# Measuring Information Processing in Neural Data: The Application of Transfer Entropy in Neuroscience

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich 12 Informatik/Mathematik
der Johann Wolfgang Goethe-Universität,
in Frankfurt am Main

von
**Patricia Wollstadt**
aus Offenbach am Main

Frankfurt am Main, Januar 2017
D30

Vom Fachbereich 12 der Johann-Wolfgang Goethe-Universität als Dissertation angenommen.

Dekan: Prof. Dr. Uwe Brinkschulte

Gutachter:
Prof. Dr. Matthias Kaschube, Frankfurt Institute for Advanced Studies, Johann Wolfgang Goethe-Universität, Frankfurt am Main
Prof. Dr. Michael Wibral, Johann Wolfgang Goethe-Universität, Frankfurt am Main

Datum der Disputation: 25. Januar 2018

# Abstract

It is a common notion in neuroscience research that the brain and neural systems in general "perform computations" to generate their complex, everyday behavior [1]. Understanding these computations is thus an important step in understanding neural systems as a whole [1–4]. It has been proposed that one way to analyze these computations is by quantifying basic information processing operations necessary for computation, namely the transfer, storage, and modification of information [5–8]. A framework for the analysis of these operations has been emerging [9], using measures from information theory [10] to analyze computation in arbitrary information processing systems (e.g. [11]). Of these measures transfer entropy (TE) [12], a measure of information transfer, is the most widely used in neuroscience today (e.g. [13–22]). Yet, despite this popularity, open theoretical and practical problems in the application of TE remain (e.g. [13, 23]). The present work addresses some of the most prominent of these methodological problems in three studies.

The first study presents an efficient implementation for the estimation of TE from non-stationary data. The statistical properties of non-stationary data are not invariant over time such that TE can not be easily estimated from these observations. Instead, necessary observations can be collected over an ensemble of data, i.e., observations of physical or temporal replications of the same process [24]. The latter approach is computationally more demanding than the estimation from observations over time. The present study demonstrates how to handles this increased computational demand by presenting a highly-parallel implementation of the estimator using graphics processing units.

The second study addresses the problem of estimating bivariate TE from multivariate data. Neuroscience research often investigates interactions between more than two (sub-)systems. It is common to analyze these interactions by iteratively estimating TE between pairs of variables, because a fully multivariate approach to TE-estimation is computationally intractable [25–27]. Yet, the estimation of bivariate TE from multivariate data may yield spurious, false-positive results [25, 28, 29]. The present study proposes that such spurious links can be identified by characteristic coupling-motifs and the timings of their information transfer delays in networks of bivariate TE-estimates. The study presents a graph-algorithm that detects these coupling

motifs and marks potentially spurious links. The algorithm thus partially corrects for spurious results due to multivariate effects and yields a more conservative approximation of the true network of multivariate information transfer.

The third study investigates the TE between pre-frontal and primary visual cortical areas of two ferrets under different levels of anesthesia. Additionally, the study investigates local information processing in source and target of the TE by estimating information storage [30] and signal entropy. Results of this study indicate an alternative explanation for the commonly observed reduction in TE under anesthesia [31–35], which is often explained by changes in the underlying coupling between areas. Instead, the present study proposes that reduced TE may be due to a reduction in information generation measured by signal entropy in the source of TE. The study thus demonstrates how interpreting changes in TE as evidence for changes in causal coupling may lead to erroneous conclusions. The study further discusses current bast-practice in the estimation of TE, namely the use of state-of-the-art estimators over approximative methods and the use of optimization procedures for estimation parameters over the use of ad-hoc choices. It is demonstrated how not following this best-practice may lead to over- or under-estimation of TE or failure to detect TE altogether.

In summary, the present work proposes an implementation for the efficient estimation of TE from non-stationary data, it presents a correction for spurious effects in bivariate TE-estimation from multivariate data, and it presents current best-practice in the estimation and interpretation of TE. Taken together, the work presents solutions to some of the most pressing problems of the estimation of TE in neuroscience, improving the robust estimation of TE as a measure of information transfer in neural systems.

# Contents

# List of Publications

1. P. Wollstadt, M. Martínez-Zarzuela, R. Vicente, F. J. Díaz-Pernas, and M. Wibral (2014). “Efficient transfer entropy analysis of non-stationary neural time series”. *PLoS ONE*, 9(7), e102833.

2. P. Wollstadt, U. Meyer, and M. Wibral (2015). “A graph algorithmic approach to separate direct from indirect neural interactions.” *PLoS ONE*, 10(10), e0140530.

3. P. Wollstadt, K. K. Sellers, L. Rudelt, V. Priesemann, A. Hutt, F. Fröhlich, and M. Wibral (2016). “The relation of local entropy and information transfer suggests an origin of isoflurane anesthesia effects in local information processing.” Submitted to *PLoS Computational Biology* (preprint available from arXiv: arXiv:1608.08387).

All articles were published under the Creative Commons Attribution (CC BY) license.

# 1 Introduction

## 1.1 Background

In neuroscience it is often stated that the brain "performs computations" or "processes information" to generate its complex, everyday behavior. The assumption underlying these statements is that the brain—and biological systems in general—represent their environment in terms of physical variables, e.g., a cell membrane potential or pheromone concentration; on the basis of these variables, "every physical system performs a computation by realizing a solution to the dynamic equations that govern its physical behaviour." [1]. Understanding these computations is thus an important step in understanding the behavior of the system as a whole—yet, analyses explicitly targeting these computations are lacking or only just emerging [1–4]. As an example, one may consider the model organism *C. elegans*: its neural architecture [36, 37] and neural dynamics (e.g., graded membrane potentials [38, 39]), i.e., the physical variables representing the organism's environment, are well described. However, despite this detailed knowledge, it remains impossible to predict the organism's behavior or learning [2]. In other words, despite detailed knowledge about physical variables and observable behavior, it is not possible to explain how behavior is generated on the basis of these variables. The example shows further that neither a detailed description of physical representations, nor a detailed description of global behavior seem to be sufficient to explain how the first gives rise to the latter. Here, the analysis of an "intermediate level [...] of neural computation" is missing to investigate how the algorithmic manipulation of physical representations give rise to the behavior of the organism as a whole [1, 2].

Yet, while the notion of computation is well defined and can be formally analyzed in traditional computing systems like digital computers, it is not clear how to find similar definitions and analysis tools for biological systems like neural systems [6, 7]. A first step to defining computations and their analysis in biological system was taken by David Marr [40] (Table 1.1): he proposed a theoretical framework for the analysis of arbitrary information processing systems, which assumes that three levels of analysis are needed to understand such a system in its entirety[1]: (1) the *functional level*, which describes the task a system solves; (2) the *algorithmic level*,

[1] A similar idea has been put forward by Pylyshyn [41]; I will restrict my explanation to the description of Marr's theory. The theory introduced by Pylyshyn [41] are conceptually similar (e.g., [42]).

Tab. 1.1 Marr's levels of analysis for information processing systems [40].

| I – Functional level | II – Algorithmic level[a] | III – Implementational level[a] |
|---|---|---|
| What the system does and why. | How the system does it. | The system's physical characteristics. |
| What is the goal of the computation, i.e., the task to be solved? This task imposes constraints on what the system does: "the resulting operation is defined uniquely by the constraints it has to satisfy". | Representation of input and output, and an algorithm that transforms the input such that it realizes the task defined on the first level, both are usually related. | Physical characteristics of the system carrying out the task. |
| *The cash register performs addition, because that's the best mathematical implementation of what is expected of it to do.* | *Different number systems require different algorithms for addition, but for a fixed representation, we can often choose a variety of different algorithms, depending on efficiency or the hardware available.* | *Addition may be implemented by a mechanical calculator, an abacus, or a digital calculator.* |

[a] Marr does not use a name for this level.

which describes how the system represents information and how it operates on these representations to generate the desired output; and (3) the *implementational level*, which describes the biophysical implementation or realization of the system.

Marr thus clearly distinguishes a "level of computations" (i.e., the algorithmic level) from the level of global behavior and function, and the level of physical phenomena[2]. He further states that these levels pose only little constraints on each other, i.e., knowledge about phenomena on one level hardly increase our knowledge about phenomena on another level. Adopting this separation into levels of analysis, it becomes clear that neuroscience is often concerned with investigating the implementational and functional level (c.f. the example of *C. elegans*) [2], while neglecting the algorithmic level, which corresponds to the level of neural computations—yet, following Marr, a transfer of knowledge from one level of analysis onto another may impossible. Hence, an independent analysis of neural computations is needed for a complete understanding of how physical phenomena enable function in neural systems.

[2] Note that Marr proposed levels of *analysis* not levels of *system organization* [43]; the term "level" may be misleading here and it should be kept in mind that "the deeper contribution made by Marr and Poggio was the idea that it is valid, fruitful, and even necessary to analyze cognition by forming abstraction barriers" [44].

An attempt to describe computations in biological systems was made by Mitchell [6]: Biological systems typically consist of collections of agents (e.g., swarms or neurons in a brain) or sub-systems (e.g., plants or cortical areas), which process information in a distributed fashion to generate collective behavior; examples are insect swarms like ants or wasps (e.g., [45]), flocks of birds (e.g., [46]), plants [47], fish (e.g., [48]), gene regulatory networks [49, 50], or neural systems (e.g., [51, 52]). To characterize information processing in these systems, Mitchell introduced the term *biological computation* [6], which describes information processing in biological systems as highly distributed, parallel, dynamic, and stochastic. In other words, global behavior arises from the non-trivial combination of collective activity of many sub-systems, distributed across space; furthermore, individual sub-system serve different computational tasks over time. Local information processing—with respect to space or time or both—performed by individual sub-systems, may not serve a human-understandable task or implement universal computation—the two common intuitions when we speak about computation. This local information processing has been termed *intrinsic computation* [7]. Classical tools like complexity theory or a description in terms of function may be ill-equipped to describe this intrinsic computation and how it gives rise to the behavior of the system as a whole. Instead, new methodological approaches are required—here, Mitchell and colleagues proposed to quantify the intrinsic computation performed by single agents or arbitrary sub-systems in terms of generic information processing operations: the transfer, storage, and modification of information. Together, these operations enable universal computation and a decomposition of computation into this building blocks has already been proposed by Alan Turing [5, 7].

However, measures for the quantification of the basic operations of computation have been lacking; only recently, a complete framework was proposed by Lizier [9], choosing information theory as the “language of computation” to quantify each of the three operations. Here, information theory—as introduced by Claude Shannon [10]—is a natural choice for measures of computation, because it provides a mathematical rigorous definition of information and derived measures in an abstract, semantics-free way—hence, we do not have to understand the “meaning” of a signal (which is highly dependent on the observer) to measure its information content. This property is especially desirable to measure the aforementioned intrinsic computation performed on arbitrary organizational levels of a computing system.

Even though information theory has been used extensively in neuroscience (e.g., [13, 53–63]), the framework proposed by Lizier [9] provides the most comprehensive and well defined approach to the generic investigation of computations in distributed systems until today. Lizier proposed the three measures transfer entropy (TE) [12] as a measure of information transfer, active information storage (AIS) [30] as a measure of information storage, and information modification [64] as a measure of the processing of information into a new form. These measures can be localized in time and space to capture distributed and dynamic aspects of computations. The framework has been used successfully to describe the computations performed in cellular automata (CA) [11]. CAs are an important example system in the analysis of distributed computation—here, the authors resolved a debate around rule 22, finding evidence for a complex rather than chaotic behavior of this CA by quantifying information storage and transfer over time. This behavior was not evident from observing the CA's activity directly, but only by investigating the performed computations using the measures presented above. Following this example, it has been proposed to apply Lizier's framework in a similar fashion to the analysis of computations in neural systems [8].

Even though the framework by Lizier [9] provides a mathematically well-defined approach to the investigation of neural computation, adapting information theoretical measures to neural data is far from trivial—among the three basic operations of computation, only TE and AIS have been successfully applied in neuroscience. Here, TE is especially popular as a measure of dependency between neural sites (see [13, 65] and references in [23] for an overview of applications), while AIS has been used to a lesser extend (e.g. [66, 67]). Information modification has not been applied yet, because no agreed-upon information-theoretic measure exists yet (see next section).

Yet, even though TE is theoretically well defined and frequently used, open practical and conceptual problems to its application in neuroscience persist: often approximations and ad-hoc solutions are used with detrimental effects to the estimated quantities (see for example Vicente et al. [13] for some practical problems); also, the interpretation of TE as a measure of computation is often not clear [68]. To some extend, these problems also concern the estimation of AIS (e.g., the setting of estimation parameters as discussed in [30], or its interpretation as discussed in [66]).

The present work addresses the most pressing open problems in the estimation of TE in neuroscience. Before I discuss these problems in more detail, I will first introduce the mathematical background of information theory (Section (1.2, *Information theoretic preliminaries*), before I introduce the information-theoretic measures of TE, AIS, and information modification used in the framework of local information dynamics [9] (Section 1.3, *Measuring information processing*). In the last two sections, I will describe open problems in the estimation of TE from neural data, which partially apply to the estimation of AIS as well (Section 1.4, *Open problems in estimating information processing measures in neuroscience*); I will close with an overview over how the present work addresses the most pressing among these challenges to allow for a robust estimation of information theoretic measures, in particular TE, in neuroscience (Section 1.5, *Contribution of the present work*).

## 1.2 Information theoretic preliminaries

**Introduction** Information theory as introduced by Claude Shannon [10] is concerned with quantifying the "information content" of a message that is transferred over some channel of communication. In neuroscience, we are most interested in the idea of measuring the information content in a "message", where often this term is defined more loosely, such that a message may be the observed spike trains sent over an axon or the membrane potential of a cell over time. To model these messages mathematically, we adopt a probabilistic approach and define a "message" as a random variable $X$ that describes the state of some physical system $\mathcal{X}$ (e.g., a cell or cortical area). A random variable is a mapping of the outcome of some event, i.e., the state of $\mathcal{X}$, onto a value $x \in \mathcal{A}_X$, where the set $\mathcal{A}_X$ is called the variable's *alphabet* with number of elements $|\mathcal{A}_X|$. A single outcome or realization $x$ is observed with some probability $p(X = x)$, while $p(X)$ is the probability density function of the random variable $X$. In the following I will write $p(x)$ as a shorthand for $p(X = x)$.

For two systems, $\mathcal{X}$ and $\mathcal{Y}$, described by two random variables $X$ and $Y$, the *joint probability* $p(x, y)$ is defined as the probability of two outcomes being observed together. To describe the probability of observing an outcome given that a second outcome was observed before, we define the *conditional probability* as $p(x|y)$, i.e., the probability of observing $x$ after $y$ was observed. Two random variables are said to be independent if $p(x, y) = p(x)p(y)$.

**Shannon information content and entropy** Shannon's *information content* quantifies the reduction in uncertainty when observing the outcome $x$ of a random event described by a random variable $X$ as

$$h(x) = -\log_b p(x). \tag{1.1}$$

In other words, $h(x)$ quantifies the amount of information we gain when observing outcome $x$. Here, the logarithm is the function of choice to define an information measure that captures our intuitive understanding of information: first, additivity of the information content of two independent events and sub-additivity for non-independent events; and second, $h$ is a continuous and monotonic function of the probability of the occurrence of an event, such that rare events are more informative than frequent ones and events with probability 1 are not informative at all ($h(x) = 0$). For a complete axiomatic definition of the information content see [69].

The choice of the base of the logarithm $b$ is arbitrary, but popular choices are $b = 2$ or $b = e$, yielding the information content in *bits* or *nats*. In the remainder of this work, we will use $b = 2$ if not stated otherwise . This allows us a second characterization of the information content as the minimal number of bits that are needed to encode or represent the outcome $x$.

From this basic measure of information content, all other information-theoretic quantities are derived: by taking the average over all outcomes $x \in \mathcal{A}_X$, weighted by their probability, we obtain the average or expected information content of $X$, the *Shannon entropy*

$$H(X) := E[h(X)] = -\sum_{x \in \mathcal{A}_X} p(x) \log_2 p(x). \tag{1.2}$$

The Shannon entropy thus describes the average amount of information we expect to obtain from observing outcomes of $X$.

For two variables $X$ and $Y$, we can define their *joint entropy* to quantify the information content of the joint distribution of $X$ and $Y$:

$$H(X, Y) = -\sum_{x \in \mathcal{A}_X, y \in \mathcal{A}_Y} p(x, y) \log_2 p(x, y). \tag{1.3}$$

The joint entropy then describes the expected reduction in uncertainty that can be obtained from the joint distribution of $X$ and $Y$.

We may further define the *conditional entropy* of $X$ given $Y$, which quantifies the average information we can still gain from $X$ after having observed $Y$, or the average uncertainty that remains about the outcome of $X$ when the outcome $Y$ is known:

$$H(X|Y) = -\sum_{x \in \mathcal{A}_X, y \in \mathcal{A}_Y} p(x, y) \log_2 p(x|y). \tag{1.4}$$

$H(X|Y)$ may take on any value between 0 (if the outcome of $X$ is completely determined by the outcome of $Y$) and $H(X)$ (if the outcome of $X$ is independent of $Y$, i.e., the two random variables are independent).

**Mutual information** Based on Shannon's measures of entropy we can define the *mutual information* (MI) as a measure of the information that is shared between two variables. From this measure, we will then derive the measures of information processing used in this work, which are all variants of the basic MI.

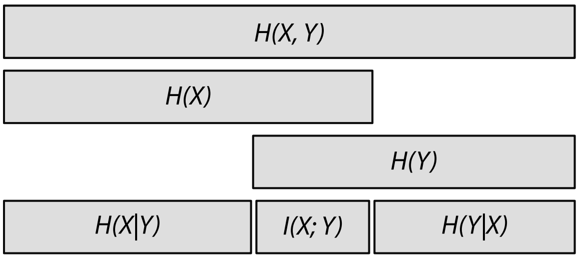


Fig. 1.1 Relationship between joint entropies, conditional entropies, and mutual information (from [70]).

The MI between $X$ and $Y$ measures the average reduction in uncertainty about $X$ that results from observing the value of $Y$, or vice versa. The MI is complementary to the conditional entropy with respect to a variable's entropy: when observing two variables $X$, $Y$ with entropies $H(X)$, $H(Y)$, the average uncertainty about $X$ that *remains* when observing $Y$ is quantified by the conditional entropy $H(X|Y)$; the average uncertainty about $X$ that is *reduced* when observing $Y$ is quantified by the MI (Fig. 1.1):

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= H(X) - H(X|Y) \\ &= \sum_{x\in\mathcal{A}_X} p(x)\log_2 p(x) + \sum_{x\in\mathcal{A}_X, y\in\mathcal{A}_Y} p(x,y)\log_2 p(x|y). \end{aligned} \tag{1.5}$$

The MI is symmetric, such that $I(X;Y) = H(X) - H(X|Y) = I(Y;X) = H(Y) - H(Y|X)$. The sum in last line can be rewritten as a fraction, which in turn can be rewritten after applying the chain rule for probabilities:

$$\begin{aligned} I(X;Y) &= \sum_{x\in\mathcal{A}_X, y\in\mathcal{A}_Y} p(x,y)\log_2 \frac{p(x|y)}{p(x)} \\ &= \sum_{x\in\mathcal{A}_X, y\in\mathcal{A}_Y} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)}. \end{aligned} \tag{1.6}$$

From this formulation, we can see that the MI may also be defined as the Kullback-Leibler divergence of the joint probability distribution of $X$ and $Y$ from the product of their marginal distributions. Hence, the MI measures the dependence between two variables by quantifying their deviation from independence:

$$I(X;Y) = D_{KL}(p(x,y)||p(x)p(y)). \tag{1.7}$$

**Conditional mutual information** We can extend the MI to define the *conditional mutual information* between $X$ and $Y$ given $Z$ as the MI between $X$ and $Y$ when $Z$ is known:

$$\begin{aligned} I(X;Y|Z) &= I(X;Y,Z) - I(X;Z) \\ &= \sum_{\substack{x\in\mathcal{A}_X, y\in\mathcal{A}_Y, \\ z\in\mathcal{A}_Z}} p(x,y,z) \log_2 \frac{p(x|y,z)}{p(x|z)}. \end{aligned} \tag{1.8}$$

Importantly, there are cases where $I(X;Y|Z) > I(X;Y)$, i.e., conditioning on a third variable will increase the MI between two variables. As an example, consider a binary `xor` gate with inputs $X$ and $Y$, and output $Z$. If $X$ and $Y$ are two random, independent inputs, the MI $I(X;Y)$ is zero, while conditioning on the output leads to a conditional mutual information of $I(X;Y|Z) = 1$. Here, observing the common effect induces a dependency between the two causes, which is also called *explaining away* in the theory of Bayesian networks (see for example [71]).

The increase in MI when conditioning on a third variable has also been described in the framework of *partial information decomposition* (PID) [72]. PID describes how the information two or more variables have about a third can be decomposed into *unique information*, i.e., information about the target that is contained solely in one variable; *shared information*, i.e., information that is redundantly present in two or more variables about the target; and *synergistic information*, i.e., information contained jointly in two or more variables, which can not be derived from looking at any subset of the variables alone (see also Fig. 1.2C). Synergistic information is present, whenever additionally considering the information from a second variable $Y$, increases the information we gain from $X$ about $Z$, hence $I(X;Z|Y) > I(X;Z)$. In the the `xor` example, the MI between one input and the output, $I(X;Z)$, is zero, while conditioning on the second input leads to $I(X;Z|Y) = 1$—thus, $X$ and $Y$ can be said to have synergistic information about $Z$. Intuitively, synergistic information can be understood as the information in $X$ that is needed to "decode" the information $Y$ has about $Z$, and vice versa (see also Section 1.3.1, *Transfer entropy* below).

All presented measures of information content readily generalize to multivariate variables $\mathbf{X}$.

## 1.3 Measuring information processing

In the next sections, I will present the three measures of information processing, namely, TE for quantifying information transfer, AIS for quantifying information storage, and information modification. While TE and AIS are readily derivable from traditional information theory, the same is not possible for information modification and here no equivalent measure exists. For the sake of completeness, I will briefly introduce the current state of research in this area.

To allow for a mathematical definition of the time-dynamics aspect of TE and AIS, I extend the basic definition from Section 1.2: if we observe two physical systems $\mathcal{X}$ and $\mathcal{Y}$ over time, observations can be described as realizations $x_n$, $y_n$ of two random processes $X = \{X_1, \dots, X_n, \dots, X_N\}$ and $Y = \{Y_1, \dots, Y_n, \dots, Y_N\}$, where each random processes is a collection random variables indexed by some integer.

### 1.3.1 Transfer entropy

Transfer entropy (TE) quantifies how much our prediction of the future of a process $Y$ improves, if we not only look at $Y$'s own past, but also the past of a second process $X$ [12] (Fig. 1.2A). TE is defined as a conditional mutual information between the future of the target process $Y$ and the past of the source process $X$, conditional on the past of $Y$

$$\begin{aligned} & TE_{SPO}\left(X \rightarrow Y, n, u, k, l\right) \\ & \quad = \lim_{k,l \to \infty} I\left(Y_n; \mathbf{X}^l_{n-u} | \mathbf{Y}^k_{n-1}\right) \\ & \quad = \lim_{k,l \to \infty} \sum_{\substack{y_n \in \mathcal{A}_{Y_n}, \mathbf{x}_{n-u} \in \mathcal{A}_{\mathbf{X}_{n-u}}, \\ \mathbf{y}_{n-1} \in \mathcal{A}_{\mathbf{Y}_{n-1}}}} p\left(y_n, \mathbf{x}_{n-u}, \mathbf{y}_{n-1}\right) \log_2 \frac{p(y_n | \mathbf{y}_{n-1}, \mathbf{x}_{n-u})}{p(y_n | \mathbf{y}_{n-1})}, \end{aligned} \tag{1.9}$$

where $Y_n$ is the future value of the random process $Y$ and $\mathbf{X}^l_{t-u}$, $\mathbf{Y}^k_{n-1}$ are the *past states* of $X$ and $Y$ respectively. Past states are collections of $k$ and $l$ past variables, respectively

$$\begin{aligned}\mathbf{Y}_{n-1}^{k} &= \{Y_{n-1}, Y_{n-1-\tau_Y}, \dots, Y_{n-(k-1)\tau_Y}\},\\ \mathbf{X}_{n-u}^{l} &= \{X_{n-u}, X_{n-u-\tau_X}, \dots, X_{n-u-(l-1)\tau_X}\},\end{aligned} \tag{1.10}$$

which form a *delay-embedding* of dimension $k$ and $l$ ($k$ and $l$ are also called the *history lengths*), where $\tau_{\cdot}$ denotes the delay between samples entering the past state [73][3]. For the target process $Y$, the embedding has to be chosen such that the constructed past state renders $Y_n$ conditionally independent of variables at time points further in the past than $n-(k-1)\tau_Y$ samples. The past state is then maximally informative about $Y_n$, which is an assumption that must hold when estimating both information storage and transfer. In the case of TE, failure to construct the two past states correctly may lead to an overestimation of TE or failure to correctly identify the direction of information transfer (see also Chapter 4).

The variable $u$ is the assumed *information transfer delay* between both processes and accounts for some physical delay $\delta_{\mathcal{X},\mathcal{Y}} \geq 1$ between the processes, $\mathcal{X}$ and $\mathcal{Y}$. By accounting for arbitrary delays between processes our estimator differs from the initial formulation of TE in [12] where TE was defined for $u = 1$ only, such that we have an identical delay of 1 time step for both past states in source and target process. In contrast, our estimator allows for a delay other than 1 between the present value and source past state, while preserving the delay of 1 between present value and target past state; thus, our estimator accounts for possible physical delays between source and target system, while preserving self prediction optimality within the target (hence the subscript $SPO$) [75]. Self prediction optimality here means that the past state is maximally informative about the future value $Y_n$. The true delay $\delta_{\mathcal{X},\mathcal{Y}}$ may be recovered by "scanning" various assumed delays and keeping the delay that maximizes $TE_{SPO}$ as shown in [75]

$$\hat{\delta}_{X,Y} = \arg\max_u \left( TE_{SPO} \left( X \rightarrow Y, n, u, k, l \right) \right). \tag{1.11}$$

TE is measured by a conditional mutual information and as mentioned in Section 1.3, conditioning on $\mathbf{Y}_{n-1}^{k}$ may lead to cases, where $\lim_{k,l\rightarrow\infty} I\left(Y_n; \mathbf{X}_{n-u}^{l} | \mathbf{Y}_{n-1}^{k}\right) > \lim_{k,l\rightarrow\infty} I\left(Y_n; \mathbf{X}_{n-u}^{l}\right)$. In this case, the past states $\mathbf{X}_{n-u}^{l}$, $\mathbf{Y}_{n-1}^{k}$ have synergistic information about $Y_n$ [72]. Hence, conditioning on $\mathbf{X}_{n-u}^{l}$ when calculating TE has two effects: first, information about $Y_t$ that is *redundantly* present in both $\mathbf{X}_{n-u}^{l}$ and $\mathbf{Y}_{n-1}^{k}$ is conditioned out or removed; while second, synergistic information jointly present in $\mathbf{X}_{n-u}^{l}$ and $\mathbf{Y}_{n-1}^{k}$ is "conditioned in".

[3] More elaborate embedding schemes exist (e.g., the non-uniform embedding [74]) and are presented in Chapter 5, *General discussion*.

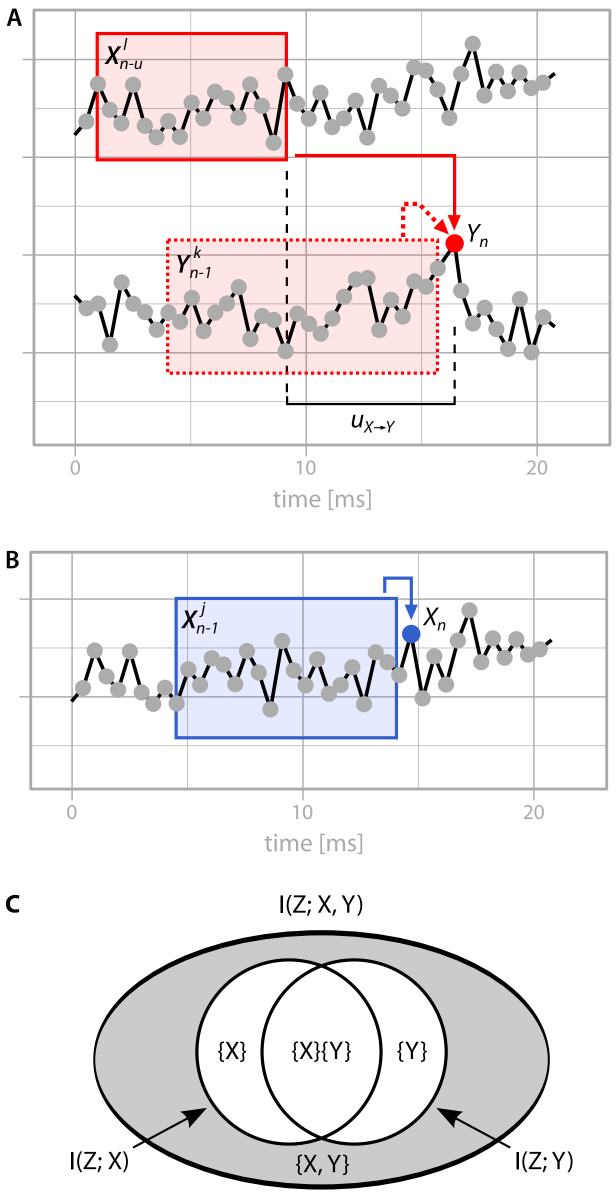


Fig. 1.2 **Information dynamics measures.** (A) Transfer entropy (TE): TE quantifies the information transfer (red arrow) from time series $X$ to time series $Y$ by quantifying how much better the present value of $Y$, $Y_n$, (red sample) can be predicted, if not only its own past, $\mathbf{Y}^k_{n-1}$, (red, dashed box), but also the past of $X$, $\mathbf{X}^l_{n-u}$, (red box) is taken into account (a delay $u_{X\to Y}$ between $\mathbf{X}^l_{n-u}$ and $Y_n$ accounts for the information transfer delay between $X$ and $Y$). (B) Active information storage (AIS): AIS quantifies how well the present value of time series $X$ (blue sample) can be predicted from its immediate past, $\mathbf{X}^j_{n-1}$ (blue box.) (C) Information modification measured by synergistic information (modified from [76]): synergistic information quantifies the information two variables, $X$ and $Y$, have jointly about a third variable $Z$, where this information can not be obtained from one of the two variables alone (gray area). Note that this information may thus be higher (represented by a larger area) than the information in the two variables $X$ and $Y$ (white circles). Synergistic information was defined in the framework of partial information decomposition [72], that proposed a decomposition of the information two variables have about a third, into the components of unique information, $\{X\}$, $\{Y\}$, shared information $\{X\}\{Y\}$, and synergistic information $\{X, Y\}$ (see main text).

### 1.3.2 Active Information Storage

Active information storage (AIS) quantifies how well we can predict the outcome of a random variable $X_n$ at a certain point in time from its immediate past (Fig. 1.2B) [30, 66]. AIS is defined as the MI between $X_n$ and its immediate past state $\mathbf{X}_{n-1}^{j}$:

$$\begin{aligned} AIS(X,n) &= \lim_{j\to\infty} I\left(\mathbf{X}_{n-1}^{j}; X_n\right) \\ &= \lim_{j\to\infty} \sum_{\mathbf{x}_{n-1}^{j}\in\mathcal{A}_{X_{n-1}^{j}}, x_n\in\mathcal{A}_{X_n}} p(x_n, \mathbf{x}_{n-1}^{j}) \log_2 \frac{p(x_n, \mathbf{x}_{n-1}^{j})}{p(x_n)p(\mathbf{x}_{n-1}^{j})}, \end{aligned} \tag{1.12}$$

where the past state $\mathbf{X}_{n-1}^{j}$ is again a collection of past random variables, e.g., a delay embedding (c.f. Eq. 1.10),

$$\mathbf{X}_{n-1}^{j} = \{X_{n-1}, X_{n-1-\tau_X}, \dots, X_{n-1-(j-1)\tau_X}\}. \tag{1.13}$$

AIS may then be interpreted as the reduction in uncertainty about the outcome $X_n$ that we gain from $X_n$'s immediate past, or the past information that is actively in use for the next state update of the system [66]; in other words, AIS may be characterized as a measure of how much information in the past of $X$ is being used to compute the next state at $X_n$, or the amount of information in $X_n$ predictable from its immediate past.

AIS has to be distinguished from related information-theoretic measures that have been used to quantify information storage in neuroscience and other areas: first, AIS is related to the so-called *excess entropy* (see [77] and references therein)

$$E(X,n) = \lim_{j-,j+\to\infty} I\left(\mathbf{X}_{n-1}^{j-}; \mathbf{X}_{n}^{j+}\right), \tag{1.14}$$

where bold fonts indicate again state variables of lengths $k+$, $k-$, with $\mathbf{X}_n^{j+} = \{X_n, X_{n+1}, \dots, X_{n+j+}\}$, describing a collection of future random variables, relative to time point $n$. The excess entropy measures the average information about the whole future of $X$ that is obtainable from $X$'s past. While, AIS quantifies the amount of past information in use at the *next* point in time, $n$, only, the excess entropy quantifies the amount of past information in use at *any* point in time $\geq n$.

Second, AIS is complementary to the so-called *entropy rate* [77]

$$H_\mu(X,n) = \lim_{j\to\infty} H(X_n|\mathbf{X}_{n-1}^j), \tag{1.15}$$

which quantifies the information in $X_n$ that can not be predicted from its immediate past. The entropy rate is complementary to AIS, i.e., $H(X_n) = AIS(X,n)+H_\mu(X,n)$ [30]. Thus, AIS quantifies "how much structure can be resolved rather than how much cannot" [66].

### 1.3.3 Information modification

Information modification describes the "interaction" between transmitted and/or stored information and its "processing into a new form" [5, 78]. A measure of information modification is lacking, because it is not directly derivable from traditional information theoretic measures [9]. Existing work instead used proxies like separable information [64].

In more recent work, Lizier et al. [78] proposed to build on the theoretical framework of PID, in particular synergistic information, for a proper definition of a measure of information modification (see also Section 1.2 and Fig. 1.2C). Synergistic information quantifies the intuitive notion of information processing, namely the non-trivial combination of two or more sources into an output that is not obtainable from one source or a sub-set of sources alone. Yet, a practical measure of unique, redundant, or synergistic information, and its estimation from experimental data, is still missing and the definition of such a measure is an area of active research[4] [8, 72, 78–81]. I discuss the current state of research in Section 5.2.1, *Quantification of information modification and its relevance to the interpretation of transfer entropy* in Chapter 5, *General discussion*).

## 1.4 Open problems in estimating information processing measures in neuroscience

The presented measures of information transfer and storage are theoretically well-defined, yet their estimation from limited, experimental data is highly non-trivial; also, their conceptual integration with other levels of research is often challenging. I will here describe the estimation of TE (and to some degree AIS) in neuroscience

[4] Note that a measure of one quantity is sufficient to calculate the remaining three.

research and outline the most pressing open problems. These problems have already been discussed in greater detail by Vicente et al. [13].

**Lack of data in the estimation of TE** As introduced in Section 1.2, information-theoretic measures are functionals of probability distributions. When applying these functionals to variables observed in experimental neuroscience—but also other fields—these probability distributions are typically not known. Hence, the distributions or the functionals have to be estimated from the collected, finite data, which are typically noisy and relatively sparse. We thus have to choose a suitable estimator that yields robust results despite the listed limitations. For TE and AIS, we either require an estimator of entropies and conditional entropies, or of MI (because TE can easily be decomposed into respective terms, see Section 2.2).

The "quality" of an estimator can be described by its bias, variance, and convergence: An estimator of a parameter $\theta$ is a function of the data that maps i.i.d. samples $x \in \{x_1, \ldots, x_N\}$ from a sample space to a set of estimates $T : x \mapsto \hat{\theta}$. The bias of this estimator is then the expected difference between the estimator and the true value to be estimated

$$B(T) := E_X \left[\hat{\theta}(x) - \theta\right]. \tag{1.16}$$

An estimator with zero bias is called *unbiased*.

The *variance* of the estimator is the expected value of the squared deviation from the expected value

$$var(T) := E_X \left[\left(\hat{\theta}(x) - E_X(\theta)\right)^2\right]. \tag{1.17}$$

and quantifies how far the set of estimates deviates on average from the expected value.

The estimator is called *consistent*, if for increasing amounts of data the estimate converges to the true value of the parameter

$$\lim_{N\to\infty} P(|T(X) - \theta| < \epsilon) = 1. \tag{1.18}$$

In other words, with increasing sample size, the probability of the estimate being close to the true value $\theta$ increases as well. See, for example, [82] for an introduction to information theory and its estimation in neuroscience.

It is desirable to chose an estimator which has low bias and variance, and quickly converges to the true value of the parameter to be estimated. Furthermore, estimators differ with respect to the type of data—continuous or discrete—they are designed for, such that an estimator appropriate for the data at hand should be chosen. Since the present work mostly concerns continuous electrophysiological data, a continuous estimator should be preferred—here, a nearest-neighbor based estimator by Kraskov, Stögbauer, and Grassberger (KSG-estimator [83]) for the estimation of MI has the most favorable bias properties [13, 84, 85]. The estimator is furthermore suitable for the estimation of TE from high-dimensional data [13, 85] and for the detection of small information transfer in noisy data [75]; it has thus become a standard for the estimation of TE and AIS [13, 86, 87].

One downside of the KSG-estimator is that it requires a considerable amount of realizations of the involved random variables for the estimation of TE. This amount of data may not always be available in neuroscience experiments, for example, where recording times have to account for tiring of subjects. A common approach is here to pool recorded data over time to obtain a sufficient amount of realizations; however, this practice requires stationarity of the underlying processes, which may not be guaranteed in neuroscience experiments (see [13]). Here, the applicability of TE suffers because of the lack of available data—if stationarity is wrongly assumed, the estimated values may be erroneous.

**Spurious TE estimates due to multivariate effects** The presented TE functional (Eq. 1.9) quantifies the information transfer between one source process $X$ and one target process $Y$. This has also been termed *apparent transfer entropy* by Lizier et al. [65], and may lead to spurious results if apparent TE is estimated in systems of more than two interacting sub-systems (which is often the case in neuroscience applications). Here, it is likely that information transfer $X \rightarrow Y$ does not happen in an isolated manner, i.e., it is influenced by third variables $\mathbf{Z}$. In this case, to detect the "true" information transfer $X \rightarrow Y$, the influence of $\mathbf{Z}$ has to be accounted for by an additional conditioning: $TE_{SPO}\left(X \rightarrow Y|\mathbf{Z}, n, u, k, l\right) = \lim_{k,l\rightarrow\infty} I\left(Y_n; \mathbf{X}^l_{n-u}|\mathbf{Y}^k_{n-1}, \mathbf{Z}\right)$. Yet, the computational demand of calculating TE in such a *multivariate* fashion increases exponentially in the problem size—hence, an exhaustive solution may not be tractable for arbitrary problem sizes. Here, approximative methods have to be found to alleviate the impact of multivariate effects on estimates of apparent TE.

**Failure to properly optimize estimation parameters** As described in Section 1.3.1, when considering random processes for the estimation of TE, we have to construct past states to fully describe the *state* of a system. Failure to do so may result in over- or under-estimation of TE, or even in the detection of spurious information transfer in the wrong direction. Furthermore, information transfer delays have to be accounted for, such that TE is not underestimated or not missed at all [75]. Also, in systems with bi-directional coupling, the reconstruction of the transfer delay is crucial, especially when comparing the strength of information transfer between directions [75]. Existing research often neglects the proper reconstruction of embedding parameters and information transfer delays, which potentially leads to erroneous results when estimating TE.

**Misinterpretation of TE as a measure of causal interaction** Lastly, TE has been used extensively in neuroscience as a connectivity measure [23, 88], i.e., as a measure to infer dependencies between neuronal sites like single cells [15] or cell assemblies in cortical areas of several millimeter in diameter [14]. In these applications TE has often been interpreted as a measure of *causal* interactions (e.g., [89–91]), i.e., a measure of a mechanistic effect between the sub-systems under investigation. This interpretation has been repeatedly criticized and it has been shown that causal connections and information transfer are two distinct concepts [68, 92]; yet, the interpretation of TE in terms of causal connections remains prevalent in neuroscience.

## 1.5 Contribution of the present work

The present work addresses the challenges listed above in three chapters:

**Chapter 2 – Efficient transfer entropy analysis of non-stationary neural time series** This chapter describes how to estimate TE from non-stationary neural data. The estimation of TE typically requires a considerable amount of data to estimate the involved—and typically unknown—probability density functions $p(\cdot)$. In neuroscience, we often assume stationarity of the observed processes and pool observed realizations of these processes over time to obtain the necessary amount of realizations per random variable. This assumption of stationarity may not always be justified—to still be able to estimate TE (but also other information-theoretic measures), it is possible to pool data over an *ensemble* of physical or temporal copies of the processes under investigation. This approach readily accommodates the data structure typically encountered in neuroscience experiments, were the same task is repeated many times to generate temporal copies of the process under investigation (so called “trials”). Trials are “cyclostationary”, i.e., they are temporal copies of the same initial process. By aligning trials, realizations can be pooled over trials to obtain a sufficient amount of data for information-theoretic estimates. Pooling over trials may be combined with temporal pooling in arbitrarily small time-windows to obtain time-resolved estimates of TE over the course of an experiment.

Yet, there is a technical disadvantage to pooling data over trials when estimating TE or other information-theoretic measures: when pooling data over time alone, such that we obtain one estimate per trial, we can exploit the trial structure to perform necessary statistical tests of our estimates. This is done by estimating TE once for each trial and once for permuted versions of the same trial; we can then perform a permutation t-test between the two data sets (original and surrogate data). When pooling data over trials, this structure gets lost, and with it the ability to perform the permutation test over trials. Instead, each estimate has to be tested against a distribution of estimates from suitable surrogate data. This means, that a sufficient number of surrogate data have to be created, for each of which the estimation has to be repeated. This multiplies the computational demand by the number of surrogates used.

Previous implementations of TE estimators did not allow for an estimation of TE from large surrogate data in feasible time. We therefore used an implementation of the core routines of TE estimation for graphical processing units (GPUs), which allow to handle simple computations in a highly parallel fashion. This highly parallel implementation allowed us to handle the increased computational demand when estimating TE from an ensemble of time series. Enabling the estimation of TE from an ensemble of time series thus allows the estimation this measure from arbitrarily small time windows while still allowing for the necessary statistical testing. The estimation of information-theoretic measures is especially relevant in neuroscience, where data can be expected to be non-stationary, such that an estimation over time does not warrant valid estimates. We presented a reference implementation of the proposed method and demonstrate its application to magnetoencephalographic data.

**Chapter 3 – A graph algorithmic approach to separate direct from indirect neural interactions** In the second chapter we present a method for the post-hoc correction for multivariate effects in apparent or bivariate TE estimates. Commonly, neuroscience data consists of multiple sources of neural activity that are recorded simultaneously. Yet, existing implementations of TE estimation mostly consider a bivariate case, where information is transferred from one source to one target in an isolated fashion. This assumption is typically not true for neuroscience data; yet, a truly multivariate approach, where all possible combinations of sources are considered, poses a NP-hard problem. However, if multivariate interactions are not taken into account, the estimated bivariate interactions may be spurious, i.e., pose false positives.

To reduce the number of false positives in bivariate TE estimates, we presented a post-hoc correction for bivariate TE estimates from multiple—potentially interacting—sources that were analyzed iteratively. The correction consists of a graph-algorithmic approach that investigates the network of bivariate interactions for characteristic timing-signatures that arise from spurious interactions. This requires the reconstruction of information transfer or interaction delays $u$ for each link in the network. In the identified sub-networks, potentially spurious links are flagged and may be removed from the network to obtain a more conservative approximation of a network of truly multivariate interactions.

The presented approximative method allows for the inference of multivariate information transfer from multiple time series. The algorithm is in theory applicable to any bivariate connectivity measure that allows for the reconstruction of interaction delays. We demonstrated the algorithm's application and discuss potential application scenarios.

**Chapter 4 – Anesthesia-related changes in information transfer may be caused by changes in local information processing** In the third chapter, we demonstrate how cortical information transfer measured by TE may be influenced by local information processing in source and target processes of the TE. We investigated TE in two ferrets under different levels of anesthesia, which is known to have an effect on long-range, cortical information transfer. Additionally, we measured local information processing in the source and target of the TE by estimating active information storage and signal entropy.

TE has been found to decrease under anesthesia and this has often been explained by a change in the underlying, physical coupling. Yet, by demonstrating that local information processing in source and target is altered as well, we provide an alternative explanation of reduced TE, which is unrelated to the underlying anatomy. We thus demonstrate that mistaking TE for a causal measure, may lead to erroneous interpretations of experimental results.

Additionally, we confirm existing findings on TE under anesthesia, while applying current best-practice in TE estimation, namely TE estimation using the KSG-estimator and by additionally applying a novel Bayesian estimator for information-theoretic measures. We discuss current best-practices for TE estimation and point out pitfalls that are especially common in anesthesia research and may have detrimental effects on estimated quantities; in particular, we show how under-embedding leads to over-

or underestimation of TE, failure to properly reconstruct interaction delays leads to failure to identify the dominant direction of information transfer, and we discuss how symbolic TE as a proxy to TE estimation on continuous data may miss actual information transfer.

Our results indicate that reduced information transfer under anesthesia may be caused by a reduction in information production in either the source or the target, rather than by changes in cortical coupling. We thus show, how an implicit mixing of levels of explanation, i.e., misinterpretation of TE as a causal measure may lead to erroneous interpretations of the obtained results.

# Efficient transfer entropy analysis of non-stationary neural time series

Patricia Wollstadt[1,*,†], Mario Martínez-Zarzuela[4,†], Raul Vicente[2,3], Francisco J. Díaz-Pernas[4], Michael Wibral[1]

**1** MEG Unit, Brain Imaging Center, Goethe University, Frankfurt am Main, Germany

**2** Frankfurt Institute for Advanced Studies (FIAS), Goethe University, Frankfurt am Main, Germany

**3** Max-Planck Institute for Brain Research, Frankfurt am Main, Germany

**4** Department of Signal Theory and Communications and Telematics Engineering, University of Valladolid, Valladolid, Spain

∗ E-mail: p.wollstadt@stud.uni-frankfurt.de

† These authors contributed equally to this work.

*Published as: P. Wollstadt, M. Martínez-Zarzuela, R. Vicente, J. D. Díaz-Pernas, M. Wibral (2014). "Efficient transfer entropy analysis of non-stationary neural time series". PLoS ONE, 9(7): e102833.*

## Abstract

Information theory allows us to investigate information processing in neural systems in terms of information transfer, storage and modification. Especially the measure of information transfer, transfer entropy, has seen a dramatic surge of interest in neuroscience. Estimating transfer entropy from two processes requires the observation of multiple realizations of these processes to estimate associated probability density functions. To obtain these necessary observations, available estimators typically assume stationarity of processes to allow pooling of observations over time. This assumption however, is a major obstacle to the application of these estimators in neuroscience as observed processes are often non-stationary. As a solution, Gomez-Herrero and colleagues theoretically showed that the stationarity assumption may be avoided by estimating transfer entropy from an ensemble of realizations. Such an ensemble of realizations is often readily available in neuroscience experiments in the form of experimental trials. Thus, in this work we combine the ensemble method with a recently proposed transfer entropy estimator to make transfer entropy estimation applicable to non-stationary time series. We present an efficient implementation of the approach that is suitable for the increased computational demand of the ensemble method's practical application.

In particular, we use a massively parallel implementation for a graphics processing unit to handle the computationally most heavy aspects of the ensemble method for transfer entropy estimation. We test the performance and robustness of our implementation on data from numerical simulations of stochastic processes. We also demonstrate the applicability of the ensemble method to magnetoencephalographic data. While we mainly evaluate the proposed method for neuroscience data, we expect it to be applicable in a variety of fields that are concerned with the analysis of information transfer in complex biological, social, and artificial systems.

## 2.1 Introduction

We typically think of the brain as some kind of information processing system, albeit mostly without having a strict definition of information processing in mind. However, more formal accounts of information processing exist, and may be applied to brain research. In efforts dating back to Alan Turing [93] it was shown that any act of information processing can be broken down into the three components of information storage, information transfer, and information modification [5, 93–95]. These components can be easily identified in theoretical or technical information processing systems, such as ordinary computers, based on the specialized machinery for and the spatial separation of these component functions. In these examples, a separation of the components of information processing via a specialized mathematical formalism seems almost superfluous. However, in biological systems in general, and in the brain in particular, we deal with a form of distributed information processing based on a large number of interacting agents (neurons), and each agent at each moment in time subserves any of the three component functions to a varying degree (see [66] for an example of time-varying storage). In neural systems it is indeed crucial to understand where and when information storage, transfer and modification take place, to constrain possible algorithms run by the system. While there is still a struggle to properly define information modification [64, 78] and its proper measure [72, 80, 81, 96, 97], well established measures for (local active) information storage [30], information transfer [12], and its localization in time and space [65, 98] exist, and are applied in neuroscience (for information storage see [66, 67, 99], for information transfer see below).

Especially the measure for information transfer, transfer entropy (TE), has seen a dramatic surge of interest in neuroscience [13–18, 22, 23, 91, 100–113], physiology [74, 114, 115], and other fields [18, 64, 65, 116, 117]. Nevertheless, conceptual and practical problems still exist. On the conceptual side, information transfer has been for a while confused with causal interactions, and only some recent studies [68, 92, 118] made clear that there can be no one-to-one mapping between causal interactions and information transfer, because causal interactions will subserve all *three* components of information processing (transfer, storage, modification). However, it is information transfer, rather than causal interactions, we might be interested in when trying to understand a computational process in the brain [68].

On the practical side, efforts to apply measures of information transfer in neuroscience have been hampered by two obstacles: (1) the need to analyze the information processing in a multivariate manner, to arrive at unambiguous conclusions that are not clouded by spurious traces of information transfer, e.g. due to effects of cascades and common drivers; (2) the fact that available estimators of information transfer typically require the processes under investigation to be stationary.

The first obstacle can in principle be overcome by conditioning TE on all other processes in a system, using a fully multivariate approach that had already been formulated by Schreiber [12]. However, the naive application of this approach normally fails because the samples available for estimation are typically too few. Therefore, recently four approaches to build an approximate representation of the information transfer network have been suggested: Lizier and Rubinov [25], Faes and colleagues [115], and Stramaglia and colleagues [119] presented algorithms for iterative inclusion of processes into an approximate multivariate description. In the approach suggested by Stramaglia and colleagues, conditional mutual information terms are additionally computed at each level as a self-truncating series expansion, following a suggestion by Bettencourt and colleagues [120]. In contrast to these approaches that explicitly compute conditional TE terms, we recently suggested an approximation based on a reconstruction of information transfer delays [75] and a graphical pruning algorithm [121]. While the first three approaches will eventually be closer to the ground truth, the graphical method may be better applicable to very limited amounts of data. In sum, the first problem of multivariate analysis can be considered solved for practical purposes, given enough data are available.

The second obstacle of dealing with non-stationary processes is also not a fundamental one, as the definition of TE relies on the availability of multiple realizations of (two or more) random processes, that can be obtained by running an ensemble of many identical copies of the processes in question, or by running one process multiple times. Only when obtaining data from such copies or repetitions is impossible, we have to turn to a stationarity assumption in order to evaluate the necessary probability density functions (PDF) based on a single realization.

Fortunately, in neuroscience we can often obtain many realizations of the processes in question by repeating an experiment. In fact, this is the typical procedure in neuroscience - we repeat trials under conditions that are kept as constant as possible (i.e we create a cyclostationary process). The possibility to use such an *ensemble* of data to estimate the time resolved TE has already been demonstrated theoretically by Gomez-Herrero and colleagues [24]. Practically, however, the statistical testing necessary for this ensemble-based method leads to an increase in computational cost by several orders of magnitude, as some shortcuts in statistical validation that can be taken for stationary data cannot be used for the ensemble approach (see [86]): For stationary data, TE is calculated per trial and *one* set of trial-based surrogate data may be used for statistical testing. The ensemble method does not allow for trial-based TE estimation as TE is estimated across trials. Instead, the ensemble method requires the generation of a sufficiently large number of surrogate data sets, for *all* of which TE has to be estimated, thus multiplying the computational demand by the number of surrogate data sets. Therefore, the use of the ensemble method has remained a theoretical possibility so far, especially in combination with the nearest neighbor-based estimation techniques by Kraskov and colleagues [83] that provide the most precise, yet computationally most heavy TE estimates. For example, the analysis of magnetoencephalographic data presented here would require a runtime of 8200 h for 15 subjects and a single experimental condition. It is easy to see that any practical application of the methods hinges on a substantial speed-up of the computation.

Fortunately, the algorithms involved in ensemble-based TE estimation, lend themselves easily to data-parallel processing, since most of the algorithm's fundamental parts can be computed simultaneously. Thus, our problem matches the massively parallel architecture of Graphics Processing Unit (GPU) devices well. GPUs were originally devised only for computer graphics, but are routinely used to speed up computations in many areas today [122, 123]. Also in neuroscience, where applied algorithms continue to grow faster in complexity than the CPU performance, the use of GPUs with data-parallel methods is becoming increasingly important [124] and GPUs have successfully been used to speedup time series analysis in neuroscientific experiments [125–130].

Thus, in order to overcome the limitations set by the computational demands of TE analysis from an ensemble of data, we developed a GPU implementation of the algorithm, where the neighbor searches underlying the binless TE estimation [83] are executed in parallel on the GPU. After parallelizing this computationally most heavy aspect of TE estimation we were able to use the ensemble method for TE estimation proposed by [24], to estimate time-resolved TE from non-stationary neural time-series in acceptable time. Using the new GPU-based TE estimation tool on a high-end consumer graphics card reduced computation time by a factor of 50 compared to the CPU optimized TE search used previously [131]. In practical terms, this speedup shortens the duration of an ensemble-based analysis for typical neural data sets enough to make the application of the ensemble method feasible for the first time.

## 2.2 Background

Our study focuses on making the application of ensemble-based estimation of TE from non-stationary data practical using a GPU-based algorithm. For the convenience of the reader, we will also present the necessary background on stationarity, TE estimation using the Kraskov-Stögbauer-Grassberger (KSG) estimator [13], and the ensemble method of Gomez-Herrero et al. [24] in condensed form in a short background section below. Readers well familiar with these topics can safely skip ahead to the *Implementation* section below.

### 2.2.1 Notation

To describe practical TE estimation from time series recorded in a system of interest $\mathcal{X}$ (e.g. a brain area), we first have to formalize these recordings mathematically: We define an observed time series $\mathbf{x} = (x_1, x_2, \ldots, x_t, \ldots, x_N)$ as a realization of a random process $\mathtt{X} = (X_1, X_2, \ldots, X_t, \ldots, X_N)$. A random process here is simply a collection of individual random variables sorted by an integer index $t \in \{1, \ldots, N\}$, representing time. TE or other information theoretic functionals are then calculated from the random variables' joint PDFs $p_{X_s Y_t}(X_s = a_i, Y_t = b_j)$ and conditional PDFs $p_{X_s|Y_t}(X_s = a_i | Y_t = b_j)$ (with $s, t \in \{1, \ldots, N\}$), where $\mathcal{A}_{X_s} = \{a_1, a_2, \ldots, a_i, \ldots, a_I\}$ and $\mathcal{B}_{Y_t} = \{b_1, b_2, \ldots, b_j, \ldots, b_J\}$ are all possible outcomes of the random variables $X_s$ and $Y_t$, and where $p_{X_s|Y_t}(X_s = a_i | Y_t = b_j) = \frac{p_{X_s Y_t}(X_s = a_i, Y_t = b_j)}{p_{Y_t}(Y_t = b_j)}$.

We call information theoretic quantities functionals as they are defined as functions that map from the space of PDFs to the real numbers. If we have to estimate the underlying probabilities from experimental data first, the mapping from the data to the information theoretic quantity (a real number) is called an estimator.

### 2.2.2 Stationarity and non-stationarity in experimental time series

PDFs in neuroscience are typically not known *a priori*, so in order to estimate information theoretic functionals, these PDFs have to be reconstructed from a sufficient amount of observed realizations of the process. How these realizations are obtained from data depends on whether the process in question is stationary or non-stationary. Stationarity of a process means that PDFs of the random variables that form the random process do not change over time, such that $p_{X_t}(X_t = a_j) = p_{X_{t'}}(X_{t'} = a_j),\ \forall t, t' \in \mathbb{N}$. Any PDF $p_{X_t}(\cdot)$ may then be estimated from one observation of process $\mathtt{X}$ by means of collecting realizations $x_{t'}$ *over time* $t' \in \{1, \ldots, N\}$.

For processes that do not fulfill the stationarity-assumption, temporal pooling is not applicable as PDFs vary over time $t$ and some random variables $X_t$, $X_s$ (at least two) are associated with different PDFs $p_{X_t}(\cdot)$, $p_{X_s}(\cdot)$ (Fig. 2.1). To still gain the necessary multiple observations of a random variable $X_t$ we may resort to either run multiple physical copies of the process `X` or—in cases where physical copies are unavailable—we may repeat a process in time. If we choose the number of repetitions large enough, i.e. there is a sufficiently large set $\mathcal{R}$ of time points $\Theta$, at which the process is repeated, we can assume that

$$\begin{gathered}\exists \mathcal{R} \subseteq \mathbb{N} \wedge \mathcal{R} \neq \emptyset : p_{X_{\Theta+t}}(a_j) = p_{X_{\Theta'+t}}(a_j) \;\; \forall t \in \mathbb{N} : t < min\left(\left|\Theta - \Theta'\right|\right), \\ \forall \Theta, \Theta' \in \mathcal{R},\; \forall a_j \in \mathcal{A}_{X_t},\end{gathered} \tag{2.1}$$

i.e. PDFs $p_{X_{\Theta+t}}(\cdot)$ at time point $t$ relative to the onset of the repetition at $\Theta$ are equal over all $R = |\mathcal{R}|$ repetitions. We call the repeated observations of a process an *ensemble* of time series. We may obtain a reliable estimation of $p_{X_{\Theta+t}}(\cdot)$ from this ensemble by evaluating $p_{\cdot}(\cdot)$ over all observations $x_{\Theta+t}, \forall \Theta \in \mathcal{R}$. For the sake of readability, we will refer to these observations from the ensemble as $x_t(r)$, where $t$ refers to a time point $t$, relative to the beginning of the process at time $\Theta$, and $r = 1, \ldots, R$ refers to the index of the repetition. If a process is repeated periodically, i.e. the repetitions are spaced by a fixed interval $T$ , we call such a process cyclostationary [132]:

$$\exists T : \forall t \; p_{X_t}(a_j) = p_{X_{nT+t}}(a_j) \;\; \forall n, t \in \mathbb{N}, t < T,\; \forall a_j \in \mathcal{A}_{X_t}. \tag{2.2}$$

In neuroscience, ensemble evaluation for the estimation of information theoretic functionals becomes relevant as physical copies of a process are typically not available and stationarity of a process can not necessarily be assumed. Gomez-Herrero and colleagues recently showed how ensemble averaging may be used to nevertheless estimate information theoretic functionals from cyclostationary processes [24]. In neuroscience for example, a cyclostationary process, and thus an ensemble of data, is obtained by repeating an experimental manipulation, e.g. the presentation of a stimulus; these repetitions are often called experimental *trials*. In the remainder of this article, we will use the term repetition, and interpret trials from a neuroscience experiment as a special case of repetitions of a random process. Building on such repetitions, we next demonstrate a computationally efficient approach to the estimation of TE using the ensemble method proposed in [24].

