# The emerging role of physical modeling in the future of structure determination

Kari Gaalswyk[a,1], Mir Ishruna Muniyat[a,1], Justin L. MacCallum[a,*]

[a]*Department of Chemistry, University of Calgary, Calgary, AB, Canada*

## Abstract

Biomolecular structure determination has long relied on heuristics based on physical insight; however, recent efforts to model conformational ensembles and to make sense of sparse, ambiguous, and noisy data have revealed the value of detailed, quantitative physical models in structure determination. We review these two key challenges, describe different approaches to physical modeling in structure determination, and illustrate several successes and emerging technologies enabled by physical modeling.

## Highlights

- Quantitative physical modeling is emerging as a key tool in structure determination

- There are different approaches to incorporate physical modeling into structure determination

- Modeling conformational ensembles and making sense of sparse, noisy, and ambiguous data are two challenges where physical modeling can play a prominent role

## Introduction

Heuristics derived from physical insight have always played an import role in biomolecular structure determination. However, more rigorous quantitative physical models are increasingly used to transform experimental data into structures and ensembles. Physical approaches become more important as the biomolecular system of study becomes more flexible and conformationally heterogeneous (Figure 1), and as experimental data becomes sparse, ambiguous, or noisy (Figure 2). Systems with these characteristics have recently come into focus, due to both the recognition of the importance of conformational heterogeneity and the emerging range of experimental techniques that can provide incomplete information about protein structures [1–5].

Physical modeling has become increasingly powerful over time, driven by improvements in computer power, improved models of energy landscapes [6–8], and improved algorithms for conformational [9–12] and data-driven [13–17] sampling. Combined with advances in experimental methodology, these developments are leading to a new era

in structural biology where physical modeling plays a pivotal role [18–20].

In this review, we outline two challenges where physical modeling can make contributions to structure determination, overview some recent successes, and provide a perspective on emerging areas where physical modeling will be important.

## There are several emerging challenges in structural biology.

*Challenge 1: Modeling conformational ensembles*

When we refer to "the structure" of a biomolecular system, we are actually referring to some continuous cloud of structures in the neighborhood of a representative structure. While historically the single structure viewpoint has dominated in structural biology, there is increasing recognition of the importance of heterogeneity and dynamics.

Most measurements in structural biology are ensemble averages, where the observed signal comes from the average across many molecules. The challenge of interpreting such averaged data increases as the conformational ensemble becomes more heterogeneous. A simple thought experiment illustrates the central concept (Figure 1), where three systems have the same average for some observable, but different conformational distributions. One system (orange) is tightly clustered, where the average conformation provides an excellent representation of the ensemble. Another system (green) has a broad distribution, where the average conformation is only somewhat representative. The final system (blue) has a multimodal distribution, where the average conformation is improbable and not representative of the underlying ensemble at all. As the experimental average is the same in each case, modeling is critical to making correct inferences about the ensemble.

---

*Corresponding author, email: justin.maccallum@ucalgary.ca
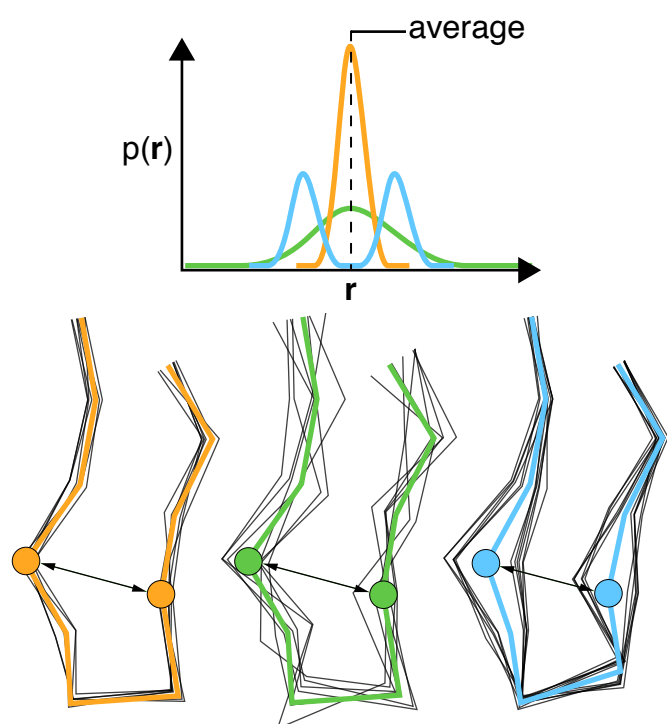[1]KG and MIM contributed equally to this work.

Figure 1: Most experiments measure ensemble averages, which poses a challenge as systems become more flexible, heterogeneous, and dynamic. This figure illustrates a thought experiment, comparing three different ensembles with the same average for some observable, but different conformational ensembles.

*Challenge 2: Making sense of sparse, ambiguous, and noisy data*

An increasing variety of experimental methods can provide partial information about the structure of a system [1–5]. While these experiments provide only an incomplete picture, their appeal is that they are often applicable to a wide range of systems, including those where traditional approaches have proven intractable.

Figure 2 shows several common pathologies. First, the data may be sparse, often only providing information about a few degrees of freedom. Second, the data may be ambiguous, where there are multiple molecular features that could explain a particular signal, e.g. an NMR experiment might tell us that two protons are close together, but not specifically which ones. Finally, experimental data is almost always corrupted by noise, which must be interpreted as such to avoid over-fitting. Noise comes in many forms, ranging from simple additive noise (often modeled by an appropriate distribution, e.g. Gaussian noise) to more challenging cases where experimental artifacts lead to the presence of false-positive and false-negative signals.
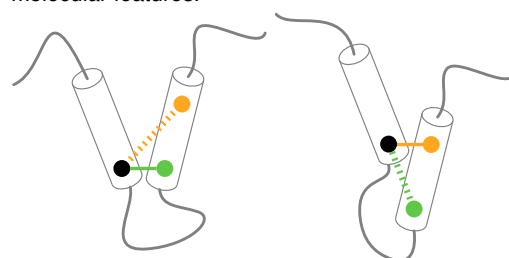
**What do we mean by physical modeling?**

The term "physical modeling" encompasses many approaches, ranging from physically-motivated heuristics to models rooted in rigorous statistical mechanics. Heuristic



(a) **Sparse:** many possible structures agree with data.
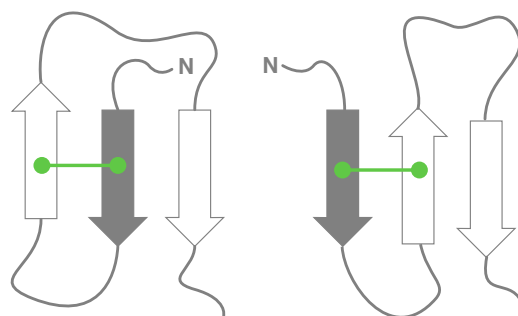
(b) **Ambiguous:** signal can be explained by multiple molecular features.

