Article

Bayesian-Maximum-Entropy reweighting of IDP ensembles based on NMR chemical shifts

Ramon Crehuet [1,2,*], Pedro J. Buigues [1,3], Xavier Salvatella [3,4] and Kresten Lindorff-Larsen [2,*]

[1] Institute for Advanced Chemistry of Catalonia (IQAC-CSIC) c/ Jordi Girona 18-26, 08034 Barcelona, Spain
[2] Structural Biology and NMR Laboratory, Linderstrøm-Lang Centre for Protein Science, Department of Biology, University of Copenhagen, 2200, Copenhagen, Denmark
[3] Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Baldiri Reixac 10, 08028, Barcelona, Spain.
[4] ICREA, Passeig Lluís Companys 23, 08010, Barcelona, Spain

* Correspondence: ramon.crehuet@iqac.csic.es (R.C.); lindorff@bio.ku.dk (K. L.-L.).

Received: date; Accepted: date; Published: date

Abstract: Bayesian and Maximum Entropy approaches allow for a statistically sound and systematic fitting of experimental and computational data. Unfortunately, assessing the relative confidence in these two types of data remains difficult as several steps add unknown error. Here we propose the use of a validation-set method to determine the balance, and thus the amount of fitting. We apply the method to synthetic NMR chemical shift data of an intrinsically disordered protein. We show that the method gives consistent results even when other methods to assess the amount of fitting cannot be applied. Finally, we also describe how the errors in the chemical shift predictor can lead to an incorrect fitting and how using secondary chemical shifts could alleviate this problem.

Keywords: Bayesian methods; Maximum Entropy; Chemical Shifts; Intrinsically Disordered Proteins; Protein Ensembles; Structural Modelling; NMR; Molecular Dynamics

1. Introduction

Intrinsically Disordered Proteins (IDPs) do not fold into stable conformations, because their free energy landscapes possess many shallow local minima. Some of these may correspond to conformations rich in secondary structure and α-helices appear to be the most common of these regular structures. Their ability to fold into different conformations allows IDPs to interact with different binding partners and tuning their populations by post-translational modifications can allow for the regulation of important cellular functions. An accurate description of the secondary structures adopted by IDPs, as well as of their populations, can therefore help us understand the cellular behaviour of this important class of proteins.

Several methods exist to generate ensembles of configurations representing the conformational heterogeneity of IDPs, including all-atom molecular dynamics simulations [1], implicit-solvent Monte Carlo simulations [2] and other sampling approaches [3–5]. As the time evolution of the conformations is not relevant in the types of ensemble reweighting we here consider, from now on we loosely refer to molecular dynamics as any simulation method that can produce an ensemble of conformations. In practice, the simulated ensembles differ from the true ensembles for three main reasons. First, because the simulation conditions may not be exactly the same as the experimental ones —for example, different ionic strength, salts or pH may be used. Second, because equilibrium sampling of the IDPs conformational space is challenging to achieve, mainly due to its enormous size. And third, because all energy functions used in molecular simulations describe particle interactions with limited accuracies. This is particularly so for water-protein interactions, that are prevalent in IDPs due to their large solvent-accessible surface area [1,6–10]. Also, because of their

"flat" energy landscapes, IDPs may be particularly prone to even minor inaccuracies in energy functions. Thus, it would be desirable to correct the simulated ensembles by improving their agreement to measured experimental data.

Some experimental data produce properties that depend on the molecular structure, and therefore can be used to correct simulated ensembles [11–13]. However, the resolution of the experimental data is lower than that of molecular dynamics so that determining conformational ensembles from experiments needs to deal with experimental uncertainties [14,15]. There are basically two different approaches to compare experiments and simulations. Either molecular coordinates are produced from the experimental data, as is done with X-ray crystallography or NMR models [16], and these can then be compared against the simulated structures; or the experimental observable is generated from the simulated conformations and compared directly to the experimentally measured quantities [17]. The first approach inevitably introduces some models and assumptions to generate a higher resolution model (molecular coordinates) from a lower resolution one (experimental outcome). Therefore, the second approach is less model dependent as long as the observable quantities can be calculated from the molecular coordinates. The latter is achieved with a predictor also known as a forward model. The need for fast and (relatively) accurate predictors is a limitation of these second approach but they exist for several nuclear magnetic resonance (NMR) techniques.

NMR chemical shifts (CS) are extremely useful to give insight into the secondary structure of IDPs [18–23]. CS measured by NMR are determined mainly by the type of residue but, for a given residue, some atoms will give different CSs depending on the local geometry of the residue, as quantified for example by its location in the Ramachandran map [23]. CS for a given protein conformation can also be predicted with a low computational cost. Therefore, CS can be used to benchmark conformational ensembles produced by molecular dynamics simulations and, if necessary, to fit them to experiment.

Maximum Entropy or Bayesian approaches are powerful tools to fit simulated ensembles to experimental CSs. Maximum Entropy produces the minimum perturbation to the original ensemble so that it fits the experimental data. In a pure Maximum Entropy approach, experimental and prediction errors are not taken into account. This limitation can be overcome by reformulating the Maximum Entropy approach within a Bayesian reasoning [24–26]. In this work, we will use the algorithm called Bayesian/Maximum Entropy (BME) put forward by Bottaro et al. [26], which is based on work by Hummer and Köfinger [25]. The CS calculated from a simulated conformational ensemble can be considered the prior distribution in a Bayesian formulation. Even if the experimental error is known, errors in the prior are difficult to quantify, including force field accuracy, lack of full convergence of the simulations and errors in the forward model used to calculate CS from the ensemble. In the BME reasoning, we can tune the confidence in the prior with respect to the likelihood of the data with a subjective hyper-parameter θ (see Eq. 1 in Methods).

The choice of the confidence parameter θ is far from trivial. θ defines how much we rely on the simulated results compared to the experimental data; i.e. how much we want to reweight the simulated data to fit the experimental ones.[12,27] Here we suggest a new technique to determine this number when fitting simulated ensembles of an IDP using CS.

Protein CS have previously been used in Bayesian approaches to structure determination [28–33]. Purely Bayesian methods require parameter sampling and are computationally expensive. Therefore, they cannot easily be applied directly to ensembles of thousands of structures. The ensembles are usually clustered into a much smaller set, but, as clustering of IDPs is far from trivial, here we use BME which can deal with tens of thousands of structures and their experimental observables. The calculations reported in this work took at most a few minutes to run - depending on the number of θ values scanned - in a python Jupyter notebook running on a workstation.

As an example of an IDP we will use the protein ACTR [34]. ACTR consists of 71 residues, and was recently used as a benchmark system in a series of molecular dynamics simulations with different force fields [35]. We here focus on the simulations with three force fields: AMBER99SB-disp, AMBER03ws, and CHARMM36m (respectively named a99SBdisp, a03ws and

C36m hereafter). As previously described [35], different force fields predict different helical content for different regions of the protein sequence (Figure 1) These force fields represent some of the newest versions of two popular force field families [1]. However, our aim here is not to evaluate their strengths or weaknesses, but to use them as three possible conformational ensembles.

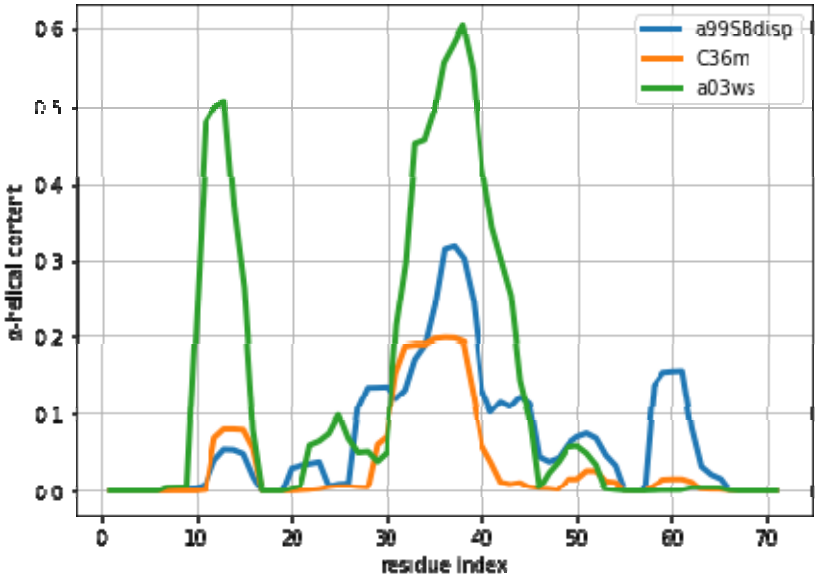

Figure 1. Amount of helical content in the ACTR trajectories simulated using different force fields. The a99SBdisp is taken as the target ensemble and the C36m and a03ws as the simulated trajectories to reweight.

