Supporting Information: AWSEM-IDP: A Coarse-Grained Force Field for Intrinsically Disordered Proteins

Hao Wu,[†] Peter G. Wolynes,[‡] and Garegin A. Papoian[*,†,¶]

†*Biophysics Program, Institute for Physical Science and Technology, University of Maryland, College Park, Maryland 20742, United States*

‡*Departments of Chemistry and Physics and Center for Theoretical Biological Physics, Rice University, Houston, Texas 77005, United States*

¶*Department of Chemistry and Biochemistry, University of Maryland, College Park, Maryland 20742, United States*

E-mail: gpapoian@umd.edu

Phone: +1 (301) 405-8867

Parametrization Procedure

In the main text we discussed the general steps to calibrate the parameters in AWSEM-IDP. Here we elaborate the detailed procedure and take H4tail and PaaA2 as examples. It is apparent that V_{R_g} is system-dependent because the residue number (N) and target radius of gyration (R_g^0) can be totally different among various systems. Hence we performed parameter calibration for V_{Hbond} and V_{FM} first. In V_{Hbond} the scaling factors of V_β, V_{P-AP} and $V_{helical}$ (λ_β, λ_{P-AP} and $\lambda_{helical}$) were modified to control the overall secondary structure propensity. It turned out that the default values of λ_β and λ_{P-AP} (both 1.0) matches the benchmark β structure level well. And $\lambda_{helical}$ was reduced to 1.2 to fit the benchmark α-helical structure level. In V_{FM}, we tuned the scaling factor (λ_{FM}) and cutoff range of ij separation along the sequence ($|i-j|_{min}$ and $|i-j|_{max}$), which set the relative intensity of V_{FM} and the range that i and j go over in the calculation of r_{ij} and r_{ij}^m. In terms of memory selection, we chose 100 snapshots from a ∼85000 frame replica-exchange atomistic simulation trajectory[1] for H4tail and all the 50 structures generated with SAXS and NMR ensemble restrictions[2] for PaaA2. Noticing that λ_{FM} also depends on the number of fragment memories, we used 0.001 for H4tail and 0.002 for PaaA2 to keep their relative weights the same. As for V_{R_g}, the parameters generally vary among different systems. The range of parameters D, α and β used in this study are described in the main text (Table 1). γ is tuned to compensate the over-compact effect from other terms in AWSEM-IDP and normally ranges from 1.1 to 1.2. The rest two parameters N and R_g^0 depend on the residue number and target R_g value. For the two IDPs we studied in this report, $D = -0.2$, $\alpha = 0.001$, $\beta = 0.003$, $\gamma = 1.16$, $N = 26$, $R_g^0 = 8.6$ for H4tail, and $D = -0.8$, $\alpha = 0.001$, $\beta = 0.0005$, $\gamma = 1.11$, $N = 71$, $R_g^0 = 20.8$ for PaaA2. We tested multiple sets of parameters on both systems and compared results with atomistic simulation or experiments concerning various structural properties. With the current set of parameters we can obtain results comparable with atomistic simulation and experiments.

RMSIP Analysis

The convergence of all the simulations is confirmed by the root mean square inner product (RMSIP) analysis,[3] which quantifies the overlap between essential subspaces with inner product of first ten principal eigenvectors of CA atom coordinates.

$$\text{RMSIP} = \Big(\frac{1}{10}\sum_{i=1}^{10}\sum_{j=1}^{10}(\boldsymbol{\eta_i}\cdot\boldsymbol{\nu_j})^2\Big)^{1/2} \tag{S1}$$

To calculate RMSIP we selected subpart with increasing time length from the whole trajectory, divided each subpart into two halves and calculated RMSIP of these pairs. Figure S1 shows that an early convergence appears even in the beginning of the simulation, with RMSIP around 0.7. Then the RMSIP curves gradually increase and become saturated at around 0.8. These results are strong proof of convergence of both H4tail and PaaA2 simulations.

Energy Analysis

$$\langle E_{secondary} \rangle = \frac{1}{N} \langle V_{Rama} + V_{Hbond} \rangle , \langle E_{tertiary} \rangle = \frac{1}{N} \langle V_{contact} + V_{burial} \rangle \tag{S2}$$

In the Results and Discussion section, we analyzed different energy terms in AWSEM-IDP Hamiltonian responsible for the protein secondary and tertiary structures. Here we provide more data on detailed contribution from each energy term (Table S1), as well as the time-evolution of $E_{secondary}$ and $E_{tertiary}$ (Figure S6). For each protein, ten separate simulation runs were performed. After cutting off the first 10 ns, those simulation runs are combined for analysis.

