Structural Characterization of N-WASP Domain V Using MD Simulations with NMR and SAXS Data

Maud Chan-Yao-Chong, Celia Deville, Louise Pinet, Carine Van Heijenoort, Dominique Durand, Tâp Ha-Duong

Structural characterization of N-WASP domain V using MD simulations with NMR and SAXS data

Maud Chan-Yao-Chong[1,2], Célia Deville[3], Louise Pinet[4], Carine van Heijenoort[4], Dominique Durand[2,*], and Tâp Ha-Duong[1,*]

[1]BioCIS, University Paris-Sud, CNRS UMR 8076, University Paris-Saclay, Châtenay-Malabry, France
[2]I2BC, University Paris-Sud, CNRS UMR 9198, University Paris-Saclay, Orsay, France
[3]IGBMC, University of Strasbourg, CNRS UMR 7104, Illkirch, France
[4]ICSN, CNRS UPR 2301, University Paris-Saclay, Gif-sur-Yvette, France

Abstract

Due to their large conformational heterogeneity, structural characterization of intrinsically disordered proteins (IDPs) is very challenging using classical experimental methods alone. In the present study, we use nuclear magnetic resonance (NMR) and small-angle X-ray scattering (SAXS) data with multiple molecular dynamics (MD) simulations to describe the conformational ensemble of the fully disordered verprolin homology domain (V) of the Neural Wiskott-Aldrick Syndrome Protein (N-WASP) involved in the regulation of actin polymerization. First, we studied several back-calculation software of SAXS scattering intensity and optimized the adjustable parameters to accurately calculate the SAXS intensity from an atomic structure. We also identified the most appropriate force fields for MD simulations of this IDP. Then, we analyzed four conformational ensembles of N-WASP domain V, two generated with the program Flexible-Meccano (FM) with or without NMR-derived information as input, and two others generated by MD simulations with two different force fields. These four conformational ensembles were compared to available NMR and SAXS data for validation. We found that MD simulations with the AMBER-03w force field and the TIP4P/2005s water model are able to correctly describe the conformational ensemble of this 67 residues IDP, at both local and global level.

Received date:

Corresponding authors: dominique.durand@i2bc.paris-saclay.fr and tap.ha-duong@u-psud.fr

1 Introduction

Intrinsically disordered proteins (IDPs) are characterized by one or several regions (longer than 30 consecutive residues) which lack stable secondary and tertiary structure in their unbound state under physiological conditions [1–5]. IDPs are common proteins in all domains of life (Bacteria, Archaea, and Eukaryota), and presumably represent more than 30% of proteins in eukaryotic cells [6–8]. They are also involved in many human diseases as shown by Uversky *et al.* who estimated that about 65% of proteins involved in cancer and diabetes, 55 % of those in cardiovascular diseases, and 50% of those in neurodegenerative diseases are IDPs [9]. IDPs frequently play important roles in the regulation of biological processes, including cell growth, cell signalling, and cell survival [10]. To exert their functions, IDPs often interact with several biomolecular partners, thanks to the large flexibility of their intrinsically disordered regions (IDRs), and are generally considered as important hub in protein-protein interaction networks.

Mechanisms of IDP association to protein partner are very diverse [11–13]. Upon binding, IDRs can retain complete or partial disordered states, or can adopt an intricate binding mechanism giving rise to ”fuzzy” complexes [14–16]. Nevertheless, in almost 30% of IDP-protein complexes, it is observed that regions, which are disordered in the unbound state, adopt well structured conformations in the bound state [17]. Regions that undergo a disorder-to-order transition upon binding are called Molecular Recognition Features (MoRFs) [4, 18–22]. Their formation can follow two limiting mechanisms, not necessarily exclusive [23]: In the first one, named ”induced fit”, the disordered region binds to the protein partner and folded into an ordered structure on its surface. In the second one, referred to as ”conformational selection”, the folded structure preexist in

0006-3495/08/09/2624/12 $2.00

the conformational ensemble of the unbound IDP and are recognized by the protein partner. To better understand the mechanisms of formation of IDP-protein complexes and how their interactions are modulated by the IDP structure and dynamics, it is thus crucial to explore the conformational ensemble of IDPs.

However, due to their large conformational heterogeneity, the structural characterization of IDPs is very challenging using classical X-ray crystallography or NMR NOE analysis. Nevertheless, secondary chemical shifts and residual dipolar couplings from NMR experiments can provide local information about the propensity of each residue to form transient secondary structures [24]. On the other hand, small-angle X-ray scattering (SAXS) can deliver global information about IDP structures in terms of average size and shape [25]. But, in order to infer a detailed conformational ensemble from these NMR and SAXS data, it is most often necessary to use complementary *in silico* approaches to generate structures at the atomic scale, such as statistical coil generator or molecular dynamics (MD) simulations [26, 27]. At this stage, the combination of experimental data and *in silico* techniques can be performed in three different ways: conformation ensembles can be generated using experimental data as input restraints, conformational ensembles can be generated by selection of structures fitting the experimental data, or conformational ensembles can simply be validated (or not) against experimental data [28, 29]. In the present study, we applied two of these combined approaches to the fully disordered verprolin homology domain (V) of the Neural Wiskott-Aldrick Syndrome Protein (N-WASP). Principally, we assessed the ability of MD simulations to generate a consistent conformational ensemble of N-WASP domain V by validation against both NMR and SAXS data. Notably, we raised the question of whether physics-based modeling can account for the formation of transient secondary structures and MoRFs even in the absence of experimental data.

N-WASP is a 65 kDa protein dominantly expressed in the brain. Its sequence shows about 50% homology to the WASP protein produced in hematopoietic cells and implicated in the Wiskott-Aldrich Syndrome [30, 31]. Like the other WASP-family proteins, N-WASP is a pivotal player in the regulation of actin cytoskeleton dynamics and organization [32]. The human N-WASP sequence (505 residues) can be decomposed into 7 domains: a primary WASP homology domain WH1 (segment 1-150), a basic domain B (186-200), a GTPase-binding domain GBD (203-274), a proline-rich domain PRD (277-392), a verprolin homology domain V (405-450), a cofilin homology domain C (451-485), and an acidic domain A (486-505). It should be noted that domain V is composed of two secondary WASP homology domains or motifs WH2, each of them being able to bind a monomer of actin [33]. Indeed, Gaucher *et al.* demonstrated that, in presence of actins, peptides constructed from N-WASP domain V or VC form 1:2 V-actin or VC-actin complexes [34]. Data reported by Gaucher *et al.* indicate as well that the second WH2 motif of N-WASP domain V has a higher binding affinity for actin than the first one. This is also suggested by the crystallographic structure of N-WASP segment VC (392-484) in complex with actin (PDB ID: 2VCP) [34], which only shows the second WH2 motif bound to an actin monomer.

In the latter X-ray structure, it could be noted that the N-terminal part of the bound WH2 motif (433-444) is folded into an amphiphilic α-helix located in the cleft between actin subdomains 1 and 3, whereas its highly conserved sequence LK[K-S]V (445-451) [35, 36] has a rather extended conformation. The absence in the crystallographic structure of atomic coordinates for regions flanking the segment 433-451 indicates that they keep a disordered conformation upon binding to actin. Thus, N-WASP domain V is a representative case of IDPs with a helical molecular recognition feature (MoRF) which binds a protein partner. To gain insight into the mechanism and physical driving forces of IDP-protein complex formation, we investigated the conformational ensemble of a 67 residues construct derived from the N-WASP domain V (Fig. 1) using MD simulations validated with NMR and SAXS data.

2 Methods

All details regarding the experimental part of this work (protein preparation, NMR spectroscopy, and SAXS experiments) are reported in Supplementary Information. We only present here the computational tools used for this study. Indeed, it is often emphasized that comparisons and validations of results from simulations against experimental data require robust and accurate methods to back-calculate experimental observables from conformational ensembles. This section particularly surveys several software for the back-calculation of SAXS intensities. It also describes the tested force fields and details of the MD simulations.

2.1 Back-calculation of NMR observables

Proton and heavy atom chemical shifts (CS) of a given protein structure can be calculated using various software, such as SHIFTS [37, 38], SHIFTX [39], SPARTA [40], or CAMSHIFT [41]. Then, structure-dependent CS can be averaged over the protein conformational ensemble and directly compared with NMR measurements. All these software were shown to be

satisfactorily accurate and sensitive [42–46]. In the present study, we chose the program SHIFTS [37] to compute CS for each structure of N-WASP domain V generated by *in silico* methods.

To identify transient secondary structures in IDPs, it is rather useful to analyze secondary chemical shifts which are defined as variations of experimental chemical shifts relative to random coil values. Several sets of residue-specific random coil values for each proton and heavy atom type are available in the literature, such as the commonly used random coil shifts from Wishart *et al.* [47] or the more recent RefDB from Zhang *et al.* [48]. We chose the latter for the present study. Furthermore, instead of analyzing secondary chemical shifts from the different nuclei of the protein, we combined those of the Cα, Cβ, CO, and N atoms into a residue-specific single secondary propensity score using the SSP software from Marsh *et al.* [49] with the default settings. For a given residue, the SSP score is between -1 and 1, the positive and negative values indicating the proportion of helical and extended structures in the protein conformational ensemble, respectively.

When an IDP is weakly aligned in a slightly anisotropic environment, such as liquid crystalline or anisotropic gel phases, NMR measurements of residual dipolar couplings (RDCs) also allow to detect transient local secondary structures [50]. To make direct comparisons with experimental values, the program PALES [51] was applied to the sampled structures of N-WASP domain V to compute residue-specific ^{1}H-^{15}N RDC theoretical values. It should be noted that RDCs from NMR measurements were scaled uniformly by a factor of 1/3 to account for an experimental 2H quadrupole splitting of 30 Hz instead of 10 Hz in the back-calculation program.

2.2 Back-calculation of SAXS intensities

The scattering intensity of a protein, obtained as the scattering curve of the solution of protein minus the scattering of the buffer, not only accounts for the protein electrons but also for those of solvent molecules in excess or in deficiency at the protein surface with respect to the buffer electronic density. Thus, back-calculations of the intensity scattered by a protein first imply to correctly calculate the volume of solvent excluded by the protein, which is not straightforward. A second difficulty is to accurately estimate the scattering contribution from the protein hydration layer [52–54]. Two types of approaches were developed to calculate SAXS intensities from atomic structures, implicit solvent and explicit solvent methods (see Table 1 of Ref. [53] for a list of the main software). Explicit solvent approaches are generally time-consuming and thus difficult to be applied to very large numbers of atomic structures. Software based on implicit models of solvation reduce the calculation computational costs but require to fix free parameters, in particular the solute hydration layer density, to best fit experimental data. Moreover, as reported by Grudinin *et al.* (see Table 9 of Ref. [55]), parameters for the hydration layer density, which were generally optimized by fitting calculated curves on data obtained on crystallographic or NMR structures, vary significantly from one protein to another. The challenge is thus to find an optimal value for the hydration layer density of the intrinsically disordered N-WASP domain V which can adopt a great variety of conformations. We thus decided to assess the hydration layer density of implicit solvent software by comparing to an explicit solvent approach which does not require to adjust this parameter.