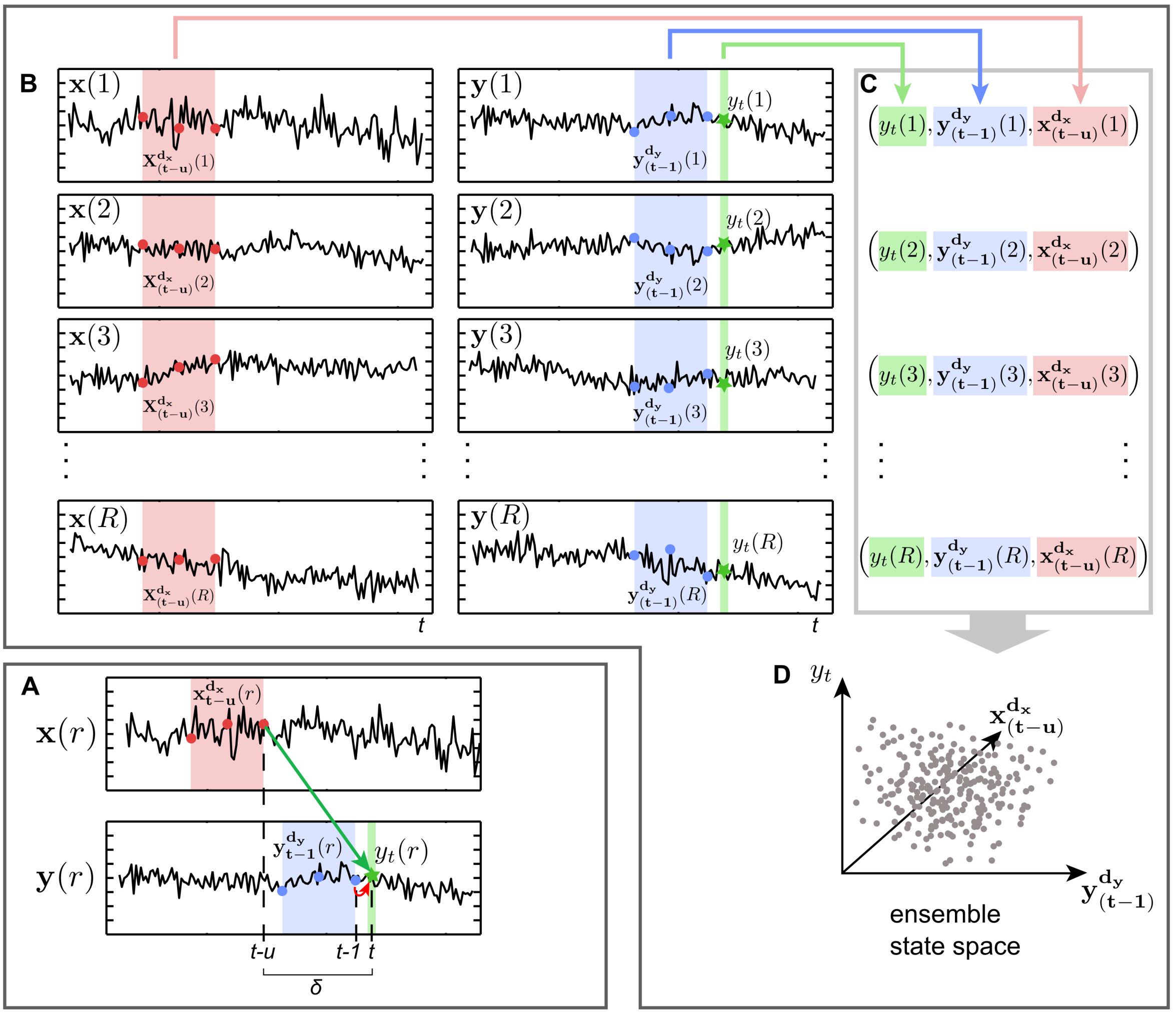

B
$\mathbf{x}(1)$
$\mathbf{x}^{\mathbf{d_x}}_{(\mathbf{t-u})}(1)$
$\mathbf{x}(2)$
$\mathbf{x}^{\mathbf{d_x}}_{(\mathbf{t-u})}(2)$
$\mathbf{x}(3)$
$\mathbf{x}^{\mathbf{d_x}}_{(\mathbf{t-u})}(3)$
$\mathbf{x}(R)$
$\mathbf{x}^{\mathbf{d_x}}_{(\mathbf{t-u})}(R)$
$t$
$\mathbf{y}(1)$
$y_t(1)$
$\mathbf{y}^{\mathbf{d_y}}_{(\mathbf{t-1})}(1)$
$\mathbf{y}(2)$
$y_t(2)$
$\mathbf{y}^{\mathbf{d_y}}_{(\mathbf{t-1})}(2)$
$\mathbf{y}(3)$
$y_t(3)$
$\mathbf{y}^{\mathbf{d_y}}_{(\mathbf{t-1})}(3)$
$\mathbf{y}(R)$
$y_t(R)$
$\mathbf{y}^{\mathbf{d_y}}_{(\mathbf{t-1})}(R)$
$t$
C
$\left(y_t(1), \mathbf{y}^{\mathbf{d_y}}_{(\mathbf{t-1})}(1), \mathbf{x}^{\mathbf{d_x}}_{(\mathbf{t-u})}(1)\right)$
$\left(y_t(2), \mathbf{y}^{\mathbf{d_y}}_{(\mathbf{t-1})}(2), \mathbf{x}^{\mathbf{d_x}}_{(\mathbf{t-u})}(2)\right)$
$\left(y_t(3), \mathbf{y}^{\mathbf{d_y}}_{(\mathbf{t-1})}(3), \mathbf{x}^{\mathbf{d_x}}_{(\mathbf{t-u})}(3)\right)$
$\left(y_t(R), \mathbf{y}^{\mathbf{d_y}}_{(\mathbf{t-1})}(R), \mathbf{x}^{\mathbf{d_x}}_{(\mathbf{t-u})}(R)\right)$
A
$\mathbf{x}(r)$
$\mathbf{x}^{\mathbf{d_x}}_{\mathbf{t-u}}(r)$
$\mathbf{y}(r)$
$\mathbf{y}^{\mathbf{d_y}}_{\mathbf{t-1}}(r)$
$y_t(r)$
t-u
t-1
t
δ
D
$y_t$
$\mathbf{x}^{\mathbf{d_x}}_{(\mathbf{t-u})}$
$\mathbf{y}^{\mathbf{d_y}}_{(\mathbf{t-1})}$
ensemble
state space

Fig. 2.1 **Pooling of data over an ensemble of time series for transfer entropy (TE) estimation.** (A) Schematic account of TE. Two scalar time series $X_r$ and $Y_r$ recorded from the $r^{th}$ repetition of processes *X* and *Y*, coupled with a delay $\delta$ (indicated by green arrow). Colored boxes indicate delay embedded states $\mathbf{x}^{\mathbf{dx}}_{\mathbf{t-u}}(r)$, $\mathbf{y}^{\mathbf{d_Y}}_{\mathbf{t-1}}(r)$ for both time series with dimension $d_X = d_Y = 3$ samples (colored dots). The star on the *Y* time series indicates the scalar observation $y_t$ that is obtained at the target time of information transfer $t$. The red arrow indicates self-information-transfer from the past of the target process to the random variable $Y_t$ at the target time. $u$ is chosen such that $u = \delta$ and influences of the state $\mathbf{x}^{\mathbf{dx}}_{\mathbf{t-u}}(r)$ arrive exactly at the information target variable $Y_t$. Information in the past state of *X* is useful to predict the future value of *Y* and we obtain nonzero TE. (B) To estimate probability density functions for $\mathbf{x}^{\mathbf{dx}}_{\mathbf{t-u}}(r)$, $\mathbf{y}^{\mathbf{d_Y}}_{\mathbf{t-1}}(r)$ and $y_t(r)$ at a certain point in time $t$, we collect their realizations from observed repetitions $r = 1, \ldots, R$. (C) Realizations for a single repetition are concatenated into one embedding vector and (D) combined into one ensemble state space. Note, that data are pooled over the ensemble of data instead of time. Nearest neighbor counts within the ensemble state space can then be used to derive TE using the Kraskov-estimator proposed in [83].

### 2.2.3 Transfer entropy estimation from an ensemble of time series

**Ensemble-based TE functional.** When independent repetitions of an experimental condition are available, it is possible to use ensemble evaluation to estimate various PDFs from an ensemble of repetitions of the time series [24]. By eliminating the need for pooling data over time, and instead pooling over repetitions, ensemble methods can be used to estimate information theoretic functionals for non-stationary time series. Here, we follow the approach of [24] and present an ensemble TE functional that extends the TE functional presented in [13, 14, 75] and also takes into account an extension of the original formulation of TE, presented in [75], guaranteeing self prediction optimality (indicated by the subscript $SPO$). In the next subsection, we will then present a practical and data-efficient estimator of this functional. The functional reads

$$TE_{SPO}\left(X \rightarrow Y, t, u\right) = I(Y_t; \mathbf{X}^{d_X}_{t-u} | \mathbf{Y}^{d_Y}_{t-1}) \; , \tag{2.3}$$

where $I(\cdot;\cdot|\cdot)$ is the conditional mutual information, and $Y_t$, $\mathbf{Y}^{d_Y}_{t-1}$, and $\mathbf{X}^{d_X}_{t-u}$ are the current value and the $d_Y$-dimensional past state variables of the target process Y, and the $d_X$-dimensional past state variable at time $t-u$ of the source process X, respectively (see next paragraph for an explanation of states).

Rewriting this, taking into account repetitions $r$ of the random processes explicitly we obtain:

$$\begin{aligned} TE_{SPO}\left(X \rightarrow Y, t, u\right) = & \sum_{\substack{y_t(r), \mathbf{y}^{d_Y}_{t-1}(r), \mathbf{x}^{d_X}_{t-u}(r) \\ \in \mathcal{A}_{Y_t, \mathbf{Y}^{d_Y}_{t-1}, \mathbf{X}^{d_X}_{t-u}}}} p\left(y_t(r), \mathbf{y}^{d_Y}_{t-1}(r), \mathbf{x}^{d_X}_{t-u}(r)\right) \\ & \log \frac{p\left(y_t(r) | \mathbf{y}^{d_Y}_{t-1}(r), \mathbf{x}^{d_X}_{t-u}(r)\right)}{p\left(y_t(r) | \mathbf{y}^{d_Y}_{t-1}(r)\right)} \; . \end{aligned} \tag{2.4}$$

Here, $u$ is the assumed delay of the information transfer between processes $X$ and $Y$ [75]; $y_t(r)$ denotes the future observation of $Y$ in repetition $r = 1, \ldots, R$; $\mathbf{y}^{d_Y}_{t-1}(r)$ denotes the past state of $Y$ in repetition $r$ and $\mathbf{x}^{d_X}_{t-u}(r)$ denotes the past state of $X$ in repetition $r$. Note, that the functional $TE_{SPO}$ used here is a modified form of the original TE formulation introduced by Schreiber [12]. Schreiber defined TE as a conditional mutual information $TE\left(X \rightarrow Y, t\right) = I(Y_t; \mathbf{X}^{d_x}_{t-1} | \mathbf{Y}^{d_y}_{t-1})$, whereas the functional in Eq. 2.3 implements the conditional mutual information $TE_{SPO}\left(X \rightarrow Y, t, u\right) = I(Y_t; \mathbf{X}^{d_x}_{t-u} | \mathbf{Y}^{d_y}_{t-1})$ [75]. The latter functional, $TE_{SPO}$, contains the definition of Schreiber as a special case for $u = 1$. Note that the two functionals are identical if $TE_{SPO}$ is used with the physically correct delay $\delta$ (i.e. $u = \delta$) and a proper embedding for the source, and the Schreiber measures is used with an over-embedding such that the source state at $(t - \delta)$ is still fully covered by the source embedding.

In addition to the original formulation of $TE_{SPO}$ in [75], here we explicitly state that the necessary realizations of the random variables in question are obtained through *ensemble evaluation* over repetitions $r$—assuming the underlying processes to be repeatable or cyclostationary. Furthermore, we note explicitly that this ensemble-based functional introduces the possibility of time resolved TE estimates.

We recently showed that the estimator presented in [75] can also be used to recover an unknown information transfer delay $\delta$ between two processes $X$ and $Y$, as $TE_{SPO}\,(X \rightarrow Y, t, u)$ is maximal when the assumed delay $u$ is equal to the true information transfer delay $\delta$ [75]. This holds for the extended estimator presented here, thus

$$\delta = \arg\max_{u} \left( TE_{SPO}\,(X \rightarrow Y, t, u) \right) . \tag{2.5}$$

**State space reconstruction and practical estimator.** Transfer entropy differs from the lagged mutual information $I(Y_t; \mathbf{X}_{t-u}^{d_x})$ by the additional conditioning on the past of the target time series, $\mathbf{Y}_{t-1}^{d_Y}$. This additional conditioning serves two important functions. First, as mentioned already by Schreiber in the original paper [12], and later detailed by Lizier [95] and Wibral and colleagues [23, 75], it removes the information about the future of the target time-series $Y_t$ that is already contained in its own past, $\mathbf{Y}_{t-1}^{d_Y}$. Second, this additional conditioning allows for a discovery of information transfer from the source $\mathbf{X}_{t-u}^{d_X}$ to the target that can only be seen when taking into account information from the past of the target $\mathbf{Y}_{t-1}^{d_Y}$ [76]. In the second case, the past information from the target serves to "decode" this information transfer, and acts like a key in cryptography. As a consequence of this importance of the past of the target process it is very important to take all the necessary information in this past into account when evaluating the TE as in Eq. 2.4.

To this end we need to form a collection of past random variables

$$\mathbf{Y}_{t-1}^{d_Y} = \left( Y_{t-1}, Y_{t-1-\tau}, \ldots, Y_{t-1-(d_Y-1)\tau} \right) , \tag{2.6}$$

such that their realizations,

$$\mathbf{y}_{t-1}^{d_Y}(r) = (y_{t-1}(r), y_{t-1-\tau}, \ldots, y_{t-1-(d_Y-1)\tau}) \ , \tag{2.7}$$

are maximally informative about the future of the target process, $Y_t$.

This task is complicated by the fact the we often deal with multidimensional systems, of which we only observe a scalar variable (here modeled as our random processes X,Y). To see this, think for example of a pendulum (which is a two dimensional system) of which we record only the current position $Y_t$. If the pendulum is at its lowest point, it could be standing still, going left, or going right. To properly describe which state the pendulum is in, we need to know at least the realization of one more random variable $Y_{t-1}$ back in time. Collections of such past random variables whose realizations uniquely describe the state of a process are called *state variables*.

Such a sufficient collection of past variables, called a delay embedding vector, can always be reconstructed from scalar observations for low dimensional deterministic systems, such as the above pendulum, as shown by Takens [73]. Unfortunately, most real world systems are high-dimensional stochastic dynamic systems (best described by non-linear Langevin equations) rather than low-dimensional deterministic ones. For these systems it is not obvious that a delay embedding similar to Takens' approach would yield the desired results. In fact, many systems can be shown to require an infinite number of past random variables when only a scalar observable of the high-dimensional stochastic process is accessible. Nevertheless, as shown by Ragwitz and Kantz [133], the behavior of scalar observables of most of these systems can be approximated very well by a finite collection of such past variables for all practical purposes; in other words, these systems can be approximated well by a finite order, one-dimensional Markov-process.

For practical TE estimation using Eq. 2.4, we therefore proceed by first reconstructing the state variables of such approximated Markov processes for the two systems $\mathcal{X}$, $\mathcal{Y}$ from their scalar time series. Then, we use the statistics of nearest ensemble neighbors with a modified KSG estimator for TE evaluation [83].

Thus, we select a delay embedding vector of the form $\mathbf{Y}_{t-1}^{d_Y} = (Y_{t-1}, Y_{t-1-\tau}, \ldots, Y_{t-1-(d_Y-1)\tau})$ from Eq. 2.6 as our collection of past random variables—with realizations in repetition $r$ given by $\mathbf{y}_{t-1}^{d_Y}(r) = (y_{t-1}(r), y_{t-1-\tau}, \ldots, y_{t-1-(d_Y-1)\tau})$. Here, $d_Y$ is called the embedding dimension and $\tau$ the embedding delay. These embedding parameters $d_Y$ and $\tau$, are chosen such that they optimize a local predictor [133], as this avoids an overestimation of TE [75]; other approaches related to minimizing non-linear prediction errors are also possible [74]. In particular, $d_Y \cdot \tau$ is chosen such that $\mathbf{Y}_t^{d_Y}$ is conditionally independent of any $\mathbf{Y}_e^{d_Y}$ with $e < t - d \cdot \tau$ given $\mathbf{Y}_{t-1}^{d_Y}$.

The same is done for the process X at time $t-u$.
Next, we decompose $TE_{SPO}$ into a sum of four individual Shannon entropies:

$$\begin{aligned} TE_{SPO}\left(X \rightarrow Y, t, u\right) = & \; H\left(\mathbf{Y}_{t-1}^{d_Y}, \mathbf{X}_{t-u}^{d_X}\right) - H\left(Y_t, \mathbf{Y}_{t-1}^{d_Y}, \mathbf{X}_{t-u}^{d_X}\right) \\ & + H\left(Y_t, \mathbf{Y}_{t-1}^{d_Y}\right) - H\left(\mathbf{Y}_{t-1}^{d_Y}\right) . \end{aligned} \tag{2.8}$$

The Shannon differential entropies in Eq. 2.8 can be estimated in a data efficient way using nearest neighbor techniques [134, 135]. Nearest neighbor estimators yield a non-parametric estimate of entropies, assuming only a smoothness of the underlying PDF. It is however problematic to simply apply a nearest neighbor estimator (for example the Kozachenko-Leonenko estimator [134]) to each term appearing in Eq. 2.8. This is because the dimensionality of each space associated with the terms differs largely over terms. Thus, a fixed number of neighbors for the search would lead to very different spatial scales (range of distances) for each term. Since the error bias of each term is dependent on these scales, the errors would not cancel each other but accumulate. We therefore use a modified KSG estimator which handles this problem by only fixing the number of neighbors $k$ in the highest dimensional space (k-nearest neighbor search, kNNS) and by projecting the resulting distances to the lower dimensional spaces as the range to look for and count neighbors there (range search, RS) (see [83], type 1 estimator, and [86, 88]). In the ensemble variant of TE estimation we proceed by searching for nearest neighbors across points from all repetitions instead of searching the same repetition as the point of reference of the search—thus we form an *ensemble search space* by combining points over repetitions. Finally, the ensemble estimator of TE reads

$$\begin{aligned} TE_{SPO}\left(X \rightarrow Y, t, u\right) = \psi\left(k\right) + \langle \psi & \left(n_{\mathbf{y_{t-1}^{d_Y}}(r)} + 1\right) \\ -\psi & \left(n_{y_t(r)\,\mathbf{y_{t-1}^{d_Y}}(r)} + 1\right) \\ -\psi\left(n_{\mathbf{y_{t-1}^{d_Y}}(r)\,\mathbf{x_{t-u}^{d_X}}(r)} + 1\right) & \rangle_r , \end{aligned} \tag{2.9}$$

where $\psi$ denotes the digamma function and the angle brackets ($< \cdot >_r$) indicate an averaging over points in different repetitions $r$ at time instant $t$. The distances to the $k$-th nearest neighbor in the highest dimensional space (spanned by $Y_t, \mathbf{Y}_{t-1}^{d_Y}, \mathbf{X}_{t-u}^{d_X}$) define the radius of the spheres for the counting of the number of points ($n_{\cdot}$) in these spheres around each state vector ($\cdot$) involved.

In cases where the number of repetitions is not sufficient to provide the necessary amount of data to reliably estimate Shannon entropies through an ensemble average, one may combine ensemble evaluation with collecting realizations over time. In these cases, we count neighbors in a time window $t' \in [t^-, t^+]$ with $t^- \leq t' \leq t^+$, where $\Delta_t = t^+ - t^-$ controls the temporal resolution of the TE estimation:

$$\begin{aligned} TE_{SPO}\left(X \rightarrow Y, t', u\right) = \psi\left(k\right) + \langle \psi & \left(n_{\mathbf{y_{t'-1}^{d_Y}}(r)} + 1\right) \\ -\psi & \left(n_{y_{t'}(r)\ \mathbf{y_{t'-1}^{d_Y}}(r)} + 1\right) \\ -\psi\left(n_{\mathbf{y_{t'-1}^{d_Y}}(r)\ \mathbf{x_{t'-u}^{d_X}}(r)} + 1\right)\rangle_{r,t'} . \end{aligned} \tag{2.10}$$

## 2.3 Implementation

The estimation of TE from finite time series consists of the estimation of joint and marginal entropies as shown in equations 2.9 and 2.10, calculated from nearest neighbor statistics, i.e. distances and the count of neighbors within these distances. In practice we obtain these neighbor counts by applying kNNS and RS to reconstructed state spaces. In particular, we use a kNNS in the highest dimensional space to determine the k-th nearest neighbor of a data point and the associated distance. This distance is then used as the range for the RS in the marginal spaces, that return the point counts $n_.$. Both searches have a high computational cost. This cost increases even further in a practical setting, where we need to calculate TE for a sufficient number of surrogate data sets for statistical testing (see [13] and below for details). To enable TE estimation and statistical testing despite its computational cost, we implemented ad-hoc kNNS and RS algorithms in NVIDIA® CUDA™ C/C++ code [136]. This allows to run thousands of searches in parallel on a modern GPU.

To allow for a better understanding of the parallelization used, we will now briefly describe the main work flow of TE analysis in the open source MathWorks® MATLAB® toolbox TRENTOOL [86], which implements the approach to TE estimation described in the *Background* section. The work flow includes the steps of data preprocessing prior to the use of the GPU algorithm for neighbor searches as well as the statistical testing of resulting TE values. In a subsequent section we will describe the core implementation of the algorithm in more detail and present its integration into TRENTOOL.

### 2.3.1 Main analysis work flow in TRENTOOL

**Practical TE estimation in TRENTOOL.** The practical GPU-based TE estimation in TRENTOOL 3.0 is divided into the two steps of data preparation and TE estimation (see Fig. 2.2 and the TRENTOOL 3.0 manual: `http://www.trentool.de`). As a first step, data is prepared by optimizing embedding parameters for state space reconstruction (Fig. 2.2A). As a second step, TE is estimated by following the approach for ensemble-based TE estimation lined out in the preceding section (Fig. 2.2B). TRENTOOL estimates $TE_{SPO}\left(X \to Y, t, u\right)$ (Eq. 2.4) for a given pair of processes $X$ and $Y$ and given values for $u$ and $t$. For each pair, we call $X$ the source and $Y$ the target process.

After data preparation $TE_{SPO}\left(X \to Y, t, u\right)$ (Eq. 2.9 and 2.10) is estimated in six steps: (1) using optimized embedding parameters, original data is embedded per repetition and repetitions are concatenated forming the ensemble search space of the original data, (2) $S$ sets of surrogate data are created from the original data by shuffling the repetitions of the target process $Y$, (3) each surrogate dataset is embedded per repetition and concatenated forming $S$ additional ensemble search spaces for surrogate data, (4) all $S+1$ search spaces of embedded original and surrogate data are passed to a wrapper function that calls the GPU functions to perform individual neighbor searches for each search space in parallel (in the following, we will refer to each of the $S+1$ ensembles as one data *chunk*), (5) TE values are calculated for original and surrogate data chunks from the neighbor counts using the KSG-estimator [83], (6) TE values for original data are tested statistically against the distribution of surrogate TE values.

The proposed GPU algorithm is accessed in step (4). As we will further explain below (see paragraph on *Input data*), the GPU implementation uses the fact that all of the necessary computations on surrogate data sets and the original data are independent and can thus be performed in parallel.

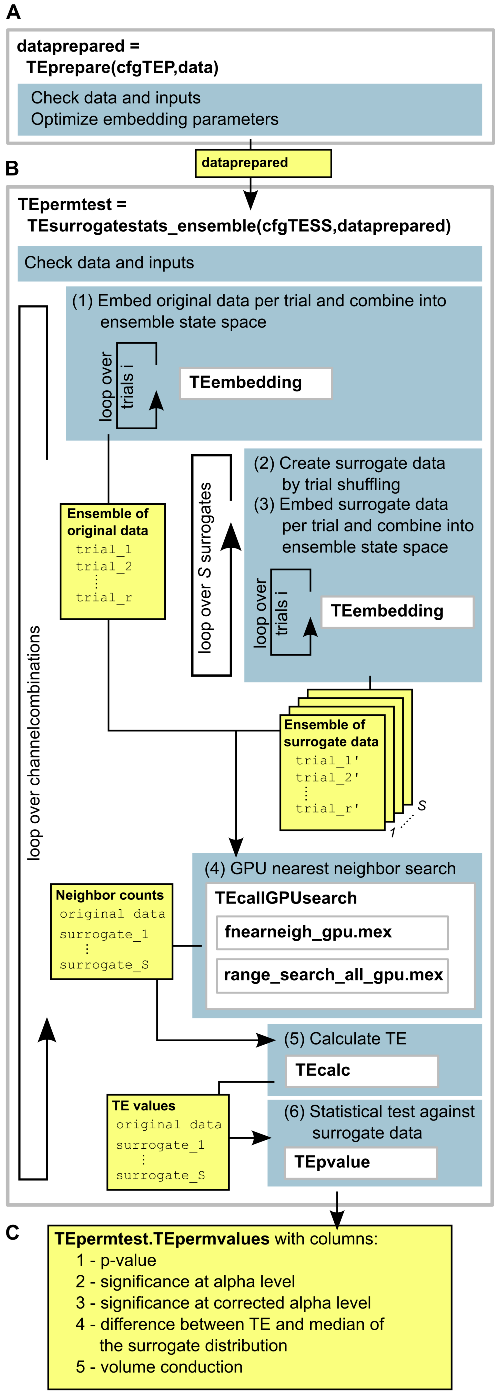

A
dataprepared =
TEprepare(cfgTEP,data)
Check data and inputs
Optimize embedding parameters
dataprepared
B
TEpermtest =
TEsurrogatestats_ensemble(cfgTESS,dataprepared)
Check data and inputs
(1) Embed original data per trial and combine into ensemble state space
loop over trials i
TEembedding
loop over channelcombinations
Ensemble of original data
trial_1
trial_2
trial_r
loop over S surrogates
(2) Create surrogate data by trial shuffling
(3) Embed surrogate data per trial and combine into ensemble state space
loop over trials i
TEembedding
Ensemble of surrogate data
trial_1'
trial_2'
trial_r'
1
S
(4) GPU nearest neighbor search
TEcallGPUsearch
fnearneigh_gpu.mex
range_search_all_gpu.mex
Neighbor counts
original data
surrogate_1
surrogate_S
(5) Calculate TE
TEcalc
TE values
original data
surrogate_1
surrogate_S
(6) Statistical test against surrogate data
TEpvalue
C
TEpermtest.TEpermvalues with columns:
1 - p-value
2 - significance at alpha level
3 - significance at corrected alpha level
4 - difference between TE and median of the surrogate distribution
5 - volume conduction

Fig. 2.2 **Transfer entropy estimation using the ensemble method in TRENTOOL 3.0.** (A) Data preparation and optimization of embedding parameters in function TEprepare.m;

**TE calculation and statistical testing against surrogate data.** Estimated TE values need to be tested for their statistical significance [86] (step (6) of the main TRENTOOL work flow). For this statistical test under a null hypothesis of *no* information

Fig. 2.2 **Transfer entropy estimation using the ensemble method in TRENTOOL 3.0 (continued).** (B) transfer entropy (TE) estimation from prepared data in `TEsurrogatestats_ensemble.m` (yellow boxes indicate variables being passed between sub-functions). TE is estimated via iterating over all channel combinations provided in the data. For each channel combination: (1) Data is embedded individually per repetition and combined over repetitions into one ensemble state space (chunk), (2) $S$ surrogate data sets are created by shuffling the repetitions of the target time series, (3) each surrogate data set is embedded per repetition and combined into one chunk (forming $S$ chunks in total), (4) $S+1$ chunks of original and surrogate data are passed to the GPU where nearest neighbor searches are conducted in parallel, (5) calculation of TE values from returned neighbor counts for original data and $S$ surrogate data sets using the KSG-estimator [83], (6) statistical testing of original TE value against distribution of surrogate TE values; (C) output of `TEsurrogatestats_ensemble.m`, an array with dimension [no. channels$\times$5], where rows hold results for all channel combinations: (1) p-value of TE for this channel combination, (2) significance at the designated alpha level (1 - significant, 0 - not significant), (3) significance after correction for multiple comparisons, (4) absolute difference between the TE value for original data and the median of surrogate TE values, (5) presence of volume conduction (this is always set to 0 when using the ensemble method as instantaneous mixing is by default controlled for by conditioning on the current state of the source time series $x_t(r)$ [137]).

transfer between a source *X* and target time series *Y*, we estimate $TE_{SPO}\left(X \rightarrow Y, t, u\right)$ and compare it to a distribution of TE values calculated from surrogate data sets. Surrogate data sets are formed by shuffling repetitions in *Y* to obtain *Y'*, such that $\mathbf{y}_t^{d_Y}(r) \rightarrow \mathbf{y}_t^{d_Y}(\phi(r))$ and $y_t(r) \rightarrow y_t(\phi(r))$, where $\phi$ denotes a random permutation of the repetitions $r$ (Fig. 2.3). From this surrogate data set, we calculate surrogate TE values $TE_{SPO}\left(X \rightarrow Y', t, u\right)$. By repeating this process a sufficient number of times $S$, we obtain a distribution of values $TE_{SPO}\left(X \rightarrow Y', t, u\right)$. To asses the statistical significance of $TE_{SPO}\left(X \rightarrow Y, t, u\right)$, we calculate a p-value as the proportion of surrogate TE values $TE_{SPO}\left(X \rightarrow Y', t, u\right)$ equal or larger than $TE_{SPO}\left(X \rightarrow Y, t, u\right)$. This p-value is then compared to a critical alpha level (see for example [86, 138]).

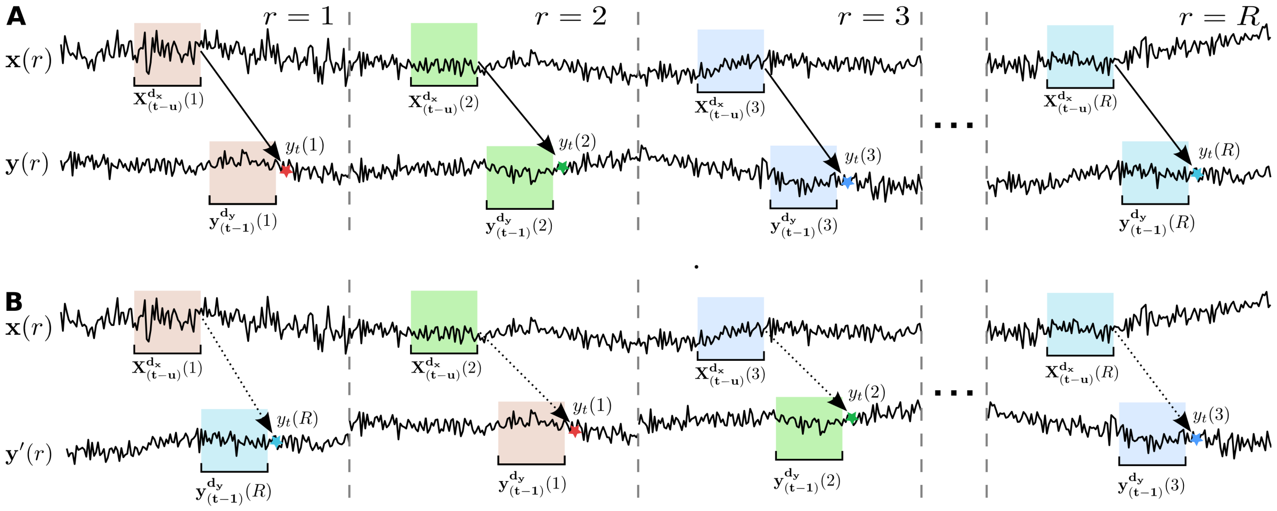

A
$r=1$
$r=2$
$r=3$
$r=R$
$\mathbf{x}(r)$
$\mathbf{y}(r)$
$\mathbf{X}^{\mathbf{d_x}}_{(\mathbf{t-u})}(1)$
$\mathbf{X}^{\mathbf{d_x}}_{(\mathbf{t-u})}(2)$
$\mathbf{X}^{\mathbf{d_x}}_{(\mathbf{t-u})}(3)$
$\mathbf{X}^{\mathbf{d_x}}_{(\mathbf{t-u})}(R)$
$y_t(1)$
$y_t(2)$
$y_t(3)$
$y_t(R)$
$\mathbf{y}^{\mathbf{d_y}}_{(\mathbf{t-1})}(1)$
$\mathbf{y}^{\mathbf{d_y}}_{(\mathbf{t-1})}(2)$
$\mathbf{y}^{\mathbf{d_y}}_{(\mathbf{t-1})}(3)$
$\mathbf{y}^{\mathbf{d_y}}_{(\mathbf{t-1})}(R)$
...
B
$\mathbf{x}(r)$
$\mathbf{y}'(r)$
$\mathbf{X}^{\mathbf{d_x}}_{(\mathbf{t-u})}(1)$
$\mathbf{X}^{\mathbf{d_x}}_{(\mathbf{t-u})}(2)$
$\mathbf{X}^{\mathbf{d_x}}_{(\mathbf{t-u})}(3)$
$\mathbf{X}^{\mathbf{d_x}}_{(\mathbf{t-u})}(R)$
$y_t(R)$
$y_t(1)$
$y_t(2)$
$y_t(3)$
$\mathbf{y}^{\mathbf{d_y}}_{(\mathbf{t-1})}(R)$
$\mathbf{y}^{\mathbf{d_y}}_{(\mathbf{t-1})}(1)$
$\mathbf{y}^{\mathbf{d_y}}_{(\mathbf{t-1})}(2)$
$\mathbf{y}^{\mathbf{d_y}}_{(\mathbf{t-1})}(3)$
...

Fig. 2.3 **Creation of surrogate data sets.** (A) Original time series with information transfer (solid arrow) from a source state $\mathbf{x}^{\mathbf{d_x}}_{(\mathbf{t-u})}(r)$ to a corresponding target time point $y_t(r)$, given the time point's history $\mathbf{y}^{\mathbf{d_y}}_{(\mathbf{t-1})}(r)$. Solid arrows indicate the direction of transfer entropy (TE) analysis, while information transfer is present. (B) Shuffled target time series, repetitions are permutes, such that $y_t(\phi(r))$ and $\mathbf{y}^{\mathbf{d_y}}_{(\mathbf{t-1})}(\phi(r))$, where $\phi$ denotes a random permutation. Dashed arrows indicate the direction of TE analysis, while no more information flow is present.

**Reconstruction of information transfer delays.** $TE_{SPO}\left(X \rightarrow Y, t, u\right)$ may be used to reconstruct the interaction transfer delay $\delta_{XY}$ between $X$ and $Y$ (Eq. 2.5, [75]). $\delta_{XY}$ may be reconstructed by *scanning* possible values for $u$: $TE_{SPO}\left(X \rightarrow Y, t, u\right)$ is estimated for all values in $u$; The value that maximizes the $TE_{SPO}\left(X \rightarrow Y, t, u\right)$ is kept as the reconstructed information transfer delay. We used the reconstruction of information transfer delays as an additional parameter when testing the proposed implementation for correctness and robustness.

## 2.3.2 Implementation of the GPU algorithm

**Parallelized nearest neighbor searches.** The KSG estimator used for estimating $TE_{SPO}\left(X \rightarrow Y, t, u\right)$ in Eq. 2.9 and 2.10 uses neighbor (distance-)statistics obtained from kNNS and RS algorithms to estimate Shannon differential entropies. Thus, the choice of computationally efficient kNNS and RS algorithms is crucial to any practical implementation of the $TE_{SPO}$ estimator. kNNS algorithms typically return a list of the k nearest neighbors for each reference point, while RS algorithms typically return a list of all neighbors within a given range for each reference point. kNNS and RS algorithms have been studied extensively because of their broad potential for application in nearest neighbor searches and related problems. Several approaches have been proposed to reduce their high computational cost: partitioning of input data into k-d Trees, Quadtrees or equivalent data structures [139] or approximation algorithms (ANN: Approximate Nearest Neighbors) [140, 141]. Furthermore, some authors have explored how to parallelize the kNNS algorithm on a GPU using different implementations: exhaustive brute force searches [142, 143], tree-based searches [144, 145] and ANN searches [145, 146].

Although performance of existing implementations of kNNS for GPU was promising, they were not applicable to TE estimation. The most critical reason was that existing implementations did not allow for the concurrent treatment of several problem instances by the GPU and maximum performance was only achieved for very large kNNS problem instances. Unfortunately, the problem instances typically expected in our application are numerous (i.e. $S + 1$ problem instances per pair of time series), but rather small compared to the main memory on a typical GPU device in use today. Thus, an implementation that handled only one instance at a time would not have made optimal use of the underlying hardware. Therefore, we designed an implementation that is able to handle several problem instances at once to perform neighbor searches for chunks of embedded original and surrogate data in parallel. Moreover, we aimed at a flexible GPU implementation of kNNS and RS that maximized the use of the GPU's hardware resources for variable configurations of data—thus making the implementation independent of the design of the neuroscientific experiment.

Our implementation is written in CUDA (Compute Unified Device Architecture) [136] (a port to OpenCL™ [147] is work in progress). CUDA is a parallel computing framework created by NVIDIA that includes extensions to high level languages such as C/C++, giving access to the native instruction set and memory of the parallel computational elements in CUDA enabled GPUs. Accelerating an algorithm using CUDA includes translating it into data-parallel sequences of operations and then carefully mapping these operations to the underlying resources to get maximum performance [122, 123]. To understand the implementation suggested here, we will give a brief explanation of these resources, i.e. the GPU's hardware architecture, before explaining the implementation in more detail (additionally, see [122, 123, 136]).

**GPU resources.** GPU resources comprise of massively parallel processors with up to thousands of cores (processing units). These cores are divided among Stream Multiprocessors (SMs) in order to guarantee automatic scalability of the algorithms to different versions of the hardware. Each SM contains 32 to 192 cores that execute operations described in the CUDA kernel code. Operations executed by one core are called a CUDA thread. Threads are grouped in blocks, which are in turn organized in a grid. The grid is the entry point to the GPU resources. It handles one kernel call at a time and executes it on multiple data in parallel. Within the grid, each block of threads is executed by one SM. The SM executes the threads of a block by issuing them in groups of 32 threads, called warps. Threads within one warp are executed concurrently, while as many warps as possible are scheduled per SM to be resident at a time, such that the utilization of all the cores is maximized.

**Input data.** As input, the proposed RS and kNNS algorithms expect a set of data points representing the search space and a second set of data points that serve as reference points in the searches. One such problem instance is considered one data chunk. Our implementation is able to handle several data chunks simultaneously to make maximum use of the GPU resources. Thus, several chunks may be combined, using an additional index vector to encode the sizes of individual chunks. These chunks are then passed at once to the GPU algorithm to be searched in parallel.

In the estimation of $TE_{SPO}\left(\mathcal{X} \rightarrow \mathcal{Y}, t, u\right)$, according to the work flow described in paragraph *Practical TE estimation in TRENTOOL*, we used the proposed implementation to parallelize neighbor searches over surrogate data sets for a given pair of time series $\mathbf{x}$ and $\mathbf{y}$ and given values for $u$ and $t$. Thus, in one call to the GPU algorithms $S+1$ data chunks were passed as input, where chunks represented the search space for the original pair of time series and $S$ search spaces for corresponding surrogate data sets. Points within the search spaces may have either been collected through temporal or ensemble pooling of embedded data points or a combination of both (Eq. 2.9 or 2.10).

**Core algorithm.** In the core GPU-based search algorithm, the kNNS implementation is mapped to CUDA threads as depicted in Fig. 2.4 (the RS implementation behaves similarly). Each chunk consists of a set of data points that represents the search space and are at the same time used as reference points for individual searches. Each individual search is handled by one CUDA thread. Parallelization of these searches on the GPU happens in two ways: (1) the GPU algorithm is able to handle several chunks, (2) each chunk can be searched in parallel, such that individual searches within one chunk are handled simultaneously. An individual search is conducted by a CUDA thread by brute-force measuring the infinity norm distance of the given reference point to any other point within the same chunk. Simultaneously, other threads measure these distances for other points in the same chunk or handle a different chunk altogether. Searching several chunks in parallel is an essential feature of the proposed solution, that maximizes the utilization of GPU resources. From the GPU execution point of view, simultaneous searches are realized by handling a variable number of kNNS (or RS) problem instances through one grid launch. The number of searches that can be executed in parallel is thus only limited by the device's global memory that holds the input data and the number of threads that can be started simultaneously (both limitations are taken into account). Furthermore, the solution is implemented such that optimal performance is guaranteed.

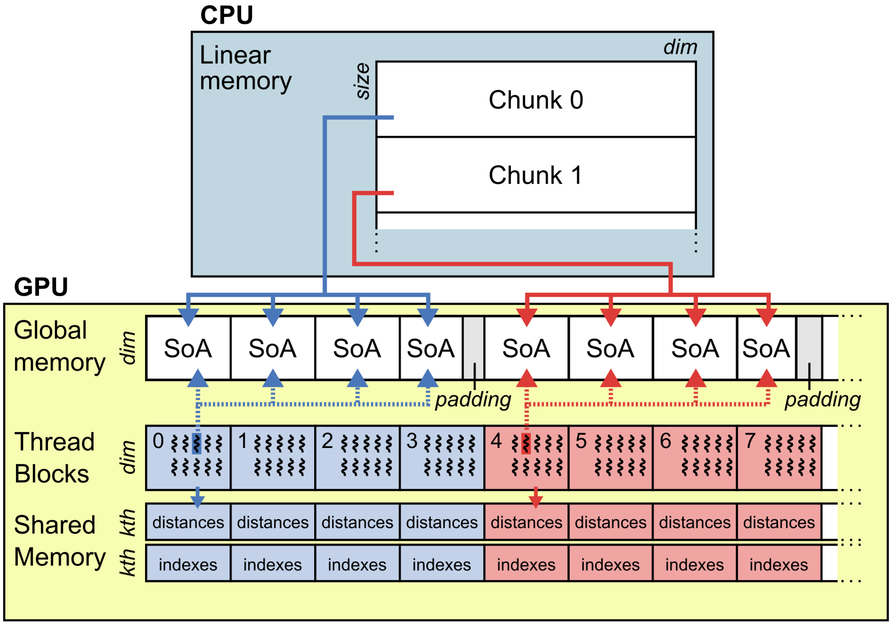


Fig. 2.4 **GPU implementation of the parallelized nearest neighbor search in TRENTOOL 3.0.** Chunks of data are prepared on the CPU (embedding and concatenation) and passed to the GPU. Data points are managed in the global memory as Structures of Arrays (SoA). To make maximum use of the memory bandwidth, data is padded to ensure coalesced reading and writing from and to the streaming multiprocessor (SM) units. Each SM handles one chunk in one thread block (dashed box). One block conducts brute force neighbor searches for all data points in the chunk and collects results in its shared memory (red and blue arrows and shaded areas). Results are eventually returned to the CPU.

**Low-level implementation details.** There are several strategies that are essential for optimal performance when implementing algorithms for GPU devices. Most important are the reduction of memory latencies and the optimal use of hardware resources by ensuring high occupancy (the ratio of number of active warps per SM to the maximum number of possible active warps [122]). To maximize occupancy, we designed our algorithm's kernels such that always more than one block of threads (ideally many) are loaded per SM [122]. We can do this since many searches are executed concurrently in every kernel launch. By maximizing occupancy, we both ensure hardware utilization and improve performance by hiding data memory latency from the GPU's global memory to the SMs' registers [136]. Moreover, in order to reduce memory latencies we take care of input data memory alignment and guarantee that memory readings issued by the threads of a warp are coalesced into as few memory transfers as possible. Additionally, with the aim of minimizing sparse data accesses to memory, data points are organized as Structures of Arrays (SoA).

Finally, we use the shared memory inside the SMs (a self-programmed intermediate cache between global memory and SMs) to keep track of nearest neighbors associated information during searches. The amount of shared memory and registers is limited in a SM. The maximum possible occupancy depends on the number of registers and shared memory needed by a block, which in turn depends on the number of threads in the block. For our implementation, we used a suitable block size of 512 threads.

**Implementation interface.** The GPU functionality is accessed through MATLAB scripts for kNNS (‘`fnearneigh_gpu.mex`’) and RS (‘`range_search_all_gpu.mex`’), which encapsulate all the associated complexity. Both scripts are called from TRENTOOL using a wrapper function. In its current implementation in TRENTOOL (see paragraph *Practical TE estimation in TRENTOOL*), the wrapper function takes all $S + 1$ chunks as input and launches a kernel that searches all chunks in parallel through the mex-files for kNNS and RS. The wrapper makes sure that the input size does not exceed the GPU device's available global memory and the maximum number of threads that can be started simultaneously. If necessary, the wrapper function splits the input into several kernel calls; it also manages the output, i.e. the neighbor counts for each chunk, which are passed on for TE calculation.

## 2.4 Evaluation

To evaluate the proposed algorithm we investigated four properties: first, whether the speedup is sufficient to allow the application of the method to real-world neural datasets; second, the correctness of results on simulated data, where the ground truth is known; third, the robustness of the algorithm for limited sample sizes; fourth, whether plausible results are achieved on a neural example dataset.

### 2.4.1 Ethics statement

The neural example dataset was taken from an experiment described in [148]. All subjects gave written informed consent before the experiment. The study was approved by the local ethics committee (Johann Wolfgang Goethe University, Frankfurt, Germany).

### 2.4.2 Evaluation of computational speedup

To test for an increase in performance due to the parallelization of neighbor searches, we compared practical execution times of the proposed GPU implementation to execution times of the serial kNNS and RS algorithms implemented in the MATLAB toolbox TSTOOL (`http://www.dpi.physik.uni-goettingen.de/tstool/`). This toolbox wraps a FORTRAN implementation of kNNS and RS, and has proven the fastest CPU toolbox for our purpose. All testing was done in MATLAB 2008b (MATLAB 7.7, The MathWorks Inc., Natick, MA, 2008). As input, we used increasing numbers of chunks of simulated data from two coupled Lorenz systems, further described below. Repetitions of simulated time series were embedded and combined to form ensemble state spaces, i.e. chunks of data (c.f. paragraph *Input Data*). To obtain increasing input sizes, we duplicated these chunks the desired number of times. While the CPU implementation needed to iteratively perform searches on individual chunks, the GPU implementation searched chunks in parallel (note that chunks are treated independently here, so that there is no speedup because of the duplicated chunk data). Note that for both, CPU and GPU implementations, data handling prior to nearest neighbor searches is identical. We were thus able to confine the testing of performance differences to the respective kNNS and RS algorithms only, as all data handling prior to nearest neighbor searches was conducted using the same, highly optimized TRENTOOL functionalities.

Analogous to TE estimation implemented in TRENTOOL, we conducted one kNNS (with $k = 4$, TRENTOOL default, see also [85]) in the highest dimensional space and used the returned distances for a RS in one lower dimensional space. Both functions were called for increasing numbers of chunks to obtain the execution time as a function of input size. One chunk of data from the highest dimensional space had dimensions [30 094×17] and size 1.952 MB (single precision); one chunk of data from the lower dimensional space had dimensions [30 094×8] and size 0.918 MB (single precision). Performance testing of the serial implementation was carried out on an Intel Xeon CPU (E5540, clocked at 2.53 GHz), where we measured execution times of the TSTOOL kNNS (functions '`nn_prepare.m`' and '`nn_search.m`') and the TSTOOL RS (function '`range_search.m`'). Testing of the parallel implementation was carried out three times on GPU devices of varying processing power (NVIDIA Tesla C2075, GeForce GTX 580 and GeForce GTX Titan). On the GPUs, we measured execution times for the proposed kNNS ('`fnearneigh_gpu.mex`') and RS ('`range_search_all_gpu.mex`') implementation. When the GPU's global memory capacity was exceeded by higher input sizes, data was split and computed over several runs (i.e. calls to the GPU). All performance testing was done by measuring execution times using the MATLAB functions `tic` and `toc`.

To obtain reliable results for the serial implementation we ran both kNNS and RS 200 times on the data, receiving an average execution time of 1.26 s for kNNS and an average execution time of 24.1 s for RS. We extrapolated these execution times to higher numbers of chunks and compared them to measured execution times of the parallel searches on three NVIDIA GPU devices. On average, execution times on the GPU compared to the CPU were faster by a factor of 22 on the NVIDIA Tesla C2075, by a factor of 33 for the NVIDIA GTX 580 and by a factor of 50 for the NVIDIA GTX Titan (Fig. 2.5).

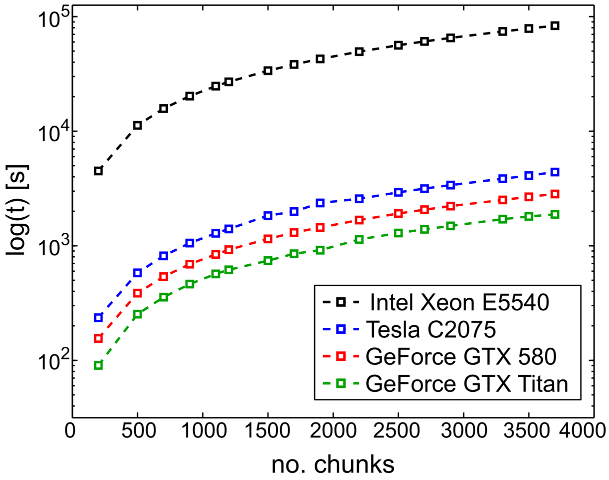


Fig. 2.5 **Practical performance measures of the ensemble method for GPU compared to CPU.** Combined execution times in s for serial and parallel implementations of k-nearest neighbor and range search as a function of input size (number of data chunks). Execution times were measured for the serial implementation running on a CPU (black) and for our parallel implementation using one of three GPU devices (blue, red, green) of varying computing power. Computation using a GPU was considerably faster than using a CPU (by factors 22, 33, and 50 respectively).

To put these numbers into perspective, we note that in a neuroscience experiment the number of chunks to be processed is the product of (typical numbers): channel pairs for TE (100) × number of surrogate data sets (1000) × experimental conditions (4) × number of subjects (15). This results in a total computational load on the order of $6 \times 10^6$ s chunks to be processed. Given an execution time of 24.1 s/50 on the NVIDIA GTX Titan for a typical test dataset, these computations will take $2.9 \times 10^6$ s or 4.8 weeks on a single GPU, which is feasible compared to the initial duration of 240 weeks on a single CPU. Even when considering a trivial parallelization of the computations over multiple CPU cores and CPUs, the GPU based solution is by far more cost and energy efficient than any possible CPU-based solution. If in addition a scanning of various possible information transfer delays is important, then parallelization over multiple GPUs seems to be the only viable option.

### 2.4.3 Evaluation on Lorenz systems

To test the ability of the presented implementation to successfully reconstruct information transfer between systems with a non-stationary coupling, we simulated various coupling scenarios between stochastic and deterministic systems. We introduced non-stationary into the coupling of two processes by varying the coupling strength over the course of a repetition (all other parameters were held constant). Simulations for individual scenarios are described in detail below. For the estimation of TE we used MathWork's MATLAB, and the TRENTOOL toolbox extended by the implementation of the ensemble method proposed above (version 3.0, see also [86] and `http://www.trentool.de`). For a detailed testing of the used estimator $TE_{SPO}$ (Eq. 2.4) refer to [75].

**Coupled Lorenz systems.** Simulated data was taken from two unidirectionally coupled Lorenz systems labeled $X$ and $Y$. Systems interacted in direction $X \rightarrow Y$ according to equations:

$$\begin{aligned}
\dot{U}_i(t) &= \sigma(V_i(t) - U_i(t)), \\
\dot{V}_i(t) &= U_i(t)(\rho_i - W_i(t)) - V_i(t) + \sum_{i,j=X,Y} \gamma_{ij} V_j^2(t - \delta_{ij}), \\
\dot{W}_i(t) &= U_i(t)V_i(t) - \beta W_i(t)\,,
\end{aligned} \tag{2.11}$$

where $i, j = X, Y$, $\delta_{ij}$ is the coupling delay and $\gamma_{ij}$ is the coupling strength; $\sigma$, $\rho$ and $\beta$ are the *Prandtl number*, the *Rayleigh number*, and a geometrical scale. Note, that $\gamma_{YX} = \gamma_{XX} = \gamma_{YY} = 0$ for the test cases (no self feedback, no coupling from $Y$ to $X$). Numerical solutions to these differential equations were computed using the *dde23* solver in MATLAB and results were resampled such that the delays amounted to the values given below. For analysis purposes we analyzed the V-coordinates of the systems.

We introduced non-stationarity in the coupling between both systems by varying the coupling strength $\gamma$ over time. In particular, a coupling $\gamma_{XY} = 0.3$ was set for a limited time interval only, whereas before and after the coupling interval $\gamma_{XY}$ was set to $0$. A constant information transfer delay $\delta_{XY} = 45ms$ was simulated for the whole coupling interval. We simulated 150 repetitions with 3000 data points each, with a coupling interval from approximately 1000 to 2000 data points (see Fig. 2.6A).

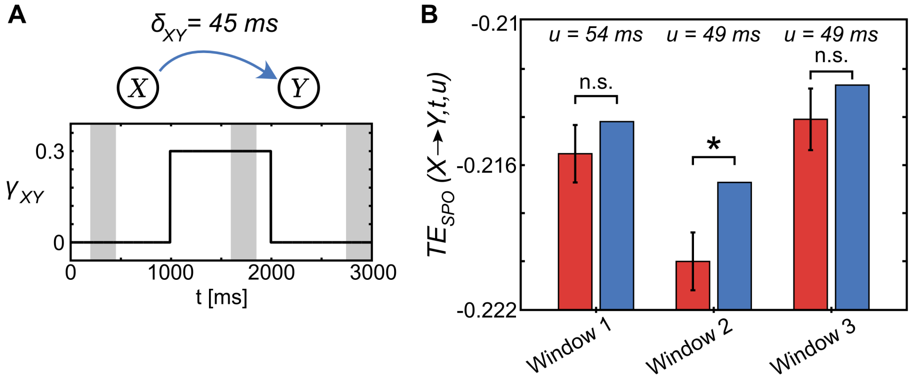


Fig. 2.6 **Transfer entropy reconstruction from non-stationary Lorenz systems.** We used two dynamically coupled Lorenz systems (A) to simulate non-stationarity in data generating processes. A coupling $\gamma_{XY} = 0.3$ was present during a time interval from 1000 ms to 2000 ms only ($\gamma_{XY} = 0$ otherwise). The information transfer delay was set to $\delta_{XY} = 45ms$. Transfer entropy (TE) values were reconstructed using the ensemble method combined with the scanning approach proposed in [75] to reconstruct information transfer delays. Assumed delays $u$ were scanned from 35 ms to 55 ms (1 ms resolution). In (B) the maximum TE values for original data over this interval are shown in blue. Red bars indicate the corresponding mean over surrogate TE values (error bars indicate 1 SD). Significant TE was found for the second time window only; here, the delay was reconstructed as $u = 49ms$.

For each scenario, 500 surrogate data sets were computed to allow for statistical testing of the reconstructed information transfer. Surrogate data were created by permutation of data points in blocks of the target time series (Fig. 2.3), leaving each repetition intact. The value $k$ for the nearest neighbor search was set to 4 for all analyses (TRENTOOL default, see also [85]).

**Results.** We analyzed data from three time windows from 200 ms to 450 ms, 1600 ms to 1850 ms and 2750 ms to 3000 ms using the estimator proposed in Eq. 2.10 with $\Delta_t = 250ms$, assuming local stationarity (Fig. 2.6A). For each time window, we scanned assumed delays in the interval $u = [35, 55]$. Fig. 2.6B, shows the maximum TE value from original data (blue) over all assumed $u$ and the corresponding mean surrogate TE value (red). Significant differences between original TE and surrogate TE were found in the second time window only (indicated by an asterisk). No significant information transfer was found during the non-coupling intervals. The information transfer delay reconstructed for the second analysis window was 49 ms (true information transfer delay $\delta_{XY} = 45ms$). Thus, the proposed implementation was able to reliably detect a coupling between both systems and reconstructed the corresponding information transfer delay with an error of less than 10 %.

### 2.4.4 Evaluation on autoregressive processes

To asses the performance of the proposed implementation on non-abrupt changes in coupling, we simulated various coupling scenarios for two autoregressive processes $X$, $Y$ of order 1 (AR(1)-processes) with variable couplings over time. In each scenario, couplings were modulated using hyperbolic functions to realize a smooth transition between uncoupled and coupled regimes. The AR(1)-processes were simulated according to the equations

$$x(t) = \alpha_X x(t-1) + \gamma_{YX}(t) y(t - \delta_{YX}) + \eta_X(t), \tag{2.12}$$

$$y(t) = \alpha_Y y(t-1) + \gamma_{XY}(t) x(t - \delta_{XY}) + \eta_Y(t), \tag{2.13}$$

where $\alpha_X$, $\alpha_Y$ are the AR parameters, $\gamma_{YX}(t)$, $\gamma_{XY}(t)$ denote coupling strength, $\delta_{YX}$, $\delta_{XY}$ are the coupling delays and $\eta_X$, $\eta_Y$ denote uncorrelated, unit-variance, zero-mean Gaussian white noise terms.

**Simulated coupling scenarios.** We simulated three coupling scenarios, where the coupling varied in strength over the course of a repetition (duration 3000 ms): (1) unidirectional coupling $X \rightarrow Y$ with a coupling onset around 1000 ms; (2) unidirectional coupling with a two-step increase in coupling $X \rightarrow Y$ at around 1000 ms and around 2000 ms; (3) bidirectional coupling $X \rightarrow Y$ with onset around 1000 ms and $Y \rightarrow X$ with onset around 2000 ms. See Table 2.1 for specific parameter values used in each scenario.

Tab. 2.1 **Parameter settings for simulated autoregressive processes.**

| Testcase | $\alpha_X$ | $\alpha_Y$ | $\beta_{YX}$ | $\beta_{XY}$ | $\delta_{YX}$ | $\delta_{XY}$ |
|---|---|---|---|---|---|---|
| Unidirectional | 0.75 | 0.35 | 0 | -0.35 | 0 | 10 |
| Two-step unidirectional | 0.75 | 0.35 | 0 | -0.35 | 0 | 10 |
| Bidirectional | 0.475 | 0.35 | -0.4 | -0.35 | 20 | 10 |

We realized a varying coupling strength $\gamma_{XY}(t)$ (and $\gamma_{YX}(t)$ for scenario (3)) by modulating coupling parameters $\beta_{YX}$, $\beta_{XY}$ with a hyperbolic tangent function. No coupling was realized by setting $\beta_{\cdot} = 0$. For scenarios (1) and (3) we used the coupling

$$\gamma_{YX} = \beta_{YX} * 0.5\left(1 + \tanh\left[0.05(t - 2000)\right]\right) \tag{2.14}$$

$$\gamma_{XY} = \beta_{XY} * 0.5\left(1 + \tanh\left[0.05(t - 1000)\right]\right), \tag{2.15}$$

where 0.05 was the slope and 2000 and 1000 are the inflection points of the hyperbolic tangent respectively. Note that we additionally scaled the $\tanh$ function such that function value ranged from 0 to 1. For coupling scenario (2), the two-step increase in $\gamma_{XY}$ was expressed as:

$$\begin{aligned} \gamma_{XY} = \beta_{XY} * 0.5[&0.5\left(1 + \tanh\left[0.05(t - 1000)\right]\right) \\ &+0.5\left(1 + \tanh\left[0.05(t - 2000)\right]\right)]. \end{aligned} \tag{2.16}$$

We chose the arguments of the hyperbolic function such that the function's slope led to a smooth increase in the coupling over an epoch of approximately 200 ms around the inflection points at 1 and 2 s respectively (Fig. 2.7A–D). For each scenario, we simulated 50 trials of length 3000 ms with a sampling rate of 1000 Hz. We then estimated time resolved TE for analysis windows of length $\Delta_t = 300\ ms$. Again, we mixed temporal and ensemble pooling according to Eq. 2.10. For the scenario with unidirectional coupling (1) we used four analysis windows to cover the change in coupling (from 0.2 s to 0.5 s, 0.5 s to 0.8 s, 0.8 s to 1.1 s, and 1.1 s to 1.4 s, see Fig. 2.7E), for the two-step increase (2) and bidirectional (3) scenarios, we used eight analysis windows each (from 0.2 s to 0.5 s, 0.5 s to 0.8 s, 0.8 s to 1.1 s, 1.1 s to 1.4 s, 1.4 s to 1.7 s, 1.7 s to 2.0 s, 2.0 s to 2.3 s, and 2.3 s to 2.6 s, see Fig. 2.7F–G). As for the Lorenz systems, 500 surrogate data sets were used for the statistical testing in each analysis. Surrogate data were created by blockwise (i.e. repetitionwise) permutation of data points in the target time series. The value $k$ for the nearest neighbor search was set to 4 for all analyses (TRENTOOL default, see also [85]).

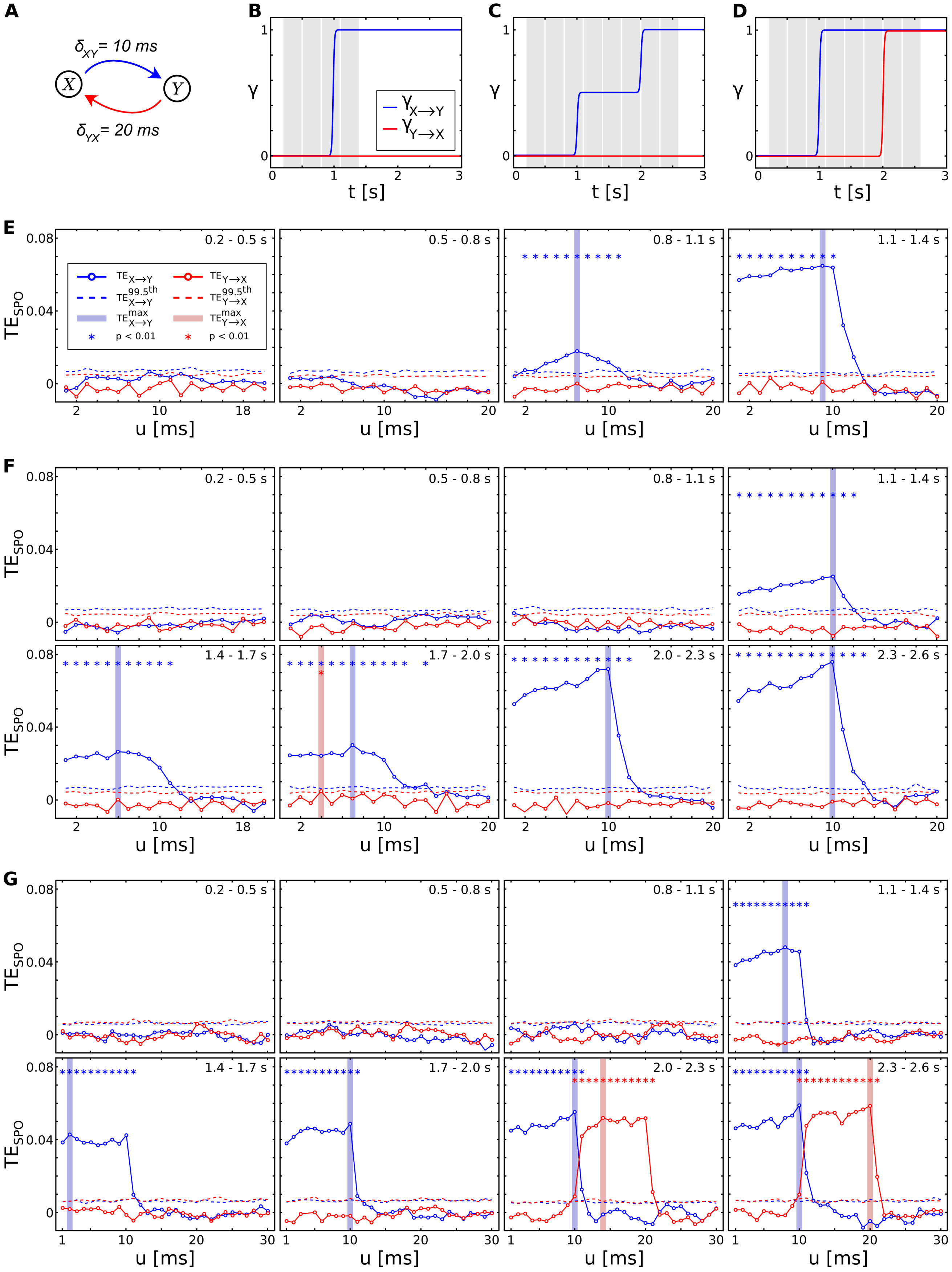


**Fig. 2.7** **Transfer entropy reconstruction from coupled autoregressive processes.** We simulated two dynamically coupled autoregressive processes (A) with coupling delays $\delta_{XY} = 10ms$ and $\delta_{YX} = 20ms$, and coupling scenarios:

Fig. 2.7 **Transfer entropy reconstruction from coupled autoregressive processes (continued).** (B) unidirectional coupling $X \rightarrow Y$ (blue line) with onset around 1 s, coupling $Y \rightarrow X$ set to 0 (red line); (C) unidirectional coupling $X \rightarrow Y$ (blue line) with onset around 1 s and an increase in coupling strength at around 2 s, coupling $Y \rightarrow X$ set to 0 (red line); (D) bidirectional coupling $X \rightarrow Y$ (blue line) with onset around 1 s and $Y \rightarrow X$ (red line) with onset around 2 s. (E-G) Time-resolved transfer entropy (TE) for both directions of interaction, blue and red lines indicate raw TE values for $X \rightarrow Y$ and $Y \rightarrow X$ respectively. Dashed lines denote significance thresholds at 0.01 % (corrected for multiple comparisons over signal combinations). Shaded areas (red and blue) indicate the maximum absolute TE values for significant information transfer (indicated by asterisks in red and blue). (E) TE values for unidirectional coupling; (F) unidirectional coupling with a two-step increase in coupling strength; (G) bidirectional coupling.

**Results – Scenario (1), unidirectional coupling.** For scenario (1) of two unidirectionally coupled AR(1)-processes with a delay $\delta_{XY} = 10ms$, we used a scanning approach [75] to reconstruct TE and the corresponding information transfer delay. We scanned assumed delays in the interval $u = [1, 20]$ and used four analysis windows of length 300 ms each, ranging from 0.2 s to 1.4 s. For the first two analysis windows, no significant information transfer was found (0.2 s to 0.5 s and 0.5 s to 0.8 s). For the third and fourth analysis window we detected significant TE, where we found a maximum significant TE value at 7 ms for the third analysis window (0.8 s to 1.1 s) and a maximum at 9 ms for the fourth window (1.1 s to 1.4 s). Thus, the proposed implementation was able to detect information transfer between both processes if present (later than 1.1 s). During the transition in coupling strength between 0.8 and 1.1 s TE was detected, but the method showed a small error in the reconstructed information transfer delay. This may be due to too little data to detect the weaker coupling at this epoch of the simulated coupling (see below).

