Bayesian mechanistic inference, statistical mechanics, and a new era for Monte Carlo

Daniel M. Zuckerman*,† and August George†,‡

†*Department of Biomedical Engineering, Oregon Health and Science University, Portland, Oregon*

‡*Current address: Pacific Northwest National Laboratory*

E-mail: zuckermd@ohsu.edu

Abstract

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., parameterizing 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 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 Perspective 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, as well as 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.

Contents

Introduction **3**

Probability theory and basic Bayesian inference formalism **5**

Elements of two dimensional probability theory 5

Higher dimensional probability theory 8

Bayesian inference as a bridge from theory to useful information 10

What is a model? 10

Recasting Bayes' theorem to define the posterior 12

The likelihood at the heart of Bayesian inference 14

A simple example 15

Considerations regarding the prior, and a contrast with statistical mechanics . . . 17

Posterior marginals and uncertainty ranges 18

Model selection and the connection to free energy calculations **20**

Defining the model selection problem 20

The marginal likelihood in model selection 20

Limitations of the partition function analogy 23

Monte Carlo in Bayesian inference **24**

Annealing algorithms 25

New and old fixed-temperature algorithms 28
Challenges for Bayesian inference and Monte Carlo 30
Tools for Bayesian inference in practice 32

Interpreting data from Bayesian Monte Carlo **33**
Confirming MC sampling quality 33
Maximum likelihood and sampling 33
Parameter credibility regions and identifiability 34
Parameter correlations 35
Synthetic data and experimental design 37

Conclusions **38**

Introduction

Bayesian inference (BI) has become a standard tool for addressing "inverse problems" in many fields of science, including chemistry and biophysics;[1–5] applications have included calorimetry analysis,[6–8] analysis of single-molecule data,[9,10] and development of Markov state models from molecular dynamics data.[11,12] BI takes observed data as input, potentially from different measurement types, and outputs a "posterior" probability distribution of parameters for a pre-set mathematical model that is consistent with the data and any prior assumptions. This multi-dimensional 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[1,7,13] – each defined by a set of equations and parameters.

Furthermore, BI can naturally be extended for nested and hierarchical models.[1]

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.[7] Procedurally, the BI inference process often uses Markov-chain Monte Carlo (MCMC) frequently borrowing methods motivated by physical science, such as parallel tempering and annealing.[4,14] There are modern molecular sampling methods that employ an implicitly Bayesian framework.[15]

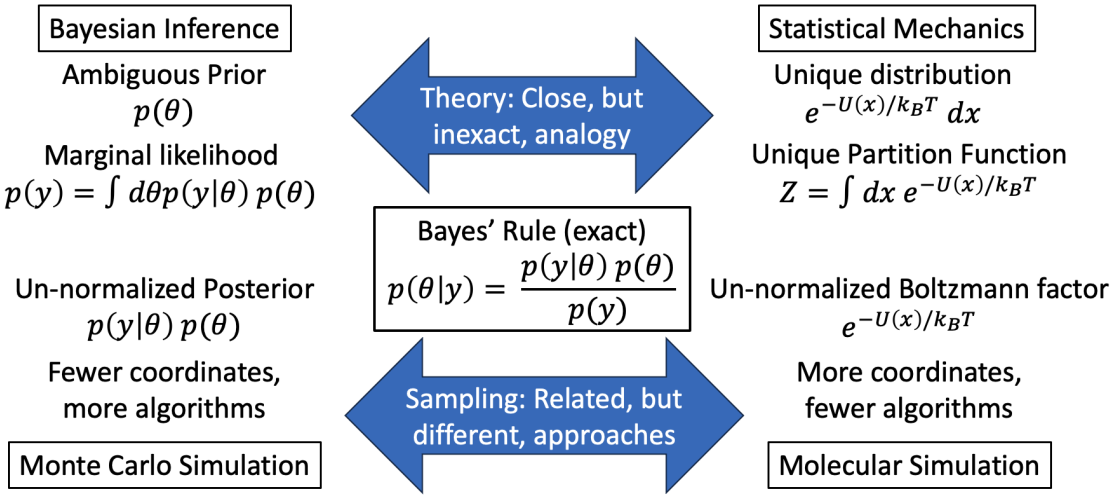

Figure 1: Relationships between theory and simulation for Bayesian inference and statistical mechanics. Although (exact) probability theory underpins both approaches, the physical basis of statistical mechanics yields unique calculations in contrast to the intrinsic uncertainty embedded in the Bayeisan 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.