Among the recent freely available explicit solvent SAXS software, we chose the program WAXSiS [56, 57] as a reference, similarly to two other recent studies [54, 58]. This approach performs short MD simulations (15-250 ps) of explicit water molecules in the solute hydration layer (of 7 Å thickness). Protein scattering intensities are then calculated by subtracting the buffer scattering which is determined from MD simulations of pure solvent. Regarding implicit solvent approaches, we applied in this work the most used software and tested several parameters for the hydration layer density:

- CRYSOL [59]. The scattering contribution of the solute hydration layer can be tuned using a parameter $\Delta\rho$ which represents the excess of electron density in the hydration layer compared to the bulk (whose density is fixed to 0.336 e/Å^3). We tested here several values of $\Delta\rho$ ranging from 0.01 to 0.03 (default value) e/Å^3.

- FoXS [60, 61]. The hydration layer density is adjusted using a parameter -2.0 $\leqslant$ c2 $\leqslant$ 4.0. According to the original paper [60], the values c2 = -2.0, c2 = 0.0 (default) and c2 = 4.0 correspond to a hydration layer density of 0.307, 0.334, and 0.388 e/Å^3, respectively. We assessed several values of this parameters between -2.0 and 1.0, corresponding to $\Delta\rho$ values varying from -0.027 to 0.014 e/Å^3.

- CRYSOL3 [62]. Compared to the original version, the last CRYSOL enables to more finely tune the solute hydration shell by using three parameters: c1 associated to solvent in the vicinity of protein convex surface, c2 associated to solvent in concave surface, and c3 associated to solvent trapped inside protein cavities. By default, c1 = c2 = 1.0 and c3 = 0.0. We tested several combinations of these parameters with 0.0 $\leqslant$ c1 $\leqslant$ 1.0, 0.6 $\leqslant$ c2 $\leqslant$ 1.0, and 0.0 $\leqslant$ c3 $\leqslant$ 1.0. It should be noted that CRYSOL3 also allows to choose either spherical harmonics or a cubic method to estimate the solvent excluded volume [63, 64]. These two options were included in our benchmark.

- Pepsi-SAXS [55]. In its new version (0.8) (https://team.inria.fr/nano-d/software/pepsi-saxs), Pepsi-SAXS allows to adjust the hydration shell contrast in percentage of the bulk density (fixed to 0.334 e/Å^3). We assessed several values of this new parameter between 1.0 and 5.0% (default value), corresponding to $\Delta\rho$ ranging from 0.0033 to 0.0167 e/Å^3.

Comparisons between WAXSiS and implicit solvent approaches were performed by calculating and averaging SAXS intensities over a limited number of N-WASP domain V structures representative of its conformational ensemble. These latter were chosen among the conformations sampled by extended MD simulations as follows: First, using the program Flexible-Meccano (see section 2.4), we generated ten thousands three-dimensional structures of N-WASP domain V and then selected 20 of them with radius of gyration regularly incremented by 1 Å from 15 to 34 Å and without any secondary structures. Then each of these 20 conformations was submitted to 100 ns MD simulations (see section 2.3.3). From each of these 20 trajectories, we picked out the frames at times $t = 0$ ns (after 2 ns of equilibration), $t = 50$ ns, and $t = 100$ ns, yielding an ensemble of 80 representative conformations (after adding the 20 initial structures provided by FM). This "semi-manual" procedure was used to build up a small pool of structures as diverse as possible, in terms of compactness, for the SAXS software benchmark. The pool distribution of radius of gyration has a bell shape centered around an average of 23.6 Å, close to the experimental value of 24.3 Å (Fig. S1 of Supplementary Information). In comparison, the distributions of radius of gyration computed over two other pools of 80 conformations generated by FM, without or with constraints on secondary structure propensities (see section 2.4), are shifted toward more compact conformations, with average values of 20.9 and 21.7 Å, respectively (Fig. S1).

SAXS intensities I(q) were calculated for $N = 101$ values of the scattering vector q from 0.0 to 0.5 Å^{-1} and averaged over the previously selected 80 conformations of N-WASP domain V. To quantify the agreement between the implicit solvent SAXS intensities ($I_{Im}(q)$) and the WAXSiS calculations ($I_{Wx}(q)$), we computed and ranked χ^2 values defined in Eq. (1). In the latter, $\sigma_{Wx}(q)$ denotes the standard deviation of $I_{Wx}(q)$ calculated by averaging individual profiles computed from MD simulation frames between which protein side chains and water molecules fluctuate whereas backbone atom positions are restrained (J. Hub, personal communication, see also Ref. [56]).

$$\chi^2 = \frac{1}{N}\sum_q \left(\frac{I_{Im}(q) * \frac{I_{Wx}(0)}{I_{Im}(0)} - I_{Wx}(q)}{\sigma_{Wx}(q)}\right)^2 \quad (1)$$

In Fig. 2 are plotted these χ^2 values obtained for each software. It is first observed that SAXS intensities calculated with the default parameters of the four tested software are not in good agreement with calculations using WAXSiS (shaded bars of Fig. 2). But the fits can be significantly improved by changing the values of these parameters: For the first version of CRYSOL, the best agreement with WAXSiS is obtained with an excess density of $\Delta\rho$ = 0.013 e/Å^3 instead of 0.03 e/Å^3. The FoXS program provides best results with the parameter c2 = 0.1 ($\Delta\rho$ = 0.001 e/Å^3) compared to the default value c2 = 0.0. The best fit observed with CRYSOL3 is obtained by using the cubic method with the parameters (c1, c2, c3) = (1.0, 0.7, 0.0). Finally, Pepsi-SAXS yielded the best agreement with an excess density of 2.2% ($\Delta\rho$ = 0.0073 e/Å^3) instead of 5%.

To check if the above values depend on the number of conformations (80) used in the hydration layer density optimization, we computed χ^2 values between WAXSiS and the two software CRYSOL and Pepsi-SAXS on various pools of size ranging from 20 to 100 conformations. (For building the pool of 100 structures, we additionally selected the 20 frames at time $t = 20$ ns of the 20 MD trajectories). Results displayed in Fig. S2 show that χ^2 values between WAXSiS and both CRYSOL and Pepsi-SAXS weakly depend on the size of the pools when the default parameters for the hydration layer density are used. When using optimized parameters, χ^2 rapidly converged, in all cases, to very small values.

Strikingly, our tests confirmed that small differences in the excess density of protein hydration layer between different programs, and even within a given one, induced large variations of χ^2 values [54]. To gain a better insight into the influence of $\Delta\rho$ values on scattering profiles, we displayed in Fig. S3 the SAXS intensities calculated using CRYSOL and Pepsi-SAXS with the default and optimized parameters for the hydration layer density. As can be seen, both CRYSOL and Pepsi-SAXS with default parameters overestimate scattering intensities at small angles, and thus reflect an overly dense protein hydration layer. Nevertheless, for both CRYSOL and Pepsi-SAXS, optimization of the hydration layer density reduces the disagreement with the explicit solvent SAXS calculations.

Comparing the four tested implicit solvent approaches, our results on N-WASP domain V indicate that an excellent agreement with WAXSiS is obtained by using CRYSOL with $\Delta\rho$ = 0.013 e/Å^3, the recent software Pepsi-SAXS with a hydration layer contrast of 2.2%, or CRYSOL3 with parameters c1 = 1.0, c2 = 0.7, and c3 = 0.0. Besides, we observed that Pepsi-SAXS was significantly faster than the other software, notably CRYSOL (0.01 s for Pepsi-SAXS versus 2 s for CRYSOL on a single structure of N-WASP domain V). More importantly, Pepsi-SAXS takes into account solvent densities in protein cavities and concave surfaces in a better way than the original version of CRYSOL [55, 59]. For these reasons, we decided to use Pepsi-SAXS thereafter for SAXS intensity back-calculations of N-WASP domain V conformations generated by *in silico* techniques.

2.3 MD simulations

2.3.1 General conditions

All molecular dynamics (MD) simulations were performed with the GROMACS software (versions 5.0.2 and 2016.1) [65]. Each initial conformation of N-WASP domain V was put and solvated in a dodecahedral rhombic box of 14.0 nm edge, then neutralized by adding 175 sodium and 176 chloride ions to reach the salt concentration of 150 mM. The non-bonded interactions were treated using the smooth PME method [66] for the electrostatic terms and a cutoff distance of 1.2 nm for the van der Waals potentials. All solute and water covalent bond lengths were kept constant using the LINCS [67] and SETTLE [68] algorithms, respectively, allowing to integrate the equations of motion with a 2 fs time step. All simulations were run in the NPT ensemble, at T = 310 K and P = 1 bar, using the Nose-Hoover and Parrinello-Rahman algorithms [69–71] with the time coupling constants $\tau_T = 0.5$ ps and $\tau_P = 2.5$ ps.

2.3.2 Protein force fields and water models

Historically, protein force fields were developed to simulate the conformational dynamics of folded proteins which generally have their non-polar residues buried in their core and protected from solvent. Recently, several force fields were improved to properly generate conformational ensembles of IDPs which non-polar residues are frequently exposed to solvent. This is generally achieved by using a four-site water model which better accounts for the electric properties of water molecule, and by accentuating the depth of the solute-solvent Lennard-Jones (LJ) potentials to better solvate non-polar residues.

This improvement was first reported by Best *et al.* who proposed to rescale by a factor γ = 1.1 the LJ parameters ϵ_{Oi} between the water oxygen of model TIP4P/2005 and the protein atoms of force field AMBER-03w [72]. (In order to differentiate these modified water-solute interactions from those arising with the original model TIP4P/2005 [73], we denoted this new water model TIP4P/2005s. The combination of AMBER-03w with TIP4P/2005s is named A03ws [72]). In the same spirit, Piana *et al.* derived from the model TIP4P/2005 another water model, named TIP4P/D, characterized by an increased dispersion coefficient c6 = 900 kcal/mol.Å^6 instead of c6 = 736 kcal/mol.Å^6, to better simulate the protein disordered states [74]. Lastly, in the communication of their improved force field CHARMM36m, Huang *et al.* alternatively suggested to increase the LJ potential depth of the TIP3P hydrogen atoms, from ϵ_H = 0.046 kcal/mol to ϵ_H = 0.10 kcal/mol, leaving the other LJ interactions unchanged [75]. (As previously, to differentiate these modified water-solute parameters from those arising with the original model TIP3P [76], we denoted this new water model TIP3Pm, and the combination of CHARMM-36m with TIP3Pm will be referred to as C36mm).

Force field code	Protein model	Water model
OPLS	OPLS-AA [77]	SPC/E [78]
A99sd	AMBER-99sb-ildn [79]	TIP4P/D [74]
A03ws	AMBER-03w [80]	TIP4P/2005s [72]
C22cd	CHARMM-22-cmap [81]	TIP4P/D [74]
C36m	CHARMM-36m [75]	TIP3P [76]
C36mm	CHARMM-36m [75]	TIP3Pm [75]

Table 1: All-atom force fields for protein and water molecules tested on N-WASP domain V.

Since these force field developments were quite recent and probably no universal protein-water models will be valid for all IDPs, we first rapidly tested and compared these new protein and water models along with classical ones, before running long MD simulations of our IDP. Preliminary tests consisted in performing short MD simulations (100 ns) with the six force fields listed in Tab. 1 from an initial conformation of N-WASP domain V with a radius of gyration equal to the one determined from SAXS data (see below section Results). Then time evolutions of the protein radius of gyration were analyzed and their deviations from the initial experimental value were compared. Results of this test (Fig. 3) show that, with OPLS or C22cd force fields, N-WASP domain V rapidly collapses into overly compact conformations, yielding an averaged radius of gyration much lower than the experimental value. With A99sd or C36m force fields, the averaged radii of gyration were closer to the experimental value than with OPLS or C22cd, but populations of extended conformations remained largely minor. Actually, only the two force fields A03ws and C36mm allowed to significantly sample both compact and extended conformations (with radius of gyration below and above the experimental value), yielding average values in fair agreement with SAXS data.

