Principles of protein structural ensemble determination

Massimiliano Bonomi[1,*,#], Gabriella Heller[1,#], Carlo Camilloni[2], Michele Vendruscolo[1,*]

[1]Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK

[2]Department of Chemistry and Institute for Advanced Study,

Technische Universität München, D-85747 Garching, Germany

ABSTRACT

The biological functions of protein molecules are intimately dependent on their conformational dynamics. This aspect is particularly evident for disordered proteins, which constitute about one-third of the human proteome. Therefore, structural ensembles often offer more useful representations of proteins than individual conformations. Here, we describe how the well-established principles of protein structure determination should be extended to the case of protein structural ensembles determination. These principles concern primarily how to deal with conformationally heterogeneous states, and with experimental measurements that are averaged over such states and affected by a variety of errors. We first review the vast literature of recent methods that combine experimental and computational information to model structural ensembles, highlighting their similarities and differences. We then address some conceptual problems in the determination of structural ensembles and define future goals towards the establishment of objective criteria for the comparison, validation, visualization, and dissemination of such ensembles.

[*]To whom correspondence should be addressed: mb2006@cam.ac.uk; mv245@cam.ac.uk

[#]These authors contributed equally to this work

Introduction: Using a structural ensemble to represent the state of a protein

As protein molecules in their natural environments experience significant conformational fluctuations, in many cases structural ensembles can effectively represent their states and provide insights into the structural basis of their biological functions [1-6]. A structural ensemble can be defined as a set of conformations together with their corresponding statistical weights. Both the conformations and their statistical weights should be determined in a manner consistent with the available experimental and theoretical information. The statistical weight defines the extent to which a particular conformation is populated by a protein under well-defined experimental conditions. An ensemble of this type is thus a ‘statistical ensemble’, rather than an ensemble made up by multiple models of a given native state, which could be called ‘uncertainty ensemble’, as the multiple models reflect the limited information available on the system, rather than an intrinsic conformational heterogeneity.

The importance of using structural ensembles is particularly evident in the case of disordered proteins, which are a class of proteins that lack well-defined structures and populate instead a large number of states [7-11]. It has been suggested that these proteins make up about one-third of the human proteome [10,12] and that their dynamic nature facilitates multiple interactions, making them particularly important in regulation and signaling processes [10,13]. In this review we discuss how the development of rigorous principles of protein structural ensemble determination will lead to a structure-based understanding of the functions of proteins that are disordered or contain disordered regions.

Challenges in the determination of structural ensembles of proteins

Experiments alone may not lead to the accurate determination of protein structural ensembles. The challenges associated with the determination of protein structural ensembles by experimental methods alone are at least threefold [14], as observations: 1) are often time and ensemble averages over conformationally heterogeneous states; 2) provide sparse and sometimes ambiguous information; 3) are always subject to random and systematic errors.

1) Averages over conformationally heterogeneous states. In principle, heterogeneous states of proteins in solution could be distinguished by an experimental technique whose observation time is faster than the dynamics of the interconversion between states. However, in most cases this observation time is shorter than the typical timescale that one is interested to resolve, and as a result the observation is an average over multiple conformational states. This is true for highly versatile techniques such as nuclear magnetic resonance (NMR) spectroscopy, which is commonly used to characterize protein conformational fluctuations, and it has recently been recognized also for X-ray crystallography [15,16]. However, certain experiments, such as single-molecule Förster resonance energy transfer (FRET) or double electron-electron resonance (DEER), can generate distributions of the experimental observable over the entire structural ensemble, rather than averaged quantities [17,18]. In addition, it is also possible, for example using relaxation-dispersion NMR spectroscopy, to obtain specific information about individual sub-states [2]. Finally, other techniques provide information that does not have a thermodynamic nature, i.e. measurements that are neither averaged over structural ensembles, nor directly reflecting the populations of individual members. A notable example of this type is provided by chemical cross-linking/mass spectrometry (XL-MS) data [19].

2) Sparse and sometimes ambiguous information. Experimental techniques are typically sensitive only to specific properties, and thus provide sparse structural information about the conformational fluctuations of proteins. For example, FRET measurements provide an indirect measure of the distance between two protein sites, small-angle X-ray scattering (SAXS) reports on the probability distribution of the distances between all pairs of atoms, XL-MS data typically probe the proximity between lysine residues, and residual dipolar couplings (RDCs) provide information about the orientations of interatomic bonds relative to an external magnetic field. Furthermore, in some cases, experimental data provides ambiguous information [14]. For example, in the analysis of XL-MS data which was collected in the presence of multiple copies of the same protein, data cannot be univocally assigned to a specific copy, or in the analysis of nuclear Overhauser effect (NOE) NMR data, previous knowledge of the structure of the molecule is required [20].

3) Random and systematic errors. All experimental observations are affected by random or systematic errors. Random errors result from statistical fluctuations across multiple observations. Systematic errors might arise in several circumstances, for example from faults in the instrumentation or in its use by the experimenter, incorrect assignment of data points, or poor sample preparation. Each experimental technique is characterized by a different level of noise in the data, which should be taken into account when constructing structural models of the system. Crucially, noise and structural heterogeneity are difficult to disentangle. For example, a Gaussian distribution of measurements might arise from a collection of observations of a static system subject to random errors or rather from repeated measurements of a dynamic system in absence of noise.

Computational techniques alone are often insufficient to determine accurate structural ensembles. Computational methods such as Monte Carlo, molecular dynamics, or combined approaches have the potential to yield a complete description of the structures and dynamics of proteins. These methods generate structural ensembles whose statistical weights are determined by a theoretical model of the physical and chemical interactions of the system. The two main problems that this approach faces are that of: 1) force fields, and 2) conformational sampling.

1) Force fields. Commonly used models range from highly-detailed *ab initio* methods, to all-atom empirical force-fields, to coarse-grained approaches. Yet, even the most detailed and accurate models are still approximations of the actual interatomic interactions and therefore they will not be able to fully predict all the properties of the systems under study. This aspect is particularly relevant for highly dynamical systems that populate multiple states, because even minor inaccuracies in the force fields may result in relatively large errors in the predicted properties [21-23].

2) Conformational sampling. Limited computational resources only allow the simulation of finite timescales (for example microseconds), which are often shorter than the time of the actual biological processes of interest. This is especially true when a very detailed representation of the system is used, because high accuracy comes at a high computational cost. As a result, it is common procedure to compromise between accuracy and efficiency, depending on the size of the system and the desired timescale. Time scale issues are further alleviated via the use of enhanced sampling methods which accelerate the exploration of the conformational space [24]. However, even with these advanced techniques, the accuracy of the generated structural ensemble is still bound by the limits of the theoretical model.

The combination of experimental and computational methods may lead to the determination of accurate structural ensembles. Since neither experimental nor computational approaches alone can generate accurate structural ensembles capable of predicting multiple biologically relevant properties, a promising strategy to achieve this goal is to combine all sources of information available (**Figure 1**). Over the last decade, significant effort has been put forth towards the development of methods that combine experimental and theoretical information for structure determination [25]. While these approaches have also been directed towards structural ensemble determination, the challenges of how to take into account the averaging over conformationally heterogeneous states and the errors in the data have only recently been addressed.