Methods to test approaches to combine experiments and simulations are often hampered by the fact that we generally do not know the true target ensemble, and thus it is then not clear what an accurate result would be. Thus, to gain full control of the reweighting procedure and to know the desired outcome, we here use synthetic data. This is an approach that has proved powerful for discovering the impact of experimental uncertainties in structure determination that we and others have previously used [36–41]. In particular, we consider the a99SBdisp as a reference ensemble for generating synthetic data (i.e. we treat calculated CS from this ensemble as the experimental data). To clarify that these structures and their predicted CS are not actually experimental, we call this ensemble *target ensemble*. Then we will fit the predicted CS of other ensembles to the CS of the target ensemble. Namely, we will fit the a03ws and the C36m trajectories, which correspond to different conformational ensembles. We will also fit the a99SBdisp derived CS into the target ensemble, i.e. into itself, to evaluate the effect of the errors of the CS predictor (see Methods section).

The use of CS derived from a target ensemble which is known will allow us to compare the reweighted distribution of CS to the target-ensemble distribution, which would not be accessible from an actual NMR experiment. The CS from an NMR experiment depend on the time-scale of exchanging conformations. If, as it is for IDPs, this time-scale is fast, the measured CS is a time and ensemble average of all conformations. Thus we only get the first moment (the mean) of the distribution of CS in the conformational ensemble. In cases when the experiment reports distributions or higher distribution moments, several approaches can extend the Maximum Entropy method to those cases [42–44].

Even if the experiment only produces the mean of the distribution, the goal of reweighting is to obtain an ensemble that better reproduces the experimental ensemble of structures, not only its (measured) mean. By using synthetic data, we can compare how the distribution of CS during the reweighting approaches the target distribution. After all, the BME procedure guarantees that we can systematically approach the mean value, but at some point, overfitting could lead to a distribution that deviates from the target ensemble distribution instead of approaching it.

In this work we will explore how similar the target and reweighted ensembles are during a reweighting procedure using the distance between the distributions (see Methods section). We will also compare the behaviour of the average value of these distributions. From that, we will propose some rules to assess the amount of reweighting (determining θ) in real situations where the target distribution is unknown because only the experimental average value is available.

2. Methods

The MD trajectories with different force fields were generated as part of a recent study aimed at developing and benchmarking force fields [35] and were kindly shared by the authors. In this work we use the trajectories generated with the Amber ff03ws force field used with the TIP4P/2005 water model and with scaled protein-water interactions (a03ws) [6], the CHARMM36m (C36m) force field with its native TIP3P water model [45], and Amber ff99SB-disp with its native TIP4P-D-based water model (a99SBdisp) [35]. All the simulations are 30 μs long and contain N = 29777 frames. In this work we take the a99SBdisp as the target distribution and we will fit the a03ws and the C36m to this one, using CS. In the main text we focus on the a03ws results, but all the analysis is reproduced for the C36m ensemble as supplementary figures.

CS can be used to reweight IDP ensembles because they are sensitive to local secondary structure conformations. Indeed, experimentally derived CS indices allow for the quantification of secondary structure based on the CS values [46,47]. Although one might use the predicted secondary structures from these algorithms to reweight simulated ensembles this would introduce unnecessary model bias, and we instead computed the CSs from the ensembles and calculate the primary data directly. For that, we used existing algorithms and software to predict CS from structures.

Methods to calculate CS can introduce both systematic and random errors in their predicted values, but because we do not have experimental CS for our target ensemble, but predicted ones, we need a procedure to model these errors, that would be present in an actual experiment. The use of a predictor with perfect precision and accuracy can give some insight into the information content of CS, as we also explore below, but it does not reflect an experimental setting. Although we could have arbitrarily introduced random and systematic error to the predicted CS of the target ensemble, we preferred an alternative approach where the errors came from using two different predictors. We believe that that magnitude of the introduced errors will be more realistic than chosen arbitrarily.

We used Sparta+ [48] and PPM [49] to simulate the target a99SBdisp CS and only PPM [49] to simulate the reweighted distributions. When not stated otherwise, the target CS come from Sparta+ and we therefore do not expect that reweighting will lead to perfect agreement with the CSs predicted from the ensembles using PPM. PPM was specifically designed to calculate CS for individual frames in a conformational ensemble, i.e. it does not include in its training the thermal fluctuations around a given structure that others methods implicitly absorb into the predicted values. We have used as the error for the PPM predictor the reported errors of the validation set (table 3 in [49]), namely 1.06 for Cα, 1.23 for Cβ and 1.32 for the carbonyl C. This choice is based on the fact that the experimental error in the CS is negligible compared to the error of the forward model, and thus that the main uncertainty when determining whether a predicted average CS is "close" to the experimental measurement comes from the error of the forward model. We note also that the estimated error of the CS predictors come from analyses of a much more heterogeneous set of structures and might be an overestimate compared to the relatively narrow distribution of CS in IDPs. Our set of CS contained 69 Cα chemical shifts, 64 Cβ chemical shifts and 69 C chemical shifts, totalling m=202 chemical shift data.

The application of the BME approach is equivalent to the minimization of the following objective function:

$$\mathcal{L} = \frac{m}{2}\chi^2_{red}(w_1 \cdots w_N) - \theta S_{REL}(w_1 \cdots w_N) \quad (1)$$

Where w_i are the reweighted weights, N is the number of structures in the ensemble and m the number of experimental observables. χ^2_{red} measures the degree of fitting:

$$\chi^2_{red} = \frac{1}{m}\sum_{i=1}^{m}\frac{\left(\sum_{i=1}^{N} w_j CS_j^i - CS_{EXP}^i\right)^2}{\sigma_i^2} \quad (2)$$

Where CS_j^i is the chemical shift of atom i in structure j and CS_{EXP}^i is the target value. And the relative entropy, defined as $S_{REL} = -\sum_{j=1}^{N} w_j \log\left(\frac{w_j}{w_j^0}\right)$ measures the deviation from the initial weights, w_j^0 , in our case taken as $1/N$. S_{REL} is closely related to the effective sample size, N_{eff}, a useful measure of how much reweighting has taken place (see Eq. 3).

In the calculation of χ_{red}^2 we divide by the number of CS data, assuming completely independent data. Each σ_i is equated to the error of the PPM predictor reported above, as errors in the convergence of the simulations and the accuracy of the force fields are difficult to quantify. Because of these assumptions, the absolute value of χ_{red}^2 is of limited utility and it should be merely interpreted as a normalized measure of the residuals.

The effective sample size varies from 1 when all structures have the same weight, i.e. before the reweighting; to 0, in case a single structure gets 100% of the weight. There are different definitions on the effective sample size, one commonly used was put forward by Kish [50], but here we use the definition based on relative entropy, as it fits better within the BME paradigm:

$$N_{eff} = \exp\left(S_{REL}\right) \quad (3)$$

The optimization of Eq. 1 can be performed with a set of auxiliary Lagrange parameters λ, which allow writing the optimal weights as [24,51,52]:

$$w_i = \frac{1}{Z}\exp\left(-\sum_{j=1}^{m} \lambda_j CS_j^i\right) \quad (4)$$

Where Z is a normalization constant. This expression highlights two important aspects of the BME approach. First, the N weights are determined by optimizing the values of m λ parameters. As N, the number of structures in the ensemble, is usually much larger than the number of experimental data m, this reduces overfitting. Second, the fitting consists of adding a linear term to the free energy associated to each observable. This linear term can be seen as the minimal perturbation needed to tilt the distribution of the computed ensemble so that its average shifts towards the experimental value. Indeed, a purely Bayesian approach introducing a linear term to the energy [53] leads to results that are equivalent to a Maximum Entropy approach including an error term [54]. The strength of the reweighting is directly determined by the values of λ, and not by θ. In other words, a given θ may lead to different λ, and thus different reweightings for two different ensembles fitted to the same target values. (see Fig. S7)

After each of the fittings, we compared the average values with Eq. 2 and the distribution of values between the reweighted and the target distribution with the Wasserstein distance, as implemented in the Python package SciPy [55]. As each pair of reweighted and target CS distributions produces a Wasserstein distance value, thus we used the mean of all these m values as a measure of the similarity between two sets of distributions.