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

We hope this article 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 article, 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

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

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 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

$$p_x(x) = \int dy\, p(x, y) \tag{2}$$

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 exactly analogous fashion and must also be normalized.

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

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

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 re-arranging (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

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

which is Bayes' rule.[1] 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.

(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 micro-features for a given macro-conformation, or vertically over macro-conformations for each set of micro-features. (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 coin-toss 'experiments'.

Looking ahead to the practical goal of parameter inference from data, if y represents the observed data and x represents a model parameter, then 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 evidently is 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 θ 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 $\vec{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ to take the place of the scalar variable x from the two dimensional case, and $\vec{y} = (y_1, y_2, \ldots, y_n)$ representing multi-dimensional data instead of the scalar y. Note that the dimensionalities of the two sets m and n generally will differ. For integrals we will use the shorthand $d\vec{\theta} = d\theta_1\, d\theta_2 \cdots d\theta_m$ and $d\vec{y} = dy_1\, dy_2 \cdots dy_n$.

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,

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

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

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

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 θ_1 and comparing to (5). Extending the marginal concept, we can define an example (normalized) two dimensional (2D) marginal via

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

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 integration. Two other marginals of great importance for BI theory consider the overall sets of parameters and data variables, namely,

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

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

Although the subscripts on these marginals may seem redundant, we include them to denote marginals defined strictly from probability theory. During Bayesian inference, the apparent marginals have different interpretations.

We can also define conditional probabilities using the marginals

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

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

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

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

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 gets 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.[6,8] 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 ΔH 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 ΔH and K_d. Therefore, two components of the parameter vector $\vec{\theta}$ will be ΔH and K_d (or its log).

Although our primary interest will be in determining the physical parameters of interest – e.g., ΔH 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 σ 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 parameters. In ITC, for example, an additional parameter ΔH_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.[6,8]

Mathematically, then, the set of parameters $\vec{\theta}$ includes both the real parameters of interest as well as 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 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

Before we recast Bayes' theorem in practical form for parameter inference, it's useful to address a question that may occur to physical scientists: Is the joint distribution of parameters $p(\vec{\theta}, \vec{y})$ a "real" thing? Is it measureable?

In deriving Bayes' theorem (12), we assumed the existence of a joint distribution of both parameters $\vec{\theta}$ and datasets $\vec{y}$. But this is not a directly observable distribution of physical variables – e.g., molecular coordinates – in the sense that physical scientists may be accustomed to considering. However, we can reconcile the joint distribution $p(\vec{\theta}, \vec{y})$ and physical intuition based on two points: (i) For every physically realizable dataset y, which is assumed to include noise and any other errors in measurement, there are indeed more and less likely model parameters based on the assumed model; these probabilities are embodied in the posterior distribution $p(\vec{\theta} \,|\, \vec{y})$, which cannot be observed via experiments but can be estimated in the inference process as shown below. (ii) Given the experimental setup including noise and possibly systematic errors, there are more and less likely data sets, embodied in $p_{\vec{y}}(\vec{y})$. We generally will not know this distribution, but we can see it has a physical basis: in principle, it could be measured based on many experiments. The product of the two functions just referenced indeed yields the full joint distribution, based on (10).

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 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 consistent with data and prior assump-

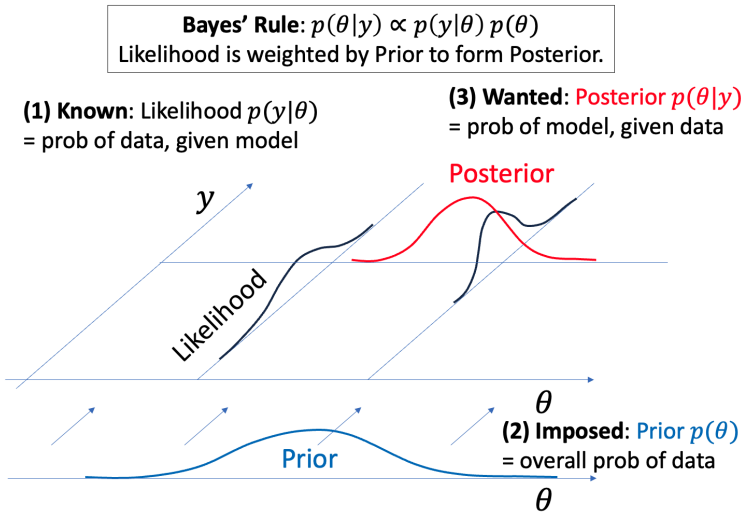

Figure 3: The components of Bayesian inference. Bayes' rule is an exact, basic result of probability theory, but Bayesian inference is a way of combining previous knowledge about a parameter of interest θ, encoded in the the Prior distribution, with data values $y_1, y_2, \ldots$. The Likelihood function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

