



Bayesian Weighting of Statistical Potentials in NMR Structure Calculation

Martin Mechelke^{1,2}, Michael Habeck^{1*}

1 Institute for Mathematical Stochastics, Georg August University Göttingen, Göttingen, Germany, **2** Department of Protein Evolution, Max Planck Institute for Developmental Biology, Tübingen, Germany

Abstract

The use of statistical potentials in NMR structure calculation improves the accuracy of the final structure but also raises issues of double counting and possible bias. Because statistical potentials are averaged over a large set of structures, they may not reflect the preferences of a particular structure or data set. We propose a Bayesian method to incorporate a knowledge-based backbone dihedral angle potential into an NMR structure calculation. To avoid bias exerted through the backbone potential, we adjust its weight by inferring it from the experimental data. We demonstrate that an optimally weighted potential leads to an improvement in the accuracy and quality of the final structure, especially with sparse and noisy data. Our findings suggest that no universally optimal weight exists, and that the weight should be determined based on the experimental data. Other knowledge-based potentials can be incorporated using the same approach.

Citation: Mechelke M, Habeck M (2014) Bayesian Weighting of Statistical Potentials in NMR Structure Calculation. PLoS ONE 9(6): e100197. doi:10.1371/journal.pone.0100197

Editor: Bruce R. Donald, Duke University Medical Center, Duke University, United States of America

Received: January 24, 2014; **Accepted:** May 23, 2014; **Published:** June 23, 2014

Copyright: © 2014 Mechelke, Habeck. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work has been supported by Deutsche Forschungsgemeinschaft (DFG) grant HA 5918/1-1, the University of Göttingen and the Max Planck Society. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* Email: mhabeck@gwdg.de

Introduction

Structural data measured by NMR spectroscopy are never complete. Even the most carefully collected data will by themselves not allow us to determine the three-dimensional structure of a biomolecule with atomic resolution. Rather, we need to interpret the data in the light of prior knowledge that is typically encoded in a potential function or force field [1].

Potential functions quantify the forces and interactions within a biomolecule and with its environment. Two fundamentally different approaches of designing potential functions are commonly used [2]. Physics-based force fields [3] aim to approximate the underlying physical laws. Statistical or knowledge-based potentials [4] are learned from a structure database and describe the effective forces resulting from all interactions including those with the solvent. Physical and statistical potentials are complementary in the sense that some interactions cannot be broken down easily into fundamental, physical contributions but are captured more effectively by potentials derived from known structures.

In NMR structure calculation, potential functions are used to guide the calculation towards structures of high quality and accuracy [1]. This guidance is needed because NMR measurements by themselves do not allow us to determine the three-dimensional structure of an entire macromolecule. To “let the data speak for themselves” and also for reasons of computational efficiency, one tends to use minimalist force fields that ignore complex effects such as electrostatic screening or solvent interactions. If additional potentials such as dihedral angle [5,6] or hydrogen bonding potentials [7] are used, their force constants are set *ad hoc* and held fixed during the structure calculation.

However, it might be necessary to adjust the force constants for each data set. Because knowledge-based potentials represent averages over large sets of structures, they are not universally transferable and may not represent the preferences of a particular structure. Think of a backbone dihedral angle potential, a “Ramachandran potential”, as an example. The minimal energy configuration of the Ramachandran potential is completely alpha-helical. Therefore it seems more appropriate to choose a higher weight for helical proteins than for all-beta proteins.

Here we introduce an objective, data-driven approach to find the optimal force constant for a given protein and data set. Our method is based on statistical mechanics and Bayesian inference and allows us to incorporate knowledge-based potential functions without biasing the structure calculation.

Results

Statistical potential for backbone dihedral angles

Protein backbone dihedral angles φ and ψ show a typical correlation pattern, an observation made by Ramachandran *et al.* [8] assuming only hard-sphere steric repulsion between atoms. Standard nonbonded energies used in NMR structure determination [9] do not fully capture all aspects of φ/ψ distributions observed in high-resolution crystal structures [10]. The dihedral angle distributions obtained from nonbonded interactions do not reproduce the empirical distribution (see the bottom row of Figure 1 and Figures S2–S4 in File S1). Depending on the quality of the data, NMR structures can show dihedral angles outside the allowed regions of the Ramachandran plot [11,12]. For this reason it is common practice [13] to assess the Ramachandran statistics of NMR structures by programs such as Procheck [14], MolProbity

[15] and WhatCheck [16]. To obtain more regular NMR structures, various dihedral angle potentials derived from database statistics have been developed [5,6]. The functional forms of these potentials range from two-dimensional histograms [17] to continuous representations based on linear interpolation, cubic splines and statistical density estimation [18–20]. Some of these models ignore that φ/ψ distributions are smooth and periodic, which can result in artifact in the refinement [21].

We use nonparametric density estimation to derive a backbone potential and expand the joint φ/ψ distribution into a Fourier series [22]. This representation is inherently smooth and periodic

and has the advantage that it can easily represent multimodal distributions. Each distribution is a linear combination of 80 two-dimensional cosine and sine functions resulting from the combination of five frequencies in the φ and ψ dimensions. The estimated distributions capture features such as the alpha-helical peak and regions corresponding to parallel and anti-parallel beta sheets. Also rare secondary structures such as left-handed helices are represented accurately. Figure 1 shows the estimated dihedral distribution of three representative amino acids and the corresponding empirical histograms (see Figure S2 in File S1 for the full set of dihedral angle distributions).

