

Review

Michael Habeck*

Bayesian methods in integrative structure modeling

<https://doi.org/10.1515/hsz-2023-0145>

Received February 27, 2023; accepted July 7, 2023;

published online July 31, 2023

Abstract: There is a growing interest in characterizing the structure and dynamics of large biomolecular assemblies and their interactions within the cellular environment. A diverse array of experimental techniques allows us to study biomolecular systems on a variety of length and time scales. These techniques range from imaging with light, X-rays or electrons, to spectroscopic methods, cross-linking mass spectrometry and functional genomics approaches, and are complemented by AI-assisted protein structure prediction methods. A challenge is to integrate all of these data into a model of the system and its functional dynamics. This review focuses on Bayesian approaches to integrative structure modeling. We sketch the principles of Bayesian inference, highlight recent applications to integrative modeling and conclude with a discussion of current challenges and future perspectives.

Keywords: Bayesian inference; biomolecular structure; integrative modeling; macromolecular assemblies; Markov chain Monte Carlo

1 Introduction

Molecular machines composed of proteins and nucleic acids are implicated in all essential processes of the living cell. To gain a fundamental understanding of how these cellular machines work, we need to transcend phenomenological descriptions and replace them with quantitative models that capture the structure and dynamics of the system, how it responds to external signals, how it assembles and disassembles functional units, and how it is embedded in the

cellular context. Currently no single experimental method is able to span all relevant spatial and temporal scales (Koukos and Bonvin 2020; Rout and Sali 2019; Sali 2021). Rather we need to study macromolecular complexes using an array of experimental techniques including X-ray crystallography, optical and electron 3D imaging, solution and solid-state NMR, cross-linking mass spectrometry (XLMS) and solution scattering (SAXS). In addition, protein structure prediction using machine learning (ML) approaches and massive training data generate complementary information. Programs such as AlphaFold2 (Jumper et al. 2021) and RoseTTAFold (Baek et al. 2021) have invigorated the field of protein structure prediction by achieving unprecedented prediction accuracy matching the quality of experimental structures. EBI's AlphaFold database (Varadi et al. 2022) is an invaluable resource for integrative modeling. Predicted protein models already play an essential part in integrative modeling (see e.g. work by Mosalaganti et al. 2022) and will become even more important in the future.

Integrated approaches to structural cell biology produce diverse datasets that show different aspects of the system on multiple spatial and temporal scales with varying quality and information content. All experimental data must be integrated into a quantitative model that captures structural and dynamical aspects of the system in the cellular environment. Computational modeling approaches are pivotal in achieving this synthesis, because our questions continue to grow in complexity. The 3D structure of a macromolecular complex only marks a point of departure. On the next level, it will become part of a dynamic network in which movable units change conformation and factors bind and dissociate. Ultimately, the model should reproduce the kinetics of the network and incorporate the cellular context.

Although integrative modeling with structural data is often still done manually, it is clear that manual model building risks to be biased and will soon be impractical in the face of the wealth of data that is being generated by emerging experimental techniques.

We should also keep in mind that all biomolecular structures, even when determined at high resolution, are models, and this is especially true for hybrid structures obtained by integrative approaches (Rout and Sali 2019).

*Corresponding author: Michael Habeck, Microscopic Image Analysis Group, Jena University Hospital, D-07743 Jena, Germany; and Max Planck Institute for Multidisciplinary Sciences, D-37077 Göttingen, Germany, E-mail: michael.habeck@uni-jena.de. <https://orcid.org/0000-0002-2188-5667>

Therefore, we need automated methods to not only estimate unknown model parameters, but also to assess parameter uncertainty and the validity of the model.

Bayesian probability theory offers a unique framework to reason about scientific models in the presence of uncertainty (Ghahramani 2015). Therefore Bayesian inference is the appropriate framework to develop computational approaches to integrative structure determination (Rout and Sali 2019). Among the first approaches to biomolecular structure determination with a full-fledged Bayesian inference approach is Inferential Structure Determination (ISD) by Rieping et al. (2005). ISD was originally formulated in the context of NMR structure determination, but the framework applies to all kinds of structural systems and data. The ISD principle is particularly adequate and productive for integrative structure determination (Rout and Sali 2019) and has been implemented in powerful modeling software such as the Integrative Modeling Platform (IMP) (Russel et al. 2012; Saltzberg et al. 2021).

Although Bayesian principles are being embraced by more and more researchers, the use of probabilistic reasoning for integrative structural modeling has by far not been exploited to its fullest potential. This review summarizes the basic principles underlying the Bayesian approach and highlights some recent applications and persistent challenges. Due to limitations in space, this review is by no means exhaustive and does not claim to describe all current developments. It only provides a cursory, at times subjective entry point for further studies. For an in-depth discussion of integrative modeling approaches bridging the atomic and cellular scale, the reader is invited to consult the recent reviews by Braitbard et al. (2019); Koukos and Bonvin (2020); Rout and Sali (2019); Sali (2021); Ziemianowicz and Kosinski (2022).

2 Bayesian integrative structural modeling

Bayesian probability theory is a general framework for reasoning under uncertainty and highly suited for scientific data analysis (Ghahramani 2015; Kinz-Thompson et al. 2021). The hallmark of Bayesian inference is to express all information including the experimental data and relevant background knowledge through probabilities. The application of Bayesian inference to structural modeling typically proceeds in two steps: (1) modeling, i.e. formulating a *probabilistic model* for the observed data, (2) computation, i.e. determining representative model realizations that are probable explanations of the observed data (see Figure 1). The following outlines the

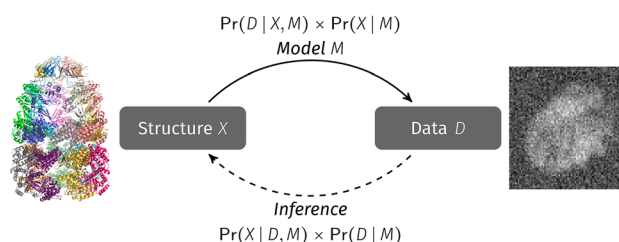


Figure 1: Bayesian integrative modeling consists of two steps. In the modeling step, a probabilistic model M is constructed that generates the observed data (right) from a structural model X (left) and additional parameters (here, a projection direction and a noise model). In the computation step, the model is “inverted” by multiplying the likelihood $\Pr(D | X, M)$ with the prior $\Pr(X | M)$ and drawing conformational samples from the resulting product.

Bayesian approach in the context of biomolecular structure determination (Rieping et al. 2005).

2.1 Probabilistic modeling

Let’s first discuss the modeling step. Here, the term *model* is understood in a very broad sense, and by M we denote all of our modeling assumptions in their entirety. The model parameters of primary interest are the 3D coordinates X of the structural representation of the biomolecular system. Because our major goal is to make inferences about X , the structural parameters are listed explicitly in our probabilistic formulation. However, the model M comprises not only the structural representation X , but all assumptions that are required to model the data.

If D denotes the structural data that we want to use for structure determination, then the modeling process involves formulating a *likelihood* $\Pr(D | X, M)$, which is the probability of measuring data D assuming that the system adopts structural coordinates X . The likelihood connects the observed data D with the structural parameters X via the model M . To establish such a connection might require additional parameters that are not of primary interest, but still necessary to formulate the model. Here, we absorb these *nuisance parameters* into the model M . Nuisance parameters are very common in integrative modeling and typically needed for a complete description of the data. For example, NMR data require the introduction of calibration factors or alignment tensors. Another important type of nuisance parameter are weighting factors that reflect the amount of noise in the data (Habeck et al. 2006).

Typically the likelihood is constructed as follows. First, we derive a *forward model* f that predicts idealized data (mock data) that should ideally match the observed data. In practice, a perfect match between observed and mock data is highly unlikely due to experimental noise or shortcomings of

the model itself (e.g. approximations in the forward model, limited resolution of the representation, etc.). Second, we choose an *error model* that accounts for deviations of observed data D and mock data $f(X)$. A common choice is a Gaussian error model, but other choices might be more appropriate such as a lognormal distribution for distance data, or a Poisson distribution for counts.

The second ingredient of a Bayesian model is the *prior* probability $\Pr(X | M)$. The prior expresses what we assume about reasonable structural models independent of the data. Depending on the representation, more or less detailed knowledge about X is available. At atomic resolution, we can use literature values for bond lengths, bond angles and other stereochemical parameters which are encoded in molecular mechanics force fields. At coarser resolution, common terms such as excluded volume or statistical potentials will be incorporated into the prior.

