Supplementary Information: Efficient Ensemble Refinement by Reweighting

Jürgen Köfinger,[*,†] Lukas S. Stelzl,[†] Klaus Reuter,[‡] César Allande,[‡] Katrin Reichel,[†] and Gerhard Hummer[*,†,¶]

†*Department of Theoretical Biophysics, Max Planck Institute of Biophysics, Max-von-Laue-Straße 3, 60438 Frankfurt am Main, Germany*

‡*Max Planck Computing and Data Facility, Gießenbachstr. 2, 85748 Garching, Germany*

¶*Institute for Biophysics, Goethe University, Max-von-Laue-Straße 9, 60438 Frankfurt am Main, Germany*

E-mail: juergen.koefinger@biophys.mpg.de; gerhard.hummer@biophys.mpg.de

1 Gradients of the Log-Posterior for Correlated Gaussian Errors

In the following, we present a detailed derivation of the expressions for the gradients of the negative log-posterior given by eq 4 of the main text in the log-weights and forces formulation for correlated Gaussian errors. Expressions for the gradients for uncorrelated errors presented in the main text are special cases of the more general expressions derived here.

For correlated Gaussian errors, the likelihood is given by $P(\{y_i\}|\boldsymbol{w}) \propto \exp\left(-\chi^2/2\right)$, with

$$\chi^2 = \boldsymbol{r}^\top \boldsymbol{S}^{-1} \boldsymbol{r} \tag{1}$$

such that eq 4 of the main text takes on the form

$$L = \theta S_{\mathrm{KL}} + \frac{\chi^2}{2}. \tag{2}$$

The components of the vector of residuals $\boldsymbol{r}$ are given by[1]

$$r_i = \sum_{\alpha=1}^{N} w_\alpha y_i^\alpha - Y_i = \sum_{\alpha=1}^{N} w_\alpha r_i^\alpha \tag{3}$$

where we introduced $r_i^\alpha = y_i^\alpha - Y_i$. $\boldsymbol{S}$ is the symmetric and positive definite covariance matrix of the statistical errors. Note that for uncorrelated errors the covariance matrix is diagonal, $\boldsymbol{S} = \mathrm{diag}\{\sigma_1^2, \ldots, \sigma_M^2\}$. Denoting the ij elements of the inverse of $\boldsymbol{S}$ as S_{ij}^{-1}, we may write

$$\chi^2 = \sum_{i=1}^{M}\sum_{j=1}^{M} S_{ij}^{-1} r_i r_j \,. \tag{4}$$

We derive the gradients of the negative log-posterior given by eq 2 by separately evaluating the gradients of the relative entropy S_{KL} and of χ^2. To derive the gradients of the relative entropy S_{KL} given by eq 3 of the main text in the log-weights and forces methods below, we use the chain rule and first derive here the gradient with respect to the weights w_α. We take into account that weights are normalized and set $w_{\mathrm{N}} = 1 - \sum_{\alpha=1}^{N-1} w_\alpha$ and write

$$S_{\mathrm{KL}} = \sum_{\alpha=1}^{N-1} w_\alpha \ln \frac{w_\alpha}{w_\alpha^0} + w_N \ln \frac{w_N}{w_N^0} \,. \tag{5}$$

The derivative of the first term of eq 5 with respect to w_γ is given by

$$\frac{\partial}{\partial w_\gamma} \sum_{\alpha=1}^{N-1} w_\alpha \ln \frac{w_\alpha}{w_\alpha^0} = \ln \frac{w_\gamma}{w_\gamma^0} + 1 \,. \tag{6}$$

The derivative of the second term of eq 5 with respect to w_γ is given by

$$\frac{\partial}{\partial w_\gamma} w_N \ln \frac{w_N}{w_N^0} = -\ln \frac{w_N}{w_N^0} - 1 \tag{7}$$

where we used that $\partial w_N/\partial w_\gamma = -1$ and $\partial \ln w_N/\partial w_\gamma = -1/w_N$. Thus, we obtain

$$\frac{\partial S_{\rm KL}}{\partial w_\gamma} = \ln \frac{w_\gamma}{w_\gamma^0} - \ln \frac{w_N}{w_N^0} \,. \tag{8}$$

1.1 Log-Weights

We derive the gradient of the negative log-posterior given by eq 2 with respect to the log-weights given by eq 12 of the main text.

