

# Bayesian Mechanistic Inference, Statistical Mechanics, and a New Era for Monte Carlo

Daniel M. Zuckerman\* and August George

Cite This: *J. Chem. Theory Comput.* 2024, 20, 2971–2984

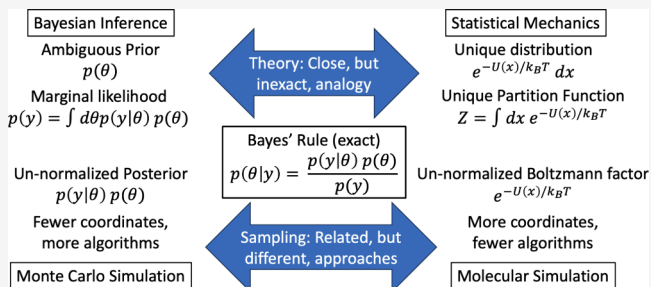
Read Online

ACCESS |

Metrics &amp; More

Article Recommendations

**ABSTRACT:** On the one hand, much of computational chemistry is concerned with “bottom-up” calculations which elucidate observable behavior starting from exact or approximated physical laws, a paradigm exemplified by typical quantum mechanical calculations and molecular dynamics simulations. On the other hand, “top down” computations aiming to formulate mathematical models consistent with observed data, e.g., parametrizing force fields, binding or kinetic models, have been of interest for decades but recently have grown in sophistication with the use of Bayesian inference (BI). Standard BI provides an estimation of parameter values, uncertainties, and correlations among parameters. Used for “model selection,” BI can also distinguish between model structures such as the presence or absence of individual states and transitions. Fortunately for physical scientists, BI can be formulated within a statistical mechanics framework, and indeed, BI has led to a resurgence of interest in Monte Carlo (MC) algorithms, many of which have been directly adapted from or inspired by physical strategies. Certain MC algorithms—notably procedures using an “infinite temperature” reference state—can be successful in a 5–20 parameter BI context which would be unworkable in molecular spaces of  $10^3$  coordinates and more. This Review provides a pedagogical introduction to BI and reviews key aspects of BI through a physical lens, setting the computations in terms of energy landscapes and free energy calculations and describing promising sampling algorithms. Statistical mechanics and basic probability theory also provide a reference for understanding intrinsic limitations of Bayesian inference with regard to model selection and the choice of priors.



## INTRODUCTION

Bayesian inference (BI) has become a standard tool for addressing “inverse problems” in many fields of science, including chemistry and biophysics;<sup>1–5</sup> applications have included calorimetry analysis,<sup>6–8</sup> analysis of single-molecule data,<sup>9,10</sup> and development of Markov state models from molecular dynamics data.<sup>11,12</sup> BI takes observed data as input, potentially from different measurement types, and outputs a “posterior” probability distribution of parameters for a preset mathematical model that is consistent with the data and any prior assumptions. This multidimensional distribution effectively scores different parameter choices, yielding not only most likely values but also possible ranges of parameters (uncertainty ranges, in effect) as well as the correlation structure among parameters. To the extent that suitable uncertainties are placed on experimentally derived inputs, BI automatically “propagates” uncertainty without assuming linear relationships among variables. Bayesian calculations additionally can perform “model selection,” which entails a quantitative comparison of candidate models<sup>1,7,13</sup>—each defined by a set of equations and parameters. Furthermore, BI can naturally be extended for nested and hierarchical models.<sup>1</sup>

There are important but inexact analogies between the probabilistic framework of BI and equilibrium statistical mechanics (Figure 1). The task of model selection, for example, requires estimation of integrals that sum over probabilities: these are akin to partition functions.<sup>7</sup> Procedurally, the BI process often uses Markov-chain Monte Carlo (MCMC), frequently borrowing methods motivated by physical science, such as parallel tempering and annealing.<sup>4,14</sup> There are also modern molecular sampling methods that employ an implicitly Bayesian framework.<sup>15</sup>

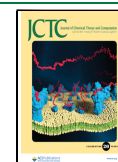
This Review aims to provide an introduction to the connections between statistical mechanics and Bayesian inference. Multidimensional probability theory provides a common basis for the two frameworks, which we discuss in a pedagogical way to demonstrate the parallels and also subtle

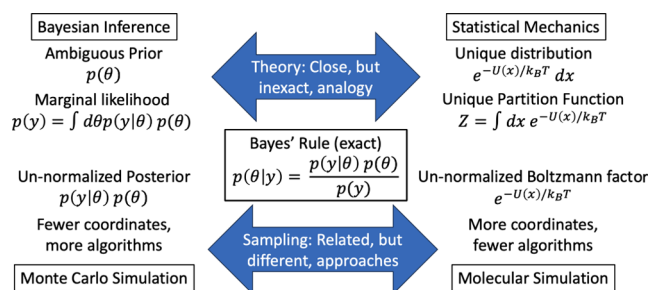
Received: January 5, 2024

Revised: March 25, 2024

Accepted: March 25, 2024

Published: April 11, 2024





**Figure 1.** Relationships between Bayesian inference and statistical mechanics for theory and simulation. Although (exact) probability theory underpins both approaches, the physical basis of statistical mechanics yields unique calculations in contrast to the intrinsic uncertainty embedded in the Bayesian prior. On the other hand, the much lower dimensional coordinate (parameter) space of Bayesian inference enables use of more diverse sampling approaches, including some developed for physical systems.

distinctions. The intrinsically subjective priors of BI are addressed, along with the related issue of parameter representation. We discuss MCMC sampling methods that employ physical ideas as well as new approaches based on machine learning.

We hope this Review will provide readers from a physical science background the confidence to engage productively in Bayesian studies, as well as to critically examine BI data using a physical lens. Because many physical scientists have extensive experience with sampling challenges, they should be well-positioned to perform high-quality Bayesian studies. In this paper, the introductory material and the range of topics covered should be of value to students and more advanced researchers.

## ■ PROBABILITY THEORY AND BASIC BAYESIAN INFERENCE FORMALISM

We review the key probability theory useful for understanding Bayesian inference, both as a refresher and to establish nomenclature. For now, the essential structures are joint and conditional probability densities, as well as marginal densities. We begin with the simplest case of two dimensions for illustration and then generalize to higher dimensions.

We will use the terms “distribution,” “probability distribution,” and “probability density” interchangeably.

### Elements of Two-Dimensional Probability Theory.

Bayesian inference relies on a division of variables into two groups, parameters, and data, so two-dimensional theory is a natural starting point. We consider the (“random”) variables  $x$  and  $y$  which are distributed according to the *joint probability density*  $p(x, y)$ . As always, a probability density must be normalized, meaning that

$$\int dx dy p(x, y) = \int_{-\infty}^{+\infty} dx \int_{-\infty}^{+\infty} dy p(x, y) = 1 \quad (1)$$

where integrals written without limits implicitly mean integration over the full domain of real numbers  $(-\infty, +\infty)$ , in this case for both  $x$  and  $y$ .

A *marginal probability density* or simply “marginal” describes the overall probability of  $x$  or  $y$  alone, which is computed by integrating over all of the probability density in the excluded coordinate. Thus, the marginal for  $x$ , denoted  $p_x$ , must integrate over all possible  $y$  values for any  $x$ , and we have

$$p_x(x) = \int dy p(x, y) \quad (2)$$

Note that the marginal is (already) normalized because if we integrate  $\int dx p_x(x)$  we arrive directly at the normalization condition for the joint distribution (1). The marginal for  $y$ , denoted  $p_y(y)$  is derived in an exactly analogous fashion and also is normalized.

Joint and marginal probabilities are shown for the two discretized examples in Figure 2. Analogous to  $x$  and  $y$

### (a) Protein conformations

|            | Helix-RotA | Helix-RotB | Coil-RotA | Coil-RotB | [Marginal] |
|------------|------------|------------|-----------|-----------|------------|
| Closed     | 0.05       | 0.40       | 0.25      | 0.10      | 0.80       |
| Open       | 0.01       | 0.15       | 0.04      | 0.0       | 0.20       |
| [Marginal] | 0.06       | 0.55       | 0.29      | 0.10      |            |

### (b) Coin tosses – fair and weighted

|            | HH    | HT    | TH    | TT    | [Marginal] |
|------------|-------|-------|-------|-------|------------|
| P(H) = 0.5 | 0.225 | 0.225 | 0.225 | 0.225 | 0.90       |
| P(H) = 0.1 | 0.001 | 0.009 | 0.009 | 0.081 | 0.10       |
| [Marginal] | 0.226 | 0.234 | 0.234 | 0.306 |            |

**Figure 2.** Joint probabilities and marginals: the discrete case. (a) An idealized joint probability distribution over “macro” (closed/open) and “micro” features (helix/coil-RotA/RotB) of protein conformations, where a certain group of residues may be in an  $\alpha$  helix or not and one residue has two alternative rotameric conformations (RotA/RotB). All possibilities must sum to 1, but we can also consider the marginal distributions by summing horizontally over microfeatures for a given macroconformation or vertically over macroconformations for each set of microfeatures. (b) A hypothetical joint distribution of coin types (based on probability of heads  $P(H)$ ) and possible coin tosses. Going across each row, the relative probabilities for each toss pair are as expected, but the absolute probabilities are based on the imposed marginal of coin types. This example takes as prior knowledge that 90% of coins are fair with  $P(H) = 0.5$  and 10% are weighted with  $P(H) = 0.1$ . Although formally identical, the analogy between joint distributions (a) and (b) is imperfect on physical grounds because (a) describes a true joint distribution of possibilities for a single experimental condition, whereas (b) encompasses two different types of coin-toss “experiments”.

integrations for the respective marginals, in the discrete case, simple sums of probability are performed in horizontal and vertical directions in the tables.

The *conditional probability density*  $p(x | y)$  describes the distribution of  $x$  values (only) for a fixed value of  $y$  which is considered a “condition.”

$$p(x|y) = \frac{p(x, y)}{\int dx p(x, y)} = \frac{p(x, y)}{p_y(y)} \quad (3)$$

where the integral in the denominator ensures normalization, and the second equality uses the definition of the  $p_y$  marginal analogous to (2). In a precise analogy, the conditional distributions for a discrete system (Figure 2) can be derived by renormalizing any row or column based on the corresponding marginal.

We can derive the key Bayesian relation by rearranging (3) to yield  $p(x, y) = p(x | y) p_y(y)$ . Combining this equality with the analogous relation based on swapping  $x$  and  $y$ , we have

$$p(x|y) = \frac{p(y | x) p_x(x)}{p_y(y)} \quad (4)$$

which is Bayes' rule.<sup>1</sup> This is an exact mathematical result that is exploited, and adjusted, in Bayesian inference. In words, the "opposing" conditional probabilities  $p(x | y)$  and  $p(y | x)$  are proportional to one another, after correction by the ratio of marginals  $p_x/p_y$ . Note that the denominator depends only on  $y$  and so does not affect relative values of  $x$ .

