HHS Public Access

Author manuscript

J Phys Chem B. Author manuscript; available in PMC 2019 November 21.

Published in final edited form as:

J Phys Chem B. 2018 November 21; 122(46): 10455–10469. doi:10.1021/acs.jpcb.8b09029.

Computational Studies of Intrinsically Disordered Proteins

Vy T. Duong[1,2], **Zihao Chen**[2], **Mahendra T. Thapa**[3], and **Ray Luo**[2,4,*]

[1]Department of Chemistry, University of California, Irvine, California 92697-3900, U.S.A.

[2]Center of Complex Biological Systems, University of California, Irvine, California 92697-3900, U.S.A.

[3]Department of Physics, California State University, Chico, Chico, California 95929, U.S.A.

[4]Departments of Molecular Biology and Biochemistry, Chemical Engineering and Materials Science, and Biomedical Engineering, University of California, Irvine, California 92697-3900, U.S.A.

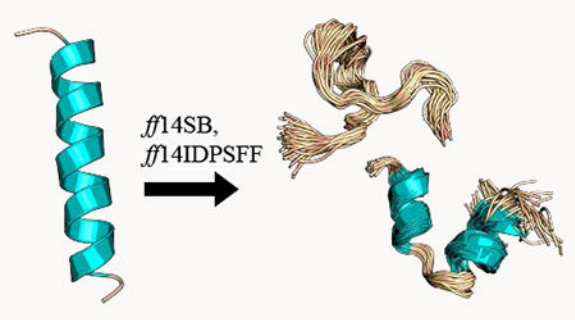
Abstract

Frequently elusive to experimental characterizations, intrinsically disordered proteins (IDPs) can be probed using molecular dynamics to provide detailed insight into their complex structure, dynamics, and function. However, previous computational studies were often found to disagree with experiment due to either force field biases or insufficient sampling. In this study, nine unstructured short peptides and the HIV-1 Rev protein were simulated and extended to microseconds to assess these limitations in IDP simulations. In short peptide simulations, a tested IDP-specific force field *ff*14IDPSFF outperforms its generic counterpart *ff*14SB as agreement of simulated NMR observables with experiment improves, though its advantages are not clear-cut in apo Rev simulations. It is worth noting that sampling is probably still not sufficient in the *ff*14SB simulations of apo Rev even if ten microseconds have been collected. This indicates that enhanced sampling techniques would greatly benefit IDP simulations. Finally, detailed structural analyses of apo Rev conformations demonstrate different secondary structural preferences between *ff*14SB (helical) and *ff*14IDPSFF (random coil). A natural next step is to ask a more quantitative question: whether *ff*14SB is too ordered or *ff*14IDPSFF is too disordered in simulations of more complex IDPs such as Rev. This requires further quantitative analyses both experimentally and computationally.

Graphical Abstract

*Author to whom correspondence should be addressed; ray.luo@uci.edu.

5. Supporting Information: Cumulative averages of NMR observables (Figures S1-S5); biphasic exponential fittings plots (Figures S6-S13); Clustering determination and additional figures (Figures S14-S16); DSSP calculations (Figures S17-S20); RMSF calculations (Figures S21).



1. INTRODUCTION

As structural data accumulates at an ever increasingly fast pace, intrinsically disordered proteins (IDPs) have garnered widespread acknowledgment for their ubiquitous presence in biochemical pathways vital to eukaryotic systems. Although the exact correlation between disordered protein regions and function remains elusive, IDPs or proteins containing both structured and intrinsically disordered regions (IDRs) have been experimentally shown to participate in DNA binding, transcription, translation, cell signaling, and the overall regulation of the cell cycle.[1–6] Mutations in IDPs/IDRs or expression pathways of IDPs/IDRs have been implicated in various neurological disorders, cancers, and other disease-related conditions.[7–9] These proteins also vary considerably in behavior, occupying a fully disordered state, exhibiting folding only upon binding (known as coupled folding and binding),[6] or existing in mixed states of structured/unstructured regions. Experimental methods to characterize IDPs and elucidate structure-function associations can therefore be arduous and challenging. To explore the dynamic structures of IDPs, computational methods can provide the expansive sampling to complement experimental measurements.

Widely used to simulate globular proteins, generic protein force fields (e.g. *ff*14SB[10] and CHARMM36[11]) have been shown to disagree with experimental observables due to biases towards structured motifs.[12] Improvements to address this bias have resulted in multiple IDP-specific force fields (CHARMM36m,[13] *ff*99IDPs,[14] *ff*14IDPs,[15] CHARMM36IDPSFF[16]) to replicate the disordered characteristics of IDPs. The *ff*14IDPs force field developed by Song *et al.*[15] included dihedral energy corrections for only eight disorder-promoting residues (A, Q, G, P, R, K, S, E).[17–19] Although this resulted in improved IDP sampling, several inconsistencies with experimental observables arose due to the limited number of residues corrected.[15] In 2017, Song *et al.*[20] extended their optimization of dihedral energy terms using grid-based energy correction maps[21–23] to all 20 amino acids resulting in the *ff*14IDPSFF force field. This new force field simulated chemical shift values in closer agreement with experimental values.[20]

Thus, our first goal of this computational study of disordered proteins is to assess the quality of both the generic protein force field (*ff*14SB[10]) and its IDP-specific counterpart (*ff*14IDPSFF[20]), However, it is notoriously difficult to obtain adequate conformational sampling for IDPs/IDRs due to the lack of one or few dominant conformations. Since microsecond timescales and multiple independent trajectories may be required, our second

goal of this study is to assess the extent of sampling that is needed for quantitative structural annotation of IDPs/IDRs and to explore how to assess the sampling convergence. Here, nine short IDP peptides of the motif EGAAXAASS (X = D, E, Q, W, Y, P, L, H, K)[24, 25] and the RNA-binding protein, HIV-1 Rev (Rev)[26–31] were chosen as test cases to assess the quality of MD simulations with the two Amber protein force fields. The EGAAXAASS short peptides were thoroughly characterized experimentally and were found to exhibit a combination of disordered behavior and local interactions between the 5X substituted residues and adjacent neutral alanine residues.[24, 25] The longer and more complex Rev protein is a more challenging and realistic system for assessment of sampling techniques and accuracy of the tested force fields. Composed of highly charged residues (10 arginines out of 23 residues), the Rev protein is a vital component in the regulation of the HIV-1 replication cycle.[26–28] Despite its short sequence the Rev protein has been shown to adopt a diverse array of conformations (α-helices, disordered, beta) and simultaneously bind to target proteins or RNA-substrates with high affinity.[27, 28, 32, 33] Once bound to its target, it was found to adopt a very stable conformation, providing a very interesting system to probe the binding-induced folding process.

By tackling issues of force field accuracy and sampling convergence, force field advancements in the realm of IDPs can be highly informative, revealing behaviors otherwise experimentally inaccessible or providing details potentially useful in guiding experimental studies. After careful analysis of the simulation sampling convergence and force field accuracy, we further analyzed the diverse conformational preferences of the Rev protein in both the apo and bound state to complete the computational analysis of this important protein.

2. METHODS

2.1 Force Fields Tested.

In this study, two Amber protein force fields (*ff*14SB and *ff*14IDPSFF) were tested to assess their quality in reproducing IDP structural properties. In the generic protein force field *ff*14SB,[10] dihedral modifications and validation relied primarily on comparison to crystal structures exhibiting ordered secondary structures. To address the limitations of increased structured propensity propagated by the *ff*14SB force field, the IDP-specific force field *ff*14IDPSFF was developed to address the deficiency of generic protein force fields by modification of the main-chain dihedral terms.[20] The *ff*14IDPSFF force field is the most recently developed AMBER IDP-specific force field, improved upon from older versions.[14, 15] Song *et al.*[20] provided the CMAP (grid-based energy correction map) parameters for *ff*14IDPSFF and a utility perl script to revise *ff*14SB-parameterized topology files into *ff*14IDPSFF topology files.

2.2 Molecular Dynamics Simulations.

The molecular dynamics package, Amber version 16, was used to generate all trajectories.[34–37] Nine short peptides with the sequence motif of EGAAXAASS (X = D, E, Q, W, Y, P, L, H, K) were tested in this study. All 9 peptides were built in the all-trans initial conformation using the Amber LEaP module, followed by minimization with the steepest

descent and conjugate gradient methods, each 500 steps. Short peptides were then simulated in the GB implicit solvent for 10 ns (time steps of 1 fs) at 450K to generate 10 random conformations per peptide per force field (Table 1). The randomized initial structures were solvated with explicit TIP3P waters in a truncated octahedron box, with a buffer of 10 Å (Table 1). Neutralization was accomplished with the addition of either Na+ or Cl− ions depending on the total charge of a peptide. All solvated structures were minimized for 20,000 steps steepest descent, heated up for 20 ps in the NVT ensemble from 0K to 298K, and were equilibrated for 20 ps in the NPT ensemble at 298K. The CUDA-accelerated PMEMD[36, 37] in Amber16 was then used to generate production trajectories in the NVT ensemble at 298K. The Langevin thermostat was used for all temperature regulation.

Force fields were also tested via simulation of a larger IDP, the HIV-1 apo Rev protein (apo Rev), by extracting the protein from its bound conformation in the crystal structure (PDB ID: 1ETF) as the initial conformation. MD preparation protocols (minimization, heating, etc.) were mostly identical to those for the nine peptides mentioned above, except that 60 random conformations per force field were generated in the GB implicit solvent. These conformations were used as the initial starting structures for two sampling strategies also outlined in Table 1: fifty 200ns simulations (short) and ten 1 μs simulations (long). Here we chose to simulate a total of 10 μs in the form of both short and long protocols to assess which strategy leads to faster convergence of tested NMR observables.

In addition to the apo Rev simulations, we also simulated the HIV-1 Rev protein bound to its RNA-binding partner Rev responsive element (RRE). Beginning with the full NMR solution structure (PDB: 1ETF), we repeated MD simulation protocol as mentioned previously, except that only five production trajectories of 200 ns each were collected.

2.3 Analyses of Simulations.

Post-simulation analysis incorporated a variety of software to extract observables for comparison with experiment. NMR observables – chemical shift and $^{3}J_{HNH\alpha}$-coupling values – were calculated to validate the performance of both tested force fields and assess the quality of MD sampling. The Amber module, cpptraj,[39] was used to remove solvent for subsequent frame-by-frame processing and analysis. All chemical shift values were calculated using the SPARTA+ package.[40] $^{3}J_{HNH\alpha}$-coupling constants were calculated using the Karplus equation that was programmed with the MDTraj python library[41] and coefficients from literature.[42] Experimental values (Figures 10C-D, 11B) were extracted from published figures in respective papers if raw data were not available from the authors (Table 1).

Time-dependent cumulative averages of both NMR observables were calculated for convergence assessment. From these cumulative average calculations, the rate of change per NMR observable ($\Delta NMR\ Observable$) was calculated to assess its rate of convergence. Rate of change datasets were fitted to a biphasic exponential-decay model:

$$\Delta\ NMR\,Observable = A_1 e^{-\frac{x}{\tau_1}} + A_2 e^{-\frac{x}{\tau_2}} + c$$

Of the fitted parameters, the slower τ_2 values were calculated and utilized to assess the rate of convergence of the observable. Kernel density estimations (KDEs) were used to analyze the detailed distribution of each predicted observable per frame. KDE's were calculated using the python packages Scikit-Learn and Seaborn.[43, 44] Epanechnikov kernels were adopted with appropriate bandwidths (h=0.5) in KDEs.[45] Initial bandwidths were determined using Scikit-Learn's grid search and cross validation function (GridSearchCV) (h=0.1) and further rescaled to h=0.5 as it yields comparable distributions with less noise.

Secondary structure propensity estimates were calculated using the DSSP program.[46] Prior to clustering, frames were pre-sorted using DSSP secondary structure assignments. Since DSSP default settings assign residues with three basic secondary structure assignments – H (α-helix, 3_{10}-helix, π-helix), E (beta ladder, isolated beta-bridge residues), C (hydrogen bond turn, bend, loops, irregular residues) – frames were first grouped into the following categories if they contained at least one of the 3 assignments: H only, E only, C only, EH only, CH only, CE only, CEH only. Frames for all simulations fell into only four of the categories: C only, CH only, CE only, and CEH only. Clustering was then restricted to a single secondary structure category (e.g. C only). This pre-clustering assortment permits filtering based on secondary structure and increases accuracy in the clustering step.

After pre-clustering, ϕ and φ torsion angles were extracted from trajectories with the MDTraj[41] module as input in our clustering methodology. Torsional data was then subjected to PCA dimensionality reduction with settings specified to retain 99% of variation in torsion angle data. Clustering was performed by generating gaussian mixture models (GMM)[47] for each secondary structure category (e.g. C only), in which each frame was clustered depending on its *likelihood* of occupying a specific component/cluster. GMMs consist of a mixture of multidimensional gaussian probability distributions from which the number of components/mixtures (number of "clusters") can be estimated using cross-validation techniques such as Bayesian information criterion (BIC).[48] The lowest BIC value was used to estimate the appropriate number of mixtures for each GMM model (Figure S14). GMMs were created using the Scikit-Learn[44] python module and implemented using the expectation-maximization algorithm[49] to fit and achieve converged mixtures/clusters.

In RRE-bound Rev (RRE-Rev) simulations, the snapshot closest to the average was used as a representative of the average structure and implemented using cpptraj.[39] Hydrogen bond occupancies were calculated using the Baker-Hubbard[50] criteria from the MDTraj[41] python module and ionic salt bridge interactions were determined with a strict distance criterion[51] (4Å) between centers of charged groups (positively charged atoms from residues Arg and Lys: NH*, NZ*; negatively charged atoms: OP* phosphate backbone atoms in the RNA-binding partner RRE). Pymol was used to generate the representative structural image and TOC image.

3. RESULTS AND DISCUSSION

Nine short peptides, EGAAXAASS (X = D, E, H, K, L, P, Q, W, Y) and the structurally dynamic apo Rev protein from type-1 HIV were simulated to illustrate the issues that must be addressed in computational studies of IDPs, namely both the accuracy of force fields and

convergence of sampling. In the following, the convergence issue of the sampling is addressed before studying the quality of the two selected force fields in reproducing NMR observables. Finally, the structural characteristics of both disordered and ordered apo Rev protein are discussed based on the expansive MD simulations in explicit solvent.

3.1 Convergence Analysis.

Previous studies of IDPs relied on backbone RMSD analysis and/or clustering of MD trajectories within hundred nanosecond timescales to confirm proper sampling and convergence of IDPs.[20, 31] In this study, we relied on direct analysis of time-dependent cumulative averages of specific NMR observables, a reasonable technique to investigate the convergence of simulated observables.

We analyzed time-dependent cumulative averages (Figure S1–S5) of simulated secondary chemical shifts and $^{3}J_{HNH\alpha}$-coupling constants to estimate the time scales at which the rates of change of the observables go to zero, an indication that convergence is achieved. A convergence decay was fitted to a biphasic exponential decay model ($\Delta\ NMR\ Observable = A_1 e^{-\frac{x}{\tau_1}} + A_2 e^{-\frac{x}{\tau_2}} + c$) thereby allowing for the determination of τ_2. Here, the parameter generated from the first rapid decay phase, τ_1 is discarded. The implementation of this technique allows us to quantitatively assess and compare the convergence rates of tested systems and sampling protocols.