tions is therefore

$$\mathrm{Post}(\vec{\theta}\,|\,\vec{y}) = \frac{\mathrm{Like}(\vec{y}\,|\,\vec{\theta})\,\mathrm{Prior}(\vec{\theta})}{p_{\vec{y}}(\vec{y})} \;\propto\; \mathrm{Like}(\vec{y}\,|\,\vec{\theta})\,\mathrm{Prior}(\vec{\theta}) \tag{13}$$

Two important points should be understood about (13). (i) Even though the prior distribution plays the formal role of the marginal in Bayes theorem (12) – more on this below – the logic here strictly describes a random process for both parameters and data: 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). However, in actuality, there is no random process generating parameters, and instead the prior is subjectively chosen. The posterior necessarily depends on the choice of prior, hopefully weakly as discussed below. (ii) 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 datasets. Hence, BI calculations safely ignore the denominator $p_{\vec{y}}(\vec{y})$, 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 we shall return to later.

The likelihood at the heart of Bayesian inference

As we can see from the fundamental Bayesian inference equation (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 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 $\vec{\theta}$, with the addition of a *noise model*. We first note that each data point y_i has been generated under conditions $\vec{c}_i$, for which the model makes a prediction – dependent on the parameters $\vec{\theta}$ – called $\hat{y}_i = \hat{y}_i(\vec{c}_i, \vec{\theta})$. With a Gaussian noise model parameterized by variance σ^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

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

This equation models the assumption that the noise for every data point is independent. Note that σ 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.

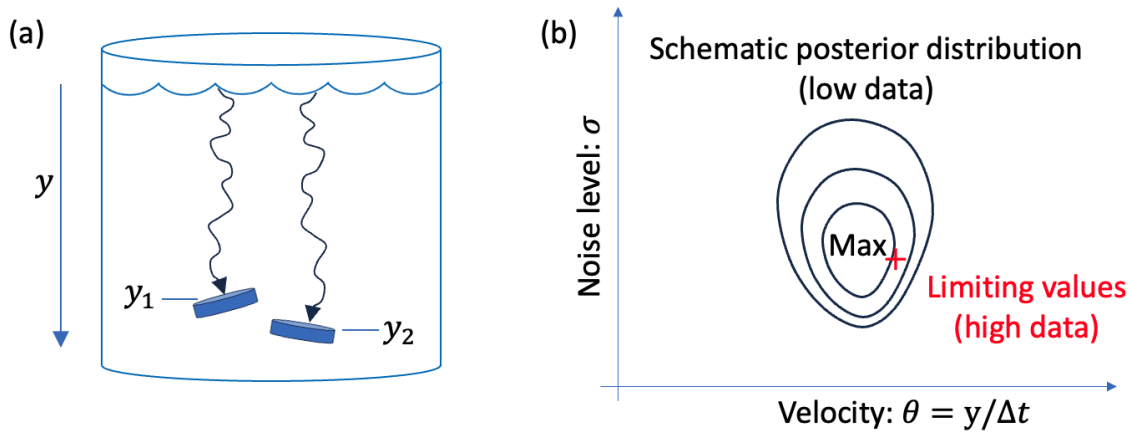

Figure 4: Bayesian inference for a simple linear model. The parameter of interest is the terminal velocity of a coin dropped in water. Measured distance (y) values are linearly related to the parameter of interest, velocity ($\theta = y/\Delta t$), given the constant observation time Δt. When a relatively 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.