**Results – Scenario (2), unidirectional coupling with two-step increase.** For scenario (2), we again used the scanning approach for TE reconstruction, using an interval of assumed delays $u = [1, 20]$, where the true delay was simulated at $\delta_{XY} = 10ms$. No TE was detected prior to the coupling onset around 1 s. TE was detected for analysis windows 4, 5, and 6 (1.1 s to 1.4 s, 1.4 s to 1.7 s, 1.7 s to 2.0 s) with reconstructed information transfer delays of 10, 4, and 7 ms respectively. Further, significant TE was found for analysis windows 7 and 8 (after the second increase in coupling strength around 2 s). Here, the correct coupling of 10 ms was reconstructed. One false positive result was obtained in window 6 (1.7 s to 2.0 s), where significant TE was found in the direction $Y \rightarrow X$.

Note, that the method's ability to recover information transfer from data depends on the strength of the coupling relative to the amount of data that is available for TE estimation. This is observable in the reconstructed TE in the third analysis window for scenario (1) and (2): in scenario (2) no TE is detected, whereas in scenario (1) weak information transfer is already reconstructed for the third window. Note, that in scenario (2) the simulated coupling between  1 and 2 s is much weaker than the coupling in the unidirectional scenario (1) (Fig. 2.7B–C). This resulted in smaller and non-significant absolute TE values and in reconstructed information transfer delays that were less precise.

**Results – Scenario (3), bidirectional coupling.** For scenario $(3)$, we used the scanning approach for TE reconstruction, using an interval of assumed delays $u = [1, 30]$, where the true delay was simulated at $\delta_{XY} = 10ms$ and $\delta_{YX} = 20ms$. No TE in either direction was detected prior to the first coupling onset around 1 s. TE for the first direction $X \rightarrow Y$ was detected after coupling onset around 1 s for analysis windows 4, 5, 6, 7, and 8. Reconstructed information transfer delays were 8 and 2 ms for analysis windows 4 and 5. For each of the following analysis windows 6 to 8 the correct delay of 10 ms was reconstructed.

TE for the second direction $Y \rightarrow X$ was detected after coupling onset around 2 s for analysis windows 7 and 8, where also the correct coupling of 20 ms was reconstructed. Thus, the proposed implementation was able to reconstruct information transfer in bidirectionally coupled systems.

### 2.4.5 Evaluation of the robustness of ensemble-based TE-estimation

We tested the robustness of the ensemble method for cases where the amount of data available for TE estimation was severely limited. We created two coupled Lorenz systems $X$, $Y$ from which we sampled a maximum number of 300 repetitions of 300 ms each at 1000 Hz, using a coupling delay of $\delta_{XY} = 45ms$ (see Eq. 2.11). We embedded the resulting data with their optimal embedding parameters for different values of the assumed delay $u$ (30 ms to 60 ms, step size of 1 ms, also see Eq. 2.4). From the embedded data, we used subsets of data points with varying size $M$ ($M = \{500, 2000, 5000, 10000, 30000\}$) to estimate TE according to Eq. 2.10 (we always used the first $M$ consecutive data points for TE estimation). For each $u$ and number of data points $M$, we created surrogate data to test the estimated TE value for statistical significance. Furthermore, we reconstructed the corresponding information transfer delay for each $M$ by finding the maximum TE value over all values for $u$. A reconstructed TE value was considered a robust estimation of the simulated coupling if the reconstructed delay value was able to recover the simulated information transfer delay of 45 ms with an error of ±5 %, i.e. 45±1.125 ms.

A sufficiently accurate reconstruction was reached for 10 000 and 30 000 data points (Fig. 2.8). For 5000 data points estimation was off by approximately 7 % (the reconstructed information transfer delay was 48 ms), less data entering the estimation led to a further decline in accuracy of the recovered information transfer delay (here, reconstructed delays were 50 ms and 54 ms for 2000 and 500 data points respectively).

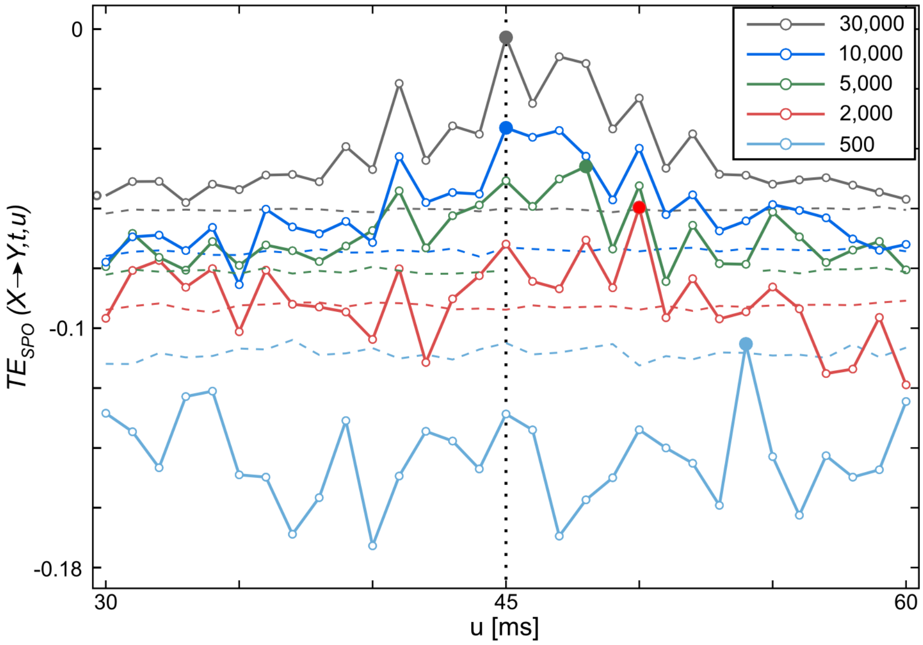


**Fig. 2.8** **Robustness of transfer entropy estimation with respect to limited amounts of data.** Estimated transfer entropy (TE) values $TE_{X\rightarrow Y}$ for estimations using varying numbers of data points (color coded) as a function of $u$. Data was sampled from two Lorenz systems *X* and *Y* with coupling $X \rightarrow Y$. The simulated information transfer delay $\delta_{XY} = 45ms$ is indicated by a vertical dotted line. Sampled data was embedded and varying numbers of embedded data points (500, 2000, 5000, 10 000, 30 000) were used for TE estimation. For each estimation, the maximum $TE_{X\rightarrow Y}$ values for all values of $u$ are indicated by solid dots. Dashed lines indicate significance thresholds ($p < 0.05$).

### 2.4.6 Evaluation on neural time series from magnetoencephalography

To demonstrate the proposed method's suitability for time-resolved reconstruction of information transfer and the corresponding delays from biological time series, we analyzed magnetoencephalographic (MEG) recordings from a perceptual closure experiment described in [148].

**Subjects.** MEG data were obtained from 15 healthy subjects (11 females; mean ± SD age, 25.4 ±5.6 years), recruited from the local community.

**Task.** Subjects were presented with a randomized sequence of degraded black and white picture of human faces [149] (Fig. 2.9A) and scrambled stimuli, where black and white patches were randomly rearranged to minimize the likelihood of detecting a face. Subjects had to indicate the detection of a face or no-face by a button press. Each stimulus was presented for 200 ms, with a random inter-repetition interval (IRI) of 3500 ms to 4500 ms (2.9E). For further analysis we used repetitions with correctly identified face conditions only.

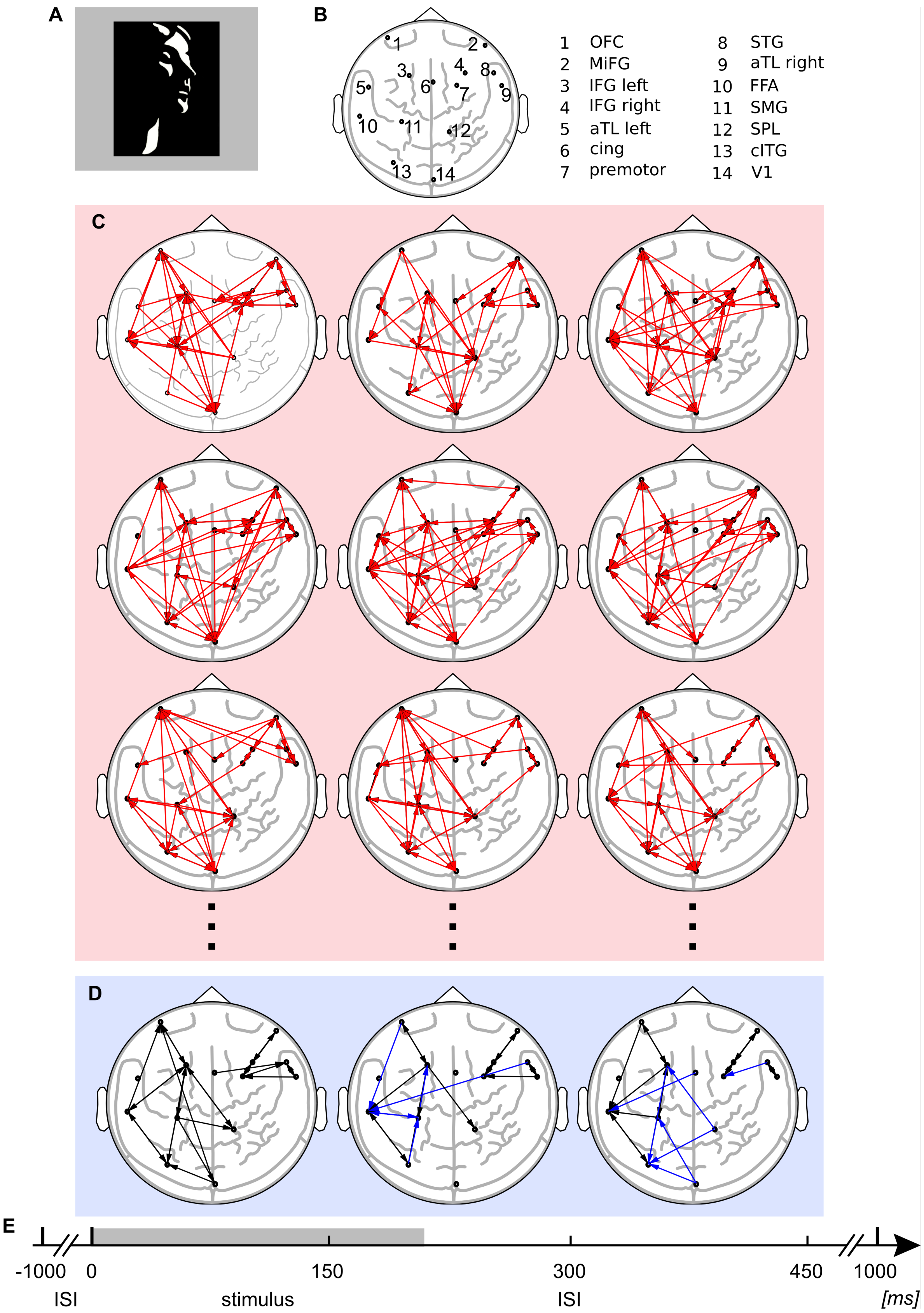


Fig. 2.9 **Transfer entropy reconstruction from electrophysiological data.** Time resolved reconstruction of transfer entropy (TE) from magnetoencephalographic (MEG) source data, recorded during a face recognition task. (A) Face stimulus [149].

Fig. 2.9 **Transfer entropy reconstruction from electrophysiological data (continued).** (B) Cortical sources after beamforming of MEG data (L, left; R, right: L orbitofrontal cortex (OFC); R middle frontal gyrus (MiFG); L inferior frontal gyrus (IFG left); R inferior frontal gyrus (IFG right); L anterior inferotemporal cortex (aTL left); L cingulate gyrus (cing); R premotor cortex (premotor); R superior temporal gyrus (STG); R anterior inferotemporal cortex (aTL right); L fusiform gyrus (FFA); L angular/supramarginal gyrus (SMG); R superior parietal lobule/precuneus (SPL); L caudal ITG/LOC (cITG); R primary visual cortex (V1)). (C) Reconstructed TE in three single subjects (red box) in three time windows (0 ms to 150 ms, 150 ms to 300 ms, 300 ms to 450 ms). Each link (red arrows) corresponds to significant TE on single subject level (corrected for multiple comparisons). (D) Thresholded TE links over 15 subjects (blue box) in three time windows (0 ms to 150 ms, 150 ms to 300 ms, 300 ms to 450 ms). Each link (black arrows) corresponds to significant TE in eight and more individual subjects ($p << 0.0001^{***}$, after correction for multiple comparisons). Blue arrows indicate differences between time windows, i.e. links that occur for the first time in the respective window. (E) Experimental design: stimulus was presented for 200 ms (gray shading), during the inter stimulus interval (ISI, 1800 ms) a fixation cross was displayed.

**MEG and MRI data acquisition.** MEG data were recorded using a 275-channel whole-head system (Omega 2005, VSM MedTech Ltd., BC, Canada) at a rate of 600 Hz in a synthetic third order axial gradiometer configuration. The data were filtered with 4th order Butterworth filters with 0.5 Hz high-pass and 150 Hz low-pass. Behavioral responses were recorded using a fiber optic response pad (Lumitouch, Photon Control Inc., Burnaby, BC, Canada).

Structural magnetic resonance images (MRI) were obtained with a 3 T Siemens Allegra, using 3D magnetization-prepared rapid-acquisition gradient echo sequence. Anatomical images were used to create individual head models for MEG source reconstruction.

**Data analysis.** MEG data were analyzed using the open source MATLAB toolboxes FieldTrip (version 2008-12-08; [150]), SPM2 (`http://www.fil.ion.ucl.ac.uk/spm`), and TRENTOOL [86]. We will briefly describe the applied analysis here, for a more in depth treatment refer to [148].

For data preprocessing, data epochs (repetitions) were defined from the continuously recorded MEG signals from −1000 ms to 1000 ms with respect to the onset of the visual stimulus. Only data repetitions with correct responses were considered for analysis. Data epochs contaminated by eye blinks, muscle activity, or jump artifacts in the sensors were discarded. Data epochs were baseline corrected by subtracting the mean amplitude during an epoch ranging from −500 ms to −100 ms before stimulus onset.

To investigate differences in source activation in the face and non-face condition, we used a frequency domain beamformer [151] at frequencies of interest that had been identified at the sensor level (80 Hz with a spectral smoothing of 20 Hz). We computed the frequency domain beamformer filters for combined data epochs ("common filters") consisting of activation (multiple windows, duration, 200 ms; onsets at every 50 ms from 0 ms to 450 ms) and baseline data (−350 ms to −150 ms) for each analysis interval. To compensate for the short duration of the data windows, we used a regularization of $\lambda = 5\%$ [152].

To find significant source activations in the face versus non-face condition, we first conducted a within-subject t-test for activation versus baseline effects. Next, the t-values of this test statistic were subjected to a second-level randomization test at the group level to obtain effects of differences between face and no-face conditions; a p-value <0.01 was considered significant. We identified 14 sources with differential spectral power between both conditions in the frequency band of interest in occipital, parietal, temporal, and frontal cortices (see Fig. 2.9B, and [148] for exact anatomical locations). We then reconstructed source time courses for TE analysis, this time using a broadband beamformer with a bandwidth of 10 Hz to 150 Hz.

We estimated TE between beamformer source time courses using our ensemble method with a mixed pooling of embedded time points over repetitions $r$ and time windows $t'$ (Eq. 2.10). We analyzed three non-overlapping time windows $\Delta_t$ of 150 ms each (0 ms to 150 ms, 150 ms to 300 ms, 300 ms to 450 ms, Fig. 2.9C). We furthermore reconstructed information transfer delays for significant information transfer by scanning over a range of assumed delays from 5 ms to 17 ms (resolution 2 ms), following the approach in [75]. We corrected the resulting information transfer pattern for cascade effects as well as common drive effects using a graph-based post-hoc correction proposed in [121].

**Results.** Time-resolved GPU-based TE analysis revealed significant information transfer at the group-level ($p \ll 0.001$ corrected for multiple comparison; binomial test under the null hypothesis of the number of occurrences $k$ of a link being $B(k|p_0, n)$-distributed, where $p_0 = 0.05$ and $n = 15$), that changed over time (Fig. 2.9D and Table 2.2 for reconstructed information transfer delays). Our preliminary findings of information transfer are in line with hypothesis formulated in [153], [154] and [148], and the time-dependent changes show the our method's sensitivity to the dynamics of information processing during experimental stimulation, in line with the simulation results above.

Tab. 2.2 **Reconstructed information transfer delays for magnetoencephalographic data.** Mean over reconstructed interaction delays for significant information transfers in three analysis windows. Information transfer delays were investigated in steps of 2 ms, from 5 ms to 17 ms. Fractional numbers arise from averaging over subjects.

| Source | Target | 0 - 150 ms | 150 - 300 ms | 300 - 450 ms |
|---|---|---|---|---|
| SPL | IFG left | 5.00 | - | 5.50 |
| SPL | cITG | - | - | 5.00 |
| cITG | IFG left | 5.00 | 5.00 | 5.00 |
| cITG | FFA | 5.00 | 5.00 | 5.00 |
| cITG | SMG | - | 5.00 | - |
| STG | aTL right | 5.00 | 5.00 | 5.00 |
| STG | Premotor | - | - | 5.83 |
| STG | FFA | - | 5.50 | - |
| aTL right | STG | 5.00 | 5.00 | 5.00 |
| aTL right | Premotor | 5.00 | 5.60 | - |
| SMG | SPL | 5.00 | - | - |
| SMG | V1 | 5.00 | - | - |
| SMG | IFG left | - | 5.20 | - |
| SMG | FFA | - | 5.22 | 5.20 |
| OFC | IFG left | 5.18 | 5.00 | 5.00 |
| OFC | FFA | - | 5.00 | 5.20 |
| MiFG | IFG right | 5.00 | 5.00 | 5.00 |
| MiFG | Premotor | 5.00 | 5.00 | 5.00 |
| IFG right | MiFG | 5.00 | 5.00 | 5.00 |
| IFG right | Premotor | 5.00 | 5.00 | 5.00 |
| IFG left | SPL | 5.00 | 5.00 | - |
| IFG left | cITG | 5.00 | - | 5.00 |
| IFG left | SMG | 5.00 | 5.40 | 5.00 |
| IFG left | OFC | 5.00 | 5.00 | 5.00 |
| IFG left | FFA | 5.00 | 5.00 | 5.00 |
| FFA | cITG | 5.00 | 5.00 | 5.22 |
| FFA | OFC | 5.00 | - | - |
| FFA | IFG left | 5.00 | - | - |
| FFA | SMG | - | 5.00 | - |
| V1 | cITG | 5.25 | - | 5.25 |
| V1 | SMG | - | - | 5.00 |
| Premotor | STG | 5.00 | - | - |
| Premotor | MiFG | 5.00 | 5.00 | 5.00 |
| Premotor | IFG right | 5.00 | 5.00 | 5.00 |
| Cing | STG | 5.25 | - | - |
| Cing | FFA | - | - | 5.67 |

## 2.5 Discussion

### 2.5.1 Efficient transfer entropy estimation from an ensemble of time series

We presented an efficient implementation of the ensemble method for TE estimation proposed by [24]. As laid out in the introduction, estimating TE from an ensemble of data allows to analyze information transfer between time series that are non-stationary and enables the estimation of TE in a time-resolved fashion. This is especially relevant to neuroscientific experiments, where rapidly changing (and thus non-stationary) neural activity is believed to reflect neural information processing. However, up until now the ensemble method has remained out of reach for application in neuroscience because of its computational cost. Only with using parallelization on a GPU, as presented here, the ensemble method becomes a viable tool for the analysis of neural data. Thus, our approach makes it possible for the first time to efficiently analyze information transfer between neural time series on short time scales. This allows us to handle the non-stationarity of underlying processes and makes a time- resolved estimation of TE possible. To facilitate the use of the ensemble method it has been implemented as part of the open source toolbox TRENTOOL (version 3.0).

Even though we will focus on neural data when discussing applications of the ensemble method for TE estimation below, this approach is well suited for applications in other fields. For example, TE as defined in [12] has been applied in physiology [74, 114, 115], climatology [155, 156], financial time series analysis [116, 157], and in the theory of cellular automata [68]. Large datasets from these and other fields may now be easily analyzed with the presented approach and its implementation in TRENTOOL.

### 2.5.2 Notes on the practical application of the ensemble method for TE estimation

**Applicability to simulated and real world experimental data.** To validate the proposed implementation of the ensemble method, we applied it to simulated data as well as MEG recordings. For simulated data, information transfer could reliably be reconstructed despite the non-stationarity in the underlying generating processes. For MEG data the obtained speed-up was large enough to analyze these data in practical time. Information transfer reconstructed in a time-resolved fashion from the MEG source data was in line with findings by [148, 153, 154], as discussed below.

Note, that even though our proposed implementation of the ensemble method reduces analysis times by a significant amount, the estimation of TE from neural time series is still time consuming relative to other measures of connectivity. For the example MEG data set presented in this paper, TE estimation for one subject and one analysis window took 93 hours on average (when scanning over seven values for the assumed information transfer delay $u$ and reconstructing TE for all possible combinations of 14 sources). Thus, for 15 subjects with three analysis windows each, the whole analysis would take approximately six months when carried out in a serial fashion on one computer equipped with a modern GPU (e.g. NVIDIA GTX Titan). This time may however be reduced by parallelizing the analysis over subjects and analysis windows on multiple GPUs, as it was done for this study.

**Available data and choice of window size.** As available data is often limited in neuroscience and other real-world applications, the user has to make sure that enough data enters the analysis, such that a reliable estimation of TE is possible. In the proposed implementation of the ensemble method for TE estimation the amount of data entering the estimation directly depends on the size of the chosen analysis window and the amount of available repetitions of the process being analyzed. Furthermore, the choice of the embedding parameters lead to varying numbers of embedded data that can be obtained from scalar time series. When estimating TE from neural data, we therefore recommend to control the amount of data in one analysis window that is available after embedding and to design experiments accordingly. For example, the presented MEG data set was sampled at 600 Hz, with 137 repetitions of the stimulus on average, which - after embedding - led to 8800 data points per analysis window of 150 ms. In comparison, for simulated data TE was reconstructed correctly for 10 000 data points and more. Thus, in our example MEG data set, shorter analysis windows would not have been advisable because of an insufficient amount of data per analysis window for reliable TE estimation. If shorter analysis windows are necessary, they will have to be counterbalanced by a higher number of experimental repetitions.

Thus, the choice of an appropriate analysis window is crucial to guarantee reliable TE estimation, while still resolving the temporal dynamics under investigation. A further data limiting factor is the need for an appropriate embedding of the scalar time series. To embed the time series at a given point $t$, enough history for this sample (embedding dimension times the embedding delay in sample points) has to be recorded. We call this epoch the *embedding window*. The need for an appropriate embedding thus constitutes another constraint for the data necessary for

TE estimation. Thus, the choice of an optimal embedding dimension (e.g. through the use of Ragwitz' criterion [133]) is crucial as the use of larger than optimal embedding dimensions wastes available data and may lead to a weaker detection rate in noisy data [86].

Note, that the embedding window should not be confused with the analysis window. The analysis window strictly describes the data points, for which neighbor statistics enter TE estimation—where neighbor counts may be averaged over an epoch $\Delta_t$ or may come from a single point in time $t$ only. The embedding window however, describes the data points that enter the embedding of a single point in time. Thus, the temporal resolution of TE analysis may still be in single time steps $t$ (i.e. only one time point entering the analysis), even though the embedding window spans several points in time that contain the history for this single point.

### 2.5.3 Repeatability of neuronal processes

When applying the ensemble method to estimate TE from neural recordings, we treat experimental repetitions as multiple realizations of the neural processes under investigation. In doing so, we assume stationarity of these processes *over repetitions*. We claim that in most cases this assumption of stationarity is justified for processes concerned with the processing of experimental stimuli and that the assumption also holds for stimulus-independent processes that contribute to neural recordings. We will first present the different contributions to neural recordings and subsequently discuss their individual statistical properties, i.e. their stationarity over repetitions. Note, that the term stationarity refers to the stability of the *probability distribution underlying* the observed realizations of contributions over repetitions and does not require individual realizations to be identical; i.e. stationarity does not preclude a variability in observed realizations, but rather implies some variance in observed realizations, that is reflective of the variance in the underlying probability distribution.

Contributions to neural recordings may either be stimulus-related (*event-related activity*) or stimulus-independent (*spontaneous ongoing activity*). Within the category of event-related activity, contributions can be further distinguished into phase-locked and non phase-locked contributions (the latter is commonly called *induced activity*). Phase-locked activity has a fixed polarity and latency with respect to the stimulus and—on averaging over repetitions—contributes to an event-related potential or field (ERP/F). Phase-locked activity is further distinguished into two types of contributions, that are discussed as mechanisms in the ERP/F-generation (e.g. [158–160]): (1) *additive evoked contributions*, i.e. neural activity that is in addition to ongoing activity and represents the stereotypical response of a neural population to the presented stimulus in each repetition [161–163]; (2) *phase- reset contributions*, i.e. the phase of ongoing activity is reset by the stimulus, such that phase-aligned signals no longer cancel each other out on averaging over repetitions [164–167]. In contrast to these two subtypes of phase-locked activity, induced activity is event-related activity that is not phase-locked to the stimulus, such that latency and polarity vary randomly over repetitions and induced activity averages out over repetitions.

We therefore have to consider four types of contributions to neural recordings: (1) additive evoked contributions, (2) phase-reset contributions, (3) induced contributions and (4) spontaneous ongoing contributions, the last being stimulus-independent. Stationarity can be assumed for all these contributions if no learning effects occur during the experiment. Learning effects may lead to slow drifts, i.e. changing mean and variances, in the recorded signal. Such learning effects may easily be tested for by comparing the first and second half of recorded repetitions with respect to equal variances and means. If variances and means are equal, learning effects can most likely be excluded. Empirically, the stationarity assumption, specifically of phase-locked contributions, can also be verified using a modified independent component analysis recently proposed in [168].

To sum up the statistical properties of different contributions to neural data and their relevance for using an ensemble approach to TE estimation, we conclude that all contributions to neural recordings can be considered stationary over repetitions by default. Non-stationarity over repetitions will only be a problem in paradigms that introduce (slow) drifts or trends in the recorded signal, for example by facilitating learning during the experiment. Testing for drifts may be done by comparing mean and variance in a split-half analysis.

### 2.5.4 Relation of the ensemble method to local information dynamics

We will now discuss the relation of the ensemble approach suggested here to the local transfer entropy (LTE) approach of Lizier [65, 95]. This may be useful as both approaches at first glance seem to have a similar goal, i.e. assessing information transfer more locally in time. As we will show, the approaches differ in what quantities they localize. From this difference it also follows that they can (and should be) combined when necessary.

In detail, the ensemble approach used here tries to exploit cyclostationarity or repeatability of random processes to obtain multiple PDFs from the different (quasi-) stationary parts of the repeated process cycle, or a PDF for each step in time from replications of a process, respectively. In contrast, local information dynamics localizes information transfer in time (and space) given the PDF of a *stationary* process.

The local information dynamics approach to information transfer computes information transfer for stationary random processes from their joint and marginal PDFs for each process step, thereby fully localizing information transfer in time. The quantity proposed for this purpose is the LTE [65]:

$$LTE(\mathtt{X} \rightarrow \mathtt{Y}, t, \delta) = log \frac{p(y_t|\mathbf{y}_{t-1}, \mathbf{x}_{t-\delta})}{p(y_t|\mathbf{y}_{t-1})} \tag{2.17}$$

LTE relates to TE in the same way Shannon information relates to Shannon entropy—by means of taking an expected value under the common PDF $p(y_t, \mathbf{y}_{t-1}, \mathbf{x}_{t-\delta})$ of the collection of random variables $\{X_t\}, \{Y_t\}$ that form the processes X, Y, which exchange information. Stationarity here guarantees that all the random variables $X_1, X_2, \ldots$ $(Y_1, Y_2, \ldots)$ have a common PDF (as the PDF is not allowed to change over time):

$$TE(\mathtt{X} \rightarrow \mathtt{Y}, t, \delta) =< LTE(\mathtt{X} \rightarrow \mathtt{Y}, t, \delta) >_{p(y_t, \mathbf{y_{t-1}}, \mathbf{x_{t-\delta}})} \tag{2.18}$$

In contrast, the approach presented here does not assume that the random processes *X*, *Y* are stationary, but that either replications if the process can be obtained, or that the process is cyclostationary. Under these constraints a *local* PDF can be obtained. The events drawn from this PDF may then be analyzed in terms of their average information transfer, i.e. using TE as presented here, or by inspecting them individually, computing LTE for each event. In this sense, the approach suggested here is aimed at extracting the proper local PDFs, while local information dynamics comes into play once these proper PDFs have been obtained. We are certain that both approaches can be fruitfully combined in future studies.

### 2.5.5 Relation of the ensemble method to other measures of connectivity for non-stationary data

Linear Granger causality (GC) is—as has been shown recently by [103]—equivalent to TE for variables with a *jointly* Gaussian distribution. Thus, for data that exhibit such a distribution, information transfer may be analyzed more easily within the GC framework. Similar to the ensemble method for TE estimation, extensions to GC estimation have been proposed that deal with non-stationary data by fitting time-variant parameters. For example, Möller and colleagues presented an approach that fitted multivariate autoregressive models (MVAR) with time-dependent parameters to an ensemble of EEG signals [169]. Similar measures, that fit time-dependent parameters in autoregressive models to data ensembles, were used by [170] and [171]. A different approach to dealing with non-stationarity was taken by Leistritz and colleagues [172]. These authors proposed to use self-exciting threshold autoregressive (SETAR) models to model neural time series within a GC framework. SETAR models extend traditional AR models by introducing state-dependent model parameters and allow for the modeling of transient components in the signal.

The presented methods for the estimation of time-variant linear GC may yield a computationally less expensive approach to the estimation of information transfer from an ensemble of data. However, linear GC is equivalent to TE regarding the full recovery of information transfer for data with a *jointly* Gaussian distribution only. For non-Gaussian data, linear GC may fail to capture higher order interactions. As neural data are most likely non-Gaussian, the application of TE may have an advantage for the analysis of information transfer in this type of data. The non-Gaussian nature of neural data can for example be seen, when comparing brain electrical source signals from physical inverse methods to time courses of corresponding ICA components [173]. Here, ICA components and extracted brain signals closely match. Given that ICA components are as non-Gaussian as possible (by definition of the ICA), we can infer that brain signals are very likely non-Gaussian.

We also note that a nonstationary measure of *coupling* between dynamic systems building on repetitions of time series and next-neighbor statistics was suggested by Andrzejak and colleagues [174]. The key difference of their approach to the ensemble method suggested here is that the previous states of the target time series are not taken into account explicitly in their method. Hence, their measure is not (and was not intended to be) a measure of information transfer (see [75] for details why a measure of information transfer needs to include the past state of target time series, and [68] for the difference between measures of (causal) coupling and information transfer). In addition, their methods explicitly tries to determine the *direction* of coupling between to systems. This implies that there should be a dominant direction of coupling in order to obtain meaningful results. Transfer entropy, in contrast, easily separates and quantifies both directions of information transfer related to bidirectional coupling, under some mild conditions related to entropy production in each of the two coupled systems [75].

### 2.5.6 Relation of the ensemble method to the direct method for the calculation of mutual information of Strong and colleagues

The ensemble method proposed here shares the use of replications (or trials) with the so called "direct method" of Strong and colleagues [175]. The authors introduced this method to calculate mutual information between a controlled stimulus set and neural responses. Similarities also exist in the sense that the surrogate data method for statistical evaluation used in our ensemble method builds on trial-to-trial variability, as does Strong's method (by looking at intrinsic variability versus variability driven by stimulus changes).

However, the two methods differ conceptually on two accounts: First, the quantity estimated is different—symmetric mutual information in Strong's method compared to inherently asymmetric conditional mutual information in the case of TE. Second, the method of Strong and colleagues requires a direct intervention in the source of information (i.e. the stimuli) to work, whereas TE in general is independent of such interventions. This has far reaching consequences for the interpretation of the two measures: The intervention inherent in Strong's method places it somewhat closer to causal measures such as Ay and Polani's causal information flow [118], whereas intervention-free TE has a clear interpretation as the information transfered in relation to distributed computation [68]. As a consequence, TE maybe easily applied to quantify neural information transfer from one neuron or brain area to another even under *constant* stimulus conditions. In contrast, using Strong's method inside a neural system in this way would require precisely setting of the activity of the source neuron or brain area, something that may often be difficult to do.

### 2.5.7 Application of the proposed implementation to other dependency measures

The use of ensemble pooling of observations for the estimation of time-resolved dependency measures has been proposed in a variety of frameworks. For example, Andrzejak and colleagues [174] use ensemble pooling of delay-embedded time series in combination with nearest neighbor statistics as a general approach to the estimation of arbitrary non-linear dependency measures. However, the practical application of ensemble pooling and nearest neighbor statistics together with the necessary generation of a sufficient amount of surrogate data sets (typically >1000 in neuroscience applications where correction for multiple comparisons is necessary) was always hindered by its high computational cost. Only with the presentation of a GPU algorithm for nearest neighbor searches, we provide an implementation of the ensemble method that allows its practical application. Note that even though we use ensemble pooling and GPU search algorithms to specifically estimate TE, the presented implementation may easily be adapted to other dependency measures that are calculated from (conditional) mutual informations estimated from nearest neighbor statistics.

### 2.5.8 Application to MEG source activity in a perceptual closure task

Application of the ensemble-based TE estimation to MEG source activities revealed a time varying pattern of information transfers, as expected in the nonstationary setting of the visual task. While a full discussion of the revealed information transfer pattern is beyond the scope of this study, we point out individual connections transferring information that underline the validity of our results. Notable connections in the first time window transfer information from the early visual cortices (V1) to the orbitofrontal cortex (OFC)—in line with earlier findings by Bar an colleagues [153], that suggest a role of the OFC in early visual scene segmentation and gist perception. Another brain area receiving information from early visual cortex is the caudal inferior temporal gyrus (cITG)[176], an area responsible for the processing of shape-from-shading information, which is thought to be essential for perception of Mooney stimuli as they were used here. Both of these areas, OFC and cITG at later stages of processing exchange information with the fusiform face area, which is essential for the processing of faces [177–179], and thereby expected to receive information from other areas in this task. Indeed, FFA seems to be an essential hub in the task-related network investigated in this study and receives increasing amounts of incoming information transfer as the task progresses in time. This is in line with the fact that the most pronounced task-related differences in FFA activity were found at latencies $>200$ ms previously [148].

Our data also clearly show a great variability in information transfer pattern across subjects, which we relate to the limited amount of data per subject, rather than to true variation. Moreover, future investigations will have to show whether more fine grained temporal segmentation of the neural information processing in this task is possible and whether it will provide additional insights.

### 2.5.9 Conclusion and further directions

We presented an implementation of the ensemble method for TE presented in [24], that uses a GPU to handle computationally most demanding aspects of the analysis. We chose an implementation that is flexible enough to scale well with different experimental designs as well as with future hardware developments. Our implementation was able to successfully reconstruct information transfer in simulated and neural data in a time-resolved fashion. Nearest neighbor searches using a GPU exhibited substantially reduced execution times. The implementation has been made available as part of the open source MATLAB toolbox TRENTOOL [86] for the use with CUDA-enabled GPU devices.

We conclude that the ensemble method in its presented implementation is a suitable tool for the analysis of non-stationary neural time series, enabling this type of analysis for the first time. It may also be applicable in other fields that are concerned with the analysis of information transfer within complex dynamic systems.

# 3

# A Graph Algorithmic Approach to Separate Direct from Indirect Neural Interactions

Patricia Wollstadt[1,*], Ulrich Meyer[2], Michael Wibral[1]

**1** MEG Unit, Brain Imaging Center, Goethe University, Frankfurt am Main, Germany

**2** Institute for Computer Science, Goethe University, Frankfurt am Main, Germany

∗ E-mail: patricia.wollstadt@stud.uni-frankfurt.de

*Published as: P. Wollstadt, U. Meyer, M. Wibral (2015). "A graph algorithmic approach to separate direct from indirect neural interactions." PLoS ONE 10(10): e0140530.*

## Abstract

Network graphs have become a popular tool to represent complex systems composed of many interacting subunits; especially in neuroscience, network graphs are increasingly used to represent and analyze functional interactions between multiple neural sources. Interactions are often reconstructed using pairwise bivariate analyses, overlooking the multivariate nature of interactions: it is neglected that investigating the effect of one source on a target necessitates to take all other sources as potential nuisance variables into account; also combinations of sources may act jointly on a given target. Bivariate analyses produce networks that may contain spurious interactions, which reduce the interpretability of the network and its graph metrics. A truly multivariate reconstruction, however, is computationally intractable because of the combinatorial explosion in the number of potential interactions. Thus, we have to resort to approximative methods to handle the intractability of multivariate interaction reconstruction, and thereby enable the use of networks in neuroscience. Here, we suggest such an approximative approach in the form of an algorithm that extends fast bivariate interaction reconstruction by identifying potentially spurious interactions post-hoc: the algorithm uses interaction delays reconstructed for directed bivariate interactions to tag potentially spurious edges on the basis of their timing signatures in the context of the surrounding network. Such tagged interactions may then be pruned, which produces a statistically conservative network approximation that is guaranteed to contain non-spurious interactions only. We describe the algorithm and present a reference implementation in MATLAB to test the algorithm's performance on simulated networks as well as networks derived from magnetoencephalographic data. We discuss the algorithm in relation to other

approximative multivariate methods and highlight suitable application scenarios. Our approach is a tractable and data-efficient way of reconstructing approximative networks of multivariate interactions. It is preferable if available data are limited or if fully multivariate approaches are computationally infeasible.

## 3.1 Introduction

Complex systems are often composed of many interacting simpler subunits. To summarize our knowledge about such a system in an accessible format we frequently draw on its representation as a network graph, where the subunits become nodes and the identified interactions become links. Indeed, this way of summarizing knowledge has become so successful that we witness a rapidly increasing interest in the graph-properties of such network depictions [180–183]. The use of networks as a tool to represent and analyze functional interactions has been gaining importance also in neuroscience [180, 184–188]. In neuroscience, however, it is often overlooked that all derived graph measures are only as good as the reconstruction of the underlying interactions. This reconstruction may suffer, because the identification of all interactions in a multi-node network is fundamentally intractable since it poses a problem in the complexity class of so called “NP-hard” problems [26, 27]. Thus, true network graphs of interactions must be recovered using approximations if we do not want to forgo the use of network representations altogether.

To see why the identification of all interactions in a multi-node network is fundamentally intractable, we have to consider that next to the interactions from one node simply to one other node (a bivariate or pairwise interaction), there may well be interactions from a set of two (or more) source nodes to a target node. Moreover, this multivariate nature of the interactions makes it necessary to control for a parallel influence from any other source in the network when trying to determine whether a particular set of source nodes interacts with the target node in question. It is the enormous number of combinations of potential sources and parallel influences that makes it impossible to search all possibilities in reasonable time for any but the smallest systems (e.g. $n < 20$, [25, 27]). In fact, it can be shown formally that the problem belongs to the class of NP-hard problems, which are believed to lack algorithms that produce solutions for arbitrary input sizes in polynomial time [189].

To nevertheless apply graph theory and network models in neuroscience we need to resort to approximate representations of the true multivariate interactions. Here, the term approximation implies that we will have to commit errors. These errors can be of two types—falsely identifying an interaction that is physically absent, or missing an interaction that is physically present. While both types of errors may have detrimental effects on interpretability of popular graph metrics, we may still ask which type of error to prefer, and how to build fast and efficient approximations that predominantly show the preferred type of error.

Here, we suggest that missing out on interactions instead of including spurious ones may be preferable because the nature of the obtained network becomes more “reliable” in the sense that all the depicted links do exist. This knowledge can then be built upon in future work. Therefore we present an algorithm that can prune the most frequent spurious interactions from graphs obtained by a simple and efficient bivariate analysis of interactions (this idea was first proposed in [121] in abstract form).

Our focus here is specifically on corrections of graphs obtained from bivariate (i.e., pairwise) analysis methods as these have most often been used to overcome the intractability of the full network reconstruction described above. Despite their popularity, iterative bivariate analyses introduce well known methodological artifacts in the reconstructed interactions [25, 28, 29]: (1) Bivariate analysis may detect *spurious interactions* (false positives) whenever the dependency between two time series is caused or mediated by one or more additional nodes in the network; (2) bivariate analysis may miss *synergistic effects* [72, 76] that two or more time series have on a third. These two problems diminish not only the reliability of individual links in a network but also compromise graph metrics of the global network.

In this study we investigate a solution to the first problem above. Our solution builds on the possibility of reconstructing the delays of interactions (e.g. [75] and similar approaches), and on specific interaction-delay based fingerprints that potentially spurious interactions must leave even in a bivariate analysis. Our method allows to tag potentially spurious interactions for further testing (e.g. by a targeted multivariate analysis) or to remove them entirely from the graph to obtain its most reliable core. Our method thus keeps the advantages of bivariate methods in terms of data efficiency and computational tractability over approaches that are approximately or fully multivariate.

In the following we first provide the necessary background on delay reconstruction by information theoretic methods, and on graphs. Then we present our algorithm and a reference implementation as part of the open source toolbox TRENTOOL ([86], `http://www.trentool.de`). Subsequently, we characterize its properties and limitations based on theoretical considerations, simulations and application to magnetoencephalographic (MEG) data. We discuss the relative merits of our approach and other possible approximations to a fully multivariate analysis of networks, and close by outlining possible strategies to deal with the identified potentially spurious links in the network.

## 3.2 Background and Implementation

Before we outline the algorithm in more detail, we will provide some background information by reviewing the recovery of interaction delays in an information theoretic framework. We will also describe the coupling motifs leading to the detection of spurious interactions. Subsequently, we will formalize the network concept in mathematical terms and complement the common undirected and unweighted network representation used for neural data [180, 190] by introducing the weighting of network connections with their respective interaction delays (using the estimator provided in [75]). We will then describe the rationale underlying the algorithm and its implementation. We conclude this section with the validation of the algorithm using simulated as well as experimental data.

### 3.2.1 Background

**Interaction delay reconstruction.** Our algorithm is based on the availability of the interaction delays for bivariately reconstructed interactions. We will systematically use the term “bivariate interaction” to indicate that in the bivariate analysis setting there is no guarantee that a reconstructed interaction is actually present in the underlying data; nevertheless, even for a spurious interaction a meaningful delay can be assigned (see examples in [75]). One possibility to obtain the delays for bivariate interactions is to use delay-sensitive measures of information transfer, i.e., transfer entropy (TE) estimators. In [75] we presented a delay-sensitive TE functional:

$$TE_{SPO}\left(X \rightarrow Y, t, u\right) = \sum_{y_t, \mathbf{y}_{t-1}, \mathbf{x}_{t-u}} p\left(y_t, \mathbf{y}_{t-1}, \mathbf{x}_{t-u}\right) \log \frac{p\left(y_t | \mathbf{y}_{t-1}, \mathbf{x}_{t-u}\right)}{p\left(y_t | \mathbf{y}_{t-1}\right)} , \qquad (3.1)$$

which quantifies the mutual information between the past state $\mathbf{x_{t-u}}$ of a source $X$ and the present value $y_t$ of a target $Y$ at a specific time delay $u$—conditional on the past state of the target, $\mathbf{y_{t-1}}$.

This functional can be used to recover the physical interaction delay $\delta_{X,Y}$ by scanning over possible values for $u$, and by taking the value of $u$ where TE reaches a maximum as the (bivariate) interaction delay:

$$\delta_{X,Y} = \arg\max_u \left(TE_{SPO}\left(X \rightarrow Y, t, u\right)\right) . \qquad (3.2)$$

A similar approach may be used for Granger causality as TE is equivalent to Granger causality for data with a jointly Gaussian distribution [103].

Note, that reconstructed interaction delays incorporate not only the mere transfer time between two neural processes, but also the time needed for local computation, if we only obtain one channel per subunit (see Fig. 3.1 for further explanation). Thus, interaction delay reconstruction as proposed in [75], captures the total delay between two measurement *points*, which in a neural system may consist of transfer time along an axonal connection but also of time needed for information transfer within the local neural microcircuit.

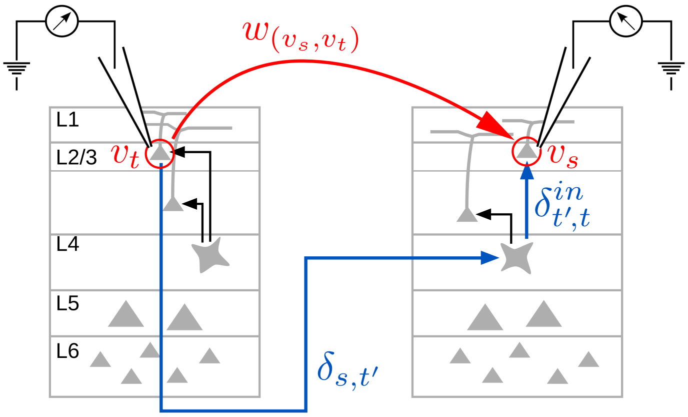


Fig. 3.1 **Reconstructing delay times from electrophysiological recordings.** The physiological delay between two measurement points $v_s$, $v_t$ consists of the time needed for information transfer via axonal connections ($\delta_{s,t'}$) and internal computation within populations of neurons ($\delta^{in}_{t',t}$), such that in a network representation of reconstructed interactions we find $w_{(v_s,v_t)} = \delta_{s,t'} + \delta^{in}_{t',t}$. Black arrows represent information transfer within the neural microcircuits.

**Spurious interactions in bivariate analysis of multivariate data sets.** Spurious interactions may arise in bivariate analysis from one of two distinct coupling motifs: In the first coupling motif (Fig. 3.2A) the dynamics of two or more nodes, representing neural sources, are simultaneously driven by processes in a third node. A bivariate analysis may detect an interaction between the two driven nodes. We term this a *common drive* (CD, also “common cause” [25]). In the second coupling motif (Fig. 3.2B) an interaction between two nodes is mediated by one or more intermediate nodes in the network and information transfer from source node to target node is routed via these intermediate nodes. A bivariate analysis may detect an interaction between the source and target node. We term this a *cascade effect* (CE, also “pathway effect” [25] or “indirect causal pathways” [28]). The detection of spurious interactions by bivariate analysis have been demonstrated for simulated data as well as in neural recordings [28, 29, 91, 191, 192].

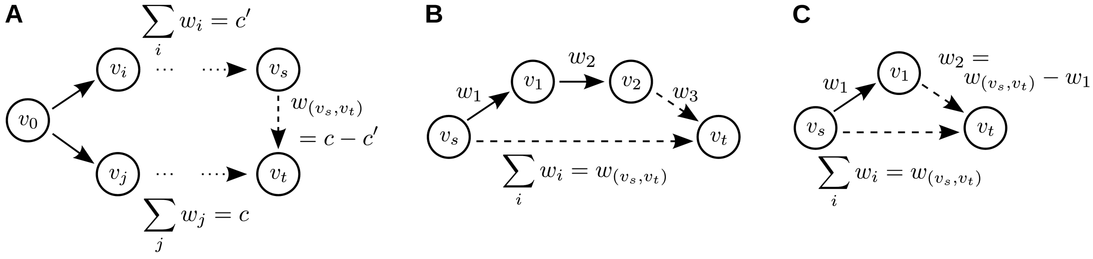


Fig. 3.2 **Spurious Interactions.** (A) Common drive effect: A spurious interaction due to common drive may be potentially present if the processes at vertices $v_s$ and $v_t$ are driven by $v_0$ with differential delays, such that the bivariate information transfer between $v_s$ and $v_t$ is a result of the common input from $v_0$; (B) Cascade effect: Spurious interaction due to cascade effects may be potentially present for all cascades of information transfer in a "chain" of sources. In the example here, the bivariate information transfer between $v_s$ and $v_t$ (edge $(s, t)$) can be explained by an alternative routing of information via vertices $v_1$ and $v_2$. The summed weight of the alternative routing is equal to $w_{(s,t)}$; (C) "Triangle" motif: This is the most simple motif potentially giving rise to either of the above spurious interactions: $(v_s, v_t)$ could be a result of a cascade effect with respect to the path $\langle v_s, v_1, v_t \rangle$, and $(v_1, v_t)$ could be the result of a common drive of $v_1$ and $v_t$ by $v_s$. At most one of these interactions can be spurious.

Coupling motifs leading to spurious interactions due to CD and CE exhibit a specific *timing signature* in the network of bivariately reconstructed interactions. We found an example for such a timing signature in experimental data recorded from the turtle (Fig. 3.3C) [75]. Here, a spurious interaction was detected between the light source and the optic tectum. This spurious interactions resulted from a CE, i.e., an actual routing of information from light source to tectum via an intermediate node, namely the retina. Information transfer delays reconstructed with the TE estimator proposed in [75], revealed this CE: The summed interaction delays in the actual routing of information equaled the delay of the spurious information transfer.

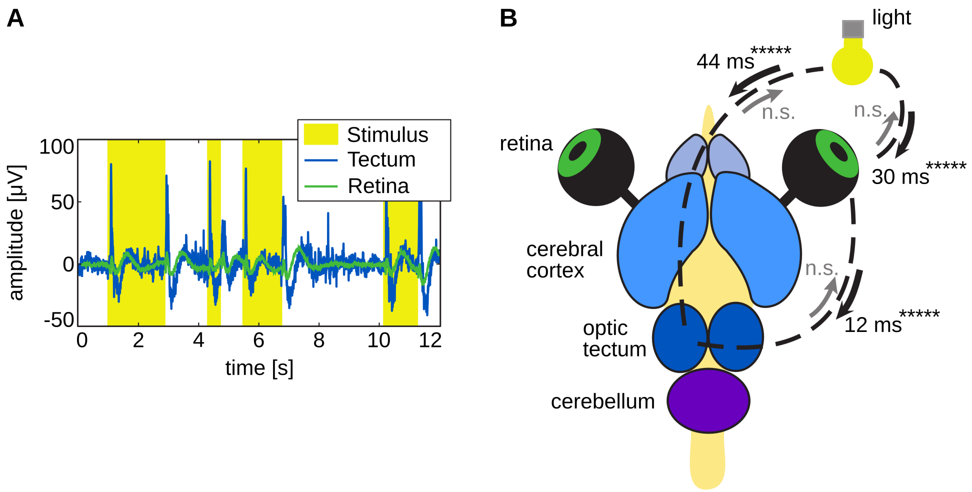


Fig. 3.3 **Directed interactions in the turtle brain during visual stimulation with random light pulses (modified from [75], creative common attribution license CC BY).** (A) Raw traces recorded in the tectum (blue) and from the retina (green) overlaid on the light pulses (yellow). (B) Turtle brain explant with eyes attached. Transfer entropy was found from the retina of the right eye to the left tectum, as well as from the light source (yellow) to the retina and to the tectum (***** denotes $p < 10^{(-5)}$). P-values for the opposite directions were not significant ($n.s.$). Note, that the interaction between light source and optic tectum shows a interaction delay roughly equal to the summed interaction delay between light source and retina and retina and optic tectum (deviation $\leq 5\,\%$).

**Graph representation of neural data and notation.** As a last preliminary, we will present the mathematical formalization of a network to give a precise account of the algorithm and its functionality in the subsequent section. Table 3.1 lists the most important variables. In mathematical terms, a network is described by a (directed) graph $\mathbf{G} = \{\mathbf{V}, \mathbf{E}\}$, where $\mathbf{V}$ denotes a set of vertices or nodes and $\mathbf{E} \subseteq \mathbf{V} \times \mathbf{V}$ represents a set of connections between nodes, called edges [193]. In a neuroscience application, $\mathbf{V}$ may represent a set of individual functional units $v_i$, e.g. neurons, sources in MEG analysis, or voxels in functional magnetic resonance imaging data; $\mathbf{E}$ may represent some sort of connection between two units, for example significant functional interactions. An edge is written as a tuple $(v_i, v_j)$, representing an edge between any two sources $v_i$ and $v_j$. Note, that such a tuple $(v_i, v_j)$ defines an edge as an ordered pair of two vertices and as such indicates a *directed* connection between these two sources (the two elements of the tuple are not interchangeable, such that $(v_i, v_j) \neq (v_j, v_i)$). We further assume that edges are weighted by a weighting function $w : \mathbf{E} \mapsto \mathbb{N}$ that maps the set of edges to the natural numbers. Here, these natural numbers are chosen proportionally to the timing information as precisely and as parsimoniously as possible. We call $w_{(v_i, v_j)}$ the weight of edge $(v_i, v_j)$.

Tab. 3.1 **Notation**

| **Graph representation** | |
|---|---|
| $\mathbf{G}$ | graph, consisting of the sets $\mathbf{V}$ and $\mathbf{E}$ |
| $\mathbf{V}$ | set of vertices |
| $\mathbf{E}$ | set of edges |
| $v_i$ | vertex with index $i$ |
| $(v_i, v_j)$ | edge from $v_i$ to $v_j$ |
| $w_{(v_i,v_j)}$ | weight of edge $(v_i, v_j)$ |
| $v_i \rightsquigarrow v_j$ | path from $v_i$ to $v_j$ |
| $v_i \overset{k}{\rightsquigarrow} v_j$ | path from $v_i$ to $v_j$ with summed weight $k$ |
| $l$ | path length, i.e. no. edges in a path |

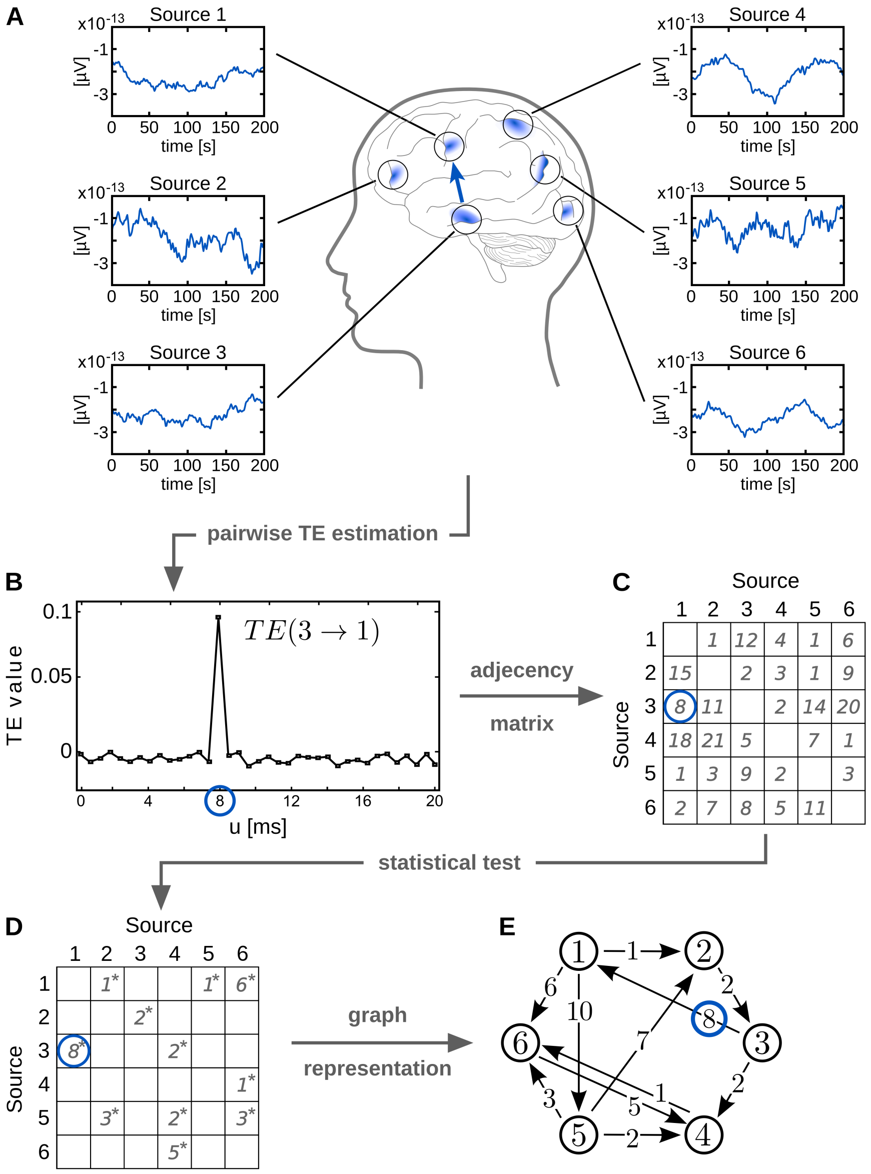
| **Algorithm** | |
|---|---|
| $(v_a, v_b)$ | edge under investigation in current algorithmic iteration |
| $w_{(v_a,v_b)}$ | weight of the edge under investigation |
| $\theta$ | threshold to account for imprecisions in interaction delay reconstruction |
| $w_{crit}$ | critical path weight, $w_{crit} = w_{(v_a,v_b)} + \theta$ (target weight of the algorithm) |
| $v_s, v_t$ | start and target node ($v_s = v_a$, $v_t = v_b$) |
| $\mathbf{L}^n_{v_s}(w_i; \rightsquigarrow v_j)$ | set of solutions in algorithmic step $n$, that solve the subproblem $v_s \overset{w_i}{\rightsquigarrow} v_j$ |

For any edge $(v_i, v_j)$, $v_i$ is considered the *predecessor* of $v_j$. $v_j$ is called the *child* of $v_i$. A *path* $v_0 \rightsquigarrow v_l$ is defined as a sequence of vertices $\langle v_0, v_1, \ldots, v_i, \ldots, v_{l-1}, v_l \rangle$, where every two consecutive vertices $v_i$ $v_{i+1}$ are connected by an edge $(v_i, v_{i+1})$. We call $l$ the length (number of edges) of the path and we will refer to this length $l$ as a path's *graphical length*, describing the number of edges used to graphically represent the path in the graph. The total weight of a path is the sum of the weights of all individual edges comprising the path, $\sum_i w_{(v_i,v_{i+1})}$.

Fig. 3.4 gives a schematic overview of the construction of a graph from time series data recorded from a set of neural sources. Edges in the graph represent significant interactions between sources (vertices); edge weights represent reconstructed interaction delays. We here use TE to analyze directed interactions, using the estimator proposed in [75] to recover significant TE and corresponding interaction delays, but other approaches are possible.

## 3.2.2 Rationale and implementation of the algorithmic solution

**Rationale of the algorithm.** Based on the graph representation of reconstructed directed interactions, their delays, and the theoretical preliminaries presented in the



Panel C:

| Source | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 |  | 1 | 12 | 4 | 1 | 6 |
| 2 | 15 |  | 2 | 3 | 1 | 9 |
| 3 | 8 | 11 |  | 2 | 14 | 20 |
| 4 | 18 | 21 | 5 |  | 7 | 1 |
| 5 | 1 | 3 | 9 | 2 |  | 3 |
| 6 | 2 | 7 | 8 | 5 | 11 |  |

Panel D:

| Source | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 |  | 1* |  |  | 1* | 6* |
| 2 |  |  | 2* |  |  |  |
| 3 | 8* |  |  | 2* |  |  |
| 4 |  |  |  |  |  | 1* |
| 5 |  | 3* |  | 2* |  | 3* |
| 6 |  |  |  | 5* |  |  |

Fig. 3.4 **Graph representation of neural data.** (A) Recorded signals from various sources in the brain; (B) Pairwise estimation of transfer entropy (TE) and reconstruction of interaction delays $u$ between any two sources; (C) Adjacency matrix: representation of estimated delay times between all source combinations, every entry represents an information transfer from the $i$th row to the $j$th column; (D) Adjacency matrix after test for statistical significance; (E) Visualization of the graph represented by the connectivity matrix: every source is represented by a vertex, every significant information transfer is represented by an edge. (The blue circle indicates the respective representation of an exemplary interaction between source 1 and source 3 throughout all steps of graph reconstruction.)

last subsection, we will now propose an algorithm that detects potentially spurious interactions by exploiting the timing signature of CE and simple CD.

As input the algorithm expects a directed, delay-weighted graph $\mathbf{G} := \{\mathbf{V}, \mathbf{E}\}$, represented by its connectivity matrix. Connections are weighted with the estimated interaction delays $w_{(v_a,v_b)} = \hat{\delta}_{(v_a,v_b)}$, i.e., the estimated physical delays between the processes represented by $v_a$ and $v_b$. Note that such a graph needs to be constructed from a connectivity measure, which is (a) directed and (b) allows for the reconstruction of interaction delays. Additionally, the user has to provide a threshold $\theta$ to account for noise in empirical measurements as well as imprecision in analysis methods (described in detail below). We furthermore assume that weights have been linearly scaled, such that they do not have any decimal places and can be represented by integer values.

As a first step, we identify potential CEs by assuming that if a CE is present in the data, the bivariate interaction represented by any edge $(v_a, v_b) \in E$ with weight $w_{(v_a,v_b)}$ can be explained by an alternative routing of information via intermediate vertices (see an example in Fig. 3.2B). Thus, iteratively for every edge $(v_a, v_b)$ in $\mathbf{E}$ the algorithm sets $v_a = v_s$ as the starting and $v_b = v_t$ as the target node of the current iteration. Then the algorithm searches for an alternative path for $(v_s, v_t)$, where a path is assumed to be an alternative path if the summed delay interaction times $\sum_i w_{(v_i,v_{i+1})}$ of all edges in the path are (approximatively) equal to $w_{(v_s,v_t)}$, i.e., $w_{(v_s,v_t)} - \theta < \sum_i w_{(v_i,v_{i+1})} < w_{(v_s,v_t)} + \theta$. For an example see Fig. 3.2B, where edge $(v_s, v_t)$ (dashed arrow) has a graphically longer, alternative path $\langle v_s, v_1, v_2, v_t \rangle$ with equal summed weight (solid arrows). If the algorithm finds such an alternative path, the edge currently under investigation is tagged as potentially spurious.

In a second step, the algorithm additionally identifies potential simple CD effects based on the results from the first analysis step. Simple CD effects occur in graph motifs that form acyclic "triangles" (Fig. 3.2C). We define an acyclic triangle as any three nodes that are acyclic, pairwise connected. We suspect a spurious link due to simple CD if a triangle motif exhibits a suspicious timing signature; these suspicious triangles are identified from the results of the first analysis step by listing all edges with alternative paths of graphical length two. We propose to tag two edges in each of these identified triangles: (1) the direct edge due to a CE (edge $(v_s, v_t)$ in Fig. 3.2C) and (2) the second edge due to a simple CD (edge $(v_1, v_t)$ in Fig. 3.2C). Note however, that of both tagged edges one has to be a non-spurious interaction for this rationale to hold: if the CE link $(v_s, v_t)$ is rejected, a possible driver-target relationship between nodes $v_s$ and $v_t$ is destroyed, thus nullifying the argument for the simultaneous rejection of edge $(v_1, v_t)$. The same argument holds vice versa: a rejection of the tagged edge $(v_1, v_t)$ due to CD destroys the information cascade from $v_s$ to $v_t$ and thus cancels the CE causing the detection of $(v_s, v_t)$. Thus, tagging both edges in a triangle motif yields a slightly too conservative approach to network representation. For the treatment of tagged links in neuroscience we refer to the *Discussion* section.

