Efficient Algorithms for Searching the Minimum Information Partition in Integrated Information Theory

Jun Kitazono [1,2,*], Ryota Kanai [1] and Masafumi Oizumi [1,3,*]

[1]   Araya, Inc., Toranomon 15 Mori Building, 2-8-10 Toranomon, Minato-ku, Tokyo 105-0001, Japan;
      kanair@araya.org
[2]   Graduate School of Engineering, Kobe University, 1-1 Rokkodai-cho, Nada-ku, Kobe-shi,
      Hyogo 657-8501, Japan
[3]   RIKEN Brain Science Institute, 2-1 Hirosawa Wako City, Saitama 351-0198, Japan
[*]   Correspondence: kitazono@araya.org (J.K.); oizumi@araya.org (M.O.); Tel.: +81-3-6550-9977 (J.K. & M.O.)

Abstract: The ability to integrate information in the brain is considered to be an essential property for cognition and consciousness. Integrated Information Theory (IIT) hypothesizes that the amount of integrated information (Φ) in the brain is related to the level of consciousness. IIT proposes that, to quantify information integration in a system as a whole, integrated information should be measured across the partition of the system at which information loss caused by partitioning is minimized, called the Minimum Information Partition (MIP). The computational cost for exhaustively searching for the MIP grows exponentially with system size, making it difficult to apply IIT to real neural data. It has been previously shown that, if a measure of Φ satisfies a mathematical property, submodularity, the MIP can be found in a polynomial order by an optimization algorithm. However, although the first version of Φ is submodular, the later versions are not. In this study, we empirically explore to what extent the algorithm can be applied to the non-submodular measures of Φ by evaluating the accuracy of the algorithm in simulated data and real neural data. We find that the algorithm identifies the MIP in a nearly perfect manner even for the non-submodular measures. Our results show that the algorithm allows us to measure Φ in large systems within a practical amount of time.

Keywords: integrated information theory; integrated information; minimum information partition; submodularity; Queyranne's algorithm; consciousness

1. Introduction

The brain receives various information from the external world. Integrating this information is an essential property for cognition and consciousness [1]. In fact, phenomenologically, our consciousness is unified. For example, when we see an object, we cannot experience only its shape independently of its color. Conversely, we cannot experience only the left half of the visual field independently of the right half. Integrated Information Theory (IIT) of consciousness considers that the unification of consciousness should be realized by the ability of the brain to integrate information [2–4]. That is, the brain has internal mechanisms to integrate information about the shape and color of an object or information of the right and left visual field, and therefore our visual experiences are unified. IIT proposes to quantify the degree of information integration by an information theoretic measure "integrated information" and hypothesizes that integrated information is related to the level of consciousness. Although the hypothesis is indirectly supported by experiments which showed the breakdown of effective connectivity in the brain during loss

of consciousness [5,6], only a few studies have directly quantified integrated information in real neural data [7–10] because of the computational difficulties described below.

Conceptually, integrated information quantifies the degree of interaction between parts or equivalently, the amount of information loss caused by splitting a system into parts [11,12]. IIT proposes that integrated information should be quantified between the least interdependent parts so that it quantifies information integration in a system as a whole. For example, if a system consists of two independent subsystems, the two subsystems are the least interdependent parts. In this case, integrated information is 0, because there is no information loss when the system is partitioned into the two independent subsystems. Such a critical partition of the system is called the Minimum Information Partition (MIP), where information is minimally lost, or equivalently where integrated information is minimized. In general, searching for the MIP requires an exponentially large amount of computational time because the number of partitions exponentially grows with the arithmetic growth of system size N. This computational difficulty hinders the application of IIT to experimental data, despite its potential importance in consciousness research and even in broader fields of neuroscience.

In the present study, we exploit a mathematical concept called submodularity to resolve the combinatorial explosion of finding the MIP. Submodularity is an important concept in set functions which is analogous to convexity in continuous functions. It is known that an exponentially large computational cost for minimizing an objective function is reduced to the polynomial order if the objective function satisfies submodularity. Previously, Hidaka and Oizumi showed that the computational cost for finding the MIP is reduced to $O(N^3)$ [13] by utilizing Queyranne's submodular optimization algorithm [14]. They used mutual information as a measure of integrated information that satisfies submodularity. The measure of integrated information used in the first version of IIT (IIT 1.0) [2] is based on mutual information. Thus, if we consider mutual information as a practical approximation of the measure of integrated information in IIT 1.0, Queyranne's algorithm can be utilized for finding the MIP. However, the practical measures of integrated information in the later versions of IIT [12,15–17] are not submodular.

In this paper, we aim to extend the applicability of submodular optimization to non-submodular measures of integrated information. We specifically consider the three measures of integrated information: mutual information Φ_{MI} [2], stochastic interaction Φ_{SI} [15,18,19], and geometric integrated information Φ_{G} [12]. Mutual information is strictly submodular but the others are not. Oizumi et al. previously showed a close relationship among these three measures [12,20]. From this relationship, we speculate that Queyranne's algorithm might work well for the non-submodular measures. Here, we empirically explore to what extent Queyranne's algorithm can be applied to the two non-submodular measures of integrated information by evaluating the accuracy of the algorithm in simulated data and real neural data. We find that Queyranne's algorithm identifies the MIP in a nearly perfect manner even for the non-submodular measures. Our results show that Queyranne's algorithm can be utilized even for non-submodular measures of integrated information and makes it possible to practically compute integrated information across the MIP in real neural data, such as multi-unit recordings used in Electroencephalography (EEG) and Electrocorticography (ECoG), which typically consist of around 100 channels. Although the MIP was originally proposed in IIT for understanding consciousness, it can be utilized to analyze any system irrespective of consciousness such as biological networks, multi-agent systems, and oscillator networks. Therefore, our work would be beneficial not only for consciousness studies but also to other research fields involving complex networks of random variables.

This paper is organized as follows. We first explain that the three measures of integrated information, Φ_{MI}, Φ_{SI}, and Φ_{G}, are closely related from a unified theoretical framework [12,20] and there is an order relation among the three measures: $\Phi_{\mathrm{MI}} \geq \Phi_{\mathrm{SI}} \geq \Phi_{\mathrm{G}}$. Next, we compare the partition found by Queyranne's algorithm with the MIP found by exhaustive search in randomly generated small networks ($N = 14$). We also evaluate the performance of Queyranne's algorithm in larger networks ($N \sim 20$ and 50 for Φ_{SI} and Φ_{G}, respectively). Since the exhaustive search is intractable, we compare Queyranne's algorithm with a different optimization algorithm called the replica exchange

Markov Chain Monte Carlo (REMCMC) method [21–24]. Finally, we evaluate the performance of Queyranne's algorithm in ECoG data recorded in monkeys and investigate the applicability of the algorithm in real neural data.

2. Measures of Integrated Information

Let us consider a stochastic dynamical system consisting of N elements. We represent the past and present states of the system as $X = (X_1, \ldots, X_N)$ and $X' = (X'_1, \ldots, X'_N)$, respectively. In the case of a neural system, the variable X can be signals of multi-unit recordings, EEG, ECoG, functional magnetic resonance imaging (fMRI), etc. Conceptually, integrated information is designed to quantify the degree of spatio-temporal interactions between subsystems. The previously proposed measures of integrated information are generally expressed as the Kullback–Leibler divergence between the actual probability distribution $p(X, X')$ and a "disconnected" probability distribution $q(X, X')$ where interactions between subsystems are removed [12].

$$\Phi = \min_q D_{KL} \left(p\left(X, X'\right) || q\left(X, X'\right) \right), \tag{1}$$

$$= \min_q \sum_{x, x'} p\left(x, x'\right) \log \frac{p\left(x, x'\right)}{q\left(x, x'\right)}. \tag{2}$$

The Kullback–Leibler divergence measures the difference between the probability distributions, and can be interpreted as the information loss when $q(X, X')$ is used to approximate $p(X, X')$ [25]. Thus, integrated information is interpreted as information loss caused by removing interactions. In Equation (2), the minimum over q should be taken to find the best approximation of p, while satisfying the constraint that the interactions between subsystems are removed [12].

There are many ways of removing interactions between units, which lead to different disconnected probability distributions q, and also different measures of integrated information (Figure 1). The arrows indicate influences across different time points and the lines without arrowheads indicate influences between elements at the same time. Below, we will show that three different measures of integrated information are derived from different probability distributions q.
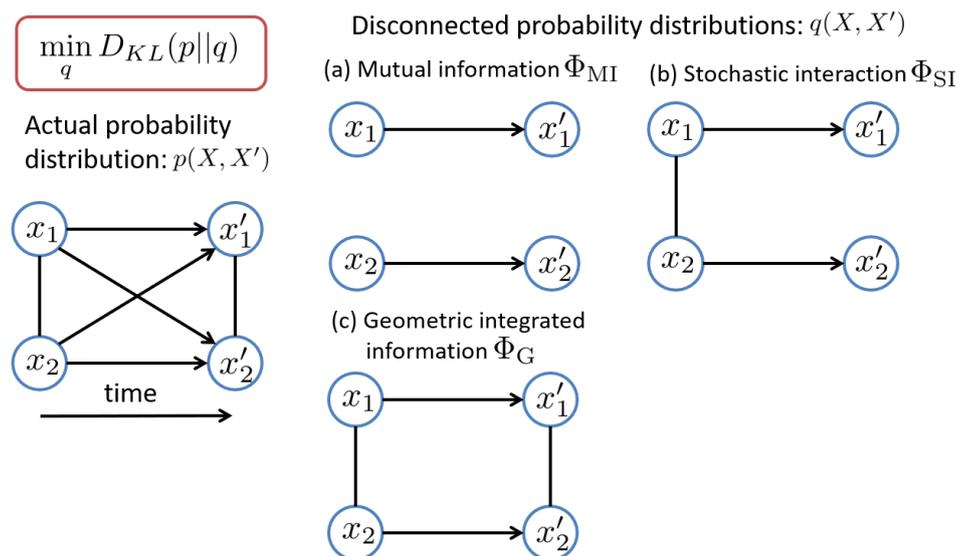

Figure 1. Measures of integrated information represented by the Kullback–Leibler divergence between the actual distribution p and q: (**a**) mutual information; (**b**) stochastic interaction; and (**c**) geometric integrated information. The arrows indicate influences across different time points and the lines without arrowheads indicate influences between elements at the same time. This figure is modified from [12].

