Cognitive Neurodynamics
https://doi.org/10.1007/s11571-021-09696-9

REVIEW PAPER

An introduction to thermodynamic integration and application to dynamic causal models

Eduardo A. Aponte[1,2] · Yu Yao[1] · Sudhir Raman[1] · Stefan Frässle[1] · Jakob Heinzle[1] · Will D. Penny[3] · Klaas E. Stephan[1,4]

Received: 31 January 2021 / Revised: 3 June 2021 / Accepted: 1 July 2021

Abstract
In generative modeling of neuroimaging data, such as dynamic causal modeling (DCM), one typically considers several alternative models, either to determine the most plausible explanation for observed data (Bayesian model selection) or to account for model uncertainty (Bayesian model averaging). Both procedures rest on estimates of the model evidence, a principled trade-off between model accuracy and complexity. In the context of DCM, the log evidence is usually approximated using variational Bayes. Although this approach is highly efficient, it makes distributional assumptions and is vulnerable to local extrema. This paper introduces the use of thermodynamic integration (TI) for Bayesian model selection and averaging in the context of DCM. TI is based on Markov chain Monte Carlo sampling which is asymptotically exact but orders of magnitude slower than variational Bayes. In this paper, we explain the theoretical foundations of TI, covering key concepts such as the free energy and its origins in statistical physics. Our aim is to convey an in-depth understanding of the method starting from its historical origin in statistical physics. In addition, we demonstrate the practical application of TI via a series of examples which serve to guide the user in applying this method. Furthermore, these examples demonstrate that, given an efficient implementation and hardware capable of parallel processing, the challenge of high computational demand can be overcome successfully. The TI implementation presented in this paper is freely available as part of the open source software TAPAS.

Keywords Model evidence · Free energy · Population MCMC · DCM · Model comparison · fMRI

Author summary

When fitting computational models to data in the setting of Bayesian inference, a user has the choice between two broad classes of algorithms: variational inference and Monte Carlo simulation. While both methods have advantages and drawbacks, variational inference has become standard in the domain of modelling directed brain connectivity due to its computational efficiency, especially when the challenge is to select between competing hypotheses that explain the observed data. By contrast, the high computational demand by Monte Carlo methods has so far prevented their widespread use for inference on brain connectivity, despite their capability to overcome some of the shortcomings of variational inference. In this paper, we introduce the user to thermodynamic integration (TI), a Monte Carlo method designed for model fitting and model selection. By covering its foundations and historical origins in statistical physics, we hope to convey an in-depth understanding of TI that goes beyond a purely technical treatment. In addition, we also provide examples for concrete applications, demonstrating that, given an efficient implementation and up-to-date hardware, the challenge of high computational demand can be overcome successfully.

Eduardo A. Aponte and Yu Yao have contributed equally.

✉ Yu Yao
yao@biomed.ee.ethz.ch

[1] Translational Neuromodeling Unit (TNU), Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Zurich, Switzerland

[2] Present Address: Roche Innovation Center, Grenzacherstrasse 124, 4070 Basel, Switzerland

[3] School of Psychology, University of East Anglia, Norwich, UK

[4] Max Planck Institute for Metabolism Research, Cologne, Germany

Published online: 25 July 2021

Introduction

Dynamic causal models (DCMs) are generative models that serve to infer latent neurophysiological processes and circuit properties—e.g., the effective connectivity between neuronal populations—from neuroimaging measurements such as functional magnetic resonance imaging (fMRI) or magneto-/electroencephalography (M/EEG) data (David et al. 2006; Friston et al. 2003). As reviewed by Daunizeau et al. (2011), DCMs consist of two hierarchically related layers: a set of state equations describing neuronal population activity, and an observation model which links neurophysiological states to observed signals and accounts for measurement noise. Equipped with a prior distribution over model parameters, a DCM specifies a full generative forward model that can be inverted using Bayesian techniques.

In addition to inference on model parameters, an important scientific problem is the comparison of competing hypotheses, for example, different network topologies, which are formalized as different models. Under the Bayesian framework, model comparison is based on the evidence or marginal likelihood of a model. The model evidence corresponds to the denominator from Bayes' theorem and represents the probability of the observed data under a given model. It is a widely used score of model quality that quantifies the trade-off between model fit and complexity (Bishop 2006; MacKay 2004).

Unfortunately, in most instances, it is not feasible to derive an analytical expression of the model evidence due to the intractable integrals that arise from the marginalization of the model parameters. While various asymptotical approximations exist, such as the Bayesian Information Criterion (BIC, Schwarz 1978) and more recently the Widely Applicable Bayesian Information Criterion (WBIC, Watanabe 2013), variational Bayes under the Laplace approximation (VBL, Friston et al. 2007) has established itself as the method of choice for DCM, partially because of its computational efficiency. Within the framework of variational Bayes (VB), a lower bound approximation of the log model evidence (LME) is obtained as a byproduct of model inversion: the variational negative free energy (which we refer to as $-F_{VB}$ throughout this paper).

While highly efficient, model comparison based on the variational free energy has several potential pitfalls. For example, under the Laplace approximation as used in the context of DCM, there is no guarantee that $-F_{VB}$ still represents a lower bound of the LME (Wipf and Nagarajan 2009). Furthermore, VB is commonly performed in combination with a mean field approximation, and the effect of this approximation on the posterior estimates can be difficult to predict (Daunizeau et al. 2011). Finally, in nonlinear models, the posterior could become a multimodal density, a condition that aggravates the application of gradient ascent methods regularly used in combination with the Laplace approximation.

For these reasons, Markov Chain Monte Carlo (MCMC) sampling has been explored as an alternative inference technique for DCM (Aponte et al. 2016; Chumbley et al. 2007; Penny and Sengupta 2016; Sengupta et al. 2015, 2016; Yao and Stephan 2020). MCMC is particularly attractive for variants of DCMs in which Gaussian assumptions might be less adequate, such as nonlinear DCMs for fMRI (Stephan et al. 2008), DCMs of electrophysiological data (Moran et al. 2013), or DCMs for layered fMRI signals (Heinzle et al. 2016). MCMC is also useful when extending DCM to more complex hierarchical models (Raman et al. 2016), in which the derivation of update equations for VB becomes difficult (but see Yao et al. 2018). MCMC is asymptotically exact and only assumes that the posterior distribution can be evaluated up to a multiplicative constant. However, in practice, its computational cost often leads to prohibitively long computation times for the datasets and models commonly encountered in neuroimaging. Furthermore, in contrast to VB, MCMC does not provide an estimate of the model evidence by default.

While several MCMC-based strategies for computing the LME in neuroimaging applications have been explored (e.g., Aponte et al. 2016; Penny and Sengupta 2016; Raman et al. 2016), one particularly powerful and theoretically attractive MCMC variant is thermodynamic integration (TI). This method, like VB, rests on the concept of free energy and has been proposed as gold standard for LME estimation (Calderhead and Girolami 2009; Lartillot and Philippe 2006). Despite strong theoretical advantages, so far, the computational costs of TI have prevented its practical use in neuroimaging.

This paper introduces the reader to thermodynamic integration (TI) and its application to DCM. In contrast to existing tutorials on TI (Annis et al. 2019), we provide an in-depth discussion of the theoretical foundations of TI and relate the tutorial specifically to DCM as a generative model that is frequently used in contemporary neuroimaging analyses. Our discussion covers the key concept of free energy starting from its historical origin in statistical physics, with the aim of conveying a deeper understanding of this method that goes beyond a purely technical treatment. In the second part, we present a series of examples involving both synthetic and real-world datasets. These include (1) a validation dataset based on a linear regression model with analytically tractable LME used to verify the accuracy of TI, (2) a synthetic fMRI dataset where the true data-generating model is known for each observation, and

(3) a real-world fMRI dataset used to demonstrate LME estimation for nonlinear DCM.

In addition to showcasing the application of TI, these examples also serve to demonstrate that, given an efficient implementation and hardware capable of parallel processing, the challenge of high computational demand of TI can be overcome successfully. The software implementation of TI and DCM used in this paper is available as part of the open-source toolbox TAPAS (Translational Neuromodeling Unit 2014).

To keep this paper short and yet accessible to a broad audience, summaries of key topics such as DCM, Bayesian model selection (BMS), or Markov chain Monte Carlo (MCMC) are offered in the supplementary material (see sections S1, S2 and S3).

