COMPUTATIONAL BIOLOGY

10.1126/sciadv.1501177

Metainference: A Bayesian inference method for heterogeneous systems

Massimiliano Bonomi,[1]*† Carlo Camilloni,[1]† Andrea Cavalli,[1,2] Michele Vendruscolo[1]*

Modeling a complex system is almost invariably a challenging task. The incorporation of experimental observations can be used to improve the quality of a model and thus to obtain better predictions about the behavior of the corresponding system. This approach, however, is affected by a variety of different errors, especially when a system simultaneously populates an ensemble of different states and experimental data are measured as averages over such states. To address this problem, we present a Bayesian inference method, called "metainference," that is able to deal with errors in experimental measurements and with experimental measurements averaged over multiple states. To achieve this goal, metainference models a finite sample of the distribution of models using a replica approach, in the spirit of the replica-averaging modeling based on the maximum entropy principle. To illustrate the method, we present its application to a heterogeneous model system and to the determination of an ensemble of structures corresponding to the thermal fluctuations of a protein molecule. Metainference thus provides an approach to modeling complex systems with heterogeneous components and interconverting between different states by taking into account all possible sources of errors.

INTRODUCTION

The quantitative interpretation of experimental measurements requires the construction of a model of the system under observation. The model usually consists of a description of the system in terms of several parameters, which are determined by requiring consistency with the experimental measurements themselves and with theoretical information, either physical or statistical in nature. This procedure presents several complications. First, experimental data (Fig. 1A) are always affected by random and systematic errors (Fig. 1B, green), which must be properly accounted for to obtain accurate and precise models. Furthermore, when integrating multiple experimental observations, one must consider that each experiment has a different level of noise so that every element of information is properly weighted according to its reliability. Second, the prediction of experimental observables from the model, which is required to assess the consistency, is often based on an approximate description of a given experiment (the so-called "forward model") and thus is intrinsically inaccurate in itself (Fig. 1B, green). Third, physical systems under equilibrium conditions often populate a variety of different states whose thermodynamic behavior can be described by statistical mechanics. In these heterogeneous systems, experimental observations depend on—and thus probe—a population of states (Fig. 1B, purple) so that one should determine an ensemble of models rather than a single one (Fig. 1C).

Among the theoretical approaches available for model building, two frameworks have emerged as particularly successful: Bayesian inference (*1*–*3*) and the maximum entropy principle (*4*). Bayesian modeling is a rigorous approach to combining prior information on a system with experimental data and to dealing with errors in such data (*1*–*3*, *5*–*8*). It proceeds by constructing a model of noise as a function of one or more unknown uncertainty parameters, which quantify the agreement between predictions and observations and which are inferred along with the model of the system. This method has a long history and is routinely used in a wide range of applications, including the reconstruction of phylogenetic trees (*9*), determination of population structures from genotype data (*10*), interpolation of noisy data (*11*), image reconstruction (*12*), decision theory (*13*), analysis of microarray data (*14*), and structure determination of proteins (*15*, *16*) and protein complexes (*17*). It has also been extended to deal with mixtures of states (*18*–*21*) by treating the number of states as a parameter to be determined by the procedure. The maximum entropy principle is at the basis of approaches that deal with experimental data averaged over an ensemble of states (*4*) and provides a link between information theory and statistical mechanics. In these methods, an ensemble generated using a prior model is minimally modified by some partial and inaccurate information to exactly match the observed data. In the recently proposed replica-averaging scheme (*22*–*26*), this result is achieved by modeling an ensemble of replicas of the system using the available information and additional terms that restrain the average values of the predicted data to be close to the experimental observations. This method has been used to determine ensembles representing the structure and dynamics of proteins (*22*–*26*).

Each of the two methods described above can deal with some, but not all, of the challenges in characterizing complex systems by integrating multiple sources of information (Fig. 1B). To simultaneously overcome all of these problems, we present the "metainference" method, a Bayesian inference approach that quantifies the extent to which a prior distribution of models is modified by the introduction of experimental data that are expectation values over a heterogeneous distribution and subject to errors. To achieve this goal, metainference models a finite sample of this distribution, in the spirit of the replica-averaged modeling based on the maximum entropy principle. Notably, our approach reduces to the maximum entropy modeling in the limit of the absence of noise in the data, and to standard Bayesian modeling when experimental data are not ensemble averages. This link between Bayesian inference and the maximum entropy principle is not surprising given the connections between these two approaches (*27*, *28*).

[1]Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK. [2]Institute for Research in Biomedicine, CH-6500 Bellinzona, Switzerland.
*Corresponding author. E-mail: mb2006@cam.ac.uk (M.B.); mv245@cam.ac.uk (M.V.)

†These authors contributed equally to this work.

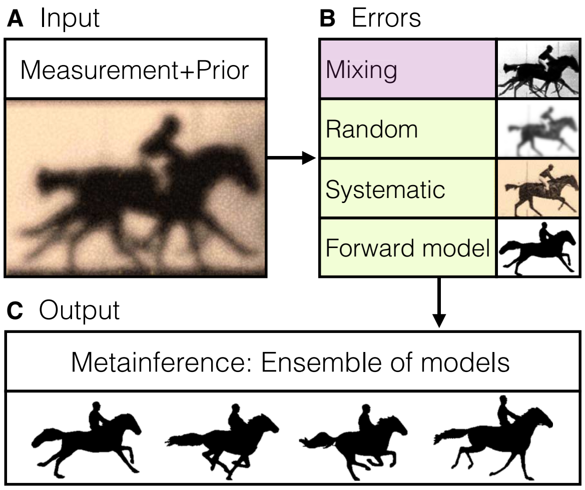

Fig. 1. Schematic illustration of the metainference method. (**A** and **B**) To generate accurate and precise models from input information (A), one must recognize that data from experimental measurements are always affected by random and systematic errors and that the theoretical interpretation of an experiment may also be inaccurate (B; green). Moreover, data collected on heterogeneous systems depend on a multitude of states and their populations (B; purple). (**C**) Metainference can treat all of these sources of error and thus it can properly combine multiple experimental data with prior knowledge of a system to produce ensembles of models consistent with the input information.