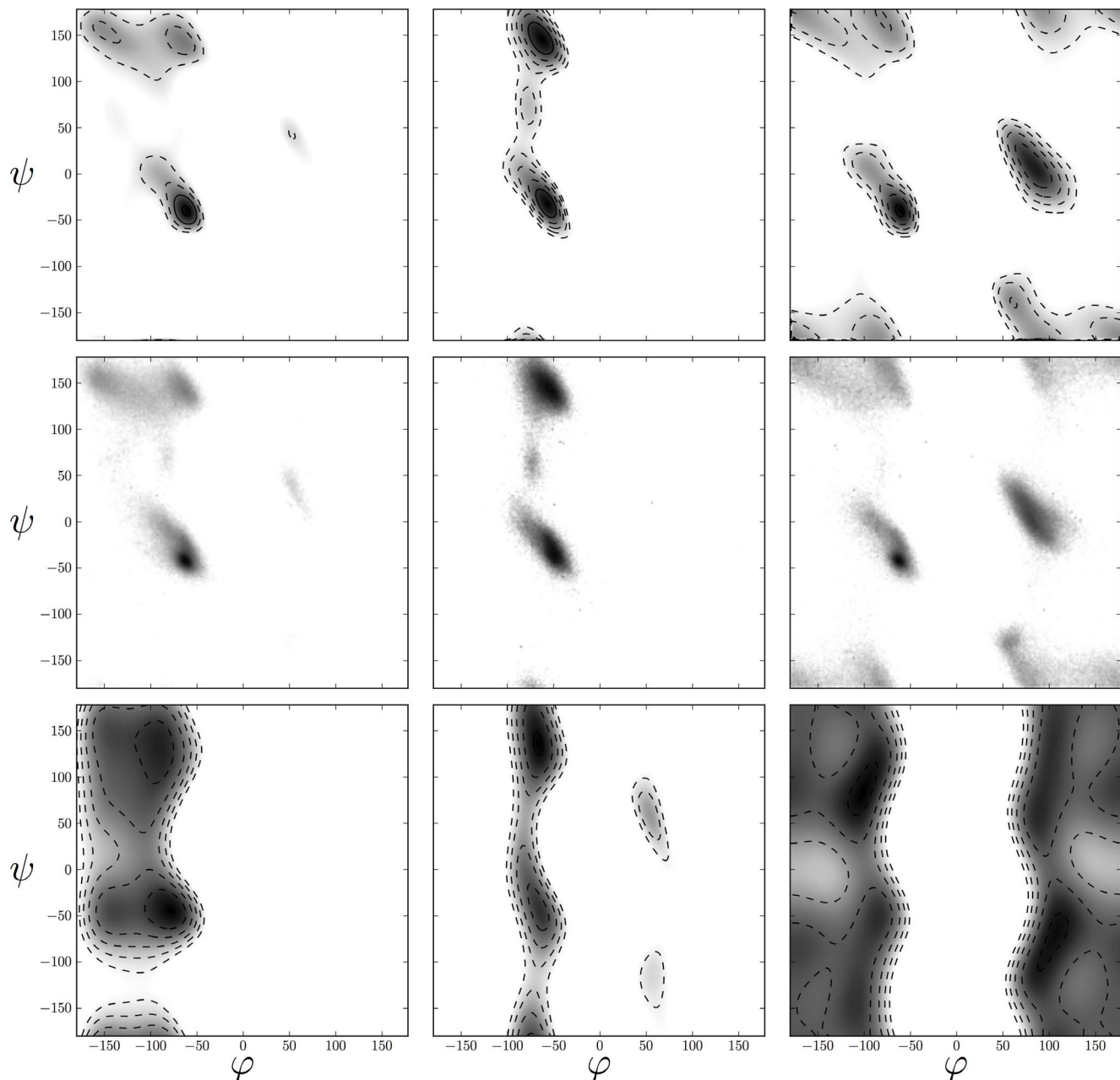


Figure 1. Backbone dihedral angle distributions of amino acids estimated from high-resolution crystal structures. Density maps of φ/ψ distributions for Alanine (left column), Proline (middle column) and Glycine (right column) as approximated by the maximum entropy distribution (top row). The middle row shows the empirical φ/ψ distribution computed over a non-redundant structure database. The bottom row shows the φ/ψ distribution obtained by sampling structures from the nonbonded force field.
doi:10.1371/journal.pone.0100197.g001

Data-driven weighting of the backbone potential

We use the probabilistic Inferential structure determination approach (ISD) [23,24] to determine protein structures from experimental data. In a standard ISD calculation, one explores the posterior probability $\Pr(\theta|D) \propto \Pr(D|\theta) \Pr(\theta)$ of all conformational degrees of freedom θ given the experimental data D . The posterior distribution itself is proportional to the likelihood function $\Pr(D|\theta) \propto \exp\{-w_{\text{data}} E_{\text{data}}(\theta)\}$ and the prior probability $\Pr(\theta)$. The likelihood is the probability of the data given the structure and involves $E_{\text{data}}(\theta)$, a measure of the goodness-of-fit between the data and a particular structure θ ; the weight w_{data} allows us to balance the data against the prior probability. The prior distribution is typically a Boltzmann distribution at inverse temperature $w_{\text{phys}} = 1/k_{\text{B}}T$, $\Pr(\theta) \propto \exp\{-w_{\text{phys}} E_{\text{phys}}(\theta)\}$ resulting from the force field E_{phys} . We incorporate the newly derived backbone potential $E_{\text{rama}}(\theta)$ by extending the prior distribution: $\Pr(\theta|w_{\text{rama}}) \propto \exp\{-w_{\text{phys}} E_{\text{phys}}(\theta) - w_{\text{rama}} E_{\text{rama}}(\theta)\}$ where w_{rama} is the weight of the backbone potential.

The weight of the backbone potential w_{rama} is unknown and has to be chosen somehow. Naively, we would set it to one ($w_{\text{rama}} = 1$). But this is problematic because some aspects of the Ramachandran plot are already captured by the force field. Figures 1 and S3 in File S1 show that structures calculated on the basis of the nonbonded force field E_{phys} already reproduce the rough outline of the φ/ψ basins. But there are more subtle aspects such as optimal hydrogen bonding geometry [25] that result in pronounced peaks, which are not reproduced by the force field alone. As a consequence, the force field and the backbone potential are not independent of each other but are positively and negatively correlated depending on the energy range (Figure 2A). Setting the φ/ψ weight to a large value risks that we overemphasize these contributions in the combined potential. In the limiting case, we will force the structure into a helical conformation, whereas with too small w_{rama} the effect of the backbone potential becomes negligible. Therefore, we need to adjust w_{rama} according to the experimental data and the structure.

We have introduced a Bayesian approach to estimate the weight of the experimental data w_{data} relative to the prior probability [26]. This approach exploits the fact that for every conformation we can calculate how well it agrees with the data and that its goodness-of-fit determines the weight of the data w_{data} . The same is not possible for w_{rama} , because the statistics reflected by the backbone potential recapitulate an ensemble property, and therefore we need to assess how well the entire *ensemble* agrees with the Ramachandran statistics. Thus it is computationally much more demanding to adjust w_{rama} than w_{data} .

To estimate w_{rama} , we compare the differences between the expected backbone energy $\langle E_{\text{rama}} \rangle$ where $\langle \cdot \rangle$ denotes an ensemble average. $\langle E_{\text{rama}} \rangle_{\text{no data}}$ summarizes how the force field and the backbone potential are correlated independent of any data. To obtain this ensemble average, structures are sampled based on the combined energy $w_{\text{phys}} E_{\text{phys}} + w_{\text{rama}} E_{\text{rama}}$. This value is contrasted with the expected backbone energy obtained with data $\langle E_{\text{rama}} \rangle_{\text{data}}$. To calculate this ensemble average, structures are sampled based on the full energy $w_{\text{data}} E_{\text{data}} + w_{\text{phys}} E_{\text{phys}} + w_{\text{rama}} E_{\text{rama}}$. If $\langle E_{\text{rama}} \rangle_{\text{data}} < \langle E_{\text{rama}} \rangle_{\text{no data}}$, the addition of the experimental data improves the backbone energy, and we can increase w_{rama} because the data comply with the Ramachandran statistics. If the data contradict the backbone potential, $\langle E_{\text{rama}} \rangle_{\text{data}} > \langle E_{\text{rama}} \rangle_{\text{no data}}$, we must lower the weight because the backbone potential biases the ensemble. The optimal weight achieves

$$\langle E_{\text{rama}} \rangle_{\text{data}} = \langle E_{\text{rama}} \rangle_{\text{no data}}. \quad (1)$$

This rule can be derived rigorously [27] by maximization of the *model evidence*, $\Pr(D|w_{\text{rama}})$, which is the probability of observing the data for a particular value of w_{rama} and whose computation involves an ensemble average. It can be shown that the derivative of the model evidence is the difference between the average Ramachandran potential under the posterior and the prior (i.e. with and without data) [27]. At the maximum evidence the derivative vanishes, therefore both energies are the same and the curves cross. If we choose the weight from this region, we bias our ensembles the least.

Bayesian weighting with high-quality data

We used Bayesian weighting to analyze the high-quality data for ubiquitin (PDB code: 1d3z). We estimated the optimal weight from the NOE-based distances and used the additional scalar and dipolar data for validation. Figure 2B shows the model evidence $\Pr(D|w_{\text{rama}})$, which peaks at the optimal weight satisfying equation (1). Assuming a uniform prior probability for w_{rama} , the estimated weight of the backbone potential is $w_{\text{rama}} = 0.94 \pm 0.09$. An additional control for evaluating the Bayesian choice of the Ramachandran weight are the residual dipolar couplings (RDCs) and scalar coupling measurements that are available for ubiquitin. For each of the 11 RDC sets, we calculated the average Q-factor obtained for different choices of w_{rama} . Figure 2C shows that the Bayesian choice of the Ramachandran weight improves the Q-factor to $Q = 0.19 \pm 0.04$. Similarly, we see an improvement in the fit with the scalar coupling measurements (Fig. 2D).