Thermodynamic integration and the origin of free energy

This section introduces TI from a statistical physics perspective. Statistical physics is a branch of physics that uses methods from probability theory and statistics to characterize the behavior of physical systems. One of the key concepts in statistical physics is that the probability of a particle being in a given state follows a probability density, and that all physically relevant quantities can be derived once this distribution is known. Starting from the free energy, we show how key concepts from information theory have developed from their counterparts in statistical physics, motivating the use of TI and providing a link to the variational Bayes approach conventionally used in DCM to approximate the log model evidence (LME).

Free energy: a perspective from statistical physics

In thermodynamics, the analogue of the model evidence is the so-called partition function Z of a system that consists of an ensemble of particles in thermal equilibrium. A classical discussion of the relationships presented here can be found in Jaynes (1957) and a more modern perspective in Ortega and Braun (2013). For example, let us consider an ideal monoatomic gas, in which the kinetic energy

$$\phi(\theta) = \frac{m\theta^2}{2} \tag{1}$$

of individual particles is a function of their velocity θ and mass m. If the system is large enough, the velocity of a single particle can be treated as a continuous random variable. The internal energy U of this ideal gas is proportional to the expected energy per particle. It is computed as the weighted sum of the energies $\phi(\theta)$ associated with all possible velocities, where the weights are given by the probability $q(\theta)$ of the particle being at a certain velocity:

$$U \stackrel{\text{def}}{=} \int q(\theta)\phi(\theta)d\theta. \tag{2}$$

A second important quantity in statistical physics is the differential entropy H of q:

$$H[q] \stackrel{\text{def}}{=} -k_B \int q(\theta) \ln q(\theta) d\theta. \tag{3}$$

Here, k_B is the Boltzmann constant with units of energy per degree temperature. For an isolated system (i.e., no exchange of matter or energy with the environment), the second law of thermodynamics states that its entropy can only increase or stay constant. Thus, the system is at equilibrium when the associated entropy is maximized, subject to the constraint that the system's internal energy is constant and equal to U, and that q is a proper density, i.e.: $q(\theta) \geq 0$ and $\int q(\theta)d\theta = 1$.

This constrained maximization problem can be solved using Lagrange multipliers (for the derivation see the supplementary material S4), yielding the following distribution:

$$q(\theta) = \frac{1}{Z}\exp\left(-\frac{\phi(\theta)}{k_B T}\right), \tag{4}$$

where Z is referred to as the partition function of the system:

$$Z \stackrel{\text{def}}{=} \int \exp\left(-\frac{\phi(\theta)}{k_B T}\right) d\theta. \tag{5}$$

In a closed system, the Helmholtz free energy F_H is defined as the difference between the internal energy U of the system and its entropy H times the temperature T:

$$F_H \stackrel{\text{def}}{=} U - TH. \tag{6}$$

The Helmholtz free energy corresponds to the work (i.e., non-thermal energy in joules that is passed from the system to its environment) that can be attained from a closed system. From Eq. 6, we see that the system with constant internal energy U is at equilibrium (i.e., maximum entropy) when the Helmholtz free energy is minimal. Substituting the internal energy (Eq. 2), the entropy (Eq. 3) and the expression of q (Eq. 4) into Eq. 6, it follows that the log of the partition function corresponds to the negative Helmholtz free energy divided by $k_B T$:

$$-\frac{F_H}{k_B T} = \ln Z. \tag{7}$$

Free energy: a perspective from statistics

In order to link perspectives on free energy from statistical physics and (Bayesian) statistics, we assume that the

system is examined at a constant temperature T such that the term $k_B T$ equals unity (normalization of temperature), allowing us to move from a physical perspective on free energy (expressed in joules) to a statistical formulation (expressed in information units proportional to bits). This is the common convention in the statistical literature, and thereby, all quantities become unit-less information theoretic terms. Under this convention, the physical concept of free energy described above gives rise to an analogous concept of free energy in statistics when the energy function is given by the negative log joint probability $-\ln p(y, \theta|m)$ (Neal and Hinton 1998):

$$\phi(\theta) = -\ln p(y, \theta|m) = -\ln p(y|\theta, m)p(\theta|m). \quad (8)$$

Hence, the log joint (which fully characterizes the system) takes the role of the kinetic energy in the ideal gas example above.

Inserting the expression for ϕ (Eq. 8) into Eq. 4, we obtain the following expression:

$$q(\theta) = \frac{1}{Z}\exp(-\phi(\theta)) = \frac{1}{Z}\exp(\ln p(y, \theta|m)), \quad (9)$$

which together with the definition of the partition function Z (Eq. 5), reveals that the equilibrium distribution of the system is the posterior distribution (i.e., the joint probability divided by the model evidence):

$$q(\theta) = \frac{\exp(\ln p(y, \theta|m))}{\int \exp(\ln p(y, \theta|m))d\theta} = \frac{p(y, \theta|m)}{p(y|m)} = p(\theta|y, m) \quad (10)$$

Based on this result, we can derive the information theoretic version of the Helmholtz free energy, by inserting the expressions for the internal energy (Eq. 2) and the entropy (Eq. 3) into Eq. 6 and making use of the expression for ϕ from Eq. 8:

$$-F_H = \int q(\theta) \ln p(y, \theta|m)d\theta - \int q(\theta) \ln q(\theta)d\theta, \quad (11)$$

In analogy to Eq. 6, the first term on the right hand side is an expectation over an energy function (cf. Equation 8); while the second term represents the differential entropy $H[q] = -\int q(\theta) \ln q(\theta)d\theta$. Notably, under the choice of the energy function in Eq. 8, the partition function (Eq. 5) corresponds to the marginal of the joint probability $p(y, \theta|m)$ with respect to θ. Comparing with Eq. 7, we see that the negative free energy is equal to the log model evidence (LME):

$$-F_H = \ln p(y|m). \quad (12)$$

Replacing the joint in Eq. 11 by the product of likelihood and prior, the negative free energy can be decomposed into two terms that have important implications for evaluating the goodness of a model:

$$-F_H = \int q(\theta) \ln p(y|\theta, m)p(\theta|m)d\theta - \int q(\theta) \ln q(\theta)d\theta, \quad (13)$$

$$-F_H = \int q(\theta) \ln p(y|\theta, m)d\theta - \int q(\theta) \ln \frac{q(\theta)}{p(\theta|m)}d\theta, \quad (14)$$

$$-F_H = accuracy - complexity \quad (15)$$

The first term (the expected log likelihood under the posterior) represents a measure of model fit or accuracy. The second term corresponds to the Kullback–Leibler (KL) divergence of the posterior from the prior, and can be viewed as an index of model complexity. Hence, maximizing the negative free energy (log evidence) of a model corresponds to finding a balance between accuracy and complexity. We will turn to this issue in more detail below and examine variations of this perspective under TI and VB, respectively.

In the following, we will explicitly display the sign of the negative free energy for notational consistency. In order to highlight similarities with statistical physics and the concepts of energy and potential, we will continue to express the free energy as a functional of a (possibly non-normalized) log density, such that

$$-F_H[\phi] = \ln \int \exp(-\phi(\theta))d\theta, \quad (16)$$

where $\phi(\theta)$ is equivalent to an energy or potential depending on θ. Figure 1 summarizes the conceptual analogies of free energy between statistical physics and Bayesian statistics.

Thermodynamic integration (TI)

We now turn to the problem of computing the negative free energy. As is apparent from Eq. 16, the free energy contains an integral over all possible θ, which is usually prohibitively expensive to compute and thus precludes direct evaluation. The basic idea behind TI is to move in small steps along a path from an initial state with known F_H to the equilibrium state and add up changes in free energy for all steps (Gelman and Meng 1998). This idea was initially introduced in statistical physics to compute the difference in Helmholtz free energy between two states of a physical system (Kirkwood 1935). Other examples for the application of TI in statistical physics are presented in Landau (2015).

