A correspondence between thermodynamics and inference

Colin H. LaMont and Paul A. Wiggins
*Departments of Physics, Bioengineering and Microbiology, University of Washington, Box 351560. 3910 15th Avenue Northeast, Seattle, WA 98195, USA**

We systematically explore a natural analogy between Bayesian statistics and thermal physics in which sample size corresponds to inverse temperature. We discover that some canonical thermodynamic quantities already correspond to well-established statistical quantities. Motivated by physical insight into thermal physics, we define two novel statistical quantities: a learning capacity and Gibbs entropy. The definition of the learning capacity leads to a critical insight: The well-known mechanism of failure of the equipartition theorem in statistical mechanics is the mechanism for anomalously-predictive or *sloppy* models in statistics. This correspondence between the learning and heat capacities provides new insight into the mechanism of machine learning. The correspondence also suggests a solution to a long-standing difficulty in Bayesian statistics: the definition of an objective prior. We propose that the Gibbs entropy provides a natural generalization of the *principle of indifference* that defines objectivity. This approach unifies the disparate Bayesian, frequentist and information-based paradigms of statistics by achieving coherent inference between these competing formulations.

I. INTRODUCTION

Scientific and technological innovations are rapidly increasing the size and scope of datasets. Accompanying this growth come new challenges in analysis, interpretation and modeling. An enduring difficulty in statistics is the existence of the competing paradigms of Bayesian, frequentist and information-based inference, each of which offers different prescriptions. Our motivation is to reconcile these competing paradigms and to understand the phenomenology of learning from a unified perspective.

A correspondence between the partition function of statistical mechanics and the Bayesian evidence has been explored by Jaynes, Balasubramanian and many others [1–6]. In this paper, we extend this correspondence by using the canonical bridge from statistical mechanics to thermodynamics to compute the standard thermodynamic properties of the system. We are able to identify several thermodynamic quantities as metrics used in statistical inference and find interesting connections between the Akaike Information Criterion (AIC) of information-based inference, [10, 11], model sloppiness in statistics [12] and the equipartition theorem in statistical mechanics [13]. We propose a novel interpretation of the Gibbs entropy in the context of statistical inference: We argue that the Gibbs entropy is log the number of indistinguishable models. This proposal provides a natural extension of the principle of indifference to models with unknown dimension *i.e.* in model selection.This *generalized principle of indifference* resolves the Lindley-Bartlett paradox and unifies the frequentist and Bayesian approaches to model selection.

* pwiggins@uw.edu; http://mtshasta.phys.washington.edu/

A. Preliminaries

We assume that a true parameter value $\boldsymbol{\theta}_0$ is drawn from the prior distribution $\varpi(\boldsymbol{\theta})$. Although $\boldsymbol{\theta}_0$ is unknown, N samples $x^N \equiv \{x_1, ..., x_N\}$ are observed, each is distributed as likelihood function $q(x|\boldsymbol{\theta}_0)$. It will be convenient to work in terms of the observation-space cross entropy H and an unbiased empirical estimator $\hat{H}_X$:

$$H(\boldsymbol{\theta};\boldsymbol{\theta}_0) \equiv -\langle \log q(X|\boldsymbol{\theta})\rangle_X\,, \tag{1}$$

$$\hat{H}_x(\boldsymbol{\theta}) \equiv -\langle \log q(X|\boldsymbol{\theta})\rangle_{X\in x^N}\,, \tag{2}$$

where $\hat{H}_x$ is an unbiased estimator of H.

To introduce the correspondence between statistical physics and Bayesian inference, we rewrite the marginal likelihood [4]:

$$Z(X^N) \equiv \int_{\boldsymbol{\Theta}} \mathrm{d}\boldsymbol{\theta}\; \varpi(\boldsymbol{\theta})\; \exp[-N\hat{H}_X(\boldsymbol{\theta})]. \tag{3}$$

The correspondence is established by comparing the equation above to the partition function in the canonical ensemble. As previously described, [1–6]: The model parameters $\boldsymbol{\theta}$ are equivalent to the variables that define the physical state vector, the cross entropy $\hat{H}_X(\boldsymbol{\theta})$ is equivalent to energy $E(\boldsymbol{\theta})$, the prior $\varpi(\boldsymbol{\theta})$ to the density of states $\rho(\boldsymbol{\theta})$ and the evidence $Z(X^N)$ to the partition function in the canonical ensemble $\mathfrak{Z}(\beta)$. (Where there is ambiguity between Bayesian and physical quantities, we will use the script font for physical quantities.) Finally, the sample size N is equivalent to inverse temperature $\beta \equiv T^{-1}$ (See Sec. C in the supplement for a more detailed discussion of this choice of β). This assignment is natural in the following sense: At small sample size N, many parameter values are consistent with the data, in analogy with the large range of states $\boldsymbol{\theta}$ occupied at high temperature T in the canonical ensemble. The correspondence is summarized in Tab. I.

Thermodynamics Quantity:	Interpretation:		Statistics Quantity:	Interpretation:
$\beta = T^{-1}$	Inverse temperature	$\leftrightarrow$	N	Sample size
$\boldsymbol{\theta}$	State variables/vector	$\leftrightarrow$	$\boldsymbol{\theta}$	Model parameters
X^N	Quenched disorder	$\leftrightarrow$	X^N	Observations
$E_X(\boldsymbol{\theta})$	State energy	$\leftrightarrow$	$\hat{H}_X(\boldsymbol{\theta})$	Cross entropy estiamtor
E_0	Disorder-averaged ground state energy	$\leftrightarrow$	$\overline{H}_0$	Shannon entropy
$\rho(\boldsymbol{\theta})$	Density of states	$\leftrightarrow$	$\varpi(\boldsymbol{\theta})$	Prior
$\mathfrak{X}$	Partition function	$\leftrightarrow$	Z	Evidence
$\mathfrak{X}^{-1}\rho\ \exp -\beta E_X$	Normalized Boltzmann weight	$\leftrightarrow$	$\varpi(\boldsymbol{\theta}\vert X^N)$	Posterior
$F = -\beta^{-1}\log \mathfrak{X}$	Free energy	$\leftrightarrow$	$F = -N^{-1}\log Z$	Minus-log-evidence
$U = \partial_\beta \beta F$	Average energy	$\leftrightarrow$	$U = \partial_N NF$	Minus-log-prediction
$C = -\beta^2\partial_\beta^2 \beta F$	Heat capacity	$\leftrightarrow$	$C = -N^2\partial_N^2 NF$	*Learning capacity*
$S = \beta^2\partial_\beta F$	Gibbs entropy	$\leftrightarrow$	$S = N^2\partial_N F$	*Complexity*

TABLE I. **Thermodynamic-Bayesian correspondence.** The top half of the table lists the correspondences that can be determined directly from the definition of the marginal likelihood as the partition function. The lower half of the table lists the implied thermodynamic expressions and their existing or proposed statistical interpretation.