Looking ahead to the practical goal of parameter inference from data, if  $y$  represents the observed data and  $x$  represents a model parameter, then the Bayes theorem (4) tells us how to calculate the so-called "posterior" distribution  $p(x | y)$  of possible parameter values given the data. This posterior is evidently proportional to the likelihood  $p(y | x)$  of observing the data given the parameter, a function we will examine in detail below. Likewise, we will delve into the other functions, especially the marginal  $p_x$  which is used to represent prior information in Bayesian inference.

A final important concept concerns *independence vs correlation*. In most cases of physical interest, two variables  $x$  and  $y$  will be correlated and thus  $p(x,y) \neq p_x(x)p_y(y)$ . If instead, equality holds for every possible  $(x,y)$  pair, that means the two variables are independent, because the value of one variable does not affect the distribution of the other.

**Higher Dimensional Probability theory.** In generalizing the preceding framework to higher dimensions, we maintain a division of variables into two groups because the separation between data and parameters is fundamental in Bayesian inference. We also adopt the standard  $\theta$  notation for "parameters", which are not variables in a physical sense (because in nature, they should have unique values for a correct physical model) but are the key variables whose distribution we would like to know, based on their consistency with available data and prior knowledge.

Thus, we define a vector for the parameter variables  $\vec{\theta} = (\theta_1, \theta_2, \dots, \theta_m)$  to take the place of the scalar variable  $x$  from the two-dimensional case, and  $\vec{y} = (y_1, y_2, \dots, y_n)$  representing multidimensional data instead of the scalar  $y$ . Note that the dimensionalities of sets  $m$  and  $n$  generally will differ. For integrals we will use the shorthand  $d\vec{\theta} = d\theta_1, d\theta_2, \dots, d\theta_m$  and  $d\vec{y} = dy_1, dy_2, \dots, dy_n$ .

Maintaining a probability theory perspective (leaving inference for later), there is no fundamental difference between the  $\vec{\theta}$  and  $\vec{y}$  variables, and we assume that a joint distribution  $p(\vec{\theta}, \vec{y})$  is a standard, normalized probability distribution. That is,

$$\int d\vec{\theta} d\vec{y} p(\vec{\theta}, \vec{y}) = 1 \quad (5)$$

With the overall dimensionality  $m + n$ , we can construct marginals of any dimensionality up to  $m + n - 1$ . An example one-dimensional (1D) marginal is

$$p_1(\theta_1) = \int d\theta_2 d\theta_3 \dots d\theta_m d\vec{y} p(\vec{\theta}, \vec{y}) \quad (6)$$

and any other 1D marginal can be written as an integral of the joint distribution over all coordinates not marginalized. Any such marginal is properly normalized, as can be seen by integrating (6) over  $\theta_1$  and comparing to (5). Extending the marginal concept, we can define an example (normalized) two-dimensional (2D) marginal via

$$p_{1,2}(\theta_1, \theta_2) = \int d\theta_3 d\theta_4 \dots d\theta_m d\vec{y} p(\vec{\theta}, \vec{y}) \quad (7)$$

Two-dimensional marginals will prove very useful in analyzing BI data. Any other variables can be chosen for marginalization by excluding the chosen set from the integration.

Two other marginals of interest for BI theory consider the overall sets of parameters and data variables, namely,

$$p_{\vec{\theta}}(\vec{\theta}) = \int d\vec{y} p(\vec{\theta}, \vec{y}) \quad (8)$$

$$p_{\vec{y}}(\vec{y}) = \int d\vec{\theta} p(\vec{\theta}, \vec{y}) \quad (9)$$

These marginals, which are useful in defining important conditional probabilities below, are abstract entities and are not easily amenable to intuition. But for completeness, we note that formally,  $p_{\vec{\theta}}(\vec{\theta})$  is the distribution of parameters considering all possible sets of data, and  $p_{\vec{y}}(\vec{y})$  is the distribution of data sets accounting for all possible systems. We emphasize that in a practical Bayesian inference process, detailed below, only a single data set  $\vec{y}$  is considered, corresponding to the data actually observed. As a minor point, although the subscripts on the marginals (8) and (9) may seem redundant, we include them to denote marginals defined strictly from probability theory.

We can also define conditional probabilities using the marginals

$$p(\vec{\theta} | \vec{y}) = \frac{p(\vec{\theta}, \vec{y})}{p_{\vec{y}}(\vec{y})} \quad (10)$$

$$p(\vec{y} | \vec{\theta}) = \frac{p(\vec{\theta}, \vec{y})}{p_{\vec{\theta}}(\vec{\theta})} \quad (11)$$

Finally, we write down the key precursor equation for Bayesian inference, the multidimensional Bayes' theorem, derived by equating the joint distributions in (10) and (11).

$$p(\vec{\theta} | \vec{y}) = \frac{p(\vec{y} | \vec{\theta}) p_{\vec{\theta}}(\vec{\theta})}{p_{\vec{y}}(\vec{y})} \quad (12)$$

We emphasize that this equation is a direct consequence of the definitions of basic probability theory.

When (12) is put to use for statistical inference of parameters, we must tread carefully to avoid confusion. Ambiguous nomenclature arises because the marginal  $p_{\vec{\theta}}$  of the full joint distribution is replaced by the "prior" distribution, as discussed below. In turn, the main distribution of interest becomes the conditional probability  $p(\vec{\theta} | \vec{y})$ , known as the "posterior" distribution, also discussed at length below.

**Bayesian Inference as a Bridge from Theory to Useful Information.** The goal of Bayesian inference is to derive the distribution of parameters  $\vec{\theta}$  consistent with the observed data  $\vec{y}$  and prior assumptions about the parameters. We now bridge between the exact probability theory discussed so far and the inference process, which necessarily involves assumptions. On the one hand, it is important to realize that (subjective) assumptions must enter Bayesian inference, so the sensitivity of the final results to any assumptions should be checked. On the other hand, the explicit inclusion of assumptions in Bayesian inference should not be taken as a weakness of the approach compared to other methods because any inference method will involve assumptions that may or may not be explicit.

**What Is a Model?** For physical scientists, the notion of a model is quite intuitive. A model is simply an equation or set of

equations describing the behavior of a system. In Bayesian inference, the model yields the numerical prediction  $\hat{y}_i(\vec{\theta})$  for measurement  $i$ , given model parameters  $\vec{\theta}$  and the experimental conditions for that measurement. A complete Bayesian model also makes explicit assumptions about the sources of uncertainty, such as the mathematical forms of measurement noise and systematic measurement bias, leading to a probabilistic description of data based on parameters.

As one example, consider the case of isothermal titration calorimetry (ITC), in which changes in heat are measured during a binding process.<sup>6,8</sup> For binding with 1:1 stoichiometry, in the simplest case, the mathematical model governing ITC data expresses a measured set of heat increments  $dQ_i$  in terms of the presumed known receptor and ligand concentrations, as well as in terms of the model thermodynamic parameters, i.e., the enthalpy of binding  $\Delta H$  and the dissociation constant  $K_d$ . That is, there is an equation that yields the set of predicted data values  $\hat{y}_i = dQ_i$  based on the known concentrations and on the model parameters  $\Delta H$  and  $K_d$ . Therefore, the two components of parameter vector  $\vec{\theta}$  will be  $\Delta H$  and  $K_d$  (or its log).

Although our primary interest will be in determining the physical parameters of interest, e.g.,  $\Delta H$  and  $K_d$  in the ITC setting, there are additional, unavoidable “nuisance” parameters. Because measurements always have noise, there will be a noise parameter, typically  $\sigma$  representing the standard deviation of Gaussian noise, assumed independent and identically distributed for each measurement, although in principle, other noise models may be used. Typically there are additional nuisance parameters. In ITC, for example, an additional parameter  $\Delta H_0$  represents heat evolved due to experimental imperfections, and a careful treatment accounts for the fact that concentrations cannot be measured perfectly and so these also should be treated as (partially) unknown model parameters.<sup>6,8</sup>

Mathematically, then, the set of parameters  $\vec{\theta}$  includes both the real parameters of interest and nuisance parameters. As we will see below, the BI distribution of nuisance parameters sometimes provides valuable lessons.

We should bear in mind that not every model will be appropriate for a given set of data and that even the best models will typically build in some unwarranted assumptions. For now, we will assume that some model, bad or good, has been chosen and examine the implications for Bayesian inference of the parameters of that model. If a model is inappropriate for the data, we can expect that a wide range of parameter values, possibly unphysical, will be evaluated to have a relatively high probability. There is indeed “signal” in the inferred parameter distribution about the validity of a given model compared to another, which we will discuss below when considering “model selection.”

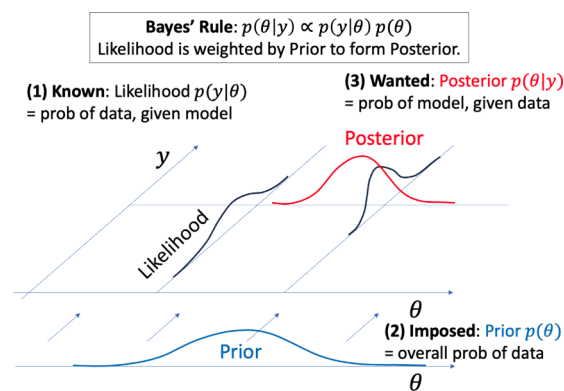
#### Recasting Bayes' Theorem to Define the Posterior.

We now rewrite Bayes' theorem very slightly to make it usable for inference. The issue with the original form (12) is that we have no idea what the marginal  $p_{\vec{\theta}}(\vec{\theta})$  should be nor any unbiased way to estimate it. Instead, with the goal of performing parameter inference where we wish to explicitly build in any assumptions about the parameters, the marginal of  $\vec{\theta}$  from (12) is replaced with an assumed *prior distribution*  $\text{Prior}(\vec{\theta})$ . We also adopt a more direct notation for the posterior and likelihood, namely,  $\text{Post}(\vec{\theta} | \vec{y}) = p(\vec{\theta} | \vec{y})$  and  $\text{Like}(\vec{y} | \vec{\theta}) = p(\vec{y} | \vec{\theta})$ .