We first benchmark the accuracy of our method on a simple heterogeneous model system, in which synthetic experimental data can be generated with different levels of noise as averages over a discrete number of states of the system. We then show its application with nuclear magnetic resonance (NMR) spectroscopy data in the case of the structural fluctuations of the protein ubiquitin in its native state, which we modeled by combining chemical shifts with residual dipolar couplings (RDCs).

RESULTS AND DISCUSSION

Metainference is a Bayesian approach to modeling a heterogeneous system and all sources of error by considering a set of copies of the system (replicas), which represent a finite sample of the distribution of models, in the spirit of the replica-averaged formulation of the maximum entropy principle (*22*–*26*). The generation of models by suitable sampling algorithms [typically Monte Carlo or molecular dynamics (MD)] is guided by a score given in terms of the negative logarithm of the posterior probability (Materials and Methods)

$$\overbrace{s(\mathbf{X},\boldsymbol{\sigma})}^{\text{score}} = \overbrace{\underbrace{-\sum_r \log P(X_r,\sigma_r)}_{\text{prior}} + \underbrace{\Delta^2(\mathbf{X})}_{\text{measurements}}}^{\text{input}} \overbrace{\underbrace{\sum_r \frac{1}{2\sigma_r^2}}_{\sigma_r^2=(\sigma_r^{SEM})^2+(\sigma_r^B)^2}}^{\text{errors}}$$

where $\mathbf{X} = [X_r]$ and $\boldsymbol{\sigma} = [\sigma_r]$ are, respectively, the sets of conformational states and uncertainties, one for each replica. σ_r includes all of the sources of errors, that is, the error in representing the ensemble with a finite number of replicas (σ_r^{SEM}), as well as random, systematic, and forward model errors (σ_r^{B}). P is the prior probability that encodes information other than experimental data, and $\Delta^2(\mathbf{X})$ is the deviation of the experimental data from the data predicted by the forward model. This schematic equation, which omits the data likelihood normalization term for the uncertainty parameters, holds for Gaussian errors and a single data point, and a more general formulation can be found in Materials and Methods (Eqs. 5 and 8).

Metainference of a heterogeneous model system

We first illustrate the metainference method for a model system that can simultaneously populate a set of discrete states, that is, a mixture. In this example, the number of states in the mixture and their population can be varied arbitrarily. We created synthetic data as ensemble averages over these discrete states (Fig. 2A), and we added random and systematic noise. We thus introduced prior information, which provides an approximate description of the system and its distribution of states and whose accuracy can also be tuned. We then used the reference data to complement the prior information and to recover the correct number and populations of the states. We tested the following approaches: metainference (with the Gaussian and outliers noise models in Eqs. 9 and 11, respectively), replica-averaging maximum entropy, and standard Bayesian inference (that is, Bayesian inference without mixtures). The accuracy of a given approach was defined as the root mean square deviation of the inferred populations from the correct populations of the discrete states. We benchmarked the accuracy as a function of the number of data points used, the level of noise in the data, the number of states and replicas, and the accuracy of the prior information. Details of the simulations, generation of data, sampling algorithm, and likelihood and model to treat systematic errors and outliers can be found in the Supplementary Materials.

Comparison with the maximum entropy method

We found that the metainference and maximum entropy methods perform equally well in the absence of noise in the data or in the presence of random noise alone (Fig. 2, B and C, gray and orange lines), as expected, given that maximum entropy is particularly effective in the case of mixtures of states (*22*, *23*). The accuracy of the two methods was comparable and, most importantly, increased with the number of data points used (Fig. 2, B and C). With 20 data points and 128 replicas, and in the absence of noise, the accuracy averaged over 300 independent simulations of a five-state system was equal to 0.4 ± 0.2% and 0.2 ± 0.1% for the metainference and maximum entropy approaches, respectively. For reference, the accuracy of the prior information alone was much lower, that is, 16%. Metainference, however, outperformed the maximum entropy approach in the presence of systematic errors (Fig. 2, B and C, green lines). The accuracy of metainference increased significantly more rapidly upon the addition of new information, despite the high level of noise. When using 20 data points, 128 replicas, and 30% outliers ratio, the accuracy averaged over 300 independent simulations of a five-state system was equal to 2 ± 2% and 14 ± 5% for the metainference and maximum entropy approaches, respectively. As systematic errors are ubiquitous both in the experimental data and in the forward model used to predict the data, this situation more closely reflects a realistic scenario. The ability of metainference to effectively deal with averaging and with the presence of systematic errors at the same time is the main motivation for introducing this method. This approach can thus leverage the substantial amount of noisy data produced by high-throughput techniques and accurately model conformational ensembles of heterogeneous systems.

