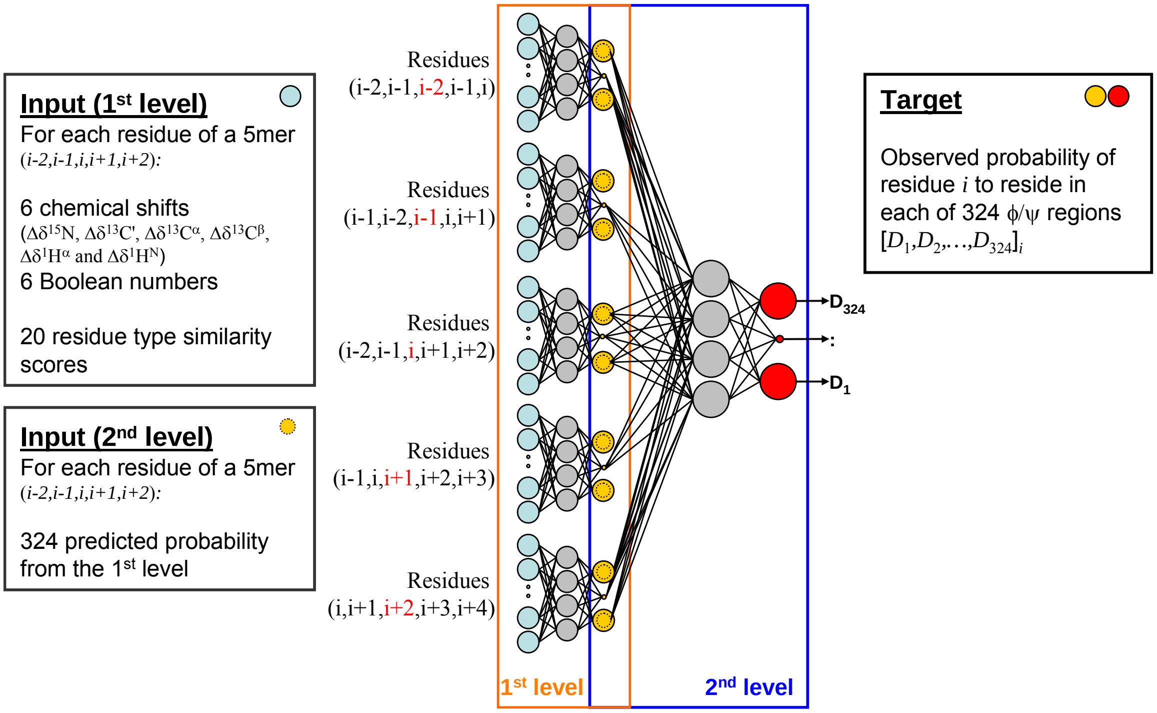
Input (1st level)
For each residue of a 5mer (i-2,i-1,i,i+1,i+2):
6 chemical shifts ($\Delta\delta^{15}$N, $\Delta\delta^{13}$C', $\Delta\delta^{13}$C$^{\alpha}$, $\Delta\delta^{13}$C$^{\beta}$, $\Delta\delta^{1}$H$^{\alpha}$ and $\Delta\delta^{1}$H^{N})
6 Boolean numbers
20 residue type similarity scores
Input (2nd level)
For each residue of a 5mer (i-2,i-1,i,i+1,i+2):
324 predicted probability from the 1st level
Residues (i-2,i-1,i-2,i-1,i)
Residues (i-1,i-2,i-1,i,i+1)
Residues (i-2,i-1,i,i+1,i+2)
Residues (i-1,i,i+1,i+2,i+3)
Residues (i,i+1,i+2,i+3,i+4)
1st level
2nd level
D_{324}
:
D_1
Target
Observed probability of residue i to reside in each of 324 ϕ/ψ regions $[D_1,D_2,\ldots,D_{324}]_i$