All together, despite the short time of these preliminary MD simulations, the two force fields A03ws and C36mm seem to be more appropriate than the other four to correctly explore the conformational space of N-WASP domain V. Because of our

limited computational resources, we thus decided to use only these two force fields to run more extended MD simulations for more exhaustive sampling.

2.3.3 Extended simulations and analyses

Extended calculations consisted in running 20 independent MD simulations starting from 20 different initial conformations of N-WASP domain V. To ensure the diversity of the 20 initial structures, we took advantage of the Flexible-Meccano program (see section 2.4) which can rapidly build various statistical coil conformations of the protein: We first generated ten thousands three-dimensional structures of N-WASP domain V and then selected 20 of them with radius of gyration regularly incremented by 1 Å from 15 to 34 Å in the pool provided by FM. It should be noted that none of the 20 selected initial conformations have any secondary structures.

Each of the 20 initial random coil conformations was submitted to 2 ns of equilibration followed by 100 ns of production within the general conditions previously described, yielding an accumulated trajectory of 2 μs. It is worthy to noted that running multiple MD simulations from diverse initial structures allows to efficiently reach convergence of the IDP conformational sampling. To verify that, we computed the residue-specific SSP scores and RDCs over the four time windows 0-40, 20-60, 40-80, and 60-100 ns of the 20 MD trajectories of N-WASP domain V. As displayed in Fig. S4, SSP profiles over the four time windows are very similar, including the first 40 ns. Regarding RDCs, profiles averaged over the last two time windows (40-80 and 60-100 ns) are slightly different from those calculated over the first two periodes (0-40 and 20-60 ns), but appear quite close to each other. This block analysis indicates that, using 20 MD simulations of 100 ns, the conformational sampling seemed to have converged after few tens of nanoseconds.

To minimize the possible bias induced by our selection of initial conformations, we only kept the last 80 ns of each MD simulation and collected data every 40 ps yielding ensembles of 40 000 structures for subsequent analyses. Most of the conformational analyses were performed using GROMACS tools. Nevertheless, it should be emphasized that protein radii of gyration were not calculated with the GROMACS tool *gmx gyrate*, but from the SAXS intensities using the Guinier approximation $Log[I(q)/I(0)] = -R_g^2 q^2/3$ when $q \to 0$ [82].

Finally, the software STRIDE [83] was used to assign secondary structure elements to each residue of each protein conformation, based on hydrogen bond criteria and backbone dihedral angle values. Then outputs from STRIDE were used to compute the probabilities for each residue to be in α-helix or β-strand within each conformational ensemble.

2.4 Flexible-Meccano

In this study, we compared N-WASP domain V conformational ensembles derived from MD simulations against those generated by a much faster statistical coil generator, Flexible-Meccano (FM) [84] which is very popular within RMN and SAXS communities. From a primary sequence, FM constructs protein three-dimensional structures by linking consecutive residues with dihedral angles ϕ / ψ randomly taken from a database of residues with only loop conformations [85, 86]. Residue-specific hard-spheres located at Cβ atoms (Cα for Gly) [87] are used to avoid steric clashes between amino acids (if a steric clash occurs between a newly added residue and the previously built polypeptide chain, another pair of dihedral angles ϕ / ψ is randomly chosen). No attractive potentials are taken into account during the generation process. Optionally, restrained conformational ensembles can be produced by specifying, in the FM inputs, fixed secondary structure propensities for given fragments of the sequence (for these residues, the dihedral angles ϕ / ψ are constrained to adopt standard α or β values) [84].

It could be noted that FM only generates structures of the protein backbone. Thus, for each protein conformation, side chains were subsequently added using the SCWRL4 program [88]. Finally, the complete atomic structures were generated by adding the lacking hydrogen atoms with the GROMACS tools *pdb2gmx* [65]. In this work, we used FM to generate two different pools of 40 000 conformations of N-WASP domain V, one without constrained secondary structure propensity and the other one by specifying two α-helical propensities of 13% and 31% for segments 10-19 and 39-47, respectively, as suggested by NMR data (see below section Results).

2.5 Conformational sub-ensemble selection

When experimental data on the studied IDP are available, it is possible to filter a given conformational ensemble and select a sub-ensemble that better agrees with experiments [29]. In this work, we used the program GAJOE from the suite EOM [89, 90] for selecting conformational sub-ensembles of N-WASP domain V that better fit the SAXS intensities. It should be noted that GAJOE was not employed here in the conventional way: We did not search for a sub-ensemble with a minimal number of conformations, as GAJOE usually does (maximum parsimony principle). We rather wanted to have a large subset of structures to account for the diversity of the disordered protein conformations (maximum entropy approach). Thus, we asked GAJOE to

perform 100 runs of genetic algorithm optimization of the fit with experiments, and to save a subset of 50 optimal structures at the end of each run, yielding a final sub-ensemble of 5000 selected conformations.

3 Results and discussion

We address here the question of whether conformational ensembles of N-WASP domain V generated by MD simulations can yield average secondary structure propensities (SSP), residual dipolar couplings (RDC), and SAXS intensities in agreement with experiments. More specifically, we assessed four ensembles, each composed of 40 000 conformations: A first ensemble generated by FM without constrained secondary structure propensity (FM_nossp), another one generated by FM with restrained α-helical propensities of 13% and 31% at positions 10-19 and 39-47, respectively (FM_ssp), a thrid one generated by MD simulations with the C36mm force field (MD_C36mm), and a last one generated by MD simulations with the A03ws model (MD_A03ws).

3.1 NMR information on local secondary structures

We first compared residue-specific secondary propensity scores computed by the SSP program [49] which combines chemical shifts (CS) of Cα, Cβ, CO, and backbone N atoms either measured by NMR or calculated by the program SHIFTS [37]. This score indicates protein regions having a propensity to form α-helix (SSP score > 0) or β-strand / coil (SSP score < 0). This information on protein local structures can also be retrieved from residue-specific N-H residual dipolar couplings (RDCs) either measured by NMR or calculated by PALES [51]. As shown in Figs. 4A and 4C, SSP score and RDC profiles averaged over the FM_ssp ensemble are in excellent agreement with NMR data, which is expected since this conformational ensemble was generated with appropriate input restraints to best fit the NMR observations. It is interesting to note that, although the N-WASP domain V helical probability profile has a well-defined crenellated shape in both input and output of the FM_ssp run (Fig 4E), the SSP profile averaged over the FM_ssp ensemble has a rather smooth mountain shape (Fig 4A). This is due to the fact that, by default, the SSP score of each residue i combines the chemical shifts of residues i-2 to i+2, and thus does not exactly reflect the probability per residue to adopt secondary structures.

Without restraints as input, FM generated structures mostly in random coil and the FM_nossp ensemble SSP score and RDC profiles are rather flat and close to zero (Figs. 4A and 4C). For the MD_C36mm ensemble, we can notice a slight tendency for secondary structures in the SSP score profile (Fig 4B). Strikingly, its RDC profile is globally shifted toward negative values (Fig 4D), suggesting the presence of transient local extended structures (Fig. 4H). Nevertheless, one should be cautious in this interpretation, since even in random coils, residue local conformations are also rather extended, yielding overall negative values of RDCs [50]. Comparatively, MD_A03ws SSP score and RDC profiles are more contrasted and show similar marked deviations from the baseline as the NMR curves (Figs. 4B and 4D). This notably indicates that, starting from 20 random coil structures, MD simulations with A03ws force field can generate significantly populated conformations with α-helices in the same regions as revealed by NMR.

The N-WASP domain V propensity to form local secondary structures was directly quantified by applying the program STRIDE [83] upon the 40 000 conformations of each ensemble (Figs. 4E to 4H). Regarding the FM_ssp ensemble, this analysis confirms that regions 9-18 and 37-46 of N-WASP domain V really have a propensity to form α-helix with probabilities very close to the input constraint values (13% and 31%, respectively) (Fig 4E). In contrast, probabilities to have α-helices in these two regions are lower than 5% and 8% in the FM_nossp and MD_C36mm conformational ensembles, respectively (Figs. 4E and 4F). Interestingly, starting from entirely random coil conformations, MD simulations with the A03ws force field is able to form transient α-helical structures in similar regions of N-WASP domain V (residues 10-15 and 37-43) (Fig 4F). However, the first helix probability (up to 30%) is at the same level as that one of the second helix, unlike observations made from the FM_ssp ensemble, and the numbers of consecutive residues of the two helices are smaller than those in FM_ssp ensemble. These shorter helices could represent intermediate states towards the more extended helices indicated by NMR data and FM_ssp ensemble but which were not sampled by MD simulations due to their arguably limited duration. This could explain the discrepancies between NMR measurements and MD estimations of SSP scores and RDCs. Our simulations also detected two segments (31-34 and 46-51) flanking the helical region 37-43 with significant propensities for α-helix higher than 10% (Fig. 4F). These two additional helical segments were not found around the α-helix of the first WH2 motif, which could contribute to differentiate the conformation and binding affinity for actin of the two N-WASP domain V WH2 motifs.

Regarding local extended structures, the two ensembles generated by FM have nearly none conformation with residues in β-strand (Fig. 4G). In contrast, MD simulations sampled several structures with residues in β conformation (Fig. 4H). This is notably the case for the MD_C36mm ensemble in which residues 20-22, 28-32, and 43-45 have β probabilities between 5 and 10% (Fig. 4). These slight propensities could account for the global shifting toward negative values of RDC profiles computed

from the MD_C36mm and MD_A03ws ensembles. However it should be reminded that, even in random coils, residue local conformations are also rather extended, yielding also overall negative values of RDCs [50].

3.2 SAXS information on global shape

Primary data collected by SAXS experiments are plotted in Figs. 5A and 5B. Kratky curves present typical profiles of a fully disordered protein (Figs. 5C and 5D) [91, 92]. From experimental data, the average radius of gyration R_g of a conformational ensemble can be estimated by using the Guinier approximation, $Log[I(q)/I(0)] = -R_g^2 q^2/3$, which is valid for very small angles [82]. A linear fit of the *Log I(q)* curve for qR_g below 0.8, using the Primus package [93], yielded an experimental average radius of gyration R_g = 24.30 $\pm$ 0.24 Å for N-WASP domain V. It could be noted that this value is quite close to the value expected from the Flory's theory which relates the average radius of gyration of polymer chains to their number of monomeric units: $R_g = R_0 N^\nu$ [94]. Using the appropriate parameters for IDPs ($R_0 = 2.54$ Å and $\nu = 0.522$) [25], a 67 residues polypeptide chain should have a radius of gyration R_g = 22.81 Å. This confirms that N-WASP domain V has a fully disordered nature. The slightly larger value of the experimental radius of gyration (24.30 Å) compared to the Flory's expected value might indicate that N-WASP domain V adopts more extended conformations than those expected from polymer theory.