2.1. Multi (Mutual) Information Φ_{MI}

First, consider the following partitioned probability distribution q,

$$q\left(X, X'\right) = \prod_{i=1}^{K} q\left(M_i, M_i'\right), \tag{3}$$

where the whole system is partitioned into K subsystems and the past and present states of the i-th subsystem are denoted by M_i and M_i', respectively, i.e., $X = (M_1, \ldots, M_K)$ and $X' = (M_1', \ldots, M_K')$. Each subsystem consists of one or multiple elements. The distribution $q\left(M_i, M_i'\right)$ is the marginalized distribution

$$q\left(M_i, M_i'\right) = \sum_{X \backslash M_i, X' \backslash M_i'} q(X, X'), \tag{4}$$

where $X \backslash M_i$ and $X' \backslash M_i'$ are the complement of M_i and M_i', that is, $(M_1, \ldots, M_{i-1}, M_{i+1}, \ldots, M_K)$ and $(M_1', \ldots, M_{i-1}', M_{i+1}', \ldots, M_K')$, respectively. In this model, all of the interactions between the subsystems are removed, i.e., the subsystems are totally independent (Figure 1a). In this case, the corresponding measure of integrated information is given by

$$\Phi_{\mathrm{MI}} = \sum_{i} H(M_i, M_i') - H(X, X'), \tag{5}$$

where $H(\cdot, \cdot)$ represents the joint entropy. This measure is called total correlation [26] or multi information [27]. As a special case when the number of subsystems is two, this measure is simply equivalent to the mutual information between the two subsystems,

$$\Phi_{\mathrm{MI}} = H(M_1, M_1') + H(M_2, M_2') - H(X, X'). \tag{6}$$

The measure of integrated information used in the first version of IIT is based on mutual information but is not identical to mutual information in Equation (6). The critical difference is that the measures in IIT are based on perturbation and those considered in this study are based on observation. In IIT, a perturbational approach is used for evaluating probability distributions, which attempts to quantify actual causation by perturbing a system into all possible states [2,4,11,28]. The perturbational approach requires full knowledge of the physical mechanisms of a system, i.e., how the system behaves in response to all possible perturbations. The measure defined in Equation (6) is based on an observational probability distribution that can be estimated from empirical data. Since we aim for the empirical application of our method, we do not consider the perturbational approach in this study.

2.2. Stochastic Interaction Φ_{SI}

Second, consider the following partitioned probability distribution q,

$$q\left(X'|X\right) = \prod_{i} q\left(M_i'|M_i\right), \tag{7}$$

which partitions the transition probability from the past X to the present X' in the whole system into the product of the transition probability in each subsystem. This corresponds to removing the causal influences from M_i to M_j' $(j \neq i)$ as well as the equal time influences at present between M_i' and M_j' $(j \neq i)$ (Figure 1b). In this case, the corresponding measure of integrated information is given by

$$\Phi_{SI} = \sum_i H(M_i'|M_i) - H(X'|X), \tag{8}$$

where $H(\cdot|\cdot)$ indicates the conditional entropy. This measure was proposed as a practical measure of integrated information by Barrett and Seth [15] following the measure proposed in the second version of IIT (IIT 2.0) [11]. This measure was also independently derived by Ay as a measure of complexity [18,19].

2.3. Geometric Integrated Information Φ_{G}

Aiming at only the causal influences between parts, Oizumi et al. [12] proposed to measure integrated information with the probability distribution that satisfies

$$q\left(M_i'|X\right) = q\left(M_i'|M_i\right), \forall i \tag{9}$$

which means the present state of a subsystem i, M_i' only depends on its past state M_i. This corresponds to removing only the causal influences between subsystems while retaining the equal-time interactions between them (Figure 1c). The constraint Equation (9) is equivalent to the Markov condition

$$q(M_i', M_i^c|M_i) = q(M_i'|M_i)q(M_i^c|M_i), \forall i \tag{10}$$

where M_i^c is the complement of M_i, that is, $M_i^c = X \backslash M_i$. This means when M_i is given, M_i' and M_i^c are independent. In other words, the causal interaction between M_i^c and M_i' is only via M_i.

There is no closed-form expression for this measure in general. However, if the probability distributions are Gaussian, we can analytically solve the minimization over q (see Appendix A).

3. Minimum Information Partition

In this section, we provide the mathematical definition of Minimum Information Partition (MIP). Then, we formulate the search for MIP as an optimization problem of a set function. The MIP is the partition that divides a system into the least interdependent subsystems so that information loss caused by removing interactions among the subsystems is minimized. The information loss is quantified by the measure of integrated information. Thus, the MIP, π_{MIP}, is defined as a partition (since the minimizer is not necessarily unique, strictly speaking, there could be multiple MIPs), where integrated information is minimized:

$$\pi_{\mathrm{MIP}} := \arg\min_{\pi \in \mathcal{P}} \Phi(\pi), \tag{11}$$

where $\mathcal{P}$ is a set of partitions. In general, $\mathcal{P}$ is the universal set of partitions, including bi-partitions, tri-partitions, and so on. In this study, however, we focus only on bi-partitions for simplicity and computational time. Note that, although Queyranne's algorithm [14] is limited to bi-partitions, the algorithm can be extended to higher-order partitions [13]. See Section 7 for more details. By a bi-partition, a whole system Ω is divided into a subset S ($S \subset \Omega, S \neq \emptyset$) and its complement $\bar{S} = \Omega \setminus S$. Since a bi-partition is uniquely determined by specifying a subset S, integrated information can be considered as a function of a set S, $\Phi(S)$. Finding the MIP is equivalent to finding the subset, S_{MIP}, that achieves the minimum of integrated information:

$$S_{\mathrm{MIP}} := \arg\min_{S \subset \Omega, S \neq \emptyset} \Phi(S). \tag{12}$$

In this way, the search of the MIP is formulated as an optimization problem of a set function.

Since the number of bi-partitions for the system with N-elements is $2^{N-1} - 1$, exhaustive search of the MIP in a large system is intractable. However, by formulating the MIP search as an optimization of a set function as above, we can take advantage of a discrete optimization technique and can reduce computational costs to a polynomial order, as described in the next section.

4. Submodular Optimization

The submodularity is an important concept in set functions, which is an analogue of convexity in continuous functions [29]. When objective functions are submodular, efficient algorithms are available for solving optimization problems. In particular, for symmetric submodular functions, there is a well-known algorithm by Queyranne which minimizes them [14]. We utilize this method for finding the MIP in this study.

4.1. Submodularity

Mathematically, the submodularity is defined as follows.

Definition 1 (Submodularrity). *Let Ω be a finite set and 2^Ω its power set. A set function $f : 2^\Omega \to \mathbb{R}$ is submodular if it satisfies the following inequality for any $S, T \subseteq \Omega$:*

$$f(S) + f(T) \geq f(S \cup T) + f(S \cap T).$$

Equivalently, a set function $f : 2^\Omega \to \mathbb{R}$ is submodular if it satisfies the following inequality for any $S, T \subseteq \Omega$ with $S \subseteq T$ and for any $u \in \Omega \setminus T$:

$$f(S \cup \{u\}) - f(S) \geq f(T \cup \{u\}) - f(T).$$

The second inequality means that the function increases more when an element is added to a smaller subset than when the element is added to a bigger subset.

4.2. Queyranne's Algorithm

A set function f is called symmetric if $f(S) = f(\Omega \setminus S)$ for any $S \subseteq \Omega$. Integrated information $\Phi(S)$ computed by bi-partition is a symmetric function, because S and $\Omega \setminus S$ specifies the same bi-partition. If a function is symmetric and submodular, we can find the minimum of the function by Queyranne's algorithm with $O(N^3)$ function calls [14].

4.3. Submodularity in Measures of Integrated Information

In a previous study, Queyranne's algorithm was utilized to find the MIP when Φ_{MI} is used as the measure of integrated information [13]. As shown previously, Φ_{MI} is submodular [13]. However, the other measures of integrated information are not submodular. In this study, we apply Queyranne's algorithm to non-submodular functions, Φ_{SI} and Φ_{G}. When the objective functions are not submodular, Queyranne's algorithm does not necessarily find the MIP. We evaluate how accurately Queyranne's algorithm can find the MIP when it is used for non-submodular measures of integrated information. There is an order relation among the three measures of integrated information [12],

$$\Phi_{\mathrm{G}} \leq \Phi_{\mathrm{SI}} \leq \Phi_{\mathrm{MI}}. \tag{13}$$

This inequality can be graphically understood from Figure 1. The more the connections are removed, the larger the corresponding integrated information (the information loss) is. That is, Φ_{G} measures only the causal influences between subsystems, Φ_{SI} measures the equal-time interactions between the present states as well as the causal influences between subsystems, and Φ_{MI} measures all the interactions between the subsystems. Thus, Φ_{SI} is closer to Φ_{MI} than Φ_{G} is. This relationship implies that Φ_{SI} would behave more similarly to a submodular measure Φ_{MI} than Φ_{G} does. Thus, one may surmise that Queyranne's algorithm would work more accurately for Φ_{SI} than for Φ_{G}. As we will show in Section 6.2, this is indeed the case. However, the difference is rather small because Queyranne's algorithm works almost perfectly for both measures, Φ_{SI} and Φ_{G}.