The desired posterior distribution of parameters  $\vec{\theta}$  consistent with the (single) set of data  $\vec{y}$  and prior assumptions is therefore

$$\text{Post}(\vec{\theta} | \vec{y}) = \frac{\text{Like}(\vec{y} | \vec{\theta}) \text{Prior}(\vec{\theta})}{p_{\vec{y}}(\vec{y})} \propto \text{Like}(\vec{y} | \vec{\theta}) \text{Prior}(\vec{\theta}) \quad (13)$$

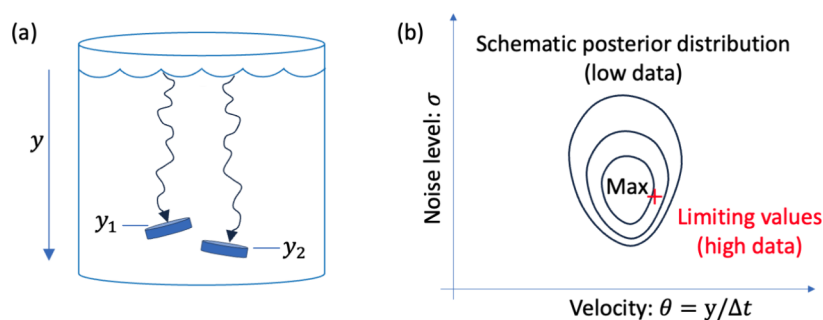
Two important points should be understood about (13), which can be visualized with help from Figure 3. (i) Even though the



**Figure 3.** Components of Bayesian inference. Bayes' rule is an exact, basic result of probability theory, but Bayesian inference is a way of combining previous knowledge about a parameter of interest  $\theta$ , encoded in the Prior distribution, with data values  $y_1, y_2, \dots$ . The Likelihood function effectively weights different parameter ( $\theta$ ) values based on the data. When prior assumptions and data likelihood are combined together, the un-normalized Posterior distribution (“after data”) is obtained.

prior distribution plays the formal role of the marginal in Bayes theorem (12)—more on this below—in actuality, the prior is imposed and not computed from a joint distribution of  $\vec{y}$  and  $\vec{\theta}$ . The logic for understanding (13) nevertheless is straightforward: the posterior probability of a given parameter set  $\vec{\theta}$  depends on the probability that  $\vec{\theta}$  occurs in the first place (the prior) weighting the conditional probability of the data  $\vec{y}$  once the parameters have been selected (the likelihood). Thus, the posterior always depends on the choice of prior, though hopefully weakly as discussed below. (ii) Despite the appearance of  $p_{\vec{y}}(\vec{y})$  in the denominator in (13), in BI, we are interested in the (posterior) distribution of parameters for a single set of data  $\vec{y}$ , and we are not concerned with all possible data sets. Hence, for BI calculations,  $p_{\vec{y}}(\vec{y})$  acts simply as a normalization constant and can be ignored in most practical calculations, as implied by the proportionality symbol in (13). This is exactly analogous to sampling a Boltzmann-factor distribution without considering the partition function, a concept to which we shall return to later.

**The Likelihood at the Heart of Bayesian Inference.** As we can see from the fundamental Bayesian inference eq 13, to obtain the desired posterior distribution of physical (and nuisance) parameters  $\text{Post}(\vec{\theta} | \vec{y})$ , we require a function called the “likelihood,”  $\text{Like}(\vec{y} | \vec{\theta})$ . The likelihood is the dominant function in Bayesian inference (at least whenever there is enough data to do reliable parameter inference, see below). Fortunately, the likelihood is also the most intuitive component of the BI pipeline: it simply tells us the probability of observing a given set of data given the physical model, including the full set of experimental conditions and the noise



**Figure 4.** Bayesian inference for a simple linear model. (a) The parameter of interest is the terminal velocity of a coin dropped in water. Measured distance ( $y$ ) values are linearly related to the parameter of interest, velocity ( $\theta = y/\Delta t$ ), given the constant observation time  $\Delta t$ . (b) When a relatively small amount of data is available, the peak of the estimated posterior likely will be displaced from “true” values based on a large amount of data.

model. In many cases, the noise is taken to have a Gaussian distribution about the model-predicted value, which is what we shall assume.

The equation for the likelihood derives from considering one measurement at a time under the assumed fixed set of conditions  $\theta$ , with the addition of a *noise model*. We first note that each data point  $y_i$  has been generated under conditions  $\tilde{c}_i$ , for which the model makes a prediction—dependent on the parameters  $\theta$ —called  $\hat{y}_i = \hat{y}_i(\tilde{c}_i, \theta)$ . With a Gaussian noise model parametrized by variance  $\sigma^2$  (assuming a mean of zero), the overall likelihood is simply the product of the probabilities for each individual point. For  $N$  data points, we have

$$\text{Like}(\vec{y}|\vec{\theta}) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i - \hat{y}_i)^2 / 2\sigma^2} \quad (14)$$

This equation models the assumption that the noise for every data point is independent. Note that  $\sigma$  is considered part of the parameter set  $\vec{\theta}$ , and the physical parameters are implicit in  $\hat{y}_i$  values because those are determined from the parameter-dependent model.

**A Simple Example.** Several key lessons regarding Bayesian inference are revealed by a simple model that permits closed-form evaluation of the likelihood function. We consider terminal velocity measurements (Figure 4) where the velocity “parameter” of interest  $\theta$  is simply linear in the measured distance value  $y$ . The model-predicted data value is simply  $\hat{y} = \theta\Delta t$  based on the single velocity parameter  $\theta$ , and it is the same for all data points, i.e.,  $\hat{y}_i = \hat{y} = \theta\Delta t$ . Thus, the likelihood is

$$\text{Like}(\vec{y}|\theta, \sigma) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i - \theta\Delta t)^2 / 2\sigma^2} \quad (15)$$

for the vector of data values  $\vec{y} = (y_1, y_2, \dots, y_N)$ . Note that by definition the likelihood is normalized for integration over  $y_i$  values, and *not* for integration over parameters  $\theta$  or  $\sigma$ .

The product of Gaussians in (15) can be trivially expressed as the exponential of a sum of terms, which in turn can be rewritten in a revealing way by completing the square:

$$\text{Like}(\vec{y}|\theta, \sigma) = \left( \frac{1}{\sqrt{2\pi}\sigma} \right)^N \exp \left\{ -\frac{1}{2\sigma^2/N} [(\theta\Delta t - \bar{y})^2 + \sigma_y^2] \right\} \quad (16)$$

where  $\sigma_y^2 = \overline{y^2} - \bar{y}^2$  is the variance in the measurements, with  $\bar{y}$  and  $\overline{y^2}$  denoting the averages of the data measurements and their squares, respectively.

Several lessons emerge readily from this form (16). (i) As we would expect intuitively, for any pair of  $(\theta, \sigma)$  parameters, the likelihood is always larger if the data has lower variance, i.e., if the strictly positive  $\sigma_y^2$  is smaller. (ii) For any fixed value of the noise parameter  $\sigma$ , the  $\theta$  distribution is a simple Gaussian with mean  $\bar{y}/\Delta t$  and standard deviation  $\sigma/(\sqrt{N}\Delta t)$ . The linearity of the model makes the  $\theta$  distribution Gaussian, even though that form was applied to the  $y_i$  distribution; therefore, this is not generally true for any model type. Nevertheless, the narrowing of the distribution of  $\theta$  with  $\sqrt{N}$  is a statistically expected behavior. (iii) Finally, examining the  $\sigma$  behavior at the most likely  $\theta = \bar{y}/\Delta t$  value, we can estimate the most likely  $\sigma^2 \sim \sigma_y^2$ , which is the intuitive result that the Bayesian noise parameter mirrors the spread in the data. Putting this together with point (ii), we estimate the most likely overall variance for  $\theta$  as  $\sigma_\theta^2 \sim (\sigma_y/\Delta t)^2/N$ .

**Considering the Prior, and a Contrast with Statistical Mechanics.** Given our derivation in terms of conditional and marginal probabilities, a potentially confusing question that naturally arises is, “Does the prior impose itself as the marginal probability observed by sampling the posterior?” A superficial glance at the equations suggests that this is the case because the prior in (13) takes the place of the true mathematical marginal for the joint distribution that appears in (12). However, this is tantamount to stating that the posterior distribution (over model parameters) must be identical with the prior distribution (over the same marginal parameters), which certainly is not true.

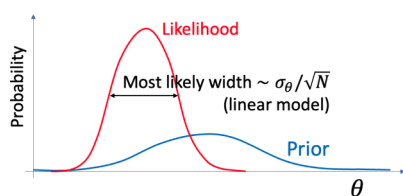
How should we understand the probabilistic logic here? The prior would indeed be the marginal if we were to integrate over (or sample) all possible data sets  $\vec{y}$ , but that is not done in Bayesian inference. BI considers only a single fixed set of data and samples parameter values consistent with that data. But the prior still affects the posterior, and this unavoidable effect of subjectivity is a contrast to statistical mechanics. Because the prior is always subjectively chosen (albeit with the aim of encoding empirical knowledge), the BI practitioner should check the degree to which their choice affects the posterior. Some sensitivity to the prior should be expected, and this does not invalidate BI findings but rather adds information to the inference results, which after all attempt to assess how precisely determined the parameters are.

It is important to realize that even a broad uniform prior, equivalent to setting a range for a parameter, intrinsically represents a subjective choice. The reason stems from the subjective nature of *parameter representation* and the multiplicity of ways to include the same parameter values. As

common examples, consider representing a rate constant or binding affinity. Both quantities are known to depend exponentially on (free) energetic quantities, so physical scientists may find the energy representations to be more “natural.” But a uniform distribution of a given parameter is very different from a uniform distribution of the log of the parameter, and there is no fundamental way to prefer one over the other in Bayesian inference. This is a contrast to statistical mechanics where the partition function is defined explicitly as an integral over Cartesian coordinates—and thus there is an unambiguous mathematical route to employing other coordinates using a Jacobian.<sup>16</sup>

A practical question that naturally arises for BI practitioners is how much data is needed so that the posterior is insensitive to the choice of prior? Superficially, one might assume that with more than 10 data points in the full likelihood (14) which is a product over single-data-point distributions, that the effects of the single prior distribution in (13) would be inconsequential. However, the situation depends on the system, data, and model—and the effects of the prior can vary significantly from parameter to parameter.<sup>8</sup> The practitioner should not assume that a certain number of data points provide immunity to the effects of the prior; instead, the sensitivity needs to be checked empirically by varying the prior. To frame this more theoretically, the real issue is whether the full likelihood is much narrower than the prior for a given parameter, and this is unknown until the distribution is sampled.

Figure 5 illustrates schematically the relative importance of prior and posterior. Although generically “enough” data should



**Figure 5.** Relative effects of likelihood and prior. Based on the available data, there is some fixed width in the likelihood function (red curve) for a parameter of interest  $\theta$  based on the most likely noise parameter(s). In a linear model, the width is simply related to the measured data variance, but more generally, there is some distribution of  $\theta$  values. The width of the likelihood will depend on the quality (noise) and quantity ( $N$ ) of the data. The prior, however, is independent of these factors and may be broader or narrower than the likelihood.

make the posterior relatively independent of the choice of the prior, there is no simple number of data points that guarantees this independence. In addition to the number of data points, the quality of the data, i.e., the level of noise or intrinsic variance) plays a key role. Thus, it is not correct, in general, to say that the prior plays the role of a single data point; this holds only if the width of the prior matches the intrinsic variance in the data. Based on this reasoning, it is sensible to quantify—on a problem-specific basis—the amount of information extracted from the data, compared to the prior.<sup>17</sup>

**Posterior Marginals and Uncertainty Ranges.** For many inference applications, the estimation of the uncertainty for parameters of interest is a key goal. The high-dimensional posterior  $\text{Post}(\vec{\theta}|\vec{y})$  embodies essentially all the information gathered in Bayesian inference including the “credibility

regions” (akin to confidence intervals) that quantify the ranges of likely parameter values, given the data and prior assumptions. These Bayesian uncertainty ranges are derived by marginalizing the posterior along the parameters of interest.