Back-calculated SAXS intensities averaged over the four conformational ensembles FM_nossp, FM_ssp, MD_C36mm, and MD_A03ws were then directly compared to experimental data (Figs. 5A and 5B). Overall, theoretical scattering intensities have similar profile than in experiments. It could be noted that, based on χ^2 values computed between simulations and experimental data, the scattering intensity calculated from the MD_A03ws ensemble ($\chi^2 = 1.75$) better fits the SAXS curve than the three other ensembles ($\chi^2 > 3$). Nevertheless, for q values above 0.2 Å^{-1}, all Kratky plots are slightly below the experimental one, indicating that *in silico* methods generated ensembles of conformations slightly more compact than the one revealed by experiments (Figs. 5C and 5D). This interpretation was confirmed by directly computing radius of gyration distributions for the four theoretical ensembles (Figs. 5E and 5F). The obtained distributions have R_g mean values equal to 22.68, 22.69, 23.59, and 23.38 Å for the FM_nossp, FM_ssp, MD_C36mm, and MD_A03ws ensembles, respectively. To sum up, FM generated conformational ensembles in moderate agreement with SAXS intensities either with or without constraints on local secondary structures. Likewise, MD simulations with C36mm or A03ws force field also yielded conformational ensembles that are fairly consistent with SAXS data. All together, it is noteworthy that starting from 20 various conformations without any secondary structures, MD simulations with the A03ws force field can generate a conformational ensemble of N-WASP domain V which well reproduces both the local and global structural characteristics revealed by NMR and SAXS experiments, respectively.

3.3 Conformational sub-ensemble selection with GAJOE

Previous comparisons between experimental and *in silico* conformations of N-WASP domain V showed that both FM_ssp and MD_A03ws approaches generated ensembles in fairly good agreement with both NMR and SAXS data. This agreement can be improved by selecting conformational sub-ensembles that minimize the χ^2 values between back-calculated and experimental data, particularly SAXS intensities. To this aim, we performed 100 runs of genetic algorithm optimization with the program GAJOE [89, 90] yielding 5000 conformations of N-WASP domain V that were subsequently analyzed.

It should be noted first that conformational sub-ensemble selections on the criterion of SAXS intensity marginally changed the SSP profiles and their agreement with NMR-derived data (Figs. 6A and 6B). Nevertheless, it is observed that deviations between calculated RDCs and NMR measurements are slightly larger for conformational sub-ensembles selected by GAJOE than for primary ensembles, particularly in the case of the FM_ssp ensemble (Figs. 6C and 6D). These variations in RDC profiles are reflected by an overall increase (after selection) in the relative populations of conformations having α-helical residues in the regions 9-18 and 37-46 of N-WASP domain V (Figs. 6E and 6F), whereas no significant change was observed for β-strand probabilities (Figs. 6G and 6H).

In contrast, and as expected, conformational sub-ensemble selections significantly improved agreements between back-calculated and experimental SAXS intensities, as indicated by the reduced residuals randomly distributed around zero (Figs. 7A and 7B). Distributions of the protein radius of gyration were found shifted toward larger values (Figs. 7E and 7F) and more satisfactorily centered around the experimental value (R_g = 24.30 Å) after selection by GAJOE than before (average values increasing from 22.69 to 24.57 Å and from 23.38 to 24.72 Å for FM_ssp and MD_A03ws, respectively). In summary, from both FM_ssp and MD_A03ws conformational ensembles, GAJOE selected sub-ensembles of structures that are in average slightly more extended than the initial ensembles, consistently with SAXS data, but which have overall higher probabilities to form α-helices than estimated by NMR measurements. Since conformations with α-helices are statistically more extended that random coil ones, the increase in helicity observed after GAJOE selection is probably an artefact of

the slightly imperfect prediction of SAXS profiles by the conformational ensembles. This could be due to either a limited accuracy of the SAXS calculation software for both compact and extended structures, or to the non fully exhaustive conformational samplings, highlighting the difficulties to provide a conformational ensemble in perfect agreement with both RDC measurements and SAXS intensities.

4 Conclusion

Intrinsically disordered proteins generally play important roles in the regulation of many biological processes and often constitute key hubs in protein-protein interaction networks. However, the detailed characterization of their conformational ensembles remains very challenging. Particularly, it is not straightforward to have a correct description of both their local and global structures. In the present report, we showed by validation against NMR and SAXS data that multiple molecular dynamics simulations with the AMBER-03w force field and the TIP4P/2005s water model are able to characterize the conformational ensemble of the 67 residues N-WASP domain V with satisfactory reliability, at both local and global level. Simulations can reproduce at expected regions of the peptide sequence the formation of transient helical structures which constitute the molecular recognition features (MoRFs) of actin binding. They can also describe accurately the extension of the polypeptide chain which ensures the accessibility of these MoRFs by actin.

These results are in line with several other MD studies of IDPs using the combination of AMBER-03w and TIP4P/2005s models proposed by Best *et al.* [72]. The latter notably reported that the conformational ensemble sampled with the A03ws force field of the activation domain (71 residues) of the activator for thyroid hormone and retinoid receptor (ACTR) is consistent with both NMR and SAXS measurements [72]. Other extensive simulations of shorter IDP, including an arginie/serine peptide (24 residues), two FG-nucleoporin peptides (16 and 50 residues) [95], the Histatin 5 (24 residues) [96], also showed that the A03ws force field can generate conformational ensembles that are not overly compact in agreement with both SAXS and NMR data. Lastly, a recent benchmark by Robustelli *et al.* on 9 disordered proteins, including the N_{TAIL} domain of the measles virus nucleoprotein (125 residues) and the α-synuclein (140 residues), confirmed that MD simulations with the A03ws force field performed quite well in terms of both secondary structure propensity and radius of gyration [97].

Thus, although our study only focused on one intrinsically disordered protein, the capability of the AMBER-03ws model to correctly reproduce the N-WASP domain V conformational ensemble seems not specific to this protein. This contributes, with other studies on other IDPs, to identify the most appropriate force fields for accurate MD simulations of IDP conformational ensembles. We believe that this physics-based computational approach can be applied to other IDPs with reasonable confidence, notably in the case where experimental data are lacking.

Supporting material

Supplementary information on experimental methods and supplementary figures can be found on the journal website.

Author contributions

C.v.H., D.D. and T.H.D. designed research; C.D., L.P., and C.v.H. produced protein and performed NMR measurements and analyses; M.C.Y.C. and D.D. carried out and analyzed SAXS experiments; M.C.Y.C. and T.H.D. conducted and analyzed MD simulations; M.C.Y.C., D.D., and T.H.D. wrote the paper with the help of all other authors.

Acknowledgements

This work is supported by the "IDI 2016" project funded by the IDEX Paris-Saclay (ANR-11-IDEX-0003-02). MD simulations were performed using HPC resources from GENCI-CINES (Grant A0040710415). We thank the staff of the Swing beamline at the SOLEIL synchrotron for assistance during SAXS experiments. We are grateful to J. Hub and collaborators for fruitful explanations about WAXSiS software. The authors declare no competing financial interests.

References

1. Wright, P. E., and H. J. Dyson, 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology* 293:321–331.
2. Dunker, A. K., J. D. Lawson, C. J. Brown, R. M. Williams, P. Romero, J. S. Oh, C. J. Oldfield, A. M. Campen, C. M. Ratliff, K. W. Hipps, J. Ausio, M. S. Nissen, R. Reeves, C. Kang, C. R. Kissinger, R. W. Bailey, M. D. Griswold, W. Chiu, E. C. Garner, and

Z. Obradovic, 2001. Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling* 19:26–59.
3. Dyson, H. J., and P. E. Wright, 2005. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol* 6:197–208.
4. Dunker, A. K., I. Silman, V. N. Uversky, and J. L. Sussman, 2008. Function and structure of inherently disordered proteins. *Current Opinion in Structural Biology* 18:756–764.
5. Habchi, J., P. Tompa, S. Longhi, and V. N. Uversky, 2014. Introducing Protein Intrinsic Disorder. *Chem. Rev.* 114:6561–6588.
6. Dunker, A. K., Z. Obradovic, P. Romero, E. C. Garner, and C. J. Brown, 2000. Intrinsic protein disorder in complete genomes. *Genome Inform Ser Workshop Genome Inform* 11:161–171.
7. Ward, J. J., J. S. Sodhi, L. J. McGuffin, B. F. Buxton, and D. T. Jones, 2004. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *Journal of Molecular Biology* 337:635–645.
8. Le Gall, T., P. R. Romero, M. S. Cortese, V. N. Uversky, and A. K. Dunker, 2007. Intrinsic disorder in the Protein Data Bank. *J. Biomol. Struct. Dyn.* 24:325–342.
9. Uversky, V. N., C. J. Oldfield, and A. K. Dunker, 2008. Intrinsically Disordered Proteins in Human Diseases: Introducing the D2 Concept. *Annu. Rev. Biophys.* 37:215–246.
10. Wright, P. E., and H. J. Dyson, 2015. Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology* 16:18–29.
11. van der Lee, R., M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, P. M. Kim, R. W. Kriwacki, C. J. Oldfield, R. V. Pappu, P. Tompa, V. N. Uversky, P. E. Wright, and M. M. Babu, 2014. Classification of Intrinsically Disordered Regions and Proteins. *Chem. Rev.* 114:6589–6631.
12. Olsen, J. G., K. Teilum, and B. B. Kragelund, 2017. Behaviour of intrinsically disordered proteins in proteinprotein complexes with an emphasis on fuzziness. *Cell. Mol. Life Sci.* 74:3175–3183.
13. Fung, H. Y. J., M. Birol, and E. Rhoades, 2018. IDPs in macromolecular complexes: the roles of multivalent interactions in diverse assemblies. *Current Opinion in Structural Biology* 49:36–43.
14. Tompa, P., and M. Fuxreiter, 2008. Fuzzy complexes: polymorphism and structural disorder in proteinprotein interactions. *Trends in Biochemical Sciences* 33:2–8.
15. Fuxreiter, M., and P. Tompa, 2012. Fuzzy complexes: a more stochastic view of protein function. *Adv. Exp. Med. Biol.* 725:1–14.
16. Sharma, R., Z. Raduly, M. Miskei, and M. Fuxreiter, 2015. Fuzzy complexes: Specific binding without complete folding. *FEBS Lett.* 589:2533–2542.
17. Zea, D. J., A. M. Monzon, C. Gonzalez, M. S. Fornasari, S. C. E. Tosatto, and G. Parisi, 2016. Disorder transitions and conformational diversity cooperatively modulate biological function in proteins. *Protein Science* 25:1138–1146.
18. Oldfield, C. J., Y. Cheng, M. S. Cortese, P. Romero, V. N. Uversky, and A. K. Dunker, 2005. Coupled Folding and Binding with -Helix-Forming Molecular Recognition Elements. *Biochemistry* 44:12454–12470.
19. Mohan, A., C. J. Oldfield, P. Radivojac, V. Vacic, M. S. Cortese, A. K. Dunker, and V. N. Uversky, 2006. Analysis of Molecular Recognition Features (MoRFs). *Journal of Molecular Biology* 362:1043–1059.
20. Vacic, V., C. J. Oldfield, A. Mohan, P. Radivojac, M. S. Cortese, V. N. Uversky, and A. K. Dunker, 2007. Characterization of Molecular Recognition Features, MoRFs, and Their Binding Partners. *J. Proteome Res.* 6:2351–2366.
21. Cheng, Y., C. J. Oldfield, J. Meng, P. Romero, V. N. Uversky, and A. K. Dunker, 2007. Mining alpha-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry* 46:13468–13477.
22. Lee, C., L. Kalmar, B. Xue, P. Tompa, G. W. Daughdrill, V. N. Uversky, and K.-H. Han, 2014. Contribution of proline to the pre-structuring tendency of transient helical secondary structure elements in intrinsically disordered proteins. *Biochimica et Biophysica Acta (BBA) - General Subjects* 1840:993–1003.
23. Kiefhaber, T., A. Bachmann, and K. S. Jensen, 2012. Dynamics and mechanisms of coupled protein folding and binding reactions. *Current Opinion in Structural Biology* 22:21–29.
24. Jensen, M. R., R. W. Ruigrok, and M. Blackledge, 2013. Describing intrinsically disordered proteins at atomic resolution by NMR. *Current Opinion in Structural Biology* 23:426–435.
25. Bernadó, P., and D. I. Svergun, 2012. Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol. BioSyst.* 8:151–167.
26. Rauscher, S., and R. Poms, 2010. Molecular simulations of protein disorder. *Biochemistry and Cell Biology* 88:269–290.
27. Schwalbe, M., V. Ozenne, S. Bibow, M. Jaremko, L. Jaremko, M. Gajda, M. Jensen, J. Biernat, S. Becker, E. Mandelkow, M. Zweckstetter, and M. Blackledge, 2014. Predictive Atomic Resolution Descriptions of Intrinsically Disordered hTau40 and α-Synuclein in Solution from NMR and Small Angle Scattering. *Structure* 22:238–249.
28. Ball, K. A., D. E. Wemmer, and T. Head-Gordon, 2014. Comparison of Structure Determination Methods for Intrinsically Disordered Amyloid-β Peptides. *J Phys Chem B* 118:6405–6416.
29. Bonomi, M., G. T. Heller, C. Camilloni, and M. Vendruscolo, 2017. Principles of protein structural ensemble determination. *Current Opinion in Structural Biology* 42:106–116.
30. Derry, J. M. J., H. D. Ochs, and U. Francke, 1994. Isolation of a novel gene mutated in Wiskott-Aldrich syndrome. *Cell* 78:635–644.
31. Miki, H., K. Miura, and T. Takenawa, 1996. N-WASP, a novel actin-depolymerizing protein, regulates the cortical cytoskeletal rearrangement in a PIP2-dependent manner downstream of tyrosine kinases. *The EMBO journal* 15:5326–5335.
32. Palma, A., C. Ortega, P. Romero, A. Garcia-V, C. Roman, I. Molina, and M. Santamaria, 2004. Wiskott-Aldrich syndrome protein (WASp) and relatives: A many-sided family. *Immunologia* 23:217–230.