Fig. S2 Architecture of the two-level feed-forward artificial neural network, (ϕ/ψ)-ANN, used to predict the voxel of the Ramachandran map in which a given residue resides. The ANN calculates the probability for any center residue of a pentapeptide fragment to reside in one of the 324 20°×20° ϕ/ψ voxels. The ANN uses as input for the first level feed-forward prediction the known parameters characterizing each of the five residues of the pentapeptide and it is trained on the 580-protein database to predict the known output ϕ/ψ state. Besides the six chemical shifts and six Boolean numbers indicating whether a given chemical shift is missing or not, input parameters for each residue of the pentapeptide are represented by a 20-dimensional vector, consisting of the coefficients of its row in the BLOSUM62 matrix (see http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=sef.figgrp.194). A total of 160 input parameters (aqua filled circles) per pentapeptide are used to predict the probability for occupation of each of the 324 ϕ/ψ voxels by its center residue (yellow), which are subsequently used as input for the second level of the ANN. 200 hidden nodes (grey) are used for the first level of the ANN. The ANN output (yellow dotted circles) of the first level for 5 sequential residues is used to fine-tune the prediction of the ϕ/ψ voxel (red), using a hidden level consisting of 360 nodes (grey). For more details, see main text.

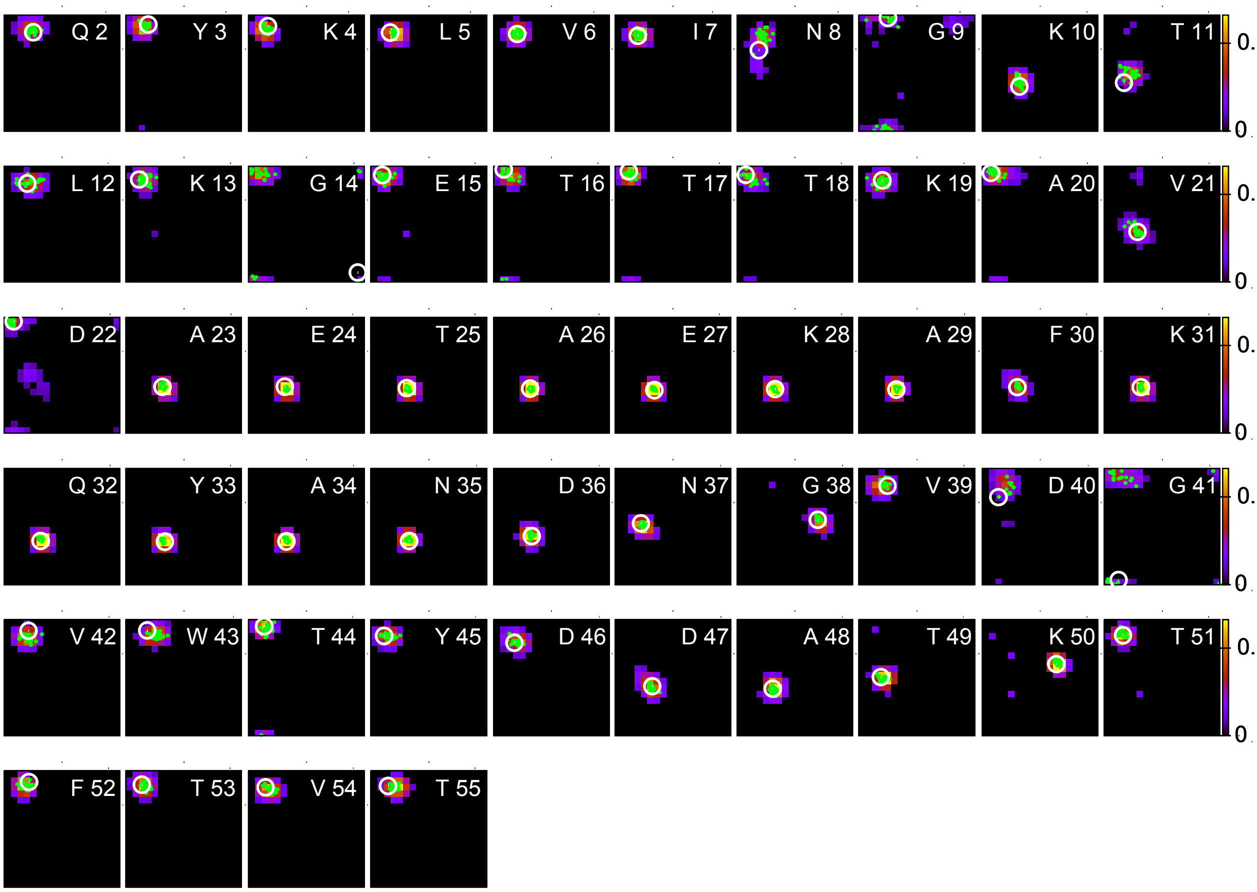

Fig. S3 (ϕ,ψ)-ANN predicted (ϕ,ψ) distribution for residues in GB3, which is not present in the ANN training dataset. For each residue in GB3 (except the first and the last residues), above average (ϕ,ψ) probability predictions, i.e., the 20°×20° ϕ/ψ regions with a high predicted density (one standard deviation above the average density of all 324 voxels, see Methods), are plotted. The ϕ/ψ angles observed in the reference structure (PDB: 2OED) are marked by white circles; the ϕ/ψ angles from the center residue of the 25 best matched database fragments are marked by green dots. Each residue panel corresponds to a Ramachandran map, with the horizontal axis corresponding to the ϕ angle (ranging from -180° to 180°), and the vertical axis depicting ψ (also ranging from -180° to 180°).

