frontiers | Frontiers in Molecular Biosciences

ORIGINAL RESEARCH
published: 18 July 2022
doi: 10.3389/fmolb.2022.910956

Integrative Conformational Ensembles of Sic1 Using Different Initial Pools and Optimization Methods

Gregory-Neal W. Gomes[1*†], *Ashley Namini*[2] *and Claudiu C. Gradinaru*[1,2*]

[1]Department of Physics, University of Toronto, Toronto, ON, Canada, [2]Department of Chemical & Physical Sciences, University of Toronto Mississauga, Mississauga, ON, Canada

Intrinsically disordered proteins play key roles in regulatory protein interactions, but their detailed structural characterization remains challenging. Here we calculate and compare conformational ensembles for the disordered protein Sic1 from yeast, starting from initial ensembles that were generated either by statistical sampling of the conformational landscape, or by molecular dynamics simulations. Two popular, yet contrasting optimization methods were used, ENSEMBLE and Bayesian Maximum Entropy, to achieve agreement with experimental data from nuclear magnetic resonance, small-angle X-ray scattering and single-molecule Förster resonance energy transfer. The comparative analysis of the optimized ensembles, including secondary structure propensity, inter-residue contact maps, and the distributions of hydrogen bond and pi interactions, revealed the importance of the physics-based generation of initial ensembles. The analysis also provides insights into designing new experiments that report on the least restrained features among the optimized ensembles. Overall, differences between ensembles optimized from different priors were greater than when using the same prior with different optimization methods. Generating increasingly accurate, reliable and experimentally validated ensembles for disordered proteins is an important step towards a mechanistic understanding of their biological function and involvement in various diseases.

Keywords: smFRET, NMR, SAXS, contact maps, molecular dynamics, IDP ensembles

OPEN ACCESS

Edited by:
Haydyn David Thomas Mertens, European Molecular Biology Laboratory Hamburg, Germany

Reviewed by:
Arnab Bhattacherjee, Jawaharlal Nehru University, India
Steven T Whitten, Texas State University, United States

****Correspondence:***
Gregory-Neal W. Gomes
gregory-neal.gomes@yale.edu
Claudiu C. Gradinaru
claudiu.gradinaru@utoronto.ca

†Present address:
Gregory-Neal W. Gomes, Department of Pathology, Yale School of Medicine, New Haven, CT, United States

Specialty section:
This article was submitted to Structural Biology, a section of the journal Frontiers in Molecular Biosciences

Received: *01 April 2022*
Accepted: *21 June 2022*
Published: *18 July 2022*

Citation:
Gomes G-NW, Namini A and Gradinaru CC (2022) Integrative Conformational Ensembles of Sic1 Using Different Initial Pools and Optimization Methods. Front. Mol. Biosci. 9:910956. doi: 10.3389/fmolb.2022.910956

1 INTRODUCTION

Important biological functions performed by intrinsically disordered proteins (IDPs), such as cell signaling and regulation (Dyson and Wright, 2005; Forman-Kay and Mittag, 2013; Oldfield and Dunker, 2014), are mediated by their interesting and nonrandom structural properties. Conversely, their dysfunction or pathological aggregation is accompanied or preceded by aberrations in these structural properties (Uversky, 2015). Describing the molecular features of IDPs at atomistic resolution would therefore provide valuable mechanistic insight into how IDPs (mal) function. Molecular dynamics (MD) simulations have recently attempted to fill this gap, including development of new force fields to accurately model disordered proteins (Best et al., 2014; Rauscher et al., 2015). However, a unique parametrization of force fields suitable for modelling IDPs is yet to emerge, and atomistic-level simulations over biologically relevant timescales remain computationally expensive. Alternatively, disordered proteins can be represented by a conformational ensemble, which is a finite set of 3D structures with corresponding statistical weights. These ensembles are commonly determined by reweighting or selecting a subset from an

initial pool of conformations according to a protocol which optimizes agreement with various experimental data, while considering experimental uncertainties and avoiding overfitting (Krzeminski et al., 2013; Jensen et al., 2014; Leung et al., 2016; Bonomi et al., 2017; Köfinger et al., 2019; Bottaro et al., 2020; Lincoff et al., 2020; Orioli et al., 2020; Lazar et al., 2021).

Recent and rapid progress in the field of protein disorder necessitates a re-examination of the ensemble determination process. Mutual consistency and complementarity have been demonstrated for the three most commonly used structural techniques for IDPs: Small Angle X-Ray Scattering (SAXS), Nuclear Magnetic Resonance (NMR) and single-molecule Förster Resonance Energy Transfer (smFRET) (Delaforge et al., 2015; Aznauryan et al., 2016; Voithenberg et al., 2016; Fuertes et al., 2017; Gomes et al., 2020; Lincoff et al., 2020; Naudi-Fabra et al., 2021). Technological advances and efforts to standardize data collection and reporting have also been made for SAXS (Martin et al., 2020), smFRET (Hellenkamp et al., 2018; Lerner et al., 2021) and NMR (Dyson and Wright, 2019; Alderson and Kay, 2021; Dyson and Wright, 2021). Improvements in the accuracy of MD force fields, which correct earlier bias toward overly compact IDP conformations (Best et al., 2014; Rauscher et al., 2015; Huang et al., 2017; Robustelli et al., 2018), have advanced their use for generating initial pools of conformers. Protocols for calculating ensembles (Leung et al., 2016; Köfinger et al., 2019; Bottaro et al., 2020; Lincoff et al., 2020) and for predicting experimental data from structures (Kalinin et al., 2012; Henriques et al., 2018; Crehuet et al., 2019; Dimura et al., 2020; Pesce and Lindorff-Larsen, 2021; Tesei et al., 2021) continue to be developed and refined. As a result of all these developments, the repository of IDP ensembles validated by agreement with experimental data, the Protein Ensemble Database, has recently undergone a major update (PED 4.0) (Lazar et al., 2021).

The high conformational entropy and extreme conformational dynamics of IDPs, however, remain the major challenges to this overall project. Experimental data provide time- and ensemble-averaged structural information which is affected by random and possibly systematic errors. As such, the number of degrees of freedom necessary to specify an ensemble of atomic resolution structures is inherently much larger than the number of experimentally determined structural restraints. Ensemble calculation is therefore a mathematically "ill-posed" or "underdetermined" problem that always has more than one solution (Marsh and Forman-Kay, 2012; Bonomi et al., 2017; Bottaro et al., 2020; Lazar et al., 2021).

Differences in how ensembles are determined, such as how an initial ensemble is generated and which ensemble optimization algorithm is used, lead to further proliferation in the number of possible solutions for the same experimental dataset. Trivially, these ensembles are distinct as they are composed of different protein conformations. However, whether these differences are significant or not remains unclear, and it will require a quantitative comparison of their impact on inferences about sequence-structure or structure-function relationships. Understanding this variability in calculated ensembles for the same system is particularly important given the renewed efforts of PED 4.0 to curate high quality ensemble structural data (Lazar et al., 2021).

To probe the intrinsic variability of this under-determined process and evaluate its effect on sequence-structure-function relationships, we examined ensembles generated from different conformational priors and using different modelling methodologies. Broadly, prior ensembles can be generated using either: 1) MD simulations, which use physics-based force fields to generate Boltzmann-weighted ensembles; or 2) statistical coil approaches, which use extensive (un) biased sampling of the complete conformational phase space. Here, we selected two MD priors, Amber ff03ws (Best et al., 2014) (a03ws) and Amber 99SBdisp (Robustelli et al., 2018) (a99SBdisp), and a statistical coil prior generated by TraDES (Feldman and Hogue, 2002), TraDES-SC.

A03ws is a force field in which the protein-water interactions in the a03w protein forcefield were rescaled by a constant factor to produce more realistic dimensions of denatured and intrinsically disordered proteins (Best et al., 2014). A99SBdisp is a recently developed force-field intended to provide accurate descriptions of both folded and disordered proteins (Robustelli et al., 2018). In a recent benchmarking study, a03ws was shown to produce global dimensions agreeing with experiment, but at the expense of residual secondary structure propensity of IDPs or stability of folded proteins (Robustelli et al., 2018). In the same study, a99SBdisp accurately described both ordered and disordered states, including global dimensions of many IDPs. However, for larger IDPs with more hydrophobic sequences (α -synuclein, N_{TAIL}, Sic1), a99SBdisp showed a bias toward overly compact global dimensions. In contrast, TraDES generates all-atom conformations in which the only physics-based interactions are excluded-volume (Feldman and Hogue, 2002).

We have selected two popular, but contrasting modeling methodologies: the Bayesian Maximum Entropy (BME) (Bottaro et al., 2020) approach and ENSEMBLE (Krzeminski et al., 2013). Although there are many specific differences between these methodologies, the major distinction is in the treatment of the prior ensemble and of experimental and prediction errors. The BME approach produces the minimum perturbation to the prior ensemble (i.e., maximum relative Shannon entropy with respect to the prior) such that it fits the experimental data, with experimental and prediction errors accounted for in a Bayesian framework. The ENSEMBLE approach, in contrast, places no restriction on the deviation from the prior ensemble while minimizing pseudo-energy terms quantifying disagreement with experimental data. These pseudo-energy terms are typically harmonic potentials with preset scaling and target energies.

We focus here on the N-terminal 90 residues of the full-length disordered protein Sic1 (henceforth Sic1) which has been extensively characterized by NMR, SAXS, and smFRET experiments (Mittag et al., 2008; Mittag et al., 2010; Liu et al., 2014; Gomes et al., 2020) and for which we have recently determined ensembles using the ENSEMBLE method (Gomes et al., 2020). In their benchmarking study and to test their recently developed a99SBdisp forcefield, Robustelli et al. (2018) produced long-timescale (30 μs) simulations of Sic1 using a03ws and a99SBdisp. The authors have kindly provided these

TABLE 1 | Optimization parameters of BME-calculated Sic1 ensembles[a].