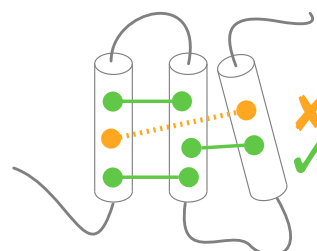(c) **Noisy:** some signals are spurious and do not correspond to true molecular features.

Figure 2: Conceptual illustration of the challenges faced in integrative structural biology and other applications where the data is sparse, ambiguous, and noisy.

approaches are motivated by physical considerations and empirical observations. One example is the use of stereochemical restraints during the refinement of X-ray crystal structures [21] that prevent physically impossible bond lengths and overlap between atoms, even though these unrealistic features might lead to naïve improvements in the agreement with experimental data. These heuristics are not a comprehensive physical description of biomolecular structure—clearly, one could not hope to predict the correct fold of a protein using only simple stereochemical restraints.

Conversely, statistical mechanics is a rigorous, comprehensive theory that connects the probability $p(\vec{r})$ of observing a particular conformation with the potential energy $V(\vec{r})$ through the Boltzmann distribution:

$$p(\vec{r}) = Z^{-1} \exp\left[-\frac{V(\vec{r})}{RT}\right], \tag{1}$$

where $R$ is the gas constant, $T$ is the absolute temperature, and $Z$ is a normalization constant called the partition function. Typically, the potential energy is modeled using an empirical approximation called a *force field* [6, 7]. Samples from $p(\vec{r})$ are generated using molecular dynamics or Monte Carlo simulations, often augmented by various enhanced sampling algorithms [10, 12, 13, 22].

Rosetta is another example of physical modeling [8]. Although the underlying philosophy and parameterization of Rosetta differ substantially from those of statistical mechanical models, the underlying goal is essentially the same—to reproduce the conformational landscape of a biomolecular system of interest.

## There are different approaches to incorporating physical models into structure determination.

Constructing structural models of a biomolecular system from one or more experimental datasets can be cast as an inference problem that can be solved through a variety of approaches (reviewed in [19, 20]). However, the distinctions between these approaches can often be subtle. In this section, we outline several characteristics that distinguish different approaches.

*Characteristic 1: What is the nature of the likelihood function?*

The likelihood, $\mathcal{L}(\theta|D) \sim \mathbb{P}(D|\theta)$, is central to many methods, where $D$ is the observed data and $\theta$ is a set of parameters specifying the structural ensemble, e.g. atomic coordinates. The likelihood is typically built by combining a *forward model*, which calculates the experimental observable from a given structure or ensemble, with a *noise model*, which calculates the probability of a given deviation from the measured value [23, 24]. Success depends critically on the quality of the forward model, as any inaccuracies lead directly to errors in the final ensemble.

One must distinguish between likelihoods that consider single structures from those that consider ensembles. We refer to these as *single-structure* and *ensemble* likelihoods; they are also referred to as likelihood *functions* and *functionals*, respectively. Use of a single-structure likelihood will make all members of the ensemble match experiment, which may be acceptable for systems with only limited heterogeneity. But ensemble likelihoods, which ensure that model averages match observations, are essential for more heterogeneous systems.

The relationship between structure and observable may be non-linear. For example, NOE and FRET measurements are proportional to $\langle r^{-6} \rangle$, leading to averages dominated by (potentially) rare short-distance conformers. Similarly, observation of chemical cross-linking indicates occasional proximity of two residues, but as the cross-links are irreversible once made, one cannot relate the average degree of cross-linking to the average distance between the residues. Furthermore, rather than only providing averages, some experiments, like EPR or single-molecule FRET, can provide *distributions*, which are potentially far richer. In all cases, the likelihood function must correctly capture these relationships in order to avoid biasing the calculated ensemble.

*Characteristic 2: What statistical formalism is used?*

Maximum likelihood (ML) and Bayesian statistical approaches are two common ways to solve the structural inference problem. Maximum likelihood methods aim to find the single best set of parameters, $\hat{\theta}$, that maximize the likelihood function. Näive ML methods determine the parameters based entirely on the data, making these methods sensitive to noise and notoriously prone to over-fitting. To mitigate this, it is common to use penalized ML methods, where the aim is to minimize:

$$\hat{\theta} = \arg\min_{\theta} \left( -\ln \mathcal{L}(\theta|D) + U(\theta) \right), \qquad (2)$$

where $U(\theta)$ is a penalty function that may include simple restraints on stereochemistry, more detailed energetic models given by force fields, or *ad hoc* penalty terms motivated by physical considerations, e.g. the use of restraints on crystallographic B-factors [25]. Such penalty terms are a form of regularization and ensure individual conformations are physically reasonable, preventing over-fitting.

The Bayesian approach offers a different perspective [23, 24], where one seeks to find the joint distribution of parameters given the data. Bayes theorem is a simple and elegant statement,

$$p(\theta|D) \propto \mathcal{L}(D|\theta)p(\theta), \qquad (3)$$

that combines prior understanding with new information in a statistically consistent way. The quantity of interest is the posterior distribution, $p(\theta|D)$, which is obtained by combining the likelihood with the prior, $p(\theta)$. The prior, often modeled as the Boltzmann distribution, plays a similar role to the penalty terms in ML methods, in that it encodes our knowledge about what structures are *a priori* probable.

A major differences between the two approaches is that ML methods provide a *point estimate* of the best parameters, whereas Bayesian methods produce a *joint distribution* of parameters—or, more commonly, a set of samples from the joint distribution generated by molecular dynamics or Monte Carlo sampling. Bayesian methods naturally provide an estimate of uncertainty that also captures the coupling between different parameters. Bayesian maximum *a posteriori* (MAP) estimators seek to find the mode of the posterior distribution, i.e. the single most likely set of parameters, blurring the lines between ML and Bayesian approaches.

Bayesian approaches readily allow for the inclusion of "nuisance parameters", which are unobserved, but nevertheless influence the results of inference. Examples include the true values of experimental observables (opposed

3

to noisy observations) or the exact values of empirical constants used in forward functions. The Bayesian approach [23, 24] is to treat these as nuisance parameters, with appropriate priors, that are inferred jointly with the rest of the model. The resulting distribution gives the likely values of the nuisance parameters and their influence on the generated ensemble.