Bayesian weighting with incomplete data

Next, we studied how the weight changes for sparse versions of the ubiquitin data. To do so, we introduce a completeness parameter λ , which controls the amount of data [28]. For $\lambda < 1$, the effective number of observations is reduced. As observed before [27], the model evidence broadens and shifts towards smaller values if we reduce the number of observations (Fig. 3). The estimated weights are: $w_{\text{rama}} = 0.94 \pm 0.09$ ($\lambda = 1.0$), $w_{\text{rama}} = 1.00 \pm 0.09$ ($\lambda = 0.1$), $w_{\text{rama}} = 0.92 \pm 0.1$ ($\lambda = 0.01$), and $w_{\text{rama}} = 0.84 \pm 0.15$ ($\lambda = 0.005$). If we reduce the number of data further, the system undergoes a phase transition because the posterior no longer peaks at the native ensemble, and w_{rama} is pushed toward zero. For all values of λ , the optimal weight falls in the range that results in a minimal RMSD to the crystal structure. The Bayesian choice of w_{rama} generates the most accurate structures.

Impact on structure ensembles from sparse and noisy NMR data

So far, we studied how Bayesian weighting of the backbone potential impacts the conformational ensemble under artificially sparsified data. We also looked at challenging real-world structure determination problems, a sparse set of distances for the Fyn-SH3 domain [23,29] and noisy distance bounds measured with solid-state NMR on the α -spectrin SH3 domain [30]. We estimated the Ramachandran weight for both data sets and obtained $w_{\text{rama}} = 0.76 \pm 0.11$ and $w_{\text{rama}} = 0.3 \pm 0.13$ for the sparse and noisy distances, respectively (see also Figure S7 in File S1).

Incorporation of the backbone potential can significantly improve the accuracy of the ensemble as measured by the RMSD

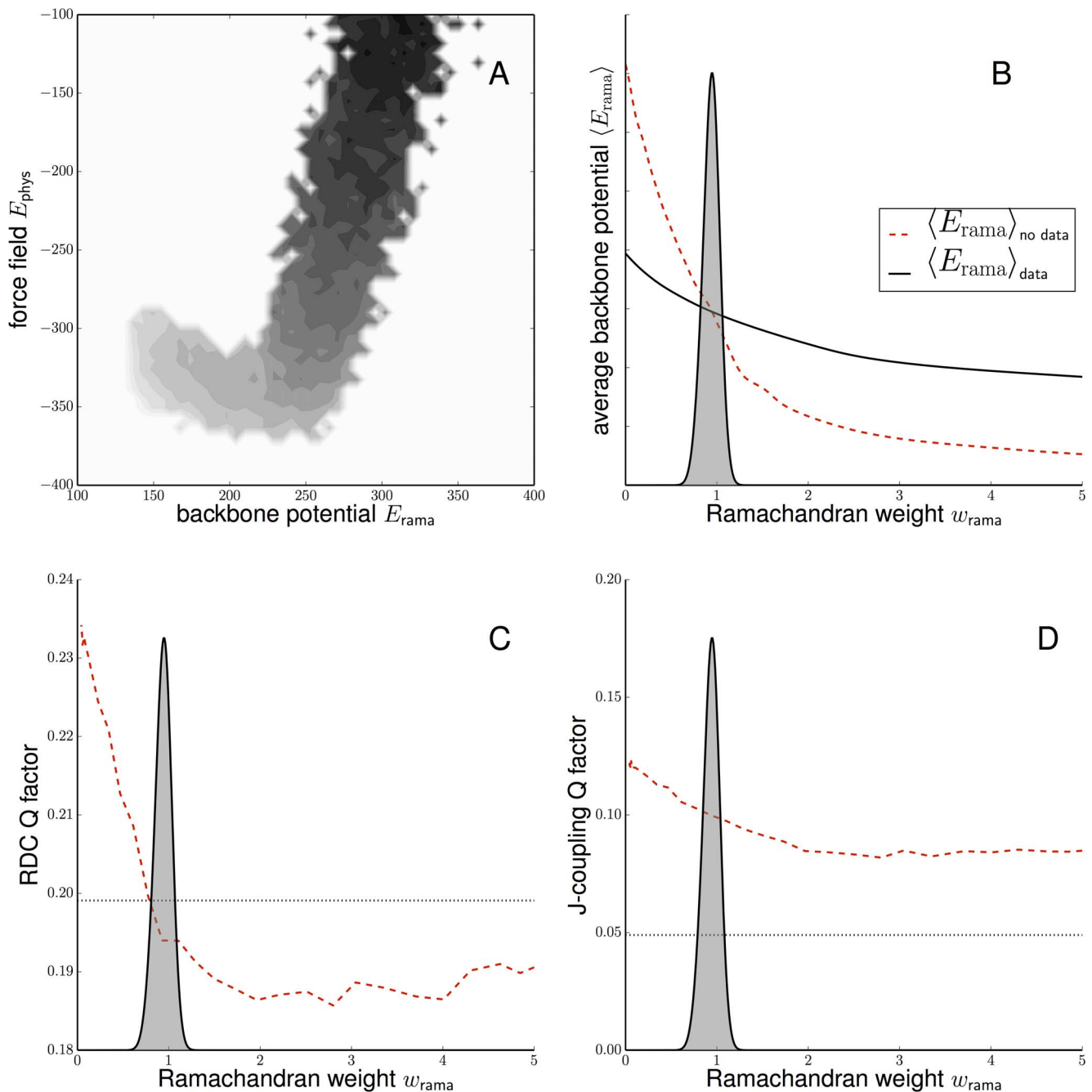


Figure 2. Bayesian weighting of the backbone potential for ubiquitin inferred from high-quality distance data. A: Correlation between backbone potential and nonbonded force field. Shown is the joint distribution of physics- and knowledge-based contributions in the absence of any structural data. (The energies of the crystal structure are $E_{\text{phys}} = -455.5$ and $E_{\text{rama}} = 187.3$.) B: Model evidence $\text{Pr}(D|w_{\text{rama}})$ as a function of the Ramachandran weight w_{rama} . C: Influence of the Ramachandran weight on the average Q-factor (red dashed line) calculated for 11 RDC data sets that were not used in the structure calculation. The Q-factor reflects the agreement between experimental and calculated RDCs. The dotted black line indicates the average Q-factor of the crystal structure (PDB code: 1ubq). D: Influence of the Ramachandran weight on the fit with scalar coupling measurements (red dashed line). Six three-bond scalar coupling data sets are available for ubiquitin and have not been used in the structure calculation. The dotted black line indicates the average Q-factor of the crystal structure (PDB code: 1ubq). The grey distribution indicates the model evidence $\text{Pr}(D|w_{\text{rama}})$. doi:10.1371/journal.pone.0100197.g002

to the crystal structure (Figure 4). But we also observe that an overly strong backbone potential can do more harm than good. For large w_{rama} values, the RMSD distribution deteriorates and even shows multiple peaks in case of the sparse data set. With an optimally weighted backbone potential also the accuracy of the

mean structure is consistently higher than the accuracy of the individual members of the structure ensemble (see Table S1 in File S1). This indicates that the structure ensembles are better defined when using the backbone potential. Figure 4 shows that the model