We introduce a special function name, Marg, for the marginals of the posterior distribution  $p(\vec{\theta}|\vec{y})$ , exemplified for one and two parameters as

$$\text{Marg}(\theta_1) = \int d\theta_2 d\theta_3 \dots d\theta_m \text{Post}(\vec{\theta}|\vec{y}) \quad (17)$$

$$\text{Marg}(\theta_1, \theta_2) = \int d\theta_3 d\theta_4 \dots d\theta_m \text{Post}(\vec{\theta}|\vec{y}) \quad (18)$$

which can be similarly defined for any subset of  $\vec{\theta}$  parameters. Note that these functions do not involve integration over  $\vec{y}$ , in contrast to (6) and (7) because  $\vec{y}$  is assumed fixed. Thus, more completely, we may write  $\text{Marg}(\theta_1) = \text{Marg}(\theta_1|\vec{y})$ , indicating that the marginals depend on the data.

The credibility regions, i.e., uncertainty ranges, are derived from the posterior marginals in a straightforward way: to obtain a 95% credibility region, for example, one takes the 2.5 and 97.5 percentile values. Because the posterior defined in (13) accounts for all possible parameter values via integration, weighted by the likelihood and prior, the credibility region also does. This contrasts with approximations sometimes employed within a maximum likelihood framework<sup>18</sup> which can fail to include information about all parameter values despite use of the likelihood function.<sup>19</sup>

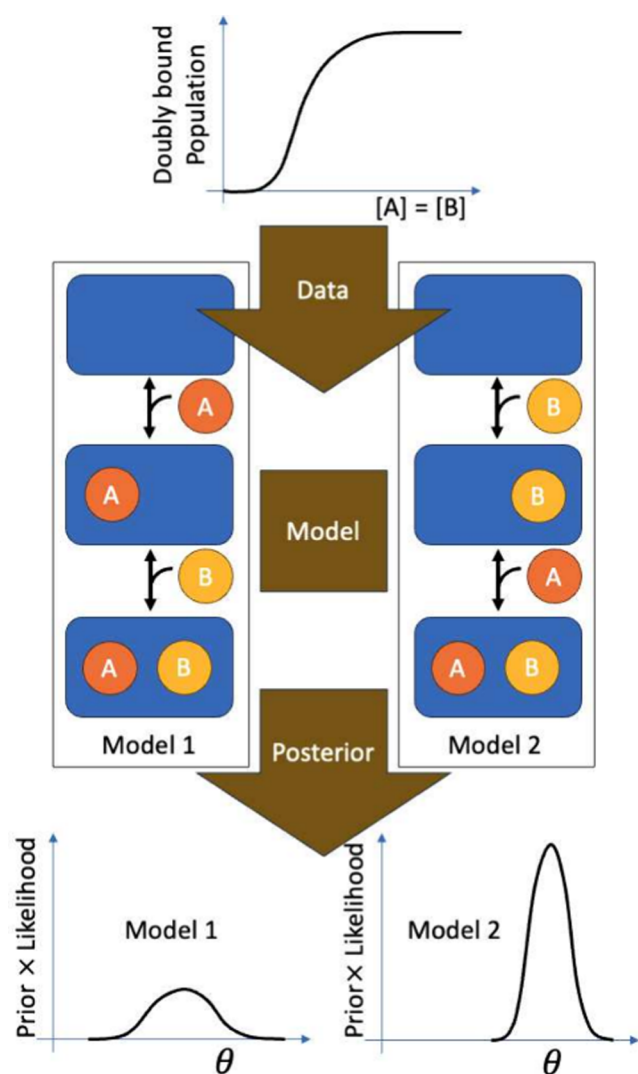
As a technical aside, note that (17) and (18) are not standard marginals if we consider  $\vec{\theta}$  and  $\vec{y}$  to be the full set of variables, as was assumed in the derivation of Bayes’ theorem (12). In the context of parameter inference, however, the real concern is the validity of different parameter sets based on a fixed data set  $\vec{y}$ , motivating the use of the posterior marginals (17) and (18).

## MODEL SELECTION AND THE CONNECTION TO FREE ENERGY CALCULATIONS

Even though parameter inference may be done systematically in a BI pipeline, by itself, knowing parameter values does not tell us whether the original model is valid. That is, once a certain model is assumed (e.g., possible states and transitions among them), Bayesian inference provides a systematic framework for evaluating possible parameter values for that model but does not automatically generate an overall assessment of the assumed model. The process of *model selection*<sup>13</sup> enables comparison among specific models, with some caveats that can be learned from statistical mechanics.

**Defining the model selection problem.** Assume we want to compare the ability of two models,  $\mathcal{M}_1$  and  $\mathcal{M}_2$ , for describing a single set of data  $\vec{y}$ . The models will be assumed to have different sets of parameters  $\vec{\theta}^{(1)}$  and  $\vec{\theta}^{(2)}$ , even though nominally some parameters could describe the same physical process in both models. For example, one model could describe a set of states with transition rate constants for parameters, and the other model could include some or all of the same states, but perhaps with fewer allowed transitions. In the schematic example of Figure 6, two different binding-event orders are compared based on a titration of both ligands simultaneously.

**The Marginal Likelihood in Model Selection.** A primary (though not unique) approach for assessing the overall suitability of a given model is via its *marginal likelihood*,



**Figure 6.** Model comparison in Bayesian inference. A hypothetical system in which a protein (blue) that binds two ligands (A and B) is analyzed based on experimental readout of the doubly bound protein population as the concentrations of A and B are simultaneously increased. A given set of experimental data (top) is analyzed using two different models (1 and 2). Each independent analysis produces its own posterior over the parameters specific to that model (bottom). Note that the posterior distributions,  $\text{Prior}(\theta) \times \text{Like}(\theta)$ , are not normalized over parameters, implying integration over parameters will yield different values known as marginal likelihoods.

also known as *Bayesian evidence*, which is a partition-function-like quantity that sums over the posterior probability associated with all possible parameter values. If the parameters  $\vec{\theta}$  and the prior distribution  $\text{Prior}(\vec{\theta}|\mathcal{M})$  are associated with the model  $\mathcal{M}$ , then given the set of data  $\vec{y}$ , the marginal likelihood for the model is

$$p(\vec{y}|\mathcal{M}) = \int d\vec{\theta} \text{Post}(\vec{\theta}|\vec{y}) = \int d\vec{\theta} \text{Like}(\vec{y}|\vec{\theta})\text{Prior}(\vec{\theta}|\mathcal{M}) \equiv Z_{\mathcal{M}} \quad (19)$$

The “marginal” nomenclature arises because if the integrand  $\text{Like}(\vec{y}|\vec{\theta}) \text{Prior}(\vec{\theta})$  is taken to be a true joint probability distribution—following (11)—then the integral over  $\theta$  represents the marginal  $p_{\vec{y}}$  as in (9).

The marginal likelihood (19) is analogous to a partition function (Figure 1) with the posterior playing the role of the Boltzmann factor. The defining integral represents the “sum” over all of the probability embodied in the posterior. Also in analogy to a partition function, the marginal likelihood  $p(\mathcal{M}|\vec{y})$  is the normalization constant for the posterior, given the fixed data set  $\vec{y}$ . Note that in (19) the model  $\mathcal{M}$  and data  $\vec{y}$  are fixed “constants” and not variables to be integrated over: in a physics context, these might be called (confusingly in the present context) “parameters”; they are analogous to  $N$ ,  $V$ ,  $T$  which are held fixed in a canonical partition function. Also, in the special but common case where the priors are uniform distributions that simply set ranges for parameters, the likelihood by itself is analogous to the Boltzmann factor.

If we want to estimate the relative probability of two models  $\mathcal{M}_1$  and  $\mathcal{M}_2$ , we can form the *Bayes factor*,<sup>13</sup> which is the simple ratio of the two marginal likelihoods.

$$\text{Bayes Factor} = \frac{p(\vec{y}|\mathcal{M}_1)}{p(\vec{y}|\mathcal{M}_2)} = \frac{Z_1}{Z_2} \quad (20)$$

where  $Z_i = Z(\mathcal{M}_i)$ . Note that the ratio (20) can be adjusted by the prior likelihood for the models themselves, if these differ;<sup>13</sup> here we assume there is no difference, so (20) is valid as written. Below, we consider computational strategies for computing Bayes factors, which are akin to partition function ratios, albeit of much lower dimensionality in typical cases of interest.

A natural, powerful feature of the comparison invoked by the Bayes factor (20) is that it automatically penalizes overfitting, i.e., models with more parameters.<sup>9</sup> To see this, consider the marginal likelihood (19), which is an integral over all parameters weighted by the (normalized) prior. On the one hand, models with more parameters may well exhibit better fits and hence higher likelihood values; however, given that the likelihood itself is essentially a Gaussian of the mean-squared error (14), there is a limit to how much the likelihood can increase once the mean-squared error is below the variance  $\sigma^2$ . On the other hand, as more parameters are included beyond what is physically appropriate to the data, we expect that the fraction of parameter space occupied by high-likelihood parameters to decrease significantly and overwhelm the slightly larger likelihood.

**Limitations of the Partition Function Analogy.** The partition function analogy has its limitations, and these could be of practical importance. We can consider whether the logic of the Bayes factor, i.e., of a ratio of integrals over the posterior distribution of parameters) is well founded. Although the marginal likelihood (19) is mathematically analogous to a partition function, this is not quite true in the physical sense. In a physical system, we wish to sum over all possible system configurations because they are expected to occur: the configurational integral of the partition function, after all, is equivalent to a time average for ergodic systems.<sup>16</sup> In contrast, in Bayesian inference, all possible sets of parameters are not expected to occur, and physically, we know there is only one true set if the model is correct. One can make a “robustness” argument, namely, that we expect more high-probability sets of parameters for a better model; further, given the noise intrinsic to data, one can argue we have no choice but to seek such robust models. However, it is not unreasonable to appeal to a maximum likelihood or maximum posterior perspective in

seeking the single best model, and this approach does play an important role in modern model selection.<sup>7</sup>

The well-defined statistical mechanics partition function also points us to a fundamental mathematical ambiguity in the marginal likelihood defined in (19). Whereas a partition function is strictly defined as an integral over Cartesian coordinates and a Jacobian correction is required for different variables,<sup>16</sup> no such fundamental integration measure can be defined for the marginal likelihood. Returning to our examples of having rate constants and/or binding affinities as BI parameters, we can equally describe these parameters by their direct values or by their energy-like logarithms. It seems impossible to consider either representation more fundamental, yet the resulting marginal likelihoods will be different. Thus, despite the appealing analogy with statistical mechanics, it seems that unavoidable ambiguities are intrinsic to Bayesian model selection.

This statistical inexactness of Bayesian model selection, as with Bayesian parameter inference, should not be interpreted as a reason to disregard the approach, but rather to proceed with caution. That is, always try more than one reasonable prior to testing the sensitivity of results to the prior.

## MONTE CARLO IN BAYESIAN INFERENCE

The main practical task in Bayesian inference is sampling the posterior distribution  $\text{Post}(\vec{\theta} | \vec{y})$  in order to obtain estimates for parameters  $\vec{\theta}$  given the data at hand  $\vec{y}$ . BI will not only provide (i) the most likely parameter values but also yields (ii) confidence intervals, technically known as “credibility regions” in BI, (iii) parameter correlations, and, when synthetic data is used, (iv) key information for experimental design.

