

A peer-reviewed version of this preprint was published in PeerJ on 24 March 2015.

[View the peer-reviewed version](https://peerj.com/articles/861) (peerj.com/articles/861), which is the preferred citable publication unless you specifically need to cite this preprint.

Bratholm LA, Christensen AS, Hamelryck T, Jensen JH. 2015. Bayesian inference of protein structure from chemical shift data. PeerJ 3:e861
<https://doi.org/10.7717/peerj.861>

Bayesian Inference of Protein Structure from Chemical Shift Data

Lars A. Bratholm¹, Anders S. Christensen^{1,3}, Thomas Hamelryck², and Jan H. Jensen¹

¹Department of Chemistry, University of Copenhagen, Copenhagen, Denmark

²Department of Biology, University of Copenhagen, Copenhagen, Denmark

³Current Affiliation: Department of Chemistry, University of Wisconsin-Madison, Madison, WI, USA

ABSTRACT

Protein chemical shifts are routinely used to augment molecular mechanics force fields in protein structure simulations, with weights of the chemical shift restraints determined empirically. These weights, however, might not be an optimal descriptor of a given protein structure and predictive model, and a bias is introduced which might result in incorrect structures. In the inferential structure determination framework, both the unknown structure and the disagreement between experimental and back-calculated data are formulated as a joint probability distribution, thus utilizing the full information content of the data. Here, we present the formulation of such a probability distribution where the error in chemical shift prediction is described by either a Gaussian or Cauchy distribution. The methodology is demonstrated and compared to a set of empirically weighted potentials through Markov chain Monte Carlo simulations of three small proteins (ENHD, Protein G and the SMN Tudor Domain) using the PROFASI force field and the chemical shift predictor CamShift. Using a clustering-criterion for identifying the best structure, together with the addition of a solvent exposure scoring term, the simulations suggests that sampling both the structure and the uncertainties in chemical shift prediction leads more accurate structures compared to conventional methods using empirical determined weights. The Cauchy distribution, using either sampled uncertainties or predetermined weights, did, however, result in overall better convergence to the native fold, suggesting that both types of distribution might be useful in different aspects of the protein structure prediction.

Keywords: Chemical shifts, Markov chain Monte Carlo, NMR, Probabilistic models, Protein structure

1 INTRODUCTION

2 Protein structures can today routinely be simulated by methods such as molecular dynamics or Monte
3 Carlo simulations, using molecular mechanics force fields (Shaw et al., 2010; Karplus and McCammon,
4 2002; Snow et al., 2002). However, this is not always a feasible method to determine a protein structure
5 by itself. To elucidate the native protein structure efficiently, the force field energy can be augmented
6 by restraints obtained from experiments. This immediately raises the question, how can this be done
7 rigorously and efficiently? One pragmatic approach to this problem is to define a hybrid energy using a
8 penalty function, which describes the agreement between experimental data and data calculated from a
9 proposed protein structure, together with a physical energy (such as from a molecular mechanics force
10 field) (Jack and Levitt, 1978). An optimal structure in this approach could then be determined for example
11 by minimizing the hybrid energy function

$$E_{\text{hybrid}} = w_{\text{data}} E_{\text{data}} + E_{\text{physical}}. \quad (1)$$

12 This approach, however, does not uniquely define neither the nature nor weight of E_{data} , and the resulting
13 protein structure will depend on the choices of these.

14 Chemical shifts have been combined with physical energies in a multitude of ways, e.g. using weighted
15 RMSD values or various types of harmonic constraints. Vendruscolo and co-workers implemented a
16 'square-well soft harmonic potential', with corresponding gradients, and were able to run a chemical shifts

17 biased MD simulation where they successfully refined slightly denatured protein structures to a C_{α} -RMSD
18 of down to 0.84 Å from the corresponding crystal structures (Robustelli et al., 2010). The groups of Bax
19 and Baker added the chi-square agreement between SPARTA (Shen and Bax, 2007) predicted chemical
20 shift values and experimental chemical shifts with an empirical weight of 0.25 to the ROSETTA all-atom
21 energy (Shen et al., 2008; Rohl et al., 2004). The CHESHIRE approach (Cavalli et al., 2007) utilizes the
22 experimental chemical shifts to predict secondary structure and backbone dihedral angles. These in turn
23 are used to score molecular fragments from a database of known structures together with the chi-square
24 agreement between the measured chemical shifts and the chemical shifts of the fragment in the database.
25 A different approach was used by Meiler and Baker (Meiler and Baker, 2003), where the contribution of
26 the experimental chemical shifts were set relative to 1 or 0 depending on whether or not the difference to
27 the PROSHIFT prediction (Meiler, 2003) exceeded a maximum tolerance. The reasoning for not using
28 a quadratic potential was that the experimental NMR data was automatically assigned and a quadratic
29 potential is more sensitive to assignment errors. In all cases the parameters, shape and weights of E_{data}
30 had to be carefully tweaked by hand, and it is obviously not clear how to choose optimal parameters.

31 The inferential structure determination (ISD) principles introduced by Rieping, Habeck and Nilges
32 (Rieping et al., 2005) defines a Bayesian formulation of Eqn. 1, which has previously been used to
33 determine protein structures based on NOE (Habeck et al., 2006; Olsson et al., 2011) and RDC restraints
34 (Habeck et al., 2008). In the following section the equations of an ISD approach for combining the
35 knowledge of experimental chemical shifts with a physical energy are presented.

36 THEORY

37 In the ISD approach we seek the probability distribution of the structure \mathbf{X} and a set of uncertainties θ ,
38 correlating experimental and predicted chemical shifts, given a set of experimentally measured chemical
39 shifts \mathbf{d} , i.e. the probability $p(\mathbf{X}, \theta | \mathbf{d})$. Using Bayes' theorem, this probability can be factored out as

$$p(\mathbf{X}, \theta | \mathbf{d}) = \frac{p(\mathbf{d} | \mathbf{X}, \theta) p(\mathbf{X}, \theta)}{p(\mathbf{d})}. \quad (2)$$

40 $p(\mathbf{d})$ merely serves as a normalization constant, which we need not evaluate.

41 We're making the basic assumption, that the deviation between predicted and experimental chemical
42 shifts, given as

$$\Delta\delta_i = \delta_{\mathbf{X},i} - \delta_{\text{exp},i} \quad (3)$$

43 approximately follows some distribution with a variance uniquely defined by the type of nuclei (C_{α} ,
44 C_{β} etc.). The relevant equations for a Gaussian distribution and a Cauchy distribution (a Student's
45 t-distribution with one degree of freedom), respectively, are presented in the next sections.

