Science Advances

AAAS

advances.sciencemag.org/cgi/content/full/2/1/e1501177/DC1

Supplementary Materials for

Metainference: A Bayesian inference method for heterogeneous systems

Massimiliano Bonomi, Carlo Camilloni, Andrea Cavalli, Michele Vendruscolo

Published 22 January 2016, *Sci. Adv.* **2**, e1501177 (2016)
DOI: 10.1126/sciadv.1501177

The PDF file includes:

Supplementary Materials

Derivation of the basic metainference equations

1) The metainference posterior in the case of a single experimental data point. Here we derive Eq. 5 of the main text, which is the general metainference equation in the case of a single experimental data point d. As discussed in the main text (Materials and Methods), we are interested in determining how the prior distribution of models (including structural states and other parameters) is affected by the introduction of experimental information. Since experimental data in equilibrium conditions are the result of ensemble averages over a distribution of states, we model a finite sample of the distribution of models, which we refer to as the set of N replicas of the system. These include: the coordinates of the system $\mathbf{X} = [X_r]$, the averages of the forward model over an infinite number of replicas $\tilde{\boldsymbol{f}} = [\tilde{f}_r]$, the uncertainty parameters that describes random and systematic errors in the experimental data as well as errors in the forward model $\boldsymbol{\sigma}^{\boldsymbol{B}} = [\sigma_r^B]$, the standard errors of the mean $\boldsymbol{\sigma}^{SEM} = [\sigma_r^{SEM}]$.
The metainference posterior probability is thus

$$p(\mathbf{X}, \tilde{\boldsymbol{f}}, \boldsymbol{\sigma}^{\boldsymbol{B}}, \boldsymbol{\sigma}^{\boldsymbol{SEM}} \mid d, I) \tag{S1}$$

We first recognize that $\mathbf{X}$ and $\boldsymbol{\sigma}^{SEM}$ do not dependent from the data d. Therefore

$$p(\mathbf{X}, \tilde{\boldsymbol{f}}, \boldsymbol{\sigma}^{\boldsymbol{B}}, \boldsymbol{\sigma}^{\boldsymbol{SEM}} \mid d, I) = p(\tilde{\boldsymbol{f}}, \boldsymbol{\sigma}^{\boldsymbol{B}} \mid d, I) \cdot p(\mathbf{X}) \cdot p(\boldsymbol{\sigma}^{\boldsymbol{SEM}}) \tag{S2}$$

At this point, we should take into account that each set $\tilde{\boldsymbol{f}} = [\tilde{f}_r]$, $\boldsymbol{\sigma}^{\boldsymbol{B}} = [\sigma_r^B]$, and $\boldsymbol{\sigma}^{SEM} = [\sigma_r^{SEM}]$ is composed of independent variables, and that the configurations $\mathbf{X} = [X_r]$ are a priori independent. Given these considerations, we can write from Eq. S2

$$p(\mathbf{X}, \tilde{\boldsymbol{f}}, \boldsymbol{\sigma}^{\boldsymbol{B}}, \boldsymbol{\sigma}^{\boldsymbol{SEM}} \mid d, I) = \prod_{r=1}^{N} p(\tilde{f}_r, \sigma_r^B \mid d, I) \cdot p(X_r) \cdot p(\sigma_r^{SEM}) \tag{S3}$$

By applying Bayes theorem to $p(\tilde{f}_r, \sigma_r^B \mid d, I)$ we can thus derive Eq. 5 of the main text

$$p(\mathbf{X}, \tilde{\boldsymbol{f}}, \boldsymbol{\sigma}^{\boldsymbol{B}}, \boldsymbol{\sigma}^{\boldsymbol{SEM}} \mid d, I) \propto \prod_{r=1}^{N} p(d \mid \tilde{f}_r, \sigma_r^B) \cdot p(\tilde{f}_r \mid \mathbf{X}, \sigma_r^{SEM}) \cdot p(\sigma_r^B) \cdot p(X_r) \cdot p(\sigma_r^{SEM}) \tag{S4}$$

2) Gaussian data noise and marginalization. We can further simply Eq. S4 in the case of Gaussian data likelihood

$$p(d \mid \tilde{f}_r, \sigma_r^B) = \frac{1}{\sqrt{2\pi}\sigma_r^B} \cdot \exp\left[-\frac{\left(d - \tilde{f}_r\right)^2}{2\left(\sigma_r^B\right)^2} \right] \tag{S5}$$