In Bayesian statistics, the same idea can be used to compute the LME of a model m. This is because the difference in free energy associated with two potentials corresponding to the negative log prior $\phi_0(\theta) = -\ln p(\theta|m)$ and the negative log joint $\phi(\theta) = -\ln p(y|\theta, m) - \ln p(\theta|m)$ (cp. Eq. 8) equals the LME. More precisely, provided the prior is properly normalized, i.e., $\int p(\theta|m)d\theta = 1$, substituting ϕ_0 and ϕ into Eq. 16 yields

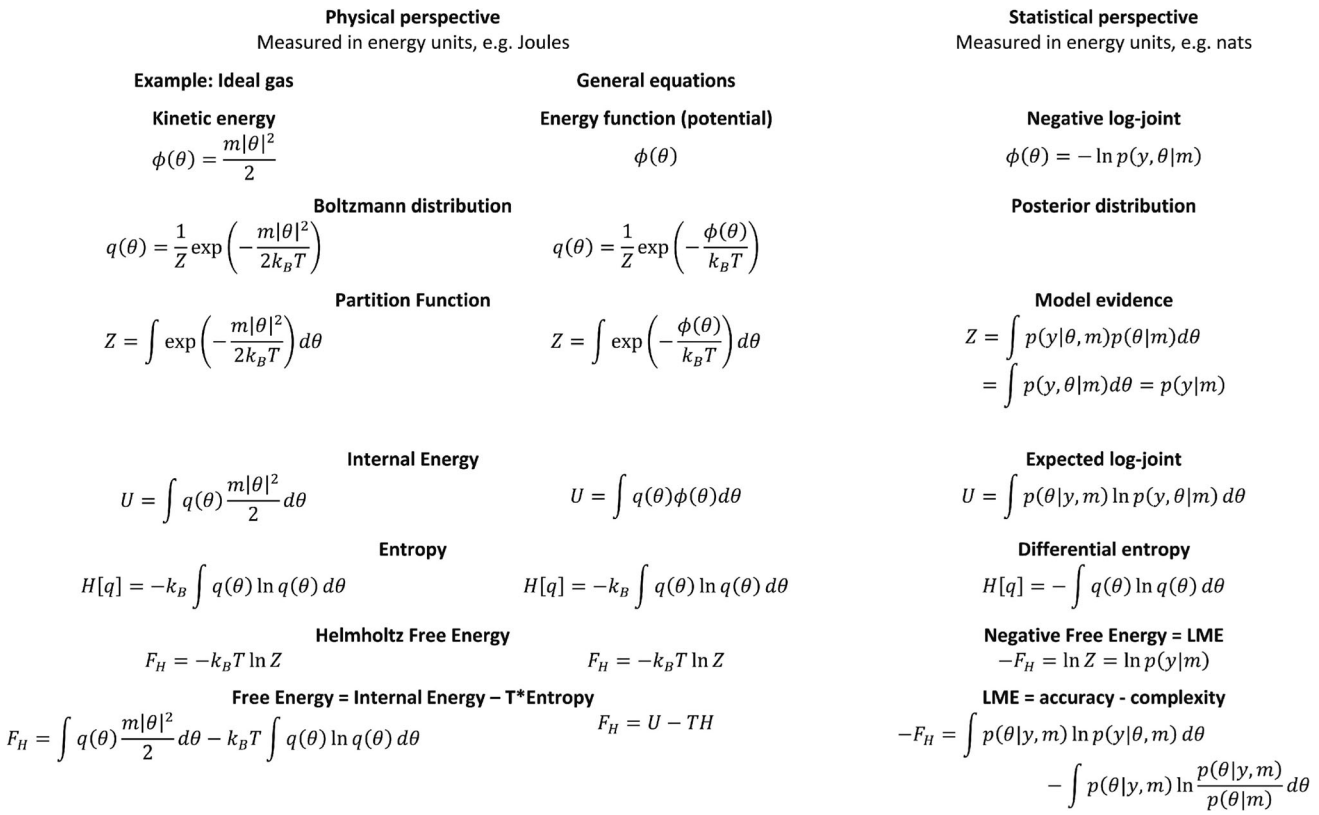

Fig. 1 Analogies between concepts of free energy in statistical physics and Bayesian statistics

$$F_H[\phi] - F_H[\phi_0] = -\ln\int p(y|\theta,m)p(\theta|m)d\theta + \ln\int p(\theta|m)d\theta = -\ln p(y|m), \tag{17}$$

The goal is now to construct a piecewise differentiable path connecting prior and posterior and then compute the LME by integrating infinitesimal changes in the free energy along this path. A smooth transition between $F[\phi]$ and $F[\phi_0]$ can be constructed by the power posteriors $p_\beta(\theta|y,m)$ (see Eq. 19 below) which are defined by the path ϕ_β :

$$\phi_\beta(\theta) = -\beta\ln p(y|\theta,m) - \ln p(\theta|m) \tag{18}$$

with $\beta\epsilon[0,1]$, such that $\phi_1 = \phi$. In the statistics literature, β is usually referred to as an inverse temperature because it has analogous properties to physical temperature in many aspects. We will use this terminology and comment on the analogy in more detail below.

The power posterior is obtained by normalizing the exponential of $-\phi_\beta(\theta)$:

$$p_\beta(\theta|y,m) = \frac{p(y|\theta,m)^\beta p(\theta|m)}{Z_\beta}, \tag{19}$$

$$Z_\beta = \int p(y|\theta,m)^\beta p(\theta|m)d\theta.$$

Combining this definition with Eq. 17, the LME can be expressed as:

$$-\ln p(y|m) = F_H[\phi] - F_H[\phi_0], \tag{20}$$

$$= \int_{\beta=0}^{\beta=1}\frac{d}{d\beta}F_H\left[\phi_\beta\right]d\beta, \tag{21}$$

$$= -\int_{\beta=0}^{\beta=1}\frac{d}{d\beta}\ln\int p(y|\theta,m)^\beta p(\theta|m)d\theta d\beta. \tag{22}$$

Applying the chain rule of differentiation (see supplementary material section S11 for a detailed derivation), the LME can be written in terms of an integral over an expectation with respect to the power posterior:

$$\ln p(y|m) = \int_{\beta=0}^{\beta=1}\int\frac{p(y|\theta,m)^\beta p(\theta|m)}{Z_\beta}\ln p(y|\theta,m)d\theta d\beta, \tag{23}$$

$$= \int_{\beta=0}^{\beta=1}\mathrm{E}[\ln p(y|\theta,m)]_{p_\beta(\theta|y,m)}d\beta. \tag{24}$$

which we refer to as the basic or fundamental TI equation (Gelman and Meng 1998).

Notably, the TI equation can also be understood in terms of the definition of the free energy (Eq. 14) by noting that the latter can be written as the sum of an expected log likelihood and a cross-entropy term (KL divergence of the power posterior from the prior):

$$-F_H(\beta) = \beta \int p_\beta(\theta|y,m) \ln p(y|\theta,m) d\theta - \int p_\beta(\theta|y,m) \ln \frac{p_\beta(\theta|y,m)}{p(\theta|m)} d\theta, \tag{25}$$

$$-F_H(\beta) = \beta A(\beta) - KL\big[p_\beta(\theta|y,m)|p(\theta|m)\big]. \tag{26}$$

The first term, $A(\beta) = -\partial F_H/\partial\beta$, is referred to as the accuracy of the model (see, for example, Penny et al. 2004a; Stephan et al. 2009), while the second term constitutes a complexity term. Note that Eq. 26 is typically presented in the statistical literature for the case of $\beta = 1$ and describes the same accuracy vs. complexity trade-off previously expressed by Eq. 14, but now from the specific perspective of TI. Also note that $A(\beta)$ is defined as the negative partial derivative of the free energy. In contrast to the full derivative, the partial derivative only considers the direct dependence of F_H on β, and ignores the indirect dependence via the KL divergence term.

Figure 2 shows a graphical representation of the relation conveyed by the fundamental TI equation (Eqs. 24) and 26. For any given β, the negative free energy at this position of the path $-F_H(\beta)$ can be interpreted as the signed area below the curve $A(\beta) = -\partial F_H/\partial\beta$ (i.e., the integral over $A(\cdot)$ from 0 to β), whereas the term $\beta \times A(\beta)$ is the rectangular area below $A(\beta)$. Equation 26 shows that the area $\beta A(\beta) + F_H(\beta)$ is the KL divergence of the corresponding power posterior from the prior.

This relationship holds because, for the power posteriors (Eq. 19), $A(\beta)$ is a monotonically increasing function of β. This is due to the fact that

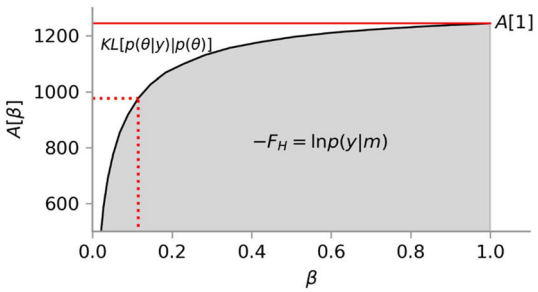

Fig. 2 Graphical representation of the TI equation. The free energy is equal to the *signed* area below $A = -\partial F_H/\partial\beta$, and thus the area $A(1) + F_H$ is equal to the KL divergence of the posterior from the prior. The same relation holds for any $\beta \in [0, 1]$

$$\frac{\partial A(\beta)}{\partial\beta} = Var[\ln p(y|\theta,m)]_{p_\beta(\theta|y,m)} > 0. \tag{27}$$

See Lartillot and Philippe (2006) for a derivation of this property. From this it follows that the negative free energy is a concave function along β.

