Supporting Information for "Evolution of all-atom force fields to improve local and global properties"

Gül H. Zerze,*,† Wenwei Zheng,*,‡ Robert B. Best,*,¶ and Jeetain Mittal*,†

†*Department of Chemical and Biomolecular Engineering, Bethlehem, PA 18015, USA*

‡*College of Integrative Sciences and Arts, Arizona State University, Mesa, AZ 85212, USA*

¶*Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892, United States*

E-mail: h.gul.ozer@gmail.com; wzheng38@asu.edu; robert.best2@nih.gov; jeetain@lehigh.edu

Simulation details

Enhanced sampling (parallel-tempering) molecular dynamics (MD) simulations are performed. While all ff03* and some of ff03w simulations were run using classical temperature replica-exchange MD[1] (REMD), most of ff03w and all of ff03ws simulations were run using parallel-tempering in the well-tempered ensemble (PTWTE), which is a variant of parallel-tempering where the fluctuations in potential energy are amplified to reduce number of replicas, therefore, cost of the parallel-tempering method.[2] Number of replicas are varied depending on the sampling technique and the system size to achieve at least 20% average exchange acceptance (at least 30% for PTWTE simulations).

We use the GROMACS 4.6.7 MD engine[3,4] and PLUMED 2.1 plugin[5] (for PTWTE simulations). The systems are solvated (using the related water model) in a truncated octahedron box with the size of at least the nearest periodic distance (protein) is less than short

range nonbonded interactions cutoff. Each system is energy minimized (steepest descents), and equilibrated in an NVT (T = 300 K) followed by an NPT (T = 300 K, P = 1 bar) ensemble. Pressure is equilibrated with a Berendsen barostat[6] and maintained constant with a Parrinello-Rahman barostat with a time constant of 2 ps. Systems are propagated using stochastic Langevin dynamics with a friction coefficient of 1/ps. Electrostatic interactions are calculated using the particle-mesh Ewald method[7] with a real space cutoff distance of 0.9 nm. A 1.2 nm cutoff distance is used for the van der Waals interactions.

Analysis

Intra-chain distances are calculated as the root-mean-square distance between alpha-carbon of every i_{th} residue and alpha-carbon of every j_{th} residue (where i is not equal to j). Chemical shift deviations are calculated using Sparta+ algorithm.[8] Chemical shift deviations with respect to random coil is calculated using the reference random coil chemical shift deviations obtained from Poulsen IDP server[9] for given sequences. The J-coupling constant ${}^3J_{HNH\alpha}$, is calculated following the Karplus equation: ${}^3J_{HNH\alpha} = Acos^2(\phi-60)+Bcos(\phi-60)+C$, where the ϕ angle is given in units of degree and the parameter set (A, B, and C) is taken from Vögeli et al.[10] ("ensemble" parameter set). χ^2 analysis is performed using the uncertainties (RMSD) reported for ensemble parameter set in reproducing observed ${}^3J_{HNH\alpha}$.[10]

References

(1) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

(2) Bonomi, M.; Parrinello, M. *Phys. Rev. Lett.* **2010**, *104*, 190601.

(3) Berendsen, H. J.; van der Spoel, D.; van Drunen, R. *Comput. Phys. Commun.* **1995**, *91*, 43–56.

(4) Hess, B.; Kutzner, C.; Van Der Spoel, D.; Lindahl, E. *J. Chem. Theory Comput.* **2008**, *4*, 435–447.

(5) Bonomi, M.; Branduardi, D.; Bussi, G.; Camilloni, C.; Provasi, D.; Raiteri, P.; Donadio, D.; Marinelli, F.; Pietrucci, F.; Broglia, R.; Parrinello, M. *Comput. Phys. Commun.* **2009**, *180*, 1961–1972.

(6) Berendsen, H. J.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. *J. Chem. Phys.* **1984**, *81*, 3684–3690.

(7) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys.* **1995**, *103*, 8577–8593.

(8) Shen, Y.; Bax, A. *J. Biomol. NMR* **2010**, *48*, 13–22.

(9) Kjaergaard, M.; Poulsen, F. M. *J. Biomol. NMR* **2011**, *50*, 157–165.

(10) Vögeli, B.; Ying, J.; Grishaev, A.; Bax, A. *J. Am. Chem. Soc.* **2007**, *129*, 9377–9385.

Tables

Table S1: Summary of simulations performed using ff03*, where N is number of residues in given protein or peptide. Details of REMD/PTWTE is given in the last three column where applicable.

Protein	N	Enhanced Sampling	Total simulation time* (ns/replica)	Simulation time analyzed* (ns/replica)	T-range (K)	# of replicas	Average acceptance (%)
$C(AGQ)_2W$	8	None	4800	4800			
$tau_{174-183}$	10	REMD	100	80	300-550	24	18
$C(AGQ)_3W$	11	None	7900	7900			
$C(AGQ)_4W$	14	None	7800	7800			
$C(AGQ)_5W$	17	None	7700	7700			
$C(AGQ)_6W$	20	None	4400	4400			
MFP	25	REMD	225	150	300-575	40	28
hIAPP	37	REMD	200	150	300-575	40	19
TDP-$43_{310-350}$	41	PTWTE	150	100	300-510	10	24
CSP	66	REMD	154	110	285-493	56	20
α-syn	140	None	824	450			

*Simulation time is in ns for no enhanced sampling cases. Pre-equilibration is done with a 50 ns PTWTE simulations for no enhanced sampling cases.

Table S2: Summary of simulations performed using ff03w, where N is number of residues in given protein or peptide. Details of REMD/PTWTE is given in the last three column where applicable.