In this case, we can write

$$p(d \mid \tilde{f}_r, \sigma_r^B) \cdot p(\tilde{f}_r \mid \mathbf{X}, \sigma_r^{SEM}) = \frac{1}{\sqrt{2\pi}\sigma_r^B} \cdot \exp\left[-\frac{\left(d - \tilde{f}_r\right)^2}{2\left(\sigma_r^B\right)^2} \right] \cdot \frac{1}{\sqrt{2\pi}\sigma_r^{SEM}} \cdot \exp\left[-\frac{\left(\tilde{f}_r - f(\mathbf{X})\right)^2}{2\left(\sigma_r^{SEM}\right)^2} \right] \quad \text{(S6)}$$

The product of the two Gaussian probability density functions (PDFs) is a scaled Gaussian PDF

$$p(d \mid \tilde{f}_r, \sigma_r^B) \cdot p(\tilde{f}_r \mid \mathbf{X}, \sigma_r^{SEM}) = \frac{S}{\sqrt{2\pi}\bar{\sigma}} \cdot \exp\left[-\frac{\left(\tilde{f}_r - \bar{f}\right)^2}{2\bar{\sigma}^2} \right] \quad \text{(S7)}$$

where

$$\bar{\sigma} = \sqrt{\frac{\left(\sigma_r^B\right)^2 \cdot \left(\sigma_r^{SEM}\right)^2}{\left(\sigma_r^B\right)^2 + \left(\sigma_r^{SEM}\right)^2}} \text{ and } \bar{f} = \frac{d \cdot \left(\sigma_r^{SEM}\right)^2 + f(\mathbf{X}) \cdot \left(\sigma_r^B\right)^2}{\left(\sigma_r^B\right)^2 + \left(\sigma_r^{SEM}\right)^2} \quad \text{(S8)}$$

The scaling factor is itself a Gaussian PDF

$$S = \frac{1}{\sqrt{2\pi\left(\left(\sigma_r^{SEM}\right)^2 + \left(\sigma_r^B\right)^2\right)}} \cdot \exp\left[-\frac{\left(d - f(\mathbf{X})\right)^2}{2\left(\left(\sigma_r^{SEM}\right)^2 + \left(\sigma_r^B\right)^2\right)} \right] \quad \text{(S9)}$$

Since typically we are not interested in determining $\tilde{f}_r$, we can marginalize it as

$$\int p(d \mid \tilde{f}_r, \sigma_r^B) \cdot p(\tilde{f}_r \mid \mathbf{X}, \sigma_r^{SEM}) \cdot d\tilde{f}_r = S = \frac{1}{\sqrt{2\pi}\sigma_r} \cdot \exp\left[-\frac{\left(d - f(\mathbf{X})\right)^2}{2\sigma_r^2} \right] \quad \text{(S10)}$$

where the effective uncertainty parameters $\sigma_r = \sqrt{\left(\sigma_r^{SEM}\right)^2 + \left(\sigma_r^B\right)^2}$ encodes all sources of error. If we incorporate Eq. S10 into Eq. S4 we obtain the marginalized version of Eq. 5 that holds for Gaussian data noise (Eq. 6 in the main text).

3) The metainference posterior in the case of multiple independent data points. We now extend Eq. S4 to the case of N_d independent data points $\mathbf{D} = [d_i]$. We thus introduce one $\tilde{\boldsymbol{f}}$, $\sigma_{r,i}^B$, and $\sigma_{r,i}^{SEM}$ per data point i and replica r. In this case $\tilde{\boldsymbol{f}} = [[\tilde{f}_{r,i}]]$, $\boldsymbol{\sigma}^{\boldsymbol{B}} = [[\sigma_{r,i}^B]]$, and $\boldsymbol{\sigma}^{SEM} = [[\sigma_{r,i}^{SEM}]]$.