46 Gaussian distribution

47 According to the principle of maximum entropy (Jaynes, 1957), the least biasing model for the error of
48 the chemical shift prediction is a Gaussian distribution with standard deviations σ_j (where j indicates
49 the nuclei type). The standard deviations are effectively describing the weight of the experimental data.
50 Assuming that each measured experimental chemical shift $\delta_{\text{exp},i}$ is conditional independent given the
51 structure, the likelihood $p(\mathbf{d} | \mathbf{X}, \theta)$ is obtained as the product of the individual probabilities of all measured
52 chemical shifts. With i iterating over all n_j measured chemical shifts of nuclei type j , this takes the form

53 of:

$$\begin{aligned}
 p(\mathbf{d} | \mathbf{X}, \theta) &= \prod_j \prod_{i=1}^{n_j} p(\delta_{\text{exp},ij} | \delta_{\mathbf{X},ij}, \sigma_j) \\
 &= \prod_j \prod_{i=1}^{n_j} \frac{1}{\sigma_j \sqrt{2\pi}} \exp\left(-\frac{\Delta\delta_{ij}^2}{2\sigma_j^2}\right) \\
 &= \prod_j \left(\frac{1}{\sigma_j \sqrt{2\pi}}\right)^{n_j} \exp\left(-\frac{\chi_j^2}{2\sigma_j^2}\right), \tag{4}
 \end{aligned}$$

54 where $\chi_j^2 = \sum_i^{n_j} \Delta\delta_{ij}^2$. The structure, \mathbf{X} , and the uncertainties in the model, θ , are assumed independent
 55 and $p(\mathbf{X}, \theta)$ can be expanded into

$$p(\mathbf{X}, \theta) = p(\mathbf{X}) p(\theta) = p(\mathbf{X}) \prod_j p(\sigma_j). \tag{5}$$

56 The prior probability for the protein structure can be expressed by the Boltzmann distribution, that is:

$$p(\mathbf{X}) = \frac{1}{Z(T)} \exp\left(-\frac{E(\mathbf{X})}{k_B T}\right), \tag{6}$$

57 where the physical energy $E(\mathbf{X})$ could for example be approximated using a molecular mechanics force
 58 field. Note that in this case, the partition function $Z(T)$ is a normalization constant and evaluation of
 59 this is not necessary. We have little prior knowledge about σ_j other than that it is a scale parameter.
 60 An uninformative choice of prior distribution is the Jeffreys prior (Jeffreys, 1946), which in this case is
 61 simply:

$$p(\sigma_j) = \sigma_j^{-1}. \tag{7}$$

62 Combining these expressions, $p(\mathbf{X}, \theta | \mathbf{d})$ is thus proportional to

$$\begin{aligned}
 p(\mathbf{X}, \theta | \mathbf{d}) &\propto p(\mathbf{d} | \mathbf{X}, \theta) p(\mathbf{X}) p(\theta) \\
 &\propto \prod_j \left[\sigma_j^{-n_j-1} \exp\left(-\frac{\chi_j^2}{2\sigma_j^2}\right) \right] \exp\left(-\frac{E(\mathbf{X})}{k_B T}\right). \tag{8}
 \end{aligned}$$

63 The resemblance to a hybrid energy such as in Eqn. 1 is obtained by (neglecting all constant terms):

$$\begin{aligned}
 E_{\text{hybrid}}(\mathbf{X}, \theta) &= -k_B T \ln(p(\mathbf{X}, \theta | \mathbf{d})) \\
 &= k_B T \sum_j \left((n_j + 1) \ln(\sigma_j) + \frac{\chi_j^2}{2\sigma_j^2} \right) + E(\mathbf{X}). \tag{9}
 \end{aligned}$$

64 This energy as a function of σ_j is depicted in Fig. 1a.

65

66 **Conjugate prior.** As discussed below, use of the Jeffrey's prior and the Gaussian model with the empirical
 67 chemical shift predictor CamShift leads to numerical problems. The problems arises if χ_j^2 converges to
 68 zero, which leads to $\sigma_j \rightarrow 0$. This can be seen from the maximum a posteriori estimator (MAP) of σ_j^2 :

$$\sigma_{j,\text{MAP}}^2 = \frac{\chi_j^2}{n_j + 1}. \tag{10}$$

69 We found that these problems can be avoided by using a weakly informative prior. The conjugate
 70 prior for the variance of the Gaussian distribution (σ_j^2), when the mean is known, can be given by an
 71 Inverse-Gamma distribution:

$$p(\sigma_j^2 | \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} (\sigma_j^2)^{-\alpha-1} \exp\left(-\frac{\beta}{\sigma_j^2}\right). \tag{11}$$

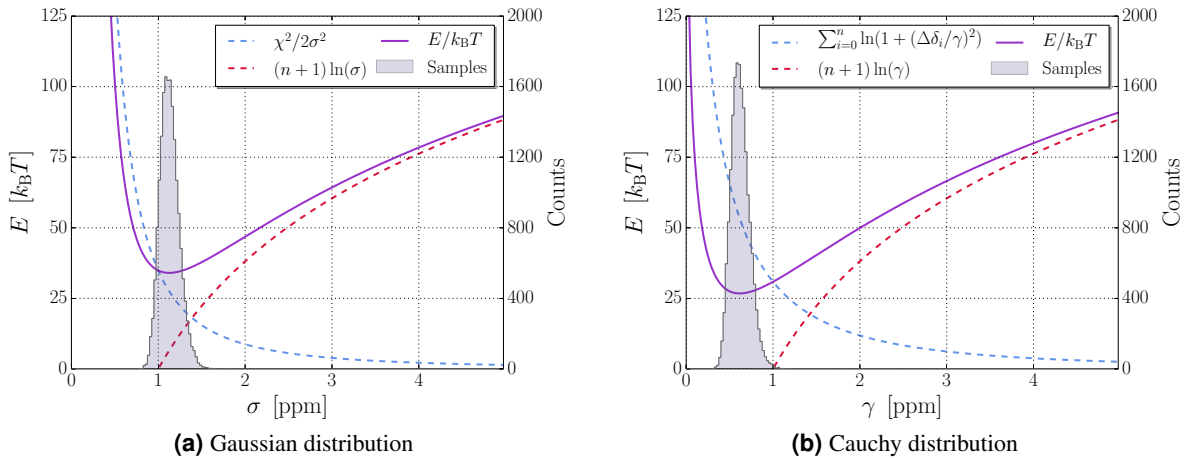


Figure 1. Sampling of σ and γ , using Jeffrey's priors, for C_α -chemical shifts of Protein G. $n_{C_\alpha} = 54$ and $\chi_{C_\alpha}^2 = 69.7 \text{ ppm}^2$.