Secondary CSs were calculated for the target and reweighted ensemble in the same way. The secondary structure of each conformation was determined with the DSSP [56] implementation in MDTraj [57]. Then, the averaged CS for residues in coiled conformations represents the random coil reference. The secondary CS is the result of subtracting the random coil reference to the CS.

3. Results

The first step in a reweighting procedure is to ensure that the ensemble can be reweighted. Ideally, the ensemble should contain values around the target value. Figure S1 shows that this is the case for the two fitted ensembles discussed in this work. If the target value lies outside the distribution of the calculated ensemble values, the convergence of the reweighting is compromised. In that case one should check if there are systematic errors in the predictor, the experiment, or the simulation. Alternatively increasing the sampling could populate the less probable regions of the distribution. Sampling of these regions can also be achieved by running simulations with restraints [33,52,58–62].

The improvement of the fit resulting from the reweighting can be measured with the χ_{red}^2 explained in the Methods section. Lower values of χ_{red}^2 indicate a better fitting. As reweighting takes place, some conformations gain weight while others lose it. This leads to a reduced entropy compared to the original ensemble, which can be converted to an effective sample size, N_{eff}, ranging

from 1 to 0. As the amount of reweighting increases to fit the data more accurately, N_{eff}, decreases. In some systems, the shape of the vs. N_{eff} curve allows for the determination of a critical value from which no further reweighting is necessary [26,63]. However, in fitting CS of ACTR we find a relatively homogeneous decrease of , which makes it difficult to decide when to stop fitting(see Figure 2). A second common procedure is the Wald-Wolfowitz run test [64] for the residuals, where one checks that they are mutually independent. Long stretches of residues with the same sign indicate that more reweighting is possible. However, when using BME, decreasing θ decreases the size of the residuals, but not their signs (see Figure S2), i.e. each reweighted CS approaches its target CS from the same side as θ is decreased. This renders the Wald-Wolfowitz run test procedure inadequate.

We note here that Figure 2 shows a low even at the start of the fitting procedure. A below one may suggest that the fit is good enough, but can also reflect problems in estimating errors in both experiments and forward models, and in the statistical independence of the data. The used in the calculation corresponds to the PPM error for a heterogeneous set of structures and their experimental CS. When compared to Sparta+ ACTR chemical shifts, the errors are much smaller (See Figure S3). Because of this error underestimation, the absolute value of is of limited use. If we use the PPM CS for the target ensemble, we are effectively using a predictor without error. In that case, is ill-defined, but we get a curve very similar to Figure 2 (see figure S4). This suggests that the CS difference reflects more the different structural composition of the ensemble (Figure 1) than the errors of the predictor.

Indeed, the lack of a clear target value for is one key reason why we need to determine the parameter θ. Independently of the origin of the small the ensembles described by the three force fields are different (Figure 1) and should be distinguishable by their differences in CS. For many residues the CS differences are small because they present a random coil structure, but the region where the target and simulated ensembles show different helical content, the CS show larger differences (Figure 3 and S5). This trend is more clear for the a03ws force field than for the C36m (Figure S5). This suggests that reweighting would improve the simulated ensembles by rendering them closer to the target ensemble. But traditional methods do not always provide a clear answer to how to choose θ to avoid over-fitting but have enough reweighting.

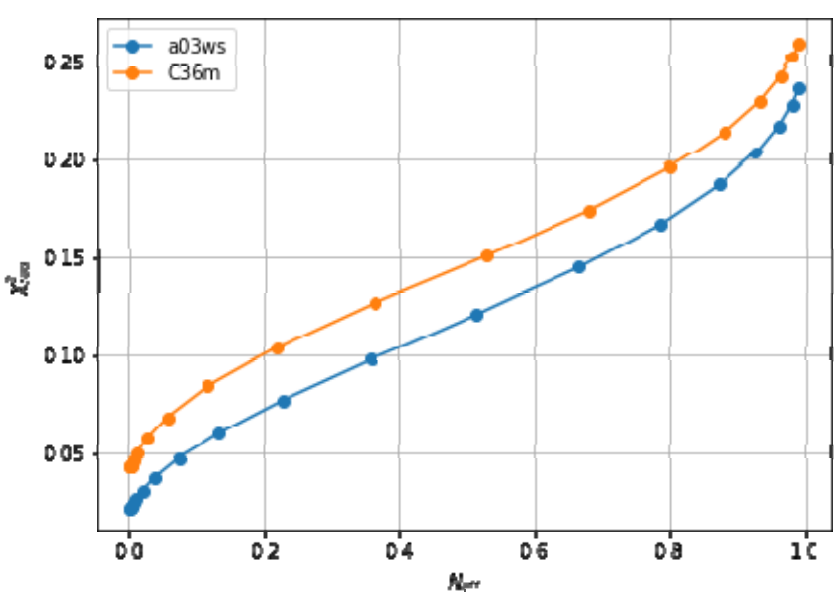

Figure 2. Evolution of with the effective sample size (N_{eff}), showing the lack of an L-shaped curve. The a99SBdisp is taken as the target ensemble and the C36m and a03ws as the simulated trajectories to reweight.

In this work we suggest using a validation set to determine the relative weights of the experimental data and the simulation prior when reweighting, i.e. determine an optimal value of θ. A similar idea was used by Cesari et al. to find the optimal balance between the distribution arising from the force field and the experimental data [65]. Here, we put forward the following procedure:

1. Split the frames of the trajectory into a training set (t) and a validation set (v) of the same size. We used odd and even frames for the training and validation set respectively. We use sets of the same size because we aim to compare average values and distributions, not individual conformations of the validation set. A validation set as large as the test set is therefore needed to minimize the standard error of the mean and discretization errors of the distribution. We here chose to use an interleaved training and validation set after confirming that the two sets have highly uncorrelated CSs as frames are sampled only once every 1 ns.

2. Fit the BME for a range of θ values to the training set and apply the optimized Lagrange parameters λ to reweight the validation set.

3. For each of the θ values, evaluate the following properties

    a. (Eq. 2) of the training ( ) and the validation set ( ).

    b. Average distance between the training and the validation sets ($D(t,v)$).

    c. Average distance between the training set and the target (goal) distribution ($D(t,g)$), and average distance between the validation set and the target (goal) distribution ($D(v,g)$).

Note that step 3c is only possible in the case of synthetic data and is used as a benchmark for a selection procedure. Ideally we would like to reweight as long as the validation set distribution approaches the target distribution, i.e., we want to minimize $D(v,g)$. In a real case scenario, as the target distribution is not known, we cannot measure this distance. Therefore we need to see if any of the accessible measures can give us information about $D(v,g)$.

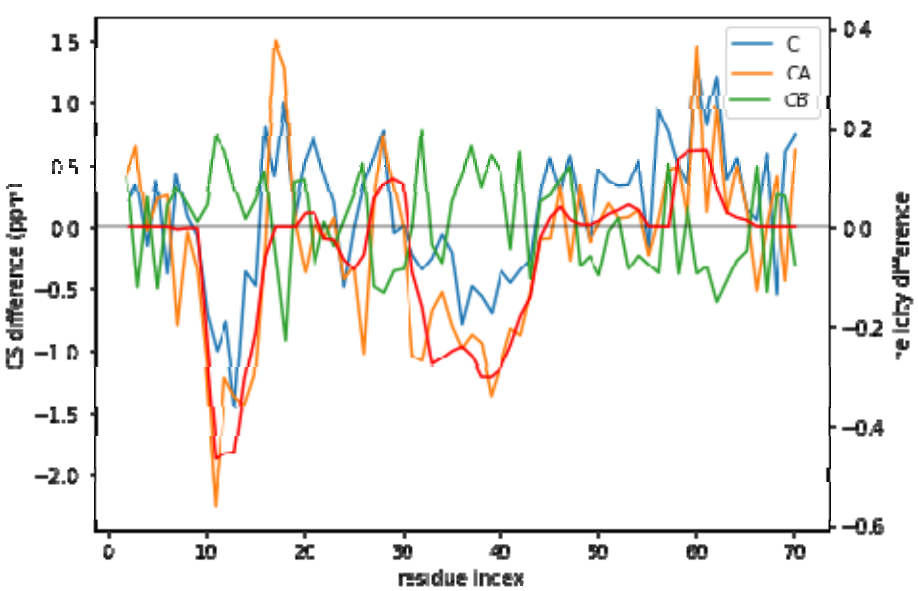

Figure 3. CS difference between the target CS and the a03ws CS for the three atoms used (C, CA, and CB) compared to the difference in helicity for these two ensembles (in red). Although for many residues the difference lies below the error of the predictor, for some regions it does not. Besides, the helicity is correlated with the CA and the CB CS difference.