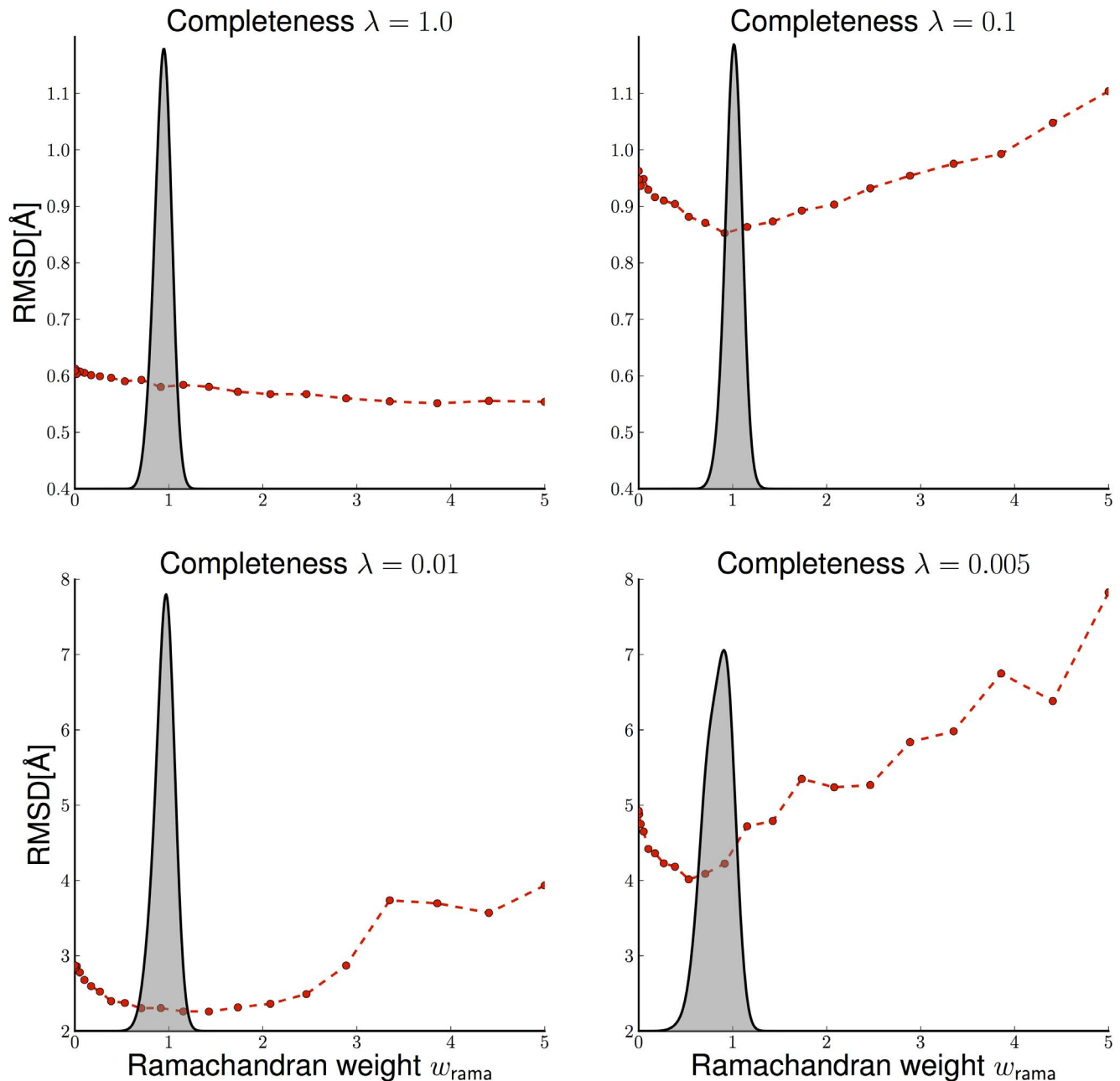


Figure 3. Impact of incomplete ubiquitin data on w_{rama} . Shown is the model evidence as a function of w_{rama} (grey) and the average RMSD (dots). The sparsity increases from the top left panel to the bottom right panel.
doi:10.1371/journal.pone.0100197.g003

evidence peaks where the average RMSD to the crystal structure shows a minimum.

Figure 5 shows the structure ensembles obtained with the sparse SH3 data for different choices of w_{rama} . If the weight is zero or too small, the ensemble is still quite heterogeneous, especially in the loops. When incorporating the backbone potential with an optimal weight, the ensemble becomes very regular and accurate: the average structure is surprisingly close to the reference structure (1.05 Å RMSD) given the sparseness of the data set. For too large weight, we introduce conflicts between the preferences of the statistical potential and the data by introducing helical structure in beta strands. The corresponding Ramachandran plots illustrate these findings. For $w_{\text{rama}} = 5$, the Ramachandran plot becomes

artificially narrow and peaks in the helical region. Another indication that the Bayesian choice of w_{rama} is optimal, is provided by the behavior of additional model parameters. Figure S8 in File S1 shows that the estimated weight of the data w_{data} [26] is largely unaffected if we incorporate the knowledge-based contribution. That is, we do not compromise the data by downweighting them upon integration of the backbone potential.

In Figure 6 we plot the effective potential function (i.e. the negative log-posterior probability) incorporating the force field, the backbone potential and the goodness-of-fit for different values of w_{rama} corresponding to the ensembles shown in Figure 5. The results indicate that optimal weighting helps to guide the simulation towards more accurate structures. Without the