Protein	N	Enhanced Sampling	Total simulation time* (ns/replica)	Simulation time analyzed* (ns/replica)	T-range (K)	# of replicas	Average acceptance (%)
$C(AGQ)_2W$	8	None	4800	4800			
$C(AGQ)_3W$	11	None	7600	7600			
$C(AGQ)_4W$	14	None	7500	7500			
A_{16}	16	REMD	200	150	270-481	36	19
G_{16}	16	REMD	200	150	270-481	36	19
$C(AGQ)_5W$	17	None	7500	7500			
$C(AGQ)_6W$	20	None	4200	4200			
G_{32}	32	REMD	200	150	270-480	64	18
CGRP F37W	37	PTWTE	200	150	300-518	16	35
hIAPP-noNH_2	37	REMD	200	150	300-575	40	19
hIAPP	37	REMD	300	200	300-575	40	19
hIAPP G24P	37	REMD	200	150	300-575	40	19
hIAPP I26P	37	REMD	200	150	300-575	40	19
hIAPP G24P/I26P	37	REMD	200	150	300-575	40	19
hIAPP S20G	37	REMD	200	150	300-575	40	19
hIAPP Y37W	37	PTWTE	200	150	300-575	40	19
rIAPP	37	REMD	200	150	300-575	40	19
GG-pHLIP	37	PTWTE	145	95	300-518	16	55
AA-pHLIP	37	PTWTE	100	50	300-518	16	55
TDP-$43_{310-350}$	41	PTWTE	200	150	300-518	16	35
TDP-$43_{310-350}$ A321V	41	PTWTE	187	137	300-518	16	35
TDP-$43_{310-350}$ A326P	41	PTWTE	187	137	300-518	16	35
$FUS_{120-163}$	44	PTWTE	260	210	300-518	16	35
PROT-C	55	REMD	200	150	285-493	56	20
PROT-N	56	REMD	300	200	285-493	56	20
N-terminal-HIV-In	57	REMD	221	170	285-493	56	23
N-terminal-HIV-In-dyes	57	REMD	200	150	285-493	56	23
CSP	66	REMD	300	200	285-493	56	20
CSP-dyes	68	REMD	200	150	285-493	56	20
LR	82	REMD	250	200	285-493	56	16
LR-dyes	82	REMD	200	150	285-493	56	16

*Simulation time is in ns for no enhanced sampling cases. Pre-equilibration is done with a 50 ns PTWTE simulations for no enhanced sampling cases.

Table S3: Summary of simulations performed using ff03ws, where N is number of residues in given protein or peptide. Details of REMD/PTWTE is given in the last three column where applicable.

Protein	N	Enhanced Sampling	Total simulation time* (ns/replica)	Simulation time analyzed* (ns/replica)	T-range (K)	# of replicas	Average acceptance (%)
$C(AGQ)_2W$	8	None	4800	4800			
$C(AGQ)_3W$	11	None	7600	7600			
$C(AGQ)_4W$	14	None	7500	7500			
GB1	16	REMD	500	unfolded only	278-595	32	20
A_{16}	16	REMD	100	80	270-481	36	19
G_{16}	16	REMD	100	80	270-481	36	19
$C(AGQ)_5W$	17	None	7400	7400			
$C(AGQ)_6W$	20	None	9600	9600			
TrpCage	20	REMD	150	unfolded only	285-455	32	20
MFP	25	PTWTE	200	150	300-500	10	20
MFP-Dopa	25	PTWTE	200	150	300-500	10	29
BBA	28	PTWTE	200	150	300-518	16	55
CGRP F37W	37	PTWTE	250	200	300-518	16	40
hIAPP	37	PTWTE	200	150	300-518	16	40
hIAPP Y37W	37	PTWTE	200	150	300-518	16	40
Aβ-40	40	PTWTE	200	150	277-525	18	35
TDP-$43_{310-350}$	41	PTWTE	200	150	300-518	16	35
TDP-$43_{310-350}$ A321G	41	PTWTE	100	50	300-518	16	35
TDP-$43_{310-350}$ G335A	41	PTWTE	200	150	300-518	16	35
TDP-$43_{310-350}$ G335D	41	PTWTE	200	150	300-518	16	35
Aβ-42	42	PTWTE	200	200	277-525	18	35
$FUS_{120-163}$	44	PTWTE	200	150	300-518	16	35
$FUS_{120-163}$-S mutant (all Y)	44	PTWTE	200	150	300-518	16	35
BBL	47	PTWTE	123	73	300-518	16	35
NTL9	52	PTWTE	100	50	300-518	16	30
ProteinG	56	PTWTE	200	150	300-518	16	26
GHM-TDP-$43_{307-360}$	57	PTWTE	200	150	300-518	16	28
CSP	66	PTWTE	232	182	300-518	16	25
UBQ	76	PTWTE	100	100	300-500	16	21
ACTR	79	None	2000	1800			
$KNL1_{1-80}$	80	PTWTE	130	80	300-518	16	18
Barstar	98	PTWTE	225	175	300-518	16	17
α-syn	140	None	386	150			

*Simulation time is in ns for no enhanced sampling cases. Pre-equilibration is done with a 50 ns PTWTE simulations for no enhanced sampling cases.

Table S4: Summary of simulations performed using ff03ws under denaturing condition, where N is number of residues in given protein or peptide.

Protein	N	Enhanced Sampling	Total simulation time* (ns/replica)	Simulation time analyzed* (ns/replica)	[Urea] [M]	[GdmCl] [M]
$C(AGQ)_2W$	8	None	4500	4000	8	6
$C(AGQ)_3W$	11	None	4500	4000	8	6
$C(AGQ)_4W$	14	None	4500	4000	8	6
$C(AGQ)_5W$	17	None	4500	4000	8	6
$C(AGQ)_6W$	20	None	4500	4000	8	6
TrpCage	20	REMD	50	50	3	
CSP M34-dyes	34	REMD	450	400	8	6

*Simulation time is in ns for no enhanced sampling cases.