Short Peptides.

Table 2 summarizes the average τ_2 values – derived from simulated $\Delta\delta C^{\alpha}$ – of the 9 short peptides. These values are further represented in boxplots detailing their ranges, medians, and lower/upper quartiles (Figure 1). Detailed fitting plots for all residues and simulation types are shown in the SI file (Figure S6–S7). Calculated average τ_2 values of EGAAXAASS simulations reveal a stark contrast between *ff*14SB- and *ff*14IDPSFF-generated simulated $\Delta\delta C^{\alpha}$ values, with *ff*14IDPSFF exhibiting lower values than the generic *ff*14SB force field, except the Q-substituted simulations, whose τ_2 values are quite similar between the two. The analysis suggests *ff*14IDPSFF simulations converge mostly faster than the *ff*14SB simulations for the chemical shifts monitored (Figure 1).

Next, we repeated the above biphasic exponential fitting to cumulative averages of a second simulated NMR observable – $^{3}J_{HNH\alpha}$-coupling constants (Figure S10–S11). Overall, the range of calculated τ_2 values is narrow and comparable between both force fields (Figure 2). Upon closer inspection, the average τ_2 (indicated by red boxes) is generally higher in *ff*14IDPSFF simulations than those in *ff*14IDPSFF simulations, different from the chemical shift analysis. Interestingly, the final $^{3}J_{HNH\alpha}$-coupling constants are comparable between the two force fields, as the average values are within standard deviations. Peptides substituted with P, Q, or W in *ff*14IDPSFF simulations, exhibit lower τ_2 values in comparison to other substituted short peptides, suggesting possible conformational preferences leading to increased convergence rate. Comparison of the τ_2 values for the two NMR observables suggests that *J*-coupling constants in general converge slower than secondary chemical shifts in our simulations, as shown in Figures 1–2 and Table 2. Nevertheless, both sets of

simulations are believed to be converged as far as both NMR observables are concerned, as the τ_2 values are much shorter than the cumulative simulation time scales sampled.

Apo Rev and RRE-Rev.

We extended the convergence analysis of the two tested force fields for the simulations of both apo and bound Rev. Biphasic exponential decay models were fitted (Figure S8–S9, S12–S13) as outlined in the *Short Peptides* subsection, using cumulative averages (Figure S3–S5) of simulated secondary Cα chemical shifts and $^3J_{HNH\alpha}$-coupling constants. A summary of τ_2 values for apo Rev in Table 3 reveals a consistent pattern in comparison to the short peptides: the τ_2 values for $\Delta\delta C^\alpha$ in *ff*14IDPSFF simulations are lower than those in *ff*14SB simulations and the τ_2 values of $^3J_{HNH\alpha}$-coupling constants in *ff*14IDPSFF simulations are higher than those in *ff*14SB simulations.

We also explored the convergence behavior of different simulation protocols in the simulations of apo Rev. Since the duration of MD simulations can significantly impact the conformational sampling, a total of 10 microseconds of MD simulation with both short (200 ns × 50) and long (1 μs × 10) protocols was generated for comparative analysis. Initial, qualitative inspection of cumulative averages (Figure S3–S4) of simulated NMR observables reveals higher fluctuations in the long protocol. Different observations in the short and long protocols suggest the two probably converged to different conformational minima, though it is clear via inspection of cumulative averages (Figure S3–S4) that the short protocol transitioned to their minima faster.

Cumulative averages were then fitted as biphasic exponential decay models (Figure S8, S12, summary of fitted τ_2 in Table 3 and Figure 3). Table 3 and Figure 3 clearly show that both NMR observables converge faster in the short protocol. This is consistent with the initial qualitative inspection of apo Rev cumulative averages (Figure S3–S4), where it appears that the short protocol produces overall better convergence trends in all cases. The τ_2 values are also consistently distributed within narrower ranges (aka smaller SDs) in the short protocol, indicating consistent convergence of simulated NMR observables. In contrast the distributions of τ_2 values from the long protocol strongly depend on force fields and observables analyzed.

Finally, convergence rates for RRE-Rev simulations in Table 3 also indicate comparable convergence between *ff*14SB and *ff*14IDPSFF simulations, although $^3J_{HNH\alpha}$-coupling-derived τ_2 values are much smaller than $\Delta\delta C^\alpha$-derived τ_2, apparently due to the much more stable Rev in the bound state. Overall the convergence rate analysis shows that it is important to monitor individual observables for their convergence trends.

3.2 Distributions of Simulated Observables.

We implemented the kernel density estimation (KDE) method to determine the probability density distributions of simulated NMR observables. There are two purposes in conducting this analysis. First, it provides a more detailed view of simulated observables. Second, it provides a means to cross-validate, in more detail, the different simulation protocols used in the simulations of the more challenging apo Rev.

Short Peptides.

Figure 4 shows KDE analyses for Cα secondary chemical shifts. The distribution in Figure 4 shows that *ff*14SB conformations (first/third columns) are concentrated into multiple peaks in regions characteristic of helices (3 ± 1 ppm) and random coil (~0 ppm).[52] As an example, peptide EGAADAASS (*ff*14SB) exhibits multiple peaks, and a higher concentration of positive secondary Cα chemical shifts. In contrast, the *ff*14IDPSFF distributions (second/fourth columns) are overall narrower, more symmetrical, and more Gaussian-like centered around 0 ppm, suggesting more uniform disordered structures in the ensemble (Figure 4).

KDEs of $^{3}J_{HNH\alpha}$-coupling scalar coupling constants are shown in Figure 5. Scalar $^{3}J_{HNH\alpha}$-coupling constants for helical structures typically average 4.2–5.6 Hz, beta sheet conformations average 8.5–10 Hz, and random coil average 5.9–7.7 Hz.[53] In Figure 5, a significant proportion of residues display peaks within the helical region, from both force fields. However, distributions in *ff*14SB simulations display higher densities characteristic of helices than those in the *ff*14IDSPFF simulations for most peptides. A high concentration of peaks can also be observed in the 8.5–10 Hz range typical of beta conformations in the *ff*14IDPSFF simulations. However only a small fraction of conformations are within values characteristic of beta conformations in the *ff*14SB simulations. We supplemented the NMR observables with a more detailed secondary structure analysis based on the DSSP[46] program. The DSSP data shows, however, that beta secondary structure is nonexistent in both simulations (Figure S17). The discrepancy is not a surprise given that the $^{3}J_{HNH\alpha}$-coupling constant calculation only considers the main-chain torsion angles while DSSP considers a range of different structural and energetic properties.

Apo Rev.

Apo Rev simulations also display similar distributions described above – increased peak densities in the helical region in the *ff*14SB simulations compared to the *ff*14IDPSFF simulations. Juxtaposition of the two distributions displays an overall heterogeneous distribution in the *ff*14SB force field, with peaks in ranges typical of helical character (3 ± 1 ppm) (Figure 6A-B). The long-protocol simulations contain higher density peaks in the 3 ± 1 ppm range, indicating that more conformations contain helical content compared to the short-protocol simulations (Figure 6B). This increased helicity observed in long *ff*14SB simulations suggests the impact of timescales (short vs. long) is more apparent in *ff*14SB simulations than *ff*14IDPSFF simulations. In the *ff*14IDPSFF simulations, both timescale types produce almost identical homogenous distributions centered ~0 ppm (Figure 6C-D).

The KDE analysis was also conducted for simulated $^{3}J_{HNH\alpha}$-coupling constants. In all simulations, we observed three general regions in the KDE distributions: helical region (average 4.2–5.6 Hz), beta region (average 8.5–10 Hz), and disordered/coiled region (average values 5.9-7.7 Hz).[53] Similar observation was also noted in experimental findings.[31] Both force fields and simulation protocols exhibit similar peaks in the helical region (broad with densities less than 0.2), but differ in the following: 1) *ff*14SB simulations peaks contain higher densities, indicating more helical content than both *ff*14IDPSFF simulations; and 2) the long-protocol *ff*14SB simulations peaks are more left-shifted indicating increased helicity than its short protocol counterpart (Figure 7A–7B). In the disordered region: 1) the

*ff*14SB simulations exhibit less disordered secondary structures as density peaks are lower than the *ff*14IDPSFF simulations; and 2) the peaks are similar between short and long-protocol simulations when apo Rev is modeled with *ff*14IDPSFF, In the beta region, density peaks in the *ff*14SB simulations are in general lower than those in the *ff*14IDPSFF simulations.

Several observations, however, are contradictory to those in the chemical-shift KDE analysis. A single peak representing residue 46R is the only density peak > 0.6 in the *ff*14SB simulations (long protocol), while all other peaks are ~0.2 density within Figure 7B. The beta region is also more readily populated with high densities in the $^{3}J_{HNH\alpha}$-coupling, distributions for all simulations whereas minimal densities were observed in the beta region (−1.48 ± 1.23 ppm)[52] in the $\Delta\delta C^{\alpha}$ distributions for the *ff*14IDPSFF simulations (Figure 6 and 7). This discrepancy might result from our uses of the $^{3}J_{HNH\alpha}$-coupling constants to infer secondary structures as discussed in the *Short Peptide* analysis.

KDE distribution analysis of simulated NMR observables is also a useful assessment of convergence quality, supplementing the convergence rate analysis in section 3.1. The distribution data show that the *ff*14SB force field is more sensitive to simulation protocols than *ff*14IDPSFF. Consistently converged distributions in the *ff*14IDPSFF simulations allow us to use the convergence rates obtained in section 3.1 to compare which protocol is better. However, the rate estimations (Table 3 and Figure 3A–3B) show that the convergence rates between the two are quite similar, within 200 ns in general, though it is clear that the short protocol converges faster than the long protocol. For *ff*14SB simulations, the different distributions presented here give us pause to claim that the sampling of the apo Rev is sufficient in either protocol even if 10 microseconds worth of sampling has been collected (Figure 6). This indicates that enhanced sampling techniques would greatly benefit IDP simulations for systems as small as 23 amino acids such as apo Rev.

3.3 Comparison of Simulated and Measured NMR Observables.

Short Peptides.

We next calculated the final averages of secondary Cα chemical shifts for both sets of simulations and compared with experimental values (Figures 8). Figure 8 shows that experimental chemical shifts[24, 25] of the 5X-substituted residues often result in a more negative ppm shift. This suggests that the 5X-substituted residues are more disordered/extended than their adjacent residues.[52] This trend can be reproduced by both force fields, with the exception of the 5W-substituted simulations (Figure 8). In 5P-substituted simulation simulations, the proline residue is expected to rigidify and increase overall order in the peptide.[24, 54] Both sets of simulations agree well with experiment, replicating the expected −2 ppm shift observed for residue 4A, with *ff*14DIPSFF generating a slightly more negative shift (Figure 8). In simulations of aromatic-substituted residues (5X = W, Y), both force fields also replicate a similar observation by Dames *et. al*,[24] a negative −0.3 ppm shift in residue 6A. Overall, the agreement between simulation and experiment is summarized in Table 4, which shows improved performance of *ff*14IDPSFF over its generic counterpart *ff*14SB in modeling the tested peptides (Table 4, Figure 8).

We also compared simulated $^{3}J_{HNH\alpha}$-coupling, constants to experimental values for these disordered peptides in Figure 9. Table 4 presents corresponding root mean square errors (RMSEs) with respect to experiment, indicating overall better agreement between experimental and *ff*14IDPSFF-simulated values (Table 4, Figure 9). In summary, both simulated chemical shifts and *J*-coupling constants demonstrates that the *ff*14IDPSFF simulations can better reproduce the two tested NMR observables than the *ff*14SB simulations in these short peptides.

Apo Rev.

In simulations of the more complex apo Rev, simulated secondary chemical shifts do not agree with experiment as well as those in the tested short peptides. For *ff*14SB simulations, short (200 ns × 50) and long (1 μs × 10) protocols overall agree with each other but not in the N-terminal portion (residues 35 to 41) (Figure 10A). Overall the long protocol agrees a bit better with experiment (Table 4). Experimental values occupy mostly positive secondary chemical shifts, indicating possible residual helical secondary structure in apo Rev and this is reproduced well in the *ff*14SB simulations. It is also worth noting experimental secondary chemical shifts are still within reasonable values typical of random coil, < 2 ppm. For *ff*14IDPSFF simulations, both short and long protocols produce nearly identical secondary chemical shift values (Figure 10B), lending support that the simulated observables converged very well. However, the agreement with experiment is not as good as the *ff*14SB simulations (Figure 10B and Table 4). Specifically, the *ff*14IDPSFF simulations may overestimate disordered structures in apo Rev.

Interestingly worse agreement is apparent between *ff*14SB-simulated $^{3}J_{HNH\alpha}$-coupling constants and experimental values (Figure 10C). Overall higher helical propensity is visible in the *ff*14SB simulations (average 4.2–5.6 Hz) versus higher disordered propensity (average 5.9–7.7 Hz) in the experiment (Figure 10C). Notably, *ff*14IDPSFF simulations agree closer to experiment in this regard with $^{3}J_{HNH\alpha}$-coupling constants in the similar range as in the experiment. Nevertheless, both experimental and simulated $^{3}J_{HNH\alpha}$-coupling constants are still within reasonable range of disordered secondary structure. These ambiguous, sometimes overlapping secondary structure boundaries used in NMR experiments highlight the difficulty in definitively assigning secondary structures based on either chemical shifts and $^{3}J_{HNH\alpha}$-coupling constants. Multiple, independent CD experiments, however, suggest the conformational landscape of apo Rev is more populated as disordered than helical.[26, 30, 31, 55] In summary, the *ff*14IDPSFF simulations agree surprisingly well with both NMR and CD experiments with disordered structures dominant in its simulations of apo Rev. These observations will be highly useful in further refining IDP-specific force fields to improve simulation of complex, dynamic IDPs such as apo Rev.

RRE-Rev.—Since the Rev protein is known to sustain a helical structure upon binding to its RNA-binding partner, Stem IIB of Rev response element (RRE), we also simulated the RRE-Rev complex (PDB: IETF) and compared to the apo Rev simulations. Experimental $\Delta\delta C^{\alpha}$ and $^{3}J_{HNH\alpha}$-coupling constant datasets were extracted from two separate literature sources and each source used different non-native residues in the N-terminal portion of otherwise identical Rev peptides.[30, 31, 38] The $^{3}J_{HNH\alpha}$-coupling dataset[31] was generated

from a Rev peptide containing a 4-residue non-native extension (GAMA) at the N-terminus, while the $\Delta\delta C^{\alpha}$ dataset[38] resulted from a Rev peptide containing a non-native, N-terminal residue Asp. The GAMA sequence was a byproduct leftover from His6-GB1 tag, and the Asp non-native sequence was used as an alternative to a synthetic N-terminal sequence from earlier experiments. Although we chose to simulate Rev bound to RRE with the N-terminal Asp from the literature,[38] the remaining 22 residues are identical between Rev peptides used in both experiments. Nevertheless experimental data show that both sequences from literature[30, 31, 38] exhibited RNA-binding specificity/activity in addition to disordered secondary structure in the apo state.