Many approaches used in structural biology do not neatly fall into either the ML or Bayesian frameworks. These are often more *ad hoc* combinations of physically-motivated scoring functions and sampling strategies that do not produce a well-defined ensemble. For example, traditional NMR refinement generates collections of low-energy structures, but these do not correspond to well-defined statistical or thermodynamic ensembles. Although these methods have less rigorous statistical underpinning, they are very common in structural biology and obviously quite useful.

*Characteristic 3: What principle is used to regularize ensembles?*

The previous section outlined how penalty terms or priors can be used to regularize *individual structures*. However, without additional regularization, *ensemble models* become prone to over-fitting due to poor data-to-parameter ratios. For example, it is uncommon to see multi-copy refinement of X-ray crystal structures [26]. Phillips and co-workers undertook a systematic study of 50 experimental structures, and found that adding up to, on average, ~ 10 copies yielded improved models [27], with over-fitting occurring beyond that. *Ensemble regularization* methods can help avoid such over-fitting.

One approach is to use the *principle of maximum parsimony*, which seeks to find a minimum representative set of conformations that adequately explain the experimental data. These methods are typically based on re-weighting or selection. A pool of conformations is generated from a prior distribution using Monte Carlo or molecular dynamics sampling. Weights are then assigned to each conformation to bring calculated averages into agreement with experimental observations. A variety of approaches are possible [28–35] with several ways to choose the number of representative conformations, including user specification [29–31], clustering [32, 33], and penalty terms or priors that favor sparse models where most weights are zero [35].

Another approach to ensemble regularization is based on the *principle of maximum entropy* (MaxEnt), which posits that the distribution that best reflects our current state of knowledge is one that agrees with experimental observations, while simultaneously maximizing entropy [36, 37]. The Boltzmann distribution (Eq. 1) is the MaxEnt distribution over conformations, subject to a constraint relating the average energy and temperature [37]. When experimental observations are included, one arrives at [38]

$$p(\vec{r}) = \frac{e^{-\frac{V(\vec{r})}{RT} - \sum_i \lambda_i f_i(\vec{r})}}{\int e^{-\frac{V(\vec{r})}{RT} - \sum_i \lambda_i f_i(\vec{r})} d\vec{r}}, \qquad (4)$$

where $V(\vec{r})$ is the potential energy, and $\{\lambda_i\}$ are Lagrange multipliers that must be determined in order satisfy agreement between model averages and observations:

$$\int p(\vec{r}) f_i(\vec{r}) d\vec{r} = f_i^{(exp)}, \qquad (5)$$

where $f_i(\vec{r})$ is the function computing the $i$th observable and $f_i^{(exp)}$ is the corresponding experimental observation. Equivalently, one can minimize the Kullback-Leibler divergence between the inferred ensemble and the Boltzmann distribution, while satisfying the constraints of Eq. 5.

MaxEnt methods fit the Lagrange multipliers, rather than fitting the conformations directly. Because there is one Lagrange multiplier for each observation, the data to parameter ratio remains constant as the number of conformations increases, which allows for large ensembles without over-fitting. The magnitude of the Lagrange multipliers provides insight into how strongly the prior was perturbed to match each experimental observation.

A variety of methods exist for MaxEnt ensemble determination, as recently reviewed in [19, 20]. Approaches include re-weighting [39, 40], iterative schemes [38, 41–43], time-dependent potentials [44, 45], and restrained ensemble or replica-averaged schemes [40, 46–52]. Pitera and Chodera [38] identified a link between Eq. 4 and restrained ensemble schemes, which has been further clarified [40, 52, 53].

Although potentially less prone to over-fitting, care is still required when using maximum parsimony and MaxEnt methods. These approaches are sensitive to the quality of the prior distribution and can be expected to perform poorly when the prior distribution differs substantially from the true distribution.

**The term "ensemble" is highly overloaded in structural biology.**

In statistical mechanics, the term "ensemble" has a specific technical meaning: the probability distribution over all possible conformations of the system under specified conditions. Unfortunately, in structural biology it has become common to refer to almost any collection of conformations as an ensemble, which can be confusing.

There are several reasons that an ensemble may be heterogeneous. First, the ensemble may be a *thermodynamic ensemble*, where the heterogeneity is intended to reflect the true physical heterogeneity of the system at equilibrium. Second, the ensemble may be an *uncertainty ensemble*, where the heterogeneity reflects that there may be many possible structures compatible with sparse, ambiguous, or noisy data. Third, one can also distinguish between uncertainty ensembles composed of samples from well-defined statistical distributions, e.g. from Bayesian approaches, and those that are simply collections of structures generated by some procedure. Caution is required

when interpreting uncertainty ensembles, as the heterogeneity present is not necessarily predictive of the true heterogeneity.

Whenever one interprets an "ensemble" there are several key characteristics that must be considered. Does the ensemble come from a maximum likelihood or Bayesian approach, or from some other method? Is a single-structure or ensemble likelihood used? Do the structures sampled come from a well defined distribution? How are errors modeled? What priors or penalty functions are used?

## Physical modeling offers solutions to challenges in structural biology.

Hummer and co-workers introduced a Bayesian ensemble refinement method BioEN, a combination of replica ensemble refinement and the Ensemble Refinement of SAXS (EROS) method, combining the principles of both restraining and reweighing [40].

Ensemble heterogeneity explains much of the difficulty in characterizing intrinsically disordered proteins (IDPs) experimentally, as they are ensembles of inter-converting conformations [54, 55]. The Bayesian weighting method is an approach for characterizing an ensemble of IDPs where the weights are defined using a Bayesian estimate from calculated chemical shift data [33]. This method has been successful in determining the relative fractions of mutated structures in an ensemble for aggregative proteins [56].

Of all biomolecules, RNA has been perhaps the most challenging to simulate with molecular dynamics, as current force fields are not accurate enough to reproduce experiments [45]. Cesari and co-workers used a MaxEnt approach to bias RNA simulations to match J-coupling experiments [45] and demonstrate that this can be used to develop a self-consistent, transferable force field correction.

High ambiguity driven biomolecular docking (HADDOCK), is a data-driven docking approach, that can take highly ambiguous data from different sources and convert them into distance restraints to guide docking processes [57, 58]. HADDOCK has been used to study protein complex interfaces using cryo-EM data [59] and protein ligand complexes using sparse intermolecular NOEs [60].