Since $\mathbf{X}$ and $\boldsymbol{\sigma}^{SEM}$ do not dependent from the data $\mathbf{D}$, the posterior can be written as

$$p(\mathbf{X}, \tilde{\boldsymbol{f}}, \boldsymbol{\sigma}^{\boldsymbol{B}}, \boldsymbol{\sigma}^{\boldsymbol{SEM}} \mid \mathbf{D}, I) = p(\tilde{\boldsymbol{f}}, \boldsymbol{\sigma}^{\boldsymbol{B}} \mid \mathbf{D}, I) \cdot p(\mathbf{X}) \cdot p(\boldsymbol{\sigma}^{\boldsymbol{SEM}}) \quad \text{(S11)}$$

Each set $\tilde{\boldsymbol{f}} = [[\tilde{f}_{r,i}]]$, $\boldsymbol{\sigma^B} = [[\sigma^B_{r,i}]]$, and $\boldsymbol{\sigma}^{SEM} = [[\sigma^{SEM}_{r,i}]]$ is composed of independent variables, and the configurations $\mathbf{X} = [X_r]$ are a priori independent. Therefore we can write

$$p(\mathbf{X}, \tilde{\boldsymbol{f}}, \boldsymbol{\sigma^B}, \boldsymbol{\sigma^{SEM}} \mid \mathbf{D}, I) = \prod_{r=1}^{N} \prod_{i=1}^{N_d} p(\tilde{f}_{r,i}, \sigma^B_{r,i} \mid \mathbf{D}, I) \cdot p(\sigma^{SEM}_{r,i}) \cdot \prod_{r=1}^{N} p(X_r) \quad \text{(S12)}$$

By applying Bayes theorem to the data likelihood $p(\tilde{f}_{r,i}, \sigma^B_{r,i} \mid \mathbf{D}, I)$, we can write

$$p(\mathbf{X}, \tilde{\boldsymbol{f}}, \boldsymbol{\sigma^B}, \boldsymbol{\sigma^{SEM}} \mid \mathbf{D}, I) \propto \prod_{r=1}^{N} \prod_{i=1}^{N_d} p(\mathbf{D} \mid \tilde{f}_{r,i}, \sigma^B_{r,i}) \cdot p(\tilde{f}_{r,i} \mid \mathbf{X}, \sigma^{SEM}_{r,i}) \cdot p(\sigma^B_{r,i}) \cdot p(\sigma^{SEM}_{r,i}) \cdot \prod_{r=1}^{N} p(X_r) \quad \text{(S13)}$$

We now use the fact that the multiple data points are independent to factorize the data likelihood

$$p(\mathbf{D} \mid \tilde{f}_{r,i}, \sigma^B_{r,i}) = \prod_{j=1}^{N_d} p(d_j \mid \tilde{f}_{r,i}, \sigma^B_{r,i}) \quad \text{(S14)}$$

and since the data point d_j depends only on $\tilde{f}_{r,j}$ and $\sigma^B_{r,j}$, we can write

$$p(\mathbf{D} \mid \tilde{f}_{r,i}, \sigma^B_{r,i}) = p(d_i \mid \tilde{f}_{r,i}, \sigma^B_{r,i}) \cdot \prod_{j=1, j \neq i}^{N_d} p(d_j) \propto p(d_i \mid \tilde{f}_{r,i}, \sigma^B_{r,i}) \quad \text{(S15)}$$

By inserting Eq. S15 into Eq. S13 we obtain the metainference equation for the case of multiple independent data points (Eq. 8 in the main text)

$$p(\mathbf{X}, \tilde{\boldsymbol{f}}, \boldsymbol{\sigma^B}, \boldsymbol{\sigma^{SEM}} \mid \mathbf{D}, I) \propto \prod_{r=1}^{N} \prod_{i=1}^{N_d} p(d_i \mid \tilde{f}_{r,i}, \sigma^B_{r,i}) \cdot p(\tilde{f}_{r,i} \mid \mathbf{X}, \sigma^{SEM}_{r,i}) \cdot p(\sigma^B_{r,i}) \cdot p(\sigma^{SEM}_{r,i}) \cdot \prod_{r=1}^{N} p(X_r) \quad \text{(S16)}$$

Details of the model system simulations.

To assess the accuracy of the different modelling approaches considered in this work, we studied a model system characterized by multiple discrete states, for which the number of states N_S and their population $[w^0]$ can be varied arbitrarily. This system captures some of the complexity of real mixtures of different species and/or conformations in which equilibrium measurements mix contributions from all states. A simulation of this model system consists of 4 steps.