	χ^2_{TOTAL}	χ^2_{SAXS}	χ^2_{FRET}	χ^2_{CS}	PRE score	N_{eff}	θ	Ω
a03ws	1.00 (0.97–1.07	0.998 (0.97–1.04)	0.07 (0.02–0.22)	0.99 (0.96–1.00)	0.313 (0.314–0.313)	0.73 (0.69–0.78)	150 (75–300)	75
a99SBdisp	1.40 (1.3–1.5)	1.96 (1.7–2.2)	0.04 (0.01–0.10)	0.635 (0.63–0.64)	0.276 (0.299–0.264)	0.74 (0.65–0.80)	300 (150–500)	75
TraDES-SC	1.46 (1.3–1.6)	1.89 (1.7–2.1)	0.09 (0.03–0.21)	0.85 (0.81–0.87)	0.273 (0.29–0.26)	0.77 (0.66–0.83)	300 (150–500)	75

[a]*The values between brackets correspond to lower and upper limits of shaded regions in* ***Figure 1****.* χ^2_{TOTAL} *is calculated as* $\chi^2_{TOTAL} = \chi^2_{SAXS} + \chi^2_{CS} + \Omega\chi^2_{FRET}$ *at the chosen value of* θ *indicated in the table and with the determined optimal* $\Omega = 75$*. The PRE score is a RMSD, calculated using DEER-PREdict.*

TABLE 2 | Optimization parameters of ENSEMBLE-calculated Sic1 ensembles[a].

	ENSEMBLE energy	χ^2_{SAXS}	PRE score	χ^2_{CS}	z-test FRET	Number of trials
a03ws	335 ± 17	0.985	0.285	0.782	0.73	327 ± 102
a99SB-disp	302 ± 6	0.976	0.264	0.478	0.67	452 ± 261
TraDES-SC	272 ± 15	0.986	0.201	0.324	1.03	141 ± 33

[a]$z_E = |E_{ens} - E_{exp}|/\sigma$*. Uncertainty in ENSEMBLE energy and Number of trials is calculated as the standard deviation of the five independent ENSEMBLE optimizations.* χ^2_{SAXS}*, PRE score and* χ^2_{CS} *are calculated for the composite 500 conformer optimized ensemble. PRE scores for the final ensembles are calculated using DEER-PREdict to calculate an RMSD, as for the BME validation scores; however, during optimization the ENSEMBLE's native PRE module was used.*

simulations to be used as prior ensembles in our calculations. Importantly, Sic1 was in the test set and not in the training set for developing a99SBdisp. The extensive experimental characterization and molecular modelling of Sic1 make it an ideal case for benchmarking both future force-field developments and ensemble modelling.

2 METHODS

The SAXS and smFRET data from our group was recently published (Gomes et al., 2020) and the NMR data was published elsewhere (Mittag et al., 2008; Mittag et al., 2010). The unoptimized MD ensembles (a03ws and a99SBdisp) were generated by resampling the original simulations (Robustelli et al., 2018) with a stride of 40 frames, resulting in $\Delta t = 7.2$ ns between consecutive frames. For BME, forward calculation of the SAXS data was performed using Pepsi-SAXS (Grudinin et al., 2017) (see SI **Section 2**); chemical shift data were calculated using ShiftX; smFRET data were calculated as described previously (Gomes et al., 2020). To accommodate smFRET measurements in BME, the J-Coupling module of BME was used, since both calculations involve a weighted linear average. For ENSEMBLE, calculations were performed as described previously (Gomes et al., 2020), using either the default conformer generation (TraDES-SC) or the resampled MD simulations (a03ws and a99SBdisp) as initial pools.

PRE intensity ratios were calculated using DEER-PREdict (Tesei et al., 2021) v0.1.8 with an effective correlation time of the spin label of $\tau_C = 2$ ns, total correlation time $\tau_t = 0.5$ ns, total INEPT time $t_d = 10$ ms, reduced transverse relaxation rate $R_2 = 10$ Hz, and proton Larmor frequency $\omega_H/2\pi = 500$ MHz. The root-mean-squared error between the calculated and experimental intensity ratios was calculated for each label location (−1, 21, 38, 64, 83, and 90) and the final PRE score is the root-mean-squared average of the six RMSDs (PRE Score, **Tables 1** and **2** and **Figure 1**).

Analysis of optimized and unoptimized ensembles (radius of gyration, scaling maps, DSSP, H-Bonds) were performed using MDTraj (McGibbon et al., 2015) v1.9.5. Pi-contact analysis was performed using scripts provided by Vernon et al. (Vernon et al., 2018) Uncertainties in the secondary structure propensities, and in the average number of each type of pi-contact were determined using bootstrapping; i.e., the calculations were performed on N conformations randomly sampled (with replacement) from the initial ensemble, with either uniform weights ($w_i^0 = 1/N$) for calculations on the prior ensembles or the ENSEMBLE-optimized ensembles, or with the BME determined weights w_i for the BME-optimized ensemble. The uncertainties were calculated as the standard deviation of the parameter of interest for 1000 bootstrapped ensembles.

3 RESULTS AND DISCUSSION

3.1 Ensemble Calculation

3.1.1 Bayesian Maximum Entropy Method

The BME method is equivalent to minimizing the function $\mathcal{L}(w_1 \ldots w_N) = \frac{1}{2}\chi^2_{TOTAL}(w_1 \ldots w_N) - \theta S_{REL}(w_1 \ldots w_N)$ where the w_i are the optimized weights for each conformer in the ensemble. Here, χ^2_{TOTAL} quantifies the total agreement with all experimental data points and $S_{REL} = -\sum w_i \ln(\frac{w_i}{w_i^0})$ is the relative entropy which quantifies the deviation from the initial weights, w_i^0, in our case all equal to $\frac{1}{N}$. The hyperparameter θ balances the confidence in the prior with respect to that of the experimental data and it is determined by tuning (discussed below), given that the combined uncertainty in the experimental data, the calculated data, and the prior ensemble is not known accurately. For more details on the theory behind BME, the reader is referred to the original author's publications (Bottaro et al., 2020; Orioli et al., 2020) and equivalent or similar approaches (Leung et al., 2016; Köfinger et al., 2019).

In this work, we use SAXS, chemical shifts (CS) and smFRET data (between residues −1 and 90C, probing approximately the

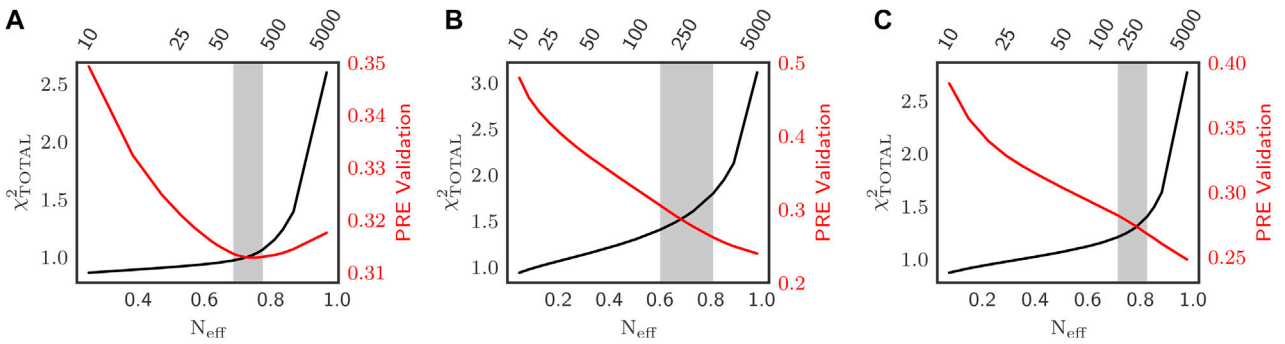

FIGURE 1 | Optimization of Sic1 conformational ensembles using BME and different initial pools: a03ws **(A)** and a99SBdisp **(B)**, and TraDES-SC **(C)**. For each value of θ (top *x*-axis), agreement with the restraining data (SAXS, CS and smFRET) is measured by χ^2_{total} (black). Agreement of the posterior ensemble with PRE data (PRE Score, red) serves as validation. For a03ws **(A)**, θ was chosen to minimize disagreement with the validation data; for a99SBdisp **(B)** and TraDES-SC **(C)**, θ was chosen at the "elbow" in the χ^2_{total} identified by L-curve analysis. In each case, a range of acceptable θ was evaluated near the optimal values (grey shaded region, **Table 1**).

end-to-end distance) as restraints, and so χ^2_{TOTAL} is the sum of the individual non-reduced χ^2 s. In principle, the relative weights of each experiment could be determined accurately if the following could be determined accurately: 1) the degrees of freedom for each experiment; 2) the statistical *and* systematic experimental uncertainties; and 3) the statistical *and* systematic uncertainties in the "forward calculation" (calculation of experimental observables from structures). In this case, χ^2_{TOTAL} would simply be the sum of each experiment's individual non-reduced χ^2, i.e., $\chi^2_{TOTAL} = \chi^2_{SAXS} + \chi^2_{CS} + \chi^2_{FRET}$. However, in practice *(i)—(iii)* are not possible to determine accurately, as discussed below. In the absence of a corrective, datatypes with many datapoints (i.e., SAXS and CSs) would overwhelm those with one or a few observations (i.e., smFRET). To compensate for the undue influence of SAXS and CS relative to FRET on the χ^2_{TOTAL} we introduced and determined a weighting factor Ω such that $\chi^2_{TOTAL} = \chi^2_{SAXS} + \chi^2_{CS} + \Omega\chi^2_{FRET}$ (SI **Section 1**, **Supplementary Figures S1, S2**). Briefly, increasing Ω from Ω = 1 to Ω ≈ 75 improves the fit to the smFRET data and a set of independent validation data (see below), without worsening the SAXS fit, and with only marginally more reweighting.