The theoretical considerations highlighted above and the relation to principles of statistical physics render TI an appealing choice for estimating the LME. However, the question remains how the LME estimator in Eq. 24 can be evaluated in practice. To solve this problem, TI relies on Monte Carlo estimates of the expected value $E[\ln p(y|\theta,m)]_{p_\beta(\theta|y,m)}$ in Eq. 24:

$$E_{MC}(\beta) := \frac{1}{K}\sum_{k=1}^{K} \ln p(y|\theta_k,m) \approx E[\ln p(y|\theta,m)]_{p_\beta(\theta|y,m)}, \tag{28}$$

where samples θ_k are drawn from the power posterior $p_\beta(\theta|y,m)$. The remaining integral over β in Eq. 24 is a one dimensional integral, which can be computed through a quadrature rule using a predefined temperature schedule for β ($0 = \beta_0 < \beta_1 < \cdots < \beta_{N-1} < \beta_N = 1$):

$$\ln p(y|m) \approx \frac{1}{2}\sum_{j=1}^{N-1} (\beta_{j+1} - \beta_j)\big(E_{MC}(\beta_{j+1}) + E_{MC}(\beta_j)\big). \tag{29}$$

The optimal temperature schedule in terms of minimal variance of the estimator and minimal error introduced by this discretization was outlined previously in the context of linear models by Gelman and Meng (1998) and Calderhead and Girolami (2009).

Note that each step β_j in the temperature schedule requires a new set of samples θ_k to be drawn from the respective power posterior $p_{\beta_j}(\theta|y,m)$, contributing to the high computational complexity of TI. However, since these sets of samples are independent from each other, this can in principle be done in parallel, provided suitable soft- and hardware capabilities are available. An efficient way to realize such a parallel sampling procedure is to adopt a population MCMC approach in which MCMC sampling is used to generate, for each β_j, a chain of samples from the respective power posterior $p_{\beta_j}(\theta|y,m)$. In addition, chains from neighboring β_j in the temperature schedule are allowed to interact by means of a "swap" accept-reject (AR) step (Swendsen and Wang 1986). This increases the sampling efficiency and speeds up convergence of the individual MCMC samplers. For readers unfamiliar with Monte Carlo methods, a primer on (population) MCMC is provided in the supplementary material S3. For a detailed treatment, we refer to McDowell et al. (2008) and Calderhead and Girolami (2009).

So far, the computational requirement of sampling from an ensemble of distributions (one for each value of β) has limited the application of TI to high performance computing environments and prevented its widespread use in neuroimaging. Luckily, the increase in computing power of stand-alone workstations and the proliferation of graphical processing units (GPU), coupled with efficient population MCMC samplers, offer possibilities to overcome this bottleneck, which will be demonstrated below for a selection of three examples involving synthetic and real-world datasets. First, however, we will complete the theoretical overview by briefly explaining the formal relationship between TI and variational Bayes.

Variational bayes

Variational Bayes (VB) is a general approach to approximate intractable integrals with tractable optimization problems. Importantly, this optimization method simultaneously yields an approximation to the posterior density and a lower bound to the LME.

The fundamental equality which underlies VB is based on introducing a tractable density $q(\theta)$ to approximate the posterior $p(\theta|y,m)$.

$$-F_H = \ln p(y|m) = \int q(\theta) \ln \frac{p(y|m)q(\theta)}{q(\theta)} d\theta, \tag{30}$$

$$= \int q(\theta) \ln \frac{p(y,\theta|m)q(\theta)}{p(\theta|y,m)q(\theta)} d\theta, \tag{31}$$

$$= \underbrace{\int q(\theta) \ln p(y|\theta,m) d\theta}_{Approx.accuracy} - \underbrace{\int q(\theta) \ln \frac{q(\theta)}{p(\theta|m)} d\theta}_{Approx.complexity} + \underbrace{\int q(\theta) \ln \frac{q(\theta)}{p(\theta|y,m)} d\theta}_{Error} \tag{32}$$

The last term in Eq. 32 is the KL divergence of the approximate density q from the unknown posterior density; this encodes the error or inaccuracy of the approximation. Given that the KL divergence is never negative, the first two terms in Eq. 32 represent a lower bound on the log evidence $-F_H$, and in the following we will refer to it as the negative variational free energy $-F_{VB}$.

In summary, the relation between the information theoretic version oHelmholtz free energy $-F_H$, log model evidence $\ln p(y|m)$, and variational negative free energy $-F_{VB}$ is therefore

$$-F_H = \ln p(y|m) = -F_{VB} + \mathrm{KL}[q(\theta)||p(\theta|y,m)] \tag{33}$$

We highlight this relationship because many readers are rightfully confused that the term 'negative free energy' is sometimes used in the literature to denote the logarithm of the partition function Z itself (i.e., $-F_H$), as we have done above, and sometimes to refer to a lower bound approximation of it (i.e., $-F_{VB}$). This is because the variational free energy $-F_{VB}$ becomes identical to the negative free energy $-F_H$ when the approximate density q equals the posterior and hence their KL divergence becomes zero. In this special case

$$\max_q [-F_{VB}[q]] = -F_H. \tag{34}$$

To maintain consistency in the notation, we will distinguish $-F_H$ and $-F_{VB}$. throughout the paper.

VB aims to reduce the KL divergence of q from the posterior density by maximizing the lower bound $-F_{VB}$ as a functional of q:

$$-F_{VB}[q] = \int q(\theta) \ln p(y|\theta,m) d\theta - \int q(\theta) \ln \frac{q(\theta)}{p(\theta|m)} d\theta. \tag{35}$$

When the functional form of q is fixed and parametrized by a vector η, VB can be reformulated as an optimization method in which η is updated according to gradient $\partial F_{VB}[q(\theta|\eta)]/\partial\eta$ (Friston et al. 2007). Thus, the path followed by η during optimization can be formulated as

$$\dot{\eta} = -\frac{\partial F_{VB}[q(\theta|\eta)]}{\partial \eta}. \tag{36}$$

This establishes a connection between TI and VB. In the former, the path of η corresponds to the path of β from 0 to 1, which was selected a priori with the conditions that

$$q(\theta|\beta = 0) = p(\theta), q(\theta|\beta = 1) \propto p(y|\theta)p(\theta) \tag{37}$$

and the gradients

$$-\frac{\partial F_H[q(\theta|\beta)]}{\partial \beta} \tag{38}$$

are used to numerically compute the free energy.

Different VB algorithms are defined by the particular functional form used for the approximate posterior. In the case of DCM, it is so far most common to use Variational Bayes under the Laplace approximation (VBL). A summary of VBL for DCM is available in the supplementary material S5, while an in-depth treatment is provided in Friston et al. (2007).

Evaluating the accuracy of TI

In this section, we investigate the accuracy of LME estimates obtained with TI, and compare the performance of TI to that of two other sampling-based LME estimators, the prior arithmetic mean (AME) and posterior harmonic mean (HME) estimators. In contrast to TI, which requires sampling from an ensemble of distributions (see Eq. 29), AME

and HME only require samples from the prior or posterior distributions, respectively. Hence, these two methods, which are described in detail in the supplementary material S6, are computationally significantly less demanding than TI.