1) Generation of states and synthetic experimental data. For each state, we randomly extracted its population w^0_k and N_d real numbers $d_{i,k}$ in the range from 1.0 to 10.0. These numbers are the pure experimental data points for each state and they will be used as forward model in the next step. The pure *observed* data points are a mixture on all states, $d_i = \sum_{k=1}^{N_S} w^0_k \cdot d_{i,k}$. We introduced two types of noise to the pure observed data points to mimic the presence of random and systematic errors. Random errors were modeled with a Gaussian noise with standard deviation equal to 0.5, while systematic errors were modeled by adding a random offset in the range from 3.0 to 5.0 to 30% of the data points. We modeled systems of 5 states using 2, 5, 10, and 20 data

points and systems of 50 states using 20, 50, 100, and 200 data points. For both model sizes, we genereted 4 datasets: (i) without errors, (ii) with only random errors, (iii) with only systematic errors, and (iv) with both random and systematic errors.

2) Scoring. In metainference, the total energy of the system is defined as

$$E = -k_B T \cdot \log p(\mathbf{X}, \boldsymbol{\sigma} \mid \mathbf{D}, I) \tag{S17}$$

where we used a Gaussian noise with one uncertainty parameter $\sigma_{r,i}$ per replica and data point (Eq. 9) or an outliers model with one uncertainty parameter per dataset (Eq. 11). In both cases, we used a Jeffrey's prior $p(\sigma) = 1/\sigma$ for the uncertainty of each data point or for the typical dataset uncertainty. σ^{SEM} was kept fixed and equal to $\tilde{\sigma}^{SEM} / \sqrt{N}$, with $\tilde{\sigma}^{SEM}$ = 5.66. For the standard Bayesian modelling, σ^{SEM} was set to zero. For the replica-averaged approach, we introduced harmonic restraints to couple forward model predictions to the observed data points. The intensity of the harmonic restraints was set to $k = N^2 \cdot k_1$, with k_1 = 0.03. We used the same prior information for the metainference, standard Bayesian modelling, and replica-averaged approaches. We randomly perturbed the exact populations w_k^0 to obtain approximate weights w_k for each state and thus we defined the energy associated to the prior information as $E_k = -k_B T \cdot \log w_k$. To study the effect of the prior accuracy, we created high and low accuracy priors, with an average population error per state equal to 0.08 and 0.16, respectively.

3) Sampling. We simulated N copies of the system to benchmark the metainference and replica-averaged approaches, and a single replica for standard Bayesian modelling. In the former case, we used 8, 16, 32, 64, and 128 replicas. The following unknown variables were sampled by Monte Carlo; a discrete index that determines which state of the system is occupied and the data uncertainty parameters for metainference and standard Bayesian modelling. The data uncertainty parameters were sampled in the range 0.00001-200, by proposing random moves at most equal to 10.0. $k_B T$ was set to 1.0. A total of 50,000 Monte Carlo steps were carried out in each simulation.

4) Analysis. During each Monte Carlo simulation we accumulated the histogram of the discrete variable that indicates which state of the system is istantaneously populated. From this histogram, we calculated the population of each state $\tilde{w}_k$ determined from prior information and experimental data. We defined as accuracy the root mean squared deviation of $[\tilde{w}_k]$ from the exact populations $[w_k^0]$. For each approach to test and choice of parameters (number of data points, level of noise in the data, and number of replicas), we run 300 indipendent simulations with random reference state populations and data points. The reported accuracy is averaged over the 300 simulations.

Details of the ubiquitin MD simulations

Classical all-atom molecular dynamics simulations of ubiquitin were perfomed using GROMACS (*33*) together with PLUMED (*34*). The X-ray structure 1UBQ (*29*) has been used as starting point in the simulations, using the CHARMM22* force field (*35*), in a cubic box of 6.3 nm of side with 7800 TIP3P water molecules (*39*). A time step of 2 fs was used together with LINCS constraints (*40*). The van der Waals and Coulomb interactions were cut-off at 0.9 nm,

while long-range electrostatic effects was treated with the particle mesh Ewald method. All simulations were carried out in the canonical ensemble by keeping the volume fixed and by thermosetting the system at 300 K with the Bussi-Donadio-Parrinello thermostat (*41*). A 1 ms molecular dynamics simulation was perfomed as a reference sampling of the *a priori* information of the CHARMM22* force field.