Bayes' theorem stipulates that the *posterior distribution* $\Pr(X | D, M)$ is proportional to the product of the prior and the likelihood:

$$\Pr(X | D, M) \Pr(D | M) = \Pr(D | X, M) \Pr(X | M) \quad (1)$$

The proportionality factor $\Pr(D | M)$ is called the *model evidence* and not needed to compute structural models. However, the model evidence is crucial when we want to compare different modeling assumptions summarized by M (Knuth et al. 2015).

Equation (1) nicely separates the two stages of applying Bayesian inference in practice: Setting up the right hand side involves making choices about the model that describes the data and the structure. These choices need to be made by the modeler and are therefore, in parts, subjective; they should be challenged by alternative models. Bayesian inference demands us to make all of the modeling assumptions explicit and estimate them as well from the data, if we are uncertain about them. The left hand side is the actual inference, which is apparent from the fact that the role of the data D and the structural coordinates X is swapped. Equation (1) is a direct consequence of the rules of probability theory. The second stage of applying Bayesian inference involves *computation* of the factors on the left hand side, i.e. drawing samples from the posterior distribution and evaluating the model evidence.

In the Bayesian view, the posterior $\Pr(X | D, M)$ captures everything that can be said about the structural model X given the experimental data D in the light of all our modeling assumptions M . Assuming that the model is correct (or at least useful) the posterior is the actual result of a structure determination exercise. The resulting posterior probability quantifies to which extent the structural coordinates are determined by all of the information that is incorporated into the model.

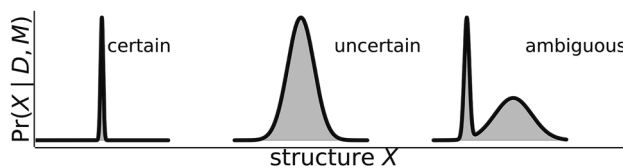


Figure 2: The posterior probability $\Pr(X | D, M)$ reflects the uncertainty in the structure when taking all data and model assumptions into account.

Figure 2 schematically shows possible posterior shapes reflecting different states of uncertainty. If the posterior peaks sharply at a single structure, then the experimental data in combination with the background information suffice to determine the structure unambiguously. This typically requires high-resolution data from X-ray, cryo-EM or NMR. Next, the posterior could still exhibit a single dominant peak which is broadened due to the lack of data. To represent the increased uncertainty, we should not only try to locate the maximum of the peak, but also generate structural models that reflect the uncertainty under the peak. These structures are suboptimal solutions which are missed by an optimization approach. In the most general situation which is typical for integrative modeling, the posterior distribution will exhibit multiple peaks. Each peak corresponds to a structure ensemble that explains the data (at least to some degree). The higher the peak, the better is the agreement between observed and mock data. But not only the height of the peak, also the probability mass under the peak is important, because it reflects how common a structure is, which again argues for the use of sampling (rather than optimization) methods for structure modeling.

2.2 Bayesian computation via posterior sampling

Often the posterior is converted into a scoring function by taking its negative logarithm (Rout and Sali 2019). Highly probable structures then correspond to structures with a low scoring function $-\log\Pr(D | X, M) - \log\Pr(X | M)$. The first term is a restraint energy that favors structural models X minimizing the discrepancy between the observed data and mock data computed from X . The second term involving the log prior acts like a regularizer. Structural models with high posterior probability can be computed by minimizing the score.

However, optimization approaches are limited in that they reduce the full posterior distribution to a few points in conformation space. From a truly Bayesian perspective, the solution of a structure determination problem is given by the posterior distribution $\Pr(X | D, M)$ itself. Therefore, we should try to capture the shape of the posterior as closely as possible. A

single or few structures cannot capture the shape of the posterior distribution unless it collapsed into a single peak. This is very unlikely to happen in integrative modeling where varied data of different quality and amount need to be combined.

An advantage of the Bayesian approach is that all sources of information about the structure of a biomolecular system enter with an appropriate strength. Likelihoods are normalized with respect to the data and therefore there is no freedom or need to set weighting factors. The width of the likelihood (which determines the weight of the log-likelihood) is determined by the quality of the data and is either known or can be estimated along with the structure parameters (Habeck et al. 2006; Rieping et al. 2005).

To infer a probabilistic model and take full advantage of the Bayesian approach, we can use statistical sampling methods and generate representative model realizations from the posterior distribution. Posterior samples can be used in several ways to assess the reliability of a structural model. The uncertainty of model parameters is readily available in the form of posterior histograms. Uncertain parameters are represented by broad histograms; ambiguous model parameters are indicated by multimodal posterior histograms.

In addition to the posterior probability, the model evidence, also known as marginal likelihood, is an important result of a Bayesian analysis. It is desirable to provide an estimate of the model evidence in addition to posterior samples, because different modeling approaches can be compared in an objective way via the estimated model evidences. Alternative models could differ in the choice of representation of a structure (e.g. its resolution), the forward model that connects structural data D with the structural representation, the error model or prior information.

3 Applications of Bayesian inference in integrative modeling

This section discusses some applications of Bayesian methods to integrative modeling from a variety of data at largely different scales. The selection of topics is limited and does not intend to cover all aspects relevant to integrative modeling. There is also no space to discuss the many impressive applications of integrative modeling such as, for example, the recent models of the nuclear pore complex (Akey et al. 2022; Kim et al. 2018; Mosalaganti et al. 2022).

3.1 Structural modeling with cryo-EM maps

Cryo-electron microscopy (cryo-EM) has emerged as one of the most powerful experimental methods for structure

determination of large macromolecular complexes. Structural models are obtained from cryo-EM maps mainly by solving one of two tasks: rigidbody docking or flexible fitting.

A common approach to obtain a structural model of a multi-component complex is to first place the subunit structures rigidly into the EM map and then refine these structures flexibly. However, this approach requires that high-resolution structures or homology models of the subunits are available, and that the subunits do not undergo a drastic conformational change during the assembly of the complex. A growing number of computational tools for structural modeling with cryo-EM maps has been developed (Villa and Lasker 2014). These are typically non-Bayesian approaches that maximize the cross-correlation coefficient.

Here, the discussion is restricted to Bayesian approaches for structural modeling with cryo-EM maps. Bonomi et al. (2019) proposed a method for rigid multi-body docking that combines Bayesian inference with coarse-grained Gaussian mixture model (GMM) representations of the experimental map and subunit structures. They use Monte Carlo sampling of the rigid transformations within the IMP software. The approach is benchmarked on 21 protein/DNA complexes consisting of 2–7 subunits at intermediate resolution of 10 Å. On 16 examples, the approach achieves a good RMSD of 2.2 Å on average. However, there are a few instances where the correct assembly is missed either partially or completely. The reason is most likely inefficient sampling of rigid body poses in multi-component complexes. A general Bayesian framework for rigid and flexible fitting of protein structures into cryo-EM maps has been introduced by Habeck (2017) and implemented in the ISD software (also see Introduction). Both atomic and coarse-grained representations of molecular systems are supported. In flexible fitting, the structural parameters are updated with Hamiltonian Monte Carlo. On the Flex-EM benchmark (Topf et al. 2008), ISD produces fits that are systematically better than Flex-EM.

With the improvement of EM reconstructions towards atomic resolution (Yip et al. 2020), full-atom modeling is the final bottleneck to also obtain an atomic model in addition to the density map. This is the domain of flexible fitting of atomic structures into EM maps. Blau et al. (2016) have combined molecular dynamics (MD) with a cryo-EM fitting score that they derive in a Bayesian fashion. Vuilleminot and Jonić (2021) combine normal-mode based fitting with cryo-EM maps within a Bayesian framework. A more recent development is CryoFold (Shekhar et al. 2021) which combines the Bayesian MELD approach (MacCallum et al. 2015) with other modeling tools. CryoFold generates an ensemble of high-quality structural models without requiring high-resolution subunit structures as

input. However, the density map needs to be of a sufficiently high resolution of about 3–5 Å.

3.2 Coarse-graining of biomolecular systems

Choosing a good representation for structural modeling is an essential part of integrative approaches. Coarse-graining (CG) lumps groups of atoms together into structural building blocks, typically spherical beads, while aiming to preserve the structural and energetic properties as best as possible. Representing biomolecules with CG models seems appropriate due to their hierarchical architecture. The goal is to bridge between multiple spatial and temporal scales. Various CG representations of proteins, lipids and other biomolecules have been developed to enable MD simulations of large systems that are not suitable for all-atom simulations (Jin et al. 2022; Noid 2013; Pak and Voth 2018).