Dealing with entropy: Current methods for structural ensemble determination

In this section we review some of the current methods for structural ensemble determination. Although this field is still in its infancy, a wide variety of techniques have already been proposed. We do not aim to provide a comprehensive summary of all the existing methods, but rather to give a concise overview of the most popular approaches (**Table 1**). For a more extensive treatment of this subject, we refer the reader to other existing reviews [26-30].

Generally speaking, methods for structural ensemble determination can be grouped into two categories [30], those following the maximum entropy principle and those inspired by the ‘Occam’s razor’ rule, which can also be called the maximum parsimony principle. Methods in the former class typically determine a large number of conformations by perturbing an initial (*a priori*) structural ensemble in order to match the experimental data. The perturbation is meant to generate a structural ensemble that is closest to the *a priori* ensemble but that also matches the experimental data. By contrast, methods based on the maximum parsimony principle are aimed at determining the minimum number of structures that can explain the experimental data. These methods require the definition of practical criteria to balance between the number of conformers of the ensemble and the quality of the fit with experimental data. The balance between number of conformers and quality of the fit is usually referred to as the problem of over-fitting in regularization algorithms.

Maximum entropy methods. Maximum entropy methods are typically based on the introduction of additional energy terms to classical molecular dynamics force fields. These additional terms are functions of the back-calculated experimental observables and their intensities are determined by Lagrange multipliers whose values should be computed to enforce agreement between experiments and simulations [31,32]. An alternative strategy to avoid these complicated calculations is the replica-averaged modelling [1,33-35]. In this approach, a set of replicas of the system is simulated in parallel and harmonic potentials are added to the molecular dynamics force field to restrain the averages of the experimental observables across the replicas close to the experimental measurements. This method has been effectively used to provide structural ensembles for a variety of systems, including disordered proteins making use of enhanced sampling techniques [34-37]. The equivalence between replica-averaged simulations and the maximum entropy Lagrange-multipliers solution was demonstrated recently [32,38]. Other more recent maximum entropy methods are aimed at matching distributions of experimental data by introducing additional restraints [39], sometimes in the form of metadynamics bias potentials, as for the case of 'ensemble-biased metadynamics' [40] and 'experiment directed metadynamics' [41].

All the maximum entropy methods described above incorporate experimental data directly into simulation by means of restraints between experimental and predicted data. Other maximum entropy approaches act *a posteriori* by reweighting a structural ensemble generated by molecular dynamics or other sampling techniques in order to determine the weights of the members of the ensemble that maximize the agreement with experimental data. In this class, we also include approaches that select a subset of components of the original ensemble and assign them identical weights. Among maximum entropy reweighting approaches we mention in particular the 'ensemble-refinement of SAXS' (EROS) method [42], the 'convex optimization for ensemble reweighting' method (COPER) [43], and the 'ENSEMBLE' method [44].

Maximum parsimony methods. Most current methods inspired by the maximum parsimony principle are based on reweighing techniques. These approaches include the 'ensemble optimization method' (EOM) [45], the 'selection tool for ensemble representations of intrinsically disordered states' (ASTEROIDS) [46], the 'sparse ensemble selection' (SES) method[47], the 'sample and select' (SAS) method [48], the 'maximum occurrence' (MaxOcc) method [49], the 'minimal ensemble search' (MES) method [50], and the 'basis-set supported SAXS reconstruction' (BSS-SAXS) method [51]. All these approaches differ in the way the initial ensemble is generated, the algorithm to select or reweight a subset of the initial ensemble to optimize the fit with experimental data, and the criterion to balance number of members of the ensemble with quality of the fit.

Some of the methods described above integrate an estimate of the error in the data in the generation of the structural ensemble, which is typically treated as a constant parameter determined by the experiments. However, in most situations this estimate only provides a lower bound on the real data uncertainty because it fails to account for systematic errors and the presence of outlier data points. Furthermore, the theoretical model to calculate an experimental observable from a structural ensemble, commonly referred to as *predictor* or *forward model*, is often inaccurate. We suggest that in an effective method each source of information used in the modelling should be weighted according to its reliability. Therefore, an accurate estimate of the level of noise and uncertainty in the measured and predicted data is a necessary condition to properly mix different experimental data with theoretical models and obtain accurate structural ensembles.

Taking errors into account: Bayesian inference methods

In a seminal paper [52], Rieping and co-workers presented a Bayesian inference method ('inferential structure determination', ISD) for single structure determination that combines prior information on a system with experimental data and accounts for errors in these data. ISD 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 are inferred along with the structure of the system. Since it can be argued that an ideal method for structural ensemble determination should account for variable and sometimes unknown errors, we regard the Bayesian method as a particularly appropriate framework to integrate multiple experimental data.

Several Bayesian *a posteriori* reweighting approaches for structural ensemble determination have been proposed in recent years. Among those inspired by the maximum entropy principle we mention the 'Bayesian ensemble refinement' method [53], the 'Bayesian ensemble SAXS' (BE-SAXS) method [54], the 'experimental inferential structure determination' (EISD) method [55], the ‘Bayesian energy landscape tilting’ (BELT) method [56], the ‘integrated Bayesian approach’ [57], the method of Sethi *et al.* [58], the ‘reference ratio’ method [59], and the ‘Bayesian inference of conformational populations’ (BICePS) method [60]. Two Bayesian reweighting methods obeying the maximum parsimony principle have also been presented, the ‘Bayesian inference of EM’ method (BioEM) [61] and the ‘Bayesian weighting’ (BW) method [62].

Only a few existing Bayesian methods incorporate experimental data directly as restraints to model structural ensemble of proteins. One of the earliest proposals, the ‘multi-state Bayesian modelling’ approach, is based on the maximum parsimony principle and was used to characterize the multiple structural states of histidine kinase PhoQ using cysteine-crosslinking data [63] and the mechanism of substrate recognition of the molecular chaperone Hsp90 [64]. More recently, two methods inspired by the maximum entropy principle have been proposed, the ‘Bayesian ensemble refinement’ method [53] and the ‘metainference’ method [65]. In both approaches, a set of N replicas of the system is simulated in parallel under the combined effect of prior information and an energy term that relates the experimental data to the average of the observable over the replicas. The intensity of this restraint on the structural ensemble is variable, depends on the unknown level of noise in the data, and scales linearly with N in presence of noise in the data. However, the specific form of the data energy term is different in the two approaches. In particular, the metainference data energy term explicitly scales more than linearly with N in absence of data noise, as requested by the maximum entropy principle. The metainference method is available in the popular open-source PLUMED library [66] and it has been combined with metadynamics in its parallel bias implementation [67] to accelerate sampling in complex biological systems [68] (**Figure 2**).

Major questions about the determination of structural ensembles of proteins

Are atomistic models of structural ensembles of proteins always needed? A common goal in structural biology is the generation of protein structures at atomic resolution. Although this is a desirable outcome, for complex proteins and their assemblies this might be a daunting task, especially in absence of a large amount of experimental and theoretical information. In these cases, an alternative but still relevant objective is to make testable predictions to shed light on the function of a given system. A valuable representation of the state of a protein is thus one that enables such predictions. Several examples of structural models suggest that interesting predictions could be made even at low resolution or coarse-grained level [69-71]. Furthermore, a coarse-grained representation of the system, and in general a poor *prior* information, can be compensated by the use of a large amount of experimental data, while simultaneously facilitating the sampling of the conformational landscape [65]. However, an atomistic representation of a system might be required to define the predictor (i.e. a forward model) of a given experimental observable. To this regard, one could choose to simplify the physico-chemical interactions while maintaining an atomistic resolution of the system [72] or alternatively to use multiple-resolutions forward models [73,74].