For the purpose of this comparison, we turn to a Bayesian linear regression model with normal prior and likelihood. This is a useful case for benchmarking because the LME can be computed analytically. This model is described by the following prior and likelihood function:

$$\begin{aligned} p(\theta) &= N\left(\theta; 0, \Pi_p^{-1}\right), \\ p(y|\theta) &= N\left(y; X\theta, \Pi_e^{-1}\right), \end{aligned} \tag{39}$$

where θ is the $[p \times 1]$ vector of regression coefficients, y is the $[M \times 1]$ vector of data points, X is the $[M \times p]$ design matrix, and Π_p^{-1} and Π_e^{-1} are the covariance matrices of the prior and errors, respectively. The analytic solution for the LME of this model is given by Penny (2012) as:

$$\begin{aligned} \ln \int p(y|\theta)p(\theta)d\theta = 0.5(&-\ln|\Pi| - N\ln 2\pi + \ln|\Pi_e| + \ln|\Pi_p| \\ &- (y - X\eta)^T\Pi_e(y - X\eta) - \eta^T\Pi_p\eta), \end{aligned} \tag{40}$$

$$\Pi = \Pi_p + X^T\Pi_e X, \tag{41}$$

$$\eta = \Pi^{-1}X^T\Pi_e y \tag{42}$$

For our simulations, we chose $M = 100$, $\Pi_p^{-1} = 16I_p$ and $\Pi_e^{-1} = 10I_e$, where I_p and I_e are the corresponding identity matrices. The design matrix was chosen to have a block structure equivalent to a design for a one-way ANOVA with p levels (for those values of p that do not exactly divide by M, the excess data points were assigned to the last cell). Synthetic data where generated by sampling from the generative model defined in Eq. 39.

By varying p from 2 to 32 in steps of 1, we created a series of models with increasing dimensionality. For each value of p, we repeated the data generation process 10 times, drawing a new set of values for the regression parameters θ each time from the prior, and generating observations y according to the likelihood. We then estimated the LME using TI, AME and HME, and compared the estimates against the analytically computed LME.

The TI approximation to the LME was computed using 64 chains with a 5th order annealing schedule, i.e. a temperature schedule with 64 steps β_j, with step size chosen according to a fifth order power rule (Calderhead and Girolami 2009). In each chain, we generated 6000 samples. We then computed AME based on the samples from the prior density and HME based on samples from the posterior. Figure 3 shows the error in the LME estimates as a function of the number of model parameters for the three approaches. Consistent with previous reports, the results show that HME overestimated the LME, while AME underestimated it (Lartillot and Philippe 2006). Only TI provided good estimates over the full range of models. This indicates that, despite comparing unfavorably in terms of computational efficiency, TI should still be preferred in practice due to the large estimation error of AME and HME, especially for higher-dimensional models (but see Penny and Sengupta (2016) for variants of AME and HME that solve some of the issues).

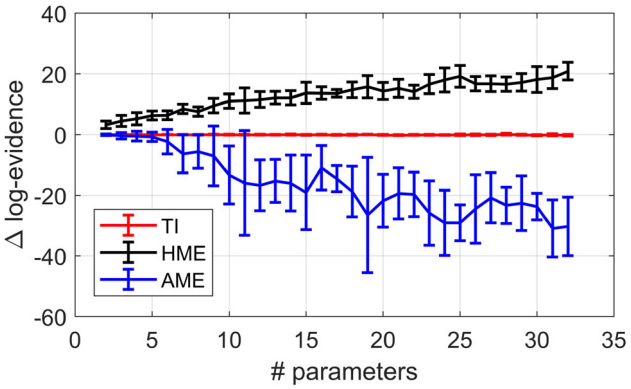

Fig. 3 Error in estimating the log evidence of linear models for three different sampling approaches. The curves show mean and standard deviation (error bars) over ten runs at each value of p (number of GLM parameters) for thermodynamic integration (TI), posterior harmonic mean estimator (HME) and prior arithmetic mean estimator (AME)

Application to DCM

Having established the accuracy of TI as an estimator for the LME in a case where the analytic solution for the LME is available, we now turn to the case of LME estimation and model selection in the context of DCM. For this purpose, we discuss two example applications. The first example considers a simulated dataset where the true model that generated each observation is known. This serves to determine the ability of TI to identify the data-generating model. In the second example, we analyze the "attention to motion" fMRI dataset (Buchel 1997), which has been analyzed by numerous previous methodological studies. Primers on DCM and Bayesian model selection are provided in the supplementary material.

DCM: simulated data

In the first experiment, we used simulated data from 5 DCMs (linear: model 1; bilinear: models 2–4; nonlinear: model 5) with two inputs (u_1 and u_2). The DCMs are

displayed in Fig. 4 and are available for download via the ETH Research Collection (ETH Zurich 2020). The numerical values of the connectivity matrices are listed in the supplementary material S7. The BOLD signal data were simulated assuming a repetition time (TR) = 2 s and 720 scans per simulation. The driving inputs were entered with a sampling rate of $2.0Hz$. Simulated time series were corrupted with Gaussian noise yielding a signal-to-noise ratio (SNR) of 1.0. Here, SNR was defined as the ratio of signal standard deviation to noise standard deviation (Welvaert and Rosseel 2013). This means that our simulated data contained identical amounts of noise and signal, representing a relatively challenging SNR scenario.

For each model, we generated 40 different datasets with different instantiations of Gaussian noise, such that the underlying time series remained constant. We then counted how often the data-generating model was assigned the largest model evidence and compared the ensuing values across the different estimators (i.e., AME, HME, TI, VBL). Notably, the absolute value of the log evidence of a given model is irrelevant for model scoring; instead, its difference to the log evidence of other models is decisive.

In a pretesting phase, we found that TI generated stable estimates of the LME using 64 chains. All simulations were executed with a burn-in phase of 1×10^4 samples, followed by an additional 1×10^4 samples used for analysis. We evaluated the convergence of the MCMC algorithm by examining the potential scale reduction factor $\hat{R}$ (Gelman & Rubin 1992) for samples of the log likelihood of all chains. We found that $\hat{R}$ was below 1.1 in all but a few instances, indicating convergence. Estimated LME values are displayed in Fig. 5 and Table 1. Consistent with the linear model analysis in the previous section, the HME was always higher and the AME always lower than the TI estimate of the LME. VBL estimates were close to the TI estimate. To test for significant differences in accuracy of recovering the correct model by the different algorithms, χ^2 tests were employed (inference method (i.e., TI, VB, HME, AME) vs inference result (i.e., number of times correct and incorrect model was selected)). TI and VBL were not significantly different ($\chi^2 = 0.3, p = 0.56$), but both TI and VBL (not shown) were significantly better than AME ($\chi^2 = 189.5, p < 10^{-5}$) and HME ($\chi^2 = 25.4, p < 10^{-5}$).

We then examined how often the data-generating model was identified correctly by model comparison, i.e., how often it showed the largest LME of all models. Of all estimators, AME failed most frequently to detect the data-generating model (Table 2). HME identified the correct

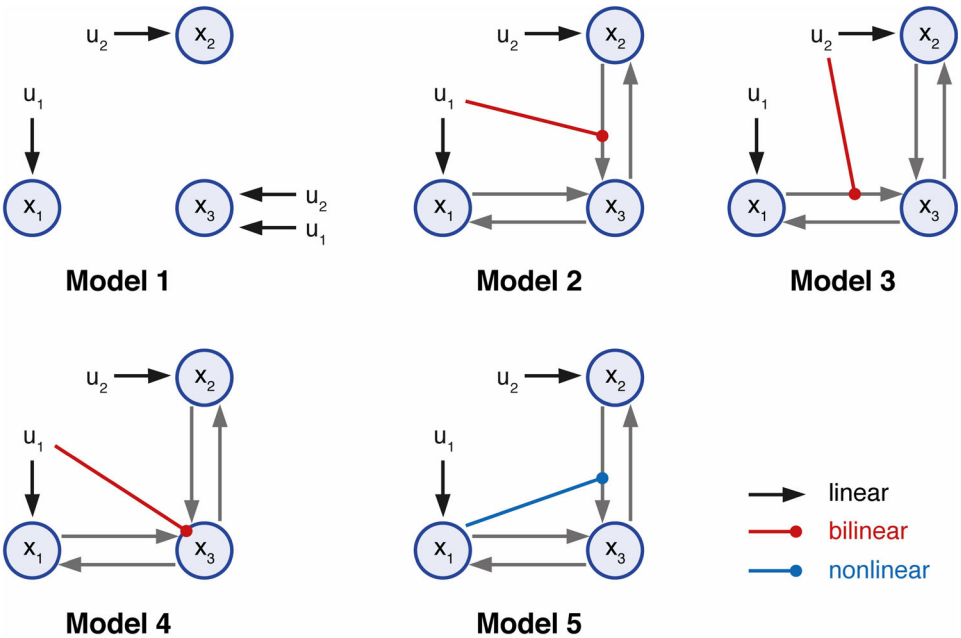

Fig. 4 Illustration of the five simulated 3-region DCMs used for cross-model comparison. Self-connections are not displayed. The variables u_1 and u_2 represent two different experimental conditions or inputs. All models represented different hypotheses of how the neuronal dynamics in area x_3 could be explained in terms of the two driving inputs and the effects of the other two regions x_1 and x_2. Model m_1 can be understood as a 'null hypothesis' in which the activity of all the areas can be explained by the driving inputs. Models m_2 and m_3 correspond to two forms of bilinear effect on the forward connection of areas x_1 and x_2. Model m_4 represents the hypothesis that input u_1 affects the self-connection of area x_3 (not displayed). Model m_5 represents a non-linear interaction between regions x_1 and x_2. Endogenous connections are depicted by gray arrows, driving inputs by black arrows, bilinear modulations by red arrows and nonlinear modulations by blue arrows. (Color figure online)