Although experimental chemical shifts fluctuate significantly, simulated values are stable and almost identical between the two force fields except terminal residues 49–52 (Figure 11). Both C-terminal experimental and simulated values seem to be decreasing to ranges characteristic of random coil (Figure 11). In analyses of $^{3}J_{HNH\alpha}$-coupling constants, experimental values and *ff*14SB-simulated values occupy typically helical ranges (< 5.6 Hz), whereas *ff*14IDPSFF-simulated values are almost identical to both *ff*14SB and experimental values until residue 49Q (Figure 11). The comparison shows that the beta-forming tendency is too strong for 49Q in the *ff*14IDPSFF simulations of the bound Rev (Figure 11B). Similar tendency is also noticeable in the *ff*14IDPSFF simulations of the apo Rev (Figure 10D) where the $^{3}J_{HNH\alpha}$-coupling constant is also overestimated for 49Q. This suggests further refinement is clearly required in the development of IDP force fields. RMSE differences between simulated NMR observables and experimental values are also rather close (Table 4), though the chemical shift agreement is not as good as those for the apo Rev simulations. This is probably because RRE was not considered in the conversion from MD conformations to chemical shifts by the SPARTA+ package.[40] Overall both *ff*14SB and *ff*14IDPSFF are adequate in the RRE-Rev simulations, with accuracy in predicted NMR observables comparable to that obtained for the NMR structure (RMSE of 2.50 ppm for $\Delta\delta C^{\alpha}$ and RMSE of 1.86 Hz for $^{3}J_{HNH\alpha}$-coupling constants).

3.4 Structural Signatures of Apo Rev Disordered State.

Despite the extensive investigation of the Rev protein, as evidenced by 1647 hits from a general Pubmed search, this highly dynamic protein only occupies a monomeric state at submicromolar concentrations,[56] thus remaining elusive to structural characterization. Previous pursuits to structurally characterize the apo form of Rev encountered difficulties ranging from protein solubility to oligomerization, preventing characterization of apo Rev in physiological conditions.[57] Early circular dichroism (CD) and mutagenesis experiments suggest that apo Rev is disordered, forming helical structure depending on terminal amino acids (e.g. amidated C-terminus, C-terminal extension AAAR).[29] Overall, attempts to characterize monomeric apo Rev have required techniques to induce ordered structure propensity, such as specific helix-inducing solution buffers (e.g. 2,2,2-trifluoroethanol), residue mutations to prevent oligomerization, or the introduction of structure-inducing binding partners.[29,58] MD simulations thus provide a useful tool to probe the highly mobile conformations of Rev in its physiological disordered state. In previous structural modeling studies and MD simulations from Song *et. al*[20] and Casu *et. al*[31], researchers observed primarily coiled secondary structure of apo Rev. These simulations however simulate apo

Rev in nanosecond timescales. Herein we generated tens of microseconds trajectories to ensure proper sampling of disordered apo Rev conformations.

Clustering and secondary structure propensity calculations are discussed hereafter, highlighting the differences between the *ff*14SB and *ff*14IDPSFF simulations (in the long protocol). Although both *ff*14SB and *ff*14IDPSFF simulations exhibit ordered and disordered characteristics, the two force fields differ in secondary structure preferences: increased helical content observations in the generic *ff*14SB simulations (Figure 12), disordered structural preferences in the *ff*14IDPSFF simulations (Figure 13). The top ten clusters between both force fields occupy similar percentages: *ff*14SB at 17.87% versus *ff*14IDPSFF at 17.41%. Further evidence from DSSP[46] (hydrogen bond estimation algorithm) calculations also suggests the majority of *ff*14IDPSFF conformations exhibit coiled secondary structure, in Figure S18. All residues in *ff*14IDPSFF simulations exhibit roughly equal probabilities of coiled secondary structure (average > 80%) in addition to some beta contents (Figure S18B-C, S19B-C). DSSP (Figures S18–S19) and clustering results (Figures S15–S16) of the short protocol simulations are also provided in the supplementary information although simulations from the long protocol are the primary focus in this section. Experimental findings ranging from secondary chemical shift, $^{3}J_{HNH\alpha}$-coupling, and CD suggests apo Rev is mainly disordered when unbound.[31] Despite the observation that both force fields replicate the average coiled secondary structure as in experiment, these clustering analyses show that each force field exhibits either disordered or ordered structural bias – observations that will be useful in future refinement of IDP-specific force fields.

3.5 Conformational Analysis of Bound Rev Ordered State.

To supplement our apo Rev simulations above, we also simulated Rev bound to its RNA binding partner, RRE Stem IIB, to assess how our simulations perform in replicating experimentally-observed behaviors such as induced fit.[56, 59] Previous studies emphasize induced fit and proper RRE binding requires the presence of a single Rev monomer, from which more Rev monomers are recruited and oligomerize.[56] The NMR solution structure depicts an α-helical Rev situated in the major groove of RRE-Stem IIB.[38] After simulating this complex, we proceeded to align the Rev peptide from the NMR solution structure (PDB: IETF) to the average Rev structure extracted from RRE-Rev simulations (Figure 14). Simulations of Rev bound to RRE yield significantly more stabilized conformations compared to apo simulations. In the *ff*14SB simulations, we observed almost entirely helical content (Figure 14). In *ff*14IDPSFF force field simulations, helical secondary structure was observed in N-terminal residues, whereas coiled, disordered structure was observed in C-terminal residues (Figure 14). We also estimated the average secondary structure propensities of each residue for all simulations using the DSSP algorithm (Figure S20). Despite some fluctuation in the last 4–5 C-terminal residues, most residues remain fairly stable, retaining the characteristic helical conformation found in the NMR solution structure (Figure S20).[31]

Unsurprisingly, *ff*14SB simulations yield a lower RMSD than *ff*14IDPSFF simulations from alignments to the experimental structure (Figure 14). This induced helical content is most

likely attributed to inherent native-structure-biases of the generic *ff*14SB protein force field.[60–63] Although the RMSD of the experimental and *ff*14IDPSFF-derived structure is larger, it is notable that the helical component is quite stable (first 16 residues), with the remaining 7 residues exhibiting multiple helix-to-coil transitions (Figure 14, S20). Chemical shift and CD data of the wild-type Rev and various mutants (oligomerization-deficient mutant V16D/I55N Rev, and L60R mutant Rev bound to Stem IIB RRE), also suggests disordered content in the C-terminus.[26, 30, 31, 55] The stable N-terminal fragment found in *ff*14SB- and *ff*14IDPSFF-simulated residues contrasts sharply with the high structural fluctuation observed in apo Rev simulations, and is consistent with experimental RRE-Rev results.[31] Alignment of average simulated complexes also generated structures similar to the experimental NMR solution structure (Figure 15).

Fluctuation of Rev backbone atoms are further explored via root-mean squared fluctuation (RMSF) analyses for apo and bound Rev simulations. In all Rev simulations, backbone atoms (Cα) fluctuate more in *ff*14IDPSFF simulations than the *ff*14SB simulations (Figure 16). Comparison of apo and bound simulations shows the bound Rev fluctuates less, due to the stabilization from binding with RRE (Figure 16C, S21). Unsurprisingly terminal residues display the highest fluctuation in all simulations, except the relatively stable N-terminal region in the bound Rev simulations. This is corroborated by hydrogen bonding populations of residues 34–36 (Figure 16, S21, Table 5), which stabilizes the N-terminal region. The observed different fluctuation trends can also be explained by the different secondary structure propensities. For instance in Figure 16B, residues 36–38 in the *ff*14SB apo Rev simulations exhibit lower RMSF values and also exhibit higher helical propensity (Figure S18A).

Inspection of intermolecular hydrogen bond and ionic salt bridge occupancies (only frequencies > 0.5 is shown) in Table 5 and 6 reveals similar interactions between simulations of both force fields, but with slight differences (Table 5). Since ionic salt bridge formations are almost identical between the two force fields (Table 6), we chose to focus primarily on differences in hydrogen bond formation. In *ff*14SB complex simulations, the hydrogen bond pair ARG46-U72 dominates compared to *ff*14IDPSFF complexes due to the increased stability and helical propensity of the C-terminal end (Table 5). While retaining mostly helical character between residues 33–46, Rev contains two hydrogen bonds (GLN36-G47, ARG41-U45) in the N-terminal region in the *ff*14IDPSFF simulations, which are less frequent in the *ff*14SB simulations, an unexpected outcome considering the stability of the *ff*14SB simulations over that of the *ff*14IDPSFF simulations (Table 5). Co-existence of stabilized N-terminal helices and coiled C-terminal components in the *ff*14IDPSFF simulations of bound Rev suggests this new force field is able to simulate disordered region in an otherwise ordered protein, while the *ff*14SB simulation retains more helical characteristics.

4. CONCLUSIONS

IDPs remain elusive by standard experimental methods due to their conformational flexibility. Molecular dynamics simulations can thus provide detailed insight into their complex structures, dynamics, and functions, if they can reproduce the available

experimental observables. However, there are several issues in computational studies. First the generic force fields were found to be biased towards ordered structures in many prior simulation studies. Second the expansive conformations occupied by IDPs is often beyond typical simulation amount needed for ordered proteins.

Thus, our first goal of this computational study is to assess the quality of both a generic protein force field (*ff*14SB) and its IDP-specific counterpart (*ff*14IDPSFF) that was intended to address the biases in the generic force field. Overall simulated average observables from *ff*14IDPSFF replicate experimental chemical shifts and $^{3}J_{HNH\alpha}$-coupling constants more accurately than those derived from *ff*14SB simulations for the tested EGAAXAASS peptides. DSSP analyses also suggest different secondary structural biases between the two force fields, increased helical content from *ff*14SB and coiled content from *ff*14IDPSFF, with the latter in higher agreement with experiment. When used to simulate more complex proteins such as Rev in apo and bound forms, computational models gravitate toward either ordered secondary structure (*ff*14SB) or disordered secondary structure (*ff*14IDPSFF) as the clustering analyses revealed. However simulated observables between the two force fields are roughly comparable to experiment, *ff*14IDPSFF simulations agree with both NMR and CD measurements slightly better.

Our second goal of this study is to assess the extent of sampling that is needed for quantitative structural annotation of IDPs and to explore how to assess the sampling convergence. This was first conducted by analyses of convergence rates of individual observables in the form of bi-phasic decays. Convergence analyses of both NMR observables show that *ff*14IDPSFF simulations converge slightly faster than *ff*14SB simulations in the chemical shift calculations for all tested systems, though they converge slightly slower for $^{3}J_{HNH\alpha}$-coupling constants for all tested systems. This is consistent with the observations that conformations in *ff*14IDPSFF simulations are more diversified, sampling a larger range of main-chain torsion angles, leading to slower convergence in $^{3}J_{HNH\alpha}$-coupling constants that solely depends on these torsion angles. The decay half times also show that the total sampling amount (in term of nanoseconds simulated) is adequate as they are much less the total amount collected.

In addition, simulation protocols were also tested by simulating apo Rev as either many short (50 × 200ns) trajectories or a few long (10 × 1 μs) trajectories. Consistently converged distributions in the *ff*14IDPSFF simulations allows us to use the convergence rates to compare which protocol is better. However, the rate estimations show that differences in the convergence rates between the two are small, within 200ns in general, though it can be said the short protocol is slightly faster than the long protocol. For *ff*14SB simulations, the different distributions give us pause to claim that the sampling of the apo Rev is sufficient in either protocol even if 10 microseconds worth of sampling has been collected. This indicates that enhanced sampling techniques would greatly benefit IDP simulations for systems as small as 23 amino acids such as apo Rev.

Despite the short sequence length of apo Rev, no monomeric disordered Rev protein has been structurally characterized as demonstrated by its absence in the Protein Data Bank (PDB). To compensate for this lack of structural characterization, we utilized a combination

of NMR and CD data for comparison to our clustering and secondary structural analyses. Chemical shift and CD studies from various different sources of oligomerization-deficient mutants and wildtype Rev conclude that monomeric Rev is mostly disordered.[26, 30, 31, 55] These experimental findings are comparable to random coil clusters and DSSP calculations from the *ff*14DIPSFF simulations of and differ from the *ff*14SB simulations where increased helical content was found. Both force fields also generate stabilized helical structure and induced fit in RRE-REV simulations, exhibiting a coiled C-terminus as shown by the chemical shift data.[30, 31, 38] These structural computational studies of apo and bound Rev stress the importance to assign the correct secondary structural biases in both force fields.

Interesting observations were also found when Rev was simulated with its RNA-binding partner RRE, *ff*14DIPSFF was able to replicate the structured regions in the bound form, despite over-representation of coiled secondary structure in the apo Rev simulations. Detailed analysis of the average conformation and secondary structures of the *ff*14IDPSFF simulations shows that both the helical N-terminal region and coiled C-terminal region are readily observed, in agreement with experimental findings, despite coiled secondary structural preferences in the apo Rev simulations. In comparison, a more stable helical structure was observed throughout the *ff*14SB simulations. A natural next step is to ask a more quantitative question: whether *ff*14SB is too stable or *ff*14IDPSFF is too unstable in the simulations of more complex IDPs such as Rev. This requires further quantitative stability analysis both experimentally and computationally.

This study articulates the difficulties of obtaining converged and expansive sampling of IDPs, though our exploration of different simulation protocols demonstrates consistent observations with the *ff*14IDPSFF force field regardless of the protocols used. Although successful in simulating short peptides and bound Rev, the advantages of *ff*14IDPSFF are not as clear-cut for the more complex apo Rev. These findings also suggest future refinements of IDP-specific force fields and reduction of force field biases are still necessary for consistent performance in modeling IDPs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We thank Drs. Song and Chen for supplying the parameter set and the perl script to implement the *ff*14IDPSFF force field. V.T.D. was supported by the Mathematical, Computational and Systems Biology Pre-doctoral Training Grant T32 EB009418-08. This work was supported in part by NIH/NIGMS (GM093040 & GM079383 to R.L.).