Figures 1A–C shows how θ was determined for the three prior ensembles: a03ws, a99SBdisp, and TraDES-SC, respectively. Lower values of θ result in greater agreement with experiment (lower χ^2_{TOTAL}) but with greater deviation from the prior ensemble, which is quantified by a lower effective number of conformations used in the posterior ensemble, $N_{eff} = \exp(S_{REL})$ (Bottaro et al., 2020; Orioli et al., 2020). Although better agreement with experiments could be achieved by letting $\theta \to 0$, within the BME framework this would 1) ignore uncertainties in the experimental and calculated values and 2) disregard information about molecular interactions encoded in the priors, e.g., the physics of the force fields in the MD priors. Enforcing too tight an agreement with the set of restraining data can also lead to overfitting. To verify if overfitting occurs and find the optimal value of θ, we assessed the agreement of the ensembles with six PRE experiments (ca. 400 data points), which were not used as input for reweighting. As such, we scanned through different values of θ and simultaneously monitored N_{eff}, χ^2_{TOTAL} and the PRE score calculated using DEER-PREdict (Tesei et al., 2021).

For all three priors, there is an initial region in which lower values of θ substantially decrease χ^2_{TOTAL} with only small decreases in N_{eff}, followed by a region in which small decreases in χ^2_{TOTAL} are accompanied by substantial decreases in N_{eff}. We therefore use L-curve analysis (Hansen and O'Leary, 1993; Orioli et al., 2020) to identify a useful region of θ corresponding to the "elbow" regions in the N_{eff} vs. χ^2_{TOTAL} plots. For a03ws, the agreement with PRE data initially improves (possibly because of distances shared by the smFRET restraint and PRE validation), then it worsens as θ is set beyond the elbow region. In contrast, for a99SBdisp and TraDES-SC priors, agreement with the PRE validation data monotonically worsens as θ decreases. We hypothesize that this is a consequence of enforcing the SAXS restraint on relatively compact prior ensembles, as discussed below.

Due to the r^{-6} averaging of nuclear-electron distances in PRE measurements, the ensemble averages are dominated by contributions from compact conformers (Ganguly and Chen, 2009). As such, the presence of few compact conformers can satisfy the PRE data (Ganguly and Chen, 2009), and decreasing the weight of these conformers to satisfy the SAXS restraint worsens agreement with the PRE validation. In contrast to a03ws, which is already in good agreement with the SAXS data before re-weighting, a99SBdisp and TraDES-SC are more compact, with fewer conformations that are expanded above the experimental radius of gyration, R_g^{EXP} (SI, **Supplementary Figure S4**). As a result, deriving ensembles for a99SBdisp and TraDES-SC that agree with the SAXS data involve significant re-weighting of the prior ensembles by reducing the weight of compact conformations and increasing the weight of expanded conformations.

For further analysis, we selected θ resulting in the minimum of the PRE validation score for a03ws, and θ corresponding to the "elbow" region of the χ^2_{TOTAL} vs. N_{eff} plots for a99SBdisp and TraDES-SC (**Figure 1**). **Table 1** shows the reduced χ^2_{TOTAL}, the

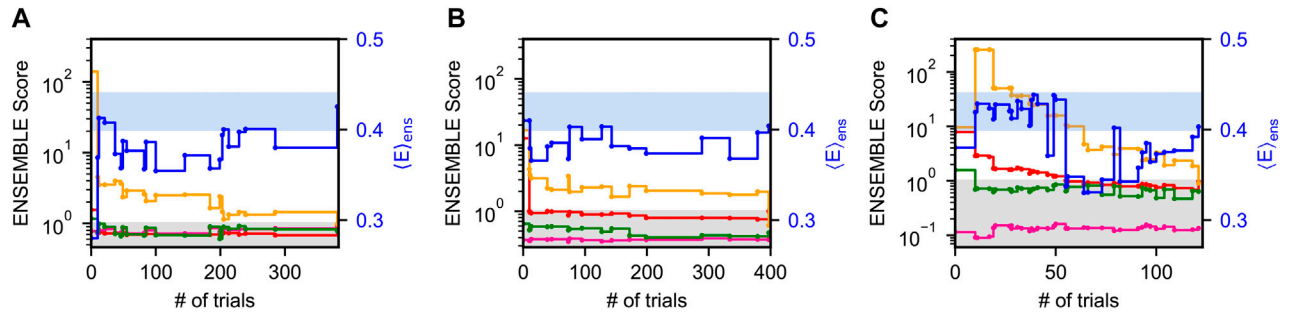

FIGURE 2 | Optimization of Sic1 conformational ensembles using ENSEMBLE and different initial pools: a03ws **(A)** and a99SBdisp **(B)**, and TraDES-SC **(C)**. Individual restraint pseudo-energies are normalized by their ENSEMBLE-defined target energy, such that a value less than one is considered "satisfied" (gray shaded region). Shown here are typical trajectories from the ENSEMBLE optimization of each prior (a03ws, a99SBdisp, TraDES-SC) using the following restraints: SAXS (red), chemical shifts (alpha—green, beta—magenta), and PRE (yellow). smFRET (blue) is used as an external validation, with the blue-shaded region showing the measured efficiency, E_{exp}, and its uncertainty.

individual restraining data reduced χ^2 s, the PRE validation score, and the N_{eff} for each prior at their corresponding optimal θ values. Although re-weighting improved the agreement of a99SBdisp and TraDES-SC ensembles with the SAXS data, it is not possible to improve it further without substantially deviating from the prior ensemble (very low N_{eff}) and incurring overfitting (e.g., poor PRE validation performance). All three posterior ensembles agree well with the smFRET data; for a03ws, this is a result of re-weighting, while prior a99SBdisp and TraDES-SC ensembles were already in reasonably good agreement. However, the end-to-end distance measured by smFRET has a high restraining strength, as in its absence the global expansion dictated by the SAXS data would lead to anomalously expanded ensembles for a99SBdisp and TraDES-SC (Fuertes et al., 2017; Gomes et al., 2020).

3.1.2 ENSEMBLE Method

The ENSEMBLE method (Krzeminski et al., 2013) minimizes a total pseudo-energy, which is the weighted sum of each individual experiment's pseudo-energy, wherein lower energies correspond to better agreement with experimental restraints. To perform this minimization, ENSEMBLE employs a switching Monte-Carlo algorithm within a simulated annealing protocol to select subsets of conformers from the initial ensemble. The optimization terminates when all experimental restraints are below their respective target energies that are set by default in ENSEMBLE (Krzeminski et al., 2013). The relative weights of different experiments are adjusted during optimization, with increased weight given to experiments that are above their target energies. We perform five independent ENSEMBLE calculations with 100 conformations and combine the results to form ensembles with 500 conformations, based on previous calculations (Marsh and Forman-Kay, 2012; Gomes et al., 2020). This ensemble size balances between the concerns of overfitting and underfitting and ideally, structural features resulting from overfitting should be averaged out in independent calculations. When applying ENSEMBLE to Sic1, we used SAXS, CS, and PRE data as experimental restraints, and reserved the smFRET data as a validation. Allocating the experimental data into restraints and validation identically for both optimization methods is not currently possible since ENSEMBLE and BME accommodate different experimental data types.

Figure 2 shows typical ENSEMBLE pseudo-energy minimizations for all three priors as a function of the number of Monte-Carlo trials. Note that because the ENSEMBLE optimization is stochastic, no two trajectories will be identical. Each pseudo-energy is normalized by its ENSEMBLE-defined target energy, such that a value less than one is considered "fit" by the program (gray shaded region). The smFRET validation is shown as a solid blue line with the right-hand axis, with the blue-shaded region corresponding to the experimental FRET efficiency E_{exp} and its uncertainty. Ensembles with a FRET value within the blue region are within one σ of the experimental value.

For a03ws (**Figure 2A**), energy minimization is largely focused on improving the agreement with the PRE data, whereas the trial ensembles agree with the CS and SAXS data either initially or after relatively few trials. For a99SBdisp (**Figure 2B**), the initial disagreement with the PRE data is less than for a03ws, though the initial disagreement with the SAXS data is greater. However, in relatively few trials the SAXS data is fit, and further energy minimization is focused on the PRE data. In contrast to a03ws and a99SBdisp, which are new MD force fields designed to accurately describe IDPs, TraDES-SC (**Figure 2C**) only accounts for excluded volume and random propensities for varying secondary structure (hence, statistical coil). Unsurprisingly, the TraDES-SC ensemble initially disagrees with most of the experimental data. Optimization first reduces the SAXS restraint below its target energy, before finally fitting the PRE data.

For all ENSEMBLE calculations, the PRE restraint was the last to be fit below its target energy, while the CS data was fit either initially or within the first few trials. This suggests that CSs are a comparatively easy experimental restraint to meet, perhaps because of the comparatively large CS calculator uncertainties. Consequently, the secondary structure propensities of the optimized ensembles will be largely dictated by the

propensities of the prior ensembles (see below). As shown in **Figure 2**, trial ensembles which fit the SAXS data but not the PRE data have overly expanded end-to-end distances resulting in $E < E_{exp}$. Jointly fitting the SAXS and PRE data places strong restraints on the end-to-end distance distribution, and consequently on E. This reinforces the conclusions drawn by Gomes et al., which were made using only the TraDES-SC prior ensemble, and emphasizes the inability of SAXS to determine specific inter-residue distances, except in the case of ideal polymer models (Gomes et al., 2020).