In the context of integrative modeling, CG representations are used to cope with the lack of data, resolution and the large size of the systems. Statistical potentials and potentials of mean force have a long tradition in protein modeling and structure prediction. Here, we focus on recent Bayesian approaches to coarse-graining. Bottom-up approaches typically run atomistic simulations from which CG models are learned. Farrell et al. (2015) have developed a general Bayesian framework to build, calibrate and validate CG models of atomistic systems. Model selection is applied to a family of CG models. Predictive coarse-graining is a Bayesian approach for learning CG models from simulations based on a combination of MAP and Monte Carlo simulation (Schöberl et al. 2017). BICePs focuses on the generation of conformational ensembles (Voelz et al. 2021). Posterior sampling is coupled to model selection techniques to optimize different types of hyperparameters. The framework can be used to optimize reference potentials or CG representations (Ge and Voelz 2018). Ultra-coarse-graining is more relevant to modeling of large biomolecular systems. Here many atoms are grouped together to form a bead (Jin and Voth 2018).

The CG approaches mentioned so far try to map the molecular representation so as to preserve the energetic parameters as best as possible. More pragmatic approaches focus on low-resolution representation of molecular structures. In cryo-EM modeling, Kawabata (2008) proposed the use of Gaussian mixture models (GMMs) to represent density maps as well as atomic structures. The GMM representation has been adopted by a several modeling approaches, most prominently by IMP (see work by Bonomi et al. (2019) and the discussion in Section 3.1 for an application). CG representations can be learned both from static atomic biomolecular

structures and cryo-EM reconstructions using a Bayesian mixture model approach (Chen and Habeck 2017). This approach estimates an optimal mapping between atoms and beads, the bead positions as well as the non-bonded radius of the particles and their interaction strengths.

Another application is to infer CG models from projection images obtained with cryoEM (Joubert and Habeck 2015; Vakili and Habeck 2021). A highly coarse-grained representation of the 3D structure is generated from 2D projection images via posterior sampling. The unknown structure is represented by hundreds to thousands of beads rather than a density map over a voxel grid. The particles are arranged such that the 2D projection images are reproduced as best as possible. Along with the particle positions the posterior sampling algorithm estimates the unknown projection directions and can therefore be used for initial structure determination in single-particle analysis. The approach samples reconstructions that are of a quality and resolution similar to reconstructions generated with other approaches for initial model generation such as PRIME (Elmlund et al. 2013). However, it remains unclear if the particle-based reconstruction approach can be scaled to a larger number of particles (>10,000 particles) and thereby yield initial reconstructions at higher resolutions.

3.3 Structural modeling of chromosomes and genomes

An emerging field is structural modeling of entire chromosomes or genomes. This development is powered by high-throughput chromosome conformation capture (Hi-C) techniques. Single-cell Hi-C (Nagano et al. 2013) probes through-space contacts between different chromosomal regions and can be used for structural modeling in a fashion similar to NMR distance bounds. Due to limitations in resolution, but also the sheer size of chromosomes only highly coarse-grained representations of chromosomes are amenable to structural modeling.

Among the multitude of chromosome structure modeling software, we only mention Bayesian approaches (Carstens et al. 2016; Hu et al. 2013; Meng et al. 2021; Rosenthal et al. 2019; Wang et al. 2015; Xie et al. 2022). Typically, the chromosome fiber is represented with a highly coarse-grained polymer model, a “beads on a string” model, a chain of beads that represent 50–500 kilobases of chromatin. Carstens et al. (2016) applies the ISD framework to sparse contacts from single-cell Hi-C. Chromosome structures are generated from the posterior distribution with Monte Carlo methods. The structural models computed with ISD are consistent with independent measurements from

fluorescence microscopy. More recently, Meng et al. (2021) have developed Si-C, a Bayesian approach for genome structure modeling from single-cell Hi-C data. Si-C works with beads at 10 kb–100 kb resolution and can utilize whole-genome data measured on single cells (Stevens et al. 2017). The data were measured on haploid cells and allow the reconstruction of the structure of an entire genome. The models are better resolved than those originally obtained by Stevens et al. (2017). Because with higher resolution the data become increasingly sparse, it is unclear which resolution (i.e. number of beads) still yields well-defined models. Using an RMSD-based criterion, the authors find that a resolution with at least 0.2 contacts per bead on average is required to generate well-defined ensembles. The calculated “super-resolution” structures are validated against independent imaging data and yield unprecedented details on the architecture of entire genomes. Although the resolution of these models is still orders of magnitude below base-pair resolution, the models are useful to study the global architecture of genome organization such as the formation of topologically associated domains (TADs) or A/B compartments.

3.4 Inference of conformational heterogeneity

Biomolecular systems are highly dynamic and undergo conformational changes. Therefore, representing the structure of a biomolecular system as a single conformational state is limited. Some data might even require modeling with a conformational ensemble composed of multiple states. For example, methods such as NMR or bulk Hi-C observe ensemble averages rather than single-state data. Multiple states are observed with XLMS or cryo-EM.

There has been an abundance of research activity aimed at developing Bayesian and maximum entropy methods for the inference of ensembles representing diverse conformational states (Bonomi et al. 2017). Typical input data are ensemble-averaged NMR or solution-scattering data. Several groups have combined the principle of maximum entropy with a Bayesian framework (Beauchamp et al. 2014; Bonomi et al. 2016; Bottaro et al. 2020; Hummer and Köfinger 2015). BioEn uses reweighting techniques in combination within a Bayesian framework and an entropic prior (Hummer and Köfinger 2015). Metainference infers a multi-state model (Bonomi et al. 2016). BME is another combination of Bayesian methods and maximum entropy (Bottaro et al. 2020) and has recently been applied to RNAs (Zhang and Frank 2021).

Various Bayesian approaches specialize on generating protein ensembles from small-angle scattering (SAS) data (Antonov et al. 2016; Pesce and Lindorff-Larsen 2021;

Potrzebowski et al. 2018; Spill et al. 2021). Model selection techniques have been used to determine the size of the ensemble (Bowerman et al. 2019; Potrzebowski et al. 2018). Ensembles with full-atom detail are generated using the Bayesian approach of Shevchuk and Hub (2017). A method to infer ensembles of disordered protein states from NMR, FRET and SAXS data has been developed by Lincoff et al. (2020).

The principle of maximum entropy has also been applied to chromosome structure modeling (Lin et al. 2021). An application of Bayesian techniques to ensemble modeling based on bulk Hi-C data chromosome modeling has been developed. To choose the number of states, Carstens et al. (2020) have used the model evidence, which is estimated with an MCMC approach.

4 Challenges

Integrative modeling is a highly complex task that currently requires a number of decisions and interventions by the modeler. Bayesian inference has the potential to automate these decisions and assign probabilities to them. Moreover, models can be criticized and ranked relative to each other within a Bayesian framework. Since Bayesian approaches separate a data analysis into a modeling and a computation task, it is also natural to think of challenges in integrative modeling that belong to either of these two categories.

4.1 Modeling challenges

Molecular representation The choice of molecular representation has a strong impact on the usefulness and success of an integrative modeling exercise. Depending on the resolution and amount of data, a more or less detailed representation is adequate. A multi-scale representation might be required when data of different quality and resolution are mixed. In particular, as we approach cellular-scale systems, there is a large degree of arbitrariness in the choice of representation. Even if high-resolution structures of the subunits of a macromolecular complex are available, it is not generally clear what the optimal model representation is. The degree of coarse-graining should be driven by the resolution and the amount of data. Likewise, the number of states should also be sampled in ensemble modeling. To address some of these challenges, Viswanath and Sali (2019) have proposed a method to optimize the molecular representation for integrative modeling.

Bridging between the molecular and cellular scale A formidable challenge is to integrate data showing the system at largely different scales. This requires the development of

multi-scale representations that describe the same molecular component at different resolutions such that they can be linked to the different types of input data. A multi-scale model needs to connect the different scales and their associated conformational degrees of freedom. Different parts of a structural model might be better resolved than others. A multi-scale model should represent a hierarchy of spatial scales and automatically adapt its maximum resolution so as to avoid overfitting. The model should also quantify non-specific interactions between biomolecules. Moreover, since the structures should be embedded into the cellular context, also structural representations of the cellular environment will become necessary (Im et al. 2016). New representations as the one proposed by Singla et al. (2022) might be useful to bridge between molecular and cellular scales. This representation builds biological structures from tetrahedra and is commonly used in numerical analysis to represent complicated three-dimensional shapes. The representation might be useful for modeling assemblies in crowded cellular environments with many diverse components.

