Supplementary Material

TALOS+: A hybrid method for predicting protein backbone torsion angles from NMR chemical shifts

Yang Shen[1], Frank Delaglio[1], Gabriel Cornilescu[2], Ad Bax[1]

Table S1. Neural network prediction statistics for the 3-state ϕ/ψ torsion angle distribution with different models

	3-5 ANN model[a]	*3-3 ANN model*[a]	*3-3 ANN (i) model*[a]	*3-3 ANN (i-1) model*[a]	*3-3 ANN (i+1) model*[a]
Total number N^{all} (A/B/P)[b]	17162 (9084/7331/747)	19894 (10297/8733/864)	19894 (10297/8733/864)	19894 (10297/8733/864)	19894 (10297/8733/864)
Q(A)/Q(B)/Q(P)[c]	0.972/0.969/0.857	0.969/0.968/0.843	0.949/0.949/0.669	0.960/0.961/0.834	0.951/0.954/0.819
Q_3[d]	0.966	0.963	0.937	0.955	0.947
TP(A)/TP(B)/TP(P)[e]	0.971/0.965/0.899	0.969/0.962/0.899	0.950/0.933/0.806	0.962/0.954/0.877	0.959/0.940/0.856
TP(All)[f]	0.966	0.963	0.937	0.955	0.947
Confidence >80%					
Total number N	15291	17482	16099	17282	16667
Q^{80}(A)/N%[g]	0.988/90.8%	0.987/90.3%	0.987/84.8%	0.987/88.8%	0.984/87.0%
Q^{80}(B)/N%[g]	0.992/89.2 %	0.992/88.6%	0.992/80.4%	0.989/86.9%	0.991/82.6%
Q^{80}(P)/N%[g]	0.943/59.4%	0.946/51.7%	0.872/38.7%	0.953/62.6%	0.939/56.2%
$Q^{80}{}_3/N_3$%[g]	0.988/89.1%	0.988/87.8%	0.987/80.9%	0.987/86.9%	0.986/83.8%
TP^{80} (A/B/P)	0.992/0.985/0.969	0.992/0.984/0.963	0.991/0.982/0.976	0.991/0.985/0.958	0.984/0.991/0.939
TP^{80}(All)	0.988	0.988	0.987	0.987	0.986

A 3-fold training and validation procedure was performed for each network; results shown in this table are obtained from the validation subsets during the validation procedure (see Methods)

[a] Models used to train the neural network, see Methods

[b] N^{all}, total number of tri-peptide or penta-peptide data in the training dataset, and number of data in each state *i* (*i* = *Alpha*, *Beta* or *Positive-*ϕ)

[c] Ratio of residues with state '*i*' that can be correctly predicted as state *i*, see eq 3

[d] Overall ratio of all residues whose state can be correctly predicted, see eq 4

[e,f] True-positive rate of neural network prediction, see eq 5; note that TP(all) is identical to Q_3

[g] N%: the percentage of data with a prediction confidence ≥0.8 for state *i* (Alpha/Beta/Positive-ϕ)

Table S2. Distribution of 3-state ϕ/ψ distribution prediction for residues in the training dataset