Figures

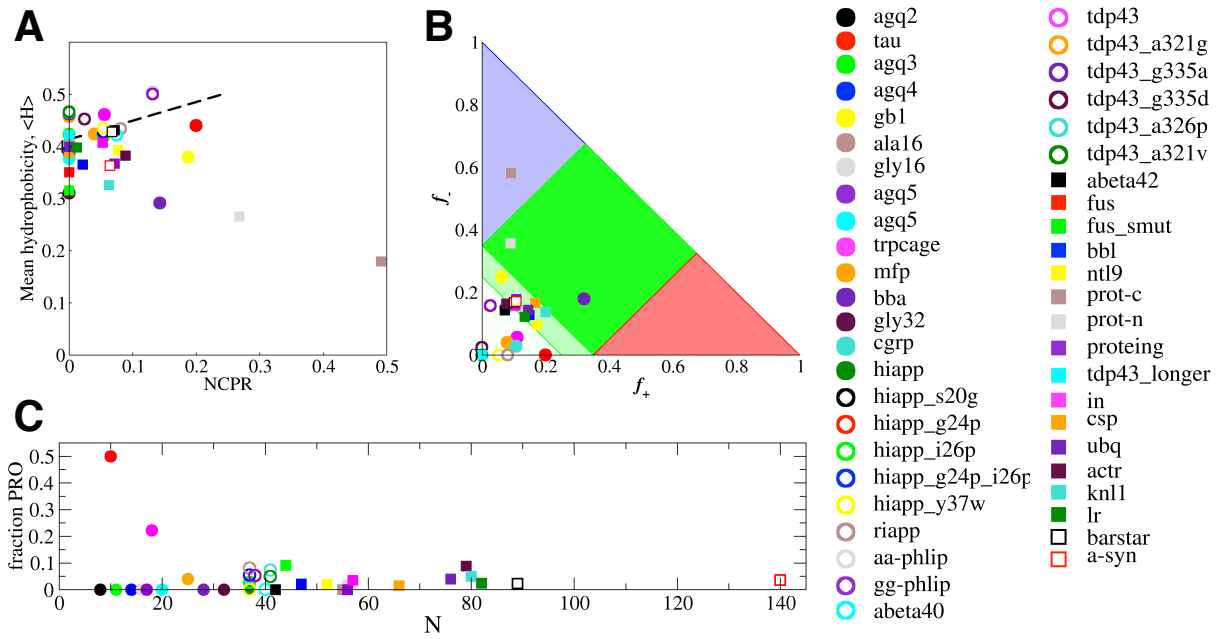

Figure S1: Sequence characteristics of the proteins and peptides studied in this work. A. Mean hydrophobicity with respect to net charge per residue (NPCR), B. Fractions of negatively and positively charged residues, C. Fraction of proline residues in the primary structures of peptides.

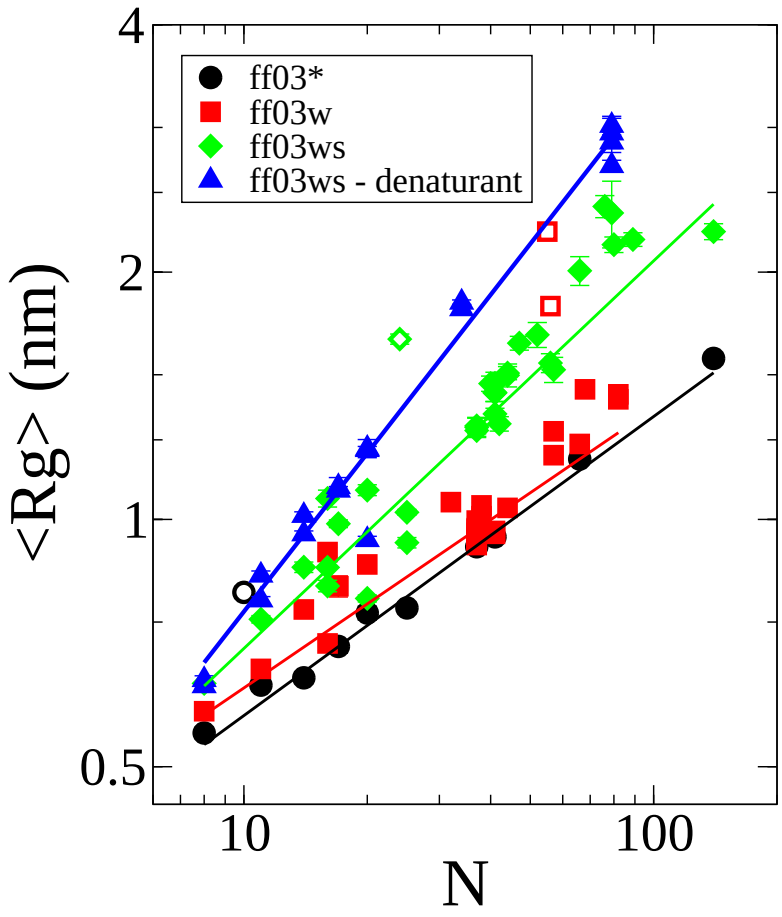

Figure S2: Radius of gyration, R_g, with respect to number of monomers, N. Filled symbols denote the ensemble averaged values used in power law fits (solid lines). Empty symbols are identified as outliers and excluded from fits. List of outlier data points: $\text{tau}_{174-183}$ due to high proline content (50% of the sequence), PROT-C and PROT-N due to high net charge per residue (NCPR>25% of the sequence).

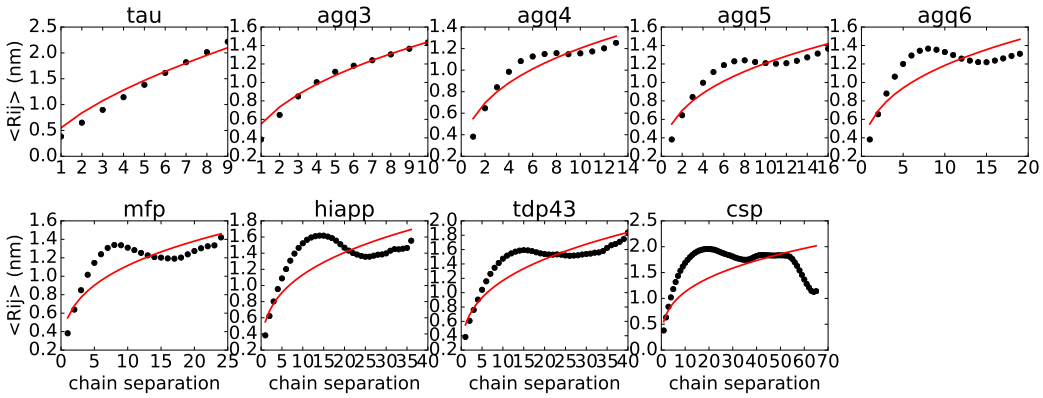

Figure S3: Intrachain distances as a function of sequence separation —i-j— (black: simulation data, red: power law fits to predict the scaling exponent) for ff03*.