To extend the previously proposed correspondence, we follow the standard prescriptions from statistical mechanics to compute thermodynamic potentials and variables for the systems [13, 14]. We would argue that the proposed correspondence is useful if (i) there is a direct one-to-one correspondence between all thermodynamics and statistical quantities and (ii) the correspondence gives us novel insight into the meaning of statistical quantities and clues to the correct way to approach ambiguities in the current formulation of (objective) Bayesian statistics.

B. Simple models

For concreteness, we consider simple but nontrivial models. In the context of statistical mechanics, we analyze a classical free particle in K-dimensions confined to a K-cube. The particle will have ground-state energy E_0 and for compactness we absorb all of the physical parameters and constants into a single inverse temperature β_0, defined in the supplement.

In the context of Bayesian inference, we analyze a normal model with a conjugate prior. Consider a K-dimensional normal model with K unknown means $\boldsymbol{\mu}$ and known variance σ^2. The true parameter $\boldsymbol{\mu}_0$ is drawn from a K-dimensional normal distribution (the prior ϖ) with mean $\boldsymbol{\mu}_\varpi$ and variance σ^2_ϖ. It will be convenient to define a critical sample size:

$$N_0 \equiv \sigma^2/\sigma^2_\varpi, \tag{4}$$

where the information content of the observations X^N is equal to the information content of the prior. (See the supplement for explicit expressions for the likelihood, cross entropy, information content of the observations, etc.) We work in K dimensions to make the results more easily comparable to a generic model. It is instructive to analyze two limiting cases of the model: (i) the $N_0 \to 0$ is the limit of the model with no prior information about the mean and (ii) the $N_0 \to \infty$ is the limit where prior information completely specifies the mean.

The normal model represents the large sample-size-limit of a regular Bayesian model of dimension K. For generality, we also study the large-sample-size limit of a singular statistical model of dimension K. Models are *singular* when parameters are *structurally unidentifiable* [15]:

$$q(x|\boldsymbol{\theta}_1) = q(x|\boldsymbol{\theta}_2) \quad \text{for} \quad \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2, \tag{5}$$

where the unidentifiability cannot be removed by coordinate transformation. A regular model is the special case where all parameters are identifiable. Using exact asymptotic results for singular models [15], the thermodynamic quantities for each model and limit are determined and shown in Tab. II.

The thermodynamic quantities depend on the particular realization of the data X^N. In the current context we are interested in the expectation over this *quenched disorder* (*i.e.* data). We define the expectation with an overbar:

$$\overline{f}(N) \equiv \langle\, f(X^N, \boldsymbol{\theta}_0)\,\rangle_{X,\boldsymbol{\theta}_0}\,, \tag{6}$$

where X^N is distributed like $q(\cdot|\boldsymbol{\theta}_0)$ and $\boldsymbol{\theta}_0$ is distributed like ϖ. This expectation is a *frequentist expectation* since it is taken over many different putative realizations of the data X^N and true parameter values $\boldsymbol{\theta}_0$.

Model	K-D-Free-particle	K-D-Normal-prior			K-D-singular
Limit	Classical	Exact	$N_0 \to 0$	$N_0 \to \infty$	$N \to \infty$
$\overline{F}$	$E_0 + \frac{K}{2\beta} \log \frac{\beta}{\beta_0}$	$\overline{H}_0 + \frac{K}{2N} \log \frac{N+N_0}{N_0}$	$\overline{H}_0 + \frac{K}{2N} \log \frac{N}{N_0}$	$\overline{H}_0$	$\overline{H}_0 + \frac{\gamma}{2N} \log N + ...$
$\overline{U}$	$E_0 + \frac{K}{2\beta}$	$\overline{H}_0 + \frac{K}{2(N+N_0)}$	$\overline{H}_0 + \frac{K}{2N}$	$\overline{H}_0$	$\overline{H}_0 + \frac{\gamma}{2N} + ...$
$\overline{C}$	$\frac{K}{2}$	$\frac{K}{2(1+N_0/N)^2}$	$\frac{K}{2}$	0	$\frac{\gamma}{2} + ...$
$\overline{S}$	$\frac{K}{2}\left(1 - \log \frac{\beta}{\beta_0}\right)$	$\frac{K}{2}\left(\frac{N}{N+N_0} - \log \frac{N_0+N}{N_0}\right)$	$\frac{K}{2}\left(1 - \log \frac{N}{N_0}\right)$	0	$-\frac{\gamma}{2} \log N + ...$

TABLE II. **Thermodynamic-Bayesian correspondence.** The thermodynamic quantities of a K-dimensional free particle are compared to a K-dimensional normal model with a conjugate prior. Two limiting cases of the normal model are also shown. Inspection reveals that the free particle is exactly equivalent to the $N_0 \to 0$ limit of the normal model, identifying the parameters as described in Tab. I. For the singular model, we supply only the two leading order contributions in the large N limit. The learning coefficient $\gamma \le K$. The special case of $\gamma = K$ is a regular model.

II. RESULTS

We follow the standard prescriptions from statistical mechanics to compute thermodynamic potentials and variables for the systems [13, 14]. We would argue that the proposed correspondence is useful if (i) there is a direct one-to-one correspondence between all thermodynamics and statistical quantities and (ii) the correspondence gives us novel insight into the meaning of statistical quantities and clues to the correct way to approach ambiguities in the current formulation of (objective) Bayesian statistics.