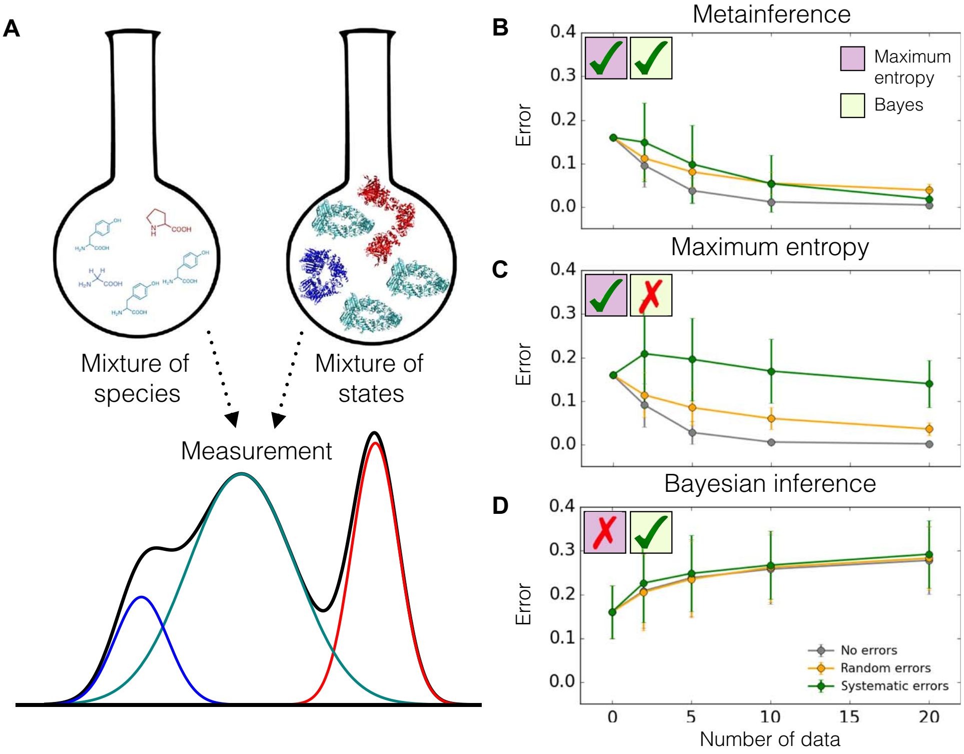

Fig. 2. Metainference of a model heterogeneous system. (**A**) Equilibrium measurements on mixtures of different species or states do not reflect a single species or conformation but are instead averaged over the whole ensemble. (**B** to **D**) We describe such a scenario using a model heterogeneous system composed of multiple discrete states on which we tested metainference (B), the maximum entropy approach (C), and standard Bayesian modeling (D), using synthetic data. We assess the accuracy of these methods in determining the populations of the states as a function of the number of data points used and the level of noise in the data. Among these approaches, metainference is the only one that can deal with both heterogeneity and errors in the data; the maximum entropy approach can treat only the former, whereas standard Bayesian modeling can treat only the latter.

Comparison with standard Bayesian modeling

In the standard Bayesian approach, one assumes the presence of a single state in the sample and estimates its probability or confidence level given experimental data and prior knowledge available. When modeling multiple-state systems with ensemble-averaged data and standard Bayesian modeling, one could be tempted to interpret the probability of each state as its equilibrium population. In doing so, however, one makes a significant error, which grows with the number of data points used, regardless of the level of noise in the data (Fig. 2D).

Role of prior information

We tested two priors of different accuracies, with an average population error per state equal to 8 and 16%, respectively. The results suggest that the number of experimental data points required to achieve a given accuracy of the inferred populations depends on the quality of the prior information (fig. S1). The more accurate the prior is, the fewer data points are needed. This is an intuitive, yet important, result. Accurate priors almost invariably require more complex descriptions of the system under study; thus, they usually come at a higher computational cost.

Scaling with the number of replicas

As the number of replicas grows, the error in estimating ensemble averages using a finite number of replicas decreases, and the overall accuracy of the inferred populations increases (fig. S2), regardless of the level of noise in the data. Furthermore, we verified numerically that, in the absence of random and systematic errors in the data, the intensity of the harmonic restraint, which couples the average of the forward model on the N replicas to the experimental data (Eq. 7), scales as N^2 (Fig. 3). This test confirms that, in the limit of the absence of noise in the data, metainference coincides with the replica-averaging maximum entropy modeling (Materials and Methods).

Scaling with the number of states

Metainference is also robust to the number of states populated by the system. We tested our model in the case of 5 and 50 states and determined that the number of data points needed to achieve a given accuracy scales less than linearly with the number of states (fig. S3).

Outliers model and error marginalization

As the numbers of data points and replicas increase, using one error parameter per replica and data point becomes computationally more and more inconvenient. In this situation, one can assume a unimodal and long-tailed distribution for the errors, peaked around a typical value for a data set (or experiment type) and replica, and marginalize all of the uncertainty parameters of the single data points (Materials and Methods). The accuracy of this marginalized error model was found to be similar to the case in which a single error parameter was used for each data point (fig. S4).

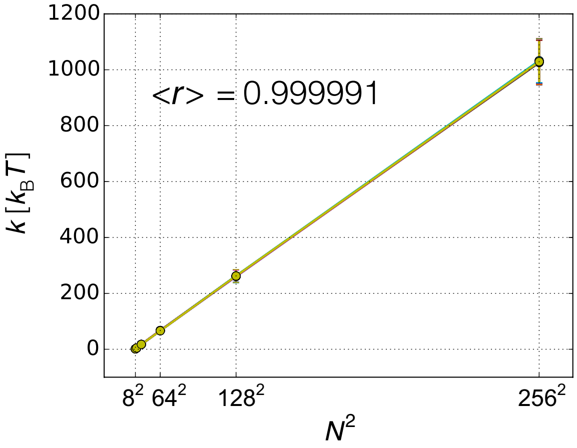

Fig. 3. Scaling of the metainference harmonic restraint intensity in the absence of noise in the data. We verified numerically that in the absence of noise in the data and with a Gaussian noise model, the intensity of the metainference harmonic restraint $k = \sum_{r=1}^{N} \frac{1}{\sigma_r^2}$, which couples the average of the forward model over the N replicas to the experimental data point (Eq. 7), scales as N^2. This test was carried out in the model system at five discrete states, with 20 data points and with the prior at 16% accuracy. For each of the 20 data points, we report the average restraint intensity over the entire Monte Carlo simulation and its SD when using 8, 16, 32, 64, 128, and 256 replicas. The average Pearson's correlation coefficient on the 20 data points is 0.999991 ± 3 × 10^{-6}, showing that metainference coincides with the replica-averaging maximum entropy modeling in the limit of the absence of noise in the data.