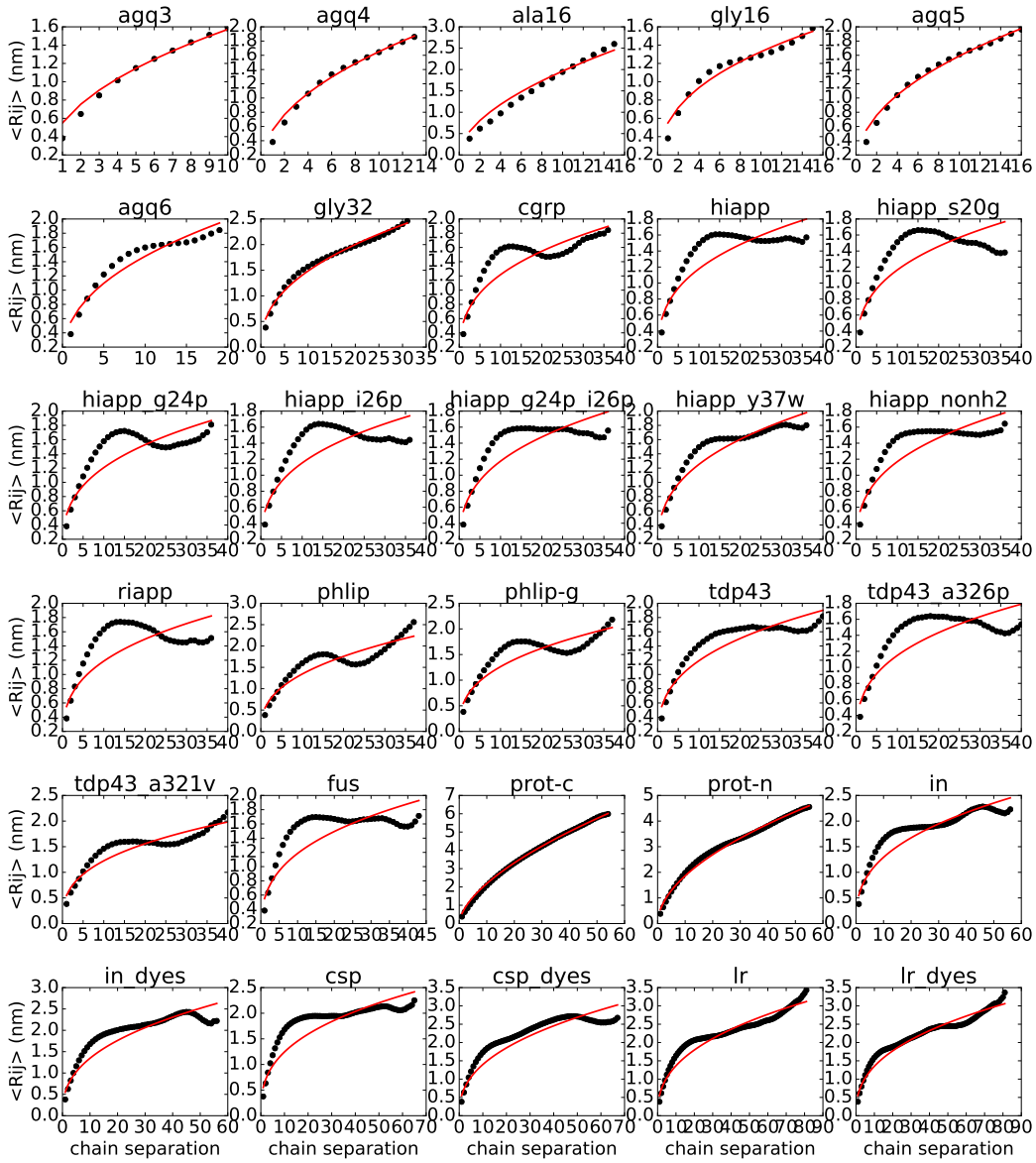

Figure S4: Intrachain distances as a function of sequence separation —i-j— (black: simulation data, red: power law fits to predict the scaling exponent) for ff03w.