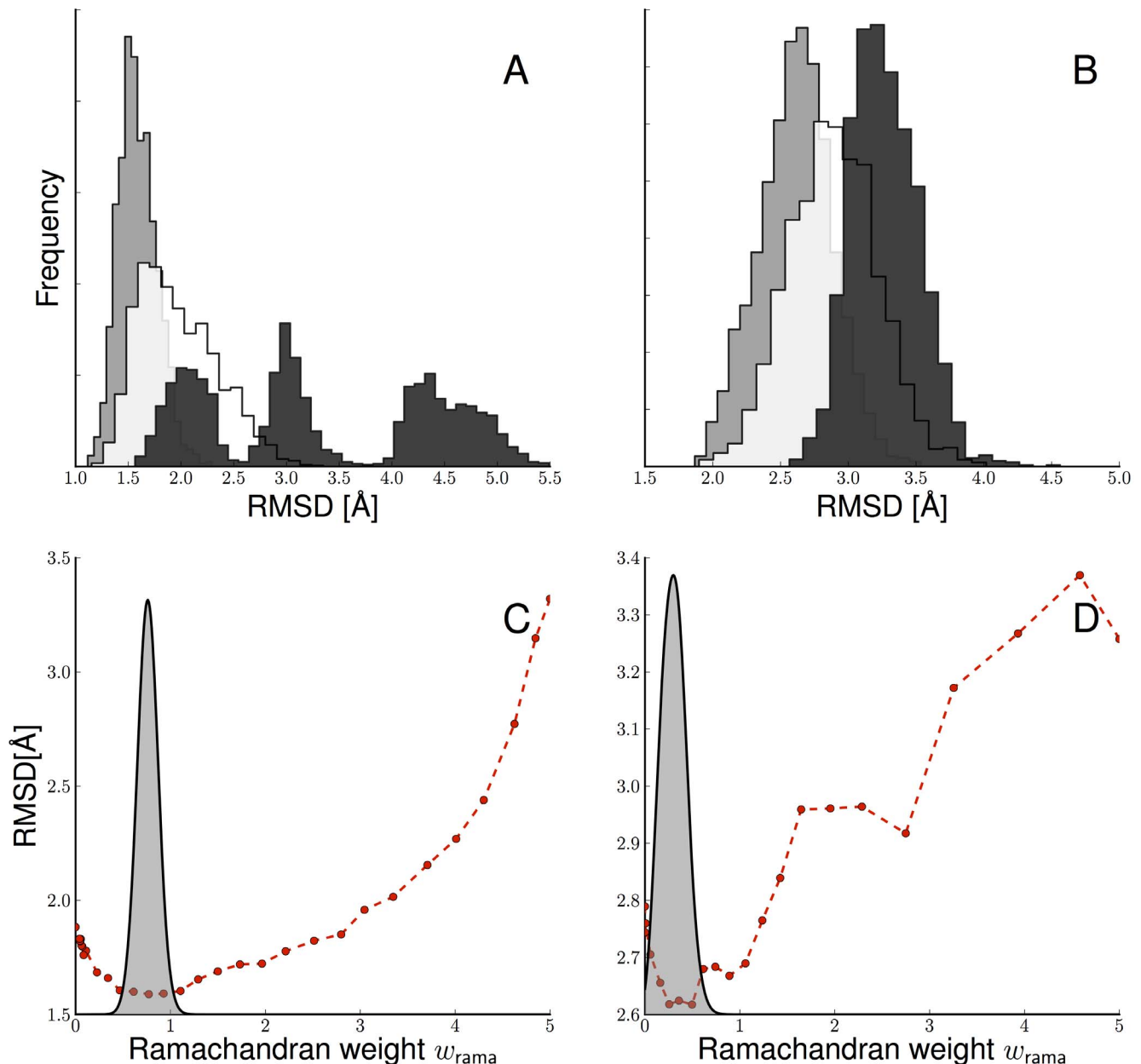


Figure 4. Impact on structure ensembles from sparse and noisy NMR data. Panels A, C show the results for the sparse Fyn-SH3 data set. Panels B, D show the results for the solid-state data. The top row displays the RMSD distributions with $w_{\text{rama}} = 0$ (white), $w_{\text{rama}} = 5$ (black) and optimal w_{rama} (grey). The grey distribution shown in the bottom panels is the model evidence as a function of the weight w_{rama} . doi:10.1371/journal.pone.0100197.g004

backbone potential, structures at the bottom of the energy funnel show a broad range of RMSDs between 1.5 and 2.5 Å. For optimal w_{rama} , the funnel narrows and selects structures with an RMSD below 1.5 Å. With too large w_{rama} we observe a negative correlation between the RMSD and the negative log-posterior probability.

Impact on structure quality

Figure 7 shows the average values of several validation criteria for structures generated at different weights (a full report of the Procheck and WhatCheck quality criteria can be found in Table S1 in File S1). All reported criteria are z-scores that provide an assessment of how a particular structure compares to the average

of all known protein structures in terms of standard deviations. That is, a z-score of -0.5 , say, means that the quality criterion of this particular structure is half a standard deviation below the average.

As expected, the Ramachandran score (RAMCHK) improves with increasing w_{rama} and shows no saturation, which demonstrates that it is not a valid indicator for selecting w_{rama} . The effect of the backbone potential on the NQACHK score is small; only for ubiquitin it agrees with the Bayesian choice of w_{rama} . WhatCheck's packing score QUACHK shows a different behavior, it increase steadily with increasing w_{rama} . BBCCHK assesses the regularity of the backbone and shows a steady increase with increasing w_{rama} indicating a high correlation with the potential

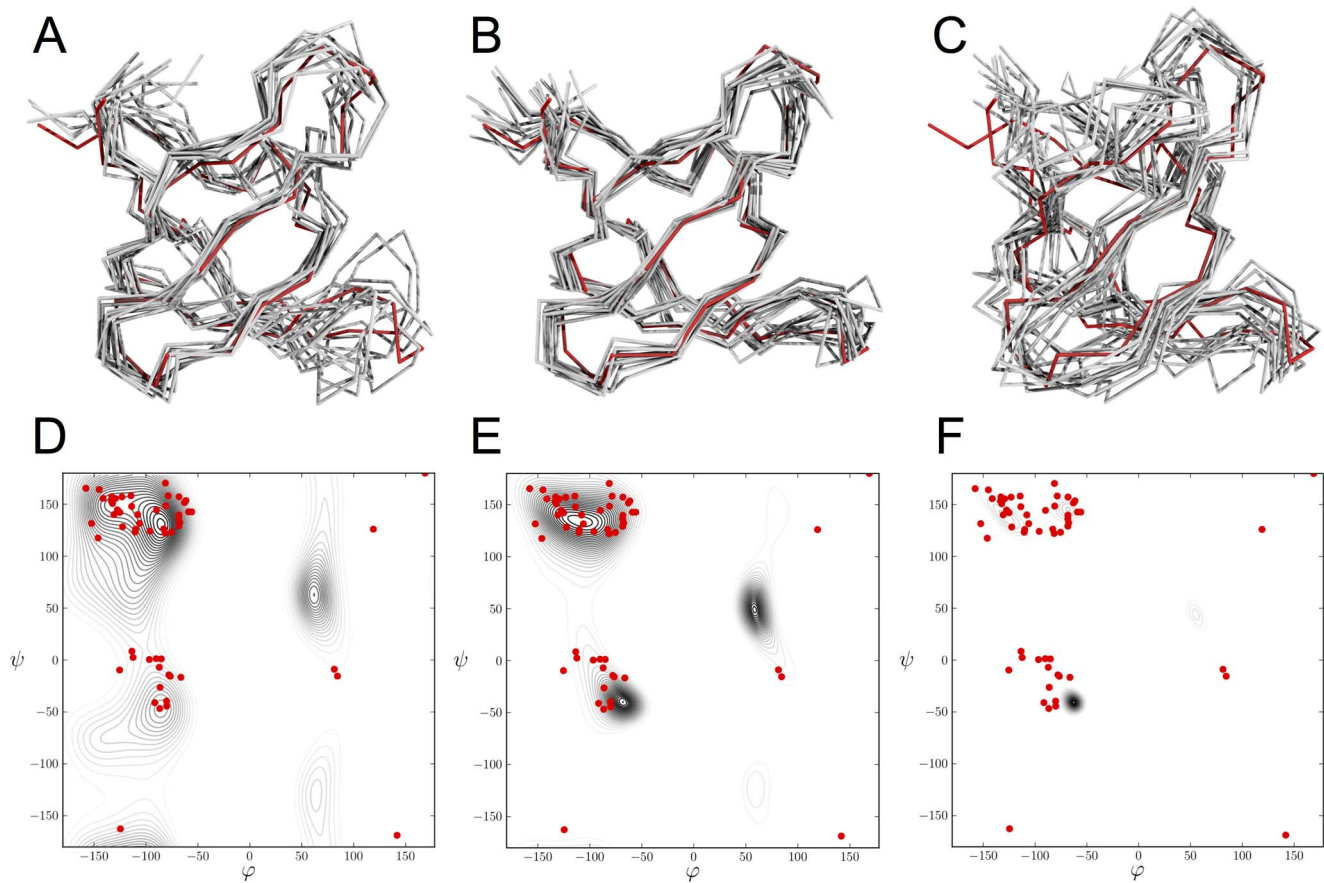


Figure 5. Influence of the weight w_{rama} on the structural ensemble of Fyn-SH3 inferred with sparse NMR data. Shown are the conformations and backbone dihedral distributions generated with different w_{rama} . Panels A–C display structure ensembles comprising ten randomly selected conformations (grey) superimposed onto the crystal structure (red). Panels D–F show in black a maximum entropy distribution fitted to the backbone torsion angles of the structures generated with ISD. The backbone dihedral angles of the crystal structure are marked by red dots. Panels A, D show the results for $w_{\text{rama}} = 0.0$, panels B, E: $w_{\text{rama}} = 0.76$ (optimal weight), panels C, F: $w_{\text{rama}} = 5.0$ (maximum weight probed during replica-exchange simulations).