Is the determination of a structural ensemble an ill-defined problem? To answer to this question, one has to distinguish two points of view. From a first point of view, the determination of the members of the ensemble from experimental data that are averaged on the entire ensemble (the so-called inverse problem) is an ill-defined problem, in the sense that it allows multiple solutions, as different structural ensembles can fit the same averaged data. From a second point of view, the determination of a structural ensemble can be seen as a well-defined problem, at least in the case when the ensembles generated are experimentally indistinguishable. This happens when all the measurable quantities are predicted to be the same from the different structural ensembles. As long as different structural ensembles generate the same observable average quantities, and thus result in similar predictions for the system properties, they should be considered equivalent. Similarly, two different samples of the same distribution, like two independent molecular simulations, can be constituted of different components, but average quantities calculated using the two sets are identical. Therefore, structural ensemble determination can be seen as a well-defined problem from the perspective of performing measurements, but as an ill-defined problem from the point of view of determining the conformations of the individual members of the ensemble (**Figure 3**).

Should one use 'maximum entropy' or 'maximum parsimony' methods? We have classified the methods for structural ensemble determination into two classes: those determining large ensemble of conformations following the maximum entropy principle, and those determining a minimal set of conformations that fit the data, in the spirit of the maximum parsimony principle. Our recommendation is for maximum parsimony methods to be used to study systems characterized by the presence of a small number of relevant states or, in other words, systems with low entropy. Maximum entropy methods are instead particularly suitable in the case of high entropy systems, *i.e.* in presence of a relevant number of significantly populated states (**Figure 4**). Our view is that for disordered proteins maximum entropy methods should be preferable despite their often greater computational cost, since it is not always easy to estimate *a priori* the amount of entropy in a system. We also note that it is not yet clear whether maximum parsimony methods can identify the regions of the conformational space of maximal probability, whereas rigorous proofs are available for maximum entropy methods [31,32,38]. In this respect, the role of the *prior* information in maximum parsimony methods is crucial, especially in reweighting approaches in which the *prior* information is used to generate a pool of candidate structures from which a minimal ensemble is extracted. It is important to observe that states at low probability might also be relevant as far as predictions are concerned, because many experimental observables can be expressed as non-linear functions of the system coordinates. For example, FRET efficiency and NOE intensity depend on the distance d between two atoms as $1/d^6$. Therefore, even low populated conformers with small values of d can significantly affect the values of these quantities. It is thus challenging to set an absolute threshold in the populations of the individual conformers and define *a priori* what is relevant and what should be ignored in a minimal ensemble determined by maximum parsimony methods.

Should one use reweighting methods or direct use data as restraint? Structural ensemble determination methods can either directly integrate the experimental data into the generation of structural ensembles or act *a posteriori* on a pre-calculated set of conformations to optimize their weights to fit the input data. Reweighting schemes are inaccurate whenever the prior distribution used to generate the initial ensemble differs greatly from the final, reweighted distribution, as the efficiency depends on the overlap between the two distributions [75]. Reweighting methods have the advantage that they can be used at any moment in time when experimental data become available to refine structural ensembles previously calculated, at convenient computational cost. Despite this feature, to generate novel structural ensembles with experimental data at hand, we believe that the methods that directly use data as restraints should be preferred.

Community goals

Methods for structural ensemble determination are becoming increasingly popular and they will certainly continue to be further developed. We believe that a collective effort of the community is now needed to establish objective criteria and standards for structural ensemble comparison, validation, visualization, and dissemination. Here, we briefly discuss what we see as the most pressing present goals.

Goal 1: To establish robust methods of structural ensembles comparison. The availability of a wide variety of methods to generate structural ensembles prompts the question of how to compare structural ensembles obtained using different techniques. While many algorithms to address this problem have been proposed, there is still no consensus on what constitutes a satisfactory answer. Generally speaking, there are three main comparison techniques, which analyse structural ensembles based on their underlying probability distributions: 1) fast harmonic algorithms for small-scale fluctuations (or harmonic ensemble similarity), 2) structural clustering based methods in which the similarity is defined by the co-occurrence of conformations in both ensembles, and 3) dimensionality reduction methods where similarity is defined by projecting the ensembles into lower dimensional spaces. For a summary of the basic algorithms underlying these techniques and their respective advantages and limitations, see [76,77]. While these methods may yield valuable insights into the conformational differences obtained when generating structural ensembles with various inputs (different force fields, water models, varying sources of experimental data, etc.) [77], we encourage the community to establish robust comparisons based on the prediction of measurable quantities from the structural ensembles.

Goal 2: To establish robust methods of structural ensemble validation. Although it is very challenging to know the accuracy and precision of a structural ensemble, it is possible to define objective criteria to assess its quality using a combination of experimental validation and data removal or remodeling techniques. Independent structural data not used in the generation of the structural ensembles, such as chemical shifts for backbone conformations, NOEs for interatomic distances, scalar couplings for backbone and side chain dihedral angles and RDCs for interatomic bond orientations, can be used to validate the structural ensembles. Furthermore, the robustness of a structural ensemble can be quantified by assessing its dependency on particular data points. Upon removing a fraction of the input data, a large modulation in the resulting structural ensemble is a sign of poor accuracy. The development of standardized, rigorous methods of validation will help increase the accuracy and reliability of the structural ensemble determination methods. These methods will also help clarify the effects of variations in the experimental conditions on the structural ensembles, which in the case of disordered proteins can be highly significant, as well as provide confidence on the functional insights that can be obtained from the structural ensembles themselves.

Goal 3: To establish effective visual representations of structural ensembles. It is rather common to represent structural ensembles by overlaying multiple conformations in a single image. While visually appealing, such representations are usually inadequate to show the weights associated with individual conformations, and thus are limited in their information content. Better methods are those that represent free energy landscapes corresponding to the structural ensembles, which can be generated using dimensionality reduction algorithms. While a detailed discussion about these methods is beyond the scope of this review, we highlight low-dimensionality reduction algorithms such as sketch-map [78], isomap [79,80], and other nonlinear manifold learning algorithms offer effective representations of complex free energy landscapes [81,82]. Establishing a standard practice for structural ensemble representation will undoubtedly facilitate unbiased comparisons between such ensembles.

Goal 4: To distribute effectively structural ensembles to the community. Analogous to the Protein Data Bank (PDB), which is the reference repository for protein structures, the 'Ensemble Protein Database' (http://www.epdb.pitt.edu/) and the 'Protein Ensemble Database' [6] (http://pedb.vib.be) have been proposed to host structural ensembles of folded and disordered proteins, respectively. Unfortunately, it has yet to become the norm to upload structural ensembles to these databases, but without these shared data, progress in the field will be hindered. Furthermore, in order to calculate statistical averages over such structural ensembles, and hence to enable predictions to be made, the statistical weights of members of the ensembles will have to be deposited as well.

