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

Heuristics based on physical insight have always been an important part of structure determination. 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, but more rigorous quantitative physical models are increasingly used to transform experimental data into structures and ensembles. These 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 in recent years, driven by improvements in computer power, improved physical models of protein structure [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]. Here, 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 can play a key role.

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 this single structure viewpoint has dominated in structural biology, there is increasing recognition of the importance of heterogeneity and dynamics, enabled by significant improvements in experimental techniques and computational capability.

Nearly all 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.

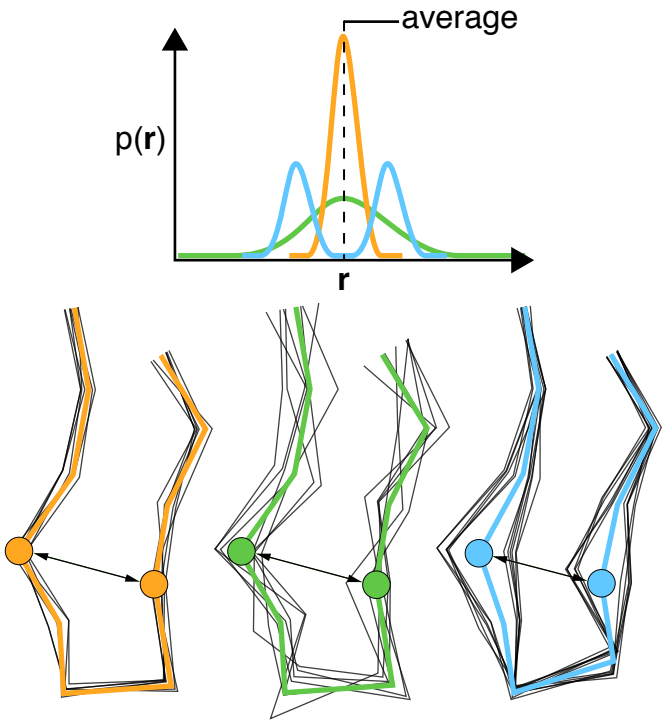

Figure 1: Most experiments measure ensemble averages, which poses a challenge as systems become more flexible, heterogeneous, and dynamic. This figure illustrates a thought experiment, comparing three different ensembles with the same average for some observable, but different conformational ensembles.

Challenge 2: Making sense of Sparse, Ambiguous, and Noisy Data

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

Figure 2 shows several common pathologies. First, the data may be sparse, often only providing information about a few degrees of freedom, e.g. an EPR experiment might measure a single distance between probes. 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 overfitting. 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.

Overcoming the dual challenges of modeling ensembles and making sense of sparse, ambiguous, and noisy data requires a synergistic combination of experiment, statistical inference, and physical modeling.

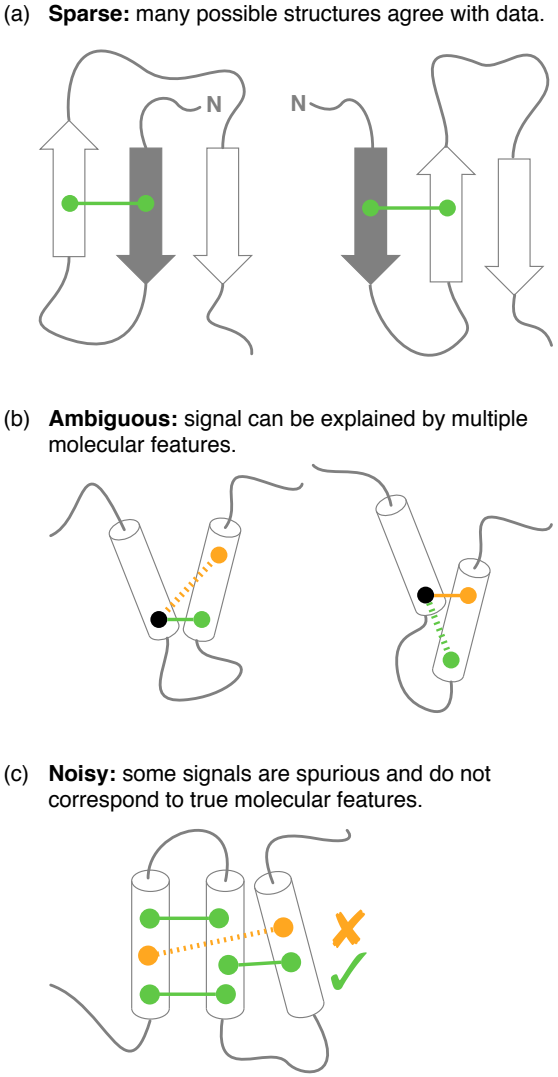

Figure 2: Conceptual illustration of the challenges faced in integrative structural biology and other applications where the data is sparse, ambiguous, and noisy.