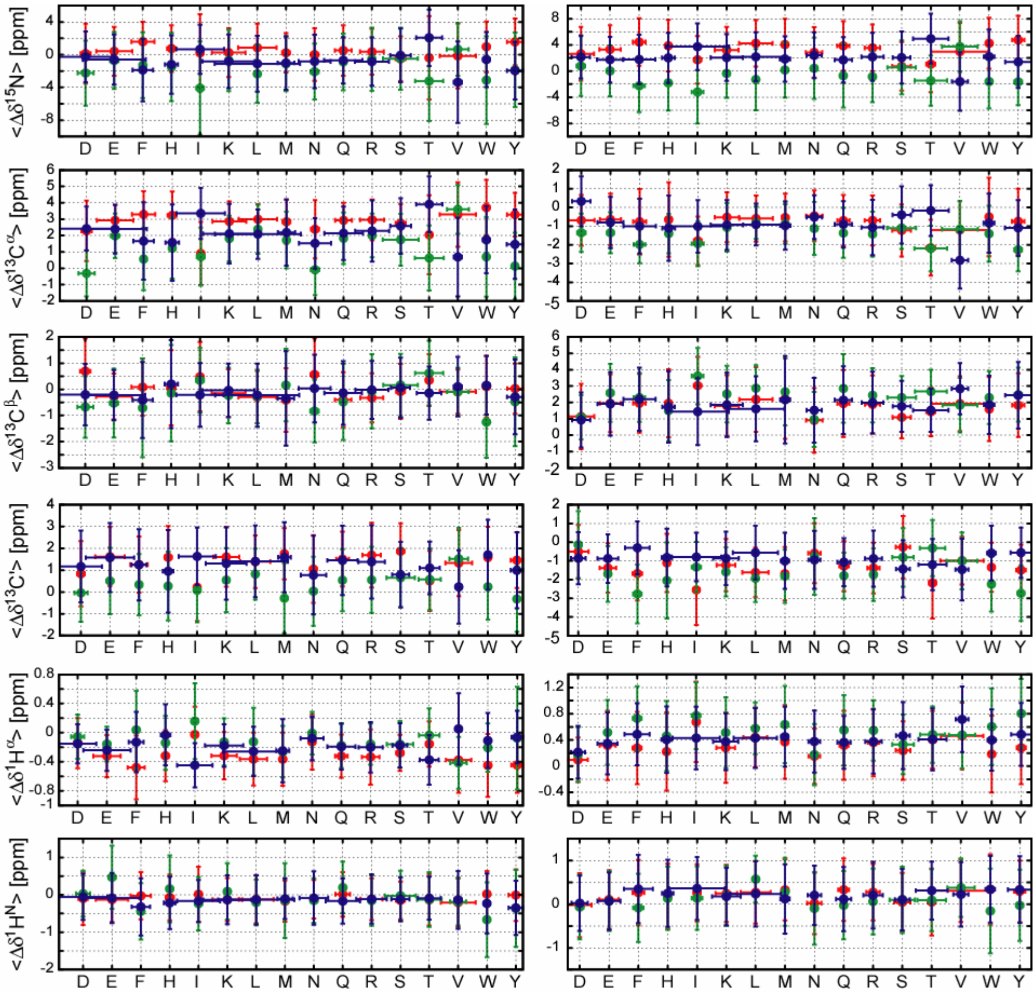

Fig. S4 χ^1-dependence of chemical shifts in proteins in the α region (left panels; $-160^{\circ}<\phi<0^{\circ}$ and $-70^{\circ}<\psi<60^{\circ}$) and β regions (right panels; $160^{\circ}<\phi<0\ ^{\circ}$ and $\psi\leq-70^{\circ}$ or $\psi\geq60^{\circ}$). For each of 17 residue types with a χ^1 torsion angle (except Pro), the average ^{15}N, ^{13}C$^{\alpha}$, ^{13}C$^{\beta}$, ^{13}C', ^{1}H$^{\alpha}$ and ^{1}H^{N} secondary chemical shifts are shown for the residues in the chemical shift database for each of three χ^1 rotamer states: *g*+ (green), *g*- (blue) and *t* (red). The length of the horizontal "error" bar represents the normalized population of the residues with a given χ^1 rotamer; the vertical error bars correspond to the standard deviations observed for the chemical shift within residues of a given χ^1 rotamer type.