5. Replica Exchange Markov Chain Monte Carlo Method

To evaluate the accuracy of Queyranne's algorithm, we compare the partition found by Queyranne's algorithm with the MIP found by the exhaustive search when the number of elements n is small enough ($n \lesssim 20$). However, when n is large, we cannot know the MIP because the exhaustive search is unfeasible. To evaluate the performance of Queyranne's algorithm in a large system, we compare it with a different method, the Replica Exchange Markov Chain Monte Carlo (REMCMC) method [21–24]. REMCMC, also known as parallel tempering, is a method to draw samples from probability distributions. REMCMC is an improved version of the MCMC methods. Here, we briefly explain how the MIP search problem is represented as a problem of drawing samples from a probability distribution. Details of the REMCMC method are given in Appendix B.

Let us define a probability distribution $p(S; \beta)$ using integrated information $\Phi(S)$ as follows:

$$p(S; \beta) \propto \exp(-\beta \Phi(S)), \tag{14}$$

where $\beta(> 0)$ is a parameter called inverse temperature. This probability is higher/lower when $\Phi(S)$ is smaller/larger. The MIP gives the highest probability by definition. If we can draw samples from this distribution, we can selectively scan subsets with low integrated information and efficiently find the MIP, compared to randomly exploring partitions independent of the value of integrated information. Simple MCMC methods such as the Metropolis method, which draw samples from Equation (14) with a single value of β, often suffer from the problem of slow convergence. That is, a sample sequence is trapped in a local minimum and the sample distribution takes time to converge to the target distribution. REMCMC aims at overcoming this problem by drawing samples in parallel from distributions with multiple values of β and by continually exchanging the sampled sequences between neighboring β (see Appendix B for more details).

6. Results

We first evaluated the performance of Queyranne's algorithm in simulated networks. Throughout the simulations below, we consider the case where the variable X obeys a Gaussian distribution for the ease of computation. As shown in Appendix A, the measures of integrated information, Φ_{SI} and Φ_{G} can be analytically computed. Note that, although Φ_{SI} and Φ_{G} can be computed in principle even when the distribution of X is not Gaussian, it is practically very hard to compute them in large systems because the computation of Φ involves summation over all possible X. Specifically, we consider the first order autoregressive (AR) model,

$$X' = AX + E, \tag{15}$$

where X and X' are present states and past states of a system, A is the connectivity matrix, and E is Gaussian noise. The stationary distribution of this AR model is considered. The stationary distribution of $p(X, X')$ is a Gaussian distribution. The covariance matrix of $p(X, X')$ consists of covariance of X, $\Sigma(X)$, and cross-covariance of X and X', $\Sigma(X, X')$. $\Sigma(X)$ is computed by solving the following equation,

$$\Sigma(X) = A\Sigma(X)A^T + \Sigma(E). \tag{16}$$

$\Sigma(X, X')$ is given by

$$\Sigma(X, X') = \Sigma(X)A^T. \tag{17}$$

By using these covariance matrices, Φ_{SI} and Φ_{G} are analytically calculated [12] (see Appendix A). The details of the parameter settings are described in each subsection.

6.1. Speed of Queyranne's Algorithm Compared With Exhaustive Search

We first evaluated the computational time of the search using Queyranne's algorithm and compared it with that of the exhaustive search when the number of elements N changed. The connectivity matrices

A were randomly generated. Each element of the connection matrix *A* was sampled from a normal distribution with mean 0 and variance $0.01/N$. The covariance of Gaussian noise *E* was generated from a Wishart distribution $\mathcal{W}(\sigma I, 2N)$ with covariance σI and degrees of freedom $2N$, where σ corresponded to the amount of noise *E* and *I* was the identity matrix. The Wishart distribution is a standard distribution for symmetric positive-semidefinite matrices [30,31]. Typically, the distribution is used to generate covariance matrices and inverse covariance (precision) matrices. For more practical details, see for example, Ref. [31]. We set σ to 0.1. The number of elements *N* was changed from 3 to 60. All computation times were measured on a machine with an Intel Xeon Processor E5-2680 at 2.70GHz. All the calculations were implemented in MATLAB R2014b.

We fitted the computational time of the search using Queyranne's algorithm for Φ_{SI} and Φ_{G} with straight lines, although the computational time for large *N* is a little deviated from the straight lines (Figure 2a,b). In Figure 2a, the red circles, which indicate the computational time of the search using Queyranne's algorithm for Φ_{SI}, are roughly approximated by the red solid line, $\log_{10} T = 3.066 \log_{10} N - 3.838$. In contrast, the black triangles, which indicate those of the exhaustive search, are fit by the black dashed line, $\log_{10} T = 0.2853N - 3.468$. This means that the computational time of the search using Queyranne's algorithm increases in polynomial order ($T \propto N^{3.066}$), while that of the exhaustive search exponentially increases ($T \propto 1.929^N$). For example, when $N = 100$, Queyranne's algorithm takes ~197 s while the exhaustive search takes 1.16×10^{25} s. This is in practice impossible to compute even with a supercomputer. Similarly, as shown in Figure 2b, when Φ_{G} is used, the search using Queyranne's algorithm roughly takes $T \propto N^{4.776}$ while the exhaustive search takes $T \propto 2.057^N$. Note that the complexity of the search using Queyranne's algorithm for Φ_{G} ($O(N^{4.776})$) is much higher than that of Queyranne's algorithm itself ($O(N^3)$). This is because the multi-dimensional equations (Equations (A20) and (A21)) need to be solved by using an iterative method to compute Φ_{G} (see Appendix A).
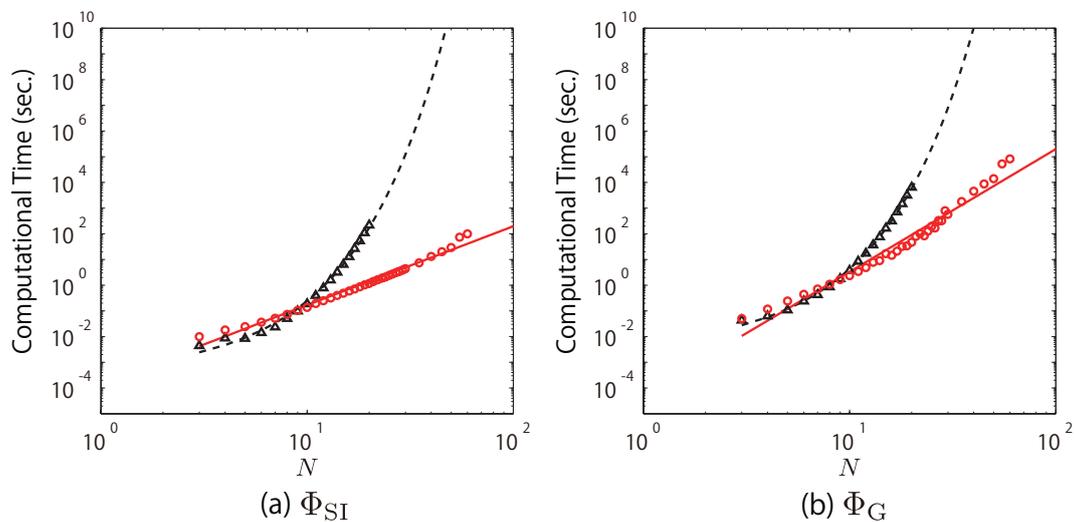

Figure 2. Computational time of the search using Queyranne's algorithm and the exhaustive search. The red circles and the red solid lines indicate the computational time of the search using Queyranne's algorithm and their approximate curves ((**a**) $\log_{10} T = 3.066 \log_{10} N - 3.838$, (**b**) $\log_{10} T = 4.776 \log_{10} N - 4.255$). The black triangles and the black dashed lines indicate the computational time of the exhaustive search and their approximate curves ((**a**) $\log_{10} T = 0.2853N - 3.468$, (**b**) $\log_{10} T = 0.3132N - 2.496$).

6.2. Accuracy of Queyranne's Algorithm

We evaluated the accuracy of Queyranne's algorithm by comparing the partition found by Queyranne's algorithm with the MIP found by exhaustive search. We used Φ_{SI} and Φ_{G} as the measures

of integrated information. We considered two different architectures in connectivity matrix A of AR models. The first one was just a random matrix: Each element of A was randomly sampled from a normal distribution with mean 0 and variance $0.01/N$. The other one was a block matrix consisting of $N/2$ by $N/2$ sub-matrices, $A_{ij}(i, j = 1, 2)$. Each element of diagonal sub-matrices A_{11} and A_{22} was drawn from a normal distribution with mean 0 and variance $0.02/N$. Off-diagonal sub-matrices A_{12} and A_{21} were zero matrices. The covariance of Gaussian noise E in the AR model was generated from a Wishart distribution $\mathcal{W}(\sigma I, 2N)$. The parameter σ was set to 0.1 or 0.01. The number of elements N was set to 14. We randomly generated 100 connectivity matrices A and $\Sigma(E)$ for each setting and evaluated performance using the following four measures. The following measures are averaged over 100 trials:

- **Correct rate (CR):** Correct rate (CR) is the rate of correctly finding the MIP.
- **Rank (RA):** Rank (RA) is the rank of the partition found by Queyranne's algorithm among all possible partitions. The rank is based on the Φ values computed at each partition. The partition that gives the lowest Φ is rank 1. The highest rank is equal to the number of possible bi-partitions, 2^{N-1}.
- **Error ratio (ER):** Error ratio (ER) is the deviation of the value of integrated information computed across the partition found by Queyranne's algorithm from that computed across the MIP, which is normalized by the mean error computed at all possible partitions. Error ratio is defined by

$$\text{Error Ratio} = \frac{\Phi_Q - \Phi_{\text{MIP}}}{\bar{\Phi} - \Phi_{\text{MIP}}}, \tag{18}$$

where Φ_{MIP}, Φ_Q, and $\bar{\Phi}$ are the amount of integrated information computed across the MIP, that computed across the partition found by Queyranne's algorithm, and the mean of the amounts of integrated information computed across all possible partitions, respectively.