The Integrative Modeling Platform (IMP), is a flexible software suite aimed at integrative structural biology, which facilitates development of integrative applications, models and methods, and allows incorporation of data from diverse sources [15]. Large protein complex structures have been modeled with IMP using *in vivo* FRET data through a Bayesian approach [61], and using a combination of cross-linking data with biochemical and EM localization data [62].

Rosetta is an extensive software suite aimed at protein structure prediction and molecular design. There are several applications of Rosetta with sparse experimental data, where Monte Carlo-based fragment assembly is guided towards native structures by data [63]. Backbone

chemical shifts and distance restraints have been used to guide structure determination [64]. Also, paramagnetic relaxation enhancement (PRE) [65], pseudo-contact shift (PCS) [66], and residual dipolar coupling (RDC) [67] restraints have been used to similar effect. Recently, the RASREC algorithm was developed, which yields better models with narrower sampling [17, 68] and has been applied to NMR on deuterated samples up to 40 kDa [69, 70].

Metainference, a recent approach based on Bayesian inference, can address statistical and systematic errors in data produced by high-throughput techniques, and can handle experimental data averaged over multiple states [14]. It is suitable for studying structural heterogeneity in complex macromolecular systems. A combination of Metainference and Parallel-bias Metadynamics (PBMetaD), an accelerated sampling technique, provides an efficient way of simultaneously treating error and sampling configuration space in all-atom simulations [9]. Coupling Metainference and Metadynamics has been particularly successful in characterizing structural ensembles of disordered peptides [71, 72].

Modeling Employing Limited Data (MELD) is a Bayesian approach that combines statistical mechanics, detailed all-atom physical models [7], and enhanced sampling to infer protein structures from sparse, ambiguous, and noisy data [13]. MELD was specifically designed to be robust in the presence of false-positive signals, and has been applied to EPR, NMR, and evolutionary data [13], *de novo* prediction of protein structures based on simple heuristics [73, 74], and mutagenesis guided peptide-protein docking [75, 76].

## Physical modeling is enabling emerging techniques in structural biology.

Advances in physical modeling will be key to enabling technologies for new approaches to structure determination. Below we outline just a few—of many—emerging techniques where the ability to model ensembles and to successfully treat sparse, ambiguous, and noisy data will be critical.

Chemical cross-linking detected by mass spectrometry is emerging as a potentially powerful tool in structure determination. Developments have focused on improvements in instrumentation [4, 77], cross-linking chemistries [78–80], and data analysis [78, 79, 81, 82]. These techniques are extremely sensitive, but the data can be highly ambiguous, both false-positive and false-negative signals are common, and one cannot relate the degree of cross-linking to the average distance. Such data has recently been used as restraints to guide Monte Carlo [83], molecular dynamics [84, 85], and integrative modeling [81, 82] approaches. The use of cross-linking restraints for structure prediction was recently assessed during the 11th round of Critical Assessment of Structure Prediction [86, 87] and various shortcomings—both in experiment and modeling—were identified.

X-ray diffuse scattering experiments can produce information about correlated motions in proteins that is complementary to the information obtained from the more typically analyzed Bragg scattering [88, 89]. Wall and co-workers found good agreement between long molecular dynamics simulations and measured diffuse scattering [89], even in the absence of any fitting. The development of suitable ensemble refinement schemes would bring the models into even better agreement with experiment and would provide a powerful new tool for studying correlated motions of proteins.

Recent work has demonstrated the utility of paramagnetic relaxation enhancement measurements in solid-state NMR [90, 91]. These experiments provide less structural information than traditional protein NMR experiments, but, combined with suitable computational modeling, represent an increasingly viable avenue for structure determination [65, 91].

Finally, recent work has demonstrated the possibility of inferring residue–residue contacts from coevolution analysis of homologous sequences [92–94], commonly referred to as evolutionary couplings. Baker and co-workers were recently able to create models for 614 protein families with unknown structures [95], several of which had folds that are not in the Protein Data Bank. Montelione and co-workers combined evolutionary couplings with sparse NMR data, which provide complementary restraints for modeling, to correctly determine structures for proteins up to 41 kDa [3].

## Conclusion and future perspectives

Physical insight has always been integral to structural biology, but the dual challenges of modeling ensembles and making sense of sparse, ambiguous, and noisy data mean that quantitative physical models will become an increasingly important part of modern structural biology. How can we avoid misinterpreting measurement noise as structural heterogeneity? How can we recognize rare, but important, conformations buried within noisy data? Making progress requires: (1) better experiments, including those that are sensitive for rare conformations [96, 97] or provide distributions [97–99]; (2) improved models of the noise and error inherent in the experimental data; (3) accurate methods to back calculate observables from structural ensembles; (4) accurate physical models that can correctly reproduce the conformational landscapes of biomolecules; and (5) suitable statistical frameworks to interpret all of this information in a coherent fashion. In the future, we anticipate that approaches combining statistical inference, physical modeling, and experiments will allow us to better understand the dynamic and heterogeneous nature of conformational ensembles, even in the presence of noisy, sparse, or ambiguous data, which will be key to addressing important biological questions.

## Important References

Papers of interest have been highlighted as:
* of special interest
** of outstanding interest
** [20] A review on approaches that combine experimental and computational methods to determine structural ensembles of dynamic proteins.

**[19] A concise review on maximum entropy approaches. The authors highlighted three papers which explored an important link between replica-averaged ensemble refinement principle and maximum entropy method.

**[18] An important perspective on the relationship between experimental data and computational techniques, and the role of integrative structural biology.

**[23] A key paper on Bayesian inference, defining the commonly applied inferential structural determination methodology and indicating the importance of developing probabilistic methods for structure determination.

**[38] This paper makes use of maximum entropy methods to develop ensemble-averaged restraints for biasing molecular simulations, noting the success of a physics-based approach compared to other refinement schemes.

*[52] This paper demonstrates the statistical equivalence of principle of restrained-ensemble simulations and the the maximum entropy approach.

*[53] This paper justifies the use of the maximum entropy approach to define experimental data-driven restraints for simulations and demonstrates its equivalence to replica-ensemble simulations.

*[14] This paper introduces a Bayesian inference method to account for different sources of error in experimental data in modeling structural ensembles of complex macromolecular systems.

*[15] This paper introduces the new and developing Integrative Modeling Platform (IMP) software package. The authors highlight its flexible capability to incorporate a variety of experimental data, and to generate and develop new models and representations.

*[13] This paper describes Modeling Employing Limited Data (MELD), highlighting its unique Bayesian methodology for determining protein structure, and demonstrating its ability to incorporate a variety of experimental data.

