Incorporation of D_2O-induced Fluorine Chemical Shift Perturbations into Ensemble-Structure Characterization of the ERalpha Disordered Region

Wenwei Zheng[*,1], *Zhanwen Du*[2], *Soo Bin Ko*[2], *Nalinda Wickramasinghe*[3], *Sichun Yang*[*,2]

[1]College of Integrative Sciences and Arts, Arizona State University, Mesa, Arizona 85212, United States

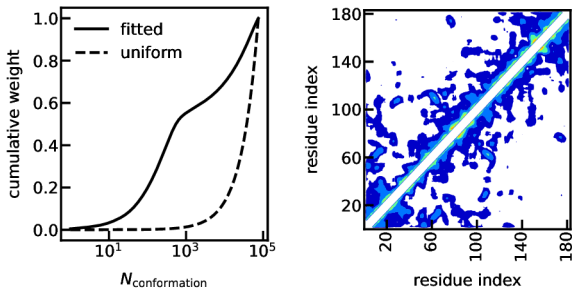
[2]Center for Proteomics and Department of Nutrition, School of Medicine, Case Western Reserve University, Cleveland, Ohio, 44106, United States

[3]Chemistry-NMR Facility, Case Western Reserve University, Cleveland, Ohio 44106, United States

KEYWORDS. Estrogen receptor; Intrinsically disordered protein; Integrative biophysics; Molecular dynamics; Ensemble structures.

ABSTRACT

Structural characterization of intrinsically disordered proteins (IDPs) requires a concerted effort between experiments and computations by accounting for their conformational heterogeneity. Given the diversity of experimental tools providing local and global structural information, constructing an experimental restraint-satisfying structural ensemble remains challenging. Here, we use the disordered N-terminal domain (NTD) of the estrogen receptor alpha (ERalpha) as a model system to combine existing small-angle X-ray scattering (SAXS) and hydroxyl radical protein footprinting (HRPF) data and newly acquired solvent accessibility data via D_2O-induced fluorine chemical shifting (DFCS) measurements. A new set of DFCS data for the solvent exposure of a set of 12 amino acid positions were added to complement previously acquired HRPF measurements for the solvent exposure of the other 16 non-overlapping amino acids, thereby improving the NTD ensemble characterization considerably. We also found that while choosing an initial ensemble of structures generated from a different atomic-level force field or sampling/modeling method can lead to distinct contact maps even when the same sets of experimental measurements were used for ensemble-fitting, comparative analyses from these initial ensembles reveal commonly recurring structural features in their ensemble-averaged contact map. Specifically, nonlocal or long-range transient interactions were found consistently between the N-terminal segments and the central region, sufficient to mediate the conformational ensemble and regulate how the NTD interacts with its coactivator proteins.



Introduction

Intrinsically disordered proteins (IDPs) are associated with various intracellular functions and pathological diseases.[1-3] Understanding the conformational properties of IDPs is of interest requiring a concerted effort between experiments and computations.[4-6] Due to the lack of a well-defined folded structure and the intrinsic conformational flexibility, it is well acknowledged that a notion of the conformational ensemble is required to characterize the heterogeneity of IDP conformations.[7-14] Obtaining such a conformational ensemble has imposed a key challenge for experimental measurements and theoretical or computational methods.

Various experimental techniques have been used to provide structural properties of IDPs. The first type of methods provides information about the global protein conformation and nonlocal pairwise distances between protein amino acids, with significant differences in the sensitivity of the distance regime between these methods. For instance, small-angle X-ray scattering (SAXS)[15, 16] and dynamic light scattering[12] can provide the protein size and/or pairwise distance distributions. In contrast, labeling techniques, such as Förster resonance energy transfer[17] and paramagnetic relaxation enhancement,[18] provide the amino acid distances or dynamics for a few or a large set of specific pairs each at their distance range of detection. The second type of methods provides residue-specific or local structural properties such as backbone chemical shifts,[19] residual dipolar coupling,[20] and relaxation measurements[21, 22] from nuclear magnetic

resonance as well as circular dichroism[23] spectroscopy for overall secondary structure assessment. These methods are often combined with the first type of methods to provide a complete picture of local and global structural properties of IDPs.[8, 24, 25] The third type of methods probes solvent exposure of individual amino acids, including hydroxyl radical protein footprinting (HRPF),[26] D_2O-induced fluorine chemical shifting (DFCS),[27] and more comprehensively, label-free solvent-PRE.[28, 29] They are sensitive to both local secondary structure and nonlocal or long-range interactions can in principle provide complementary information to existing methods, although it remains elusive how these diverse sparse data can be combined for structural interpretation, the subjection of this work.

Computationally, a large set of degrees of freedom for an IDP makes it impossible to sample its vast conformational space without the help of experimental restraints. Combining the collective knowledge from computations and experimental measurements offers an alternative strategy for investigating the conformational ensemble of IDPs.[30] Typically, this integration is achieved in the two following approaches: 1) ensemble fitting by first generating an ensemble of structural candidates using molecular simulations and then selecting a subset of conformations from the initial ensembles to fit against experimental restraints;[7, 9, 12] and 2) generating the restraint-satisfying ensemble via biased molecular simulations.[31] Ensemble fitting methods are often used for their ease and simplicity without invoking the complication of simultaneous biasing for a relatively large set of structures on the fly. One clear advantage of this ensemble-fitting approach is that the initial ensemble only needs to be generated once and can be used for fitting different experimental restraints[32]. However, a disadvantage is that if the initial ensemble fails to describe the experimental restraints accurately, the ensemble fitting methods would require an enlarged pool of structural candidates for their fitting procedure, often in an iterative

fashion. Given sparse experimental restraints available, it remains unclear how the choice of the initial ensemble of structures and related ensemble fitting method, affects an accurate description of the IDPs[33].

Specifically, we here investigate the strategy of integrating experimental measurements providing residue-specific solvent accessibility and size of the protein. To this end, the N-terminal domain (NTD) of the estrogen receptor, critical for its hormone-independent ER activation,[34] was used as a model system of intrinsic disorder.[35, 36] Our previous work has pointed out the compaction of this NTD disorder and its important nonlocal/long-range interactions by integrating SAXS and HRPF data using an ensemble fitting method.[37] While HRPF provided the solvent accessibilities of 16 amino acids along its sequence (Fig. 1A), it is always a question of whether additional restraints can be incorporated to improve the accuracy of such an IDP conformational ensemble. In this work, a relatively new probing method DFCS was first explored to provide extra information about site-specific solvent accessibility, complementing a different set of amino acids probed by HRPF previously.[37] As such, a key question becomes the extent to which the DFCS data improve the structural description of the NTD compared to the HRPF alone. Together with the existing SAXS data,[37] we further evaluate the robustness of the ensemble fitting method by using both atomic-level simulations and coarse-grained modeling for initial structure generation and examine the extent to which the combination of these different initial ensembles better interprets those experimental findings. Finally, we predict specific nonlocal or long-range interactions from the conformational ensembles integrating multiple sources of experimental measurements of the NTD.

Methods

Experimental measurements. Site-directed mutagenesis was conducted to introduce a cysteine residue at comparable serine sites of the NTD. Protein expression and purification were described previously[37]. BTFA (3-Bromo-1,1,1-trifluoroacetone; Alfa Aesar Cat# A14948) was used for the attachment of a trifluoromethyl ($–CF_3$) group to the end of cysteine sidechains by adding BTFA (1:100) into a resolubilization buffer (10 mM sodium phosphate, pH 7.2, 0.5 mM EDTA, and 0.1 mM PMSF) with overnight incubation at 4 °C. Proteins were prepared in a range of 1.2-2.2 mg/ml, each with a titration at 10%, 20%, 30%, 40%, and 50% D_2O. One-dimensional ^{19}F chemical shift spectra were recorded on a Bruker Ascend III HD 500 MHz spectrometer equipped with a nitrogen-cooled ^{19}F tuned BBO probe. All spectra were acquired with 512 scans, 131K data points in the direct dimension, a pulse length of 15.0 μsec, a spectral width of 468,750 Hz (^{19}F), a digital resolution of 3.5 Hz/point, and a relaxation delay of 1.0 s at 8 °C. Topspin 3.5 pl 7 was used for data processing, and free induction decay signals were apodized with an exponential window function as to line broadening of 0.30 Hz.

Theoretical methods. Three different simulation methods were used for generating the initial ensembles (see Supporting Methods for details). The first ensemble (hereafter AA-1 with 35,240 structures from a total accumulative time of 35 μs) was generated from the combination of replica exchange and Gaussian-accelerated molecular dynamics simulations (using the Amber ff99sb force field[38] and the TIP3P water model[39]) as described in the previous literature.[37] The second ensemble (hereafter AA-2 with 20,000 conformations from a total time of 4 μs) was generated using a brute force simulation (using the Amber99sbws force field[40] and the TIP4P/2005 water model[41]). The third ensemble (hereafter FM with 20,000 structures) was generated using the Flexible-Meccano model[42] to obtain a coarse-grained representation including C_α and C_β atoms. For each residue, its missing atoms including the rest of its sidechain,

were added via the FASPR algorithm.[43] Theoretical calculations for the SAXS intensity of each simulated structure in the ensemble were achieved using Fast-SAXS-pro.[44] HRPF data were interpreted via the solvent accessible surface area (SA) of individual residue sidechains.[26] DFCS data were evaluated via the SA of the CF_3 group attached to each corresponding cysteine sidechain. Details of computational methods were provided in the Supporting Methods.