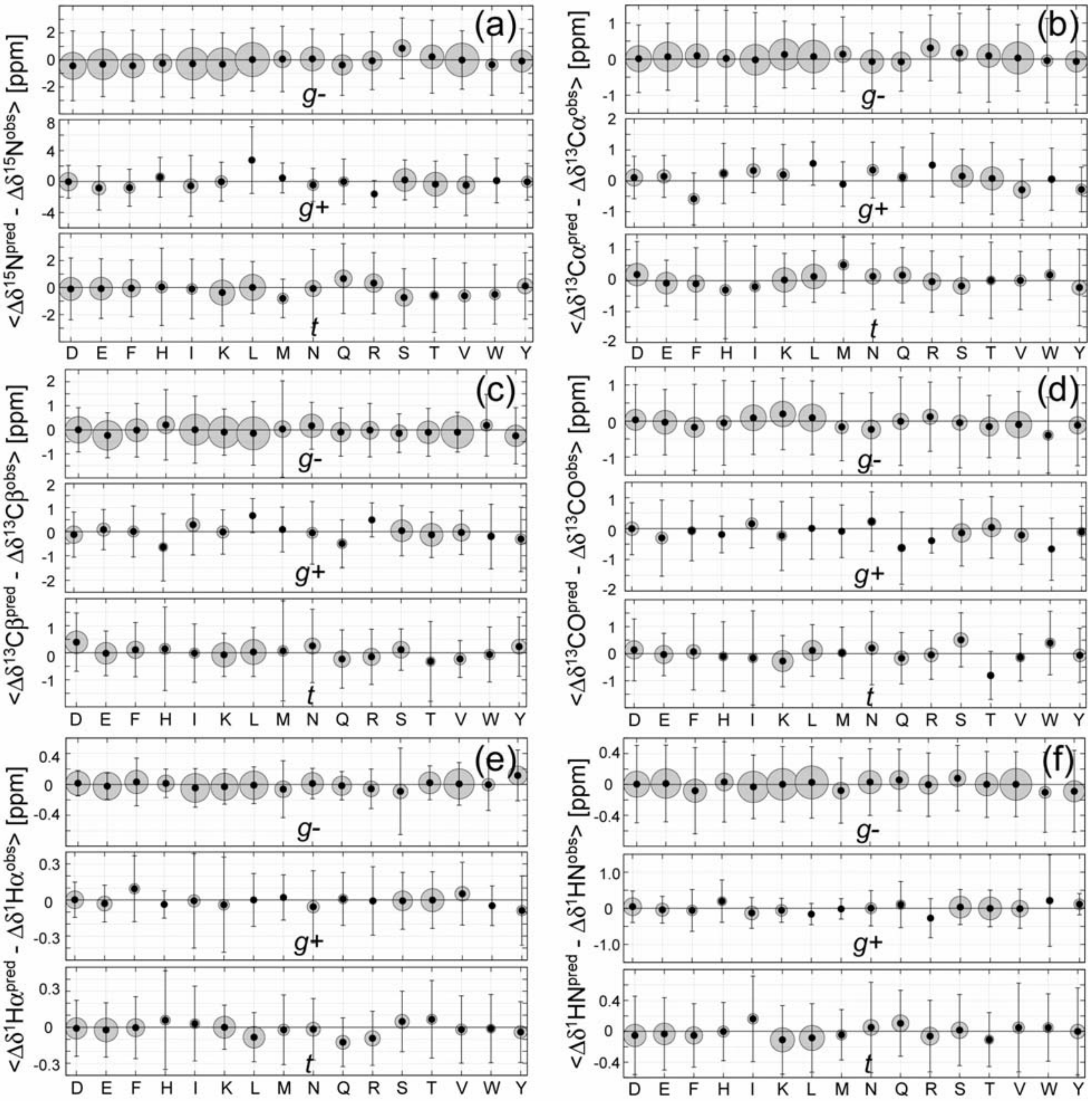

Fig. S5 Absence of systematic χ^1-dependent errors in SPARTA+ predicted chemical shifts. For each of 17 residue types acids with a χ^1 torsion angle (except Pro), the average difference between the observed and the SPARTA+ derived ^{15}N, ^{13}C$^{\alpha}$, ^{13}C$^{\beta}$, ^{13}C', ^{1}H$^{\alpha}$ and ^{1}H^{N} secondary chemical shifts (a-f) are calculated for all residues in the 34-protein validation database, and displayed according to their observed χ^1 rotamer state, i.e., *g*- (upper panels), *g*+ (middle) and *t* (lower). The area of the grey bubbles represents the normalized population of the residues with a given χ^1 rotamer.