*[70] This paper introduces the RASREC Rosetta approach, describing its improvements over regular CS-Rosetta in detail, and exhibiting its capability to develop models closer to the native structure.

## Acknowledgements

[1] Shahidul M. Islam, Richard A. Stein, Hassane S. Mchaourab, and Benoît Roux. Structural refinement from restrained-ensemble simulations based on EPR/DEER data: Application to T4 lysozyme. *The Journal of Physical Chemistry B*, 117(17):4740–4754, 2013.

[2] Carl Öster, Simone Kosol, Christoph Hartlmüller, Jonathan M. Lamley, Dinu Iuga, Andres Oss, Mai-Liis Org, Kalju Vanatalu, Ago Samoson, Tobias Madl, and Józef R. Lewandowski. Characterization of protein–protein interfaces in large complexes by solid-state NMR solvent paramagnetic relaxation enhancements. *Journal of the American Chemical Society*, 139(35):12165–12174, 2017.

[3] Yuefeng Tang, Yuanpeng J. Huang, Thomas A. Hopf, Chris Sander, Debora S. Marks, and Gaetano T. Montelione. Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nature Methods*, 12(8):751–754, 2015.

[4] Shahid Mehmood, Timothy M. Allison, and Carol V. Robinson. Mass spectrometry of protein complexes: From origins to applications. *Annual Review of Physical Chemistry*, 66(1):453–474, 2015.

[5] Alexey G. Kikhney and Dmitri I. Svergun. A practical guide to small angle x-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Letters*, 589(19):2570–2577, 2015.

[6] Jing Huang, Sarah Rauscher, Grzegorz Nawrocki, Ting Ran, Michael Feig, Bert L. de Groot, Helmut Grubmüller, and Alexander D. MacKerell. Charmm36m: An improved force field for folded and intrinsically disordered proteins. *Nature Methods*, 14(1):71–73, 2017.

[7] James A. Maier, Carmenza Martinez, Koushik Kasavajhala, Lauren Wickstrom, Kevin E. Hauser, and Carlos Simmerling. ff14SB: Improving the accuracy of protein side chain and backbone parameters from ff99SB. *Journal of Chemical Theory and Computation*, 11(8):3696–3713, 2015.

[8] Rebecca F. Alford, Andrew Leaver-Fay, Jeliazko R. Jeliazkov, Matthew J. O'Meara, Frank P. DiMaio, Hahnbeom Park, Maxim V. Shapovalov, P. Douglas Renfrew, Vikram K. Mulligan, Kalli Kappel, Jason W. Labonte, Michael S. Pacella, Richard Bonneau, Philip Bradley, Roland L. Dunbrack, Rhiju Das, David Baker, Brian Kuhlman, Tanja Kortemme, and Jeffrey J. Gray. The Rosetta all-atom energy function for macromolecular modeling and design. *Journal of Chemical Theory and Computation*, 13(6):3031–3048, 2017.

[9] Massimiliano Bonomi, Carlo Camilloni, and Michele Vendruscolo. Metadynamic metainference: Enhanced sampling of the metainference ensemble using metadynamics. *Scientific Reports*, 6:31232, 2016.

[10] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: A smoothly converging and tunable free-energy method. *Physical Review Letters*, 100(2):20603, 2008.

[11] John D. Chodera and Frank Noé. Markov state models of biomolecular conformational dynamics. *Current Opinion in Structural Biology*, 25:135–144, 2014.

[12] Yinglong Miao, Victoria A. Feher, and J. Andrew McCammon. Gaussian accelerated molecular dynamics: Unconstrained enhanced sampling and free energy calculation. *Journal of Chemical Theory and Computation*, 11(8):3584–3595, 2015.

[13] Justin L. MacCallum, Alberto Perez, and Ken A. Dill. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proceedings of the National Academy of Sciences*, 112(22):6985–6990, 2015.

[14] Massimiliano Bonomi, Carlo Camilloni, Andrea Cavalli, and Michele Vendruscolo. Metainference: A Bayesian inference method for heterogeneous systems. *Science Advances*, 2(1):e1501177, 2016.

[15] Daniel Russel, Keren Lasker, Ben Webb, Javier Velázquez-Muriel, Elina Tjioe, Dina Schneidman-Duhovny, Bret Peterson, and Andrej Sali. Putting the pieces together: Integrative modeling platform software for structure determination of macromolecular assemblies. *PLOS Biology*, 10(1):e1001244, 2012.

[16] Simon Olsson, Hao Wu, Fabian Paul, Cecilia Clementi, and Frank Noé. Combining experimental and simulation data of molecular processes via augmented markov models. *Proceedings of the National Academy of Sciences*, 114(31):8265–8270, 2017.

[17] Oliver F. Lange and David Baker. Resolution-adapted recombination of structural features significantly improves sampling in restraint-guided structure calculation. *Proteins: Structure, Function, and Bioinformatics*, 80(3):884–895, 2012.

[18] Andrew B. Ward, Andrej Sali, and Ian A. Wilson. Integrative structural biology. *Science*, 339(6122):913–915, 2013.

[19] Wouter Boomsma, Jesper Ferkinghoff-Borg, and Kresten Lindorff-Larsen. Combining experiments and simulations using the maximum entropy principle. *PLOS Computational Biology*, 10(2):e1003406, 2014.

[20] Massimiliano Bonomi, Gabriella T. Heller, Carlo Camilloni, and Michele Vendruscolo. Principles of protein structural ensemble determination. *Current Opinion in Structural Biology*, 42:106–116, 2017.

[21] Alexander Wlodawer, Wladek Minor, Zbigniew Dauter, and Mariusz Jaskolski. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS Journal*, 275(1):1–21, 2008.

[22] Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical Physics Letters*, 314(1):141–151, 1999.

[23] Wolfgang Rieping, Michael Habeck, and Michael Nilges. Inferential structure determination. *Science*, 309(5732):303–306, 2005.

[24] Wolfgang Rieping, Michael Habeck, and Michael Nilges. *Probabilistic structure calculation*, pages 81–98. NATO Security through Science Series. Springer, Berlin, Germany, 2007.

[25] Dale E. Tronrud. Knowledge-based B-factor restraints for the refinement of proteins. *Journal of Applied Crystallography*, 29(2):100–104, 1996.

[26] Rahel A. Woldeyes, David A. Sivak, and James S. Fraser. E pluribus unum, no more: From one crystal, many conformations. *Current Opinion in Structural Biology*, 28:56–62, 2014.