1. *Free energy*

A relation between the partition function and Bayesian evidence has long been discussed [1–6]. The Free energy F represents the canonical Bayesian model preference: the minus-log-evidence per observation, or message length per observation. In Tab II it is seen that F breaks up into two parts. The first term is the code length per observation using the optimal encoding and the second term the length of the code required to encode the model parameters using the prior (per observation) [16, 17]. The model that maximizes the evidence and therefore minimizes F is selected in the canonical approach to Bayesian model selection.

2. *Average energy*

The average energy U measures the predictive accuracy of the model. (See Tabs. I and II.) The thermodynamic prescription involves a derivative with respect to sample size. We will formally interpret this derivative using a finite difference definition so that the statistical meaning of average energy is

$$U(X^N) \equiv -\left\langle \log q(X_i|X^{\neq i})\right\rangle_{i=1..N}, \tag{7}$$

where the empirical expectation is taken by averaging all values of the predicted observation i in the N samples and $q(X_i|X^{\neq i}) \equiv Z(X^N)/Z(X^{\neq i})$ is the Bayes-predictive distribution. The RHS is a well-known statistical object: it is the Leave-One-Out-Cross-Validation (LOOCV) estimator of model performance (*e.g.* [18]). The statistical interpretation of average energy U is therefore the minus-expected-predictive-performance of the model.

For our analysis of the normal model, we analytically continue N to a continuous variable to perform the derivative. As shown in the examples in Tab. II, the averaged energy can be written as the sum of two contributions: The first term $\overline{H}_0$ is the performance of the model if the true parameterization was known. The second term represents the loss associated with predicting a new observation X using estimated model parameters rather than $\boldsymbol{\theta}_0$. Comparing this to the free-particle in statistical physics indicates that $\overline{H}_0$ corresponds to a ground state energy and the *loss term* to the thermal energy. The loss term follows the typical behavior predicted by the equipartition theorem: *there is a half k_BT of thermal energy per harmonic degree of freedom* [13]. This thermal energy is characterized by the heat capacity, a key observable quantity in thermodynamics (See Tabs. I and II.) The universal form of the predictive loss in the large-sample-size limit of regular models has an important statistical precedent and is the mechanism by which the Akaike Information Criterion (AIC) estimates the predictive performance [10, 11].

3. *Learning Capacity*

Physically, the heat capacity measures the rate of increase in thermal energy with temperature ($\overline{C}$ in Tab. I). The statistical analogue of the heat capacity, a *learning capacity*, is a measure of the rate of increase in predictive performance with sample size. The equipartition theorem implies that both the heat capacity and the learning capacity obey the same universal behavior: They are equal to half the number of degrees of freedom and therefore both the heat and learning capacities can be interpreted as one-half the effective number of degrees of freedom. (See Tab. II.)

The applicability and failure of the equipartition theorem are well understood phenomena in statistical mechanics. At high or low temperature, degrees of freedom can become non-harmonic, altering their contribution to the heat capacity [13]. For instance, due to the discrete structure of the quantum energy levels, degrees of freedom can *freeze out* at low temperature. (See Fig. 1A.) Degrees of freedom can also become irrelevant at high temperature. For instance, the position degrees of freedom of a gas do not contribute to the heat capacity [19]. We see that an analogous high-temperature feeze-out mechanism is at work in the context of inference in the normal model. The learning capacity for the normal model is plotted in Fig. 1B. In the large-sample-size limit, the normal model is well-described by the equipartition theorem (Tab. II: $N_0 \to 0$), but at small sample size (Tab. II: $N_0 \to \infty$), the learning capacity shrinks to zero. In the small-sample-size limit, the K degrees of freedom of the model are frozen out. These degrees of freedom un-freeze at $N \approx N_0$. For larger sample size, equipartition applies.

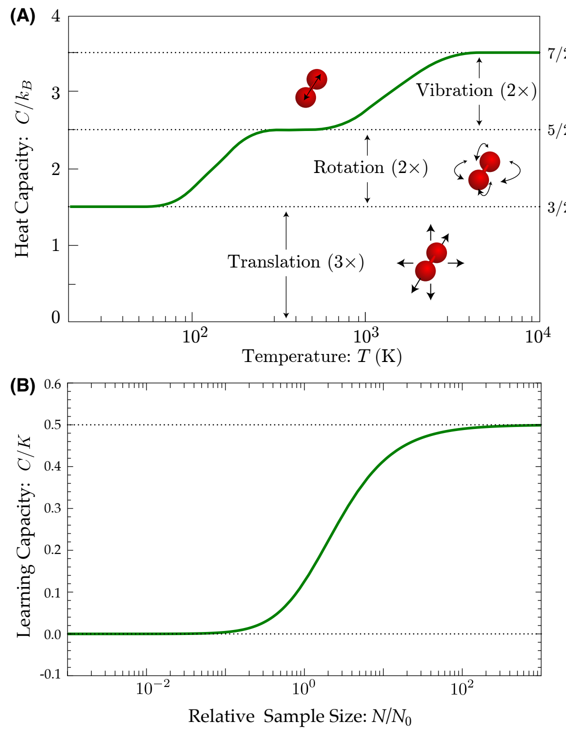

FIG. 1. **The failure of equipartition.** The behaviors of the heat capacity and learning capacity are compared and related to the applicability or inapplicability of the Equipartition theorem in different regimes. **Panel A: Low-temperature freeze-out in a quantum system.** The heat capacity is plotted as a function of temperature. Equipartition predicts that the reduced heat capacity should be constant, equal to half the degrees of freedom in the system. Plateaus can clearly be observed at half-integer values, but the number of degrees of freedom is temperature dependent due to the discrete nature of quantum energy levels. At low temperature, some degrees of freedom are frozen out since the first excited state is thermally inaccessible. This discrete topology of the energy levels implies anharmonicity in the potential and therefore failure of the equipartition theorem. **Panel B: High-temperature freeze-out in the Learning capacity.** Analogous to the statistical mechanics system, the statistical learning capacity transitions between half integer plateaus, reflecting a temperature-dependent number of degrees of freedom. At low sample size N (high temperature), the parameters are completely specified by model constraints (the prior) and therefore the parameters do not contribute to the learning capacity. At large sample size N, the parameters become data dominated and therefore the learning capacity is predicted by equipartition ($\frac{1}{2}K$).