doi:10.1371/journal.pone.0100197.g005

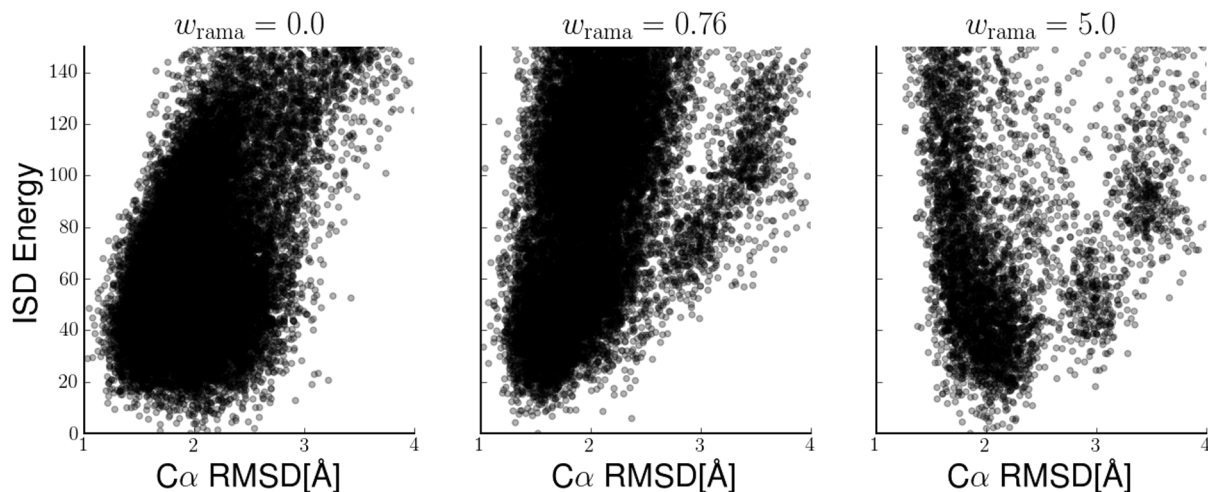


Figure 6. Energy funnels obtained with the sparse Fyn-SH3 data at different Ramachandran weights (left: $w_{\text{rama}} = 0.0$, middle: $w_{\text{rama}} = 0.76$, right: $w_{\text{rama}} = 5.0$). The full ISD energy (negative log-posterior probability) is plotted against the RMSD to the Fyn-SH3 crystal structure.

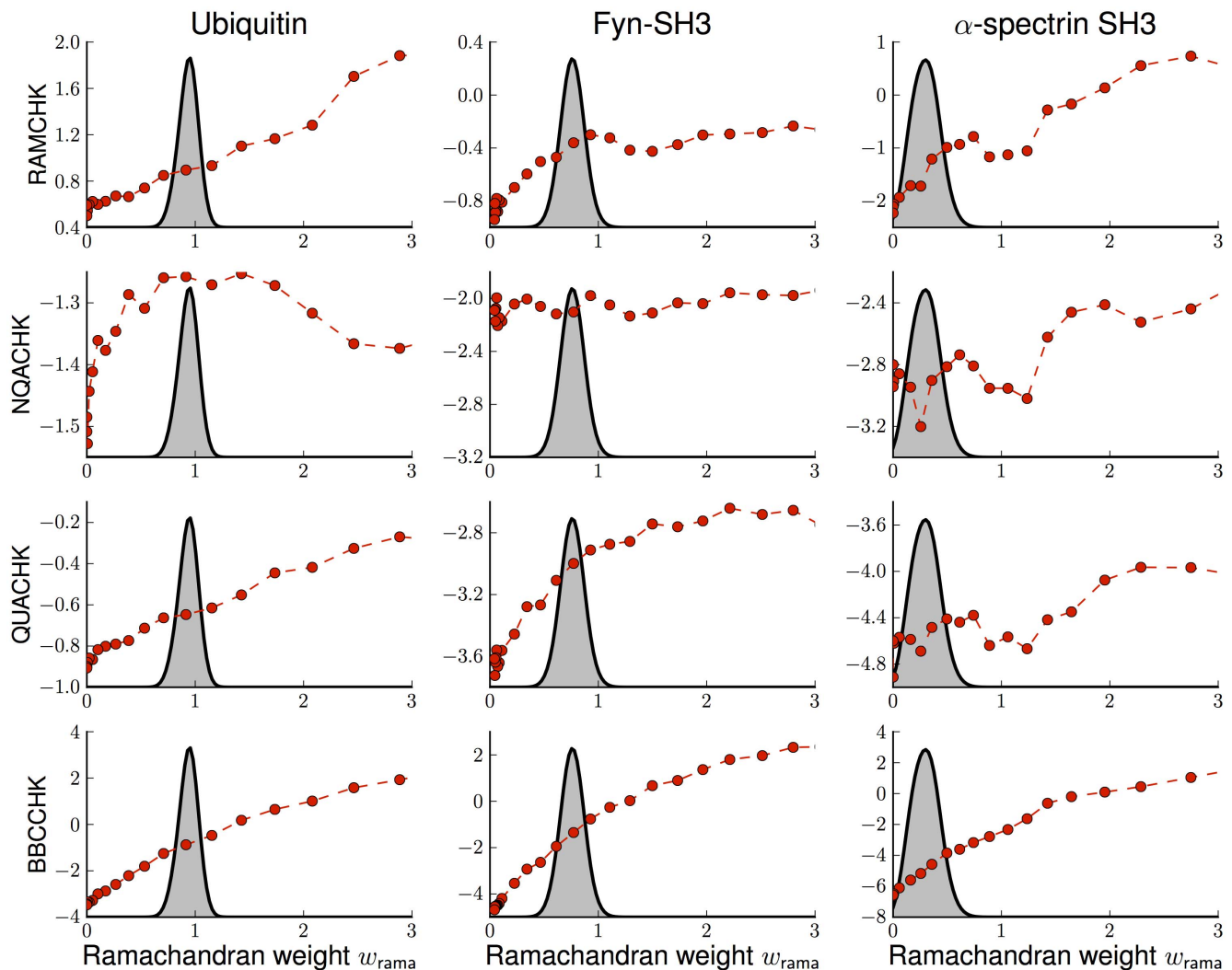


Figure 7. Influence of the Ramachandran weight on various quality criteria. Shown is the impact of w_{rama} on WhatCheck validation criteria. Each column reports the results for a different data set (left column: ubiquitin, middle column: Fyn-SH3 domain, right column: α -spectrin SH3 domain). Each row shows the evolution of a quality score with increasing w_{rama} (each dot marks the average over 100 structures that were randomly selected from the ISD ensemble, dashed lines are added to guide the eye). The first row reports the Ramachandran appearance as assessed by RAMCHK. The second and third row show WhatCheck's packing scores. The last row reports the regularity of the backbone (BBCCHK). The grey distribution indicates the model evidence $\Pr(D|w_{\text{rama}})$ as a function of w_{rama} . doi:10.1371/journal.pone.0100197.g007

Ramachandran potential. Although the ubiquitin data are of a high quality, the structure ensemble still improves upon incorporation of the backbone potential. As for ubiquitin, we observe an overall improvement in the quality of the ensembles obtained with the sparse and noisy SH3 data. The slight decrease of the average NQACHK score for the optimal weight ensemble in the case alpha-spectrin is within the ensemble spread; again the score is only weakly affected by the Ramachandran weight (see Figure S5 in File S1).

The WHATCHECK validation criteria by themselves are unable to choose a w_{rama} that would lead to a global improvement of the structural quality as well as the accuracy of the ensemble. Moreover, although some of the scores seem to be highly correlated (e.g. RAMCHK and BBCCHK, see Figure S6 in File S1) it is not clear whether it is possible to maximize all scores simultaneously. Rather we have to find a compromise between the different quality criteria, and this is exactly what our weighting scheme achieves.