7. REFERENCES

1. Weiss MA; Ellenberger T; Wobbe CR; Lee JP; Harrison SC; Struhl K Folding transition in the DNA-binding domain of GCN4 on specific binding to DNA. Nature 1990, 347, 575–578. [PubMed: 2145515]
2. Wright PE; Dyson HJ Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol 1999, 293, 321–331. [PubMed: 10550212]
3. Liu J; Perumal NB; Oldfield CJ; Su EW; Uversky VN; Dunker AK Intrinsic disorder in transcription factors. Biochemistry 2006, 45, 6873–6888. [PubMed: 16734424]

4. Iakoucheva LM; Brown CJ; Lawson JD; Obradovic Z; Dunker AK Intrinsic disorder in cell-signaling and cancer-associated proteins. J. Mol. Biol 2002, 323 573–584. [PubMed: 12381310]
5. Galea CA; Wang Y; Sivakolundu SG; Kriwacki RW Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. Biochemistry 2008, 47 7598–7609. [PubMed: 18627125]
6. Spolar RS; Record MT, Jr. Coupling of local folding to site-specific binding of proteins to DNA. Science 1994, 263, 777–784. [PubMed: 8303294]
7. Vavouri T; Semple JI; Garcia-Verdugo R; Lehner B Intrinsic protein disorder and interaction promiscuity are widely associated with dosage sensitivity. Cell 2009, 138, 198–208. [PubMed: 19596244]
8. Gsponer J; Futschik ME; Teichmann SA; Babu MM Tight regulation of unstructured proteins: from transcript synthesis to protein degradation. Science 2008, 322, 1365–1368. [PubMed: 19039133]
9. Babu MM; van der Lee R; de Groot NS; Gsponer J Intrinsically disordered proteins: regulation and disease. Curr. Opin. Struct. Biol 2011 21, 432–440. [PubMed: 21514144]
10. Maier JA; Martinez C; Kasavajhala K; Wickstrom L; Hauser KE; Simmerling C ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J. Chem. Theory Comput 2015, 11, 3696–3713. [PubMed: 26574453]
11. Best RB; Zhu X; Shim J; Lopes PE; Mittal J; Feig M; Mackerell AD, Jr. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone phi, psi and side-chain chi(1) and chi(2) dihedral angles. J. Chem. Theory Comput 2012, 8, 3257–3273. [PubMed: 23341755]
12. Best RB; Zheng W; Mittal J Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. J. Chem. Theory Comput 2014, 10, 5113–5124. [PubMed: 25400522]
13. Huang J; Rauscher S; Nawrocki G; Ran T; Feig M; de Groot BL; Grubmuller H; MacKerell AD, Jr. CHARMM36m: An improved force field for folded and intrinsically disordered proteins. Nat. Methods 2017, 14, 71–73. [PubMed: 27819658]
14. Ye W; Ji D; Wang W; Luo R; Chen HF Test and Evaluation of ff99IDPs Force Field for Intrinsically Disordered Proteins. J. Chem. Inf. Model 2015, 55, 1021–1029. [PubMed: 25919886]
15. Song D; Wang W; Ye W; Ji D; Luo R; Chen HF ff14IDPs force field improving the conformation sampling of intrinsically disordered proteins. Chem. Biol. Drug Des 2017, 89, 5–15. [PubMed: 27484738]
16. Liu H; Song D; Lu H; Luo R; Chen HF Intrinsically disordered protein-specific force field CHARMM36IDPSFF. Chem. Biol. Drug Des 2018.
17. Dunker AK; Lawson JD; Brown CJ; Williams RM; Romero P; Oh JS; Oldfield CJ; Campen AM; Ratliff CM; Hipps KW; et al. Intrinsically disordered protein. J. Mol. Graph. Model 2001 19, 26–59. [PubMed: 11381529]
18. Romero P; Obradovic Z; Li X; Garner EC; Brown CJ; Dunker AK Sequence complexity of disordered protein. Proteins 2001, 42, 38–48. [PubMed: 11093259]
19. Williams RM; Obradovi Z; Mathura V; Braun W; Garner EC; Young J; Takayama S; Brown CJ; Dunker AK The protein non-folding problem: amino acid determinants of intrinsic order and disorder. Pac. Symp. Biocomput 2001, 89–100. [PubMed: 11262981]
20. Song D; Luo R; Chen HF The IDP-Specific Force Field ff14IDPSFF Improves the Conformer Sampling of Intrinsically Disordered Proteins. J. Chem. Inf. Model 2017, 57, 1166–1178. [PubMed: 28448138]
21. MacKerell AD, Jr.; Feig M; Brooks CL, 3rd Improved treatment of the protein backbone in empirical force fields. J. Am. Chem. Soc 2004, 126, 698–699. [PubMed: 14733527]
22. Mackerell AD, Jr.; Feig M; Brooks CL, 3rd Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. J. Comput. Chem 2004, 25, 1400–1415. [PubMed: 15185334]
23. MacKerell AD; Bashford D; Bellott M; Dunbrack RL; Evanseck JD; Field MJ; Fischer S; Gao J; Guo H; Ha S;et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B 1998, 102, 3586–3616. [PubMed: 24889800]

24. Dames SA; Aregger R; Vajpai N; Bernado P; Blackledge M; Grzesiek S Residual dipolar couplings in short peptides reveal systematic conformational preferences of individual amino acids. J. Am. Chem. Soc 2006, 128, 13508–13514. [PubMed: 17031964]
25. Leung HT; Bignucolo O; Aregger R; Dames SA; Mazur A; Berneche S; Grzesiek S A Rigorous and Efficient Method to Reweight Very Large Conformational Ensembles Using Average Experimental Data and to Determine Their Relative Information Content. J. Chem. Theory Comput 2016, 12, 383–394. [PubMed: 26632648]
26. Daugherty MD; D'Orso I; Frankel AD A solution to limited genomic capacity: using adaptable binding surfaces to assemble the functional HIV Rev oligomer on RNA. Mol. Cell 2008, 31, 824–834. [PubMed: 18922466]
27. Malim MH; Cullen BR HIV-1 structural gene expression requires the binding of multiple Rev monomers to the viral RRE: Implications for HIV-1 latency Cell 1991 65 241–248. [PubMed: 2015625]
28. Mann DA; Mikaelian I; Zemmel RW; Green SM; Lowe AD; Kimura T; Singh M; Butler PJ; Gait MJ; Karn J A molecular rheostat. Co-operative rev binding to stem I of the rev-response element modulates human immunodeficiency virus type-1 late gene expression. J. Mol. Biol 1994, 241, 193–207. [PubMed: 8057359]
29. Tan R; Chen L; Buettner JA; Hudson D; Frankel AD RNA recognition by an isolated alpha helix. Cell 1993, 73 1031–1040. [PubMed: 7684657]
30. Battiste JL; Mao H; Rao NS; Tan R; Muhandiram DR; Kay LE; Frankel AD; Williamson JR Alpha helix-RNA major groove recognition in an HIV-1 rev peptide-RRE RNA complex. Science 1996, 273, 1547–1551. [PubMed: 8703216]
31. Casu F; Duggan BM; Hennig M The arginine-rich RNA-binding motif of HIV-1 Rev is intrinsically disordered and folds upon RRE binding. Biophysical Journal 2013, 105, 1004–1017. [PubMed: 23972852]
32. Smith CA; Calabro V; Frankel AD An RNA-binding chameleon. Mol. Cell 2000, 6, 1067–1076. [PubMed: 11106746]
33. Tan R; Frankel AD Structural variety of arginine-rich RNA-binding peptides. Proc. Natl. Acad. Sci. U. S. A 1995, 92, 5282–5286. [PubMed: 7777498]
34. Case DA; Cheatham TE, 3rd; Darden T; Gohlke H; Luo R; Merz KM, Jr.; Onufriev A; Simmerling C; Wang B; Woods RJ The Amber biomolecular simulation programs. J. Comput Chem 2005, 26, 1668–1688. [PubMed: 16200636]
35. Case DA; Walker RC; Cheatham III TE; Simmerling CL; Wang J; Duke RE; Luo R; Crowley M; Walker RC; Zhang W; et al. AMBER 2017 Reference Manual. University of California, San Francisco: 2017.
36. Gotz AW; Williamson MJ; Xu D; Poole D; Le Grand S; Walker RC Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 1. Generalized Born. J. Chem. Theory Comput 2012, 8, 1542–1555. [PubMed: 22582031]
37. Salomon-Ferrer R; Gotz AW; Poole D; Le Grand S; Walker RC Routine Microsecond Molecular Dynamics Simulations with AMBER on GPUs. 2. Explicit Solvent Particle Mesh Ewald. J. Chem. Theory Comput 2013, 9, 3878–3888. [PubMed: 26592383]
38. Battiste JL Structure determination of an HIV-1 RRE RNARev peptide complex by NMR spectroscopy. Ph.D Thesis, Massachusetts Institute of Technology, Cambridge, MA, 1996.
39. Roe DR; Cheatham TE, 3rd PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. J. Chem. Theory Comput 2013, 9, 3084–3095. [PubMed: 26583988]
40. Shen Y; Bax A SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network J. Biomol. NMR 2010, 48, 13–22. [PubMed: 20628786]
41. McGibbon RT; Beauchamp KA; Harrigan MP; Klein C; Swails JM; Hernandez CX; Schwantes CR; Wang LP; Lane TJ; Pande VS MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. Biophys. J 2015, 109, 1528–1532. [PubMed: 26488642]
42. Vogeli B; Ying J; Grishaev A; Bax A Limits on variations in protein backbone dynamics from precise measurements of scalar couplings. J. Am. Chem. Soc 2007, 129, 9377–9385. [PubMed: 17608477]

43. Hunter JD Matplotlib: A 2D graphics environment. Comput. Sci. Eng 2007, 9, 90–95.
44. Pedregosa F; Varoquaux G; Gramfort A; Michel V; Thirion B; Grisel O; Blondel M; Prettenhofer P; Weiss R; Dubourg V Scikit-learn: Machine learning in Python. J. Mach. Learn. Res 2011 12, 2825–2830.
45. Epanechnikov VA Non-parametric estimation of a multivariate probability density. Theory Probab. Its Appl 1969, 14, 153–158.
46. Kabsch W; Sander C Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983, 22, 2577–2637. [PubMed: 6667333]
47. Duda RO; Hart PE Pattern classification and scene analysis. A Wiley-Interscience Publication, New York: Wiley, 1973 1973.
48. Schwarz G Estimating Dimension of a Model. Ann. Stat 1978, 6, 461–464.
49. Dempster AP; Laird NM; Rubin DB Maximum likelihood from incomplete data via the EM algorithm. J. Royal Stat. Soc. Series B (methodological) 1977, 1–38.
50. Baker EN; Hubbard RE Hydrogen bonding in globular proteins. Prog. Biophys. Mol. Biol 1984, 2, 97–179.
51. Barlow DJ; Thornton JM Ion-pairs in proteins. J. Mol. Biol 1983, 168, 867–885. [PubMed: 6887253]
52. Spera S; Bax A Empirical correlation between protein backbone conformation and C. alpha. and C. beta. J. Am. Chem. Soc 1991 113, 5490–5492.
53. Smith LJ; Bolin KA; Schwalbe H; MacArthur MW; Thornton JM; Dobson CM Analysis of main chain torsion angles in proteins: prediction of NMR coupling constants for native and random coil conformations. J. Mol. Biol 1996, 255, 494–506. [PubMed: 8568893]
54. Louhivuori M; Fredriksson K; Paakkonen K; Permi P; Annila A Alignment of chain-like molecules. J. Biomol. NMR 2004, 29, 517–524. [PubMed: 15243182]
55. Daugherty MD; Booth DS; Jayaraman B; Cheng Y; Frankel AD HIV Rev response element (RRE) directs assembly of the Rev homooligomer into discrete asymmetric complexes. Proc. Natl. Acad. Sci. U. S. A 2010, 107, 12481–12486. [PubMed: 20616058]
56. Cole JL; Gehman JD; Shafer JA; Kuo LC Solution oligomerization of the rev protein of HIV-1: implications for function. Biochemistry 1993, 32, 11769–11775. [PubMed: 8218247]
57. Pond SJ; Ridgeway WK; Robertson R; Wang J; Millar DP HIV-1 Rev protein assembles on viral RNA one molecule at a time. Proc. Natl. Acad. Sci. U. S. A 2009, 106, 1404–1408. [PubMed: 19164515]
58. Scanlon MJ; Fairlie DP; Craik DJ; Englebretsen DR; West ML NMR solution structure of the RNA-binding peptide from human immunodeficiency virus (type 1) Rev. Biochemistry 1995, 34, 8242–8249. [PubMed: 7599117]
59. Williamson JR Induced fit in RNA-protein recognition. Nat. Struct. Biol 2000, 7, 834–837. [PubMed: 11017187]
60. Best RB; Buchete NV; Hummer G Are current molecular dynamics force fields too helical Biophys. J 2008, 95, L07–09. [PubMed: 18456823]
61. Freddolino PL; Liu F; Gruebele M; Schulten K Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. Biophys. J 2008, 94, L75–77. [PubMed: 18339748]
62. Garcia AE; Sanbonmatsu KY Alpha-helical stabilization by side chain shielding of backbone hydrogen bonds. Proc. Natl. Acad. Sci. U. S. A 2002, 99, 2782–2787. [PubMed: 11867710]
63. Hornak V; Abel R; Okur A; Strockbine B; Roitberg A; Simmerling C Comparison of multiple Amber force fields and development of improved protein backbone parameters. Proteins 2006, 65, 712–725. [PubMed: 16981200]

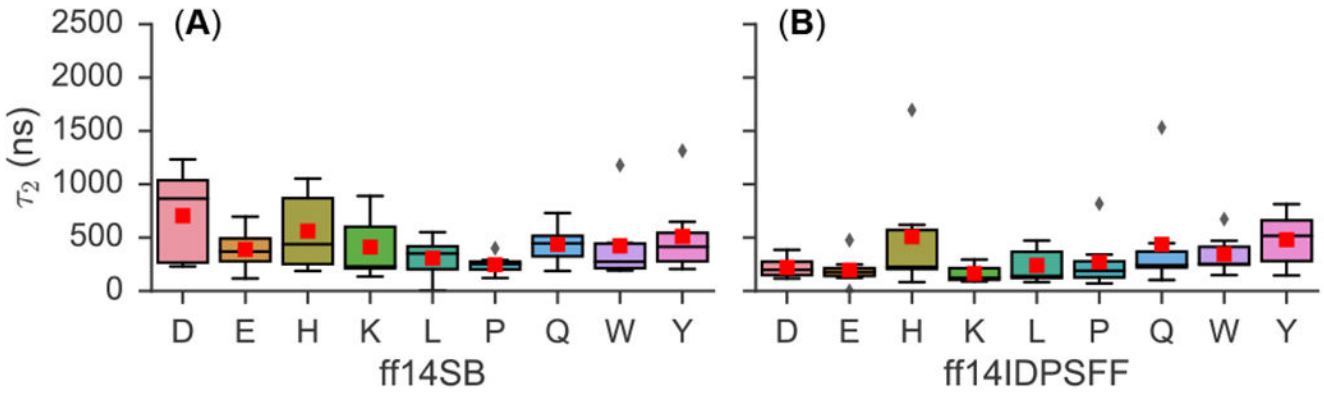


Figure 1.

Summary of τ_2 values (medians, ranges, quartiles, outliers) for peptides of EGAAXAASS (X=D, E, H, K, L, P, Q, W, Y), derived from $\Delta\Delta\delta C^{\alpha}$ calculations. Simulations are labeled by peptide and force field: (A) *ff*14SB and (B) *ff*14IDPSFF. Diamonds indicate outliers and a red box denotes the average τ_2 value. Fitted plots from which boxplots were derived can be found in the SI (Figure S6–S7).