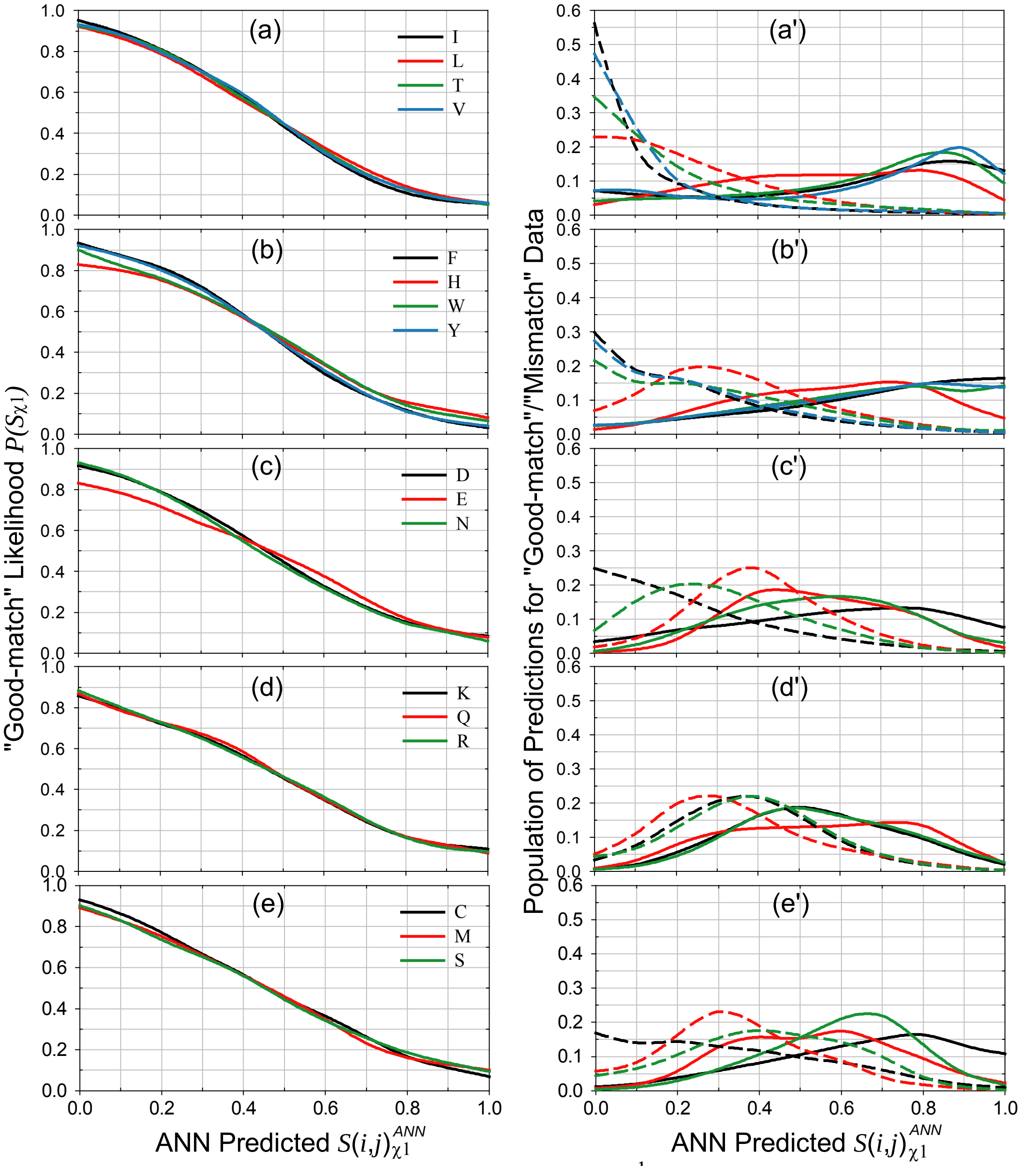

Fig. S6 Prediction performance of 17 residue-specific χ^1-ANNs. For each of 17 amino acids with a χ^1 torsion angle, the prediction performance of the trained $(\chi^1)_a$-ANN is evaluated by using the validation dataset (see Methods). The validation data is separated into two groups, the "good-match" data which have an observed χ^1 rotamer matching score $S(i,j)_{\chi1} = 0$, and the "mismatch" data with $S(i,j)_{\chi1} = 1$ (see Methods). The trained $(\chi^1)_a$-ANN is used to calculate back for all data the χ^1 rotamer matching score $S(i,j)_{\chi1}^{ANN}$. All predictions are then binned into ten groups, spread evenly from 0 to 1, according to their $S(i,j)_{\chi1}^{ANN}$ score. For each bin, the prediction performance is evaluated in terms of: (1) the ratio of the correct predictions (left panels, a-e), or the likelihood $P(S_{\chi1})$ that the center residue of the query fragment has the same χ^1 rotamer state as the center residue of the database fragment; (2) the population of the true (correct) predictions in this bin relative to the predictions for all "good-match" data (dashed lines in right panels, a'-e'); and (3) the population of the predictions for the "mismatch" data relative to the predictions for all "mismatch" data (solid lines in right panels, a'-e').