72 $p(\mathbf{X}, \theta | \mathbf{d})$ is thus proportional to

$$\begin{aligned}
 p(\mathbf{X}, \theta | \mathbf{d}) &\propto p(\mathbf{d} | \mathbf{X}, \theta) p(\mathbf{X}) p(\theta) \\
 &\propto \prod_j \left[\sigma_j^{-n_j - 2\alpha - 2} \exp\left(-\frac{2\beta + \chi_j^2}{2\sigma_j^2}\right) \right] \exp\left(-\frac{E(\mathbf{X})}{k_B T}\right).
 \end{aligned} \tag{12}$$

73 In contrast to Eqn 10, the maximum a posteriori estimator of σ_j^2 does not equal zero in the limit of $\chi_j^2 \rightarrow 0$
 74 with a non-zero choice of β :

$$\sigma_{j,\text{MAP}}^2 = \frac{2\beta + \chi_j^2(\mathbf{X})}{2\alpha + 2 + N_j} \tag{13}$$

75 In all the simulations where σ_j was sampled we use Eqn 12 and $\alpha = \beta = 0.001$ (Gelman, 2006) unless
 76 stated otherwise.

77

78 **Marginal likelihood.** Alternatively one can use the marginal likelihood where σ_j is integrated out:

$$\begin{aligned}
 p(\mathbf{d} | \mathbf{X}) &= \prod_j \int_0^\infty p(\mathbf{d} | \mathbf{X}, \sigma_j) p(\sigma_j) d\sigma_j \\
 &\propto \prod_j (\chi_j^2)^{\frac{-n_j}{2}}
 \end{aligned} \tag{14}$$

79 This results in a hybrid energy of the form:

$$\begin{aligned}
 E_{\text{hybrid}}(\mathbf{X}) &= -k_B T \ln(p(\mathbf{X} | \mathbf{d})) \\
 &= k_B T \sum_j \left(\frac{n_j}{2} \ln(\chi_j^2) \right) + E(\mathbf{X})
 \end{aligned} \tag{15}$$

80 Cauchy distribution

81 The Cauchy and Gaussian distribution are both special cases of the Student's t-distribution, with degrees of
 82 freedom $\nu = 1$ and $\nu = \infty$ respectively. Compared to the Gaussian distribution, the Cauchy distribution has
 83 much heavier tails meaning that it will be less penalizing of single predictions far from the experimental
 84 values.

85 $p(\mathbf{d}|\mathbf{X}, \theta)$ is again obtained as the product of the individual probabilities of all measured chemical
 86 shifts, with scale parameters γ_j (equivalent to σ_j of the Gaussian distribution):

$$\begin{aligned}
 p(\mathbf{d}|\mathbf{X}, \theta) &= \prod_j \prod_{i=1}^{n_j} p(\delta_{\text{exp},ij} | \delta_{\mathbf{X},ij}, \gamma_j) \\
 &= \prod_j \left\{ (\pi\gamma_j)^{-n_j} \prod_{i=1}^{n_j} \left[1 + \left(\frac{\Delta\delta_{ij}}{\gamma_j} \right)^2 \right]^{-1} \right\}
 \end{aligned}
 \tag{16}$$

87 Note that the Cauchy distribution does not reduce into an expression that depends on the χ_j^2 differences
 88 (in contrast to the Gaussian). The Jeffreys prior is the same as for the Gaussian distribution:

$$p(\gamma_j) = \gamma_j^{-1}. \tag{17}$$

89 $p(\mathbf{X}, \theta | \mathbf{d})$ is thus proportional to

$$p(\mathbf{X}, \theta | \mathbf{d}) \propto \prod_j \left\{ \gamma_j^{-(n_j+1)} \prod_{i=1}^{n_j} \left[1 + \left(\frac{\Delta\delta_{ij}}{\gamma_j} \right)^2 \right]^{-1} \right\} \exp\left(-\frac{E(\mathbf{X})}{k_B T}\right) \tag{18}$$

90 The resemblance to a hybrid energy such as in Eqn. 1 is obtained by (neglecting all constant terms):

$$\begin{aligned}
 E_{\text{hybrid}}(\mathbf{X}, \theta) &= -k_B T \ln(p(\mathbf{X}, \theta | \mathbf{d})) \\
 &= k_B T \sum_j \left\{ \left((n_j + 1) \ln(\gamma_j) + \sum_{i=1}^{n_j} \ln \left[1 + \left(\frac{\Delta\delta_{ij}}{\gamma_j} \right)^2 \right] \right) \right\} + E(\mathbf{X})
 \end{aligned}
 \tag{19}$$

91 METHODOLOGY

92 Computational methodology

93 Markov chain Monte Carlo simulations were carried out with PHAISTOS v1.0 (Boomsma et al., 2013) us-
 94 ing either the multicanonical generalized ensemble via MUNINN (Feringhoff-Borg, 2002) or Metropolis-
 95 Hastings (Metropolis et al., 1953). Chemical shift predictions were performed with an implementation
 96 of CamShift (Kohlhoff et al., 2009) and the physical energy was approximated using the computational
 97 efficient PROFASI force field (Irbäck and Mohanty, 2006). The conformational degrees of freedom
 98 explored in the simulations were restricted to the backbone and side-chain dihedral angles (ϕ, ψ, χ) as
 99 well as the backbone bond angles. Backbone moves had torsion and bond angles biased by CS-Torus
 100 (Boomsma et al., 2014) and Engh-Huber statistics (Engh and Huber, 1991) respectively, which both
 101 introduces an implicit energy. Chemical shifts were only utilized by CS-Torus for biased sampling
 102 in reference simulations where no CamShift energy term was used. The simulations were performed
 103 on AMD Opteron 2.1 GHz CPU's at ~12M steps/day or on Intel Xeon 3.07 GHz CPU's at ~18M steps/day.

104
 105 **Convergence simulations.** The Protein G convergence simulations were initialized from the experi-
 106 mental structure (PDB-id: 2OED). The simulations were run for 10M MC steps at 300K using Metropolis-
 107 Hastings. The physical move set was comprised of 50% local, uniform single side chain moves, 25%
 108 CRISP local moves (Bottaro et al., 2012) and 25% semilocal biased Gaussian step (BGS) backbone moves
 109 (Favrin et al., 2001).