To calculate the gradient of the relative entropy we apply the chain rule, i.e.,

$$\frac{\partial S_{\rm KL}}{\partial g_\gamma} = \sum_{\alpha=1}^{N} \frac{\partial S_{\rm KL}}{\partial w_\alpha} \frac{\partial w_\alpha}{\partial g_\gamma} \,. \tag{9}$$

Inserting eq 8 into eq 9 and using

$$\frac{\partial w_\alpha}{\partial g_\gamma} = w_\alpha \left[\delta_{\alpha\gamma} - w_\gamma\right] \tag{10}$$

and

$$\frac{\partial \ln \sum_{\alpha=1}^{N} e^{g_\alpha}}{\partial g_\gamma} = w_\gamma \,, \tag{11}$$

we obtain

$$\frac{\partial S_{\rm KL}}{\partial g_\gamma} = w_\gamma \left(g_\gamma - \langle g \rangle - G_\gamma + \langle G \rangle\right) \tag{12}$$

where $g_\gamma = \ln w_\gamma$, $G_\gamma = \ln w_\gamma^0$, $\langle g \rangle = \sum_{\alpha=1}^{N} w_\alpha g_\alpha$, and $\langle G \rangle = \sum_{\alpha=1}^{N} w_\alpha^0 G_\alpha$.

To calculate the gradient of χ^2 given by eq 4 with respect to the log-weights we use the chain rule,

$$\begin{aligned} \frac{\partial r_i}{\partial g_\gamma} &= \sum_{\alpha=1}^{N} \frac{\partial r_i}{\partial w_\alpha} \frac{\partial w_\alpha}{\partial g_\gamma} \\ &= \sum_{\alpha=1}^{N} r_i^\alpha w_\alpha \left(\delta_{\alpha\gamma} - w_\gamma\right) \\ &= w_\gamma (r_i^\gamma - r_i) \,. \end{aligned} \tag{13}$$

Thus, the gradient of eq 4 with respect to the log-weights becomes

$$\frac{\partial \chi^2}{\partial g_\gamma} = w_\gamma \sum_{i,j=1}^{M} S_{ij}^{-1} \left(r_i^\gamma r_j + r_j^\gamma r_i - 2 r_i r_j \right) \ . \tag{14}$$

Consequently, the gradient of the negative log-posterior with respect to the log-weights for correlated Gaussian errors is given by

$$\begin{aligned} \frac{\partial L}{\partial g_\gamma} &= \theta w_\gamma \left(g_\gamma - \langle g \rangle - G_\gamma + \langle G \rangle \right) \\ &\quad + \frac{w_\gamma}{2} \sum_{i,j=1}^{M} S_{ij}^{-1} \left(r_i^\gamma r_j + r_j^\gamma r_i - 2 r_i r_j \right) \ . \end{aligned} \tag{15}$$

For uncorrelated errors the covariance matrix is diagonal and the gradient of χ^2 simplifies to

$$\frac{\partial \chi^2}{\partial g_\gamma} = 2 \sum_{i=1}^{M} \frac{r_i (r_i^\gamma - r_i)}{\sigma_i^2} \ , \tag{16}$$

such that we recover eq 14 of the main text as expected.