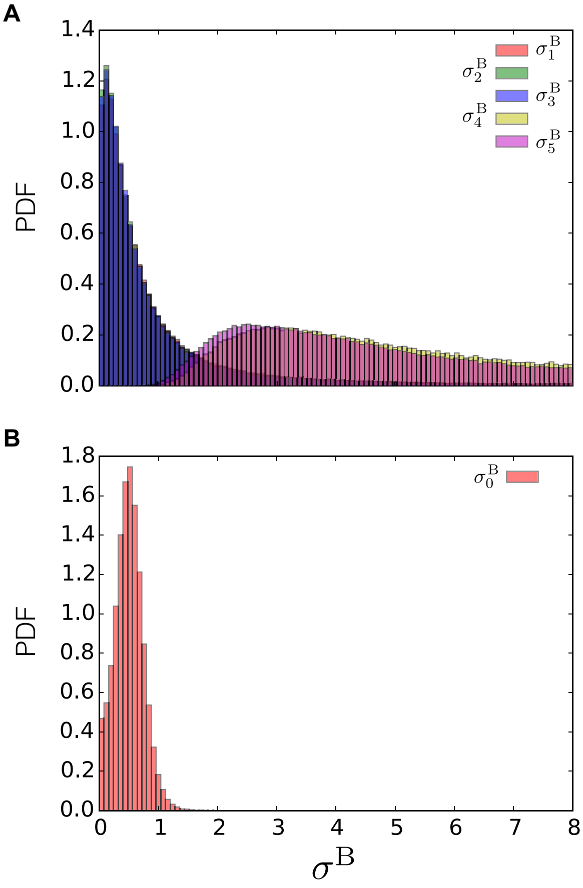

Fig. 4. Analysis of the inferred uncertainties. (**A** and **B**) Distributions of inferred uncertainties (PDF) in the presence of systematic errors, using (A) a Gaussian data likelihood with one uncertainty per data point and (B) the outliers model with one uncertainty per data set. This test was carried out in the model system at five discrete states, with 20 data points (of which eight were outliers), 128 replicas, and the prior at 16% accuracy. For the Gaussian noise model, we report the distributions of three representative points not affected by noise ($\sigma_{1-3}^{\mathrm{B}}$) and of two representative points affected by systematic errors (σ_4^{B} and σ_5^{B}). For the outliers model, we report the distribution of the typical data set uncertainty (σ_0^{B}).

Analysis of the inferred uncertainties

We analyzed the distribution of inferred uncertainties σ^{B} in the presence of systematic errors (outliers) when using a Gaussian data likelihood with one uncertainty per data point (Eq. 9) and the outliers model with one uncertainty per data set (Eq. 11). In the former case, metainference was able to automatically detect the data points affected by systematic errors, assign a higher uncertainty unto them, and thus downweight the associated restraints (Fig. 4A). In the latter case, the inferred typical data set uncertainty was somewhere in between the uncertainty inferred using the Gaussian likelihood on the data points with no noise and the uncertainty inferred using the Gaussian likelihood on the outliers (Fig. 4B). In this specific test (five states, 20 data points including eight outliers, prior accuracy equal to 16%, and 128 replicas), both data noise models generated an ensemble of comparable accuracy (3%).

Metainference in integrative structural biology

We compared the metainference and maximum entropy approaches using NMR experimental data on a classical example in structural biology—the structural fluctuations in the native state of ubiquitin (*22*, *29*, *30*). A conformational ensemble of ubiquitin was modeled using CA, CB, CO, HA, HN, and NH chemical shifts combined with RDCs collected in a steric medium (*30*) (Fig. 5A). The ensemble was validated by multiple criteria (table S1). The stereochemical quality was assessed by PROCHECK (*31*); data not used for modeling, including $^3J_{\mathrm{HNC}}$ and $^3J_{\mathrm{HNHA}}$ scalar couplings and RDCs collected in other media (*32*), were backcalculated and compared with the experimental data. Exhaustive sampling was achieved by 1-μs-long MD simulations performed with GROMACS (*33*) equipped with PLUMED (*34*). We used the CHARMM22* force field as prior information (*35*). Additional details of these simulations can be found in the Supplementary Materials.

The quality of the metainference ensemble (Fig. 5B) was higher than that of the maximum entropy ensemble, as suggested by the better fit with the data not used in the modeling (Fig. 5C and table S1) and by the stereochemical quality (table S2). Data used as restraints were also more accurately reproduced by metainference. One of the major differences between the two approaches is that metainference can deal more effectively with the errors in the chemical shifts calculated on different nuclei. The more inaccurate HN and NH chemical shifts were detected by metainference and thus automatically downweighted in constructing the ensemble (Fig. 6).

We also compared the metainference ensemble with an ensemble generated by standard MD simulations and with a high-resolution NMR structure. The metainference ensemble obtained by combining chemical shifts and RDCs reproduced all of the experimental data not used for the modeling better than the MD ensemble and the NMR