110
 111 **Structure determination simulations.** The structure determination simulations were each run on 32
 112 threads for 100M iterations. The temperature range explored with MUNINN were set to 273K - 500K. The
 113 physical move set was comprised of 50% local, uniform single side chain moves, 40% CRISP backbone
 114 moves and 10% backbone-DBN pivot moves (Boomsma et al., 2008). In the simulations where the
 115 uncertainties were dynamically adjusted, an extra 10M Monte Carlo steps were added which sampled a
 116 change in σ_j or γ_j as described below. Note that these moves are essentially computationally costless,

117 since neither chemical shifts or force field energy terms need be recomputed.

118

119 **Clustering of sampled structures.** To make clustering feasible for the large amount of structures
120 generated (320,000 structures for each combination of potential and protein), the sampled structures were
121 converted to GIT vectors (Røgen and Fain, 2003) with PHAISTOS. The structures from each individual
122 thread were subsequently divided into sets of 15 clusters with the Pleiades module (Harder et al., 2012)
123 using K-means clustering (Lloyd, 1982). The choice of using 15 clusters is based on the suggestion
124 of the Pleiades authors of creating 10 - 20 clusters. Since the clustering process is stochastic it was
125 performed 10 times for each thread and the optimal clustering according to the sum of squared errors
126 were used for further analysis. From each of these clusters, a subset consisting of the 100 structures
127 closest to the cluster centroid were selected for energy and RMSD evaluation and the median energy
128 structures were chosen as cluster representatives. The GIT vectors can be created as output observables
129 directly from the simulations, but in this case they were created from the simulation trajectories using the
130 pdb2git application in PHAISTOS with the program GNU Parallel (Tange, 2011) used to parallelize the
131 jobs. Re-weighting from the generalized ensemble to approximate the canonical ensemble were done
132 automatically with Pleiades using the weighted k-means option.

133 Monte Carlo move in uncertainty parameter space

134 The ξ -move which re-samples the value of the uncertainties (i.e. σ or γ) was constructed by multiplying
135 the previous value of ξ by a sampled constant centered around 1. Detailed balance is maintained by
136 proposing a small change, $\xi \rightarrow \xi'$, by:

$$\xi' = \xi \cdot \exp(\text{rnom}(\sigma_\mu)), \quad (20)$$

137 where $\text{rnom}(\sigma_\mu)$ is a random number from a normal distribution with zero mean and standard deviation
138 σ_μ . A value of $\sigma_\mu = 0.1$ was found to yield a rapid and stable convergence for both the Gaussian and the
139 Cauchy distribution.

140 Issues with CamShift prediction

141 It was observed that CamShift predictions of C_β chemical shifts for Isoleucine were consistently off
142 by 3 - 8 ppm. This was observed using both the CamShift implementation in PHAISTOS as well as
143 with the standalone predictor. CamShift was trained on high quality X-ray structures where missing
144 Hydrogens were added in accordance with the CHARMM22 topology file (Brooks et al., 2009). Letting
145 the CamShift program optimize Hydrogen placement before prediction brought the accuracy of predicted
146 Isoleucine C_β chemical shifts in range with the prediction for the remaining amino-acids. For reference,
147 the RMSD for C_β chemical shift prediction of all amino-acids of a Chymotrypsin Inhibitor-II protein
148 (CI2) structure were found to be 1.90 ppm including predictions for Isoleucine and 1.25 ppm if these
149 predictions were excluded. As bond lengths and side-chain bond angles are not degrees of freedom in
150 the simulations performed with PHAISTOS, the β -Hydrogen placements relative to the C_β atoms are
151 constant and prediction for Isoleucine C_β chemical shifts was disabled.

152 RESULTS AND DISCUSSION

153 Problems with Gaussian weighting scheme when using a Jeffreys prior

154 Attempts to use predicted chemical shifts from CamShift while sampling σ using a Gaussian model
155 (Eqn. 9) initially proved unsuccessful. Using any structure (compact or unfolded) as starting point for the
156 Monte Carlo simulation, it was often observed that the χ^2 agreement between predicted and experimental
157 chemical shifts would converge to zero after only a few million iterations. Naturally this leads to $\sigma \rightarrow 0$,
158 which in turn essentially freezes the structure in the simulation, since any MC move that causes the
159 slightest increase in chi-square will result in an enormous change in energy. If several types of chemical
160 shifts were included in the simulation (possible chemical shift types from CamShift are H_α , C_α , H, N, C
161 and C_β) the χ^2 for one (random) of the included types would quickly converge to zero. One suspected
162 reason was that the prior distribution was not well described by the more coarse grained PROFASI force

163 field. CamShift calculations were therefore redone using the OPLS-AA/L force field (Kaminski and
164 Friesner, 2001). This, however, led to identical results.

165 On this basis we conclude that the problem is due to CamShift (and most likely other choices of
166 predictors) being able to make relatively large changes in prediction, from a small perturbation in the
167 structure. Combined with sampling of σ , this can drive the simulation into an energy minimum with
168 essentially zero error in the chemical shift prediction, even though the structure may or may not be
169 anything like the native structure. We found the Cauchy distribution to be less sensitive to divergence of
170 the scale parameter and to perform better as an uninformative model in our case. As an alternative to the
171 Jeffreys prior, a weakly informative conjugate prior for the Gaussian model did not show these sampling
172 issues.

173 Convergence of scale parameters

174 The convergence of the scale parameters for the Gaussian and Cauchy distributions (σ and γ respectively),
175 with chemical shifts predictions by CamShift (Kohlhoff et al., 2009), were explored by starting a simulation
176 with PHAISTOS (Boomsma et al., 2013) from the native structure of Protein G (PDB: 2OED (Ulmer et al.,
177 2003)). Experimental chemical shifts were obtained from Ref-DB (Zhang et al., 2003) (RefDB:2575
178 (Orban et al., 1992)). For each model a 10^7 MC step simulation was performed keeping the structure fixed,
179 only sampling uncertainties (frozen), and a simulation where the atomic coordinates (\mathbf{X}) was sampled as
180 well (free). Tables 1 and 2 shows the mean of the sampled parameters from the last 10^6 steps together
181 with the maximum likelihood values obtained from the CamShift training set for reference.

Table 1. Maximum likelihood estimates of σ (or root-mean-square deviation (RMSD)) obtained from the CamShift training set, compared to means extracted from a 10^7 MC step simulation using the Gaussian model (see text). Shown values are in units of ppm.