4. *Learning capacity in more general models*

To investigate whether the properties of the learning capacity generalize, we consider the large-sample-size limit of a generic model on a continuous parameter manifold for which asymptotic results are known [15]. (See Tab. II.) Like the normal model, the learning capacity has the canonical equipartition form for a regular model ($\overline{C} = K/2$) and a smaller effective dimension for singular models ($\overline{C} = \gamma/2$ where $\gamma \leq \dim \boldsymbol{\Theta}$). Therefore, the reported plateauing behavior at large sample size is expected to be quite general. Based on empirical evidence, we expect structural mechanisms can also result in reduced-effective-dimension due to *practically unidentifiable* parameters for which Eqn. 5 is only approximately realized at finite sample size. We shall discuss the significance of these results below.

5. *The Gibbs entropy*

In physics, the Gibbs entropy generalizes the Boltzmann formula: $S = \log \Omega$ where Ω is the number of accessible states. We propose that the Gibbs entropy has the analogous meaning in the context of Bayesian statistics: The Gibbs entropy is the log number of models consistent with the data. The finite sample size expression for the entropy is

$$S(X^N) \equiv N(U - F), \tag{8}$$

where Eqn. 7 provides an explicit expression for U. The Gibbs entropy of a normal model is shown in Fig. 2. At small sample size all models specified by the prior are consistent with the data and therefore the Gibbs entropy is zero. Above the critical sample size N_0, the data is informative to the parameter values and therefore the number of models consistent with the data is reduced. As a result the Gibbs entropy becomes increasingly negative as sample size N grows.

A. The principle of indifference

In statistical physics, the density of states is known (*i.e.* measured) but in inference the selection of a prior is often subjective. A long standing problem in Bayesian statistics is how to construct an objective or uninformative prior. A poor prior choice can lead to poor predictive performance. The relative Bayesian preference between models is the difference in free energy ΔF, which can be written as the sum of the relative predictive performance ΔU minus the scaled-difference in the entropy $N^{-1}\Delta S$:

$$\Delta F \equiv \Delta U - \tfrac{1}{N}\Delta S. \tag{9}$$

Clearly a significant predictive advantage can be abrogated by model multiplicity when the weight of poorly performing models is astronomically large: *i.e.* $\Delta U < N^{-1}\Delta S$. This construction suggests that large differences in Gibbs entropy, caused by the presence of a high density of indistinguishable models, can result in a loss of predictive performance. What insight does the proposed correspondence provide for prior choice?

An important lesson from statistical mechanics is the concept of *indistinguishability*. A surprising result from the perspective of classical physics (the Gibbs paradox) is that Nature makes no distinction between states with identical particles exchanged (*e.g.* electrons) and counts only distinguishable states. Following V. Balasubramanian [4], we proposed that the concept of indistinguishability must be applied to the context of objective Bayesian inference. We propose that the collective weight of all models with the same likelihood function should be identified and collectively assigned the weight of a single distinguishable model. This concept already has significant statistical precedent and is called the *Principle of Indifference*: All *mutually exclusive* and *exhaustive* possibilities should be assigned equal prior probability [20, 21]. One interpretation of this prescription is that it maximizes entropy [1, 22].

Although the principle of indifference is easily interpreted for discrete models, it is more difficult to interpret in the context of continuous parameters and models of different dimension. For example, are normal models with means μ and $\mu + \mathrm{d}\mu$ mutually exclusive (distinguishable)? Even if we were to constrain the mean to be an integer ($\mu \in \mathbb{Z}$) to define *mutually exclusive*, the exhaustive condition is itself problematic. Exhaustive would correspond to a uniform weighting over all integers. This vanishing prior weight ($1/\infty$) on the non-compact set $\mathbb{Z}$ is acceptable in some scenarios (*e.g.* posterior on μ), but results in paradoxical results in others (*e.g.* the Bartlett-Lindley paradox which describes the automatic rejection of high-dimensional models[1]) [23, 24].

[1] In fact we have already seen this divergent contribution ($\log N_0$) in both the free energy and Gibbs entropy in the $N_0 \to 0$ limit of the normal model (Tab. II).

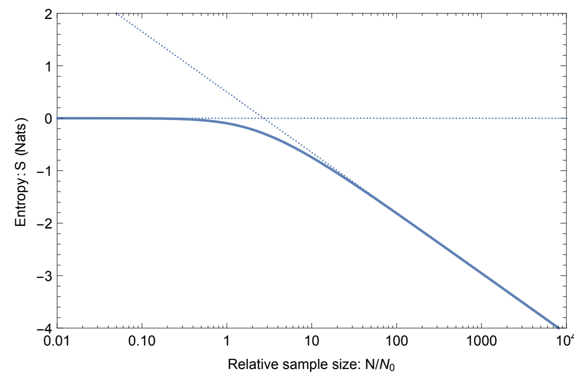

FIG. 2. **Understanding Gibbs entropy.** The Gibbs entropy for the normal-model-with-prior is plotted as a function of sample size. The Gibbs entropy can be understood heuristically as the log ratio of the model consistent with the data to allowed models. At small sample size, the model structure determines the parameterization and therefore all models allowed are consistent with the data and there is zero Gibbs entropy. As the sample size grows beyond the critical sample size N_0, fewer and fewer of the allowed models are consistent with the data and the entropy decreases like $-\frac{1}{2}K\log N$. The non-positivity of the Gibbs entropy is a direct consequence of the normalization of the prior, which forces the Gibbs entropy to have a maximum value of zero. A prior determined by the principle of indifference avoids this non-physical result.

B. A generalized principle of indifference

The correspondence offers a natural mechanism for applying the principle of indifference: Since the Gibbs entropy is understood as the log-number of indistinguishable models at sample size N, constant entropy is understood as equal model weighting. To study the weighting of each model, *we must prepare the data using a different procedure*. We distribute X^N using the likelihood at an unknown true parameter $\boldsymbol{\theta}_0$, $X^N \sim q(\cdot|\boldsymbol{\theta}_0)$, as before but now we omit the expectation over $\boldsymbol{\theta}_0$ [2]. A generalized principle of indifference states that the prior ϖ should be chosen such that:

$$\overline{S}(\boldsymbol{\theta}_0; N, \varpi) \approx \text{const}, \tag{10}$$

at sample size N, where the Gibbs entropy is now a function of $\boldsymbol{\theta}_0$. Eqn. 10 realizes the condition of equal model weighting on mutually exclusive models using a statistically principled definition.