A simple example

Several key lessons regarding Bayesian inference are revealed by a simple model which permits closed-form evaluation of the likelihood function. We consider terminal velocity measurements (Figure 4) where the velocity "parameter" of interest θ 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 θ, and it is the same for all datapoints, i.e., $\hat{y}_i = \hat{y} = \theta\Delta t$. Thus, the likelihood is

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

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

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

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

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

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

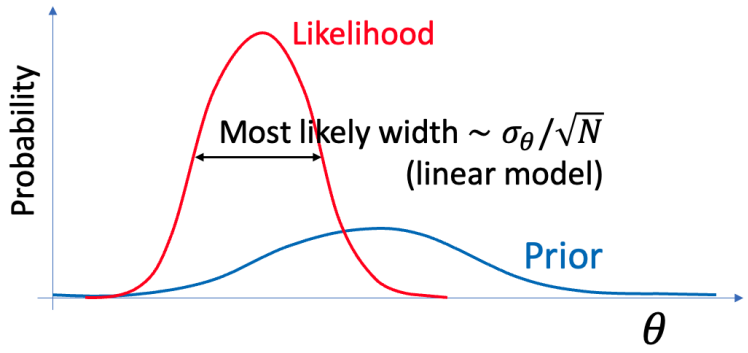

Figure 5: Relative effects of likelihood and prior. Based on the available data, there is some fixed width in the likelihood function (red curve) for a parameter of interest θ based on the most likely noise parameter(s). In a linear model, the width is simply related to the measured data variance, but more generally there is some distribution of θ 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 likelihood.

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 to 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 datasets $\vec{y}$, 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.[16]

A practical question that natural 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.[8] 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 make the posterior relatively independent of the choice of the prior, there is no simple number of data points which 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 only holds 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.[17]

Posterior marginals and uncertainty ranges

For many inference applications, the estimation of uncertainty for parameters of interest is a key goal. The high-dimensional posterior $\mathrm{Post}(\vec{\theta}\,|\,\vec{y})$ 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 as-

sumptions. 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

$$\mathrm{Marg}(\theta_1) = \int d\theta_2\, d\theta_3 \cdots d\theta_m\ \mathrm{Post}(\vec{\theta}\,|\,\vec{y}) \tag{17}$$

$$\mathrm{Marg}(\theta_1, \theta_2) = \int d\theta_3\, d\theta_4 \cdots d\theta_m\ \mathrm{Post}(\vec{\theta}\,|\,\vec{y}) \tag{18}$$

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 $\mathrm{Marg}(\theta_1) = \mathrm{Marg}(\theta_1|\vec{y})$, indicating 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[18] which can fail to include information about all parameter values despite use of the likelihood function.[19]

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 dataset $\vec{y}$, 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*[13] 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*, 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