Probabilistic models for structural data A diverse array of structural data is already used to build integrative structures (Rout and Sali 2019) and will become even more heterogeneous in the future. Each method requires a probabilistic model that links the data to the structural representation of the system. The probabilistic model should be realistic in the sense that the main characteristics of the data and the noise are captured adequately. At the same time, the model should be computationally efficient. For example, SAXS curves can be computed with high accuracy and detail, but evaluating these models is computationally expensive. Efficient models for computing approximate EM maps exist, but the noise models are typically very limited in that they ignore correlations between neighboring voxels.

4.2 Computational challenges

Software Bayesian integrative modeling is enabled by software packages such as the Integrative Modeling Platform (IMP) (Russel et al. 2012; Saltzberg et al. 2021), BioEn (Hummer and Köfinger 2015) or PLUMED-ISDB (Bonomi and Camilloni 2017). In the future, it will be important to make Bayesian integrative modeling available to non-experts. Given the diversity of the data and the systems there are many challenges that a software has to meet. The software should be easy to install, run and used with diverse structural data. A related issue is to train users of Bayesian integrative modeling software in probabilistic reasoning and modeling.

Efficient conformational sampling Conformational sampling refers to the process of generating representative

structures X from the posterior $\Pr(X | D, M)$. Since X tends to be high-dimensional, direct sampling is inefficient and not viable. Therefore, one typically resorts to *Markov chain Monte Carlo* (MCMC) sampling which explores the conformation space by running a Markov chain. Conformational sampling is more or less challenging depending on the quality and amount of data. MCMC algorithms tend to converge very slowly. Faster MCMC algorithms and implementations are required to enable Bayesian integrative modeling of ever larger systems. Also pragmatic approaches such as jump-starting an MCMC simulation from optimized structures should be supported.

Automation of posterior inference MCMC algorithms typically come with a number of parameters that need to be optimized by the user. An example is the temperature schedule of parallel tempering simulations, which is one of the working horses of exhaustive posterior sampling in integrative modeling. Often, a large fraction of computation time is spent on optimizing algorithmic parameters so as to guarantee the convergence of the MCMC simulation. Adaptive MCMC methods that optimize algorithmic parameters in the course of simulation need to be developed to enable Bayesian integrative modeling also for non-experts with limited computational resources.

Validation of posterior sampling MCMC is an approximate method and can fail. A problem with all sampling methods is that they rely on asymptotic guarantees, meaning that they will produce correct samples only in the long run. But there is no sufficient measure that assesses whether a Markov chain has already converged. An MCMC simulation can fail in several ways. The simulation might miss an important peak of the posterior distribution, i.e. a structure that is a good explanation of the data. The simulation might misrepresent the probability mass under a peak, meaning that the frequency with which a structure is visited by the Markov chain might be biased and fail to reflect the true importance of the structure. Viswanath et al. (2017) have developed several techniques to assess the exhaustiveness of conformational sampling. These need to be extended in the future.

4.3 Deposition, validation and analysis of Bayesian integrative structures

Finally, there are a couple of challenges that are related to the deposition, validation and analysis of integrative structures obtained with a Bayesian approach.

Deposition of posterior ensembles In the Bayesian view, the solution of an integrative structure determination is the full posterior distribution $\Pr(X | D, M)$ rather than a single structure or a few structures. But how can we

adequately represent a posterior distribution with finite means? How many structures suffice to capture the posterior? These questions are even more pressing, if ensembles of structures are inferred that represent diverse conformational states. Thinning approaches that reduce the number of MCMC samples could help us compress the large amount of data and information contained in posterior samples. The Protein Data Bank (PDB) has created an archive for integrative structures (Vallat et al. 2021). PDB-Dev provides timely support for integrative structure modeling, but might require future extensions to represent all relevant aspects of Bayesian integrative modeling. For example, it is desirable to report the complete posterior sampling protocol as well as samples of additional nuisance parameters. Moreover, the modeling software and deposition system should aim to ensure that the structure determination can be reproduced.

Validation of integrative structures Statistical methods, in particular those based on Bayesian inference, enable us to make sound statements about the quality of an integrative structure. New quality measures for the validation of integrative structures need to be developed based on concepts and techniques from Bayesian inference. A prerequisite for statistical model assessment is that we sample conformational space exhaustively. Therefore, an important aspect is to also improve and validate conformational sampling techniques (Viswanath et al. 2017).

Although the number of integrative structures continues to grow, the community still lacks a generally accepted quality measure for structures obtained by integrating data from multiple experimental sources. Estimated data weights (Habeck et al. 2006) should prove useful for the validation of integrative structures. From a Bayesian perspective, the data weights have an intuitive interpretation: They are the inverse squared error of a dataset, so high-quality data with low noise levels will be assigned a high weight.

Other statistical figures of merit should also be reported along with the structure. For example, the entropy of the posterior measures the effective number of structures needed to represent the posterior. Another largely unexplored quantity is the *model evidence* $\Pr(D | M)$. The model evidence is the goodness of fit averaged over all possible conformations X . Since it balances goodness of fit against model complexity, the model evidence embodies Occam's razor and can be used to select among competing models of explaining the data.

Arroyuelo et al. (2021) have recently proposed various Bayesian tools to assess the quality of an NMR ensemble. These tools report per residue quality measures that allow the authors to identify problematic residues. The approach requires atomic structures and NMR chemical shifts. But some of the strategies might be transferable to integrative structures. We envision that based on PDB-Dev similar

reference distributions can be derived for coarse-grained representations and other types of data.

Analysis of posterior samples A Bayesian integrative structure is encoded in the marginal posterior distribution of the structural parameters. This distribution needs to be approximated with efficient clustering methods that analyze the large amount of structural samples. Based on a cluster analysis we can then report the number of clusters, their weights and precisions to obtain insights into the shape and quality of the posterior ensemble. High-quality data will result in a single, narrow cluster, whereas ambiguous and sparse data result in several, broad clusters. This information should feed back into the validation of integrative structures. To validate structural models at atomic resolution it is also important to compare the models with related structures in the PDB. Bayesian methods for structural alignment of proteins have been developed by Rodriguez and Schmidler (2014) and Fallaize et al. (2020). In addition, powerful probabilistic methods for characterizing structure ensembles are also available (Theobald and Wuttke 2008). However, when working at non-atomic resolution, new tools for assessing the precision of structural models need to be developed. PrISM identifies regions of low and high precision in ensembles of bead models (Ullanat et al. 2022) and color-codes them in a graphical representation. That way, reliable and ambiguous regions can easily be identified. In addition, it is desirable to devise new visual analytics tools for Bayesian integrative structures. These tools could show distributions of structural degrees of freedom and additional model parameters.

5 Perspectives

5.1 High-resolution integrative modeling powered by ML-based structure prediction

Protein structures predicted with ML-based systems such as AlphaFold2 (AF2) and RoseTTAFold are already used in integrative modeling. For example, the combination of cryo-ET maps with high-resolution structure prediction has enabled near-atomic modeling of the nuclear pore complex (Mosalganti et al. 2022). Another example is the integration of intermediate-resolution cryo-EM and AF2 to obtain a detailed model of the cytoplasmic ring of the nuclear pore complex (Fontana et al. 2022). A natural next step in protein structure prediction was to use ML-systems to predict protein complexes. In their assessment of the impact of AF2 on structural biology, Akdel et al. (2022) found that AF2 is already capable of predicting the structure of some protein complexes and

outperforms standard docking approaches. Several extensions of AF2 and other ML systems have been proposed that will also facilitate integrative modeling. AF2 has recently been extended to specifically target protein complexes (Evans et al. 2021). AlphaFold-multimer outperforms AF2 on predicting protein complexes. While the performance is already quite impressive, the predictions of protein complexes tends to be more accurate for symmetric complexes rather than heteromeric assemblies. To improve the prediction of heteromeric complexes, FoldDock (Bryant et al. 2022a) combines AF2 with docking methods. AF2Complex predicts physical contacts between subunit structures without requiring paired sequence alignments (Gao et al. 2022). Several large-scale modeling studies have been carried out to predict the structures of entire interactomes. Humphreys et al. (2021) applied RoseTTAFold and AF2 to predict protein complexes and model their 3D structures. FoldDock has been used to predict complex structures of the human protein interaction network (Burke et al. 2023). Bryant et al. (2022b) use AF-Multimer and FoldDock to predict the assemblies that comprise as many as 30 chains. Nonetheless large assemblies still pose a formidable challenge and typically require additional information. Another challenge is that systems like AF2 do not predict all accessible conformations of a biomolecule (Lane 2023) and miss out on the structure of nucleic acids (Tunyasuvunakool 2022). RoseTTAFoldNA predicts complexes of proteins and nucleic acids (Baek et al. 2022). More applications of AF2 and RoseTTAFold relevant to challenges in integrative modeling are under development. For example, Terwilliger et al. (2022) combined AF2 with experimental information. Stein and Mchaourab (2022) have modified AF2 so as to generate conformational ensembles. There is no foreseeable end to the development of ML methods in protein structure prediction in the near future. For example, language models have recently be used to enable rapid and highly reliable prediction of protein structures (Lin et al. 2023). For integrative modeling, we expect that physics-based modeling in combination with experimental data and ML-based models of subunits will continue to be the most promising approach given the scarcity and heterogeneity of training data.