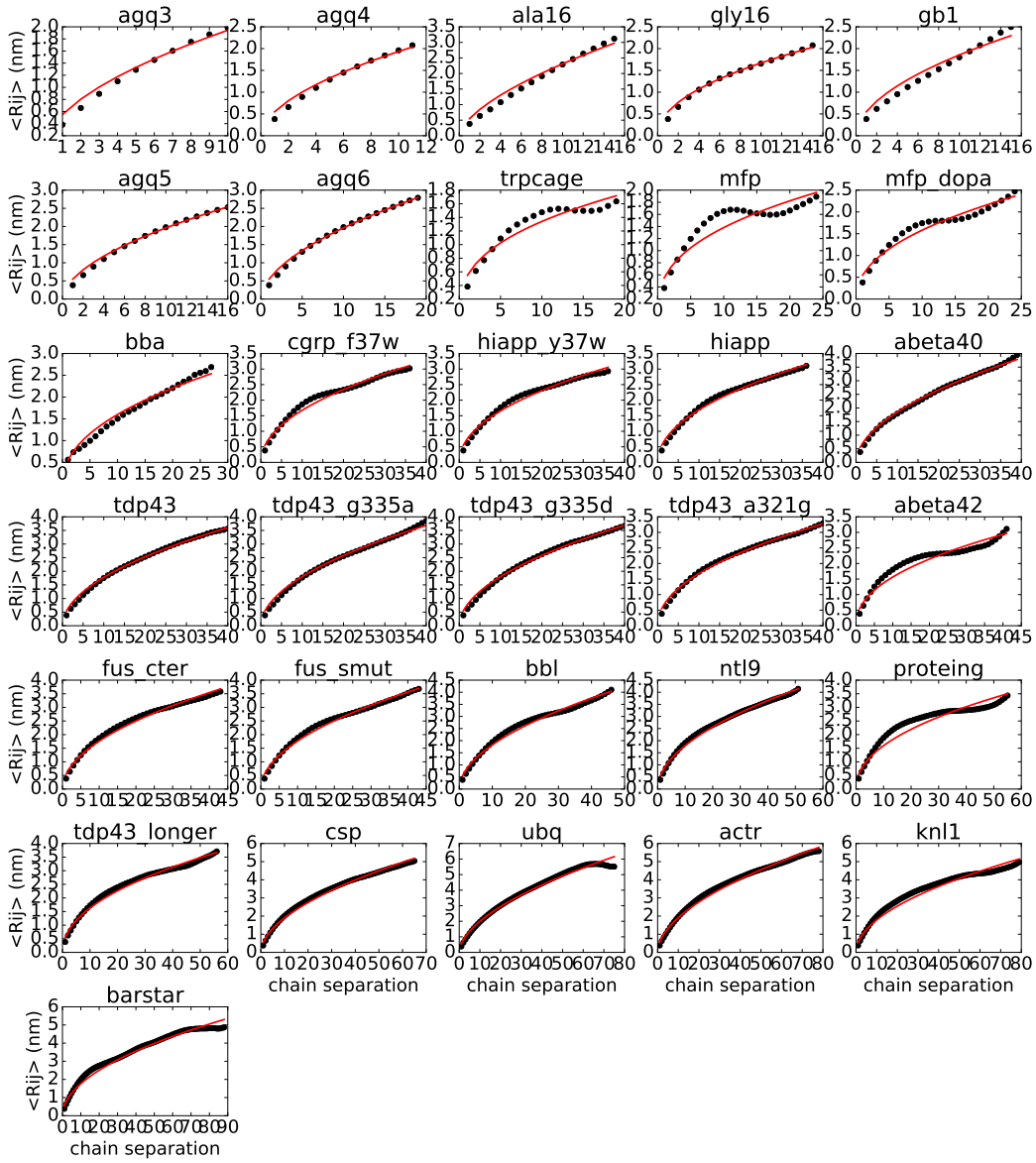

Figure S5: Intrachain distances as a function of sequence separation —i-j— (black: simulation data, red: power law fits to predict the scaling exponent) for ff03ws.

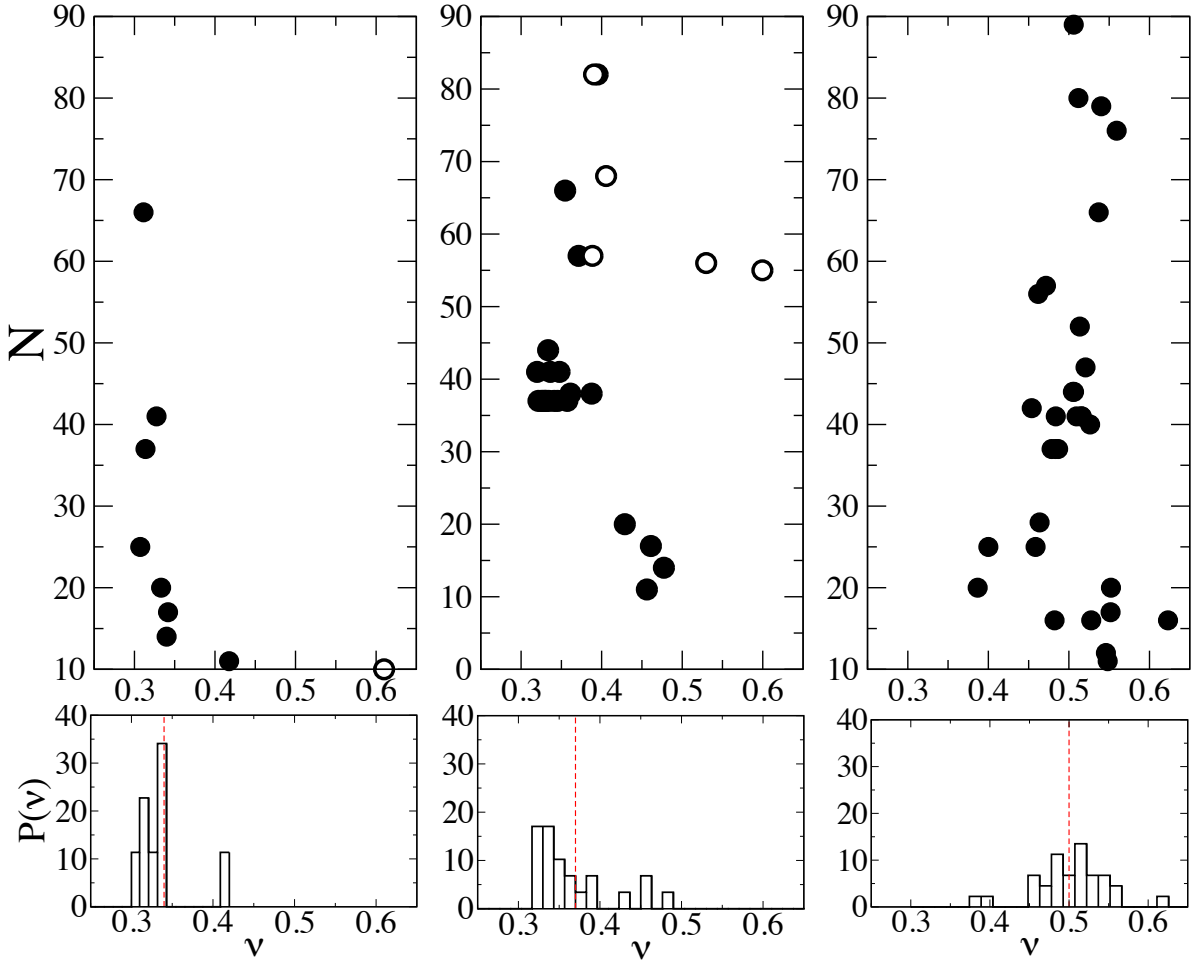

Figure S6: Intra-chain distance scaling characteristics. Bottom panels show the distributions of the scaling exponents, ν, and top panels show the each contributing data point (filled symbols) as function of peptide length, N. Empty symbols represent the proteins identified as outliers (Figure S2) and also excluded from the distributions of ν. Averages are denoted by broken red lines in bottom panels. Left, middle and right columns respectively shows the data for ff03*, ff03w and ff03ws.