Intuition about sampling energy landscapes applies directly to sampling Bayesian posteriors. Mathematically, the connection is straightforward: the posterior is a probability akin to a Boltzmann factor, and so the negative logarithm of the posterior is like an energy. Probability peaks in the posterior plane are directly analogous to energy minima. In one dimension, if we suppress the data  $\vec{y}$  for simplicity, in BI we want to sample  $\text{Post}(\theta)$  where  $\theta$  is the model parameter, and this is analogous to the role of the Boltzmann factor  $e^{-U(x)/k_B T}$  where  $x$  is the physical coordinate. The simplest energy landscape would have a single, smooth basin, which could readily be sampled; likewise, it would be easy to sample a posterior with a single peak, e.g., the Gaussian likelihood (16) from our simple example in combination with a uniform prior. In complex BI problems, not surprisingly, multimodal distributions can occur.

In principle, a standard Monte Carlo method such as the Metropolis–Hastings algorithm<sup>20</sup> can be used to sample any Bayesian posterior, whether it is simple or complicated. In one dimension, the Metropolis acceptance probability would be

$$P_{\text{accept}} = \min[1, \text{Post}(\theta_{\text{new}})/\text{Post}(\theta_{\text{old}})] \quad (21)$$

in analogy to the physical Metropolis probability  $\min[1, e^{-U(x_{\text{new}})/k_B T}/e^{-U(x_{\text{old}})/k_B T}]$  in the simple but typical case of a symmetric attempt distribution, where “new” and “old” denote the trial and previous configurations, respectively. In the most interesting and complex problems, however, simple approaches are not sufficient, and therefore, our discussion of MC methods focuses on more cutting-edge methods.

Although sampling the BI posterior is formally identical with equilibrium molecular simulation, the much lower dimension-

ality in BI dramatically alters the slate of algorithms that can be considered. Instead of trying to sample the Boltzmann-factor distribution  $e^{-U(\vec{x})/k_B T}$  where  $U$  is the potential energy and  $\vec{x}$  is the full set of configurational coordinates—typically  $10^5$  or more coordinates—we are sampling  $\text{Post}(\vec{\theta}) = \text{Post}(\vec{\theta} | \vec{y})$ , where typically there are often 10–100 components of  $\vec{\theta}$ , and  $\vec{y}$  is fixed data that is not sampled.

Here, we primarily restrict ourselves to Markov chain Monte Carlo (MCMC) algorithms, which are less formally called Monte Carlo (MC) in much of the physical sciences. As described in greater detail later, there are many classes of algorithms, such random walk Metropolis–Hastings,<sup>20</sup> Hamiltonian MC,<sup>7,21</sup> and ensemble methods that sample the posterior, each with their own performance characteristics. Of particular interest in this paper are “temperature”-based methods, as described below.

The relatively low dimensionality of Bayesian inference problems, compared to biomolecular systems, opens up the possibility to exploit high and effectively infinite temperature in MC sampling for BI. Although high-temperature-based sampling protocols are common for biomolecular systems,<sup>22,23</sup> there are intrinsic limitations in that arena, including the challenge of sampling an unfolded biomolecule and the potentially minimal overlap between distributions at different temperatures.<sup>24,25</sup> The authors are unaware of sampling protocols that employ infinite temperature, i.e., a uniform distribution in all coordinates) for biomolecular systems. Because of its importance in BI, “temperature”-based annealing is discussed below in greater detail.

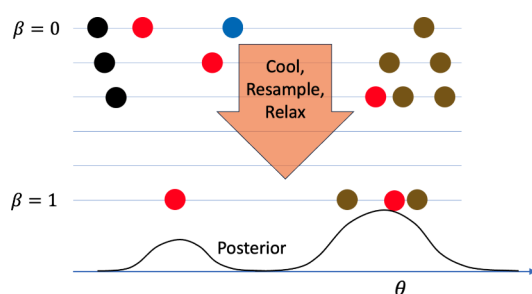
Finally, although the topic has not been investigated systematically to our knowledge, we expect that the choice of parameter representation could affect sampling. In a molecular system, one would expect internal coordinates to be more natural and effective for Monte Carlo sampling, as opposed to Cartesian coordinates.<sup>26,27</sup> Likewise, in Bayesian inference, it may prove more natural to consider a transformed coordinate, e.g., based on an exponential, logarithm, or trigonometric function.

**Annealing Algorithms.** Although temperature plays no direct role in Bayesian inference (even though experimental data being analyzed has been collected at finite temperature), the BI problem can usefully be reformulated to include an artificial temperature.<sup>14</sup> Instead of considering only the true posterior, we introduce an inverse temperature parameter  $\beta$  and consider distributions of the form

$$p_{\beta}(\vec{\theta}) \equiv [\text{Like}(\vec{y} | \vec{\theta})]^{\beta} \text{Prior}(\vec{\theta}) \quad (22)$$

With this definition,  $p_1(\vec{\theta}) = \text{Like}(\vec{y} | \vec{\theta}) \cdot \text{Prior}(\vec{\theta}) = \text{Post}(\vec{\theta})$  corresponds to the true posterior and  $p_0(\vec{\theta}) = \text{Prior}(\vec{\theta})$  represents the infinite temperature limit and yields the prior. The prior is typically a distribution that can be sampled exactly, such as a uniform distribution over a finite range for each parameter in  $\vec{\theta}$  or a multidimensional Gaussian.

Annealed importance sampling (AIS),<sup>14,28</sup> in physical terms, uses a valid high-temperature (low  $\beta$ ) sample—i.e., an ensemble of many systems—that is gradually cooled in order to obtain a representative sample at a lower temperature ( $\beta = 1$ ). See Figure 7. Cooling is gradual in the sense that finite temperature decreases are alternated with a constant-temperature simulation (e.g., MC or molecular dynamics). Each cooling step adjusts a weight factor that is tracked for every system in the ensemble, and a resampling process may be used



**Figure 7.** Annealed importance sampling (AIS) algorithm. A set of independent systems is started at a high temperature (colored circles, top row), followed by three steps. (i) The systems are instantaneously “cooled” to a slightly lower temperature (larger  $\beta$ , second row from top), which results in the association of a weight to each system—the ratio of probabilities for the two  $\beta$  values, given the  $\theta$  value at the preceding temperature. (ii) The systems are then resampled according to their weights, leading to the elimination of some systems and multiple copies of others (color corresponds to “parent” system of top row). (iii) A brief amount of dynamics or MC sampling for each system is then performed to “relax” or decorrelate the systems (horizontal displacements along the  $\theta$  axis). These three steps of cooling-resampling-relaxing are repeated until the target temperature of interest ( $\beta = 1$ , bottom row) is reached. In Bayesian inference, the parameter  $\beta$  is nonphysical but acts to smooth the landscape in the same manner as the physical analog  $\beta = 1/k_B T$ .

to suppress low-weight samples and replicate those with higher weights.<sup>28</sup>

Importantly, AIS is formally equivalent to a stochastic Jarzynski-relation<sup>29</sup> protocol, and permits the calculation of free energies or partition functions.<sup>14</sup> AIS therefore can yield estimates of the marginal likelihood (19) for Bayesian model selection.

When an annealing protocol is used for sampling, the infinite-temperature limit ( $\beta = 0$ ) is attractive for several reasons. First, it can be sampled exactly, enabling the use of an annealing process initiated from truly independent  $\beta = 0$  samples. In BI problems with  $\sim 10$  or tens of parameters, typically, there are finite values of  $\beta > 0$  with sufficient overlap with the  $\beta = 0$  ensemble to permit practical sampling in an annealing protocol that gradually “cools” the system until  $\beta = 1$  is reached. In a biomolecular system, generically, one would expect effectively zero overlap with any numerically realizable finite  $\beta$  because of the huge number of degrees of freedom. A second point is that the ability to conduct multiple independent runs using fully random initial samples for  $\beta = 0$  readily enables the confirmation of the sampling quality. This type of complete independence of runs is essentially out of reach for complex biomolecular systems.

Third, when it is necessary to perform model selection via the Bayes factor (20), the AIS algorithm has the unique advantage of being able to compute the partition-function-like marginal likelihood (19) simultaneously as it estimates the posterior for parameter inference.<sup>14</sup> That is, based on the trivially computable  $\beta = 0$  reference system, AIS can yield what amounts to “absolute” partition functions (not partition function ratios) at the target  $\beta = 1$  value. Note that both the partition function and the marginal likelihood are integrals over (un-normalized) probability distributions. Thus, in practice, separate AIS simulations can be run for different models, each yielding the usual posterior for parameter inference and marginal likelihood, without the need for additional calculations to generate the ratios of marginal

likelihoods. This is a significant strength of AIS, given the importance of model selection in biophysical systems.<sup>7,9</sup>

In principle, AIS can be used with any constant-temperature canonical sampler, and some of these are discussed below. Note that the use of “walkers” at multiple temperatures is a feature shared in common with parallel tempering (PT) methods<sup>4,22,30,31</sup> which also are commonly used for Bayesian inference. In practical situations, PT sampling typically cannot exploit the  $\beta = 0$  limit,<sup>24</sup> unlike AIS.

**New and Old Fixed-Temperature Algorithms.** As noted above, methods such as AIS or parallel tempering are metasamplers and can use any one of a host of fixed-temperature (fixed  $\beta$ ) methods to generate probability distributions of interest. Fixed-temperature algorithms are stand-alone sampling procedures that can be integrated into a larger multi-temperature sampler. While we only discuss a few selected algorithms and their trade-offs, it is important to note that there are hybrid approaches that combine multiple methods.<sup>32</sup>

The random walk Metropolis–Hastings (MH) algorithm<sup>20</sup> is a well established, basic method for Markov chain Monte Carlo sampling. Briefly, for each iteration, a random jump from the current parameter set to a new parameter set is performed using a proposal (jump) distribution. The new candidate parameter set is either accepted or rejected based on the comparative probabilities of the old and proposed parameter sets. This generates a list of  $\theta$  values that eventually will converge to the target (in this case Posterior) distribution; however, the method is well-known to require significant tuning leading to potential convergence problems.<sup>1,33</sup>

A more advanced approach commonly used in BI is Hamiltonian Monte Carlo (HMC), which constructs trial moves based on approximate “dynamics” computed from gradients of the effective energy landscape (negative log of the probability distribution).<sup>7,21</sup> Hamiltonian-based methods build on MH, utilizing information on the posterior’s shape via its gradient to efficiently generate the proposed moves. Briefly, HMC introduces a momentum variable associated with each parameter that is combined to create a fictitious Hamiltonian. The Hamiltonian dynamics is simulated (via integration) for a given step size and number of steps, generating a new proposal. A popular extension of HMC is the No-U-Turn sampler<sup>34</sup> which adaptively determines the simulation step size and length, reducing the need for hyperparameter tuning. However, while Hamiltonian-based methods are able to efficiently sample from high dimensional posterior distributions, they require gradient calculations which may be infeasible to calculate. Furthermore, these methods may also struggle with multimodal distributions (i.e., multiple energy basins). We refer the interested reader to a conceptual introduction of HMC<sup>35</sup> for more details on this popular approach.