Confidence level	*Alpha Prediction*[a]				*Beta Prediction*[a]				*Positive-ϕ Prediction*[a]			
	N%[b]	A%[c]	B%[c]	P%[c]	N%[b]	A%[c]	B%[c]	P%[c]	N%[b]	A%[c]	B%[c]	P%[c]
0.0-0.1	0.3	58.8	29.4	11.8	0.4	34.3	51.4	14.3	2.9	36.4	13.6	50.0
0.1-0.2	0.6	51.7	29.3	19.0	0.7	41.9	50.0	8.1	2.7	28.6	14.3	57.1
0.2-0.3	0.7	50.0	32.4	17.6	0.9	25.9	61.7	12.3	3.9	43.3	16.7	40.0
0.3-0.4	0.6	58.7	30.2	11.1	1.1	29.8	61.7	8.5	4.3	6.1	24.2	69.7
0.4-0.5	1.0	56.7	32.7	10.6	1.1	20.4	72.4	7.1	5.2	15.0	17.5	67.5
0.5-0.6	1.5	73.2	17.2	9.6	1.5	18.5	73.1	8.5	6.8	11.5	9.6	78.8
0.6-0.7	1.8	78.8	15.9	5.3	2.3	18.6	77.0	4.4	8.6	15.2	4.5	80.3
0.7-0.8	3.2	85.0	9.8	5.2	3.5	11.7	84.4	3.9	13.7	4.8	9.5	85.7
0.8-0.9	**5.8**	**93.3**	**4.2**	**2.5**	**8.2**	**7.2**	**91.8**	**1.0**	**22.8**	**2.9**	**4.0**	**93.1**
0.9-1.0	**84.5**	**99.6**	**0.3**	**0.1**	**80.4**	**0.7**	**99.2**	**0.1**	**28.9**	**0.5**	**0.0**	**99.5**

Results are obtained from a 3-fold cross validation using a two-level neural network with a 3-3 ANN model *(Table S1, 3rd column)*; the predictions with high true-positive ratio (or with confidence≥0.8) are shown in bold.

[a] For all residues predicted as "Alpha", "Beta" and "Positive-ϕ", respectively

[b] The percentage of predictions in a given confidence range

[c] Percentage of experimentally observed "Alpha", "Beta" and "Positive-ϕ" residues, respectively, for predictions in a given confidence range

Table S3. Statistics of neural network predictions for test proteins which are not included in the TALOS database

Protein Name	ϕ/ψ distribution prediction				Secondary structure prediction						
					TALOS+				CSI	PSSI	PsiCSI
	Q^{80}(A)	Q^{80}(B)	Q^{80}(P)	Q_3^{80}	Q(H)	Q(E)	Q(L)	Q_3	Q_3	Q_3	Q_3
gb3	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*0.706*	*0.911*	*0.839*	*0.857*	*0.911*
DinI	*1.000*	*1.000*	*1.000*	*1.000*	*0.972*	*0.895*	*0.913*	*0.914*	*0.852*	*0.802*	*0.914*
BAF	*1.000*	*1.000*	*1.000*	*1.000*	*0.949*	-	*0.962*	*0.933*	*0.944*	*0.933*	*0.921*
tolR	*1.000*	*1.000*	*1.000*	*1.000*	*1.000*	*0.789*	*0.850*	*0.912*	*0.926*	*0.897*	*0.897*
HR2106	0.978	0.976	1.000	0.977	1.000	0.852	0.724	0.938	0.844	0.896	0.906
TM1112	1.000	0.984	1.000	0.988	0.700	0.941	0.957	0.911	0.744	0.867	0.889
TM1442	1.000	1.000	1.000	1.000	0.851	1.000	0.659	0.824	0.704	0.815	0.806
XcR50	1.000	1.000	1.000	1.000	1.000	0.875	0.852	0.960	0.827	0.907	0.893
MrR110	0.900	1.000	1.000	0.987	0.750	0.911	0.813	0.875	0.760	0.854	0.854
Spo0F	1.000	0.974	1.000	0.990	0.980	0.889	0.906	0.957	0.853	0.853	0.914
Paxillin	1.000	0.958	1.000	0.991	0.938	-	0.727	0.884	0.845	0.837	0.845
CtR107	0.953	0.974	1.000	0.967	0.834	0.889	0.880	0.876	0.765	0.810	0.863
HR41	0.962	0.981	1.000	0.970	0.952	0.774	0.910	0.893	0.843	0.836	0.881
Average				**0.990**				**0.907**	**0.826**	**0.859**	**0.884**