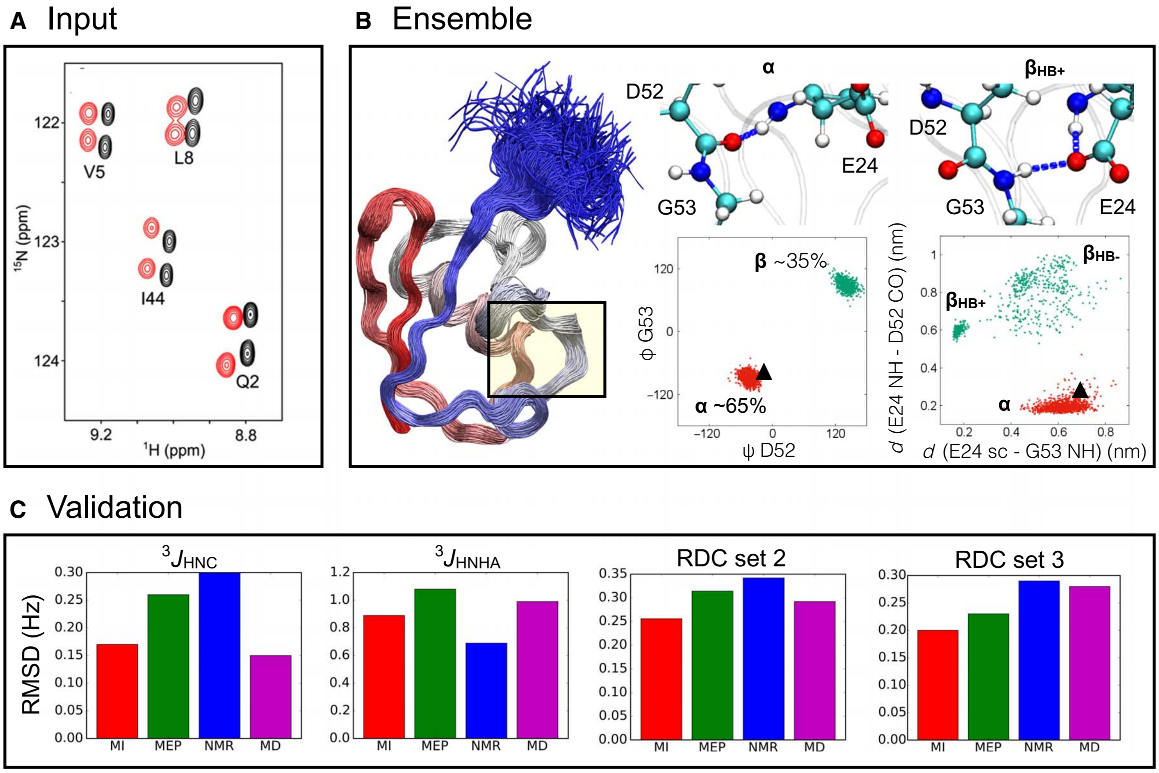

Fig. 5. Example of the application of metainference in integrative structural biology. (**A**) Comparison of the metainference and maximum entropy approaches by modeling the structural fluctuations of the protein ubiquitin in its native state using NMR chemical shifts and RDC data. (**B**) The metainference ensemble supports the finding (*36*) that a major source of dynamics involves a flip of the backbone of residues D^{52}-G^{53} (B; left scatterplot), which interconverts between an α state with a 65% population and a β state with a 35% population. This flip is coupled with the formation of a hydrogen bond between the side chain of E^{24} and the backbone of G^{53} (B; right scatterplot); the state in which the hydrogen bond is present (β_{HB+}) is populated 30% of the time, and the state in which the hydrogen bond is absent (β_{HB-}) is populated 5% of the time. By contrast, the NMR structure (Protein Data Bank code 1D3Z) provides a static picture of ubiquitin in this region in which the α state is the only populated one (black triangle). (**C**) Validation of the metainference (MI; red) and maximum entropy principle (MEP; green) ensembles, along with the NMR structure (blue) and the MD ensemble (purple), by the backcalculation of experimental data not used in the modeling: $^{3}J_{HNC}$ and $^{3}J_{HNHA}$ scalar couplings and two independent sets of RDCs (RDC sets 2 and 3).

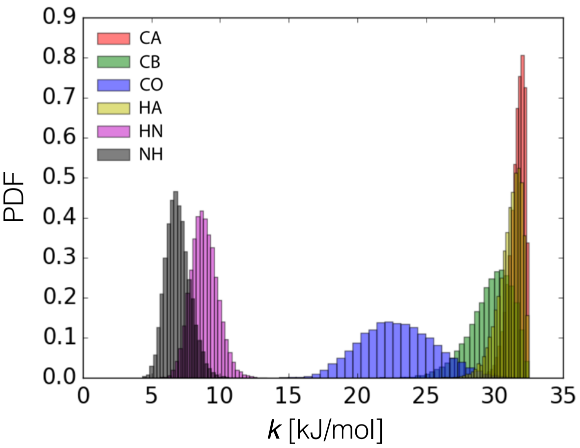

Fig. 6. Distributions (PDF) of restraint intensities for different chemical shifts of ubiquitin. When combining data from different experiments, metainference automatically determines the weight of each piece of information. In the case of ubiquitin, the NH and HN chemical shifts were determined as the less reliable data and thus were downweighted in the construction of the ensemble of models. From this procedure it is not possible to determine whether these two specific data sets have a higher level of random or systematic noise, or whether instead the CAMSHIFT predictor (*38*) is less accurate for these specific nuclei.

structure. The only exception were the $^{3}J_{HNC}$ scalar couplings, which were slightly more accurate in the MD ensemble, and the $^{3}J_{HNHA}$ scalar couplings, which were better predicted by the NMR structure (Fig. 5C and table S1).

The NMR structure, which was determined according to the criterion of maximum parsimony, accurately reproduced most of the available experimental data. Ubiquitin, however, exhibits rich dynamical properties over a wide range of time scales averaged in the experimental data (*36*). In particular, a main source of dynamics involves a flip of the backbone of residues D^{52}-G^{53} coupled with the formation of a hydrogen bond between the side chain of E^{24} and the backbone of G^{53}. Although metainference was able to capture the conformational exchange between these two states, the static representation provided by the NMR structure could not (Fig. 5B).

In conclusion, we have presented the metainference approach, which enables the building of an ensemble of models consistent with experimental data when the data are affected by errors and are averaged over mixtures of the states of a system. Because complex systems and experimental data almost invariably exhibit both heterogeneity and errors, we anticipate that our method will find applications across a wide variety of scientific fields, including genomics, proteomics, metabolomics, and integrative structural biology.

MATERIALS AND METHODS

The quantitative understanding of a system involves the construction of a model M to represent it. If a system can occupy multiple possible states, one should determine the distribution of models $p(M)$ that specifies in which states the system can be found and the corresponding probabilities. To construct this distribution of models, one should take into account the consistency with the overall knowledge that one has about the system. This includes theoretical knowledge (called the "prior" information I) and the information acquired from experimental measurements (that is, the "data" D) (*1*). In Bayesian inference, the probability of a model given the information available is known as the posterior probability $p(M|D, I)$ of M given D and I, and it is given by

$$p(M|D,I)\propto\ p(D|M,I)p(M|I) \qquad (1)$$

where the likelihood function $p(D|M, I)$ is the probability of observing D given M and I, and the prior probability $p(M|I)$ is the probability of M given I. To define the likelihood function, one needs a forward model $f(M)$ that predicts the data that would be observed for model M and a noise model that specifies the distribution of the deviations between the observed data and the predicted data. In the following, we assumed that the forward model depends only on the conformational state X of the system and that the noise model is defined in terms of unknown parameters σ that are part of the model $M = (X, \sigma)$. These parameters quantify the level of noise in the data, and they are inferred along with the state X by sampling the posterior distribution. The sampling is usually carried out using computational techniques such as Monte Carlo, MD, or combined methods based on Gibbs sampling (*1*).

Mixture of states

Experimental data collected under equilibrium conditions are usually the result of ensemble averages over a large number of states. In metainference, the prior information $p(X)$ of state X provides an a priori description of the distribution of states. To quantify the fit with the observed data and to determine to what extent the prior distribution is modified by the introduction of the data, we needed to calculate the expectation values of the forward model over the distribution of states. Inspired by the replica-averaged modeling based on the maximum entropy principle (*22*–*26*), we considered a finite sample of this distribution by simultaneously modeling N replicas of the model $\mathbf{M} = [M_r]$, and we calculated the forward model as an average over the states $\mathbf{X} = [X_r]$

$$f(\mathbf{X})=\frac{1}{N}\sum_{r=1}^{N}f(X_r) \qquad (2)$$

Typically, we have information only about expectation values on the distribution of states X, and not on the other parameters of the model, such as σ. However, we were also interested in determining how the prior distributions of these parameters are modified by the introduction of the experimental data. Therefore, we modeled a finite sample of the joint probability distribution of all parameters of the model.