Fig. 4 plots all the values described in the previous procedure. As expected, the BME produces an average CS that approaches the target CS, so that it systematically reduces . At θ≈3 the validation set fits best the target CS and then starts increasing. This is a sign of overfitting. Interestingly the behaviour of these averaged values is followed very similarly by the distributions, as represented by the shape of $D(t,g)$ and $D(v,g)$. The fact that the distributions, which are what we aim to fit, and the average values have parallel behaviours is very positive as this tells us that we can use the average value as a proxy of the distribution. At a value close to $\theta \approx 3$, $D(t,v)$ has a high curvature and starts growing. The training and the validation distributions, which had been very similar for values $\theta > 3$, start diverging. This confirms that at this point, overfitting occurs. The fit to the C36m ensemble shows a very similar trend (Fig. S6). A different measure also suggests the start of the overfitting regime. Fig. S7 shows the root mean square of the Lagrange λ terms that the BME introduces for the fit (see Eq. 4). At $\theta < 2$ the λ parameters start growing at a much faster speed, showing that from this point the reweighting becomes very strong and very sensitive to the θ value.

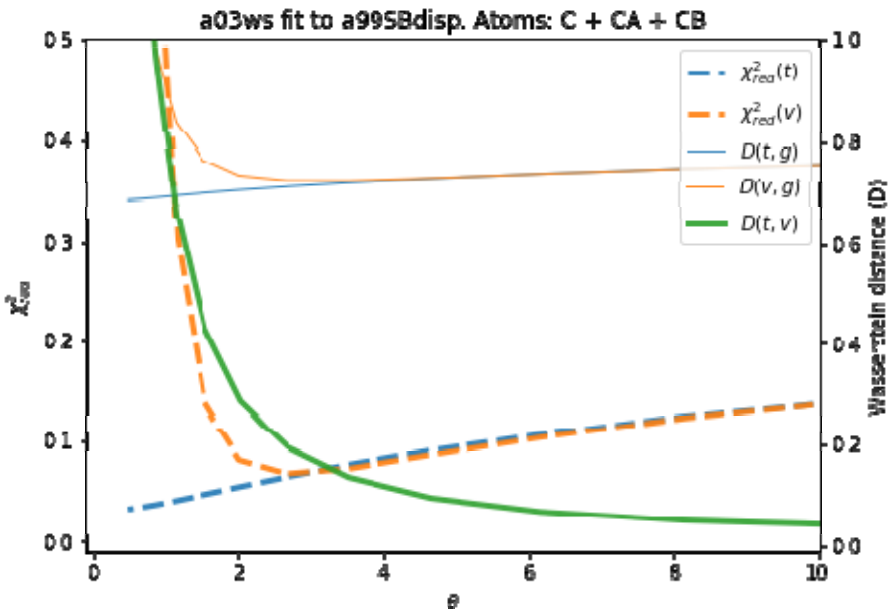

Figure 4. Behaviour of different quantities during the reweighting procedure of the a03ws ensemble. The quantities defined as thin lines would not be measurable in a real case scenario, but their behaviour can be inferred from the quantities in thick lines.

To evaluate the importance of the error introduced by the predictor and whether this was leading to overfitting, we reproduced the same analysis with target CS coming from the PPM predictor, i.e. excluding the predictor error. As Figure S8 depicts, the behaviour of the different quantities is equivalent to Figure 4, with the only difference than the difference in distributions $D(t,g)$ and $D(v,g)$ and shifted to lower values, as expected from the use of the same predictor. The two distributions are more similar, because the predictor error is missing, and they become diverging at a slightly smaller value of θ, as also does. This shows that the stochastic error of the predictor averages out and that BME is robust to it. Systematic errors in the predictor could be important, as will be shown below, but in this case their influence is much smaller than the differences in CS arising from conformational differences.

Once an optimal value for θ has been determined, other properties of the ensemble can be calculated. As an example, here we calculate the α-helical content and plot it in Figure 5 and S9. We stress that the reweighting improves the fit, but does not lead to a perfect agreement. There are several sources for the disagreement. The major source of the disagreement is the fact that the predictor (PPM software) has an error in the prediction with respect to the target value (Sparta+ software). Even for the same ensemble of configurations predicted and target CS differ, thus leading to a disagreement between secondary structure content, as is discussed below. Further, although the

two are tightly related, there is not a one-to-one relationship between the secondary structure and the CS values. Also, the reweighted ensemble may not contain structures with the optimal α-helical content. For example, the a03ws ensemble contains a negligible amount of α-helical structures around residue number 60, therefore, even after the reweighting, the helicity of this region cannot increase and remains too low. Finally, as long as the experimental data does not uniquely specify the target distribution, procedures such as this will always be affected by the choice of a reference (prior) distribution and will thus not give the exact target distribution.

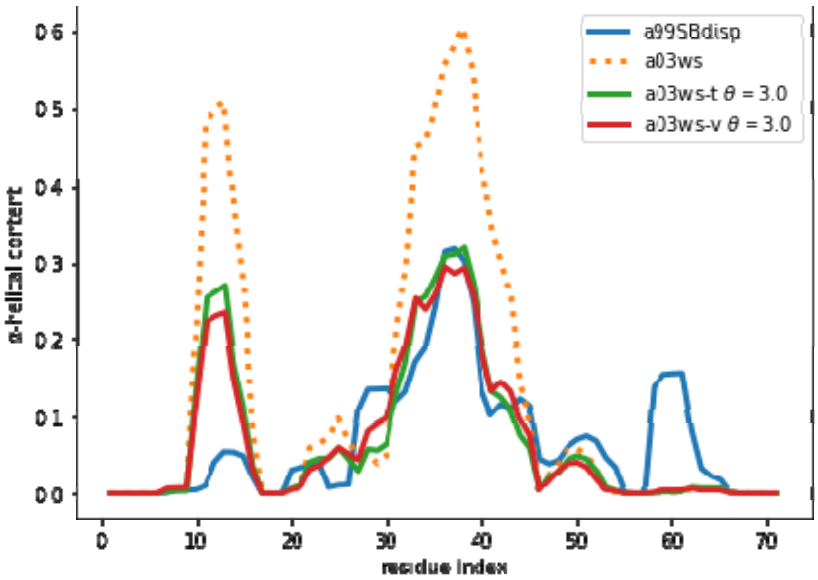

Figure 5. α-helical content for the reweighted and target ensembles. For the reweighted ensembles the train (a03ws-t) and validation (a03ws-v) sets are shown. The original ensembles before the reweighting is also shown.

When we examine the α-helical content of the validation ensemble after reweighting we find that it is further away from the initial ensemble than the training ensemble. This makes sense as we are using a Maximum Entropy algorithm, which ensures minimal perturbation of the training ensemble from the original one. The minimal perturbation, expressed as minimization of the Kullback-Leibler divergence, is true only for the training ensemble. Therefore, the parameters of the training ensembles applied to the validation ensemble lead to a further divergence from the original ensemble. Consequently, the described procedure should be used to determine the optimal θ, but once it is known, one should re-fit the complete ensemble with that optimal θ.

Figure 5 and S9 show that the reweighting changed the helicity the most in the region between residues 30 and 45. This is because CSs are not only sensitive to local residue conformations, but also to neighbouring residue conformations. This results in the calculated data in long helical stretches differing most from a random coil than short helices. Consequently, the reweighting is able better to 'see' long helices and reweight those regions accordingly.

