Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

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

Kari Gaalswyk[2], Mir Ishruna Muniyat[2] and Justin L MacCallum

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.

Address
Department of Chemistry, University of Calgary, Calgary, AB, Canada

Corresponding author: MacCallum, Justin L (justin.maccallum@ucalgary.ca)
[2] KG and MIM contributed equally to this work.

Current Opinion in Structural Biology 2018, **49**:145-153

This review comes from a themed issue on **Theory and simulation**

Edited by **Kresten Lindorff-Larsen** and **Robert Best**

https://doi.org/10.1016/j.sbi.2018.03.005

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•,14•,15•,16,17] 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••,19••,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.

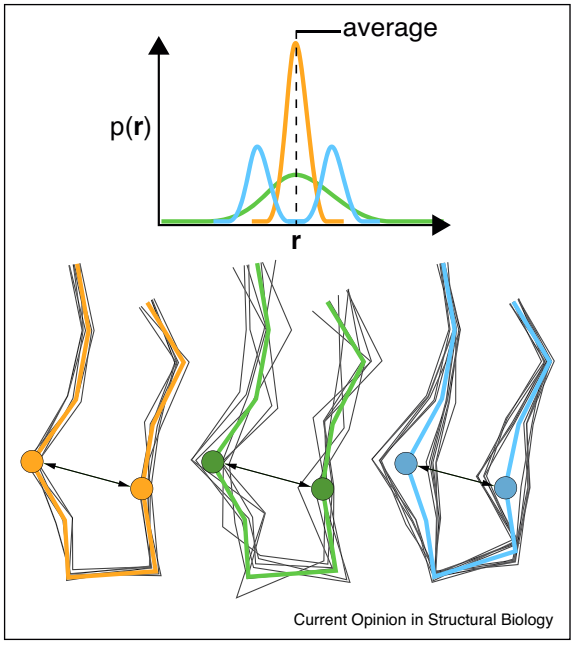
Most measurements in structural biology are ensemble averages, where the observed signal comes from the average across many molecules. The challenge of interpreting such averaged data increases as the conformational ensemble becomes more heterogeneous. A simple thought experiment illustrates the central 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.

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

Figure 1



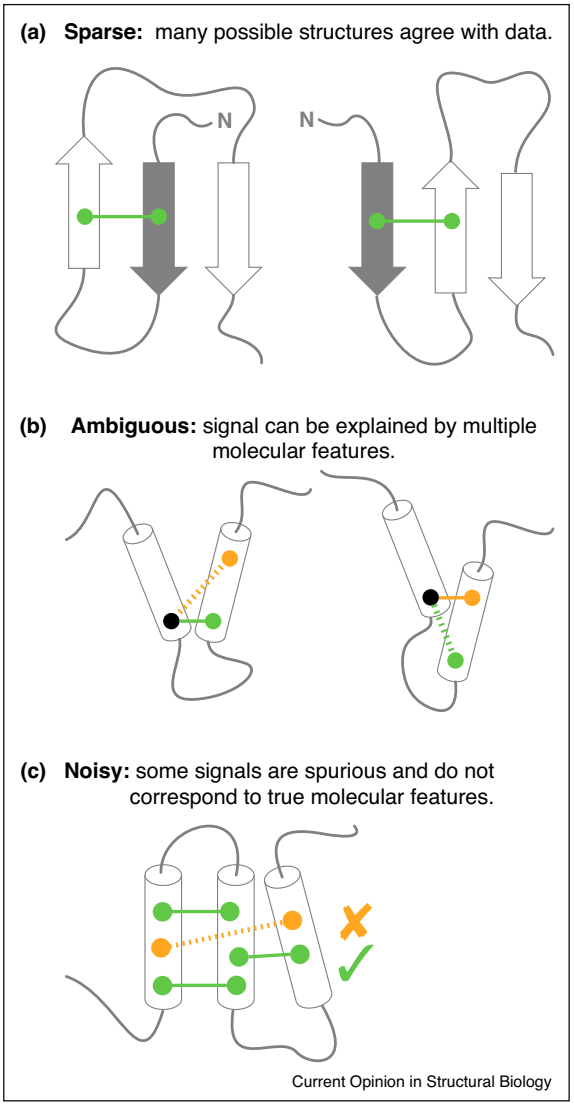
Most experiments measure ensemble averages, which poses a challenge as systems become more flexible, heterogeneous, and dynamic. This figure illustrates a thought experiment, comparing three different ensembles with the same average for some observable, but different conformational ensembles.

degrees of freedom. Second, the data may be ambiguous, where there are multiple molecular features that could explain a particular signal, for example, 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 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.

Figure 2



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

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:

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

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