- **Correlation (CORR):** Correlation (CORR) is the correlation between the partition found by Queyranne's algorithm and the MIP found by the exhaustive search. Let us represent a bi-partition of N-elements as an N-dimensional vector $\sigma = (\sigma_1, \ldots, \sigma_N) \in \{-1, 1\}^N$, where ± 1 indicates one of the two subgroups. The absolute value of the correlation between the vector given by the MIP (σ^{MIP}) and that given by the partition found by Queyranne's algorithm (σ^Q) is computed:

$$|\text{corr}(\sigma^{\text{MIP}}, \sigma^Q)| = \left| \frac{\sum_{i=1}^N (\sigma_i^{\text{MIP}} - \bar{\sigma}^{\text{MIP}})(\sigma_i^Q - \bar{\sigma}^Q)}{\sqrt{\sum_{i=1}^N (\sigma_i^{\text{MIP}} - \bar{\sigma}^{\text{MIP}})^2 \sum_{i=1}^N (\sigma_i^Q - \bar{\sigma}^Q)^2}} \right|, \tag{19}$$

where $\bar{\sigma}^{\text{MIP}}$ and $\bar{\sigma}^Q$ are the means of σ_i^{MIP} and σ_i^Q, respectively.

The results are summarized in Table 1. This table shows that, when Φ_{SI} was used, Queyranne's algorithm perfectly found the MIPs for all 100 trials, even though Φ_{SI} is not strictly submodular. Similarly, when Φ_G was used, Queyranne's algorithm almost perfectly found the MIPs. The correct rate was 100% for the normal models and 97% for the block structured models. Additionally, even when the algorithm missed the MIP, the rank of the partition found by the algorithm was 2 or 3. The averaged rank over 100 trials were 1.03 and 1.05 for the block structured models. In addition, the error ratio in error trials were around 0.1 and the average error ratios were very small. See Appendix C for box plots of the values of the integrated information at all the partitions. Thus, such miss trials would not affect evaluation of the amount of integrated information in practice. However, in terms of partitions, the partitions found by Queyranne's algorithm in error trials were markedly different from the MIPs. In the block structured model, the MIP for Φ_G was the partition that split the system in halves. In contrast, the partitions found by Queyranne's algorithm were one-vs-all partitions.

Table 1. Accuracy of Queyranne's algorithm.

Model		Φ_{SI}				Φ_{G}			
A	σ	CR	RA	ER	CORR	CR	RA	ER	CORR
Normal	0.01	100%	1	0	1	100%	1	0	1
	0.1	100%	1	0	1	100%	1	0	1
Block	0.01	100%	1	0	1	97%	1.05	2.38×10^{-3}	0.978
	0.1	100%	1	0	1	97%	1.03	9.11×10^{-4}	0.978

In summary, Queyranne's algorithm perfectly worked for Φ_{SI}. With regards to Φ_{G}, although Queyranne's algorithm almost perfectly evaluated the amount of integrated information, we may need to treat partitions found by the algorithm carefully. This slight difference in performance between Φ_{SI} and Φ_{G} can be explained by the order relation in Equation (13). Φ_{SI} is closer to the strictly submodular function Φ_{MI} than Φ_{G} is, which we consider to be why Queyranne's algorithm worked better for Φ_{SI} than Φ_{G}.

6.3. Comparison between Queyranne's Algorithm and REMCMC

We evaluated the performance of Queyranne's algorithm in large systems where an exhaustive search is impossible. We compared it with the Replica Exchange Markov Chain Monte Carlo Method (REMCMC). We applied the two algorithms to AR models generated similarly as in the previous section. The number of elements was 50 for Φ_{SI} and 20 for Φ_{G}, respectively. The reason for the difference in *N* is because Φ_{G} requires much heavier computation than Φ_{SI} (see Appendix A). We randomly generated 20 connectivity matrices *A* and $\Sigma(E)$ for each setting. We compared the two algorithms in terms of the amount of integrated information and the number of evaluations of Φ. REMCMC was run until a convergence criterion was satisfied. See Appendix B.3 for details of the convergence criterion.

The results are shown in Tables 2 and 3. "Winning percentage" indicates the fraction of trials each algorithm won in terms of the amount of integrated information at the partition found by each algorithm. We can see that the partitions found by the two algorithms exactly matched for all the trials. We consider that the algorithms probably found the MIPs for the following three reasons. First, it is well known that REMCMC can find a minima if it is run for a sufficiently long time in many applications [24,32–34]. Second, the two algorithms are so different that it is unlikely that they both incorrectly identified the same partitions as the MIPs. Third, Queyranne's algorithm successfully finds the MIPs in smaller systems as shown in the previous section. This fact suggests that Queyranne's algorithm worked well also for the larger systems. Note that, in the case of Φ_{G}, the half-and-half partition is the MIP in the block structured model because $\Phi_{\mathrm{G}} = 0$ under the half-and-half partition. We confirmed that the partitions found by Queyranne's algorithm and REMCMC were both the half-and-half partition for all the 20 trials. Thus, in the block structured case, it is certain that the true MIPs were successfully found by both algorithms.

Table 2. Comparison of Queyranne's algorithm with REMCMC (Φ_{SI}, $N = 50$).

Model		Winning Percentage			Number of Evaluations of Φ		
						REMCMC (Mean $\pm$ std)	
A	σ	Queyranne's	Even	REMCMC	Queyranne's	Converged	Solution Found
Normal	0.01	0%	100%	0%	41,699	$274{,}257 \pm 107{,}969$	8172.6 ± 6291.0
	0.1	0%	100%	0%	41,699	$315{,}050 \pm 112{,}205$	9084.9 ± 7676.4
Block	0.01	0%	100%	0%	41,699	$308{,}976 \pm 110{,}905$	7305.6 ± 6197.0
	0.1	0%	100%	0%	41,699	$339{,}869 \pm 154{,}161$	4533.4 ± 3004.8

Table 3. Comparison of Queyranne's algorithm with REMCMC (Φ_G, $N = 20$).

Model		Winning Percentage			Number of Evaluations of Φ		
						REMCMC (Mean $\pm$ std)	
A	σ	Queyranne's	Even	REMCMC	Queyranne's	Converged	Solution Found
Normal	0.01	0%	100%	0%	2679	$136{,}271 \pm 46{,}624$	862.4 ± 776.3
	0.1	0%	100%	0%	2679	$122{,}202 \pm 46{,}795$	894.3 ± 780.2
Block	0.01	0%	100%	0%	2679	$129{,}770 \pm 88{,}483$	245.2 ± 194.3
	0.1	0%	100%	0%	2679	$146{,}034 \pm 61{,}880$	443.2 ± 642.1

We also evaluated the number of evaluations of Φ in both algorithms before the end of the computational processes. In our simulations, the computational process of Queyranne's algorithm ended much faster than the convergence of REMCMC. Queyranne's algorithm ends at a fixed number of evaluations of Φ depending only on N. In contrast, the number of the evaluations before the convergence of REMCMC depends on many factors such as the network models, the initial conditions, and pseudo random number sequences. Thus, the time of convergence varies among different trials. Note that, by "retrospectively" examining the sequence of the Monte Carlo search, the solutions turned out to be found at earlier points of the Monte Carlo searches than Queyranne's algorithm (which are indicated as "solution found" in Tables 2 and 3). However, it is impossible to stop the REMCMC algorithm at these points where the solutions were found because there is no way to tell whether these points reach the solution until the algorithm is run for enough amount of time.

6.4. Evaluation with Real Neural Data

Finally, to ensure the applicability of Queyranne's algorithm to real neural data, we similarly evaluated the performance with electrocorticography (ECoG) data recorded in a macaque monkey. The dataset is available at an open database, Neurotycho.org (http://neurotycho.org/) [35]. One hundred twenty-eight channel ECoG electrodes were implanted in the left hemisphere. The electrodes were placed at 5 mm intervals, covering the frontal, parietal, temporal, and occipital lobes, and medial frontal and parietal walls. Signals were sampled at a rate of 1k Hz and down-sampled to 100 Hz for the analysis. The monkey "Chibi" was awake with the eyes covered by an eye-mask to restrain visual responses. To remove line noise and artifacts, we performed bipolar re-referencing between nearest neighbor electrode pairs. The number of re-referenced electrodes was 64 in total.

In the first simulation, we evaluated the accuracy. We extracted a 1 min length of the signals of the 64 electrodes. Each 1 min sequence consists of 100 Hz $\times$ 60 s = 6000 samples. Then, we randomly selected 14 electrodes 100 times. We approximated the probability distribution of the signals with multivariate Gaussian distributions. The covariance matrices were computed with a time window of 1 min and a time step of 10 ms. We applied the algorithms to the 100 randomly selected sets of electrodes and measured the accuracy similarly as in Section 6.2. The results are summarized in Table 4. We can see that Queyranne's algorithm worked perfectly for both Φ_{SI} and Φ_G.

Next, we compared Queyranne's algorithm with REMCMC. We applied the two algorithms to the 64 re-referenced signals, and evaluated the performance in terms of the amount of integrated information and the number of evaluations of Φ, as in Section 6.3. We segmented 15 non-overlapping sequences of 1 min each, and computed covariance matrices with a time step of 10 ms. We measured the average performance over the 15 sets. Here, we only used Φ_{SI}, because Φ_G requires heavy computations for 64 dimensional systems. The results are shown in Table 5. We can see that the partitions selected by the two algorithms matched for all 15 sequences. In terms of the amount of computation, Queyranne's algorithm ended much faster than the convergence of REMCMC.

Table 4. Accuracy of Queyranne's algorithm in ECoG data. Randomly-selected 14 electrodes were used.

Φ_{SI}				Φ_{G}			
CR	**RA**	**ER**	**CORR**	**CR**	**RA**	**ER**	**CORR**
100%	1	0	1	100%	1	0	1

Table 5. Comparison of Queyranne's algorithm with REMCMC in ECoG data (SI).