The reweighting will only improve properties that are sensitive to the experimental data used. In this case, the CS of different atoms are sensitive to the helical conformations to a different degree. Thus, reweighting leads to an ensemble with improved helical content. Other properties, such as the radius of gyration (R_g), are sensitive to the global mass distribution of the conformations. For expanded ensembles, an increase of helicity leads to more collapsed ensembles, as helical structures are compact [66,67]. For the a03ws, reweighting leads to an overall decrease of helicity (Fig. S10), and therefore to a larger R_g. For the C36m, the slight increase in helicity does not correlate with the R_g, presumably because the initial ensemble is already rather compact. As also emphasized by Best as co-workers, both local (such as secondary structure) and global (such as R_g) properties should be used to characterize IDPs ensembles, and one should not expect the reweighting based on one set of

these properties to improve the other [1]. Using these other properties as cross-validation, as it is sometimes done, may lead to an incorrect perception of the amount of reweighting needed. As an example, the C36m fitting shows a minimum that could suggest that the evolution of R_g could be used as a cross-validating property, but the a36m shows a monotonic increase, for the reasons previously mentioned. If several experimental data are known, they should be included in the reweighting to obtain more realistic ensembles with improved local and global properties.

A good way to estimate the effect of the predictor error is to fit the target ensemble to itself. In that case, the disagreement between predicted and target CS comes exclusively from the use of different predictors. We will refer to predictor error as the difference between the predictor used for the reweighting ensemble (PPM) and the predictor used as a target (Sparta+) without any assumption of which one gives CS closer to the true experimental ones.

If the predictor error was a Gaussian noise with zero mean, it would average to zero for an ensemble of tens of thousands of structures. However, the predictor has systematic deviations that depend on the residue type and its conformation (Fig S11). The systematic deviations also correspond to what higher helical content would give: negative C and CA and positive CB (see Fig. 3). This suggests that the reweighting will tend to decrease the helical content.

By following the same procedure as before, we obtain an optimal value of $\theta \approx 2$ (Fig. S12). It may seem surprising that when the reweighted ensemble matches exactly the target ensemble the θ value is lower than when fitting the a03ws and C36m ensembles. Two things need to be considered. First, the amount of disagreement between the CSs is not much lower than when fitting the a03ws or the C36m ensembles. This shows that the predictor is a major source of errors, but even with considerable unknown systematic errors in the predictor, the reweighting of the a03ws and C36m ensembles gave consistently improved ensembles. Second, θ values between ensembles are not comparable because θ does not determine the strength of the reweighting. The strength of the reweighting is determined from the optimized values of the weights, which themselves arise from the optimized values of the Lagrange (λ) parameters (see Eq. 4). For a given θ, the reweighting is weaker for the current a99SBdisp ensemble than for the a03ws and C36m (see Fig. S7). As Fig. 6 shows, even after the reweighting with $\theta = 2$, the helicity has changed, by at most ~ 0.05, from 0.31 to 0.26 in the region of residue 37. For all the other regions the changes in helicity are smaller. After all, the reweighting procedure 'sees' CSs that would correspond to lower helicity, and therefore results in an ensemble with lower helicity.

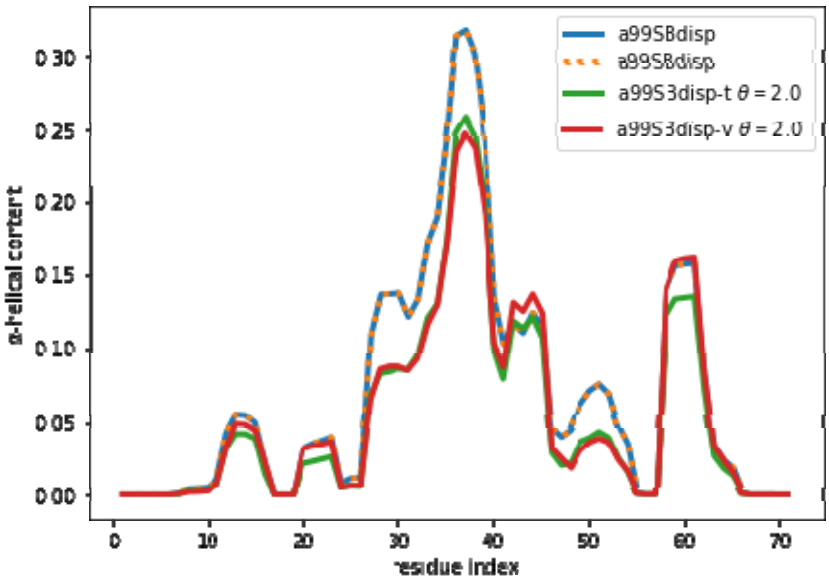

Figure 6. α-helical content for the reweighted and target ensembles. For the reweighted ensembles the train (a99SBdisp-t) and validation (a99SBdisp-v) sets are shown. The original ensembles before

the reweighting is also shown and, as expected, it corresponds exactly to the target ensemble as they are the same.

Although the reweighting from the target a99SBdisp ensemble was not large, deviating from the initially correct ensemble is not satisfactory. We therefore looked for alternatives to minimize the reweighting. One alternative is to improve the matching between calculated and target CS. One would be to use the best possible predictor as its accuracy when comparing with actual experimental CS should not be overlooked.

Considering that a source of error in the predictor may come from the baseline CS for each residue, we hypothesized that the use of secondary CSs could lead to a cancellation of errors. Secondary CS are defined as the difference between measured CSs and CSs from random coil conformations [20]. They measure the change of CS when a given residue adopts a secondary structure conformation. If predictors are good at describing this change then secondary CS could lead to more sensitive reweightings. As secondary structures can be assigned at a residue level for each conformation of a computational ensemble, the generation of secondary CS from computed ensembles is fast and simple (see Methods section).

The use of secondary CS leads to essentially no-reweighting. We repeated the optimization procedure described above with calculated and target secondary CSs. The optimization of θ leads to flat curves (Fig. S13), that at first sight could suggest that the optimal value of θ is difficult to determine. To a certain extent this is true, the reason being that the reweighting is mostly insensitive to θ. What is more interesting is that the reweighting is negligible even for extremely small values of θ, as Fig. S13 shows, so that the reweighted ensemble remains essentially unchanged and thus, close to target ensemble as it was desirable.

As expected, the negligible reweighting does not lead to overfitting. Fig S15 shows that in the fitting of secondary CS, the effective sample size stabilizes to a value of 0.42 for $\theta<10^{-3}$. A value of N_{eff} far from 0 shows that the method is not overfitting. We remind the reader that the amount of reweighting is not determined by θ itself but by the Lagrange λ parameters that result from the optimization procedure. The result of this procedure tells us that, whatever our relative confidence in the computed ensemble and the target CS, the reweighting will be essentially zero.

The secondary CS before any reweighting already show an excellent agreement with the target ones, so that reweighting is not necessary. The χ^2_{red} before reweighting in one order of magnitude smaller for secondary CS than for CS. However, we wanted to test our method in an extreme case. We conclude that even when it could easily lead to overfitting, the nature of BME prevents that, leading to stable results. But it also shows that when there are alternative methods to reduce the errors of the predictors, as the use of secondary CSs, the reweighting becomes simpler (in the sense that one does not need to tune θ) and more accurate.

4. Discussion

The Maximum Entropy approach is a simple, yet powerful method to fit computed ensembles to experimental data [68]. Its extension into a Bayesian approach allows the treatment of uncertainties arising from both experiment and computation [17,25,26]. It is often difficult to determine the optimal balance between the prior information, encoded in the force field, and the experimental data. These difficulties arise in particular because computational errors are difficult to assess. They arise from the predictor (the forward model), the convergence of the simulation, the accuracy of the force field used and the computational reproducibility of the experimental conditions. Also, in many cases it is difficult to determine accurate estimates of experimental errors. Together, these effects result in difficulties in determining a unique value of θ, or that a wide range of possible θ values appear comparably good. Because of uncertainties in the experimental and computational errors it is difficult to choose the level of reweighting simply by the magnitude of χ^2_{red}, thus making it important to find objective and robust criteria to determine the values of θ.

Here we suggest using a validation set that arises from the same distribution, and therefore belongs to the same model as the training set. As we show in the case of comparing results from CS

and radius of gyration, the use of other physical observables for cross-validation can sometimes lead to a wrong choice of θ. If the observables are strongly correlated to the ones being fitted, it will be hard to detect any overfitting, as the overfitted ensemble, even if having a small effective size, will also reproduce the correlated observables. On the contrary, if the physical observable is not correlated, then there is no reason why the reweighted ensemble should improve its fitting, as we showed for the radius of gyration.