Table 2 shows the mean and standard deviation of the non-normalized ENSEMBLE total energy upon termination for the five independent trials. Although ENSEMBLE minimizes an ENSEMBLE-defined energy term for each experimental data type, **Table 2** shows the reduced χ^2 for SAXS and CS data (Cα and Cβ combined) to facilitate comparison of the fits with those done by BME. ENSEMBLE optimization considers PRE restraints as $\langle r^{-6} \rangle^{-\frac{1}{6}}$ distance restraints and approximates the electron location of the paramagnetic probe to the position of the Cβ atom of the spin-labelled residue (Krzeminski et al., 2013). In **Table 2**, the PRE score is the RMSD between the calculated and experimental PRE intensity ratios using the more accurate rotamer library approach, DEER-PREdict (Tesei et al., 2021), which was also used to calculate BME validation scores (see above). Additionally, the agreement with the smFRET validation data is reported using a z-test.

Interestingly, the ENSEMBLE-optimized TraDES-SC ensemble is in better agreement with the PRE and CS data than the ENSEMBLE-optimized a03ws and a99SBdisp ensembles. This may be due to the much larger conformational diversity in the TraDES-SC initial pool. When optimizing for a03ws and a99SBdisp, no new conformations are generated, and ENSEMBLE must select from the fixed initial pool of MD-generated conformers. For TraDES-SC, we used ENSEMBLE's built-in conformer generation and management (Krzeminski et al., 2013), in which new conformations are regularly replenished using TraDES. The conformer management algorithm favors conformers that have been selected fewer times in Monte-Carlo trials. Moreover, conformations in the MD prior ensembles will naturally have some degree of structural correlation as they are generated by the system's time-evolution. The increased sampling of conformational space for the ENSEMBLE optimized TraDES-SC ensemble might explain the more rapid approach to the final solution (fewer trials, see **Table 2**), and the lower PRE score and χ^2_{CS} when compared to the optimization using MD priors.

3.2 Analysis of Optimized Ensembles

3.2.1 Secondary Structure Propensity

Secondary structures of proteins are defined by specific patterns of hydrogen bonds, dihedral angles and other geometrical restraints. Based on the continuously expanding library of 3D structures in the Protein Data Bank (PDB, www.rcsb.org), various algorithms were developed to classify and predict secondary structure motifs in proteins (Reeb and Rost, 2019). Define Secondary Structure of Proteins (DSSP) annotates secondary structure elements to one of eight possible states and groups them into three classes: helical (α-, 3_{10}-, and π-helices), strand/extended (β-bridges and β-bulges) and loop/coil (turn, bend and other) (Kabsch and Sander, 1983; Touw et al., 2015). **Figure 3** shows the DSSP distributions of the three classes of secondary structure (helical, extended and coil) for 6 optimized Sic1 ensembles (2 methods and 3 priors).

The TraDES-SC ensembles stand out as almost exclusively consisting of coil structures (>90% for BME, >95% for ENSEMBLE), with essentially null fraction of extended elements, and at most 10% of helical fraction quasi-uniformly distributed throughout the sequence (**Figures 3C,F**). At the other end of the spectrum, the a03ws ensembles exhibit much larger helical propensities at the expense of the coil fraction. There are six 5–10 residue helical patches distributed throughout the sequence around serine residues, with propensities ranging from ~0.2 near the C-terminus (S87) to ~0.5 around S26 (**Figures 3A,D**). The ENSEMBLE optimization allows the experimental restraints to act on the prior more aggressively, leading to a reduction of the helical propensities by ~0.05 for each patch, although the sequence distribution is preserved. While the extended structure propensities are higher than when using the TraDES-SC prior, they do not exceed 0.05 and appear as short patches interleaved with the larger helical patches.

The ensembles calculated from the a99SBdisp prior reveal an intermediate picture between the two other cases (**Figure 3B,E**). The 6 helical patches present in the a03ws ensembles are still present here, albeit at a reduced propensity (~0.1–~0.25), with the BME method again exhibiting slightly larger values. Notably, a higher beta/extended propensity is observed at various points throughout the sequence, with the BME ensemble showing more of them and with larger values (~0.2) than the one obtained by ENSEMBLE (~0.1).

To a large extent, the differences in the DSSP maps reflect inherent differences in the structural ensembles used as priors. The impact of the optimization method on secondary is limited, with BME (by design) effecting a smaller bias of the prior than ENSEMBLE. TraDES-SC, which we used in a recent study of Sic1 (Gomes et al., 2020), is the least sophisticated prior of the three studied here, as it includes only excluded-volume interactions between chain residues. It is not surprising that imposing averaged size and chemical shift restraints on this ensemble cannot create "de novo" secondary structure. The chance of bringing patches of residues within hydrogen bond contact with peptide backbone forming specific dihedral angles is infinitesimally small, especially given the level of imprecision in the back calculators and the error margins of the experimental values.

Robustelli et al. (2018) benchmarked several MD force fields to describe the properties (size, secondary structure, etc.,) of both folded and disordered proteins, including Sic1. Among those, a03ws, which empirically optimized protein-water dispersion interactions for disordered protein (Best et al., 2014), reproduced R_g^{EXP} most accurately and exhibited relatively large helical propensities. However, in addition to experimentally observed helices, it also populated regions where helical propensity were not observed experimentally. As such, it is not surprising that the a03ws prior contains the highest

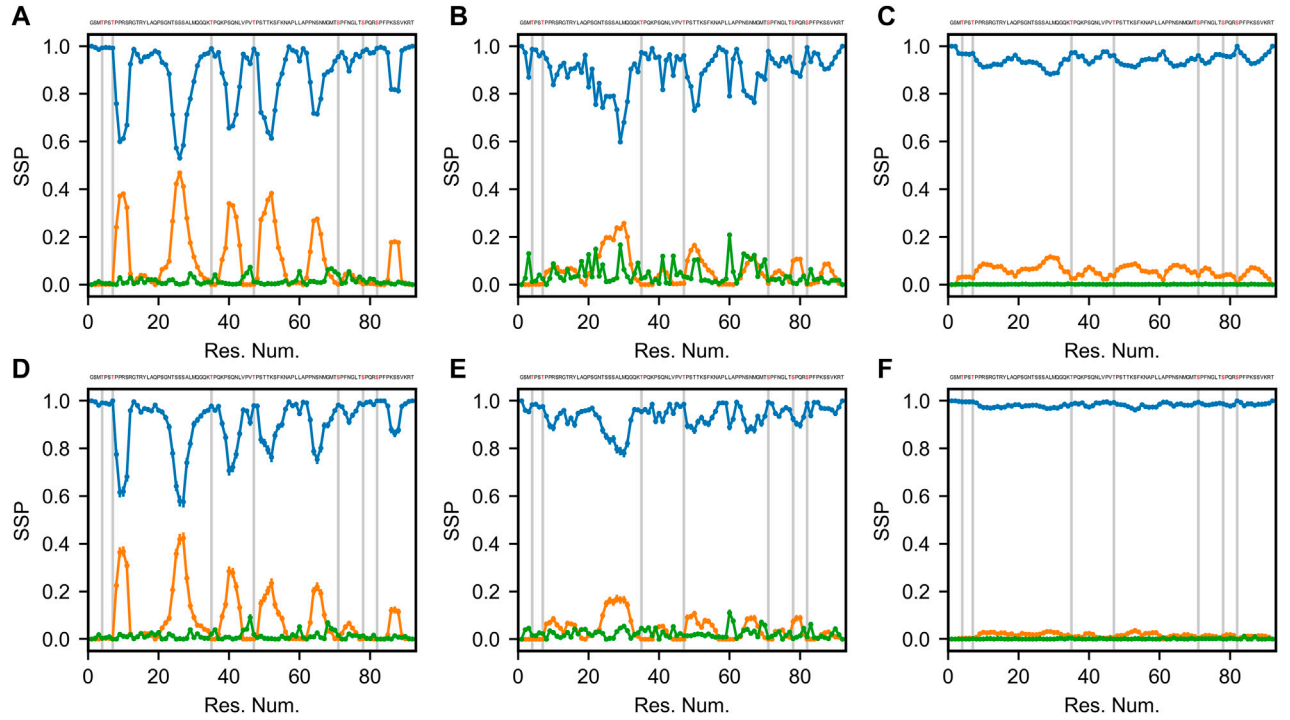

FIGURE 3 | Secondary structure propensity (SSP) of optimal Sic1 ensembles using the DSSP algorithm (Kabsch and Sander, 1983; McGibbon et al., 2015). The ensembles were calculated using BME **(A–C)** or ENSEMBLE **(D–F)**, and different initial ensembles: a03ws **(A,D)** a99SBdisp **(B,E)**, and TraDES-SC **(C,F)**. For each ensemble, secondary structure elements shown are coil (blue), helix (orange), and beta/extended (green). Error bars were calculated using bootstrapping. Phosphorylation sites (S and T residues) are shown in the Sic 1 sequence at the top of each panel in red, and as grey vertical lines in each panel.

fraction of helicity of all the priors used here. On the other hand, a99SBdisp is a force field with modified Lennard-Jones parameters (Nerenberg et al., 2012) and optimized torsion angles and van der Waals parameters, which achieved best scores in matching the experimental observables for both folded and unfolded proteins in the benchmark set (Robustelli et al., 2018). For disordered proteins (e.g., α-synuclein) a99SBdisp shows less helical propensity than a03ws, a trend that is also observed for Sic1. As mentioned above for TraDES, the impact of experimental restraints on biasing the secondary structure in ensemble calculations is minor (a slight decrease for a03ws and a slight increase for a99SBdisp). The physics model of the prior, i.e., the force field parametrization, is by far the most important factor that drives formation of stable/transient secondary structure motifs.