5.2 Learning of coarse-grained representations and forcefields

In principle, Bayesian methods can be used to infer any type of parameter including parameters of the forcefield used in the prior distribution $\Pr(X | M)$. The technical difficulty of learning these hyperparameters is that the normalizing constant of the prior depends on the parameters, but is not available in closed form resulting in a doubly intractable model. Approximate

MCMC algorithms have been developed to infer doubly intractable models. Various approaches to learn forcefields with Bayesian methods have been proposed (Bottaro and Lindorff-Larsen 2018; Ge and Voelz 2018; Habeck 2014; Köfinger and Hummer 2021; Madin et al. 2022). These need to be improved and should become a part of integrative modeling itself.

Machine learning also impacts coarse-graining of proteins and their forcefields (Ding et al. 2023; Durumeric et al. 2023). A promising development is the use of representation learning in the context of molecular structures. These models could serve as prior distributions over the structural degrees of freedom. Particle-based simulator models have been inferred for deformable objects and fluids (Li et al. 2018) and might also be learnable for biomolecular systems. Wang et al. (2022) use generative models to train CG representations of molecules including short peptides.

5.3 Improved posterior sampling via coarse-graining and variational inference

Uncertainty quantification requires posterior sampling. Therefore, drawing structural models from the posterior distribution is at the core of a truly Bayesian approach to integrative modeling. It is also a major bottleneck in scaling the Bayesian approach to large complexes and cellular systems. Coarse-graining might not only help to cope with the scarcity of data, but also benefit posterior sampling and should be exploited for sampling in a more systematic fashion.

Posterior sampling typically relies on Markov chain Monte Carlo simulation. The current working horse in Bayesian integrative modeling are MCMC methods such as Hamiltonian Monte Carlo and parallel tempering (Rieping et al. 2005; Saltzberg et al. 2021). A viable alternative might be to learn a variational approximation of the posterior. Therefore, ML techniques powered by deep generative models could also help overcome the sampling problem in integrative modeling. For example, Boltzmann generators introduced by Noé et al. (2019) use deep architectures for sampling the Boltzmann distribution of biomolecules. These should be applicable to posterior distributions arising in integrative structure determination. More, recently Monroe and Shen (2022) have demonstrated learning of efficient, collective Monte Carlo moves with variational autoencoders.

5.4 Modeling the dynamics of biomolecular systems

Biological function involves dynamical changes in the structure and composition of biomolecular systems.

Therefore, integrative modeling has to go beyond static structures and infer structure and dynamics simultaneously from the data. Inference of conformational ensembles and multi-state models is a first step to go beyond static models. Simultaneous determination of protein structure and dynamics from cryo-EM data has been proposed by Bonomi et al. (2018). Giraldo-Barreto et al. (2021) extract free-energy profiles from cryo-EM micrographs using a Bayesian approach called BIFE. BIFE recovers free-energy profiles along a path collective variable from cryo-EM images. On a real data set, not only the most probable conformation but also metastable states are found and activation barriers are estimated thereby providing a more complete characterization of the thermodynamic ensemble. Recent developments in representation learning also enable the reconstruction of heterogeneous structures directly from cryo-EM data. CryoDRGN uses a variational autoencoder (VAE) to learn ensembles of cryo-EM structures from projection images (Zhong et al. 2021).

What are efficient representations of dynamic structures? An explicit representation is a trajectory of successive conformational states. Also in the context of representing biomolecular dynamics, deep generative models inspired by recent developments in representation learning are a promising approach (Hoseini et al. 2021). It is becoming apparent that going beyond structures and searching for efficient and expressive representations will become important. In the light of the ongoing success of ML-based approaches to structure prediction, also more conceptual questions such as “What is a structural model?” will become relevant. We are used to think of structural models as an explicit set of 3D coordinates. Implicit models that *generate* coordinate arrays, potentially at varying resolution, might replace explicit representations. This would have vast implications on the way biomolecular structures are represented, stored and used. Guo et al. (2021) generate protein structures with graph VAEs. Diffusion models such as torsional diffusion (Jing et al. 2022) or EigenFold (Jing et al. 2023) can generate protein structures and could be efficient representations of conformationally heterogeneous and flexible protein structures. More generally, differential programs such as simulators will play an important role in modeling biophysical processes (AlQuraishi and Sorger 2021). For example, Ingraham et al. (2019) propose a differentiable simulator to model protein structures.

5.5 Emerging data

A number of emerging experimental techniques will provide additional sources for integrative modeling. These data need to be supported by new probabilistic models linking the data

to the structural degrees of freedom. New types of data will help us bridge the gap between the atomic and cellular scale. For example, tomographic methods such as cryo-ET and soft X-ray tomography provide 3D maps of organelles and other cellular compartments (Loconte et al. 2022, 2023). The 4D nucleome project collects detailed information about the structural organization of the cell nucleus at various stages of the cell cycle (Dekker et al. 2017). A multitude of proximity ligation and imaging methods enables the study of genome organization (Jerkovic and Cavalli 2021; Misteli 2020) including ChromEMT (Ou et al. 2017). These developments are complemented by light microscopy techniques that allow for structural structures of protein complexes and larger structural units (Sieben et al. 2018; Sigal et al. 2018). Correlative microscopy combining fluorescence with electron microscopy allows the visualization of whole cells at many scales (Hoffman et al. 2020). Multiscale modeling of biological systems will require neighborhood information obtained by mapping physical and functional proximities (Schaffer and Ideker 2021). There is a growing array of proximity ligation assays that allows the characterization the structure of genomes, chromosomes or large RNAs. Advanced cross-linking techniques uncover interactions within and between interacting proteins (Braberg et al. 2020; Graziadei and Rappsilber 2022; Mintseris and Gygi 2020; O'Reilly et al. 2020; Sae-Lee et al. 2022; Yperman et al. 2021). We also see a continued improvement of cryo-EM maps towards atomic resolution (Yip et al. 2020) and the emergence of single-molecule diffraction techniques based on, for example, free electron lasers (von Ardenne et al. 2018).

5.6 Structural cell modeling

The long-term goal of integrative modeling is the creation of physical molecular models of cellular components and of entire cells (Singla et al. 2018; Zhong et al. 2022). To achieve this goal, a large array of data and additional information needs to be fed into the modeling procedure to produce a consistent model of the cell. A promising approach is to couple simulation software such as cellPACK (Johnson et al. 2015) to cellular data. An impressive demonstration of this approach is the structural model of a mycoplasma cell (Maritan et al. 2022). Structural cell modeling is facing a number of challenges (Im et al. 2016). Bayesian metamodelling of cellular systems is a highly exciting development that aims to tackle these challenges within a Bayesian framework (Raveh et al. 2021).

Acknowledgments: The author gratefully acknowledges funding by the German Research Foundation (DFG) within

SFB 860, project B09 and the Carl Zeiss Foundation within the program “CZS Stiftungsprofessuren”.

Author contributions: The author has accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: None declared.

Conflict of interest statement: The author declares no conflicts of interest regarding this article.