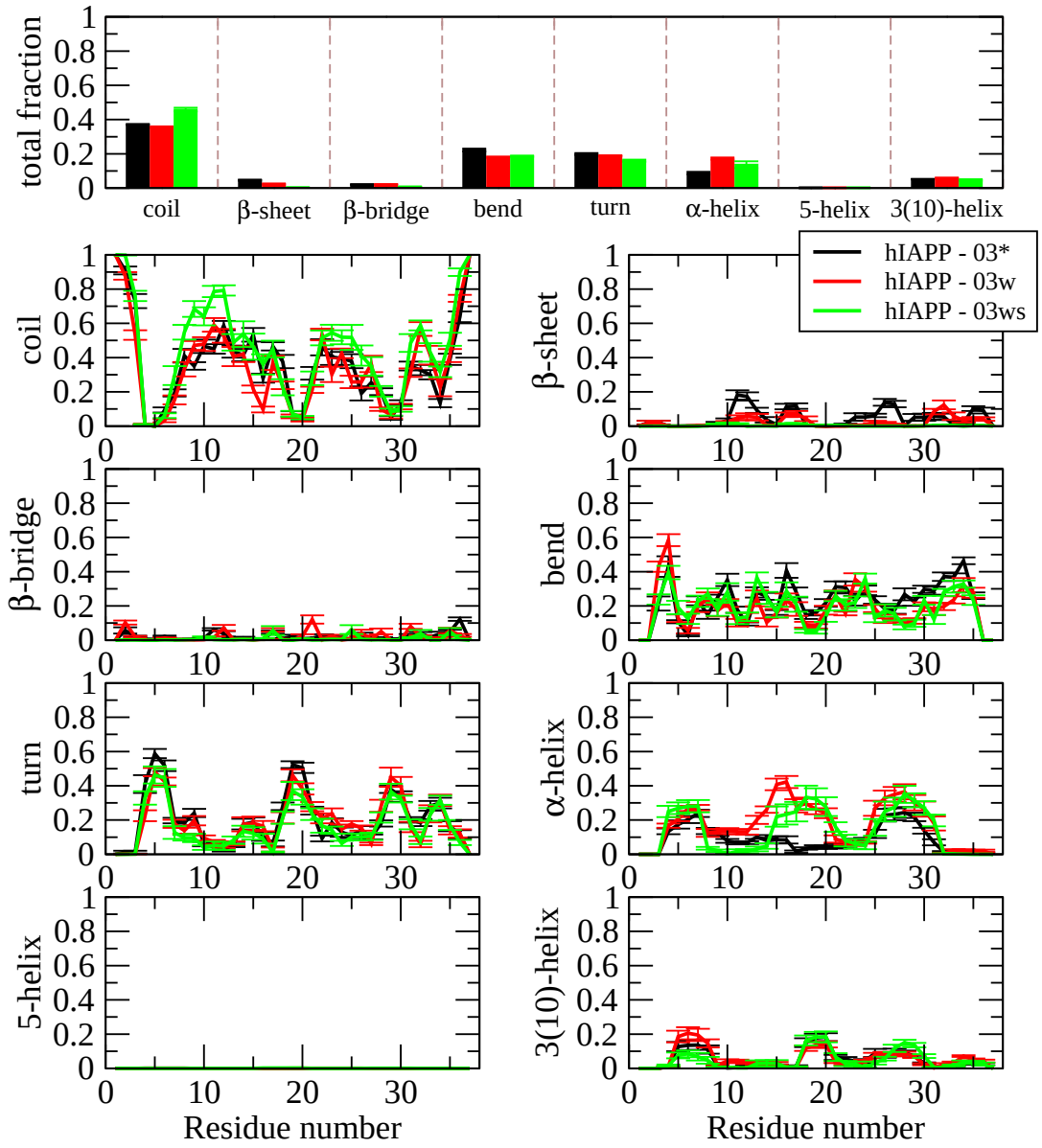

Figure S7: DSSP based secondary propensities for IAPP, for all eight DSSP types of structures.