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. The former have always been an integral part of biomolecular structure determination, while the latter are becoming increasingly important in modern structural biology.

Heuristic 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 widely used 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.

The aim of integrative structural biology is to construct a structural model of a biomolecular system from one or more experimental datasets, which is a problem of statistical inference that can be approached from a variety of perspectives, including maximum likelihood, maximum entropy, maximum parsimony, and Bayesian approaches.

The likelihood, $\mathcal{L}(\theta|D) \sim \mathbb{P}(D|\theta)$, is central to many methods, where D is the observed data and θ is a set of parameters specifying the structural ensemble, e.g. atomic coordinates and B-factors. This probabilistic relationship encapsulates the experimental measurement and relates the model to experimental observables. The likelihood function is often evaluated on single structures. However, newer ensemble refinement methods [23–25] use likelihood functionals to evaluate distributions of structures, which, as described later, is more suitable for conformationally heterogeneous ensembles.

Maximum likelihood (ML) methods seek to find the single set of parameter values with maximum likelihood. Näive ML methods rely entirely on the data, making these methods sensitive to noise and notoriously prone to overfitting. To mitigate this, ML methods are often augmented by *ad hoc* penalty terms motivated by physical considerations, e.g. the use of restraints on crystallographic B-factors which ensure that variations in flexibility between nearby atoms are physically plausible [26]. However, even after augmentation with penalty terms, ML methods are still prone to over-fitting as the data to parameter ratio becomes increasingly poor.

In contrast to ML, maximum entropy (MaxEnt) methods seek to find a *distribution* of parameters, $p(\theta|D)$, to explain the observed data. Although there are many possible distributions that could match the observed data, there is a unique maximum entropy distribution [27, 28], providing a powerful basis for statistical inference. An ensemble generated using Eq. 1 alone may not agree with experiment. MaxEnt methods seek to minimally perturb (in a well-defined MaxEnt sense) this ensemble, either through biasing [23–25] or reweighting [29, 30], to bring the results into agreement with experimental measurements.

Maximum parsimony methods [20, 31] have many similarities with MaxEnt approaches. A key distinction is that maximum parsimony aims for *simple* models, e.g. describing an ensemble with a minimal number of representative conformations.

The Bayesian approach offers a different perspective [32] that partially encompasses both MaxEnt and maximum parsimony methods. Bayes theorem is a simple and elegant statement,

$$p(\theta|D) \propto \mathcal{L}(D|\theta)p(\theta), \tag{2}$$

which 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 function, $\mathcal{L}(D|\theta)$, with the prior, $p(\theta)$.

Bayesian methods differ from ML in several key respects. First, the prior, often given by Eq. 1, represents our knowledge of protein structures in the absence of data. The prior, rather than *ad hoc* penalty terms, provides a means to make sense of otherwise sparse, ambiguous, or noisy data. Second, Bayesian methods generate an ensemble from the posterior distribution, rather than a single sample, as in ML. The assumption of maximum entropy [27, 28] underlying Eq. 1 leads to ensembles that are as broad as possible given both the data and energetic considerations from the prior, which mitigates over-fitting. Finally, the prior may include "nuisance parameters", like the level of noise corrupting a particular observable. During sampling, these parameters are jointly inferred with the other parameters describing the model, leading to a statistically consistent ensemble without the need to specify the exact values of nuisance parameters.

The lines between these different approaches are often blurred, and many methods do not clearly fall into any of the categories. These are often more *ad hoc* combinations of physically-motivated scoring functions and sampling strategies that do not produce a well-defined ensemble. However, although these methods have less rigorous statistical underpinning, they are often quite successful.

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

In statistical mechanics, an ensemble has a specific technical meaning: the probability distribution over all possible configurations of a 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 key characteristics of these pseudo-ensembles that must be considered. Does the likelihood consider only individual structures, or properties of the distribution as a whole? Do the structures sampled come from a well-defined distribution, e.g. a Boltzmann distribution, or are they simply a set of low-energy conformations, e.g. as in traditional NMR refinement? How are experimental errors handled? What priors are used? What is sampled over?—is it just atomic coordinates, or are there other parameters like error magnitudes? It is only through consideration of these questions that the correct interpretation of the "ensemble" can be arrived at.

Maximum entropy and related methods can be robust against over-fitting.

Maximum likelihood methods become prone to overfitting as the data to parameter ratio becomes poor. For example, it is uncommon to see multi-copy refinement of X-ray crystal structures, where heterogeneity is represented using multiple copies of the system [33], as the data to parameter ratio decreases linearly with number of copies. Phillips and co-workers undertook a systematic study of 50 experimental structures, and found that adding up to, on average, $\sim$ 10 copies yielded improved models [34]. However, ensembles from maximum entropy or Bayesian methods can easily have thousands of models. How are these models not grossly over-fit?