Discussion

We outline a new formalism to integrate physics- with knowledge-based potential functions in biomolecular structure calculation. Our approach is founded on Bayesian principles and allows us to incorporate prior knowledge derived from structure databases. The new method is data-driven and adaptively weighs knowledge-based contributions relative to the force field and the experimental data thereby reducing potential bias and artifact. We show that even simple terms such as knowledge-based backbone potentials have a significant effect on the quality of the structure ensemble. The optimally weighted dihedral angle potential improves not only the Ramachandran appearance but also the backbone normality and packing scores. Moreover, it systematically produces more accurate and more precise structure ensembles. The combination of physics- and knowledge-based potential functions is particularly powerful for sparse and noisy NMR data and shifts the ensemble closer to the native structure.

Our findings suggest that there is no universal weight that can be transferred to all proteins and data sets. Rather, the method of choice is to estimate the weight in the course of the structure calculation.

For the solid-state SH3 data, we find the smallest weight $w_{\text{rama}} \approx 0.3$; for the ubiquitin data we estimate the largest weight $w_{\text{rama}} \approx 1$. We can rationalize this behavior as follows. For high-quality data (high completeness, low noise level), the optimal weight adopts a large value and drops as noise and sparseness increase [27]. This behavior seems counterintuitive at first sight but is sensible: The forces that pull the ensemble towards the correct structure are weaker with low-quality than with high-quality data. We therefore have to soften the backbone potential to not overwhelm the data.

In the future, we plan to extend our method to weigh multiple statistical energy terms simultaneously in the course of a structure calculation. However, this will require a more efficient algorithm for estimating multi-dimensional densities of states, because the computational complexity of the approach becomes prohibitive. The naive extension of the presented approach would involve a multi-dimensional replica-exchange simulation in which replicas are introduced for every combination of the weights that we want to estimate. Therefore the computational burden grows exponentially in the number of weights such that with our current algorithm it is only possible to estimate up to two or three weights. The final goal is to design an efficient, unbiased but highly informative conformational prior distribution that allows the calculation of high quality ensembles from very sparse data sets.

Materials and Methods

Data sets

Backbone dihedral angles were extracted from PDBselect25 [31] and used to estimate angular distributions for all amino acids using the maximum entropy method (see next section). We illustrate the impact of the backbone potential on three NMR data sets. The first data set (PDB code: 1d3z) comprises high-quality data for ubiquitin. All distance data were reduced to 1444 non-redundant restraints, additional data (scalar coupling constants, residual dipolar couplings) were not included in the structure calculation but used for validation. The second data set comprises sparse distance data for the Fyn-SH3 domain [23,29] (PDB code: 1zbj). The third data set has been measured with solid-state NMR on the α -spectrin SH3 domain [30] (PDB code: 1m8m). The solid-state data are very generous distance bounds ranging from 4.5 Å to 7.5 Å out of which 90% are equal or greater than 6 Å, which is the largest distance bound obtained in standard solution NMR.

Maximum entropy distributions for backbone dihedral angles

Following Pertsemlidis *et al.* [22], we use a maximum entropy distribution with a Fourier basis to describe the distribution of backbone dihedral angles:

$$p(\varphi, \psi) = \frac{1}{Z(a, b, c, d)} \exp\{-E_{\text{rama}}(\varphi, \psi)\} \quad (2)$$

where the Ramachandran potential $E_{\text{rama}}(\varphi, \psi)$ is given by

$$E_{\text{rama}}(\varphi, \psi) = \sum_{i,j=1}^k [a_{ij} \cos(i\varphi) \cos(j\psi) + b_{ij} \cos(i\varphi) \sin(j\psi) + c_{ij} \sin(i\varphi) \cos(j\psi) + d_{ij} \sin(i\varphi) \sin(j\psi)] \quad (3)$$

$Z(a, b, c, d)$ normalizes the dihedral angle distribution, and $k = 5$ is the order of the Fourier expansion (see Figure S1 in File S1). We fit the expansion coefficients a, b, c, d to observed φ/ψ pairs by using the maximum entropy method, which, in this case, is equivalent to maximum likelihood parameter estimation. We approximate the normalization constant $Z(a, b, c, d)$ using the two-dimensional trapezoidal rule. To avoid over-fitting of the φ/ψ distributions, we introduce a Gaussian prior with unknown precision λ over the expansion coefficients:

$$p(a, b, c, d | \lambda) = \left(\frac{\lambda}{2\pi}\right)^{2k(k-1)} \exp\left\{-\frac{\lambda}{2} \sum_{i=1}^k \sum_{j=1}^k (a_{ij}^2 + b_{ij}^2 + c_{ij}^2 + d_{ij}^2)\right\} \quad (4)$$

The precision of the prior λ is not known and is estimated simultaneously with the expansion coefficients. We use an iterative scheme in which we cycle through updates of the expansion coefficients and of the precision. For fixed precision, the negative log-posterior probability of the expansion coefficients is a convex function, which we optimize using the Powell minimizer [32]. The update of the precision can be calculated analytically.

Inferential structure determination

Inferential structure determination (ISD) [23,24] is a probabilistic framework for biomolecular structure calculation from experimental data. ISD uses Bayes's theorem to obtain a posterior distribution over all unknown parameters including the conformational degrees of freedom θ (typically main and side chain torsions) and all additional model parameters α [33]. The posterior distribution is proportional to the product of two terms, the likelihood function and the prior probability distribution. The likelihood function, $\Pr(D|\theta, \alpha)$, is the conditional probability of the experimental data D viewed as a function of the parameters θ and α . Here we consider NMR distance measurements, which we model using the lognormal distribution [34]. This introduces two model parameters, the distance scale and error, that we estimate simultaneously with the structure. We use the lognormal model for the analysis of the 1d3z and 1zbj data, which provide distance restraints. In case of the solid-state data (PDB code: 1m8m), only lower and upper distance bounds are given. We apply a new probabilistic model (Habeck, in preparation) that estimates a set of experimental distances falling between the lower and upper bounds. For given estimated experimental distances, the lognormal model is used to relate the distance data with the structure. This model has in addition to the set of unknown experimental distances only one model parameter, the distance error.

Optimization of the potential function

We aim to find an optimal potential function for each dataset by adjusting the influence of the backbone potential E_{rama} . The combined potential function is given by $w_{\text{phys}} E_{\text{phys}} + w_{\text{rama}} E_{\text{rama}}$ where $w_{\text{phys}} = 1/k_B T$ is the reciprocal temperature involving Boltzmann's constant k_B and the absolute temperature T . Here E_{phys} was set to the Lennard-Jones potential adapted from the Rosetta software [35]. The only free parameter is the weight of the

backbone potential w_{rama} . The model evidence $\Pr(D|w_{\text{rama}})$ can be interpreted as the probability of the experimental data for a particular w_{rama} . The optimization of $\Pr(D|w_{\text{rama}})$ is demanding as the calculation depends on an intractable high-dimensional integral:

$$\Pr(D|w_{\text{rama}}) = \int \Pr(D|\theta, \alpha, w_{\text{rama}}) \Pr(\theta|w_{\text{rama}}) \Pr(\alpha) d\theta d\alpha.$$

over all model parameters α and conformational degrees of freedom θ . Here $\Pr(\theta|w_{\text{rama}}) \propto \exp\{-w_{\text{phys}}E_{\text{phys}}(\theta) - w_{\text{rama}}E_{\text{rama}}(\theta)\}$ is the combined prior probability of conformation θ for a given weight w_{rama} . We can reduce the computation to a low-dimensional integral by expressing the above equation using the density of states $g_z(E_{\text{rama}})$

$$\Pr(D|w_{\text{rama}}) = \frac{\int g_1(E_{\text{rama}}) e^{-w_{\text{rama}}E_{\text{rama}}} dE_{\text{rama}}}{\int g_0(E_{\text{rama}}) e^{-w_{\text{rama}}E_{\text{rama}}} dE_{\text{rama}}}$$

The density of states is given as

$$g_z(E_{\text{rama}}) = \int \delta(E_{\text{rama}} - E_{\text{rama}}(\theta)) [\Pr(D|\theta, \alpha)]^z \times \Pr(\alpha) e^{-w_{\text{phys}}E_{\text{phys}}(\theta)} d\theta d\alpha$$

where $\delta(\cdot)$ denotes the Dirac delta function. Estimates of the density of states are obtained by applying multiple histogram reweighting [36,37] as outlined in [27].