In the case of Sic1, it is worth comparing the DSSP maps of the optimized ensembles with the SSP scores calculated using chemical shift data (Mittag et al., 2010). Three of the six helical patches observed in the DSSP maps are also present in the SSP map (around res. #26, 50, and 65), however, in contrast to SSP, regions of extended secondary structure were not significantly populated by DSSP for any of the 6 cases examined. Notably, each of the seven phosphorylation sites (indicated by vertical grey bars in **Figure 3**) lies outside the helical patches, in the coil regions of Sic1. This may ensure access of kinase enzymes to these sites, and favor a binding model in which the multiple CPD sites in Sic1 engage the single receptor site of Cdc4 in a fast dynamic equilibrium. On the other hand, the ubiquitination sites in Sic1 (Lys 32, 36, 50, 53, 84, and 88) must bridge a 64 Å between the binding site on Cdc4 and a catalytic cysteine residue on Cdc34 within the SCF^{Cdc4} ubiquitin ligase dimer (Tang et al., 2007). These sites lie predominantly in non-helical regions (all except 53 and 84), which seems consistent with the prerequisite for Sic1 to simultaneously be docked on Cdc4 and reach the ubiquitination site on Cdc34.

3.2.2 Inter-residue Contact Maps

Inter-residue distance maps are two-dimensional representations of IDP structural propensities. Here, for each pair of residues in the Sic1 sequence, we calculated the average Cα—Cα distances in the optimized ensembles and normalized them to the respective distances in a random coil (RC) state (**Figure 4**). This type of analysis identifies regional biases for expansion (red) or compaction (blue). Alternatively, the inter-residue distances of the optimized ensemble can be normalized by the prior ensemble (see SI **Supplementary Figure S5**). Ensembles which agree well with the SAXS data (**Figure 4A,D-F**) have inter-residue distances r_{ij} that are overall more expanded than that of a RC, since $R_g^{EXP} > R_g^{RC}$ and $R_g = \sqrt{\frac{1}{2n^2}\sum_{ij}^{n}\langle r_{ij}^2\rangle}$, where n is the number of residues. Conversely, the BME-optimized a99SBdisp and TraDES-SC ensembles (**Figure 4B,C**) are overall more

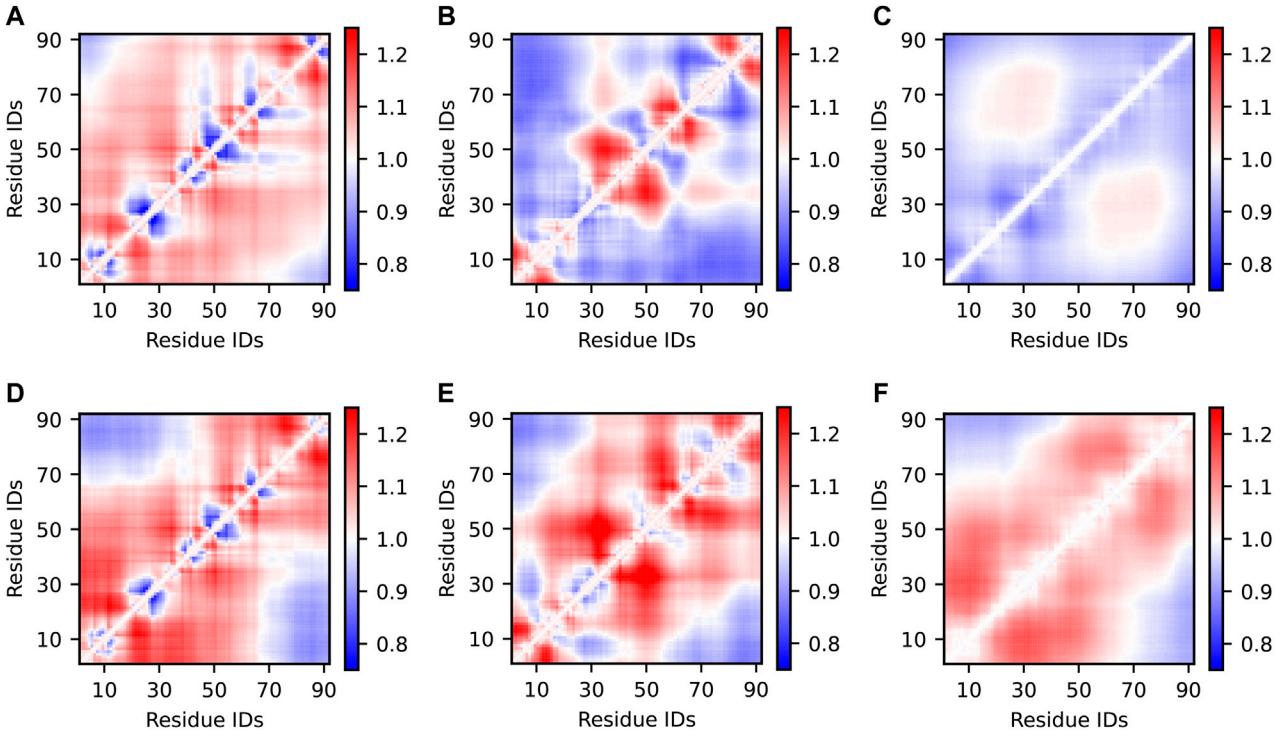

FIGURE 4 | Inter-residue scaling maps of optimized Sic1 ensembles relative to a reference random-coil Sic1 ensemble (TraDES-RC). The ensembles were calculated using BME **(A–C)** or ENSEMBLE **(D–F)**, and different initial ensembles: a03ws **(A,D)** a99SBdisp **(B,E)**, and TraDES-SC **(C,F)**. Ensemble-averaged distances between the Cα atoms of every unique pair of residues are normalized by the respective distances in a random coil. Regions in red are expanded relative to a random coil, while those in blue are more compact.

compact. Compact regions near the diagonal indicate a propensity for secondary structure (see also **Figure 3**). Since all ensembles agree with the FRET data between residues -1 and 90C, this region is similarly compact across all ensembles.

Notably, ENSEMBLE-optimization with different priors leads to different patterns of intermediate- and long-range distances, despite identical experimental restraints which included PRE measurements from six sites throughout the Sic1 sequence. This suggests, as Naudi-Fabra et al. have recently demonstrated (Naudi-Fabra et al., 2021), that multiple FRET and PRE measurements, which sufficiently sample the linear sequence of the protein, are needed to accurately reproduce intermediate- and long-range distances. Incorporating additional FRET restraints is expected to make ensembles optimized from different priors more similar in this respect. Indeed, certain distances (e.g., approximately 1,40) appear to be less restrained (vary more between optimized ensembles) than other distances (e.g., approximately 40,70), suggesting these pairs for future FRET labelling locations.

3.2.3 Molecular Interactions

Determining which specific molecular interactions determine the observed structural properties of IDP ensembles is an important goal. Knowledge of these interactions connect sequence properties to structural properties, allow testable predictions for the effects of mutations, and aid the rational design of molecules that bind disordered protein sequences with high affinity and specificity, stabilizing distinct IDP conformations (Ambadipudi and Zweckstetter, 2016; Robustelli et al., 2021). However, the experimental data, which are spatially and temporally averaged and are affected by noise, are insufficient to restrain distances and angles between groups of atoms, such that specific molecular interactions in conformations could be identified (e.g., hydrogen bond).

Including information from a force field which describes bonded and nonbonded interactions between the atoms, partially removes the degeneracy of the problem. The BME approach, which produces the minimum perturbation to the prior ensemble so that it fits the experimental data, is expected to retain the maximum amount of this information possible. Conversely, in the ENSEMBLE and similar approaches, which do not explicitly consider deviation from the prior ensemble, it is unclear in what capacity information about specific molecular interactions is retained. We therefore sought to compare the specific interactions in the resulting optimized ensembles. It is important to note that in our use of ENSEMBLE, PRE data was used as a restraint, whereas in our use of BME, PRE data was used as validation. This is expected to affect the inferred patterns of molecular interactions, in addition to the differences between optimization methods.

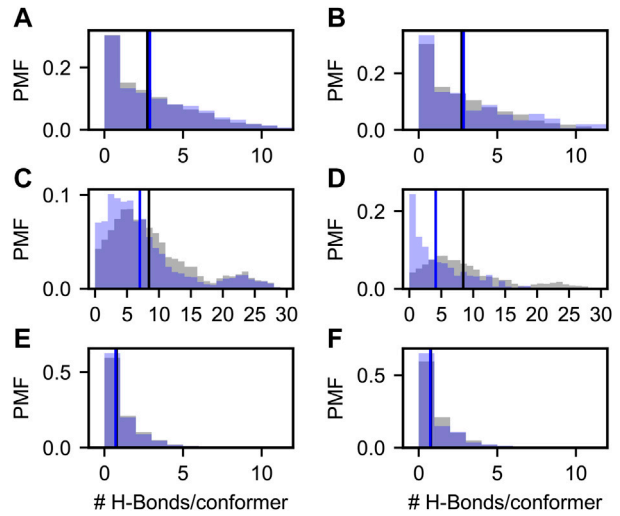

FIGURE 5 | Probability mass functions (PMFs) of hydrogen bonds per conformer in prior (gray) and posterior (blue) Sic1 ensembles. The ensembles were calculated using BME **(A,C,E)** or ENSEMBLE **(B,D,F)**, and different priors: a03ws **(A,B)** a99SBdisp **(C,D)**, and TraDES-SC **(E,F)**. Note that the *x*-axis for the a99SBdisp prior is larger than that of the a03ws and TraDES-SC priors.

Excluding hydrophobic contacts, which are relatively non-specific, we hypothesized that the most likely interactions were hydrogen bonds (**Figure 5**) and pi-contacts (**Figure 6**). The Sic1 sequence has a high fraction of polar and charged residues (~54%) that can participate in hydrogen bonding. Sic1 also has a high fraction of residues with sidechain pi bonds (~23%) and small residues with relatively exposed backbone peptide bonds (~52%) (Vernon et al., 2018). Prior to phosphorylation, Sic1 does not have any negatively charged residues, thus excluding salt-bridges and electrostatic attraction. To distinguish short-range/secondary structure (**Figure 3**) from long-range tertiary contacts, we examine only those interactions with a sequence separation $|i - j| > 10$.