References

- Akdel, M., Pires, D.E., Pardo, E.P., Jänes, J., Zalevsky, A.O., Mészáros, B., Bryant, P., Good, L.L., Laskowski, R.A., Pozzati, G., et al (2022). A structural biology community assessment of alphafold2 applications. *Nat. Struct. Mol. Biol.* 29: 1–12, <https://doi.org/10.1038/s41594-022-00849-w>.
- Akey, C.W., Singh, D., Ouch, C., Echeverria, I., Nudelman, I., Varberg, J.M., Yu, Z., Fang, F., Shi, Y., Wang, J., et al. (2022). Comprehensive structure and functional adaptations of the yeast nuclear pore complex. *Cell* 185: 361–378.
- AlQuraishi, M. and Sorger, P.K. (2021). Differentiable biology: using deep learning for biophysics-based and data-driven modeling of molecular mechanisms. *Nat. Methods* 18: 1169–1180.
- Antonov, L.D., Olsson, S., Boomsma, W., and Hamelryck, T. (2016). Bayesian inference of protein ensembles from SAXS data. *Phys. Chem. Chem. Phys.* 18: 5832–5838.
- Arroyuelo, A., Vila, J.A., and Martin, O.A. (2021). Exploring the quality of protein structural models from a Bayesian perspective. *J. Comput. Chem.* 42: 1466–1474.
- Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* 373: 871–876.
- Baek, M., McHugh, R., Anishchenko, I., Baker, D., and DiMaio, F. (2022). Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA. *bioRxiv* 1–16.
- Beauchamp, K.A., Pande, V.S., and Das, R. (2014). Bayesian energy landscape tilting: towards concordant models of molecular ensembles. *Biophys. J.* 106: 1381–1390.
- Blau, C., Lenner, N., Kutzner, C., Grubmüller, H., and Lindahl, E. (2016). From cryo-EM densities to atom coordinates and ensembles with bayes approach. *Biophys. J.* 110: 156a–157a.
- Bonomi, M. and Camilloni, C. (2017). Integrative structural and dynamical biology with PLUMED-ISDB. *Bioinformatics* 33: 3999–4000.
- Bonomi, M., Camilloni, C., Cavalli, A., and Vendruscolo, M. (2016). Metainference: a Bayesian inference method for heterogeneous systems. *Sci. Adv.* 2: e1501177.
- Bonomi, M., Heller, G.T., Camilloni, C., and Vendruscolo, M. (2017). Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* 42: 106–116.
- Bonomi, M., Pellarin, R., and Vendruscolo, M. (2018). Simultaneous determination of protein structure and dynamics using cryo-electron microscopy. *Biophys. J.* 114: 1604–1613.
- Bonomi, M., Hanot, S., Greenberg, C.H., Sali, A., Nilges, M., Vendruscolo, M., and Pellarin, R. (2019). Bayesian weighing of electron cryo-microscopy data for integrative structural modeling. *Structure* 27: 175–188.
- Bottaro, S. and Lindorff-Larsen, K. (2018). Biophysical experiments and biomolecular simulations: a perfect match? *Science* 361: 355–360.
- Bottaro, S., Bengtson, T., and Lindorff-Larsen, K. (2020). Integrating molecular simulation and experimental data: a Bayesian/maximum entropy reweighting approach. In: *Structural Bioinformatics: Methods and Protocols*, pp. 219–240.
- Bowerman, S., Curtis, J.E., Clayton, J., Brookes, E.H., and Wereszczynski, J. (2019). BEES: bayesian ensemble estimation from SAS. *Biophys. J.* 117: 399–407.
- Braberg, H., Echeverria, I., Bohn, S., Cimermancic, P., Shiver, A., Alexander, R., Xu, J., Shales, M., Dronamraju, R., Jiang, S., et al. (2020). Genetic interaction mapping informs integrative structure determination of protein complexes. *Science* 370: eaaz4910.
- Braitbard, M., Schneidman-Duhovny, D., and Kalisman, N. (2019). Integrative structure modeling: overview and assessment. *Annu. Rev. Biochem.* 88: 113–135.
- Bryant, P., Pozzati, G., and Elofsson, A. (2022a). Improved prediction of protein-protein interactions using alphafold2. *Nat. Commun.* 13: 1265.
- Bryant, P., Pozzati, G., Zhu, W., Shenoy, A., Kundrotas, P., and Elofsson, A. (2022b). Predicting the structure of large protein complexes using AlphaFold and Monte Carlo tree search. *Nat. Commun.* 13: 6028.
- Burke, D.F., Bryant, P., Barrio-Hernandez, I., Memon, D., Pozzati, G., Shenoy, A., Zhu, W., Dunham, A.S., Albanese, P., Keller, A., et al. (2023). Towards a structurally resolved human protein interaction network. *Nat. Struct. Mol. Biol.* 30: 216–225.
- Carstens, S., Nilges, M., and Habeck, M. (2016). Inferential structure determination of chromosomes from single-cell Hi-C data. *PLoS Comput. Biol.* 12: e1005292.
- Carstens, S., Nilges, M., and Habeck, M. (2020). Bayesian inference of chromatin structure ensembles from population-averaged contact data. *Proc. Natl. Acad. Sci. U.S.A.* 117: 7824–7830.
- Chen, Y.-L. and Habeck, M. (2017). Data-driven coarse graining of large biomolecular structures. *PloS One* 12: e0183057.
- Dekker, J., Belmont, A.S., Guttman, M., Leshyk, V.O., Lis, J.T., Lomvardas, S., Mirny, L.A., O’shea, C.C., Park, P.J., Ren, B., et al. (2017). The 4d nucleome project. *Nature* 549: 219–226.
- Ding, Y., Yu, K., and Huang, J. (2023). Data science techniques in biomolecular force field development. *Curr. Opin. Struct. Biol.* 78: 102502.
- Durumeric, A.E., Charron, N.E., Templeton, C., Musil, F., Bonneau, K., Pasos-Trejo, A.S., Chen, Y., Kelkar, A., Noé, F., and Clementi, C. (2023). Machine learned coarse grained protein force-fields: are we there yet? *Curr. Opin. Struct. Biol.* 79: 102533.
- Elmlund, H., Elmlund, D., and Bengio, S. (2013). PRIME: probabilistic initial 3D model generation for single-particle cryo-electron microscopy. *Structure* 21: 1299–1306.
- Evans, R., O’Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Židek, A., Bates, R., Blackwell, S., Yim, J., et al (2021). Protein complex prediction with AlphaFold-Multimer. *BioRxiv* 1–25.
- Fallaize, C.J., Green, P.J., Mardia, K.V., and Barber, S. (2020). Bayesian protein sequence and structure alignment. *J. Roy. Stat. Soc. C Appl. Stat.* 69: 301–325.
- Farrell, K., Oden, J.T., and Faghihi, D. (2015). A Bayesian framework for adaptive selection, calibration, and validation of coarse-grained models of atomistic systems. *J. Comput. Phys.* 295: 189–208.
- Fontana, P., Dong, Y., Pi, X., Tong, A.B., Hecksel, C.W., Wang, L., Fu, T.-M., Bustamante, C., and Wu, H. (2022). Structure of cytoplasmic ring of nuclear pore complex by integrative cryo-EM and AlphaFold. *Science* 376: eabm9326.