An alternative method that does not require gradients, the affine invariant ensemble sampler<sup>36</sup> evolves a set of walkers (MCMC chains) through the posterior landscape. For each sampling step, each walker is proposed a new position in the parameter space based on the position of other randomly selected walkers. In the “stretch” move, for example, a walker is moved in a direct line toward or away from another randomly selected walker, based on a random scale factor. The proposed move is then accepted or rejected based on the MH acceptance probability criterion. Critically, this proposal strategy is invariant to stretching, rotation, and translation, i.e., affine-invariant. In effect, the affine invariant ensemble sampler is able

to transform the complex geometry of a posterior distribution into a simpler distribution that can be sampled more efficiently. In our hands, this approach proved effective at sampling small-to-medium sized models of isothermal titration calorimetry (<10 parameters)<sup>8</sup> but required extensive computing. Relatedly, this algorithm requires an adequate number of walkers to densely cover the posterior, which can be well suited for multimodal distributions, but may become computationally infeasible for large dimensionality.

Finally, recent machine learning advances for density estimation using normalizing flows<sup>37,38</sup> have been incorporated into MCMC sampling methods. These generative models use neural networks with a series of differentiable and invertible transformations to learn a mapping from a simple distribution (e.g., Gaussian) to a complex distribution (e.g., the Posterior). One interesting application is the use of normalizing flows for preconditioning<sup>39</sup> within an annealed importance sampler. Here the normalizing flows are trained using the previous temperature stage and used to transform the current  $\beta$ -specific distribution into a standard Normal distribution. Metropolis-Hastings with a standard Normal proposal distribution is used, enabling efficient sample generation subject to the accuracy of transformation. We have found this method to be valuable for the Bayesian inference of dynamic models of membrane transporters with 10–20 parameters.<sup>40</sup>

While we have only scratched the surface of the available sampling algorithms here, we advise the reader that each algorithm has trade-offs and that the “best” sampler will depend on the specific problem at hand, such as the feasibility of gradient calculations or multimodality of the posterior.

### Challenges for Bayesian Inference and Monte Carlo.

While Bayesian inference and Markov-chain Monte Carlo are robust approaches to quantifying parameter estimates and uncertainty, they are not without challenges. First, as we have described previously, BI methods require the choice of a model, prior, and likelihood for a given data set in order to generate a posterior, which inherently requires some subjectivity and assumptions. However, the sensitivity of the inference results to these choices can be readily analyzed, and we find explicitly integrating and testing our modeling assumptions to be a valuable aspect of Bayesian inference.

Another potential challenge for Bayesian inference is how to deal with complex data structures and models. For hierarchical data such as data generated from a population and individuals, BI can be extended into hierarchical Bayesian inference.<sup>1</sup> This approach uses priors, likelihoods, and models for both the individual and population levels and introduces hyperparameters that govern how tightly coupled the different levels are. In effect, this systematically embeds individual BI models into a larger interconnected framework that can account for the uncertainties and dependencies of the data. A related concern is how to incorporate multiple data sets while accounting for batch effects (i.e., unique nuisance parameters for each data set). Two possible solutions to this are to use the posterior of one data set as the priors of the next data set, and use importance sampling between the posterior of the one data set and next.<sup>41</sup>

Moving to MCMC sampling algorithms specifically, the primary challenge is efficiently generating independent samples of the posterior distribution, due to complex posterior geometries and/or poor hyperparameter selection of the MCMC method. In particular, sampling challenges arising from strong correlations, multiple modes, and heavy tails in the

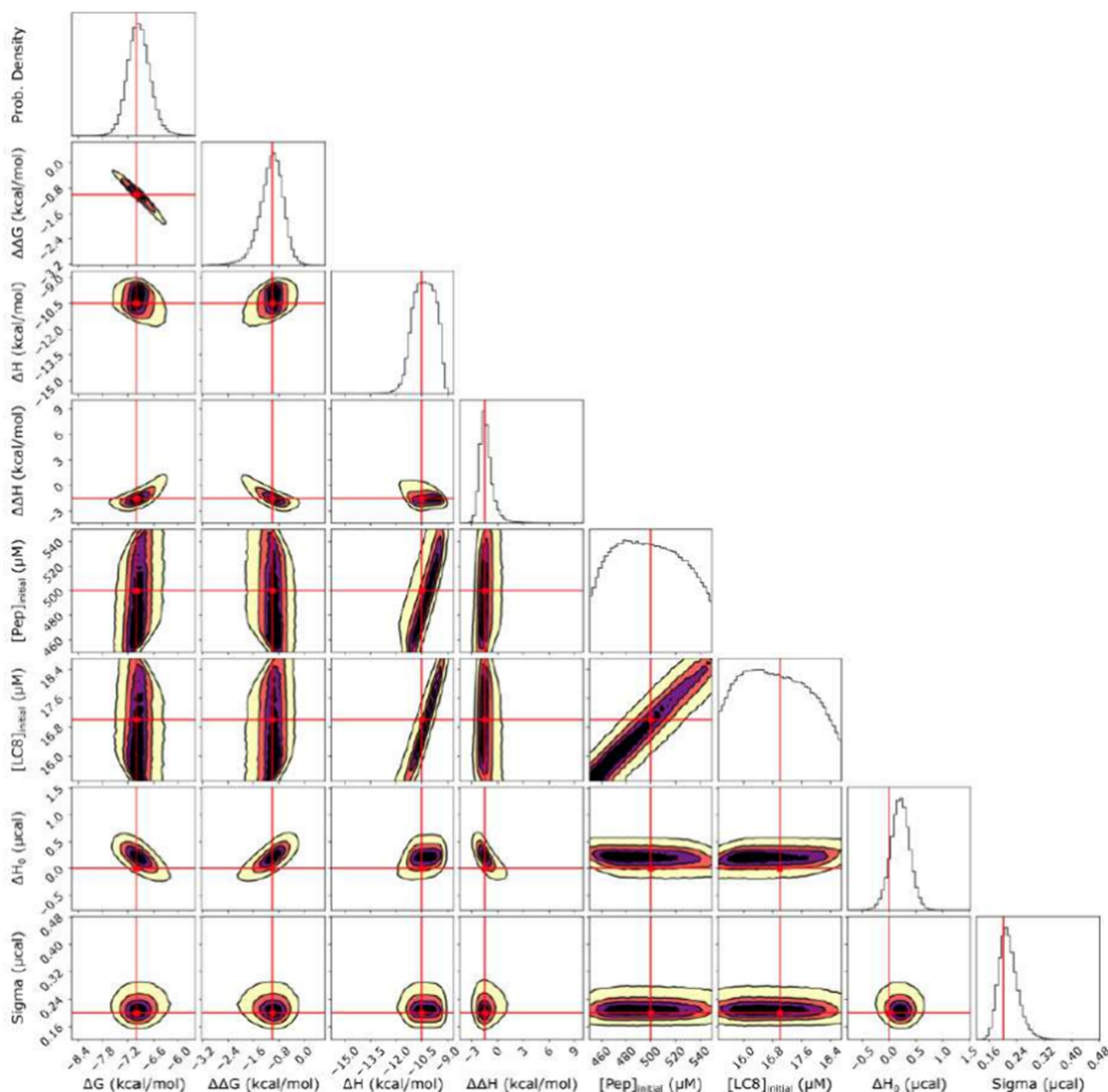
posterior region are exacerbated by high dimensionality and will prove difficult for any sampling method. Unfortunately, there is no silver bullet approach. Relatedly, MCMC algorithms may be susceptible to initial state bias, which can be tested by using multiple independent runs.

As an alternative to MCMC, variational inference provides a deterministic optimization approach to estimate intractable or high dimensional posteriors.<sup>1,42</sup> Here a tractable distribution (e.g., Gaussian) is selected to represent the posterior. Then the parameters of the tractable distribution, such as the mean and variance of a Gaussian, are adjusted such that the difference between the tractable distribution and the true posterior is minimized. Since the true posterior is not known, we cannot directly calculate this divergence but instead use an approximation: the evidence (marginal likelihood) as a lower bound. Maximizing this quantity will generate the optimal set of model parameters for the chosen surrogate posterior distribution. While computationally efficient, these methods are approximate and introduce stronger assumptions on the posterior distribution that may yield significant inaccuracies if not validated.

For certain problems, the likelihood function may be unknown or complex enough that it is intractable to calculate. In this case, there are “simulation-based” inference methods that calculate the posterior by comparing simulated data under different conditions to the observed data.<sup>43</sup> Approximate Bayesian computation<sup>1,44</sup> typically computes a similarity measure between the simulated and observed data or related summary statistics, only keeping the parameter sets that yield data below some threshold of similarity. These methods are sensitive to the choice of similarity measure and cutoff threshold, suffer from the curse of dimensionality, and may yield less informative and accurate estimates of the posterior than typical MCMC methods. However, recent neural-network based methods for density estimation such as normalizing flows<sup>45,46</sup> noted above have shown promise to address the limitations of approximate Bayesian computation.

Finally, while this review focuses on BI methods, we note alternative frequentist methods for parameter estimation and uncertainty quantification, such as maximum likelihood estimation with likelihood profiling,<sup>47</sup> or bootstrapping.<sup>48</sup> These approaches may be more computationally efficient than BI but do not generate an exact posterior distribution that fully and transparently describes the uncertainties and correlations of the model and data.

**Tools for Bayesian Inference in Practice.** We have found that successful BI in practice requires an appropriate choice of sampling algorithm and hyperparameters for a given model and data of interest. We recommend utilizing modern probabilistic programming frameworks, which provide an array of efficiently implemented sampling algorithms. Popular examples include STAN,<sup>21</sup> pyMC,<sup>49</sup> Pyro,<sup>50</sup> and Turing.jl.<sup>51</sup> In certain cases, more specialized sampling implementations may be useful, such as preconditioned Monte Carlo with normalizing flows<sup>39,52</sup> or the affine invariant ensemble sampler.<sup>36,53</sup> In either case, we suggest examining parameter identifiability<sup>54</sup> and correlations when doing BI. Also, we suggest validating the sampling results with ground truth values (i.e., synthetic data), checking for convergence using independent runs, comparing predicted posterior models with the observed data, and evaluating the sensitivity of the posterior to your modeling assumptions. Fortunately, many of the above frameworks can aid with exploring and diagnosing BI



**Figure 8.** Corner plot showing 2D and 1D marginals from Bayesian inference.<sup>8</sup> The plots show various projections of the full eight-dimensional posterior distribution for an example system based on MC simulation of synthetic experimental binding data mimicking isothermal titration calorimetry. Contour lines indicate probability densities (darker = higher probability) for the corresponding parameters with red lines indicating “known true values”. Correlations among parameters are clearly visible in the diagonal contours of 2D marginals. The 1D marginals reveal how “identifiable” each parameter is based on the width of the distribution. The prior distribution for each parameter was taken to be uniform over the range shown. Reproduced from ref 8. Available under a CC BY 4.0 DEED license. Copyright 2023 Estelle, George, Barbar, and Zuckerman.

models and sampling data, in addition to stand-alone tools such as ArviZ.<sup>55</sup> Finally, we highlight recent work that aims to integrate many of these practices into a comprehensive workflow for dynamic models.<sup>56</sup>

## ■ INTERPRETING DATA FROM BAYESIAN MONTE CARLO

**Confirming MC Sampling Quality.** Although true equilibrium sampling is effectively out of reach for many atomistic biomolecular systems, with some possible exceptions,<sup>57</sup> in Bayesian inference the guiding mindset should be to

obtain validated full sampling. The effective energy landscapes encountered in BI will not often be trivial, but we have a chance to sample them. If we trust the posterior distribution, then we can in turn trust the uncertainty ranges, “credibility regions”, provided by BI.