Results and Discussion

The NTD structural properties probed by multiple experimental techniques

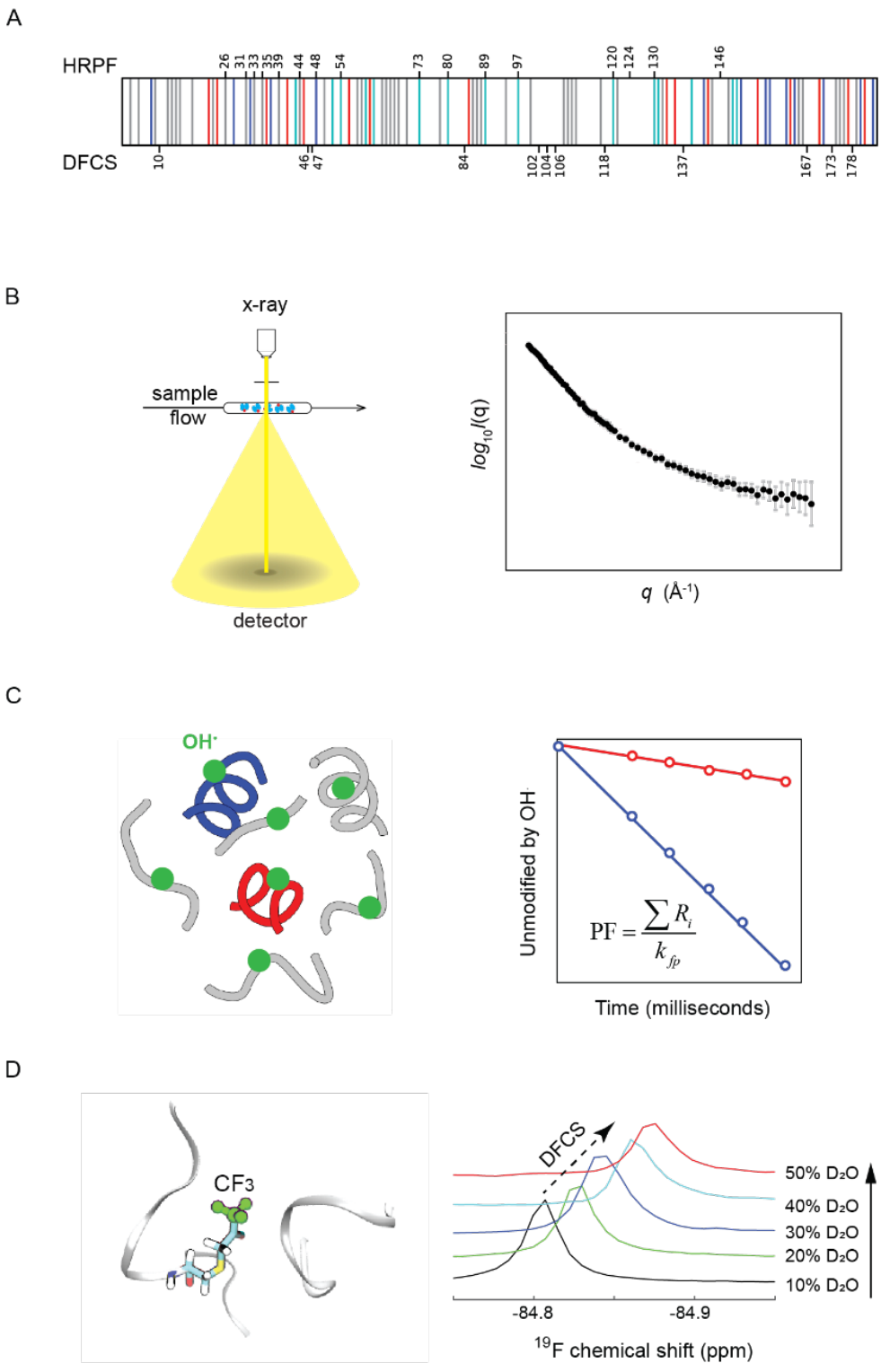


Figure 1. Biophysical probing of the NTD disorder. A) The NTD consists of heterogeneous amino acids with positive and negative charges highlighted in blue and red, respectively,

aromatic residues in cyan and hydrophobic residues in grey. Top, labeling positions by HRPF; Bottom, labeling sites by DFCS. **B)** SAXS data acquisition. Left: Schematic of the SAXS setup with an X-ray source, a biomolecular sample flow, and a photon detector. Right: A typical one-dimensional SAXS profile as a function of q, the amplitude of momentum transfer. C) Solvent exposure measurements via HRPF. Left: Schematic of HRPF measurements. Green dots, reactive hydroxyl radicals generated from X-ray hydrolysis and their attachment with amino acid sidechains. Right: Representative HRPF measurements via a dose-response curve before conversion into a structure-based protection factor (PF). D) Site-specific solvent exposure probed via DFCS. Left, schematic for a cysteine with a CF_3 tag (in green) attached to its sidechain; Right, peak changes in D_2O-induced fluorine chemical shifts as a function of increasing D_2O concentrations from 10% to 50%. DFCS, the linear slope of the peak shift as a function of D_2O concentration. More details can be found in Methods.

We first inspected the sequence characteristics of the NTD (Fig. 1A), specifically the distribution of charged and aromatic amino acids. They are clustered in three regions, the N-terminal region before residue 50, the central region between residue 50-140, and the C-terminal region after residue 140. We calculated a series of sequence descriptors of these three regions to investigate the conformational properties of these regions shown in Table S1. Within both N-terminal and C-terminal regions, positively and negatively charged amino acids are abundant with a fraction of charged amino acids (FCR) of 22.0% and 34.1%, respectively, while the central region only has five negatively charged amino acids with an FCR of 6.7%, suggesting a more extensive electrostatic influence in the two terminal regions. Also, the charge patterning using sequence charge decoration (SCD)[45] indicates net attractive interactions in the N-terminal region, net repulsive interactions in the central region but relatively weak repulsive interactions

in the C-terminal region. Under the assumption that electrostatic interactions between these charged amino acids are the sole driving force, the NTD would adopt a dumbbell-like conformation where its two heads are each dominated by intra-region attractions but connected via a long central region hanging with mostly repulsive electrostatic interactions. On the other hand, a high percentage of the hydrophobic amino acids, especially abundant aromatic residues, are localized within the central region (*i.e.*, 13.3%), suggesting its regional tendency to collapse. Sequence hydropathy decoration (SHD)[46], which is based on the hydropathy patterning, further suggests a much stronger hydrophobic effect in the central region, compared with the two terminal regions. Using SCD and SHD analysese[46], the scaling exponent for the central region is predicted at 0.443, which is considerably smaller than the two terminal regions (*i.e.*, 0.492 and 0.543, respectively). Of note, the overall scaling exponent predicted for the entire sequence is 0.415, which is even smaller than the central region of 0.443 alone, consistent with the overall compaction we previously reported[37]. This collapse of the entire chain appears to be driven by the hydrophobic interactions within the central region (residues 51-140), further mediating nonlocal electrostatic attractions between charged amino acids of the two terminal regions (see Table S1). These simple sequence-based analyses provide an inspiring set of testable predictions on nonlocal interactions within the NTD for further experimental validation.

In a recent work, we applied the SAXS and HRPF techniques to investigate the NTD structural properties.[37] SAXS suggested roughly the size of an IDP (Fig. 1B). As for the NTD, the obtained radius of gyration R_g of 31.0 Å is smaller than that of a typical IDP with 184 residues (~ 38.6 Å), estimated via a power law derived from the experimental observations of a large number of IDPs,[15] consistent with our previous observation regarding the attractive interactions based on the sequence properties. The HRPF further probed the solvent accessibility of 16

amino acids along the sequence (Fig. 1A top and Fig. 1C), and a few of them were observed to be less solvent accessible (e.g., residues 33 and 124), departing from the behavior of a random coil expected with high solvent exposure. These SAXS and HRPF data suggested transient intramolecular interactions, whereas the challenge is to define a structural ensemble illustrating these essential interactions, especially for specific amino acids. In the same literature, we applied an ensemble fitting method to derive a reweighted conformational ensemble and found a few nonlocal interactions, e.g., between residues 33 and 118, which were confirmed by site-directed mutagenesis and ^{19}F-NMR. However, whether additional experimental restraints can lead to an improved structural ensemble with other significant transient interactions remains to be seen.

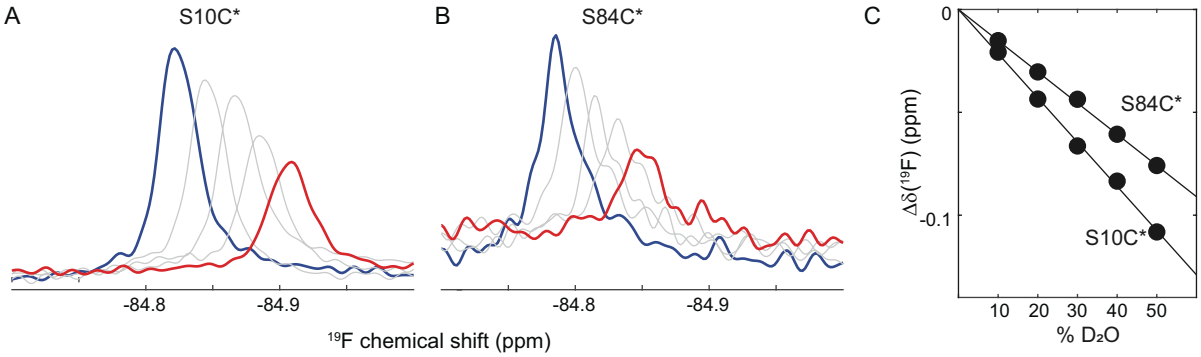


Figure 2. DFCS measurements for solvent exposure of individual NTD amino acids. A and B) The ^{19}F spectrum of S10C* (A) and S84C* (B) at different D_2O concentrations ranging from 10% D_2O (blue) to 50% (red). C) The DFCS values as the slope of ^{19}F spectral peaks illustrated for two representative S10C* and S84C* examples. Δδ, peak values relative to that in the absence of D_2O.

TABLE 1. The DFCS values of 12 amino acid positions probed. * denotes ^{19}F labeling. Numbers in parentheses denote the errors in the last few digits.

S10C* -0.211 (5)	S46C* -0.218 (24)	S47C* -0.186 (7)	S84C* -0.167 (7)	S102C* -0.124 (8)	S104C* -0.185 (14)
S106C* -0.184 (6)	S118C* -0.192 (17)	S137C* -0.208 (9)	S167C* -0.169 (9)	S173C* -0.183 (10)	S178C* -0.163 (7)

We further sought additional information on position-specific solvent accessibility via a ^{19}F labeling technique, which we termed D_2O-induced fluorine chemical shifting (DFCS). The DFCS probes the solvent accessibility at amino acid positions not overlapped with HRPF (Fig. 1A). It uses a small molecule known as BTFA (3-Bromo-1,1,1-trifluoroacetone), so a trifluoromethyl ($–CF_3$) group is attached to the end of a cysteine sidechain with minimal structural perturbation (Fig. 1D).[26, 47] Advantageously, the NTD has no native cysteine residue. A fluorine tag can be attached at any serine amino acid position (comparable to cysteine) after mutating to cysteine. More importantly, because of an isotopic effect of D_2O water, increasing concentrations of the isotopic D_2O water lead to a peak shift of fluorine spectra, denoted as $\Delta\delta$ (see Fig. 2), dependent on the local solvent exposure.[27, 48] Thus, the DFCS reports the exposure to the solvent D_2O via $\Delta\delta$ (relative to that in the absence of D_2O, i.e., 100% H_2O). For instance, a larger $\Delta\delta$ value of $S10C^{BTFA}$ indicates higher solvent exposure (Fig. 2A), while a smaller $\Delta\delta$ value of $S84C^{BTFA}$ indicates otherwise (Fig. 2B). We denote the DFCS value to represent the linear slope between $\Delta\delta$ and D_2O concentrations (Fig. 2C), an indicator of the solvent exposure of individual sites.

By taking advantage of the well-distributed serine amino acids along the NTD sequence, we repeated this ^{19}F labeling and related DFCS determination procedure for the set of 12 amino acid positions one by one (Fig. 1A bottom), one by one. Their DFCS values are listed in Table 1. We found that most ^{19}F-tagged positions are fully solvent-exposed, except for positions S84C* S102C*, S167C*, and S178C*, which are relatively solvent-protected. This high solvent

exposure is in line with the HRPF measurements that only a few amino acids show little solvent protection, and they are thus likely to be involved in transient intramolecular interactions while most other amino acids remain as flexible as in a typical IDP. Nonetheless, DFCS and HRPF provide a non-overlapping coverage of amino acid positions along the sequence, which offers an unparalleled opportunity to evaluate whether these DFCS measurements provide additional and complementary information to improve the NTD structural ensemble.