[27] Elena J. Levin, Dmitry A. Kondrashov, Gary E. Wesenberg, and George N. Phillips Jr. Ensemble refinement of protein crystal structures: Validation and application. *Structure*, 15(9):1040–1052, 2007.

[28] Bernadó Pau, Efstratios Mylonas, Maxim V. Petoukhov, Martin Blackledge, and Dmitri I. Svergun. Structural characterization of flexible proteins using small-angle x-ray scattering. *Journal of the American Chemical Society*, 129(17):5656–5664, 2007.

[29] Yiwen Chen, Sharon L. Campbell, and Nikolay V. Dokholyan. Deciphering protein dynamics from NMR data using explicit structure sampling and selection. *Biophysical Journal*, 93(7):2300–2306, 2007.

[30] Martin Pelikan, Greg L. Hura, and Michal Hammel. Structure and flexibility within proteins as identified through small angle x-ray scattering. *General Physiology and Biophysics*, 28(2):174–189, 2009.

[31] Gabrielle Nodet, Loïc Salmon, Valéry Ozenne, Sebastian Meier, Malene Ringkjøbing Jensen, and Martin Blackledge. Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from NMR residual dipolar couplings. *Journal of the American Chemical Society*, 131(49):17908–17918, 2009.

[32] Sichun Yang, Lydia Blachowicz, Lee Makowski, and Benoît Roux. Multidomain assembled states of hck tyrosine kinase in solution. *Proceedings of the National Academy of Sciences*, 107(36):15757–15762, 2010.

[33] Charles K. Fisher, Austin Huang, and Collin M. Stultz. Modeling intrinsically disordered proteins with bayesian statistics. *Journal of the American Chemical Society*, 132(42):14919–14927, 2010.

[34] Ivano Bertini, Andrea Giachetti, Claudio Luchinat, Giacomo Parigi, Maxim V. Petoukhov, Roberta Pierattelli, Enrico Ravera, and Dmitri I. Svergun. Conformational space of flexible biological macromolecules from average data. *Journal of the American Chemical Society*, 132(38):13553–13558, 2010.

[35] Konstantin Berlin, Carlos A. Castañeda, Dina Schneidman-Duhovny, Andrej Sali, Alfredo Nava-Tudela, and David Fushman. Recovering a representative conformational ensemble from underdetermined macromolecular structural data. *Journal of*

*the American Chemical Society*, 135(44):16595–16609, 2013.

[36] Edwin T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620, 1957.

[37] Steve Pressé, Kingshuk Ghosh, Julian Lee, and Ken A. Dill. Principles of maximum entropy and maximum caliber in statistical physics. *Reviews of Modern Physics*, 85(3):1115–1141, 2013.

[38] Jed W. Pitera and John D. Chodera. On the use of experimental observations to bias simulated ensembles. *Journal of Chemical Theory and Computation*, 8(10):3445–3451, 2012.

[39] Bartosz Różycki, Young C. Kim, and Gerhard Hummer. SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. *Structure*, 19(1):109–116, 2011.

[40] Gerhard Hummer and Jürgen Köfinger. Bayesian ensemble refinement by replica simulations and reweighting. *Journal of Chemical Physics*, 143(24):12B634_1, 2015.

[41] Simon Olsson, Jes Frellsen, Wouter Boomsma, Kanti V. Mardia, and Thomas Hamelryck. Inference of structure ensembles of flexible biomolecules from sparse, averaged data. *PLOS ONE*, 8(11):e79439, 2013.

[42] Simon Olsson, Dariusz Ekonomiuk, Jacopo Sgrignani, and Andrea Cavalli. Molecular dynamics of biomolecules through direct analysis of dipolar couplings. *Journal of the American Chemical Society*, 137(19):6270–6278, 2015.

[43] Michael Habeck. Bayesian approach to inverse statistical mechanics. *Physical Review E*, 89(5):052113, 2014.

[44] Fabrizio Marinelli and José D. Faraldo-Gómez. Ensemble-biased metadynamics: A molecular simulation method to sample experimental distributions. *Biophysical Journal*, 108(12):2779–2782, 2015.

[45] Andrea Cesari, Alejandro Gil-Ley, and Giovanni Bussi. Combining simulations and solution experiments as a paradigm for RNA force field refinement. *Journal of Chemical Theory and Computation*, 12(12):6192–6200, 2016.

[46] Robert B. Best and Michele Vendruscolo. Determination of protein structures consistent with NMR order parameters. *Journal of the American Chemical Society*, 126(26):8090–8091, 2004.

[47] Kresten Lindorff-Larsen, Sigridur Kristjansdottir, Kaare Teilum, Wolfgang Fieber, Christopher M. Dobson, Flemming M. Poulsen, and Michele Vendruscolo. Determination of an ensemble of structures representing the denatured state of the bovine acyl-coenzyme a binding protein. *Journal of the American Chemical Society*, 126(10):3291–3299, 2004.

[48] Kresten Lindorff-Larsen, Robert B. Best, Mark A. DePristo, Christopher M. Dobson, and Michele Vendruscolo. Simultaneous determination of protein structure and dynamics. *Nature*, 433(7022):128, 2005.

[49] Matthew M. Dedmon, Kresten Lindorff-Larsen, John Christodoulou, Michele Vendruscolo, and Christopher M. Dobson. Mapping long-range interactions in $\alpha$-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *Journal of the American Chemical Society*, 127(2):476–477, 2005.

[50] Michele Vendruscolo. Determination of conformationally heterogeneous states of proteins. *Current Opinion in Structural Biology*, 17(1):15–20, 2007.

[51] Sunhwan Jo and Wonpil Im. Transmembrane helix orientation and dynamics: Insights from ensemble dynamics with solid-state NMR observables. *Biophysical Journal*, 100(12):2913–2921, 2011.

[52] Benoît Roux and Jonathan Weare. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *Journal of Chemical Physics*, 138(8):02B616, 2013.

[53] Andrea Cavalli, Carlo Camilloni, and Michele Vendruscolo. Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *Journal of Chemical Physics*, 138(9):03B603, 2013.

[54] Stephanie J. Hirst, Nathan Alexander, Hassane S. Mchaourab, and Jens Meiler. RosettaEPR: An integrated tool for protein structure determination from sparse EPR data. *Journal of Structural Biology*, 173(3):506–514, 2011.

[55] David E. Kim, Frank DiMaio, Ray Yu-Ruei Wang, Yifan Song, and David Baker. One contact for every twelve residues allows robust and accurate topology-level protein structure modeling. *Proteins*, 82(0):208–218, 2014.