1.2 Generalized Forces

For correlated Gaussian errors, the generalized forces are given by

$$F_i = -\frac{1}{\theta} \sum_{j=1}^{M} S_{ij}^{-1} f_j \tag{17}$$

where $f_j = \langle y_j \rangle - Y_j$. These forces determine the weights via eq 19 of the main text. To calculate the gradient of L given by eq 2 with respect to the forces we use the chain rule,

$$\frac{\partial L}{\partial F_k} = \sum_{\alpha=1}^{N} \frac{\partial L}{\partial w_\alpha} \frac{\partial w_\alpha}{\partial F_k} \tag{18}$$

with

$$\frac{\partial w_\alpha}{\partial F_k} = w_\alpha \left(r_k^\alpha - r_k\right) \ . \tag{19}$$

By applying the chain rule, we obtain the gradient of the relative entropy with respect to the forces,

$$\frac{\partial S_{\mathrm{KL}}}{\partial F_k} = \sum_{\alpha=1}^{N-1} \left(\ln \frac{w_\alpha}{w_\alpha^0} - \ln \frac{w_N}{w_N^0} \right) w_\alpha \left(r_k^\alpha - r_k\right) \ . \tag{20}$$

where we used eqs 8 and 19.

Next, we calculate the gradient of χ^2 given by eq 4 with respect to w_α. Because of the normalization condition $\sum_{\alpha=1}^{N} w_\alpha = 1$, we only have $N-1$ independent variables. Using that $w_N = 1 - \sum_{\alpha=1}^{N-1} w_\alpha$, we write

$$r_i = \sum_{\alpha=1}^{N-1} w_\alpha r_i^\alpha + \left(1 - \sum_{\alpha=1}^{N-1} w_\alpha\right) r_i^N \, . \tag{21}$$

Consequently, $\partial r_i / \partial w_\gamma = r_i^\gamma - r_i^N$ for $\gamma < N$. We obtain for the gradient of eq 4 with respect to w_γ for $\gamma < N$:

$$\frac{\partial \chi^2}{\partial w_\gamma} = \sum_{i,j=1}^{M} S_{ij}^{-1} \left[(r_i^\gamma - r_i^N) r_j + (r_j^\gamma - r_j^N) r_i \right] \tag{22}$$

and $\partial \chi^2 / \partial w_N = 0$. By applying the chain rule and using eqs 19 and 22, we obtain

$$\frac{\partial \chi^2}{\partial F_k} = \sum_{\gamma=1}^{N-1} \sum_{i,j=1}^{M} S_{ij}^{-1} \left[(r_i^\gamma - r_i^N) r_j + (r_j^\gamma - r_j^N) r_i \right] \times w_\gamma \left(r_k^\gamma - r_k\right) \ . \tag{23}$$

Consequently, for correlated Gaussian errors the gradient of the negative log-posterior in eq

2 with respect to the generalized forces is given by

$$\frac{\partial L}{\partial F_k} = \sum_{\gamma=1}^{N-1} \left[\theta \left(\ln \frac{w_\gamma}{w_\gamma^0} - \ln \frac{w_N}{w_N^0} \right) + \frac{1}{2} \sum_{i,j=1}^{M} S_{ij}^{-1} \left[(r_i^\gamma - r_i^N) r_j + (r_j^\gamma - r_j^N) r_i \right] \right] \times w_\gamma \left(r_k^\gamma - r_k \right) . \quad (24)$$

For uncorrelated errors, eq 23 simplifies to

$$\frac{\partial \chi^2}{\partial F_k} = 2 \sum_{\gamma=1}^{N-1} \sum_{i=1}^{M} \frac{1}{\sigma_i^2} (r_i^\gamma - r_i^N) r_i w_\gamma \left(r_k^\gamma - r_k \right) \ , \quad (25)$$

i.e., we recover eq 20 of the main text, which in the notation used in here in Supplementary Information takes on the form

$$\frac{\partial L}{\partial F_k} = \sum_{\gamma=1}^{N-1} \left[\theta \left(\ln \frac{w_\gamma}{w_\gamma^0} - \ln \frac{w_N}{w_N^0} \right) + \sum_{i=1}^{M} \frac{1}{\sigma_i^2} (r_i^\gamma - r_i^N) r_i \right] \times w_\gamma \left(r_k^\gamma - r_k \right) .$$

2 Refinement of Ala-5 using J-Couplings

Comparison of Optimization using Generalized Forces and Log-Weights

Ensemble refinements using generalized forces and log-weights give very similar results across the full range of the confidence parameter θ. The correlation of the optimal weights for Ala-5 refined against J-couplings is shown in Figure S1A. The DFT2 set of Karplus parameters was used for this comparison. Small deviations are seen at small values of θ. The effective log-likelihoods from optimization with the two methods agree very well over the full range of θ values (Figure S1B).