This review is not a manual for performing BI or assessing sampling, but we do recommend performing multiple independent MC simulations from different start points to ensure the same distributions are obtained. Visual comparison of the one and two-dimensional marginals of the posterior for all parameters is a key step, we have found.<sup>8</sup>

**Maximum Likelihood and Sampling.** It is worthwhile to revisit the maximum-likelihood approach to characterizing a model, analogous to global energy minimization, in the context of challenging-to-sample distributions, i.e., rugged landscapes. On the one hand, in contrast to molecular systems, there is no true physical heterogeneity in BI parameter space: if the model structure is correct, then there is a single set of correct parameters, not an equilibrium ensemble as in a physical system. Of course, in reality, finite data, noise, and experimental bias are likely to obscure the truly most likely parameters.

But aside from data-centric issues, there is a key practical issue in a maximum-likelihood/energy-minimization strategy: it may be extremely difficult to do global maximization in a landscape that is rugged and challenging to sample. In other words, can you trust the maximum likelihood estimates? In our hands, working with systems that are challenging—but possible—to sample, optimization algorithms do not provide reproducible, truly optimal results.<sup>40</sup> We know this because we can simply compare optimization results to the maximum likelihood observed in MC sampling, which is reproducible when sampling is adequate. While computationally expensive, we have found that good MC sampling is a robust way to find globally optimal parameters.

**Parameter Credibility Regions and Identifiability.** The main strength of Bayesian inference, in the authors' view, is the ability to provide not only parameter values but also uncertainty ranges consistent with the observed data. Any fitting (optimization) method can give a point estimate of most likely parameters for an assumed model, but uncertainties that do not include information about all possible parameter values make an uncontrolled approximation.<sup>19</sup> In BI, the posterior distribution of parameters consistent with model and data (13) tells us whether we really “know” the parameter value: the parameter may be characterized by a narrow posterior marginal (17), suggesting it is indeed *identified* by the data, or the posterior marginal may be very broad, indicating the most likely value is nearly meaningless. See Figure 8. The rationale here is in direct analogy to molecular simulation estimates, which always provide average values of observables, but the associated uncertainties indicate whether the averages are reliable.<sup>58,59</sup>

What “narrow” and “broad” mean will depend on the context, and domain expertise will play a key role in evaluating this, but comparison with the prior distribution provides a natural reference point. If the prior distribution correctly characterizes the prior state of knowledge before considering the data, then it is highly appropriate to characterize what is learned from the data via the concepts of information theory.<sup>17</sup>

It is important to note that it may be strictly impossible to gain information on certain parameters, based on the mathematically formal concept of identifiability.<sup>60</sup> Intuitively, the concept is straightforward for physical systems based on examples: equilibrium binding measurements cannot reveal on- or off-rates but can determine their (equilibrium) ratio; measurement of binding in one domain of a decoupled two-domain system cannot reveal properties of the second domain. There is, however, a gray zone framed as “practical identifiability.” For example, in a two-step binding process, if the first binding step is much weaker than the second, binding parameters for the first step may be practically nonidentifiable given insufficient data.<sup>8</sup> It is important to be aware of these issues, although the posterior implicitly includes identifiability

information even if it is unknown beforehand: if a posterior marginal does not differ from the prior for a given parameter, then the data has not helped to identify that parameter.<sup>17</sup> For further discussion of identifiability analysis in a Bayesian pipeline, see the discussion in recent work.<sup>5</sup>

We also note that BI credibility regions can help to reveal bias in experimental measurements. This can be done directly if an explicit “nuisance” parameter is used to represent bias in experimental measurements.<sup>40</sup> Alternatively, bias can be assessed implicitly based on nuisance parameters describing uncertainty for input experimental values: if the posterior marginal for such a nuisance parameter is found to have significant probability at the extremes of the assumed range, that suggests the original experimental range for the parameter was underestimated (assuming the model itself is correct).

**Parameter Correlations.** The posterior may contain significantly more information than can be gleaned from 1D posterior marginals, because some parameters will exhibit correlations. Such correlations, if present, are easily noticeable in the 2D marginals typically shown in “corner plots” (Figure 8) which can be visualized as contour plots.<sup>8</sup> Bear in mind that correlations may be linear or nonlinear.<sup>16</sup>

Such correlations provide important information in two ways. First, if two parameters are correlated, that suggests that a different experiment measuring one of the parameters separately could be used as prior information to identify the second by narrowing its marginal. Consider for example, the two concentrations [Pep] and [LC8] which lie along a clear diagonal in their 2D marginal in Figure 8. This posterior marginal indicates that, based on the data used, neither concentration is well identified, but if one could be refined, then the other would also be better determined. Note that these two concentrations are considered nuisance parameters: they are not the binding parameters of interest, but a full analysis requires accounting for realistic uncertainty in experimentally determined input parameters.<sup>8</sup>

A second, related value of correlations concerns different representations of parameters, which may prove to be revealing. That is, depending on the specific nature of the correlations, it may be the case that a function of two (or more) parameters may be well-identified, i.e., exhibit a narrower marginal than either of the original parameters. In the example data of Figure 8, the strong correlations between  $\Delta G$  (standard free energy change of first binding event) and  $\Delta\Delta G$  (free energy difference between first and second binding events) ultimately reveal that the marginal distribution of the total free energy for double binding ( $2\Delta G + \Delta\Delta G$ ) is much narrower than either of the components.<sup>8</sup> Thus, correlations may have readily interpretable physical meaning.

**Synthetic Data and Experimental Design.** While Bayesian inference methods are agnostic to the type of data used, we have found it valuable to utilize computer-generated “synthetic” data for several reasons. First, unlike with experimental data, synthetic data provide ground truth values since they are generated using a known model and parameter set. These ground truth parameter values help facilitate the testing and validation of a particular model and BI sampling implementation, such as confirming that the most likely parameter estimates correspond to their expected “true” values (within noise). Similarly, synthetic data may be used to examine the sensitivity to different choices of priors or model parametrizations.

A second use for synthetic data is the cost-effective design of the experiments. As discussed above, the width of the parameter estimates from the posterior distribution can be compared to that of the prior distributions in order to gauge how informative a given data set is. In other words, if the data does not significantly change the posterior as compared to the prior, then we have not gained information, as seen from Bayes' law (13). Operationally, BI is performed using the same physical model and priors but different synthetic data sets generated across a range of experimental conditions. Quantitative measures such the Kullback–Leibler (KL) divergence (i.e., relative entropy) may be used to compare separation of the posterior and prior distributions.<sup>17,40</sup> These quantities can then be used to determine the optimal experiment or set of experiments that maximizes the information gain. Here the use of synthetic data saves the potentially significant experimental costs needed to generate the data.

Generally speaking, BI can be broadly applied for optimal decision making with a given objective function (e.g., KL divergence). For further details we refer the interested reader to texts on Bayesian experimental design.<sup>1,61</sup>

## CONCLUSIONS

This review has attempted to connect the dots between Bayesian inference and statistical mechanics with a practical focus on Monte Carlo sampling (Figure 1). On the one hand, multidimensional probability theory—of which Bayes' rule is one example—is a point of rigorous commonality. On the other, whereas classical statistical mechanics is uniquely defined by the Boltzmann factor and Cartesian integration measure over phase space, the Bayesian framework is intrinsically ambiguous due to the need for prior specification of existing knowledge and of the parameters themselves. Yet regardless of those differences, the key message is that knowledge of statistical mechanics in the realm of physical chemistry/chemical physics provides a solid basis for understanding of Bayesian inference theory and practice.

The relationship between practical Monte Carlo calculations in these two realms is perhaps more interesting. Approaches based on physical ideas, such as faster sampling at higher temperatures, Hamiltonian dynamics, and the annealing idea, have proven to be of great value in Bayesian inference, arguably more than for the physical systems that motivated the development of some of the approaches. Fundamentally, this is due to the vast gap in dimensionality between typical Bayesian and biomolecular systems: the relatively low dimensionality (often 10–100 parameters) typically encountered in exact Bayesian calculations enables the practical use of high, and even infinite, effective temperatures in sampling.

## AUTHOR INFORMATION

### Corresponding Author

**Daniel M. Zuckerman** – Department of Biomedical Engineering, Oregon Health and Science University, Portland, Oregon 97239, United States; [orcid.org/0000-0001-7662-2031](https://orcid.org/0000-0001-7662-2031); Email: [zuckermd@ohsu.edu](mailto:zuckermd@ohsu.edu)

### Author

**August George** – Department of Biomedical Engineering, Oregon Health and Science University, Portland, Oregon 97239, United States; Present Address: Pacific Northwest National Laboratory, Richland, Washington, 99354, USA

Complete contact information is available at:  
<https://pubs.acs.org/10.1021/acs.jctc.4c00014>

## Notes

The authors declare no competing financial interest.

## ACKNOWLEDGMENTS

The authors are grateful for grant support from the NSF (MCB 2119837) and the NIH (GM141733), and appreciate valuable input from Jeremy Copperman, Aidan Estelle, and Harry Ryu.