Table S1: Detailed AWSEM Hamiltonian for all the simulated proteins

	H4 tail		**PaaA2**		**1UZC**		**1R69**		**1UBQ**	
Residue #	26		71		69		63		76	
$\boldsymbol{E_{con}}$	30.91	± 4.32	84.33	± 7.1	81.88	± 7.02	74.76	± 6.68	90.26	± 7.37
$\boldsymbol{E_{chain}}$	15.19	± 2.96	48.64	± 5.11	44.93	± 4.81	39.87	± 4.56	49.28	± 5.03
$\boldsymbol{E_{\chi}}$	2.63	± 1.09	10.96	± 2.18	10.86	± 2.19	9.13	± 2.01	12.42	± 2.42
$\boldsymbol{E_{excl}}$	1.34	± 1.10	4.10	± 1.80	4.32	± 1.82	4.24	± 1.82	6.3	± 2.25
$\boldsymbol{E_{rama}}$	-32.47	± 3.13	-159.74	± 5.19	-157.65	± 4.29	-129.87	± 4.07	-161.04	± 3.95
$\boldsymbol{E_{contact}}$	-5.31	± 2.00	-17.02	± 4.76	-18.78	± 4.36	-35.34	± 3.95	-64.29	± 4.62
$\boldsymbol{E_{burial}}$	-22.86	± 0.62	-61.43	± 0.98	-58.12	± 1.24	-53.21	± 1.28	-67.03	± 1.00
$\boldsymbol{E_{\beta}}$	-3.34	± 4.27	-0.82	± 1.51	-0.73	± 1.82	-0.03	± 0.36	-22.17	± 4.00
$\boldsymbol{E_{P-AP}}$	-12.32	± 6.20	-5.9	± 3.49	-6.31	± 3.71	-12.76	± 3.18	-18.73	± 1.89
$\boldsymbol{E_{helix}}$	-0.41	± 1.08	-41.84	± 5.81	-66.9	± 6.63	-49.9	± 5.62	-14.33	± 2.26
$\boldsymbol{E_{FM}}$	-13.93	± 0.84	-118.41	± 3.21	-171.34	± 3.36	-136.1	± 2.89	-295.8	± 3.61
$\boldsymbol{E_{R_g}}$	-20.54	± 0.27	-55.93	± 0.78	-54.63	± 0.61	-50.38	± 0.04	-60.8	± 0.00

[a] Energies in kcal/mol.
[b] All simulations performed at 300 K.