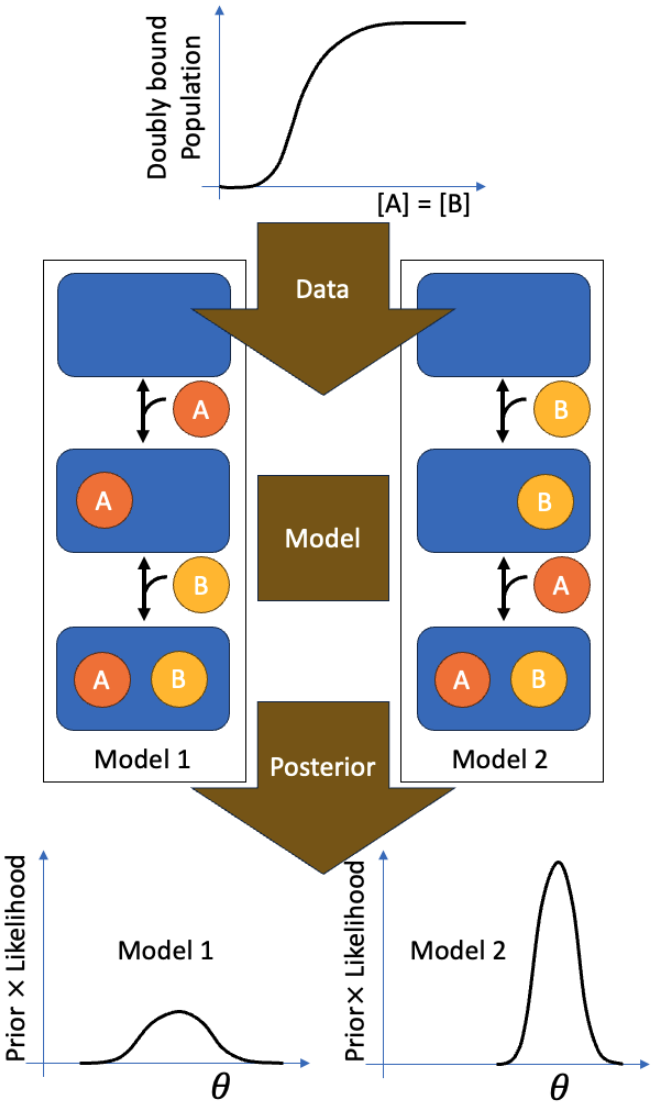

Figure 6: Model comparison in Bayesian inference. A hypothetical system in which a protein (blue) that binds two ligands (A and B) is analyzed based on experimental readout of the doubly-bound protein population as the concentrations of A and B are simultaneously increased. A given set of experimental 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, Prior(θ) $\times$ Like(θ), are not normalized over parameters, implying integration over parameters will yield different values known as marginal likelihoods.

model $\mathcal{M}$, then given the set of data $\vec{y}$, the marginal likelihood for the model is

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

The "marginal" nomenclature arises because if the integrand $\mathrm{Like}(\vec{y}\,|\vec{\theta})\,\mathrm{Prior}(\vec{\theta})$ is taken to be a true joint probability distribution – following (11) – then the integral over $\vec{\theta}$ 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 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 dataset $\vec{y}$. 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*,[13] which is the simple ratio of the two marginal likelihoods.

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

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;[13] 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.[9] 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 σ^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.[16] 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, that given the noise intrinsic to data, one can argue we have no choice but 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.[7]

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,[16] 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 test the sensitivity of results to the prior.

Monte Carlo in Bayesian inference

The main practical task in Bayesian inference is sampling the posterior distribution $\mathrm{Post}(\vec{\theta}|\vec{y})$ 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.

Sampling the BI posterior is formally identical to equilibrium molecular simulation, but the much lower dimensionality in BI dramatically alters the slate of algorithms that can be considered. Instead of trying to sample the Boltzmann-factor distribution $e^{-U(\vec{x})/k_BT}$ 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 $\mathrm{Post}(\vec{\theta}) = \mathrm{Post}(\vec{\theta}|\vec{y})$, where typically there are often 10-100 components of $\vec{\theta}$ and $\vec{y}$ is fixed data that is not sampled. By analogy to molecular simulation, we can say that $\vec{y}$, along with the prior and noise model, set the effective "energy

landscape" for the BI coordinates $\vec{\theta}$. See (13) and (14).

Here, we primarily restrict ourselves to Markov chain Monte Carlo (MCMC) algorithms, 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,[20] Hamiltonian MC,[7,21] 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,[22,23] 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.[24,25] 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 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.[26,27] Likewise, in Bayesian inference, it may prove more natural to consider a transformed coordinate, e.g., based on 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.[14] Instead of considering only the true

posterior, we introduce an inverse temperature parameter β and consider distributions of the form

$$p_\beta(\vec{\theta}) \equiv \left[\text{Like}(\vec{y}|\vec{\theta})\right]^\beta \text{Prior}(\vec{\theta}) \tag{21}$$

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 multi-dimensional Gaussian.

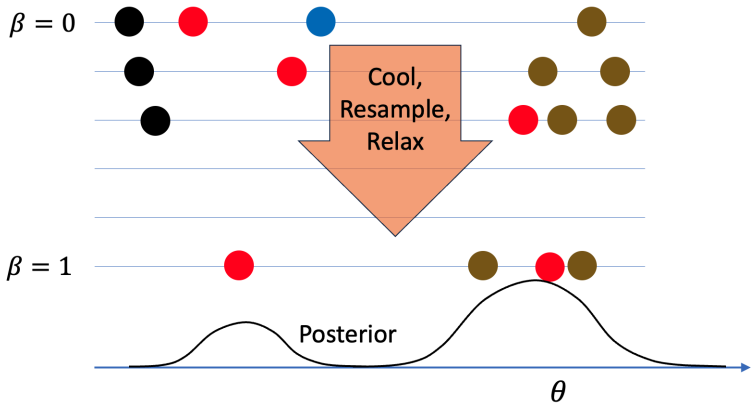

Figure 7: The annealed importance sampling (AIS) algorithm. A set of independent systems is started at a high temperature (colored circles, top row), followed by three steps. (i) The systems are instantaneously "cooled" to a slightly lower temperature (larger β, second row from top), which results in the association of a weight to each system – the ratio of probabilities for the two β values, given the θ 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 θ 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 β is non-physical but acts to smooth the landscape in the same manner as the physical analog $\beta = 1/k_B T$.