Goal 5: To improve force fields used in molecular simulations. One could envision iterative refinement procedures to improve the force fields, for example by modifying selected energy terms of a force field in order to reproduce increasingly well a 'target' structural ensemble determined from experimental data. While progress has already been made in this direction [83,84], a goal for the community is to establish automatic procedures to map back structural ensembles into corrections to the underlying force field and ensure that such modifications are portable to several different systems beyond those used in the refinement process.

Goal 6: To understand the role of dynamical effects in protein behaviour. In the definition of structural ensembles, we have not included the transition rates between different conformations. We should thus point out that such a definition is not suitable to describe phenomena that depend on dynamical effects. To study such phenomena, it will be necessary to develop additional methods for the determination of transition rates [2].

References

1. Lindorff-Larsen K, Best RB, Depristo MA, Dobson CM, Vendruscolo M: **Simultaneous determination of protein structure and dynamics**. *Nature* 2005, **433**:128-132.
2. Baldwin AJ, Kay LE: **NMR spectroscopy brings invisible protein states into focus**. *Nat Chem Biol* 2009, **5**:808-814.
3. Henzler-Wildman K, Kern D: **Dynamic personalities of proteins**. *Nature* 2007, **450**:964-972.
4. Boehr DD, Nussinov R, Wright PE: **The role of dynamic conformational ensembles in biomolecular recognition**. *Nat Chem Biol* 2009, **5**:789-796.
5. Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, et al.: **Atomic-level characterization of the structural dynamics of proteins**. *Science* 2010, **330**:341-346.
6. Varadi M, Kosol S, Lebrun P, Valentini E, Blackledge M, Dunker AK, Felli IC, Forman-Kay JD, Kriwacki RW, Pierattelli R, et al.: **pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins**. *Nucl Acids Res* 2014, **42**:D326-D335.

   •• This article presents a repository of protein structural ensembles, which will drive the improvement the methods for determining such ensembles.
7. Tompa P: **Intrinsically disordered proteins: a 10-year recap**. *Trends Biochem Sci* 2012, **37**:509-516.
8. Oldfield CJ, Dunker AK: **Intrinsically disordered proteins and intrinsically disordered protein regions**. *Annu Rev Bioch* 2014, **83**:553-584.

   • A comprehensive review of both early and more recent examples of disordered systems, the functions of these proteins, and characterization techniques.

9. Habchi J, Tompa P, Longhi S, Uversky VN: **Introducing protein intrinsic disorder**. *Chem Rev* 2014, **114**:6561-6588.

- A broad review of disordered proteins which introduces structural and conformational characteristics of these proteins, experimental techniques and computational prediction methods, function of disorder, among other relevant topics in the field.

10. van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT: **Classification of intrinsically disordered regions and proteins**. *Chem Rev* 2014, **114**:6589-6631.

11. Bhowmick A, Brookes DH, Yost S, Dyson HJ, Forman-Kay JD, Gunter D, Head-Gordon M, Hura GL, Pande VS, Wemmer DE: **Finding Our Way in the Dark Proteome**. *J Am Chem Soc* 2016, **138**:9730-9742.

12. Heller GT, Sormanni P, Vendruscolo M: **Targeting disordered proteins with small molecules using entropy**. *Trends Biochem Sci* 2015, **40**:491-496.

13. Wright PE, Dyson HJ: **Intrinsically disordered proteins in cellular signalling and regulation**. *Nat Rev Mol Cell Biol* 2015, **16**:18-29.

14. Schneidman-Duhovny D, Pellarin R, Sali A: **Uncertainty in integrative structural modeling**. *Curr Op Struct Biol* 2014, **28**:96-104.

15. Konrat R: **NMR contributions to structural dynamics studies of intrinsically disordered proteins**. *J Mag Res* 2014, **241**:74-85.

- A summary of the major advances in NMR methodologies for the study of structure and dynamics of disordered proteins, including the limitations and opportunities for future development of these techniques.

16. Fraser JS, van den Bedem H, Samelson AJ, Lang PT, Holton JM, Echols N, Alber T: **Accessing protein conformational ensembles using room-temperature X-ray crystallography**. *Proc Natl Acad Sci USA* 2011, **108**:16247-16252.

17. Schuler B: **Single-molecule FRET of protein structure and dynamics - a primer**. *J Nanobiotechnol* 2013, **11**.

18. Jeschke G: **DEER Distance Measurements on Proteins**. *Annu Rev Phys Chem* 2012, **63**:419-446.

19. Rappsilber J: **The beginning of a beautiful friendship: Cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes**. *J Struct Biol* 2011, **173**:530-540.

20. Nilges M: **Ambiguous distance data in the calculation of NMR structures**. *Folding Des* 1997, **2**:S53-S57.

21. Palazzesi F, Prakash MK, Bonomi M, Barducci A: **Accuracy of Current All-Atom Force-Fields in Modeling Protein Disordered States**. *J Chem Theory Comput* 2015, **11**:2-7.

22. Rauscher S, Gapsys V, Gajda MJ, Zweckstetter M, de Groot BL, Grubmuller H: **Structural Ensembles of Intrinsically Disordered Proteins Depend Strongly on Force Field: A Comparison to Experiment**. *J Chem Theory Comput* 2015, **11**:5513-5524.

23. Henriques J, Cragnell C, Skepo M: **Molecular Dynamics Simulations of Intrinsically Disordered Proteins: Force Field Evaluation and Comparison with Experiment**. *J Chem Theory Comput* 2015, **11**:3420-3431.

24. Abrams C, Bussi G: **Enhanced Sampling in Molecular Dynamics Using Metadynamics, Replica-Exchange, and Temperature-Acceleration**. *Entropy* 2014, **16**:163-199.

25. Ward AB, Sali A, Wilson IA: **Biochemistry. Integrative structural biology**. *Science* 2013, **339**:913-915.

26. Vendruscolo M: **Determination of conformationally heterogeneous states of proteins**. *Curr Op Struct Biol* 2007, **17**:15-20.

27. Fisher CK, Stultz CM: **Constructing ensembles for intrinsically disordered proteins**. *Curr Op Struct Biol* 2011, **21**:426-431.

28. Boomsma W, Ferkinghoff-Borg J, Lindorff-Larsen K: **Combining Experiments and Simulations Using the Maximum Entropy Principle**. *PLoS Comput Biol* 2014, **10**.

    • A perspective that, by summarising the state-of-the-art of maximum entropy methods, challenges the community to explore further in the direction of new methods to include error estimates and contribution.

29. van den Bedem H, Fraser JS: **Integrative, dynamic structural biology at atomic resolution-it's about time**. *Nat Methods* 2015, **12**:307-318.

    •• A review of the opportunities for integrative, dynamic structural biology and the potential of integrating X-ray crystallography, NMR and computer simulations to reveal the structural basis of protein conformational dynamics at atomic resolution.

30. Ravera E, Sgheri L, Parigi G, Luchinat C: **A critical assessment of methods to recover information from averaged data**. *Phys Chem Chem Phys* 2016, **18**:5686-5701.

•• A comprehensive review of methods of combining experimental and computational methods to determine structural ensembles of proteins with a specific focus on the classification between maximum entropy principle approaches and those modelling ensembles composed on a restricted number of conformations.