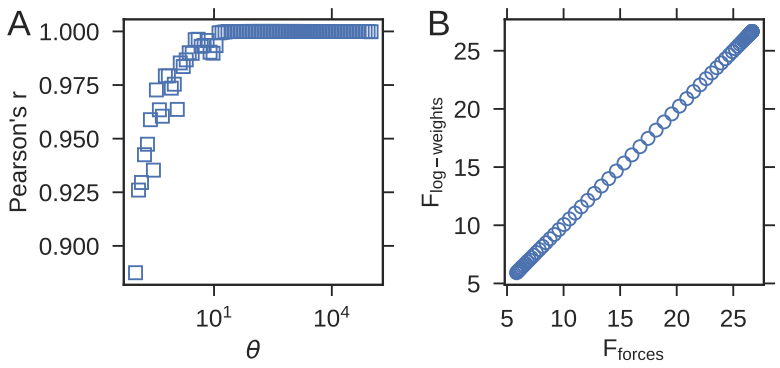

Figure S1: Comparison of forces and log-weights optimization. (A) Correlation between optimal weights for different values of the confidence parameter θ using log-weights or generalized forces in the BioEn optimization problem. (B) Effective log-likelihoods from forces and log-weights optimization.

Effect of the Choice of Karplus Parameters on Ala-5 Ensemble Refinement

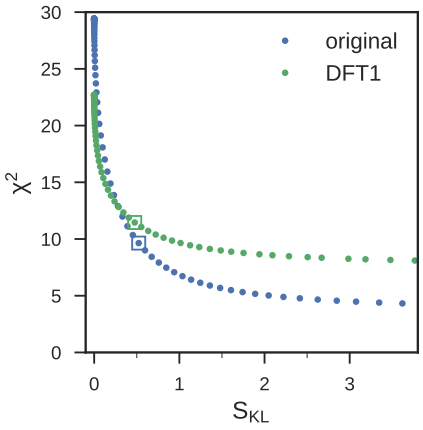

Figure S2: Variation of S_{KL} and χ^2 with θ, to determine the optimal value of the confidence parameter θ for the reweighting of Ala-5 using the original (blue) and DFT1 sets of Karplus parameters (green). Values of θ of 9.43 and 5.58, respectively, provide a compromise between minimizing χ^2 and small changes to the reference weights for BioEn S_{KL} of about 0.5, as highlighted by squares.

BioEn ensemble refinement produced very similar trends no matter which Karplus parameters were used to calculate the J-couplings. We performed independent Ala-5 ensemble refinement with three different set of Karplus parameters: the empirical parameters[2] (origi-

nal) and two set of parameters obtained from density functional theory[3] (DFT1 and DFT2). For further analysis of optimization with the original and DFT1 parameter sets we picked refined ensembles with $S_{\mathrm{KL}} = 0.5$. Irrespective of which set of Karaplus parameters we used to calculate J-couplings the polyproline-II conformation becomes more populated and the α-helical like conformation less populated (Figure 3D in main text, Figure S4D and Figure S3D). The changes in the populations of the conformational states, as defined previously,[4] are summarized in Figure S5. The main difference between the results with different Karplus parameters is the reduction in the β-strand like conformations seen when calculating the J-couplings with the original set of Karplus parameters. The left-handed helical α_L conformation becomes somewhat less prominent after the refinement. Leaving out the $^3J_{\mathrm{CC'}}$ coupling does not change the trends either (Figure S5).