Maximum entropy replica-averaged simulations and metainference replica-averaged simulations were perfomed using backbone chemical shifts (bmr17760) and residual dipolar couplings measured in a liquid-crystallin phase (N-H, Cα-Hα, Cα-C′, C′-N, C′-H and Cα-Cb bonds) as structural restraints modelled with CamShifts and the exact ϑ-method, respectively. Maximum entropy and metainference simulations were performed using 8 replicas, in all cases for a total simulation time of 1 ms, consistently with the reference sampling. A Gaussian noise model with one error parameter per nucleus was used in the metainference approach, along with a Jeffrey's prior on each error parameter.

From the resulting ensembles we back-calculated chemical shifts using SPARTA+ (*42*), RDCs measured in a large number of conditions (*32*) using PALES in the SVD approximation (*43*) using only data for residues 1 to 70 to obtain the alignment tensor. Scalar couplings across hydrogen bonds have been calculated as $^{h3}JNC = (-357\ Hz)exp(-3.2r_{HO}/Å)cos^2\theta$ where θ represents the H…O=C angle (*44*), while H-Hα scalar coupling have been calculated using the Karplus equation with previously reported parameters (*45*). In addition the presence of distorted geometries have been tested with PROCHECK (*31*). The ensembles have been also compared with the 1UBQ X-ray (*29*) and the 1D3Z NMR (*30*) structures.

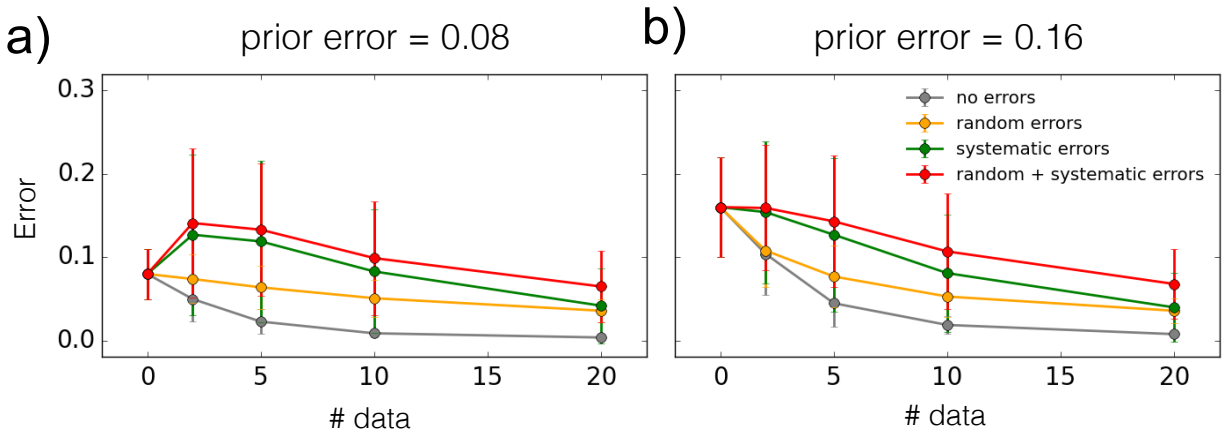

Figure S1. Effect of prior accuracy on the error of the metainference method. Metainference error as a function of the number of data points and for varying levels of noise in the data in the case of prior with average error in the state populations equal to 0.08 (**A**) and 0.16 (**B**). The quality of the prior information influences the number of data points required to achieve a given accuracy of the inferred state populations. The more accurate is the prior, the fewer data points are needed. These simulations were carried out on a 5-state model, using 128 replicas.