A further important consideration is that once the algorithm has detected an alternative path $v_a \rightsquigarrow v_b$ for an edge $(v_a, v_b)$, this alternative path stays intact, even if, at a later step, some edge in $v_a \rightsquigarrow v_b$ is tagged (and probably removed). Assume, that we have an edge $(v_a, v_b)$ with an alternative path $v_a \rightsquigarrow v_b$ and that at a later step, one or more edges in this path get tagged. Then, for each of these tagged edges, leaving a "gap" in $v_a \rightsquigarrow v_b$, an alternative path of equal summed weight exists by definition of the algorithm. This alternative path closes the "gap" in $v_a \rightsquigarrow v_b$, such that nodes $v_a$ and $v_b$ are still connected by a new path $v_{a'} \rightsquigarrow v_{b'}$. This new path has the same summed delay as $v_a \rightsquigarrow v_b$, but is graphically longer. Therefore, alternative paths identified at some step in the algorithm remain intact even if links within the paths are tagged at a later point in the algorithm's execution.

**Implementation overview.** We now present the implementation of the strategy described above; more precisely, we present an algorithm that finds all alternative paths for any given edge $(v_a, v_b) \in \mathbf{E}$, given a directed, weighted graph $\mathbf{G} = \{\mathbf{V}, \mathbf{E}\}$. See Fig. 3.5 for an overview of the algorithm.

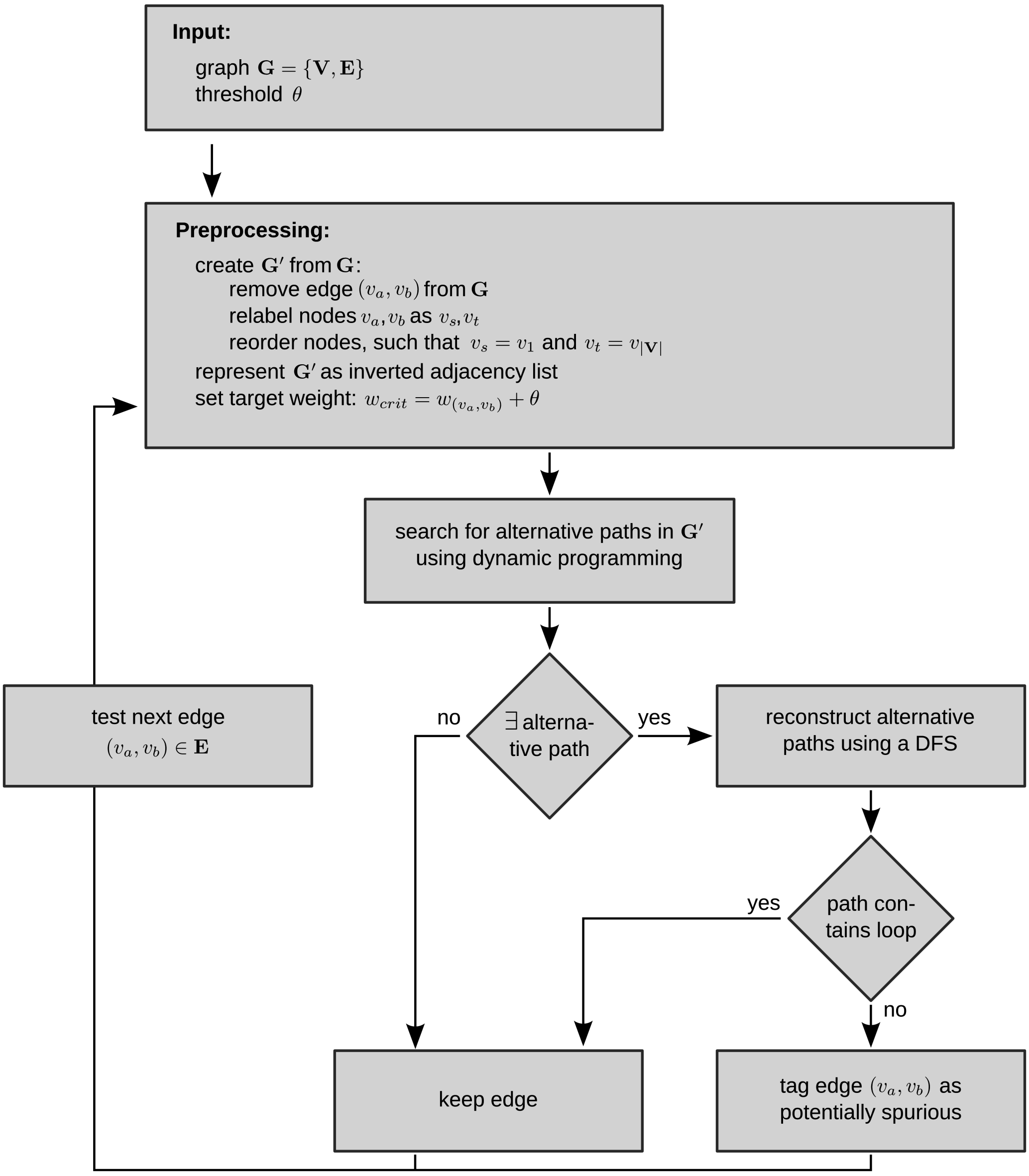

Input:
graph $\mathbf{G} = \{\mathbf{V}, \mathbf{E}\}$
threshold $\theta$
Preprocessing:
create $\mathbf{G}'$ from $\mathbf{G}$:
remove edge $(v_a, v_b)$ from $\mathbf{G}$
relabel nodes $v_a, v_b$ as $v_s, v_t$
reorder nodes, such that $v_s = v_1$ and $v_t = v_{|\mathbf{V}|}$
represent $\mathbf{G}'$ as inverted adjacency list
set target weight: $w_{crit} = w_{(v_a,v_b)} + \theta$
search for alternative paths in $\mathbf{G}'$ using dynamic programming
$\exists$ alternative path
no
yes
reconstruct alternative paths using a DFS
path contains loop
yes
no
test next edge $(v_a, v_b) \in \mathbf{E}$
keep edge
tag edge $(v_a, v_b)$ as potentially spurious

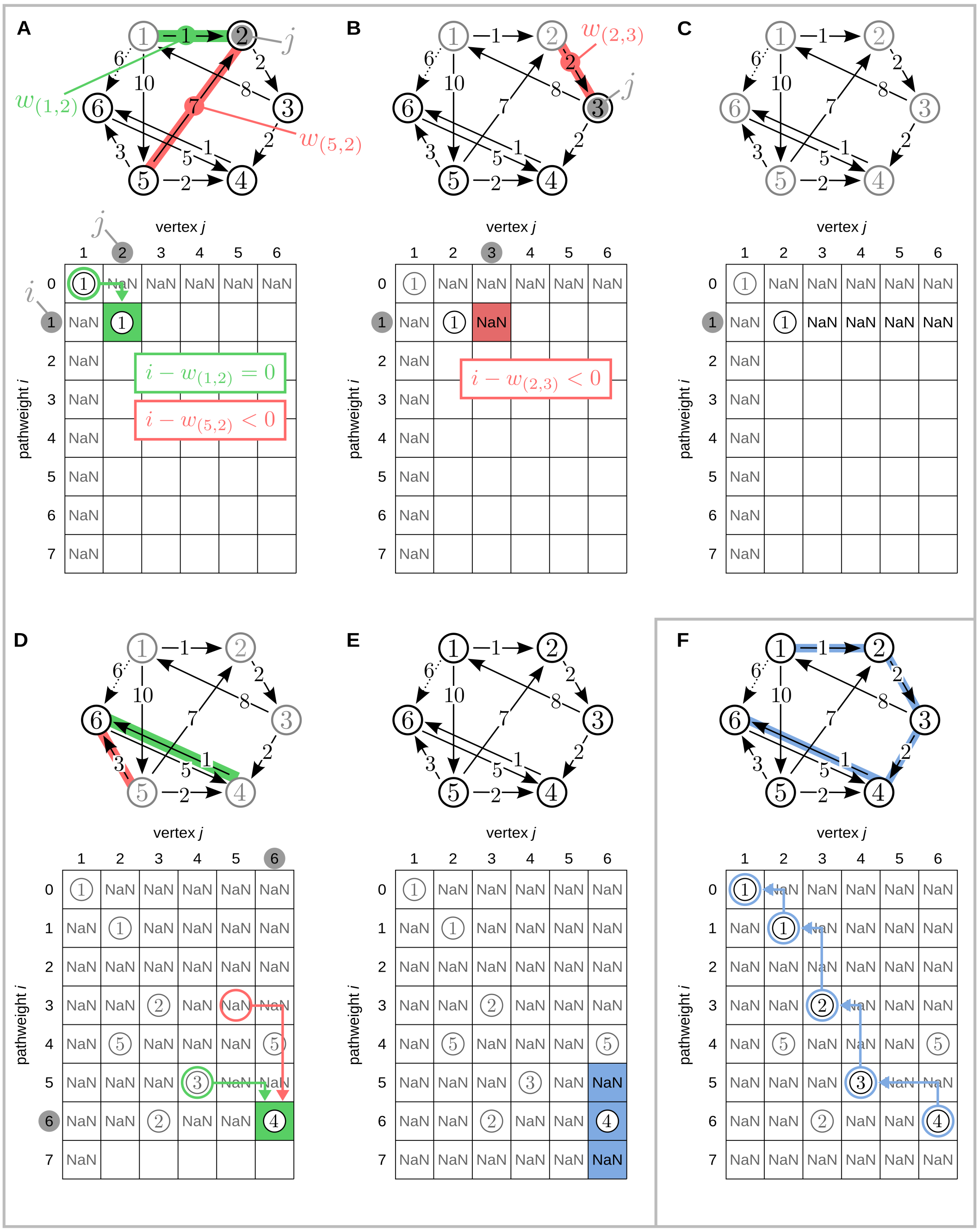
Fig. 3.5 **Overview of the proposed algorithm.** The algorithm expects a weighted and directed graph $\mathbf{G} = \{\mathbf{V}, \mathbf{E}\}$ and a threshold $\theta$ as input.
In a preprocessing step, the algorithm creates graph **G'** from input **G**, as an input for the dynamic programming algorithm, by removing edge $(v_a, v_b)$ and by relabeling and reordering nodes. Then, in the next step, alternative paths for $(v_a, v_b)$, are searched through dynamic programming (see also main text). If at least one alternative path is found, paths are reconstructed using a depth first search (DFS, [194]) to ensure that alternative paths do not contain loops. If an alternative path contains no loops, the currently investigated edge $(v_a, v_b)$ is tagged as potentially spurious. If no alternative edge is found, $(v_a, v_b)$ is considered non-spurious.
The algorithm then enters the next iteration, in which the next edge $(v_a, v_b) \in \mathbf{E}$ is investigated for alternative paths.

As a preprocessing step, we construct a graph $\mathbf{G}'$ by removing the edge $(v_a, v_b)$ from $\mathbf{G}$ and relabeling nodes $v_a$, $v_b$ as starting and target nodes $v_s$, $v_t$ of the current iteration of the algorithm. The target weight of the alternative path is set to $w_{crit} = w_{(v_a,v_b)} + \theta$. Furthermore, $\mathbf{G}'$ is represented as an inverted adjacency list, i.e., a list of all nodes in $\mathbf{V}$, where for every node all of its predecessors are listed. $v_s$ is set as the first node in this list (note that we assume that $v_s$ has index 1), $v_t$ is set as the last node (has index $|\mathbf{V}|$). Therefore, for all other nodes $v_j$ in the adjacency list it now holds that $2 \leq j \leq |\mathbf{V}|$.

After preprocessing the input, alternative paths $v_s \rightsquigarrow v_t$ with weight $w_{(v_s,v_t)} \pm \theta$ are detected in two steps: (1) a memoized dynamic programming approach [194] is used to determine, whether any path $v_s \rightsquigarrow v_t$ of a total weight $w_{(v_s,v_t)} \pm \theta$ exists; (2) a modified depth first search (DFS, see [194] for a description) is used to reconstruct all paths with weights in the interval $w_{(v_s,v_t)} \pm \theta$ from the solution obtained in step (1). The second step is necessary to reject paths that contain loops and to allow for further analysis (for example the identification of triangle motifs).

The algorithm was implemented as part of the open source toolbox TRENTOOL [86].

**Dynamic programming.** We use a dynamic programming approach in step (1) to handle the inherent complexity of the problem at hand (see *Discussion*). Dynamic programming allows for the solution of a complex problem by decomposing it into easily solvable subproblems. By starting with trivial base cases, subproblems of increasing complexity are solved iteratively by taking recourse to tabulated solutions of previous (more simple) subproblems. This reduces computational demand and is repeated until the algorithm reaches the most complex subproblem, which represents the input problem. In the following, we will first define the subproblems to the present problem, and then describe the algorithmic solution to an individual subproblem. A detailed, graphical account of both steps is presented in Fig. 3.6.



A

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 0 | ① | NaN | NaN | NaN | NaN | NaN |
| 1 | NaN | ① | | | | |
| 2 | NaN | | | | | |
| 3 | NaN | | | | | |
| 4 | NaN | | | | | |
| 5 | NaN | | | | | |
| 6 | NaN | | | | | |
| 7 | NaN | | | | | |

B

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 0 | ① | NaN | NaN | NaN | NaN | NaN |
| 1 | NaN | ① | NaN | | | |
| 2 | NaN | | | | | |
| 3 | NaN | | | | | |
| 4 | NaN | | | | | |
| 5 | NaN | | | | | |
| 6 | NaN | | | | | |
| 7 | NaN | | | | | |

C

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 0 | ① | NaN | NaN | NaN | NaN | NaN |
| 1 | NaN | ① | NaN | NaN | NaN | NaN |
| 2 | NaN | | | | | |
| 3 | NaN | | | | | |
| 4 | NaN | | | | | |
| 5 | NaN | | | | | |
| 6 | NaN | | | | | |
| 7 | NaN | | | | | |

D

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 0 | ① | NaN | NaN | NaN | NaN | NaN |
| 1 | NaN | ① | NaN | NaN | NaN | NaN |
| 2 | NaN | NaN | NaN | NaN | NaN | NaN |
| 3 | NaN | NaN | ② | NaN | NaN | NaN |
| 4 | NaN | ⑤ | NaN | NaN | NaN | ⑤ |
| 5 | NaN | NaN | NaN | ③ | NaN | NaN |
| 6 | NaN | NaN | ② | NaN | NaN | ④ |
| 7 | NaN | | | | | |

E

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 0 | ① | NaN | NaN | NaN | NaN | NaN |
| 1 | NaN | ① | NaN | NaN | NaN | NaN |
| 2 | NaN | NaN | NaN | NaN | NaN | NaN |
| 3 | NaN | NaN | ② | NaN | NaN | NaN |
| 4 | NaN | ⑤ | NaN | NaN | NaN | ⑤ |
| 5 | NaN | NaN | NaN | ③ | NaN | NaN |
| 6 | NaN | NaN | ② | NaN | NaN | ④ |
| 7 | NaN | NaN | NaN | NaN | NaN | NaN |

F

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 0 | ① | NaN | NaN | NaN | NaN | NaN |
| 1 | NaN | ① | NaN | NaN | NaN | NaN |
| 2 | NaN | NaN | NaN | NaN | NaN | NaN |
| 3 | NaN | NaN | ② | NaN | NaN | NaN |
| 4 | NaN | ⑤ | NaN | NaN | NaN | ⑤ |
| 5 | NaN | NaN | NaN | ③ | NaN | NaN |
| 6 | NaN | NaN | ② | NaN | NaN | ④ |
| 7 | NaN | NaN | NaN | NaN | NaN | NaN |

Fig. 3.6 **Visualization of the proposed algorithm.** Search for alternative paths to edge $(1, 6)$ (dotted arrow), i.e., $v_s = 1$ and $v_t = 6$. Solutions $\mathbf{L}^n_{v_s}(w_i; \rightsquigarrow v_j)$ to subproblems are managed in a two dimensional solution array indexed by path weight $w_i$ and vertex number $v_j$. Solutions are calculated iteratively over $w_i$ (rows) and $v_j$ (columns). (A) Solution matrix after first iterative step (subproblem $\mathbf{L}^1_{v_s}(1; \rightsquigarrow v_2)$): There are two edges leading to vertex 2 of which only edge $(1, 2)$ yields a valid solution by pointing to an earlier solved subproblem (green box), whereas edge $(5, 2)$ has a weight of 7 leading to a negative difference in weights $i - w_{(5,2)}$ (red box) for which no earlier solution exists; (B) Solution matrix after third iterative step $\mathbf{L}^3_{v_s}(1; \rightsquigarrow v_3)$: Here, no valid solution exists (none of the arrows leading to vertex 2 are part of a path with summed weight 1); (C) Solution array after iteration over all vertices $v_j$ for $w_i = 1$ (all vertices have been checked for a path of weight 1, originating from the start vertex 1); (D) Solution to subproblem $\mathbf{L}^n_{v_s}(6; \rightsquigarrow v_6)$: edge $(4, 6)$ together with the solution $\mathbf{L}_{v_s}(5; \rightsquigarrow v_4)$ form a valid path, whereas edge $(5, 6)$ is not part of a valid solution as $\mathbf{L}_{v_s}(3; \rightsquigarrow v_5)$ is empty;

Fig. 3.6 **Visualization of the proposed algorithm (continued).** (E) The algorithm terminates after iteration over all vertices $v_j$ and path weights up to $w_{(v_s,v_t)} + \theta$, where $\theta$ is a user defined threshold of 1. Backtracking is conducted for all entries in the reconstruction interval $w_{(v_s,v_t)} \pm \theta$ (entries marked blue); (F) Reconstructed alternative path by backtracking of subproblems.

**Formulation and ordering of subproblems.** The overall problem of finding a path from $v_s$ to $v_t$ with weight $w_{(v_s,v_t)} \pm \theta$ is divided into subproblems by asking, whether a "simpler" path $v_s \overset{w_i}{\rightsquigarrow} v_j$ exists; $v_j$ is any node in $\mathbf{V}$ and $w_i$ is any path weight $w_i \leq w_{crit} = w_{(v_s,v_t)} + \theta$. For example, finding a path $v_1 \overset{5}{\rightsquigarrow} v_3$ is a subproblem to finding path $v_1 \overset{9}{\rightsquigarrow} v_{10}$ where $v_1 = v_s$ and $v_{10} = v_t$.

The presented algorithm solves subproblems iteratively for increasing complexity using a dynamic programming approach. It is thus necessary to order the subproblems by their complexity and to make their solutions immediately accessible for reuse in subsequent algorithmic steps. This is realized by organizing solutions in a two-dimensional *solution array*, in which solutions are ordered by path weights and node indices, the two parameters determining complexity. Starting with the most simple base case where $w_i = 0$ and $v_s = v_1 = v_j$ (this subproblem describes a path $v_s \overset{0}{\rightsquigarrow} v_s$), subsequent subproblems are formulated by increasing path weights and node indices in integer steps. Thus subproblems are formulated for all combinations $w_i = 1, 2, \ldots, w_{crit}$ and $v_j \in \mathbf{V}$ (see for example Fig. 3.6C for the first iteration of path weights and a complete iteration over all vertices). Individual solutions to subproblems are tabulated in the solution array of size $[0; w_{crit}]$ by $[1; |\mathbf{V}|]$ and are indexed by the current values for $w_i$ and $v_j$. This organization of subproblems allows for the easy retrieval of solutions from earlier iterations to solve the subproblem currently at hand (see below and Fig. 3.6).

**Finding solutions to subproblems.** For any given subproblem $v_s \overset{w_i}{\rightsquigarrow} v_j$ the algorithm determines whether a path of weight $w_i$ leads into node $v_j$. The algorithm does this by testing whether any single edge leading into $v_j$ together with the solution to a simpler subproblem forms the path $v_s \overset{w_i}{\rightsquigarrow} v_j$. In particular, for every edge $(v_p, v_j)$ leading into $v_j$ (where $v_p$ is a predecessor of $v_j$), the algorithm checks if this edges extends a path leading into $v_p$ such that the resulting path solves the current subproblem $v_s \overset{w_i}{\rightsquigarrow} v_j$ (see Fig. 3.7A). We call the treatment of one edge $(v_p, v_j)$ an *algorithmic step*. In one algorithmic step, the algorithm checks (1) if there *exists* a path to $v_p$ and if so, (2) whether the path has length $w_p = w_i - w_{(v_p,v_j)}$. If both conditions are met, the currently considered edge $(v_p, v_j)$ together with the path to predecessor $v_p$ solves the current subproblem $v_s \overset{w_i}{\rightsquigarrow} v_j$ (because a path $v_s \overset{w_p}{\rightsquigarrow} v_p$ exists, that together with $(v_p, v_j)$ forms a path of summed weight $w_i = w_p + w_{(v_p,v_j)}$). The algorithm terminates once the most complex subproblem has been investigated, i.e., it has been tested if a path of length $w_{crit}$, connecting node $v_s$ to $v_t$ can be found.

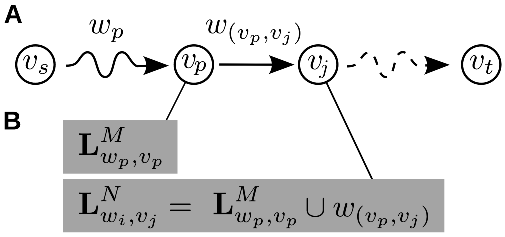

A
$w_p$
$w_{(v_p,v_j)}$
$v_s$
$v_p$
$v_j$
$v_t$
B
$\mathbf{L}^M_{w_p,v_p}$
$\mathbf{L}^N_{w_i,v_j} = \mathbf{L}^M_{w_p,v_p} \cup w_{(v_p,v_j)}$

Fig. 3.7 **Schematic example of a subproblem of the proposed algorithm.** (A) Example subproblem $\mathbf{L}^n_{w_i,v_j}$: At the $n^{th}$ algorithmic step, we search for all paths of weight $w_i$ leading to node $v_j$; (B) Finding a solution for the current subproblem by investigating solutions to prior subproblems: We investigate all predecessors $v_p$ of the current node $v_j$; if there exists a solution to $\mathbf{L}^M_{w_p,v_p}$, i.e., there is a solution to the prior subproblem $v_s \overset{w_p}{\rightsquigarrow} v_p$ of finding a path of weight $w_p$ leading from $v_s$ to $v_p$, and $w_p + w_{(v_p,v_j)} = w_i$, we find a solution to the current subproblem $\mathbf{L}^n_{w_i,v_j}$.

Note that "checking" if a path of weight $w_p$ to node $v_p$ exists corresponds to looking up whether a solution to the subproblem $v_s \overset{w_p}{\rightsquigarrow} v_p$ exists in the solution array, i.e., the algorithm has to look up the entry in the solution array at row $p$ (for node $v_p$) and column $w_i - w_{(v_p,v_j)}$ (for weight $w_p$). In doing so, the algorithm solves the subproblem by reusing earlier solutions. Note also, that relevant subproblems are guaranteed to have been solved at an earlier point in the execution of the algorithm as the algorithm treats subproblems in the order of increasing complexity: subproblems are solved by first iterating over all path weights $w_i = 1, 2, \ldots, w_{crit}$ in an outer loop, and second iterating over all nodes from $v_s := v_1$ to $v_t := v_{|\mathbf{V}|}$ in an inner loop; a relevant subproblem $v_s \overset{w_p}{\rightsquigarrow} v_p$ is guaranteed to have been solved because it always holds that $w_p < w_i$ per definition of $w_p$.

To tabulate solutions to subproblems, for every edge solving the subproblem and its weight, we add a tuple $\left(w_{(v_p,v_j)}, v_p\right)$ to a set of solutions. More specifically, when iterating over all $N$ potential incoming edges $(v_p, v_j)$, we enter valid solutions in a sequence of sets indexed by the current algorithmic step $n$: $\mathbf{L}^0_{w_i,v_j} \subseteq \ldots \subseteq \mathbf{L}^n_{w_i,v_j} \subseteq \mathbf{L}^{n+1}_{w_i,v_j} \subseteq \ldots \subseteq \mathbf{L}^N_{w_i,v_j}$. Each tuple in $\mathbf{L}^N_{w_i,v_j}$ then indicates the existence of one alternative path for the subproblem. The collection later allows the reconstruction of all alternative paths for the overall problem.

Formally, after initially setting $\mathbf{L}^0_{w_i,v_j} = \emptyset$, we can define each algorithmic step $n$ as

*For every edge $(v_p, v_j)$ leading into $v_j$:*

$$\mathbf{L}^{n+1}_{w_i,v_j} := \begin{cases} \mathbf{L}^{n}_{w_i,v_j} \cup (w_{(v_p,v_j)}, v_p) & \text{if } v_j \neq v_s \wedge \mathbf{L}^{M}_{w_p,v_p} \neq \emptyset \\ \mathbf{L}^{n}_{w_i,v_j} & \text{if } \mathbf{L}^{M}_{w_p,v_p} = \emptyset \\ (0, v_s) & \text{if } v_j = v_s. \end{cases} \tag{3.3}$$

Here, $\mathbf{L}^{n}_{w_i,v_j}$ denotes the current set of tuples contributing to the solution of the subproblem $v_s \overset{w_i}{\rightsquigarrow} v_j$, i.e., a path leading from $v_s$ to $v_j$, that has a summed weight of $w_i$ in algorithmic step $n$. $\mathbf{L}^{M}_{w_i,v_j}$ is a set of solutions to the subproblem $v_s \overset{w_p}{\rightsquigarrow} v_p$ investigated earlier (where solutions were collected over $M$ algorithmic steps). Formula 3.3 expressed that for every edge $(v_p, v_j)$, it is tested if two conditions are met (see also Fig. 3.7B):

(1) there exists a solution to the previous subproblem $\mathbf{L}^{M}_{w_p,v_p}$, i.e., a path to the predecessor $v_p$ with weight $w_p = w_i - w_{(v_p,v_j)}$;

(2) the edge $(v_p, v_j)$ from predecessor to current node has a weight such that $w_p + w_{(v_p,v_j)} = w_i$.

If both conditions are met, then the tuple $\left(w_{(v_p,v_j)}, v_p\right)$ is added to $\mathbf{L}^{n}_{w_i,v_j}$. When all edges $(v_p, v_j)$ have been tested, the next subproblem $v_s \overset{w_i}{\rightsquigarrow} v_{j+1}$ (inner loop) or $v_s \overset{w_i+1}{\rightsquigarrow} v_j$ (outer loop) is considered. The algorithm terminates once all subproblems have been investigated.

**Backtracking.** In a second step, the algorithm uses backtracking to reconstruct relevant paths from the solution array returned by the dynamic programming step described in the last subsection (see Fig. 3.6F). Paths are considered relevant, if they lead to the current target node $v_t$ and have a summed weight of $w_{(v_s,v_t)} \pm \theta$. Thus, paths are reconstructed from all entries that correspond to the solutions to subproblems $v_s \rightsquigarrow v_t$ with weight $w_{(v_s,v_t)} \pm \theta$ (we call these relevant entries in the solution array *reconstruction interval*).

The backtracking algorithm uses depth first search (DFS, for a more detailed description see for example [194]) to reconstruct all paths starting from one entry in the reconstruction interval at a time; it is therefore called for each entry in the reconstruction interval individually (for example, in Fig. 3.6F, this corresponds to three calls of the backtracking algorithm for fields $[5, 6]$, $[6, 6]$ and $[7, 6]$). The backtracking is done by recursively expanding each entry in the currently considered field (i.e., visiting the next field, indicated by the currently considered solution to a subproblem). For example, in Fig. 3.6F, the field $[6, 6]$ points to the field $[5, 4]$, which points to $[3, 3]$ and so on. While expanding one path, the algorithm checks, whether a node in the currently reconstructed path has already been visited during the recursion. If this is the case, the path contains a loop, i.e., a node is visited twice in one path, and the respective path is discarded as it is not a valid solution to the overall problem.

All remaining reconstructed paths are considered alternative paths to the edge $(v_s, v_t)$. If at least one such alternative path exists, $(v_s, v_t)$ is considered potentially spurious due to a CE.

**Additional analysis of triangle motifs.** As laid out in the subsection *Rationale of the algorithm*, the algorithm identifies simple CD in an additional step. Simple CD occurs in triangle motifs, which can be identified by listing all edges with valid alternative paths of length two. In each triangle, the second edge of the alternative path is considered as potentially spurious due to a simple CD effect additionally to the edge considered spurious due to CE (see coupling motifs shown in Fig. 3.2C).

**Output.** The algorithm returns a list of potentially spurious edges in $\mathbf{E}$, and tags these spurious edges as a CE (identified by an alternative path) or a simple common drive effect (identified in a triangle motif).

## 3.3 Evaluation

To test the proposed algorithm for correctness and performance in terms of execution times, we simulated networks of different sizes and densities to serve as input graphs. To further demonstrate the algorithm's applicability to neuroscience data, we applied it to networks derived from electrophysiological recordings during a face recognition task.

### 3.3.1 Performance of the algorithm in simulated networks

We simulated networks of different types: small-world networks [195], scale-free networks [196] and random networks of different densities [197, 198]. We chose these topologies because small-worldness and scale-freeness have frequently been reported to occur in functional and anatomical networks derived from neuroscience experiments [190, 199, 200] (for a review see also [186] and [201]).

The performance of the proposed algorithm on the simulated networks was tested by (1) varying the size $|\mathbf{V}|$ of simulated networks and (2) varying the critical path weight for alternative paths $w_{crit}$. Higher values for both parameters increased computational demand by either increasing input size directly (higher values for $|\mathbf{V}|$) or by increasing the likelihood for the detection of an alternative path (higher values for $w_{crit}$).

The dependency of the likelihood of detecting an alternative path on $w_{crit}$ is especially relevant in random network topologies: here, the number of possible alternative paths increases exponentially in $w_{crit}$, as higher values for $w_{crit}$ (relative to individual path weights) increase the number of possible combinations of edges that form a path of weight $w_{crit}$. More precisely, given a random graph $\mathbf{G} = \{\mathbf{V}, \mathbf{E}\}$ and $w_{crit} > |\mathbf{V}|$, the probability for the existence of an edge is $p(e) = \rho = \frac{|\mathbf{E}|}{|\mathbf{V}|(|\mathbf{V}|-1)}$ (where $\rho$ is the density of $\mathbf{G}$); furthermore, the probability of the edge having weight $w$ is uniformly distributed with $p(w) = \frac{1}{w_{max}}$. The number of possible alternative paths can then be calculated as

$$\sum_{w'=1}^{w_{crit}} \left[ \sum_{j=1}^{w'} \left( \binom{w'-1}{j-1} \cdot p(w)^j p(e)^j \right) \right], \tag{3.4}$$

where $\sum_{j=1}^{w'} \binom{w'-1}{j-1}$ is the number of compositions, i.e., the number of ordered sequences of integers that sum to $w'$, the currently considered summed edge weight. The inner sum is dominated by the growth of the number of compositions given by $\sum_{j=1}^{i} \binom{i-1}{j-1} = 2^{i-1}$. Thus, the number of alternative paths grows in $\Omega(2^{w_{crit}})$, i.e., it grows exponentially in the critical path weight given a sufficiently dense, random network topology.

**Small-world networks.** For the simulation of small-world networks we modified the rule for network generation proposed by Watts and Strogatz in [195]. The network generation was done in two steps with parameters $|\mathbf{V}|$ (number of nodes), $n$ (neighborhood coefficient) and $p$ (rewiring probability):

1. Construct a regular ring lattice with $\mathbf{V}$ nodes, where every node $v_i \in \mathbf{V}$ is connected to its $n$ nearest neighbors $v_j$ (such that $(v_i, v_j) \in \mathbf{E}$). Given that $\mathbf{V} = v_1, \ldots, v_{|\mathbf{V}|}$, each node $v_i$ is connected to its neighbors $v_{i+1}, \ldots, v_{i+n/2}$ and $v_{i-1}, \ldots, v_{i-n/2}$.

2. For every node $v_i$ all edges $(v_i, v_j)$ are rewired with probability $p$ by replacing $(v_i, v_j)$ with $(v_i, v_k)$ where $n$ is chosen randomly with uniform probability from $\mathbf{V} \backslash v_i$ while avoiding loops and multiple edges.

3. Every edge is weighted by a random weight $w_{(v_i,v_j)}$ drawn from an interval $[1; w_{max}]$ with uniform probability

Note, that we made two extensions to the original generation rule proposed in [195]: we simulated network edges as directed and weighted, while the original work by Watts and Strogatz assumes undirected and unweighted edges. We defined connections from every node $v_i$ to its $n$ nearest neighbors as directed and weighted them with values randomly drawn from an interval $[1; w_{max}]$. We set $w_{max}$ to the maximum interaction delay found in an MEG data set also used as a second test case described below. Thus, strictly speaking, only the undirected and unweighted network underlying our test cases had small-world properties (i.e., a high clustering coefficient and a short characteristic path length). We used this approximation of small-world properties in a weighted and directed network, as there is no agreement over how directedness and edge weights are to be incorporated into the original notion of small-worldness (as both parameters may alter the global behavior commonly observed in undirected and unweighted small-world networks) [202].

**Scale-free networks.** Scale-free networks were simulated using an algorithm proposed in [203], following an implementation of the rationale by Barabási and Albert[196] in [204] (see also [205]).

Scale-free networks resemble small-world networks in their topology, i.e., they exhibit high local clustering and low characteristic path lengths. Both network types differ however in their degree distributions $p(n)$, the probability that a node interacts with $n$ other nodes in the network: in small-world networks the degrees are normally distributed, while in scale-free networks the degrees follow a power law $p(n) \sim n^{-\gamma}$ (where $\gamma$ may vary for different networks) [196].

**Random networks.** We created random networks of size $|\mathbf{V}|$ by independently including weighted edges with probability $\rho$ [198], where $\rho$ denotes the density or edge probability of a graph:

$$\rho = \frac{|\mathbf{E}|}{|\mathbf{V}|(|\mathbf{V}| - 1)}, \qquad (3.5)$$

i.e., the the ratio of edges actually present in a graph to the number of possible edges [206]. Compared to small-world networks, a random graph typically exhibits a small average minimum path length and small average clustering coefficient [202], depending on the graph's density. In the present study, we created two test cases with $\rho = 0.25$ and $\rho = 0.50$ respectively.

**Performance results for simulated networks.** Running times of the algorithm increased as a function of graph size $|\mathbf{V}|$ and critical path weight $w_{crit}$ (Fig. 3.8). For the dynamic programming part of the algorithm, running times increased in a linear fashion in both network size $|\mathbf{V}|$ (Fig. 3.8A) and critical path weight (Fig. 3.8B). Running times thus correspond to theoretically expected running times. The time needed for backtracking the obtained solutions grew exponentially in $|\mathbf{V}|$ (Fig. 3.8C) and critical path weight (Fig. 3.8D) for random networks and scale-free networks, where running times were least favorable for random networks. For small-world networks on the other hand, running times did not increase dramatically for higher values $|\mathbf{V}|$ and $w_{crit}$.

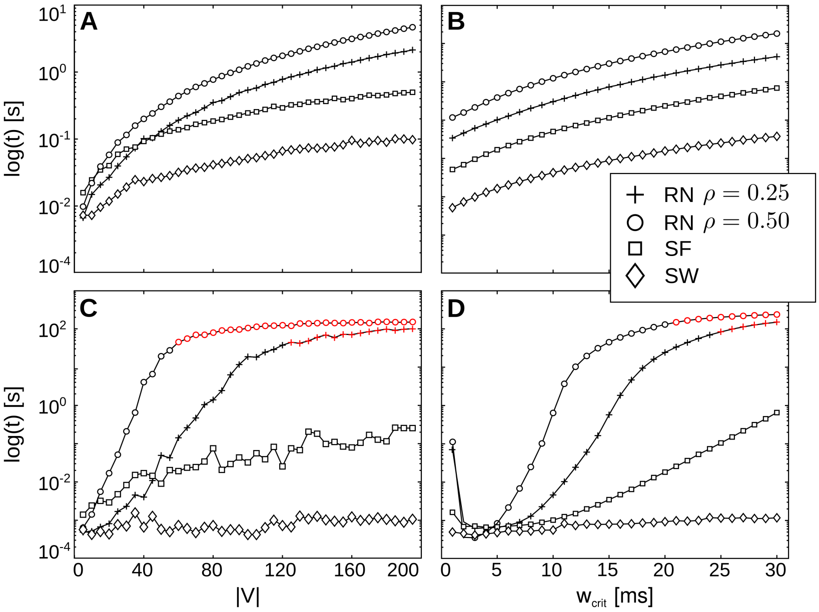


Fig. 3.8 **Results running time.** Running times [log(s)] for dynamic programming (A, B) and backtracking (C, D) by number of vertices $|\mathbf{V}|$ and maximum path weight $w_{crit}$. Running times are shown for different graph types (SW: small-world, SF: scale-free, RN: random networks with density $\rho$). Red markers indicate cases of intractability (execution was aborted after a pre-defined limit of reconstructed alternative paths was reached).

Running times for backtracking depend on the number of alternative paths to be reconstructed from the solution array. Since the number of paths increases exponentially in $w_{crit}$ (given a sufficient graph size $|\mathbf{V}|$), exponential running times were expected for higher values for both parameters. We therefore defined an a priori limit of 20 000 for the number of alternative paths to be reconstructed. If this limit was reached, the algorithm's execution was aborted. These problem instances were considered intractable (red markers in Fig. 3.8). Intractable test cases were found for random graphs only and occurred earlier for graphs with higher densities of $\rho = 0.50$ (for comparison: the scale-free graphs had a density of approx. 0.15, small-world graphs of approx 0.5). Cases of intractability were found for random graphs of sizes $|\mathbf{V}| \geq 65$ with a density $\rho = 0.50$ and $|\mathbf{V}| \geq 130$ for graphs with density $\rho = 0.25$ respectively. Furthermore, intractable cases occurred for path weights $w_{crit} \geq 21$ for random graphs with density $\rho = 0.50$ and for path weights $w_{crit} \geq 25$ for random graphs with density $\rho = 0.25$.

Thus, network size as well as network structure influenced the computational demand of a given input to the presented algorithm. Note that intractable cases may well occur in a neuroscience application, where inputs can not be assumed to be bounded in any respect (e.g. in terms of graph density or graph size). Here, the network size may be used to determine whether an input may prove intractable. In the present simulation, network sizes smaller than 25 nodes posed no problem for the algorithm; of course, these limits are subject to moderate changes with increasing computational power.

### 3.3.2 Detection of spurious interactions in networks derived from electrophysiological time series

**Ethics statement.** To test the algorithms applicability to biological time series, we used MEG data recorded from 15 healthy human subjects during a face recognition task as described in [148]. All subjects gave written informed consent before the experiment. The study was approved by the local ethics committee (Johann Wolfgang Goethe University, Frankfurt, Germany).

**Preparation and MEG data acquisition.** MEG data was obtained from 30 healthy subjects, recruited from the local community. All participants had normal or corrected-to-normal vision and were right-handed (assessed by the Edinburgh Handedness Inventory [207]).

MEG data were recorded using a 275-channel whole-head system (Omega 2005, VSM MedTech Ltd., BC, Canada) at a rate of 600 Hz in a synthetic third order axial gradiometer configuration (Data Acquisition Software Version 5.4.0, VSM MedTech Ltd., BC, Canada). Data were filtered with 4th order butterworth filters with 0.5 Hz high-pass and 150 Hz low-pass. Behavioral responses were recorded using a fiber-optic response pad (Lumitouch, Photon Control Inc., Burnaby, BC, Canada). Trials with excessive head movement (more than 5 mm) were excluded from further analysis.

Structural magnetic resonance images were obtained with a 3 T Siemens Allegra, using 3D magnetization-prepared rapid-acquisition gradient echo sequence. Anatomical images were used to create individual head models for MEG source reconstruction.

**Task.** The participants were presented with a randomized sequence of degraded two tone images of human faces (Mooney Faces, [149], see Fig. 3.9C for an example stimulus) and scrambled stimuli, where black and white patches were randomly rearranged to minimize the likelihood of detecting a face. The participants had to indicate the detection of a face by button press. Only trials in which faces were correctly identified entered further analysis.

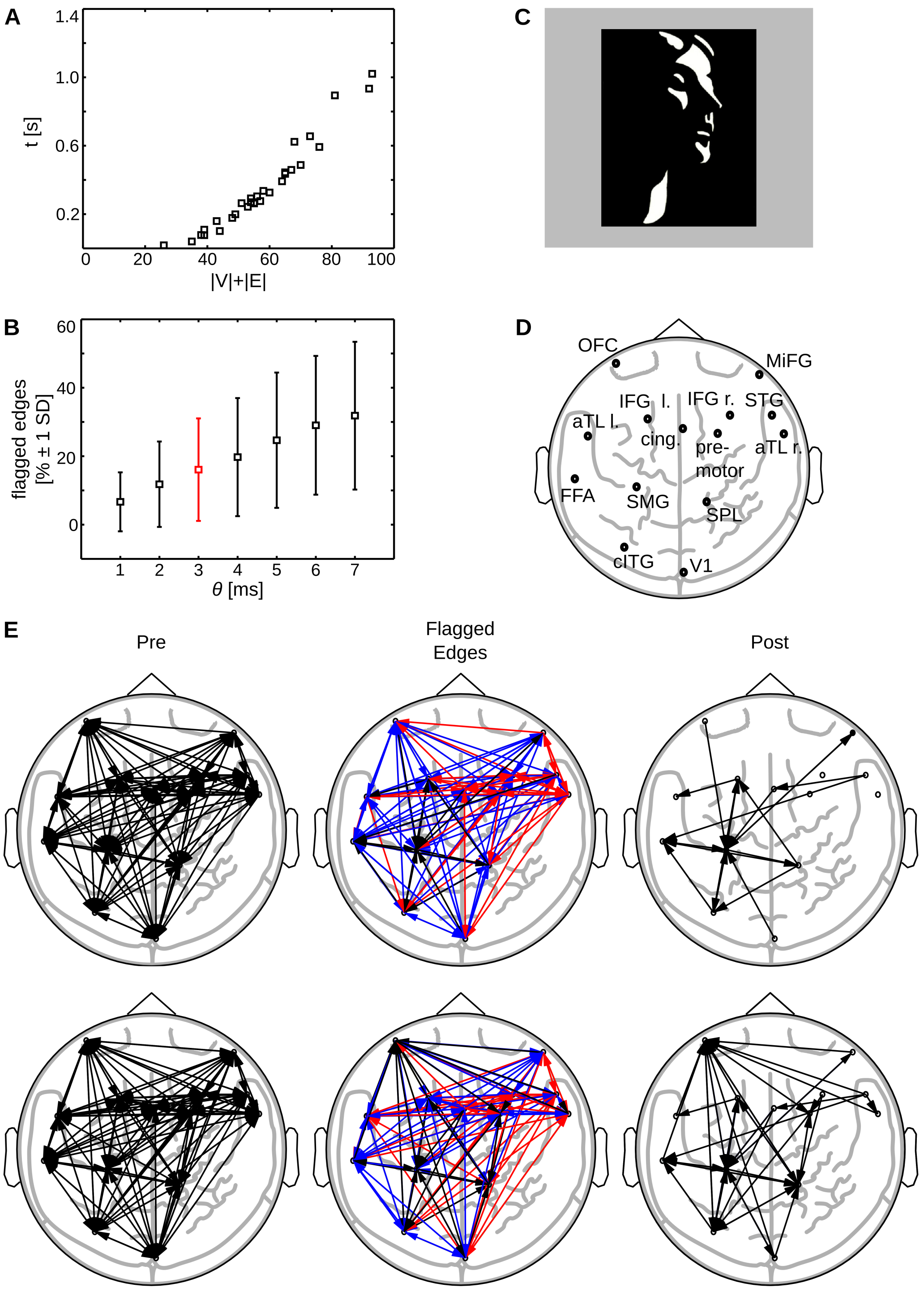


Fig. 3.9 **Results empirical data sets.** (A) Running time of the complete algorithm by number of nodes plus number of edges $|\mathbf{V}| + |\mathbf{E}|$; (B) Mean percentage of tagged, potentially spurious edges by chosen threshold $\theta$ after application of the algorithm, error bars indicate 1 standard deviation (SD); the value for $\theta$ obtained from bootstrapping in two example data sets is marked in red; (C) Mooney Stimulus [149]; (D) Cortical sources after beamforming of MEG data (l.,left; r., right: l. orbitofrontal cortex (OFC); r. middle frontal gyrus (MiFG); l. inferior frontal gyrus (IFG left); r. inferior frontal gyrus (IFG right); l. anterior inferotemporal cortex (aTL left); l. cingulate gyrus (cing); r. premotor cortex (premotor); r. superior temporal gyrus (STG); r. anterior inferotemporal cortex (aTL right); l. fusiform gyrus (FFA); l. angular/supramarginal gyrus (SMG); r. superior parietal lobule/precuneus (SPL); l. caudal ITG/LOC (cITG); r. primary visual cortex (V1)), see also [148];

Fig. 3.9 **Results empirical data sets (continued).**
(E) Example of removal of tagged edges: MEG data of a face detection task in two subjects. First column shows transfer entropy values prior to detection of potentially spurious edges (**Pre**). The second column shows color-coded tagged edges (red: Potential cascade effects, blue: potential common drive effects; $\theta = 3ms$). The third column shows the network of directed interactions after removal of all tagged edges (**Post**).

**Data analysis.** MEG data were analyzed using MathWorks® MATLAB® (2008b, The MathWorks, Natick, MA) and the open source MATLAB® toolboxes FieldTrip (version 2008-12-08; [150]), SPM2 (`http://www.fil.ion.ucl.ac.uk/spm`), and TRENTOOL [86]. We will briefly describe the applied analysis here, for a more in depth treatment refer to [148].

For data preprocessing, we defined experimental trials from the continuously recorded MEG data. A trial was defined as the epoch from −1000 ms prior to stimulus presentation until 1000 ms after stimulus presentation. Trials contaminated by artifacts (eye blinks, muscle activity, or jump artifacts in the sensors) as well as trials with wrong responses were discarded. Trials were baseline corrected by subtracting the mean amplitude between −500 ms to −100 ms before stimulus onset.

To investigate differences in source activation in the face and non-face condition, we used a frequency domain beamformer [151] at frequencies of interest identified at the sensor level (80 Hz with a spectral smoothing of 20 Hz). We computed the frequency domain beamformer filters for combined trials ("common filters") consisting of activation (multiple windows, duration, 200 ms; onsets at every 50 ms from 0 ms to 450 ms) and baseline data (−350 ms to −150 ms). To compensate for the short duration of the data windows, we used a regularization of $\lambda = 5\ \%$ [152].

To find significant source activation in the face versus non-face condition, we first conducted a within-subject t-test for activation versus baseline effects. Next, the t-values of this test statistic were subjected to a second-level randomization test at the group level to obtain effects of differences between face and no-face conditions; a p-value <0.01 was considered significant. We identified 14 sources with differential spectral power between both conditions in the frequency band of interest in occipital, parietal, temporal, and frontal cortices (see Fig. 3.9D and [148] for exact anatomical locations). Namely, our network representing information flow between sources has 14 nodes. We then reconstructed source time courses for TE analysis, this time using a broadband beamformer with a bandwidth of 10 Hz to 150 Hz.

We estimated TE between beamformer source time courses [86, 208] within an analysis window of 500 ms ($-50$ ms to 450 ms) and tested resulting TE values for their statistical significance [86]. We furthermore reconstructed information transfer delays for significant information transfer by scanning over a range of assumed interaction delays from 5 ms to 17 ms (resolution 2 ms), following the approach in [75] and parameters used in a similar analysis in [208]. We thus obtained a delay weighted, directed network of information transfer during a face recognition task, consisting of 14 nodes and edges with weights in the range from 5 to 17. We then applied the proposed algorithm to the resulting delay-weighted networks of directed interactions. For two example data sets, we used bootstrapping (1000 resampled cases) to obtain an estimate of the standard error of the delay estimation [209] and used this estimated standard error as input parameter $\theta$ for a more detailed example application of the algorithm.

**Performance results for “empirical” networks.** For empirical data running times increased almost linearly in $|\mathbf{V}| + |\mathbf{E}|$ (Fig. 3.9A). We chose to present running times as function of the sum of the number of nodes and number of edges because a systematic variation of network size was not possible here (rather, network size was determined by previous source reconstruction). Cases of intractability did not occur even though some data sets exhibited high network densities (ranging from 0.07 to 0.43 with a mean of 0.24 and a SD of 0.09).

The percentage of potentially spurious edges increased with higher thresholds $\theta$ up to 32 % of potentially spurious edges for $\theta = 7ms$ (Fig. 3.9B). Edges were considered potentially spurious if at least one alternative path existed or a simple CD was present. Note, that the threshold serves to adjust the algorithm’s sensitivity and may lead to the erroneous exclusion of edges if chosen too high. Thus, the value for $\theta$ should be chosen such that imprecisions in interaction reconstruction are accounted for, while the false discovery rate is not increased. As a rule of thumb, a user may use prior knowledge about the minimum interaction delay to be expected in the data as an upper bound for $\theta$ or use bootstrapping to obtain an error estimate for reconstructed delays.

In Fig. 3.9E, we show results for two example MEG data sets from the validation test-set before and after analysis with the proposed algorithm. We used bootstrapping to estimate the standard error in the reconstruction of the interaction delays. We found an average error over channels of 2.6 ms for subject one and 2.9 ms for subject two. Accordingly, we set $\theta = 3ms$ as this corresponded to the next integer value in ms. The average reconstructed interaction delay was found to be 7.06 ms (SD: 3.57 ms) for subject one and 6.94 ms (SD: 3.32 ms) for subject two. We also calculated the average path length as the average weight of the shortest path for each node to every other node in the network; the average path length was 6.84 ms for subject one and 6.81 ms for subject two. Note that the graphs are highly connected prior to the application of the algorithm, such that the shortest path between any two nodes consists of just one edge; thus the average path length is close to the average edge weight.

After application of the algorithm, the recovered networks of information transfer consisted of 20 links for subject one and 34 links for subject two; networks showed an overlap of nine edges, which corresponds to an 45 % overlap and is 20 times higher than an overlap of 2 % expected purely by chance. Thus, the network can be considered highly consistent.

## 3.4 Discussion

### 3.4.1 Algorithmic detection of potentially spurious edges in delay weighted networks

We have presented an algorithm that finds potentially spurious links arising from bivariate analysis of multi-node networks based on interaction timing. The algorithm identifies the most common motifs causing the reconstruction of spurious links, such that identified links can be subjected to further testing, or removed. By removing all potentially spurious edges, the user obtains a sub-network that is guaranteed to contain only non-spurious edges; this improves the validity of the network representation itself as well as the validity of potential subsequent network analysis. The algorithm thus allows the user to find an approximate representation of multivariate interactions in the data, using only bivariate interaction reconstruction and avoiding the computationally heavy problem of an approximately or even fully multivariate approach.

The presented algorithm may be used in neuroscience to post-process any network of reconstructed bivariate interactions, where interactions are directed and weighted by their estimated delays. We demonstrated the application of the algorithm using a reference implementation in MATLAB® as part of the open source toolbox TRENTOOL [86].

### 3.4.2 Application in neuroscience

Based on findings in [75], we propose to identify spurious links by their characteristic timing signatures in networks of reconstructed bivariate interactions [121]. In particular, we propose that a link is likely to be spurious if an alternative path with identical timing exists (Fig. 3.2).

We assume that a bivariate information transfer between two nodes and a corresponding alternative path constitute a redundant routing of information. Such a redundant routing conflicts with the hypothesis that the brain evolves under the objective of maximizing economy and efficacy [185, 210] while minimizing biological costs [211, 212] (see for example the “save-wire hypothesis” in [213]). We thus argue that any redundant routing of information between two sources of neural activity—with identical timing—would be implausible, given the brain’s organizational principles. Therefore, whenever a redundant routing for a bivariate information transfer is found, our rationale implies spuriousness of either the bivariate information transfer *or* the alternative path. Of the two, we consider the bivariate interaction spurious, because (1) spurious bivariate interactions are a likely artifact in bivariate analysis of multi-node networks; and (2) if a bivariate and thereby *direct* means of information transfer between a source and a target existed, the maintenance of a physiologically more costly alternative path of identical information transfer would be unlikely.

Note, that this rationale exclusively applies to neural systems. Also remember, that the algorithm does not tag *all* alternative routings of information, but only those with a certain timing signature; alternative routings with different delays than the bivariate interaction are not considered redundant and are not tagged.

### 3.4.3 Treatment of tagged links in neuroscience applications

Our algorithm tags potentially spurious edges to let the user decide if a tagged edge should be ultimately excluded from the network representation. To minimize erroneous exclusions, the user may inform the decision by additional evidence, e.g. previous anatomical or functional findings. If such previous findings do not exist, we recommend the exclusion of tagged edges.

We consider the erroneous rejection of links favorable over erroneous inclusion, i.e., we suggest to rather commit a false negative error if in doubt. We favor false negative errors because in statistical terms, false negatives are considered less severe than false positives (erroneously including a spurious link), as they yield more conservative results. If the user removes all tagged links, the resulting network is guaranteed to contain non-spurious links only, but some links may be missing from the network.

In triangle motifs, the exclusion of all tagged links will definitely lead to false negatives: Here, two links are tagged but the exclusion of both edges is mutually exclusive; more precisely, the exclusion of one of the two tagged edges destroys the motif giving rise to the second, potentially spurious link. This is illustrated in Fig. 3.2C, where the exclusion of link $(v_1, v_t)$ destroys the CE leading to the tagging of edge $(v_s, v_t)$, and on the other hand, the exclusion of $(v_s, v_t)$ destroys the CD leading to the tagging of $(v_1, v_t)$. In triangle motifs, the rejection of both edges thus produces a false negative error; here, prior anatomical or functional evidence are required to decide which of the two tagged edges is non-spurious.

It is further possible to use modeling approaches to test if tagged links are actually present in the network of bivariate interactions. For example dynamic causal modeling (DCM, [214–216]) may be used to test whether a model containing a certain link is favorable given the observed data over a second model missing this link.

### 3.4.4 Types of multivariate effects not identified by the algorithm

The correction performed by the presented algorithm is not exhaustive with respect to all types of multivariate interactions potentially occurring in neuroscience data. The interactions not targeted by the algorithm are of two types: (1) more general cases of CD; (2) synergistic effects [72], i.e., combined effects of two or more sources on a third source. In the following we will discuss the conceptual and practical limitations that prevent an exhaustive algorithmic correction for these two types of multivariate interactions.

**Detection of general common drive effects.** The presented algorithm detects *simple* CD in triangle motifs by listing all links with alternative paths of length two. Our algorithm can theoretically be extended to explicitly search for general cases of CD, where two nodes are commonly driven via arbitrarily long cascades of information transfer (Fig. 3.2A).

General CD may be identified by searching for paths of equal summed interaction delays that have a common source and target node. We again assume that equal summed delays hint at redundant and therefore spurious information transfer. It can then be tested if the source node is a common driver for the last and second to last node in one of the two paths by looking at the information transfer delay between these last two nodes. For an example, see Fig. 3.2A, where the bivariate information transfer in the network forms two paths of equal delays connecting nodes $v_0$ and $v_t$: The link $(v_s, v_t)$ is tagged as spurious because its weight corresponds to the difference in the summed path weights $v_0 \overset{c'}{\rightsquigarrow} v_s$ and $v_0 \overset{c}{\rightsquigarrow} v_t$: $w_{(v_s,v_t)} = c - c'$. In this scenario, $v_0$ is a common driver of nodes $v_s$ and $v_t$. This approach also allows to test for higher order CD, i.e., one source driving three or more nodes simultaneously. A similar algorithm was proposed by Marinazzo and colleagues [217] to identify spurious bivariate links using a network of multivariately reconstructed interactions (see next section).

Even though an extension to general CD is hypothetically possible, its realization is not feasible in practice: The extension requires that for each network node all originating paths of arbitrary length need to be listed. An algorithm fulfilling this task would have an asymptotic running time many times higher than the algorithm presented in this work: For each node in $\mathbf{V}$, $O(|\mathbf{V}-1|!)$ paths of arbitrary length and weight exist (in the first step $O(|\mathbf{V}|-1)$ nodes can be reached from the current starting node, in the second step $O(|\mathbf{V}|-2)$ nodes can be reached, and so forth). The asymptotic running time of such an algorithm thus amounts to $O(|\mathbf{V}|\cdot|\mathbf{V}-1|!)$. Such a factorial running time is commonly not considered feasible in practice and would limit the application to networks of very small size.

**Detection of synergistic effects.** Synergistic effects describe information that is transferred from a set of sources to a common target, whereby information is combined in a non-trivial fashion [72, 76, 79, 81]. In this case, looking at the set of sources simultaneously provides information about the target that is not obtainable from looking at each source separately. As a toy example, one can think of three nodes implementing a logical XOR operator, where two nodes serve as binary input and the third node serves as output node. Each state of the output node is the exclusive OR of the two previous input states. A bivariate analysis of every pairwise interaction between the three nodes will not detect any significant interaction, because the pairwise mutual information between any two nodes is 0. Analyzing the triplet of nodes simultaneously will however detect an interaction; e.g. the conditional mutual information (TE) will be greater than zero, because it "decodes" the information in one source by conditioning on the second source.

Consequently, synergistic effects between a set of sources and a target node can only be revealed if the whole set is considered simultaneously in some multivariate reconstruction of interactions or by explicitly reconstructing synergistic interactions [76, 80]. Such synergistic effects are not targeted by design of our algorithm as it simply post-processes results from bivariate network analyses. To include synergistic effects, a multivariate interaction analysis would have to replace the estimation of bivariate interactions. Note however, that any fully multivariate method for interaction reconstruction would need to identify the optimal subset of sources that exert some meaningful influence on a given target node. The identification of such an optimal set of sources would require the exhaustive testing of the power set $\mathcal{P}(\mathbf{V})$ of all network sources, due to the non-additivity of information contributions from individual sources (because of redundant and synergistic effects). The power set has size $|\mathcal{P}(\mathbf{V})| = 2^{|\mathbf{V}|}$, i.e., testing all sources brute force has a theoretical running time of $O\left(2^{|\mathbf{V}|}\right)$. In fact, it has been shown that optimal subset selection in regression is an NP-hard problem [27]. This proof extends to source selection for TE due to the equality of TE with Granger causality for *jointly* Gaussian variables [103]. Thus, the reconstruction of truly multivariate interactions in arbitrarily large networks poses a computationally intractable problem (if P $\neq$ NP). In the next subsection we will present approaches that try to approximate fully multivariate methods to circumvent the inherent computational complexity of the problem at hand.

### 3.4.5 Comparison to other approximative methods for multivariate network reconstruction

The proposed algorithm provides an approximative method for the inference of networks of multivariate interactions to handle the computational intractability of exact network reconstruction. Methods with a similar purpose have been proposed by various authors. In the following we will review some of these methods and list scenarios that may benefit from the application of our algorithm.

**Multivariate reconstruction of effective networks by Lizier and Rubinov.** Lizier and Rubinov [25] proposed a greedy algorithm which for each network node $Y$ (the target) infers a set of influential source variables $\mathbf{V_Y}$. A source is considered influential if it adds significantly to the information transfer from $\mathbf{V_Y}$ to $Y$. The set $\mathbf{V_Y}$ is thus built by iteratively adding sources, which have significant information transfer into $Y$, conditional on the previously included sources. Finally, information transfer is re-evaluated conditional on the complete set of included sources:

$$TE(X \rightarrow Y|\mathbf{V_Y}) = I\left(X;Y|\mathbf{V_Y}\backslash X\right),$$

i.e., the mutual information between each source $X \in \mathbf{V_Y}$ and target $Y$ while conditioning on all remaining relevant sources in $\mathbf{V_Y}$, except $X$. If a source fails to provide statistically significant information about the present of $Y$, it is removed from $\mathbf{V_Y}$. After this "pruning step", the set $\mathbf{V_Y}$ consists of all relevant sources that contribute information about the target. The approach is robust against the detection of spurious interactions due to CD and CE, because for each interaction reconstruction it conditions on all relevant sources in the network.

Note that the greedy strategy used by Lizier and Rubinov is approximative insofar as it does not guarantee a maximal informative set $\mathbf{V_Y}$ over all sets $\mathcal{P}(\mathbf{V})$. For example, purely synergistic effects between two or more sources may be missed. The authors propose to extend their greedy method by also testing tuples and higher order combinations of sources, but they note that this requires considerable computational resources, which may not be worth the gain in information. The testing of tuples may however be feasible for small networks.

**Partial conditioning of information transfer by Marinazzo and colleagues.** Marinazzo and colleagues [217] proposed a greedy algorithm resembling the approach by Lizier and Rubinov. Again, the algorithm tries to account for other relevant network sources when evaluating the information transfer from a source $X$ to a target $Y$ in a multivariate system. To identify relevant sources, Marinazzo et al. propose to iteratively construct a "partial conditioning set" $\mathbf{Z}$ from all sources $\mathbf{V}\backslash\{X,Y\}$. In each iterative step $k$ the algorithm includes the source $Z_k$ that maximizes the mutual information between $\mathbf{Z_k}$ and the source $X$, i.e., $Z_k = \max_Z\left(I(X;\mathbf{Z_k})\right)$, where $\mathbf{Z_k} = \mathbf{Z_{k-1}} \cup Z$.

The partial conditioning approach may miss interactions if sources share a lot of redundant information about a target: If an existing source-target relationship is evaluated while conditioning on sources providing redundant information about the target, the source-target relationship under investigation is not detected. Therefore, Stramaglia and colleagues [218] extended partial conditioning with a graph algorithm that identifies these missed interactions. The authors proposed to reconstruct interactions multivariately (e.g. with partial conditioning) and bivariately. The multivariate network is then used to algorithmically separate bivariate links in two sets: (1) links explained by an indirect path of information transfer in the multivariate network (CE), and (2) links not explained by an indirect path. Bivariate links in the second group, which are missing from the multivariate network, are assumed to reflect non-spurious information transfer that was missed by the multivariate approach. These bivariate links are then merged with the multivariate network.

Note that the rationale underlying the algorithm proposed in [218] resembles the rationale presented in this paper because spurious bivariate interactions are identified by alternative paths; however, the aim of both approaches differs: The algorithm presented by Stramaglia and colleagues improves multivariate interaction reconstruction, while the algorithm presented in this paper tries to approximate a multivariate approach from bivariate interaction reconstruction alone. The approach proposed by Stramaglia thus still requires the potentially intractable reconstruction of multivariate interactions from data.

**Non-uniform multivariate embedding by Faes and colleagues.** Faes and colleagues [74] proposed a non-uniform embedding technique to estimate the information transfer from one source variable to a target in the presence of further potentially relevant sources of information transfer. The authors propose to iteratively build a non-uniform embedding vector from a set of candidate time points from the past of all sources in a network up to a certain predefined limit. Points are included in the vector if they add significant information about the next state of the target. Information transfer between a source and a target may then be estimated while conditioning on this non-uniform embedding vector.

The iterative construction of the embedding vector follows a greedy strategy that is similar to the strategies discussed above [25, 217]. Accordingly, the returned embedding vector is not guaranteed to yield the set of maximally informative source time points with respect to the target, as it will miss purely synergistic contributions of two or more points to the target. As said above, an exhaustive testing of all possible subsets of source time points poses an intractable computational problem. This is explained next.

**Exhaustive brute force analysis.** A brute force analysis of interactions between all possible subsets ($m$-tuples) in a set of nodes would yield an exact solution to the problem of inferring the network of multivariate interactions from data. For the example of $m = 3$, one would enumerate all 3-tuples or triplets in the set of nodes and for each triplet evaluate the six possible interaction motifs—three potential targets and for each target two possible combinations of source and conditioning node. Note that here the mandatory conditioning takes care of potential synergies. For the general case of $m$-tuples, this would generalize to $\binom{|\mathbf{V}|}{m}$ possible subsets of size $m$, where for each tuple $m \cdot (m-1)$ possible interaction motifs exist. As for approximative approaches, such an analysis is feasible for small numbers of $m$ only.