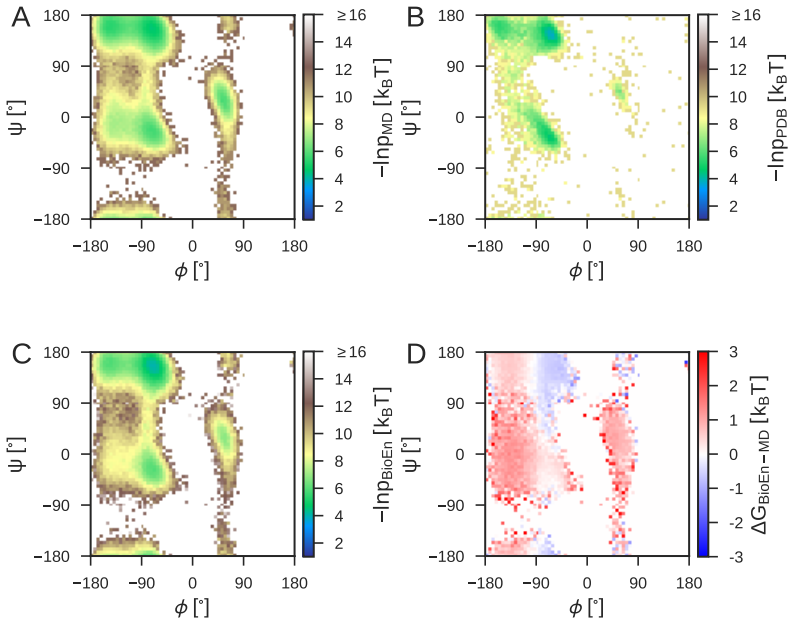

Figure S3: Free energy surface $G(\phi, \psi) = -\ln p(\phi, \psi)$ for Ala-5 from BioEn optimization, with J-couplings calculated with the original Karplus parameters from Graf et al.[2] Free energy surface for AMBER99SB*-ildn-q for the central residues 2-4. (B) Ramachandran map for Ala residues outside of regular secondary structure from the PDB.[5] (C) Free energy surfaces for the optimal BioEn ensemble. (D) Free energy differences between initial ensemble and the optimal BioEn ensemble.

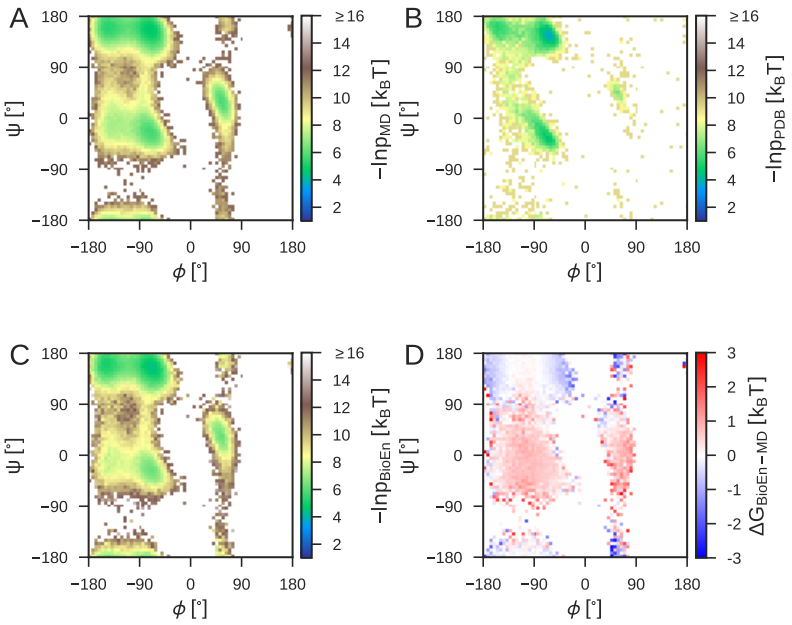

Figure S4: Free energy surface $G(\phi, \psi) = -\ln p(\phi, \psi)$ for Ala-5 from BioEn optimization, with J-couplings calculated with DFT1 Karplus parameters. Free energy surface for AMBER99SB*-ildn-q for the central residues 2-4. (B) Ramachandran map for Ala residues outside of regular secondary structure from the PDB.[5] (C) Free energy surfaces for the optimal BioEn ensemble. (D) Free energy differences between initial ensemble and the optimal BioEn ensemble.