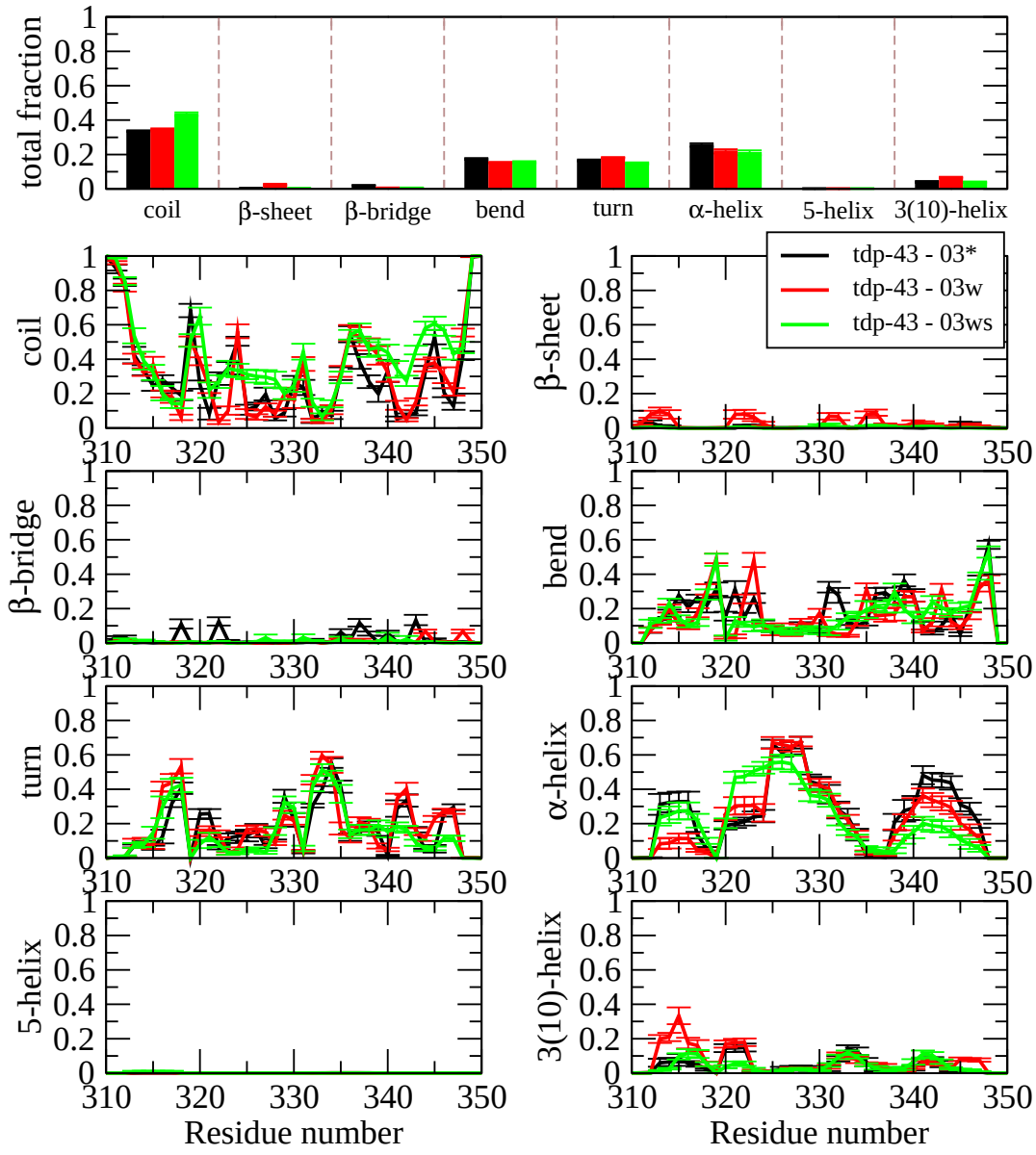

Figure S8: DSSP based secondary propensities for TDP-43$_{310-350}$, for all eight DSSP types of structures.

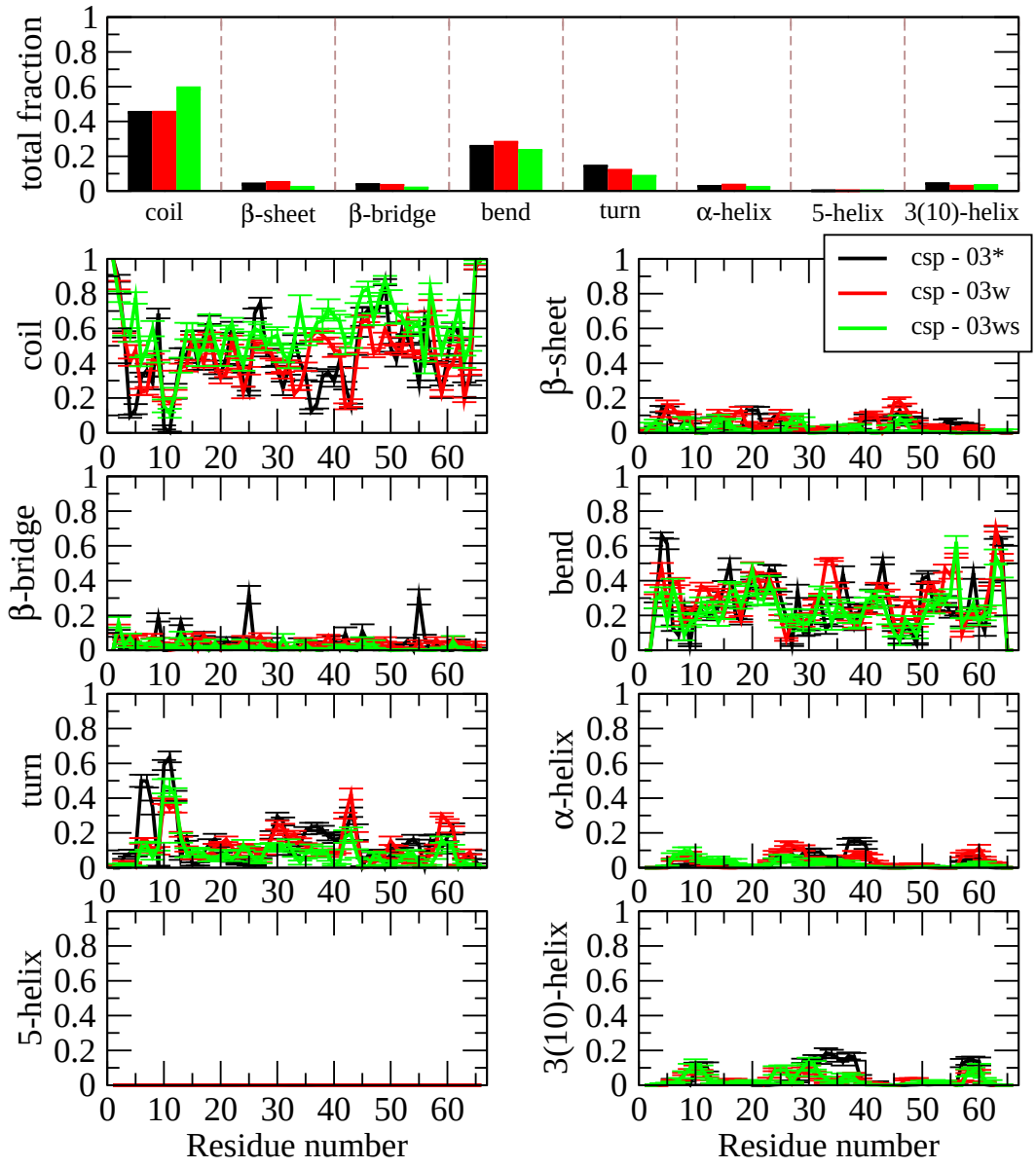

Figure S9: DSSP based secondary propensities for CSP, for all eight DSSP types of structures.