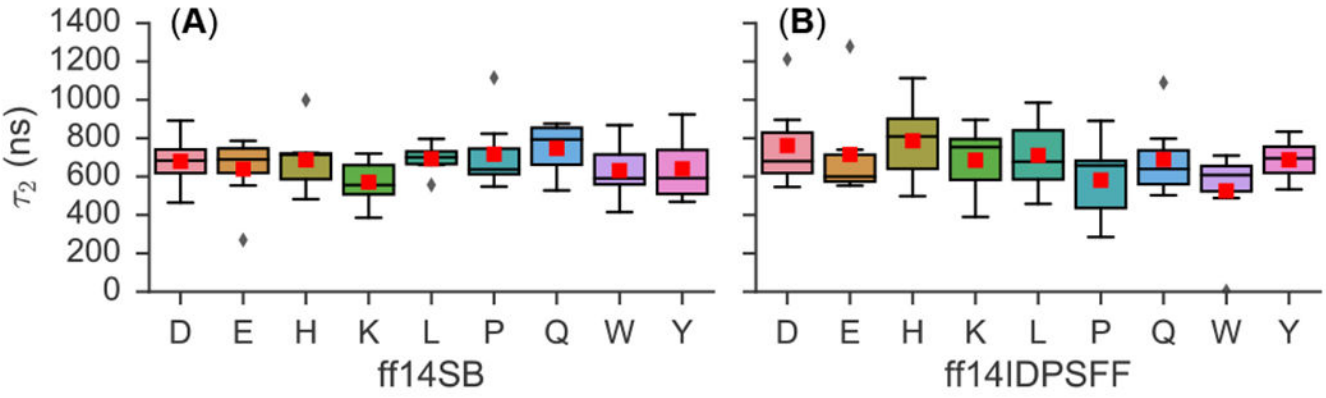


Figure 2.

Summarization of τ_2 values (median, range, quartiles, outliers) for peptides of EGAAXAASS (X=DEHKLPQWY), derived from $^3J_{HNHa}$-coupling constants. Diamonds indicate outliers and a red box denotes the average τ_2 value. Fitted plots from which boxplots were derived can be found in the SI (Figure S10–S11).

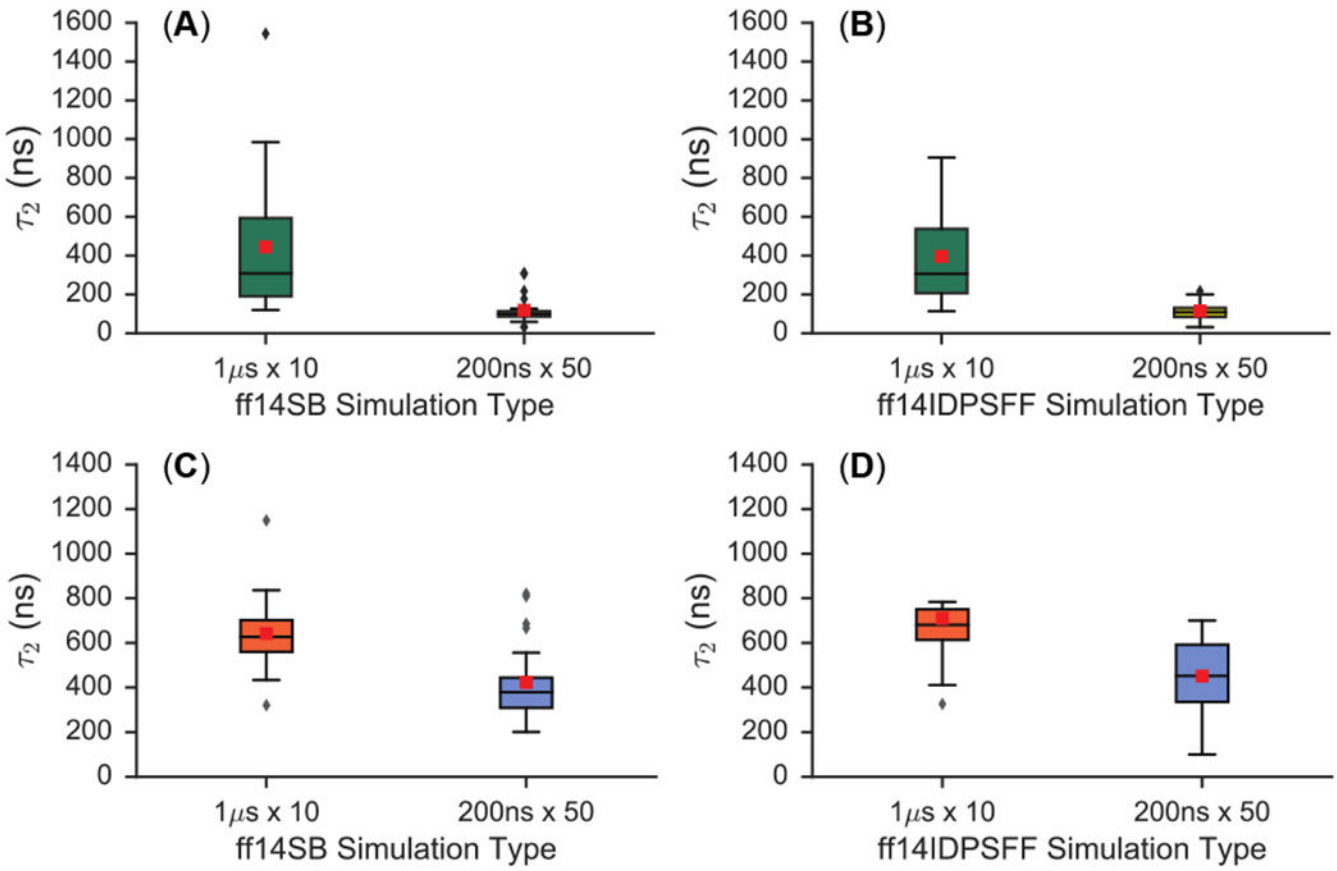


Figure 3.

Summarization of τ_2 values derived from cumulative averages of $\Delta\delta C^{\alpha}$ and $^{3}J_{HNH\alpha}$-coupling constants for apo Rev. Boxplots depict median, range, quartiles, outliers, and averages (red box). (A) Details only *ff*14SB-parameterized simulations of $\Delta\delta C^{\alpha}$-derived τ_2 values. (B) Details only *ff*14IDPSFF-parameterized simulations of $\Delta\delta C^{\alpha}$-derived τ_2 values. (C) Only *ff*14SB-parameterized simulations of $^{3}J_{HNH\alpha}$-coupling-derived τ_2 values are shown. (D) Details only *ff*14IDPSFF-parameterized simulations of $^{3}J_{HNH\alpha}$-coupling-derived τ_2 values.

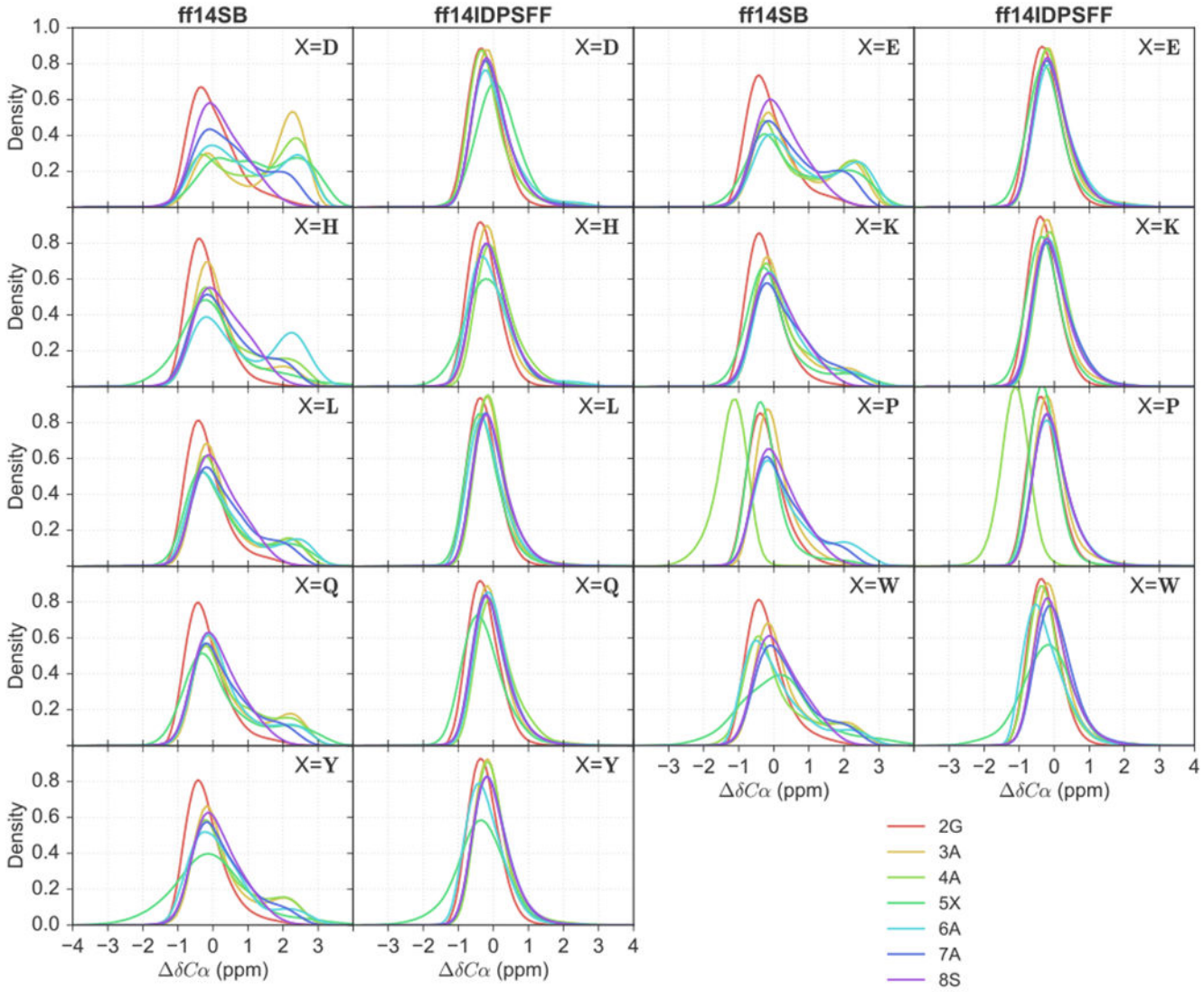


Figure 4.

Kernel density estimations (KDEs) of secondary Cα chemical shift values for 9 short peptides of EGAAXAASS (X = D, E, H, K, L, P, Q, W, Y) and residues 2-8. Residues are colored as indicated in the legend.

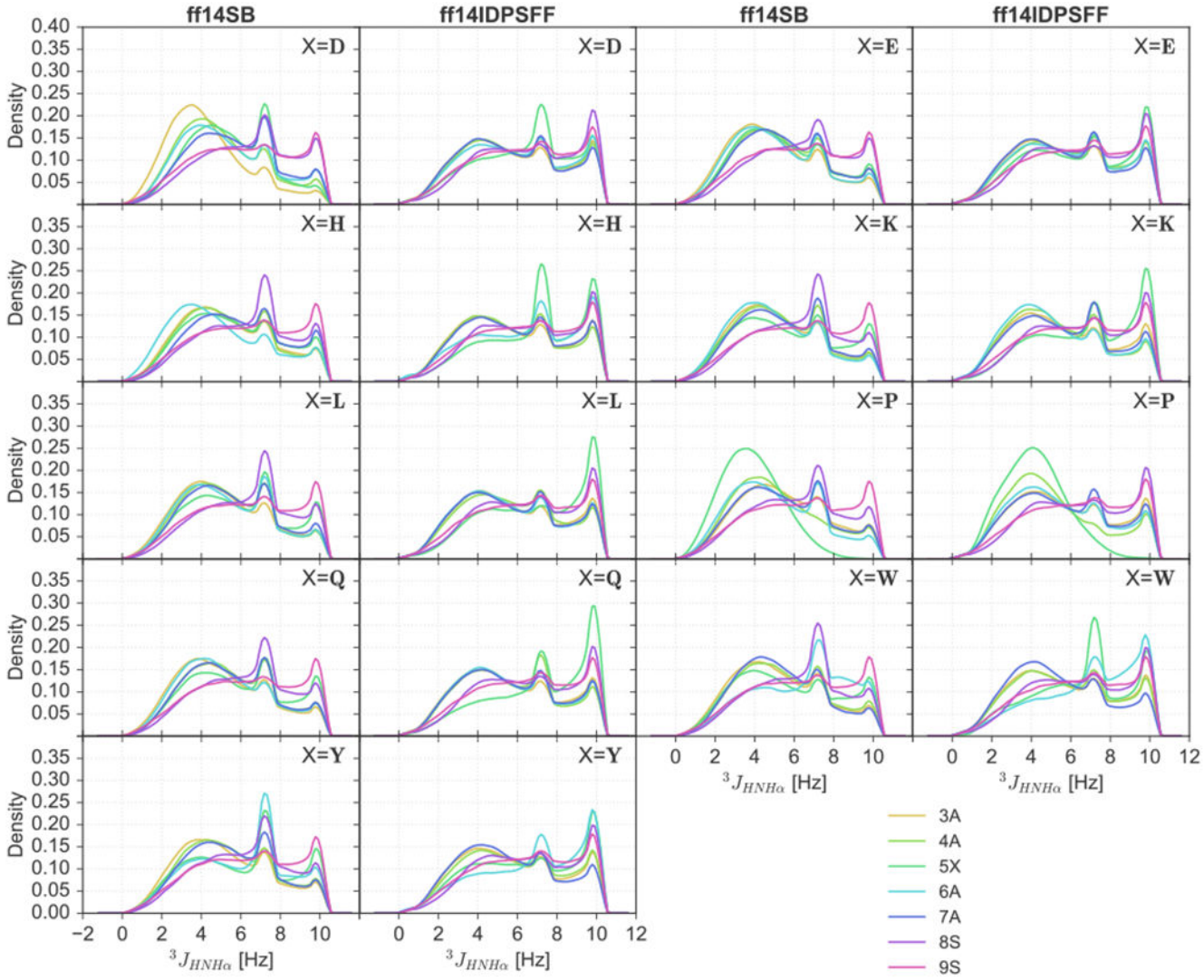


Figure 5.

KDEs of $^3J_{HNH\alpha}$-coupling constants for 9 short peptides of EGAAXAASS (X = D, E, H, K, L, P, Q, W, Y) and residues 3-9.