	C_α	H_α	N	H	C	C_β
CamShift training set	1.22	0.26	2.78	0.56	1.12	1.19
Frozen simulation ^a	1.13	0.26	3.53	0.52	1.06	1.21
Free simulation ^a	1.03	0.20	2.92	0.46	1.16	1.23

^a Estimated over the last 10^6 MC steps.

Table 2. Maximum likelihood estimates of γ obtained from the CamShift training set, compared to means extracted from a 10^7 MC step simulation using the Cauchy model (see text). Shown values are in units of ppm.

	C_α	H_α	N	H	C	C_β
CamShift training set	0.70	0.19	1.87	0.31	0.74	0.77
Frozen simulation ^a	0.62	0.17	1.90	0.32	0.64	0.69
Free simulation ^a	0.43	0.05	1.57	0.25	0.67	0.55

^a Estimated over the last 10^6 MC steps.

182 Using a Gaussian distribution, the parameters in the 'frozen' simulation all converged within 0.1 ppm
183 to the reported values from the CamShift training set, with the exception of the N nuclei which deviated
184 by 0.75 ppm. The RMSDs presented in Table 1 for the CamShift training set were based on predictions
185 on 7 proteins, and using a larger data set of 28 proteins, the average RMSD for the N nucleus increased
186 from 2.78 ppm to 3.01 ppm (Kohlhoff et al., 2009). Thus the slightly higher mean for N seems reasonable.
187 Allowing the structure and weight parameters to be sampled simultaneously in the 'free' simulation
188 overall lowered the RMSD of the prediction as expected, since the accepted structures in the Monte Carlo
189 simulation will be biased by the correlation of predicted and experimental chemical shifts. However the
190 RMSD increased moderately for the C nucleus and slightly for C_β , indicating that the chemical shift
191 prediction of C and C_β are less sensitive to changes in local structure than the four other nuclei.

192 In the simulations using a Cauchy distribution, the 'frozen' values were seen to be similar to the
 193 CamShift data set (within 0.1 ppm). When physical moves were introduced in the 'free' simulation, the
 194 sampled parameters were again found to be lowered, but remained within 0.3 ppm. Surprisingly γ for
 195 H_α went from 0.17 ppm to 0.05 ppm with similar values found when repeating the simulation. The χ^2
 196 error in the prediction of H_α chemical shifts were similar to that obtained with the Gaussian potential,
 197 indicating that the error in prediction for H_α atoms had several outliers. Since the Cauchy distribution
 198 is less sensitive to outlier values, these will have a lesser effect on the sampled parameters than for the
 199 Gaussian.

200 Comparison of weighting schemes in structure determination

201 A series of simulations starting from an unfolded state were performed on ENHD (PDB: 1ENH (Clarke
 202 et al., 1994), BMRB:15536 (Religa, 2008)), Protein G and the SMN Tudor Domain (PDB: 1MHN
 203 (Sprangers et al., 2003), RefDB:4899 (Selenko et al., 2001)) to compare how different weighting schemes
 204 performed for structure determination. The probabilistic schemes used included three Gaussian models:
 205 One using the maximum likelihood estimates of σ from the CamShift training set (Gaussian / fixed). One
 206 where the values of σ were sampled (Gaussian / sampled) and one using the marginalized distribution
 207 (Gaussian / marginalized). Similarly two Cauchy models were tested: One using maximum likelihood
 208 values for γ from the CamShift training set (Cauchy / fixed), and one where the values for γ were sampled
 209 (Cauchy / sampled). As reference, the square well potential of Robustelli et. al., which was made
 210 specifically for refinement with the CamShift model, were included in the simulations with different
 211 weights (Square well / $\alpha = 1$, Square well / $\alpha = 5$) (Robustelli et al., 2010).

212 In all simulations, the generative predictive model CS-Torus (Boomsma et al., 2014) was used to
 213 sample backbone dihedral angles from a distribution biased by the amino-acid sequence. Chemical shifts
 214 can provide local information to the CS-Torus model to further improve the biased sampling, but this was
 215 not utilized in any simulations using CamShift predictions. Although including chemical shifts in the
 216 sampling would most likely improve the simulation results, we chose to keep the CamShift energy terms
 217 as the only bias from the experimental chemical shifts. To display the effect of using a non-local chemical
 218 shift predictor like CamShift instead of relying on local information alone in the sampling, simulations
 219 using chemical shifts in the CS-Torus model, rather than with CamShift prediction, were run as well.

Table 3. Different weighting schemes used in the protein folding simulations. In the columns to the left, the number of threads, out of a total of 32, sampling structures below 2 and 4 Å C_α -RMSD respectively to the reference structure is shown. The sampled structures from each thread were divided into clusters and representative structures for each cluster were selected as the structure median in PROFASI+CamShift energy, from the 100 structures closest to the cluster centroid. The C_α -RMSD in Å of the lowest-energy cluster representative is shown below in the columns to the right.

	Threads (out of 32) sampling below 2Å (left) and 4Å (right)						Lowest-energy RMSD (Å)		
	ENHD		Protein G		SMN		ENHD	Protein G	SMN
Gaussian / fixed	32	32	0	7	29	30	3.67	3.11	3.11
Gaussian / sampled	32	32	4	15	13	20	2.15	3.03	5.88
Gaussian / marginalized	32	32	1	16	7	14	4.24	2.72	6.06
Cauchy / fixed	32	32	9	25	15	21	1.94	1.15	2.58
Cauchy / sampled	32	32	13	24	11	16	1.87	2.82	5.51
Square well / $\alpha = 1^a$	19	22	2	12	14	18	2.29	3.14	3.71
Square well / $\alpha = 5^a$	32	32	0	1	1	5	3.82	5.83	1.91
CS-Torus ^b	4	27	8	25	0	0	19.2	3.01	8.33

^a Weights, α , of 1 and 5 were used by Robustelli et. al.

^b Lowest-energy cluster representatives for the CS-Torus simulations were selected from PROFASI energy alone.