Proteins with an NMR reference structure are shown in *italic*

Table S4. Statistics of "problematic" TALOS/TALOS+ predictions for protein FluA

residue	*TALOS predicted ϕ/ψ*	*RCI-S^2 value*	*ANN predicted ϕ/ψ distribution*[a]	*TALOS+ predicted ϕ/ψ*	*ϕ/ψ from reference structure*[b]		
					1KXO	1N0S	1BPP
16N	-106/134	0.778	0.18/0.80/0.02	ambiguous	-140/28	-132/48	-128/36
34S	ambiguous	0.791	0.02/0.98/0.00	-121/122	-61/-29 (D)	**-127/160**	**-146/160 (N)**
37G	80/16	0.733	0.03/0.01/0.96	80/16	-65/-54 (T)	**58/25**	-99/9 (E)
39Y	-132/151	0.832	0.01/0.99/0.00	-131/149	-122/4	**-160/156**	74/-4
41K	-118/138	0.874	0.02/0.98/0.00	-121/144	**-91/167**	-121/30	**-125/167**
42c	-136/147	0.891	0.01/0.99/0.00	-134/147	64/31	**-148/94**	48/53
76V	-61/-34	0.784	0.88/0.09/0.03	-87/-25	-111/112	**-72/-37**	**-70/-40**
121K	-92/8	0.854	0.97/0.03/0.00	-93/6	-44/144	-80/171	**-104/-5**
122K	58/39	0.864	0.04/0.00/0.96	58/38	-72/32	-85/80	**53/47**

The corresponding graphical presentation of TALOS+ predictions is provided in Figure S4.

[a] The 3-state (alpha/beta/positive-ϕ) probability predicted from the neural network.

[b] The ϕ/ψ torsion angles from three different reference X-ray structures; those matching TALOS+ predicted ϕ/ψ values are in boldface; mutated residues are included in parentheses.

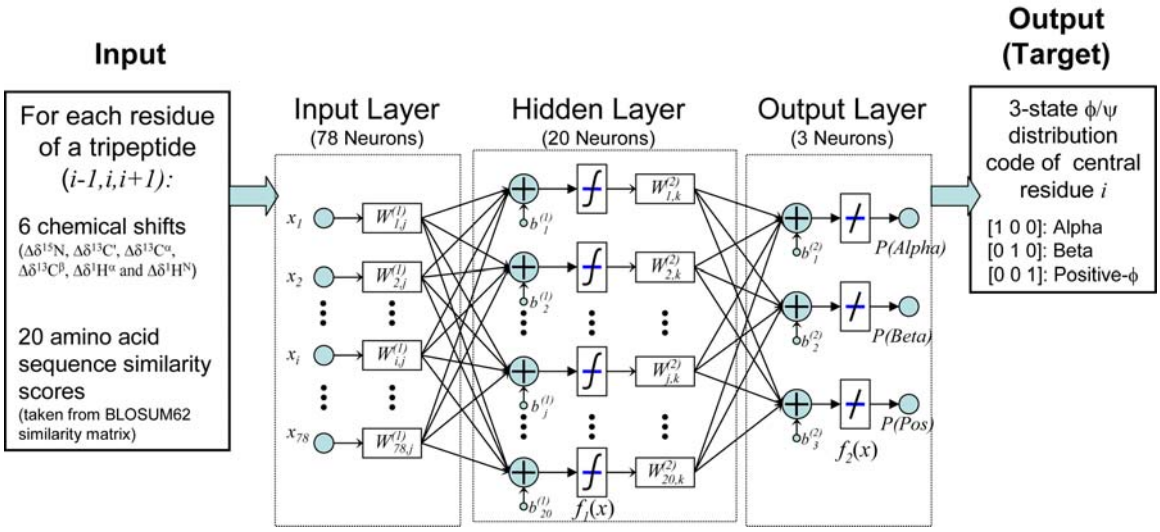

Figure S1. Architecture of the first level of the artificial neural network used in this work. The parameters x_1, .., x_{78} represent the chemical shifts and 20 BLOSUM62 coefficients describing each of the three residues of the input tripeptide. Two weight matrices $W^{(1)}$ and $W^{(2)}$, the 20-dimensional bias vector, $b^{(1)}$, and the three-dimensional bias vector $b^{(2)}$ are optimized during the ANN training process (see eq 1, main text).