33. Dominguez, R., 2009. Actin filament nucleation and elongation factors structurefunction relationships. *Critical Reviews in Biochemistry and Molecular Biology* 44:351–366.
34. Gaucher, J.-F., C. Maug, D. Didry, B. Guichard, L. Renault, and M.-F. Carlier, 2012. Interactions of Isolated C-terminal Fragments of Neural Wiskott-Aldrich Syndrome Protein (N-WASP) with Actin and Arp2/3 Complex. *Journal of Biological Chemistry* 287:34646–34659.
35. Chereau, D., F. Kerff, P. Graceffa, Z. Grabarek, K. Langsetmo, and R. Dominguez, 2005. Actin-bound structures of Wiskott–Aldrich syndrome protein (WASP)-homology domain 2 and the implications for filament assembly. *Proceedings of the National Academy of Sciences* 102:16644–16649.
36. Kollmar, M., D. Lbik, and S. Enge, 2012. Evolution of the eukaryotic ARP2/3 activators of the WASP family: WASP, WAVE, WASH, and WHAMM, and the proposed new family members WAWH and WAML. *BMC Research Notes* 5:88.
37. Xu, X.-P., and D. A. Case, 2001. Automated prediction of 15N, 13C, 13C and 13C chemical shifts in proteins using a density functional database. *J Biomol NMR* 21:321–333.
38. Xu XiaoPing, and Case David A., 2002. Probing multiple effects on 15N, 13C, 13C, and 13C chemical shifts in peptides using density functional theory. *Biopolymers* 65:408–423.
39. Neal, S., A. M. Nip, H. Zhang, and D. S. Wishart, 2003. Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *J. Biomol. NMR* 26:215–240.
40. Shen, Y., and A. Bax, 2007. Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. *J Biomol NMR* 38:289–302.
41. Kohlhoff, K. J., P. Robustelli, A. Cavalli, X. Salvatella, and M. Vendruscolo, 2009. Fast and Accurate Predictions of Protein NMR Chemical Shifts from Interatomic Distances. *J. Am. Chem. Soc.* 131:13894–13895.
42. Seidel, K., M. Etzkorn, R. Schneider, C. Ader, and M. Baldus, 2009. Comparative analysis of NMR chemical shift predictions for proteins in the solid phase. *Solid State Nuclear Magnetic Resonance* 35:235–242.
43. Vila, J. A., Y. A. Arnautova, O. A. Martin, and H. A. Scheraga, 2009. Quantum-mechanics-derived 13C chemical shift server (CheShift) for protein structure validation. *Proc Natl Acad Sci U S A* 106:16972–16977.
44. Shen, Y., and A. Bax, 2010. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J Biomol NMR* 48:13–22.
45. Christensen, A. S., S. P. A. Sauer, and J. H. Jensen, 2011. Definitive Benchmark Study of Ring Current Effects on Amide Proton Chemical Shifts. *J Chem Theory Comput* 7:2078–2084.
46. Christensen, A. S., T. E. Linnet, M. Borg, W. Boomsma, K. Lindorff-Larsen, T. Hamelryck, and J. H. Jensen, 2013. Protein structure validation and refinement using amide proton chemical shifts derived from quantum mechanics. *PLoS ONE* 8:e84123.
47. Wishart, D. S., C. G. Bigam, A. Holm, R. S. Hodges, and B. D. Sykes, 1995. 1H, 13C and 15N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. *Journal of Biomolecular NMR* 5:67–81.
48. Zhang, H., S. Neal, and D. S. Wishart, 2003. RefDB: A database of uniformly referenced protein chemical shifts. *Journal of Biomolecular NMR* 25:173–195.
49. Marsh, J. A., V. K. Singh, Z. Jia, and J. D. Forman-Kay, 2006. Sensitivity of secondary structure propensities to sequence differences between - and -synuclein: Implications for fibrillation. *Protein Sci* 15:2795–2804.
50. Mohana-Borges, R., N. K. Goto, G. J. A. Kroon, H. J. Dyson, and P. E. Wright, 2004. Structural Characterization of Unfolded States of Apomyoglobin using Residual Dipolar Couplings. *Journal of Molecular Biology* 340:1131–1142.
51. Zweckstetter, M., 2008. NMR: prediction of molecular alignment from structure using the PALES software. *Nat. Protocols* 3:679–690.
52. Trewhella, J., A. P. Duff, D. Durand, F. Gabel, J. M. Guss, W. A. Hendrickson, G. L. Hura, D. A. Jacques, N. M. Kirby, A. H. Kwan, J. Prez, L. Pollack, T. M. Ryan, A. Sali, D. Schneidman-Duhovny, T. Schwede, D. I. Svergun, M. Sugiyama, J. A. Tainer, P. Vachette, J. Westbrook, and A. E. Whitten, 2017. 2017 publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution: an update. *Acta Crystallogr D Struct Biol* 73:710–728.
53. Hub, J. S., 2018. Interpreting solution X-ray scattering data using molecular simulations. *Current Opinion in Structural Biology* 49:18–26.
54. Henriques, J., L. Arleth, K. Lindorff-Larsen, and M. Skep, 2018. On the Calculation of SAXS Profiles of Folded and Intrinsically Disordered Proteins from Computer Simulations. *Journal of Molecular Biology* .
55. Grudinin, S., M. Garkavenko, and A. Kazennov, 2017. Pepsi-SAXS: an adaptive method for rapid and accurate computation of small-angle X-ray scattering profiles. *Acta Crystallogr D Struct Biol* 73:449–464.
56. Chen, P.-c., and J. Hub, 2014. Validating Solution Ensembles from Molecular Dynamics Simulation by Wide-Angle X-ray Scattering Data. *Biophysical Journal* 107:435–447.
57. Knight, C. J., and J. S. Hub, 2015. WAXSiS: a web server for the calculation of SAXS/WAXS curves based on explicit-solvent molecular dynamics. *Nucl Acids Res* 43:W225–W230.
58. Cordeiro, T., P. Chen, A. DeBiasio, N. Sibille, F. Blanco, J. Hub, R. Crehuet, and P. Bernadó, 2017. Disentangling polydispersity in the PCNAp15PAF complex, a disordered, transient and multivalent macromolecular assembly. *Nucleic Acids Res.* 45:1501.
59. Svergun, D., C. Barberato, and M. H. J. Koch, 1995. CRYSOL a Program to Evaluate X-ray Solution Scattering of Biological Macromolecules from Atomic Coordinates. *J Appl Cryst, J Appl Crystallogr* 28:768–773.
60. Schneidman-Duhovny, D., M. Hammel, J. Tainer, and A. Sali, 2013. Accurate SAXS Profile Computation and its Assessment by Contrast Variation Experiments. *Biophys J* 105:962–974.

61. Schneidman-Duhovny, D., M. Hammel, J. A. Tainer, and A. Sali, 2016. FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucl Acids Res* 44:W424–W429.
62. Franke, D., M. V. Petoukhov, P. V. Konarev, A. Panjkovich, A. Tuukkanen, H. D. T. Mertens, A. G. Kikhney, N. R. Hajizadeh, J. M. Franklin, C. M. Jeffries, and D. I. Svergun, 2017. ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J Appl Crystallogr* 50:1212–1225.
63. Fedorov, B. A., O. B. Ptitsyn, and L. A. Voronin, 1974. X-ray diffuse scattering by proteins in solution. Consideration of solvent influence. *J Appl Cryst, J Appl Crystallogr* 7:181–186.
64. Pavlov, M. Y., and B. A. Fedorov, 1983. Improved technique for calculating Xray scattering intensity of biopolymers in solution: Evaluation of the form, volume, and surface of a particle. *Biopolymers* 22:1507–1522.
65. Abraham, M. J., T. Murtola, R. Schulz, S. Pll, J. C. Smith, B. Hess, and E. Lindahl, 2015. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 12:19–25.
66. Essmann, U., L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, 1995. A smooth particle mesh Ewald method. *Journal of chemical physics* 103:8577–8593.
67. Hess, B., 2008. P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *Journal of Chemical Theory and Computation* 4:116–122.
68. Miyamoto, S., and P. A. Kollman, 1992. SETTLE: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of computational chemistry* 13:952–962.
69. Nosé, S., 1984. A unified formulation of the constant temperature molecular dynamics methods. *Journal of Chemical Physics* 81:511–519.
70. Hoover, W. G., 1985. Canonical dynamics: Equilibrium phase-space distributions. *Physical Review A* 31:1695–1697.
71. Parrinello, M., and A. Rahman, 1981. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* 52:7182–7190.
72. Best, R. B., W. Zheng, and J. Mittal, 2014. Balanced ProteinWater Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *Journal of Chemical Theory and Computation* 10:5113–5124.
73. Abascal, J. L. F., and C. Vega, 2005. A general purpose model for the condensed phases of water: TIP4P/2005. *Journal of Chemical Physics* 123:234505.
74. Piana, S., A. G. Donchev, P. Robustelli, and D. E. Shaw, 2015. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *Journal of Physical Chemistry B* 119:5113–5123.
75. Huang, J., S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmller, and A. D. MacKerell, 2017. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nature Methods* 14:71–73.
76. Jorgensen, W. L., J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, 1983. Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics* 79:926–935.
77. Kaminski, G. A., R. A. Friesner, J. Tirado-Rives, and W. L. Jorgensen, 2001. Evaluation and Reparametrization of the OPLS-AA Force Field for Proteins via Comparison with Accurate Quantum Chemical Calculations on Peptides. *The Journal of Physical Chemistry B* 105:6474–6487.
78. Berendsen, H. J. C., J. R. Grigera, and T. P. Straatsma, 1987. The missing term in effective pair potentials. *Journal of Physical Chemistry* 91:6269–6271.
79. Lindorff-Larsen, K., S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, 2010. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Structure, Function, and Bioinformatics* 78:1950–1958.
80. Best, R. B., and J. Mittal, 2010. Protein Simulations with an Optimized Water Model: Cooperative Helix Formation and Temperature-Induced Unfolded State Collapse. *The Journal of Physical Chemistry B* 114:14916–14923.
81. Mackerell, A. D., M. Feig, and C. L. Brooks, 2004. 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. *Journal of Computational Chemistry* 25:1400–1415.
82. Guinier, A., 1939. La diffraction des rayons X aux trs petits angles : application l'tude de phnomnes ultramicroscopiques. *Ann. Phys.* 11:161–237.
83. Heinig, M., and D. Frishman, 2004. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res* 32:W500–W502.
84. Ozenne, V., F. Bauer, L. Salmon, J.-r. Huang, M. R. Jensen, S. Segard, P. Bernadó, C. Charavay, and M. Blackledge, 2012. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics* 28:1463–1470.
85. Hus, J.-C., D. Marion, and M. Blackledge, 2001. Determination of Protein Backbone Structure Using Only Residual Dipolar Couplings. *Journal of the American Chemical Society* 123:1541–1542.
86. Lovell, S. C., I. W. Davis, W. B. Arendall, P. I. de Bakker, J. M. Word, M. G. Prisant, J. S. Richardson, and D. C. Richardson, 2003. Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. *Proteins: Structure, Function, and Bioinformatics* 50:437–450.
87. Levitt, M., 1976. A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of Molecular Biology* 104:59–107.
88. Krivov, G. G., M. V. Shapovalov, and R. L. Dunbrack, 2009. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* 77:778–795.