31. Pitera JW, Chodera JD: **On the Use of Experimental Observations to Bias Simulated Ensembles**. *J Chem Theory Comput* 2012, **8**:3445-3451.
32. Roux B, Weare J: **On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method**. *J Chem Phys* 2013, **138**.
33. Best RB, Vendruscolo M: **Determination of protein structures consistent with NMR order parameters**. *J Am Chem Soc* 2004, **126**:8090-8091.
34. Dedmon MM, Lindorff-Larsen K, Christodoulou J, Vendruscolo M, Dobson CM: **Mapping long-range interactions in alpha-synuclein using spin-label NMR and ensemble molecular dynamics simulations**. *J Am Chem Soc* 2005, **127**:476-477.
35. Allison JR, Varnai P, Dobson CM, Vendruscolo M: **Determination of the Free Energy Landscape of alpha-Synuclein Using Spin Label Nuclear Magnetic Resonance Measurements**. *J Am Chem Soc* 2009, **131**:18314-18326.
36. Camilloni C, Cavalli A, Vendruscolo M: **Replica-Averaged Metadynamics**. *J. Chem. Theory Comput.* 2013, **9**:5610-5617.
37. Cabrita LD, Cassaignau AM, Launay HM, Waudby CA, Wlodarski T, Camilloni C, Karyadi ME, Robertson AL, Wang X, Wentink AS, et al.: **A structural ensemble of a ribosome-nascent chain complex during cotranslational protein folding**. *Nat Struct Mol Biol* 2016, **23**:278-285.
38. Cavalli A, Camilloni C, Vendruscolo M: **Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle**. *J. Chem. Phys.* 2013, **138**:094112.
39. Roux B, Islam SM: **Restrained-Ensemble Molecular Dynamics Simulations Based on Distance Histograms from Double Electron-Electron Resonance Spectroscopy**. *J Phys Chem B* 2013, **117**:4733-4739.
40. Marinelli F, Faraldo-Gomez JD: **Ensemble-Biased Metadynamics: A Molecular Simulation Method to Sample Experimental Distributions**. *Biophys J* 2015, **108**:2779-2782.
41. White AD, Dama JF, Voth GA: **Designing Free Energy Surfaces That Match Experimental Data with Metadynamics**. *J Chem Theory Comput* 2015, **11**:2451-2460.

42. Rozycki B, Kim YC, Hummer G: **SAXS Ensemble Refinement of ESCRT-III CHMP3 Conformational Transitions**. *Structure* 2011, **19**:109-116.

43. Leung HTA, Bignucolo O, Aregger R, Dames SA, Mazur A, Berneche S, Grzesiek S: **A Rigorous and Efficient Method To Reweight Very Large Conformational Ensembles Using Average Experimental Data and To Determine Their Relative Information Content**. *J Chem Theory Comput* 2016, **12**:383-394.

44. Choy WY, Forman-Kay JD: **Calculation of ensembles of structures representing the unfolded state of an SH3 domain**. *J Mol Biol* 2001, **308**:1011-1032.

45. Bernado P, Mylonas E, Petoukhov MV, Blackledge M, Svergun DI: **Structural characterization of flexible proteins using small-angle X-ray scattering**. *J Am Chem Soc* 2007, **129**:5656-5664.

46. Nodet G, Salmon L, Ozenne V, Meier S, Jensen MR, Blackledge M: **Quantitative Description of Backbone Conformational Sampling of Unfolded Proteins at Amino Acid Resolution from NMR Residual Dipolar Couplings**. *J Am Chem Soc* 2009, **131**:17908-17918.

47. Berlin K, Castaneda CA, Schneidman-Duhovny D, Sali A, Nava-Tudela A, Fushman D: **Recovering a Representative Conformational Ensemble from Underdetermined Macromolecular Structural Data**. *J Am Chem Soc* 2013, **135**:16595-16609.

48. Chen Y, Campbell SL, Dokholyan NV: **Deciphering protein dynamics from NMR data using explicit structure sampling and selection**. *Biophys J* 2007, **93**:2300-2306.

49. Bertini I, Giachetti A, Luchinat C, Parigi G, Petoukhov MV, Pierattelli R, Ravera E, Svergun DI: **Conformational Space of Flexible Biological Macromolecules from Average Data**. *J Am Chem Soc* 2010, **132**:13553-13558.

50. Pelikan M, Hura GL, Hammel M: **Structure and flexibility within proteins as identified through small angle X-ray scattering**. *Gen Physiol Biophys* 2009, **28**:174-189.

51. Yang SC, Blachowicz L, Makowski L, Roux B: **Multidomain assembled states of Hck tyrosine kinase in solution**. *Proc Natl Acad Sci USA* 2010, **107**:15757-15762.

52. Rieping W, Habeck M, Nilges M: **Inferential structure determination**. *Science* 2005, **309**:303-306.

53. Hummer G, Kofinger J: **Bayesian ensemble refinement by replica simulations and reweighting**. *J Chem Phys* 2015, **143**:243150.

•• A comprehensive review of methods of combining experimental and computational methods to determine structural ensembles of proteins with a specific focus on the classification between maximum entropy principle approaches and those modelling ensembles composed on a restricted number of conformations.

54. Antonov LD, Olsson S, Boomsma W, Hamelryck T: **Bayesian inference of protein ensembles from SAXS data**. *Phys Chem Chem Phys* 2016, **18**:5832-5838.
55. Brookes DH, Head-Gordon T: **Experimental Inferential Structure Determination of Ensembles for Intrinsically Disordered Proteins**. *J Am Chem Soc* 2016, **138**:4530-4538.
56. Beauchamp KA, Pande VS, Das R: **Bayesian Energy Landscape Tilting: Towards Concordant Models of Molecular Ensembles**. *Biophys J* 2014, **106**:1381-1390.
57. Xiao X, Kallenbach N, Zhang YK: **Peptide Conformation Analysis Using an Integrated Bayesian Approach**. *J Chem Theory Comput* 2014, **10**:4152-4159.
58. Sethi A, Anunciado D, Tian JH, Vu DM, Gnanakaran S: **Deducing conformational variability of intrinsically disordered proteins from infrared spectroscopy with Bayesian statistics**. *Chem Phys* 2013, **422**:143-155.
59. Olsson S, Frellsen J, Boomsma W, Mardia KV, Hamelryck T: **Inference of Structure Ensembles of Flexible Biomolecules from Sparse, Averaged Data**. *PLoS One* 2013, **8**.
60. Voelz VA, Zhou GF: **Bayesian Inference of Conformational State Populations from Computational Models and Sparse Experimental Observables**. *J Comput Chem* 2014, **35**:2215-2224.
61. Cossio P, Hummer G: **Bayesian analysis of individual electron microscopy images: Towards structures of dynamic and heterogeneous biomolecular assemblies**. *J Struct Biol* 2013, **184**:427-437.
62. Fisher CK, Huang A, Stultz CM: **Modeling Intrinsically Disordered Proteins with Bayesian Statistics**. *J Am Chem Soc* 2010, **132**:14919-14927.
63. Molnar KS, Bonomi M, Pellarin R, Clinthorne GD, Gonzalez G, Goldberg SD, Goulian M, Sali A, DeGrado WF: **Cys-scanning disulfide crosslinking and bayesian modeling probe the transmembrane signaling mechanism of the histidine kinase, PhoQ**. *Structure* 2014, **22**:1239-1251.
64. Street TO, Zeng X, Pellarin R, Bonomi M, Sali A, Kelly MJ, Chu F, Agard DA: **Elucidating the mechanism of substrate recognition by the bacterial Hsp90 molecular chaperone**. *J. Mol. Biol.* 2014, **426**:2393-2404.
65. Bonomi M, Camilloni C, Cavalli A, Vendruscolo M: **Metainference: a Bayesian inference method for heterogeneous systems**. *Sci Adv* 2016, **2**:e1501177.