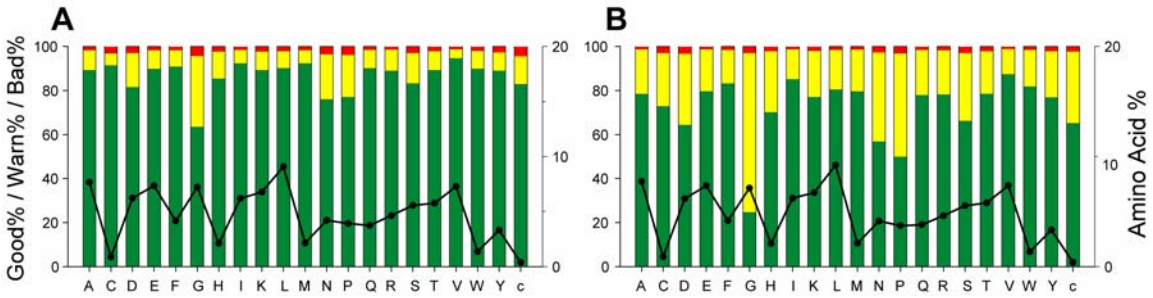

Figure S2. Residue type specific prediction results of TALOS+ (A) and original TALOS (B). The percentage of "Good", "Ambiguous" and "Bad" predictions are shown as green, yellow and red bars, respectively. Percentages of "Bad" predictions are calculated relative to the total number of predictions. The fraction of each amino acid in the dataset is also marked by the black line and dots (with the scales shown to the right).

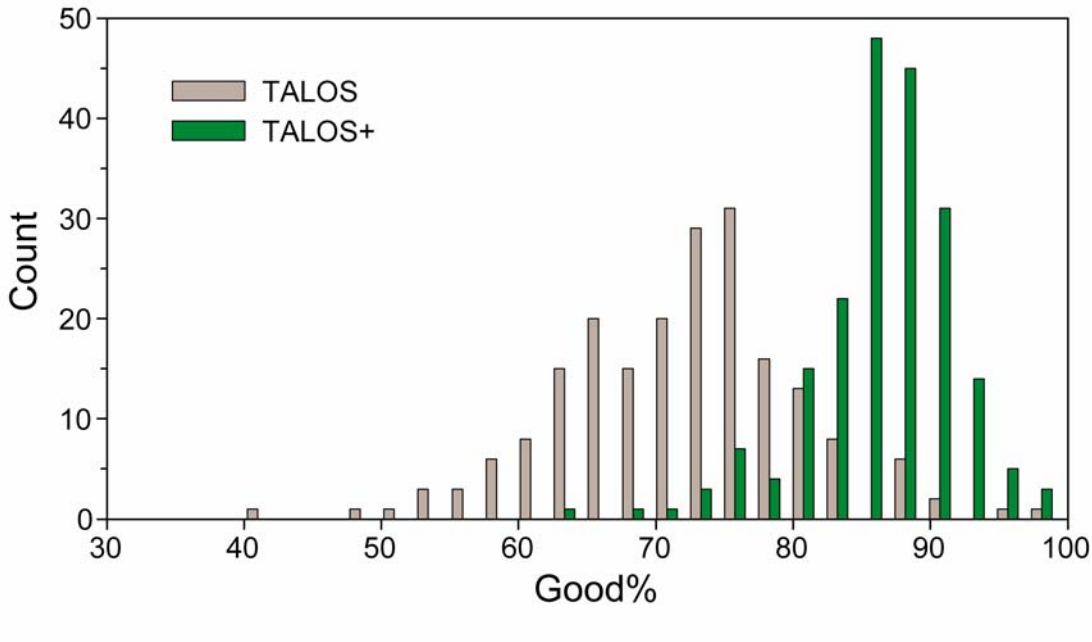

Figure S3. Histogram of the number of proteins with a given percentage of "Good" predictions for the 200 proteins in the database. Predictions by TALOS are in grey and by TALOS+ in green.