- Gao, M., Nakajima An, D., Parks, J.M., and Skolnick, J. (2022). Af2complex predicts direct physical interactions in multimeric proteins with deep learning. *Nat. Commun.* 13: 1744.
- Ge, Y. and Voelz, V.A. (2018). Model selection using BICePs: a Bayesian approach for force field validation and parameterization. *J. Phys. Chem. B* 122: 5610–5622.
- Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature* 521: 452–459.
- Giraldo-Barreto, J., Ortiz, S., Thiede, E.H., Palacio-Rodriguez, K., Carpenter, B., Barnett, A.H., and Cossio, P. (2021). A Bayesian approach to extracting free-energy profiles from cryo-electron microscopy experiments. *Sci. Rep.* 11: 13657.
- Graziadei, A. and Rappsilber, J. (2022). Leveraging cross-linking mass spectrometry in structural and cell biology. *Structure* 30: 37–54.
- Guo, X., Du, Y., Tadepalli, S., Zhao, L., and Shehu, A. (2021). Generating tertiary protein structures via interpretable graph variational autoencoders. *Bioinform. Adv.* 1: vbab036.
- Habeck, M. (2014). Bayesian approach to inverse statistical mechanics. *Phys. Rev. E* 89: 052113.
- Habeck, M. (2017). Bayesian modeling of biomolecular assemblies with cryo-EM maps. *Front. Mol. Biosci.* 4: 15.
- Habeck, M., Rieping, W., and Nilges, M. (2006). Weighting of experimental evidence in macromolecular structure determination. *Proc. Natl. Acad. Sci. U.S.A.* 103: 1756–1761.
- Hoffman, D.P., Shtengel, G., Xu, C.S., Campbell, K.R., Freeman, M., Wang, L., Milkie, D.E., Pasolli, H.A., Iyer, N., Bogovic, J.A., et al. (2020). Correlative three-dimensional super-resolution and block-face electron microscopy of whole vitreously frozen cells. *Science* 367: eaaz5357.
- Hoseini, P., Zhao, L., and Shehu, A. (2021). Generative deep learning for macromolecular structure and dynamics. *Curr. Opin. Struct. Biol.* 67: 170–177.
- Hu, M., Deng, K., Qin, Z., Dixon, J., Selvaraj, S., Fang, J., Ren, B., and Liu, J.S. (2013). Bayesian inference of spatial organizations of chromosomes. *PLoS Comput. Biol.* 9: e1002893.
- Hummer, G. and Köfinger, J. (2015). Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* 143: 243150-1–243150-14.
- Humphreys, I.R., Pei, J., Baek, M., Krishnakumar, A., Anishchenko, I., Ovchinnikov, S., Zhang, J., Ness, T.J., Banjade, S., Bagde, S.R., et al. (2021). Computed structures of core eukaryotic protein complexes. *Science* 374: eabm4805.
- Im, W., Liang, J., Olson, A., Zhou, H.-X., Vajda, S., and Vakser, I.A. (2016). Challenges in structural approaches to cell modeling. *J. Mol. Biol.* 428: 2943–2964.
- Ingraham, J., Riesselman, A., Sander, C., and Marks, D. (2019). Learning protein structure with a differentiable simulator. In: *International conference on learning representations*.
- Jerkovic, I. and Cavalli, G. (2021). Understanding 3d genome organization by multidisciplinary methods. *Nat. Rev. Mol. Cell Biol.* 22: 511–528.
- Jin, J. and Voth, G.A. (2018). Ultra-coarse-grained models allow for an accurate and transferable treatment of interfacial systems. *J. Chem. Theory Comput.* 14: 2180–2197.
- Jin, J., Pak, A.J., Durumeric, A.E., Loose, T.D., and Voth, G.A. (2022). Bottomup coarse-graining: principles and perspectives. *J. Chem. Theory Comput.* 18: 5759–5791.
- Jing, B., Corso, G., Chang, J., Barzilay, R., and Jaakkola, T. (2022). *Torsional diffusion for molecular conformer generation*, arXiv preprint arXiv: 2206.01729.
- Jing, B., Erives, E., Pao-Huang, P., Corso, G., Berger, B., and Jaakkola, T. (2023). *Eigenfold: Generative protein structure prediction with diffusion models*, arXiv preprint arXiv:2304.02198.
- Johnson, G.T., Autin, L., Al-Alusi, M., Goodsell, D.S., Sanner, M.F., and Olson, A.J. (2015). cellPACK: a virtual mesoscope to model and visualize structural systems biology. *Nat. Methods* 12: 85–91.
- Joubert, P. and Habeck, M. (2015). Bayesian inference of initial models in cryo-electron microscopy using pseudo-atoms. *Biophys. J.* 108: 1165–1175.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596: 583–589.
- Kawabata, T. (2008). Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a Gaussian mixture model. *Biophys. J.* 95: 4643–4658.
- Kim, S.J., Fernandez-Martinez, J., Nudelman, I., Shi, Y., Zhang, W., Raveh, B., Herricks, T., Slaughter, B.D., Hogan, J.A., Upla, P., et al. (2018). Integrative structure and functional anatomy of a nuclear pore complex. *Nature* 555: 475–482.
- Kinz-Thompson, C.D., Ray, K.K., and Gonzalez, R.L., Jr. (2021). Bayesian inference: the comprehensive approach to analyzing single-molecule experiments. *Ann. Rev. Biophys.* 50: 191–208.
- Knuth, K.H., Habeck, M., Malakar, N.K., Mubeen, A.M., and Placek, B. (2015). Bayesian evidence and model selection. *Digit. Signal Process.* 47: 50–67.
- Köfinger, J. and Hummer, G. (2021). Empirical optimization of molecular simulation force fields by Bayesian inference. *Eur. Phys. J. B* 94: 245.
- Koukos, P. and Bonvin, A. (2020). Integrative modelling of biomolecular complexes. *J. Mol. Biol.* 432: 2861–2881.
- Lane, T.J. (2023). Protein structure prediction has reached the single-structure Frontier. *Nat. Methods* 20: 1–4.
- Li, Y., Wu, J., Tedrake, R., Tenenbaum, J.B., and Torralba, A. (2018). *Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids*, arXiv preprint arXiv:1810.01566.
- Lin, X., Qi, Y., Latham, A.P., and Zhang, B. (2021). Multiscale modeling of genome organization with maximum entropy optimization. *J. Chem. Phys.* 155: 010901.
- Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379: 1123–1130.
- Lincoff, J., Haghighatlari, M., Krzeminski, M., Teixeira, J.M., Gomes, G.-N.W., Gradinaru, C.C., Forman-Kay, J.D., and Head-Gordon, T. (2020). Extended experimental inferential structure determination method in determining the structural ensembles of disordered protein states. *Commun. Chem.* 3: 74.
- Loconte, V., Singla, J., Li, A., Chen, J.-H., Ekman, A., McDermott, G., Sali, A., Le Gros, M., White, K.L., and Larabell, C.A. (2022). Soft X-ray tomography to map and quantify organelle interactions at the mesoscale. *Structure* 30: 510–521.
- Loconte, V., Chen, J.-H., Vanslebrouck, B., Ekman, A.A., McDermott, G., Le Gros, M.A., and Larabell, C.A. (2023). Soft X-ray tomograms provide a structural basis for whole-cell modeling. *FASEB J.* 37: e22681.
- MacCallum, J.L., Perez, A., and Dill, K.A. (2015). Determining protein structures by combining semi-reliable data with atomistic physical models by bayesian inference. *Proc. Natl. Acad. Sci. U.S.A.* 112: 6985–6990.