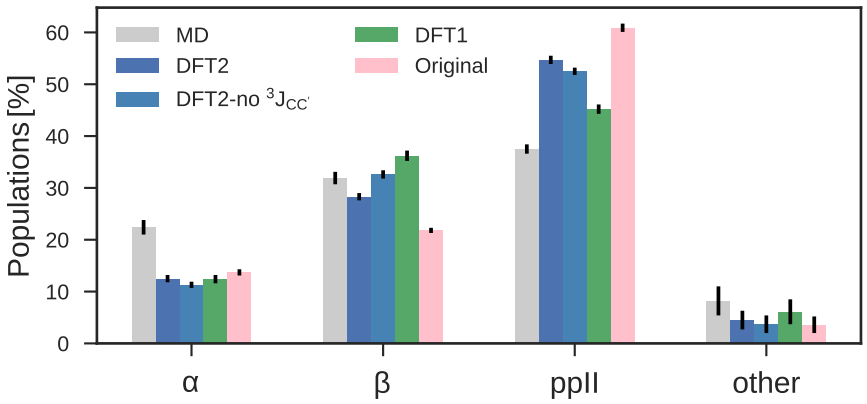

Figure S5: Populations for the conformational states for the initial ensemble and for the optimal ensembles. Here optimal ensembles for DFT1, DFT2, DFT2 (and excluding $^3J_{CC'}$ coupling) and original Karplus parameters are compared to each other and the original molecular dynamics (MD) simulations.

Agreement for Individual J-Couplings

Comparing the agreement between the simulated ensemble and experiments for individual observables (Figure S6) shows which data points drive the ensemble refinement. Here we focus at ensemble refinement using J-couplings calculated with the DFT2 of Karplus parameters. For ${}^3J_{\mathrm{CC}'}$ (Figure S6D), ${}^3J_{\mathrm{HNH}\alpha}$ (Figure S6A) and ${}^3J_{\mathrm{H}\alpha\mathrm{C}'}$ (Figure S6C) couplings the agreement between experiment and simulations improves considerably with the optimal ensemble at $\theta = 6.65$. ${}^3J_{\mathrm{CC}'}$ was measured only for residue 2 of Ala-5 and χ^2 was decreased from ≈ 8 to ≈ 2. For ${}^3J_{\mathrm{H}\alpha\mathrm{C}'}$ the improvement is driven by residue 4 which fits poorly in the initial ensemble, whereas for the other residues the agreement is already very good in the initial ensemble. Some improvement in the fit was obtained for ${}^2J_{\mathrm{NC}\alpha}$ (Figure S6G) and ${}^3J_{\mathrm{HNC}\alpha}$ (Figure S6H), with χ^2 reduced from 3 to < 1 and 2 to ≈ 0.5 respectively. Only very small changes were seen for ${}^1J_{\mathrm{NC}\alpha}$ (Figure S6F). Note that the ${}^1J_{\mathrm{NC}\alpha}$ coupling for residue 5 is uninformative in our analysis as evidenced by the flat χ^2 across the full-range of θ values. The ψ dihedral angle is not defined for the terminal residue and the calculated ${}^1J_{\mathrm{NC}\alpha}$ depends on ψ in the current parameterization. For ${}^3J_{\mathrm{HNC}'}$ (Figure S6B) and ${}^3J_{\mathrm{HNC}\beta}$ (Figure S6E) the agreement is extremely good to start with and deteriorates somewhat with the refinement. Importantly, as discussed in the main text, the refinement removes systematic offsets for ${}^3J_{\mathrm{HNH}\alpha}$, ${}^3J_{\mathrm{H}\alpha\mathrm{C}'}$ and ${}^2J_{\mathrm{NC}\alpha}$.

Chemical Shifts

Experimental chemical shifts for Ala-5 from Graf et al.[2] were compared to the chemical shifts calculated from the initial and reweighted, optimal ensemble Figure S7. The error in the comparison of calculated and measured shifts is dominated by the forward model. Hence the error bars shows the root mean square error for SPARTA+[6] predictions for the respective nuclei previously determined.