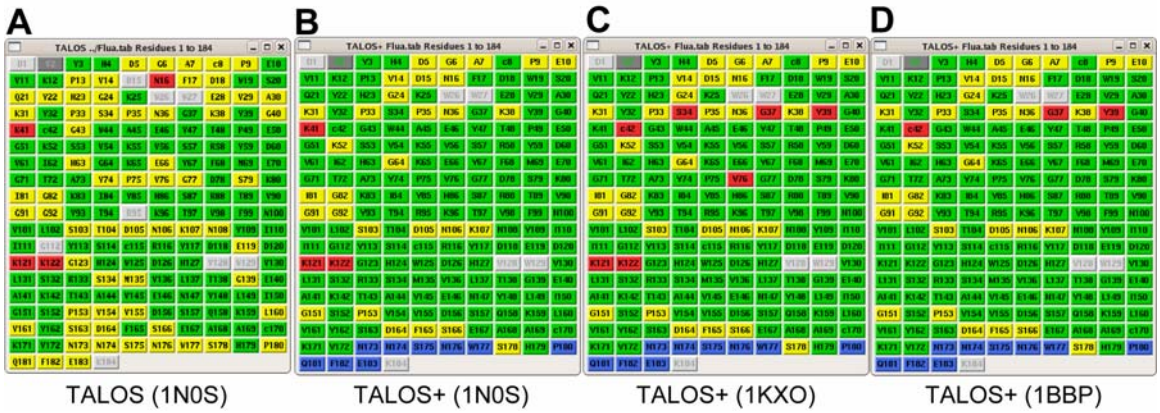

Figure S4. TALOS and TALOS+ predictions for protein FluA. (A) Graphical interface of TALOS predictions, using the 1N0S reference X-ray structure (2.0 Å resolution). The "Good" (106), "Ambiguous" (64) and "Bad" (4) predictions are shown in green, yellow and red, respectively. (B-D) TALOS+ predictions, using X-ray reference structures of (B) 1N0S, (C) 1KXO (1.8 Å) and (D)1BPP (2.0 Å); the number of consensus "Good", "Ambiguous" and "Bad" predictions are 131, 30 and 0, respectively. Note that residues 34 and 37 of 1KXO and 1BBP are mutated relative to the sequence of 1N0S.

User guide for the TALOS+ System

The TALOS+ core database search system is implemented in C++; a graphical interface to display the prediction results and apply manual manipulation of the output is implemented using a TCL script based on NMRWish, which is a companion to the NMRPipe program for multidimensional spectral processing and analysis.

There are two major scripts comprising the TALOS+ system:
1. **TALOS+** (`talos+`) : Searches the database for shift matches.
2. **RAMA** (`rama+.tcl`): Used to display and analyze the predictions.

Any of the scripts can be invoked with the "-help" command-line argument to generate a complete list of options.

The `'rama+.tcl'` script requires definitions of the following environment variables, which will usually be established automatically by the (original) TALOS or NMRPipe installation procedure:

```
TALOS_DIR    /disk1/NMRPipe/talos
TCLPATH      /disk1/NMRPipe/com
TCL_LIBRARY  /disk1/NMRPipe/nmrtxt/tcl7.6
TK_LIBRARY   /disk1/NMRPipe/nmrtxt/tk4.2
```

Other files of the TALOS+ system include:

talos+/tab/talos.tab
The compiled database of residue triplets with their corresponding secondary shifts and PHI/PSI values.

talos+/tab/randcoil.tab
The table of random coil shifts used in the prediction process.

talos+/tab/homology.tab
The residue type homology factors used in the prediction process.

talos+/tab/weight.tab
The weighting factors of the 18 secondary shifts used in the prediction process.

talos+/tab/*level*.tab
The weighting factors and biases of the neural network used in the prediction process.

talos+/src/*.*
The C++ source code of the neural network based database search and prediction process.

talos+/rama.gif
Image of the populated regions of the TALOS database, used as a background for the RAMA Ramachandran plot display.