For a reduced computational cost, a relatively small number of replicas are typically used in the modeling. In this situation, the estimate $f(\mathbf{X})$ of the forward model deviated from the average $\tilde{f}$ that would be obtained using an infinite number of replicas. This was an unknown quantity, which we added to the parameters of our model. However, the central limit theorem provided a strong parametric prior because it guaranteed that the probability of having a certain value of $\tilde{f}$ given a finite number of states $\mathbf{X}$ is a Gaussian distribution

$$p(\tilde{f}\,|\mathbf{X},\sigma^{\mathrm{SEM}})=\frac{1}{\sqrt{2\pi}\sigma^{\mathrm{SEM}}}\exp\left[-\frac{(\tilde{f}-f(\mathbf{X}))^2}{2(\sigma^{\mathrm{SEM}})^2}\right] \qquad (3)$$

where the standard error of the mean σ^{SEM} decreases with the square root of the number of replicas

$$\sigma^{\mathrm{SEM}}\propto\frac{1}{\sqrt{N}} \qquad (4)$$

We recognized that, in considering a finite sample of our distribution of states, we introduced an error in the calculation of expectation values. Therefore, experimental data should be compared to the (unknown) average of the forward model over an infinite number of replicas $\tilde{f}$, which is then related to the average over our finite sample $f(\mathbf{X})$ via the central limit theorem of Eq. 3. From these considerations, we can derive the posterior probability of the ensemble of N replicas representing a finite sample of our distribution of models $\mathbf{M}=(\mathbf{X},\tilde{\boldsymbol{f}},\boldsymbol{\sigma}^{\mathbf{B}},\boldsymbol{\sigma}^{\mathbf{SEM}})$. In the case of a single experimental data point d, this can be expressed as (Supplementary Materials)

$$p(\mathbf{X},\tilde{\boldsymbol{f}},\boldsymbol{\sigma}^{\mathbf{B}},\boldsymbol{\sigma}^{\mathbf{SEM}}|d,I)\propto\prod_{r=1}^{N}p(d|\tilde{f}_r,\sigma_r^{\mathrm{B}})p(\tilde{f}_r|\mathbf{X},\sigma_r^{\mathrm{SEM}})p(\sigma_r^{\mathrm{B}})p(X_r)p(\sigma_r^{\mathrm{SEM}}) \qquad (5)$$

The data likelihood $p(d|\tilde{f}_r,\sigma_r^{\mathrm{B}})$ relates the experimental data d to the average of the forward model over an infinite number of replicas, given the uncertainty σ_r^{B}. This parameter describes random and systematic errors in the experimental data and errors in the forward model. The functional form of $p(d|\tilde{f}_r,\sigma_r^{\mathrm{B}})$ depends on the nature of the experimental data, and it is typically a Gaussian or lognormal distribution. As noted above, $p(\tilde{f}_r|\mathbf{X},\sigma_r^{\mathrm{SEM}})$ is the parametric prior on $\tilde{f}_r$ that relates the (unknown) average $\tilde{f}_r$ to the estimate $f(\mathbf{X})$ computed with a finite number of replicas N via the central limit theorem of Eq. 3, and thus it is always a Gaussian distribution. $p(\sigma_r^{\mathrm{SEM}})$ is the prior on the standard error of the mean σ_r^{SEM} and encodes Eq. 4. $p(\sigma_r^{\mathrm{B}})$ is the prior on the uncertainty parameter σ_r^{B}, and $p(X_r)$ is the prior on the structure X_r.

Gaussian noise model

We can further simplify Eq. 5 in the case of Gaussian data likelihood $p(d|\tilde{f}_r,\sigma_r^{B})$. In this situation, $\tilde{f}_r$ can be marginalized (Supplementary Materials), and the posterior probability can be written as

$$p(\mathbf{X},\boldsymbol{\sigma}|d,I)\propto\prod_{r=1}^{N}\frac{1}{\sqrt{2\pi}\sigma_r}\exp\left[-\frac{(d-f(\mathbf{X}))^2}{2\sigma_r^2}\right]p(\sigma_r)p(X_r) \qquad (6)$$

where the effective uncertainty $\sigma_r=\sqrt{(\sigma_r^{\mathrm{SEM}})^2+(\sigma_r^{\mathrm{B}})^2}$ encodes all sources of errors: the statistical error due to the use of a finite number of replicas, experimental and systematic errors, and errors in the forward model. The associated energy function in units of $k_B T$ becomes

$$E = (d - f(\mathbf{X}))^2 \sum_{r=1}^{N} \frac{1}{2\sigma_r^2} + \sum_{r=1}^{N} [\log \sigma_r - \log p(\sigma_r) - \log p(X_r)] \quad (7)$$

This equation shows how metainference includes different existing modeling methods in limiting cases. In the absence of data and forward model errors ($\sigma_r^B = 0$), our approach reduces to the replica-averaged maximum entropy modeling, in which a harmonic restraint couples the replica-averaged observable to the experimental data. The intensity of the restraint $k = \sum_{r=1}^{N} \frac{1}{\sigma_r^2}$ scales with the number of replicas as N^2, that is, more than linearly, as required by the maximum entropy principle (*24*). We numerically verified this behavior in our heterogeneous model system in the absence of any errors in the data (Fig. 3). In the presence of errors ($\sigma_r^B \neq 0$), the intensity k scales as N, and it is modulated by the data uncertainty σ_r^B. Finally, in the case in which the experimental data are not ensemble averages ($\sigma_r^{SEM} = 0$), we recover the standard Bayesian modeling.