- Madin, O.C., Boothroyd, S., Messerly, R.A., Fass, J., Chodera, J.D., and Shirts, M.R. (2022). Bayesian-inference-driven model parametrization and model selection for 2CLJQ fluid models. *J. Chem. Inf. Model.* 62: 874–889.
- Maritan, M., Autin, L., Karr, J., Covert, M.W., Olson, A.J., and Goodsell, D.S. (2022). Building structural models of a whole mycoplasma cell. *J. Mol. Biol.* 434: 167351.
- Meng, L., Wang, C., Shi, Y., and Luo, Q. (2021). Si-C is a method for inferring superresolution intact genome structure from single-cell Hi-C data. *Nat. Commun.* 12: 4369.
- Mintseris, J. and Gygi, S.P. (2020). High-density chemical cross-linking for modeling protein interactions. *Proc. Natl. Acad. Sci. U.S.A.* 117: 93–102.
- Misteli, T. (2020). The self-organizing genome: principles of genome architecture and function. *Cell* 183: 28–45.
- Monroe, J.I. and Shen, V.K. (2022). Learning efficient, collective Monte Carlo moves with variational autoencoders. *J. Chem. Theory Comput.* 18: 3622–3636.
- Mosalaganti, S., Obarska-Kosinska, A., Siggel, M., Taniguchi, R., Turoňová, B., Zimmerli, C.E., Buczak, K., Schmidt, F.H., Margiotta, E., Mackmull, M.-T., et al. (2022). AI-based structure prediction empowers integrative structural analysis of human nuclear pores. *Science* 376: eabm9506.
- Nagano, T., Lubling, Y., Stevens, T.J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E.D., Tanay, A., and Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature* 502: 59–64.
- Noé, F., Olsson, S., Köhler, J., and Wu, H. (2019). Boltzmann generators: sampling equilibrium states of many-body systems with deep learning. *Science* 365: eaaw1147.
- Noid, W.G. (2013). Perspective: coarse-grained models for biomolecular systems. *J. Chem. Phys.* 139: 09B201 1.
- Ou, H.D., Phan, S., Deerinck, T.J., Thor, A., Ellisman, M.H., and O'Shea, C.C. (2017). ChromEMT: visualizing 3D chromatin structure and compaction in interphase and mitotic cells. *Science* 357: eaag0025.
- O'Reilly, F.J., Xue, L., Graziadei, A., Sinn, L., Lenz, S., Tegunov, D., Blötz, C., Singh, N., Hagen, W.J., Cramer, P., et al. (2020). In-cell architecture of an actively transcribing translating expressome. *Science* 369: 554–557.
- Pak, A.J. and Voth, G.A. (2018). Advances in coarse-grained modeling of macromolecular complexes. *Curr. Opin. Struct. Biol.* 52: 119–126.
- Pesce, F. and Lindorff-Larsen, K. (2021). Refining conformational ensembles of flexible proteins against small-angle x-ray scattering data. *Biophys. J.* 120: 5124–5135.
- Potrzebowski, W., Trehwella, J., and Andre, I. (2018). Bayesian inference of protein conformational ensembles from limited structural data. *PLoS Comput. Biol.* 14: e1006641.
- Raveh, B., Sun, L., White, K.L., Sanyal, T., Tempkin, J., Zheng, D., Bharath, K., Singla, J., Wang, C., Zhao, J., et al. (2021). Bayesian metamodeling of complex biological systems across varying representations. *Proc. Natl. Acad. Sci. U.S.A.* 118: e2104559118.
- Rieping, W., Habeck, M., and Nilges, M. (2005). Inferential structure determination. *Science* 309: 303–306.
- Rodriguez, A. and Schmidler, S.C. (2014). Bayesian protein structure alignment. *Ann. Appl. Stat.* 8: 2068.
- Rosenthal, M., Bryner, D., Huffer, F., Evans, S., Srivastava, A., and Neretti, N. (2019). Bayesian estimation of three-dimensional chromosomal structure from single-cell HiC data. *J. Comput. Biol.* 26: 1191–1202.
- Rout, M.P. and Sali, A. (2019). Principles for integrative structural biology studies. *Cell* 177: 1384–1403.
- Russel, D., Lasker, K., Webb, B., Velazquez-Muriel, J., Tjioe, E., Schneidman-Duhovny, D., Peterson, B., and Sali, A. (2012). Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.* 10: e1001244.
- Sae-Lee, W., McCafferty, C.L., Verbeke, E.J., Havugimana, P.C., Papoulas, O., McWhite, C.D., Houser, J.R., Vanuytsel, K., Murphy, G.J., Drew, K., et al. (2022). The protein organization of a red blood cell. *Cell Rep.* 40: 111103.
- Sali, A. (2021). From integrative structural biology to cell biology. *J. Biol. Chem.* 296: 100743-1–100743-15.
- Saltzberg, D.J., Viswanath, S., Echeverria, I., Chemmama, I.E., Webb, B., and Sali, A. (2021). Using Integrative Modeling Platform to compute, validate, and archive a model of a protein complex structure. *Protein Sci.* 30: 250–261.
- Schaffer, L.V. and Ideker, T. (2021). Mapping the multiscale structure of biological systems. *Cell Syst.* 12: 622–635.
- Schöberl, M., Zabarar, N., and Koutsourelakis, P.-S. (2017). Predictive coarse-graining. *J. Comput. Phys.* 333: 49–77.
- Shekhar, M., Terashi, G., Gupta, C., Sarkar, D., Debussche, G., Sisco, N.J., Nguyen, J., Mondal, A., Vant, J., Fromme, P., et al. (2021). CryoFold: determining protein structures and data-guided ensembles from cryo-EM density maps. *Matter* 4: 3195–3216.
- Shevchuk, R. and Hub, J.S. (2017). Bayesian refinement of protein structures and ensembles against SAXS data using molecular dynamics. *PLoS Comput. Biol.* 13: e1005800.
- Sieben, C., Douglass, K.M., Guichard, P., and Manley, S. (2018). Super-resolution microscopy to decipher multi-molecular assemblies. *Curr. Opin. Struct. Biol.* 49: 169–176.
- Sigal, Y.M., Zhou, R., and Zhuang, X. (2018). Visualizing and discovering cellular structures with super-resolution microscopy. *Science* 361: 880–887.
- Singla, J., McClary, K.M., White, K.L., Alber, F., Sali, A., and Stevens, R.C. (2018). Opportunities and challenges in building a spatio-temporal multi-scale model of the human pancreatic β cell. *Cell* 173: 11–19.
- Singla, J., Burdsall, K., Cantrell, B., Halsey, J.R., McDowell, A., McGregor, C., Mittal, S., Stevens, R.C., Su, S., Thomopoulos, A., et al. (2022). A new visual design language for biological structures in a cell. *Structure* 30: 485–497.
- Spill, Y.G., Karami, Y., Maisonneuve, P., Wolff, N., and Nilges, M. (2021). Automatic Bayesian weighting for SAXS data. *Front. Mol. Biosci.* 8: 671011.
- Stein, R.A. and Mchaourab, H.S. (2022). Speach AF: sampling protein ensembles and conformational heterogeneity with AlphaFold2. *PLoS Comput. Biol.* 18: e1010483.
- Stevens, T.J., Lando, D., Basu, S., Atkinson, L.P., Cao, Y., Lee, S.F., Leeb, M., Wohlfahrt, K.J., Boucher, W., O'Shaughnessy-Kirwan, A., et al. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* 544: 59–64.
- Terwilliger, T.C., Poon, B.K., Afonine, P.V., Schlicksup, C.J., Croll, T.I., Millán, C., Richardson, J.S., Read, R.J., and Adams, P.D. (2022). Improved AlphaFold modeling with implicit experimental information. *Nat. Methods* 19: 1–7.
- Theobald, D.L. and Wuttke, D.S. (2008). Accurate structural correlations from maximum likelihood superpositions. *PLoS Comput. Biol.* 4: e43.
- Topf, M., Lasker, K., Webb, B., Wolfson, H., Chiu, W., and Sali, A. (2008). Protein structure fitting and refinement guided by cryo-EM density. *Structure* 16: 295–307.
- Tunyasuvunakool, K. (2022). The prospects and opportunities of protein structure prediction with AI. *Nat. Rev. Mol. Cell Biol.* 23: 445–446.
- Ullanat, V., Kasukurthi, N., and Viswanath, S. (2022). PriSM: precision for integrative structural models. *Bioinformatics* 38: 3837–3839.

- Vakili, N. and Habeck, M. (2021). Bayesian random tomography of particle systems. *Front. Mol. Biosci.* 8: 658269.
- Vallat, B., Webb, B., Fayazi, M., Voinea, S., Tangmunarunkit, H., Ganesan, S.J., Lawson, C.L., Westbrook, J.D., Kesselman, C., Sali, A., et al. (2021). New system for archiving integrative structures. *Acta Crystallogr. Section D Struct Biol.* 77: 1486–1496.
- Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., et al. (2022). Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 50: D439–D444.
- Villa, E. and Lasker, K. (2014). Finding the right fit: chiseling structures out of cryoelectron microscopy maps. *Curr. Opin. Struct. Biol.* 25: 118–125.
- Viswanath, S. and Sali, A. (2019). Optimizing model representation for integrative structure determination of macromolecular assemblies. *Proc. Natl. Acad. Sci. U.S.A.* 116: 540–545.
- Viswanath, S., Chemmama, I.E., Cimermancic, P., and Sali, A. (2017). Assessing exhaustiveness of stochastic sampling for integrative modeling of macromolecular structures. *Biophys. J.* 113: 2344–2353.
- Voelz, V.A., Ge, Y., and Raddi, R.M. (2021). Reconciling simulations and experiments with BICePs: a review. *Front. Mol. Biosci.* 8: 661520.
- von Ardenne, B., Mechelke, M., and Grubmüller, H. (2018). Structure determination from single molecule X-ray scattering with three photons per image. *Nat. Commun.* 9: 2375.
- Vuillemot, R. and Jonić, S. (2021). Combined bayesian and normal mode flexible fitting with hamiltonian monte carlo sampling for cryo electron microscopy. In: *29th European signal processing conference (EUSIPCO) 2021*. IEEE, pp. 1211–1215.
- Wang, S., Xu, J., and Zeng, J. (2015). Inferential modeling of 3D chromatin structure. *Nucleic Acids Res.* 43: e54.
- Wang, W., Xu, M., Cai, C., Miller, B.K., Smidt, T., Wang, Y., Tang, J., and Gomez-Bombarelli, R. (2022). Generative coarse-graining of molecular conformations. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (Eds.), *Proceedings of the 39th international conference on machine learning. Proceedings of machine learning research volume 162*, pp. 23213–23236.
- Xie, Q., Han, C., Jin, V., and Lin, S. (2022). HiCImpute: a Bayesian hierarchical model for identifying structural zeros and enhancing single cell Hi-C data. *PLoS Comput. Biol.* 18: e1010129.
- Yip, K.M., Fischer, N., Paknia, E., Chari, A., and Stark, H. (2020). Atomic-resolution protein structure determination by cryo-EM. *Nature* 587: 157–161.
- Yperman, K., Wang, J., Eeckhout, D., Winkler, J., Vu, L.D., Vandorpe, M., Grones, P., Mylle, E., Kraus, M., Merceron, R., et al. (2021). Molecular architecture of the endocytic TPLATE complex. *Sci. Adv.* 7: eabe7999.
- Zhang, K. and Frank, A.T. (2021). Probabilistic modeling of RNA ensembles using NMR chemical shifts. *J. Phys. Chem. B* 125: 9970–9978.
- Zhong, E.D., Bepler, T., Berger, B., and Davis, J.H. (2021). Cryodrgn: reconstruction of heterogeneous cryo-em structures using neural networks. *Nat. Methods* 18: 176–185.
- Zhong, X., Zhao, J., and Sun, L. (2022). Integrative modeling of the cell. *Acta Biochim. Biophys. Sin.* 54: 1213–1221.
- Ziemianowicz, D.S. and Kosinski, J. (2022). New opportunities in integrative structural modeling. *Curr. Opin. Struct. Biol.* 77: 102488.