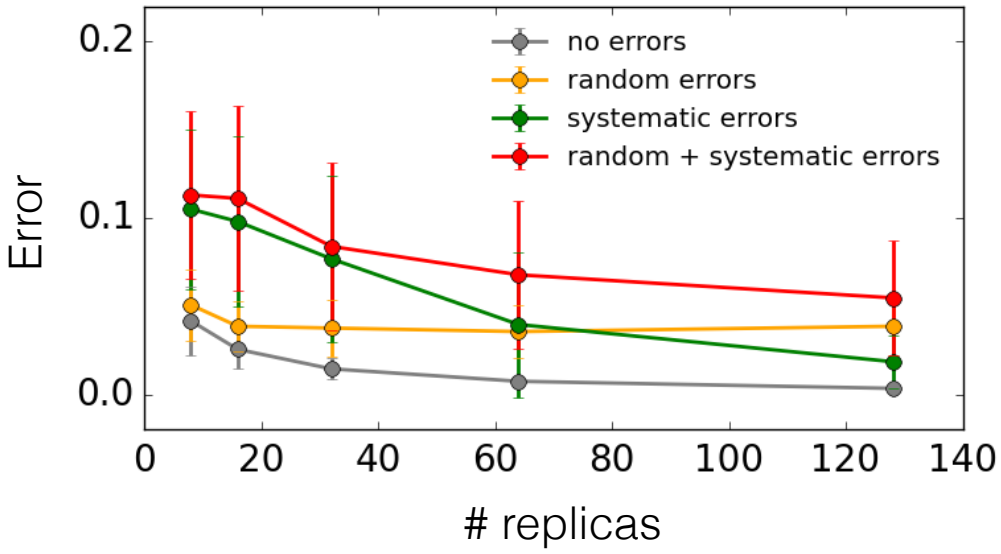

Figure S2. Scaling of metainference error with number of replicas for varying level of noise in the data. As the number of replicas increases, the statistical error in calculating ensemble averages with a finite number of replicas converges to zero, and the overall accuracy of metainference increases. These simulations were carried out on a 5-state model, using 20 data points and the prior with average error equal to 0.16.

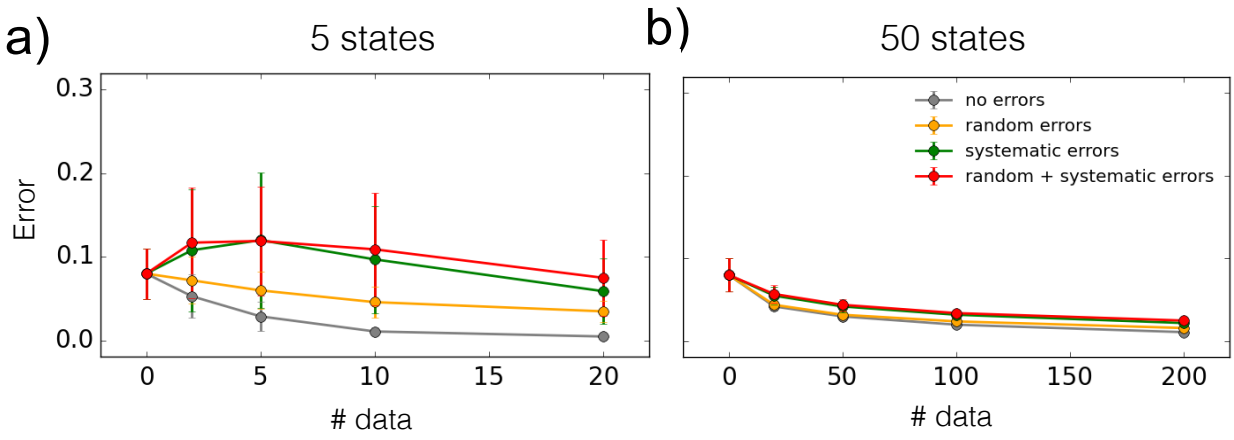

Figure S3. Scaling of metainference error with number of states. The metainference error as a function of the number of data points and for varying levels of noise in the data for a system composed of 5 (**A**) and 50 (**B**) states. These simulations were carried using 64 replicas and the prior with average error equal to 0.08.