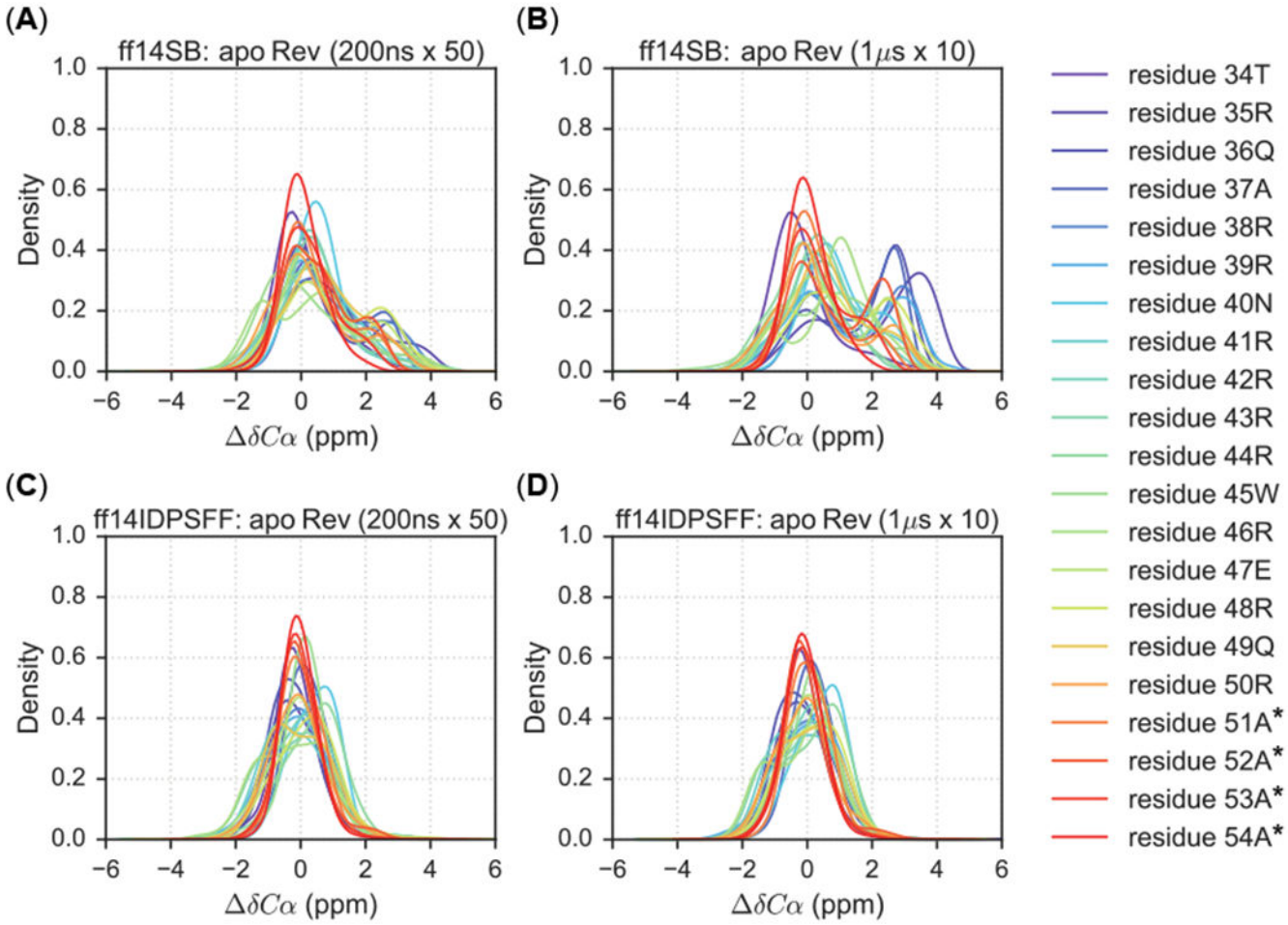


Figure 6.

KDEs of secondary Cα chemical shift values for 1 μs × 10 (long) simulations and 200ns × 50 (short) simulations. Residues are colored according to the legend and simulations are plotted according to the following combination of force field and timescale types: (A) Short simulations using the *ff*14SB force field. (B) Long simulations using the *ff*14SB force field. (C) Short simulations using the *ff*14IDPSFF force field. (D) Long simulations using the *ff*14IDPSFF force field. Asterisks (*) indicate non-native residues.

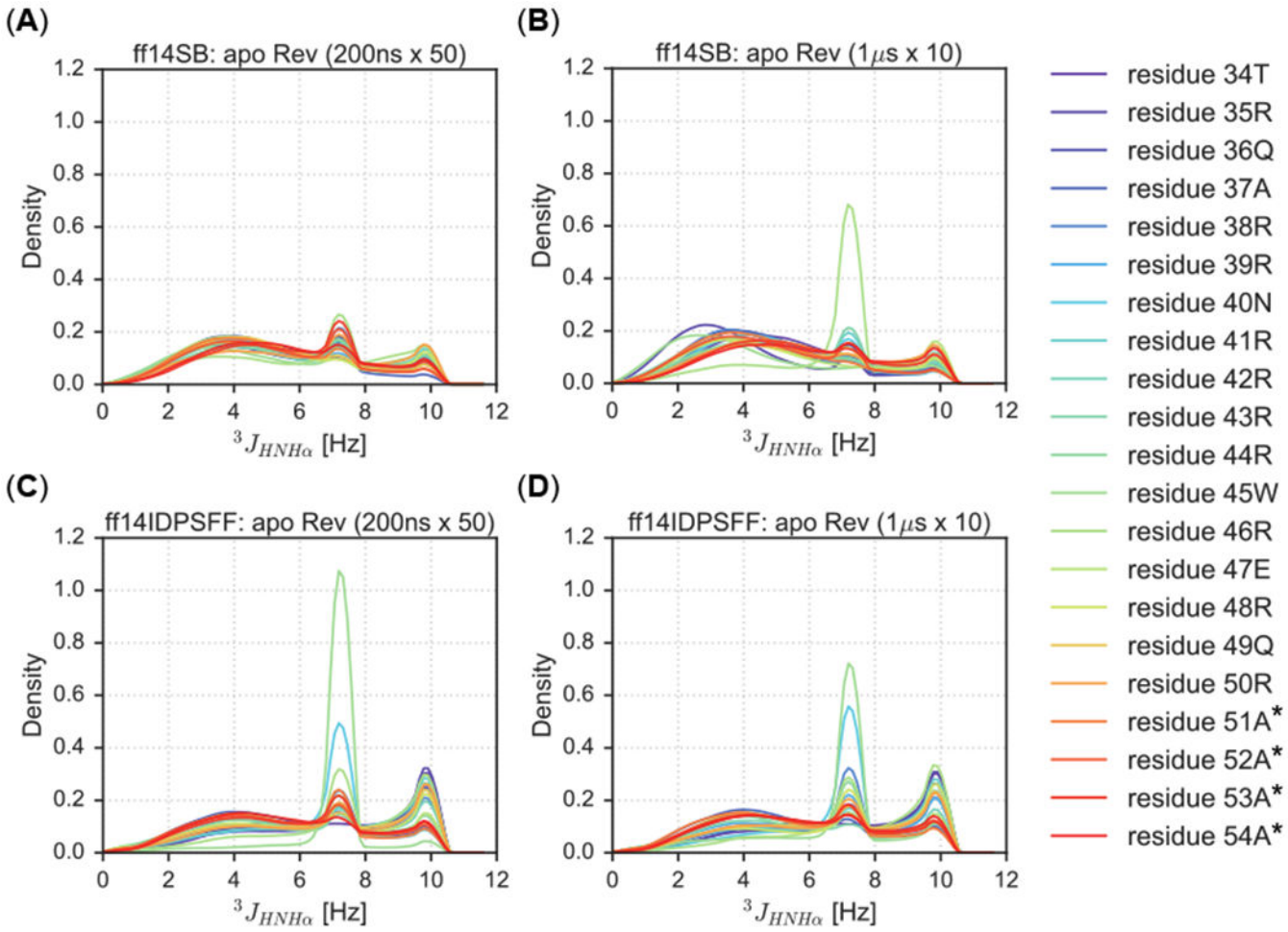


Figure 7.

KDEs of $^3J_{HNH\alpha}$-coupling constants of short (200 ns × 50) and long (1μs × 10) simulations types. Residues are colored according to the legend and simulations are plotted according to the following combination of force field and timescale types: (A) Short simulations using the *ff*14SB force field. (B) Long simulations using the *ff*14SB force field. (C) Short simulations using the *ff*14IDPSFF force field. (D) Long simulations using the *ff*14IDPSFF force field. Asterisks (*) indicate non-native residues.

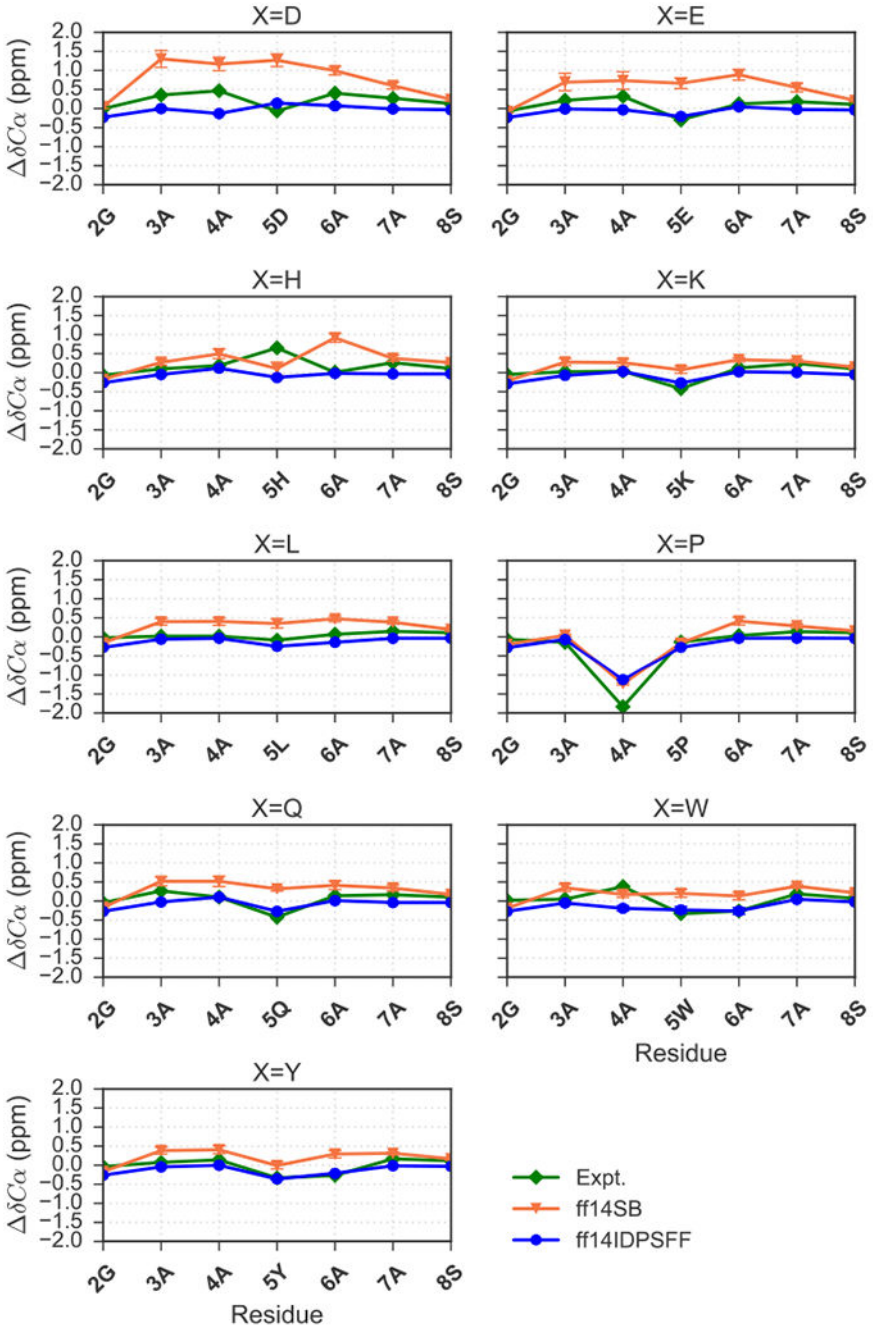


Figure 8.

Comparison of experimental[24, 25] secondary Cα chemical shift values and simulated chemical shifts for the 9 short peptides (EGAAXAASS, X = D, E, H, K, L, P, Q, W, Y). Experimental and simulated values are colored as indicated in the legend. Standard deviation error bars are also visible for simulated values.

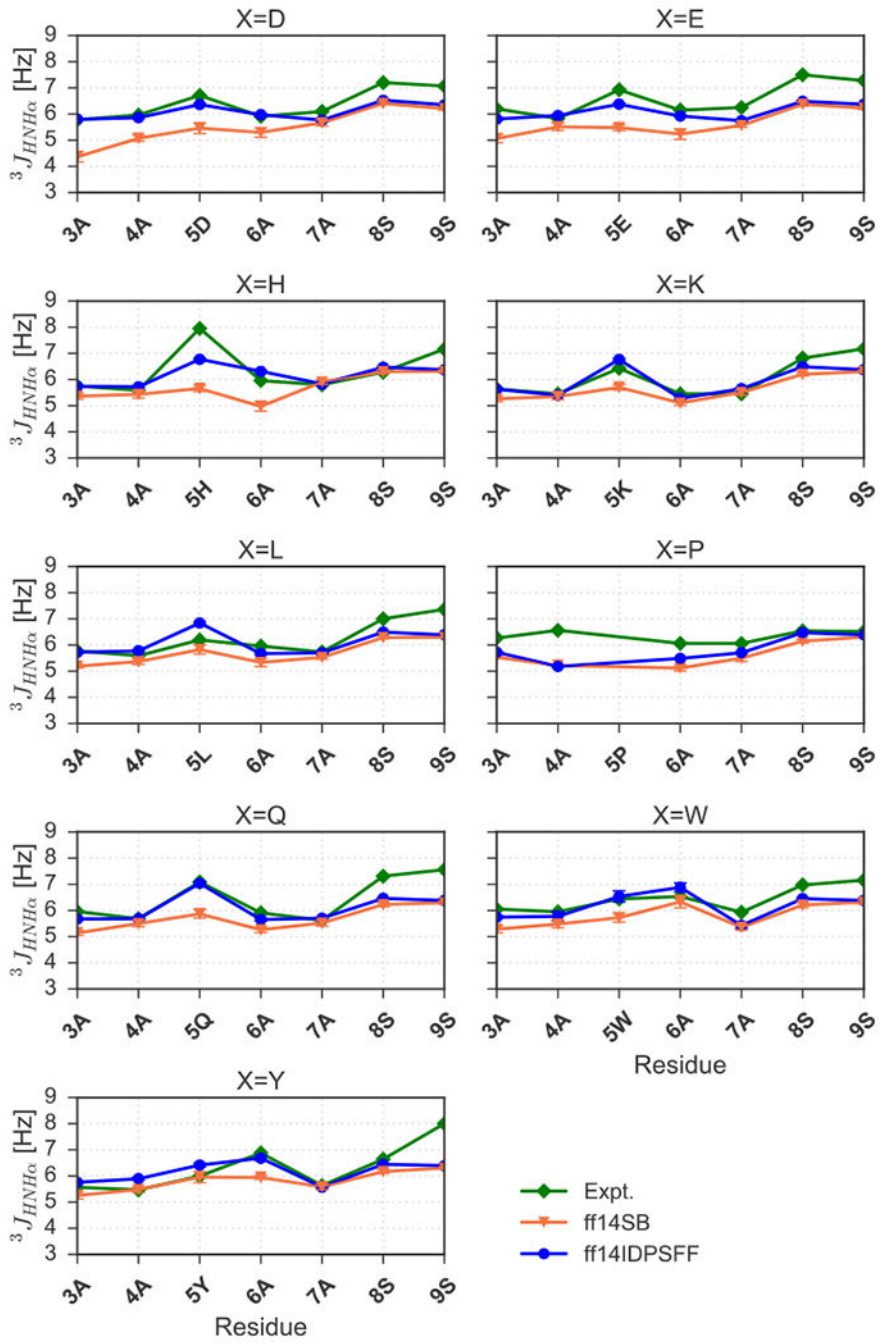


Figure 9.