[56] Thomas Gurry, Orly Ullman, Charles K. Fisher, Iva Perovic, Thomas Pochapsky, and Collin M. Stultz. The dynamic structure of $\alpha$-synuclein multimers. *Journal of the American Chemical Society*, 135(10):3865–3872, 2013.

[57] Biswaranjan Mohanty, Martin L. Williams, Bradley C. Doak, Mansha Vazirani, Olga Ilyichova, Geqing Wang, Wolfgang Bermel, Jamie S. Simpson, David K. Chalmers, Glenn F. King, Mehdi Mobli, and Martin J. Scanlon. Determination of ligand binding modes in weak protein-ligand complexes using sparse NMR data. *Journal of Biomolecular NMR*, 66(3):195–208, 2016.

[58] Cyril Dominguez, Rolf Boelens, and Alexandre M. J. J. Bonvin. HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125(7):1731–1737, 2003.

[59] Gydo C. P. van Zundert, Adrien S. J. Melquiond, and Alexandre M. J. J. Bonvin. Integrative modeling of biomolecular complexes: HADDOCKing with cryo-electron microscopy data. *Structure*, 23(5):949–960, 2015.

[60] Dipen M. Shah, Eiso Ab, Tammo Diercks, Mathias A. S. Hass, Nico A. J. van Nuland, and Gregg Siegal. Rapid protein–ligand costructures from sparse NOE data. *Journal of Medicinal Chemistry*, 55(23):10786–10790, 2012.

[61] Massimiliano Bonomi, Riccardo Pellarin, Seung Joong Kim, Daniel Russel, Bryan A. Sundin, Michael Riffle, Daniel Jaschob, Richard Ramsden, Trisha N. Davis, Eric G. D. Muller, and Andrej Sali. Determining protein complex structures based on a Bayesian model of in vivo Förster resonance energy transfer (FRET) data. *Molecular & Cellular Proteomics*, 13(11):2812–2823, 2014.

[62] Alex Zelter, Massimiliano Bonomi, Jae ook Kim, Neil T. Umbreit, Michael R. Hoopmann, Richard Johnson, Michael Riffle, Daniel Jaschob, Michael J. MacCoss, Robert L. Moritz, and Trisha N. Davis. The molecular architecture of the Dam1 kinetochore complex is defined by cross-linking based structural modelling. *Nature Communications*, 6:8673, 2015.

[63] Binchen Mao, Roberto Tejero, David Baker, and Gaetano T. Montelione. Protein NMR structures refined with Rosetta have higher accuracy relative to corresponding x-ray crystal structures. *Journal of the American Chemical Society*, 136(5):1893–1906, 2014.

[64] James M. Thompson, Nikolaos G. Sgourakis, Gaohua Liu, Paolo Rossi, Yuefeng Tang, Jeffrey L. Mills, Thomas Szyperski, Gaetano T. Montelione, and David Baker. Accurate protein structure modeling using sparse NMR data and homologous structure information. *Proceedings of the National Academy of Sciences*, 109(25):9875–9880, 2012.

[65] Hajime Tamaki, Ayako Egawa, Kouki Kido, Tomoshi Kameda, Masakatsu Kamiya, Takashi Kikukawa, Tomoyasu Aizawa, Toshimichi Fujiwara, and Makoto Demura. Structure determination of uniformly 13C, 15N labeled protein using qualitative distance restraints from MAS solid-state 13C-NMR observed paramagnetic relaxation enhancement. *Journal of Biomolecular NMR*, 64(1):87–101, 2016.

[66] Christophe Schmitz, Robert Vernon, Gottfried Otting, David Baker, and Thomas Huber. Protein structure determination from pseudocontact shifts using ROSETTA. *Journal of Molecular Biology*, 416(5):668–677, 2012.

[67] Lisa R. Warner, Krisztina Varga, Oliver F. Lange, Susan L. Baker, David Baker, Marcelo C. Sousa, and Arthur Pardi. Structure of the BamC two-domain protein obtained by Rosetta with a limited NMR data set. *Journal of Molecular Biology*, 411(1):83–95, 2011.

[68] Katrin Reichel, Olivier Fisette, Tatjana Braun, Oliver F. Lange, Gerhard Hummer, and Lars V. Schäfer. Systematic evalua-

tion of CS-Rosetta for membrane protein structure prediction with sparse NOE restraints. *Proteins: Structure, Function, and Bioinformatics*, 85(5):812–826, 2017.

[69] Gijs van der Schot, Zaiyong Zhang, Robert Vernon, Yang Shen, Wim F. Vranken, David Baker, Alexandre M. J. J. Bonvin, and Oliver F. Lange. Improving 3D structure prediction from chemical shift data. *Journal of Biomolecular NMR*, 57(1):27–35, 2013.

[70] Oliver F. Lange, Paolo Rossi, Nikolaos G. Sgourakis, Yifan Song, Hsiau-Wei Lee, James M. Aramini, Asli Ertekin, Rong Xiao, Thomas B. Acton, Gaetano T. Montelione, et al. Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. *Proceedings of the National Academy of Sciences*, 109(27):10873–10878, 2012.

[71] Thomas Löhr, Alexander Jussupow, and Carlo Camilloni. Metadynamic metainference: Convergence towards force field independent structural ensembles of a disordered peptide. *Journal of Chemical Physics*, 146(16):165102, 2017.

[72] Gabriella T. Heller, Francesco A. Aprile, Massimiliano Bonomi, Carlo Camilloni, Alfonso De Simone, and Michele Vendruscolo. Sequence specificity in the entropy-driven binding of a small molecule and a disordered peptide. *Journal of Molecular Biology*, 429(18):2772–2779, 2017.

[73] Alberto Perez, Justin L. MacCallum, and Ken A. Dill. Accelerating molecular simulations of proteins using Bayesian inference on weak information. *Proceedings of the National Academy of Sciences*, 112(38):11846–11851, 2015.

[74] Alberto Perez, Joseph A. Morrone, Emiliano Brini, Justin L. MacCallum, and Ken A. Dill. Blind protein structure prediction using accelerated free-energy simulations. *Science Advances*, 2(11):e1601274, 2016.

[75] Joseph A. Morrone, Alberto Perez, Qiaolin Deng, Sookhee N. Ha, M. Katharine Holloway, Tomi K. Sawyer, Bradley S. Sherborne, Frank K. Brown, and Ken A. Dill. Molecular simulations identify binding poses and approximate affinities of stapled $\alpha$-helical peptides to MDM2 and MDMX. *Journal of Chemical Theory and Computation*, 13(2):863–869, 2017.