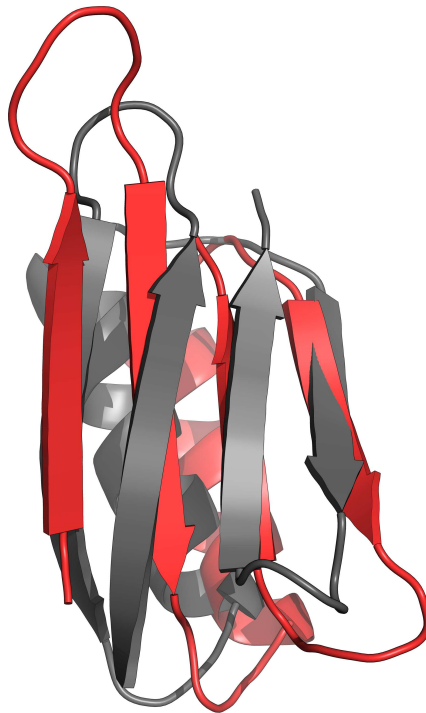


Figure 2. Crystal structure (grey) and local energy-minimum conformation (red) of Protein G. Figure made with PyMOL (Schrödinger, LLC, 2010)

220 32 folding simulations were run for each potential and protein for 100M MC steps using the PROFASI
 221 (Irbäck and Mohanty, 2006) force field and a CamShift energy term. For each set of simulations, the
 222 sampled structures from each thread were subsequently split into clusters as described in the Methodology
 223 section, and cluster representatives were selected as the structures median in energy, from the 100
 224 structures closest to the cluster centroid. Table 3 shows the number of threads sampling structures below
 225 2 and 4 Å C_{α} -RMSDs to the native structures as well as the RMSDs for the cluster representative with
 226 the lowest PROFASI+CamShift energy. The residue ranges used to calculate the RMSDs were 5-54 for
 227 ENHD, all residues for Protein G and 4-56 for the SMN Tudor Domain.

228 **Convergence of sampling**

229 The data in Table 3 shows that for certain potentials and proteins, several threads failed to sample near-
 230 native structures. For ENHD all potentials but the CS-Torus model and square well / $\alpha = 1$ potential
 231 sampled structures below 2 Å C_{α} -RMSD for all threads. While more than 20 threads sampled structures
 232 below 4 Å for both the CS-Torus and square well model, only 4 threads sampled structures below 2 Å for
 233 CS-Torus. For Protein G no threads for the Gaussian / fixed and square well / $\alpha = 5$ potentials sampled
 234 structures below 2 Å. The square well / $\alpha = 1$, Gaussian / marginalized and Gaussian / sampled potentials
 235 only sampled these near-native states with a few threads, while the Cauchy potentials and the CS-Torus
 236 model showed the fewest sampling issues.

237 Looking closer at the threads never sampling structures close to native for Protein G, it is found that
 238 the majority of these never progressed past a local energy-minimum with an alternative conformation
 239 where two β -strands have interchanged position (Fig. 2). Taking the median structure of the most dense
 240 cluster as representative for each thread, 27 of these shows this incorrect fold for the Gaussian / fixed
 241 potential and 26 for the square well / $\alpha = 1$ potential. The Cauchy distributions shows the opposite trend
 242 with 25 correct folds for both potentials, while the structures from the Gaussian / sampled and Gaussian
 243 / marginalized simulations had 14 and 11 correctly folded respectively. For all of these potentials, the
 244 densest clusters of each thread have either this misfold or the correct structure. While the square well /
 245 $\alpha = 5$ potential seem to find completely incorrect structures, the CS-Torus simulations finds the correct
 246 overall fold in 20 threads. The remaining CS-Torus threads are partly unfolded and none of them have the

247 misfolded structure found in the simulations with CamShift energy terms. Finally for the SMN Tudor
248 Domain, the Gaussian / fixed model sampled structures below 2 Å for nearly all threads. The CS-Torus
249 model and square well / $\alpha = 5$ potential for 0 and 1 thread(s) respectively, while the remaining potentials
250 sampled below 2 Å for around a third of the threads.

251 Ideally the simulations with a given potential samples structures close to native consistently well for all
252 proteins, which was not the case for the Gaussian / fixed model, square well / $\alpha = 5$ potential, the CS-Torus
253 reference model and to a lesser extent the Gaussian / sampled model. The two Cauchy potentials was
254 most likely to sample low-RMSD structures across the three proteins. Due to limitations of the MUNINN
255 implementation in PHAISTOS at the time the simulations were run, the multicanonical generalized
256 ensembles from each thread can not be re-weighted to approximate a single canonical ensemble, and
257 clustering of structures must be done on a per-thread basis. Since cluster densities can't readily be
258 compared across threads, the structure clusters are evaluated from the force field and CamShift energy.

259 **Lowest-energy clusters**

260 Table 3 shows for each potential and protein the C_{α} -RMSDs to native for the lowest-energy structures
261 found by clustering. There is no clear consensus of which potentials results in the most accurate structures
262 overall based on the RMSD values. Visually (Fig. S1-6) all but CS-Torus has the correct fold for ENHD,
263 with the Gaussian / fixed, Gaussian / marginalized and square well / $\alpha = 5$ structures being less compact
264 than the crystal structure. For protein G only the square well / $\alpha = 5$ potential shows a slight misfold,
265 and the overall somewhat high RMSDs is again due to slightly less compact structures, as well as a small
266 displacement of beta-sheet positions for all but the CS-Torus and Cauchy / fixed models. Although the
267 misfold shown in Fig. 2 was prevalent in the simulations in many threads, none of the lowest-energy
268 structures have these interchanged β -strand positions. For the SMN Tudor Domain the difference in
269 RMSDs between the potentials is mainly due to the protein tails not being correctly placed in a compact
270 structure.

271 As mentioned above, the obtained structures from the lowest-energy clusters are in general less
272 compact than the crystal structures. This is a result of additional compactness terms being excluded in
273 the simulations such that the effect of using different potentials for modelling the discrepancy between
274 observed and predicted chemical shifts might be more clear. In nearly all of the simulations higher energy
275 clusters exists that have lower RMSDs to the native structure, suggesting that near-native structures are
276 sampled, but the compactness of the protein isn't properly described by the force field. Evaluating sampled
277 structures with energy terms not included in the Monte Carlo simulations is problematic, since the energy
278 can fluctuate greatly with small changes in local structure. However when entire clusters of structures are
279 evaluated this becomes less of a problem, especially when coarse grained energy terms is used in addition
280 to the energies obtained from the simulations. The half-sphere exposure mixture model (HSEMM),
281 implemented in PHAISTOS for modelling solvent exposure, is a variation of the multibody multinomial
282 model (MuMu) (Johansson and Hamelryck, 2013) with the environment of residue i described by four
283 features: The secondary structure according to CS-Torus, the backbone hydrogen bond network and
284 the half sphere exposure up and down measure (Hamelryck, 2005). For every cluster, the energy from
285 HSEMM was calculated and added to the total energy of the structures, with the hydrogen bond network
286 feature integrated out to enforce the coarse grained characteristics of the model.