•• A Bayesian modelling approach that can be used to model heterogeneous systems by combining prior information with ensemble-averaged, noisy experimental data. The method is inspired by the maximum entropy replica-averaged modelling and incorporates data directly as restraint in a molecular dynamics simulation.

66. Tribello GA, Bonomi M, Branduardi D, Camilloni C, Bussi G: **PLUMED 2: New feathers for an old bird**. *Comp. Phys. Comm.* 2014, **185**:604-613.

• An open source library that integrates a wide variety of enhanced sampling techniques and analysis tools into popular molecular dynamics simulation packages. Among the order parameters that can be used in combination with the available biasing methods, there are several experimental observables, including NMR scalar couplings, chemical shifts and RDCs.

67. Pfaendtner J, Bonomi M: **Efficient sampling of high-dimensional free-energy landscapes with Parallel Bias Metadynamics**. *J. Chem. Theory Comput.* 2015, **11**:5062-5067.

68. Bonomi M, Camilloni C, Vendruscolo M: **Metadynamic metainference: Enhanced sampling of the metainference ensemble using metadynamics**. *Sci Rep* 2016, **6**:31232.

69. Alber F, Dokudovskaya S, Veenhoff L, Zhang W, Kipper J, Devos D, Suprapto A, Karni-Schmidt O, Williams R, Chait B, et al.: **The molecular architecture of the nuclear pore complex**. *Nature* 2007, **450**:695-701.

70. Zelter A, Bonomi M, Hoopman MR, Johnson R, Kim J, Riffle M, Umbreit NT, Moresco JJ, Yates JR, MacCoss MJ, et al.: **The molecular architecture of the Dam1 kinetochore complex is defined by cross-linking based structural modeling**. *Nat Comm* 2015, **6**:8673.

71. Deckert A, Waudby CA, Wlodarski T, Wentink AS, Wang X, Kirkpatrick JP, Paton JF, Camilloni C, Kukic P, Dobson CM, et al.: **Structural characterization of the interaction of alpha-synuclein nascent chains with the ribosomal surface and trigger factor**. *Proc Natl Acad Sci U S A* 2016, **113**:5012-5017.

72. Kukic P, Kannan A, Dijkstra MJJ, Abeln S, Camilloni C, Vendruscolo M: **Mapping the Protein Fold Universe Using the CamTube Force Field in Molecular Dynamics Simulations**. *PLoS Comput Biol* 2015, **11**.

73. Erzberger JP, Stengel F, Pellarin R, Zhang S, Schaefer T, Aylett CH, Cimermancic P, Boehringer D, Sali A, Aebersold R, et al.: **Molecular architecture of the 40SeIF1eIF3 translation initiation complex**. *Cell* 2014, **158**:1123-1135.

74. Frank AT, Law SM, Ahlstrom LS, Brooks CL: **Predicting Protein Backbone Chemical Shifts From C alpha Coordinates: Extracting High Resolution Experimental Observables from Low Resolution Models**. *J Chem Theory Comput* 2015, **11**:325-331.

75. Ceriotti M, Brain GAR, Riordan O, Manolopoulos DE: **The inefficiency of re-weighted sampling and the curse of system size in high-order path integration**. *Proc R Soc A* 2012, **468**:2-17.

76. Lindorff-Larsen K, Ferkinghoff-Borg J: **Similarity measures for protein ensembles**. *PLoS One* 2009, **4**:e4203.

77. Tiberti M, Papaleo E, Bengtsen T, Boomsma W, Lindorff-Larsen K: **ENCORE: Software for Quantitative Ensemble Comparison**. *PLoS Comput Biol* 2015, **11**.

78. Ceriotti M, Tribello GA, Parrinello M: **Simplifying the representation of complex free-energy landscapes using sketch-map**. *Proc Natl Acad Sci USA* 2011, **108**:13023-13028.

79. Tenenbaum JB, de Silva V, Langford JC: **A global geometric framework for nonlinear dimensionality reduction**. *Science* 2000, **290**:2319-+.

80. Das P, Moll M, Stamati H, Kavraki LE, Clementi C: **Low-dimensional, free-energy landscapes of protein-folding reactions by nonlinear dimensionality reduction**. *Proc Natl Acad Sci USA* 2006, **103**:9885-9890.

81. Roweis ST, Saul LK: **Nonlinear dimensionality reduction by locally linear embedding**. *Science* 2000, **290**:2323-2326.

82. Ferguson AL, Panagiotopoulos AZ, Debenedetti PG, Kevrekidis IG: **Systematic determination of order parameters for chain dynamics using diffusion maps**. *Proc Natl Acad Sci USA* 2010, **107**:13597-13602.

83. Giorgetti L, Galupa R, Nora EP, Piolot T, Lam F, Dekker J, Tiana G, Heard E: **Predictive Polymer Modeling Reveals Coupled Fluctuations in Chromosome Conformation and Transcription**. *Cell* 2014, **157**:950-963.

84. Vasile F, Civera M, Belvisi L, Potenza D, Tiana G: **Thermodynamically-Weighted Conformational Ensemble of Cyclic RGD Peptidomimetics from NOE Data**. *J Phys Chem B* 2016, **120**:7098-7107.