89. Bernadó, P., E. Mylonas, M. V. Petoukhov, M. Blackledge, and D. I. Svergun, 2007. Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering. *Journal of the American Chemical Society* 129:5656–5664.
90. Tria, G., H. D. T. Mertens, M. Kachala, and D. I. Svergun, 2015. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ* 2:207–217.
91. Receveur-Brchot, V., and D. Durand, 2012. How Random are Intrinsically Disordered Proteins? A Small Angle Scattering Perspective. *Current Protein Peptide Science* 13:55–75.
92. Kikhney, A. G., and D. I. Svergun, 2015. A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Letters* 589:2570–2577.
93. Konarev, P. V., V. V. Volkov, A. V. Sokolova, M. H. J. Koch, and D. I. Svergun, 2003. PRIMUS: a Windows PC-based system for small-angle scattering data analysis. *J Appl Cryst, J Appl Crystallogr* 36:1277–1282.
94. Flory, P., 1953. Principles of Polymer Chemistry. Cornell Univ. Press, Ithaca, New York, 1953.
95. Rauscher, S., V. Gapsys, M. J. Gajda, M. Zweckstetter, B. L. de Groot, and H. Grubmller, 2015. Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.* 11:5513–5524.
96. Henriques, J., and M. Skep, 2016. Molecular Dynamics Simulations of Intrinsically Disordered Proteins: On the Accuracy of the TIP4P-D Water Model and the Representativeness of Protein Disorder Models. *J. Chem. Theory Comput.* 12:3407–3415.
97. Robustelli, P., S. Piana, and D. E. Shaw, 2018. Developing a molecular dynamics force field for both folded and disordered protein states. *PNAS* 201800690.

Figure legends

Figure 1: Sequence alignment of N-WASP domain VC crystallized with actin by Gaucher *et al.* (2VCP_D) with the 67 residues peptide used in the present study (DomainV). The upper amino acid numbering comes from the human sequence. The solid line box indicates the 19 residues of the second WH2 motif present in the X-ray structure [34]. The dashed line box highlights the homologous 19 residues of the first WH2 motif.

Figure 2: Comparison between SAXS intensity back-calculation software using implicit solvent models versus WAXSiS which uses an explicit solvent approach. Orange, green, blue, and pink bars are results for CRYSOL, FoXS, CRYSOL3, and Pepsi-SAXS, respectively. Shaded bars indicate results when using default values for the hydration layer density.

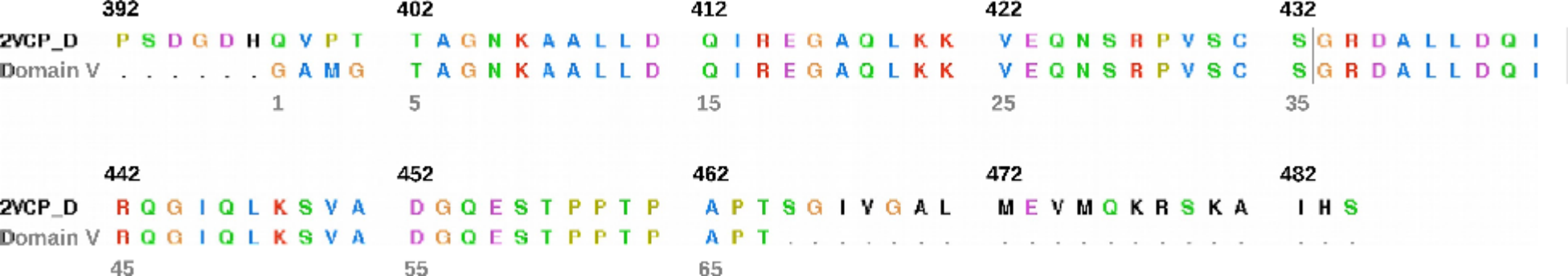
Figure 3: Time evolution of N-WASP domain V radius of gyration computed from short MD simulations using 6 different force fields. Horizontal coloured dashed lines and black solid lines indicate average radii of gyration in simulations and in experiments, respectively.

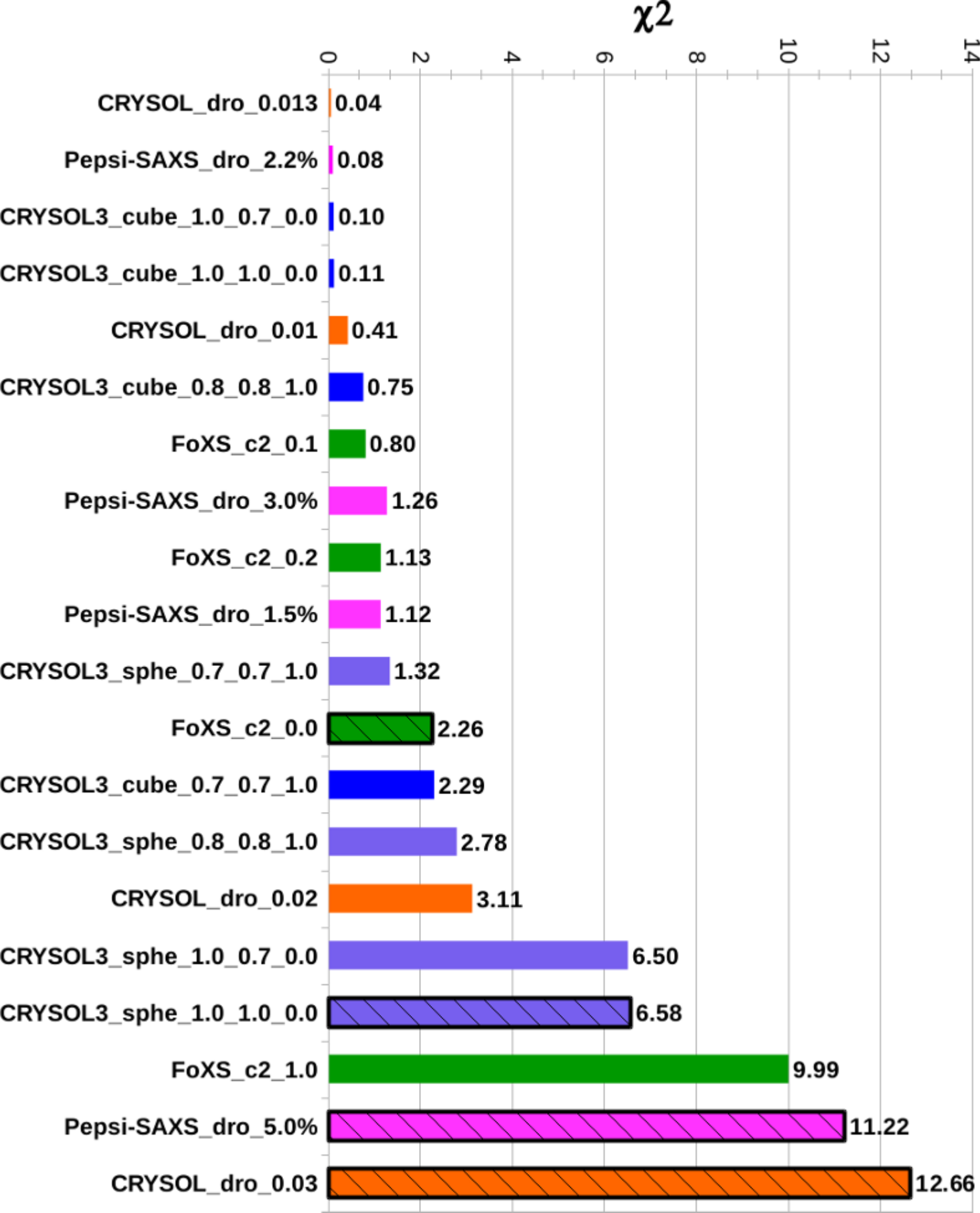
Figure 4: Direct comparison of SSP scores (**A** and **B**) and N-H RDCs (**C** and **D**) averaged over FM_nossp, FM_ssp, MD_C36mm, and MD_A03ws ensembles with NMR measurements. Probabilities for N-WASP domain V residues to be in α-helix (**E** and **F**) and β-strand (**G** and **H**) structures computed from the FM_nossp, FM_ssp, MD_C36mm, and MD_A03ws ensembles. Dashed and dotted brown vertical lines represent protein regions in α-helix (as indicated by the X-ray structure 2VCP [34]) and the highly conserved sequences LK[K-S]V [35, 36], respectively.