## 3.4.6 Application scenarios for the proposed algorithm

The most basic application scenario for the proposed algorithm is as post-processing step after bivariate reconstruction of directed, delay-weighted interactions from a set of neural sources. Here, the algorithm helps to prune potentially spurious edges to obtain an approximative, statistically conservative network representation of the physical interactions. In this scenario, our algorithmic correction is favorable over multivariate interaction reconstruction whenever available data is limited or high-dimensional, such that data points are not sufficient to estimate highly multivariate interactions. Here, our algorithm is more data-efficient because it relies on bivariate interaction reconstruction only. Such data-efficiency is especially relevant for information theoretic measures, where quantities are often estimated using kernel estimators or neighbor methods (as for example proposed in [83]); applying these kernel estimators to the estimation of highly multivariate data may lead to very high dimensional search spaces, which suffer from the curse of dimensionality, hindering a reliable estimation of the quantities of interest.

The application of our algorithm may prove beneficial prior to calculating graph theoretical measures from networks of reconstructed interactions. We argue that these measures are more reliable when applied to a statistically conservative network representation.

The algorithm may further be used in conjunction with modeling approaches such as DCM, where it serves to limit the model space to be tested. DCM may also help to identify the most plausible network representation from models, after in- and excluding individual tagged edges respectively.

The presented algorithm may further serve as a preprocessing step for trivariate estimation of information transfer: By testing only the triangle motifs identified by the algorithm, the number of necessary information transfer estimations reduces drastically compared to the brute-force approach discussed in the last subsection. Necessary estimations include bivariate interaction reconstruction ($|\mathbf{V}| \cdot (|\mathbf{V}| - 1)$ calculations for $|\mathbf{V}|$ nodes) and subsequent multivariate interaction reconstruction for identified triangle motifs. The actual number of multivariate reconstructions depends on the number of motifs: The approach is asymptotically faster if 90 % or less of the possible $\binom{|\mathbf{V}|}{3}$ triangle motifs are actually present in the data (for network sizes $|\mathbf{V}| > 12$). Trivariate estimation of TE has been implemented in TRENTOOL [86], which also includes the reference implementation of the proposed algorithm. Thus, both analyses may be used in conjunction to estimate multivariate TE of order three. An extension to higher orders is theoretically possible although not implemented as it is not deemed feasible for practical purposes.

Finally, the algorithm is especially suitable for the application in simulated networks where all information transfers have a delay of unity, such as elementary cellular automata, and spurious interactions are therefore easily found.

# The relation of local entropy and information transfer suggests an origin of isoflurane anesthesia effects in local information processing

Patricia Wollstadt[1,*] Kristin K. Sellers[2,3], Lucas Rudelt[4], Viola Priesemann[4,5,*], Axel Hutt[6], Flavio Fröhlich[2,3,7,8,9,10], Michael Wibral[1]

**1** MEG Unit, Brain Imaging Center, Goethe University, Frankfurt am Main, Germany

**2** Department of Psychiatry, University of North Carolina at Chapel Hill, Chapel Hill NC, USA

**3** Neurobiology Curriculum, University of North Carolina at Chapel Hill, Chapel Hill NC, USA

**4** Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany

**5** Bernstein Center for Computational Neuroscience, BCCN, Göttingen, Germany

**6**Deutscher Wetterdienst, Section FE 12 - Data Assimilation, Offenbach am Main, Germany

**7** Department of Cell Biology and Physiology, University of North Carolina at Chapel Hill, Chapel Hill NC, USA

**8** Department of Biomedical Engineering, University of North Carolina at Chapel Hill, Chapel Hill NC, USA

**9** Neuroscience Center, University of North Carolina at Chapel Hill, Chapel Hill NC, USA

**10** Department of Neurology, University of North Carolina at Chapel Hill, Chapel Hill NC, USA

∗ patricia.wollstadt@stud.uni-frankfurt.de

∗ viola@nld.ds.mpg.de

*Submitted as: P. Wollstadt, K. K. Sellers, L. Rudelt, V. Priesemann, A. Hutt, F. Fröhlich, M. Wibral (2016). "The relation of local entropy and information transfer suggests an origin of isoflurane anesthesia effects in local information processing". PLoS Computational Biology (preprint available from arXiv: arXiv:1608.08387).*

# Abstract

The disruption of coupling between brain areas has been suggested as the mechanism underlying loss of consciousness in anesthesia. This hypothesis has been tested previously by measuring the information transfer between brain areas, and by taking reduced information transfer as a proxy for decoupling. Yet, information transfer is influenced by the amount of information available in the information source—such that transfer decreases even for unchanged coupling when less source information is available. Therefore, we reconsidered past interpretations of reduced information transfer as a sign of decoupling, and asked whether impaired local information processing leads to a loss of information transfer. An important prediction of this alternative hypothesis is that changes in locally available information (signal entropy) should be at least as pronounced as changes in information transfer. We tested this prediction by recording local field potentials in two ferrets after administration of isoflurane in concentrations of 0.0 %, 0.5 %, and 1.0 %.

We found strong decreases in the source entropy under isoflurane in visual area V1 and the prefrontal cortex (PFC)—as predicted by the alternative hypothesis. The decrease in source entropy was more pronounced in PFC compared to V1. In addition, information transfer between V1 and PFC was reduced bidirectionally, but with a more pronounced decrease from PFC to V1. This links the stronger decrease in information transfer to the stronger decrease in source entropy—suggesting that the reduced source entropy reduces information transfer. Thus, changes in information transfer under isoflurane seem to be a consequence of changes in local processing more than of decoupling between brain areas. Our results fit the observation that the synaptic targets of isoflurane are located in local cortical circuits rather than on the synapses formed by interareal axonal projections.

## 4.1 Introduction

To this day it is an open question in anesthesia research how general anesthesia leads to loss of consciousness (LOC). Several recent theories agree in proposing that anesthesia-induced LOC may be caused by the disruption of long range inter-areal information transfer in cortex [31, 219–222]—a hypothesis supported by a series of recent studies [31–35]. In all of these studies, information transfer is quantified using transfer entropy [12], an information theoretic measure, which has become a quasi-standard for the estimation of information transfer in anesthesia research, or by transfer entropy's linear implementation as a Granger causality. In many of these studies reduced information transfer has been interpreted as a sign of inter-areal long range connectivity being disrupted by anesthesia.

Yet, information transfer between a source of information and a target depends on information (entropy) being available at the source in the first place. Considering constraints of this kind we can easily conceive of cases where a decrease in information transfer under anesthesia is observed despite unchanged long range coupling, e.g., when the available information at the source decreases due to an anesthesia-related change in *local* information processing. Ultimately, this dissociation between information transfer and causal coupling just reflects that information transfer is *one* possible consequence of physical coupling, but not identical to it [68, 92, 118].

Therefore, we consider it necessary to evaluate the hypothesis that the reduced inter-areal information transfer observed under anesthesia possibly originates from disrupted information processing in local circuits rather than from disrupted long range connectivity. This alternative hypothesis receives additional support for the case of isoflurane, which potentiates agonist actions at $GABA_A$-receptors and inhibits nicotinic acetylcholine (nAChR) receptors. Conversely, evidence on direct inhibitory effects of isoflurane on AMPA and NMDA synapses, which are the dominant mediators of long-range cortico-cortical interactions, is sparse at best (see Table 2 in [223]). Under the alternative hypothesis of changed local information processing, a decrease in transfer entropy under anesthesia must be accompanied by:

1. a reduction in locally available information per brain area, i.e. in the sources of information transfer,

2. and the fact that the strongest decrease in locally available information is found at the source of the link with the strongest decrease in information transfer, rather than at its target (i.e. the end point).

Here, we perform tests of these predictions by estimating local information processing in and information transfer between local field potentials (LFPs) simultaneously recorded from primary visual cortex (V1) and prefrontal cortex (PFC) of two ferrets under different levels of isoflurane. We quantify local information processing by estimating the signal entropy (measuring available information) and quantified information transfer between recording sites by estimating transfer entropy. Additionally, to demonstrate the effect of reduced source entropy on transfer entropy, we estimated transfer entropy on simulated data with a constant coupling between processes, but a varying source entropy.

To better understand potential changes in local information processing we also quantified the active information storage, a measure of the information available at a recording site that can be predicted from past signals at that site, i.e. the information stored from past to present.

Because the estimation of such information theoretic quantities from finite data is difficult in general, we employ two complementary strategies: (i) probability density estimation based on nearest-neighbor searches in continuous data, and (ii) Bayesian estimation based on discretized data.

We also test whether the previously reported decrease of transfer entropy under anesthesia can indeed be replicated when avoiding some recently identified pitfalls in estimation of information transfer related to the use of symbolic time series, suboptimal embedding of the time series, and the use of net transfer entropy without identification of the individual information transfer delays (for problems related to these approaches see [75]).

Our results provide first evidence for the alternative hypothesis of altered local information processing causing reduced information transfer, as the above predictions were indeed met. We suggest to consider the alternative hypothesis as a serious candidate mechanism for LOC, and to use causal interventions to gather further experimental evidence.

Preliminary results for this study were published in abstract form in [224].

## 4.2 Results

### 4.2.1 Existence of information transfer between recording sites

Before analyzing differences in information transfer induced by isoflurane, we tested for the existence of significant information transfer between the recording sites (Table 4.1). For both animals, $TE_{SPO}$ was significant in the top-down direction, while the bottom-up direction was significant for animal 1 only. A non-significant information transfer in bottom-up direction for animal 2 may be explained by the dark experimental environment, i.e., the lack of visual input.

Tab. 4.1 Results significance test of $TE_{SPO}$ estimates in both animals and for both directions of interaction.

| **direction of interaction** | **$p$-value** |
|---|---|
| PFC → V1 | <0.01$^{**}$ |
| V1 → PFC | <0.05$^{*}$ |
| PFC → V1 | <0.05$^{*}$ |
| V1 → PFC | 0.2262 n.s. |

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$; Bonferroni-corrected

### 4.2.2 Changes in information theoretic measures under anesthesia

Overall, for higher isoflurane levels both locally available information and information transfer were decreased, while information storage in local activity increased.

As the estimation of information theoretic measures from finite length neural recordings poses a considerable challenge we present detailed, converging results from two complementary strategies to deal with this challenge—nearest-neighbor based estimators, and a Bayesian approach to entropy estimation suggested by Nemenman, Shafe, and Bialek (NSB-estimator) [225, 226]. This latter approach required a discretization of the continuous-valued LFP data, but yields principled control of bias, while the first approach allows the estimation of information-theoretic measures directly from continuous data, and thus conserves the information originally present in those data. Statistical testing was performed using a nonparametric permutation ANOVA (pANOVA), and a linear mixed model (LMM). The LMM approach was used in addition to the main pANOVA for the purpose of comparison to older studies using parametric statistics.

**Results based on next neighbor-based estimation from continuous data.** For higher isoflurane levels, we found an overall reduction in the locally available information ($H$), and in the information transfer ($TE_{SPO}$). We found an increase in the locally predictable (stored) information ($AIS$) (Table 4.2 and Fig. 4.1).

Tab. 4.2 Results of permutation analysis of variance for information theoretic measures ($p$-values).

| measure | effect | Ferret 1 | Ferret 2 |
|---|---|---|---|
| $TE_{SPO}$ | *isoflurane level* | <0.0001*** | <0.0001*** |
| | *direction* | 0.0003*** | 0.9345 |
| | *interaction* | 0.0052** | 0.2486[a] |
| $AIS$ | *isoflurane level* | 0.0017** | <0.0001*** |
| | *recording site* | 0.6774 | <0.0001*** |
| | *interaction* | <0.0001*** | 0.0029** |
| $H$ | *isoflurane level* | 0.0010** | <0.0001*** |
| | *recording site* | 0.9326 | <0.0001*** |
| | *interaction* | <0.0001*** | 0.0243* |

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$;

[a]This effect was significant when using LMM for statistical testing

In general, $H$ decreased in both animals under higher isoflurane levels (main effect *isoflurane level*, $p < 0.01^{**}$ for ferret 1 and $p < 0.001^{***}$ for ferret 2), indicating a reduction of locally available information for higher isoflurane concentrations.

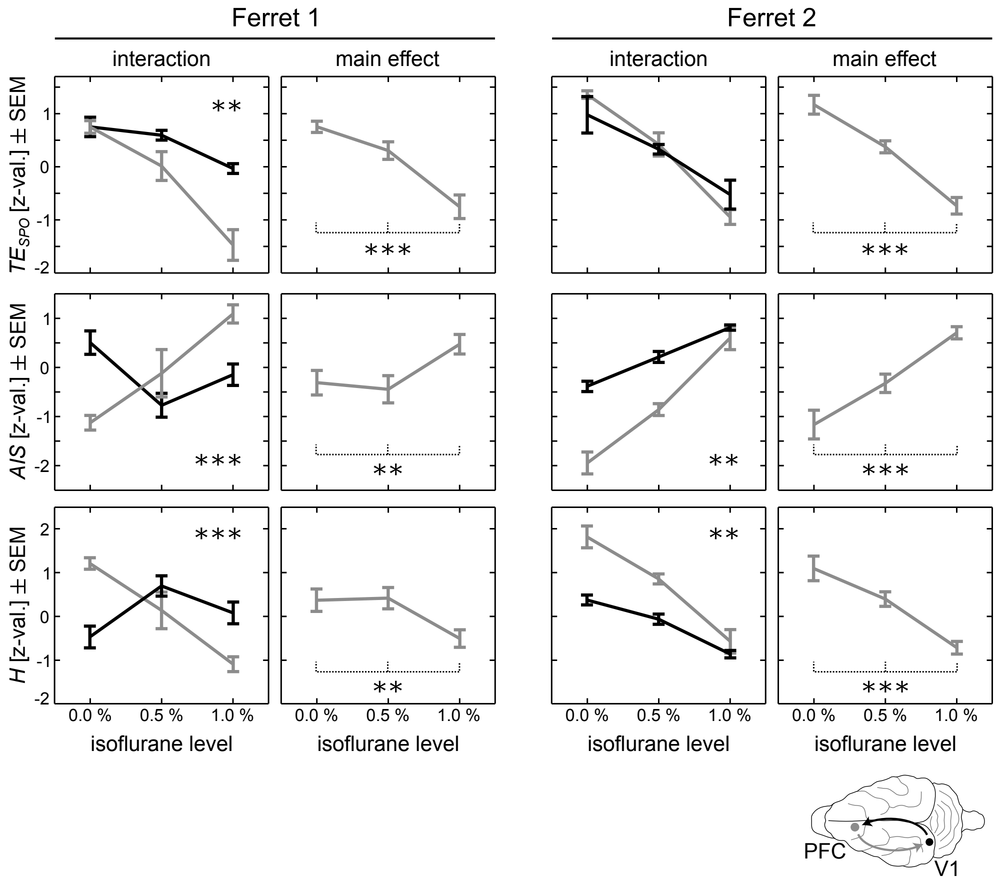


Fig. 4.1 **pANOVA results for nearest-neighbor based estimates of transfer entropy ($TE_{SPO}$), active information storage ($AIS$), and entropy ($H$).** Left columns show interactions *isoflurane level x direction* and *isoflurane level x recording site* for both animals; right columns show main effects *isoflurane level*. Grey lines in interaction plots indicate $TE_{SPO}$ from prefrontal cortex (PFC) to primary visual areas (V1), or $H$ and $AIS$ in PFC; black lines indicate $TE_{SPO}$ from V1 to PFC, or $H$ and $AIS$ in V1. Error bars indicate the standard error of the mean (SEM); stars indicate significant interactions or main effects ($^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$). Axis units for all information theoretic measures based on continuous variables are z-normalized values across conditions.

Yet, an isoflurane concentration of 0.5 % (abbreviated as *iso* 0.5 %, with other concentrations abbreviated accordingly) led to a slight increase in $H$ for ferret 1, followed by a decrease for concentration *iso* 1.0 %, which was below initial entropy values. This rise in $H$ in condition *iso* 0.5 % was present only in V1 of ferret 1, while $H$ decreased monotonically in PFC. In ferret 2, $H$ increased monotonically in both recording sites, with a stronger decrease in PFC. The interaction effect (*isoflurane level x brain region*) was significant for both animals ($p < 0.001^{***}$ for ferret 1 and $p < 0.01^{**}$ for ferret 2).

The information transfer as measured by the self prediction optimal transfer entropy ($TE_{SPO}$, [75]) decreased significantly with higher levels of isoflurane in both animals (main effect *isoflurane level*, $p < 0.001^{***}$), indicating an overall reduction in information transfer. This reduction was stronger in the top-down direction $PFC \rightarrow V1$ (significant interaction effect *isoflurane level x direction* in ferret 1, $p < 0.01^{**}$). In ferret 2 this interaction was not significant in the permutation ANOVA (pANOVA) on aggregated data, but was highly significant using the LMM approach (see next section).

The stored information as measured by the active information storage ($AIS$, [30]) increased in both animals under higher isoflurane levels (main effect *isoflurane level*, $p < 0.01^{**}$ for ferret 1 and $p < 0.001^{***}$ for ferret 2), indicating more predictable information in LFP signals under higher levels of isoflurane. In ferret 1 the concentration *iso* 0.5 % led to a slight decrease in $AIS$, followed by an increase compared to initial levels for concentration *iso* 1.0 %. This initial decrease in $AIS$ in condition *iso* 0.5 % was present only in V1 of this animal, while in its PFC $AIS$ increased monotonically. In ferret 2, $AIS$ increased monotonically in both recording sites, with a stronger increase in PFC. The interaction effect was significant for both animals ($p < 0.001^{***}$ for ferret 1 and $p < 0.01^{**}$ for ferret 2). Overall, $AIS$ behaved complementary to $H$ for all animals and isoflurane levels, despite the fact that AIS is one component of $H$ [30].

**Alternative statistical testing using linear mixed models** We additionally performed a parametric test, using linear mixed models (LMM) on non-aggregated data from individual epochs of recording sessions (adding *recording* as additional random factor to encode the recording session) to enable a comparison to results from earlier studies using parametric testing.

For both animals, model comparison showed a significant main effect of factor *isoflurane level* on $TE_{SPO}$, $AIS$, and $H$ as well as a significant interaction of factors *isoflurane level* and *direction* (see Table 4.3). The only exception to this was the factor direction in the evaluation of $TE_{SPO}$ in Ferret 2. Thus, the results of this alternative statistical analysis were in agreement with those from the pANOVA. The detailed Table 4.3 reports the Bayesian Information Criterion with $BIC = -2\log p(\mathbf{x}|M, \hat{a}) + k(a)log(N)$, where $\mathbf{x}$ are the observed realizations of the data, $\hat{a}$ are the parameters that optimize the likelihood for a given model $M$, $k(a)$ is the number of parameters and $N$ the number of data points. The $BIC$ becomes smaller with better model fit. In addition, Table 4.3 gives the deviance, $-2\log p(\mathbf{x}|M, \hat{a})$, which is higher for better model fit, and the $\chi^2$ with the corresponding $p$-value, describing the likelihood ratio, which follows a $\chi^2$-distribution.

**Bayesian estimation on discretized data** In addition to the neighbor-distance based estimators for $TE_{SPO}$, $AIS$, and $H$ used above, we also applied Bayesian estimators recently proposed by Nemenman, Shafe, and Bialek (NSB) [225, 226] to our data.

Tab. 4.3 Results of parametric statistical testing using model comparison between linear mixed models. Simple effects of factors *isoflurane level* and *direction* or *recording site* were tested against the null model; interaction effects *isoflurane level* $times$ *direction* and *isoflurane level* $times$ *recording site* were tested against the additive models *isoflurane level* + *direction* and *isoflurane level* + *recording site*, respectively. *** as defined in Table 4.2

| animal | measure | effect | BIC | deviance | $\mathcal{X}^2$ | $\mathcal{X}$ df | $p$ |
|---|---|---|---|---|---|---|---|
| Ferret 1 | $TE_{SPO}$ | *null model* | 345430.8 | 345401.7 | NA | NA | NA |
| | | *isoflurane level* | 345421.5 | 345373.1 | 28.58 | 2 | $<<0.000^{***}$ |
| | | *direction* | 342828.4 | 342789.7 | 2612.04 | 1 | $<<0.000^{***}$ |
| | | *isoflurane l. + direction* | 342819.3 | 342761.3 | NA | NA | NA |
| | | *isoflurane l. × direction* | 342139.9 | 342062.4 | 698.84 | 2 | $<<0.000^{***}$ |
| | $AIS$ | *null model* | 37581.2 | 37551.8 | NA | NA | NA |
| | | *isoflurane level* | 37555.8 | 37506.8 | 46.00 | 2 | $<<0.000^{***}$ |
| | | *recording site* | 33296.2 | 33256.9 | 4294.82 | 1 | $<<0.000^{***}$ |
| | | *isoflurane l. + recording site* | 33270.9 | 33212.1 | NA | NA | NA |
| | | *isoflurane l. × recording site* | 30754.3 | 30675.8 | 2536.29 | 2 | $<<0.000^{***}$ |
| | $H$ | *null model* | 353650.8 | 353621.8 | NA | NA | NA |
| | | *isoflurane level* | 353653.1 | 353604.7 | 17.03 | 2 | $<0.000^{***}$ |
| | | *recording site* | 353446.8 | 353408.1 | 213.64 | 1 | $<<0.000^{***}$ |
| | | *isoflurane l. + recording site* | 353449.1 | 353391.1 | NA | NA | NA |
| | | *isoflurane l. × recording site* | 347204.7 | 347127.2 | 6263.85 | 2 | $<<0.000^{***}$ |
| Ferret 2 | $TE_{SPO}$ | *null model* | 131164.7 | 131135.3 | NA | NA | NA |
| | | *isoflurane level* | 131152.3 | 131103.2 | 32.06 | 2 | $<<0.000^{***}$ |
| | | *direction* | 131168.9 | 131129.6 | 5.66 | 1 | 0.020 |
| | | *isoflurane l. + direction* | 131109.5 | 131097.5 | NA | NA | NA |
| | | *isoflurane l. × direction* | 130871.5 | 130793.0 | 304.56 | 2 | $<<0.000^{***}$ |
| | $AIS$ | *null model* | 109203.1 | 109173.7 | NA | NA | NA |
| | | *isoflurane level* | 109183.1 | 109134.1 | 39.63 | 2 | $<<0.000^{***}$ |
| | | *recording site* | 105827.2 | 105788.0 | 3385.68 | 1 | $<<0.000^{***}$ |
| | | *isoflurane l. + recording site* | 105807.3 | 105748.4 | NA | NA | NA |
| | | *isoflurane l. × recording site* | 104173.5 | 104095.0 | 1653.46 | 2 | $<<0.000^{***}$ |
| | $H$ | *null model* | -14074.8 | -14104.2 | NA | NA | NA |
| | | *isoflurane level* | -14088.6 | -14137.7 | 33.43 | 2 | $<0.000^{***}$ |
| | | *recording site* | -18423.0 | -18462.3 | 4358.04 | 1 | $<<0.000^{***}$ |
| | | *isoflurane l. + recording site* | -18436.7 | -18495.6 | NA | NA | NA |
| | | *isoflurane l. × recording site* | -19867.2 | -19945.7 | 1450.09 | 2 | $<<0.000^{***}$ |

The Bayesian approach promises to yield unbiased estimators when priors are chosen appropriately. However, one has to keep in mind that these estimators currently require the discretization of continuous data, and therefore may loose important information.

When applying the NSB estimator to the discretized LFP time series with $N_{bins}$ discretization steps, we observed that for $N_{bins} \geq 8$ the results were qualitatively consistent for different choices of numbers of bins. We present results for $N_{bins} = 12$ (Fig. 4.2), which provides a reasonable resolution of the signal while still allowing for a reliable estimation of entropies within the scope of available data.

We confirmed the reliability of the estimator by systematically reducing the sample size and found no substantial impact on our estimates (Fig. 4.3).

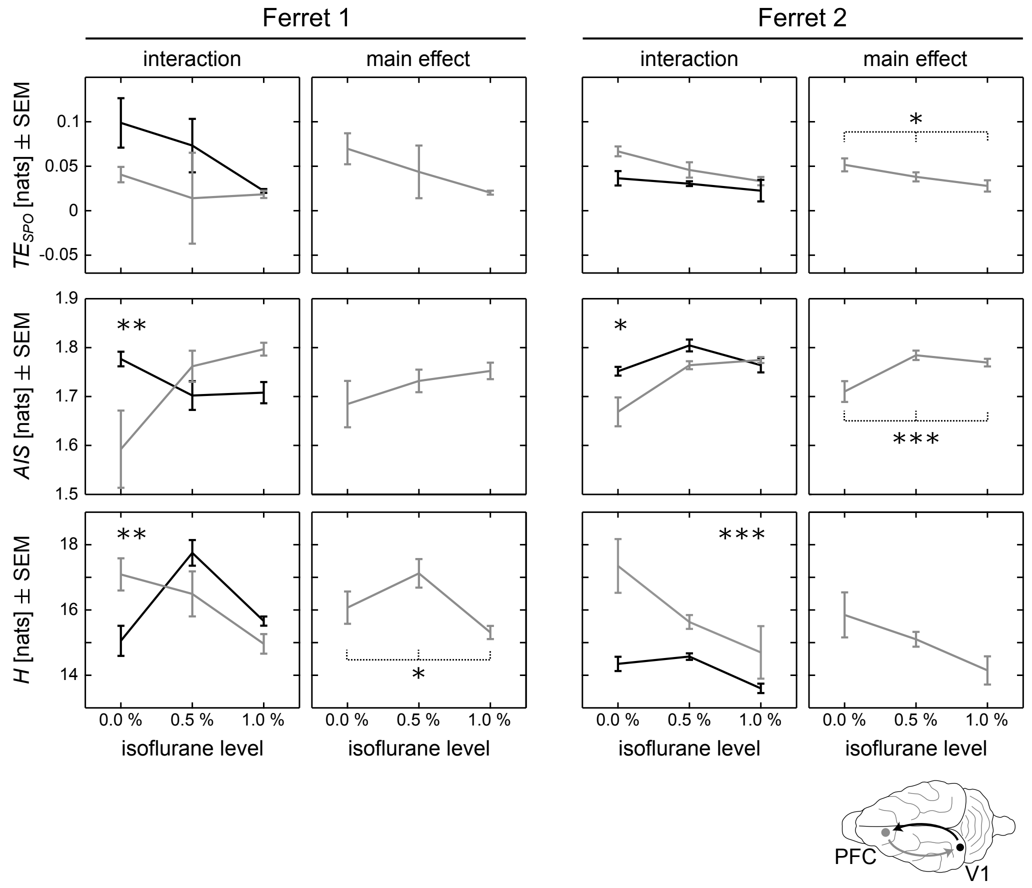


Fig. 4.2 **pANOVA results for Bayesian estimates of transfer entropy ($TE_{SPO}$), active information storage ($AIS$), and entropy ($H$).** Left columns show interactions *isoflurane level x direction* and *isoflurane level x recording site* for both animals; right columns show main effects *isoflurane level*. Grey lines in interaction plots indicate $TE_{SPO}$ from prefrontal cortex (PFC) to primary visual areas (V1), or $H$ and $AIS$ in PFC; black lines indicate $TE_{SPO}$ from V1 to PFC, or $H$ and $AIS$ in V1. Error bars indicate the standard error of the mean (SEM); stars indicate significant interactions or main effects ($^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$).

The estimation of $H$, $AIS$, and $TE_{SPO}$ by Bayesian techniques for the binned signal representations provided results that were qualitatively consistent with results from the neighbor-distance based estimation techniques (compare Figs. 4.1 and 4.2, and Tables 4.2 and 4.4). While the Bayesian estimates showed larger variances across different recordings and sample sizes, on average, we saw a decrease of $TE_{SPO}$ and $H$ for higher concentrations of isoflurane (main effect *isoflurane level*), while $AIS$ increased for higher concentrations. For ferret 1 we also found an interaction effect for $TE_{SPO}$, with a stronger reduction in information transfer in top-down direction.

Note that we also applied an alternative Bayesian estimation scheme based on Pitman-Yor-process priors [227]. However, for this estimation procedure, we observed that the data were insufficient to allow for a robust estimation of the tailing

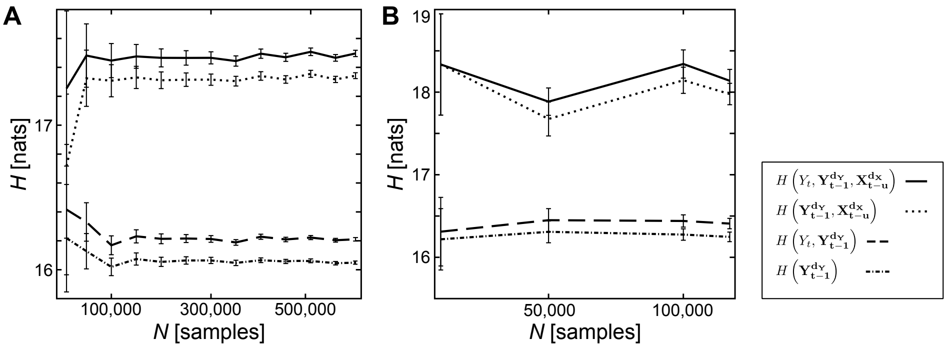


Fig. 4.3 **Examples for estimates of entropy ($H$) terms for transfer entropy calculation (Eq. 4.12) by number of data points $N$, using the Nemenman-Shafee-Bialek-estimator (NSB).** (A) entropies for ferret 1, V1, *iso* 0.5 %: estimates are stable for $N \geq 100,000$; (B) entropies for ferret 2, PFC, *awake*: an insufficient number of data points does not allow for verification of the estimate's robustness (recording was excluded from further analysis). Variables labeled $X$ reflect data from the source variable, $Y$ from the target variable. $t$ is an integer time index, and bold typeface indicates the state of a system (see Methods).

Tab. 4.4 Results of permutation analysis of variance for information theoretic measures obtained through Bayesian estimation ($p$-values).

| measure | effect | Ferret 1 | Ferret 2 |
|---|---|---|---|
| $TE_{SPO}$ | *isoflurane level* | 0.0549 | 0.0013* |
| | *direction* | 0.0272* | 0.0002*** |
| | *interaction* | 0.6627 | 0.0616 |
| $AIS$ | *isoflurane level* | 0.2148 | <0.0001*** |
| | *direction* | 0.0788 | 0.0032** |
| | *interaction* | 0.0026** | 0.0153* |
| $H$ | *isoflurane level* | 0.0184* | <0.0001*** |
| | *recording site* | 0.2738 | <0.0001*** |
| | *interaction* | 0.0017** | 0.0053 |

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

behavior of the distribution as indicated by large variances and unreasonably high estimates across different sample sizes.

**Simulated effects of changed source entropy on transfer entropy** To test the effect of reduced source entropy on $TE_{SPO}$ between a source and target process, we simulated two test cases with high and low source entropy respectively. In both test cases signals were based on the original recordings, and the coupling between source and target was held constant (see Methods).

We found significantly higher $TE_{SPO}$ in the case with high source entropy than in the case with low source entropy (Fig. 4.4A). The simulation thus demonstrated that a lower source entropy does indeed lead to a reduction in $TE_{SPO}$ despite an unchanged coupling. Information transfer in the high-entropy case was similar to the average information transfer found in recordings under an isoflurane concentration of 0.0 %, indicating that the simulation scheme mirrored information transfer found in the experimental recordings.

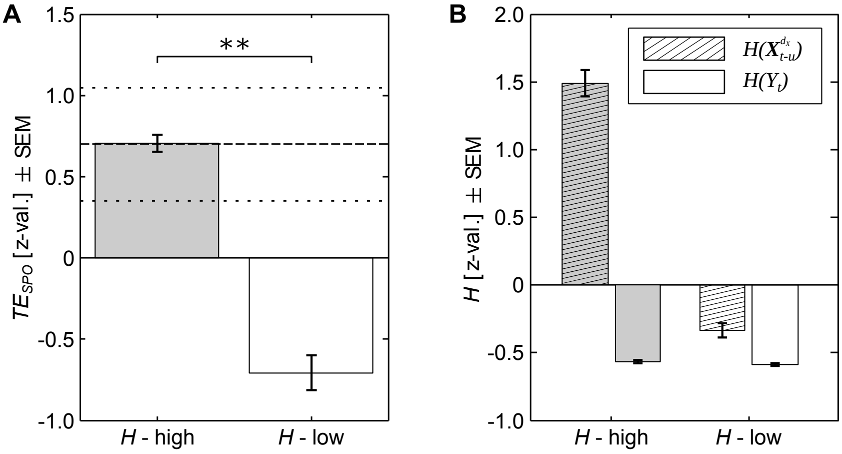


Fig. 4.4 **Simulated effects of changed source entropy on transfer entropy ($TE_{SPO}$).** (A) $TE_{SPO}$ for two simulated cases of high ($H$-*high*) and low ($H$-*low*) source entropy ($^{**}p < 0.01$, error bars indicate the standard error of the mean, SEM, over data epochs). Dashed lines indicate the average $TE_{SPO}$ estimated from the original data under 0.0 % isoflurane ± SEM; (B) source entropy $H(\mathbf{X}^{dx}_{t-u})$ (dashed bars) and target entropy $H(Y_t)$ (empty bars) for the two simulated test cases of high (gray bars) and low entropy (white bars), error bars indicate the SEM over data epochs. Source entropy was higher in the high-entropy simulation, while target entropy was approximately the same for both cases. Results are given as z-values across estimates for all epochs from both simulations.

### 4.2.3 Optimized embedding parameters for $TE_{SPO}$ and $AIS$ estimation

As noted in the introduction, the estimation of information theoretic measures from finite data is challenging. For the measures that describe distributed computation in complex systems, such as transfer entropy, estimation is further complicated because the available data are typically only scalar observations of a process with multidimensional dynamics. This necessitates the approximate reconstructions of the process' states via a form of embedding [133], where a certain number of past values of the scalar observations spaced by an embedding delay are taken into account (e.g. for a pendulum swinging through its zero position one additional past position values will clarify whether it's going left or right). An important part of proper transfer entropy estimation is, thus, optimization of this number of past values (embedding dimension), and of the embedding delay. These two embedding parameters then approximately define past states, whose correct identification is crucial for the estimation of transfer entropy [13], but also for the estimation of active information storage. Without it information storage may be underestimated and, erroneous values of the information transfer will be obtained; even a detection of information transfer in the wrong direction is likely (see [13, 75, 86] and Methods section). Existing studies using transfer entropy often omitted the optimization of embedding parameters, and instead used ad-hoc choices, which may have had a detrimental effect on the estimation of transfer entropy—hence the need for a confirmation of previous results in this study.

In the present study, we therefore used a formal criterion proposed by Ragwitz [133], to find an optimal embedding defined by the embedding dimension $d$ (the number of values collected) and delay $\tau$ (the temporal spacing between them), to find embeddings for the $TE_{SPO}$ and $AIS$ past states. We used the implementation of this criterion provided by the TRENTOOL toolbox. Since the bias of the estimators used depends on $d$, we used the maximum $d$ over all conditions and directions of interaction as the common embedding dimension for estimation from each epoch to make the estimated values statistically comparable. The resulting dimension used was 15 samples, which is considerably higher than the value commonly used in the literature [32, 33], when choosing the embedding dimension ad-hoc or using other criteria as for example the sample size [35]. The embedding delay $\tau$ was optimized individually for single epochs in each condition and animal as it has no influence on the estimator bias.

### 4.2.4 Relevance of individual information transfer delay estimation

Several previous studies on information transfer under anesthesia reported the sign of the so called net transfer entropy ($TE_{net}$) as a measure of the predominant direction of information transfer between two sources. $TE_{net}$ is essentially just the normalized difference between the transfer entropies measured in the two direction connecting a pair of recording sites (see Methods). When calculating $TE_{net}$, it is particularly important to individually account for physical delays $\delta$ in the two potential directions of information transfer, because otherwise the sign of $TE_{net}$ may become meaningless (see examples in [75]). As this delay is unknown a priori it has to be found prior to the actual estimation of information transfer. We have recently shown that this can be done by using a variable delay parameter $u$ in a delay sensitive transfer entropy estimator $TE_{SPO}$ (see Methods); here, the $u$ that maximizes the transfer entropy over a range of assumed values for $u$ reflects the physical delay [75].

The necessity to individually optimize $u$ for each interaction is not a mere theoretical concern but was clearly visible in the present study: Fig. 4.5C shows representative results from ferret 1 under 0.0 % isoflurane, where apparent $TE_{SPO}$ values strongly varied as a function of the delay $u$. As a consequence, $TE_{net}$ values also varied if a *common* delay $u$ was chosen for both directions. In other words, the sign of the $TE_{net}$ varied as a function of individual choices for $u_{PFC\to V1}$ and $u_{V1\to PFC}$ for each direction of information transfer and hence became uninterpretable (Fig. 4.5D).

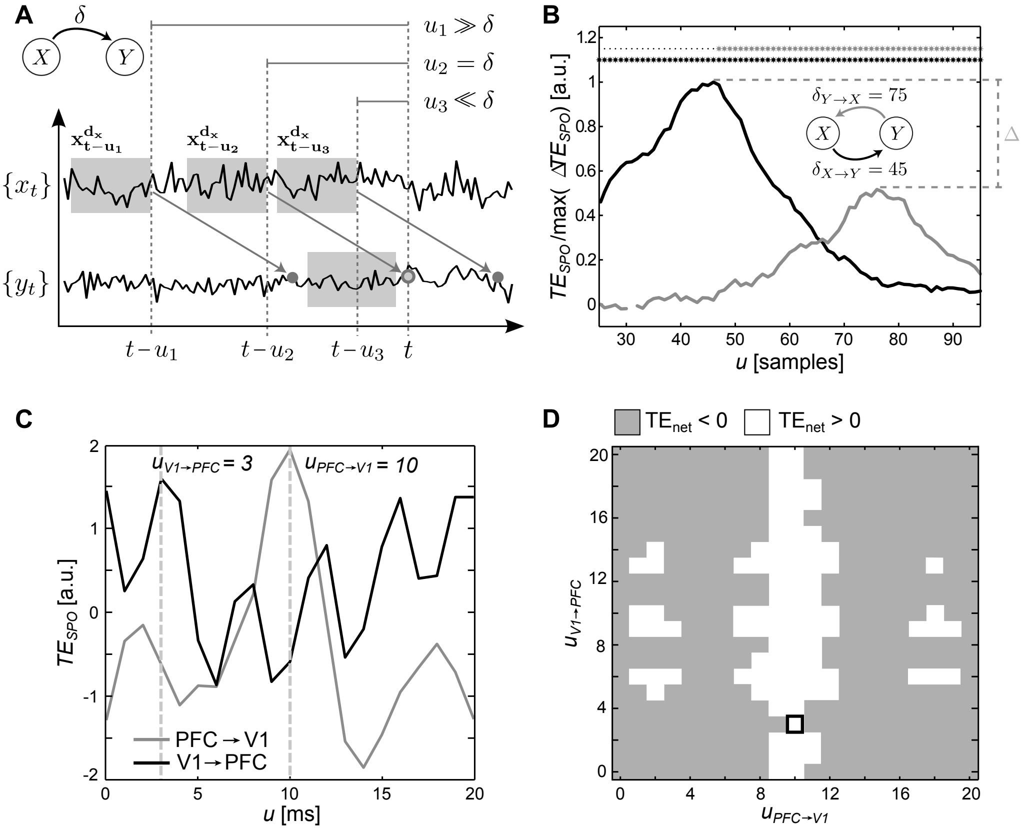


Fig. 4.5 **Estimation of information transfer delays.** (A, modified from [75]) estimation of transfer entropy ($TE_{SPO}$) depends on the choice of the delay parameter $u$, if $u$ is much smaller or bigger than the true delay $\delta$, information arrives too late or too early in the target time series and information transfer is not correctly measured; (B, modified from [75]) $TE_{SPO}$ values estimated from two simulated, bidirectionally coupled Lorenz systems (see [75] for details) as a function of $u$ for both directions of analysis, $TE_{SPO}(X \to Y, t, u)$ (black line) and $TE_{SPO}(Y \to X, t, u)$ (gray line): $TE_{SPO}$ values vary with the choice of $u$ and so does the absolute difference between values; using individually optimized transfer delays for both directions of analysis yields the meaningful difference $\Delta$ (dashed lines), where $TE_{SPO}(X \to Y, t, u_{opt}) > TE_{SPO}(Y \to X, t, u_{opt})$; (C) example transfer entropy analysis for one recorded epoch: $TE_{SPO}$ values vary greatly as a function of $u$, optimal choices of $u$ are marked by dashed lines; (D) sign (gray: negative, white: positive) of $TE_{net}$ for different values of $u_{PFC\to V1}$ (x-axis) and $u_{V1\to PFC}$ (y-axis), calculated from $TE_{SPO}$ values shown in panel C: the sign varies with individual choices of $u$; the black frame marks the combination of individually optimal choices for both parameters that yields the correct result.

As a consequence of the above, we here individually optimized $u$ for each direction of information transfer in each condition and animal to estimate the true delay of information transfer following the mathematical proof in [75]. We individually optimized $u$ to obtain estimates of transfer entropy that were not biased by a non-optimal choice for $u$. We used the implementation in TRENTOOL [86] and scanned values for $u$ ranging from 0 ms to 20 ms. Averages for optimized delays ranged from 4 ms to 7 ms across animals and isoflurane levels (Fig. 4.6).

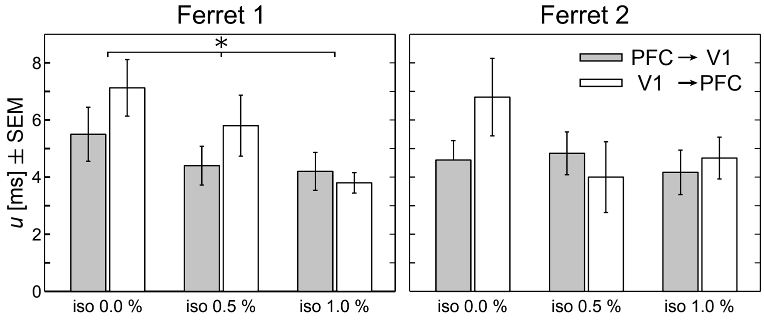


Fig. 4.6 **Optimized information transfer delays $u$ for both directions of interaction and three levels of isoflurane, by animal.** Bars denote averages over recordings per condition; error bars indicate the standard error of the mean (SEM). There was a significant main effect of *isoflurane level* for ferret 1 ($p < 0.05$).

Note that in Fig. 4.5C, $TE_{SPO}$ as a function of $u$ shows multiple peaks, especially for the direction $V1 \rightarrow PFC$. Since these peaks resembled an oscillatory pattern close to the low-pass filter frequency used for preprocessing the data, we investigated the influence of filtering on delay reconstruction by simulating two coupled time series for which we reconstructed the delay with and without prior band-pass filtering (Fig. 4.7). For filtered, simulated data $TE_{SPO}$ as a function of $u$ indeed showed an oscillatory pattern for certain filter settings. However, broadband filtering—as used here for the original data—did not lead to the reconstruction of an incorrect value for the delay $\delta$ (Fig. 4.7C). Yet, when filtering using narrower bandwidths, the reconstruction of the correct information transfer delay failed (Fig. 4.7D–F). This finding supports the general notion that filtering, especially narrow-band filtering, may only be applied with caution before estimating connectivity measures [228, 229].

Yet, since broad-band filtering did not lead to the observed oscillatory patterns in the simulated data, alternative explanations for multiple peaks in $TE_{SPO}$ are more likely: one alternative cause of multiple peaks in the $TE_{SPO}$ are multiple information transfer channels with various delays between source and target (see [75], especially Fig. 6). Each information transfer channel with its individual delay will be detected when estimating $TE_{SPO}$ as a function of $u$ (see simulations in [75]). A further possible cause for multiple peaks in the $TE_{SPO}$ is the existence of a feedback loop between the source of $TE_{SPO}$ and a third source of neural activity [75]. In such a feedback loop, information in the source is fed back to the source such that it occurs

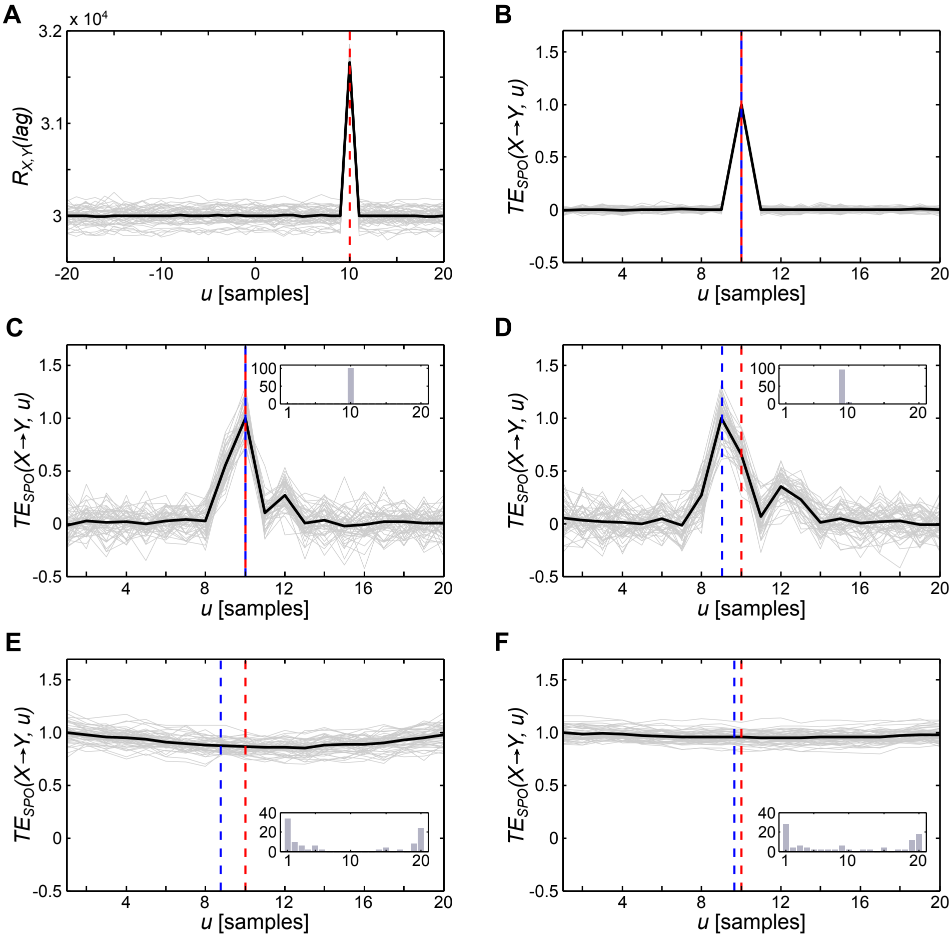


Fig. 4.7 **Simulated effects of filtering on the reconstruction of information transfer delays ($u$).** (A) Cross-correlation $R_{X,Y}$ between two simulated, coupled time series ($N = 100000$, drawn from a uniform random-distribution over the open interval $(0, 1)$, true coupling delay 10 samples, indicated by the red dashed line), the simulation was run 50 times, black lines indicate the mean over simulation runs, gray lines indicate individual runs; (B) $TE_{SPO}$ as a function of information transfer delay $u$ before filtering the data; (C–D) $TE_{SPO}$ as a function of $u$ after filtering the data with a bandpass filter (fourth order, causal Butterworth filter, implemented in the MATLAB toolbox FieldTrip [150]); red dashed lines indicate the simulated delay $\delta$, blue lines indicate the average reconstructed delay $u$ over simulation runs, histograms (inserts) show the distribution of reconstructed values for $u$ over simulation runs in percent; (C) broadband filtering (0.1 Hz to 300 Hz) introduced additional peaks in $TE_{SPO}$ values, however, the maximum peak indicating the optimal $u$ was still at the simulated delay for all simulated runs; (D) filtering within a narrower band (0.1 Hz to 200 Hz) led to an imprecise reconstruction of the correct $\delta$ in each run with an error of 1 sample, i.e. $\delta$ tended to be underestimated; (E–F) narrow-band filtering in the beta range (12 Hz to 30 Hz) and theta range (4 Hz to 8 Hz) led to a wide distribution of $u$ with a large absolute error of up to 10 samples.

again at a later point in time, leading to the repeated transfer of identical bits of information. Both causes are potential explanations for the occurrence of multiple peaks in the $TE_{SPO}$ in the present study—yet, deciding between these potential causes requires additional research, specifically interventional approaches to the causal structure underlying the observed information transfer.

In sum, our results clearly indicate the necessity to individually optimize information transfer delays and that often employed ad-hoc choices for $u$ may result in spurious results in information transfer.The additional simulations of filtering effects showed that narrow-band filtering may have detrimental effects on this optimization procedure and should thus be avoided.

### 4.2.5 Relation of information-theoretic measures with other time-series properties

The information-theoretic measures used in this study are relatively novel and have been applied in neuroscience research rarely. It is, thus, conceivable that these measures quantify signal properties that are more easily captured by traditional time-series measures, namely, the autocorrelation decay time (ACT), the signal variance, and the power in individual frequency bands. To investigate the overlap between information-theoretic and traditional measures, we calculated correlations between $AIS$ and ACT, between $H$ and signal variance, between $TE_{SPO}$ and ACT, between $TE_{SPO}$ and signal variance, and between $AIS$, $H$, and $TE_{SPO}$ and the power in various frequency bands, respectively. Fig. 4.8 shows average ACT, signal variance, and power over recordings, for individual isoflurane levels and recording sites, and for both animals.

Detailed tables showing correlation coefficients and explained variances are provided as supporting information S2. In summary, for $AIS$ and ACT we found significant correlations in both animals, however the median of the variance explained, $\widetilde{R}^2$, over individual recordings was below 0.1 for each significant correlation in an animal, recording site and isoflurane level. Hence, even though there was some shared variance between ACT and $AIS$, there remained a substantial amount of unexplained variance in $AIS$, indicating that $AIS$ quantified properties other than the signal's ACT—in line with results from our earlier studies [67]. This is expected because the autocorrelation (decay) time ACT measures how long information persists in a time series in a linear encoding, while $AIS$, in contrast, measures how much information is stored (at any moment), and also reflects nonlinear transformations of this information. The fact that there is still some shared variance between the two measures here may be explained by the construction of the $AIS$'s past state embedding, where the time steps between the samples in the embedding vector are defined as a fractions of the ACT.

Between $H$ and signal variance, we found no significant correlation; moreover, correlations for individual recordings were predominantly negative.

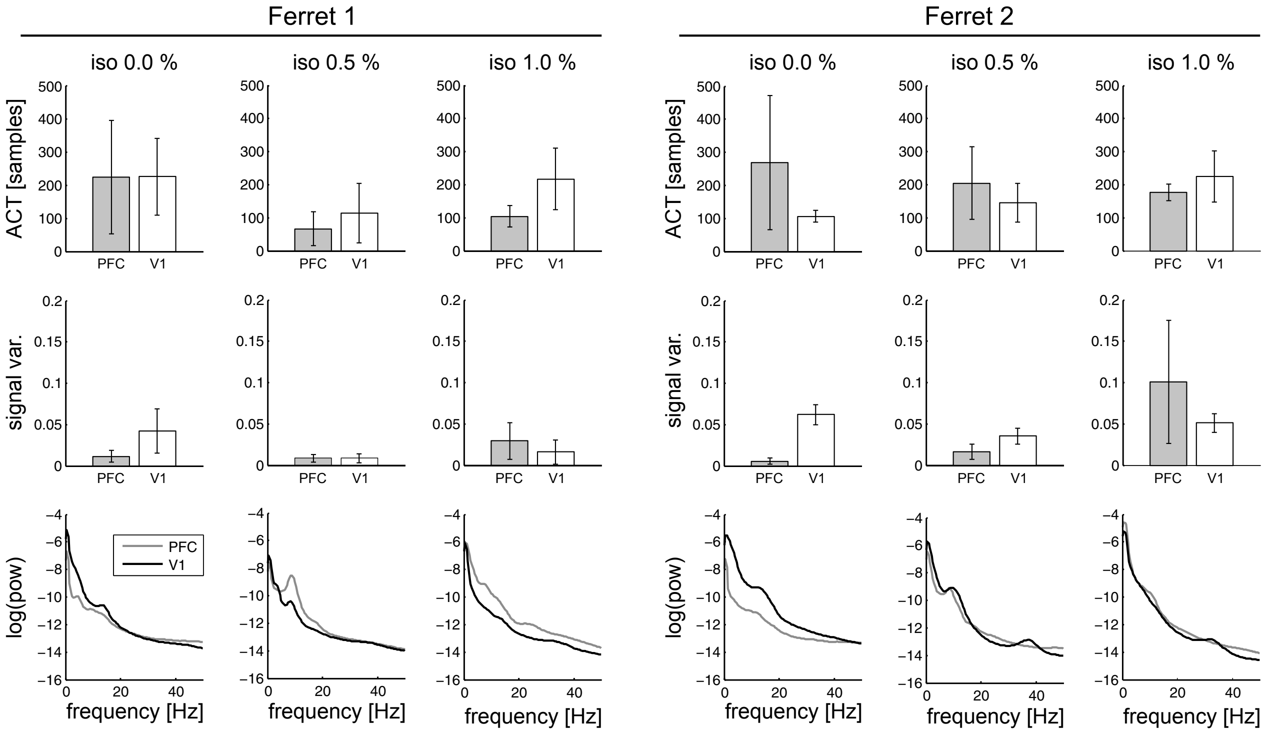


Fig. 4.8 **Autocorrelation decay time (ACT), signal variance, and power by isoflurane level and recording site for both animals.** Averages over recordings per isoflurane level and recording site; error bars indicate one standard deviation.

The $TE_{SPO}$ was significantly correlated with source ACT for an isoflurane level of 1.0 %. Also for the correlation of $TE_{SPO}$ and the source's variance, we found significant correlations for the bottom-up direction under 0.5 % and 1.0 % isoflurane in animal 1, and for both directions under 1.0 % isoflurane in animal 2. For all significant correlations we found $\widetilde{R}^2 \leq 0.015$, again indicating a substantial amount of variance in $TE_{SPO}$ that was not explained by ACT or signal variance.

When correlating band power with information-theoretic measures, we found significant correlations between $AIS$ and all bands in both animals animals (see Table 4.8 as part of supporting information S2, $\widetilde{R}^2 < 0.16$); for $H$ and band power, we found correlations predominantly in higher frequency bands (beta and gamma, see Table 4.9 as part of supporting information S2, $\widetilde{R}^2 < 0.1$); for $TE_{SPO}$ and band power, we found significant correlations predominantly in the gamma band and one significant correlation in the delta band for animal 2 and in the beta band for animal 1, respectively (see Table 4.10 as part of supporting information S2, $\widetilde{R}^2 < 0.024$). In sum, we found relationships between the power in individual frequency bands and all three information-theoretic measures. For all measures, the variance explained was below 0.2, indicating again that band power did not fully capture the properties measured by $TE_{SPO}$, $AIS$, and $H$.

## 4.3 Discussion

We analyzed long-range information transfer between areas V1 and PFC in two ferrets under different levels of isoflurane. We found that transfer entropy was indeed reduced under isoflurane and that this reduction was more pronounced in top-down directions. These results validate earlier findings made using different estimation procedures [31–35]. As far as information transfer alone was concerned our results are compatible with an interpretation of reduced long-range information transfer due to reduced inter-areal coupling. Yet, this interpretation provides no direct explanation for the findings of reduced locally available information as explained below. In contrast, the alternative hypothesis that the reduced long range information transfer is a secondary effect of changes in local information processing provides a concise explanation for our findings both with respect to locally available information, and information transfer.

### 4.3.1 Reduction in transfer entropy may be caused by changes in local information processing

To test our alternative hypothesis we evaluated two of its predictions about changes in locally available information, as measured by signal entropy, under administration of different isoflurane concentrations. First, entropy should be reduced; second, the strongest decrease in information transfer should originate from the source node with the strongest decrease in entropy, rather than end in this node. Indeed, we found that signal entropy decreased; the most pronounced decrease in signal entropy was found in PFC. In accordance with our prediction, we found that PFC—the node with the larger decrease in entropy—was also at the source, not the target, of the most pronounced decrease in transfer entropy. This is strong evidence against the theoretical possibility that in a recording site, entropy decreased due to a reduced influx of information—because in this latter scenario the strongest reduction in entropy should have been found at the target (end point) of the most pronounced decrease in information transfer. Hence, our hypothesis that long-range cortico-cortical information transfer is reduced because of changes in local processing must be taken as a serious alternative to the currently prevailing theories of anesthetic action based on disruptions of long range interactions. We suggest that a renewed focus on local information processing in anesthesia research will be pivotal to advance our understanding of how consciousness is lost.

Our predictions that reductions in Entropy should potentially be reflected in reduced information transfer derives from the simple principle that information that is not available at the source cannot be transferred. We may thus in principle reduce

$TE_{SPO}$ to arbitrarily low values by reducing the entropy of one of the involved processes, without changing the physical coupling between the two systems, just by changing their internal information processing. This trivial but important fact has been neglected in previous studies when interpreting changes in information transfer as changes in coupling strength. Even when this bound is not attained, e.g. because only a certain fraction of the local information is transferred even under 0.0 % isoflurane, it seems highly plausible that reductions in the locally available information affect the amount of information transfered.

A possible indication of how exactly local information processing is changed is given by the observation of increased active information storage in PFC and V1 (also see the next paragraph). This means that more "old" information is kept stored in a source's activity under anesthesia, rather than being dynamically generated. Such stored source information will not contribute to a measurable transfer entropy under most circumstances because it is already known at the target (see [8], section 5.2.3, for an illustrative example).

### 4.3.2 Relation between changed locally available information and information storage

In general, $AIS$ increases if a signal becomes more predictable when knowing it's past, but is unpredictable otherwise. This also means that the absolute $AIS$ is upper-bounded by the system's entropy $H$ (see Methods, Eq. 4.5). Thus, a decrease in $H$ can in principle lead to a decrease in $AIS$, i.e., fewer possible system states may lead to a decrease in absolute $AIS$. However, in the present study, we observed an *increase* in $AIS$ while $H$ decreased—this indicates an increase in predictability that more than compensates for the decrease in locally available information. In other words, the system visited fewer states in total but the next state visited became more predictable from the system's past. Thus, a reduction in entropy and increase in predictability points at highly regular neural activity for higher isoflurane concentrations. Such a behavior in activity is in line with existing electrophysiological findings: under anesthesia signals have been reported to become more uniform, exhibiting repetitive patterns, interrupted by bursting activity (see [222] for a review). For example, Purdon and colleagues observed a reduction of the median frequency and an overall shift towards high-power, low-frequency activity during LOC [230]. In particular, slow-wave oscillatory power was more pronounced during anesthesia-induced LOC. LOC was also accompanied by a significant increase in power and a narrowing of oscillatory bands in the alpha frequency range in their study.

### 4.3.3 Limitations of the applied estimators and measures

Unfortunately a quantitative comparison between different information theoretic estimates that would directly relate $H$, $AIS$ and $TE_{SPO}$ is not possible using the continuous estimators applied in this study. Their estimates are not comparable because the bias properties of each estimator differ. For each estimator, the bias depends on the number of points used for estimation as well as on the dimensionality of involved variables—however, the exact functional relationship between these two quantities and the bias is unknown and may differ between estimators. (In our application, the dimensionality is determined mainly by the dimension of the past state vectors, and by how many different state variables enter the computation of a measure).

This lack of comparability makes it impossible to normalize estimates; for example, transfer entropy is often normalized by the conditional entropy of the present target state to compare the fraction of transferred information to the fraction of stored information. We forgo this possibility of comparison for a greater sensitivity and specificity in the detection of changes in the individual information theoretic measures here. New estimators, e.g. Bayesian estimators like the ones tested here, promise more comparable estimates by tightly controlling the biases. Yet, these estimators were not as reliable as expected in our study, displaying a relatively high variance.

A further important point to consider when estimating transfer entropy between recordings from neural sites, is the possibility of third unobserved sources influencing the information transfer. In the present study, third sources (e.g., in the thalamus) may influence the information transfer between PFC and V1—for example, if a third source drove the dynamics in both areas, the areas would become correlated, leading to non-zero estimates of $TE_{SPO}$. This $TE_{SPO}$ is then attributable to the correlation between source and target, but does not measure an actual information transfer between sources. Hence, information transfer estimated by transfer entropy should in general not be directly equated with a causal connection or causal mechanism existing between the source and target process (see also [68] for a detailed discussion of the difference between transfer entropy and measures of causal interactions).

### 4.3.4 Potential physiological causes for altered information transfer under anesthesia

We here tested the possibility that changes in local information processing lead to the frequently observed reduction in information transfer between cortical areas under isoflurane administration, instead of altered long-range coupling. Results on entropies and active information storage suggest that this is a definite possibility from a mathematical point of view.

This is supported by the neurophysiology related to the mode of action of isoflurane, because a dominant influence of altered long-range coupling on $TE_{SPO}$ would mandate that synaptic terminals of the axons mediating long range connectivity should be targets of isoflurane. Such long range connectivity is thought to be dominated by glutamatergic AMPA receptors for inter-areal bottom-up connections, and glutamatergic NMDA receptors for inter-areal top-down connections, building upon findings in [231] and [232] (but see [233] for some evidence of GABAergic long range connectivity). Yet, evidence for isoflurane effects on AMPA and NMDA receptors is sparse to date (Table 2 in [223]). In contrast, the receptors most strongly influenced by isoflurane seem to be $GABA_A$ and nicotinic acetylcholine (nAChR) receptors. More specifically, isoflurane potentiates agonist interactions at the former, while inhibiting the latter.

Thus, if one adopts the current state of knowledge on the synapses involved in long-range inter-areal connectivity, evidence speaks against a dominant effect of modulation of effective long-range connections by isoflurane. This, in turn, points at local information processing as a more likely reason for changed transfer entropy under isoflurane anesthesia. This interpretation is perfectly in line with our finding that decreases in source entropy seem to determine the transfer entropy decreases, instead of decreases in transfer entropy determining the target entropies.

Nevertheless, targeted local interventions by electrical or optogenetic activation of projection neurons, combined with the set of information theoretic analyses used here, will most likely be necessary to reach final conclusions on the causal role of local entropy changes in reductions of transfered information.

### 4.3.5 How may altered long-range information transfer lead to loss of consciousness?

Investigating long-range information transfer under anesthesia is motivated by the question how changed information transfer may cause LOC. To that effect, our findings—a dominant decrease in top-down information transfer under anesthesia, and a decrease in locally available information possibly driving it—may be interpreted in the framework of predictive coding theory [3, 234, 235]. Predictive coding proposes that the brain learns about the world by constructing and maintaining an internal model of the world, such that it is able to *predict* future sensory input at lower levels of the cortical hierarchy. Whether predictions match actual future input is then used to further refine the internal model. It is thus assumed that top-down information transfer serves the propagation of predictions [236]. Theories of conscious perception within this predictive coding framework propose that conscious perception is “determined” by the internal prediction (or “hypothesis”) that matches the actual input best [234, p. 201]. It may be conversely assumed that the absence of predictions leads to an absence of conscious perception.

In the framework of predictive coding theory two possible mechanisms for LOC can be inferred from our data: (1) the disruption of information transfer, predominantly in top-down direction, may indicate a failure to propagate predictions to hierarchically lower areas; (2) the decrease in locally available information and in entropy rates in PFC may indicate a failure to integrate information in an area central to the generation of a coherent model of the world and the generation of the corresponding predictions. These hypotheses are in line with findings reviewed in [221] and [237], which discuss activity in frontal areas and top-down modulatory activity as important to conscious perception.