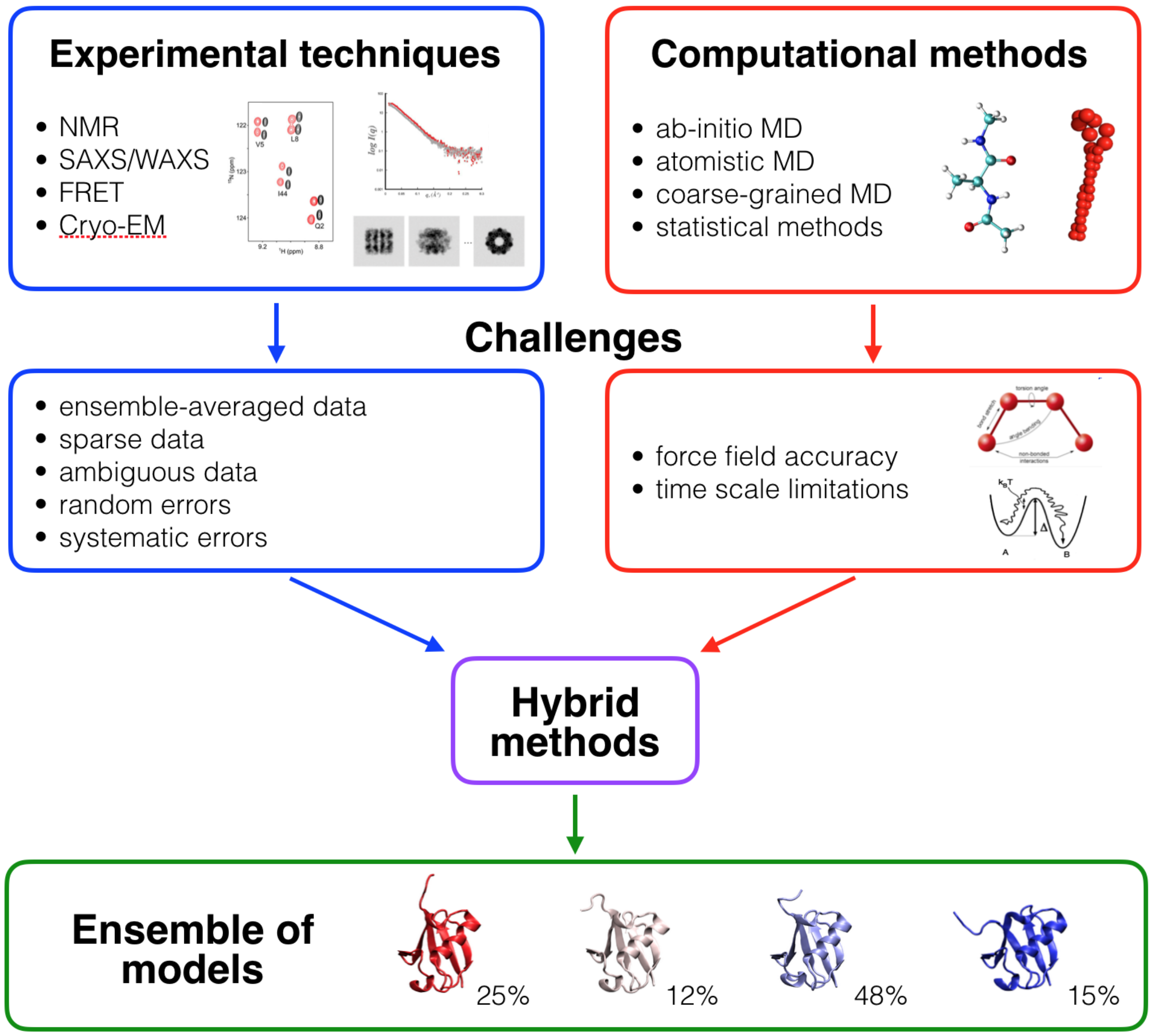

Figure 1. Determination of protein structural ensembles by combining experimental and theoretical methods. The combined use of experimental and computational techniques can lead to the accurate determination of structural ensembles of proteins by overcoming the limitations of individual techniques. Experimental methods on conformationally heterogeneous states of proteins, such as NMR spectroscopy, SAXS/WAXS and cryo-EM, typically provide ensemble-averaged, sparse, and sometimes ambiguous data affected by random and systematic errors. Computational methods, such as molecular dynamics simulations, are affected by the inaccuracies of the underlying force fields and by the limited timescales accessible to the simulations.

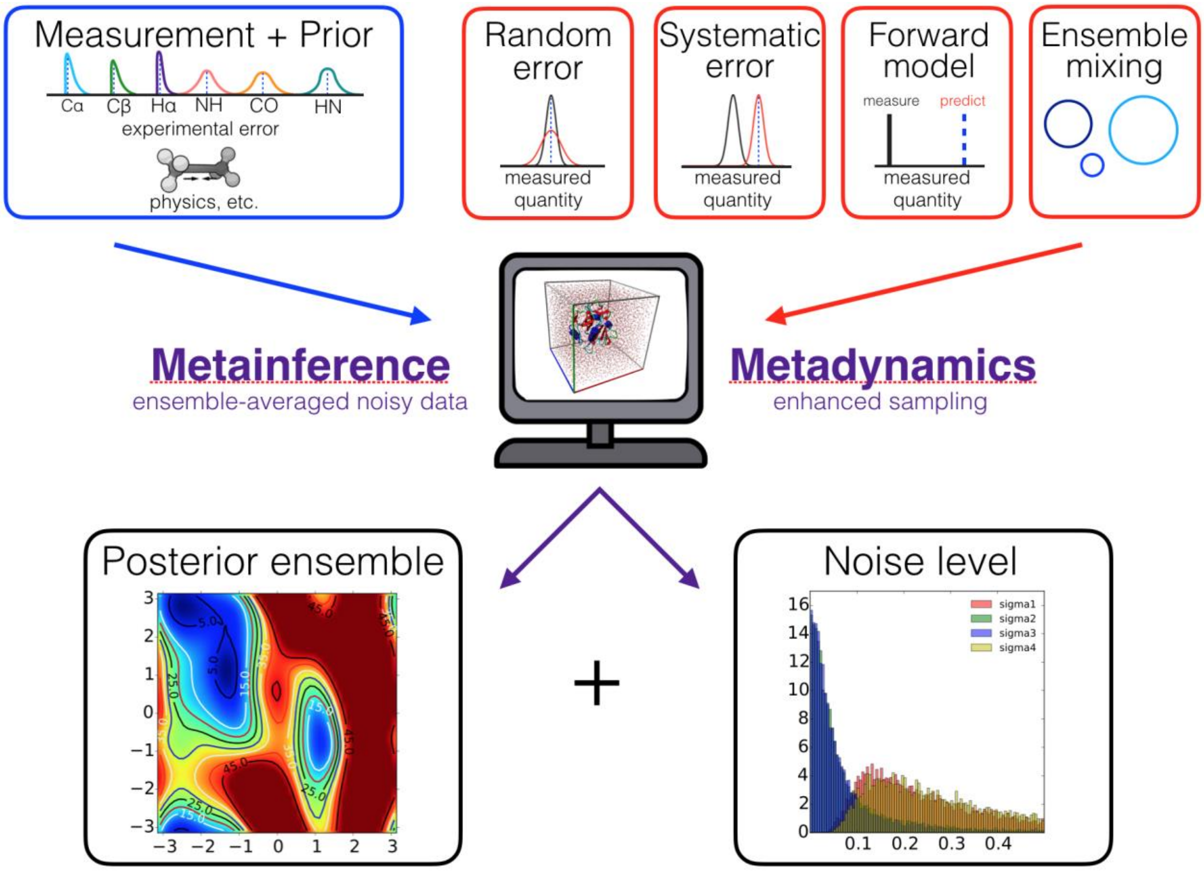

Figure 2. Determination of protein structural ensembles using the 'metadynamic metainference' method. The metadynamic metainference approach [68] combines the 'metainference' method [65], which models heterogeneous systems by integrating noisy, ensemble-averaged experimental data and prior knowledge, with the enhanced conformational sampling provided by the 'parallel bias metadynamics' method [67] to yield a posterior ensemble and the respective noise levels.