Calculated *ff*14IDSPFF- and *ff*14SB-parameterized $^3J_{HNH\alpha}$-coupling constants compared to experimentally-derived[24, 25] constants. Experimental and simulated values are colored as indicated in the legend. Standard deviation error bars are also visible for simulated values.

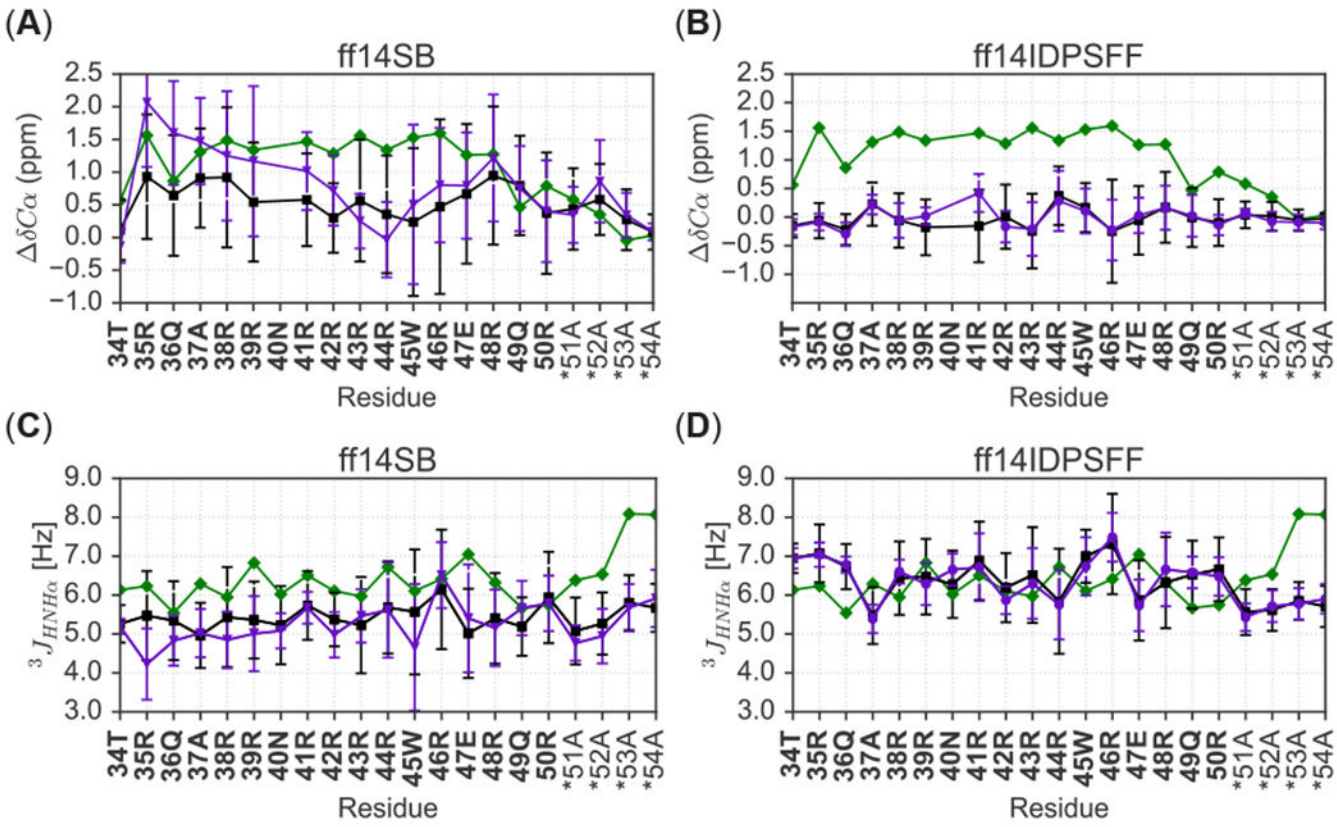


Figure 10.

Comparison of force field and simulation types of apo Rev to experimental results. Colors are labeled according to experiment (green), short simulations (black), and long simulations (purple) and an asterisk (*) denotes non-native residues. (A) Comparison of short and long *ff*14SB-derived secondary chemical shifts with experiment.[31] (B) Comparison of short and long *ff*14IDPSFF-derived secondary chemical shifts to experiment.[31] (C) Comparison of short and long *ff*14SB-derived *J*-coupling constants with experiment.[31] (D) Comparison of short and long *ff*14IDPSFF-derived *J*-coupling constants with experiment.[31]

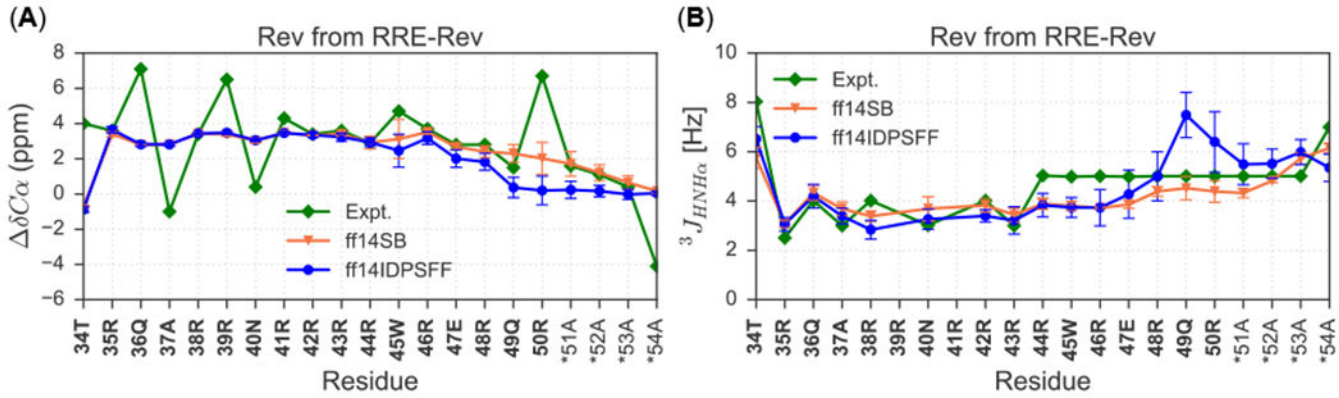


Figure 11.

Simulated NMR observables are superimposed with experimental NMR values of Rev bound to the Stem IIB of RNA-binding partner, Rev-response element. Bold residues indicate native residues and asterisk (*) denotes non-native residues. (A) Comparison of experimental[30, 38] and average simulated $\Delta\delta C^{\alpha}$values. (B) Comparison of experimental[38] and average simulated $^{3}J_{HNH\alpha}$-coupling constants.

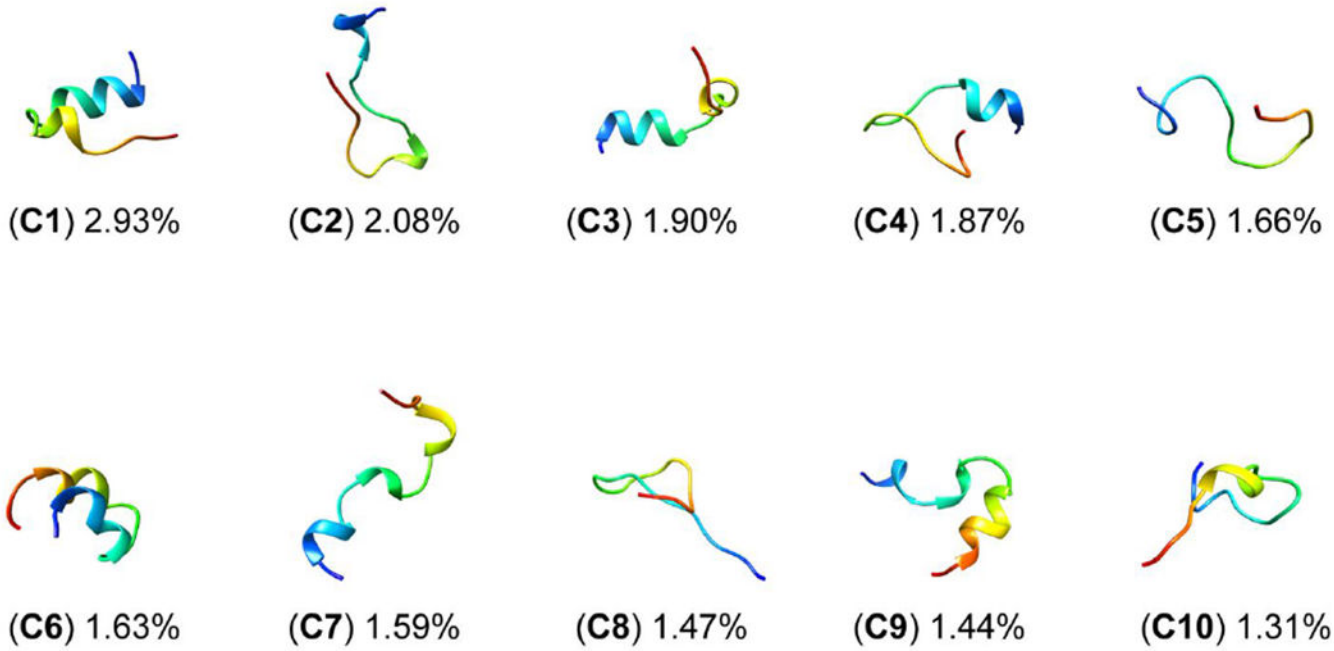


Figure 12.

Top 10 clusters of *ff*14SB-parameterized simulations encompass 17.87% of all frames. Clusters are labeled C1-C10 and colored according to N- to C-termini sequence (red to blue).

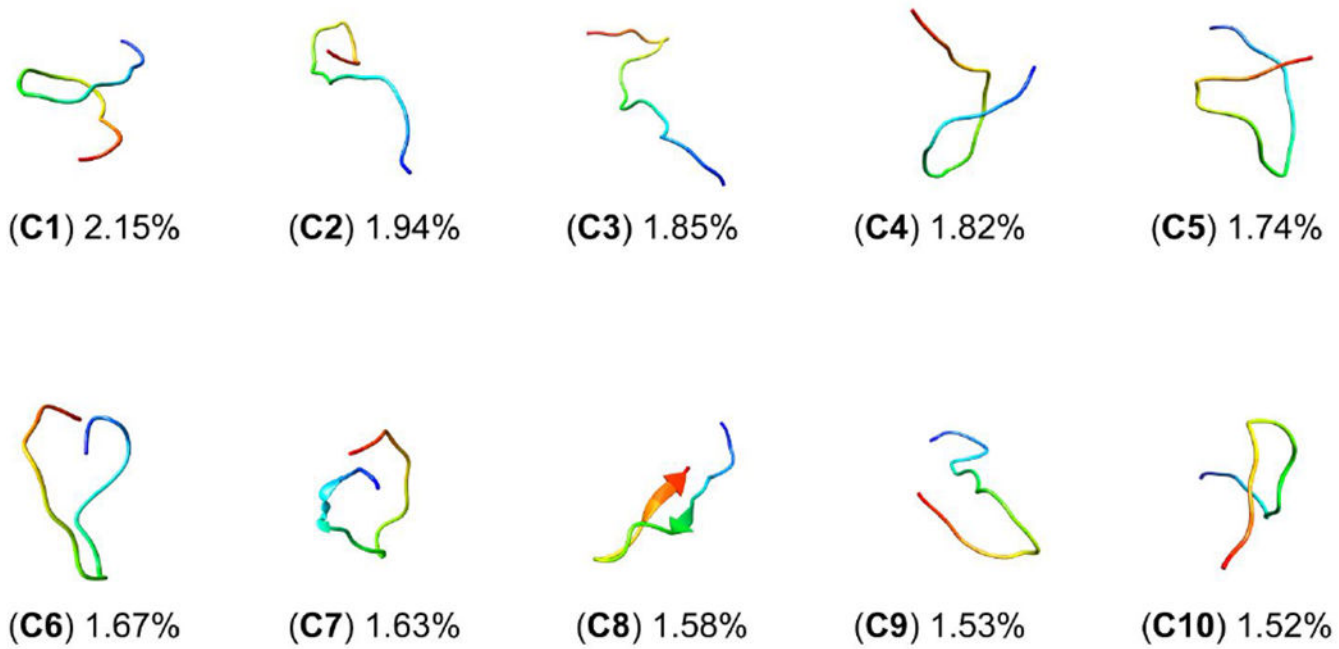


Figure 13.

Top 10 clusters of *ff*14IDPSFF-parameterized simulations encompass 17.41% of all frames. Clusters are labeled C1-C10 and colored according to N- to C-termini sequence (red to blue).

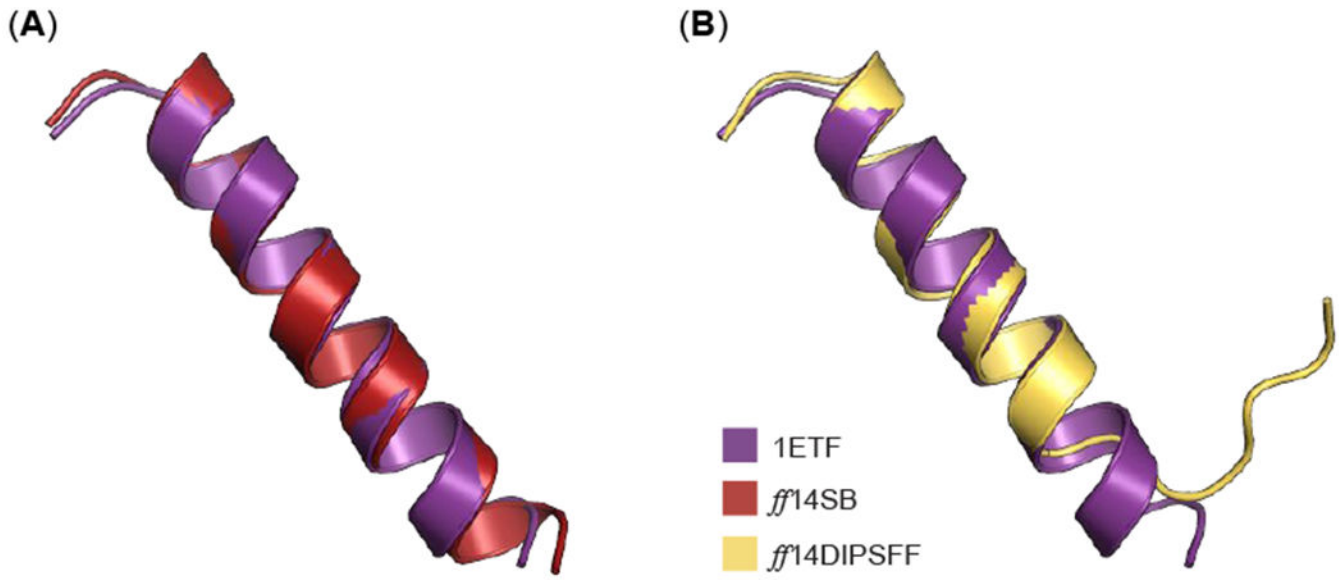


Figure 14.