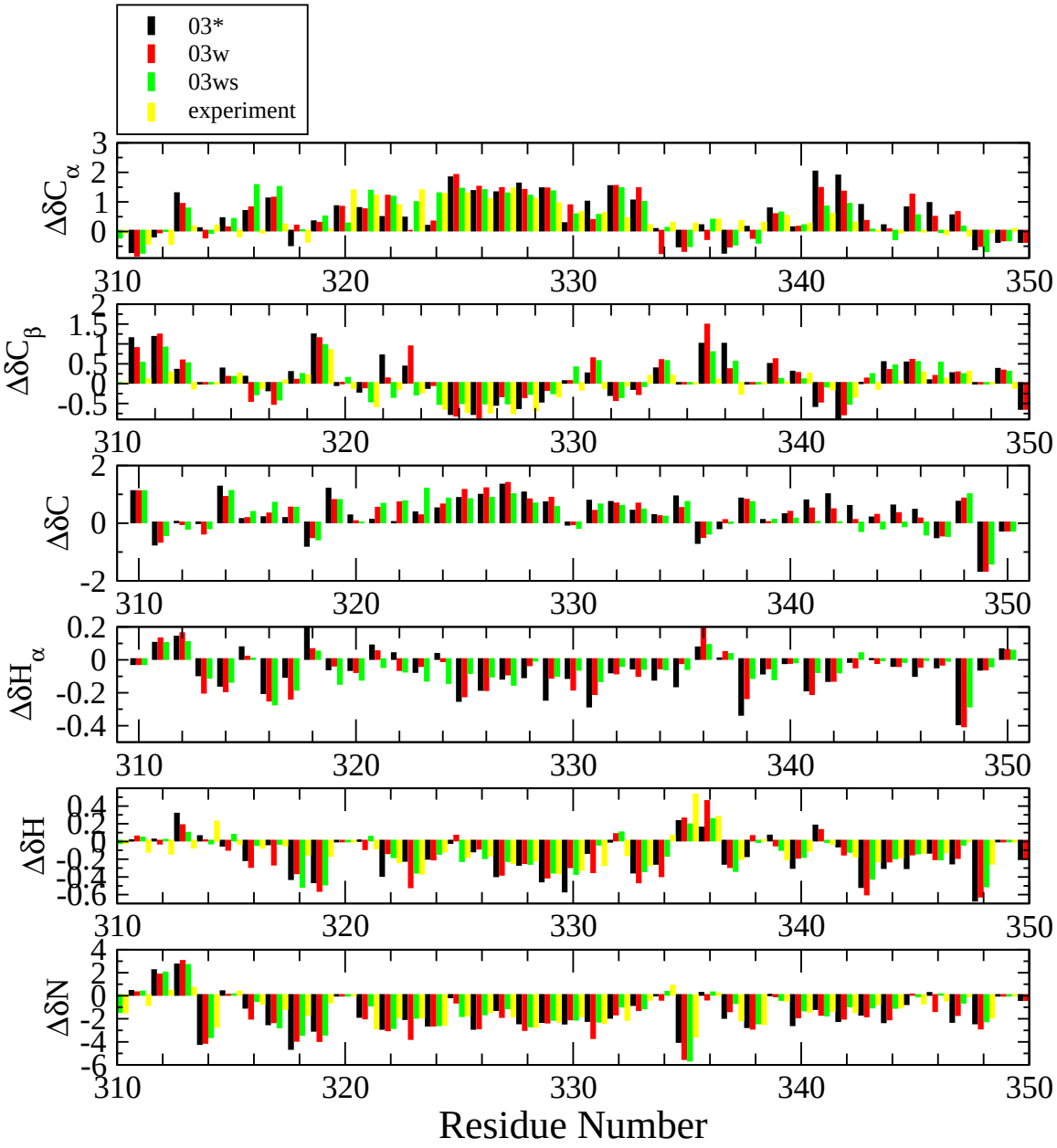

Figure S10: All calculated chemical shift deviations with respect to random coil for TDP-$43_{310-350}$. Same random coil reference is used both for simulation and experimental data as described in supporting text.

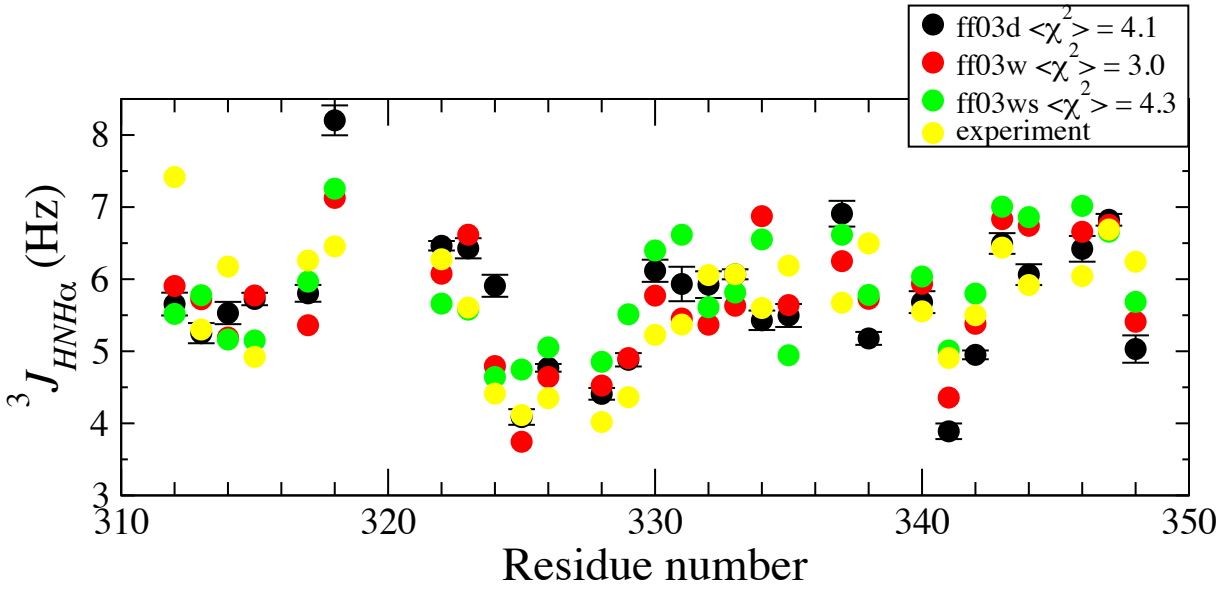

Figure S11: Scalar coupling constant ${}^3J_{HNH\alpha}$ for each force field.