Winning Percentage			Number of Evaluations of Φ		
				REMCMC (Mean $\pm$ std)	
Queyranne's	**Even**	**REMCMC**	**Queyranne's**	**Converged**	**Solution Found**
0%	100%	0%	87,423	607,797 $\pm$ 410,588	15,859 $\pm$ 10,497

7. Discussion

In this study, we proposed an efficient algorithm for searching for the Minimum Information Partition (MIP) in Integrated Information Theory (IIT). The computational time of an exhaustive search for the MIP grows exponentially with the arithmetic growth of system size, which has been an obstacle to applying IIT to experimental data. We showed here that by using a submodular optimization algorithm called Queyranne's algorithm, the computational time was reduced to $O(N^{3.066})$ and $O(N^{4.776})$ for stochastic interaction Φ_{SI} and geometric integrated information Φ_{G}, respectively. These two measures of integrated information are non-submodular, and thus it is not theoretically guaranteed that Queyranne's algorithm will find the MIP. We empirically evaluated the accuracy of the algorithm by comparing it with an exhaustive search in simulated data and in ECoG data recorded from monkeys. We found that Queyranne's algorithm worked perfectly for Φ_{SI} and almost perfectly for Φ_{G}. We also tested the performance of Queyranne's algorithm in larger systems ($N = 20$ and 50 for Φ_{SI} and Φ_{G}, respectively) where the exhaustive search is intractable by comparing it with the Replica Exchange Markov Chain Monte Carlo method (REMCMC). We found that the partitions found by these two algorithms perfectly matched, which suggests that both algorithms most likely found the MIPs. In terms of the computational time, the number of evaluations of Φ taken by Queyranne's algorithm was much smaller than that taken by REMCMC before the convergence. Our results indicate that Queyranne's algorithm can be utilized to effectively estimate MIP even for non-submodular measures of integrated information. Although the MIP is a concept originally proposed in IIT for understanding consciousness, it can be utilized to general network analysis irrespective of consciousness. Thus, the method for searching MIP proposed in this study will be beneficial not only for consciousness studies but for other research fields.

Here, we discuss the pros and cons of Queyranne's algorithm in comparison with REMCMC. Since the partitions found by both algorithms perfectly matched in our experiments, they were equally good in terms of accuracy. With regards to computational time, Queyranne's algorithm ended much faster than the convergence of REMCMC. Thus, Queyranne's algorithm would be a better choice in rather large systems ($N \sim 20$ and 50 for Φ_{SI} and Φ_{G}, respectively). Note that, if we retrospectively examine the sampling sequence in REMCMC, we find that REMCMC found the partitions much earlier than its convergence and that the estimated MIPs did not change in the later parts of sampling process. Thus, if we could introduce a heuristic criterion to determine when to stop the sampling based on the time course of the estimated MIPs, REMCMC could be stopped earlier than its convergence. However, setting such a heuristic criterion is a non-trivial problem. Queyranne's algorithm ends within a fixed number of function calls regardless of the properties of data. If the system size is much larger ($N \gtrsim 100$), Queyranne's algorithm will be computationally very demanding because of $O(N^3)$ time complexity and may not practically work. In that case, REMCMC would work better if the above-mentioned heuristics are introduced to stop the algorithm earlier than the convergence.

As an alternative interesting approach for approximately finding the MIP, a graph-based algorithm was proposed by Toker and Sommer [36]. In their method, to reduce the search space, candidate partitions are selected by a spectral clustering method based on correlation. Then Φ_G is calculated for those candidate partitions, and the best partition is selected. A difference between our method and theirs is whether the search method is fully based on the values of integrated information or not. Our method uses no other quantities than Φ for searching the MIP, while their method uses a graph theoretic measure, which may significantly differ from Φ in some cases. It would be an interesting future work to compare our method and the graph-theoretic methods or combine these methods to develop better search algorithms.

In this study, we considered the three different measures of integrated information, Φ_{MI}, Φ_{SI}, and Φ_G. Of these, Φ_{MI} is submodular but the other two measures, while Φ_{SI} and Φ_G, are not. As we described in Section 4.3, there is a clear order relation among them (Equation (13)). Φ_{SI} is closer to a submodular function Φ_{MI} than Φ_G is. This relation implies that Queyranne's algorithm would work better for Φ_{SI} than for Φ_G. We found that it was actually the case in our experiments because there were a few error trials for Φ_G whereas there were no miss trials for Φ_{SI}. For the practical use of these measures, we note that there are two major differences among the three measures. One is what they quantify. As shown in Figure 1, Φ_G measures only causal interactions between units across different time points. In contrast, Φ_{SI} and Φ_{MI} also measure equal time interactions as well as causal interactions. Φ_G best follows the original concept of IIT in the sense that it measures only the "causal" interactions. One needs to acknowledge the theoretical difference whenever applying one of these measures in order to correctly interpret the obtained results. The other difference is in computational costs. The computational costs of Φ_{MI} and Φ_{SI} are almost the same while that of Φ_G is much larger, because it requires multi-dimensional optimization. Thus, Φ_G may not be practical for the analysis of large systems. In that case, Φ_{MI} or Φ_{SI} may be used instead with care taken of the theoretical difference.

Although in this study we focused on bi-partitions, Queyranne's algorithm can be extended to higher-order partitions [13]. However, the algorithm becomes computationally demanding for higher-order partitions, because the computational complexity of the algorithm for *K*-partitions is $O(N^{3(K-1)})$. This is the main reason why we focused on bi-partitions. Another reason is that there has not been an established way to fairly compare partitions with different *K*. In IIT 2.0, it was proposed that the integrated information should be normalized by the minimum of the entropy of partitioned subsystems [3], while, in IIT 3.0, it was not normalized [4]. Note that, when integrated information is not normalized, the MIP is always found in bi-partitions because integrated information becomes larger when a system is partitioned into more subsystems.

Whether the integrated information should be normalized and how the integrated information should be normalized are still open questions. In our study, the normalization used in IIT 2.0 is not appropriate, because the entropy can be negative for continuous random variables. Additionally, regardless of whether random variables are continuous or discrete, normalization significantly affects the submodularity of the measures of integrated information. For example, if we use normalization proposed in IIT 2.0, even the submodular measure of integrated information, Φ_{MI}, no longer satisfies submodularity. Thus, Queyranne's algorithm may not work well if Φ is normalized.

Although we resolved one of the major computational difficulties in IIT, an additional issue still remains. Searching for the MIP is an intermediate step in identifying the informational core, called the "complex". The complex is the subnetwork in which integrated information is maximized, and is hypothesized to be the locus of consciousness in IIT. Identifying the complex is also represented as a discrete optimization problem which requires exponentially large computational costs. Queyranne's algorithm cannot be applied to the search for the complex because we cannot formulate it as a submodular optimization. We expect that REMCMC would be efficient in searching for the complex and will investigate its performance in a future study.

An important limitation of this study is that we only showed the nearly perfect performance of Queyranne's algorithm in limited simulated data and real neural data. In general, we cannot

tell whether Queyranne's algorithm works well for other data beforehand. For real data analysis, we recommend that the procedure below should be applied. First, as we did in Section 6.2, accuracy should be checked by comparing it with the exhaustive search in small randomly selected subsets. Next, if it works well, the performance should be checked by comparing it with REMCMC in relatively large subsets, as we did in Section 6.3. If Queyranne's algorithm works better than or equally as well as REMCMC, it is reasonable to use Queyranne's algorithm for the analysis. By applying this procedure, we expect that Queyranne's algorithm could be utilized to efficiently find the MIP in a wide range of time series data.

Author Contributions: Jun Kitazono and Masafumi Oizumi conceived and designed the experiments; Jun Kitazono performed the experiments; Jun Kitazono and Masafumi Oizumi analyzed the data; and Jun Kitazono, Ryota Kanai and Masafumi Oizumi wrote the paper.

Abbreviations

The following abbreviations are used in this manuscript:

IIT	integrated information theory
MIP	minimum information partition
MCMC	Markov chain Monte Carlo
REMCMC	replica exchange Markov chain Monte Carlo
EEG	electroencephalography
ECoG	electrocorticography
AR	autoregressive
CR	correct rate
RA	rank
ER	error ratio
CORR	correlation
MCS	Monte Carlo step

Appendix A. Analytical Formula of Φ for Gaussian Variables

We describe the analytical formula of three measures of integrated information, multi information (Φ_{MI}), stochastic interaction (Φ_{SI}) and geometric integrated information (Φ_{G}), when the probability distribution is Gaussian. For more details about the theoretical background, see [12,15,18,19].

First, let us introduce the notation. We consider a stochastic dynamical system consisting of N elements. We represent the past and present states of the system as $X = (X_1, \ldots, X_N)$ and $X' = (X'_1, \ldots, X'_N)$, respectively, and define a joint vector

$$\tilde{X} = (X, X'). \tag{A1}$$

We assume that the joint probability distribution $p(X, X')$ is Gaussian:

$$p(x, x') = \exp\left\{ -\frac{1}{2}\left(\tilde{x}^T \Sigma(\tilde{X}) \tilde{x} - \psi \right) \right\}, \tag{A2}$$

where ψ is the normalizing factor and $\Sigma(\tilde{X})$ is the covariance matrix of $\tilde{X}$. Note that we can assume the mean of the Gaussian distribution is zero without loss of generality because the mean value does not affect the values of integrated information. This covariance matrix $\Sigma(\tilde{X})$ is given by

$$\Sigma(\tilde{X}) = \begin{pmatrix} \Sigma(X) & \Sigma(X, X') \\ \Sigma(X, X')^T & \Sigma(X') \end{pmatrix}, \tag{A3}$$

where $\Sigma(X)$ and $\Sigma(X')$ are the equal time covariance at past and present, respectively, and $\Sigma(X, X')$ is the cross covariance between X and X'. Below we will show the analytical expression of Φ_{MI}, Φ_{SI} and Φ_{G}.