Generation of initial structural ensembles

Given the availability of additional site-specific information, a key question is whether the initial ensemble of candidate structures is sufficiently diverse to satisfy the collective experimental restraints from SAXS, HRPF, and DFCS data. We previously used two major advanced sampling methods with a standard all-atom force field Amber ff99sb[38] and the TIP3P water model[39] to generate the initial all-atom ensemble (see Methods) as our first option of testing (referred to as AA-1). This combination of such an all-atom force field and a water model is known to generate an overly compacted conformation[40, 49]. In our previous work, however, a sufficient number of conformations were found close to the experimental radii of gyration from SAXS.[37] Such enhanced sampling was deemed to be helpful, without which the simulation was often stuck in a few conformations due to an overestimation of amino acid interactions. While no direct control was done, the fact we are able to find a restraint-satisfying ensemble by the current sampled space indicates the sampling is well poised to provide a sufficiently large pool of conformations with diverse amino acid interactions.

Due to the recent advance in the all-atom force field optimized specifically for IDP structures, we also utilized an optimized Amber99sbws[40] and the TIP4P/2005 water model[41] to generate an alternative all-atom ensemble (referred to as AA-2). We are aware that there exist multiple sets

of new all-atom force fields that are optimized for intrinsically disordered proteins.[50-54] However, the main focus of this work is to examine to which extent an all-atom ensemble can best fit the experimental data for ensemble-structure characterization, as opposed to evaluating the performance of different all-atom force fields. These previously developed force fields still work reasonably well, as shown in the Result section; testing the performance of an initial ensemble generated with newer force fields will be a topic of future studies. Since a coarse-grained model is less computationally demanding, there is also a growing interest in using such a model to investigate IDPs; we generated a third initial ensemble (hereafter FM) using the Flexible-Meccano method.[42]

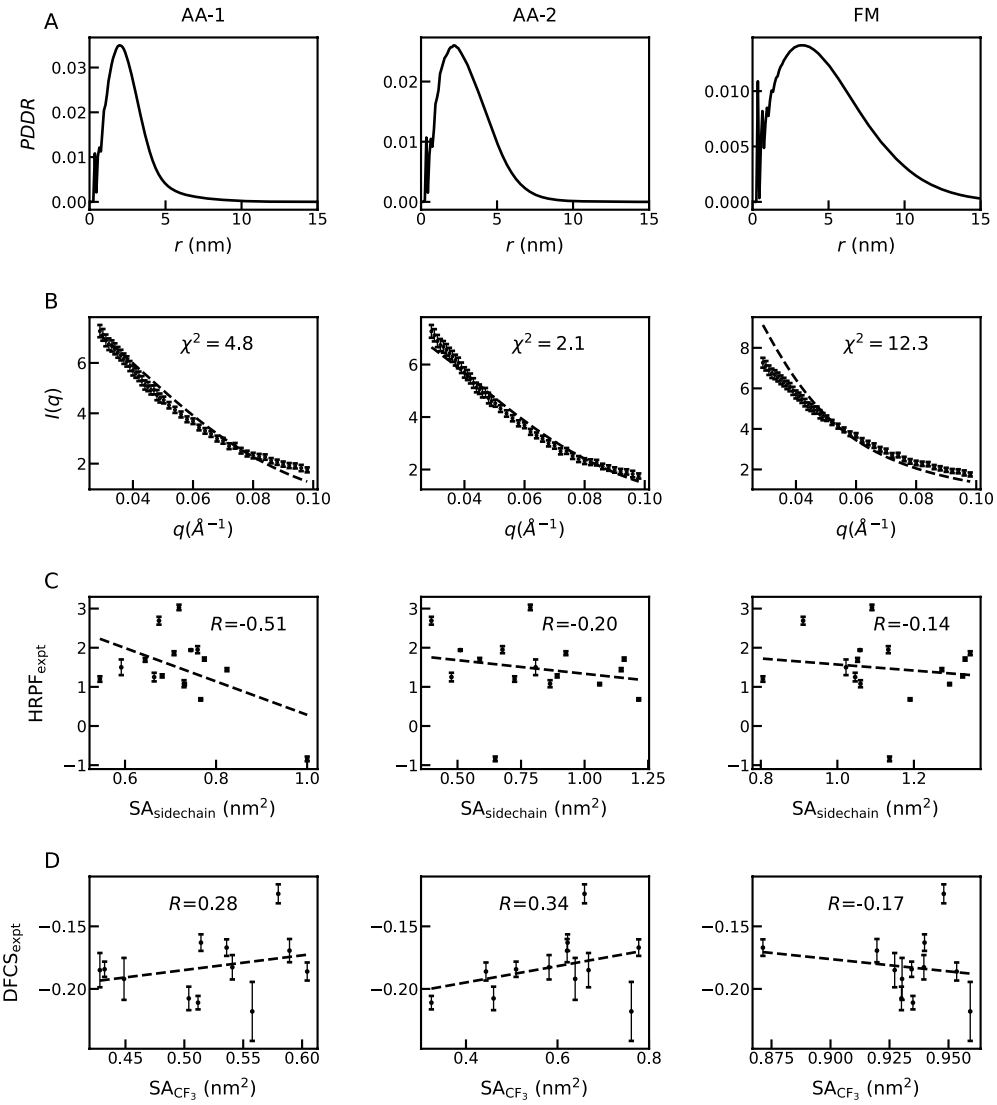


Figure 3. Comparison of the initial ensembles using three computational methods of AA-1 (all-atom simulations with advanced sampling), AA-2 (brute-force all-atom simulations with an IDP-optimized force field, and FM (the Flexible-Meccano method[42]) (see Methods). A) Pair distance distribution function (PDDR). B) Comparison of SAXS data with χ^2 shown in the legend. Dashed line, theoretical SAXS curve; q, the scattering vector amplitude; $I(q)$, scattering intensity. C) Comparison between the calculated SA value of each residue sidechain ($SA_{sidechain}$) with experimental HRPF data. R, Pearson correlation coefficient. D) Comparison between the solvent

accessible surface area of each CF_3 group (SA_{CF3}) with experimental DFCS data. Dots with error bars, experimental data with uncertainty.

Fig. 3A shows that the pair distance distribution function (PDDF) for these three initial ensembles looks quite different. Notably, the AA-1 ensemble is relatively collapsed with an averaging radius of gyration (R_g) of ~22 Å, the FM is more expanded (~42 Å), and the AA-2 in-between (~25 Å). None is close to the experimental estimate of 31 Å, suggesting the necessity of including SAXS data as experimental restraints in the ensemble fitting. Given the difference among these three initial ensembles, we evaluate their goodness of fit against experimental measurements.

The ensemble fitting is achieved by minimizing the difference between experimental measurements and theoretical counterparts for the structural ensemble. For theoretical SAXS calculations, we used the in-house *Fast-SAXS-pro* algorithm[44] by taking advantage of its capability of computing SAXS profiles from atomistic and coarse-grained structures. Fig. 3B compares the calculated SAXS profiles and experimental SAXS curves. For the HRPF comparison, the SA values of amino acid sidechains were calculated and averaged for the comparison with the HRPF measurements that provide a measure of solvent protection, also known as a "protection factor."[26] The correlation or the goodness of fit between these HRPF measurements and the averaged SA values of individual amino acids is shown in Fig. 3C. In terms of DFCS comparison, the solvent accessibility of the CF_3 tag, as opposed to the entire sidechain, was used after each site was computationally mutated to cysteine attached with a CF_3 tag (see Methods). The correlation between the DFCS measurements and calculated solvent accessibility for the set of amino acid positions probed is shown in Fig. 3D.

On its own, each initial ensemble is best at capturing a specific experimental technique. The AA-1 set matches the HRPF well by capturing the high solvent protection of specific residues in part due to the more collapsed structures sampled while maintaining reliable goodness of fit with SAXS data. The AA-2 reproduces the SAXS experimental data (regarding R_g) as expected since the force field was parameterized to address overly compact conformations for IDP simulations. When comparing with the DFCS data, such an initial ensemble is not in perfect agreement with experimental measurements, *i.e.*, the lack of an apparent correlation between the solvent accessibility of the probed amino acids and the DFCS datasets. For this exact reason, ensemble fitting for such initial ensembles against the experimental datasets becomes a logical step by reweighting the conformations in the initial ensemble to generate an optimal new ensemble that better reproduces the experimental DFCS data.

Complementary information on site-specific solvent accessibility from HRPF and DFCS

To quantify the complementarity between the HRPF and DFCS measurements, each was combined separately with SAXS data in the ensemble fitting (i.e., HRPF+SAXS and DFCS+SAXS). First, SAXS effectively describes the global-conformation information about the ensemble-averaging size and its pairwise distance distribution. In contrast, the HRPF and DFCS information is at a site-specific level about the solvent accessibility of probed amino acid positions, reflecting the intramolecular interactions of these amino acids involved. It is an area of active research to develop ensemble fitting methods where calculated and experimental measurements agree.[7, 9, 12] Notably, a recent study found that the outcome of the ensemble fitting depends on the choice of initial ensembles.[33] As such, it is necessary to test multiple initial ensembles during ensemble fitting.

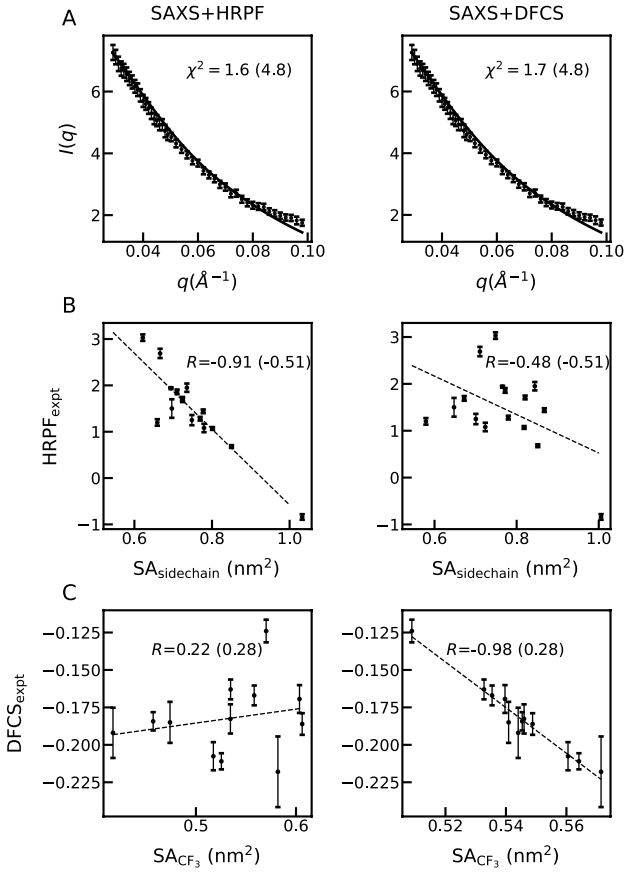


Figure 4. Refining the AA-1 ensemble using SAXS, and HRPF (left) or DFCS (right). A) Comparison of SAXS data with χ^2 shown in the legend. B) Comparison between the calculated SA values of individual sidechains with experimental HRPF data. C) Comparison between the solvent accessible surface area of the CF_3 group with the experimental DFCS data. *R*, Pearson correlation. (See Fig. S1 and S2 for similar results on refining the AA-1 and FM ensembles)

Fig. 4A demonstrates the ensemble fitting results using the AA-1 set as an example, where SAXS can be reproduced equally well in both cases of the SAXS/HRPF combination and the SAXS/DFCS combination. Of note, the outcome from the SAXS/HRPF combination alone is unable to achieve a good agreement with the DFCS measurements (Fig. 4 left) and vice versa

when using SAXS+DFCS (Fig. 4B, right). This room of improvement indicates that the HRPF and DFCS indeed provide complementary information in part due to their non-overlapping nature in terms of the amino acid positions probed (Fig. 1A). The other two initial ensembles (AA-2 and FM) were also used to evaluate the extent to which the ensemble fitting of the two sets of data (SAXS+HRPF or SAXS+DFCS); we found they cannot blindly predict the third dataset (DFCS or HRPF), demonstrating that the new DFCS data yield new knowledge to improve the ensemble fitting (see SI Figs. S1 and S2).