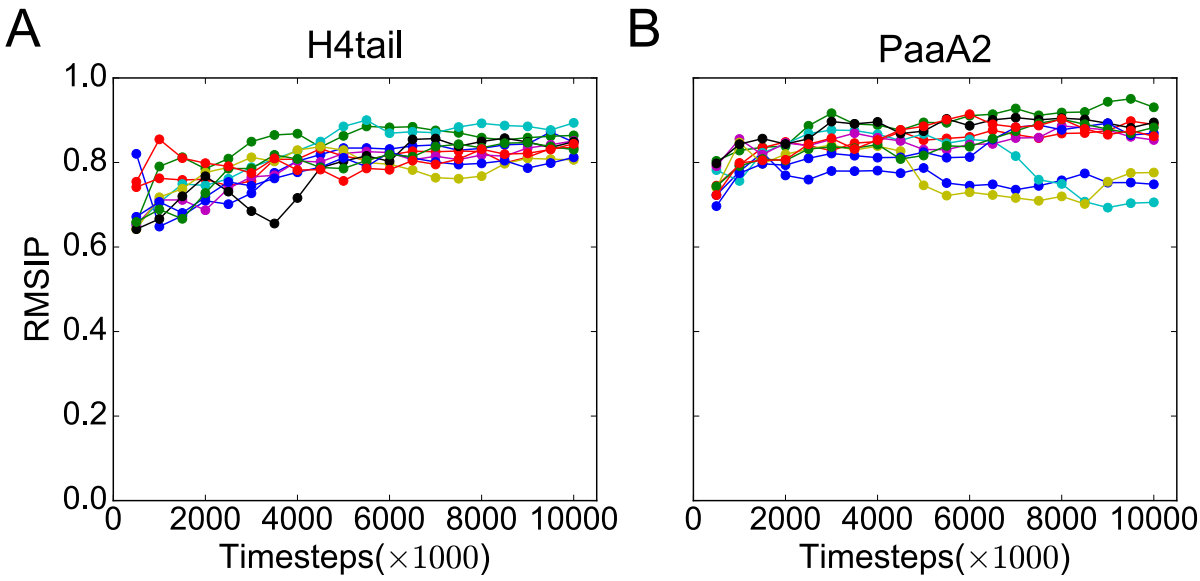

Figure S1: **RMSIP analysis demonstrates convergence of AWSEM-IDP simulations.** RMSIP curves of both H4tail (A) and PaaA2 (B) rises steadily with increasing time length and all the RMSIP values above 0.6, showing all the simulations are converged. Results of all the 10 runs are labeled with different colors.