References

- Brünger AT, Nilges M (1993) Computational challenges for macromolecular structure determination by X-ray crystallography and solution NMR-spectroscopy. *Q Reviews of Biophys* 26: 49–125.
- Skolnick J (2006) In quest of an empirical potential for protein structure prediction. *Curr Opin Struct Biol* 16: 166–171.
- Ponder JW, Case DA (2003) Force fields for protein simulations. *Adv Protein Chem* 66: 27–85.
- Sippl MJ (1995) Knowledge-based potentials for proteins. *Curr Opin Struct Biol* 5: 229–235.
- Kuszewski J, Gronenborn AM, Clore GM (1996) Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. *Protein Sci* 5: 1067–1080.
- Kuszewski J, Gronenborn AM, Clore GM (1997) Improvements and extensions in the conformational database potential for the refinement of NMR and X-ray structures of proteins and nucleic acids. *J Magn Reson* 125: 171–177.
- Grishaev A, Wu J, Trewhella J, Bax A (2005) Refinement of multidomain protein structures by combination of solution small-angle X-ray scattering and NMR data. *J Am Chem Soc* 127: 16621–16628.
- Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations. *J Mol Biol* 7: 95–99.
- Linge JP, Nilges M (1999) Influence of non-bonded parameters on the quality of NMR structures: a new force-field for NMR structure calculation. *J Biomol NMR* 13: 51–59.
- Hovmöller S, Zhou T, Ohlson T (2002) Conformations of amino acids in proteins. *Acta Cryst sect D* 58: 768–776.
- Doreleijers JF, Rullmann JA, Kaptein R (1998) Quality assessment of NMR structures: a statistical survey. *J Mol Biol* 281: 149–64.
- Nederveen AJ, Doreleijers JF, Vranken W, Miller Z, Spronk CA, et al. (2005) RECOORD: a recalculated coordinate database of 500+ proteins from the PDB using restraints from the BioMagResBank. *Proteins* 59: 662–672.
- Markley JL, Bax A, Arata Y, Hilbers CW, Kaptein R, et al. (1998) Recommendations for the presentation of NMR structures of proteins and nucleic acids. *J Mol Biol* 280: 933–952.
- Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Cryst* 26: 283–291.
- Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, et al. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* 66: 12–21.
- Hooft RW, Vriend G, Sander C, Abola EE (1996) Errors in protein structures. *Nature* 381: 272.
- Gong H, Shen Y, Rose GD (2007) Building native protein conformation from NMR backbone chemical shifts using Monte Carlo fragment assembly. *Protein Sci* 16: 1515–1521.
- Mardia KV, Taylor CC, Subramaniam GK (2007) Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics* 63: 505–512.
- Boomsma W, Mardia K, Taylor C, Ferkinghoff-Borg J, Krogh A, et al. (2008) A generative, probabilistic model of local protein structure. *Proc Natl Acad Sci USA* 105: 8932–8937.
- Ting D, Wang G, Shapovalov M, Mitra R, Jordan MI, et al. (2010) Neighbor-dependent Ramachandran probability distributions of amino acids developed from a hierarchical Dirichlet process model. *PLoS Comput Biol* 6: e1000763.
- Kuszewski J, Clore GM (2000) Sources of and solutions to problems in the refinement of protein NMR structures against torsion angle potentials of mean force. *J Magn Reson* 146: 249–254.
- Pertselmidis A, Zelinka J, Fondon JW, Henderson RK, Otwinowski Z (2005) Bayesian statistical studies of the Ramachandran distribution. *Stat Appl Genet Mol Biol* 4: Article35.
- Rieping W, Habeck M, Nilges M (2005) Inferential Structure Determination. *Science* 309: 303–306.
- Rieping W, Nilges M, Habeck M (2008) ISD: a software package for Bayesian NMR structure calculation. *Bioinformatics* 24: 1104–1105.
- Porter LL, Rose GD (2011) Redrawing the Ramachandran plot after inclusion of hydrogen-bonding constraints. *Proc Natl Acad Sci USA* 108: 109–113.
- Habeck M, Rieping W, Nilges M (2006) Weighting of experimental evidence in macromolecular structure determination. *Proc Natl Acad Sci USA* 103: 1756–1761.
- Mechelke M, Habeck M (2012) Calibration of boltzmann distribution priors in bayesian data analysis. *Physical Review E* 86: 066705.
- Habeck M (2011) Statistical mechanics analysis of sparse data. *J Struct Biol* 173: 541–548.
- Mal TK, Matthews SJ, Kovacs H, Campbell ID, Boyd J (1998) Some NMR experiments and a structure determination employing a $f15N$, $2H$ enriched protein. *J Biomol NMR* 12: 259–276.
- Castellani F, van Rossum B, Diehl A, Schubert M, Rehbein K, et al. (2002) Structure of a protein determined by solid-state magic-angle-spinning NMR spectroscopy. *Nature* 420: 98–102.

Replica-exchange Monte Carlo

To estimate the density of states we generate conformations using an extended replica-exchange Monte Carlo scheme (REMC). The idea behind REMC is to simulate a system in parallel at different temperatures [38]. In our scheme, two temperature parameters control the force field and the likelihood function independently [39]. We treat w_{rama} as a third temperature parameter. Across the first 30 replicas w_{rama} decreases from five to zero while the force field and the data are fully taken into account. In the remaining 50 replicas, the force field and the data are gradually switched off as described by Habeck *et al.* [39]. The convergence of the method depends on the size of the system as well as the quality and quantity of the data. In case of the tested systems, convergence was achieved after 30000 to 100000 replica transitions corresponding to 2–3 days worth of computation on a 80 node cluster depending on the size of the system.

Supporting Information

File S1 Supporting file including supporting text, Figures S1–S8, and Table S1.
(PDF)

Author Contributions

Conceived and designed the experiments: MH. Performed the experiments: MM. Analyzed the data: MM MH. Contributed reagents/materials/analysis tools: MM MH. Wrote the paper: MM MH.

31. Hobohm U, Scharf R M, Schneider, Sander C (1992) Selection of a representative set of structures from the Brookhaven protein data bank. *Protein Sci* 1: 409–417.
32. Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1989) *Numerical Recipes: The Art of Scientific Computing*. Cambridge UK: Cambridge University Press.
33. Habeck M, Nilges M, Rieping W (2005) Bayesian inference applied to macromolecular structure determination. *Phys Rev E* 72: 031912.
34. Rieping W, Habeck M, Nilges M (2005) Modeling errors in NOE data with a lognormal distribution improves the quality of NMR structures. *J Am Chem Soc* 127: 16026–16027.
35. Kuhlman B, Dantas G, Ireton GC, Varani G, Stoddard BL, et al. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science* 302: 1364–1368.
36. Ferrenberg AM, Swendsen RH (1989) Optimized Monte Carlo Data Analysis. *Phys Rev Lett* 63: 1195–1198.
37. Habeck M (2012) Bayesian estimation of free energies from equilibrium simulations. *Phys Rev Lett* 109: 100601.
38. Swendsen RH, Wang JS (1986) Replica Monte Carlo simulation of spin glasses. *Phys Rev Lett* 57: 2607–2609.
39. Habeck M, Nilges M, Rieping W (2005) Replica-Exchange Monte Carlo scheme for Bayesian data analysis. *Phys Rev Lett* 94: 0181051–0181054.