287 The results are summarized in Table 4 and show that the lowest-energy clusters re-scored with
288 the solvent exposure term all have lower or similar RMSDs to the clusters evaluated with just the
289 PROFASI+CamShift energies. Sampling of the uncertainty when using the Gaussian distribution results in
290 the structures closest to native, with RMSDs below 1.5 Å for all three proteins. For the Cauchy distribution,
291 sampling the uncertainties does not seem to be an improvement over using predetermined weights, but
292 both approaches gives better structures overall than the remaining potentials. Furthermore it is clear that
293 the non-local information provided by the CamShift model greatly improves structure sampling, as shown
294 by the relatively poor performance of the simulations using only CS-Torus.

Table 4. C_α -RMSDs in Å of the lowest-energy cluster representative, when a solvent exposure energy term (HSEMM) is added to re-score the structures.

	Lowest-re-scored-energy RMSD		
	ENHD	Protein G	SMN
Gaussian / fixed	1.40	2.45	2.23
Gaussian / sampled	1.03	1.29	1.24
Gaussian / marginalized	1.11	1.00	3.81
Cauchy / fixed	1.40	1.16	1.55
Cauchy / sampled	1.86	0.86	2.50
Square well potential / $\alpha = 1^a$	1.15	1.37	3.05
Square well potential / $\alpha = 5^a$	0.96	4.35	1.91
CS-Torus ^b	3.88	1.57	9.18

^a Weights, α , of 1 and 5 were used by Robustelli et. al.

^b Lowest-energy cluster representatives for the CS-Torus simulations were selected from PROFASI+HSEMM energy alone.

CONCLUSION

We present a probabilistic method for biasing protein structure simulations with experimentally measured chemical shifts, based on the inferential structure determination formalism (ISD). (Rieping et al., 2005) In this formalism, the weighting of experimental data can be determined entirely by the data itself, the predictive model and the physical force field.

Simulations were performed on three small proteins (ENHD, Protein G and SMN Tudor Domain) for a Gaussian and Cauchy-based probability distribution, using the chemical shift predictor CamShift (Kohlhoff et al., 2009). The ISD-determined uncertainties were found to correspond well to the empirically determined uncertainties in the CamShift predictions. Furthermore sampling the uncertainties as part of the protein structure determination simulations, lead to improved accuracy of the predicted structures when a Gaussian potential was used. Using a Cauchy potential with either sampled or fixed uncertainties did, however, show overall better convergence to the native fold, suggesting that the simulations are less likely to get stuck in local minima with these potentials. Additionally the importance of capturing non-local information from experimental chemical shifts have been shown by comparing the use of the CamShift predictor to the local-only CS-Torus model.

REFERENCES

- Boomsma, W., Frellsen, J., Harder, T., Bottaro, S., Johansson, K. E., Tian, P., Stovgaard, K., Andreetta, C., Olsson, S., Valentin, J. B., Antonov, L. D., Christensen, A. S., Borg, M., Jensen, J. H., Lindorff-Larsen, K., Ferkinghoff-Borg, J., and Hamelryck, T. (2013). Phaistos: A framework for markov chain monte carlo simulation and inference of protein structure. *Journal of Computational Chemistry*, 34(19):1697–1705.
- Boomsma, W., Mardia, K., Taylor, C., Ferkinghoff-Borg, J., Krogh, A., and Hamelryck, T. (2008). A generative, probabilistic model of local protein structure. *Proc. Natl. Acad. Sci.*, 106(26):8932–8937.
- Boomsma, W., Tian, P., Frellsen, J., Ferkinghoff-Borg, J., Hamelryck, T., Lindorff-Larsen, K., and Vendruscolo, M. (2014). Equilibrium simulations of proteins using molecular fragment replacement and nmr chemical shifts. *Proceedings of the National Academy of Sciences*, 111(38):13852–13857.
- Bottaro, S., Boomsma, W., E. Johansson, K., Andreetta, C., Hamelryck, T., and Ferkinghoff-Borg, J. (2012). Subtle monte carlo updates in dense molecular systems. *Journal of Chemical Theory and Computation*, 8(2):695–702.
- Brooks, B. R., Brooks, C. L., Mackerell, A. D., Nilsson, L., Petrella, R. J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S., Caffisch, A., Caves, L., Cui, Q., Dinner, A. R., Feig, M., Fischer, S., Gao, J., Hodoscek, M., Im, W., Kuczera, K., Lazaridis, T., Ma, J., Ovchinnikov, V., Paci, E., Pastor, R. W.,