The split of the data between training and validation sets can be done in different ways. The two sets would be maximally uncorrelated if the sets arise from the first and the second half of the trajectory. But unless each of the halves is fully equilibrated –which is not the case in our trajectory– the two sets would correspond to different distributions, and one should not expect that the same Lagrange λ parameters should apply to the two sets. For instance, if the first half samples more helical conformations than the second, the reweighting could lead to a reduction of helicity for the first half and an increase for the second. On the contrary, by using interleaved frames to create training and validation sets we ensure they represent samples of the same distribution, whether fully equilibrated or not. Therefore, the same Lagrange λ parameters should improve both training and test until the moment we start overfitting the training set. Therefore, there is a range of θ values before the overfitting that are dependent on the distribution to be fitted but independent of the particular structures chosen from that distribution. This is the basis of our selection of the optimal θ value. However, if the validation set is strongly correlated to the training set, the overfitting could not be detected in the validation set. We therefore recommend checking the correlation between neighbouring frames. Alternatively, the reweighting and validation suggested procedure can be repeated for a subsample of the trajectory and it should lead to a similar optimal θ value.

Even if we do not incur in overfitting, a large amount of reweighting indicates either that the prior is considerably different from the target distribution or that the predictor has important systematic errors. In both cases the reweighted ensemble will still be far from the actual ensemble, and one should be especially cautious in not over-interpreting it, especially in extracting structural information from it that is not sensitive to chemical shifts, as we have discussed for the radius of gyration.

We also showed that the errors of the predictor should not be underestimated. In the specific case of NMR CS, it appears that even when the predictors were designed and parameterized to remove systematic errors, they may still arise in applications to a particular system. These unknown systematic errors contribute to the reweighting and are a major source of overfitting that our method avoids to a certain extent.

The use of secondary CS seems a promising approach to reduce the errors of the predictors. Its calculation from a computational ensemble is simple, but the experimental calculation is not trivial. One approach is to use tabulated CS for protein random coils [20], including corrections depending on neighbouring residues as well as pH and temperature [69–73]. A more precise but more time consuming approach is to completely denature an IDP and measure the resulting CS [74,75]. Both approaches introduce some uncertainties that may render the secondary CS less accurate. In future work we will explore whether they still remain a better alternative when reweighing computational ensembles of IDPs.

We would like to end this discussion by reminding the reader that the BME and related approaches are particularly suited to avoid overfitting. Maximization of the entropy implies that the reweighted distribution is minimally perturbed, unlike in other reweighting approaches [45,46,46–49,49–53]. One may think that still, small values of θ, which means a strong confidence on the experimental data, would lead to an overfit of the finite number of computed structures. However, the fitting procedure fits a small number of parameters (equal to the number of CSs), in our case m=202, to a much larger number of structures, N=29977. If each weight was fitted individually, the procedure would be highly prone to overfitting, with N=29777 parameters. Instead, BME determines weights from a much smaller set of parameters, the number of observables (m=202) plus θ. The effect of this is that the possibility of overfitting is substantially reduced.

5. Conclusions

In this work we have shown how chemical shifts can be used to improve the configurations arising from molecular dynamics simulations of intrinsically disordered proteins. We have introduced a systematic method to assess the amount of reweighting (θ) needed to fit the experimental chemical shifts based on cross-validation. This approach allows to circumvent the difficulty of knowing the errors associated to the simulations and the chemical shift predictors. We have also shown how the predictor errors lead to an incorrect reweighting as they include a systematic bias. This error seems to be greatly reduced by using secondary chemical shifts, something that we will further explore in future work.

Supplementary Materials: The following are available online.

Figure S1: Distribution of the difference between the ensemble CS and the average target CS for the 2 ensembles discussed in this work. The x-values correspond to a concatenation of C, CA and CB CS. To be able to reweight the values should be distributed at both sides of y=0, as is the case.

Figure S2: Evolution of the CS difference (residuals) for the reweighting procedure for a range of θ values from 100 (purple values) to 0.1 (yellow values). As the residuals have the same sign for most of the reweighting procedure, one cannot use the Wald-Wolfowitz run test to define the amount of reweighting.

Figure S3: Difference in the chemical shifts predicted with Sparta+ and PPM for the a99SBdisp ensemble. The left plots show the value for each residue and the right plots show their distribution for a better visualization of their spread and shift. Atom types are CA (top), C (middle) and CB) bottom.

Figure S4: Evolution of χ^2_{red} with the effective sample size (N_{eff}), showing the lack of an L-shaped curve. The blue and orange lines are the same as in Figure 1. The NE suffix stands for "no-error". It corresponds to the values where the target CS are calculated with PPM. For the sake of clarity χ^2_{red} uses the same error as in the case with errors, even though χ^2_{red} is ill-defined in this case and should be regarded as a scaled root-mean-square error.

Figure S5: The CS difference between the target CS and the computed CS for the C36m ensemble compared to the difference in helicity for these two ensembles.

Figure S6: Behaviour of different quantities during the reweighting procedure of the C36m ensemble.

Figure S7: Root mean square of the Lagrange parameters λ during the reweighing procedure for all three ensembles. See Eq. 4.

Figure S8: Behaviour of different quantities during the reweighting procedure of the a03ws ensemble with no predictive errors (NE suffix). As was done in figure S3, χ^2_{red} uses the same error as in the case with errors, even though χ^2_{red} is ill-defined in this case and should be regarded as a scaled root-mean-square error.

Figure S9: Alpha-helical content for the reweighted and target ensembles for the C36m force field. For the reweighted ensembles the train (C36m-t) and validation (C36m-v) sets are shown. The original ensembles before the reweighting is also shown.

Figure S10: Evolution of the radius of gyration for different reweightings.

Figure S11: Error in the predictor (PPM) with respect to the target chemical shift (Sparta+) for different secondary structure elements of the a99SBdisp ensemble. For atoms CA and C there is a systematic underestimation of the chemical shift, whereas for CB, there is a systematic overestimation. The errors also depend on the type of secondary structure. The codes are the following: 'C': Loops and irregular elements, 'B': Residue in isolated beta-bridge, 'E': Extended strand, participates in beta ladder, 'G': 3-helix (3/10 helix), 'H': Alpha helix, 'I': 5 helix (pi helix), 'S': bend, and 'T': hydrogen bonded turn, as determined by the DSSP algorithm implemented in MDtraj.

Figure S12: Behaviour of different quantities during the reweighting procedure of the a99SBdisp ensemble. The quantities defined as thin lines would not be measurable in a real case scenario, but their behaviour can be inferred from the quantities in thick lines.

Figure S13: Behaviour of different quantities during the reweighting procedure of the a99SBdisp ensemble using secondary chemical shifts. The quantities defined as thin lines would not be measurable in a real case scenario, but their behaviour can be inferred from the quantities in thick lines. Remark that the χ^2_{red} values are very small for all θ values. D(t,g) and D(t,v) have been scaled by 1/20 so that the shape of the curves could be seen.

Figure S14: Alpha-helical content for the reweighted and target ensembles using secondary chemical shifts. For the reweighted ensembles the train (a99SBdisp-t) and validation (a99SBdisp-v) sets are shown. The original ensembles before the reweighting is also shown and, as expected, it corresponds exactly to the target ensemble as they are the same. Figure S12: Evolution of the effective sample size (N_{eff}) with θ for the secondary chemical shift fitting of a99SBdisp ensemble to the a99SBdisp target ensemble.

Figure S15. Evolution of the effective sample size (N_{eff}) with θ for the secondary chemical shift fitting of a99SBdisp ensemble to the a99SBdisp target ensemble.

Author Contributions: Conceptualization, R.C., X.S., and K.L.-L.; Methodology, R.C. and P.B..; Software, R.C.; Investigation, R.C.and P.B.; Writing – Original Draft Preparation, R.C.; Writing – Review & Editing, R.C., P.B.,X.S., and K.L.-L.

Funding: R.C. acknowledges funding from the MINECO (CTQ2016-78636-P) and to AGAUR, together with X.S. (2017 SGR 324). X.S. acknowledges funding from MINECO (BIO2015-70092-R), and the European Research Council (CONCERT, contract number 648201). K.L.-L. acknowledges funding from the Lundbeck Foundation BRAINSTRUC initiative.