The key to understanding this apparent paradox is to realize that the atomic coordinates *are not* free parameters in maximum entropy and related methods. Consider a simple maximum entropy reweighting procedure [30]. First, an unbiased ensemble is generated using Eq. 1, say with 1000 conformers, giving $1000 \times 3 \times N_{\text{atoms}}$ coordinates. But these coordinates are now fixed, and instead the weights for each conformation, 1000 in total, are used to bring the computed averages into accordance with experimental observations. However, even these 1000 weights are not free parameters, as the maximum entropy principle prescribes a particular set of weights that simultaneously maximize entropy and bring compute average quantities into agreement with their experimentally observed counterparts. In practice, there is one Lagrange multiplier to be determined for each experimental observation, so the data to parameter ratio is essentially one-to-one, regardless of the number of conformers in the ensemble. Similar ideas apply to the ensemble refinement schemes discussed in the next section.

Physical modeling offers solutions to several key challenges in structural biology.

Challenge 1: Modeling Conformational Ensembles

The form of the likelihood function is of critical importance in ensemble refinement. If the likelihood function considers only single structures, there is little hope of reproducing the correct ensemble, as the likelihood function "cannot see the big picture". Single structure-based likelihoods have the effect of forcing all structures to satisfy the average data, rather than reflecting the true distribution (blue vs orange systems in Figure 1). However, in many cases an "ensemble" of structures is still produced. For example, Bayesian single copy refinement [30] will produce an ensemble of structures, but the resulting heterogeneity arises from the non-zero temperature and sampling over nuisance parameters, rather than necessarily reflecting the true underlying ensemble.

A variety of replica-based approaches use restraints that couple the behavior of many replicas or copies of the system to the measured averages from experiment, as recently reviewed in [19, 20].

Replica-averaged ensemble approaches simulate several replicas of the system in parallel, which are coupled through a harmonic potential that restrains properties averaged over all replicas to the corresponding experimental quantities [35]. While successful [36–39], these methods lack a formal connection to maximum entropy or Bayesian principles.

Pitera and Chodera [23] derived an expression for the maximum entropy biasing potential to bring calculated averages from a single simulation into agreement with experiment. This formulation is difficult to use in practice, as it requires determining Lagrange multipliers through trial and error. Nevertheless, Pitera and Chodera were able to identify an important link between their maximum entropy formalism and replica-averaged restraints—as the number of replicas and the harmonic force constant both increase, the replica-averaged ensemble approach converges to the correct MaxEnt distribution. This link was made rigorous in several follow up papers [24, 25] and now forms the backbone of a number of approaches. 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 [30].

Ensemble heterogeneity explains much of the difficulty in characterizing intrinsically disordered proteins (IDPs) experimentally, as they are ensembles of inter-converting conformations [40, 41]. 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 [42]. This method has been successful in determining the relative fractions of mutated structures in an ensemble for aggregative proteins [43].

Challenge 2: Making sense of Sparse, Ambiguous, and Noisy Data

Data from some experimental techniques can often be sparse, ambiguous, and noisy, due to inherent limitations of the technique, or the number and difficulty of the experiments that must be performed. Nevertheless, such data can still be highly valuable in inferring the structures of biomolecules and complex. A number of computational methods have been developed over the past decade which can translate such low information data into meaningful structural models.

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 [44, 45]. Among its many applications, HADDOCK has been used to study protein complex interfaces using cryo-EM data [46] and protein ligand complexes using sparse intermolecular NOEs [47].

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]. Among many applications, protein complex structures have been defined with IMP using *in vivo* FRET data through a Bayesian approach [48], and using a combination of cross-linking data with biochemical and EM localization data [49].

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 [50]. Backbone chemical shifts and distance restraints have been used to guide structure determination [51]. Also, paramagnetic relaxation enhancement (PRE) [52], pseudo-contact shift (PCS) [53], and residual dipolar coupling (RDC) [54] restraints have been used to similar effect. Recently, the RASREC (resolution-adapted structural recombination) algorithm was developed, which yields better models with narrower sampling [17, 55]. RASREC enriches the structure pool by re-using structural features that were frequently observed in previous runs. It requires fewer restraints, and develops models that are closer to the native structure, including for NMR on deuterated samples up to 40 kDa [56, 57].

A newer approach based on Bayesian inference, Metainference, 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 [58, 59].

Modeling Employing Limited Data (MELD) is a Bayesian approach that combines statistical mechanics (Eq. 1), 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 [60, 61], and mutagenesis guided peptide-protein docking [62, 63].