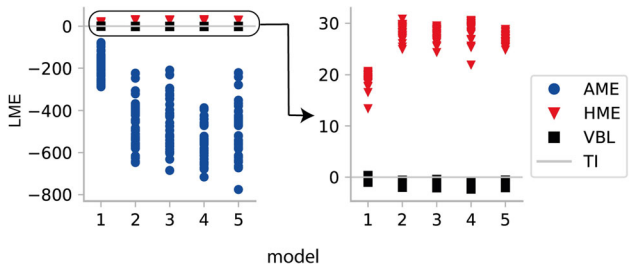

Fig. 5 Estimated LME for all models relative to TI when inverted with the corresponding data-generating model under *SNR* = 1for 40 different models. Right panel zooms in the left panel. Red triangles correspond to the HME, blue circles to the AME, and black squares to VBL. HME was always higher and AME always lower than the TI estimate. All LME estimates are shown after subtracting the TI-based estimate for the same model

Table 3 Cross-model comparison results for HME in the case of synthetic data (SNR = 1)

HME: synthetic data					
	Generation				
	m_1	m_2	m_3	m_4	m_5
Inversion					
m_1	39				
m_2		40			12
m_3			40		12
m_4				40	4
m_5	1				12

The row label indicates the data-generating model, the column index is the inferred model

Table 1 LME estimated with TI and VBL

	TI					VBL				
	Generating model					Generating model				
	m_1	m_2	m_3	m_4	m_5	m_1	m_2	m_3	m_4	m_5
Inverting model										
m_1	**762**	0	0	0	0	**746**	18	16	− 3	11
m_2	10	**11155**	10676	4242	4359	− 34	**11101**	10614	4182	4294
m_3	6	10360	**11683**	2903	4349	− 39	10298	**11627**	2842	4288
m_4	0	10038	9102	**5163**	4376	− 49	9984	9028	**5097**	4310
m_5	84	9045	9190	2931	**4430**	44	9034	9181	2877	**4375**

Tables display the LME (summed across 40 simulations) of each combination of inverting and generating models. Columns have been normalized by the lowest LME: according to TI. Columns on the right and left tables share the same normalization and their absolute values can be directly compared. On most, but not all occasions, VBL underestimated the LME compared to TI. However, for both VBL and TI the data-generating model obtained the highest LME (marked in bold)

Table 2 Cross-model comparison results for AME in the case of synthetic data (SNR = 1)

HME: synthetic data					
	Generation				
	m_1	m_2	m_3	m_4	m_5
Inversion					
m_1	39	14	10	34	29
m_2	1	13	11		3
m_3		6	11	3	
m_4		3	4	2	5
m_5		4	4	1	3

The row label indicates the data-generating model, the column index is the inferred model

Table 4 Cross-model comparison results for VBL in the case of synthetic data (SNR = 1)

VBL: synthetic data					
	Generation				
	m_1	m_2	m_3	m_4	m_5
Inversion					
m_1	40				
m_2		40			
m_3			40		
m_4				40	1
m_5					39

The row label indicates the data-generating model, whereas the column index is the inferred model

model more consistently (Table 3). Both VBL and TI displayed a similar behavior (Tables 4 and 5), although model m_5 was identified slightly more consistently identified by VBL. However, as displayed in Table 1, according

Table 5 Cross-model comparison results for TI in the case of synthetic data (SNR = 1).

TI: synthetic data					
	Generation				
	m_1	m_2	m_3	m_4	m_5
Inversion					
m_1	40				
m_2		40			
m_3			40		
m_4				40	2
m5					38

The row label indicates the data-generating model, the column index is the inferred model

to both inversion schemes, the data-generating model was consistent with the model showing the highest LME.

Empirical data: attention to motion

In this example, we demonstrated TI-based parameter estimation and model comparison for DCM on an empirical dataset. Since the previous two examples have shown that TI consistently outperforms the other sampling-based LME estimators, AME and HME, we limit our comparison to TI and VB, from here on.

For the analysis of empirical data, we selected the "attention to motion" fMRI dataset (Buchel 1997) that has been analyzed in numerous previous methodological studies (e.g., Friston et al. 2003; Marreiros et al. 2008; Penny et al. 2004a; Penny et al. 2004b; Stephan et al. 2008). The original study investigated the effect of attention on motion perception (Buchel 1997); in particular, the authors examined attentional effects on the connectivity between primary visual cortex (V1), motion-sensitive visual area (V5) and posterior parietal cortex (PPC). In brief, the experimental paradigm consisted of four conditions (all under constant fixation): fixation only (F), presentation of stationary dots (S), passive observation of radially moving dots (N), or attention to the speed of these dots (A). Four sessions were recorded and concatenated, yielding a total of 360 volumes ($T_E = 40ms, TR = 3.22s$). Three inputs were constructed using a combination of the three conditions: $stimulus = S + N + A$, $motion = N + A$, $attention = A$. Driving inputs were resampled at $0.8Hz$, requiring a total of 1440 integration steps. Further details of the experimental design and analysis can be found in Buchel (1997).

One reason for selecting this dataset is that Stephan et al. (2008) previously demonstrated that a nonlinear model (model 4 in Fig. 6) had higher evidence than comparable bilinear models (model 1–3 in Fig. 6). This case is of interest for evaluating the quality of different LME estimators, as one would expect that the introduction of nonlinearities represents a challenging case for VBL.

For TI-based LME estimation, 16×10^3 samples were collected from 64 chains, of which 8×10^3 were discarded in the burn-in phase. The convergence of the algorithm was evaluated using the $\hat{R}$ statistic of the samples of the log likelihood of each chain and model. In all but one chain, $\hat{R}$ was below 1.1, indicating convergence.

Table 6 summarizes the evidence estimates obtained with TI and VBL. In comparison to previous results see Table 8 in Stephan et al. (2008), three findings are worth highlighting. First, as shown in Table 6, the VBL algorithm reproduced the ranking of models reported in Stephan et al. (2008), although an earlier version of the VBL algorithm with different prior parameters and a different integration scheme was used by Stephan et al. (2008). Moreover, our TI implementation produced the same ranking as the one obtained under VBL.

Second, the difference between the VBL free energy estimates and the TI estimates varied considerably across models. To investigate this variability, we compared TI and VBL with regard to the accuracy term. The results are summarized in the lower section of Table 7. Table 6 shows that the discrepancies between VBL and TI varied across models, and the difference was particularly pronounced for the nonlinear model m_4 (> 40 log units).

Third, while VBL detected the most plausible model, the findings from this dataset suggest that VBL-based inversion of DCMs might not always be fully robust. In particular, the difference between the algorithms could be attributed to the VBL algorithm converging to a local extremum. To assess the differences between TI and VBL more systematically, we initialized each algorithm 10 times from different starting values that were randomly sampled from the prior density. Figure 7 depicts the estimated model evidence and accuracy and Fig. S1 in the supplementary material S10 displays the predicted BOLD signal. VBL estimates of the accuracy and LME displayed much larger variance than the TI estimates. This suggests that the greater variance of the VBL estimates is due to the propensity of the gradient ascent used in VBL to converge to local maxima.