Multiple experimental data points

Equation 5 can be extended to the case of N_d independent data points $\mathbf{D} = [d_i]$, possibly gathered in different experiments at varying levels of noise (Supplementary Materials)

$$p(\mathbf{X}, \tilde{f}, \boldsymbol{\sigma}^{\mathbf{B}}, \boldsymbol{\sigma}^{\mathbf{SEM}} | \mathbf{D}, I) \propto \prod_{r=1}^{N} \prod_{i=1}^{N_d} p(d_i | \tilde{f}_{r,i}, \sigma_{r,i}^B)$$
$$\times\, p(\tilde{f}_{r,i} | \mathbf{X}, \sigma_{r,i}^{SEM}) p(\sigma_{r,i}^B) p(\sigma_{r,i}^{SEM}) \prod_{r=1}^{N} p(X_r) \quad (8)$$

Outliers model

To reduce the number of parameters that need to be sampled in the case of multiple experimental data points, one can model the distribution of the errors around a typical data set error and marginalize the error parameters for the individual data points. For example, a data set can be defined as a set of chemical shifts or RDCs on a given nucleus. In this case, it is reasonable to assume that the level of error of the individual data points in the data set is homogeneous, except for the presence of few outliers. Let us consider, for example, the case of Gaussian data noise. In the case of multiple experimental data points, Eq. 6 becomes

$$p(\mathbf{X}, \boldsymbol{\sigma} | \mathbf{D}, I) \propto \prod_{r=1}^{N} p(X_r) \prod_{i=1}^{N_d} \frac{1}{\sqrt{2\pi}\sigma_{r,i}} \exp\left[-\frac{(d_i - f_i(\mathbf{X}))^2}{2\sigma_{r,i}^2}\right] p(\sigma_{r,i}) \quad (9)$$

The prior $p(\sigma_{r,i})$ can be modeled using a unimodal distribution peaked around a typical data set effective uncertainty $\sigma_{r,0}$ and with a long tail to tolerate outliers data points (*37*)

$$p(\sigma_{r,i}) = \frac{2\sigma_{r,0}}{\sqrt{\pi}\sigma_{r,i}^2} \exp\left(-\frac{\sigma_{r,0}^2}{\sigma_{r,i}^2}\right) \quad (10)$$

where $\sigma_{r,0} = \sqrt{(\sigma^{SEM})^2 + (\sigma_{r,0}^B)^2}$, with σ^{SEM} as the standard error of the mean for all data points in the data set and replicas and with $\sigma_{r,0}^B$ as the typical data uncertainty of the data set. We can thus marginalize $\sigma_{r,i}$ by integrating over all its possible values, given that all of the data uncertainties $\sigma_{r,i}^B$ range from 0 to infinity

$$\begin{aligned} p(\mathbf{X}, \boldsymbol{\sigma_0} | \mathbf{D}, I) &\propto \prod_{r=1}^{N} p(X_r) \\ &\times \prod_{i=1}^{N_d} \int_{\sigma^{SEM}}^{+\infty} d\sigma_{r,i} \frac{\sqrt{2}\sigma_{r,0}}{\pi\sigma_{r,i}^3} \exp\left[-\frac{0.5(d_i - f_i(\mathbf{X}))^2 + \sigma_{r,0}^2}{\sigma_{r,i}^2}\right] \\ &= \prod_{r=1}^{N} p(X_r) \prod_{i=1}^{N_d} \frac{\sqrt{2}\sigma_{r,0}}{\pi} \frac{1}{(d_i - f_i(\mathbf{X}))^2 + 2\sigma_{r,0}^2} \\ &\times \left\{1 - \exp\left[-\frac{0.5(d_i - f_i(\mathbf{X}))^2 + \sigma_{r,0}^2}{(\sigma^{SEM})^2}\right]\right\} \end{aligned} \quad (11)$$

After marginalization, we are left with just one parameter $\sigma_{r,0}^B$ per replica that needs to be sampled.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/2/1/e1501177/DC1
Derivation of the basic metainference equations
Details of the model system simulations
Details of the ubiquitin MD simulations
Fig. S1. Effect of prior accuracy on the error of the metainference method.
Fig. S2. Scaling of metainference error with the number of replicas at varying levels of noise in the data.
Fig. S3. Scaling of metainference error with the number of states.
Fig. S4. Accuracy of the outliers model.
Table S1. Comparison of the quality of the ensembles obtained using different modeling approaches in the case of the native state of the protein ubiquitin.
Table S2. Comparison of the stereochemical quality of the ensembles or single models generated by the approaches defined in table S1.
References (*39*–*45*)

REFERENCES AND NOTES

1. G. E. Box, G. C. Tiao, *Bayesian Inference in Statistical Analysis* (John Wiley & Sons, New York, 2011), vol. 40.
2. J. M. Bernardo, A. F. Smith, *Bayesian Theory* (John Wiley & Sons, New York, 2009), vol. 405.
3. P. M. Lee, *Bayesian Statistics: An Introduction* (John Wiley & Sons, New York, 2012).
4. E. T. Jaynes, Information theory and statistical mechanics. *Phys. Rev.* **106**, 620–630 (1957).
5. S. Tavaré, D. J. Balding, R. C. Griffiths, P. Donnelly, Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518 (1997).
6. J. K. Pritchard, M. T. Seielstad, A. Perez-Lezaun, M. W. Feldman, Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Mol. Biol. Evol.* **16**, 1791–1798 (1999).
7. D. Poole, A. E. Raftery, Inference for deterministic simulation models: The Bayesian melding approach. *J. Am. Stat. Assoc.* **95**, 1244–1255 (2000).
8. M. C. Kennedy, A. O'Hagan, Bayesian calibration of computer models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **63**, 425–464 (2001).
9. J. P. Huelsenbeck, F. Ronquist, MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).
10. J. K. Pritchard, M. Stephens, P. Donnelly, Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
11. D. J. C. MacKay, Bayesian interpolation. *Neural Comp.* **4**, 415–447 (1992).
12. S. Geman, D. Geman, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **6**, 721–741 (1984).
13. J. O. Berger, *Statistical Decision Theory and Bayesian Analysis* (Springer Science & Business Media, New York, 2013).
14. P. Baldi, A. D. Long, A Bayesian framework for the analysis of microarray expression data: Regularized *t*-test and statistical inferences of gene changes. *Bioinformatics* **17**, 509–519 (2001).
15. W. Rieping, M. Habeck, M. Nilges, Inferential structure determination. *Science* **309**, 303–306 (2005).