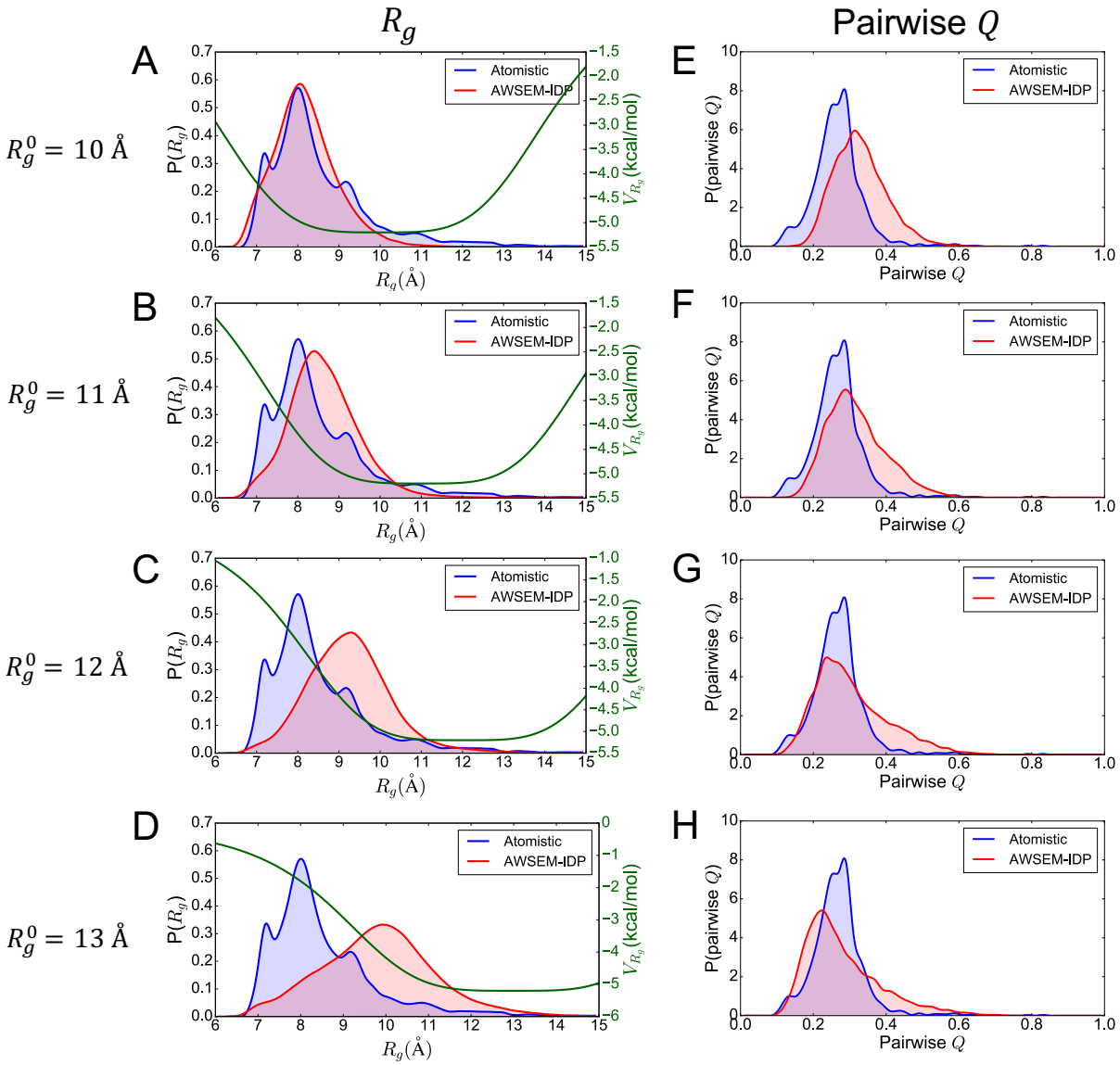

Figure S2: **A wide range of conformations can be sampled via parameter tuning of V_{R_g}.** We simulated H4 tail with AWSEM-IDP with different $R_g^0 = 10, 11, 12, 13$ in the R_g potential and calculated the corresponding R_g (A-D) and pairwise q (E-H) distributions. When R_g^0 increases, we can observe a higher average value and wider distribution of R_g, accompanied with a smaller average value of pairwise q. All the rest parameters in V_{R_g} are the same. The corresponding V_{R_g} curves are shown in green solid lines (A-D).

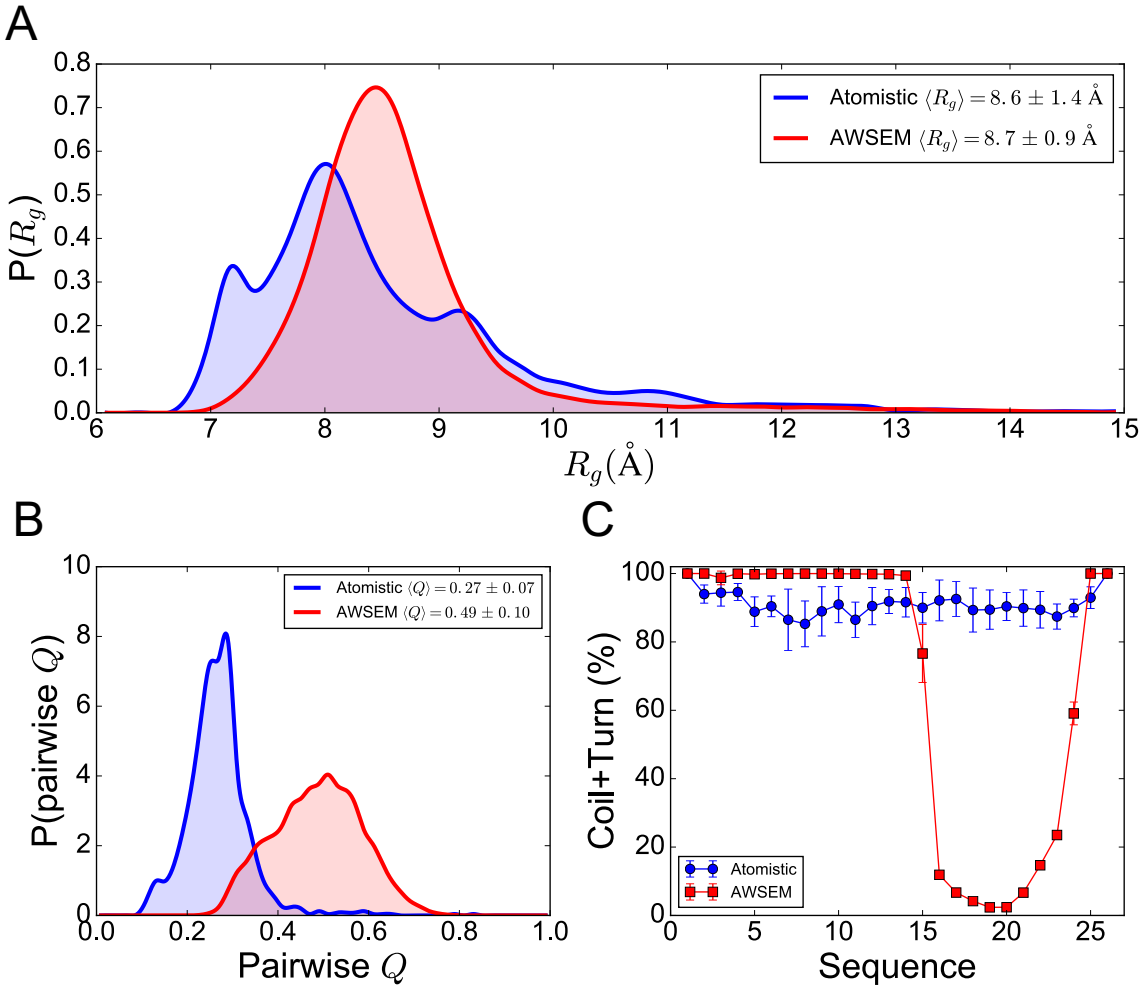

Figure S3: **Simulations with the standard AWSEM are less accurate in describing the structure of H4 tail.** The structural metrics and analyzing approaches are the same as in the main text (Figure 4).