The observations listed above highlight two important challenges faced by VBL. Due to restricting the approximate posterior to a normal distribution, the negative free energy obtained with VBL is a lower bound approximation and hence, will always be smaller than the actual log-model evidence, especially for nonlinear models with non-normal posteriors. Independently from this, the presence of local maxima in the optimization process means that VBL

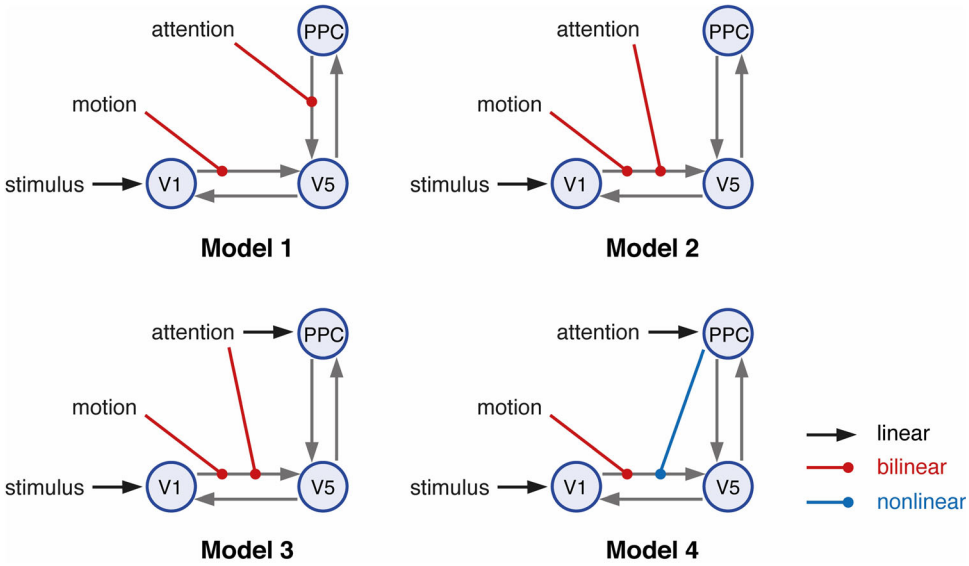

Fig. 6 Illustration of the four models used in Stephan et al. (2008) representing different hypotheses of the putative mechanisms underlying attention-related effects in the motion-sensitive area V5. The first three models are bilinear whereas the fourth model is a nonlinear DCM. Endogenous connections are depicted by gray arrows, driving inputs by black arrows, bilinear modulations by red arrows and nonlinear modulations by blue arrows. Inhibitory self-connections are not displayed. V1: primary visual area, V5 = motion sensitive visual area, PPC: posterior parietal cortex. (Color figure online)

Table 6 Results of model comparison, in terms of log evidence differences with respect to the worst model (m_1), from Stephan et al. (2008), who used a different prior and integrator as in here

m_1	m_2	m_3	m_4
Stephan et al. (2008) *(VBL)*			
0.0	3.1	5.6	13.6
VBL			
0.0	11.4	13.4	15.2
TI			
0.0	11.5	14.8	43.5

Table 7 Log model evidence, accuracy and log likelihood at the MAP estimate using both TI and VBL

Attention to motion dataset				
	m_1	m_2	m_3	m_4
Log model evidence				
VBL	− 1790.0	− 1778.6	− 1776.6	− 1774.8
TI	− 1772.6	− 1761.1	− 1757.8	− 1729.1
Accuracy				
VBL	− 1547.6	− 1538.5	− 1531.6	− 1530.7
TI	− 1525.6	− 1520.2	− 1511.8	− 1483.5

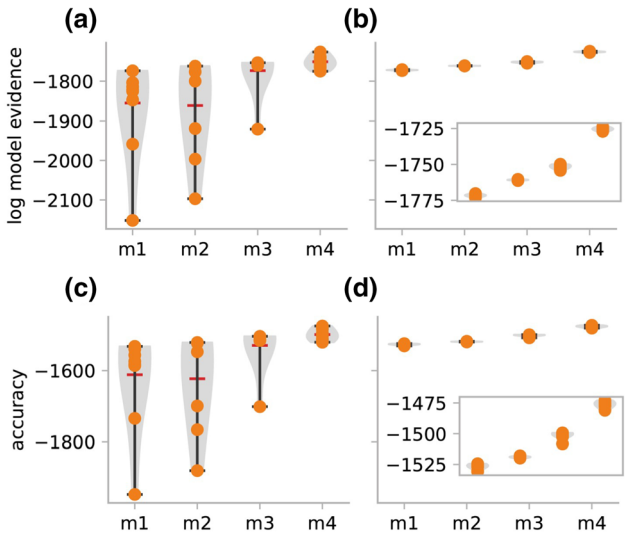

Fig. 7 Estimates of the LME and accuracy in the attention to motion dataset after initializing VBL and TI from 10 different starting points (yellow points) drawn from the prior. The inset on the right panel zooms into the range of TI estimates. **a** LME estimates from VBL. **b** LME estimates from TI. **c** Accuracy component of the LME estimates from VBL. **d** Accuracy component of the LME estimates from TI. The results demonstrate that TI estimates show much lower variability as compared to VBL estimates. (Color figure online)

may in practice even fail to find the best lower bound, i.e. the global maximum of the negative free energy. TI is able to address both these issues, which will be important for performing reliable subject-level inference.

Conclusion

In this paper, we have reviewed the theoretical foundation of thermodynamic integration. In the process, we have introduced the concept of free energy, which has found its

way into information theory and Bayesian statistics from its origin in statistical mechanics. Approaching TI from this dual perspective allowed us to highlight the parallels and analogous concepts shared between these different scientific fields.

A key result was obtained in Eq. 24 (the TI equation), which provided (1) a graphical interpretation of the LME as the signed area under the curve given by the accuracy $A(\beta) = -\partial F_H/\partial\beta$; and (2) a reliable method for estimating LME via Monte Carlo samples drawn from the power posteriors. The application of this method was demonstrated in the second part of this paper on synthetic and real-world datasets.

Specifically, we started with an experiment involving synthetic data from a linear regression model with analytical solutions for LME. This experiment demonstrated that TI produces accurate LME estimates and outperforms computationally less complex sampling-based LME estimators (AME and HME), justifying the additional complexity.

Finally, we used synthetic and real-world fMRI data to compare TI to VB, which is the current gold standard in the context of model inversion and LME estimation for DCM. Although VB was robust in most instances, we found evidence for variability in the estimates due to local optima in the objective function—especially in the case of the real-world dataset, where the model space included nonlinear DCMs, and for challenging scenarios where the number of network nodes and free parameters is high. While this problem can be ameliorated by initializing the VB algorithm from different starting points or using global optimization methods (see Lomakina et al. 2015), this would reduce computational efficiency, which is the main justification for VB as the default choice for standard applications of DCM.

Hence, sampling-based approaches like TI might become the method of choice when the robustness and validity of single-subject inference is paramount. For example, the utility of generative models for clinical applications, such as differential diagnosis based on model comparison or prediction of individual treatment responses (Stephan et al. 2017), depends on our ability to draw reliable and accurate conclusions from model-based estimates.

In addition, the experiments presented in this paper also demonstrated the practical feasibility of applying TI to complex generative models like DCM, which are characterized by high computational cost for evaluation of the likelihood function. This is made possible by an implementation that relies on parallel computing techniques, offering reasonable execution times on stand-alone workstations. Specifically, the computations for this paper were performed on a workstation equipped with an Intel Core i7 4770 K (CPU) and a Nvidia Geforce GTX 1080 (GPU), with a software implementation that allow obtaining as many as 10^5 samples of realistic DCMs in only a few minutes. Here, the important implication is that TI is no longer a method that is exclusively reserved for users with access to high performance computing clusters. The TI and DCM implementations used in this paper is available to the community as open source software (Translational Neuromodeling Unit 2014).

Finally, we would like to point out that although this paper mainly focuses on the estimation of the model evidence, which is the intended purpose of TI, TI also provides samples from the posterior distribution over model parameters. This is due to the fact that TI's temperature schedule always includes $\beta_N = 1$ (cf. Equation 29), meaning that the last power posterior being sampled from is equivalent to the posterior distribution. While this is conceptually different from the parametric distribution which VB provides as an approximation to the true posterior distribution, the posterior samples obtained by TI can be used to calculate summary statistics, such as posterior mean and variance or Bayesian credible intervals for the DCM parameters. This enables the user to obtain quantitative estimates of DCM parameters, such as the connection strength between brain regions or the strength of contextual modulations, in addition to the inference over network structure based on the comparison of LME.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11571-021-09696-9.

Funding Open Access funding provided by ETH Zurich. This work was funded by the René and Susanne Braginsky Foundation (KES), the Clinical Research Priority Program "Multiple Sclerosis" (KES), the Swiss National Science Foundation, grant number 320030_179377 (KES) and the ETH Zurich Postdoctoral Fellowship Program and the Marie Curie Actions for People COFUND Program (SF).