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, 64], cross-linking chemistries [65–67], and data analysis [65, 66, 68, 69]. These techniques are extremely sensitive, but the data can be highly ambiguous and both false-positive and false-negative signals are common. Such data has recently been used as restraints to guide Monte Carlo [70], molecular dynamics [71], and integrative modeling [68, 69] approaches. The use of cross-linking restraints for structure prediction was recently assessed during the 11th round of Critical Assessment of Structure Prediction [72, 73] 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 [74, 75]. Wall and co-workers found good agreement between long molecular dynamics simulations and measured diffuse scattering [75], 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 [76, 77]. 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 [52, 77].

Transition metal ion FRET (tmFRET) measures the distance between small-molecule fluorophores and a non-fluorescent transition metal. Because it provides short

range distances, and because different metals have different absorptions, the method is tunable for a range of distances (10–20 Å) [78] and has been used to study membrane proteins [79].

Finally, recent work has demonstrated the possibility of inferring residue–residue contacts from coevolution analysis of homologous sequences [80–82], commonly referred to as evolutionary couplings. Baker and co-workers were recently able to create models for 614 protein families with unknown structures [83], 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. Driven by faster computers, advances in theoretical understanding, and better algorithms, detailed physical modeling is enabling new methods in structural biology, which are essential to addressing exciting 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.

**[32] 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.

**[23] 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.

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

*[25] This paper justifies the use of the maximum entropy approach to define experimental data-driven restraints for 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.

*[57] 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

This work is supported by funding from the Natural Sciences and Engineering Research Council of Canada. JLM is a Tier 2 Canada Research Chair.

[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 Hartlmller, Jonathan M. Lamley, Dinu Iuga, Andres Oss, Mai-Liis Org, Kalju Vanatalu, Ago Samoson, Tobias Madl, and Jzef 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. OMeara, 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, G 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] 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.

[24] 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.

[25] 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.

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

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

[28] 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.

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

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

[31] 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.

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

[33] 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.

[34] Elena J Levin, Dmitry A Kondrashov, Gary E Wesenberg, and George N Phillips Jr. Ensemble Refinement of Protein Crystal Structures: Validation and Application. *Structure (London, England : 1993)*, 15(9):1040–1052, 2007.

[35] Michele Vendruscolo. Determination of conformationally heterogeneous states of proteins. *Current opinion in structural biology*, 17(1):15–20, 2007.

[36] 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.

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

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

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

[44] 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 proteinligand complexes using sparse NMR data. *Journal of Biomolecular NMR*, 66(3):195–208, 2016.

[45] 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.

[46] Gydo C. P. vanZundert, 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] Katrin Reichel, Olivier Fisette, Tatjana Braun, Oliver F. Lange, Gerhard Hummer, and Lars V. Schäfer. Systematic evaluation of CS-rosetta for membrane protein structure prediction with sparse NOE restraints. *Proteins: Structure, Function, and Bioinformatics*, 85(5):812–826, 2017.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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 α-helical peptides to MDM2 and MDMX. *Journal of Chemical Theory and Computation*, 13(2):863–869, 2017.

[63] 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.

[64] 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.

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

[66] 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.

[67] Daniel S Ziemianowicz, Ryan Bomgarden, Chris Etienne, and David C Schriemer. Amino Acid Insertion Frequencies Arising from Photoproducts Generated Using Aliphatic Diazirines. pages 1–11, 2017.

[68] Martial Rey, Vladimir Sarpe, Kyle M. Burns, Joshua Buse, CharlesA. H. Baker, Marc vanDijk, Linda Wordeman, AlexandreM. J. J. Bonvin, and DavidC. Schriemer. Mass spec studio for integrative structural biology. *Structure*, 22(10):1538–1548, 2014.

[69] 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.

[70] Argyris Politis, Florian Stengel, Zoe Hall, Helena Hernndez, Alexander Leitner, Thomas Walzthoeni, Carol V. Robinson, and Ruedi Aebersold. A mass spectrometrybased hybrid method for structural modeling of protein complexes. *Nature Methods*, 11(4):403, 2014.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

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

[78] William N. Zagotta, Moshe T. Gordon, Eric N. Senning, Mika A. Munari, and Sharona E. Gordon. Measuring distances between TRPV1 and the plasma membrane using a noncanonical amino acid and transition metal ion FRET. *The Journal of General Physiology*, 147(2):201–206, 2016.

[79] Sharona E. Gordon, Eric N. Senning, Teresa K. Aman, and William N. Zagotta. Transition metal ion FRET to measure short-range distances at the intracellular surface of the plasma membrane. *The Journal of General Physiology*, 147(2):189–200, 2016.

[80] 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.

[81] 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 Of The United States Of America*, 2012.

[82] 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.

[83] 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.