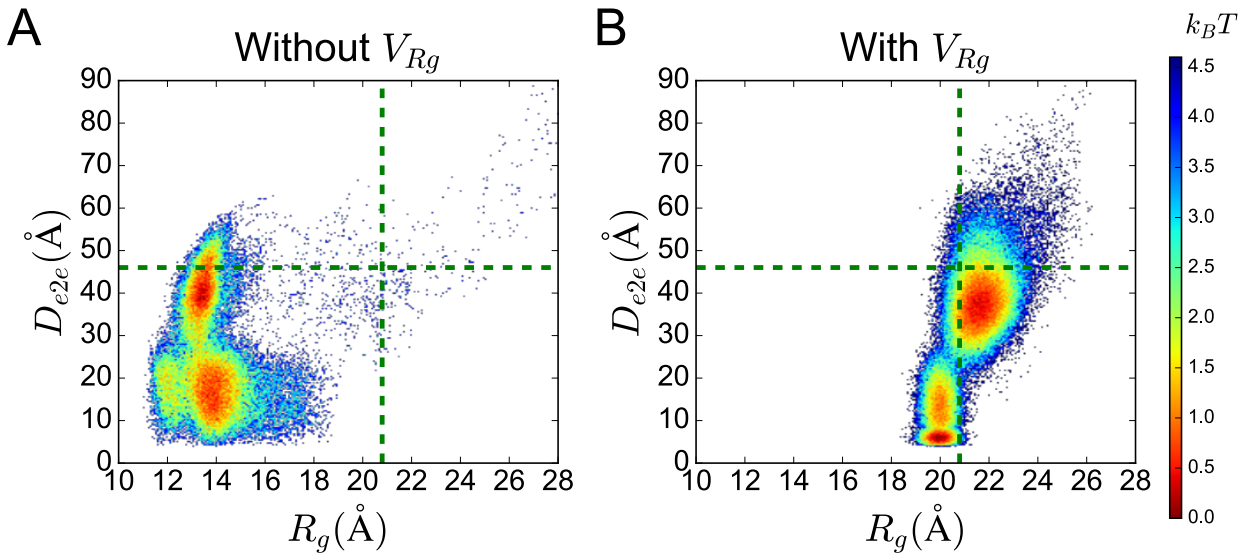

Figure S4: V_{R_g} **efficiently prevents artificially collapsed conformations of PaaA2.** The effect of V_{R_g} is highlighted by the comparison between free energy landscape with R_g and D_{e2e} as reaction coordinates simulated with (A) and without (B) V_{R_g}. After V_{R_g} with proper parameters is applied, the locations of the major energy minimums shift closer to the NMR average values[2] (green dotted lines).