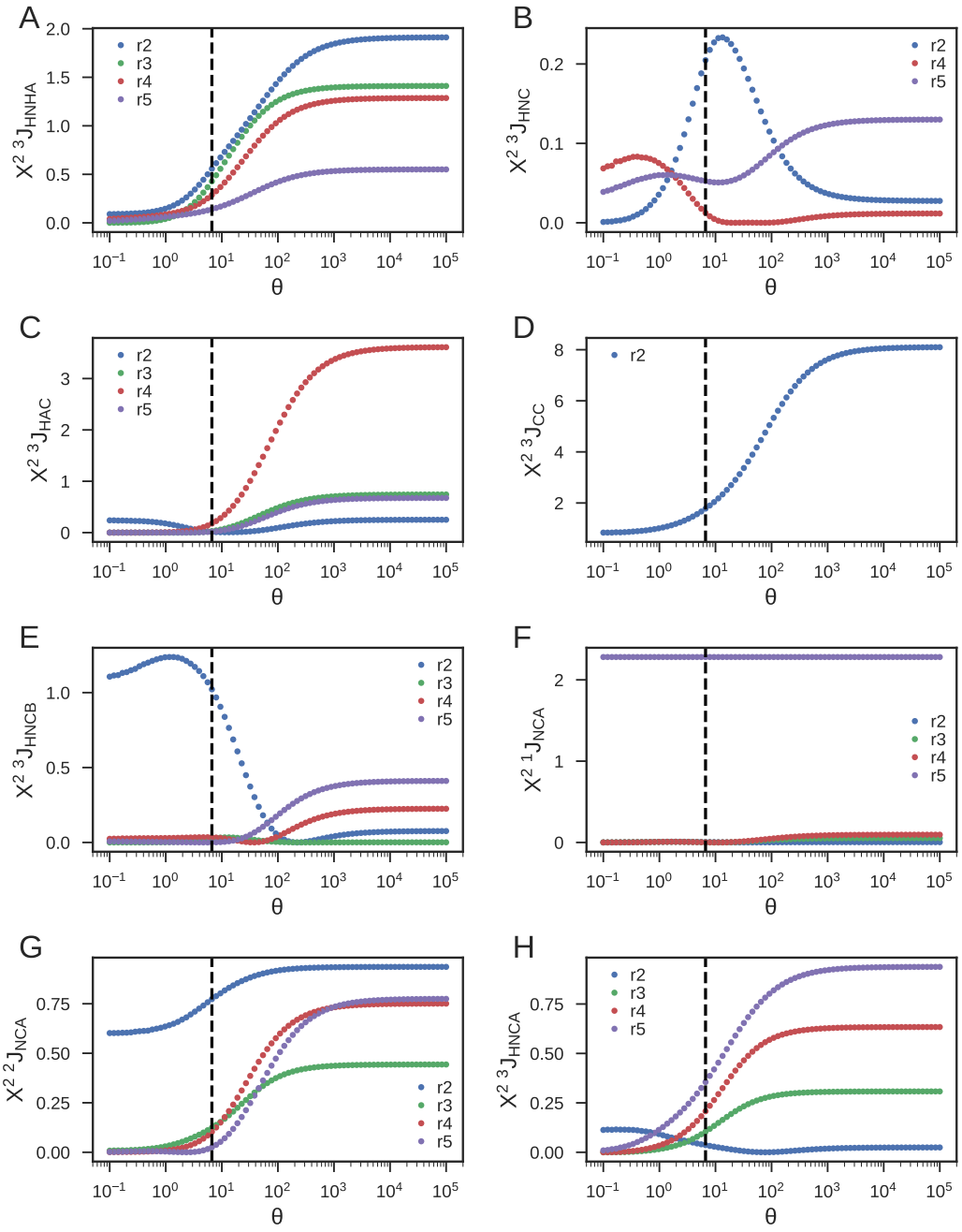

Figure S6: Comparison of the agreement with experiments for J-couplings calculated with the DFT2 set of Karplus parameters. The grey circles indicated the sum of χ^2 for different residues for a type of J-couplings at given value of the confidence parameter θ. The dashed line indicates $\theta = 6.65$ chosen for further analysis.