3.2.3.1 Hydrogen Bonds

Figure 5 shows the probability mass function (PMF) of the number of hydrogen bonds (H-bonds) per conformer in the prior (grey) and in the BME and ENSEMBLE optimized (blue) ensembles. H-bond contacts were defined using the distance and angle criteria established previously (Baker and Hubbard, 1984) and implemented in MDTraj (McGibbon et al., 2015). Vertical lines show the first moments of the corresponding PMFs. The TraDES-SC prior, for which there is no force-field describing non-bonded interactions, has very few H-bonds (i.e., average number of H-bonds per conformer $n_{H-bonds} = 0.7$), corresponding to the small but finite probability of meeting the H-bond criteria by chance. As such, both BME and ENSEMBLE optimization have a very small effect on the H-bond propensity (**Figure 5E,F**).

As expected, there are significantly more H-bond contacts in the a03ws prior compared to the statistical noise in TraDES-SC (**Figure 5A,B**). Optimization using either BME or ENSEMBLE slightly increases the average number of H-bonds per conformer. This is consistent with the slight decrease in R_g and root-mean-squared end-to-end distance R_{ee}, as the individual conformer radii of gyration r_g are inversely correlated with the number of H-bonds/conformer (see SI **Supplementary Figure S6**).

The a99SBdisp prior has an even higher average number of H-bonds/conformer than the a03ws prior, and the PMF is bimodal (**Figure 5C,D**). The differences in hydrogen bonding between a03ws and a99SBdisp may reflect parameterization choices in a99SBdisp to maintain accuracy for folded proteins (Robustelli et al., 2018). Alternatively, this may reflect incomplete sampling of extended structures, as simulations of Sic1 using enhanced-sampling techniques and the a99SBdisp force field produced R_g similar to that of a03ws and experiment (Shrestha et al., 2021). Interestingly, while a03ws has higher helical propensities than a99SBdisp and helical stability is largely driven by hydrogen bonding, it shows lower propensity for long-range H-bond interactions.

Whereas for a03ws both optimization methods result in qualitatively similar H-Bond PMFs, for a99SBdisp they differ considerably. Both optimizations reduce the average number of H-bonds per conformer; however, ENSEMBLE optimization removes the highly H-bonded subpopulation, and the resulting PMF is monotonically decreasing and similar to that of a03ws. Conversely, BME optimization retains this minor subpopulation, and shifts the center of the major subpopulation.

One reason for discrepant ENSEMBLE and BME H-bond inferences is how they achieve agreement with the SAXS data. The subpopulation of highly H-bonded conformations has a very compact radius of gyration (~2 nm, see SI **Supplementary Figure S6**) compared to the experimental radius of gyration (~3 nm). ENSEMBLE optimization prioritizes agreement with experimental data by eliminating the compact and highly H-bonded subpopulation. BME optimization seeks a balance between agreement with experiments and deviation from the prior, retaining this subpopulation at the expense of SAXS agreement, but smaller deviation from the a99SBdisp prior.

Experimental data is known to make ensembles more similar to one another (Tiberti et al., 2015; Larsen et al., 2020; Ahmed et al., 2021). Our results show that ensembles that agree with experimental data and were generated from an MD prior (a03ws-BME, a03ws-ENSEMBLE, a99SBdisp-ENSEMBLE) have similar H-bond PMFs. They are monotonically decreasing and have an average number of H-bonds per conformer $n_{H-bonds}$, between 3 and 4. However, our experimental data alone is insufficient to define H-bonds, as shown by the ENSEMBLE posterior ensembles (**Figure 5B,D,F**). Overall, this analysis suggests that the specific tertiary contacts and the nature of their molecular interactions in ensembles should be interpreted with caution. The type and amount of experimental data used here is insufficient, however incorporation of information about non-bonded interactions from MD force fields removes at least some of this degeneracy.

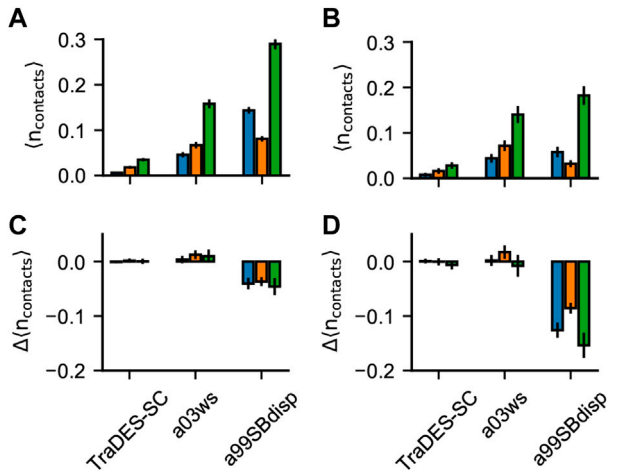

FIGURE 6 | Pi-interactions in optimized Sic1 ensembles. Average number of pi-contacts per conformer using BME **(A)** and ENSEMBLE **(B)** and different priors. Pi contacts are separated into sidechain—sidechain (sc-sc, blue), back-bone—back-bone (bb-bb, orange) and sidechain—backbone (sc-bb, green). The change in the average number of pi-contacts per conformer induced by ensemble optimization via BME **(C)** and ENSEMBLE **(D)**. Error bars are estimated by bootstrapping.

3.2.3.2 Pi Interactions

Although fixed charge atomistic MD force-fields do not explicitly include polarization and quantum effects to describe pi-interactions, they are valuable for understanding the relative importance of pi-interactions *vs.* other modes of interactions in stabilizing liquid-liquid phase separation in IDPs (Murthy et al., 2019; Schuster et al., 2020; Zheng et al., 2020). In folded protein structures, Vernon et al. (2018) found that the frequency of planar pi contacts strongly correlates with the quantity and quality of the experimental data and with the quality of the fit of the structure to the data. This suggests that current force fields may underestimate the relative importance of pi-pi interactions, and thus they appear more frequently when structures are more experimentally constrained. We therefore sought to determine: 1) whether the experimental data on Sic1 would refine the average number of planar pi-pi contacts per conformer in the ensembles and 2) whether BME and ENSEMBLE optimization would result in different pi-pi contact frequencies.

Figure 6 shows the average number of planar pi-pi contacts per conformer, n_{pi-pi},, in the optimized ensembles and the differences upon optimization. Pi contacts were defined using the distance and angle criteria presented by Vernon et al. (2018) and calculated using custom Python scripts provided by the authors. These contacts are classified as interactions between backbone amide groups (bb-bb), side chain amide, carboxyl, guanidinium or aromatic groups (sc-sc), or between backbone and side chain (bb-sc).

The TraDES-SC ensembles show the average number of each type of pi-contacts formed by chance, since there is no force field describing non-bonded interactions. Like for H-bonds, optimization of prior TraDES ensembles using either BME or ENSEMBLE did not change the frequency of pi contacts. A03ws exhibits higher n_{pi-pi} than TraDES-SC, suggesting that this force-field can somehow reproduce the pi-interaction geometries even without explicitly including polarization and quantum effects. Both BME and ENSEMBLE optimization did not significantly alter n_{pi-pi}. Like for H-bonds, the correlation between compactness, especially r_g, with the number of contacts is the major driver for changes in n_{pi-pi}, upon optimization. In contrast to H-bonds, there are few conformations with more than one pi-interaction, and so the effect of optimization is more attenuated. Moreover, planar pi-pi interactions often involve groups with H-bond donors and acceptors, presenting an additional degree of degeneracy since the current experimental data do not directly report on pi-interactions.

Like for H-bonds, a99SBdisp has a higher average number of all types of pi-contacts than does a03ws and conformations with more than one pi-interaction are more compact (e.g., $r_g < R_g^{prior} < R_g^{EXP}$). Consequently, both optimization methods significantly reduce the average number of pi-contacts. However, ENSEMBLE optimization, which results in better fits to the SAXS data, reduces n_{pi-pi} more than BME optimization, which balances agreement with the SAXS data with deviation from the prior.

Intuitively, experimental data alone is insufficient to meaningfully describe pi-interactions in the absence of a force field (e.g., the TraDES-SC optimized ensembles). When the prior is constructed using a force field that describes the interaction geometries/strengths and the optimized ensembles agree with experimental data (e.g., BME-a03ws, ENSEMBLE-a03ws, and ENSEMBLE-a99SBdisp) the resulting ensembles have similar average numbers of pi-contacts (see also **Supplementary Figure S7**). As previously mentioned, experimental data makes the ensembles more similar, only when there exist interactions which can be reweighted, and they are correlated with experimental data.

CONCLUSION

Conformational ensembles for the disordered Sic1 protein were obtained by using experimental data (SAXS, CS, PRE, and smFRET) as restraints and validation on three prior ensembles that were generated using either MD force fields or a statistical coil approach. The ensembles were optimized for agreement with the experiment using two contrasting modeling methodologies, Bayesian Maximum Entropy (Bottaro et al., 2020) (BME) and ENSEMBLE (Krzeminski et al., 2013). We compared the six different outcomes by examining global dimensions (e.g., R_g), secondary structure propensities, inter-residue distances and specific non-local interactions, i.e., H-bonds and pi-interactions. Overall, differences between ensembles optimized using different priors were greater than when using the same prior with different optimization methods. Differences between methods were greatest when the priors were in poor agreement with experimental data, as BME balances perturbation of the prior ensemble with experimental agreement, whereas ENSEMBLE only focuses on the latter.

An advantage of MD priors is that they contain explicit information about specific molecular interactions (e.g., H-bonds and pi-interactions) that can be modulated, though not uniquely

determined by experimental data. However, a disadvantage of MD priors is that they, by design and/or due to computational limitations, only sample a limited region of the entire conformational landscape. If incorrectly biased (e.g., overly compact) this will result in more significant reweighting and experimental data may be insufficient to debias the ensemble. Future work would benefit from priors which are in better agreement with more than one type of experimental data prior to optimization.