The correspondence also offers a natural mechanism for resolving statistical anomalies arising from the *exhaustive* condition in the principle of indifference that gives rise to the Lindley-Bartlett paradox [3]. In statisti-

[2] We must make the important distinction between the *inference prior* ϖ, used to perform inference, and the *true prior*, used to generate the parameters $\boldsymbol{\theta}_0$. This approach has a long and important statistical precedent: *e.g.* [15, 25].

[3] A canonical approach to the problems associated with objective

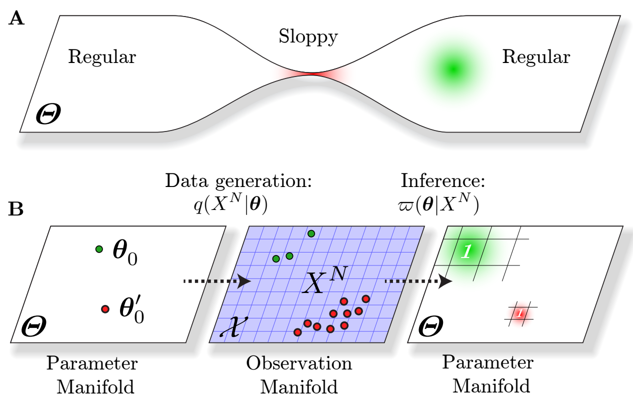

FIG. 3. **Panel A: Sloppiness is determined by parameter manifold geometry and posterior width.** Parameters are defined on a compact manifold $\boldsymbol{\Theta}$. In sloppy regions of parameter space, the parameters are model-structure dominated (red posterior) whereas in regular regions of parameter space parameters are data dominated (green posterior). From the perspective of the learning capacity, the model is effectively one dimensional in proximity to the red posterior and two dimensional in proximity to the green posterior. **Panel B: Generalized principle of indifference.** The posterior distribution $\varpi(\boldsymbol{\theta}_0|X^N)$ is shown schematically for two different sample sizes N. The resolution increases with sample size as the posterior shrinks. In the w prior (Eqn. 11), all parameter values consistent with the posterior are assigned unit prior weight collectively.

cal mechanics, the partition function $\mathfrak{X}$ is not normalized by construction since the density of states ρ is a density but not a probability density. Therefore, a natural solution to statistical anomalies arising from the exaustive condition is to re-interpret the objective inference prior as a *density of models*. To specify a consistent density of models between different parameter values and model families, we replace the inference prior $\varpi(\boldsymbol{\theta})$ with a model density $w_i(\boldsymbol{\theta})$ such that:

$$\overline{S}_i(\boldsymbol{\theta}_0; N, w_i) \approx 0, \tag{11}$$

assigning unit multiplicity to all parameters $\boldsymbol{\theta}_0$ and model families i. We avoid specifying Eqns. 10 and 11 as equalities since the condition is typically not exactly realizable for all $\boldsymbol{\theta}_0$ at finite sample size N. A precise formulation will be described elsewhere, but is analogous to the approach of Kashyap [25]. We shall refer to Eqn. 11 as the *Generalized Principle of Indifference* which realizes both the mutually-exclusive and exhaustive conditions using a principled statistical approach, regardless of the nature of the parameter space.

III. DISCUSSION

A. Learning capacity and model sloppiness

One valuable feature of the proposed correspondence is the potential to gain new insights into statistical phenomenology using physical insights into the thermodynamic properties of physical systems. The connection between the learning and heat capacities has great potential to provide novel insights into the mechanism of learning. Artificial Neural Networks (ANN) and systems-biology models are two examples of systems with a large number of poorly-specified parameters that none-the-less prove qualitatively predictive. This phenomena has been discovered empirically and has been termed *model sloppiness* [12, 26]. These models often have a logarithmic distribution of Fisher information eigenvalues and this characteristic has been used as a definition of sloppiness [12]. But, this definition is unsatisfactory since it is not reparameterization invariant. But, the correspondence suggests a definition directly written in terms of the predictive performance of the model and the equipartition theorem. Sloppy models have a smaller learning capacity than estimated from the model dimension:

$$\overline{C} < \tfrac{1}{2}\dim \boldsymbol{\Theta}, \tag{12}$$

as illustrated by the normal-model-with-prior. At small sample size, the parameters are model-structure rather than data dominated. See Fig. 3. The correspondence indicates that it is only when the model has frozen out degrees of freedom that it can be expected to exhibit sloppiness. Small Fisher information eigenvalues are not a sufficient conditions since even exponentially small eigenvalues could still result in models with the canonical predictive performance [4].

Although the analogy between the learning and heat capacity is most natural in the Bayesian context, the learning capacity is defined in terms of the predictive performance. It therefore generalizes to other paradigms of inference. The fact that the learning capacity is equal to the model dimension in the large sample size limit suggests that this quantity is a natural measure of learning performance and has the potential to offer new insights into the mechanism of learning in more complex systems, including ANNs. Our preliminary investigations suggest some training algorithms may map to physical systems with well-understood

Bayesian inference on a non-compact parameter space is to adopt an improper prior (*i.e.* do not enforce the normalization condition). Although this can result in sensible posteriors over model parameters, the Bayesian evidence Z_i cannot be compared between model families i since Z cannot be interpreted as a probability. The proposed correspondence provides a natural solution. If a normalized weighting of macro-state i (*i.e.* family of models) is required, it is achieved by dividing by the total partition function ($\mathfrak{X}_i/\mathfrak{X}_{\mathrm{tot}}$).

[4] For instance, the normal model with $N_0 \to \infty$ is not sloppy for any finite Fisher information matrix eigenvalue.

thermodynamic properties. The detailed physical understanding of the complex phenomenology of physical systems, including phase transitions, renormalization, *etc.*, have great promise for increasing our understanding of the fundamental mechanisms of learning.