[76] Joseph A. Morrone, Alberto Perez, Justin MacCallum, and Ken A. Dill. Computed binding of peptides to proteins with MELD-accelerated molecular dynamics. *Journal of Chemical Theory and Computation*, 13(2):870–876, 2017.

[77] Francesco Lanucara, Stephen W. Holman, Christopher J. Gray, and Claire E. Eyers. The power of ion mobility-mass spectrometry for structural characterization and the study of conformational dynamics. *Nature Chemistry*, 6(4):281, 2014.

[78] Lutz Fischer, Zhuo A. Chen, and Juri Rappsilber. Quantitative cross-linking/mass spectrometry using isotope-labelled cross-linkers. *Journal of Proteomics*, 88(C):120–128, 2013.

[79] Zhuo A. Chen, Lutz Fischer, Jürgen Cox, and Juri Rappsilber. Quantitative cross-linking/mass spectrometry using isotope-labeled cross-linkers and MaxQuant. *Molecular & Cellular Proteomics*, 15(8):2769–2778, 2016.

[80] Daniel S. Ziemianowicz, Ryan Bomgarden, Chris Etienne, and David C. Schriemer. Amino acid insertion frequencies arising from photoproducts generated using aliphatic diazirines. *Journal of the American Society for Mass Spectrometry*, 28(10):2011–2021, 2017.

[81] Martial Rey, Vladimir Sarpe, Kyle M. Burns, Joshua Buse, Charles A. H. Baker, Marc van Dijk, Linda Wordeman, Alexandre M. J. J. Bonvin, and David C. Schriemer. Mass spec studio for integrative structural biology. *Structure*, 22(10):1538–1548, 2014.

[82] Vladimir Sarpe, Atefeh Rafiei, Morgan Hepburn, Nicholas Ostan, Anthony B. Schryvers, and David C. Schriemer. High sensitivity crosslink detection coupled with integrative structure modeling in the mass spec studio. *Molecular & Cellular Proteomics*, 15(9):3071–3080, 2016.

[83] Argyris Politis, Florian Stengel, Zoe Hall, Helena Hernández, Alexander Leitner, Thomas Walzthoeni, Carol V. Robinson, and Ruedi Aebersold. A mass spectrometry-based hybrid

method for structural modeling of protein complexes. *Nature Methods*, 11(4):403, 2014.

[84] Eric D. Merkley, Steven Rysavy, Abdullah Kahraman, Ryan P. Hafen, Valerie Daggett, and Joshua N. Adkins. Distance restraints from crosslinking mass spectrometry: Mining a molecular dynamics simulation database to evaluate lysine–lysine distances. *Protein Science*, 23(6):747–759, 2014.

[85] Nicholas I. Brodie, Konstantin I. Popov, Evgeniy V. Petrotchenko, Nikolay V. Dokholyan, and Christoph H. Borchers. Solving protein structures using short-distance cross-linking constraints as a guide for discrete molecular dynamics simulations. *Science Advances*, 3(7):e1700479, 2017.

[86] Adam Belsom, Michael Schneider, Oliver Brock, and Juri Rappsilber. Blind evaluation of hybrid protein structure analysis methods based on cross-linking. *Trends in Biochemical Sciences*, 41(7):564–567, 2016.

[87] Michael Schneider, Adam Belsom, Juri Rappsilber, and Oliver Brock. Blind testing of cross-linking/mass spectrometry hybrid methods in CASP11. *Proteins: Structure, Function, and Bioinformatics*, 84:152–163, 2016.

[88] Andrew H. Van Benschoten, Lin Liu, Ana Gonzalez, Aaron S. Brewster, Nicholas K. Sauter, James S. Fraser, and Michael E. Wall. Measuring and modeling diffuse scattering in protein x-ray crystallography. *Proceedings of the National Academy of Sciences*, 113(15):4069–4074, 2016.

[89] Michael E. Wall, Andrew H. Van Benschoten, Nicholas K. Sauter, Paul D. Adams, James S. Fraser, and Thomas C. Terwilliger. Conformational dynamics of a crystalline protein from microsecond-scale molecular dynamics simulations and diffuse x-ray scattering. *Proceedings of the National Academy of Sciences*, 111(50):17887–17892, 2014.

[90] Ishita Sengupta, Philippe S. Nadaud, and Christopher P. Jaroniec. Protein structure determination with paramagnetic solid-state NMR spectroscopy. *Accounts of Chemical Research*, 46(9):2117–2126, 2013.

[91] Christopher P. Jaroniec. Structural studies of proteins by paramagnetic solid-state NMR spectroscopy. *Journal of Magnetic Resonance*, 253:50–59, 2015.

[92] Debora S. Marks, Lucy J. Colwell, Robert Sheridan, Thomas A. Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. Protein 3D structure computed from evolutionary sequence variation. *PLOS ONE*, 6(12):e28766, 2011.

[93] Timothy Nugent and David T. Jones. Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. *Proceedings of the National Academy of Sciences*, 109(24):E1540–E1547, 2012.

[94] Hetunandan Kamisetty, Sergey Ovchinnikov, and David Baker. Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. *Proceedings of the National Academy of Sciences*, 110(39):15674–15679, 2013.

[95] Sergey Ovchinnikov, Hahnbeom Park, Neha Varghese, Po-Ssu Huang, Georgios A. Pavlopoulos, David E. Kim, Hetunandan Kamisetty, Nikos C. Kyrpides, and David Baker. Protein structure determination using metagenome sequence data. *Science*, 355(6322):294–298, 2017.

[96] Lewis E. Kay. New views of functionally dynamic proteins by solution NMR spectroscopy. *Journal of Molecular Biology*, 428(Part A):323–331, 2016.

[97] Haim Yuval Aviram, Menahem Pirchi, Yoav Barak, Inbal Riven, and Gilad Haran. Two states or not two states: Single-molecule folding studies of protein L. *The Journal of Chemical Physics*, 148(12):123303–7, 2018.

[98] Eitan Lerner, Thorben Cordes, Antonino Ingargiola, Yazan Alhadid, SangYoon Chung, Xavier Michalet, and Shimon Weiss. Toward dynamic structural biology: Two decades of single-molecule Förster resonance energy transfer. *Science (New York, NY)*, 359(6373):eaan1133–14, 2018.

[99] Ashok A. Deniz. Deciphering complexity in molecular biophysics with single-molecule resolution. *Journal of Molecular*

9

*Biology*, 428(Part A):301–307, 2016.