Dependence on the initial ensemble of structures

To investigate how the use of different initial ensemble structures affects the fitting outcomes, we evaluate the goodness of fit for each of the three initial ensembles (AA-1, AA-2, and FM) using the combination of all experimental SAXS, HRPF, and DFCS measurements. First, these initial ensembles describe the SAXS data reasonably well, given the small χ^2 values (Fig. 5A), even though the FM has a larger deviation at the low-q region. Of note, the ensemble-averaged R_g values of the AA-1 and the FM are quite different from the R_g derived from experimental SAXS data, and these conformations are still sufficient to reproduce a reasonable R_g value from the refined ensemble. At last, about 1000 conformations dominate the ensemble properties (see Fig. S3).

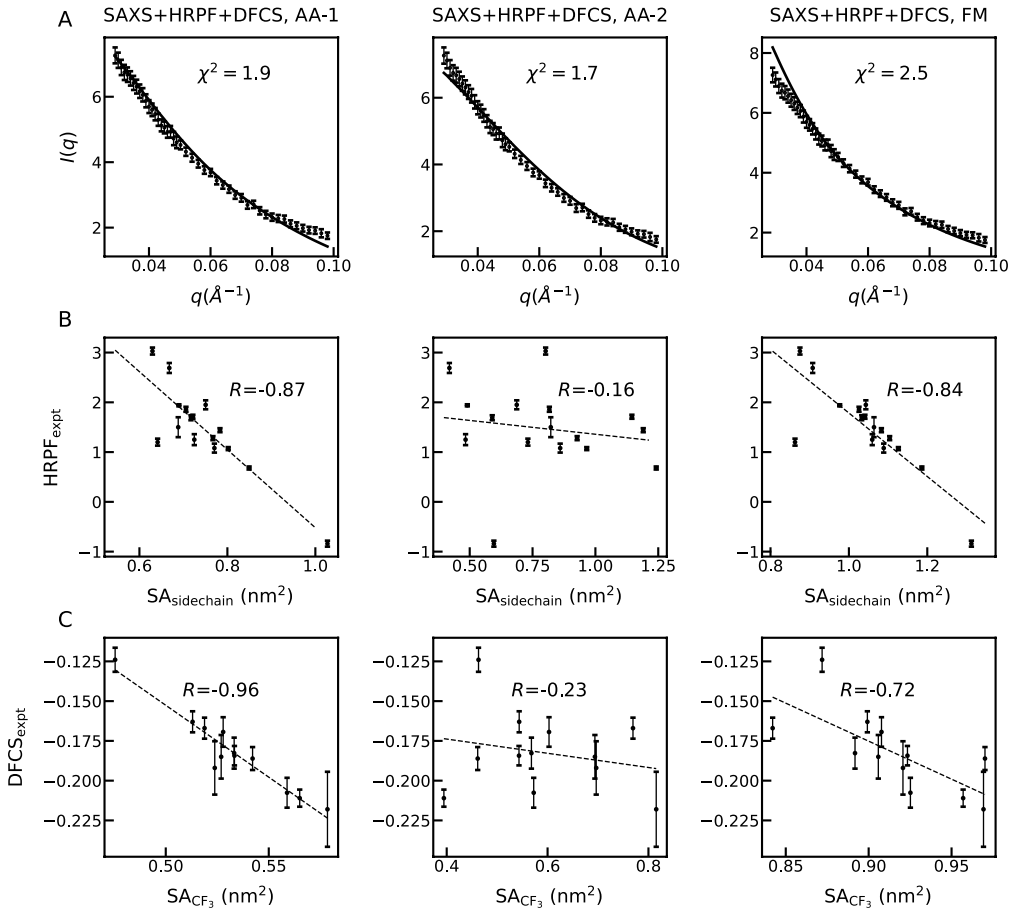


Figure 5. Ensemble fitting using three initial ensembles (AA-1, AA-2, and FM). A) Comparison to SAXS data. χ^2, the deviation between the theoretical and experimental SAXS data. B) Comparison between calculated SA values of residue sidechains with experimental HRPF data. C) Comparison between calculated SA values of CF_3 groups with experimental DFCS data. *R*, Pearson correlation coefficient.

On its own, it is difficult for each initial ensemble to reach convergence with reasonable goodness of fit with HRPF and DFCS data simultaneously due to the relatively broad coverage of the combined non-overlapping amino acid positions probed. While both AA-1 and FM fit well with HRPF and DFCS data, the FM-based SA values seem outside the range obtained from

AA-1 and AA-2, partly due to overly expanded conformations with virtually no amino acid solvent-protected. Notably, the AA-2 ensemble alone cannot simultaneously match SAXS, HRPF, and DFCS data. This issue could arise from the imperfect force field, insufficient sampling, or both. Of note, the observation that AA-2 with an updated force field alone does not perform well in capturing NTD experimental data may not be generalized to other IDPs. It is worth pointing out that both HRPF and DFCS probe more residue-specific local microenvironment, while the force field samples more expanded conformations resulting in less residue-residue contact formation. As such, the force field accuracy is vital to characterize these residue-specific interactions. The other issue could be related to the IDP conformational sampling of these conformations, which is a challenging task.[55]

Nonetheless, it is becoming apparent that the outcomes of ensemble fitting are likely dependent on the initial ensemble (e.g., compact or expanded), as in Fig. 2A. Furthermore, the current highly residue-specific structural information seems to require high-resolution knowledge of amino acid interactions and sufficient sampling in the conformational space, as we shall see next. If there is no or few conformation with specific amino acid interactions in the initial ensemble, it is unlikely to achieve a good fit with experimental data. There can be an alternative way to allow the conformations in the ensemble to adapt to the experimental data instead of fixing the conformations in the initial ensemble, e.g., experiment-biased simulations[31] or evolving ensemble structures during ensemble fitting,[56] which is beyond the scope of this work.

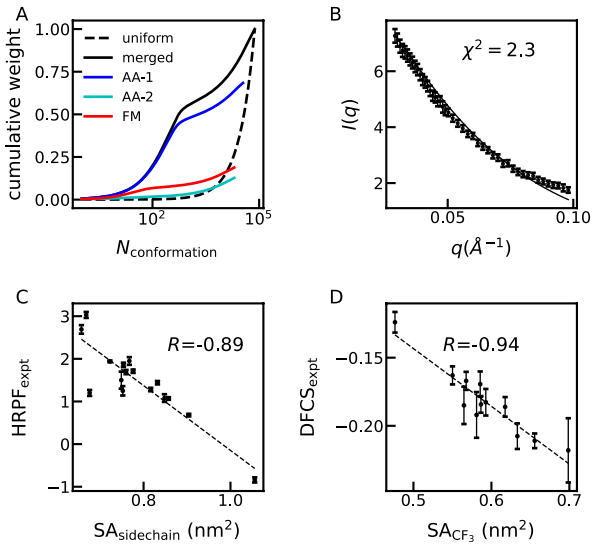


Figure 6. Ensemble fitting against a merged/enlarged ensemble of structures with a combination of SAXS, HRPF, and DFCS data. A) Cumulative weight as a function of the number of conformations included. B) Comparison to SAXS data. C) Comparison between calculated SA values of residue sidechains with experimental HRPF data. D) Comparison between calculated SA values of CF_3 groups with experimental DFCS data. *R*, Pearson correlation coefficient.

To test the extent to which an enlarged pool of candidate structures improves the experimental data interpretation, we evaluate the goodness of fit via the combination of the three pools of candidate structures from these initial ensembles (AA-1, AA-2, and FM). The final ensemble is merged from all three initial ensembles. To evaluate the minimal number of conformations that contribute most to the final ensemble, we calculated the cumulative weights as a function of the conformation number after ensemble fitting. As shown in the black line of Fig. 6A, approximately 1,000 conformations contribute to more than 50% of the total weight of the entire ensemble, suggesting that while there are more than 75,000 conformations in the initial merged

ensemble, only a small portion of these conformations contributes to reproducing the experimental restraints. It is true that the more experimental restraints used in ensemble fitting would lead to a less number of important conformations in the final ensemble, although the reasonably good agreement between the merged ensemble and the experimental SAXS, HRPF, and DFCS data (Figs. 6B-D) suggests that one of the best-fit ensembles is reached. Also, the cumulative weights analyses showed that the AA-2 and the FM still contribute more than 30% of the conformations in the final merged ensemble, suggesting their contributions of improvement over the AA-1 alone. However, the contribution of each initial ensemble is quite different in terms of reproducing the experimental data. For instance, AA-1 contributes the most and AA-2 the least to the final merged ensemble, presumably because the AA-1 alone can capture most of the experimental datasets while the AA-2 cannot, as shown in Fig. 5. This difference suggests the ensemble fitting method is effective in selecting meaningful conformations from a large pool of conformations.