B. Unification of statistical paradigms

A second valuable feature of the proposed correspondence is that it suggests novel interpretations of statistical quantities. We have proposed that the Gibbs entropy can be used to formulate a generalized principle of indifference and therefore to define an objective prior (Eqn. 11). Schematically, this procedure assigns equal prior weighting to all models that can be distinguished at finite sample size N. As the sample size increases, the prior must be modified to accommodate the increased resolution (shrinking of the posterior support). (See Fig. 3.) It is important to stress that this procedure gives rise to a *sample-size-dependent prior* and therefore is not Bayesian in a classical sense since it violates Lindley's dictum: *today's posterior is tomorrow's prior*. But, the generalized principle of indifference naturally addresses many problems with existing approaches.

An important shortcoming with existing objective Bayesian approaches relates to the compactness of the parameter manifold and the automatic rejection of higher-dimensional models (the Bartlett-Lindley paradox [23, 24]). The mechanism for this failure is easy to understand from the perspective of the density of models. For instance, the reference prior [27, 28] enforces a generalized-principle-of-indifference-like condition to enforce equal prior-weighting between *different parameter values* but not between different *model families* [5]! In contrast, the generalized principle of indifference consistently enforces equal prior-weighting over both parameter values and model families. As a result, the canonical Bayes factor depends on *ad hoc* modeling decisions, like the range of allowed parameter values, whereas the generalized-principle-of-indifference-based approach does not. Therefore, the generalized principle of indifference circumvents the Bartlett-Lindley paradox. The absence of the Lindley-Bartlett paradox implies coherent inference between paradigms [29, 30] and therefore the generalized principle of indifference naturally unifies objective Bayesian inference with the other two paradigms of inference.

To understand the significance of this unification, consider model selection. Bayesian statistics contains a natural Occam's razor—*parsimony increases probability*—a mechanism for favoring parsimonious models [31, 32]. A sensible question is whether the generalized principle of indifference contains any endogenous mechanism for model selection to favor parsimonious models. To see that such a mechanism exists, it is useful to rewrite Eqn. 11:

$$-N\overline{F} \approx -N\overline{U}, \tag{13}$$

which in statistical language corresponds to using a prior that makes the log partition function (LHS) an unbiased estimator of the log predictive performance (RHS). Since the Akaike Information Criterion (AIC) is an unbiased estimator of RHS at large sample size N, the generalized principle of indifference encodes an AIC-like model selection [33] and an information-based (AIC) realization of Occam's razor: *parsimony increases predictivity* [11].

The generalized principle of indifference also naturally unifies two seemingly disparate approaches to Bayesian statistics: uninformative priors and predictive methods. In the large sample size limit, Eqn. 13 implies that the new prior is proportional to the well known Jeffreys prior: $w \propto (\det \boldsymbol{I})^{1/2}$ where $\boldsymbol{I}$ is the Fisher Information Matrix [34–36]. The Jeffreys prior has been advocated as *uninformative* by Bernardo and Berger [6] in the development of the reference prior [27, 28]. But Eqn. 13 has a second interesting Bayesian interpretation: The log predictive performance (RHS Eqn. 13) has been advocated in the context of model selection through the use of pseudo Bayes factors by Gelman and coworkers [37–41]. The generalized principle of indifference therefore results in both a prior that is uninformative and model weighting that is predictive, bridging two seemingly unrelated Bayesian approaches.

C. Conclusion

Nature reveals an elegant formulation of statistics in the thermal properties of physical systems. Measurements of the heat capacity, compressibility or susceptibility reveal unambiguously how Nature enumerates states and defines entropy. Through the correspondence, these physical insights provide clues to the definition of novel statistical quantities and the resolutions of ambiguities in the formulation of Bayesian statistics. We have developed on a previously proposed correspondence between the Bayesian marginal likelihood and the partition function of statistical physics and between the sample size and the inverse temperature. We demonstrate a novel and substantive mapping between the average energy, heat capacity and entropy and statistical quantities. This correspondence gives new insight into the Akaike Information Criterion (AIC) and model sloppiness through a correspondence with the

[5] Instead, the posterior probabilities of the two model families are typically equalized, even if one of families describes many more distinguishable models than the other.

[6] At least in one dimension.

equipartition theorem of statistical mechanics. The newly-defined learning capacity is a natural quantity for characterizing and understanding learning algorithms and demands further study. Finally, we use the Gibbs entropy to define a generalized principle of indifference and an objective Bayesian prior with the property that all distributions have equal prior weight. This approach provides a natural unification and coherent inference between the three paradigms of statistics as well as unifying seemingly disparate approaches within Bayesian statistics itself.

Acknowledgements: PAW and CHL acknowledge helpful discussions with M. Linden, J. Kinney, D. Mayo and C. Heilig. This work was support by NSF grants NSF-PHY-084845 and NSF-MCB-1151043-CAREER.

---

[1] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[2] E. T. Jaynes. *Probability Theory: The Logic of Science.* Cambridge University Press., 2003.

[3] Ole E Barndorff-Nielsen and Peter E Jupp. Statistics, yokes and symplectic geometry. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 6, pages 389–427, 1997.

[4] V. Balasubramanian. Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Computation*, 9:349–368, 1997.

[5] Hidetoshi Nishimori. *Statistical physics of spin glasses and information processing: an introduction*, volume 111. Clarendon Press, 2001.

[6] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[7] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová. Phase transitions in sparse pca. In *Information Theory (ISIT), 2015 IEEE International Symposium on*, pages 1635–1639. IEEE, 2015.

[8] Lenka Zdeborová and Florent Krzakala. Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552, 2016.

[9] Florent Krzakala, Marc Mézard, François Sausset, YF Sun, and Lenka Zdeborová. Statistical-physics-based reconstruction in compressed sensing. *Physical Review X*, 2(2):021005, 2012.

[10] H. Akaike. Information theory and an extension of the maximum likelihood principle. In Petrov B. N. and E. Csaki, editors, *2nd International Symposium of Information Theory.*, pages 267–281. Akademiai Kiado, Budapest., 1973.

[11] K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference.* Springer-Verlag New York, Inc., 2nd. edition, 1998.