Noting that ensembles optimized from different priors make different predictions regarding secondary structure, intermediate- and long-range distances, it appears that additional experimental data is needed, either as restraints or post-hoc validation. For secondary structure, this could include RDC data, which has been published for Sic1 but was not used in this analysis (Mittag et al., 2008; Mittag et al., 2010), and fluorescence anisotropy decay, which reports on segmental dynamics of IDPs (Milles and Lemke, 2014). For intermediate- and long-range contacts, the $C\alpha$—$C\alpha$ distance maps can be used to design maximally informative FRET label locations. Lastly, development of more rigorous ensemble optimization tools that integrate complementary biophysical data on multiple scales will lead to more accurate descriptions of conformational ensembles of IDPs and enable a mechanistic understanding of their biological function in implication in pathologies.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding authors.

AUTHOR CONTRIBUTIONS

G-NG and CG designed and coordinated the research and wrote the manuscript. AN and G-NG performed data analysis and made figures and tables for the manuscript. All authors have given approval to the final version of the manuscript.

FUNDING

This work has been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC RGPIN 2017–06030 to CG).

ACKNOWLEDGMENTS

The authors are grateful to DR. J.-D. Forman-Kay from Sick Kids Hospital for providing the NMR data of Sic1 used in this study and to DR. M. Krezminski from her lab for his support in using the ENSEMBLE program. We thank DR. T. Mittag and DR. E. Martin from St. Jude Hospital for providing SAXS data of Sic1. The authors are grateful to DR. D.E. Shaw for providing molecular dynamics simulation data of Sic1. We also thank DR. K. Lindorff-Larsen and DR. S. Bottaro for technical help with using the BME program.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmolb.2022.910956/full#supplementary-material

REFERENCES

Ahmed, M. C., Skaanning, L. K., Jussupow, A., Newcombe, E. A., Kragelund, B. B., Camilloni, C., et al. (2021). Refinement of α-Synuclein Ensembles against SAXS Data: Comparison of Force Fields and Methods. *Front. Mol. Biosci.* 8, 216. doi:10.3389/fmolb.2021.654333

Alderson, T. R., and Kay, L. E. (2021). NMR Spectroscopy Captures the Essential Role of Dynamics in Regulating Biomolecular Function. *Cell* 184, 577–595. doi:10.1016/j.cell.2020.12.034

Ambadipudi, S., and Zweckstetter, M. (2016). Targeting Intrinsically Disordered Proteins in Rational Drug Discovery. *Expert Opin. Drug Discov.* 11, 65–77. doi:10.1517/17460441.2016.1107041

Aznauryan, M., Delgado, L., Soranno, A., Nettels, D., Huang, J.-r., Labhardt, A. M., et al. (2016). 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.* 113, E5389–E5398. doi:10.1073/pnas.1607193113

Baker, E. N., and Hubbard, R. E. (1984). Hydrogen Bonding in Globular Proteins. *Prog. Biophys. Mol. Biol.* 44, 97–179. doi:10.1016/0079-6107(84)90007-5

Best, R. B., Zheng, W., and Mittal, J. (2014). Balanced Protein-Water Interactions Improve Properties of Disordered Proteins and Non-Specific Protein Association. *J. Chem. Theory Comput.* 10, 5113–5124. doi:10.1021/ct500569b

Bonomi, M., Heller, G. T., Camilloni, C., and Vendruscolo, M. (2017). Principles of Protein Structural Ensemble Determination. *Curr. Opin. Struct. Biol.* 42, 106–116. doi:10.1016/j.sbi.2016.12.004

Bottaro, S., Bengtsen, T., and Lindorff-Larsen, K. (2020). "Integrating Molecular Simulation and Experimental Data: A Bayesian/Maximum Entropy Reweighting Approach," in *Structural Bioinformatics: Methods and Protocols, Methods in Molecular Biology*. Editor Z. Gáspári (New York, NY: Springer US), 219–240. doi:10.1007/978-1-0716-0270-6_15

Crehuet, R., Buigues, P. J., Salvatella, X., and Lindorff-Larsen, K. (2019). Bayesian-Maximum-Entropy Reweighting of IDP Ensembles Based on NMR Chemical Shifts. *Entropy* 21, 898. doi:10.3390/e21090898

Delaforge, E., Milles, S., Bouvignies, G., Bouvier, D., Boivin, S., Salvi, N., et al. (2015). Large-Scale Conformational Dynamics Control H5N1 Influenza Polymerase PB2 Binding to Importin α. *J. Am. Chem. Soc.* 137, 15122–15134. doi:10.1021/jacs.5b07765

Dimura, M., Peulen, T.-O., Sanabria, H., Rodnin, D., Hemmen, K., Hanke, C. A., et al. (2020). Automated and Optimally FRET-Assisted Structural Modeling. *Nat. Commun.* 11, 5394. doi:10.1038/s41467-020-19023-1

Dyson, H. J., and Wright, P. E. (2005). Intrinsically Unstructured Proteins and Their Functions. *Nat. Rev. Mol. Cell Biol.* 6, 197–208. doi:10.1038/nrm1589

Dyson, H. J., and Wright, P. E. (2021). NMR Illuminates Intrinsic Disorder. *Curr. Opin. Struct. Biol.* 70, 44–52. doi:10.1016/j.sbi.2021.03.015

Dyson, H. J., and Wright, P. E. (2019). Perspective: the Essential Role of NMR in the Discovery and Characterization of Intrinsically Disordered Proteins. *J. Biomol. NMR* 73, 651–659. doi:10.1007/s10858-019-00280-2

Feldman, H. J., and Hogue, C. W. V. (2002). Probabilistic Sampling of Protein Conformations: New Hope for Brute Force? *Proteins* 46, 8–23. doi:10.1002/prot.1163

Forman-Kay, J. D., and Mittag, T. (2013). From Sequence and Forces to Structure, Function, and Evolution of Intrinsically Disordered Proteins. *Structure* 21, 1492–1499. doi:10.1016/j.str.2013.08.001

Fuertes, G., Banterle, N., Ruff, K. M., Chowdhury, A., Mercadante, D., Koehler, C., et al. (2017). Decoupling of Size and Shape Fluctuations in Heteropolymeric Sequences Reconciles Discrepancies in SAXS vs. FRET Measurements. *Proc. Natl. Acad. Sci. U. S. A.* 114, E6342–E6351. doi:10.1073/pnas.1704692114

Ganguly, D., and Chen, J. (2009). Structural Interpretation of Paramagnetic Relaxation Enhancement-Derived Distances for Disordered Protein States. *J. Mol. Biol.* 390, 467–477. doi:10.1016/j.jmb.2009.05.019

Gomes, G.-N. W., Krzeminski, M., Namini, A., Martin, E. W., Mittag, T., Head-Gordon, T., et al. (2020). Conformational Ensembles of an Intrinsically Disordered Protein Consistent with NMR, SAXS, and Single-Molecule FRET. *J. Am. Chem. Soc.* 142, 15697–15710. doi:10.1021/jacs.0c02088

Grudinin, S., Garkavenko, M., and Kazennov, A. (2017). Pepsi-SAXS: An Adaptive Method for Rapid and Accurate Computation of Small-Angle X-Ray Scattering Profiles. *Acta Cryst. Sect. D. Struct. Biol.* 73, 449–464. doi:10.1107/S2059798317005745

Hansen, P. C., and O'Leary, D. P. (1993). The Use of the L-Curve in the Regularization of Discrete Ill-Posed Problems. *SIAM J. Sci. Comput.* 14, 1487–1503. doi:10.1137/0914086

Hellenkamp, B., Schmid, S., Doroshenko, O., Opanasyuk, O., Kühnemuth, R., Rezaei Adariani, S., et al. (2018). Precision and Accuracy of Single-Molecule FRET Measurements-A Multi-Laboratory Benchmark Study. *Nat. Methods* 15, 669–676. doi:10.1038/s41592-018-0085-0

Henriques, J., Arleth, L., Lindorff-Larsen, K., and Skepö, M. (2018). On the Calculation of SAXS Profiles of Folded and Intrinsically Disordered Proteins from Computer Simulations. *J. Mol. Biol.* 430, 2521–2539. doi:10.1016/j.jmb.2018.03.002

Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B. L., et al. (2017). CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Methods* 14, 71–73. doi:10.1038/nmeth.4067

Jensen, M. R., Zweckstetter, M., Huang, J.-r., and Blackledge, M. (2014). Exploring Free-Energy Landscapes of Intrinsically Disordered Proteins at Atomic Resolution Using NMR Spectroscopy. *Chem. Rev.* 114, 6632–6660. doi:10.1021/cr400688u

Kabsch, W., and Sander, C. (1983). Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* 22, 2577–2637. doi:10.1002/bip.360221211

Kalinin, S., Peulen, T., Sindbert, S., Rothwell, P. J., Berger, S., Restle, T., et al. (2012). A Toolkit and Benchmark Study for FRET-Restrained High-Precision Structural Modeling. *Nat. Methods* 9, 1218–1225. doi:10.1038/nmeth.2222

Köfinger, J., Stelzl, L. S., Reuter, K., Allande, C., Reichel, K., and Hummer, G. (2019). Efficient Ensemble Refinement by Reweighting. *J. Chem. Theory Comput.* 15, 3390–3401. doi:10.1021/acs.jctc.8b01231

Krzeminski, M., Marsh, J. A., Neale, C., Choy, W.-Y., and Forman-Kay, J. D. (2013). Characterization of Disordered Proteins with ENSEMBLE. *Bioinformatics* 29, 398–399. doi:10.1093/bioinformatics/bts701