Appendix A.1. Multi Information

Let us consider the following partitioned probability distribution q,

$$q\left(X, X'\right) = \prod_i q\left(M_i, M_i'\right), \tag{A4}$$

where M_i and M_i' are the past and present states of i-th subsystem. Then multi information is defined as

$$\Phi_{\text{MI}} = \sum_i H(M_i, M_i') - H(X, X'). \tag{A5}$$

When the distribution is Gaussian, Equation (A5) is transformed to

$$\Phi_{\text{MI}} = \sum_i \log |\Sigma(\tilde{M}_i)| - \log |\Sigma(\tilde{X})|, \tag{A6}$$

where $\tilde{M}_i = (M_i, M_i')$ and $\Sigma(\tilde{M}_i)$ is the covariance of $\tilde{M}_i$.

Appendix A.2. Stochastic Interaction

We consider the following partitioned probability distribution q,

$$q\left(X'|X\right) = \prod_i q\left(M_i'|M_i\right). \tag{A7}$$

Then, stochastic interaction [12,15,18,19] is defined as

$$\Phi_{\text{SI}} = \sum_i H(M_i'|M_i) - H(X'|X). \tag{A8}$$

When the distribution is Gaussian, Equation (A8) is transformed to

$$\Phi_{\text{SI}} = \sum_i \log |\Sigma(M_i'|M_i)| - \log |\Sigma(X'|X)|, \tag{A9}$$

where $\Sigma(M_i'|M_i)$ and $\Sigma(X'|X)$ are covariance matrices of conditional distributions. These matrices are represented as

$$\begin{aligned} \Sigma(M_i'|M_i) &= \Sigma(M_i') - \Sigma(M_i, M_i')^T \Sigma(M_i)^{-1} \Sigma(M_i, M_i'), \\ \Sigma(X'|X) &= \Sigma(X') - \Sigma(X, X')^T \Sigma(X)^{-1} \Sigma(X, X'), \end{aligned} \tag{A10}$$

where $\Sigma(M_i)$ and $\Sigma(M_i')$ are the equal time covariance of subsystem i at past and present, respectively, and $\Sigma(M_i, M_i')$ is the cross covariance between M_i and M_i'.

Appendix A.3. Geometric Integrated Information

To calculate the geometric integrated information [12], we first transform Equation (A2). Equation (A2) is equivalently represented as an autoregressive model:

$$X' = AX + E, \tag{A11}$$

where A is the connectivity matrix and E is Gaussian random variables, which are uncorrelated over time. By using this autoregressive model, the joint distribution $p(X, X')$ is expressed as

$$p(x, x') = \exp\left\{ -\frac{1}{2}\left(x^T \Sigma(X)x + (x' - Ax)^T \Sigma(E)^{-1}(x' - Ax) - \psi \right) \right\}, \tag{A12}$$

and the covariance matrices as

$$\begin{aligned} \Sigma(X, X') &= \Sigma(X)A^T, \\ \Sigma(X') &= \Sigma(E) + A\Sigma(X)A^T, \end{aligned} \tag{A13}$$

where $\Sigma(E)$ is the covariance of E. Similarly, the joint probability distribution in a partitioned model is given by

$$\begin{aligned} q(x, x') &= \exp\left\{ -\frac{1}{2}\left(\tilde{x}^T \Sigma(\tilde{X})_\mathrm{p} \tilde{x} - \psi \right) \right\} \\ &= \exp\left\{ -\frac{1}{2}\left(x^T \Sigma(X)_\mathrm{p} x + (x' - A_\mathrm{p}x)^T \Sigma(E)_\mathrm{p}^{-1}(x' - A_\mathrm{p}x) - \psi \right) \right\}, \end{aligned} \tag{A14}$$

where $\Sigma(X)_\mathrm{p}$ and $\Sigma(E)_\mathrm{p}$ are the covariance matrices of X and E in the partitioned model, respectively, and A_p is the connectivity matrix in the partitioned model.

The geometric integrated information is defined as

$$\Phi_\mathrm{G} = \min_q D_\mathrm{KL}\left(p(X, X') \,||\, q(X, X') \right), \tag{A15}$$

$$D_\mathrm{KL}\left(p(X, X') \,||\, q(X, X') \right) = \frac{1}{2}\left(\log \frac{|\Sigma(\tilde{X})_\mathrm{p}|}{|\Sigma(\tilde{X})|} + \mathrm{Tr}(\Sigma(\tilde{X})\Sigma(\tilde{X}_\mathrm{p})^{-1}) - 2N \right), \tag{A16}$$

such that

$$q\left(M'_i | X \right) = q\left(M'_i | M_i \right), \forall i. \tag{A17}$$

This constraint (Equation (A17)) corresponds to setting the between-subsystem blocks of A_p to 0:

$$(A_\mathrm{p})_{ij} = 0 \; (i \neq j). \tag{A18}$$

By transforming stationary point conditions, $\partial D_\mathrm{KL}/\partial \Sigma(\tilde{X})_\mathrm{p}^{-1} = 0$, $\partial D_\mathrm{KL}/\partial (A_\mathrm{p})_{ii} = 0$, and $\partial D_\mathrm{KL}/\partial \Sigma(E)_\mathrm{p}^{-1} = 0$, we get

$$\Sigma(X)_\mathrm{p} = \Sigma(X), \tag{A19}$$

$$(\Sigma(X)(A - A_\mathrm{p})\Sigma(E)_\mathrm{p}^{-1})_{ii} = 0, \tag{A20}$$

$$\Sigma(E)_\mathrm{p} = \Sigma(E) + (A - A_\mathrm{p})\Sigma(X)(A - A_\mathrm{p})^T. \tag{A21}$$

By substituting Equations (A19) and (A21) into Equation (A15), Φ_G is simplified as

$$\Phi_\mathrm{G} = \frac{1}{2}\log\frac{|\Sigma(E)_\mathrm{p}|}{|\Sigma(E)|}. \tag{A22}$$

To obtain the value of Equation (A22), we need to find the value of $\Sigma(E)_\mathrm{p}$. The computation of $\Sigma(E)_\mathrm{p}$ requires solving Equations (A20) and (A21) for $\Sigma(E)_\mathrm{p}$ and A_p simultaneously. However, it is difficult to express Equations (A20) and (A21) as closed-form expressions. Therefore, we need to solve the multi-dimensional equations (Equations (A20) and (A21)) using an iterative method. This iterative

process increased the complexity of the search using Queyranne's algorithm up to roughly $O(N^{4.776})$ (see Section 6.1). The MALAB codes for this computation of Φ_G are available at [37].

Appendix B. Details of Replica Exchange Markov Chain Monte Carlo Method

The Replica Exchange Markov Chain Monte Carlo (REMCMC) method was originally proposed to investigate physical systems [21–23], and was then rapidly utilized in other applications, including combinatorial optimization problems [32–34,38,39]. For a more detailed history of REMCMC, see, for example, [24].

We first briefly explain how the MIP search problem is dealt with by the Metropolis method. Then, as an improvement of Metropolis method, we introduce REMCMC to more effectively search for the global minimum while avoiding being trapped around at a local minimum. Next, we describe the convergence criterion of MCMC sampling. Finally, we present the parameter settings in our experiments.

Appendix B.1. Metropolis Method

We consider the way to sample subsets from the probability distribution in Equation (14). An initial subset $S^{(0)}$ is randomly selected, and then a sample sequence is drawn as follows.

- **Propose a candidate of the next sample** An element e is randomly selected and if it is in the current subset $S^{(t)}$, the candidate S_c is $S^{(t)} \setminus \{e\}$. If not, the candidate is $S^{(t)} \cup \{e\}$.
- **Determine whether to accept the candidate or not** The candidate S_c is accepted ($S^{(t+1)} = S_c$) or not accepted ($S^{(t+1)} = S^{(t)}$) according to the following probability $a(S^{(t)} \to S_c)$:

$$a(S^{(t)} \to S_c) = \min(1, r),$$
$$r = \frac{p(S_c; \beta)}{p(S^{(t)}; \beta)} = \exp\left[\beta \left\{\Phi(S^{(t)}) - \Phi(S_c)\right\}\right]. \tag{A23}$$

This probability means that if the integrated information decreases by stepping from $S^{(t)}$ to S_c, the candidate S_c is always accepted, and otherwise it is accepted with the probability r.

By iterating these two steps with sufficient time, the sample distribution converges to the probability distribution given in Equation (14). N steps of the sampling is referred to as one Monte Carlo step (MCS), where N is the number of elements. In one MCS, each element is attempted to be added or removed once on average.

Depending on the value of β, the behavior of the sample sequence changes. If β is small, the probability distribution given by Equation (14) is close to a uniform distribution and subsets are sampled nearly independently of the value of $\Phi(S)$. If β is large, the candidate is more likely to be accepted when the integrated information decreases. The sample sequence easily falls to a local minimum and cannot explore many subsets. Thus, smaller and larger β have an advantage and a disadvantage: Smaller β is better for exploring around many subsets while larger β is better for finding a (local) minimum. In the Metropolis method, we need to set β to an appropriate value taking account of this trade-off, but it is generally difficult.

Appendix B.2. Replica Exchange Markov Chain Monte Carlo

To overcome the difficulty in setting inverse temperature β, REMCMC samples from distributions at multiple values of β in parallel and the sampled sequences are exchanged between nearby values of β. By this exchange, the sampled sequences at high inverse temperatures can escape from local minima and can explore many subsets.