Acknowledgments: We would like to thank D.E. Shaw Research for sharing the molecular dynamics trajectories with us.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. 1. Zerze, G.H.; Zheng, W.; Best, R.B.; Mittal, J. Evolution of All-atom Protein Force Fields to Improve Local and Global Properties. *J. Phys. Chem. Lett.* **2019**, *10*, 2227–2234.
2. 2. Vitalis, A.; Pappu, R.V. ABSINTH: a new continuum solvation model for simulations of polypeptides in aqueous solutions. *J. Comput. Chem.* **2009**, *30*, 673–99.
3. 3. Krzeminski, M.; Marsh, J.A.; Neale, C.; Choy, W.-Y.; Forman-Kay, J.D. Characterization of disordered proteins with ENSEMBLE. *Bioinforma. Oxf. Engl.* **2013**, *29*, 398–399.
4. 4. Ozenne, V.; Bauer, F.; Salmon, L.; Huang, J.-R.; Jensen, M.R.; Segard, S.; Bernadó, 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*, 1463–70.
5. 5. Estaña, A.; Sibille, N.; Delaforge, E.; Vaisset, M.; Cortés, J.; Bernadó, P. Realistic Ensemble Models of Intrinsically Disordered Proteins Using a Structure-Encoding Coil Database. *Struct. Lond. Engl. 1993* **2018**, *0*.
6. 6. 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*, 5113–5124.
7. 7. Best, R.B. Computational and theoretical advances in studies of intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **2017**, *42*, 147–154.
8. 8. Anandakrishnan, R.; Izadi, S.; Onufriev, A.V. Why Computed Protein Folding Landscapes Are Sensitive to the Water Model. *J. Chem. Theory Comput.* **2019**, *15*, 625–636.
9. 9. 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*, 5113–23.
10. 10. 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*, 2620–2634.
11. 11. Bonomi, M.; Heller, G.T.; Camilloni, C.; Vendruscolo, M. Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* **2017**, *42*, 106–116.
12. 12. Köfinger, J.; Różycki, B.; Hummer, G. Inferring Structural Ensembles of Flexible and Dynamic Macromolecules Using Bayesian, Maximum Entropy, and Minimal-Ensemble Refinement Methods. In

Biomolecular Simulations: Methods and Protocols; Bonomi, M., Camilloni, C., Eds.; Methods in Molecular Biology; Springer New York: New York, NY, 2019; pp. 341–352 ISBN 978-1-4939-9608-7.

13. 13. Ravera, E.; Sgheri, L.; Parigi, G.; Luchinat, C. A critical assessment of methods to recover information from averaged data. *Phys Chem Chem Phys* **2015**.
14. 14. Schneidman-Duhovny, D.; Pellarin, R.; Sali, A. Uncertainty in integrative structural modeling. *Curr. Opin. Struct. Biol.* **2014**, *28*, 96–104.
15. 15. Fenwick, R.B.; Esteban-Martín, S.; Salvatella, X. Influence of Experimental Uncertainties on the Properties of Ensembles Derived from NMR Residual Dipolar Couplings. *J. Phys. Chem. Lett.* **2010**, *1*, 3438–3441.
16. 16. Schröder, G.F. Hybrid methods for macromolecular structure determination: experiment with expectations. *Curr. Opin. Struct. Biol.* **2015**, *31*, 20–27.
17. 17. Bottaro, S.; Lindorff-Larsen, K. Biophysical experiments and biomolecular simulations: A perfect match? *Science* **2018**, *361*, 355–360.
18. 18. Jensen, M.R.; Zweckstetter, M.; Huang, J.-R.; Blackledge, M. Exploring Free-Energy Landscapes of Intrinsically Disordered Proteins at Atomic Resolution Using NMR Spectroscopy. *Chem. Rev.* **2014**.
19. 19. Kashtanov, S.; Borcherds, W.; Wu, H.; Daughdrill, G.W.; Ytreberg, F.M. Using Chemical Shifts to Assess Transient Secondary Structure and Generate Ensemble Structures of Intrinsically Disordered Proteins. In *Intrinsically Disordered Protein Analysis: Volume 1, Methods and Experimental Tools*; Uversky, V.N., Dunker, A.K., Eds.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2012; pp. 139–152 ISBN 978-1-61779-927-3.
20. 20. Kjaergaard, M.; Poulsen, F.M. Disordered proteins studied by chemical shifts. *Prog. Nucl. Magn. Reson. Spectrosc.* **2012**, *60*, 42–51.
21. 21. Kragelj, J.; Ozenne, V.; Blackledge, M.; Jensen, M.R. Conformational Propensities of Intrinsically Disordered Proteins from NMR Chemical Shifts. *ChemPhysChem* **2013**, *14*, 3034–3045.
22. 22. 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*, 1270–1272.
23. 23. Mantsyzov, A.B.; Shen, Y.; Lee, J.H.; Hummer, G.; Bax, A. MERA: a webserver for evaluating backbone torsion angle distributions in dynamic and disordered proteins from NMR data. *J. Biomol. NMR* **2015**, *63*, 85–95.
24. 24. Cesari, A.; Reißer, S.; Bussi, G. Using the Maximum Entropy Principle to Combine Simulations and Solution Experiments. *Computation* **2018**, *6*, 15.
25. 25. Hummer, G.; Köfinger, J. Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* **2015**, *143*, 243150.
26. 26. Bottaro, S.; Bengtsen, T.; Lindorff-Larsen, K. Integrating Molecular Simulation and Experimental Data: A Bayesian/Maximum Entropy reweighting approach. *bioRxiv* **2018**, 457952.
27. 27. Escobedo, A.; Topal, B.; Kunze, M.B.A.; Aranda, J.; Chiesa, G.; Mungianu, D.; Bernardo-Seisdedos, G.; Eftekharzadeh, B.; Gairí, M.; Pierattelli, R.; et al. Side chain to main chain hydrogen bonds stabilize a polyglutamine helix in a transcription factor. *Nat. Commun.* **2019**, *10*, 2034.
28. 28. Fisher, C.K.; Ullman, O.; Stultz, C.M. Efficient construction of disordered protein ensembles in a bayesian framework with optimal selection of conformations. *Pac. Symp. Biocomput.* **2012**, 82–93.
29. 29. Fisher, C.K.; Stultz, C.M. Constructing ensembles for intrinsically disordered proteins. *Curr. Opin. Struct. Biol.* **2011**, *21*, 426–431.
30. 30. Fisher, C.K.; Huang, A.; Stultz, C.M. Modeling intrinsically disordered proteins with bayesian statistics. *J. Am. Chem. Soc.* **2010**, *132*, 14919–27.
31. 31. Bratholm, L.A.; Christensen, A.S.; Hamelryck, T.; Jensen, J.H. Bayesian inference of protein structure from chemical shift data. *PeerJ* **2015**, *3*, e861.
32. 32. Potrzebowski, W.; Trewhella, J.; Andre, I. Bayesian inference of protein conformational ensembles from limited structural data. *PLOS Comput. Biol.* **2018**, *14*, e1006641.
33. 33. Bonomi, M.; Camilloni, C.; Cavalli, A.; Vendruscolo, M. Metainference: A Bayesian inference method for heterogeneous systems. *Sci. Adv.* **2016**, *2*, e1501177–e1501177.
34. 34. Iešmantavičius, V.; Jensen, M.R.; Ozenne, V.; Blackledge, M.; Poulsen, F.M.; Kjaergaard, M. Modulation of the Intrinsic Helix Propensity of an Intrinsically Disordered Protein Reveals Long-Range Helix–Helix Interactions. *J. Am. Chem. Soc.* **2013**, *135*, 10155–10163.