Annealed importance sampling (AIS),[14,28] in physical terms, uses a valid high-temperature (low β) 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 constant-temperature simulation (e.g., MC or MD). Each cooling step adjusts a weight factor that is tracked for

every system in the ensemble, and a resampling process may be used to suppress low-weight samples and replicate those with higher weights.[28]

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

When using an annealing protocol 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 β 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 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).[14] That is, based on the trivially computable $\beta = 0$ reference system, AIS can yield what amounts to absolute free energies at the target $\beta = 1$ value. This is an important strength of AIS, given the importance of model selection in biophysical systems.[7,9]

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[4,22,30,31] which also are commonly used for Bayesian inference. In practical situations, PT sampling typically cannot exploit the $\beta = 0$ limit,[24] unlike AIS.

New and old fixed-temperature algorithms

As noted above, methods such as AIS or parallel tempering are meta-samplers, and can use any one of a host of fixed-temperature (fixed β) 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.[32]

The random walk Metropolis-Hastings (MH) algorithm[20] 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 $\vec{\theta}$ 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.[1,33]

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).[7,21] Hamiltonian-based methods build off of MH, utilizing information of the posterior's shape via its gradient to efficiently generate independent samples to inform the proposed moves. Briefly, HMC introduces a momentum variable associated with each parameter, that are combined to create a fictitious Hamiltonian. The Hamiltonian dynamics are 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[34] which adaptively determines that simulate 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 infeasible to calculate. Furthermore, these methods

may also struggle with multi-modal distributions (i.e., multiple energy basins). We refer the interested reader to a conceptual introduction of HMC[35] for more details on this popular approach.

An alternative method that does not require gradients, the affine invariant ensemble sampler[36] 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 towards or away from another randomly selected walker, based on a random scale factor. The proposed move is then accepted/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)[8] but required extensive computing. Relatedly, this algorithm requires an adequate number of walkers to densely cover the posterior, which can be well suited for multi-modal distributions, but may become computationally infeasible for large dimensionality.

Finally, recent machine learning advances for density estimation using normalizing flows[37,38] can be 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[39] within an annealed importance sampler. Here the normalizing flows are trained using the previous temperature stage and used to transform the current posterior into a standard Normal distribution. Metropolis-Hastings with a standard Normal distribution is used, enabling efficient independant samples to be drawn from the posterior – 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.[40]

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 uncertainities, 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 dataset 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 that 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 heirarchical data such as data generated from a population and individuals, BI can be extended into hierarchical Bayesian inference.[1] 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 dataset). Two possible solutions to this are to use the posterior of one dataset as the priors of the next dataset, and use importance sampling between the posterior of the one dataset and next.[41]

Moving to MCMC sampling algorithms specifically, the primary challenge is efficiently generating independent samples of the posterior distribution which are often 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 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 using multiple independent runs.

As an alternative to MCMC, variational inference provides a deterministic optimization approach to estimate intractable or high dimensional posteriors.[1,42] 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 true posterior is minimized. Since the true posterior isn't known we cannot directly calculate this divergence, but instead use an approximation – the evidence (marginal likelihood) 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.[43] Approximate Bayesian computation[1,44] 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[45,46] described previously have shown promise to address the limitations of approximate Bayesian computation.

Finally, while this perspective focuses on BI methods, we note alternative frequentist methods for parameter estimation and uncertainty quantification, such as maximum likelihood estimation with likelihood profiling,[47] or bootstrapping.[48] These approaches may be more computationally efficient than BI but do not generate an exact posterior distribution which 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,[21] pyMC,[49] Pyro,[50] and Turing.jl.[51] In certain cases, more specialized sampling implementations may be useful, such as preconditioned Monte Carlo with normalizing flows[39,52] or the affine invariant ensemble sampler.[36,53] In either case, we suggest examining parameter identifiability[54] 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 data with the observed data, and evaluating the sensitivity of posterior to your modeling assumptions. Fortunately, many of the above frameworks can aid with exploring and diagnosing BI models and sampling data, in addition to stand-alone tools such as ArviZ.[55] Finally, we highlight recent work that aims to integrate many of these practices into a comprehensive workflow for dynamic models.[56]

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,[57] 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 article 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.[8]

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 maximum likelihood estimates? In our hands, working with systems that are challenging – but possible – to sample, optimization algorithms do not provide reproducible, truly optimal re-

sults.[40] We know this because we can simply compare 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.[19] 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.[58,59]