How to Use TALOS+

Similar to TALOS, use of TALOS+ to predict phi and psi angles involves the following steps:

1. Create a directory for the prediction session; all subsequent commands will be executed from this directory.
2. Prepare the input table of shift assignments (for example "`myshifts.tab`"), according to the format given below. A UNIX shell script, “bmrb2talos.com”, is also included in the distribute package for the purpose of converting BMRB chemical shift tables to TALOS format.
3. Run TALOS+ (`talos+`) to perform the database searches. Most commonly, this will simply require a command such as:

```
talos+ -in myshifts.tab
```

During the database search, a series of files "pred/res*.tab" will be created. Each one of these files tallies the 10 best database matches for a given residue in the target protein. Before exiting, a file "`pred.tab`" will also be created, which includes an initial summary of the prediction results. Additionally, two files with default name of "`pred.ss.tab`" and "`pred.abp.tab`" will be created to store the RCI-derived order parameter (S^2) values and ANN-predicted secondary structure. The database search will typically take about 15-20 sec per 100 residues.
Unlike the original TALOS, a summarization step (originally performed by `vina.tcl`) is now part of the TALOS+ database search procedure.

4. Run RAMA (`rama+.tcl`) to inspect and adjust the predictions. The simplest invocations are:

```
rama+.tcl -in myshifts.tab
rama.tcl+ -in myshifts.tab -ref mystruct.pdb
```

During this inspection, you will:

- Examine the phi/psi distributions of the center residues of the best 10 database matches for a given query residue, and decide which ones should be included in the prediction, and which are "outliers". (**NOTE:** in most cases, the initial automated classifications performed by the current version of the TALOS+ program should be acceptable with no manual adjustment needed).
- Classify the results for a given residue as "Good", "Ambiguous", or (if a reference structure is known) "Bad".

The files "`pred/res*.tab`" will be adjusted along the way to reflect any changes made interactively, and a new "`pred.tab`" summary file will be created on exiting. When the above steps are completed, the final "`pred.tab`" file will include the classification ("Good" etc) and predictions (averages and standard deviations) for phi and psi at each residue.

Inspecting and Refining the Prediction Results

The final step in interpreting the results of the TALOS database search is to inspect and classify the matches so that useful predictions can be formed; however, in most cases, the

initial automated classifications performed by the current version of the TALOS+ program should be acceptable with no manual adjustment needed.
The refinement of predictions is done via the graphical interface RAMA. The simplest invocation of RAMA is:

```
rama+.tcl -in myshifts.tab
```

If a proposed structure is available, first run TALOS+ with it to generate a prediction summary:

```
talos+ -in myshifts.tab -ref mystruct.pdb
```

Then, invoke RAMA so that the reference structure is included in the display of prediction data:

```
rama+.tcl -in myshifts.tab -ref mystruct.pdb
```

How to Select "Good" Predictions

The standard TALOS+ rules for defining "Good" predictions are:

1. All 10 best database matches fall in a consistent Alpha, Beta or Positive-Phi region of the Ramachandran map.
2. The confidence of the ANN 3-state Phi/Psi distribution prediction for this residue must be above 0.6.
3. The predicted order parameter S^2 value must be above 0.5.

All the cases with predicted S^2 value <0.5 are likely to be "Dynamic", and will not be considered as unambiguous predictions.

All other cases are automatically classified as "Ambiguous".

Definition of "Bad" Predictions

When a reference structure is available, predictions will be flagged as "Bad" if either the condition $\{[|\phi_{obs} - \phi_{pred}|>60^o$ or $|\psi_{obs} - \psi_{pred}|>60^o]$ and $|\phi_{obs} - \phi_{pred} + \psi_{obs} - \psi_{pred}| >60^o\}$ or the condition $|\phi_{obs} - \phi_{pred}|>90^o$ or $|\psi_{obs} - \psi_{pred}|>90^o$ applies.
Cases where $|\phi_{obs} - \phi_{pred} + \psi_{obs} - \psi_{pred}| <60^o$ cause the peptide chain to continue in roughly the correct direction, and larger tolerance limits (up to $\pm 90^o$) are accepted for ϕ and ψ in these cases.