Larsen, A. H., Wang, Y., Bottaro, S., Grudinin, S., Arleth, L., and Lindorff-Larsen, K. (2020). Combining Molecular Dynamics Simulations with Small-Angle X-Ray and Neutron Scattering Data to Study Multi-Domain Proteins in Solution. *PLoS Comput. Biol.* 16, e1007870. doi:10.1371/journal.pcbi.1007870

Lazar, T., Martínez-Pérez, E., Quaglia, F., Hatos, A., Chemes, L. B., Iserte, J. A., et al. (2021). PED in 2021: A Major Update of the Protein Ensemble Database for Intrinsically Disordered Proteins. *Nucleic Acids Res.* 49, D404–D411. doi:10.1093/nar/gkaa1021

Lerner, E., Barth, A., Hendrix, J., Ambrose, B., Birkedal, V., Blanchard, S. C., et al. (2021). FRET-Based Dynamic Structural Biology: Challenges, Perspectives and an Appeal for Open-Science Practices. *eLife* 10, e60416. doi:10.7554/eLife.60416

Leung, H. T. A., Bignucolo, O., Aregger, R., Dames, S. A., Mazur, A., Bernèche, S., et al. (2016). A Rigorous and Efficient Method to Reweight Very Large Conformational Ensembles Using Average Experimental Data and to Determine Their Relative Information Content. *J. Chem. Theory Comput.* 12, 383–394. doi:10.1021/acs.jctc.5b00759

Lincoff, J., Haghighatlari, M., Krzeminski, M., Teixeira, J. M. C., Gomes, G.-N. W., Gradinaru, C. C., et al. (2020). Extended Experimental Inferential Structure Determination Method in Determining the Structural Ensembles of Disordered Protein States. *Commun. Chem.* 3, 1–12. doi:10.1038/s42004-020-0323-0

Liu, B., Chia, D., Csizmok, V., Farber, P., Forman-Kay, J. D., and Gradinaru, C. C. (2014). The Effect of Intrachain Electrostatic Repulsion on Conformational Disorder and Dynamics of the Sic1 Protein. *J. Phys. Chem. B* 118, 4088–4097. doi:10.1021/jp500776v

Marsh, J. A., and Forman-Kay, J. D. (2012). Ensemble Modeling of Protein Disordered States: Experimental Restraint Contributions and Validation. *Proteins* 80, 556–572. doi:10.1002/prot.23220

Martin, E. W., Hopkins, J. B., and Mittag, T. (2020). Small Angle X-Ray Scattering Experiments of Monodisperse Samples Close to the Solubility Limit. arXiv: 2003.01278 [q-bio].

McGibbon, R. T., Beauchamp, K. A., Harrigan, M. P., Klein, C., Swails, J. M., Hernández, C. X., et al. (2015). MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical J.* 109, 1528–1532. doi:10.1016/j.bpj.2015.08.015

Milles, S., and Lemke, E. A. (2014). Mapping Multivalency and Differential Affinities within Large Intrinsically Disordered Protein Complexes with Segmental Motion Analysis. *Angew. Chem. Int. Ed.* 53, 7364–7367. doi:10.1002/anie.201403694

Mittag, T., Marsh, J., Grishaev, A., Orlicky, S., Lin, H., Sicheri, F., et al. (2010). Structure/Function Implications in a Dynamic Complex of the Intrinsically Disordered Sic1 with the Cdc4 Subunit of an SCF Ubiquitin Ligase. *Structure* 18, 494–506. doi:10.1016/j.str.2010.01.020

Mittag, T., Orlicky, S., Choy, W.-Y., Tang, X., Lin, H., Sicheri, F., et al. (2008). Dynamic Equilibrium Engagement of a Polyvalent Ligand with a Single-Site Receptor. *Proc. Natl. Acad. Sci. U. S. A.* 105, 17772–17777. doi:10.1073/pnas.0809222105

Murthy, A. C., Dignon, G. L., Kan, Y., Zerze, G. H., Parekh, S. H., Mittal, J., et al. (2019). Molecular Interactions Underlying Liquid–liquid Phase Separation of the FUS Low-Complexity Domain. *Nat. Struct. Mol. Biol.* 26, 637–648. doi:10.1038/s41594-019-0250-x

Naudi-Fabra, S., Tengo, M., Jensen, M. R., Blackledge, M., and Milles, S. (2021). Quantitative Description of Intrinsically Disordered Proteins Using Single-Molecule FRET, NMR, and SAXS. *J. Am. Chem. Soc.* 143, 20109–20121. doi:10.1021/jacs.1c06264

Nerenberg, P. S., Jo, B., So, C., Tripathy, A., and Head-Gordon, T. (2012). Optimizing Solute-Water van der Waals Interactions To Reproduce Solvation Free Energies. *J. Phys. Chem. B* 116, 4524–4534. doi:10.1021/jp2118373

Oldfield, C. J., and Dunker, A. K. (2014). Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. *Annu. Rev. Biochem.* 83, 553–584. doi:10.1146/annurev-biochem-072711-164947

Orioli, S., Larsen, A. H., Bottaro, S., and Lindorff-Larsen, K. (2020). "Chapter Three - How to Learn from Inconsistencies: Integrating Molecular Simulations with Experimental Data," in *Progress in Molecular Biology and Translational Science, Computational Approaches for Understanding Dynamical Systems: Protein Folding and Assembly*. Editors B. Strodel and B. Barz (Academic Press), 123–176. doi:10.1016/bs.pmbts.2019.12.006

Pesce, F., and Lindorff-Larsen, K. (2021). Refining Conformational Ensembles of Flexible Proteins against Small-Angle X-Ray Scattering Data. Preprint. doi:10.1101/2021.05.29.446281

Rauscher, S., Gapsys, V., Gajda, M. J., Zweckstetter, M., de Groot, B. L., and Grubmüller, H. (2015). Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment. *J. Chem. Theory Comput.* 11, 5513–5524. doi:10.1021/acs.jctc.5b00736

Reeb, J., and Rost, B. (2019). "Secondary Structure Prediction," in *Encyclopedia of Bioinformatics and Computational Biology*. Editors S. Ranganathan, M. Gribskov, K. Nakai, and C. Schönbach (Oxford: Academic Press), 488–496. doi:10.1016/B978-0-12-809633-8.20267-7

Robustelli, P., Ibanez-de-Opakua, A., Campbell-Bezat, C., Giordanetto, F., Becker, S., Zweckstetter, M., et al. (2021). Molecular Basis of Small-Molecule Binding to α-Synuclein. Preprint. doi:10.1101/2021.01.22.426549

Robustelli, P., Piana, S., and Shaw, D. E. (2018). Developing a Molecular Dynamics Force Field for Both Folded and Disordered Protein States. *Proc. Natl. Acad. Sci. U. S. A.* 115, E4758–E4766. doi:10.1073/pnas.1800690115

Schuster, B. S., Dignon, G. L., Tang, W. S., Kelley, F. M., Ranganath, A. K., Jahnke, C. N., et al. (2020). Identifying Sequence Perturbations to an Intrinsically Disordered Protein that Determine its Phase-Separation Behavior. *Proc. Natl. Acad. Sci. U. S. A.* 117, 11421–11431. doi:10.1073/pnas.2000223117

Shrestha, U. R., Smith, J. C., and Petridis, L. (2021). Full Structural Ensembles of Intrinsically Disordered Proteins from Unbiased Molecular Dynamics Simulations. *Commun. Biol.* 4, 1–8. doi:10.1038/s42003-021-01759-1

Tang, X., Orlicky, S., Lin, Z., Willems, A., Neculai, D., Ceccarelli, D., et al. (2007). Suprafacial Orientation of the SCFCdc4 Dimer Accommodates Multiple Geometries for Substrate Ubiquitination. *Cell* 129, 1165–1176. doi:10.1016/j.cell.2007.04.042

Tesei, G., Martins, J. M., Kunze, M. B. A., Wang, Y., Crehuet, R., and Lindorff-Larsen, K. (2021). DEER-PREdict: Software for Efficient Calculation of Spin-Labeling EPR and NMR Data from Conformational Ensembles. *PLoS Comput. Biol.* 17, e1008551. doi:10.1371/journal.pcbi.1008551

Tiberti, M., Papaleo, E., Bengtsen, T., Boomsma, W., and Lindorff-Larsen, K. (2015). ENCORE: Software for Quantitative Ensemble Comparison. *PLoS Comput. Biol.* 11, e1004415. doi:10.1371/journal.pcbi.1004415

Touw, W. G., Baakman, C., Black, J., te Beek, T. A. H., Krieger, E., Joosten, R. P., et al. (2015). A Series of PDB-Related Databanks for Everyday Needs. *Nucleic Acids Res.* 43, D364–D368. doi:10.1093/nar/gku1028

Uversky, V. N. (2015). Intrinsically Disordered Proteins and Their (Disordered) Proteomes in Neurodegenerative Disorders. *Front. Aging Neurosci.* 7, 18. doi:10.3389/fnagi.2015.00018

Vernon, R. M., Chong, P. A., Tsang, B., Kim, T. H., Bah, A., Farber, P., et al. (2018). Pi-Pi Contacts are an Overlooked Protein Feature Relevant to Phase Separation. *eLife* 7, e31486. doi:10.7554/eLife.31486

Voithenberg, L. V. V., Sánchez-Rico, C., Kang, H.-S., Madl, T., Zanier, K., Barth, A., et al. (2016). Recognition of the 3′ Splice Site RNA by the U2AF Heterodimer Involves a Dynamic Population Shift. *Proc. Natl. Acad. Sci. U. S. A.* 113, E7169–E7175. doi:10.1073/pnas.1605873113

Zheng, W., Dignon, G. L., Jovic, N., Xu, X., Regy, R. M., Fawzi, N. L., et al. (2020). Molecular Details of Protein Condensates Probed by Microsecond Long Atomistic Simulations. *J. Phys. Chem. B* 124, 11671–11679. doi:10.1021/acs.jpcb.0c10489

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.