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.[17]

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.[60] 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 non-identifiable given insufficient data.[8] 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.[17] For further discussion of identifiability analysis in a Bayesian pipeline, see the discussion in recent work.[5]

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.[40] 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 under-estimated (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.[8] Bear in mind that correlations may be linear or non-linear.[16]

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

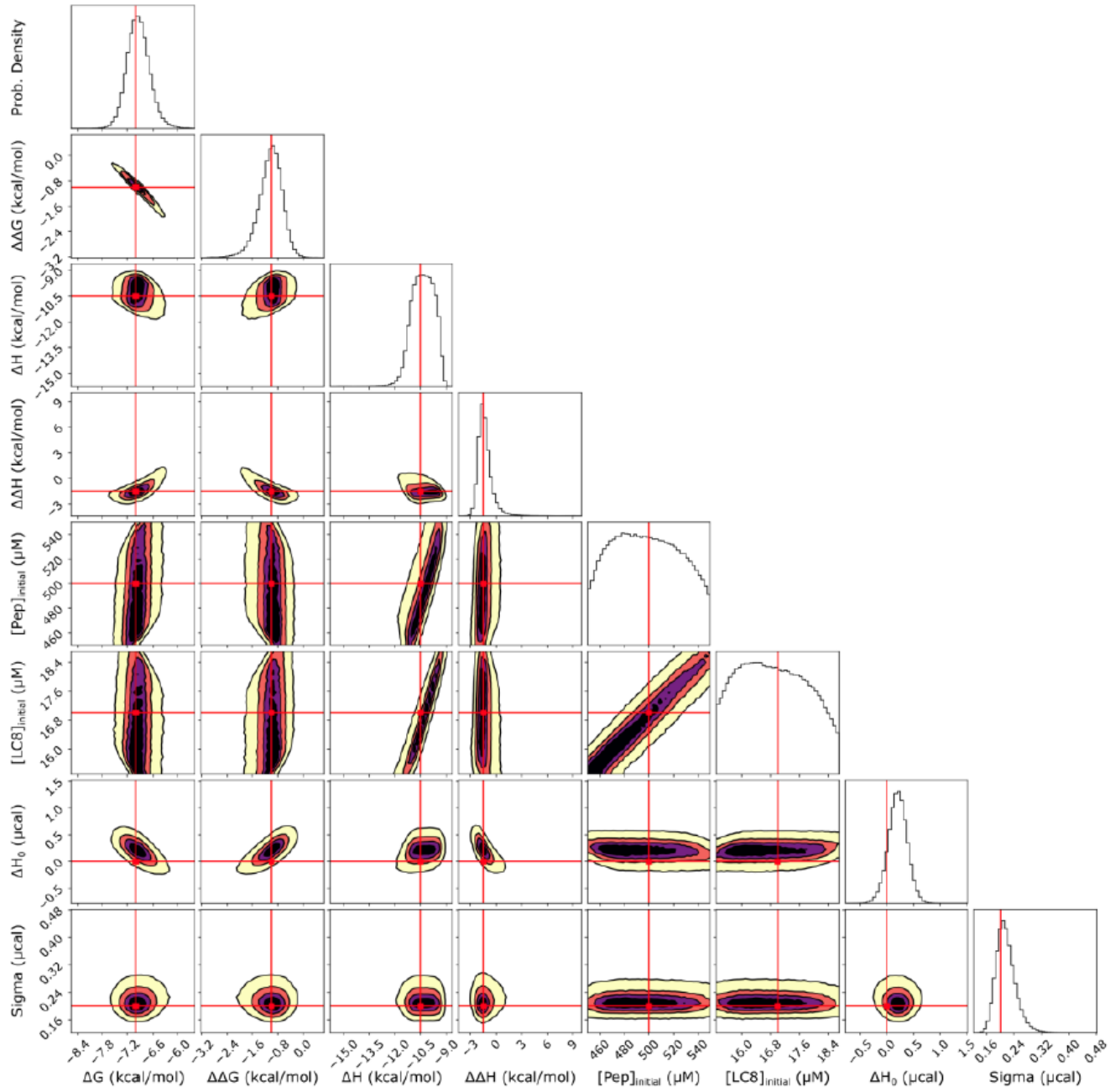

Figure 8: Corner plot showing 2D and 1D marginals from Bayesian inference.[8] The plots show various projections of the full eight-dimensional posterior distribution for an example system based on MC simulation of synthetic experimental binding data mimicking isothermal titration calorimetry. Contour lines indicate probability densities (darker = higher probability) for the corresponding parameters noted, with red lines indicating known true values. Correlations among parameters are clearly visible in 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.