References

Annis J, Evans NJ, Miller BJ, Palmeri TJ (2019) Thermodynamic integration and steppingstone sampling methods for estimating

Bayes factors: a tutorial. J Math Psychol 89:67–86. https://doi.org/10.1016/j.jmp.2019.01.005

Aponte EA, Raman S, Sengupta B, Penny W, Stephan KE, Heinzle J (2016) mpdcm: a toolbox for massively parallel dynamic causal modeling. J Neurosci Methods 257:7–16. https://doi.org/10.1016/j.jneumeth.2015.09.009

Bishop C (2006) Pattern recognition and machine learning. Springer, Cambridge

Buchel C (1997) Modulation of connectivity in visual pathways by attention: cortical interactions evaluated with structural equation modelling and fMRI. Cerebral Cortex 7(8):768–778. https://doi.org/10.1093/cercor/7.8.768

Calderhead B, Girolami M (2009) Estimating Bayes factors via thermodynamic integration and population MCMC. Comput Stat Data Anal 53:4028–4045. https://doi.org/10.1016/j.csda.2009.07.025

Chumbley JR, Friston KJ, Fearn T, Kiebel SJ (2007) A Metropolis-Hastings algorithm for dynamic causal models. NeuroImage 38(3):478–487. https://doi.org/10.1016/j.neuroimage.2007.07.028

Daunizeau J, David O, Stephan KE (2011) Dynamic causal modelling: a critical review of the biophysical and statistical foundations. Neuroimage 58(2):312–322. https://doi.org/10.1016/j.neuroimage.2009.11.062

David O, Kiebel SJ, Harrison LM, Mattout J, Kilner JM, Friston KJ (2006) Dynamic causal modeling of evoked responses in EEG and MEG. Neuroimage 30(4):1255–1272. https://doi.org/10.1016/j.neuroimage.2005.10.045

ETH Zurich (2020). ETH Research Collection. Retrieved from https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/301664/simulation_dcms.zip

Friston KJ, Harrison L, Penny W (2003) Dynamic causal modelling. Neuroimage 19(4):1273–1302. https://doi.org/10.1016/S1053-8119(03)00202-7

Friston KJ, Mattout J, Trujillo-Barreto N, Ashburner J, Penny W (2007) Variational free energy and the Laplace approximation. Neuroimage 34(1):220–234. https://doi.org/10.1016/j.neuroimage.2006.08.035

Gelman A, Meng XL (1998) Simulating normalizing constants: from importance sampling to bridge sampling to path sampling. Stat Sci 13(2):163–185

Gelman A, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Stat Sci 7(4):457–472. https://doi.org/10.1214/ss/1177011136

Heinzle J, Koopmans PJ, den Ouden HEM, Raman S, Stephan KE (2016) A hemodynamic model for layered BOLD signals. NeuroImage 125:556–570. https://doi.org/10.1016/j.neuroimage.2015.10.025

Jaynes ET (1957) Information theory and statistical mechanics. Phys Rev 106(4):620–630. https://doi.org/10.1103/physrev.106.620

Kirkwood JG (1935) Statistical mechanics of fluid mixtures. J Chem Phys 3(5):300–313. https://doi.org/10.1063/1.1749657

Landau DP (2015) A guide to monte carlo simulations in statistical physics. University Press, Cambridge

Lartillot N, Philippe H (2006) Computing bayes factors using thermodynamic integration. Syst Biol 55(2):195–207. https://doi.org/10.1080/10635150500433722

Lomakina EI, Paliwal S, Diaconescu AO, Brodersen KH, Aponte EA, Buhmann JM, Stephan KE (2015) Inversion of hierarchical bayesian models using gaussian processes. Neuroimage 118:133–145. https://doi.org/10.1016/j.neuroimage.2015.05.084

MacKay DJC (2004) Information theory, inference, and learning algorithms. University Press, Cambridge

Marreiros AC, Kiebel SJ, Friston KJ (2008) Dynamic causal modelling for fMRI: a two-state model. NeuroImage 39(1):269–278. https://doi.org/10.1016/j.neuroimage.2007.08.019

McDowell JE, Dyckman KA, Austin BP, Clementz BA (2008) Neurophysiology and neuroanatomy of reflexive and volitional saccades: evidence from studies of humans. Brain Cogn 68(3):255–270. https://doi.org/10.1016/j.bandc.2008.08.016

Moran R, Pinotsis DA, Friston K (2013) Neural masses and fields in dynamic causal modeling. Front Comput Neurosci 7:57–57. https://doi.org/10.3389/fncom.2013.00057

Neal RM, Hinton GE (1998) A view of the em algorithm that justifies incremental, sparse, and other variants. In: Jordan MI (ed) Learning in graphical models. Springer, Dordrecht, pp 355–368

Ortega PA, Braun DA (2013) Thermodynamics as a theory of decision-making with information-processing costs. Proc R Soc A Math Phys Eng Sci 469(2153):20120683. https://doi.org/10.1098/rspa.2012.0683

Penny W (2012) Comparing dynamic causal models using AIC BIC and free energy. Neuroimage 59(1):319–330. https://doi.org/10.1016/j.neuroimage.2011.07.039

Penny W, Sengupta B (2016) Annealed importance sampling for neural mass models. PLoS Comput Biol 12(3):e1004797–e1004797. https://doi.org/10.1371/journal.pcbi.1004797

Penny W, Stephan KE, Mechelli A, Friston KJ (2004a) Comparing dynamic causal models. Neuroimage 22(3):1157–1172. https://doi.org/10.1016/j.neuroimage.2004.03.026

Penny W, Stephan KE, Mechelli A, Friston KJ (2004b) Modelling functional integration: a comparison of structural equation and dynamic causal models. Neuroimage 23:S264–S274. https://doi.org/10.1016/j.neuroimage.2004.07.041

Raman S, Deserno L, Schlagenhauf F, Stephan KE (2016) A hierarchical model for integrating unsupervised generative embedding and empirical Bayes. J Neurosci Methods 269:6–20. https://doi.org/10.1016/j.jneumeth.2016.04.022

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6(2):461–464. https://doi.org/10.1214/aos/1176344136

Sengupta B, Friston KJ, Penny W (2015) Gradient-free MCMC methods for dynamic causal modelling. NeuroImage 112(C):375–381. https://doi.org/10.1016/j.neuroimage.2015.03.008

Sengupta B, Friston KJ, Penny W (2016) Gradient-based MCMC samplers for dynamic causal modelling. NeuroImage 125:1107–1118. https://doi.org/10.1016/j.neuroimage.2015.07.043

Stephan KE, Kasper L, Harrison LM, Daunizeau J, den Ouden HEM, Breakspear M, Friston KJ (2008) Nonlinear dynamic causal models for fMRI. Neuroimage 42(2):649–662. https://doi.org/10.1016/j.neuroimage.2008.04.262

Stephan KE, Penny W, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. Neuroimage 46(4):1004–1017. https://doi.org/10.1016/j.neuroimage.2009.03.025

Stephan KE, Schlagenhauf F, Huys QJM, Raman S, Aponte EA, Brodersen KH, Heinz A (2017) Computational neuroimaging strategies for single patient predictions. NeuroImage 145:180–199. https://doi.org/10.1016/j.neuroimage.2016.06.038

Swendsen RH, Wang J-S (1986) Replica monte carlo simulation of spin-glasses. Phys Rev Lett 57(21):2607–2609. https://doi.org/10.1103/physrevlett.57.2607

Translational Neuromodeling Unit (2014). TAPAS: Translational algorithms for psychiatry-advancing science. Retrieved from http://www.translationalneuromodeling.org/tapas

Watanabe S (2013) A Widely applicable bayesian information criterion. J Mach Learn Res 14:867–897

Welvaert M, Rosseel Y (2013) On the definition of signal-to-noise ratio and contrast-to-noise ratio for fmri data. PLoS ONE 8(11):e77089. https://doi.org/10.1371/journal.pone.0077089

Wipf D, Nagarajan S (2009) A unified Bayesian framework for MEG/EEG source imaging. Neuroimage 44(3):947–966. https://doi.org/10.1016/j.neuroimage.2008.02.059

Yao Y, Raman SS, Schiek M, Leff A, Frässle S, Stephan KE (2018) Variational Bayesian inversion for hierarchical unsupervised generative embedding (HUGE). Neuroimage 179:604–619. https://doi.org/10.1016/j.neuroimage.2018.06.073

Yao Y, Stephan KE (2021) Markov chain Monte Carlo methods for hierarchical clustering of dynamic causal models. Hum Brain Mapp 42:2973–2989