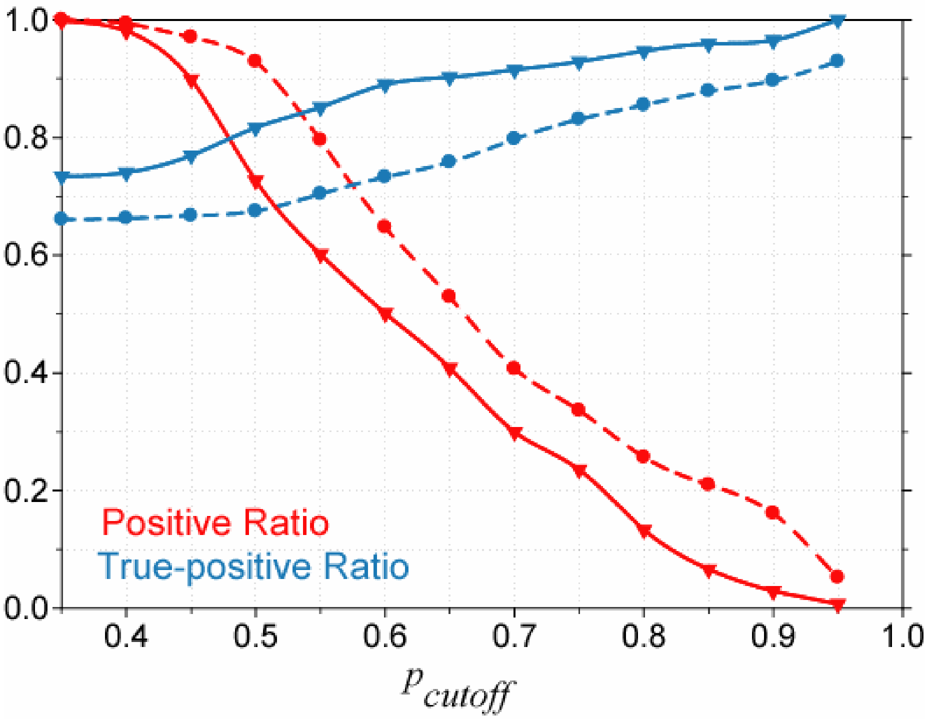

Fig. S7 Plot of χ^1 rotamer prediction performance versus the p_{cutoff} cutoff value used. The solid curves correspond to the TALOS-N prediction performance for the 34-protein validation dataset, with the red line marking the fraction of residues for which the $p(<S(i)_{\chi 1,c}>)$ falls above the cut-off value, p_{cutoff}, and the solid blue line marking the fraction of the corresponding residues for which the TALOS-N prediction of χ^1 agrees with the X-ray reference structure. Dashed lines correspond to the analogous numbers when using the backbone-dependent χ^1 rotamer probabilities tabulated by Shapovalov and Dunbrack (see http://dunbrack.fccc.edu/bbdep2010/).

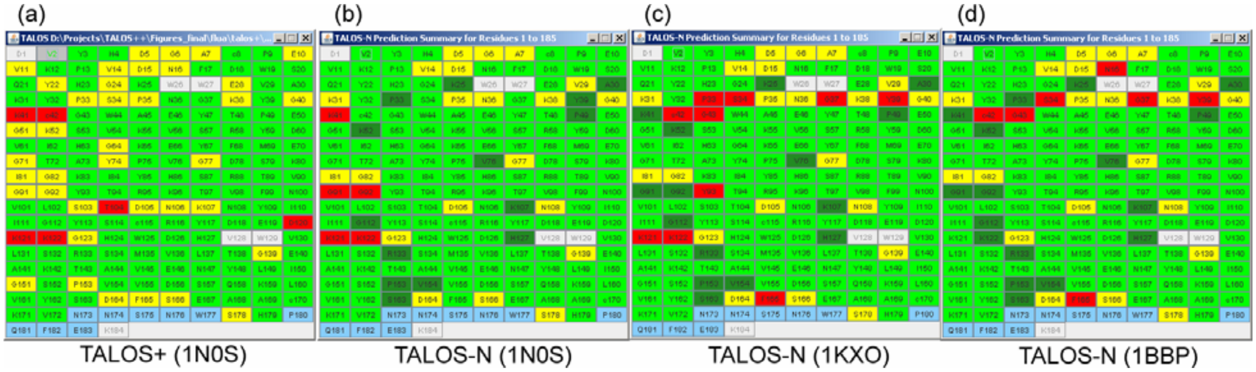
(a)
(b)
(c)
(d)
TALOS D:\Projects\TALOS++\Figures_final\flua\talos+\...
TALOS-N Prediction Summary for Residues 1 to 185
TALOS-N Prediction Summary for Residues 1 to 185
TALOS-N Prediction Summary for Residues 1 to 185
TALOS+ (1N0S)
TALOS-N (1N0S)
TALOS-N (1KXO)
TALOS-N (1BBP)