## REFERENCES

- (1) Gelman, A.; Carlin, J. B.; Stern, H. S.; Dunson, D. B.; Vehtari, A.; Rubin, D. B. *Bayesian data analysis*; CRC Press, 2013.
- (2) Von Toussaint, U. Bayesian inference in physics. *Rev. Mod. Phys.* **2011**, *83*, 943.
- (3) Hines, K. E. A primer on Bayesian inference for biophysical systems. *Biophysical journal* **2015**, *108*, 2103–2113.
- (4) Gupta, S.; Lee, R. E.; Faeder, J. R. Parallel Tempering with Lasso for model reduction in systems biology. *PLoS computational biology* **2020**, *16*, e1007669.
- (5) Linden, N. J.; Kramer, B.; Rangamani, P. Bayesian parameter estimation for dynamical models in systems biology. *PLOS Computational Biology* **2022**, *18*, e1010651.
- (6) Nguyen, T. H.; Rustenburg, A. S.; Krimmer, S. G.; Zhang, H.; Clark, J. D.; Novick, P. A.; Branson, K.; Pande, V. S.; Chodera, J. D.; Minh, D. D. Bayesian analysis of isothermal titration calorimetry for binding thermodynamics. *PLoS One* **2018**, *13*, e0203224.
- (7) Nguyen, T. H.; La, V. N.; Burke, K.; Minh, D. D. Bayesian regression and model selection for isothermal titration calorimetry with enantiomeric mixtures. *PLoS One* **2022**, *17*, e0273656.
- (8) Estelle, A. B.; George, A.; Barbar, E. J.; Zuckerman, D. M. Quantifying cooperative multisite binding in the hub protein LC8 through Bayesian inference. *PLoS computational biology* **2023**, *19*, e1011059.
- (9) Bronson, J. E.; Fei, J.; Hofman, J. M.; Gonzalez, R. L.; Wiggins, C. H. Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule FRET data. *Biophysical journal* **2009**, *97*, 3196–3205.
- (10) Persson, F.; Lindén, M.; Unoson, C.; Elf, J. Extracting intracellular diffusive states and transition rates from single-molecule tracking data. *Nat. Methods* **2013**, *10*, 265–269.
- (11) Schütte, C.; Noé, F.; Lu, J.; Sarich, M.; Vanden-Eijnden, E. Markov state models based on milestoning. *J. Chem. Phys.* **2011**, *134*, 204105.
- (12) Trendelkamp-Schroer, B.; Wu, H.; Paul, F.; Noé, F. Estimation and uncertainty of reversible Markov models. *J. Chem. Phys.* **2015**, *143*, 174101.
- (13) Wasserman, L. Bayesian model selection and model averaging. *Journal of mathematical psychology* **2000**, *44*, 92–107.
- (14) Neal, R. M. Annealed importance sampling. *Statistics and computing* **2001**, *11*, 125–139.
- (15) MacCallum, J. L.; Perez, A.; Dill, K. A. Determining protein structures by combining semireliable data with atomistic physical models by Bayesian inference. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112*, 6985–6990.
- (16) Zuckerman, D. M. *Statistical physics of biomolecules: an introduction*; CRC Press, 2010.
- (17) Huber, H. A.; Georgia, S. K.; Finley, S. D. Systematic Bayesian posterior analysis guided by Kullback–Leibler divergence facilitates hypothesis formation. *J. Theor. Biol.* **2023**, *558*, 111341.
- (18) Bevington, P. R.; Robinson, D. K.; Blair, J. M.; Mallinckrodt, A. J.; McKay, S. Data reduction and error analysis for the physical sciences. *Computers in Physics* **1993**, *7*, 415–416.
- (19) Zuckerman, D. Maximum Likelihood vs. Bayesian estimation of uncertainty. 2022, <https://osf.io/ajuvf/> (access date: Sept. 1, 2023).
- (20) Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **1970**, *57*, 97–109.

- (21) Carpenter, B.; Gelman, A.; Hoffman, M. D.; Lee, D.; Goodrich, B.; Betancourt, M.; Brubaker, M. A.; Guo, J.; Li, P.; Riddell, A. Stan: A probabilistic programming language. *J. Statistical Software* **2017**, *76*, 1–32.
- (22) Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters* **1999**, *314*, 141–151.
- (23) Hénin, J.; Lelièvre, T.; Shirts, M.; Valsson, O.; Delemotte, L. Enhanced Sampling Methods for Molecular Dynamics Simulations [Article v1.0]. *Living Journal of Computational Molecular Science* **2022**, *4*, 1583–1583.
- (24) Zuckerman, D. M.; Lyman, E. A second look at canonical sampling of biomolecules using replica exchange simulation. *J. Chem. Theory Comput.* **2006**, *2*, 1200–1202.
- (25) Zuckerman, D. M. Equilibrium sampling in biomolecular simulations. *Annual review of biophysics* **2011**, *40*, 41–62.
- (26) Chang, G.; Guida, W. C.; Still, W. C. An internal-coordinate Monte Carlo method for searching conformational space. *J. Am. Chem. Soc.* **1989**, *111*, 4379–4386.
- (27) Jorgensen, W. L.; Tirado-Rives, J. Monte Carlo vs molecular dynamics for conformational sampling. *J. Phys. Chem.* **1996**, *100*, 14508–14513.
- (28) Huber, G. A.; McCammon, J. A. Weighted-ensemble simulated annealing: Faster optimization on hierarchical energy surfaces. *Phys. Rev. E* **1997**, *55*, 4822.
- (29) Jarzynski, C. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.* **1997**, *78*, 2690.
- (30) Swendsen, R. H.; Wang, J.-S. Replica Monte Carlo simulation of spin-glasses. *Physical review letters* **1986**, *57*, 2607.
- (31) Earl, D. J.; Deem, M. W. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3910–3916.
- (32) Kilic, Z.; Schweiger, M.; Moyer, C.; Pressé, S. Monte Carlo Samplers for Efficient Network Inference. *PLOS Computational Biology* **2023**, *19*, e1011256.
- (33) *Handbook of Markov Chain Monte Carlo*, 1st ed.; Brooks, S., Gelman, A., Jones, G., Meng, X.-L., Eds.; Chapman and Hall/CRC: New York, 2011; p 619.
- (34) Homan, M. D.; Gelman, A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.* **2014**, *15*, 1593–1623.
- (35) Betancourt, M. A Conceptual Introduction to Hamiltonian Monte Carlo. *arXiv* **2018**, DOI: 10.48550/arXiv.1701.02434.
- (36) Goodman, J.; Weare, J. Ensemble samplers with affine invariance. *Communications in Applied Mathematics and Computational Science* **2010**, *5*, 65–80.
- (37) Papamakarios, G.; Nalisnick, E.; Rezende, D. J.; Mohamed, S.; Lakshminarayanan, B. Normalizing Flows for Probabilistic Modeling and Inference. *J. Mach. Learn. Res.* **2021**, *22*, 1–64.
- (38) Gabrié, M.; Rotskoff, G. M.; Vanden-Eijnden, E. Adaptive Monte Carlo augmented with normalizing flows. *Proc. Natl. Acad. Sci. U. S. A.* **2022**, *119*, e2109420119.
- (39) Karamanis, M.; Beutler, F.; Peacock, J. A; Nabergoj, D.; Seljak, U. Accelerating Astronomical and Cosmological Inference with Preconditioned Monte Carlo. *Mon. Not. R. Astron. Soc.* **2022**, *516*, 1644–1653.
- (40) George, A.; Zuckerman, D. M. From average transient transporter currents to microscopic mechanism-A Bayesian analysis. *J. Phys. Chem. B* **2023**, *128*, 1830–1842.
- (41) Zuckerman, D. *Combining datasets for Bayesian inference*; OSF Preprints hv7yd; Center for Open Science, 2023.
- (42) Blei, D. M.; Kucukelbir, A.; McAuliffe, J. D. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association* **2017**, *112*, 859–877.
- (43) Cranmer, K.; Brehmer, J.; Louppe, G. The frontier of simulation-based inference. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 30055–30062.
- (44) Beaumont, M. A.; Zhang, W.; Balding, D. J. Approximate Bayesian Computation in Population Genetics. *Genetics* **2002**, *162*, 2025–2035.
- (45) Gonçalves, P. J.; Lueckmann, J.-M.; Deistler, M.; Nonnenmacher, M.; Öcal, K.; Bassetto, G.; Chintaluri, C.; Podlaski, W. F.; Haddad, S. A.; Vogels, T. P.; Greenberg, D. S.; Macke, J. H. Training deep neural density estimators to identify mechanistic models of neural dynamics. *eLife* **2020**, *9*, e56261.
- (46) Furia, C. S.; Churchill, R. M. Normalizing flows for likelihood-free inference with fusion simulations. *Plasma Physics and Controlled Fusion* **2022**, *64*, 104003.
- (47) Cole, S. R.; Chu, H.; Greenland, S. Maximum Likelihood, Profile Likelihood, and Penalized Likelihood: A Primer. *American Journal of Epidemiology* **2014**, *179*, 252–260.
- (48) Efron, B. Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association* **1987**, *82*, 171–185.
- (49) Abril-Pla, O.; Andreani, V.; Carroll, C.; Dong, L.; Fonnesbeck, C. J.; Kochurov, M.; Kumar, R.; Lao, J.; Luhmann, C. C.; Martin, O. A.; Osthege, M.; Vieira, R.; Wiecki, T.; Zinkov, R. PyMC: A Modern, and Comprehensive Probabilistic Programming Framework in Python. *PeerJ. Computer Science* **2023**, *9*, e1516.
- (50) Bingham, E.; Chen, J. P.; Jankowiak, M.; Obermeyer, F.; Pradhan, N.; Karaletsos, T.; Singh, R.; Szerlip, P.; Horsfall, P.; Goodman, N. D. Pyro: Deep Universal Probabilistic Programming. *J. Mach. Learn. Res.* **2018**, *20*, 1–6.
- (51) Ge, H.; Xu, K.; Ghahramani, Z. Turing: a language for flexible probabilistic inference. *International Conference on Artificial Intelligence and Statistics, AISTATS 2018*, Playa Blanca, Lanzarote, Canary Islands, Spain, April 9–11, 2018; PMLR, 2018; pp 1682–1690.
- (52) Karamanis, M.; Nabergoj, D.; Beutler, F.; Peacock, J. A.; Seljak, U. pocoMC: A Python package for accelerated Bayesian inference in astronomy and cosmology. *JOS* **2022**, *7*, 4634.
- (53) Foreman-Mackey, D.; Hogg, D. W.; Lang, D.; Goodman, J. emcee: The MCMC Hammer. *PASP* **2013**, *125*, 306–312.
- (54) Dong, R.; Goodbrake, C.; Harrington, H. A.; Pogudin, G. Differential Elimination for Dynamical Models via Projections with Applications to Structural Identifiability. *SIAM J. Appl. Algebra Geometry* **2023**, *7*, 194–235.
- (55) Kumar, R.; Carroll, C.; Hartikainen, A.; Martin, O. ArviZ: a unified library for exploratory analysis of Bayesian models in Python. *Journal of Open Source Software* **2019**, *4*, 1143.
- (56) Linden, N. J.; Kramer, B.; Rangamani, P. Bayesian parameter estimation for dynamical models in systems biology. *PLOS Computational Biology* **2022**, *18*, e1010651.
- (57) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How fast-folding proteins fold. *Science* **2011**, *334*, 517–520.
- (58) Grossfield, A.; Zuckerman, D. M. Quantifying uncertainty and sampling quality in biomolecular simulations. *Annual reports in computational chemistry* **2009**, *5*, 23–48.
- (59) Grossfield, A.; Patrone, P. N.; Roe, D. R.; Schultz, A. J.; Siderius, D. W.; Zuckerman, D. M. Best practices for quantification of uncertainty and sampling quality in molecular simulations [Article v1.0]. *Living journal of computational molecular science* **2018**, *1*, S067.
- (60) Raue, A.; Karlsson, J.; Saccomani, M. P.; Jirstrand, M.; Timmer, J. Comparison of approaches for parameter identifiability analysis of biological systems. *Bioinformatics* **2014**, *30*, 1440–1448.
- (61) Chaloner, K.; Verdinelli, I. Bayesian Experimental Design: A Review. *Statistical Science* **1995**, *10*, 273–304.