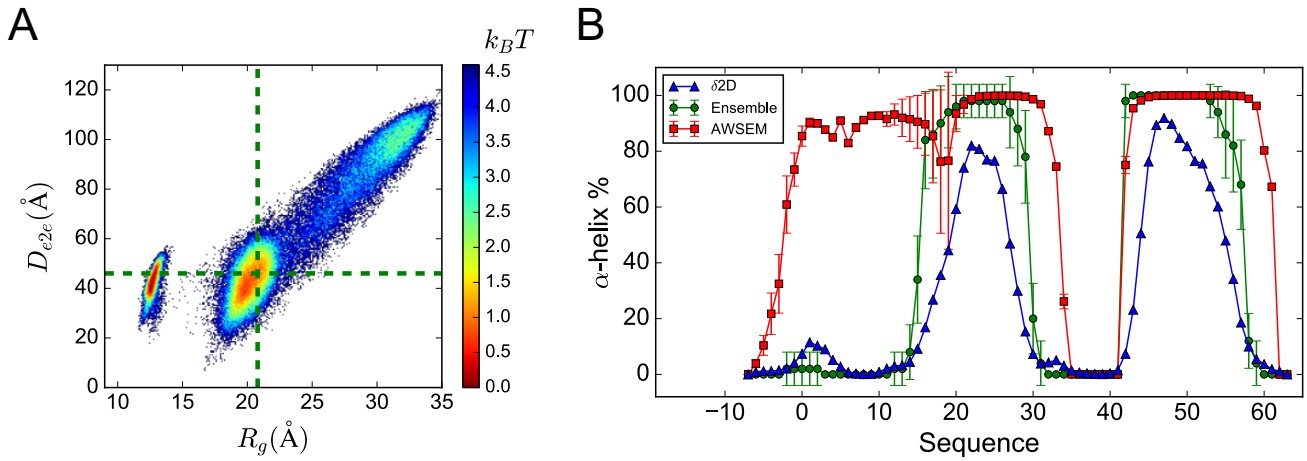

Figure S5: **Standard AWSEM simulations are less accurate in describing the structure of PaaA2.** (A) Two of the three free energy minima in simulations are distant from NMR average values (green dotted lines). (B) The helical probabilities near N-terminal region in simulations are much higher than that in experiments.