[12] Benjamin B Machta, Ricky Chachra, Mark K Transtrum, and James P Sethna. Parameter space compression underlies emergent theories and predictive models. *Science*, 342(6158):604–7, Nov 2013.

[13] R.K. Pathria and P.D. Beale. *Statistical Mechanics*. Elsevier Science, 1996.

[14] Josiah Willard Gibbs. Elementary principles of statistical mechanics. *Compare*, 289:314, 1902.

[15] S. Watanabe. *Algerbraic geometry and statistical learning theory.* Cambridge Univeristy Press, 2009.

[16] J. Rissanen. Modeling by the shortest data description. *Automatica*, 14:465–471, 1978.

[17] Peter D. Grünwald. *The Minimum Description Length Principle.* MIT, Cambridge, MA, 2007.

[18] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001.

[19] F. Reif. *Statistical Physics.* McGraw-Hill (New York), 1967.

[20] P. S. Laplace. *Theorie analytique des probabilites*. Courcier Imprimeur, Paris, 3rd edition, 1812.

[21] J. M. Keynes. *A Treatise on Probability.* Macmillan Limited, London, 1921.

[22] J. Shore and R. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *Information Theory, IEEE Transactions on*, 26(1):26–37, January 1980.

[23] M. S. Bartlett. A comment on D. V. Lindley's statistical paradox. *Biometrika*, 44(3/4):533–534, 1957.

[24] D. V. Lindley. A statistical paradox. *Biometrika*, 44(1/2):187–192, 1957.

[25] R. L. Kashyap. Prior probability and uncertainty. *IEEE Transactions on information theory*, IT-17(6):?641–650, ? 1971.

[26] Mark K Transtrum, Benjamin B Machta, Kevin S Brown, Bryan C Daniels, Christopher R Myers, and James P Sethna. Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *J Chem Phys*, 143(1):010901, Jul 2015.

[27] J. O. Berger and J.-M. Bernardo. On the development of the reference prior method. Techincal Report 91-15C, Purdue University, 1991.

[28] J. M. Bernardo. *Bayesian statistics 6: Nested Hypothesis Testing: The Bayesian Reference Criterion*. Oxford University Press, 1999.

[29] Robert D. Cousins. The Jeffreys–Lindley paradox and discovery criteria in high energy physics. *Synthese*, pages 1–38, 2014.

[30] Colin H. LaMont and Paul A. Wiggins. The lindley paradox: The loss of resolution in bayesian inference. *Unbder review. (arXiv:1610.09433)*, 2017.

[31] D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

[32] D. J. C. MacKay. A practical framework for backpropagation networks. *Neural Computation*, 4(3):448–472, 1992.

[33] M. Stone. An asymptotic equivalence of choice of model by cross-validation and Akaike's Criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):44–47., 1977.

[34] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 186(1007):453–461, 1946.

[35] H. Jeffreys. *The theory of probability.* Oxford University Press, 1939.

[36] Robert E Kass and Larry Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 1996.

[37] Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

[38] Alan E Gelfand and Dipak K Dey. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 501–514, 1994.

[39] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society*, B64:583–639, 2002.

[40] Aki Vehtari and Janne Ojanen. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6:142–228, 2012.

[41] Kenneth P. Burnham and David R. Anderson. Multimodel inference: Understanding AIC and BIC in model selection. *Socialogical Methods & Research*, 33(2):261–304, 2004.

[42] Ronald Aylmer Fisher. Moments and product moments of sampling distributions. *Proceedings of the London Mathematical Society*, 2(1):199–238, 1930.

[43] John W Tukey. Keeping moment-like sampling computations simple. *The Annals of Mathematical Statistics*, pages 37–54, 1956.

[44] Jun Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264, 1997.

Appendix A: Effective temperature of confinement

To calculate the free energy $\mathcal{F}$ of a free particle confined to a volume $V = L^3$, we calculate the partition function by integrate over available phase space:

$$\mathcal{Z}(\beta) = \int \frac{d^K\boldsymbol{p}\, d^K\boldsymbol{x}}{(2\pi\hbar)^K} e^{-\beta H(\boldsymbol{p},\boldsymbol{x})} \tag{A1}$$

$$= \frac{e^{-\beta E_0} L^K}{(2\pi\hbar)^K} \left(\int dp\, e^{\frac{-\beta p^2}{2m}} \right)^K = \left(\frac{mL^2}{2\pi\hbar^2\beta} \right)^{K/2} e^{-\beta E_0} \tag{A2}$$

The Free enregy is then

$$\mathcal{F}(\beta) = E_0 + \frac{K}{2\beta} \log \frac{mL^2}{2\pi\hbar^2\beta} = E_0 + \frac{K}{2\beta} \log \frac{\beta_0}{\beta} \tag{A3}$$

where we have made the identifications

$$\beta_0 = \frac{mL^2}{2\pi\hbar^2} \quad \text{and} \quad K = 3. \tag{A4}$$

β_0 can be interpreted as the inverse of the (typically negligibly small) temperature at which the thermal de Broglie wavelength of the confined particle is on the order of the width of the confining box.

Appendix B: Normal Model sample and Shannon entropies

The normal model has a likelihood which is conveniently written as (in the parametrization where the known covariance matrix is proportional to the identity):

$$\hat{H}_x(\boldsymbol{\theta}) = \frac{K}{2} \log 2\pi\sigma^2 + \frac{(\boldsymbol{\mu} - \overline{\boldsymbol{\mu}})^2}{2\sigma^2} + \frac{K(N-1)}{2N} \frac{\overline{\sigma^2}}{\sigma^2} \tag{B1}$$

where we have written the sample average likelihood in terms of the unbiased estimates for the mean and variance (k-statistics). This simplifies derivatives with respect to sample size because the k-statistics, being unbiased, are not expected to change with sample size [42, 43]. The Shannon entropy

$$\overline{H_0} = \frac{K}{2} \log 2\pi\sigma^2 + \frac{K}{2} = \overline{\hat{H}_x(\hat{\boldsymbol{\theta}}_X)} + \frac{K}{2N}, \tag{B2}$$

follows from Eqn. B1 as the squared deviation of the sample mean from the true mean obeys a χ^2_K which has expected value of K.