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 experimental measurements.[8]

A second, related value of correlations concerns different representations of parameters, which may prove 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 ΔG (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.[8] Thus, correlations may have a 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. Firstly, unlike with experimental data, synthetic data provides ground truth values since it is 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 parameterizations.

A second use for synthetic data is the cost-effective design of 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 dataset 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 datasets 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.[17,40] These quantities can then be used to determine the optimal experiment that maximizes the information gain. Here the use of synthetic data saves the valuable 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.[1,61]

Conclusions

This Perspective 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, multi-dimensional 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 the two realms is perhaps more interesting. Approaches based on physical ideas, such as faster sampling at higher temperatures and the annealing idea, have proven of great value in Bayesian inference, arguably more than for the physical systems which motivated 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 practical use of high, and even infinite, effective temperatures in sampling.

Acknowledgement

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. *Reviews of Modern Physics* **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. *Nature methods* **2013**, *10*, 265–269.

(11) Schütte, C.; Noé, F.; Lu, J.; Sarich, M.; Vanden-Eijnden, E. Markov state models based on milestoning. *The Journal of chemical physics* **2011**, *134*.

(12) Trendelkamp-Schroer, B.; Wu, H.; Paul, F.; Noé, F. Estimation and uncertainty of reversible Markov models. *The Journal of chemical physics* **2015**, *143*.

(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. *Proceedings of the National Academy of Sciences* **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. *Journal of Theoretical Biology* **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/.

(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. *Journal of statistical software* **2017**, *76*.

(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. *Journal of chemical theory and computation* **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. *Journal of the American Chemical Society* **1989**, *111*, 4379–4386.

(27) Jorgensen, W. L.; Tirado-Rives, J. Monte Carlo vs molecular dynamics for conformational sampling. *The Journal of Physical Chemistry* **1996**, *100*, 14508–14513.

(28) Huber, G. A.; McCammon, J. A. Weighted-ensemble simulated annealing: Faster optimization on hierarchical energy surfaces. *Physical Review E* **1997**, *55*, 4822.

(29) Jarzynski, C. Nonequilibrium equality for free energy differences. *Physical Review Letters* **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) Brooks, S., Gelman, A., Jones, G., Meng, X.-L., Eds. *Handbook of Markov Chain Monte Carlo*, 1st ed.; 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. 2018.

(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*.

(38) Gabrié, M.; Rotskoff, G. M.; Vanden-Eijnden, E. Adaptive Monte Carlo augmented with normalizing flows. *Proceedings of the National Academy of Sciences* **2022**, *119*, e2109420119.

(39) Karamanis, M.; Beutler, F.; Peacock, J. A.; Nabergoj, D.; Seljak, U. Accelerating Astronomical and Cosmological Inference with Preconditioned Monte Carlo. *Monthly Notices of the Royal Astronomical Society* **2022**, *516*, 1644–1653.

(40) George, A.; Zuckerman, D. M. From average transient transporter currents to microscopic mechanism–A Bayesian analysis. *bioRxiv* **2023**, 2023–10.

(41) Zuckerman, D. *Combining datasets for Bayesian inference*; OSF Preprints hv7yd, 2023.

(42) David M. Blei, A. K.; 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. *Proceedings of the National Academy of Sciences* **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* **2013**, *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. *Journal of Machine Learning Research* **2018**,

(51) Ge, H.; Xu, K.; Ghahramani, Z. Turing: a language for flexible probabilistic inference. International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain. 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. 2022.

(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.; G., P. Differential Elimination for Dynamical Models via Projections with Applications to Structural Identifiability. *SIAM Journal on Applied Algebra and 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*, 1–48.

(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*.

(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.

TOC Graphic

Some journals require a graphical entry for the Table of Contents. This should be laid out "print ready" so that the sizing of the text is correct.
Inside the `tocentry` environment, the font used is Helvetica 8 pt, as required by *Journal of the American Chemical Society*.
The surrounding frame is 9 cm by 3.5 cm, which is the maximum permitted for *Journal of the American Chemical Society* graphical table of content entries. The box will not resize if the content is too big: instead it will overflow the edge of the box.
This box and the associated title will always be printed on a separate page at the end of the document.