35. 35. 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**, 201800690.
36. 36. Robustelli, P.; Cavalli, A.; Dobson, C.M.; Vendruscolo, M.; Salvatella, X. Folding of Small Proteins by Monte Carlo Simulations with Chemical Shift Restraints without the Use of Molecular Fragment Replacement or Structural Homology. *J. Phys. Chem. B* **2009**, *113*, 7890–7896.
37. 37. Esteban-Martín, S.; Fenwick, R.B.; Ådén, J.; Cossins, B.; Bertoncini, C.W.; Guallar, V.; Wolf-Watz, M.; Salvatella, X. Correlated inter-domain motions in adenylate kinase. *PLoS Comput. Biol.* **2014**, *10*, e1003721.
38. 38. De Simone, A.; Richter, B.; Salvatella, X.; Vendruscolo, M. Toward an Accurate Determination of Free Energy Landscapes in Solution States of Proteins. *J. Am. Chem. Soc.* **2009**, *131*, 3810–3811.
39. 39. Schneider, T.R.; Brünger, A.T.; Nilges, M. Influence of internal dynamics on accuracy of protein NMR structures: derivation of realistic model distance data from a long molecular dynamics trajectory. *J. Mol. Biol.* **1999**, *285*, 727–740.
40. 40. Lindorff-Larsen, K.; Ferkinghoff-Borg, J. Similarity measures for protein ensembles. *PloS One* **2009**, *4*, e4203.
41. 41. Camilloni, C.; Robustelli, P.; Simone, A.D.; Cavalli, A.; Vendruscolo, M. Characterization of the Conformational Equilibrium between the Two Major Substates of RNase A Using NMR Chemical Shifts. *J. Am. Chem. Soc.* **2012**, *134*, 3968–3971.
42. 42. Lou, H.; Cukier, R.I. Reweighting ensemble probabilities with experimental histogram data constraints using a maximum entropy principle. *J. Chem. Phys.* **2018**, *149*, 234106.
43. 43. White, A.D.; Dama, J.F.; Voth, G.A. Designing Free Energy Surfaces That Match Experimental Data with Metadynamics. *J. Chem. Theory Comput.* **2015**, *11*, 2451–2460.
44. 44. Marinelli, F.; Faraldo-Gómez, J.D. Ensemble-Biased Metadynamics: A Molecular Simulation Method to Sample Experimental Distributions. *Biophys. J.* **2015**, *108*, 2779–2782.
45. 45. Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B.L.; Grubmüller, H.; MacKerell, A.D. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **2016**, *14*, 71–73.
46. 46. Marsh, J.A.; Singh, V.K.; Jia, Z.; Forman-Kay, J.D. Sensitivity of secondary structure propensities to sequence differences between α- and γ-synuclein: Implications for fibrillation. *Protein Sci.* **2006**, *15*, 2795–2804.
47. 47. Camilloni, C.; De Simone, A.; Vranken, W.F.; Vendruscolo, M. Determination of Secondary Structure Populations in Disordered States of Proteins Using Nuclear Magnetic Resonance Chemical Shifts. *Biochemistry* **2012**, *51*, 2224–2231.
48. 48. Shen, Y.; Bax, A. SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR* **2010**, *48*, 13–22.
49. 49. Li, D.-W.; Brüschweiler, R. PPM: a side-chain and backbone chemical shift predictor for the assessment of protein conformational ensembles. *J. Biomol. NMR* **2012**, *54*, 257–65.
50. 50. Leslie Kish *Survey Sampling*; John Wiley & Sons, Inc.: New York, 1965;
51. 51. Pitera, J.W.; Chodera, J.D. On the Use of Experimental Observations to Bias Simulated Ensembles. *J. Chem. Theory Comput.* **2012**, *8*, 3445–3451.
52. 52. Weare, J. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. **2013**, *084107*.
53. 53. Beauchamp,Kyle A.; Pande, V.S.; Das, R. Bayesian energy landscape tilting: towards concordant models of molecular ensembles. *Biophys. J.* **2014**, *106*, 1381–90.
54. 54. Sanchez-Martinez, M.; Crehuet, R. Application of the maximum entropy principle to determine ensembles of intrinsically disordered proteins from residual dipolar couplings. *Phys. Chem. Chem. Phys. PCCP* **2014**, *16*, 26030–9.
55. 55. Jones, E.; Oliphant, T.; Peterson, P.; others *SciPy: Open Source Scientific Tools for Python*; 2001;
56. 56. Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.
57. 57. McGibbon, R.T.; Beauchamp, K.A.; Harrigan, M.P.; Klein, C.; Swails, J.M.; Hernández, C.X.; Schwantes, C.R.; Wang, L.P.; Lane, T.J.; Pande, V.S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*.
58. 58. Bonomi, M.; Camilloni, C. Integrative structural and dynamical biology with PLUMED-ISDB. *Bioinformatics* **2017**, *33*, 3999–4000.

59. 59. Marsh, J.A.; Forman-Kay, J.D. Ensemble modeling of protein disordered states: Experimental restraint contributions and validation. *Proteins* **2012**, *80*, 556–572.
60. 60. Richter, B.; Gsponer, J.; Várnai, P.; Salvatella, X.; Vendruscolo, M. The MUMO (minimal under-restraining minimal over-restraining) method for the determination of native state ensembles of proteins. *J. Biomol. NMR* **2007**, *37*, 117–35.
61. 61. Rangan, R.; Bonomi, M.; Heller, G.T.; Cesari, A.; Bussi, G.; Vendruscolo, M. Determination of Structural Ensembles of Proteins: Restraining vs Reweighting. *J. Chem. Theory Comput.* **2018**, *14*, 6632–6641.
62. 62. Olsson, S.; Vögeli, B.R.; Cavalli, A.; Boomsma, W.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K.; Hamelryck, T. Probabilistic Determination of Native State Ensembles of Proteins. *J. Chem. Theory Comput.* **2014**, *10*, 3484–3491.
63. 63. Hansen, P.; O'Leary, D. The Use of the L-Curve in the Regularization of Discrete Ill-Posed Problems. *SIAM J. Sci. Comput.* **1993**, *14*, 1487–1503.
64. 64. Wald, A.; Wolfowitz, J. On a test whether two samples are from the same population. *Ann Math Stat.* *11*, 147–162.
65. 65. Cesari, A.; Bottaro, S.; Lindorff-Larsen, K.; Banáš, P.; Šponer, J.; Bussi, G. Fitting Corrections to an RNA Force Field Using Experimental Data. *J. Chem. Theory Comput.* **2019**, *15*, 3425–3431.
66. 66. Piana, S.; Lindorff-Larsen, K.; Dirks, R.M.; Salmon, J.K.; Dror, R.O.; Shaw, D.E. Evaluating the Effects of Cutoffs and Treatment of Long-range Electrostatics in Protein Folding Simulations. *PLoS ONE* **2012**, *7*, e39918.
67. 67. Tian, C.; Kasavajhala, K.; Belfon, K.; Raguette, L.; Huang, H.; Migues, A.; Bickel, J.; Wang, Y.; Pincay, J.; Wu, Q.; et al. ff19SB: Amino-Acid Specific Protein Backbone Parameters Trained Against Quantum Mechanics Energy Surfaces in Solution. **2019**.
68. 68. Boomsma, W.; Ferkinghoff-Borg, J.; Lindorff-Larsen, K. Combining Experiments and Simulations Using the Maximum Entropy Principle. *PLoS Comput. Biol.* **2014**, *10*, e1003406.
69. 69. Tamiola, K.; Acar, B.; Mulder, F. a a Sequence-specific random coil chemical shifts of intrinsically disordered proteins. *J. Am. Chem. Soc.* **2010**, *132*, 18000–18003.
70. 70. De Simone, A.; Cavalli, A.; Hsu, S.-T.D.; Vranken, W.; Vendruscolo, M. Accurate random coil chemical shifts from an analysis of loop regions in native states of proteins. *J. Am. Chem. Soc.* **2009**, *131*, 16332–16333.
71. 71. Nielsen, J.T.; Mulder, F.A.A. POTENCI: prediction of temperature, neighbor and pH-corrected chemical shifts for intrinsically disordered proteins. *J. Biomol. NMR* **2018**, *70*, 141–165.
72. 72. Kjaergaard, M.; Brander, S.; Poulsen, F.M. Random coil chemical shift for intrinsically disordered proteins: effects of temperature and pH. *J. Biomol. NMR* **2011**, *49*, 139–49.
73. 73. Kjaergaard, M.; Poulsen, F.M. Sequence correction of random coil chemical shifts: correlation between neighbor correction factors and changes in the Ramachandran distribution. *J. Biomol. NMR* **2011**, *50*, 157–165.
74. 74. Modig, K.; Jürgensen, V.W.; Lindorff-Larsen, K.; Fieber, W.; Bohr, H.G.; Poulsen, F.M. Detection of initiation sites in protein folding of the four helix bundle ACBP by chemical shift analysis. *FEBS Lett.* **2007**, *581*, 4965–4971.
75. 75. Haxholm, G.W.; Nikolajsen, L.F.; Olsen, J.G.; Fredsted, J.; Larsen, F.H.; Goffin, V.; Pedersen, S.F.; Brooks, A.J.; Waters, M.J.; Kragelund, B.B. Intrinsically disordered cytoplasmic domains of two cytokine receptors mediate conserved interactions with membranes. *Biochem. J.* **2015**, *468*, 495–506.
76.

© 2019 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).