Alignment of average Rev structure from *ff*14SB and *ff*14IDPSFF RRE-Rev simulations to chain B in the NMR solution structure (PDB: 1ETF). (A) The average structure from *ff*14SB simulations is superimposed to Rev protein from IETF, with an RMSD of 0.57 Å (Cα atoms). (B) The average structure from *ff*14IDPSFF simulations is superimposed to Rev protein from IETF, with an RMSD of 1.14 Å (Cα atoms).

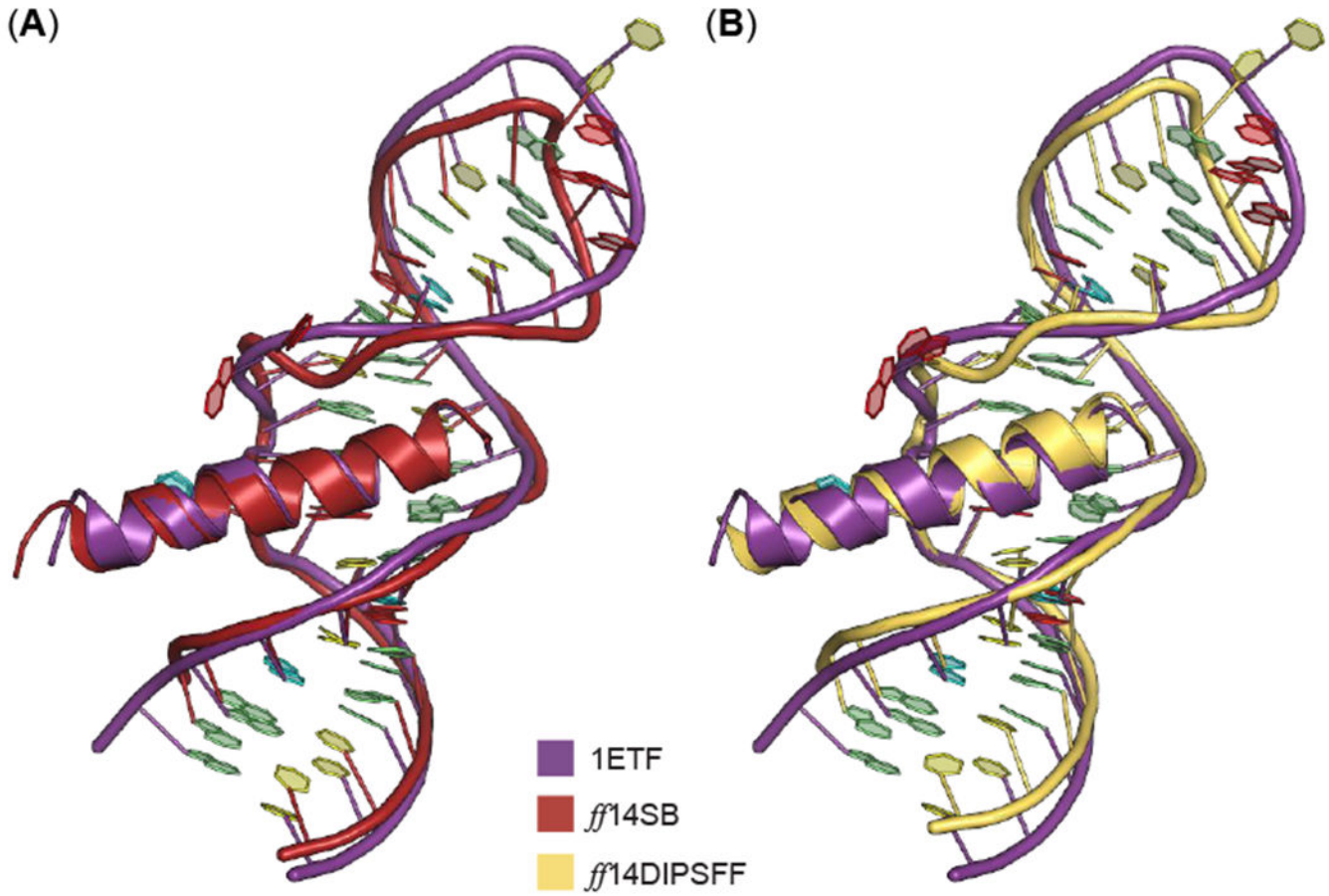


Figure 15.

Alignment of average complex structure from *ff*14SB and *ff*14IDPSFF RRE-Rev simulations to the full NMR solution structure (PDB: 1ETF). Nitrogenous bases are colored according to Nucleic Acid Database convention: A – red, U – cyan, C – yellow, and G – green. (**A**) The average structure from *ff*14SB simulations (red) is superimposed to RRE-Rev from 1ETF, with an RMSD of 1.48 Å (backbone atoms: CA, P, O5′, O3′, C3′, C4′, C5′). (**B**) The average structure from *ff*14IDPSFF simulations is superimposed to RRE-Rev from 1ETF, with an RMSD of 1.9 Å (backbone atoms: CA, P, O5′, O3′, C3′, C4′, C5′).

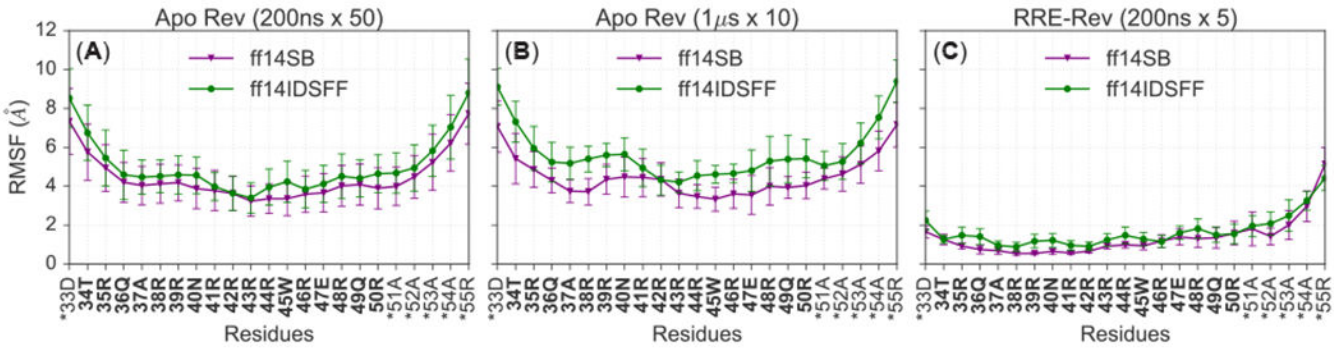


Figure 16.

RMSF analyses of backbone Cα atoms per force field and simulation type. (A) Average RMSF of backbone atoms between fifty, 200ns apo Rev simulations. Asterisks (*) indicate non-native residues. (B) Average RMSF of backbone atoms between ten, 1 μs apo Rev simulations. (C) Average RMSF of backbone atoms between five, 200ns RRE-Rev simulations.

Table 1.

Summary of Simulation Setups

Short peptide	Citations BMRB, PDB	Force fields	Simulation number	Length per simulation	Ions	Waters
EGAA**D**AAS	[24]	*ff14SB*	10	1 μs	1 Na+	1532-2178
S		*ff14IDPSFF*	10	1 μs	1 Na+	1465-2569
EGAA**E**AAS	[24]	*ff14SB*	10	1 μs	1 Na+	1628-2622
S		*ff14IDPSFF*	10	1 μs	1 Na+	1464-3151
EGAA**Q**AAS	[24]	*ff14SB*	10	1 μs	1 Na+	1299-2752
S		*ff14IDPSFF*	10	1 μs	1 Na+	1520-3668
EGAA**W**AAS	24, 25	*ff14SB*	10	1 μs	0	1574-2637
S		*ff14IDPSFF*	10	1 μs	0	1876-3092
EGAA**Y**AAS	[24]	*ff14SB*	10	1 μs	0	1804-2867
S		*ff14IDPSFF*	10	1 μs	0	1888-3141
EGAA**L**AAS	[24]	*ff14SB*	10	1 μs	0	1373-3224
S		*ff14IDPSFF*	10	1 μs	0	1606-3131
EGAA**P**AAS	[24]	*ff14SB*	10	1 μs	0	1751-2713
S		*ff14IDPSFF*	10	1 μs	0	1693-2885
EGAA**H**AAS	[24]	*ff14SB*	10	1 μs	0	1498-2675
S		*ff14IDPSFF*	10	1 μs	0	1430-3159
EGAA**K**AAS	[24]	*ff14SB*	10	1 μs	1 Cl−	1733-2434
S		*ff14IDPSFF*	10	1 μs	1 Cl−	1633-2399
apo Rev (23 amino acids)	($\Delta\delta C^{\alpha}$),[31]	*ff14SB*	10/50	1 μs / 200 ns	9 Cl−	3727-11638
	($^{3}J_{HNHa}$),[31] BRMB:18851[31]	*ff14IDPSFF*	10/50	1 μs / 200 ns	9 Cl−	4424-13224
RRE - Rev complex	($\Delta\delta C^{\alpha}$),[30, 38]	*ff14SB*	5	200 ns	53 Na + 29 Cl−	10928
	($^{3}J_{HNHa}$),[38] PDB:1ETF[30]	*ff14IDPSFF*	5	200 ns	53 Na + 29 Cl−	10928

Table 2.

Average τ_2 Values $\Delta\delta C^{\alpha}$ and $^{3}J_{HNH\alpha}$-coupling constants) of 9-Residue EGAAXAASS with Standard Deviations (SDs)

Protein	Avg. τ_2 ± SD from $\Delta\delta C^{\alpha}$ (ns)		Avg. τ_2 ± SD from $^{3}J_{HNH\alpha}$ (ns)	
	*ff*14SB	*ff*14IDPSFF	*ff*14SB	*ff*14IDPSFF
EGAA**D**AASS	705 ± 134	221 ± 22	679 ± 242	761 ± 187
EGAA**E**AASS	389 ± 63	195 ± 25	639 ± 179	715 ± 179
EGAA**H**AASS	561 ± 104	508 ± 107	686 ± 193	786 ± 279
EGAA**K**AASS	412 ± 68	163 ± 21	570 ± 130	685 ± 183
EGAA**L**AASS	307 ± 50	239 ± 36	692 ± 163	710 ± 185
EGAA**P**AASS	247 ± 31	270 ± 40	716 ± 205	581 ± 181
EGAA**Q**AASS	435 ± 68	437 ± 74	747 ± 225	689 ± 154
EGAA**W**AASS	423 ± 60	343 ± 40	631 ± 113	525 ± 136
EGAA**Y**AASS	511 ± 93	480 ± 77	641 ± 173	687 ± 250

Table 3.

Average τ_2 Values ($\Delta\delta C^{\alpha}$ and $^{3}J_{HNH\alpha}$-coupling constants) of apo Rev and RRE-Rev with SDs

Protein	Avg. τ_2 ± SD from $\Delta\delta C^{\alpha}$ (ns)		Avg. τ_2 ± SD from $^{3}J_{HNH\alpha}$ (ns)	
	*ff*14SB	*ff*14IDPSFF	*ff*14SB	*ff*14IDPSFF
Apo Rev (1 μs × 10)	445 ± 75	396 ± 70	642 ± 166	710 ± 209
Apo Rev (200 ns × 50)	119 ± 73	115 ± 58	422 ± 71	451 ± 67
RRE-Rev (200 ns × 5)	21.8 ± 1.6	24.0 ± 2.3	3.4 ± 0.3	3.6 ± 0.3

Table 4.

RMSE of Calculated Cα Chemical Shifts and $^{3}J_{HNHa}$-coupling Constants with Respect to Experimental Values

Protein	$\Delta\delta C^{\alpha}$ RMSE (ppm)		$^{3}J_{HNH\alpha}$-coupling RMSE (Hz)	
	*ff*14SB	*ff*14IDPSFF	*ff*14SB	*ff*14IDPSFF
EGAA**D**AASS	0.72	0.34	0.95	0.42
EGAA**E**AASS	0.54	0.20	1.01	0.61
EGAA**H**AASS	0.43	0.33	1.01	0.56
EGAA**K**AASS	0.25	0.16	0.53	0.36
EGAA**L**AASS	0.32	0.17	0.61	0.50
EGAA**P**AASS	0.29	0.30	0.79	0.67
EGAA**Q**AASS	0.36	0.18	0.88	0.57
EGAA**W**AASS	0.31	0.26	0.65	0.44
EGAA**Y**AASS	0.30	0.14	0.76	0.66
Apo Rev (1 μs × 10)	0.64	1.16	1.34	1.03
Apo Rev (200 ns × 50)	0.68	1.19	1.17	1.02
RRE-Rev (200 ns × 5)	2.35	2.62	0.90	1.08

Table 5.

Intermolecular Hydrogen Bond Occupancy (criteria: $\theta > 120°$, distance < 2.5Å)[50]

Row Number	Donor Residue	Acceptor Residue	Freq. (*ff*14SB)	Freq. (*ff*14IDPSFF)
0	THR34	G47	0.5926	0.6576
1	ARG35	C65	0.753	0.5848
2	ARG35	U66	0.8388	0.7287
3	GLN36	G48	0.7831	0.6025
4	ARG38	U66	0.9777	0.9303
5	ARG38	G67	0.7867	0.7301
6	ARG39	G70	0.9918	0.9702
7	ASN40	G47	0.8201	0.9814
8	ASN40	G46	0.6765	0.8927
9	ARG41	G46	0.6674	0.7484
10	ARG42	G67	0.8515	0.8345
11	ARG42	A68	0.764	0.8502
12	ARG44	U45	0.7013	0.728
13	ARG46	U72	0.6373	0.4805
14	ARG48	U43	0.8294	0.7139
15	ARG48	C44	0.6949	0.6611
16	GLN36	G47	0.3891	0.5076
17	ARG41	U45	0.4667	0.5766

Table 6.

Intermolecular Ionic Salt Bridge Occupancy (criterion: distance < 4Å)[51]

Row Number	Acidic Residue	Basic Residue	Freq. (*ff*14SB)	Freq. (*ff*14IDPSFF)
0	U43	ARG48	0.8611	0.7314
1	C44	ARG48	0.7535	0.8062
2	U45	ARG44	0.5244	0.5146
3	G46	ARG41	0.7136	0.8019
4	C65	ARG35	0.7934	0.6226
5	U66	ARG35	0.8821	0.7189
6	U66	ARG38	0.9821	0.9527
7	G67	ARG38	0.7981	0.7406
8	G67	ARG42	0.9152	0.9513
9	A68	ARG42	0.7722	0.8712
10	U72	ARG46	0.6879	0.6017
11	U45	ARG41	0.4922	0.5960