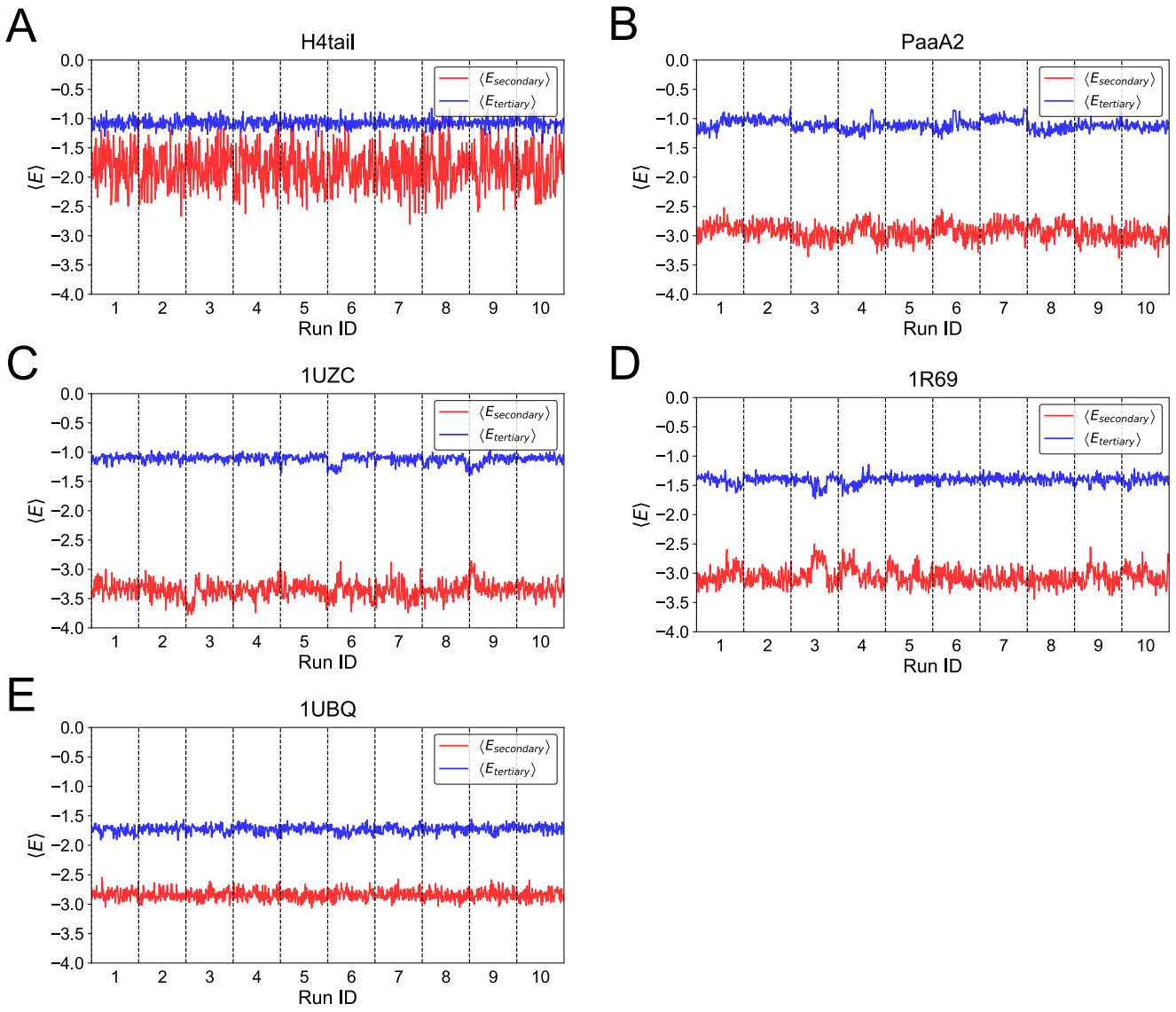

Figure S6: **The secondary and tertiary structure energy vs time in the simulations of the two IDPs and three ordered proteins.** As IDPs, H4tail and PaaA2 have either higher secondary or tertiary structural energy (A, B). As a comparison, the overall ordered 1UZC, with a disordered tail, has similar level of tertiary structure, but much lower secondary structure energy (C). The two totally ordered proteins (1R69 and 1UBQ) have both lower secondary and tertiary structural energy (D, E).

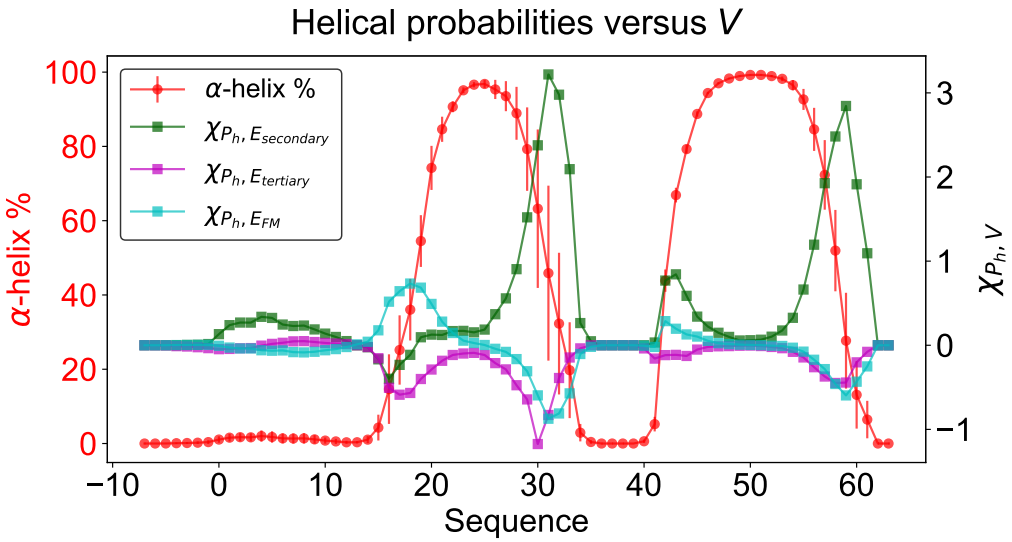

Figure S7: **The sensitivity of helical probability versus energies are not highly correlated with helical occupations, but their fluctuations.**

References

(1) Winogradoff, D.; Echeverria, I.; Potoyan, D. A.; Papoian, G. A. The Acetylation Landscape of the H4 Histone Tail: Disentangling the Interplay between the Specific and Cumulative Effects. *J. Am. Chem. Soc.* **2015**, *137*, 6245–6253.

(2) Sterckx, Y. G.; Volkov, A. N.; Vranken, W. F.; Kragelj, J.; Jensen, M. R.; Buts, L.; Garcia-Pino, A.; Jové, T.; Van Melderen, L.; Blackledge, M. et al. Small-Angle X-Ray Scattering- and Nuclear Magnetic Resonance-Derived Conformational Ensemble of the Highly Flexible Antitoxin PaaA2. *Structure* **2014**, *22*, 854–865.

(3) Amadei, A.; Ceruso, M. A.; Di Nola, A. On the Convergence of the Conformational Coordinates Basis Set Obtained by the Essential Dynamics Analysis of Proteins' Molecular Dynamics Simulations. *Proteins: Struct., Funct., Genet.* **1999**, *36*, 419–424.