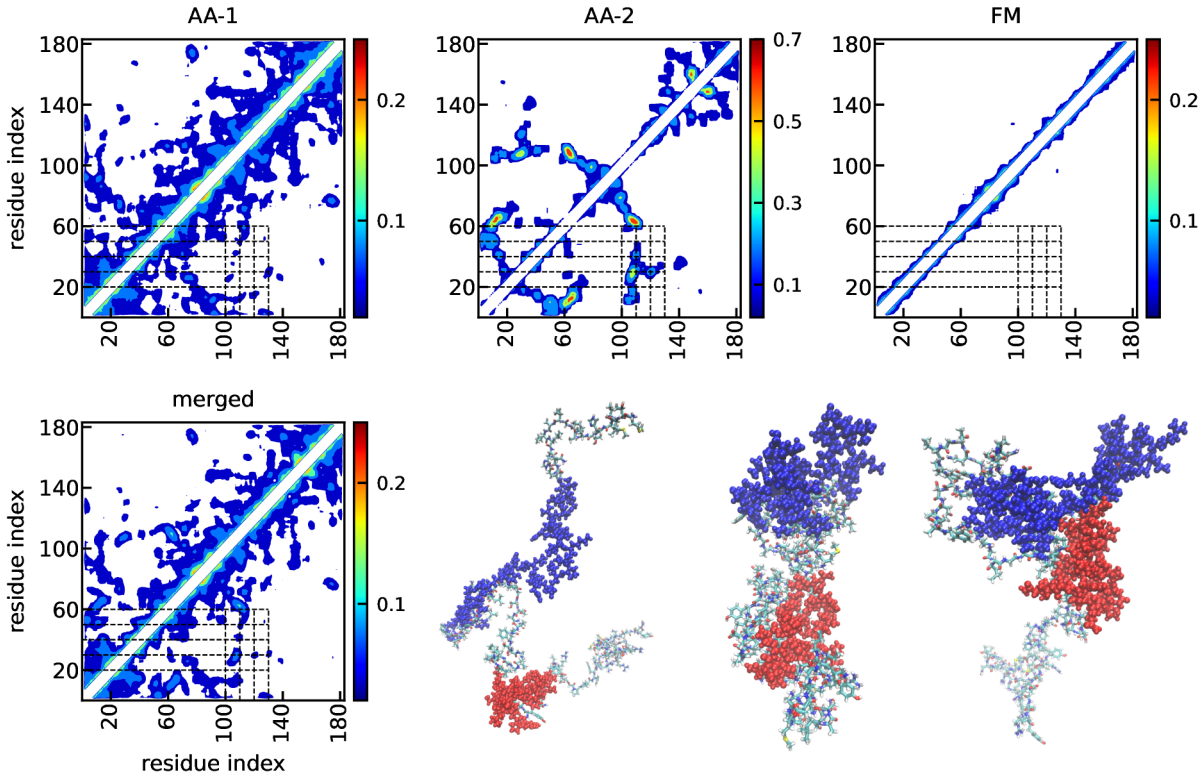


Figure 7. Ensemble-averaged contact maps using different initial ensembles of AA-1, AA-2, FM, and the merged pool with all three initial structures. All three experimental measurements of SAXS, HRPF, and DFCS data were used for the ensemble-structure fitting. The top three conformations in the final ensemble using the merged pool as the initial ensemble are shown in the bottom with blue highlighting residues 20 to 60 and red highlighting residues 100 to 130.

Only looking at the agreement with SAXS, HRPF and DFCS cannot tell the role of additional structures selected from the AA-2 and FM to characterize the experimental data since the AA-2 ensemble almost matches the experimental data same as the merged ensemble. Additional physical variables besides the experimental measurements must be calculated from the ensemble for further comparison. We now examine the NTD structural features via the ensemble-averaged contact maps obtained using three different initial ensembles (Fig. 7). Because of the vast

difference in the conformational space covered by each initial ensemble (Fig. 2A), it is not surprising that their resulting contact maps are quite different despite their optimization against the exactly same experimental restraints. This differentiation suggests that there is room for improvement when using the AA-1 ensemble alone or currently available experimental restraints.

However, by comparing these contact maps, there are commonly shared structural features among these refined ensembles, except for the one from FM with less long-range contact formation. Transient long-range interactions are formed between the N-terminal region centering around residue 33 and the central region around residue 118, as we previously validated using site-specific mutagenesis.[37] We note that such mutagenesis cannot rule out the role of other amino acids in proximity or other specific amino acids far away; as we move forward along this NTD study, it is expected that additional long-range contacts will be uncovered and communicated elsewhere. Here by comparing the contact maps from the AA-1 and the AA-2, it is consistent that interactions are abundant between two regions of amino acids, one from residues 20-60 and one from residues 100-130. This prediction confirms and extends the notion of the contact formation from a specific pair of amino acids, pointing to a broader picture that such transient interactions within an IDP are not only due to very few specific pairs of amino acids but also the collective cooperativity of multiple amino acids from for instance short linear motifs.[57, 58] Nonetheless, while comprehensive validation is needed for such a notion (e.g., via a secondary biophysical technique), the current ensemble-averaged contact maps provide an exemplary structural feature for the NTD disorder associated with regulating its hormone-independent ER activation.

Conclusions

Integrative biophysics involving multiple experimental methods has been a mainstay in the structural investigation of IDPs. In this work, we demonstrate the predictive power of combining experimental SAXS, HRPF, and DFCS measurements and computational methods (e.g., ensemble fitting and all-atom simulations) to interpret transient interactions of the NTD disorder. In doing so, we examine such ensemble-fitting and make notable observations from three different perspectives. First, extra information from site-specific solvent accessibility from DFCS (i.e., a set of 12 residues in terms of their solvent exposure) provides non-overlapping and complementary knowledge beyond previously existing HRPF data (a set of 16 residues in terms of their solvent exposure), thereby improving the ensemble-structure characterization considerably. Second, the role of three different initial ensembles (each generated from a different force field or a sampling/modeling method) was investigated to report that the choice of an initial ensemble of structures can lead to a distinct outcome even when each of them is fit against the same experimental measurements. As such, it is non-trivial to correlate the fitted ensemble with experimental restraints before one can confidently conclude the convergence and robustness of the ensemble. An attempt has been made to merge multiple initial ensembles. While it is not entirely conclusive to say such an effort substantially improves data interpretation, the comparison of their corresponding contact maps with shared structural features appears to increase the level of confidence. Finally, nonlocal or long-range transient interactions for the NTD disorder were found between two sequence segments/motifs, each containing tens of amino acids at the N-terminal and central regions. These transient interactions do not form with high probability compared with stable native contacts as in a folded protein. However, they could be sufficient to alter the conformational ensemble and thus regulate how the NTD interacts with its coactivator proteins critical for transactivation function.

ASSOCIATED CONTENT

Supporting Information. Supporting methods, figures, and tables for molecular dynamics simulation and ensemble fitting.

AUTHOR INFORMATION

Corresponding Author

*Email: wenweizheng@asu.edu

*Email: sichun.yang@case.edu

Acknowledgment

We acknowledge the support from the National Science Foundation (MCB-2015030 to W.Z.) and the National Institutes of Health (R01GM114056 to. S.Y. and R35GM146814 to W.Z.) as well as the research computing facility at Arizona State University.

ABBREVIATIONS

Intrinsically disordered protein, IDP; small-angle X-ray scattering, SAXS; hydroxyl radical protein footprinting, HRPF; D_2O-induced fluorine chemical shifting, DFCS; 3-Bromo-1,1,1-trifluoroacetone, BTFA

REFERENCES

(1) Wright, P. E.; Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Bio* **2015**, *16* (1), 18-29. DOI: 10.1038/nrm3920.

(2) Tompa, P. Intrinsically disordered proteins: a 10-year recap. *Trends Biochem Sci* **2012**, *37* (12), 509-516. DOI: 10.1016/j.tibs.2012.08.004.
(3) Dunker, A. K.; Silman, I.; Uversky, V. N.; Sussman, J. L. Function and structure of inherently disordered proteins. *Curr. Opin. Struct. Biol.* **2008**, *18*, 756-764.
(4) Forman-Kay, J. D.; Mittag, T. From sequence and forces to structure, function, and evolution of intrinsically disordered proteins. *Structure* **2013**, *21* (9), 1492-1499. DOI: 10.1016/j.str.2013.08.001.
(5) Best, R. B. Computational and theoretical advances in studies of intrinsically disordered proteins. *Curr Opin Struct Biol* **2017**, *42*, 147-154. DOI: 10.1016/j.sbi.2017.01.006.
(6) Das, R. K.; Ruff, K. M.; Pappu, R. V. Relating sequence encoded information to form and function of intrinsically disordered proteins. *Curr Opin Struct Biol* **2015**, *32*, 102-112. DOI: 10.1016/j.sbi.2015.03.008.
(7) Brookes, D. H.; Head-Gordon, T. Experimental Inferential Structure Determination of Ensembles for Intrinsically Disordered Proteins. *J Am Chem Soc* **2016**, *138* (13), 4530-4538. DOI: 10.1021/jacs.6b00351.
(8) Gomes, G. W.; Krzeminski, M.; Namini, A.; Martin, E. W.; Mittag, T.; Head-Gordon, T.; Forman-Kay, J. D.; Gradinaru, C. C. Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. *J Am Chem Soc* **2020**, *142* (37), 15697-15710. DOI: 10.1021/jacs.0c02088 From NLM Medline.
(9) Kofinger, J.; Stelzl, L. S.; Reuter, K.; Allande, C.; Reichel, K.; Hummer, G. Efficient Ensemble Refinement by Reweighting. *J Chem Theory Comput* **2019**, *15* (5), 3390-3401. DOI: 10.1021/acs.jctc.8b01231.
(10) Lazar, T.; Martinez-Perez, E.; Quaglia, F.; Hatos, A.; Chemes, L. B.; Iserte, J. A.; Mendez, N. A.; Garrone, N. A.; Saldano, T. E.; Marchetti, J.; et al. PED in 2021: a major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res* **2021**, *49* (D1), D404-D411. DOI: 10.1093/nar/gkaa1021.
(11) Jensen, M. R.; Salmon, L.; Nodet, G.; Blackledge, M. Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. *J Am Chem Soc* **2010**, *132* (4), 1270-1272. DOI: 10.1021/ja909973n.
(12) Borgia, A.; Zheng, W.; Buholzer, K.; Borgia, M. B.; Schuler, A.; Hofmann, H.; Soranno, A.; Nettels, D.; Gast, K.; Grishaev, A.; et al. Consistent View of Polypeptide Chain Expansion in Chemical Denaturants from Multiple Experimental Methods. *J Am Chem Soc* **2016**, *138* (36), 11714-11726. DOI: 10.1021/jacs.6b05917.
(13) Rauscher, S.; Gapsys, V.; Gajda, M. J.; Zweckstetter, M.; de Groot, B. L.; Grubmuller, H. Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J Chem Theory Comput* **2015**, *11* (11), 5513-5524. DOI: 10.1021/acs.jctc.5b00736.
(14) Fisher, C. K.; Stultz, C. M. Constructing ensembles for intrinsically disordered proteins. *Curr Opin Struct Biol* **2011**, *21* (3), 426-431. DOI: 10.1016/j.sbi.2011.04.001.
(15) Bernado, P.; Svergun, D. I. Structural analysis of intrinsically disordered proteins by small-angle X-ray scattering. *Mol Biosyst* **2012**, *8* (1), 151-167. DOI: 10.1039/c1mb05275f.
(16) Yang, S. Methods for SAXS-based structure determination of biomolecular complexes. *Adv Mater* **2014**, *26* (46), 7902-7910. DOI: 10.1002/adma.201304475.
(17) Nettels, D.; Müller-Späth, S.; Küster, F.; Hofmann, H.; Haenni, D.; Rüegger, S.; Reymond, L.; Hoffmann, A.; Kubelka, J.; Heinz, B.; et al. Single molecule spectroscopy of the temperature-induced collapse of unfolded proteins. *Proc Natl Acad Sci USA* **2009**, *106*, 20740-20745.