Future research should investigate top-down information transfer more closely; for example, recent work suggests that neural activity in separate frequency bands may be responsible for the propagation of predictions and prediction errors respectively [236, 238]. Future experiments may target information transfer within a specific band to test if the disruption of top-down information transfer happens in the frequency band responsible for the propagation of predictions.

Last, it should be kept in mind that different anesthetics may lead to loss of consciousness by vastly different mechanisms. Ketamine, for example, seems to increase, rather than decrease, overall information transfer—at least in sub-anesthetic doses [21].

### 4.3.6 Comparison of the alternative approaches to the estimation of information theoretic measures and to statistical testing

The main analysis of this study was based on information theoretic estimators relying on distances between neighboring data points and on a permutation ANOVA. As each of these has its weaknesses we used additional alternative approaches, first, a Bayesian estimator for information theoretic measures, and second, statistical testing of non-aggregated data using LMMs. Both approaches returned results qualitatively very similar to those of the main analysis. Specifically, replacing neighbor-distance based estimators with Bayesian variants, we replicated the main finding of our study—a reduction in information transfer and locally available information, and an increase in information storage under anesthesia. However, using the Bayesian estimators we did not find a predominant reduction of top-down compared to bottom up information transfer. In contrast, using LMMs for statistical testing instead of a pANOVA additionally revealed a more pronounced reduction in top-down information transfer also in ferret 2 (in this animal, this effect was not significant when performing permutation tests based on the aggregated data).

The Bayesian estimators performed slightly worse than the next neighbor-based estimators in terms of their higher variance in estimates across recordings. Thus, even though Bayesian estimators are currently the best available estimators for discrete data, they may be a non-optimal choice for continuous data. A potential reason for this is the destruction of information on neighborhood relationships through data binning. This is, however, necessary to make the current Bayesian estimators applicable to continuous data.

In sum, we obtained similar results through three different approaches. This makes us confident that locally available information and information transfer indeed decrease under anesthesia, while the amount of predictable information increases.

### 4.3.7 On information theoretic measures obtained from from continuous time processes via time-discrete sampling

When interpreting the information theoretic measures presented in this work, it must be kept in mind that they were obtained via time-discrete sampling of processes that unfold in continuous time (such as the LFP, or spike trains). This time-discrete sampling was taken into account in recent work by Spinney and colleagues [239], who could show that the classic transfer entropy as defined by Schreiber is indeed ambiguous when applied to sampled data from time-continuous processes—it can either be seen as the integral of a transfer entropy rate in continuous time over one sampling interval, and is then given in bits, or it can be seen as an approximation to the continuous time transfer entropy rate itself, and be given in bits/sec. In our study we stick with the first notion of transfer entropy, and note that our results will numerically change when using different sampling intervals. In contrast to the case of transfer entropy covered in [239], corresponding analytical results for AIS are not known at present, as this is still a field of active research. To nevertheless elucidate the practical dependency of AIS on the sampling rate, when using our estimators, we have taken the original data and have up- and down-sampled them. Indeed the empirical AIS depended on the sampling rate (supporting information S3). Yet, all relations of AIS values between different isoflurane concentrations, i.e., the qualitative results, are independent of sampling. Thus, sampling effects do not affect the conclusions of the current study.

### 4.3.8 A cautionary note on the interpretation of information theoretic measures evaluated on local field potential data

When interpreting the results obtained in this study it should be kept in mind that LFP signals are not in themselves the immediate carriers of information relevant to individual neurons. This is because from a neuron's perspective information arrives predominantly in the form postsynaptic potentials generated by incoming spikes and chemical transmission in the synaptic cleft (but see [240] for a potential influence of LFPs on neural dynamics and computation). Thus, LFP signals merely reflect a coarse grained view of the underlying neural information processing. As a consequence, our results only hold in as far as at least some relevant information about the underlying information processing survives this coarse graining in the recording process, and little formal mathematical work has been carried out to estimate bounds on the amount of information available after coarse graining (but see [241]).

Yet, the enormous success that brain reading approaches had when based on local field potentials or on even more coarse grained magnetoencephalography (MEG) recordings (e.g. [242]) indicates that relevant information on neural information processing is indeed available at the level of these signals. However, successful attempts at decoding neural representations of stimuli or other features of the experimental setting should not lead us to misinterpret the information captured by information theoretic measures of neural processing as necessarily *being about something we can understand and link to the outside world*. Quite to the contrary, the larger part of information captured by these measures may be related to intrinsic properties of the unfolding neural computation.

### 4.3.9 Conclusion

Using two different methods for transfer entropy estimation, and two different statistical approaches, we found that locally available information and information transfer are reduced under isoflurane administration. The larger decrease in the locally available information was found at the source of the larger decrease of information transfer, not at its end point, or target. Therefore, previously reported reductions in information transfer under anesthesia may be caused by changes in local information processing rather than a disruption of long range connectivity. We suggest to put this hypothesis more into the focus of future research effort to understand the loss of consciousness under anesthesia. This suggestion receives further support from the fact that the synaptic targets of the anesthetic isoflurane, as used in this study, are most likely located in local circuits.

# 4.4 Methods

## 4.4.1 Electrophysiological Recordings

We conducted simultaneous electrophysiological recordings of the local field potential (LFP) in primary visual cortex (V1) and prefrontal cortex (PFC) of two female ferrets (17 to 20 weeks of age at study onset) under different levels of isoflurane (Fig. 4.9). The choice of the animal model is discussed further in [243]. Recordings were made in a dark environment during multiple, individual sessions of max. 2 h length, during which the animals' heads were fixed. For recordings, we used single metal electrodes acutely inserted in putative layer IV, measured 0.3 mm to 0.6 mm from the surface of cortex (tungsten micro-electrode, 250 $\mu$m shank diameter, 500 k$\Omega$ impedance, FHC, Bowdoin, ME). The hardware high pass filter was 0.1 Hz and the low pass filter was 5000 Hz. A silver chloride wire placed between the skull and soft tissue was used as the reference electrode. The reference electrode was located approximately equidistant between the recording electrodes. This location was selected in order to have little shared activity with either recording electrode. The same reference was used for both recording locations; also the same electrode position was used for both animals and all isoflurane concentrations. To verify that electrode placement was indeed in V1, we mapped receptive fields by eliciting visually evoked potentials in a separate series of experiments. We confirmed electrode placement in PFC by lesioning through the recording electrode after completion of data collection and post-mortem histology (as described in [243]). Details on surgical procedures can be found in [244]. Unfiltered signals were amplified with gain 1000 (model 1800, A-M Systems, Carlsborg, WA), digitized at 20 kHz (Power 1401, Cambridge Electronic Design, Cambridge, UK), and digitally stored using Spike2 software (Cambridge Electronic Design). For analysis, data were low pass filtered (300 Hz cutoff) and down-sampled to 1000 Hz.

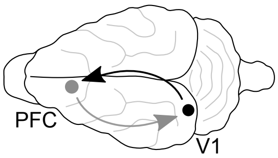


Fig. 4.9 **Recording sites in the ferret brain.** Prefrontal cortex (PFC, gray dot): anterior sigmoid gyrus; primary visual area (V1, black dot): lateral gyrus. Arrows indicate analyzed directions of information transfer (gray: top-down; black: bottom-up).

All procedures were approved by the University of North Carolina-Chapel Hill Institutional Animal Care and Use Committee (UNC-CH IACUC) and exceed guidelines set forth by the National Institutes of Health and U.S. Department of Agriculture.

LFPs were recorded during wakefulness (condition *iso* 0.0 %, number of recording sessions: 8 and 5 for ferret 1 and 2, respectively) and with different concentrations of anesthetic: 0.5 % isoflurane with xylazine (condition *iso* 0.5 %, number of sessions: 5 and 6), as well as 1.0 % isoflurane with xylazine (condition *iso* 1.0 %, number of sessions: 10 and 11). In the course of pilot experiments, both concentrations *iso* 0.5 % and *iso* 1.0 % lead to a loss of the righting reflex; however, a systematic assessment of this metric during recordings was technically not feasible. Additionally, animals were administered 4.25 mL/h 5 % dextrose lactated Ringer and 0.015 mL/h xylazine via IV.

LFP recordings from each session were cut into epochs of 4.81 s length to be able to remove segments of data if they were contaminated by artifacts (e.g., due to movement). We chose a relatively short epoch length to avoid removing large chunks of data when there was only a short transient artifact. This resulted in 196 to 513 epochs per recording (mean: 428.6) for ferret 1, and 211 to 526 epochs (mean: 472.8) for ferret 2. epochs with movement artifacts were manually rejected (determined by extreme values in the LFP raw traces). In the *iso* 0.0 % condition, infrared videography was used to verify that animals were awake during the whole recording; additionally, *iso* 0.0 % epochs with a relative delta power (0.5 Hz to 4.0 Hz) of more than 30 % of the total power from 0.5 Hz to 50 Hz were rejected to ensure that only epochs during which the animal was truly awake entered further analysis.

### 4.4.2 Information theoretic measures

To measure information transfer between recording sites V1 and PFC, we estimated the transfer entropy [12] in both directions of possible interactions, $PFC \rightarrow V1$ and $V1 \rightarrow PFC$. To investigate local information processing within each recording site, we estimated active information storage ($AIS$) [30] as a measure of predictable information, and we estimated differential entropy ($H$) [245] as a measure of information available locally. We will now explain the applied measures and estimators in more detail, before we describe how these estimators were applied to data from electrophysiological recordings in the next section. To mathematically formalize the estimation procedure from these data, we assume that neural time series recorded from two systems $\mathcal{X}$ and $\mathcal{Y}$ (e.g. cortical sites) can be treated as collections of realizations $x_t$ and $y_t$ of random variables $X_t$, $Y_t$ of two random processes $X = \{X_1, \ldots, X_t, \ldots, X_N\}$ and $Y = \{Y_1, \ldots, Y_t, \ldots, Y_N\}$. The index $t$ here indicates samples in time, measured in units of the dwell time (inverse sampling rate) of the recording.

**Transfer entropy** Transfer entropy [12, 13] is defined as the mutual information between the future of a process $Y$ and the past of a second process $X$, conditional on the past of $Y$. Transfer entropy thus quantifies the information we obtain about the future of $Y$ from the past of $X$, taking into account information from the past of $Y$. Taking this past of $Y$ into account here removes information redundantly available in the past of both $X$ and $Y$, and reveals information provided synergistically by them [76]. In this study, we used an improved estimator of transfer entropy presented in [75], which accounts for arbitrary information transfer delays:

$$TE_{SPO}(X \rightarrow Y, t, u) = I\left(Y_t; \mathbf{X}_{t-u}^{d_X} | \mathbf{Y}_{t-1}^{d_Y}\right), \tag{4.1}$$

where $I$ is the conditional mutual information (or the differential conditional mutual information for continuous valued variables) between $Y_t$ and $\mathbf{X}_{t-u}^{d_X}$, conditional on $\mathbf{Y}_{t-1}^{d_Y}$; $Y_t$ is the future value of random process $Y$, and $\mathbf{X}_{t-u}^{d_X}$, $\mathbf{Y}_{t-1}^{d_Y}$ are the past states of $X$ and $Y$, respectively. Past states are collections of past random variables

$$\mathbf{Y}_{t-1}^{d_Y} = \left(Y_{t-1}, Y_{t-1-\tau}, \ldots, Y_{t-1-(d_Y-1)\tau}\right), \tag{4.2}$$

that form a delay embedding of length $d_Y$ [73], and that render the future of the random process conditionally independent of all variables of the random process that are further back in time than the variables forming the state. Parameters $\tau$ and $d$ denote the embedding delay and embedding dimension and can be found through optimization of a local predictor as proposed in [133] (see next section on the estimation of information theoretic measures). Past states constructed in this manner are then maximally informative about the present variable of the target process, $Y_t$, which is an important prerequisite for the correct estimation of transfer entropy (see also [75]).

In our estimator (Eq. 4.1), the variable $u$ describes the assumed information transfer delay between the processes $X$ and $Y$, which accounts for a physical delay $\delta_{X,Y} \geq 1$ [75]. The estimator thus accommodates arbitrary physical delays between processes. The true delay $\delta_{X,Y}$ must be recovered by "scanning" various assumed delays and keeping the delay that maximizes $TE_{SPO}$ [75]:

$$\hat{\delta}_{X,Y} = \arg\max_{u} \left( TE_{SPO} \left( X \rightarrow Y, t, u \right) \right). \tag{4.3}$$

**Active Information Storage** $AIS$ [30] is defined as the (differential) mutual information between the future of a signal and its immediate past state

$$AIS(Y_t) = I\left(Y_t; \mathbf{Y}_{t-1}^{d_Y}\right), \tag{4.4}$$

where $Y$ again is a random process with present value $Y_t$ and past state $\mathbf{Y}_{t-1}^{d_Y}$ (see Eq. 4.2). $AIS$ thus quantifies the amount of predictable information in a process or the information that is currently in use for the next state update [30]. $AIS$ is low in processes that produce little information or are highly unpredictable, e.g., fully stochastic processes, whereas $AIS$ is highest for processes that visit many equi-probable states in a predictable sequence, i.e., without branching. In other words, $AIS$ is high for processes with "rich dynamics" that are predictable from the processes' past [66]. A reference implementation of $AIS$ can be found in the Java Information Dynamics Toolkit (JIDT) [87]. As for $TE_{SPO}$ estimation, an optimal delay embedding $\mathbf{Y}_{t-1}^{d_Y}$ may be found through optimization of the local predictor proposed in [133].

Note, that $AIS$ is upper bounded by the entropy as:

$$\begin{aligned} AIS(Y) &= I\left(Y_t; \mathbf{Y}_{t-1}^{d_Y}\right) \\ &= H\left(Y_t\right) - H\left(Y_t | \mathbf{Y}_{t-1}^{d_Y}\right). \end{aligned} \tag{4.5}$$

**Differential entropy** The differential entropy $H$ (see for example [245]) expands the classical concept of Shannon's entropy for discrete variables to continuous variables:

$$H = -\int_{\mathbb{X}_{t-u}^{d_X}} f(\mathbf{X}_{t-u}^{d_X}) \log f(\mathbf{X}_{t-u}^{d_X}) \, d\mathbf{X}_{t-u}^{d_X}, \tag{4.6}$$

where $f(Y_t)$ is the probability density function of $Y_t$ over the support $\mathbb{Y}$. Entropy quantifies the average information contained in a signal. Based on the differential entropy the corresponding measures for mutual and conditional mutual information and, thereby, active information storage and transfer entropy can be defined.

### 4.4.3 Entropy as an upper bound on information transfer

The transfer entropy from Eq. 4.1 can be rewritten as:

$$\begin{aligned} TE_{SPO}(X \rightarrow Y, t, u) =& I\left(\mathbf{X}_{t-u}^{d_X} : Y_t | \mathbf{Y}_{t-1}^{d_Y}\right) \\ =& H(\mathbf{X}_{t-u}^{d_X} | \mathbf{Y}_{t-1}^{d_Y}) - H(\mathbf{X}_{t-u}^{d_X} | \mathbf{Y}_{t-1}^{d_Y}, Y_t)\ . \end{aligned} \tag{4.7}$$

By dropping the negative term on the right hand side we obtain an upper bound (as already noticed by [113, p. 65]), and by realizing that a conditional entropy is always smaller than the corresponding *unconditional* one, we arrive at

$$TE_{SPO}(X \rightarrow Y, t, u) \leq H(\mathbf{X}_{t-u}^{d_X}). \tag{4.8}$$

This indicates that the overall entropy of the source states is an upper bound. Several interesting other bounds on information transfer exist as detailed in [113], yet these are considerably harder to interpret and were not the focus of the current presentation.

### 4.4.4 Estimation of information theoretic measures

In this section we will describe how the information theoretic measures presented in the last section may be estimated from neural data. In doing so, we will also describe the methodological pitfalls mentioned in the introduction in more detail and we will describe how these were handled here. If not stated otherwise, we used implementations of all presented methods in the open source toolboxes TRENTOOL [86] and JIDT [87], called through custom MATLAB® scripts (MATLAB 8.0, The MathWorks® Inc., Natick, MA, 2012). Time series were normalized to zero mean and unit variance before estimation.

**Estimating information theoretic measures from continuous data** Estimation of information theoretic measures from continuous data is often handled by simply discretizing the data. This is done either by binning or the use of symbolic time series—mapping the continuous data onto a finite alphabet. Specifically, the use of symbolic time series for transfer entropy estimation was first introduced by [246] and maps the continuous values in past state vectors with length $d$ (Eq. 4.2) onto a set of rank vectors. Hence, the continuous-valued time series is mapped onto an alphabet of finite size $d!$. After binning or transformation to rank vectors transfer entropy and the other information theoretic measures can then be estimated using plug-in estimators for discrete data, which simply evaluate the relative frequency of occurrences of symbols in the alphabet. Discretizing the data therefore greatly simplifies the estimation of transfer entropy from neural data, and may even be necessary for very small data sets. Yet, binning ignores the neighborhood relations in the continuous data and the use of symbolic times series destroys important information on the absolute values in the data. An example where transfer entropy estimation fails due to the use of symbolic time series is reported in [156] and discussed in [75]: In this example, information transfer between two coupled logistic maps was not detected by symbolic transfer entropy [156]; only when estimating $TE_{SPO}$ directly using an estimator for continuous data, the information transfer was identified correctly [75]. To circumvent the problems with binned or symbolic time series, we here used a nearest-neighbor based $TE_{SPO}$-estimator for continuous data, the Kraskov-Stögbauer-Grassberger (KSG) estimator for mutual information described in [83]. At present, this estimator has the most favorable bias properties compared to similar estimators for continuous data. The KSG-estimator leads to the following expression for the estimation of $TE_{SPO}$ as introduced in Eq. 4.1 [13, 247]:

$$\begin{aligned} TE_{SPO}(X \rightarrow Y, t, u) &= I(Y_t : \mathbf{X}_{t-u}^{d_X} | \mathbf{Y}_{t-1}^{d_Y}) \\ &= \psi(k) + \langle \psi(n_{\mathbf{y}_{t-1}^{d_Y}} + 1) - \psi(n_{y_t \mathbf{y}_{t-1}^{d_Y}} + 1) - \psi(n_{\mathbf{y}_{t-1}^{d_Y} \mathbf{x}_{t-u}^{d_X}} + 1) \rangle_r, \end{aligned} \tag{4.9}$$

where $\psi$ denotes the digamma function, $k$ is the number of neighbors in the highest-dimensional space spanned by variables $Y_t$, $\mathbf{Y}_{t-1}^{d_Y}$, $\mathbf{X}_{t-u}^{d_X}$, and is used to determine search radii for the lower dimensional subspaces; $n_.$ are the number of neighbors within these search radii for each point in the lower dimensional search spaces spanned by the variable indicated in the subscript. Angle brackets indicate the average over realizations $r$ (e.g. observations made over an ensemble of copies of the systems or observations made over time in case of stationarity, which we assumed here). We used a $k$ of 4 as recommended by Kraskov [85, p. 23], such as to balance the estimator's bias—which decreases for larger $k$—and variance—which increases for larger $k$ (see also [84] for similar recommendations based on simulation studies). For a detailed derivation of $TE_{SPO}$-estimation using the KSG-estimator see [13, 83, 247].

The KSG-estimator comes with a bias that is not analytically tractable [85], hence, estimates can not be interpreted at face value, but have to be tested for their statistical significance against the null-hypothesis of no information transfer [13, 86]. We thus performed a permutation test on the $TE_{SPO}$ estimates against surrogate data to assess the statistical significance of the estimated information transfer [86]. We used the ensemble-method for transfer entropy estimation [208], implemented in TRENTOOL [86]. The ensemble method allows to pool data over epochs, which maximizes the amount of data entering the estimation, while providing an efficient implementation of this estimation procedure using graphics processing units (GPU). We tested the statistical significance of $TE_{SPO}$ in both directions of interaction in the *iso* 0.0 % condition. We only tested this condition, because $TE_{SPO}$ was expected to be reduced for higher isoflurane levels based on the results of existing studies. Because the estimation of $TE_{SPO}$ is computationally heavy, we used a random subset of 50 epochs to reduce the running time of this statistical test (see supporting information S4 for theoretical and practical running times of the used estimators). We first tested information transfer estimates ($TE_{SPO}$) for their significance within individual recording sessions, and then used a binomial test to establish the statistical significance *over* recordings. We used a one-sided Binomial test under the null hypothesis of no significant $TE_{SPO}$ estimates, where individual estimates $l$ were assumed to be $B(l, p_0, n)$-distributed, with $p_0 = 0.05$ and $n = 5$ for animal 1 and $p_0 = 0.05$ and $n = 8$ for animal 2.

As the KSG-estimator used for estimating $TE_{SPO}$ (Eq. 4.9) is an estimator of mutual information it can also be used for the estimation of $AIS$:

$$\begin{aligned} AIS(Y) &= I(Y_t : \mathbf{Y}_{t-1}^{d_Y}) \\ &= \psi(k) - 1/k + \psi(N) - \langle \psi(n_{y_t}) + \psi(n_{\mathbf{y}_{t-1}^{d_Y}}) \rangle_r, \end{aligned} \tag{4.10}$$

where again $\psi$ denotes the digamma function, $k$ is the number of neighbors in the highest-dimensional space, $N$ is the number of realizations, and $n_\cdot$ denotes the number of neighbors for each point in the respective search space. Again, we chose $k = 4$ for the estimation of AIS (see above). Note that the sampling rate has an effect on the estimated absolute values of AIS (a simulation of the effect of sampling on AIS estimates are shown as supporting information S3)—however, qualitative results are not influenced by the choice of sampling rate; in other words, relative differences between estimates are the same for different choices of sampling rates. As a consequence, for AIS estimation the sampling rate should be constant over data sets if the aim is to compare these estimates.

A conceptual predecessor of the KSG-estimator for mutual information is the Kozachenko-Leonenko (KL) estimator for differential entropies [134]. The KL-estimator also allows for the estimation of $H$ from continuous data and reads

$$H(X) = -\psi(1) + \psi(N) + \sum_{i=1}^{N} \log(\epsilon(i)), \tag{4.11}$$

where $\epsilon(i)$ is twice the distance from data point $i$ to its $k$-th nearest neighbor in the search space spanned by all points.

**Bayesian Estimators for discretized data** For Bayesian estimation we converted the continuous LFP time series to discrete data by applying voltage bins as follows: The voltages $\pm 3$ standard deviations around the mean of the LFP were subdivided into $N$ equally spaced bins. We added two additional bins containing all the values that were either smaller or larger than the 3 SD region, amounting to a total number of bins $N_{bins} = N + 2$. We then calculated $H$, $AIS$ and $TE_{SPO}$ for the discrete data. For $AIS$ and $TE_{SPO}$ estimation states were defined using the same dimension $d$ and $\tau$ as for the KSG-estimator, optimized using the Ragwitz criterion. We decomposed $TE_{SPO}$ into four entropies (Eq. 4.9), and $AIS$ into three entropies, which we then estimated individually [248]. To reduce the bias introduced by the limited number of observed states, we used the NSB-estimator by Nemenman, Shafee, and Bialek [225], which is based on the construction of an almost uniform prior over the expected entropy using a mixture of symmetric Dirichlet priors $\mathcal{P}_\beta$. The estimator has been shown to be unbiased for a broad class of distributions that are typical in $\mathcal{P}_\beta$ [226].

We further applied the recently proposed estimator by Archer et al. [227] that uses a prior over distributions with infinite support based on Pitman-Yor-processes. In contrast to the NSB prior, this prior also accounts for heavy-tailed distributions that one encounters frequently in neuronal systems and does not require knowledge of the support of the distribution.

When estimating entropies for the embedding dimensions used here, the number of possible states or “words” is between $K = 12^{14}$ and $K = 12^{31}$. This is much larger than the typical number of observed states per recording of around $N =$

$5 \cdot 10^5$. As a consequence, a precise estimation of entropies is only possible if the distribution is sparse, i.e. most words have vanishing probability. In this case, however, the estimates should be independent of the choice of support $K$ as long as $K$ is sufficiently large and does not omit states of finite probability. We chose $K' = 10^{12}$ for the results shown in this paper, which allowed a robust computation of the NSB estimator instead of the maximum support $K$ that results from simple combinatorics.

**Finding optimal embedding parameters** The second methodological problem raised in the introduction was the choice of embedding parameters for transfer entropy estimation. One important parameter here is the choice of the total signal history when constructing past states for source and target signal (see Eq. 4.1 and 4.2). Failure to properly account for signal histories may lead to a variety of errors, such as underestimating transfer entropy, failure to detect transfer entropy altogether, or the detection of spurious transfer entropy. Transfer entropy is underestimated or missed if the past state of the *source* time series does not cover all the relevant history, i.e., the source is under-embedded. In contrast, spurious transfer entropy may be detected if the past state of the *target* time series is under-embedded, such that spurious detection of transfer entropy is a false positive and therefore the most serious error. One scenario where spurious transfer entropy results from under-embedding is shown in Fig. 4.10.

The choice of an optimal embedding is also relevant for the estimation of $AIS$, where under-embedding leads to underestimation of the true $AIS$. Note that on the other hand, we can not increase the embedding length to arbitrarily high values because this leads to computationally intractable problem sizes and requires exponentially more data for estimation.

Optimal embedding parameters $d$ and $\tau$ may be found through the optimization of a local predictor proposed by Ragwitz [133]. Ragwitz' criterion tests different combinations of a range of values for $d$ and $\tau$. The current combination is used to embed each point in a time series, then, the future state for each point is predicted from the future states of its neighbors. The parameter combination that leads to the best prediction on average is used as the optimal embedding. To determine the neighbors of a point, we used a $k$-nearest-neighbor search with $k = 4$, i.e., the same value for $k$ as was used for $k$-nearest-neighbor searches when estimating information-theoretic measures. We minimized the mean squared error when optimizing Ragwitz' local predictor. Further details on Ragwitz' criterion can be found in the documentations of the TRENTOOL [86] and JIDT [87] toolbox.

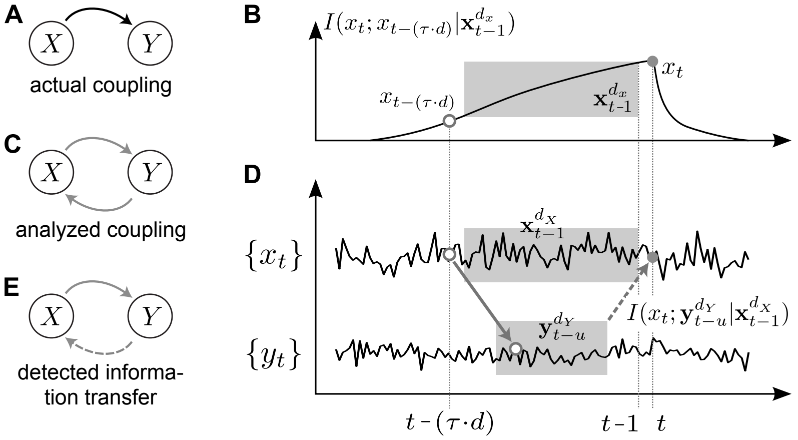


Fig. 4.10 **Spurious information transfer resulting from under-embedding (modified from [13]).** (A) Actual coupling between processes $X$ and $Y$. (B) Mutual information between the present value in $X$, $x_t$, and a value in the far past of $X$, $x_{t-(\tau\cdot d)}$, conditional on all intermediate values $\mathbf{x}^{d_X}_{t-1}$ (shaded box), the mutual information is non-zero, i.e., $x_{t-(\tau\cdot d)}$ holds some information about $x_t$. (C) Both directions of interaction are analyzed; (D) Information in $x_{t-(\tau\cdot d)}$ (white sample point) is transferred to $Y$ (solid arrow), because of the actual coupling $X \rightarrow Y$. The information in $x_{t-(\tau\cdot d)}$ about $x_t$ is thus transferred to the past of $Y$, $\mathbf{y}^{d_Y}_{t-u}$, which thus becomes predictive of $x_t$ as well. Assume now, we analyzed information transfer from $Y$ to $X$, $I(x_t : \mathbf{y}^{d_Y}_{t-u}|\mathbf{x}^{d_X}_{t-1})$, without a proper embedding of $X$, $\mathbf{x}^{d_X}_{t-1}$: Because of the actually transferred information from $x_{t-(\tau\cdot d)}$ to $\mathbf{y}^{d_Y}_{t-u}$, the mutual information $I(x_t : \mathbf{x}^{d_X}_{t-1})$ is non-zero. If we now under-embed $X$, such that the information in $x_{t-(\tau\cdot d)}$ is not contained in $\mathbf{x}^{d_X}_{t-u}$ and is not conditioned out, $I(x_t : \mathbf{y}^{d_Y}_{t-1}|\mathbf{x}^{d_X}_{t-u})$ will be non-zero as well. In this case, under-embedding of the target $X$ will lead to the detection of spurious information transfer in the non-coupled direction $Y \rightarrow X$. (E) Information transfer is falsely detected for both directions of interaction, the link from $Y$ to $X$ is spurious (dashed arrow).

Other approaches for embedding parameter optimization have been proposed, see for example non-uniform embedding using mutual information to determine all relevant past samples as proposed by [115].

**Reconstruction of information transfer delays** The third methodological problem raised in the introduction was failure to account for a physical delay $\delta$ between neural sites when estimating transfer entropy. In our estimator $TE_{SPO}$ (Eq. 4.9) we account for $\delta$ by introducing the parameter $u$. The delay $u$ needs to be optimized to correctly estimate $TE_{SPO}$. If $u$ is not optimal, i.e., $u$ is not sufficiently close to $\delta$ (Fig. 4.5 and [75]), information transfer may be underestimated or not measured at all. This is because choosing the parameter $u$ too large ($u \gg \delta$) means that the information present in the evaluated samples of the source is also present in the history of the target already, and conditioned away. In contrast, choosing the parameter $u$ too small means that the information of the evaluated samples of the source will only arrive in the future of the current target sample, and is useless for providing information about it (Fig. 4.5A).

It can be proven for bivariate systems that $TE_{SPO}$ becomes maximal when the true delay $\delta$ is chosen for $u$ [75]. Therefore, the true delay $\delta$—and thus an optimal choice for $u$—can be found by using the value for $u$ that maximizes $TE_{SPO}$ [75]. This optimal $u$ can be found by scanning a range of assumed values. In the present study, we scanned values ranging from 0 ms to 20 ms. (Note that assumed values should be physiologically plausible to keep the computations practically feasible.)

Accounting for the information transfer $\delta$ by finding optimal parameters $u$ for $TE_{SPO}$ estimation has important consequences when calculating indices from estimated $TE_{SPO}$, such as $TE_{net}$:

$$TE_{net} = \frac{TE_{SPO}(X \rightarrow Y, t, u_{X\rightarrow Y}) - TE_{SPO}(Y \rightarrow X, t, u_{Y\rightarrow X})}{TE_{SPO}(X \rightarrow Y, t, u_{X\rightarrow Y}) + TE_{SPO}(Y \rightarrow X, t, u_{Y\rightarrow X})}, \tag{4.12}$$

or variations of this measure. The $TE_{net}$ is popular in anesthesia research [33, 35] and indicates the predominant direction of information transfer between two bidirectionally coupled processes $X$ and $Y$ ($TE_{net} \geq 0$ if $TE_{SPO}(X \rightarrow Y, t, u_{X\rightarrow Y}) \geq TE_{SPO}(Y \rightarrow X, t, u_{Y\rightarrow X})$ and $TE_{net} < 0$ if $TE_{SPO}(Y \rightarrow X, t, u_{Y\rightarrow X}) > TE_{SPO}(X \rightarrow Y, t, u_{X\rightarrow Y})$). However, if values for $u_{X\rightarrow Y}$ and $u_{Y\rightarrow X}$ are not optimized individually, $TE_{net}$ may take on arbitrary signs: In Fig. 4.5B, we show a toy example of two coupled Lorenz systems [75], where the absolute difference between raw $TE_{SPO}$ values changes as a function of a common $u$ for both directions and where the difference even changes signs for values $u > 65$. To obtain a meaningful value from $TE_{net}$ we thus need to find the individually optimal choices of $u$ for both directions of transfer—in the example, these optima are found at $u_{X\rightarrow Y} = 46$ and $u_{Y\rightarrow X} = 76$, leading to the "true" difference.

**Simulating the effect of filtering on information transfer delay reconstruction** To simulate the effects of filtering as a preprocessing technique on the ability of the $TE_{SPO}$ estimator to reconstruct the correct information transfer delay, we simulated two coupled time series, for which we estimated transfer entropy before filtering and after band-pass filtering with different bandwidths. The simulation was repeated 50 times.

We simulated two time series with 100 000 samples each, which were drawn from a uniform random distribution over the open interval (0,1). We introduced a coupling between the time series by adding a scaled version (factor 0.2) of the first time series to the second with a delay of 10 samples. We estimated $TE_{SPO}$ from the first to the second time series using the KSG-estimator implemented in the JIDT toolbox, with $k = 4$ and a history of one sample for both source and target.

We estimated $TE_{SPO}$ with and without band-pass filtering the data for different values of $u$, ranging from 1 to 20 samples. We filtered the data using a fourth order, causal Butterworth filter, implemented in the MATLAB toolbox FieldTrip [150]. We filtered the data using four different bandwidths: from 0.1 Hz to 300 Hz (corresponding to the filtering done in this study), from 0.1 Hz to 200 Hz, from 12 Hz to 30 Hz (corresponding to the beta frequency range), and from 4 Hz to 8 Hz (corresponding to the theta frequency range).

### 4.4.5 Statistical testing using permutation testing

To test for statistically relevant effects of isoflurane levels and direction of interaction or recording site on estimated measures, we performed two-factorial permutation analyses of variance (pANOVA) for each animal and estimated measure [238, 249–251]. We used a MATLAB® implementation of the test described in [251], which is compatible with the FieldTrip toolbox data format [150].

The permutation ANOVA can be used if the normality assumptions of parametric ANOVA are violated or—as in the present study—if assumptions are not testable due to too few data points per factor level. The permutation ANOVA evaluates the significance of the main effect of individual factors or their interaction effect under the null-hypothesis of no experimental effect at all. The significance is evaluated by calculating a F-ratio for the effect from the original data; this original F-ratio is then compared against a distribution of F-ratios obtained from permuted data. The F-ratio's p-value is calculated as the fraction of ratios obtained from permuted data that is bigger than the original F-ratio. When permuting data, it is crucial to only permute data in such a way that the currently investigated effect is destroyed while all other effects are kept intact [249]: for example, consider a two-factorial design with factors $A$ and $B$—if the main effect of factor $A$ is to be tested, the assignment of levels of $A$ to data points has to be permuted; yet, the permutation of levels of $A$ can only happen *within* levels of factor $B$ such as to not simultaneously destroy the effect of factor $B$. Thus, when testing for the effect of one factor, the effect of all other factors is preserved, making sure that variability due to one factor is tested for while the variability due to the other factor is held constant. However, this permutation scheme is not applicable to the interaction effect because it leaves no possible permutations—destroying the interaction effect through permutation always also destroys the main effects of individual factors. Here, both factors have to be permuted, which yields an approximative test of the interaction effect (see [251] for a discussion of permutation strategies and simulations).

The factors in the permutation ANOVA were *isoflurane level* and *direction* for $TE_{SPO}$ estimates, and *isoflurane level* and *recording site* for $AIS$ and $H$ estimates. The number of permutations was set to 10 000. In the present study, data was recorded in different sessions and segmented into epochs. As a result, estimates for individual epochs should not be pooled over recordings for statistical analysis (see for example [252], and also the next paragraph for a discussion). We therefore aggregated estimates over epochs to obtain one estimate per recording session. We used the median to aggregate the estimated values for each recording session over individual epochs, because the distribution of measures over epochs was skewed and we considered the median a more exact representation of the distributions' central tendencies. The aggregation of data resulted in relatively few observations per ANOVA cell and also in unequal number of observations between cells, thus violating two basic assumptions of parametric ANOVA (too few observations make it impossible to test for parametric assumptions like homogeneity of variances). Therefore, we used the non-parametric permutation approach over a parametric one, because the permutation ANOVA does not make any assumptions on data structure.

### 4.4.6 Statistical testing using linear mixed models

As described above, LFP recordings were conducted in epochs over multiple recording sessions. This introduces a so-called "nested design" [252], i.e., a hierarchical structure in the data, where data epochs are nested within recordings, which are nested in animals. Such structures lead to systematic errors or dependence within the data. This violates the assumption of uncorrelated errors made by the most common tests derived from the general linear model and leads to an inflation of the type I error [252]. A measure of the degree of dependence is the intraclass correlation coefficient (ICC), which was 0.35 for ferret 1 and 0.19 for ferret 2, indicating a significant dependency within the data (see [252] for a discussion of this measure). Thus, we performed an additional statistical tests where we again tested for a significant effect of the two factors *isoflurane level* and *direction* as well as their interaction, but we additionally modeled the nested structure in our data by a random factor *recordings*. Such a model is called a linear mixed effects model [253], and may yield higher statistical power than aggregating data within one level of the nested design (as was done for the permutation ANOVA). We used the following model for both animals separately:

$$y_{ij} = \beta_0 + \gamma_j + \beta_1 D_i + \beta_2 A_i^{(0.5)} + \beta_3 A_i^{(1.0)} + \epsilon_{ij},$$

where $y_{ij}$ is the $TE_{SPO}$ value from the $i$-th epoch in recording $j$, modeled as a function of isoflurane level and direction of interaction, where $\beta_0$ describes the model intercept, $\beta_k$ are the regression coefficients and describe fixed effects, $\gamma_j$ is the random deviation of recording $j$ from the intercept $\beta_0$, and $\epsilon_{ij}$ describes random noise. $D_i$, $A_i^{(0.5)}$, and $A_i^{(1.0)}$ are predictor variables, encoding factors *direction* as

$$D_i := \begin{cases} 1 & \text{if } \textit{direction} \text{ is } PFC \rightarrow V1, \\ -1 & \text{if } \textit{direction} \text{ is } V1 \rightarrow PFC, \end{cases} \tag{4.13}$$

and factor *isoflurane level* as

$$A_i^{(0.5)} := \begin{cases} 1 & \text{if } \textit{isoflurane level} \text{ is } \textit{iso 0.5 \%}, \\ 0 & \text{else}, \end{cases} \tag{4.14}$$

$$A_i^{(1.0)} := \begin{cases} 1 & \text{if } \textit{isoflurane level} \text{ is } \textit{iso 1.0 \%}, \\ 0 & \text{else}. \end{cases} \tag{4.15}$$

Note that we used dummy coding for factor *direction* so that the the estimated effect $\beta_1$ can be interpreted like the simple or main effect in a standard ANOVA framework. We used contrast coding for factor *isoflurane level* which allows to interpret estimated effects $\beta_2$ and $\beta_3$ as deviations from a reference group (in this case the condition *iso* 0.0 %). We further assume that noise was i.i.d. and $\epsilon_{ij} \sim \mathcal{N}\left(0, \sigma_\epsilon^2\right)$ and $\gamma_j \sim \mathcal{N}\left(0, \sigma_\gamma^2\right)$.

We used the R language [254] and the function `lmer` from the `lme4`-package [255] for model fitting. We assessed statistical significance of individual factors by means of model comparison using the maximum likelihood ratio between models [256]. To allow for this model comparison, we used maximum likelihood estimation, instead of restricted maximum likelihood estimation, of random and fixed effects. To test for main effects, we compared models including individual factors *isoflurane level* ($fm\_a$) and *direction* ($fm\_d$) to a Null model ($fm\_0$) including only the random effect; to furthermore test for an interaction effect, we compared the model including an interaction term ($fm\_axb$) to a model where both factors only entered additively ($fm\_ab$).

The models were fitted to 15 973 $TE_{SPO}$ values from ferret 1 and 18 202 $TE_{SPO}$ values from ferret 2, respectively.

### 4.4.7 Simulating the effect of reduced source entropy on transfer entropy

To test whether changes in source entropy influenced the transfer entropy despite unchanged coupling, we simulated two test cases, with high and low source entropy, respectively, while the coupling between source and target process were held constant. To simulate the two test cases, we randomly selected two recordings—one from the *iso* 0.0 % condition, which on average showed higher source entropy, and one from the *iso* 1.0 % condition, which on average showed lower source entropy (Figs. 4.1 and 4.2). In the recording from the *iso* 0.0 % condition, we permuted epochs in the target time series to destroy all information transfer present in the original data. From the permuted data, we simulated two cases of artificial coupling, first, using the high-entropy source time course from the *iso* 0.0 % condition; and second, using the low-entropy source time course from the *iso* 1.0 % condition. The coupling was simulated by adding a filtered, scaled and delayed version of the respective source time course to the target time course for each epoch. For filtering, we used a Gaussian filter with a smoothing of 10 samples; for the scaling factor, we used a value of 0.2, which resulted in a $TE_{SPO}$ value for the high-entropy test case that was close to the $TE_{SPO}$ in the *iso* 0.0 % condition. By replacing the original coupling in both recordings with a simulated coupling, we made sure that the coupling was constant for both test cases.

For both test cases, we estimated $TE_{SPO}$ and $H(\mathbf{X}_{t-u}^{d_X})$ following the estimation procedures described above. We tested differences in $TE_{SPO}$ using a permutation independent samples t-test with 10 000 permutations. To make sure that our simulation reflected information transfer found in the original data, we further tested for a significant difference between $TE_{SPO}$ in the original data from the *iso* 0.0 % recording and $TE_{SPO}$ in the high-entropy test case (using the source time *iso* 0.0 % condition). Here, we found no significant difference, indicating that the transfer entropy in our test case did not differ significantly from the transfer entropy found in the original data.

### 4.4.8 Correlating information-theoretic measures with other time-series properties

We correlated $TE_{SPO}$, $AIS$, and $H$ with more conventional measures from time-series analysis, namely, the autocorrelation decay time (ACT), signal variance, and power in individual frequency bands. This correlation was performed to investigate whether information-theoretic measures captured signal properties that could be described equally well by more simple measures.

The ACT was calculated by finding the lag at which the autocorrelation coefficient decayed below $e^{-1}$. The signal variance was calculated as the variance over time for each recording after subtracting the mean of the time series. The band power was calculated for individual frequency bands (delta = 0.5 Hz to 4 Hz, theta = 4 Hz to 8 Hz, alpha = 8 Hz to 12 Hz, beta = 12 Hz to 30 Hz, gamma = 30 Hz to 40 Hz, following [244]), using the multitaper method with discrete prolate spheroidal sequences (Slepian sequences) as tapers (smoothing = 1 Hz) implemented in the MATLAB toolbox FieldTrip [150].

We calculated correlations between information-theoretic measures and conventional measures for individual isoflurane levels and recording sites or direction of interaction, respectively. For each correlation, we pooled data over recording sessions for the respective isoflurane level and calculated Spearman's rank correlation. We tested the correlation for significance using a restricted permutation test, where permutations were allowed only *within* one recording, accounting for the nested experimental design (see [252] and section *Statistical testing using linear mixed models*, below). Additionally, we calculated the correlation as well as the variance explained, $R^2$, for individual recordings, because calculating $R^2$ for the coefficient calculated from *pooled* data does not yield interpretable results.

## 4.5 Supporting Information

**S1 Dataset Group statistical data and estimated information-theoretic values entering statistical tests.**

**S2 Correlation tables Correlation between information-theoretic measures and traditional measures of time-series properties.**

Tab. 4.5 Correlation of $AIS$ with autocorrelation decay time (ACT). $\widetilde{R}^2$ and $max(R^2)$ indicate the median and maximum of $R^2$ over recordings per condition, respectively.

| animal | isoflurane level | recording site | $p$ | $\widetilde{R}^2$ ($max(R^2)$) |
|---|---|---|---|---|
| 1 | iso 0.0 % | PFC | 1.0000 | 0.024 (0.289) |
| | | V1 | 0.1028 | 0.003 (0.153) |
| | iso 0.5 % | PFC | 0.0013** | 0.001 (0.122) |
| | | V1 | 0.0000*** | 0.014 (0.114) |
| | iso 1.0 % | PFC | 0.0000*** | 0.005 (0.271) |
| | | V1 | 0.9998 | 0.004 (0.051) |
| 2 | iso 0.0 % | PFC | 0.0000*** | 0.022 (0.116) |
| | | V1 | 0.0000*** | 0.029 (0.175) |
| | iso 0.5 % | PFC | 0.7672 | 0.007 (0.018) |
| | | V1 | 0.0000*** | 0.017 (0.076) |
| | iso 1.0 % | PFC | 0.0000*** | 0.008 (0.121) |
| | | V1 | 0.8247 | 0.009 (0.214) |

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Tab. 4.6 Correlation of $H$ with signal variance. $\widetilde{R}^2$ and $max(R^2)$ indicate the median and maximum of $R^2$ over recordings per condition, respectively.

| animal | isoflurane level | recording site | $p$ | $\widetilde{R}^2$ ($max(R^2)$) |
|---|---|---|---|---|
| 1 | iso 0.0 % | PFC | 1.0000 | 0.087 (0.705) |
| | | V1 | 1.0000 | 0.079 (0.452) |
| | iso 0.5 % | PFC | 1.0000 | 0.036 (0.093) |
| | | V1 | 1.0000 | 0.062 (0.217) |
| | iso 1.0 % | PFC | 1.0000 | 0.010 (0.474) |
| | | V1 | 1.0000 | 0.018 (0.083) |
| 2 | iso 0.0 % | PFC | 1.0000 | 0.042 (0.699) |
| | | V1 | 1.0000 | 0.162 (0.363) |
| | iso 0.5 % | PFC | 1.0000 | 0.041 (0.130) |
| | | V1 | 1.0000 | 0.031 (0.289) |
| | iso 1.0 % | PFC | 1.0000 | 0.118 (0.602) |
| | | V1 | 1.0000 | 0.018 (0.578) |

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Tab. 4.7 Correlation of $TE_{SPO}$ with source autocorrelation decay time (ACT) and source signal variance. $\widetilde{R}^2$ and $max(R^2)$ indicate the median and maximum of $R^2$ over recordings per condition, respectively.

| **animal** | **isoflurane level** | **direction** | ACT | | signal variance | |
|---|---|---|---|---|---|---|
| | | | $p$ | $\widetilde{R}^2$ $(max(R^2))$ | $p$ | $\widetilde{R}^2$ $(max(R^2))$ |
| 1 | iso 0.0 % | PFC → V1 | 1.0000 | 0.035 (0.110) | 0.9861 | 0.009 (0.060) |
| | | V1 → PFC | 0.9996 | 0.029 (0.215) | 0.6644 | 0.024 (0.089) |
| | iso 0.5 % | PFC → V1 | 1.0000 | 0.016 (0.110) | 1.0000 | 0.024 (0.033) |
| | | V1 → PFC | 0.7310 | 0.003 (0.014) | 0.0063* | 0.005 (0.011) |
| | iso 1.0 % | PFC → V1 | 1.0000 | 0.006 (0.047) | 0.5836 | 0.002 (0.037) |
| | | V1 → PFC | 0.0069* | 0.001 (0.032) | 0.0000*** | 0.005 (0.092) |
| 2 | iso 0.0 % | PFC → V1 | 1.0000 | 0.063 (0.105) | 0.8315 | 0.008 (0.025) |
| | | V1 → PFC | 0.9996 | 0.006 (0.041) | 0.0434 | 0.005 (0.031) |
| | iso 0.5 % | PFC → V1 | 1.0000 | 0.115 (0.155) | 0.0440 | 0.017 (0.103) |
| | | V1 → PFC | 1.0000 | 0.022 (0.125) | 0.9985 | 0.005 (0.086) |
| | iso 1.0 % | PFC → V1 | 0.0000** | 0.011 (0.111) | 0.0000*** | 0.012 (0.058) |
| | | V1 → PFC | 0.0076* | 0.009 (0.031) | 0.0000*** | 0.015 (0.032) |

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Tab. 4.8 Correlation of $AIS$ with band power. $\widetilde{R}^2$ and $max(R^2)$ indicate the median and maximum of $R^2$ over recordings per condition, respectively.

| **animal** | **level** | **site** | | $\delta$ | | $\theta$ | | $\alpha$ |
|---|---|---|---|---|---|---|---|---|
| | | | $p$ | $\widetilde{R}^2$ $(max(R^2))$ | $p$ | $\widetilde{R}^2$ $(max(R^2))$ | $p$ | $\widetilde{R}^2$ $(max(R^2))$ |
| 1 | iso 0.0 % | PFC | 1.0000 | 0.068 (0.418) | 0.0000*** | 0.021 (0.201) | 0.0000*** | 0.022 (0.258) |
| | | V1 | 0.0000*** | 0.041 (0.275) | 0.0003** | 0.020 (0.160) | 0.0000*** | 0.029 (0.166) |
| | iso 0.5 % | PFC | 0.0000*** | 0.019 (0.174) | 0.0000*** | 0.005 (0.040) | 0.0000*** | 0.003 (0.075) |
| | | V1 | 0.0000*** | 0.008 (0.129) | 0.0093 | 0.024 (0.060) | 0.0986 | 0.009 (0.090) |
| | iso 1.0 % | PFC | 0.0000*** | 0.006 (0.036) | 0.8710 | 0.005 (0.093) | 0.0445 | 0.003 (0.028) |
| | | V1 | 0.2751 | 0.010 (0.303) | 0.0000*** | 0.002 (0.054) | 0.0000*** | 0.002 (0.030) |
| 2 | iso 0.0 % | PFC | 0.0000*** | 0.158 (0.359) | 0.0000*** | 0.044 (0.089) | 0.0000*** | 0.025 (0.075) |
| | | V1 | 0.0000*** | 0.005 (0.303) | 0.0000*** | 0.023 (0.228) | 0.0000*** | 0.042 (0.238) |
| | iso 0.5 % | PFC | 0.0000*** | 0.019 (0.255) | 0.0000*** | 0.007 (0.018) | 0.0000*** | 0.006 (0.074) |
| | | V1 | 0.0000*** | 0.025 (0.098) | 0.6476 | 0.024 (0.068) | 1.0000 | 0.017 (0.045) |
| | iso 1.0 % | PFC | 0.0000*** | 0.005 (0.298) | 0.9729 | 0.005 (0.021) | 0.9999 | 0.004 (0.151) |
| | | V1 | 0.0063* | 0.043 (0.302) | 0.7462 | 0.001 (0.039) | 1.0000 | 0.003 (0.051) |

| **animal** | **level** | **site** | | $\beta$ | | $\gamma$ |
|---|---|---|---|---|---|---|
| | | | $p$ | $\widetilde{R}^2$ $(max(R^2))$ | $p$ | $\widetilde{R}^2$ $(max(R^2))$ |
| 1 | iso 0.0 % | PFC | 0.1179 | 0.016 (0.154) | 1.0000 | 0.036 (0.089) |
| | | V1 | 0.0001*** | 0.039 (0.163) | 1.0000 | 0.018 (0.331) |
| | iso 0.5 % | PFC | 0.0000*** | 0.003 (0.023) | 0.5147 | 0.007 (0.068) |
| | | V1 | 0.9981 | 0.025 (0.081) | 1.0000 | 0.005 (0.009) |
| | iso 1.0 % | PFC | 0.7781 | 0.001 (0.044) | 0.9392 | 0.004 (0.063) |
| | | V1 | 0.1122 | 0.003 (0.045) | 0.8076 | 0.003 (0.061) |
| 2 | iso 0.0 % | PFC | 0.0000*** | 0.005 (0.047) | 1.0000 | 0.006 (0.077) |
| | | V1 | 0.2450 | 0.015 (0.185) | 1.0000 | 0.027 (0.060) |
| | iso 0.5 % | PFC | 0.0000*** | 0.007 (0.094) | 0.0000*** | 0.014 (0.088) |
| | | V1 | 1.0000 | 0.036 (0.087) | 1.0000 | 0.010 (0.080) |
| | iso 1.0 % | PFC | 1.0000 | 0.007 (0.073) | 1.0000 | 0.008 (0.227) |
| | | V1 | 1.0000 | 0.001 (0.062) | 1.0000 | 0.017 (0.147) |

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Tab. 4.9 Correlation of $H$ with band power. $\widetilde{R}^2$ and $max(R^2)$ indicate the median and maximum of $R^2$ over recordings per condition, respectively.

| **animal** | **level** | **site** | | $\delta$ | | $\theta$ | | $\alpha$ |
|---|---|---|---|---|---|---|---|---|
| | | | $p$ | $\widetilde{R}^2$ ($max(R^2)$) | $p$ | $\widetilde{R}^2$ ($max(R^2)$) | $p$ | $\widetilde{R}^2$ ($max(R^2)$) |
| 1 | iso 0.0 % | PFC | 1.0000 | 0.095 (0.458) | 1.0000 | 0.015 (0.201) | 1.0000 | 0.020 (0.271) |
| | | V1 | 1.0000 | 0.091 (0.706) | 0.9998 | 0.037 (0.109) | 1.0000 | 0.041 (0.097) |
| | iso 0.5 % | PFC | 1.0000 | 0.046 (0.202) | 1.0000 | 0.004 (0.036) | 1.0000 | 0.003 (0.059) |
| | | V1 | 1.0000 | 0.035 (0.133) | 0.9100 | 0.040 (0.051) | 0.8029 | 0.002 (0.086) |
| | iso 1.0 % | PFC | 1.0000 | 0.020 (0.079) | 0.0359 | 0.005 (0.082) | 0.9198 | 0.004 (0.023) |
| | | V1 | 1.0000 | 0.010 (0.461) | 0.9941 | 0.003 (0.061) | 0.9991 | 0.004 (0.032) |
| 2 | iso 0.0 % | PFC | 1.0000 | 0.171 (0.395) | 1.0000 | 0.058 (0.110) | 1.0000 | 0.024 (0.060) |
| | | V1 | 1.0000 | 0.031 (0.723) | 1.0000 | 0.036 (0.135) | 1.0000 | 0.048 (0.170) |
| | iso 0.5 % | PFC | 1.0000 | 0.035 (0.298) | 1.0000 | 0.011 (0.016) | 1.0000 | 0.015 (0.088) |
| | | V1 | 1.0000 | 0.038 (0.135) | 0.0428 | 0.021 (0.073) | $0.0000^{***}$ | 0.016 (0.045) |
| | iso 1.0 % | PFC | 1.0000 | 0.015 (0.580) | $0.0000^{***}$ | 0.006 (0.076) | $0.0000^{***}$ | 0.008 (0.219) |
| | | V1 | 1.0000 | 0.112 (0.601) | 0.1329 | 0.003 (0.062) | $0.0000^{***}$ | 0.009 (0.094) |

| **animal** | **level** | **site** | | $\beta$ | | $\gamma$ |
|---|---|---|---|---|---|---|
| | | | $p$ | $\widetilde{R}^2$ ($max(R^2)$) | $p$ | $\widetilde{R}^2$ ($max(R^2)$) |
| 1 | iso 0.0 % | PFC | 0.9665 | 0.018 (0.165) | $0.0000^{***}$ | 0.052 (0.097) |
| | | V1 | 0.9999 | 0.046 (0.121) | $0.0000^{***}$ | 0.014 (0.358) |
| | iso 0.5 % | PFC | 1.0000 | 0.007 (0.020) | 0.5510 | 0.013 (0.064) |
| | | V1 | $0.0000^{***}$ | 0.028 (0.089) | $0.0000^{***}$ | 0.006 (0.011) |
| | iso 1.0 % | PFC | 0.2144 | 0.005 (0.056) | 0.0269 | 0.006 (0.082) |
| | | V1 | 0.4394 | 0.002 (0.040) | 0.2250 | 0.003 (0.053) |
| 2 | iso 0.0 % | PFC | 1.0000 | 0.011 (0.043) | $0.0000^{***}$ | 0.001 (0.068) |
| | | V1 | 0.5631 | 0.022 (0.094) | $0.0000^{***}$ | 0.037 (0.061) |
| | iso 0.5 % | PFC | 1.0000 | 0.013 (0.107) | 1.0000 | 0.012 (0.095) |
| | | V1 | $0.0000^{***}$ | 0.034 (0.096) | $0.0000^{***}$ | 0.010 (0.102) |
| | iso 1.0 % | PFC | $0.0000^{***}$ | 0.007 (0.051) | $0.0000^{***}$ | 0.012 (0.290) |
| | | V1 | $0.0017^{*}$ | 0.011 (0.072) | $0.0000^{***}$ | 0.013 (0.103) |

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

Tab. 4.10 Correlation of $TE_{SPO}$ with band power. $\widetilde{R}^2$ and $max(R^2)$ indicate the median and maximum of $R^2$ over recordings per condition, respectively.

| **animal** | **level** | **direction** |  | $\delta$ |  | $\theta$ |  | $\alpha$ |
|---|---|---|---|---|---|---|---|---|
|  |  |  | $p$ | $\widetilde{R}^2$ $(max(R^2))$ | $p$ | $\widetilde{R}^2$ $(max(R^2))$ | $p$ | $\widetilde{R}^2$ $(max(R^2))$ |
| 1 | iso 0.0 % | PFC → V1 | 1.0000 | 0.024 (0.101) | 1.0000 | 0.034 (0.122) | 1.0000 | 0.061 (0.146) |
|  |  | V1 → PFC | 0.9662 | 0.030 (0.230) | 0.8389 | 0.009 (0.045) | 1.0000 | 0.008 (0.047) |
|  | iso 0.5 % | PFC → V1 | 1.0000 | 0.033 (0.090) | 1.0000 | 0.073 (0.132) | 1.0000 | 0.013 (0.117) |
|  |  | V1 → PFC | 0.1319 | 0.001 (0.010) | 1.0000 | 0.018 (0.058) | 0.9097 | 0.001 (0.015) |
|  | iso 1.0 % | PFC → V1 | 1.0000 | 0.009 (0.058) | 1.0000 | 0.009 (0.029) | 1.0000 | 0.012 (0.038) |
|  |  | V1 → PFC | 0.0947 | 0.002 (0.024) | 1.0000 | 0.003 (0.116) | 0.9996 | 0.004 (0.058) |
| 2 | iso 0.0 % | PFC → V1 | 1.0000 | 0.060 (0.102) | 1.0000 | 0.055 (0.121) | 1.0000 | 0.036 (0.052) |
|  |  | V1 → PFC | 0.9998 | 0.006 (0.038) | 1.0000 | 0.012 (0.056) | 0.9998 | 0.010 (0.040) |
|  | iso 0.5 % | PFC → V1 | 1.0000 | 0.108 (0.157) | 1.0000 | 0.115 (0.166) | 1.0000 | 0.079 (0.160) |
|  |  | V1 → PFC | 1.0000 | 0.026 (0.149) | 1.0000 | 0.015 (0.052) | 0.6675 | 0.003 (0.009) |
|  | iso 1.0 % | PFC → V1 | 0.0000$^{***}$ | 0.009 (0.095) | 1.0000 | 0.010 (0.037) | 1.0000 | 0.007 (0.030) |
|  |  | V1 → PFC | 0.0153 | 0.008 (0.023) | 1.0000 | 0.006 (0.035) | 0.7369 | 0.001 (0.017) |

| **animal** | **level** | **site** |  | $\beta$ |  | $\gamma$ |
|---|---|---|---|---|---|---|
|  |  |  | $p$ | $\widetilde{R}^2$ $(max(R^2))$ | $p$ | $\widetilde{R}^2$ $(max(R^2))$ |
| 1 | iso 0.0 % | PFC → V1 | 1.0000 | 0.069 (0.157) | 0.0352 | 0.005 (0.044) |
|  |  | V1 → PFC | 0.9612 | 0.010 (0.040) | 0.0009$^{**}$ | 0.010 (0.028) |
|  | iso 0.5 % | PFC → V1 | 1.0000 | 0.033 (0.121) | 0.0611 | 0.031 (0.038) |
|  |  | V1 → PFC | 0.9991 | 0.010 (0.039) | 0.0014$^{**}$ | 0.021 (0.028) |
|  | iso 1.0 % | PFC → V1 | 0.0000$^{***}$ | 0.004 (0.019) | 0.0000$^{***}$ | 0.007 (0.054) |
|  |  | V1 → PFC | 0.0912 | 0.002 (0.063) | 0.0000$^{***}$ | 0.006 (0.109) |
| 2 | iso 0.0 % | PFC → V1 | 1.0000 | 0.021 (0.040) | 0.3062 | 0.015 (0.036) |
|  |  | V1 → PFC | 0.9998 | 0.002 (0.039) | 0.2046 | 0.002 (0.041) |
|  | iso 0.5 % | PFC → V1 | 1.0000 | 0.107 (0.132) | 0.6054 | 0.006 (0.013) |
|  |  | V1 → PFC | 0.0186 | 0.004 (0.019) | 0.0000$^{***}$ | 0.022 (0.041) |
|  | iso 1.0 % | PFC → V1 | 0.9940 | 0.004 (0.039) | 0.0025$^{*}$ | 0.003 (0.027) |
|  |  | V1 → PFC | 0.9122 | 0.002 (0.013) | 0.0000$^{***}$ | 0.002 (0.058) |

$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

**S3 The effect of sampling on AIS-estimation** **AIS estimates for data sampled at different rates.**

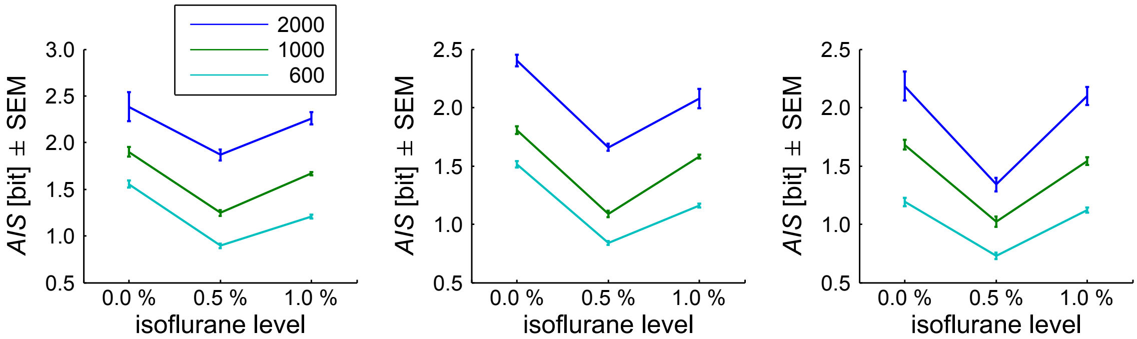


Fig. 4.11 **Estimates of active information storage (AIS) from data sampled at different rates.** AIS estimates from three random recordings in animal 1 under three levels of Isoflurane; estimated from data with the sampling rate used for analysis in the present work (1000 Hz) and re-sampled at 2000 and 600 Hz respectively; note that qualitative results did not change due to re-sampling, but absolute estimates increased for higher sampling rates; the number of data points was held approximately constant by selecting a subset of trials for estimation such that the number of points entering the analysis was equal to the smallest number of points over all isoflurane levels.

**S4 Theoretical and practical running time of nearest-neighbor based estimators** The KSG-estimator used in this work is computationally demanding, because as a nearest-neighbor based estimator it requires the execution of $k$-nearest-neighbor as well as range searches for all data points $N$. Hence, the algorithms used to perform these searches determine the asymptotic time complexity of the estimator. In this work, we used two different implementations published as part of the TRENTOOL toolbox: a CPU-based estimator for the estimation of $TE_{SPO}$ from epochs or trials of data [86], and a GPU-based estimator for the estimation of $TE_{SPO}$ from data pooled over epochs or trials [208]. We used the CPU-based work-flow for the epoch-wise estimation of $TE_{SPO}$ and the GPU-based work-flow when testing for statistical significance of $TE_{SPO}$ to maximize the number of points entering the estimation. The CPU-based estimator uses a neighbor-search algorithm that—depending on data-properties—requires for neighbor-searches on all $N$ points at maximum $O(kN\log(N))$ time. The GPU-based estimator does not make use of fast data structures used for the CPU-based estimator, but uses a linear search which results in a worse time complexity of $O(dN^2)$. However, in the implementation used in this study [208] the linear neighbor search is performed in parallel over points and problem instances, which significantly improves the overall running time when using the estimator on multiple problem instances—a comparison of the running times of the serial CPU-algorithm and the parallel GPU-algorithm can be found in [208].