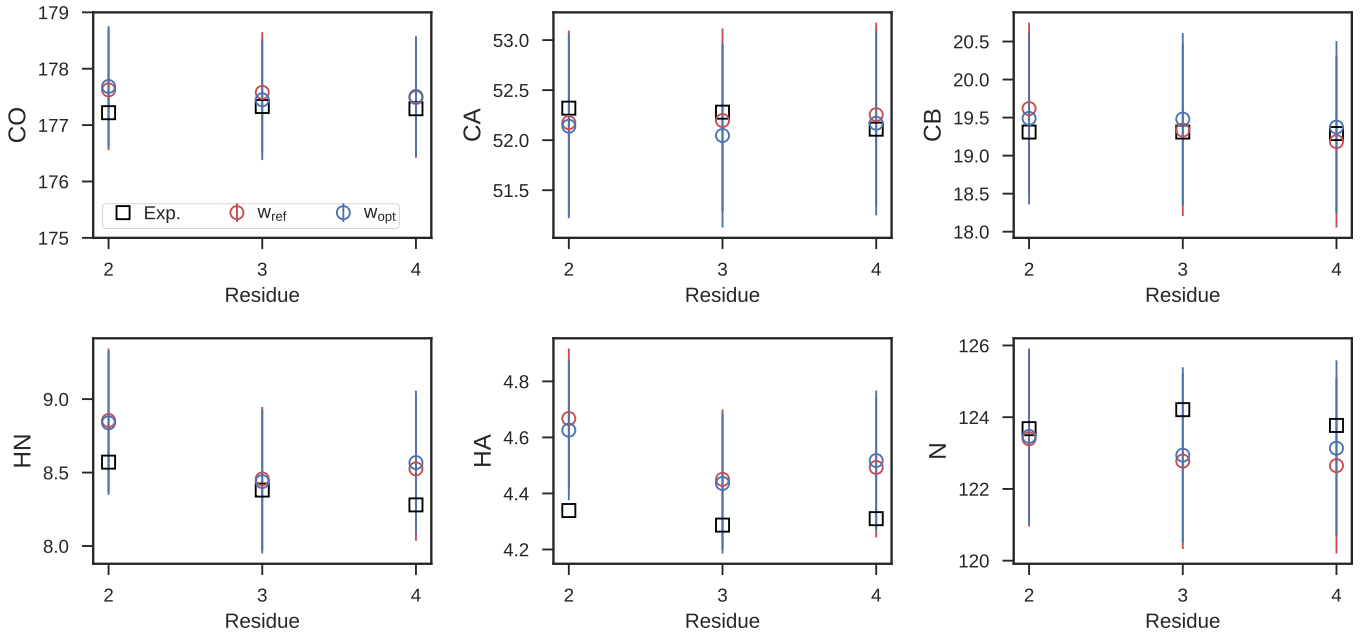

Figure S7: Ala-5 chemical shifts calculated from the initial and optimal ensemble with the DFT2 Karplus parameters.

References

(1) Hummer, G.; Köfinger, J. Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* **2015**, *143*, 243150.

(2) Graf, J. et al. Structure and Dynamics of the Homologous Series of Alanine Peptides: A Joint Molecular Dynamics/NMR Study. *J. Am. Chem. Soc.* **2007**, *129*, 1179–1189.

(3) Case, D. A.; Scheurer, C.; Brüschweiler, R. Static and Dynamic Effects on Vicinal Scalar J Couplings in Proteins and Peptides: A MD/DFT Analysis. *J. Am. Chem. Soc.* **2000**, *122*, 10390–10397.

(4) Best, R. B.; Buchete, N.-V.; Hummer, G. Are Current Molecular Dynamics Force Fields too Helical? *Biophys. J.* **2008**, *95*, L07 – L09.

(5) Mantsyzov, A. B. et al. MERA: a webserver for evaluating backbone torsion angle dis-

tributions in dynamic and disordered proteins from NMR data. *J. Biomol. NMR* **2015**, 1–11.

(6) 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.