(18) Clore, G. M.; Iwahara, J. Theory, practice, and applications of paramagnetic relaxation enhancement for the characterization of transient low-population states of biological macromolecules and their complexes. *Chem Rev* **2009**, *109* (9), 4108-4139. DOI: 10.1021/cr900033p.
(19) Wishart, D. S.; Sykes, B. D.; Richards, F. M. Relationship between nuclear magnetic resonance chemical shift and protein secondary structure. *J Mol Biol* **1991**, *222* (2), 311-333. DOI: 10.1016/0022-2836(91)90214-q.
(20) Mantsyzov, A. B.; Maltsev, A. S.; Ying, J.; Shen, Y.; Hummer, G.; Bax, A. A maximum entropy approach to the study of residue-specific backbone angle distributions in alpha-synuclein, an intrinsically disordered protein. *Protein Sci* **2014**, *23* (9), 1275-1290. DOI: 10.1002/pro.2511.
(21) Klein-Seetharaman, J.; Oikawa, M.; Grimshaw, S. B.; Wirmer, J.; Duchardt, E.; Ueda, T.; Imoto, T.; Smith, L. J.; Dobson, C. M.; Schwalbe, H. Long-range interactions within a nonnative protein. *Science* **2002**, *295*, 1719-1722.
(22) Abyzov, A.; Salvi, N.; Schneider, R.; Maurin, D.; Ruigrok, R. W.; Jensen, M. R.; Blackledge, M. Identification of Dynamic Modes in an Intrinsically Disordered Protein Using Temperature-Dependent NMR Relaxation. *J Am Chem Soc* **2016**, *138* (19), 6240-6251. DOI: 10.1021/jacs.6b02424.
(23) Kelly, S. M.; Jess, T. J.; Price, N. C. How to study proteins by circular dichroism. *Biochim Biophys Acta* **2005**, *1751* (2), 119-139. DOI: 10.1016/j.bbapap.2005.06.005.
(24) Wells, M.; Tidow, H.; Rutherford, T. J.; Markwick, P.; Jensen, M. R.; Mylonas, E.; Svergun, D. I.; Blackledge, M.; Fersht, A. R. Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. *Proc Natl Acad Sci U S A* **2008**, *105* (15), 5762-5767. DOI: 10.1073/pnas.0801353105 From NLM Medline.
(25) Aznauryan, M.; Delgado, L.; Soranno, A.; Nettels, D.; Huang, J. R.; Labhardt, A. M.; Grzesiek, S.; Schuler, B. Comprehensive structural and dynamical view of an unfolded protein from the combination of single-molecule FRET, NMR, and SAXS. *Proc Natl Acad Sci U S A* **2016**, *113* (37), E5389-5398. DOI: 10.1073/pnas.1607193113.
(26) Huang, W.; Ravikumar, K. M.; Chance, M. R.; Yang, S. Quantitative mapping of protein structure by hydroxyl radical footprinting-mediated structural mass spectrometry: a protection factor analysis. *Biophys J* **2015**, *108* (1), 107-115. DOI: 10.1016/j.bpj.2014.11.013.
(27) Kitevski-LeBlanc, J. L.; Prosser, R. S. Current applications of 19F NMR to studies of protein structure and dynamics. *Prog Nucl Magn Reson Spectrosc* **2012**, *62*, 1-33. DOI: 10.1016/j.pnmrs.2011.06.003.
(28) Hocking, H. G.; Zangger, K.; Madl, T. Studying the structure and dynamics of biomolecules by using soluble paramagnetic probes. *Chemphyschem : a European journal of chemical physics and physical chemistry* **2013**, *14* (13), 3082-3094. DOI: 10.1002/cphc.201300219 From NLM Medline.
(29) Gong, Z.; Gu, X. H.; Guo, D. C.; Wang, J.; Tang, C. Protein Structural Ensembles Visualized by Solvent Paramagnetic Relaxation Enhancement. *Angew Chem Int Ed Engl* **2017**, *56* (4), 1002-1006. DOI: 10.1002/anie.201609830 From NLM Medline.
(30) Yang, S.; Bernado, P. Integrative Biophysics: Protein Interaction and Disorder. *J Mol Biol* **2020**, *432* (9), 2843-2845. DOI: 10.1016/j.jmb.2020.04.001 From NLM Medline.
(31) Lindorff-Larsen, K.; Best, R. B.; Depristo, M. A.; Dobson, C. M.; Vendruscolo, M. Simultaneous determination of protein structure and dynamics. *Nature* **2005**, *433*, 128-132.

(32) Huang, W.; Ravikumar, K. M.; Parisien, M.; Yang, S. Theoretical modeling of multiprotein complexes by iSPOT: Integration of small-angle X-ray scattering, hydroxyl radical footprinting, and computational docking. *J Struct Biol* **2016**, *196* (3), 340-349. DOI: 10.1016/j.jsb.2016.08.001.
(33) Gomes, G.-N. W.; Namini, A.; Gradinaru, C. C. Integrative conformational ensembles of Sic1 using different initial pools and optimization methods. *Front. Mol. Biosci.* **2022**, *9*, 910956.
(34) Kato, S.; Endoh, H.; Masuhiro, Y.; Kitamoto, T.; Uchiyama, S.; Sasaki, H.; Masushige, S.; Gotoh, Y.; Nishida, E.; Kawashima, H.; et al. Activation of the estrogen receptor through phosphorylation by mitogen-activated protein kinase. *Science* **1995**, *270* (5241), 1491-1494. DOI: 10.1126/science.270.5241.1491.
(35) Warnmark, A.; Wikstrom, A.; Wright, A. P.; Gustafsson, J. A.; Hard, T. The N-terminal regions of estrogen receptor alpha and beta are unstructured in vitro and show different TBP binding properties. *J Biol Chem* **2001**, *276* (49), 45939-45944. DOI: 10.1074/jbc.M107875200.
(36) Rajbhandari, P.; Finn, G.; Solodin, N. M.; Singarapu, K. K.; Sahu, S. C.; Markley, J. L.; Kadunc, K. J.; Ellison-Zelski, S. J.; Kariagina, A.; Haslam, S. Z.; et al. Regulation of estrogen receptor alpha N-terminus conformation and function by peptidyl prolyl isomerase Pin1. *Mol Cell Biol* **2012**, *32* (2), 445-457. DOI: MCB.06073-11 [pii] 10.1128/MCB.06073-11.
(37) Peng, Y.; Cao, S.; Kiselar, J.; Xiao, X.; Du, Z.; Hsieh, A.; Ko, S.; Chen, Y.; Agrawal, P.; Zheng, W.; et al. A Metastable Contact and Structural Disorder in the Estrogen Receptor Transactivation Domain. *Structure* **2019**, *27* (2), 229-240. DOI: 10.1016/j.str.2018.10.026.
(38) 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.
(39) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79* (2), 926-935.
(40) Best, R. B.; Zheng, W.; Mittal, J. Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *J Chem Theory Comput* **2014**, *10* (11), 5113-5124. DOI: 10.1021/ct500569b.
(41) Abascal, J. L. F.; Vega, C. A general purpose model for the condensed phases of water: TIP4P/2005. *J. Chem. Phys* **2005**, *123*, 234505.
(42) Ozenne, V.; Bauer, F.; Salmon, L.; Huang, J. R.; Jensen, M. R.; Segard, S.; Bernado, P.; Charavay, C.; Blackledge, M. Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics* **2012**, *28* (11), 1463-1470. DOI: 10.1093/bioinformatics/bts172.
(43) Huang, X.; Pearce, R.; Zhang, Y. FASPR: an open-source tool for fast and accurate protein side-chain packing. *Bioinformatics* **2020**, *36* (12), 3758-3765. DOI: 10.1093/bioinformatics/btaa234.
(44) Ravikumar, K. M.; Huang, W.; Yang, S. Fast-SAXS-pro: a unified approach to computing SAXS profiles of DNA, RNA, protein, and their complexes. *J Chem Phys* **2013**, *138* (2), 024112. DOI: 10.1063/1.4774148.
(45) Sawle, L.; Ghosh, K. A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *J Chem Phys* **2015**, *143* (8), 085101. DOI: 10.1063/1.4929391.

(46) Zheng, W.; Dignon, G.; Brown, M.; Kim, Y. C.; Mittal, J. Hydropathy Patterning Complements Charge Patterning to Describe Conformational Preferences of Disordered Proteins. *J Phys Chem Lett* **2020**, *11* (9), 3408-3415. DOI: 10.1021/acs.jpclett.0c00288.
(47) Campos-Olivas, R.; Aziz, R.; Helms, G. L.; Evans, J. N.; Gronenborn, A. M. Placement of 19F into the center of GB1: effects on structure and stability. *FEBS letters* **2002**, *517* (1-3), 55-60.
(48) Chrisman, I. M.; Nemetchek, M. D.; de Vera, I. M. S.; Shang, J.; Heidari, Z.; Long, Y.; Reyes-Caballero, H.; Galindo-Murillo, R.; Cheatham, T. E., 3rd; Blayo, A. L.; et al. Defining a conformational ensemble that directs activation of PPARgamma. *Nat Commun* **2018**, *9* (1), 1794. DOI: 10.1038/s41467-018-04176-x.
(49) Piana, S.; Donchev, A. G.; Robustelli, P.; Shaw, D. E. Water Dispersion Interactions Strongly Influence Simulated Structural Properties of Disordered Protein States. *J Phys Chem B* **2015**, *119* (16), 5113-5123. DOI: 10.1021/jp508971m.
(50) Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B. L.; Grubmuller, H.; MacKerell, A. D. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* **2017**, *14* (1), 71-73. DOI: 10.1038/Nmeth.4067.
(51) Robustelli, P.; Piana, S.; Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc Natl Acad Sci U S A* **2018**, *115* (21), E4758-E4766. DOI: 10.1073/pnas.1800690115.
(52) Wu, H. N.; Jiang, F.; Wu, Y. D. Significantly Improved Protein Folding Thermodynamics Using a Dispersion-Corrected Water Model and a New Residue-Specific Force Field. *J Phys Chem Lett* **2017**, *8* (14), 3199-3205. DOI: 10.1021/acs.jpclett.7b01213.
(53) Song, D.; Liu, H.; Luo, R.; Chen, H. F. Environment-Specific Force Field for Intrinsically Disordered and Ordered Proteins. *J Chem Inf Model* **2020**, *60* (4), 2257-2267. DOI: 10.1021/acs.jcim.0c00059.
(54) Shabane, P. S.; Izadi, S.; Onufriev, A. V. General Purpose Water Model Can Improve Atomistic Simulations of Intrinsically Disordered Proteins. *J Chem Theory Comput* **2019**, *15* (4), 2620-2634. DOI: 10.1021/acs.jctc.8b01123.
(55) Lincoff, J.; Sasmal, S.; Head-Gordon, T. The combined force field-sampling problem in simulations of disordered amyloid-beta peptides. *J Chem Phys* **2019**, *150* (10), 104108. DOI: 10.1063/1.5078615.
(56) Zhang, O.; Haghighatlari, M.; Li, J.; Teixeira, J. M. C.; Namini, A.; Liu, Z.-H.; Forman-Kay, J. D.; Head-Gordon, T. Learning to Evolve Structural Ensembles of Unfolded and Disordered Proteins Using Experimental Solution Data. *arXiv* **2022**, https://arxiv.org/abs/2206.12667.
(57) Bhowmick, P.; Guharoy, M.; Tompa, P. Bioinformatics Approaches for Predicting Disordered Protein Motifs. *Adv Exp Med Biol* **2015**, *870*, 291-318. DOI: 10.1007/978-3-319-20164-1_9.
(58) Cohan, M. C.; Shinn, M. K.; Lalmansingh, J. M.; Pappu, R. V. Uncovering non-random binary patterns within sequences of intrinsically disordered proteins. *J Mol Biol* **2022**, *434* (2), 167373.