In the expression for the cross-entropy, the χ^2_K is non-central

$$H(\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \frac{K}{2} \log 2\pi\sigma^2 + \frac{(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^2}{2\sigma^2} + \frac{K}{2}. \tag{B3}$$

1. Jeffreys prior satisfies the Generalized Principle of Indifference in the large-sample-size limit

In the large-sample-size limit, the partition function can be evaluated using the Lapalce (saddle-point) approximation and the resulting prior is proportional to the Jeffreys prior. The integral is evaluated by expanding around the minimum of $\hat{H}_X(\boldsymbol{\theta})$, the maximum likelihood estimator: $\hat{\boldsymbol{\theta}}_X$,

$$Z(X^N) = \int_{\boldsymbol{\Theta}} \mathrm{d}\boldsymbol{\theta}\; \varpi(\boldsymbol{\theta})\; \exp[-N\hat{H}_X(\boldsymbol{\theta})] \tag{B4}$$

$$= e^{-N\hat{H}_X(\hat{\boldsymbol{\theta}}_X)} \int_{\boldsymbol{\Theta}} \mathrm{d}\boldsymbol{\theta}\; \varpi(\boldsymbol{\theta})\; \exp[-\frac{N}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_X)^\intercal I(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_X) + \ldots] \tag{B5}$$

$$\approx e^{-N\hat{H}_X(\hat{\boldsymbol{\theta}}_X)} \left(\frac{2\pi}{N(\det I)^{1/K}} \right)^{K/2} \varpi(\boldsymbol{\theta}_X) \tag{B6}$$

So that

$$\log Z(X^N) = -N\hat{H}_X(\hat{\boldsymbol{\theta}}_X) + \frac{K}{2}\log\frac{2\pi}{N(\det I)^{1/K}} + \log\varpi(\boldsymbol{\theta}_X) + O(N^{-1}) \tag{B7}$$

$$\langle\log Z(X^N)\rangle_{\boldsymbol{\theta}_0} = -N\overline{H}_0 + \frac{K}{2} + \frac{K}{2}\log\frac{2\pi}{N(\det I)^{1/K}} + \log\varpi(\boldsymbol{\theta}_0) + O(N^{-1}) \tag{B8}$$

$$N\langle U(X^N)\rangle_{\boldsymbol{\theta}_0} = N\overline{H}_0 + \frac{K}{2} + O(N^{-1}) \tag{B9}$$

$$\overline{S}(\boldsymbol{\theta}_0,\varpi,N) = \frac{K}{2}\log\frac{2\pi}{N(\det I)^{1/K}} + K + \log\varpi(\boldsymbol{\theta}_0) + O(N^{-1}) \tag{B10}$$

If we enforce the generalized principle of indifference, ignoring small factors of N.

$$0 = S(\boldsymbol{\theta}_0, w, N) \tag{B11}$$

$$\log w(\boldsymbol{\theta}_0) = -K - \frac{K}{2}\log\frac{2\pi}{N(\det I)^{1/K}} \tag{B12}$$

$$w(\boldsymbol{\theta}_0) = (\det I)^{1/2}\left(\frac{N}{2\pi}\right)^{K/2} e^{-K} \tag{B13}$$

The generalized principle of indifference results in the Jeffries prior in the large sample size limit, but with a constant factor which is important in model selection. The factor of e^{-K} expresses the Akaike weighting.

This constant factor shows a characteristic of the w-prior: it has an important sample-size dependence. This sample size dependence will in general break the de-Finetti likelihood principle: that the prior should not depend on the nature of the data-generating procedure (including the sample size). The departure from the likelihood principle is the origin of the departure from the conventional Bayesian model selection behavior.

2. Cross Validation as finite difference approximation

The log-predictive distribution can be written as a finite difference

$$\log q(X_i|X^{\neq i}) = \log Z(X^N) - \log Z(X^{\neq i}) \tag{B14}$$

$$\tag{B15}$$

We can interpret the $\log q(X_i|X^{\neq i})$ as a finite difference estimate of the the sample size derivative of the free energy. We take the mean over all permutations of the data so that this estimate is symmetric with respect to all data points. Under expectation, analytically continuing sample size, the LOOCV relationship to the internal energy is clear:

$$\langle\log q(X_i|X^{\neq i})\rangle \approx \frac{\partial}{\partial N}\langle\log Z(X^N)\rangle + O(N^{-1}) \tag{B16}$$

Appendix C: Alternate Statistical Mechanics / Inference Correspondence

In establishing the correspondence between inference and statistical mechanics, we identify the partition function Z as the marginal likelihood and $N \leftrightarrow \beta$. This is by no means the only choice, or even the most common choice. For instance Watanabe [15] instead chooses to define the inverse-temperature β so that the likelihood is given by $q^\beta(X^N|\boldsymbol{\theta})$, that is raised to an arbitrary power. This identification has two advantages: i.) It seems to be more closely related to the physical temperature, which can be varied independently with the strength of the quenched disorder ii.) It allows one to very simply interpolate between a Bayesian posterior (given by $\beta = 1$) and the point estimates of the MLE's (given by $\beta = \infty$). This temperature has also been applied in tempering schemes in MCMC methods, and simulated annealing—increasing the temperature promotes a better exploration of the sample space (chain-mixing) that can be used to better sample multimodal distributions, or find the minima in a rough function.

On the other hand, there are two disadvantages of a power β relative to $N \leftrightarrow \beta$ which we believe outweigh the advantages: i.) First, it is not a preexisting statistical parameter within the Bayesian framework. ii.) Second, the internal energy under this other choice of β is not the predictive performance U. Consequently, the principle of indifference which results from a likelihood-power β does not induce the Akaike weights as the model averaging procedure. Instead $\overline{U} = \overline{H}_0$, which does not encode a realization of Occam's razor.

Thermodynamic expressions using both definitions may give somewhat complementary information, and which is useful will depend on the context. We do not believe that statistical mechanics prescribes a uniquely-correct procedure for objective Bayesian inference. It is the reproduction of a principled model selection criteria, AIC with its proven asymptotic efficiency [44] that justifies the proposed correspondence in the context of model selection.