- 327 Post, C. B., Pu, J. Z., Schaefer, M., Tidor, B., Venable, R. M., Woodcock, H. L., Wu, X., Yang, W.,
328 York, D. M., and Karplus, M. (2009). Charmm: The biomolecular simulation program. *Journal of*
329 *Computational Chemistry*, 30(10):1545–1614.
- 330 Cavalli, A., Salvatella, X., Dobson, C. M., and Vendruscolo, M. (2007). Protein structure determination
331 from nmr chemical shifts. *Proceedings of the National Academy of Sciences*, 104(23):9615–9620.
- 332 Clarke, N. D., Kissinger, C. R., Desjarlais, J., Gilliland, G. L., and Pabo, C. O. (1994). Structural studies
333 of the engrailed homeodomain. *Protein Science*, 3(10):1779–1787.
- 334 Engh, R. A. and Huber, R. (1991). Accurate bond and angle parameters for X-ray protein structure
335 refinement. *Acta Crystallographica Section A*, 47(4):392–400.
- 336 Favrin, G., Irbäck, A., and Sjunnesson, F. (2001). Monte carlo update for chain molecules: Biased
337 gaussian steps in torsional space. *The Journal of Chemical Physics*, 114(18):8154–8158.
- 338 Ferkinghoff-Borg, J. (2002). Optimized monte carlo analysis for generalized ensembles. *The European*
339 *Physical Journal B - Condensed Matter and Complex Systems*, 29(3):481–484.
- 340 Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*,
341 1(3):15–533.
- 342 Habeck, M., Nilges, M., and Rieping, W. (2008). A unifying probabilistic framework for analyzing
343 residual dipolar couplings. *J. Biomol. NMR*, 40:135–144.
- 344 Habeck, M., Rieping, W., and Nilges, M. (2006). Weighting of experimental evidence in macromolecular
345 structure determination. *Proc. Natl. Acad. Sci.*, 103(6):1756–1761.
- 346 Hamelryck, T. (2005). An amino acid has two sides: A new 2d measure provides a different view of
347 solvent exposure. *Proteins: Structure, Function, and Bioinformatics*, 59(1):38–48.
- 348 Harder, T., Borg, M., Boomsma, W., Røgen, P., and Hamelryck, T. (2012). Fast large-scale clustering of
349 protein structures using gauss integrals. *Bioinformatics*, 28(4):510–515.
- 350 Irbäck, A. and Mohanty, S. (2006). Profasi: A monte carlo simulation package for protein folding and
351 aggregation. *Journal of Computational Chemistry*, 27(13):1548–1555.
- 352 Jack, A. and Levitt, M. (1978). Refinement of large structures by simultaneous minimization of energy
353 and *R* factor. *Acta Crystallographica Section A*, 34(6):931–935.
- 354 Jaynes, E. T. (1957). Information theory and statistical mechanics. *Phys. Rev.*, 106:620–630.
- 355 Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond.*
356 *A*, 186:453–461.
- 357 Johansson, K. E. and Hamelryck, T. (2013). A simple probabilistic model of multibody interactions in
358 proteins. *Proteins: Structure, Function, and Bioinformatics*, 81(8):1340–1350.
- 359 Kaminski, G. A. and Friesner, R. A. (2001). Evaluation and reparametrization of the OPLS-AA force
360 field for proteins via comparison with accurate quantum chemical calculations on peptides. *J. Phys.*
361 *Chem. B*, 105:6474–6487.
- 362 Karplus, M. and McCammon, J. A. (2002). Molecular dynamics simulations of biomolecules. *Nature*
363 *Structural Biology*, 9(9).
- 364 Kohlhoff, K. J., Robustelli, P., Cavalli, A., Salvatella, X., and Vendruscolo, M. (2009). Fast and accurate
365 predictions of protein NMR chemical shifts from interatomic distances. *J. Am. Chem. Soc.*, 131:13894–
366 13895.
- 367 Lloyd, S. (1982). Least squares quantization in pcm. *Information Theory, IEEE Transactions on*,
368 28(2):129–137.
- 369 Meiler, J. (2003). PROSHIFT: Protein chemical shift prediction using artificial neural networks. *J. Biomol.*
370 *NMR*, 26:25–37.
- 371 Meiler, J. and Baker, D. (2003). Rapid protein fold determination using unassigned nmr data. *Proceedings*
372 *of the National Academy of Sciences*, 100(26):15404–15409.
- 373 Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of
374 state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092.
- 375 Olsson, S., Boomsma, W., Frellesen, J., Bottaro, S., Harder, T., Ferkinghoff-Borg, J., and Hamelryck, T.
376 (2011). Generative probabilistic models extend the scope of inferential structure determination. *J.*
377 *Magn. Reson.*, 213:182–186.

378 Orban, J., Alexander, P., and Bryan, P. (1992). Sequence-specific proton nmr assignments and secondary
379 structure of the streptococcal protein g b2-domain. *Biochemistry*, 31(14):3604–3611.

380 Religa, T. (2008). Comparison of multiple crystal structures with nmr data for engrailed homeodomain.
381 *Journal of Biomolecular NMR*, 40(3):189–202.

382 Rieping, W., Habeck, M., and Nilges, M. (2005). Inferential structure determination. *Science*, 308:303–
383 306.

384 Robustelli, P., Kohlhoff, K., Cavalli, A., and Vendruscolo, M. (2010). Using nmr chemical shifts as
385 structural restraints in molecular dynamics simulations of proteins. *Structure*, 18:923–933.

386 Rohl, C. A., Strauss, C. E., Misura, K. M., and Baker, D. (2004). Protein structure prediction using
387 rosetta. In Brand, L. and Johnson, M. L., editors, *Numerical Computer Methods, Part D*, volume 383
388 of *Methods in Enzymology*, pages 66 – 93. Academic Press.

389 Røgen, P. and Fain, B. (2003). Automatic classification of protein structure by using gauss integrals.
390 *Proceedings of the National Academy of Sciences*, 100(1):119–124.

391 Schrödinger, LLC (2010). The PyMOL molecular graphics system, version 1.3r1.

392 Selenko, P., Sprangers, R., Stier, G., Bühler, D., Fischer, U., and Sattler, M. (2001). Smn tudor domain
393 structure and its interaction with the sm proteins. *Nature Structural Biology*, 8(1):27.

394 Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. A.,
395 Jumper, J. M., Salmon, J. K., Shan, Y., and Wriggers, W. (2010). Atomic-level characterization of the
396 structural dynamics of proteins. *Science*, 330(6002):341–346.

397 Shen, Y. and Bax, A. (2007). Protein backbone chemical shifts predicted from searching a database for
398 torsion angle and sequence homology. *J. Biomol. NMR.*, 38:289–302.

399 Shen, Y., Lange, O., Delaglio, F., Rossi, P., Aramini, J. M., Liu, G., Eletsky, A., Wu, Y., Singarapu, K. K.,
400 Lemak, A., Ignatchenko, A., Arrowsmith, C. H., Szyperski, T., Montelione, G. T., Baker, D., and Bax,
401 A. (2008). Consistent blind protein structure generation from nmr chemical shift data. *Proceedings of*
402 *the National Academy of Sciences*, 105(12):4685–4690.

403 Snow, C. D., Nguyen, H., Pande, V. S., and Gruebele, M. (2002). Absolute comparison of simulated and
404 experimental protein-folding dynamics. *Nature*, 420:102–106.

405 Sprangers, R., Groves, M. R., Sinning, I., and Sattler, M. (2003). High-resolution x-ray and {NMR}
406 structures of the {SMN} tudor domain: Conformational variation in the binding site for symmetrically
407 dimethylated arginine residues. *Journal of Molecular Biology*, 327(2):507 – 520.

408 Tange, O. (2011). Gnu parallel—the command-line power tool. *The USENIX Magazine*, 36(1):42–47.

409 Ulmer, T. S., Ramirez, B. E., Delaglio, F., and Bax, A. (2003). Evaluation of backbone proton positions
410 and dynamics in a small protein by liquid crystal nmr spectroscopy. *Journal of the American Chemical*
411 *Society*, 125(30):9179–9191.

412 Zhang, H., Neal, S., and Wishart, D. (2003). RefDB: a database of uniformly referenced protein chemical
413 shifts. *J. Biomol. NMR.*, 25:173–195.