We consider M-probabilities at different inverse temperatures $\beta_1 > \beta_2 > \cdots > \beta_M$ and introduce the following joint probability:

$$p(S_1, \ldots, S_M; \beta_1, \ldots, \beta_M) = \prod_{m=1}^{M} p(S_m; \beta_m). \tag{A24}$$

Then, the simulation process of the REMCMC consists of the following two steps:

- **Sampling from each distribution:** Samples are drawn from each distribution $p(S_m; \beta_m)$ separately by using the Metropolis method as described in the previous subsection.
- **Exchange between neighboring inverse temperatures:** After a given number of samples are drawn, subsets at neighboring inverse temperatures are swapped, according to the following probability $p(S_m \leftrightarrow S_{m+1})$:

$$
\begin{aligned}
p(S_m \leftrightarrow S_{m+1}) &= \min(1, r'), \\
r' &= \frac{p(S_{m+1}; \beta_m) p(S_m; \beta_{m+1})}{p(S_m; \beta_m) p(S_{m+1}; \beta_{m+1})} \\
&= \exp\left[(\beta_{m+1} - \beta_m) \left\{ \Phi(S_{m+1}) - \Phi(S_m) \right\} \right].
\end{aligned}
\tag{A25}
$$

This probability indicates that if the integrated information at a higher inverse temperature is larger than that at a lower inverse temperature, subsets are always swapped; otherwise, they are swapped with the probability r'.

By iterating these two steps for sufficient time, the sample distribution converges to the joint distribution in Equation (A24).

To maximize the efficiency of the REMCMC, it is important to appropriately set the multiple inverse temperatures. If the neighboring temperatures are far apart, the acceptance ratio of exchange (Equation (A25)) becomes too small. The REMCMC is then reduced to just separately simulating distributions at different temperatures without any exchange. In a previous study [40], it was recommended to keep the average ratio higher than 0.2 for every temperature pair. At the same time, the highest/lowest inverse temperatures should be high/low enough so that sample sequence at the highest inverse temperature can reach the tips of (local) minima and that at the lowest one can search around many subsets. To satisfy these constraints, a sufficient number M of inverse temperatures are accommodated and the inverse temperatures are optimized to equalize the average of the acceptance ratio of exchanges at all temperature pairs [40–43]. Details of temperature setting are described below.

Appendix B.2.1. Initial Setting

Inverse temperatures $\beta_m (m = 1, \ldots, M)$ are initially set as follows. First, a subset is randomly selected for each m. Then, a randomly chosen element is added to or eliminated from each subset, and the absolute value of the change $\Delta\Phi_m$ in the amount of integrated information is taken. By using these absolute values, the highest and lowest inverse temperatures are determined by a bisection method so that the respective averages of the acceptance ratio $\exp(-\beta\Delta\Phi_1)$ and $\exp(-\beta\Delta\Phi_M)$ match the predefined values. The intermediate inverse temperatures are set to be a geometric progression:
$\beta_m = \beta_1 \left(\frac{\beta_M}{\beta_1} \right)^{\frac{m-1}{M-1}}$.

Appendix B.2.2. Updating

The difference in the amount of integrated information between the candidate subset $\Phi(S_c)$ and the current subset $\Phi(S^{(t)})$ is stored when the difference is positive ($\Phi(S_c) - \Phi(S^{(t)}) \geq 0$). Then, by using the stored values at all the inverse temperatures, the highest and lowest inverse temperatures are determined by a bisection method so that the average of the acceptance ratio $\exp\left[\beta \left\{ \Phi(S^{(t)}) - \Phi(S_c) \right\} \right]$ matches the predefined value, as in the initial setting. The intermediate inverse temperatures are set

to approximately equalize the expected values of acceptance ratio of the exchange at all temperature pairs [40–43]. The expected value is represented as a sum of two probabilities:

$$\mathbb{E}\left[p(S_m \leftrightarrow S_{m+1})\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left\{ p(\Phi_m \geq \Phi_{m+1}) \right.$$
$$\left. + p(\Phi_m < \Phi_{m+1})e^{(\beta_m - \beta_{m+1})(\Phi_m - \Phi_{m+1})} \right\} d\Phi_m d\Phi_{m+1}. \tag{A26}$$

In [43], this expected value is approximated as

$$\mathbb{E}\left[p(S_m \leftrightarrow S_{m+1})\right] \approx \frac{1}{2}\text{erfc}\left(\frac{\mu(T_{m+1}) - \mu(T_m)}{\sqrt{2\left\{\sigma^2(T_{m+1}) + \sigma^2(T_m)\right\}}}\right)$$
$$+ \left\{1 - \frac{1}{2}\text{erfc}\left(\frac{\mu(T_{m+1}) - \mu(T_m)}{\sqrt{2\left\{\sigma^2(T_{m+1}) + \sigma^2(T_m)\right\}}}\right)\right\} e^{(\beta_m - \beta_{m+1})(\mu(T_m) - \mu(T_{m+1}))}, \tag{A27}$$

where $\mu(T)$ and $\sigma^2(T)$ are the mean and variance of Φ, represented as functions of temperature T. In [43], these functions are given by interpolating the sample mean and variance. In this study, these functions are estimated using regression, because the sample mean and variance are highly variable. The mean and variance at each temperature are computed at every update, and these means and variances are regressed on temperature using a continuous piecewise linear function, the T-axis of anchor points of which are current temperatures. The anchor points are interpolated using piecewise cubic Hermite interpolating polynomials. Then, to roughly equalize the expected values of the acceptance ratio of the exchange at all temperature pairs, we minimize the following cost function by varying temperatures [43]:

$$\text{Cost} = \sum_{m=1}^{M-1} \mathbb{E}\left[p(S_m \leftrightarrow S_{m+1})\right]^{-4}. \tag{A28}$$

The minimization is performed by a line-search method.

Appendix B.3. Convergence Criterion

One of the most commonly used MCMC convergence criteria is potential scale reduction factor (PSRF), which was proposed by Gelman and Rubin (1992) [44], and modified by Brooks and Gelman (1998) [45]. In this criterion, multiple MCMC sequences are run. If all of them converge, statistics of the sequences must be about the same. This is assessed by comparing between-sequence variance and within-sequence variance of a random variable and calculating the PSRF, $\hat{R}_c$. Large $\hat{R}_c$ suggests that some of the sequences do not converge yet. If $\hat{R}_c$ is close to 1, we can diagnose them as converged. In this study, we cut the sequence at each inverse temperature into the former and the latter halves, and applied the criterion to these two half sequences. If $\hat{R}_c$ of all the temperatures were below a predefined threshold, we regarded the sequences as converged.

Appendix B.4. Parameter Settings

The number of inverse temperatures M was fixed at 6 throughout out the experiments. The highest/lowest inverse temperatures were set so that the averages of acceptance ratio become 0.01 and 0.5, respectively. The exchange process was done every 5 MCSs. The update of inverse temperatures was performed every 5 MCSs for the 200 initial MCSs. The threshold of $\hat{R}_c$ was set to 1.01. When computing $\hat{R}_c$, we discarded the first 200 MCSs as a burn-in period and started to computing it after 300 MCSs.

Appendix C. Values of Φ

We show some examples of the distributions of the values of Φ in the experiments in Section 6.2. Figure A1a,b are the box plots of Φ_{SI} and Φ_{G} for the block-structured models at $\sigma = 0.01$, respectively. We can see that in Figure A1a, Φ_{SI} computed at the partition found by Queyranne's algorithm perfectly matched with that at the MIPs. In Figure A1b, Φ_{G} computed at the partition found by Queyeranne's algorithm did not match that at the MIPs in 3 trials (the trial numbers 11, 54 and 83) but the deviations were very small.
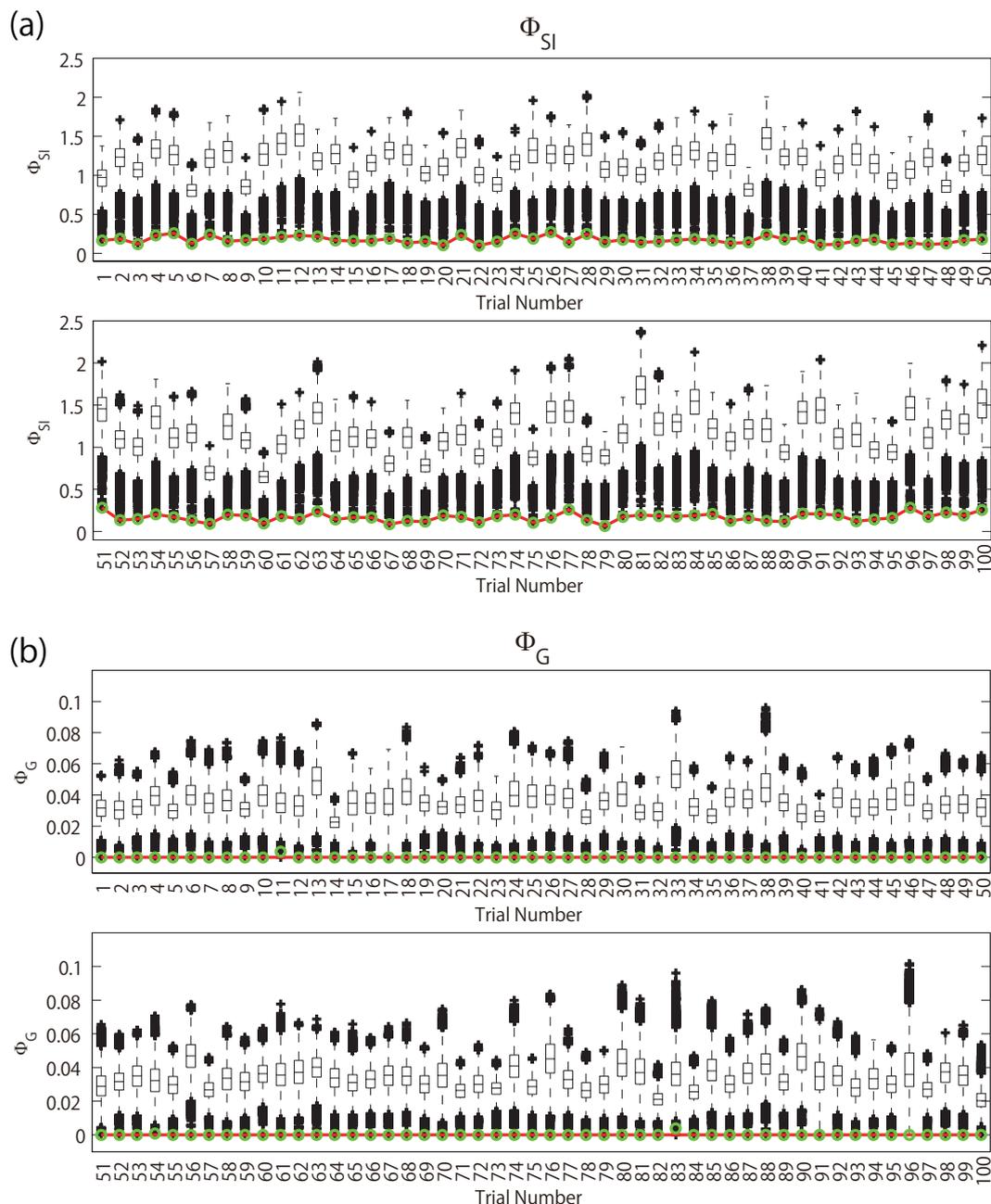

Figure A1. The values of Φ for the block-structured models at $\sigma = 0.01$. The box plots represent the distribution of Φ at all the partitions. The red solid line indicates Φ at the MIP. The green circles indicate Φ at the partitions found by Queyranne's algorithm. (**a**) Φ_{SI}, (**b**) Φ_{G}.