Supplementary information for "Incorporation of D_2O-induced Fluorine Chemical Shift Perturbations into Ensemble-Structure Characterization of the ERalpha Disordered Region"

Wenwei Zheng[1], Zhanwen Du[2], Soo Bin Ko[2], Nalinda Wickramasinghe[3], Sichun Yang[2]

[1]College of Integrative Sciences and Arts, Arizona State University, Mesa, Arizona 85212, United States

[2]Center for Proteomics and Department of Nutrition, School of Medicine, Case Western Reserve University, Cleveland, Ohio, 44106, United States

[3]Chemistry-NMR Facility, Case Western Reserve University, Cleveland, Ohio 44106, United States

Supporting Methods

Molecular simulations for the generation of initial ensemble-structures. The first all-atom ensemble (AA-1) is from the previous literature[1]. A pool of 35,240 NTD candidate structures was generated from molecular dynamics simulations with a total accumulative time of 35 μs in a 1-ns recording frequency. Two advanced algorithms were used: Gaussian accelerated molecular dynamics (GaMD)[2] and replica exchange solute tempering (REST2).[3] First, a set of 25 GaMD trajectories (using the software AMBER16,[4] each starting with a random configuration and lasting 1 ns) resulted in a total of 25 μs. Second, a set of 64 replicas in REST2 simulations, ranging from 300 K to 600 K were performed at the Argonne Leadership Computing Facility using the software NAMD as previously described.[5, 6] Each replica lasted 160 ns, which resulted in 10 μs. In both GaMD and REST2 simulations, a molecular Amber ff99sb[7] force field and a TIP3P water model[8] were used.

The second ensemble (AA-2) was generated using an all-atom force field Amber99sbws[9] with a TIP4P/2005 water model.[10] The simulations were performed using Gromacs 4.6.7[11] at a constant temperature of 300 K maintained by a Langevin thermostat with a friction coefficient of 1 ps^{-1} and pressure of 1 bar using a Parrinello-Rahman barostat.[12] The time step was 2 fs. Electrostatic energies and forces were computed with particle-mesh Ewald[13] using a 0.12-nm grid spacing and real-space cutoff of 0.9nm. Lennard-Jones interactions were calculated using a twin-range scheme with inner and outer cutoffs of 0.9 and 1.4 nm, respectively. The box was set to be rhombic dodecahedron and the shortest distance between periodic image was 12 nm. The simulation was run for a total of 4 μs, and 20,000 candidate structures were evenly selected as the initial ensemble.

The third ensemble (FM) was generated using the Flexible-Meccano model[14] to obtain a coarse-grained representation including C_α and C_β atoms (https://www.ibs.fr/research/scientific-output/software/flexible-meccano. The method was developed by Blackledge and colleagues to generate a large set of conformations in a time-efficient fashion by accounting for residue-specific backbone dihedral angle propensities derived from high-resolution X-ray crystallographic structures. For each residue, missing atoms including the rest of its sidechain were added via the FASPR algorithm.[15] A total of 20,000 candidate structures were generated in this initial ensemble.

Theoretical calculations for DFCS data interpretation. To calculate the solvent accessible surface area of the CF_3 group attached onto the cysteine sidechain, the force field parameters for cysteine with a -CH_2COCF_3 group attached were parameterized using the Amber99sbws force field. The partial charges were determined via RESP using the ANTECHAMBER program.[16] To be consistent with the Amber force field,[17] electrostatic potentials were determined with a restricted Hartree-Fock method and the 6-31G∗ basis set, using the GAUSSIAN software package.[18] The obtained partial charges and atom types are shown in Table S2. For each individual labeling position, the serine was mutated to cysteine with -CH_2COCF_3 using tleap in Ambertools[4], and each resulting configuration was energy minimized before the solvent accessible surface area of each corresponding CF_3 group was calculated.

Ensemble fitting. The fitting against different sources of experimental measurements was achieved by varying a probability distribution of $\{w_i\}$ for each structure i from the ensemble by minimizing a scoring function of

$$F(\{w_i\}) = \chi^2_{SAXS}(\{w_i\}) + \alpha\chi^2_{HRPF}(\{w_i\}) + \beta\chi^2_{DFCS}(\{w_i\}) - T_{fit}S(\{w_i\}).$$

$\chi^2_{SAXS}(\{w_i\})$ is the goodness of fit to characterize the deviation between the calculated ensemble-weighted SAXS profile and experimental SAXS data via

$$\chi^2_{SAXS}(\{w_i\}) = \frac{1}{N_q}\Sigma_q \frac{\{logI_{exp}(q) - \Sigma_{i=1}^{N} \log [w_i \cdot I_{calc,i}(q)]\}^2}{\sigma^2_{exp}(q)},$$

where N is the total number of structures in the ensemble. $I_{calc}(q)$ is the calculated SAXS profile via Fast-SAXS-pro[19], while $I_{exp}(q)$ is its experimental SAXS profile with a total of N_q data points. $\chi^2_{HRPF}(\{w_i\})$ is the deviation between the ensemble-weighted solvent accessible surface area of each residue sidechain $SA_{sidechain}$ and experimental HRPF measurements via

$$\chi^2_{HRPF}(\{w_i\}) = \frac{1}{N_{PF}}\Sigma_s \frac{[\log PF(s) - \Sigma_{i=1}^{N} w_i \cdot SA_{sidechain,i}(s)]^2}{\sigma^2_{PF}(s)},$$

where s represents each specific amino acid probed by HRPF with a total of N_{HRPF} =16 amino acids. The ensemble-weighted $SA_{sidechain}$ ($\Sigma_{i=1}^{N} w_i \cdot SA_{sidechain,i}$) was scaled and offset via the linear regression with experimental PF (protection factor) values prior to χ^2_{HRPF} calculations. χ^2_{DFCS} describes the deviation between the ensemble-weighted solvent accessible surface area of each CF_3 group (SA_{CF3}) and its experimental DFCS data (i.e., the linear slope of peak shift as a function of D_2O concentrations as shown in Fig. 2C) via

$$\chi^2_{DFCS}(\{w_i\}) = \frac{1}{N_{DFCS}}\Sigma_c \frac{[\,DFCS(c) - \Sigma_{i=1}^{N} w_i \cdot SA_{CF_3,i}(c)]^2}{\sigma^2_{DFCS}(c)},$$

where c represents one CF_3-tagged amino acid position probed by DFCS with a total number of N_{DFCS} =12 amino acid positions. The ensemble-weighted SA_{CF3} ($\Sigma_{i=1}^{N} w_i \cdot SA_{CF_3,i}$) was scaled and offset using the linear regression with the experimental DFCS values prior to the χ^2_{DFCS} ($\{w_i\}$) calculations. In addition to the three deviation terms, a Shannon entropy term $T_{fit}S(\{w_i\})$ was added to prevent overfitting. An effective temperature T_{fit} is used as a variable to control the biasing weights to satisfy a minimal deviation from the initial ensemble while

counterintuitively attempting to achieve a minimal number of structures needed for best-fitting. The parameter $T_{fit}$was chosen to have the lowest $F(\{w_i\})$ score, practically at the point where the summation of the three deviation terms begins to sharply increase along *S* (see Fig. S4). The parameters α and β are the relative weight of contribution of different experimental measurements and were selected so that all the three χ^2 terms can be minimized reasonably well. The final parameters were shown in Table S3. Simulated annealing in the $\{w_i\}$ space was conducted to minimize the $F(\{w_i\})$ score. The annealing was repeated 50 times for each condition and their averaging weights were reported.

Supporting Figures

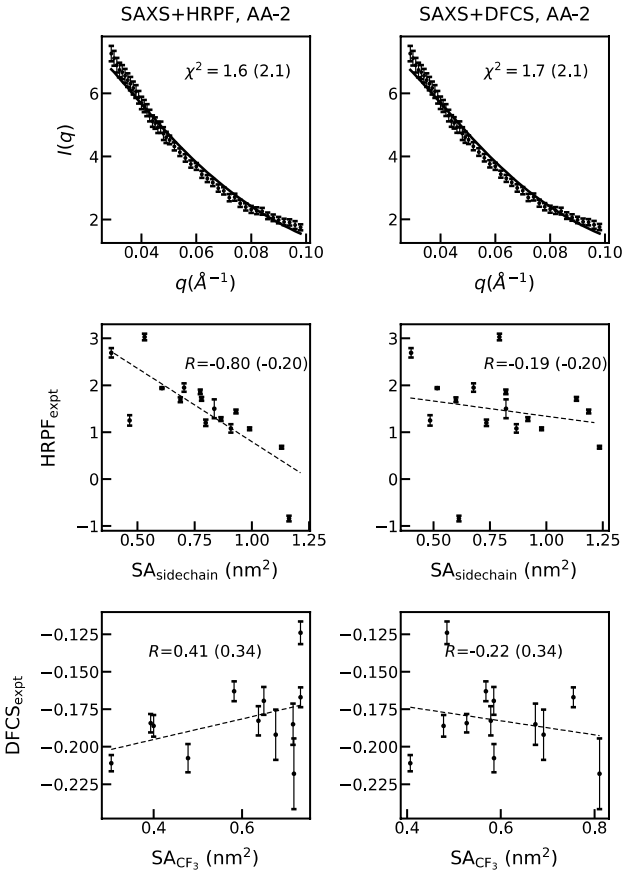


Figure S1. Refining the AA-2 ensemble using experimental SAXS+HRPF (left) and SAXS+DFCS data (right). A) Comparison with SAXS data with χ^2 shown in the legend. B) Comparison between the calculated solvent accessible surface areas of residue sidechains with experimental HRPF data. *R*, Pearson correlation coefficient. C) Comparison between the calculated solvent accessible surface area of each CF_3 group with experimental DFCS data.

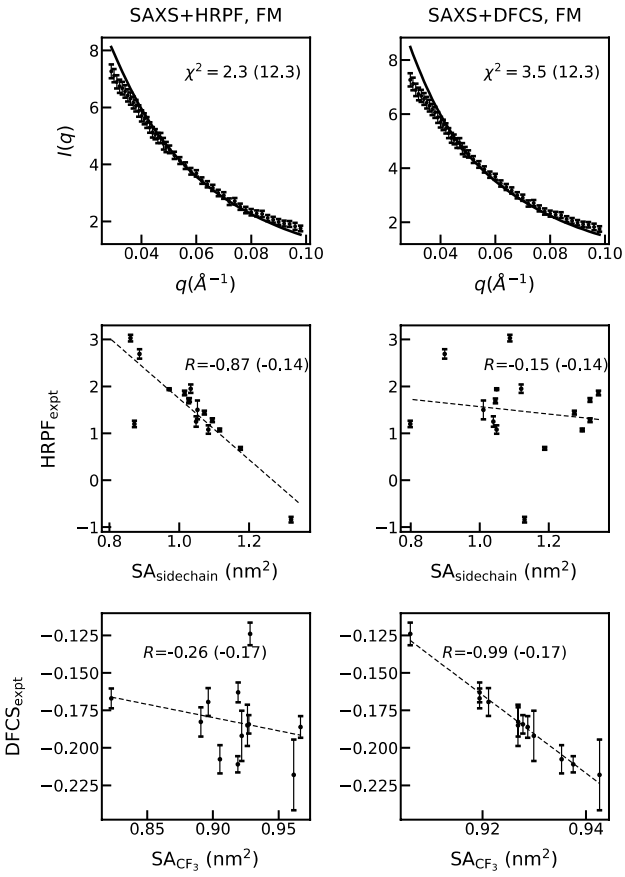


Figure S2. Refining the FM ensemble using SAXS+HRPF (left) and SAXS+DFCS (right). A) Comparison with SAXS data with χ^2 shown in the legend. B) Comparison between the solvent accessible surface area of sidechain from simulations with the experimental HRPF data. Pearson correlation coefficient is shown in the legend. C) Comparison between the calculated solvent

accessible surface area of each CF_3 group with experimental DFCS data. R, Pearson correlation coefficient.