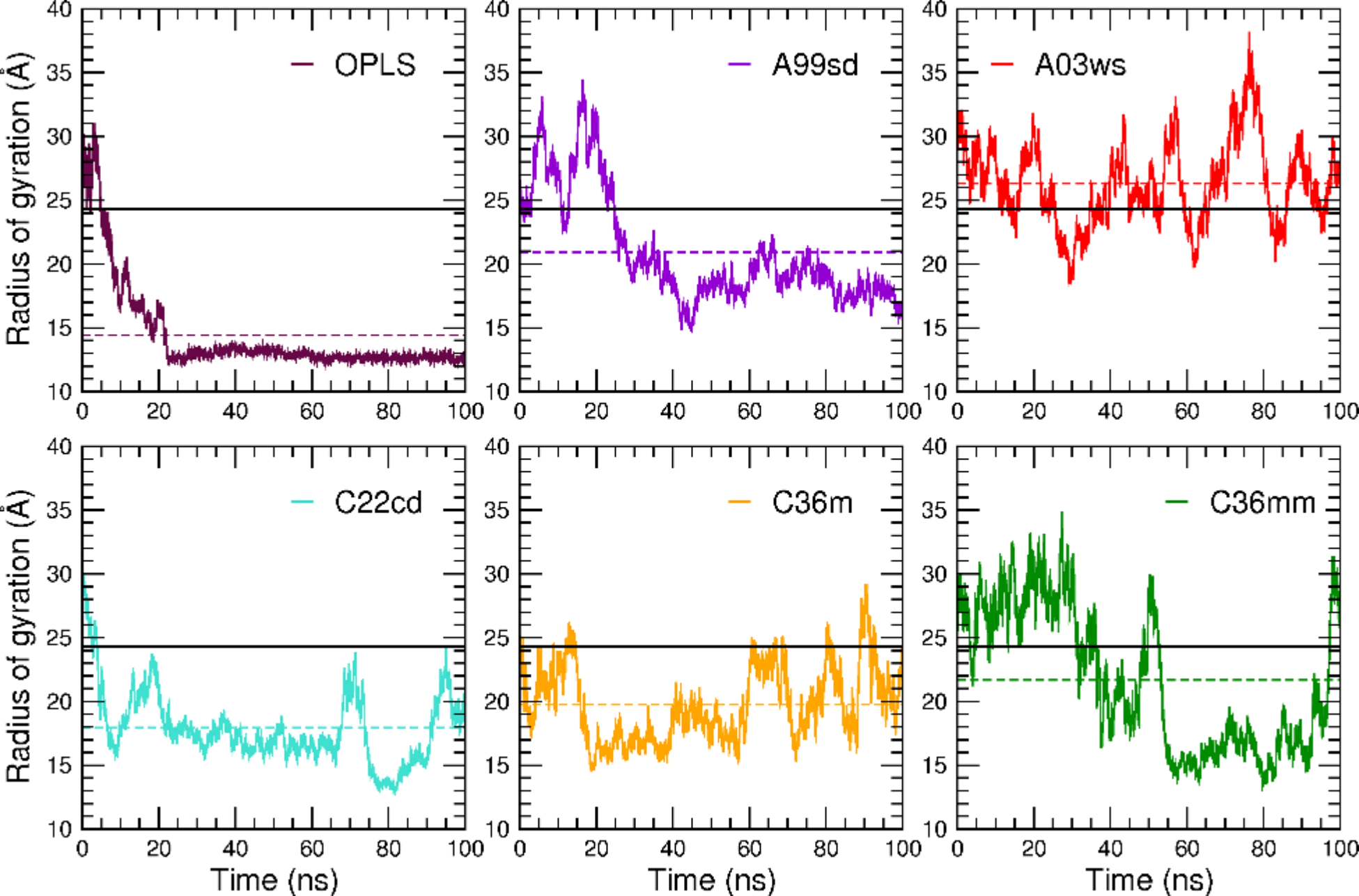
Figure 5: Comparison of SAXS intensities Log I(q) (**A** and **B**) and Kratky plots (**C** and **D**) as a function of scattering vector q calculated from the FM_nossp, FM_ssp, MD_C36mm, and MD_A03ws ensembles against experimental data. Reduced residuals $\Delta/\sigma = [I_{calc}(q) - I_{exp}(q)]/\sigma_{exp}(q)$. Probability of the N-WASP domain V radius of gyration (**E** and **F**) computed for the F_nossp, FM_ssp, MD_C36mm, and MD_A03ws conformational ensembles. Vertical black solid and coloured dashed lines indicate radius of gyration measured by experiments and mean values found in simulations, respectively.

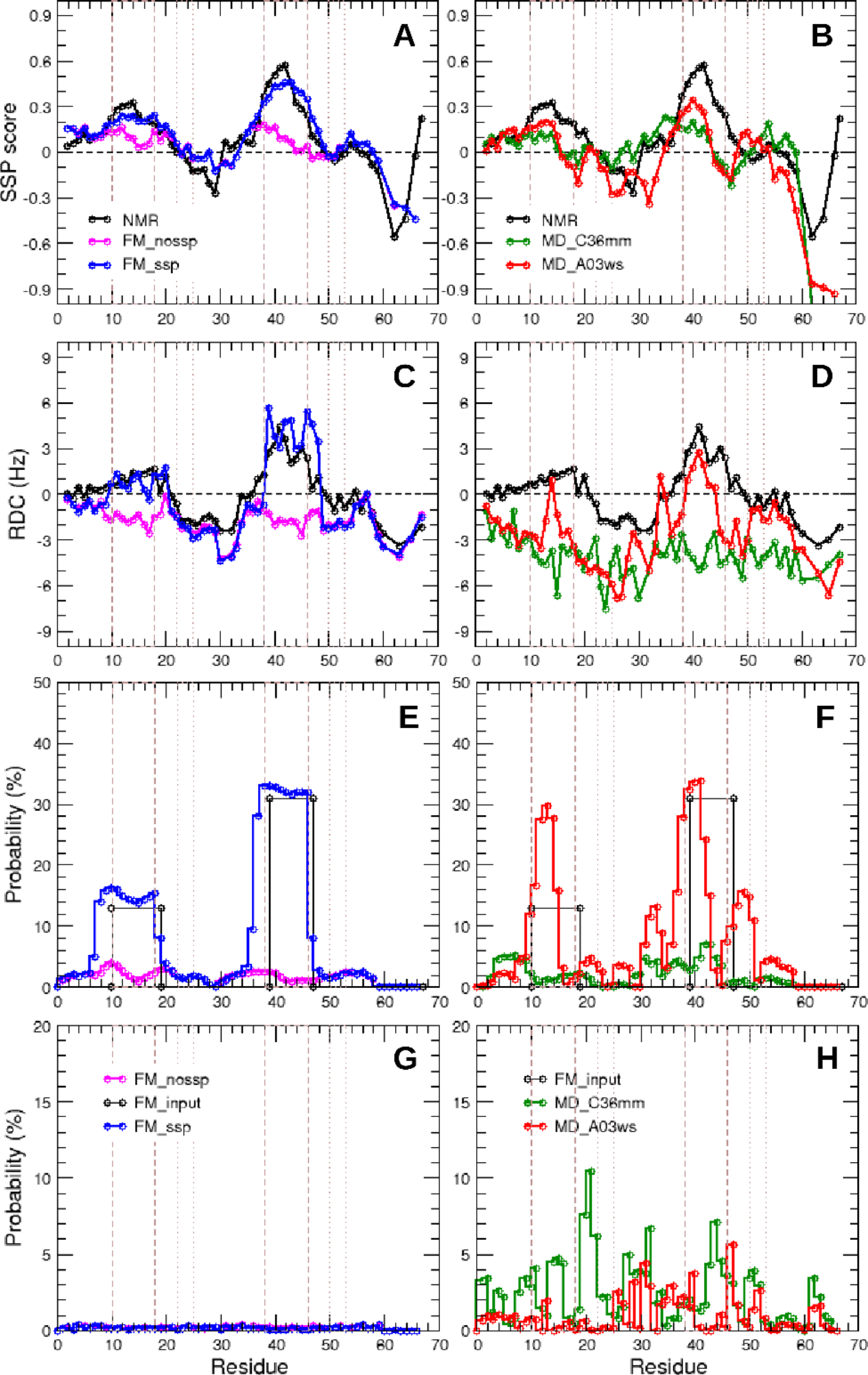
Figure 6: Comparison with NMR experiments of SSP scores (**A** and **B**) and N-H RDCs (**C** and **D**) averaged over FM_ssp and MD_A03ws ensembles, before and after selection by GAJOE. Residue-specific probability to be in α-helix (**E** and **F**) or in β-strand (**G** and **H**) calculated from the FM_ssp and MD_A03ws ensembles, before and after selection by GAJOE.

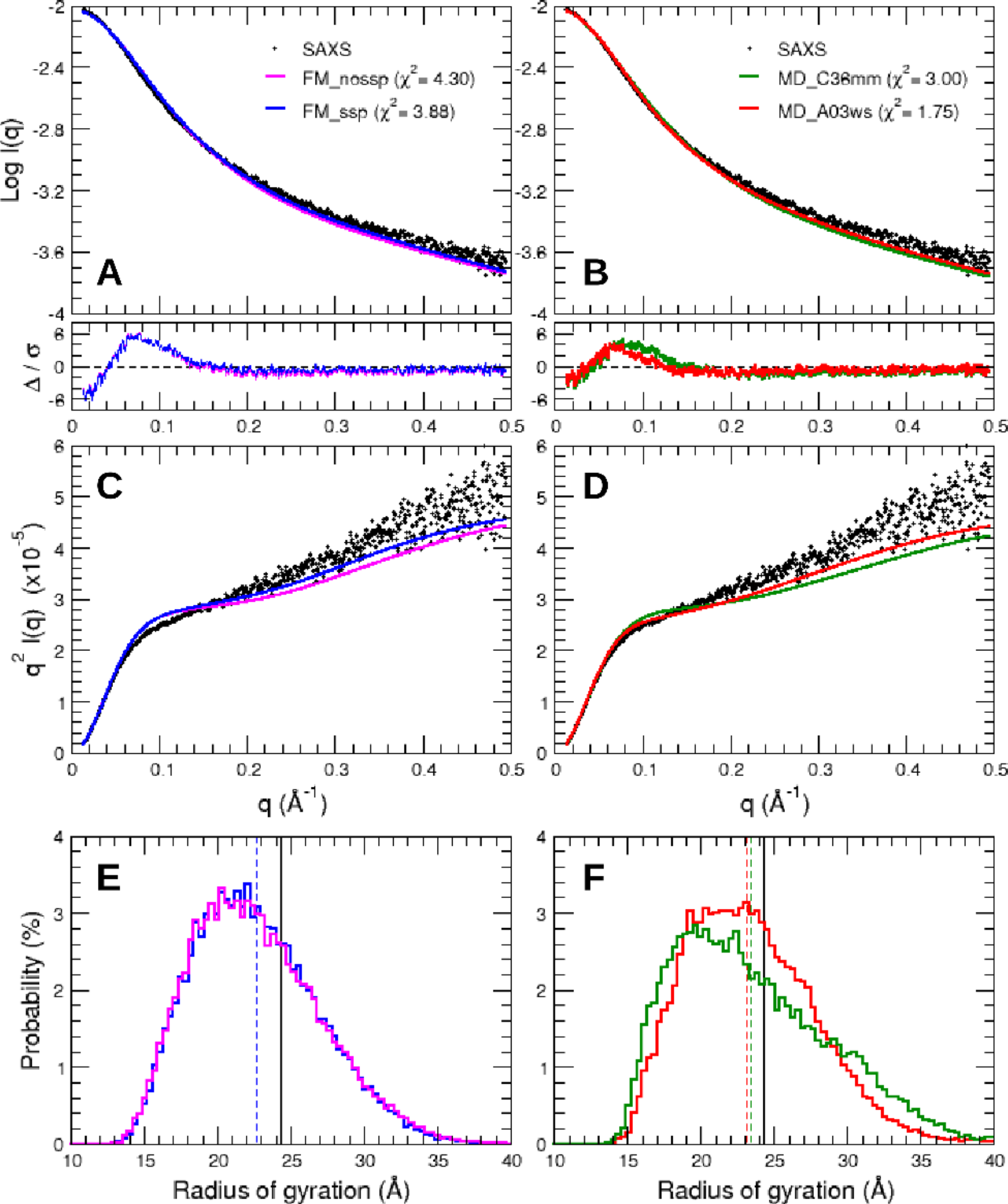
Figure 7: Comparison with SAXS experiments of Log I(q) (**A** and **B**) and Kratky plots (**C** and **D**) averaged over the FM_ssp and MD_A03ws ensembles, before and after selection with GAJOE. Reduced residuals $\Delta/\sigma = [I_{calc}(q) - I_{exp}(q)]/\sigma_{exp}(q)$. Probability of the N-WASP domain V radius of gyration (**E** and **F**) computed for the FM_ssp and MD_A03ws conformational ensembles, before and after selection by GAJOE.