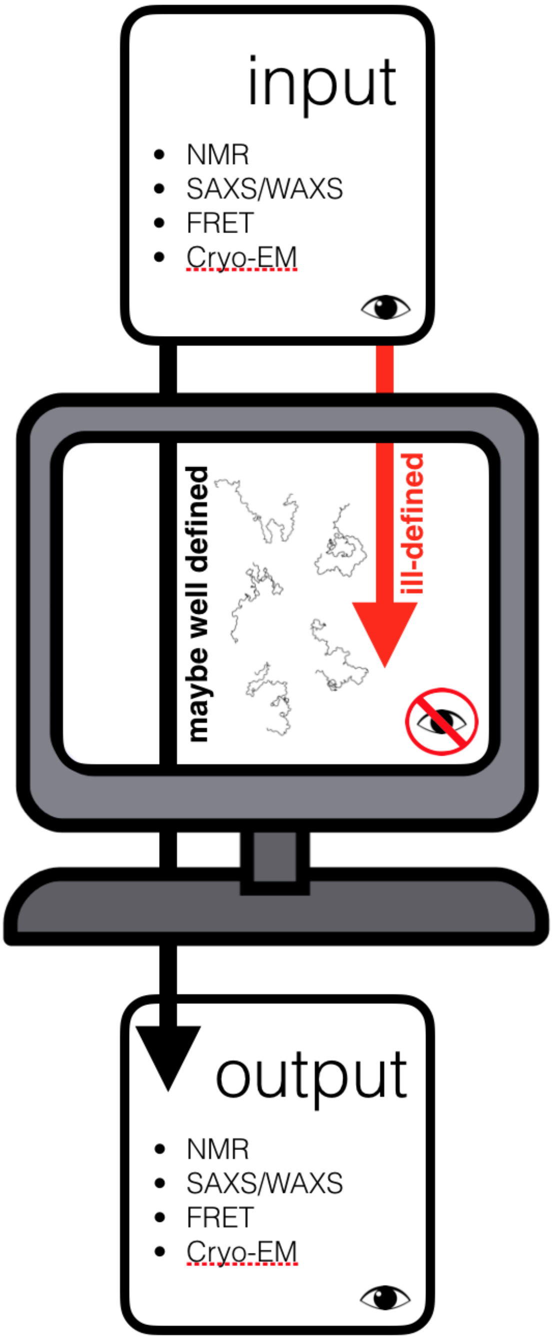

Figure 3. The determination of a structural ensemble can be a well-defined problem. The determination of a structural ensemble from a given set of experimental data may be considered as an ill-defined problem (i.e. a problem that admits multiple solutions) because there are many ensembles that fit equally well the data. However, when these ensembles are not directly distinguishable by experimental measurements, the determination of a structural ensemble can be in fact considered a well-defined problem. In this case, the many ensembles that fit the data used to determine them are effectively indistinguishable because they give rise to the same predictions for independent measurable quantities not used to determine them.

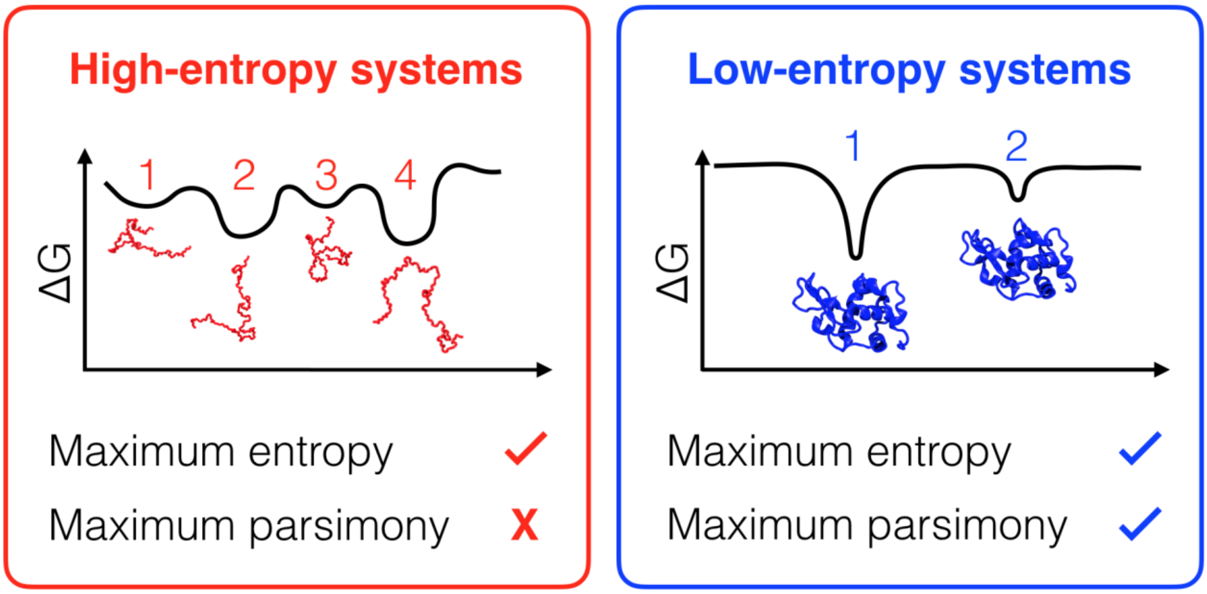

Figure 4. Comparison of maximum entropy and maximum parsimony methods. Maximum entropy methods are particularly suitable to study protein states characterized by high degree of conformational heterogeneity (i.e. with high entropy), while maximum parsimony methods can be used for ordered protein states (i.e. with low entropy).

Table 1. Summary of available methods for structural ensemble determination. For each method, we report the name, the year of the original paper, if it is inspired by the maximum entropy (ME) or maximum parsimony (MP) principles, if it is based on Bayesian statistics, if it deals with errors in the data and with ensemble-averaged data, and if the data are used directly as restraints or in *a posteriori* reweighting procedure.

ID	Name	Year	ME	MP	Bayes	Data errors	Ensemble averaged data	Restraint	Reweight	Ref.
1	Maximum entropy restraints	2012	x				x	x		[31]
2	Maximum entropy restraints	2013	x				x	x		[32]
3	Replica-averaged metadynamics	2013	x				x	x		[36]
4	Maximum entropy restraints for distance histograms	2013	x					x		[39]
5	Ensemble-Biased Metadynamics	2015	x					x		[40]
6	Experiment directed metadynamics	2015	x					x		[41]
7	EROS	2011	x			x	x		x	[42]
8	COPER	2015	x			x	x		x	[43]
9	ENSEMBLE	2001	x				x		x	[44]
10	EOM	2007		x			x		x	[45]
11	ASTEROIDS	2009		x			x		x	[46]
12	SES	2013		x		x	x		x	[47]
13	SAS	2007		x			x		x	[48]
14	MaxOcc	2010		x			x		x	[49]
15	MES	2009		x			x		x	[50]
16	BSS-SAXS	2010		x			x		x	[51]

17	Bayesian ensemble refinement	2015	x		x	x	x	x	x	[53]
18	BE-SAXS	2016	x		x	x	x		x	[54]
19	EISD	2016	x		x	x	x		x	[55]
20	BELT	2014	x		x	x	x		x	[56]
21	Integrated Bayesian Approach	2014	x		x	x	x		x	[57]
22	Sethi et al.	2013	x		x	x	x		x	[58]
23	Reference ratio method	2013	x		x	x	x		x	[59]
24	BICePS	2014	x		x	x	x		x	[60]
25	BioEM	2013		x	x	x	x		x	[61]
26	BW	2010		x	x	x	x		x	[62]
27	Multi-state Bayesian modeling	2014		x	x	x	x	x		[63]
28	Metainference metadynamics	2016	x		x	x	x	x		[68]