Additionally, we used the KL-estimator for entropy estimation. This estimator requires the execution of a $k$-nearest-neighbor search for all data points $N$—we used the estimator's implementation published as part of the JIDT toolbox [87], which has a time complexity of $O(kN\log(N))$.

We also measured practical running times of the estimators to provide a point of reference when planning similar analyses. Approximate, average running times for the estimation of $AIS$, $TE_{SPO}$, and $H$ are presented in Table 4.11. The presented running times include the estimation of each measure for both directions of interactions or recording site, for one recording; presented running times are averages over recordings. We measured the total time needed for estimation, including data preparation (e.g., the optimization of embedding parameters for the estimation of $TE_{SPO}$). All estimation procedures that did not require a GPU, were executed on a Intel(R) Xeon(R) CPU clocked at 2.90 GHz. The GPU-implementation of the $TE_{SPO}$ estimator was run on a Intel(R) Xeon(R) CPU clocked at 2.00 GHz and a NVIDIA GeForce GTX TITAN. Both machines were running 64-bit Ubuntu Linux. Note that the GPU estimation was performed on 50 trials only, because here the computational demand was higher due to more data points entering one estimation of $TE_{SPO}$ (pooled over epochs).

Tab. 4.11 Practical average running times for estimation of information-theoretic measures from one recording session and two recording sites or directions of interaction in animal 1.

| measure | toolbox/implementation | mean running time [min] |
|---|---|---:|
| $TE_{SPO}$ | TRENTOOL/GPU-implementation | 2235.93 (314.27 SD) |
| $TE_{SPO}$ | TRENTOOL/CPU-implementation | 1871.51 (926.56 SD) |
| $AIS$ | TRENTOOL/CPU-implementation | 2.83 (1.51 SD) |
| $H$ | JIDT/CPU-implementation | 1.55 (0.56 SD) |

SD = standard deviation

# 5 General discussion

The present work introduced three studies on the application of TE in neuroscience. In the first two studies, we presented two improvements for the estimation of TE, which solved two of the most pressing problems in the estimation of bivariate TE in neuroscience: first, we presented an approach to estimate TE from non-stationary data, which is commonly encountered in neuroscience experiments; second, we presented an algorithmic correction for multivariate effects when estimating bivariate TE in a multivariate setting. In the third study, we discussed current best-practice in the estimation of TE and its interpretation, also providing an example of how the misinterpretation of TE as a causality measure may lead to erroneous explanations of neural phenomena.

In the first section of the general discussion (5.1, *Application of transfer entropy in neuroscience*), I will discuss the current state of TE as a measure of information transfer in neuroscience research, in particular in the light of the improved estimation techniques presented in Chapters 2 and 3. In the second section (5.2, *Future directions*), I will present possible directions for future research. In the third section (5.3, *Application of the proposed methods*), I will present applications of the proposed methods in neuroscience and briefly discuss applications of TE outside neuroscience. I will close the general discussion with an outlook (5.4, *Conclusion and outlook*).

## 5.1 Application of transfer entropy in neuroscience – review of the current status

Since its first formulation by Shannon [10], information theory has been adopted as a tool for data analysis in a variety of disciplines. Also neuroscience has used information theory to perform a variety of tasks, one of them being the quantification of information transfer between two sources of neural activity by estimating TE [257]. Today, TE has become an important analysis method in a variety of neuroscience sub-fields: it is frequently applied to investigate the neurophysiological correlates of different stages of consciousness [20, 258] and anesthesia [31, 221, 222, 237, 259]; it is used to investigate oscillatory activity in MEG data [22], brain-organ interactions [19], auditory perception [15], selective attention [260], and auditory short-term memory [14].

Yet, the transfer of information theoretic measures from man-made communication systems to biological, especially neural systems, is not trivial resulting in ongoing debates about the correct application of TE [75, 261] (see also Chapter 4, *The relation of local entropy and information transfer suggests an origin of isoflurane anesthesia effects in local information processing*), its interpretation [68, 262], and whether it is justifiable to use information theory at all when investigating information processing in neural systems [263]. In the following, I will review key problems in the application of TE in neuroscience research, discuss existing and potential solutions, and how the present work contributes to solving these problems: I will first discuss conceptual problems when applying TE to neuroscience data (Section 5.1.1), before I go on to discuss more practical problems when applying TE (Section 5.1.2).

### 5.1.1 Operationalizing neural information transfer as transfer entropy—conceptual considerations

**Existence of a channel** In its original form, Shannon's information theory is concerned with quantifying information and its communication from a sender to a receiver over some channel [10]. However, many of the more recent practical applications of information theory do not concern man-made systems, hence it is not immediately clear if a channel between assumed sender and a receiver exists (here, "channel" means any form of physical mechanism that enables the transfer of information). If no channel exists, TE or other information-theoretic dependency measures may still be non-zero for reasons discussed below. In these scenarios, estimated TE has to be considered spurious. Consequently, estimated TE may not reflect true information transfer and should thus in general not be used to infer the existence of an underlying channel.

In general, TE may be detected between any two correlated source, irrespective of whether the correlation is caused by an actual information transfer or by effects third processes have on the two sources in question (e.g., a common driving input). The detection of spurious TE due to the effect of third sources may only be prevented by accounting for all relevant processes in the system (see Chapter 3, *A Graph Algorithmic Approach to Separate Direct from Indirect Neural Interactions*). Yet, accounting for all sources requires the estimation of TE in a multivariate fashion, which is—when done exhaustively—a NP-hard problem—thus, estimating fully multivariate TE for arbitrary input sizes is not feasible if $\mathsf{P} \neq \mathsf{NP}$. Consequently, we can only use approximations when estimating multivariate TE and thus minimize the risk of detecting spurious TE, as presented in Chapter 3.

Thus, in practice, the erroneous detection of spurious, bivariate TE due to correlated sources can not be avoided with absolute certainty in multivariate settings—accordingly, we can not infer the existence of a channel from non-zero TE. More generally speaking, estimating bivariate TE does not allow for the inference of the causal structure underlying the transfer of information between multiple sources (see also next section). Instead, additional methods, targeted directly at the investigation of causal structure, like tract tracing or dynamic causal modeling (DCM) [214], should be used. Gathering additional *causal* evidence, in turn, provides additional evidence for information transfer, because it infers the existence of an underlying channel. This approach is presented in detail in Section 5.3.2, *Transfer entropy estimation as preprocessing step in DCM analysis [264]*, below.

A further potential cause for spurious TE are estimation problems due to limited data (this problem has practical reasons which will be discussed in more detail in the next section). In short, estimators of TE or MI come with a finite sample bias (see Section 1.4, *Open problems in estimating information processing measures in neuroscience*), hence, non-zero estimates may be due to the bias rather than information transfer. This problem is solved by using the estimated quantity as a statistic that is tested against a suitable surrogate distribution to infer the statistical significance of the estimate.

In summary, non-zero estimates of information-theoretic dependency measures do occur in the absence of a channel or other means of communication—like for all correlative measures, it holds for information theoretic quantities that a correlation does not indicate causation. Thus, in general we can not infer the existence of a channel—the causal mechanism enabling information transfer—from non-zero estimates of bivariate TE. Accordingly, we also have to be careful when interpreting estimated TE as information transfer if no additional knowledge about an underlying channel exists (e.g., from anatomical studies), or if the potential error due to multivariate effects has not been minimized.

**Transfer entropy is not a measure of causality** TE quantifies how much information we can gain about the next state of a target process, if we not only look at the target process's past alone, but also at the past of a second source process. This definition of TE implements a notion of "causality" introduced by Norbert Wiener [265]. Wiener was the first to formally describe a measure of causality for experimental observations, also termed "observational causality" [75, 266]; his definition required, first, temporal precedence of cause over effect and, second, an increase in the predictability of the effect given the observation of the cause. Wiener's abstract idea was later realized by Granger in the form of a statistical, autoregressive model, the so-called *Granger causality* [267]. Granger causality has been shown to be equivalent to TE for jointly Gaussian variables [103]—hence, both measures are implementations of observational causality as defined by Wiener.

Accordingly, both TE and Granger causality have been used as measures of "causality" in the past (see for example [89–91, 268, 269]). Yet, in newer research, competing definitions of causality have been formulated. These definitions replaced Wiener's observational causality by definitions and measures that rely on physical interventions, i.e., the mechanistic *manipulation* of causes [118, 270]. Following this definition, Ay and Polani [118] presented a measure of *causal information flow*, which relies on the evaluation on so-called *interventional conditional probabilities*, $p(a|\hat{s})$, where the $\hat{\cdot}$-operator indicates the *imposing* of the value of $s$. In the light of this newer definitions, TE (and Granger causality) can no longer be considered measures of "causality", because they do not rely on physical intervention [68, 92, 271]: While TE quantifies the amount of information "being transferred into the computation taking place at the destination" [68] and thus quantifies *predictive information transfer* [68], causal measures like causal information flow [118] aim at inferring the physical structure underlying these computations. In the terms of Marr [40], TE is a measure of computations performed on the algorithmic level, while causal measures are measures of the physical structure on the implementational level.

Differentiating between these two measures and levels of analysis is highly beneficial for neuroscience research, because it allows for the analysis of two distinct phenomena: first, physical structure (implementational level), and second, computational tasks performed on this structure (algorithmic level). Mixing these two levels when analyzing or interpreting neuroscience data by trying to answer questions on one level using evidence collected on another level, may lead to erroneous results, because knowing a system's physical structure does not reveal which type of computational tasks is being performed on it [68, 272]. Rather, one physical structure may serve different computational tasks over time. This has already been stated by Marr [40], who observed that levels hardly constrain each other (see Chapter 1, *Introduction*). Accordingly, transferring findings on one level of analysis to answer questions on another level of analysis may remain speculative and may lead to erroneous conclusions.

As an example for such erroneous conclusions, one may consider an axonal connection between two neurons, over which spikes are conveyed. It may seem likely that the connection serves the transfer of information, i.e., without making the different levels of explanation explicit, it may be self-evident that neurons traveling along the axon (the implementational level) serve the transfer of information (the algorithmic level). Yet, this conclusion may be false—for example, if the two neurons are part of a recurrently connected ensemble of neurons, which together serve the computational task of storing information. Thus, whether spikes traveling along an axon serve information transfer or storage is not answered by knowing the causal structure alone (i.e., the axonal connection and spikes traveling along this connection). Rather, the computational task being performed on this structure has to be investigated separately. One approach to accomplish this is by estimating TE between the two neurons and estimating AIS within the two neurons. If the connection serves information storage, TE will be zero for a sufficient embedding because information travels recurrently through the two neurons (the next state in the target neuron will be perfectly predictable from its own past, so observing the source neuron does not add to the prediction and no new information is transferred from the source to the target); furthermore, AIS will be high, because given a proper embedding, information in one neuron is highly predictable from its own past. In sum, the transfer of information requires causal interactions [68], yet, the existence of a causal interaction does not imply information transfer.

An example for an artificial neural network where connections serve information storage, rather than transfer has been presented in [273]. In the present work, we further demonstrated the need to differentiate between physical structure and computations in the analysis of real-world data in Chapter 4, *The relation of local entropy and information transfer suggests an origin of isoflurane anesthesia effects in local information processing*: we showed that a measured reduction in TE under anesthesia may *not* be caused by changes in the underlying physical structure, but in changes in the information processing performed *on* this physical structure. The reduction in TE was often explained by changes in underlying coupling in existing anesthesia research, thus mixing the two explanatory levels of computation and implementation. However, by distinguishing these levels and by explicitly analyzing the information processing performed, potential alternative explanations for the loss of consciousness under anesthesia could be generated.

In summary, in neuroscience, it is fruitful to adopt a definition of causality that is based on physical interventions and to distinguish measures of causality from measures of computation. When adopting this distinction, TE is a measure of information processing and thus a measures of computations. Furthermore—as discussed in the last subsection—non-causal or correlative measures of TE are not suitable to infer the existence of a channel underlying information transfer, i.e., they are not able to infer the causal structure enabling the transfer of information. Hence, the independent investigation of the implementational and algorithmic level provides complementary insights into information processing systems and is crucial to their understanding.

**Investigating neural computation with transfer entropy** In the last section, I motivated an independent analysis of computations, i.e., the algorithmic level, in neural systems. Yet, even though the notion of computation has been widely adapted for neural systems in the sense that these systems represent and process information, and that this information processing generates the observable function or behavior of the system [1, 2, 4, 6], a clear definition of *computation* and its explicit analysis is often missing [1, 2, 4]. Instead, neuroscience typically analyzes the implementational and the functional level of neural systems, aiming to link the findings on both levels—these links are then often formulated in terms of "neural computations" or "algorithms" that enable function on the basis of the implementation [2, 44]. Yet, as already noted by Marr, these levels of explanation pose only little constraints on each other, such that a transfer of knowledge from one level to the other remains mostly speculative [40]. Hence, a novel approach to the explicit investigation of neural computations is needed to understand how neural dynamics give rise to function. Such an approach requires, first, a definition of computation applicable to neural systems, and second, methods for the quantitative analysis of computation in neural data.

As briefly described in the introduction, a definition of computations performed by neural systems—and biological systems in general—has been provided by Melanie Mitchell [6], who coined the term *biological computation*. Biological computation is described as "massively parallel, stochastic, inexact, and on-going, with no clean notion of a mapping between 'inputs' and 'outputs'." [6]. This distributed nature is the defining property of neural computation, where single agents or arbitrary subsystems perform computations independently and in parallel [7]. The computations performed by single agents may—opposed to computation performed by traditional computing systems—not be understood in terms of the two most common notions of computation: (1) the solution of a specific task, where an input is transformed into an output and both states have a clear mapping to some human-understandable problem and its solution (e.g., visual input and the recognition of a face from this input); (2) universal computation, a system that, depending on the input, is capable of computing any function that is computable by a Turing machine. Mitchell terms this computation performed by a single agent "intrinsic computation" [7]. An alternative way to describe or analyze intrinsic computation may be to decompose it into so-called "generic structural elements", i.e., the basic information-processing operations of information storage, transfer, and modification [6]. Decomposing computation in such a fashion has already been proposed by Alan Turing (see Langton [5]).

Following the proposal to investigate information processing by quantifying its basic operations, it has been proposed that information theory provides natural quantitative measures of these operations [9]. Information theory provides a mathematically rigorous description of information and its processing in a semantics-free fashion. Hence information-theoretic measures do not require knowledge about the *meaning* of the information being transferred or stored, to quantify the amount of information being processed. This property is especially desirable when trying to investigate information processing in distributed systems, because it allows to analyze intrinsic computation, which may not be describable in terms of function or semantics. The first comprehensive framework for quantitatively measuring all three information-processing operations named by Mitchell [6] has been proposed in the form of the information-theoretic framework of local information dynamics [9].

Information theory has been popular in neuroscience since its formulation by Shannon. Attempts to quantify information processing date back to the first application of information theory in neuroscience by Attneave [53] and Barlow [54], who investigated the encoding of given stimuli in neural responses of single cells or cell populations (see for example [55–59] for reviews). Further applications quantified the entropy of neural signals as an upper limit to the information about a stimulus that can be transferred by a neural signal (e.g. [274]). These early applications of information theory are however limited to the analysis of neural representations of external stimuli and—in terms of Marr's framework—tried to link the implementational level (spike trains) to the functional level (representing objects in the external world). Only newer applications extended the application of information theory beyond the analysis of neural representations and began to investigate information processing in neural populations more directly, for example, by deriving dependency-measures between activity in individual neural sub-systems [13, 23, 60]. A further prominent example for the application of information theory to investigate neural computations is the formulation of the *Infomax principle* [61–63, 275]. Infomax proposes a generic algorithm for cortical information processing; it states that the brain aims at maximizing the MI between its sensory input and internal representation, while maximizing the efficiency of this representation (minimizing redundancy). Here, information theory is used to formulate an algorithmic description of neural activity.

In contrast to these earlier applications of information theory in neuroscience, the framework of local information dynamics [9] presents a comprehensive approach to the quantification of *generic* information processing operations. For example, in comparison to the Infomax principle, the framework allows to quantitatively *describe* information processing, while, the Infomax principle *explains* neural function using information theory, i.e., in the latter application information theory is used as a "normative theory" [276]. The framework of local information dynamics in contrast can thus be used to define how *any* neural computation should be reflected by its information-processing profile (e.g., is information storage in a sub-system expected to be high, given a certain algorithm); this in turn allows to formulate testable hypotheses about the dynamics of information processing measures. Hence, local information dynamics allows to formulate constraints under which an arbitrary algorithm is performed by a system—this approach has successfully been applied to artificial systems like cellular automata [11] or boolean networks [277]. It has since been proposed to transfer this approach to neural systems [8].

We provided a first demonstration of how this approach can be applied to neural systems in a recent study [278], where we applied the measures proposed by Lizier [9] to investigate neural computations, following the theoretical considerations of Mitchell [6] and Wibral et al. [8]. We used local versions of AIS and TE to investigate the "algorithm" governing the information processing at the retinogeniculate synapse of the cat. We tested two alternative algorithms by formulating hypotheses about their respective information processing profiles.

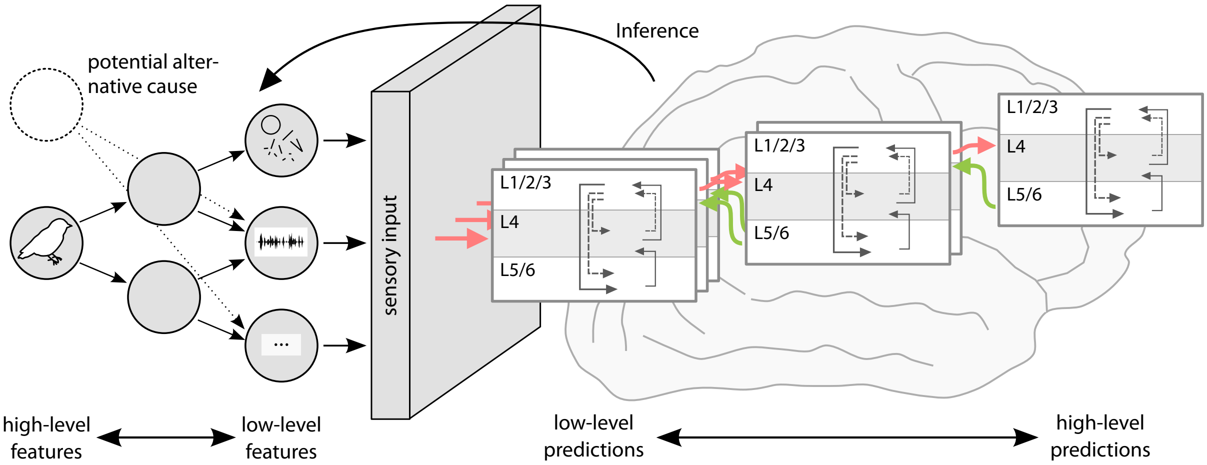


Fig. 5.1 **Predictive coding theory.** Schematic representation of predictive coding theory (PCT, see also main text): The brain tries to infer the high-level causes of its sensory input in the outside world (gray circles). It accomplishes this by building an internal, hierarchical model (white boxes), from which—on each level—it tries to predict incoming input (see also [236]). The model is refined by matching actual input (red arrows) against predictions (green arrows). Within this hierarchical model, lower areas generate predictions about lower level input, while higher levels generate predictions about higher-level input.

The assumed algorithms were derived from predictive coding theory (PCT, Fig. 5.1), which is arguably, the most comprehensive theory on brain function today [3, 235, 279]. PCT tries to explain how the brain solves the problem of representing the external causes of its sensory input. The difficulty of this task lies in the fact that the brain never has direct access to these causes, but only to their effects—namely, the neural signals generated at the sensory organs. PCT proposes that the brain—instead of trying to *reconstruct* the causes of the sensory signals—tries to *predict* signals generated in the future. The basis of this theory is that natural input exhibits statistical regularities, which are reflected in the generated sensory signals. These regularities are learned by the brain to build an internal model of the world. From this model, the brain generates predictions about future input. Predicted and actual future input are then compared to further refine the internal model with the goal of minimizing the error in the prediction. For this abstract idea of error minimization as a computational goal of brain function, several concrete realizations have been proposed. These realizations differ in how the brain updates its internal model when comparing prediction and actual future input (Fig. 5.2B): either, on the basis of the *mismatching* features (e.g., [280–282]), or on the basis of the *matching* features (e.g., [283–285]). Both realizations have been shown to be equivalent in terms of the computational task they solve—the minimization in the error between prediction and input [286–289]. The two realizations only differ in the signal that is propagated up the cortical hierarchy: the signal either represents predicted or unpredicted features of the input. Hence, even though evidence for predictive coding has been found on the behavioral (e.g., [282]) and the implementational level (e.g., [236]), none of these findings could resolve the conflict in theories of how this mechanism is realized algorithmically. In our study, we aimed at solving this conflict by directly investigating the algorithmic level by quantifying the information processing at the retinogeniculate synapse of the cat.

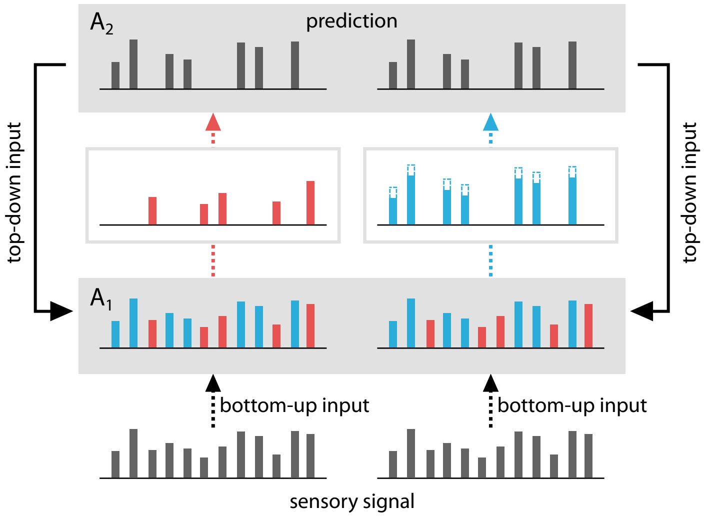


Fig. 5.2 **Realization of predictive coding theory in cortex (adapted from [290]).** Possible algorithms realizing PCT in the cortex: Grey boxes denote cortical areas $A_1$ and $A_2$. $A_1$ receives bottom-up input from a lower processing stage in the hierarchy and top-down input from $A_2$, which is higher in the processing hierarchy. $A_1$ compares predictions coming from $A_1$ against actual input. Two theories describe how this comparison is used to refine the model and predictions in area $A_2$: propagation of the prediction error, where non-matching input is communicated up the hierarchy (red bars and arrows); and propagation of the predicted input, where matching input is enhanced and communicated up the hierarchy (blue bars and arrows).

To test the two potential algorithms against each other, we formulated hypotheses about the information processing resulting from either of the two algorithms [278]: if predictable features of the synapse's input were predominantly transferred up the cortical hierarchy, transfer should be high when predictability was high; if non-predicable features of the input were predominantly transferred, transfer should be high when predictability was low. We measured transfer and predictability with local versions of TE and AIS, respectively. We hypothesized that by evaluating the correlation between local AIS and TE, the two algorithms should be distinguishable. We found a positive correlation in all investigated cell pairs, indicating that transfer was higher when predictability was high—hence favoring a transfer of predictable input. We thus demonstrated how information-theoretic measures of information processing can be used to directly investigate computations in neural systems and to infer constraints on the algorithms being performed.

Formulating testable hypotheses about the neural computations performed is a promising alternative to existing approaches to the investigation of neural computations in the context of PCT. These existing approaches were criticized by De-Wit et al. [291]: the authors discuss a study by Alink et al. [292], who aimed to demonstrate the transfer of mismatching over the transfer of matching input by investigating neural correlates of violated predictions using functional magnetic resonance imaging (fMRI). The authors tried to generate a mismatch between assumed prediction and actual input, and analyzed the neural dynamics accompanying the thus generated "prediction error"—yet, as De-Wit et al. [291] point out, the interpretation of neural activity as an error or confirmed prediction is, in principle, a circular argument, because the interpretation depends on the a-priori point of view, about what the brain should predict, which is then experimentally "validated". Whether the observed neural activity actually represents a mismatch or match is not clear from the recorded data or quantities computed from it [291]. In their study, Alink et al. [292] thus attempted to interpret findings on the implementational level in terms of their functional meaning [291]. This approach to the interpretation of findings on the implementational level is common in neuroscience (e.g., our own study, [238]), however, it contradicts the theoretical considerations described earlier: first, computations performed by neural systems, especially arbitrary sub-systems, may not be describable in terms of a human-understandable computation [4, 6, 7]; second, phenomena on different levels of explanation may hardly constraint each other, such that findings on one level do not increase our understanding of another level [40, 44]. Here, the direct investigation of neural computations, demonstrated in [278], may be a promising alternative.

In summary, the application of information-theoretic measures to formulate and test hypotheses on information processing necessary to perform assumed computations in neural systems is a promising approach to investigate the algorithmic level of neural systems. Information theory as a tool to investigate generic computational operations may provide ways to directly investigate computations in neural systems, an approach that has been claimed to be missing in neuroscience research [2, 4]. In particular, the definition of biological computation by Mitchell [6], together with the framework of local information dynamics by Lizier [9] provide a comprehensive approach to the analysis of neural computation, which has been applied to a variety of systems, i.a., neural systems [278]. Future research should aim at extending the application of information theory and local information dynamics in neuroscience to gain a broader understanding of neural computations.

**Utilization of transferred information** When quantifying information transfer in a neural system by estimating TE, we typically make the implicit assumption that the receiver of the information (e.g., a single neuron) makes use of the whole amount of estimated information transfer; in other words, we assume that the numerical value of information transfer we estimate is actually what the receiver utilizes for its local computation. Yet, this is not necessarily the case in neural systems, especially under experimental conditions: here, the probability density functions estimated from data collected during an experiment may differ from the *natural* probability density functions that the receiver learned under natural conditions over its lifetime [293]. Accordingly, TE calculated from distributions estimated during an experiment may over- or underestimate the information truly *usable* by the receiver.

To illustrate this, assume that we quantify information transferred by spike trains traveling from a source neuron $X$ to a target neuron $Y$. The transfer of information is quantified by estimating TE as $I(Y_t; \mathbf{X}^l_{t-u} | \mathbf{Y}^k_{t-1})$ (Eq. 1.9), where information is transfered from a realization $\mathbf{x}^l_{t-u}$ to a realization $y_t$, if $p(y_n | \mathbf{y}^k_{t-1}, \mathbf{x}^l_{t-u}) > p(y_n | \mathbf{y}^k_{t-1})$, i.e., information is transferred if it is more likely to observe outcome $y_n$ after observing outcomes $\mathbf{x}^l_{t-u}$, $\mathbf{y}^k_{t-1}$ together, than observing $y_n$ after observing $\mathbf{y}^k_{t-1}$ alone. The amount of information transferred depends on the fraction $p(y_n | \mathbf{y}^k_{t-1}, \mathbf{x}^l_{t-u}) / p(y_n | \mathbf{y}^k_{t-1})$ and its weighting by the joint probability $p(y_n, \mathbf{x}_{n-u}, \mathbf{y}_{n-1})$. Hence, the amount of information transferred depends on some *reference distribution* $p(\cdot)$ from which the conditional probabilities and the weighting factor are derived.

The reference distribution is encoded in the target neuron(s) [293] (see [294] and [295] for potential encoding mechanisms for single neurons and neural populations, respectively) and is typically unknown when estimating TE from experimental data. Hence, probabilities entering TE estimation have to be estimated from experimental observations to obtain $\hat{p}\left(y_n, \mathbf{x}_{n-u}, \mathbf{y}_{n-1}\right)$, $\hat{p}(y_n | \mathbf{y}^k_{t-1}, \mathbf{x}^l_{t-u})$, and $\hat{p}(y_n | \mathbf{y}^k_{t-1})$. Accordingly, only states occurring under experimental stimulation and their observed frequencies enter the estimated distributions. The estimated distributions may thus differ from the Neuron's true or "natural" reference distribution, $p_{nat}(\cdot)$, formed by biological learning from natural inputs over the neuron's lifespan [293]. As a consequence, the estimated distributions, $\hat{p}(\cdot)$, may differ substantially from the true distributions under $p_{nat}(\cdot)$, leading to differences in the amount of information transfer estimated and the amount of information transfer actually "seen" by the target neuron.

To solve this problem, the distributions used to evaluate TE may be estimated from samples collected under natural stimulation of the neuron to obtain an estimate of the natural reference distribution, $\hat{p}_{nat}(\cdot)$, yielding the so-called *neurally accessible* information transfer [293]. $\hat{p}_{nat}(\cdot)$ may then be used to evaluate the data collected during an experiment by using $\hat{p}_{nat}(\cdot)$ to evaluate the probability of individual experimental outcomes weighted by their probability to occur during the experiment, yielding the *neurally accessible TE*

$$
\begin{aligned}
&TE^{neuro}_{SPO}\left(X \to Y, n, u, k, l\right) \\
&\quad = \lim_{k,l\to\infty} \sum_{\substack{y_n \in \mathcal{A}_{Y_n}, \mathbf{x}_{n-u} \in \mathcal{A}_{\mathbf{X}_{n-u}}, \\ \mathbf{y}_{n-1} \in \mathcal{A}_{\mathbf{Y}_{n-1}}}} \hat{p}\left(y_n, \mathbf{x}_{n-u}, \mathbf{y}_{n-1}\right) \log_2 \frac{\hat{p}_{nat}(y_n|\mathbf{y}_{n-1}, \mathbf{x}_{n-u})}{\hat{p}_{nat}(y_n|\mathbf{y}_{n-1})}. \quad (5.1)
\end{aligned}
$$

The necessity to quantify neurally accessible information has also been recently formulated by De-Wit et al. [4]: the authors criticize that often, neuroscience research focuses on how neural activity can be interpreted by the experimenter (termed the "experimenter-as-receiver" perspective), instead of focusing on how information encoded in this activity is used by computations in other cortical areas (termed "cortex-as-receiver" perspective). The latter perspective may be more relevant to the understanding of neural computations. Measuring neurally accessible TE implements this perspective, because it aims at quantifying the amount of information other cortical areas can utilize instead of quantifying information transfer with respect to the experimental condition alone.

Yet, also the estimation of TE from purely experimental data is legitimate as an approximation of the neurally accessible TE [4, 293], while its estimation may be more feasible in practice. How well the experimental TE approximates the neural accessible TE depends on the ecological validity of the experiment—i.e., whether the frequencies of neural states occurring during the experiment mirror the frequencies of their occurrence under natural stimulation. In sum, estimating TE from experimental data alone is justified, yet, using approaches like the one proposed by Wibral et al. [293] may provide additional insight into neural computations.

**Alternative measures of information transfer** Above, I discussed that TE is equivalent to Granger causality for jointly Gaussian variables [103]. Granger causality is thus a measure of information transfer, imposing a linear model on the measured interaction. This raises the question, whether information transfer may be quantified by other measures of statistical dependency than the implementation of TE used in this work—measures that may have more desirable properties, like less computational demand or less requirements in term of data size.

In general, measures of statistical dependency may be categorized into two classes: directed and non-directed measures of dependency (see for example [296] for a "taxonomy" of measures; see also [297–299] for reviews). The class of non-directed measures includes linear as well as non-linear measures of dependency. These measures have been used in the past in early attempts to quantify information transfer between two processes, $X$ and $Y$: for example, MI and lagged MI have been used to quantify the dependency between two neural areas (e.g., [300–302]). However, MI—like all non-directed, correlative measures—is a static measure of information shared between two processes; hence, this shared information does not have to be the result of a transfer of information, but may also be the result of a common driving input or inner dynamics that are similar. This problem holds for all non-directed dependency measures. In an attempt to tackle this problem, the lagged MI, $I(X_{t-u}; Y_t)$, was introduced. However, also the lagged MI will not reveal the full information transfer in cases, where information transferred from $X$ to $Y$ is mostly synergistic [23] (see also [303] for a practical comparison of lagged MI and TE). Also, in bidirectionally coupled systems, the lagged MI and other measures of cross-correlation may be hard to interpret, because of multiple, significant peaks in these measures for both directions of interaction [296]. In sum, non-directed correlative measures are ill-suited to measure the *transfer* of information because they fail to model the sender-receiver relationship of information transfer.

In contrast to non-directed measures, directed measures explicitly model the sender and the receiver of information transferred, when estimating the dependency between two processes. Amongst these directed measures, TE (and also Granger causality) are the most popular measures of information transfer, because they explicitly quantify how much information from a source process $X$ is used in computing the next state of a target process $Y$; in other words, TE and Granger causality quantify the influence of $X$ on the transition probabilities of $Y$ [12]. This influence is quantified by estimating the mutual information between the past of $X$ and the present value of $Y$ while conditioning on the past of $Y$—in doing so, both measures introduce a directional aspect and allow to infer which of two interacting processes is the sender and which the receiver of the information transfer, or if information transfer is bidirectional such that both processes send and receive information.

The difference between TE and Granger causality is that Granger causality imposes a linear model on the dependency between two processes, while TE is model-free. The linearity-assumption underlying Granger causality makes its calculation computationally less demanding and thus attractive as an alternative measure of information transfer. However, even though Granger causality is the preferable method if linear interactions can be expected, linearity can in general not be assumed for neuroscience data. We discussed this limitation in Section 2.5.5, *Relation of the ensemble method to other measures of connectivity for non-stationary data*. Hence, for neural data, linear approaches like Granger causality and derived measures (e.g., the directed transfer function [304] or partial directed coherence [305]) may fail in the detection of *non-linear* information transfer.

A further alternative to measuring information transfer with TE is the use of symbolic TE. Symbolic TE is often used because—like Granger causality—its estimation is computationally less demanding. However, as discussed in Chapter 4, *The relation of local entropy and information transfer suggests an origin of isoflurane anesthesia effects in local information processing*, symbolic TE fails in some cases, where data discretization destroys important information about neighborhood relationships (see also [75]). Hence, TE in its original formulation is preferable over symbolic TE for continuous data in most application scenarios.

Further alternative directed information-theoretic measures been proposed for measuring information transfer. Two examples are the momentary information transfer and the directed information. However, momentary information transfer has been shown to fail to identify the correct information transfer delays in some cases [75, 257], while directed information has been shown to be equivalent to TE when applied in practice [257].

A last class of alternative measures of information transfer are measures investigating neighborhood-relationships in the reconstructed state spaces of observed systems. Measures include convergent cross mapping [306, 307] or generalized synchronization (e.g., [308]). These measures seem promising for the application in neuroscience—for example, an implementation of convergent cross mapping introduced in Schumacher et al. [307] models the influence of an unobserved global driving system on information transfer. This assumption may address the problem of unobserved sources, often present in neuroscience research. Yet, these measures are novel and have been so far less often applied than TE—hence it remains open whether their application is feasible in neuroscience research.

In summary, TE estimated from continuous data is the most robust estimator of information transfer in neuroscience today (see simulation studies in [84, 268]). Other measures of dependency are conceptually different and do not measure information transfer in the sense of Wiener's principle of *predictive information transfer* [68, 265]. They may also fail due to methodological problems. The current practical drawbacks of TE estimation, for example, its high computational demand, may be improved by novel estimation techniques; these problems and potential solutions are discussed in the next section.

### 5.1.2 Practical problems in the estimation of transfer entropy

When estimating TE from experimental data, not only conceptual, but also practical problems arise. In the following I will review the most pressing of these problems and discuss potential solutions.

**Estimator bias and bias-correction methods** Central to the inference of TE from experimental data is the choice of a suitable estimator of entropy or MI. The quality of an estimator is judged by its bias-properties, variance, and consistency (see Section 1.4, *Open problems in estimating information processing measures in neuroscience*). Also—specifically when estimating TE or AIS using an embedding to represent past states—a further criterion is the estimator's ability to perform well on high-dimensional data. These requirements have to be considered when choosing an estimator to obtain valid and robust estimates of information-theoretic quantities.

The listed requirements are met by the KSG-estimator for MI used throughout the present work. The KSG-estimator is an estimator for continuous data, which was the data type primarily investigated in the present work (MEG recordings and LFPs), and is a data type frequently encountered in neuroscience research. Amongst continuous estimators for MI, the KSG-estimator shows the most favorable bias properties [83, 84], is sensitive to small dependencies in noisy data [75], performs well on high-dimensional data [85], and showed to be consistent in simulation studies [85, p. 25] (see also [90])—however, its consistency has not yet been formally proven. Overall, in simulation studies the KSG-estimator has performed well in comparison to other continuous estimators across various data sizes and noise levels [84]. The estimator relies on the setting of one parameter only, namely, the number of nearest-neighbors $k$, while being relatively robust against various values of $k$ [83, 261, 309].

A downside of the KSG-estimator is its bias that is not analytically tractable. Through simulation studies, it has been demonstrated that the bias depends on the dimensionality of the data $d$, the number of the nearest neighbors $k$, and the number of samples $N$ [85]; also bias and variance are reciprocally influenced by $k$, where smaller values for $k$ lead to smaller bias and larger variance, while larger values lead to smaller variance and larger bias [84, 85]—yet, a functional relationship between these parameters and the bias could not be established [85].

A consequence of the intractability of the bias is that no correction exists, such that obtained estimates can not be interpreted at face value. Instead, estimates are commonly used as a statistic and tested for their statistical significance [13, 101]. Such a test requires the bias to be constant over all sets of data tested against each other, such that differences in the estimates can be attributed to differences in information transfer rather than to differences in estimator bias. The bias can be held constant by using the same estimation parameters (see last paragraph) for the estimation in each group.

When testing estimates for their statistical significance, two tests are common [86]: first, estimates can be tested for statistical significance under a null-hypothesis of no information transfer, by testing them against estimates from surrogate data; second, estimates from two or more sets of data can be tested for significant differences by comparing them against each other. Tests are usually performed as permutation tests, because estimated TE values are not known to follow a certain distribution [13]. Surrogate data for permutation testing are typically created by shuffling realizations $\mathbf{x}_{t-u}$ against realizations of $(y_t, \mathbf{y}_{t-1})$ to keep dynamics within processes intact, while breaking up the dependencies between processes [86, 101]. Testing estimates for their statistical significance instead of interpreting them directly, shifts the research question from quantitative to qualitative—instead of quantifying the precise amount of information transfer, it is asked if information transfer is higher in one group over the other, or if information is transferred *at all*. For an example application of a permutation test between groups of data, see [21].

Next to the remaining bias, an important limitation of the KSG-estimator (but also other continuous estimators of MI) is its high demand in data and computing time (see also Chapter 2, *Efficient transfer entropy analysis of non-stationary neural time series*). Due to these limitations, studies often revert to discretization of continuous data to apply discrete estimators, which require less data and computing time. However, as discussed in Section 4.4.4, *Estimation of information theoretic measures*, the use of discrete estimators should be avoided, because the necessary discretization of continuous data causes data-loss, namely, the loss of neighborhood-relationships between samples. Furthermore, discrete estimators also have unfavorable bias-properties—it has been shown that the maximum likelihood or plug-in entropy estimator is negatively biased, albeit consistent, while the MI estimator has a positive bias [310]. Even though various bias-correction methods have been proposed (see [311] for a review), bias-problems remain, especially in the so-called under-sampled regime, where $N < |\mathcal{A}|$. Simulation studies showed that in the under-sampled regime bias-corrected plug-in estimators as well as the Bayesian NSB-estimator showed a bias when estimating MI [57, 274]. Only in the asymptotic sampling-regime, where the number of observations is close to the alphabet size of the sampled variable, estimates approached the true value. Here, estimators required a number of samples $N \geq 2 - 4 \cdot |\mathcal{A}|$ to obtain unbiased estimates [274]. Such a sample size may not always be obtainable in neuroscience experiments—this is especially the case when estimating information-theoretic measures using state-variables, for which the alphabet size grows exponentially in the state's dimension.

In conclusion, the KSG-estimator provides a powerful approach to the estimation of TE from noisy data when its practical limitations are accounted for. It has often been applied in neuroscience research [14, 18–22, 312] and implementations in numerous software packages exist [86, 87, 313]. The estimator's limitations have to be carefully considered: for example, because of the intractable bias, it excludes the ability to quantify the exact amount of information transfer, which may be desirable to relate different information-theoretic measures (as discussed in Chapter 4); furthermore, the estimator requires a considerable amount of data to obtain reliable estimates. Here, novel developments such as the Bayesian estimator proposed by Nemenman et al. [225] (see also [226]), as used in Chapter 4, *The relation of local entropy and information transfer suggests an origin of isoflurane anesthesia effects in local information processing*, promise an alternative approach that yields interpretable values. Yet, further research is needed to establish the bias properties of this estimator (see [274]) and to provide recommendations for parameter settings, e.g., for the discretization of continuous data. Here, data loss has to be balanced against feasible alphabet sizes, such that important information is retained while a proper sampling of variables is guaranteed (see also Section 4.4.4, *Estimation of information theoretic measures*). In the application presented in Chapter 4, KSG- and NSB-estimators showed qualitatively similar results, indicating that the NSB-estimator may be a viable alternative to the estimation of information-theoretic measures in future research.

**Monotonicity of used estimators** As discussed in the last subsection, the bias of the KSG-estimator is typically handled by using estimates as a test-statistic in a comparison between groups of data. Here, the question asked shifts from precise estimates to *relative* statements about information transfer in two or more groups. This approach tacitly assumes that estimates are monotone in the strength of the dependency, i.e., stronger dependencies will result in higher estimates and vice versa, such that the order of dependencies is preserved in the order of the respective estimates. Yet, this assumption may not hold for the KSG-estimator.

In a recent study, Gao et al. [314] demonstrated that the relationship between the dependency in the data and estimates from the KSG-estimator is potentially non-monotone. The authors showed that the KSG-estimator demands exponentially more data to correctly estimate stronger relationships between two variables (given constant dimensionality $d$ and a number of nearest neighbors $k \geq 1$): the estimator required a number of samples $N \geq C \exp\left(\frac{I(x)-\epsilon}{d-1}\right) + 1$, where $C = exp(-\frac{k-1}{k})$ to approximate the true MI (i.e., $|\hat{I}_{KSG,k}(x) - I(x)| \leq \epsilon$). Thus, for two sets of data of equal size and a sufficient difference in the true MI, the error of the KSG-estimator may be higher for estimates of the higher MI, than for estimates of the weaker MI. This may be problematic when comparing the estimates from groups of data if the error in the estimate of a stronger dependency overlaps with the estimate of a weaker dependency—in such a case, stronger dependencies may not necessarily result in higher estimates of MI.

Gao et al. [314] furthermore introduced an algorithmic correction for the estimator bias—yet, their solution is computationally costly. Thus, before adopting this solution, further research and simulations are needed to determine the impact of the bias on neuroscience data. Here, dependencies may be typically weak enough such that the available data sizes are sufficient to not suffer from estimation bias due to stronger true dependencies. Gao et al. [314] provide theoretical and empirical lower bounds for $N$ for different values of $I$, where 100 to 1000 samples were sufficient to estimate a MI of 2.5 nats or approx. 3.6 bit. Hence, if the true dependency is weak such that sufficient data can be obtained, the bias in the estimate is not dependent on the strength of the dependency (but only determined by the choice of $k$, $d$, and $N$, see last section). Also, the KSG-estimator is especially suitable to detect independence [85]; other estimation techniques may be more suitable to distinguish the strength of dependency between groups. One may first test for a deviation from independence with the KSG-estimator and, in a second step, test for a difference in the strength of dependency with another estimator, e.g., the NSB-estimator [225, 226].

**Required data sizes** To robustly estimate TE from finite data it is important for the collected samples to contain a *sufficient* amount of realizations of each involved random variable. Yet, it may not be clear what number of realizations is sufficient if—as is often the case in neuroscience—the ground truth is unknown. In theory, a sufficient amount of samples is defined such that, first, observations cover the whole process of interest, i.e., span the whole attractor of a dynamical system or its "characteristic period"; second, the sampling rate should be high enough to capture relevant high-frequent features of the system, i.e., the sampling rate should be higher than the Nyquist rate [84]. Even though these criteria provide a theoretical basis for defining what amount of data is sufficient for reliable estimation, they may be difficult to evaluate in practice—as for neural processes this ground truth is unknown. Here, data obtained from computational models may help to validate measures like TE, but respective frameworks just begin to evolve [315]. In the present work, we used the reconstructed delay $u$ as a criterion for the exactness of estimated values (Section 2.4.5, *Evaluation of the robustness of ensemble-based TE-estimation*), given different amounts of data; here 10 000 data points collected from two coupled Lorenz systems, sampled at 1000 Hz, were required for a sufficiently accurate reconstruction.

Obtaining several thousand samples in the observation of a single process is difficult in neuroscience, because processes are typically non-stationary and evolve on a millisecond time scale (see Chapter 2, *Efficient transfer entropy analysis of non-stationary neural time series*). Typical sampling rates range from 600 Hz to 1200 Hz such that only a few hundred samples can be recorded from a process of interest; also, processes may only be stationarity within short time-windows. The present work provided a practical solution to this problem by presenting an efficient implementation of the ensemble method proposed by Gomez-Herrero et al. [24] (Chapter 2). The implementation allows to efficiently estimate TE—or other information-theoretic measures—from an ensemble of temporal repetitions of a process. The ensemble estimation increases the amount of available samples per estimate by the number of repetitions, which drastically improves the robustness of the estimate and allows the estimation from small data-windows. Furthermore, the implementation allows to estimate measures from short time-windows in which stationarity can be assumed. As an example, in human studies the length of an experiment is designed such as to prevent tiring of the subject; here, the application of the ensemble method can increase the number of available data points for estimation by an order of magnitude. Respective applications of the method are described in Sections 5.3.1, *Transfer entropy in resting state networks under ketamine [21]* and 5.3.3, *Transfer entropy estimation during a face-recognition task [316]*, below.

The required data size when estimating TE and also AIS also heavily depends on the dimensionality of the variables used, which in turn depends on the embedding of the past states in both measures. For discrete variables, higher embedding dimensions lead to an exponential increase in the alphabet size, $|\mathcal{A}_X| = O(b^j)$ (where $b$ is the base of the discrete random variable $X$ and $j$ is the embedding dimension). Increasing $j$ and thus $|\mathcal{A}_X|$ while keeping $N$ fixed leads to an under-sampling; for an example, see Fig. 5.3, where AIS estimates are plotted as a function of history length $j$: in general, after an initial increase and saturation in AIS, AIS begins to slowly increase again, indicating the onset of overestimation of AIS with a plug-in estimator due to increasing under-sampling.

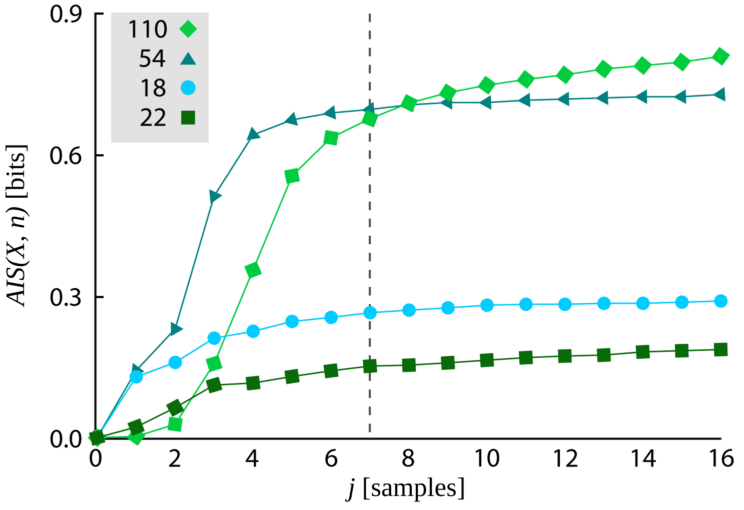


Fig. 5.3 **Active information storage (AIS) as a function of history length $j$, modified from [30].** AIS estimates are shown for different rules for elementary cellular automata. The dashed line indicates the history length for rule 110, for which the majority of the information storage seems to be captured ($j \geq 7$) according to [30].

For continuous data, higher embedding dimensions lead to similar estimation problems: when using nearest-neighbor estimators like the KSG-estimator, higher embedding dimensions lead to higher-dimensional search spaces for nearest-neighbor searches. High-dimensional search spaces may lead to searches becoming "unstable" (this may already be the case for $d \geq 15$) [317]. A neighbor search becoming unstable means that the differences in distances between a query point and most other points approach zero—making the concept of a *nearest* neighbor meaningless; more formally, "a nearest neighbor query is unstable for a given $\epsilon$ if the distance from the query point to most data points is less than $(1 + \epsilon)$ times the distance from the query point to its nearest neighbor" [317]. A further consequence of high-dimensional spaces is an increase in the required amount of data, i.e., the data needed to "fill" a space increases exponentially in the number of dimensions until the space becomes virtually empty for a very high number of dimensions [318]. These effects of a high number of dimensions on neighbor searches are also called the "curse of dimensionality" [318, p. 33].

To reduce the problems due to high-dimensional data, embedding schemes that lead to sparse, yet, maximally informative past state vectors should be used. In the present work, we used a uniform embedding with an embedding delay $\tau$ to be able to cover a relatively long history while not including all samples up to the sample furthest in the past. Future work may use more flexible embedding schemes as, for example, non-uniform embedding proposed by Faes and colleagues [319]. In non-uniform embedding, a candidate set of past samples is defined (e.g., all samples from $n-1$ to samples at maximum lag of $n-k_{max}$). Each of these samples is then iteratively tested for a significant contribution in information about the target's present at time $n$, conditional on all samples already included in the embedding vector. In an average-case scenario, this scheme may produce more sparse and hence lower-dimensional embedding vectors (see also Section 5.2.2, *Multivariate approaches to the estimation of transfer entropy* below).

**Stationarity and (truly) local transfer entropy** As discussed in the last paragraph, the estimation of information-theoretic quantities requires a sufficient amount of realizations to estimate underlying probability distributions. In practice, these realizations are often collected over time, that is, by pooling data over an observation period assuming stationarity. However, stationarity can often not be assumed in neuroscience processes, hence, we introduced the ensemble-method for estimating information-theoretic quantities from an ensemble of repetitions, which allows for the estimation from short temporal windows (Chapter 2).

Being able to estimate information-theoretic measures from an ensemble allows to estimate from time-windows that are arbitrarily small (given the ensemble is sufficiently large). Hence, the approach allows for a time-resolved estimation of information-theoretic quantities, where—in the extreme case—the resolution can be set to produce sample-wise estimates. Sample-wise estimates—as discussed in Chapter 2—enable the estimation ofso-called *local* measures of information processing [9].

Local measures of information processing are measures that quantify information storage or transfer locally in time or space, i.e., for individual samples [9]. It has been shown that these measures have meaningful interpretations in their own right and that they can be calculated by evaluating AIS and TE for individual realizations of the involved random variables (see [9] for the full derivation of this relationship). Local activity information storage (LAIS) is then written as (cf. Eq. 1.12)

$$LAIS(X, n, j) = \lim_{j\to\infty} \log_2 \frac{p(\mathbf{x}^j_{n-1}, x_n)}{p(\mathbf{x}^j_{n-1})\, p(x_n)} \tag{5.2}$$

and local transfer entropy (LTE) is written as (cf. Eq. 1.9)

$$LTE_{SPO}(X \to Y, n, u, k, l) = \lim_{l\to\infty} \log_2 \frac{p(y_n|\mathbf{x}^l_{n-u}, \mathbf{y}^k_{n-1})}{p(y_n|\mathbf{y}^k_{n-1})}. \tag{5.3}$$

When estimating local measures, like for the estimation of their non-local counterparts, probability density functions $p(\cdot)$ have to be estimated from multiple realizations of the random variables in question. If probability density functions are estimated from realizations pooled over time (assuming stationarity), local measures become non-local in the sense that their estimation relies on realizations collected from different points in time. In contrast, if instead of temporal pooling ensemble-estimation is used, local measures become "truly" local because then their estimation relies on local realizations only.

Being able to compute truly local measures of information processing from neuroscience data allows to quantify the dynamics of information processing over time, and thus adds an important aspect to the investigation of neural computation (see for example the application in Wibral et al. [278], discussed above).

**Common input and other multivariate effects** A major practical problem in the estimation of TE in neuroscience is the estimation of TE between more than two processes that interact in a multivariate fashion. These multivariate interactions comprise of common drive effects, cascade effects, and synergistic information transfer from two or more sources to a target (see also Chapter 3, *A Graph Algorithmic Approach to Separate Direct from Indirect Neural Interactions*). The detection of these interactions requires a fully multivariate approach to the estimation of TE—yet, such an approach is computationally complex, hence, often bivariate TE between all pairs of processes is estimated as an approximation. When taking this approximative approach (as was done in the present study), two classes of errors have to be expected: first, the detection of spurious information transfer between sources that are correlated due to common drive and cascade effects (see Chapter 3 and Section 5.1.1); second, the failure to detect synergistic information transfer from two or more sources to a target.

Errors of the first class, i.e., spurious bivariate TE, can be partially corrected for. We presented such a correction in Chapter 3, and discussed further correction methods. These corrections increase the validity of TE as a measure of information transfer. Yet—as the underlying problem of inferring the true multivariate interactions is NP-hard—we can not hope to find an exact solution to the inference of multivariate TE in all but the smallest problem instances in polynomial time if P $\neq$ NP. Accordingly, we can not be sure if bivariate TE estimated from observational data is non-spurious and does represent true information transfer; rather, we can only decrease the probability of detecting false positives.

Errors of the second class, i.e., failure to detect synergistic information transfer, are not as easily corrected for post-hoc. The detection of synergistic information requires the simultaneous consideration of multiple sources when estimating information transfer. Hence, detecting synergy necessarily requires the estimation of some form of multivariate TE. Since a fully multivariate approach is not feasible, because of its computational complexity, approximative solutions have to be developed. Hence, future efforts should aim at improving existing approximations of multivariate TE (presented in Chapter 3). I will discuss potential future developments in Section 5.2.2, *Multivariate approaches to the estimation of transfer entropy*, below.

**Information transfer delays** In practical applications, the physical transfer of information between two processes requires time, causing a delay $\delta$ between information leaving the source process and arriving at the target process. For example, in neural systems information is transferred by spike trains traveling along axonal connections with specific conduction velocities. Hence, when estimating TE, the delay between source and target process has to be accounted for, such as to not underestimate information transfer (discussed in Chapter 4).

The problem of accounting for $\delta$ was solved by extending the original TE-functional by a delay parameter $u$ [75]. This parameter accounts for the delay between informative observations in the source's past state and the target's present [75] (see also Eq. 1.11 in Section 1.3.1, *Transfer entropy* and Section 4.4.4, *Estimation of information theoretic measures*). An optimal value for $u$ can be found by estimating $TE_{SPO}$ using a range of candidate values, and using the $u$ that maximizes $TE_{SPO}$ (see Eq. 1.11); this $u$ then reconstructs the true delay $\delta$, given the true delay is amongst the candidate values.

Optimizing $u$ when estimating TE from experimental data is crucial to avoid the underestimation of TE (discussed in Section 4.4.4, *Estimation of information theoretic measures*). Furthermore, the detection of the primary direction of information transfer hinges on optimizing $u$ for both directions of information transfer independently. In sum, the information transfer delay should always be taken into account when estimating TE in practice.

**Optimization of embedding parameters** Lastly, when estimating TE and AIS past states have to be accounted for by constructing an embedding from the variables in the past of the process under investigation. We discussed the impact of under-embedding of past states in Chapter 4, *The relation of local entropy and information transfer suggests an origin of isoflurane anesthesia effects in local information processing*, namely, the under-estimation of TE or AIS, and—for TE—the detection of spurious information transfer in the wrong direction. Optimization procedures to construct past states exist (e.g., [74, 133]) and have been discussed in Chapter 4 as well as in this section—these optimization procedures should always be chosen over ad-hoc choices for embedding parameters.

### 5.1.3 Summary

I have discussed conceptual and practical problems in the operationalization of neural information transfer as TE. In summary, TE is a robust measure of information transfer in neural systems; it has been established as a measure of (neural) information processing and as such directly targets a level of analysis that has to be distinguished from levels of causal and functional analysis, but has been addressed rarely. TE thus fills an important methodological gap in the analysis of neural systems and information processing systems in general.

The estimation of TE from neural data is non-trivial—it requires the prior optimization of delay and embedding parameters, as well as the careful evaluation of available data in terms of sample size, stationarity, and dimensionality. Optimization procedures for these parameters exist and have been discussed extensively in the present work. If data requirements are met, estimators such as the KSG-estimator provide a robust way of estimating TE from neural data, while being virtually parameter-free and exhibiting favorable bias-properties that are controlled for by permutation testing. The present work provided an important contribution to the estimation of TE by presenting an efficient implementation dealing with non-stationary data. Emerging alternative estimators like the Bayesian NSB-estimator may further improve the estimation of TE in the future, especially in the asymptotically- and under-sampled regime.

A problem that remains partially open are multivariate effects when estimating information transfer from neural data—despite the correction method proposed in the present work and by others, not all multivariate effects, e.g., synergistic information transfer, are considered by current estimation methods. Here, more research is needed to extend existing multivariate approaches.

## 5.2 Future directions

In the last section, I discussed how the present work improved the estimation and interpretation of TE in neuroscience research. In the following, I will point out directions for future research.

### 5.2.1 Quantification of information modification and its relevance to the interpretation of transfer entropy

As described in Chapter 1, *Introduction*, measures of information transfer and storage are theoretically well-defined in the framework of information theory [9]; however, for information modification an equivalent definition and measure is lacking. Finding such a measure is desirable in neuroscience, because it would allow to quantify the third of the three proposed building blocks of computation, information transfer, storage, and modification [5, 7]. Taken together, these measures would allow to fully describe how the next state of a system is computed from information transferred into the system, information stored in the system, and the modification of stored or transferred information or both into a new form. Thus, finding a measure of information modification allows for a complete description of information processing in the framework of local information dynamics [9].

A candidate measure for information modification has been proposed in the theoretical framework of partial information decomposition (PID), namely, *synergistic mutual information* [72]. Synergistic information captures the intuitive notion of information modification—"a non-trivial processing of information from two or more (storage or transfer) sources" into a new form [78]—yet, a practical measure of this quantity is lacking. Measures of synergistic information have been proposed [76, 79, 80], however, none of these measures meets all theoretical requirements for a measure of information modification as formulated in the framework of local information dynamics [9, 78]. The most promising candidate measure at the time of writing is a measure of synergistic information recently introduced by Bertschinger et al. [81]. Yet, the measure is limited by the fact that it is currently only applicable to the decomposition of the joint mutual information of two input and one output variable (which can both be multivariate themselves, however) [8]. Hence, the development of a theoretically sound and computationally feasible measure, as well as its estimation, is still an area of active research.

Operationalizing information modification as synergistic information and developing an appropriate measure is not only desirable in its own right, but also has important implications for the interpretation of TE: when calculating TE, information shared by both past states is "conditioned out"; on the other hand, synergistic information between both past states is "conditioned in" [76] (Fig. 5.4). Hence, using PID, TE can be decomposed into two information contributions about the target: first, the unique information present only in the source past state; second, the synergistic information present jointly in the source and target past states. These two components have also been termed state-independent TE (SITE)—the unique information from the source—and state-dependent TE (SDTE)—the synergistic information in source and target [76]. Being able to distinguish these two contributions allows for additional insights into performed computations; for example, Williams and Beer [76] provide an example application of the decomposition of TE to the analysis of heart and breath rates—the authors found that TE was almost entirely due to synergistic information transfer from heart to breath rate, which was highest for low chest volumes—thus adding substantially to the interpretation of TE from one physiological marker to the other.

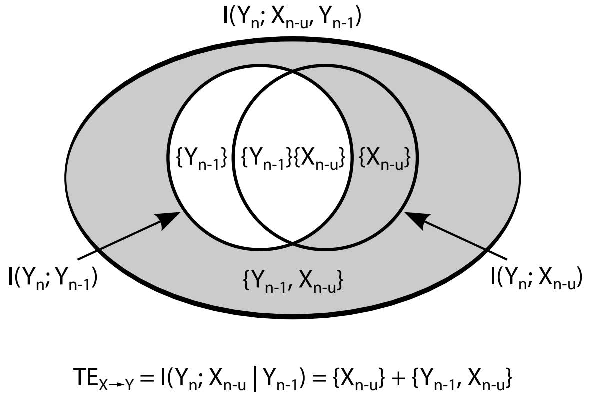


Fig. 5.4 **Decomposition of transfer entropy (TE) into state-dependent and state-independent information transfer.** Information that two variables, $\mathbf{Y}_{n-1}$, $\mathbf{X}_{n-u}$, have about a third, $Y_n$, can be decomposed into *unique information* contributed by each of the two variables, $\{\mathbf{Y}_{n-1}\}$, $\{\mathbf{X}_{n-u}\}$; *synergistic information* contributed jointly by the two variables, $\{\mathbf{Y}_{n-1}, \mathbf{X}_{n-u}\}$; and *shared information* contributed redundantly be the two variables, $\{\mathbf{Y}_{n-1}\}\{\mathbf{X}_{n-u}\}$. TE from process $X$ to process $Y$ (gray area) can be decomposed into the unique information from the source, $\{\mathbf{Y}_{n-1}\}$, called state-independent TE (SITE); and the synergistic information, $\{\mathbf{Y}_{n-1}, \mathbf{X}_{n-u}\}$, called state-dependent TE (SDTE). Shared information and unique information in the target's past is conditioned out (white area).

Furthermore, separating TE into SITE and SDTE may help to resolve a problem in the interpretation of TE recently raised by James et al. [262]: the authors criticize the common interpretation of TE as information transfer into a target from one source *alone*, because this interpretation "localizes" the source of information transfer in a *single* process. However, this localization may be misleading if TE is mainly due to synergistic information the past states of source and target have about the target. James et al. [262] state that in such a scenario, the interpretation of TE as information originating from the source alone may be flawed. They further state that this localization has consequences for the network representation of neural activity, because synergistic information transfer is falsely labeled as transfer from a single source: when adding this transfer as a directed edge to the network, a spurious link is introduced. However, in the terminology of Williams and Beer [76], synergistic information—SDTE—may be viewed as information that is only transferred if the target process is in a given state (represented by its past state); in other words, information is only transferred from a source if it can be "decoded" by the past state of the target. When adapting this view on synergistic information, "localizing" the source of information transfer seems justified.

In summary, finding a measure of synergistic information as defined in the PID framework [72], complements existing measures of information transfer and storage, and holds potential for further insights into (neural) information processing. Furthermore, being able to measure information modification allows for additional insight into information transfer measured by TE, because it allows to quantify the information transfer due to information synergistically present in both past states.