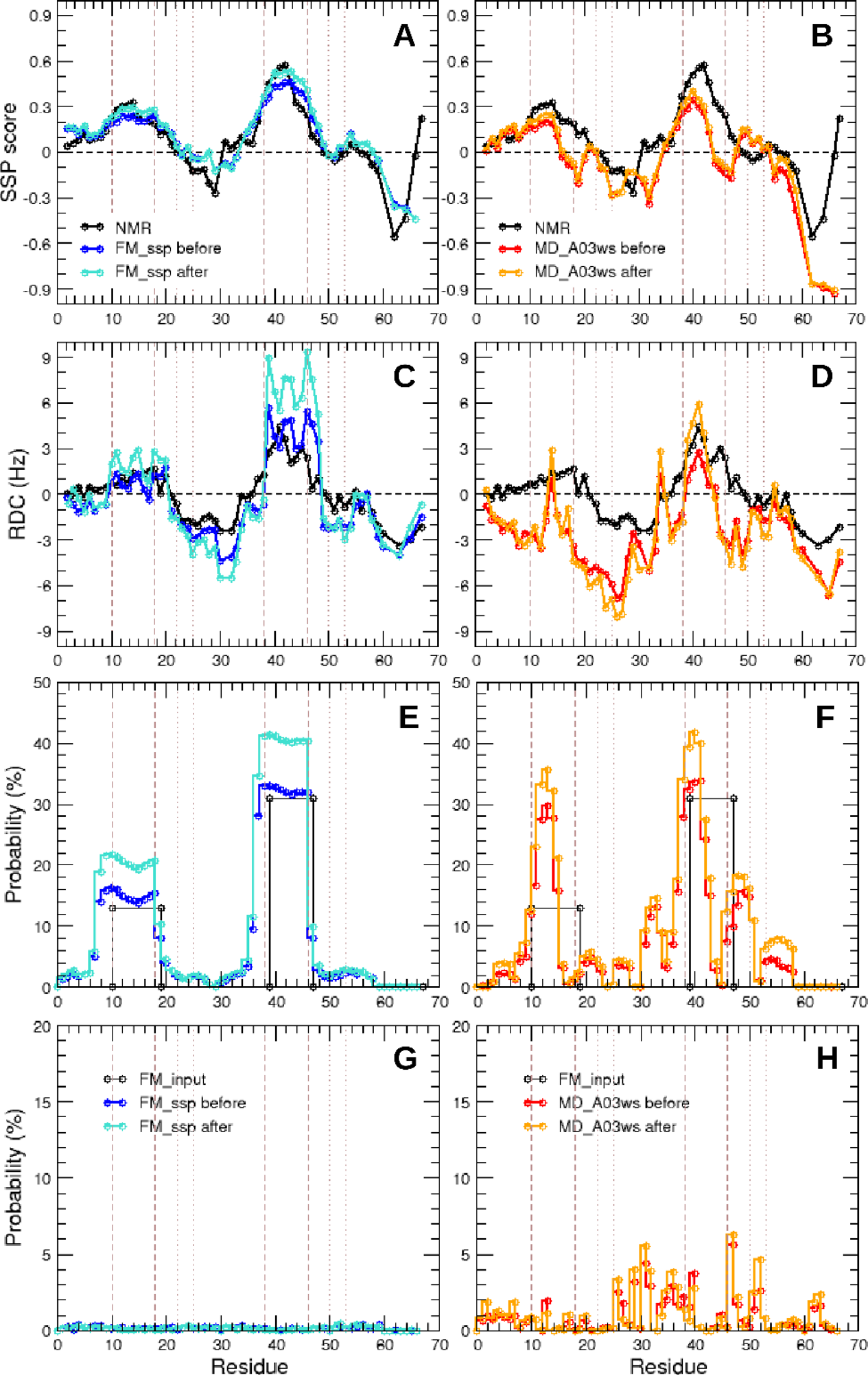

392
402
412
422
432
2VCP_D P S D G D H Q V P T T A G N K A A L L D Q I R E G A Q L K K V E Q N S R P V S C S G R D A L L D Q I
Domain V . . . . . G A M G T A G N K A A L L D Q I R E G A Q L K K V E Q N S R P V S C S G R D A L L D Q I
1
5
15
25
35
442
452
462
472
482
2VCP_D R Q G I Q L K S V A D G Q E S T P P T P A P T S G I V G A L M E V M Q K R S K A I H S
Domain V R Q G I Q L K S V A D G Q E S T P P T P A P T . . . .
45
55
65

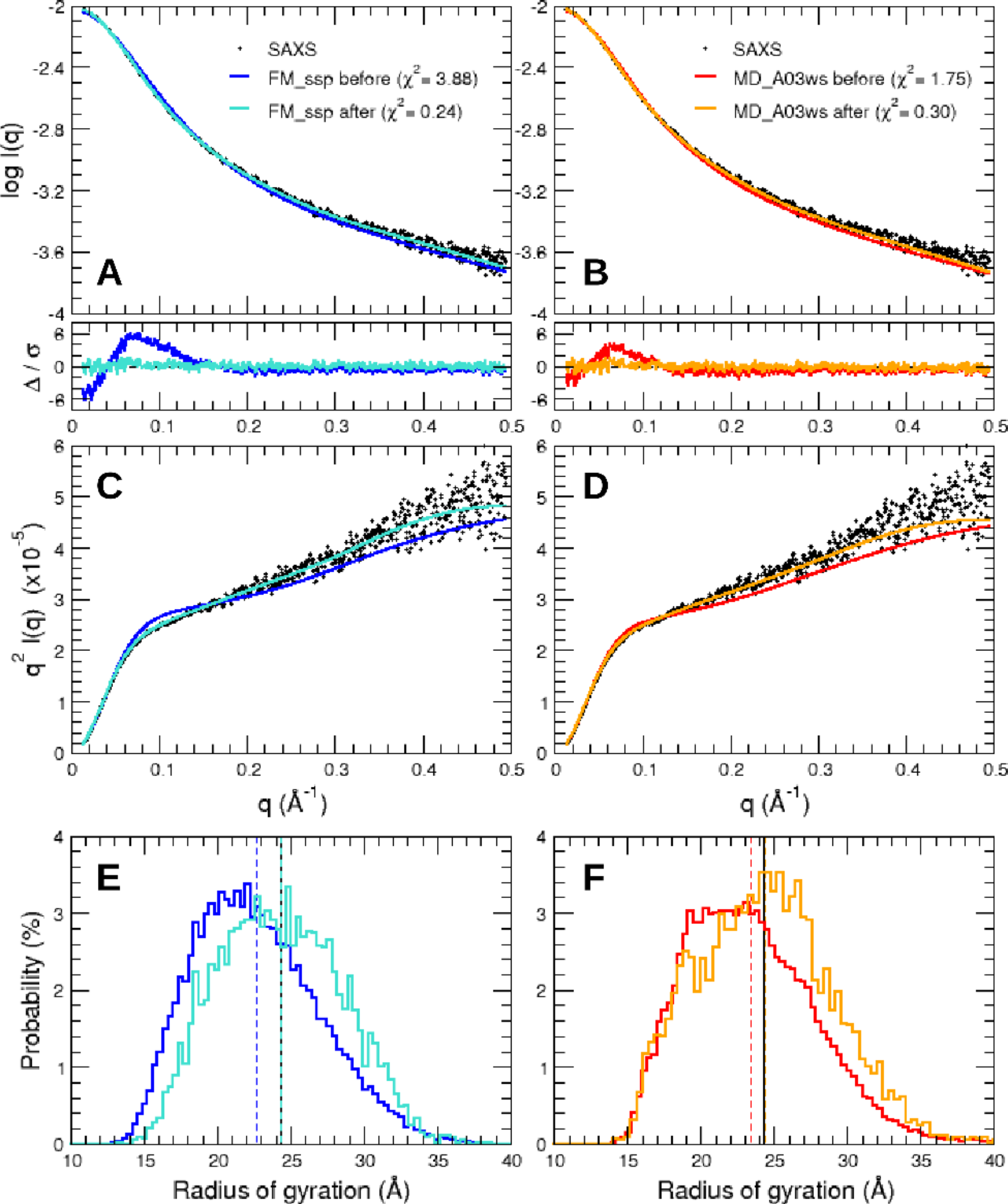

χ2
0
2
4
6
8
10
12
14
CRYSOL_dro_0.013
0.04
Pepsi-SAXS_dro_2.2%
0.08
CRYSOL3_cube_1.0_0.7_0.0
0.10
CRYSOL3_cube_1.0_1.0_0.0
0.11
CRYSOL_dro_0.01
0.41
CRYSOL3_cube_0.8_0.8_1.0
0.75
FoXS_c2_0.1
0.80
Pepsi-SAXS_dro_3.0%
1.26
FoXS_c2_0.2
1.13
Pepsi-SAXS_dro_1.5%
1.12
CRYSOL3_sphe_0.7_0.7_1.0
1.32
FoXS_c2_0.0
2.26
CRYSOL3_cube_0.7_0.7_1.0
2.29
CRYSOL3_sphe_0.8_0.8_1.0
2.78
CRYSOL_dro_0.02
3.11
CRYSOL3_sphe_1.0_0.7_0.0
6.50
CRYSOL3_sphe_1.0_1.0_0.0
6.58
FoXS_c2_1.0
9.99
Pepsi-SAXS_dro_5.0%
11.22
CRYSOL_dro_0.03
12.66


OPLS
A99sd
A03ws
C22cd
C36m
C36mm
Radius of gyration (Å)
Time (ns)
10
15
20
25
30
35
40
0
20
40
60
80
100


A
B
C
D
E
F
G
H
SSP score
RDC (Hz)
Probability (%)
Probability (%)
NMR
FM_nossp
FM_ssp
NMR
MD_C36mm
MD_A03ws
FM_nossp
FM_input
FM_ssp
FM_input
MD_C36mm
MD_A03ws
Residue
Residue


A
B
C
D
E
F
SAXS
FM_nossp (χ^2 = 4.30)
FM_ssp (χ^2 = 3.88)
MD_C36mm (χ^2 = 3.00)
MD_A03ws (χ^2 = 1.75)
Log I(q)
Δ / σ
q^2 I(q) (x10^{-5})
q (Å^{-1})
Probability (%)
Radius of gyration (Å)


A
B
C
D
E
F
G
H
SSP score
RDC (Hz)
Probability (%)
Probability (%)
NMR
FM_ssp before
FM_ssp after
NMR
MD_A03ws before
MD_A03ws after
FM_input
FM_ssp before
FM_ssp after
FM_input
MD_A03ws before
MD_A03ws after
Residue
Residue


A
B
C
D
E
F
SAXS
FM_ssp before (χ^2 = 3.88)
FM_ssp after (χ^2 = 0.24)
MD_A03ws before (χ^2 = 1.75)
MD_A03ws after (χ^2 = 0.30)
log I(q)
Δ / σ
q^2 I(q) (x10^{-5})
q (Å^{-1})
Probability (%)
Radius of gyration (Å)