16. J. L. MacCallum, A. Perez, K. A. Dill, Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc. Natl. Acad. Sci. U.S.A.* **112**, 6985–6990 (2015).
17. J. P. Erzberger, F. Stengel, R. Pellarin, S. Zhang, T. Schaefer, C. H. S. Aylett, P. Cimermančič, D. Boehringer, A. Sali, R. Aebersold, N. Ban, Molecular architecture of the 40S·eIF1·eIF3 translation initiation complex. *Cell* **158**, 1123–1135 (2014).
18. P. Cossio, G. Hummer, Bayesian analysis of individual electron microscopy images: Towards structures of dynamic and heterogeneous biomolecular assemblies. *J. Struct. Biol.* **184**, 427–437 (2013).
19. M. T. Marty, A. J. Baldwin, E. G. Marklund, G. K. A. Hochberg, J. L. P. Benesch, C. V. Robinson, Bayesian deconvolution of mass and ion mobility spectra: From binary interactions to polydisperse ensembles. *Anal. Chem.* **87**, 4370–4376 (2015).
20. K. S. Molnar, M. Bonomi, R. Pellarin, G. D. Clinthorne, G. Gonzalez, S. D. Goldberg, M. Goulian, A. Sali, W. F. DeGrado, Cys-scanning disulfide crosslinking and Bayesian modeling probe the transmembrane signaling mechanism of the histidine kinase, PhoQ. *Structure* **22**, 1239–1251 (2014).
21. T. O. Street, X. Zeng, R. Pellarin, M. Bonomi, A. Sali, M. J. S. Kelly, F. Chu, D. A. Agard, Elucidating the mechanism of substrate recognition by the bacterial Hsp90 molecular chaperone. *J. Mol. Biol.* **426**, 2393–2404 (2014).
22. K. Lindorff-Larsen, R. B. Best, M. A. DePristo, C. M. Dobson, M. Vendruscolo, Simultaneous determination of protein structure and dynamics. *Nature* **433**, 128–132 (2005).
23. A. Cavalli, C. Camilloni, M. Vendruscolo, Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *J. Chem. Phys.* **138**, 094112 (2013).
24. B. Roux, J. Weare, On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *J. Chem. Phys.* **138**, 084107 (2013).
25. J. W. Pitera, J. D. Chodera, On the use of experimental observations to bias simulated ensembles. *J. Chem. Theory Comput.* **8**, 3445–3451 (2012).
26. W. Boomsma, J. Ferkinghoff-Borg, K. Lindorff-Larsen, Combining experiments and simulations using the maximum entropy principle. *PLOS Comput. Biol.* **10**, e1003406 (2014).
27. A. Giffin, A. Caticha, Updating probabilities with data and moments. arXiv preprint arXiv:0708.1593 (2007).
28. A. Caticha, Entropic inference. arXiv preprint arXiv:1011.0723 (2010).
29. S. Vijay-Kumar, C. E. Bugg, W. J. Cook, Structure of ubiquitin refined at 1.8 Å resolution. *J. Mol. Biol.* **194**, 531–544 (1987).
30. G. Cornilescu, J. L. Marquardt, M. Ottiger, A. Bax, Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. *J. Am. Chem. Soc.* **120**, 6836–6837 (1998).
31. R. A. Laskowski, M. W. MacArthur, D. S. Moss, J. M. Thornton, PROCHECK: A program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**, 283–291 (1993).
32. O. F. Lange, N. A. Lakomek, C. Farès, G. F. Schröder, K. F. Walter, S. Becker, J. Meiler, H. Grubmüller, C. Griesinger, B. L. de Groot, Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science* **320**, 1471–1475 (2008).
33. S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, E. Lindahl, GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845–854 (2013).
34. G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, G. Bussi, PLUMED 2: New feathers for an old bird. *Comput. Phys. Commun.* **185**, 604–613 (2014).
35. S. Piana, K. Lindorff-Larsen, D. E. Shaw, How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* **100**, L47–L49 (2011).
36. N. Salvi, S. Ulzega, F. Ferrage, G. Bodenhausen, Time scales of slow motions in ubiquitin explored by heteronuclear double resonance. *J. Am. Chem. Soc.* **134**, 2481–2484 (2012).
37. D. Sivia, J. Skilling, *Data Analysis: A Bayesian Tutorial* (Oxford Univ. Press, Oxford, NY, 1996).
38. K. J. Kohlhoff, P. Robustelli, A. Cavalli, X. Salvatella, M. Vendruscolo, Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. *J. Am. Chem. Soc.* **131**, 13894–13895 (2009).
39. W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, M. L. Klein, Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
40. B. Hess, H. Bekker, H. J. Berendsen, J. G. Fraaije, LINCS: A linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).
41. G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
42. Y. Shen, A. Bax, SPARTA+: A modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. *J. Biomol. NMR* **48**, 13–22 (2010).
43. M. Zweckstetter, A. Bax, Prediction of sterically induced alignment in a dilute liquid crystalline phase: Aid to protein structure determination by NMR. *J. Am. Chem. Soc.* **122**, 3791–3792 (2000).
44. M. Barfield, Structural dependencies of interresidue scalar coupling ${}^{h3}J_{NC'}$ and donor ^{1}H chemical shifts in the hydrogen bonding regions of proteins. *J. Am. Chem. Soc.* **124**, 4158–4168 (2002).
45. B. Vögeli, J. Ying, A. Grishaev, A. Bax, Limits on variations in protein backbone dynamics from precise measurements of scalar couplings. *J. Am. Chem. Soc.* **129**, 9377–9385 (2007).

Acknowledgments: We would like to thank A. Mira for useful discussions on the Bayesian method. **Funding:** This work received no funding. **Author contributions:** M.B., C.C., A.C., and M.V. designed the research and analyzed the data. M.B. and C.C. performed the research. M.B., C.C., and M.V. wrote the paper. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper itself and in the Supplementary Materials or are available upon request from the authors.

Submitted 27 August 2015
Accepted 21 November 2015
Published 22 January 2016
10.1126/sciadv.1501177

Citation: M. Bonomi, C. Camilloni, A. Cavalli, M. Vendruscolo, Metainference: A Bayesian inference method for heterogeneous systems. *Sci. Adv.* **2**, e1501177 (2016).