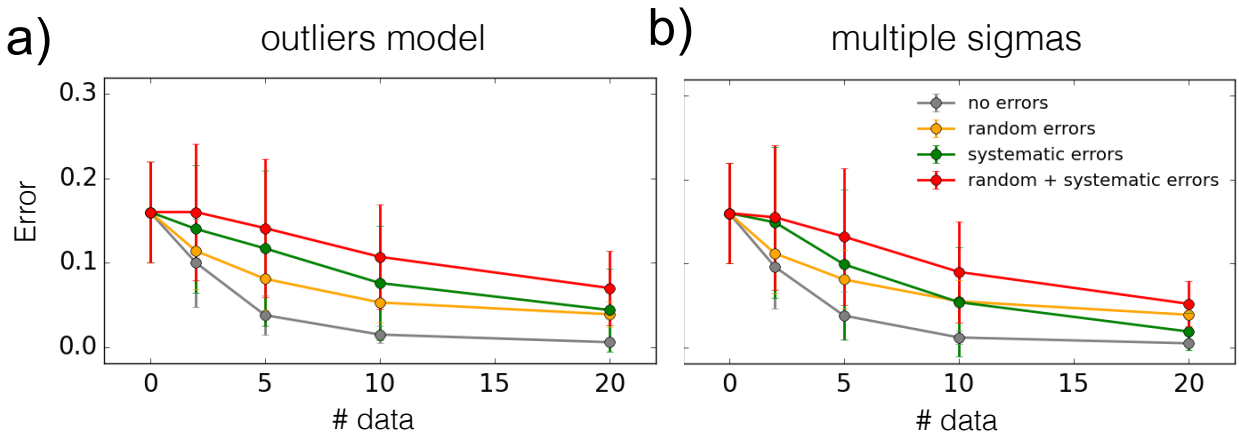

Figure S4. Accuracy of the outliers model. Metainference error as a function of the number of data points and for varying levels of noise in the data with an outlier model for the errors that uses a single error parameter per dataset (**A**) and with one error parameter per data point (**B**). These simulations were carried out on a 5-state model, using 128 replicas and the prior with average error equal to 0.16.

Score	Maximum entropy	Metainference	NMR	MD	X-ray
Modelling					
Chemical shifts					
CA	0.76	0.72	0.63	0.90	0.71
CB	0.89	0.93	0.89	1.12	0.94
CO	0.80	0.81	0.75	0.93	0.80
HA	0.13	0.13	0.21	0.23	0.17
HN	0.34	0.39	0.40	0.44	0.39
NH	2.51	2.32	1.86	2.77	2.03
RDC set 1					
NH	0.16	0.15	0.19	0.27	0.21
CAC	0.13	0.13	0.27	0.23	0.31
CAHA	0.15	0.15	0.13	0.23	0.28
CN	0.15	0.14	0.23	0.24	0.21
CH	0.52	0.18	0.29	0.31	0.32
Validation					
$^3J_{HNC}$					
RMSD	0.26	0.17	0.30	0.15	0.22
$^3J_{HNHA}$					
RMSD	1.08	0.89	0.69	0.99	0.89
RDC set 2					
NH(36)	0.23	0.20	0.29	0.28	0.29
RDC set 3					
NH	0.32	0.24	0.24	0.24	0.29
CAC	0.27	0.22	0.28	0.24	0.32
CAHA	0.37	0.33	0.40	0.32	0.42
CN	0.27	0.23	0.28	0.32	0.33
CH	0.34	0.26	0.51	0.34	0.47

Table S1. Comparison of the quality of the ensembles obtained using different modelling approaches in the case of the native state of the protein ubiquitin. Maximum entropy and metainference indicate the ensembles generated in this work using 8 replicas and chemical shifts combined with RDCs. NMR, MD and X-ray indicate a structure determined using high-resolution NMR methods (PDB code 1D3Z (*30*)), an ensemble determined by standard molecular dynamics simulations, and a X-ray structure (1UBQ) (*29*), respectively. In the upper part of the Table ("Modelling") we report the fit with the data used in the modelling, in the lower part ("Validation") the fit with independent data not used in the modelling.

Score	Maximum Entropy	Metainference	NMR	MD	X-ray
Procheck					
RAMA	1.6	1.2	1.0	1.0	1.0
HBGEO	2.3	1.9	1.4	2.1	1.7
CHI-1	1.4	1.4	1.0	1.3	2.0
CHI-2	1.0	1.0	1.0	1.0	1.4
OMEGA	2.5	2.0	1.0	2.0	1.0

Table S2. Comparison of the stereochemical quality of the ensembles or single models generated by the approaches defined in Table S1. The quality was assessed with PROCHECK (*31*).