References

1. Tononi, G.; Sporns, O.; Edelman, G.M. A measure for brain complexity: Relating functional segregation and integration in the nervous system. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 5033–5037.
2. Tononi, G. An information integration theory of consciousness. *BMC Neurosci.* **2004**, *5*, 42, doi:10.1186/1471-2202-5-42.
3. Tononi, G. Consciousness as integrated information: A provisional manifesto. *Biol. Bull.* **2008**, *215*, 216–242.
4. Oizumi, M.; Albantakis, L.; Tononi, G. From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Comput. Biol.* **2014**, *10*, e1003588, doi:10.1371/journal.pcbi.1003588.
5. Massimini, M.; Ferrarelli, F.; Huber, R.; Esser, S.K.; Singh, H.; Tononi, G. Breakdown of cortical effective connectivity during sleep. *Science* **2005**, *309*, 2228–2232.
6. Casali, A.G.; Gosseries, O.; Rosanova, M.; Boly, M.; Sarasso, S.; Casali, K.R.; Casarotto, S.; Bruno, M.A.; Laureys, S.; Tononi, G.; et al. A theoretically based index of consciousness independent of sensory processing and behavior. *Sci. Transl. Med.* **2013**, *5*, 198ra105, doi:10.1126/scitranslmed.3006294.
7. Lee, U.; Mashour, G.A.; Kim, S.; Noh, G.J.; Choi, B.M. Propofol induction reduces the capacity for neural information integration: Implications for the mechanism of consciousness and general anesthesia. *Conscious. Cogn.* **2009**, *18*, 56–64.
8. Chang, J.Y.; Pigorini, A.; Massimini, M.; Tononi, G.; Nobili, L.; Van Veen, B.D. Multivariate autoregressive models with exogenous inputs for intracerebral responses to direct electrical stimulation of the human brain. *Front. Hum. Neurosci.* **2012**, *6*, 317, doi:10.3389/fnhum.2012.00317.
9. Boly, M.; Sasai, S.; Gosseries, O.; Oizumi, M.; Casali, A.; Massimini, M.; Tononi, G. Stimulus set meaningfulness and neurophysiological differentiation: A functional magnetic resonance imaging study. *PLoS ONE* **2015**, *10*, e0125337, doi:10.1371/journal.pone.0125337.
10. Haun, A.M.; Oizumi, M.; Kovach, C.K.; Kawasaki, H.; Oya, H.; Howard, M.A.; Adolphs, R.; Tsuchiya, N. Conscious Perception as Integrated Information Patterns in Human Electrocorticography. *eNeuro* **2017**, *4*, 1–18, doi:10.1523/ENEURO.0085-17.2017.
11. Balduzzi, D.; Tononi, G. Integrated information in discrete dynamical systems: Motivation and theoretical framework. *PLoS Comput. Biol.* **2008**, *4*, e1000091, doi:10.1371/journal.pcbi.1000091.
12. Oizumi, M.; Tsuchiya, N.; Amari, S.i. Unified framework for information integration based on information geometry. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 14817–14822.
13. Hidaka, S.; Oizumi, M. Fast and exact search for the partition with minimal information loss. *arXiv* **2017**, arXiv:1708.01444.
14. Queyranne, M. Minimizing symmetric submodular functions. *Math. Program.* **1998**, *82*, 3–12.
15. Barrett, A.B.; Barnett, L.; Seth, A.K. Multivariate Granger causality and generalized variance. *Phys. Rev. E* **2010**, *81*, 041907, doi:10.1103/PhysRevE.81.041907.
16. Oizumi, M.; Amari, S.; Yanagawa, T.; Fujii, N.; Tsuchiya, N. Measuring integrated information from the decoding perspective. *PLoS Comput. Biol.* **2016**, *12*, e1004654, doi:10.1371/journal.pcbi.1004654.
17. Tegmark, M. Improved measures of integrated information. *PLoS Comput. Biol.* **2016**, *12*, e1005123, doi:10.1371/journal.pcbi.1005123.
18. Ay, N. Information geometry on complexity and stochastic interaction. *MIP MIS Preprint 95* **2001**. Available online: http://www.mis.mpg.de/publications/preprints/2001/prepr2001-95.html (accessed on 6 March 2018).
19. Ay, N. Information geometry on complexity and stochastic interaction. *Entropy* **2015**, *17*, 2432–2458.
20. Amari, S.; Tsuchiya, N.; Oizumi, M. Geometry of information integration. *arXiv* **2017**, arXiv:1709.02050.
21. Swendsen, R.H.; Wang, J.S. Replica Monte Carlo simulation of spin-glasses. *Phys. Rev. Lett.* **1986**, *57*, 2607–2609.
22. Geyer, C.J. Markov chain Monte Carlo maximum likelihood. In Proceedings of the 23rd Symposium on the Interface, Seattle, WA, USA, 21–24 April 1991; Interface Foundation of North America: Fairfax Station, VA, USA, 1991; pp. 156–163.
23. Hukushima, K.; Nemoto, K. Exchange Monte Carlo method and application to spin glass simulations. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604–1608.
24. Earl, D.J.; Deem, M.W. Parallel tempering: Theory, applications, and new perspectives. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3910–3916.
25. Burnham, K.P.; Anderson, D.R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*; Springer: New York, NY, USA, 2003.

26. Watanabe, S. Information theoretical analysis of multivariate correlation. *IBM J. Res. Dev.* **1960**, *4*, 66–82.

27. Studený, M.; Vejnarová, J. *The Multiinformation Function as a Tool For Measuring Stochastic Dependence*; MIT Press: Cambridge, MA, USA, 1999.

28. Pearl, J. *Causality*; Cambridge University Press: Cambridge, UK, 2009.

29. Iwata, S. Submodular function minimization. *Math. Program.* **2008**, *112*, 45–64.

30. Wishart, J. The generalised product moment distribution in samples from a normal multivariate population. *Biometrika* **1928**, *20A*, 32–52.

31. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006.

32. Pinn, K.; Wieczerkowski, C. Number of magic squares from parallel tempering Monte Carlo. *Int. J. Mod. Phys. C* **1998**, *9*, 541–546.

33. Hukushima, K. Extended ensemble Monte Carlo approach to hardly relaxing problems. *Computer Phys. Commun.* **2002**, *147*, 77–82.

34. Nagata, K.; Kitazono, J.; Nakajima, S.; Eifuku, S.; Tamura, R.; Okada, M. An Exhaustive Search and Stability of Sparse Estimation for Feature Selection Problem. *IPSJ Online Trans.* **2015**, *8*, 25–32.

35. Nagasaka, Y.; Shimoda, K.; Fujii, N. Multidimensional recording (MDR) and data sharing: an ecological open research and educational platform for neuroscience. *PLoS ONE* **2011**, *6*, e22561, doi:10.1371/journal.pone.0022561.

36. Toker, D.; Sommer, F. Information Integration in Large Brain Networks. *arXiv* **2017**, arXiv:1708.02967.

37. Kitazono, J.; Oizumi, M. *phi_toolbox.zip*, version 6; Figshare, 6 September 2017. Available online: https://figshare.com/articles/phi_toolbox_zip/3203326/6 (accessed on 6 March 2018).

38. Barthel, W.; Hartmann, A.K. Clustering analysis of the ground-state structure of the vertex-cover problem. *Phys. Rev. E* **2004**, *70*, 066120, doi:10.1103/PhysRevE.70.066120.

39. Wang, C.; Hyman, J.D.; Percus, A.; Caflisch, R. Parallel tempering for the traveling salesman problem. *Int. J. Mod. Phys. C* **2009**, *20*, 539–556, doi:10.1142/S0129183109013893.

40. Rathore, N.; Chopra, M.; de Pablo, J.J. Optimal allocation of replicas in parallel tempering simulations. *J. Chem. Phys.* **2005**, *122*, 024111, doi:10.1063/1.1831273.

41. Sugita, Y.; Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **1999**, *314*, 141–151.

42. Kofke, D.A. On the acceptance probability of replica-exchange Monte Carlo trials. *J. Chem. Phys.* **2002**, *117*, 6911–6914; Erratum in **2004**, *120*, 10852, doi:10.1063/1.1738103.

43. Lee, M.S.; Olson, M.A. Comparison of two adaptive temperature-based replica exchange methods applied to a sharp phase transition of protein unfolding-folding. *J. Chem. Phys.* **2011**, *134*, 244111, doi:10.1063/1.3603964.

44. Gelman, A.; Rubin, D.B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **1992**, *7*, 457–472.

45. Brooks, S.P.; Gelman, A. General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* **1998**, *7*, 434–455.