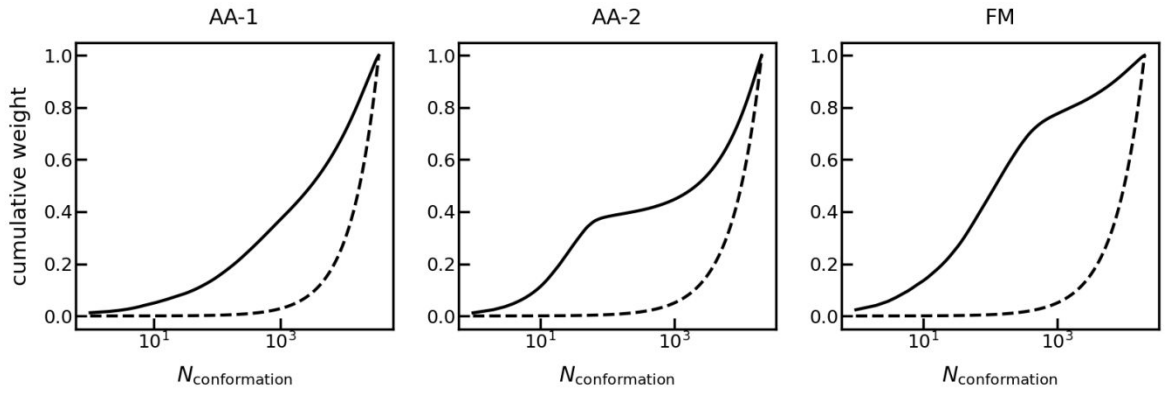


Figure S3. Cumulative weight as a function of the number of conformations in the refined ensembles using the AA-1 (left), AA-2 (middle), and FM (right). Dashed line, a reference ensemble with uniform weights.

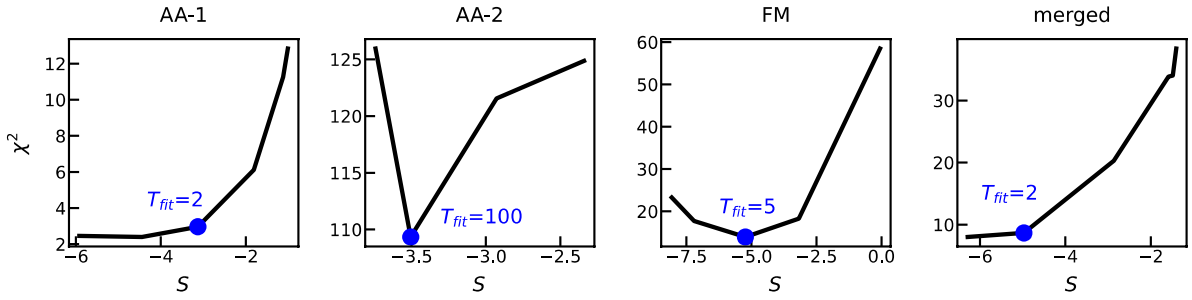


Figure S4. Deviation of the weighted ensemble from the experimental measurements $\chi^2(\{w_i\}) = \chi^2_{SAXS}(\{w_i\}) + \alpha\chi^2_{HRPF}(\{w_i\}) + \beta\chi^2_{DFCS}(\{w_i\})$ as a function of the entropy S in each ensemble. Solid blue symbol, the T_{fit} of the ensemble reported for each case.

Supporting Tables

Table S1. Sequence descriptors for the NTD. f+: fraction of positively charged amino acids. f-: fraction of negatively charged amino acids; FCR: fraction of charged amino acids; NCPR: net charge per residue[20]; κ: charge patterning parameter[21]; SCD: sequence charge decoration[22]; $f_{aromatic}$: fraction of aromatic amino acids; <λ>: mean hydropathy using a scale from a previous literature[23]; SHD: sequence hydropathy decoration[24]; $\nu_{predicted}$: scaling exponent predicted using SCD and SHD[24].

	NTD	Residue 1-50	Residue 51-140	Residue 141-184
f+	0.087	0.120	0.011	0.205
f-	0.087	0.100	0.056	0.136
FCR	0.174	0.220	0.067	0.341
NCPR	0.000	0.020	-0.044	0.068
κ	0.135	0.225	0.301	0.068
SCD	-0.383	-0.242	0.384	0.075
$f_{aromatic}$	0.087	0.020	0.133	0.068
<λ>	0.693	0.702	0.753	0.561
SHD	6.677	4.899	6.159	3.759
$\nu_{predicted}$	0.415	0.492	0.443	0.543

Table S2. The force field parameters for cysteine with -CH_2COCF_3 attached.

Atom name	Atom type	Partial charge
N	N	-0.29579
H	H	0.23899
CA	CT	-0.23163
HA	H1	0.14578
CB	CT	-0.20188
HB1	H1	0.15933
HB2	H1	0.15933
SG	S	-0.18586
CD	CT	-0.26195
HD1	H1	0.14093
HD2	H1	0.14093
CE	C	0.54917
OE	O	-0.44204
CF	CT	0.57915
FF1	F	-0.20900

FF2	F	-0.20900
FF3	F	-0.20900
C	C	0.71584
O	O	-0.58330

Table S3. Ensemble fitting parameters.

Initial ensemble	α	β	T_{fit}
AA-1	0.03	0.3	2
AA-2	0.2	5.0	100
FM	0.2	0.2	5
merged	0.1	1.0	2

REFERENCES

1. Peng, Y.; Cao, S.; Kiselar, J.; Xiao, X.; Du, Z.; Hsieh, A.; Ko, S.; Chen, Y.; Agrawal, P.; Zheng, W., et al., A Metastable Contact and Structural Disorder in the Estrogen Receptor Transactivation Domain. *Structure* **2019,** *27* (2), 229-240.
2. Miao, Y.; Feher, V. A.; McCammon, J. A., Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. *J Chem Theory Comput* **2015,** *11* (8), 3584-3595.
3. Wang, L.; Friesner, R. A.; Berne, B. J., Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2). *J Phys Chem B* **2011,** *115* (30), 9431-8.
4. Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E., III; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P., AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comp. Phys. Comm.* **1995,** *91*, 1-41.
5. Jo, S.; Jiang, W., A generic implementation of replica exchange with solute tempering (REST2) algorithm in NAMD for complex biophysical simulations. *Comput Phys Commun* **2015,** *197*, 304-311.
6. Phillips, J. C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R. D.; Kale, L.; Schulten, K., Scalable molecular dynamics with NAMD. *J Comput Chem* **2005,** *26* (16), 1781-1802.
7. 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.
8. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D., Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983,** *79* (2), 926-935.
9. Best, R. B.; Miller, C.; Mittal, J., Role of solvation in pressure-induced helix stabilization. *J. Chem. Phys.* **2014,** *141* (22), 22D522.

10. Abascal, J. L. F.; Vega, C., A general purpose model for the condensed phases of water: TIP4P/2005. *J. Chem. Phys* **2005,** *123*, 234505.
11. Hess, B.; Kutzner, C.; Van der Spoel, D.; Lindahl, E., GROMACS4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theor. Comput.* **2008,** *4* (3), 435-447.
12. Parrinello, M.; Rahman, A., Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* **1981,** *52* (12), 7182-7190.
13. Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G., A smooth particle mesh ewald potential. *J. Chem. Phys* **1995,** *103*, 8577-8592.
14. Ozenne, V.; Bauer, F.; Salmon, L.; Huang, J. R.; Jensen, M. R.; Segard, S.; Bernado, P.; Charavay, C.; Blackledge, M., Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics* **2012,** *28* (11), 1463-70.
15. Huang, X.; Pearce, R.; Zhang, Y., FASPR: an open-source tool for fast and accurate protein side-chain packing. *Bioinformatics* **2020,** *36* (12), 3758-3765.
16. Bayly, C. I.; Cieplak, P.; Cornell, W.; Kollman, P. A., A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the RESP model. *J. Phys. Chem.* **1993,** *97*, 10269-10280.
17. Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P. A., A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.* **1995,** *117* (19), 5179-5197.
18. Frisch, M. J. T., G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Montgomery, J., J. A.; Vreven, T.; Kudin, K. N.; Burant, J. C.; Millam, J. M.; Iyengar, S. S.; Tomasi, J.; Barone, V.; Mennucci, B.; Cossi, M.; Scalmani, G.; Rega, N.; Petersson, G. A.; Nakatsuji, H.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Klene, M.; Li, X.; Knox, J. E.; Hratchian, H. P.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Ayala, P. Y.; Morokuma, K.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Zakrzewski, V. G.; Dapprich, S.; Daniels, A. D.; Strain, M. C.; Farkas, O.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Ortiz, J. V.; Cui, Q.; Baboul, A. G.; Clifford, S.; Cioslowski, J.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Challacombe, M.; Gill, P. M. W.; Johnson, B.; Chen, W.; Wong, M. W.; Gonzalez, C.; Pople, J. A. Gaussian, Inc, Wallingford CT: 2004.
19. Ravikumar, K. M.; Huang, W.; Yang, S., Fast-SAXS-pro: a unified approach to computing SAXS profiles of DNA, RNA, protein, and their complexes. *J Chem Phys* **2013,** *138* (2), 024112.
20. Mao, A. H.; Crick, S. L.; Vitalis, A.; Chicoine, C. L.; Pappu, R. V., Net charge per residue modulates conformational ensembles of intrinsically disordered proteins. *Proc Natl Acad Sci USA* **2010,** *107* (18), 8183-8188.
21. Das, R. K.; Pappu, R. V., Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc Natl Acad Sci U S A* **2013,** *110* (33), 13392-7.

22. Sawle, L.; Ghosh, K., A theoretical method to compute sequence dependent configurational properties in charged polymers and proteins. *J Chem Phys* **2015,** *143* (8), 085101.
23. Kapcha, L. H.; Rossky, P. J., A simple atomic-level hydrophobicity scale reveals protein interfacial structure. *J Mol Biol* **2014,** *426* (2), 484-498.
24. Zheng, W.; Dignon, G.; Brown, M.; Kim, Y. C.; Mittal, J., Hydropathy Patterning Complements Charge Patterning to Describe Conformational Preferences of Disordered Proteins. *J Phys Chem Lett* **2020,** *11* (9), 3408-3415.