Fig. S8 TALOS+ and TALOS-N predictions for FluA. (A) Graphical interface of TALOS+ predictions, using the 1N0S reference X-ray structure (2.0 Å resolution). (B-D) TALOS-N predictions, using X-ray reference structures of (B) 1N0S, (C) 1KXO (1.8 Å) and (D) 1BPP (2.0 Å); the number of consensus "Strong", "Generous", "Ambiguous" and "Bad" predictions are 122, 12, 21 and 0, respectively. Note that both TALOS+ and TALOS-N "Bad" predictions are assigned based on the new tighter criterion introduced in this work, and residues 34 and 37 of 1KXO and 1BBP are mutated relative to the sequence of 1N0S. None of the TALOS-N "Bad" predictions disagree with all three X-ray structures, and the annotation "Bad" therefore is likely to reflect true differences between the crystalline and solution states.

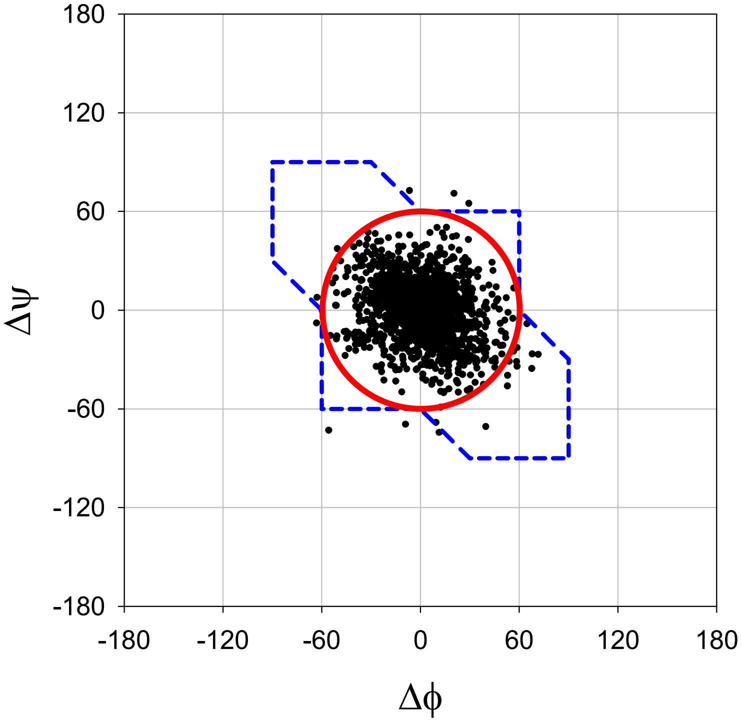

Fig. S9 Range of differences in (ϕ,ψ) torsion angles observed for protein fragments with very similar backbone conformations. For each of seventy 7-residue fragments in ubiquitin, the backbone coordinates are taken from a reference X-ray structure (PDB entry 1ubq). The protein structural database is then searched for the first 50 fragments that match the backbone N, C^{α}, C' and O atoms to better than 0.8 Å. The ϕ and ψ torsion angle difference between the center residues of the ubiquitin fragments and the database fragments (Δϕ and Δψ) are plotted as black dots. About 99% of these dots fall within a circle of 60° radius (red circle), or $\sqrt{\Delta\phi^2 + \Delta\psi^2} \leq 60°$, indicating that a $\sqrt{\Delta\phi^2 + \Delta\psi^2}$ cutoff of 60° is an appropriate criterion for TALOS-N to assign an unambiguously predicted ϕ/ψ angles as "Good" (the allowed Δϕ and Δψ is shown as the area covered by the red circle) or "Bad" (the area not covered by the red circle), where Δϕ and Δψ are $\phi_{obs}-\phi_{pred}$ and $\psi_{obs}-\psi_{pred}$, respectively. For comparison, the region marked by the dashed blue line marks the tolerance originally used by TALOS+, using the criteria $|\Delta\phi| \leq 60°$ and $|\Delta\psi| \leq 60°$, or ($90° \geq |\Delta\phi| > 60°$ or $90° \geq |\Delta\psi| > 60°$ and $|\Delta\phi + \Delta\psi\ | \leq 60°$). The area covered by these looser TALOS+ criteria is 1.94 times larger than that defined by the tighter criterion of TALOS-N.