### 5.2.2 Multivariate approaches to the estimation of transfer entropy

When discussing limitations of the TE-estimator used in the present work (section 5.1), I discussed its restriction to the estimation of bivariate TE. Bivariate TE does not address multivariate effects in information transfer; in particular, it misses synergistic information transfer from two or more sources to a target. This omission can not be easily corrected for; thus, approaches to the estimation of multivariate TE are needed. However, as stated before, a fully multivariate approach to finding multivariate information transfer into a target is intractable due to the computational complexity of identifying the subset of all informative sources from a set of candidate processes (Chapter 3, *A Graph Algorithmic Approach to Separate Direct from Indirect Neural Interactions* and Section 5.1.2, *Practical problems in the estimation of transfer entropy*). Hence, only approximative approaches to the estimation of multivariate TE are feasible in practice. We reviewed existing approximative methods in Chapter 3—yet, these approaches have conceptual limitations or implementations are lacking.

Hence, future research should aim at improving approximative methods for multivariate TE estimation. Among the methods reviewed in Chapter 3, the most promising approach is the greedy algorithm proposed by Lizier and Rubinov [25]: it calculates multivariate TE by identifying informative sources for a given target process by iteratively testing each candidate source for significant information about the target's present, $y_n$. If a source has significant information about $y_n$ (conditional on all sources already included), it is added to the set of informative sources. Then the next source is tested and so on. If no candidate provides any new information about $y_n$, the algorithm terminates. Multivariate TE is then calculated from the set of informative sources by estimating TE between a single informative source and the target, conditional on all remaining informative sources. Note that the algorithm considers individual source *samples*, which, in the average case, results in a maximally sparse representation of the informative sources. A sparse representation is crucial, because the estimation of TE may become intractable if the dimension of the involved variables is too high (see Section 5.1.2).

The greedy algorithm by Lizier and Rubinov [25] shares similarities with an algorithm for non-uniform multivariate embedding proposed by Faes et al. [74]. The non-uniform embedding also selects informative sources from a set of candidates by iteratively testing for the contribution in information about some present value $y_n$. However, when searching for informative sources, past samples from the sources' and target's past are simultaneously tested for their information contribution. Hence, the algorithm does not first test exhaustively for informative past samples in the *target's* past, before testing the samples in the *sources'* past. The algorithm thus jeopardizes self-prediction optimality which is an important requirement for the correct estimation of TE and its interpretability as predictive information transfer in the Wiener-Granger sense [75]. If the full information present in the target's own past is not included before evaluating the information present in other sources, information transfer from these other sources may be overestimated.

An implementation of the non-uniform embedding [74] can be found in the MuTE toolbox [313]. In addition to the conceptual problems when interpreting the estimated quantity as a TE, the implementation has two shortcomings: first, a correction for multiple statistical testing during the iterative evaluation of individual samples is lacking; second, estimators do not make use of the multi-trial structure commonly encountered in experimental data (especially in neuroscience). The latter shortcoming prevents an application of the ensemble method (Chapter 2, *Efficient transfer entropy analysis of non-stationary neural time series*) to allow for the estimation of TE from small time-windows, for example, in the case of non-stationary data.

Thus, future research should aim at developing an efficient and conceptually sound implementation of approximative, multivariate TE. A promising starting point for such a development is the greedy algorithm by Lizier and Rubinov [25][1], which provides a tractable approach to the representation of multiple sources of information transfer, while implementing the Wiener-Granger principle. Furthermore, future software implementations should make use of the methods presented in Chapter 2 to allow for the estimation from ensembles of time series and to handle the costly, iterative statistical testing. In general, implementations could benefit from the massively parallel capacities provided by GPUs as presented in Chapter 2, to ensure a favorable scaling for bigger data size[2]. In this work, we have shown that the respective GPU-implementations significantly lessen the computing time required

[1] Note that the greedy algorithm is similar to the approach by Marinazzo et al. [217], which was also presented in Chapter 3, *A Graph Algorithmic Approach to Separate Direct from Indirect Neural Interactions*.

[2] Note that in the meantime also a OpenCl implementation of the presented neighbor searches for GPU was developed [320]

for TE estimation; future software should expand this approach to allow for the estimation of multivariate TE and information-theoretic measures.

On a cautionary note, the approach to multivariate TE estimation proposed here, is approximative because, first, sources are tested iteratively, i.e., single sources will be missed if they provide significant synergistic information together with other sources, but no significant unique information; secondly, because information transfer from a candidate source is tested conditionally on all sources already included, redundant information transfer will be missed [25]. The presented approaches aim at producing a so-called *minimal computationally equivalent* network [25]. This means, the approaches reconstruct the minimal information transfer needed to reproduce the dynamics found in the data. In other words, the approach infers the minimal network of information transfer that is computationally equivalent to the original network inasmuch as the reconstructed network is able to produce the observed dynamics in the data. As an example for missed redundant information transfer, consider information transfer $A \to Y$ and $B \to Y$, where both sources have a lot of redundant information, i.e., both links transfer redundant information. Depending on which link is reconstructed first, the respective second link will not be detected, because the information transferred is already accounted for. Hence, the approach will find a network that is equivalent with respect to the dynamics it enables, yet, not all links present are necessarily reconstructed. In contrast, the approach presented by Marinazzo et al. [217] and extended by Stramaglia et al. [218] aims at post-hoc reconstructing interactions missed due to shared information. Thus, the latter approach aims at inferring a network that is more true to the underlying *causal information flow* [118] and thus closer to the actual causal connections. However, we here propose to opt for a more computational approach, where the reconstruction of information transfer is given precedence over the reconstruction of causal interactions. The computational approach may be more desirable, because we are ultimately interested in investigating the computations carried out by the observed dynamics, rather than reconstructing the underlying causal structure—the latter being most likely not attainable by the estimation of TE (see section 5.1.1).

## 5.3 Application of the proposed methods

The methods proposed in Chapters 2, *Efficient transfer entropy analysis of non-stationary neural time series*, and 3, *A Graph Algorithmic Approach to Separate Direct from Indirect Neural Interactions*, have successfully been applied in neuroscience studies, reviewed in the following.

### 5.3.1 Transfer entropy in resting state networks under ketamine [21]

Both methods presented in Chapters 2 and 3 were applied in the analysis of resting state MEG activity under administration of ketamine and under administration of a placebo [21]. Bivariate TE was estimated between all identified, relevant sources of neural activity. Here, the ensemble method was used to allow for the estimation of TE from a significantly higher amount of data by pooling data over experimental trials. The resulting TE network showed a high initial connectivity. This connectivity could not be explained by volume conduction or other zero-lag interactions, because this was accounted for by using a method proposed by Faes et al. [137] that additionally conditions on the present value in the source process, $X_t$, when estimating $TE_{SPO}(X \rightarrow Y)$, and thus conditions out any influence with a zero lag relative to time point $t$. However, when applying the post-hoc correction for multivariate effects a substantial amount of potentially spurious links were identified. By removing these links, the network could be reduced to an interpretable size.

In sum, using the ensemble method together with the post-hoc graph-correction led to estimates with higher robustness, due to more data entering the estimation and due to the removal of all links flagged as potentially spurious. The study was the first application of both methods proving their feasibility in the analysis of real-world neuroscience data with TE. TE was estimated using the TRENTOOL toolbox, including the implementations of both presented methods [86, 321].

### 5.3.2 Transfer entropy estimation as preprocessing step in DCM analysis [264]

In a second study [264], we used bivariate TE estimation together with the graph-algorithm (Chapter 3) as a preprocessing method for dynamic causal modeling (DCM) for MEG/EEG data [214, 215]. DCM is a method that tries to infer causal connectivity between neural areas, underlying the neural activity measured during an experiment. DCM achieves this through a combination of realistic, biophysical modeling and statistical data analysis [214, 322, 323]. We used TE estimation to inform the a-priori choice of a model, as has been proposed in Chapter 3 and by Friston et al. [324].

When trying to infer causal connectivity from experimental data with DCM analysis, first, a biophysical network model of involved areas and their connections has to be formulated; second, through Bayesian model inversion, this model is tested for its likelihood to have generated the observed experimental data. In the biophysical model, neural areas are represented by network nodes, whose activity is described by a “neural mass model” [325]. Neural mass models describe the combined activity of thousands of neurons that underlies the measured MEG/EEG signal. The variation of this activity under experimental manipulation is described by ordinary differential equations. Through a further set of equations, the nodes’ activity is then mapped onto MEG/EEG signals measurable at the scalp. Through Bayesian model inversion, parameters for both sets of equations are inferred. DCM thus formulates a realistic, biophysical model of brain areas and their connections, and combines this causal model with a statistical model of how the biophysical architecture gives rise to signals measured during an experiment. DCM thus allows for two analyses: first, parameter estimation for a given model to answer how connectivity in this model changes under the given experimental manipulation; second, comparison of various hypothesized models to find the model architecture most likely to have generated the observed data.

The latter analysis strategy is important in cases where multiple possible models exist. Models are typically formulated based on existing, anatomical evidence about brain connectivity. However, this evidence may be lacking for certain connections, in which case multiple models become possible (if no evidence about a single connection exists, two candidate models, one with and one without the connection in question, can be formulated). Hence, often multiple candidate models for DCM analysis exist. If this is the case, it is possible to use *Bayesian model comparison* to test these candidate models against each other. In this test, Bayesian model comparison tries to identify from a set of candidate models the model with the highest model evidence, which is the probability of observing the data given this particular model.

Bayesian model comparison finds the model with the highest *relative* evidence given the set of candidate models and the observed activity [322, 326]. In other words, Bayesian model comparison does not yield an absolute assessment of whether an individual model is “true” or even adequate, but it will only determine the “best” model given the candidate models and the data. The relative nature of model comparison has to be considered such as to avoid the selection of a poor model. Here, poor means that the model either has a bad absolute model fit, or that the model has high fit but is highly implausible in the context of neural data. In the former scenario a model with poor fit is selected because it is compared to models with even worse fit. In the latter scenario, an implausible model is selected because it has higher model evidence than all other (plausible) models or is over-fitted to the data—in general, for every formulated (plausible) model it is probable that the theoretical search space of all possible (implausible) models contains many more models with very similar or better fit to the observed data (see [326] and references therein for a critical discussion of this issue). In sum, it is central to Bayesian model comparison that the the a-priori formulated candidate models are biologically *plausible* such as to minimize the risk of unknowingly selecting a poor model either in terms of fit or interpretability [322].

A straightforward strategy for the formulation of plausible candidate models is the use of existing, anatomical evidence—but if no such evidence is available, alternative approaches have to be found. One alternative proposed is the sampling of the full space spanned by all potential models [322]. However, this potential model space may quickly become too large and the number of models to be tested intractable. Also, not all models in the theoretical space are plausible ones, violating the central premise of DCM discussed above. Hence, it has been proposed instead, to use "exploratory", i.e., data-driven, methods like Granger causality or TE to define an initial model [327, 328].

In the second application study, we therefore used bivariate TE estimation to inform the formulation of a set of candidate models for DCM-analysis to analyze MEG data recorded during the presentation of an audio-visual stimulus (sound-induced flash illusion (SiFi)) [264]. We estimated TE between all reconstructed sources of neural activity using the ensemble method presented in Chapter 2 and its implementation in [86, 321]. We used the algorithm presented in Chapter 3 to identify coupling motifs characteristic of potentially spurious information transfer. The set of candidate models fro DCM was then defined by iteratively removing potentially spurious links from the TE network. We then used DCM to test these models against each other through Bayesian model comparison. Because in a triangle motif, the link to be removed is ambiguous, we encoded simultaneously removable links as a Boolean function: the function was formulated such that it evaluated to true if no "forbidden" links were removed, when removed links were represented by assigning "false" to a variable. We thus made sure not to violate the rationale underlying our algorithm (i.e., at any given time, one of the two links leading to a common drive or cascade effect was preserved while the other was removed from the triangle).

In conclusion, in this study we not only used TE as a preprocessing step for DCM analysis, but we also used DCM to validate bivariate TE results: as stated in Section 5.1, *Application of transfer entropy in neuroscience*, bivariate TE may be spurious due to multivariate effects. Here, DCM allowed us to test for spuriousness of individual TE-links by inferring their underlying causal connection—if no causal connection existed, TE could be assumed to be spurious, because the causal connection necessary for information transfer was lacking. Thus, we successfully demonstrated the combination of the two approaches for the first time—TE meaningfully reduced the model space prior to DCM analysis in the absence of additional evidence for causal connectivity; DCM validated results from bivariate TE estimation by testing for the existence of potentially spurious links with DCM.

### 5.3.3 Transfer entropy estimation during a face-recognition task [316]

In a third study [316], we used AIS and TE to investigate the theory of predictive coding (PCT) in data collected in the MEG during a face detection task. TE was estimated using the ensemble method presented in Chapter 2, using the implementation in [86, 321]. AIS was estimated using the JAVA information dynamics toolkit [87].

PCT assumes that the brain constantly infers the external causes of its sensory inputs by internally maintaining a model of the potential causes and trying to predict future inputs under this model [3, 235, 279]. The model is constantly refined by matching predictions against novel input, thus incorporating novel knowledge into the model about the world. Future input is then in turn matched against predictions derived from the updated model, i.e, it is matched against "prior knowledge". Thus, PCT assumes that the brain processes input in an iterative fashion, where the brain tries to predict novel input by activating "prior knowledge".

We investigated the activation of prior knowledge in MEG activity, recorded during a face/house detection task to test PCT as a theory of brain function. We used AIS to quantify activated prior knowledge in task-relevant areas, because AIS quantifies the amount of information actively in use to compute the next state of the system—AIS thus captures the notion of knowledge actively in use to predict the next state of a neural area, representing novel, incoming input. Hence, AIS should be high in content-specific areas (areas concerned with the detection of houses or faces), and AIS should facilitate behavioral performance. Additionally, we used TE to quantify the amount of information being transferred into areas concerned with the detection of a face or house. TE should be high in top-down direction, especially for areas with high AIS, indicating the transfer of information related to predictions. Results supported PCT as a general principle underlying the detection of faces: increases in AIS were found in task-related areas for face processing, also increased AIS was related to better performance in face detection.

In summary, information theory allowed for the formulation of precise and testable hypotheses about information processing in service of PCT: first, about the storage of information for the prediction of future events; second, about the transfer of novel information for the update of activated knowledge. The use of information-theoretic measures allowed to formulate hypotheses that did not rely on assumptions on the semantics of neural activity, i.e., they did not require assumptions regarding which features encoded or represented predictions or novel input (discussed in Section 5.1.1, *Operationalizing neural information transfer as transfer entropy—conceptual* *considerations*). The analysis was made possible by using the ensemble method (Chapter 2), which enabled the precise estimation of TE during the baseline period (prior to the task). A future extension of the TRENTOOL toolbox should include a similar implementation for AIS estimation.

## 5.4 Conclusion and outlook

In this work, we presented methods for the estimation and interpretation of TE and, partly, other information theoretic measures in neuroscience data. We showed how to efficiently estimate TE from non-stationary data and how to correct for multivariate effects when estimating bivariate TE in a multivariate setting. We thus contributed significant improvements in the estimation of TE and solved two of the most pressing practical problems in the estimation of bivariate TE in neuroscience. Implementations of all developed methods are published in an open-source software package to be used by other researchers. We furthermore presented work on current best-practice in applying and interpreting TE in neuroscience research, and discussed consequences of violating this practice.

I discussed these methodological improvements as well as the necessity of an information-theoretic analysis framework in neuroscience. Future research should aim at further improving the estimation of TE by developing new estimators and porting the presented methods to high-performance computing hardware. Also future work should extend the theoretical framework of information dynamics, in particular by developing measures of information modification, to enable an exhaustive analysis of information processing operations in neuroscience data.

# Bibliography

[1] M. J. Schnitzer. “Biological computation: amazing algorithms”. *Nature*, 416(6882) (2002), pp. 683–683 (cit. on pp. iii, 1, 152, 205).

[2] M. Carandini. “From circuits to behavior: a bridge too far?” *Nature Neuroscience*, 15(4) (2012), pp. 507–509 (cit. on pp. iii, 1, 2, 152, 157, 205).

[3] A. Clark. “Whatever next? Predictive brains, situated agents, and the future of cognitive science”. *Behavioral and Brain Sciences*, 36(03) (2013), pp. 181–204 (cit. on pp. iii, 1, 119, 154, 179, 205).

[4] L. De-Wit, D. Alexander, V. Ekroll, and J. Wagemans. “Is neuroimaging measuring information in the brain?” *Psychological Bulletin & Review* (2016), pp. 1–14 (cit. on pp. iii, 1, 152, 156–158, 205).

[5] C. G. Langton. “Computation at the edge of chaos: phase transitions and emergent computation”. *Physica D: Nonlinear Phenomena*, 42(1) (1990), pp. 12–37 (cit. on pp. iii, 3, 13, 22, 152, 171, 206).

[6] M. Mitchell. “Ubiquity symposium: Biological computation”. *Ubiquity*, 2011(February) (2011), p. 3 (cit. on pp. iii, 1, 3, 152–154, 156, 157, 205, 206).

[7] M. Mitchell, P. Hraber, and J. P. Crutchfield. “Revisiting the edge of chaos: evolving cellular automata to perform computations”. *Complex Systems*, 7 (1993), pp. 89–130 (cit. on pp. iii, 1, 3, 152, 156, 171, 205, 206).

[8] M. Wibral, J. T. Lizier, and V. Priesemann. “Bits from brains for biologically inspired computing”. *Frontiers in Robotics and AI*, 2 (2015) (cit. on pp. iii, 4, 13, 117, 154, 171, 206).

[9] J. T. Lizier. “The local information dynamics of distributed computation in complex systems”. PhD thesis. The University of Sydney, Sydney, Australia, 2010 (cit. on pp. iii, 3, 4, 13, 152–154, 157, 167, 171, 206, 212).

[10] C. E. Shannon. “A mathematical theory of communication”. *The Bell System’s Technical Journal*, 27(I) (1948), pp. 379–423 (cit. on pp. iii, 3, 5, 147, 148, 206).

[11] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya. “Coherent information structure in complex computation”. *Theory in Biosciences*, 131 (2012), pp. 193–203 (cit. on pp. iii, 4, 154, 206).

[12] T. Schreiber. “Measuring information transfer”. *Physical Review Letters*, 85(2) (2000), pp. 461–464 (cit. on pp. iii, 3, 9, 10, 22, 23, 28, 29, 56, 100, 124, 125, 159, 206).

[13] R. Vicente, M. Wibral, M. Lindner, and G. Pipa. “Transfer entropy—a model-free measure of effective connectivity for the neurosciences.” *Journal of Computational Neuroscience*, 30(1) (2011), pp. 45–67 (cit. on pp. iii, 3, 4, 14, 15, 22, 25, 28, 32, 109, 110, 125, 128, 129, 132, 153, 162, 207).

[14] M. Wibral, B. Rahm, M. Rieder, et al. “Transfer entropy in magnetoencephalographic data: quantifying information flow in cortical and cerebellar networks.” *Progress in Biophysics and Molecular Biology*, 105(1-2) (2011), pp. 80–97 (cit. on pp. iii, 16, 22, 28, 147, 163, 206).

[15] B. Gourévitch and J. J. Eggermont. “Evaluating information transfer between auditory cortical neurons”. *Journal of Neurophysiology*, 97(3) (2007), pp. 2533–2543 (cit. on pp. iii, 16, 22, 147, 206).

[16] V. A. Vakorin, N. Kovacevic, and A. R. McIntosh. “Exploring transient transfer entropy based on a group-wise ICA decomposition of EEG data.” *Neuroimage*, 49(2) (2010), pp. 1593–1600 (cit. on pp. iii, 22, 206).

[17] M. Besserve, B. Scholkopf, N. K. Logothetis, and S. Panzeri. “Causal relationships between frequency bands of extracellular signals in visual cortex revealed by an information theoretic analysis.” *Journal of Computational Neuroscience*, 29(3) (2010), pp. 547–566 (cit. on pp. iii, 22, 206).

[18] J. T. Lizier, J. Heinzle, A. Horstmann, J.-D. Haynes, and M. Prokopenko. “Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity”. *Journal of Computational Neuroscience*, 30(1) (2011), pp. 85–107 (cit. on pp. iii, 22, 163, 206).

[19] C. G. Richter, M. Babo-Rebelo, D. Schwartz, and C. Tallon-Baudry. “Phase-amplitude coupling at the organism level: the amplitude of spontaneous alpha rhythm fluctuations varies with the phase of the infra-slow gastric basal rhythm”. *Neuroimage (in press)* (2016) (cit. on pp. iii, 147, 163, 206).

[20] C.-S. Huang, N. R. Pal, C.-H. Chuang, and C.-T. Lin. “Identifying changes in EEG information transfer during drowsy driving by transfer entropy.” *Frontiers in Human Neuroscience*, 9 (2015), p. 570 (cit. on pp. iii, 147, 163, 206).

[21] D. Rivolta, T. Heidegger, B. Scheller, et al. “Ketamine dysregulates the amplitude and connectivity of high-frequency oscillations in cortical–subcortical networks in humans: evidence from resting-state magnetoencephalography-recordings”. *Schizophrenia Bulletin*, 41(5) (2015), pp. 1105–1114 (cit. on pp. iii, 120, 162, 163, 165, 176, 206).

[22] F. Roux, M. Wibral, W. Singer, J. Aru, and P. J. Uhlhaas. “The phase of thalamic alpha activity modulates cortical gamma-band activity: evidence from resting-state MEG recordings.” *Journal of Neuroscience*, 33(45) (2013), pp. 17827–17835 (cit. on pp. iii, 22, 147, 163, 206).

[23] M. Wibral, R. Vicente, and M. Lindner. “Transfer entropy in neuroscience”. In: *Directed Information Measures in Neuroscience*. Ed. by M. Wibral, R. Vicente, and J. T. Lizier. Berlin, Heidelberg: Springer, 2014, pp. 3–36 (cit. on pp. iii, 4, 16, 22, 29, 153, 159, 207).

[24] G. Gomez-Herrero, W. Wu, K. Rutanen, et al. “Assessing coupling dynamics from an ensemble of time series”. *arXiv preprint: arXiv:1008.0539* (2010) (cit. on pp. iii, 24–26, 28, 56, 63, 165, 208).

[25] J. T. Lizier and M. Rubinov. “Multivariate construction of effective computational networks from observational data. Preprint”. *Technical Report 25/2012, Max Planck Institute for Mathematics in the Sciences*, 25 (2012). URL: `http://www.mis.mpg.de/preprints/2012/preprint2012_25.pdf` (cit. on pp. iii, 23, 66, 67, 69, 95, 97, 173–175, 207, 209).

[26] A. Das and D. Kempe. “Algorithms for subset selection in linear regression”. In: *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*. 2008, pp. 45–54 (cit. on pp. iii, 66, 209).

[27] W. J. Welch. “Algorithmic complexity: three NP-hard problems in computational statistics”. *Journal of Statistical Computation and Simulation*, 15(1) (1982), pp. 17–25 (cit. on pp. iii, 66, 95, 209).

[28] M. Kamiński, M. Ding, W. Truccolo, and S. L. Bressler. "Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance". *Biological Cybernetics*, 85(2) (2001), pp. 145–157 (cit. on pp. iii, 67, 69, 70, 207, 209).

[29] K. Blinowska, R. Kuś, and M. Kamiński. "Granger causality and information flow in multivariate processes". *Physical Review E*, 70(5) (2004), p. 050902 (cit. on pp. iii, 67, 70, 207, 209).

[30] J. T. Lizier, M. Prokopenko, and A. Zomaya. "Local measures of information storage in complex distributed computation". *Information Sciences* (2012) (cit. on pp. iv, 3, 4, 12, 13, 22, 105, 124, 126, 166, 210).

[31] O. A. Imas, K. M. Ropella, B. D. Ward, J. D. Wood, and A. G. Hudetz. "Volatile anesthetics disrupt frontal-posterior recurrent information transfer at gamma frequencies in rat". *Neuroscience Letters*, 387(3) (2005), pp. 145–150 (cit. on pp. iv, 100, 116, 147, 210, 211).

[32] S.-W. Ku, U. Lee, G.-J. Noh, I.-G. Jun, and G. A. Mashour. "Preferential inhibition of frontal-to-parietal feedback connectivity is a neurophysiologic correlate of general anesthesia in surgical patients". *PLoS ONE*, 6(10) (2011), e25155 (cit. on pp. iv, 100, 110, 116, 207, 210, 211).

[33] U. Lee, S. Ku, G. Noh, et al. "Disruption of frontal-parietal communication by ketamine, propofol, and sevoflurane." *Anesthesiology*, 118(6) (2013), pp. 1264–1275 (cit. on pp. iv, 100, 110, 116, 133, 207, 210, 211).

[34] D. Jordan, R. Ilg, V. Riedl, et al. "Simultaneous electroencephalographic and functional magnetic resonance imaging indicate impaired cortical top-down processing in association with anesthetic-induced unconsciousness." *Anesthesiology*, 119(5) (2013), pp. 1031–1042 (cit. on pp. iv, 100, 116, 210, 211).

[35] G. Untergehrer, D. Jordan, E. F. Kochs, R. Ilg, and G. Schneider. "Fronto-parietal connectivity is a non-static phenomenon with characteristic changes during unconsciousness". *PLoS ONE*, 9(1) (2014), e87498 (cit. on pp. iv, 100, 110, 116, 133, 210, 211).

[36] J. G. White, E. Southgate, J. N. Thomson, and B. S. "The structure of the nervous system of the nematode Caenorhabditis elegans". *Philosophical Transactions of the Royal Society B: Biological Sciences*, 314(1165) (1986), pp. 1–340 (cit. on pp. 1, 205).

[37] L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, and D. B. Chklovskii. "Structural properties of the Caenorhabditis elegans neuronal network". *PLoS Computational Biology*, 7(2) (2011) (cit. on pp. 1, 205).

[38] M. B. Goodman, D. H. Hall, L. Avery, and S. R. Lockery. "Active currents regulate sensitivity and dynamic range in C. elegans Neurons". *Neuron*, 20(4) (1998), pp. 763–772 (cit. on pp. 1, 205).

[39] S. R. Lockery and M. B. Goodman. "The quest for action potentials in C. elegans neurons hits a plateau". *Nature Neuroscience*, 12(4) (2009), pp. 377–378 (cit. on pp. 1, 205).

[40] D. C. Marr. "The philosophy and the approach". In: *Vision*. 1982. Chap. 1, pp. 8–38 (cit. on pp. 1, 2, 150, 152, 156, 205).

[41] Z. Pylyshyn. *Computation and Cognition*. Cambridge, Mass.: MIT Press, 1984 (cit. on p. 1).

[42] R. McClamrock. "Marr's three levels: a re-evaluation". *Minds and Machines*, 1 (1991), pp. 185–196 (cit. on p. 1).

[43] W. Bechtel and O. Shagrir. "The non-redundant contributions of Marr's three levels of analysis for explaining information-processing mechanisms". *Topics in Cognitive Science*, 7(2) (2015), pp. 312–322 (cit. on p. 2).

[44] T. L. Griffiths, F. Lieder, and N. D. Goodman. “Rational use of cognitive resources: levels of analysis between the computational and the algorithmic”. *Topics in Cognitive Science*, 7 (2015), pp. 217–229 (cit. on pp. 2, 152, 156).

[45] C. Detrain, J. L. Deneubourg, and J. M. Pasteels, eds. *Information Processing in Social Insects*. Basel, Boston, Berlin: Birkhäuser, 1999 (cit. on pp. 3, 205).

[46] R. Lukeman, Y.-X. Li, and L. Edelstein-Keshet. “Inferring individual rules from collective behavior”. *Proceedings of the National Academy of Sciences*, 107(28) (2010), pp. 12576–12580 (cit. on pp. 3, 205).

[47] D. Peak, J. D. West, S. M. Messinger, and K. A. Mott. “Evidence for complex, collective dynamics and emergent, distributed computation in plants”. *Proceedings of the National Academy of Sciences*, 101(4) (2004), pp. 918–922 (cit. on p. 3).

[48] J. K. Parrish. “Complexity, pattern, and evolutionary trade-offs in animal aggregation”. *Science*, 284(5411) (1999), pp. 99–101 (cit. on p. 3).

[49] S. A. Kauffman. “Metabolic stability and epigenesis in randomly constructed genetic nets”. *Journal of Theoretical Biology*, 22(3) (1969), pp. 437–467 (cit. on pp. 3, 205).

[50] D. Bray. “Protein molecules as computational elements in living cells”. *Nature*, 376(6538) (1995), pp. 307–312 (cit. on pp. 3, 205).

[51] H. T. Siegelmann. “Complex systems science and brain dynamics: a frontiers in computational neuroscience special topic”. *Frontiers in Computational Neuroscience*, 4 (2010), pp. 4–5 (cit. on pp. 3, 205).

[52] Q. K. Telesford, S. L. Simpson, and E. D. Kolaczyk. “Editorial: complexity and emergence in brain network analyses”. *Frontiers in Computational Neuroscience*, (65) (2015) (cit. on pp. 3, 205).

[53] F. Attneave. “Some informational aspects of visual perception”. *Psychological Review*, 3(61) (1954), pp. 183–193 (cit. on pp. 3, 153).

[54] H. B. Barlow. “Possible principles underlying the transformation of sensory messages”. In: *Sensory Communication*. Ed. by W. Rosenblith. Cambridge, MA: MIT Press, 1961, pp. 217–234 (cit. on pp. 3, 153).

[55] A. Borst and F. E. Theunissen. “Information theory and neural coding.” *Nature Neuroscience*, 2(11) (1999), pp. 947–957 (cit. on pp. 3, 153).

[56] D. A. Butts. “How much information is associated with a particular stimulus?” *Network: Computation in Neural Systems*, 14(2) (2003), pp. 177–187 (cit. on pp. 3, 153).

[57] J. D. Victor. “Approaches to information-theoretic analysis of neural activity”. *Biological Theory*, 1(3) (2006), pp. 302–316 (cit. on pp. 3, 153, 162).

[58] R. Quian Quiroga and S. Panzeri. “Extracting information from neuronal populations: information theory and decoding approaches.” *Nature Reviews Neuroscience*, 10(3) (2009), pp. 173–85 (cit. on pp. 3, 153).

[59] R. A. A. Ince, R. Senatore, E. Arabzadeh, et al. “Information-theoretic methods for studying population codes”. *Neural Networks*, 23(6) (2010), pp. 713–727 (cit. on pp. 3, 153).

[60] D. Ostwald and A. P. Bagshaw. “Information theoretic approaches to functional neuroimaging”. *Magnetic Resonance Imaging*, 29(10) (2011), pp. 1417–1428 (cit. on pp. 3, 153).

[61] R. Linsker. “Self-organization in a perceptual network”. *IEEE Computer*, 21(3) (1988), pp. 105–117 (cit. on pp. 3, 153).

[62] R. Linsker. “Perceptual neural organization: some approaches based on network models and information theory”. *Annual Review of Neuroscience*, 13 (1990), pp. 257–281 (cit. on pp. 3, 153).

[63] J. W. Kay and W. A. Phillips. “Coherent infomax as a computational goal for neural systems”. *Bulletin of Mathematical Biology*, 73(2) (2011), pp. 344–372 (cit. on pp. 3, 153).

[64] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya. “Information modification and particle collisions in distributed computation.” *Chaos*, 20(3) (2010), pp. 037109–037109 (cit. on pp. 3, 13, 22).

[65] J. T. Lizier, M. Prokopenko, and A. Zomaya. “Local information transfer as a spatiotemporal filter for complex systems”. *Physical Review E*, 77(2) (2008), p. 026110 (cit. on pp. 4, 15, 22, 59, 60).

[66] M. Wibral, J. T. Lizier, S. Vögler, V. Priesemann, and R. Galuske. “Local active information storage as a tool to understand distributed neural information processing”. *Frontiers in Neuroinformatics*, 8 (2014) (cit. on pp. 4, 12, 13, 22, 126).

[67] C. Gómez, J. T. Lizier, M. Schaum, et al. “Reduced predictable information in brain signals in autism spectrum disorder.” *Frontiers in Neuroinformatics*, 8 (2014), p. 9 (cit. on pp. 4, 22, 114).

[68] J. T. Lizier and M. Prokopenko. “Differentiating information transfer and causal effect”. *The European Physical Journal B*, 73 (2010), pp. 605–615 (cit. on pp. 4, 16, 22, 23, 56, 61, 62, 101, 118, 148, 150, 151, 161, 207).

[69] I. Csiszár. “Axiomatic characterizations of information measures”. *Entropy*, 10(3) (2008), pp. 261–273 (cit. on p. 6).

[70] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press, 2005 (cit. on p. 7).

[71] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Francisco, CA: Morgan Kaufmann, 1988 (cit. on p. 8).

[72] P. L. Williams and R. D. Beer. “Nonnegative decomposition of multivariate information”. *arXiv preprint: arXiv:1004.2515* (2010) (cit. on pp. 8, 10, 11, 13, 22, 67, 93, 94, 171, 173).

[73] F. Takens. “Detecting Strange Attractors in Turbulence”. In: *Dynamical Systems and Turbulence, Warwick 1980. Lecture Notes in Mathematics*. Heidelberg, Berlin: Springer, 1981, pp. 366–381 (cit. on pp. 10, 30, 125).

[74] L. Faes, G. Nollo, and A. Porta. “Information-based detection of nonlinear Granger causality in multivariate processes via a nonuniform embedding technique”. *Physical Review E*, 83(5) (2011), p. 051112 (cit. on pp. 10, 22, 30, 56, 96, 169, 174).

[75] M. Wibral, N. Pampu, V. Priesemann, et al. “Measuring information-transfer delays”. *PLoS ONE*, 8(2) (2013), e55809 (cit. on pp. 10, 15, 16, 23, 28–30, 36, 43, 44, 48, 54, 61, 67–72, 90, 91, 102, 104, 110–112, 125, 128, 132, 133, 148, 149, 160, 161, 169, 174).

[76] P. L. Williams and R. D. Beer. “Generalized measures of information transfer”. *arXiv preprint: arXiv:1102.1507* (2011) (cit. on pp. 11, 29, 67, 94, 125, 171, 172).

[77] J. P. Crutchfield and D. P. Feldman. “Regularities unseen, randomness observed: levels of entropy convergence”. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 13(1) (2003), pp. 25–54 (cit. on pp. 12, 13).

[78] J. T. Lizier, B. Flecker, and P. L. Williams. “Towards a synergy-based approach to measuring information modification”. *arXiv preprint: arXiv:1303.3440* (2013) (cit. on pp. 13, 22, 171).

[79] V. Griffith and C. Koch. “Quantifying synergistic mutual information”. In: *Guided Self-Organization: Inception*. Berlin, Heilderberg: Springer, 2014, pp. 159–190 (cit. on pp. 13, 94, 171).

[80] M. Harder, C. Salge, and D. Polani. “A bivariate measure of redundant information”. *Physical Review E*, 87(1) (2013), p. 012130 (cit. on pp. 13, 22, 94, 171).

[81] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay. “Quantifying unique information”. *Entropy*, 16(4) (2014), pp. 2161–2183 (cit. on pp. 13, 22, 94, 171).

[82] F. Effenberger. “A Primer on Information Theory with Applications to Neuroscience”. *Computational Medicine in Data Mining and Modeling* (2013), pp. 1–376 (cit. on p. 15).

[83] A. Kraskov, H. Stögbauer, and P. Grassberger. “Estimating mutual information”. *Physical Review E*, 69(6) (2004), p. 066138 (cit. on pp. 15, 24, 27, 30, 31, 33, 35, 97, 128, 129, 161, 209, 211).

[84] S. Khan, S. Bandyopadhyay, A. R. Ganguly, et al. “Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data”. *Physical Review E*, 76(2) (2007), p. 026209 (cit. on pp. 15, 129, 160, 161, 164).

[85] A. Kraskov. “Synchronization and Interdependence measures and their application to the electroencephalogram of epilepsy patients and clustering of data”. PhD thesis. University of Wuppertal, Wuppertal, Germany, 2004 (cit. on pp. 15, 41, 44, 46, 128, 129, 161, 164, 208).

[86] M. Lindner, R. Vicente, V. Priesemann, and M. Wibral. “TRENTOOL: a Matlab open source toolbox to analyse information flow in time series data with transfer entropy”. *BMC Neuroscience*, 12(1) (2011), p. 119 (cit. on pp. 15, 24, 31, 32, 34, 35, 43, 53, 58, 64, 67, 77, 89–91, 98, 110, 112, 127, 129, 131, 144, 162, 163, 176, 178, 179, 208).

[87] J. T. Lizier. “JIDT: an information-theoretic toolkit for studying the dynamics of complex systems”. *Frontiers in Robotics and AI*, 1(11) (2014) (cit. on pp. 15, 126, 127, 131, 145, 163, 179).

[88] R. Vicente and M. Wibral. “Efficient estimation of information transfer”. In: *Directed Information Measures in Neuroscience*. Ed. by M. Wibral, R. Vicente, and J. T. Lizier. Berlin, Heidelberg: Springer, 2014, pp. 37–58 (cit. on pp. 16, 31).

[89] H. Sumioka, Y. Yoshikawa, and M. Asada. “Causality detected by transfer entropy leads acquisition of joint attention”. In: *IEEE 6th International Conference on Development and Learning. ICDL 2007*. 2007, pp. 264–269 (cit. on pp. 16, 150, 207).

[90] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya. “Causality detection based on information-theoretic approaches in time series analysis”. *Physics Reports*, 441(1) (2007), pp. 1–46 (cit. on pp. 16, 150, 161, 207).

[91] V. A. Vakorin, O. A. Krakovska, and A. R. McIntosh. “Confounding effects of indirect connections on causality estimation”. *Journal of Neuroscience Methods*, 184(1) (2009), pp. 152–160 (cit. on pp. 16, 22, 70, 150, 207).

[92] D. Chicharro and A. Ledberg. “When two become one: the limits of causality analysis of brain dynamics”. *PLoS ONE*, 7(3) (2012), e32466 (cit. on pp. 16, 22, 101, 150, 207).

[93] A. M. Turing. “On computable numbers, with an application to the Entscheidungsproblem”. *Proceedings of the London Mathematical Society*, 42(2) (1936), pp. 230–265 (cit. on p. 22).

[94] M. Mitchell. “Computation in Cellular Automata: a selected review”. In: *Non-Standard Computation*. Ed. by T. Gramß, S. Bornholdt, M. Groß, M. Mitchell, and T. Pellizzari. Weinheim: VCH Verlagsgesellschaft, 1998, pp. 95–140 (cit. on p. 22).

[95] J. T. Lizier. *The local information dynamics of distributed computation in complex systems*. Springer Theses Series. Berlin, Heidelberg: Springer, 2013 (cit. on pp. 22, 29, 59).

[96] N. Bertschinger, J. Rauh, E. Olbrich, and J. Jost. “Shared information—new insights and problems in decomposing information in complex systems”. *arXiv preprint: arXiv:1210.5902* (2012) (cit. on p. 22).

[97] V. Griffith and C. Koch. “Quantifying synergistic mutual information”. *arXiv preprint: arXiv:1205.4265* (2012) (cit. on p. 22).

[98] J. T. Lizier. “Measuring the dynamics of information processing on a local scale in time and space”. In: *Directed Information Measures in Neuroscience*. Ed. by M. Wibral, R. Vicente, and J. T. Lizier. Berlin, Heidelberg: Springer, 2014, pp. 161–193 (cit. on p. 22).

[99] S. Dasgupta, F. Wörgötter, and P. Manoonpong. “Information dynamics based self-adaptive reservoir for delay temporal memory tasks”. *Evolving Systems* (2013), pp. 1–15 (cit. on p. 22).

[100] M. Paluš. “Synchronization as adjustment of information rates: detection from bivariate time series”. *Physical Review E*, 63 (2001), p. 046211 (cit. on p. 22).

[101] M. Chávez, J. Martinerie, and M. Le Van Quyen. “Statistical assessment of nonlinear causality: application to epileptic EEG signals.” *Journal of Neuroscience Methods*, 124(2) (2003), pp. 113–28 (cit. on pp. 22, 162).

[102] P. O. Amblard and O. J. Michel. “On directed information theory and Granger causality graphs.” *Journal of Computational Neuroscience*, 30(1) (2011), pp. 7–16 (cit. on p. 22).

[103] L. Barnett, A. Barrett, and A. Seth. “Granger causality and transfer entropy are equivalent for Gaussian variables”. *Physical Review Retters*, 103(23) (2009), p. 238701 (cit. on pp. 22, 60, 69, 95, 150, 159).

[104] A. Buehlmann and G. Deco. “Optimal information transfer in the cortex through synchronization.” *PLoS Computational Biology*, 6(9) (2010), e1000934 (cit. on p. 22).

[105] M. Garofalo, T. Nieus, P. Massobrio, and S. Martinoia. “Evaluation of the performance of information theory-based methods and cross-correlation to estimate the functional connectivity in cortical networks.” *PLoS ONE*, 4(8) (2009), e6482 (cit. on p. 22).

[106] N. Lüdtke, N. K. Logothetis, and S. Panzeri. “Testing methodologies for the nonlinear analysis of causal relationships in neurovascular coupling.” *Magnetic Resonance Imaging*, 28(8) (2010), pp. 1113–1119 (cit. on p. 22).

[107] S. A. Neymotin, K. M. Jacobs, A. A. Fenton, and W. W. Lytton. “Synaptic information transfer in computer models of neocortical columns.” *Journal of Computational Neuroscience*, 30(1) (2011), pp. 69–84 (cit. on p. 22).

[108] S. Sabesan, L. B. Good, K. S. Tsakalis, et al. “Information flow and application to epileptogenic focus localization from intracranial EEG.” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 17(3) (2009), pp. 244–53 (cit. on p. 22).

[109] M. Staniek and K. Lehnertz. “Symbolic transfer entropy: inferring directionality in biosignals”. *Biomedizinische Technik. Biomedical engineering*, 54(6) (2009), pp. 323–8 (cit. on pp. 22, 211).

[110] V. A. Vakorin, B. Misic, O. Kraskovska, and A. R. McIntosh. “Empirical and theoretical aspects of generation and transfer of information in a neuromagnetic source network”. *Frontiers in Systems Neuroscience*, 5 (2011) (cit. on p. 22).

[111] N. C. Pampu, R. Vicente, R. C. Muresan, et al. “Transfer entropy as a tool for reconstructing interaction delays in neural signals”. In: *2013 International Symposium on Signals, Circuits and Systems (ISSCS)*. 2013, pp. 1–4 (cit. on p. 22).

[112] D. Marinazzo, G. Wu, M. Pellicoro, and S. Stramaglia. “Information transfer in the brain: insights from a unified approach”. In: *Directed Information Measures in Neuroscience*. Ed. by M. Wibral, R. Vicente, and J. T. Lizier. Berlin, Heidelberg: Springer, 2014, pp. 87–110 (cit. on p. 22).

[113] L. Faes and A. Porta. “Conditional entropy-based evaluation of information dynamics in physiological systems”. In: *Directed Information Measures in Neuroscience*. Ed. by M. Wibral, R. Vicente, and J. T. Lizier. Berlin, Heidelberg: Springer, 2014, pp. 61–86 (cit. on pp. 22, 127).

[114] L. Faes and G. Nollo. “Bivariate nonlinear prediction to quantify the strength of complex dynamical interactions in short-term cardiovascular variability.” *Medical & Biological Engineering & Computing*, 44(5) (2006), pp. 383–392 (cit. on pp. 22, 56).

[115] L. Faes, G. Nollo, and A. Porta. “Non-uniform multivariate embedding to assess the information transfer in cardiovascular and cardiorespiratory variability series.” *Computers in Biology and Medicine*, 42 (2011), pp. 290–297 (cit. on pp. 22, 23, 56, 132).

[116] O. Kwon and J.-S. Yang. “Information flow between stock indices”. *Europhysics Letters*, 82(6) (2008), p. 68003 (cit. on pp. 22, 56).

[117] J. Kim, K. Gunn, A. Sungbae, K. Young-Kyun, and Y. Sungroh. “Entropy-based analysis and bioinformatics-inspired integration of global economic information transfer”. *PLoS ONE*, 8(1) (2013), e51986 (cit. on p. 22).

[118] N. Ay and D. Polani. “Information flows in causal networks”. *Advances in Complex Systems*, 11(1) (2008), pp. 17–42 (cit. on pp. 22, 62, 101, 150, 175).

[119] S. Stramaglia, G.-R. Wu, M. Pellicoro, and D. Marinazzo. “Expanding the transfer entropy to identify information circuits in complex systems”. *Physical Review E*, 86(6) (2012), p. 066211 (cit. on p. 23).

[120] L. M. A. Bettencourt, G. J. Stephens, M. I. Ham, and G. W. Gross. “Functional structure of cortical neuronal networks grown in vitro”. *Physical Review E*, 75(2) (2007), p. 021915 (cit. on p. 23).

[121] M. Wibral, P. Wollstadt, U. Meyer, et al. “Revisiting Wiener’s principle of causality-interaction-delay reconstruction using transfer entropy and multivariate analysis on delay-weighted graphs”. In: *Engineering in Medicine and Biology Society (EMBC), 2012 Annual International Conference of the IEEE*. 2012, pp. 3676–3679 (cit. on pp. 23, 54, 67, 91).

[122] J. D. Owens, M. Houston, D. Luebke, et al. “GPU Computing”. *Proceedings of the IEEE*, 96(5) (2008), pp. 879–899 (cit. on pp. 24, 37, 39).

[123] A. R. Brodtkorb, T. R. Hagen, and M. L. Sætra. “Graphics processing unit (GPU) programming strategies and trends in GPU computing”. *Journal of Parallel and Distributed Computing*, 73(1) (2013), pp. 4–13 (cit. on pp. 24, 37).

[124] D. Lee, I. Dinov, B. Dong, et al. “CUDA optimization strategies for compute- and memory-bound neuroimaging algorithms”. *Computer Methods and Programs in Biomedicine*, 106(3) (2012), pp. 175–187 (cit. on p. 24).

[125] M. Martínez-Zarzuela, C. Gómez, F. J. Díaz-Pernas, A. Fernández, and R. Hornero. “Cross-Approximate Entropy parallel computation on GPUs for biomedical signal analysis. Application to MEG recordings”. *Comput Methods Programs Biomed*, 112(1) (2013), pp. 189–199 (cit. on p. 24).

[126] E. I. Konstantinidis, C. A. Frantzidis, C. Pappas, and P. D. Bamidis. “Real time emotion aware applications: a case study employing emotion evocative pictures and neuro-physiological sensing enhanced by Graphic Processor Units”. *Computer Methods and Programs in Biomedicine*, 107(1) (2012), pp. 16–27 (cit. on p. 24).

[127] A. S. Arefin, C. Riveros, R. Berretta, and P. Moscato. “GPU-FS-kNN: A software tool for fast and scalable kNN computation using GPUs.” *PLoS ONE*, 7(8) (2012), e44000 (cit. on p. 24).

[128] J. A. Wilson and J. C. Williams. “Massively parallel signal processing using the graphics processing unit for real-time brain-computer interface feature extraction.” *Frontiers in Neuroengineering*, 2 (2009), p. 11 (cit. on p. 24).

[129] D. Chen, L. Wang, G. Ouyang, and X. Li. “Massively parallel neural signal processing on a many-core platform”. *Computing in Science & Engineering*, 13(6) (2011), pp. 42–51 (cit. on p. 24).

[130] Y. Liu, B. Schmidt, W. Liu, and D. L. Maskell. “CUDA-MEME: accelerating motif discovery in biological sequences using CUDA-enabled graphics processing units”. *Pattern Recognition Letters*, 31(14) (2010), pp. 2170–2177 (cit. on p. 24).

[131] C. Merkwirth, U. Parlitz, and W. Lauterborn. “Fast nearest-neighbor searching for nonlinear signal processing”. *Physical Review E*, 62 (2 2000), pp. 2089–2097 (cit. on p. 24).

[132] W. A. Gardner, A. Napolitano, and L. Paura. “Cyclostationarity: Half a century of research”. *Signal Process*, 86(4) (2006), pp. 639–697 (cit. on p. 26).

[133] M. Ragwitz and H. Kantz. “Markov models from data by simple nonlinear time series predictors in delay embedding spaces.” *Physical Review E*, 65(5) (2002), p. 056201 (cit. on pp. 30, 58, 109, 110, 125, 126, 131, 169).

[134] L. F. Kozachenko and N. N. Leonenko. “On statistical estimation of entropy of random vector”. *Problems of Information Transmission*, 23 (1987), pp. 95–101 (cit. on pp. 31, 130).

[135] J. D. Victor. “Binless strategies for estimation of information from neural data”. *Physical Review E*, 72 (2005), p. 051903 (cit. on p. 31).

[136] NVIDIA Corporation. *CUDA toolkit documentation*. Available: http://docs.nvidia.com/cuda. Accessed 7 November 2013. 2013 (cit. on pp. 32, 37, 39).

[137] L. Faes, G. Nollo, and A. Porta. “Compensated transfer entropy as a tool for reliably estimating information transfer in physiological time series”. *Entropy*, 15(1) (2013), pp. 198–219 (cit. on pp. 35, 176).

[138] E. Maris and R. Oostenveld. “Nonparametric statistical testing of EEG- and MEG-data.” *Journal of Neuroscience Methods*, 164(1) (2007), pp. 177–90 (cit. on p. 35).

[139] J. L. Bentley and J. H. Friedman. “Data structures for range searching”. *ACM Computing Surveys*, 11(4) (1979), pp. 397–409 (cit. on p. 37).

[140] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu. “An optimal algorithm for approximate nearest neighbor searching fixed dimensions”. *Journal of the ACM*, 45(6) (1998), pp. 891–923 (cit. on p. 37).

[141] M. Muja and D. G. Lowe. “Fast approximate nearest neighbors with automatic algorithm configuration”. In: *In VISAPP International Conference on Computer Vision Theory and Applications*. 2009, pp. 331–340 (cit. on p. 37).

[142] V. Garcia, E. Debreuve, F. Nielsen, and M. Barlaud. “K-nearest neighbor search: fast GPU-based implementations and application to high-dimensional feature matching.” In: *2010 17th IEEE International Conference on Image Processing (ICIP)*. 2010, pp. 3757–3760 (cit. on p. 37).

[143] N. Sismanis, N. Pitsianis, and X. Sun. “Parallel search of k-nearest neighbors with synchronous operations”. In: *IEEE Conference on High Performance Extreme Computing, HPEC 2012, Waltham, MA, USA, September 10-12, 2012*. 2012, pp. 1–6 (cit. on p. 37).

[144] S. Brown and J. Snoeyink. “GPU nearest neighobr searches using a minimal kd-tree”. In: *GPU Technology Conference (GTC 2010)*. 2010 (cit. on p. 37).

[145] S. L. Amenta, L. C. Simons, J. B. Pakaravoor, et al. “kANN on the GPU with shifted sorting”. In: *Proceedings of the Fourth ACM SIGGRAPH/Eurographics conference on High-Performance Graphics*. Ed. by C. D. Pantaleoni, J. Munkberg, and Jacopo. High Performance Graphics 2012. Paris, France: The Eurographics Association, 2012, pp. 39–47 (cit. on p. 37).

[146] J. Pan and D. Manocha. “Bi-level locality sensitive hashing for k-nearest neighbor computation”. In: *2012 IEEE 28th International Conference on Data Engineering (ICDE)*. 2012, pp. 378–389 (cit. on p. 37).

[147] Khronos OpenCL Working Group and A. Munshi. “The OpenCL Specification Version: 1.0 Document Revision: 48” (2009). `http://www.khronos.org/registry/cl/specs/opencl-1.0.pdf`. Accessed 30 May 2014 (cit. on p. 37).

[148] C. Grützner, P. J. Uhlhaas, E. Genc, et al. “Neuroelectromagnetic correlates of perceptual closure processes”. *Journal of Neuroscience*, 30(24) (2010), pp. 8342–8352 (cit. on pp. 40, 50, 53, 54, 57, 63, 86, 88, 89).

[149] C. Mooney and G. Ferguson. “A new closure test”. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, 5(3) (1951), pp. 129–133 (cit. on pp. 51, 52, 87, 88).

[150] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen. “FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data.” *Computational Intelligence and Neuroscience*, 2011 (2011), pp. 1–9 (cit. on pp. 53, 89, 113, 134, 138).

[151] J. Groß, J. Kujala, M. Hämäläinen, et al. “Dynamic imaging of coherent sources: studying neural interactions in the human brain”. *Proceedings of the National Academy of Sciences*, 98(2) (2001), pp. 694–699 (cit. on pp. 53, 89).

[152] M. J. Brookes, J. Vrba, S. E. Robinson, et al. “Optimising experimental design for MEG beamformer imaging”. *Neuroimage*, 39(4) (2008), pp. 1788–1802 (cit. on pp. 54, 89).

[153] M. Bar, K. S. Kassam, A. S. Ghuman, et al. “Top-down facilitation of visual recognition”. *Proceedings of the National Academy of Sciences*, 103(2) (2006), pp. 449–454 (cit. on pp. 54, 57, 63).

[154] P. Cavanagh. “What’s up in top-down processing”. In: *Representations of Vision: Trends and Tacit Assumptions in Vision Research*. Ed. by A. Gorea. Cambridge: Cambridge University Press, 1991, pp. 295–304 (cit. on pp. 54, 57).

[155] P. F. Verdes. “Assessing causality from multivariate time series”. *Physical Review E*, 72 (2 2005), p. 026222 (cit. on p. 56).

[156] B. Pompe and J. Runge. “Momentary information transfer as a coupling measure of time series”. *Physical Review E*, 83(5) (2011), p. 051122 (cit. on pp. 56, 128).

[157] R. Marschinski and H. Kantz. “Analysing the information flow between financial time series”. *The European Physical Journal B*, 30(2) (2002), pp. 275–281 (cit. on p. 56).

[158] P. Sauseng, W. Klimesch, W. R. Gruber, et al. “Are event-related potential components generated by phase resetting of brain oscillations? A critical discussion”. *Neuroscience*, 146(4) (2007), pp. 1435–1444 (cit. on p. 58).

[159] S. Makeig, S. Debener, J. Onton, and A. Delorme. “Mining event-related brain dynamics.” *Trends in Cognitive Sciences*, 8(5) (2004), pp. 204–210 (cit. on p. 58).

[160] A. S. Shah, S. L. Bressler, K. H. Knuth, et al. “Neural dynamics and the fundamental mechanisms of event-related brain potentials”. *Cerebral Cortex*, 14(5) (2004), pp. 476–483 (cit. on p. 58).

[161] B. W. Jervis, M. J. Nichols, T. E. Johnson, E. Allen, and N. R. Hudson. “A fundamental investigation of the composition of auditory evoked potentials”. *IEEE Transactions on Biomedical Engineering*, (1) (1983), pp. 43–50 (cit. on p. 59).

[162] G. R. Mangun. “Human brain potentials evoked by visual stimuli: induced rhythms or time-locked components?” In: *Induced Rhythms in the Brain*. Ed. by E. Basar and T. H. Bullock. Boston, MA: Birkhauser, 1992, pp. 217–231 (cit. on p. 59).

[163] C. E. Schroeder, M. Steinschneider, D. C. Javitt, et al. “Localization of ERP generators and identification of underlying neural processes”. *Electroencephalography and Clinical Neurophysiology Suppl.* 44 (1995), pp. 55–75 (cit. on p. 59).

[164] B. M. A. Sayers, H. A. Beagley, and W. R. Henshall. “The mechanism of auditory evoked EEG responses.” *Nature*, 247 (1974), pp. 481–483 (cit. on p. 59).

[165] S Makeig, M Westerfield, T.-P. Jung, et al. “Dynamic brain sources of visual evoked responses”. *Science*, 295(5555) (2002), pp. 690–694 (cit. on p. 59).

[166] B. H. Jansen, G. Agarwal, A. Hegde, and N. N. Boutros. “Phase synchronization of the ongoing EEG and auditory EP generation”. *Clinical Neurophysiology*, 114(1) (2003), pp. 79–85 (cit. on p. 59).

[167] W. Klimesch, B. Schack, M. Schabus, et al. “Phase-locked alpha and theta oscillations generate the P1–N1 complex and are related to memory performance”. *Cognitive Brain Research*, 19(3) (2004), pp. 302–316 (cit. on p. 59).

[168] G. Turi, S. Gotthardt, W. Singer, et al. “Quantifying additive evoked contributions to the event-related potential”. *Neuroimage*, 59 (2012), pp. 2607–2624 (cit. on p. 59).

[169] E. Möller, B. Schack, M. Arnold, and H. Witte. “Instantaneous multivariate EEG coherence analysis by means of adaptive high-dimensional autoregressive models”. *Journal of Neuroscience Methods*, 105(2) (2001), pp. 143–158 (cit. on p. 61).

[170] M. Ding, S. L. Bressler, W. Yang, and H. Liang. “Short-window spectral analysis of cortical event-related potentials by adaptive multivariate autoregressive modeling: data preprocessing, model validation, and variability assessment”. *Biological Cybernetics*, 83(1) (2000), pp. 35–45 (cit. on p. 61).

[171] W. Hesse, E. Möller, M. Arnold, and B. Schack. “The use of time-variant EEG Granger causality for inspecting directed interdependencies of neural assemblies”. *Journal of Neuroscience Methods*, 124(1) (2003), pp. 27–44 (cit. on p. 61).

[172] L. Leistritz, W. Hesse, M. Arnold, and H. Witte. “Development of interaction measures based on adaptive non-linear time series analysis of biomedical signals.” *Biomedical Engineering/Biomedizinische Technik*, 51(2) (2006), pp. 64–69 (cit. on p. 61).

[173] M. Wibral, G. Turi, D. E. J. Linden, J. Kaiser, and C. Bledowski. “Decomposition of working memory-related scalp ERPs: crossvalidation of fMRI-constrained source analysis and ICA.” *International Journal of Psychophysiology*, 67(3) (2008), pp. 200–211 (cit. on p. 61).

[174] R. G. Andrzejak, A. Ledberg, and G Deco. “Detecting event-related time-dependent directional couplings”. *New Journal of Physics*, 8(1) (2006), p. 6 (cit. on pp. 61, 62).

[175] S. P. Strong, R. R. de Ruyter van Steveninck, W. Bialek, and R. Koberle. “On the application of information theory to neural spike trains”. *Pacific Symposium on Biocomputing* (1998), pp. 621–632 (cit. on p. 62).

[176] S. S. Georgieva, J. T. Todd, R. Peeters, and G. A. Orban. “The extraction of 3D shape from texture and shading in the human brain”. *Cerebral Cortex*, 18 (2008), pp. 2416–2438 (cit. on p. 63).

[177] N. Kanwisher, F. Tong, and K. Nakayama. "The effect of face inversion on the human fusiform face area". *Cognition*, 68(1) (1998), B1–B11 (cit. on p. 63).

[178] T. J. Andrews and D. Schluppeck. "Neural responses to Mooney images reveal a modular representation of faces in human visual cortex". *Neuroimage*, 21(1) (2004), pp. 91–98 (cit. on p. 63).

[179] T. J. McKeeff and F. Tong. "The timing of perceptual decisions for ambiguous face stimuli in the human ventral visual cortex". *Cerebral Cortex*, 17(3) (2007), pp. 669–678 (cit. on p. 63).

[180] E. Bullmore and O. Sporns. "Complex brain networks: graph theoretical analysis of structural and functional systems". *Nature Reviews Neuroscience*, 10(3) (2009), pp. 186–198 (cit. on pp. 66, 68).

[181] M. Rubinov and O. Sporns. "Complex network measures of brain connectivity: uses and interpretations". *Neuroimage*, 52(3) (2010), pp. 1059–1069 (cit. on p. 66).

[182] C. J. Stam, P Tewarie, E Van Dellen, et al. "The trees and the forest: characterization of complex brain networks with minimum spanning trees". *International Journal of Psychophysiology*, 92(3) (2014), pp. 129–138 (cit. on p. 66).

[183] P. Tewarie, E. Van Dellen, A. Hillebrand, and C. J. Stam. "The minimum spanning tree: an unbiased method for brain network analysis". *Neuroimage*, 104 (2015), pp. 177–188 (cit. on p. 66).

[184] L. Amaral and J. Ottino. "Complex networks". *The European Physical Journal B - Condensed Matter and Complex Systems*, 38(2) (2004), pp. 147–162 (cit. on p. 66).

[185] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. "Complex networks: structure and dynamics". *Physics Reports*, 424 (2006), pp. 175–308 (cit. on pp. 66, 92).

[186] C. Stam and J. Reijneveld. "Graph theoretical analysis of complex networks in the brain". *Nonlinear Biomedical Physics*, 1(1) (2007), p. 3 (cit. on pp. 66, 82).

[187] K. Börner, S. Sanyal, and A. Vespignani. "Network science". *Annual Review of Information Science and Technology*, 41(1) (2008), pp. 537–607 (cit. on p. 66).

[188] O. Sporns. "From simple graphs to the connectome: networks in neuroimaging." *Neuroimage*, 62(2) (2012), pp. 881–886 (cit. on p. 66).

[189] M. Garey and D. Johnson. *Computers and Intractability*. Vol. 174. San Francisco, CA: Freeman, 1979 (cit. on p. 66).

[190] D. Bassett and E. Bullmore. "Small-world brain networks". *The Neuroscientist*, 12(6) (2006), pp. 512–523 (cit. on pp. 68, 82).

[191] Y. Chen, G. Rangarajan, J. Feng, and M. Ding. "Analyzing multiple nonlinear time series with extended Granger causality". *Physics Letters A*, 324(1) (2004), pp. 26–35 (cit. on p. 70).

[192] K. Blinowska. "Review of the methods of determination of directed connectivity from multichannel data". *Medical and Biological Engineering and Computing*, 49(5) (2011), pp. 521–529 (cit. on p. 70).

[193] R. Diestel. *Graph Theory*. 4th ed. New York, NY: Springer, 2010 (cit. on p. 71).

[194] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. 3rd ed. Cambridge, MA: The MIT Press, 2009 (cit. on pp. 76, 77, 81).

[195] D. Watts and S. Strogatz. "Collective dynamics of small-world networks". *Nature*, 393(6684) (1998), pp. 440–442 (cit. on pp. 82–84).

[196] A. Barabási and R. Albert. “Emergence of scaling in random networks”. *Science*, 286(5439) (1999), pp. 509–512 (cit. on pp. 82, 84).

[197] P. Erdős and A. Rényi. “On random graphs I”. *Publicationes Mathematicae Debrecen*, 6 (1959), pp. 290–297 (cit. on p. 82).

[198] E. N. Gilbert. “Random graphs”. *Annals of Mathematical Statistics*, 30 (1959), pp. 1141–1144 (cit. on pp. 82, 84).

[199] O. Sporns, D. Chialvo, M. Kaiser, and C. Hilgetag. “Organization, development and function of complex brain networks”. *Trends in Cognitive Sciences*, 8(9) (2004), pp. 418–425 (cit. on p. 82).

[200] J. Power, D. Fair, B. Schlaggar, and S. Petersen. “The development of human functional brain networks”. *Neuron*, 67(5) (2010), pp. 735–748 (cit. on p. 82).

[201] C. J. Stam. “Characterization of anatomical and functional connectivity in the brain: a complex networks perspective”. *International Journal of Psychophysiology*, 77(3) (2010), pp. 186–194 (cit. on p. 82).

[202] R. Albert and A.-L. Barabási. “Statistical mechanics of complex networks”. *Reviews of Modern Physics*, 74(1) (2002), pp. 47–97 (cit. on p. 84).

[203] V. Batagelj and U. Brandes. “Efficient generation of large random networks”. *Physical Review E*, 71(3) (2005), p. 036113 (cit. on p. 84).

[204] B. Bollobás, O. Riordan, J. Spencer, G. Tusnády, et al. “The degree sequence of a scale-free random graph process”. *Random Structures & Algorithms*, 18(3) (2001), pp. 279–290 (cit. on p. 84).

[205] B. Bollobás and O. Riordan. “The diameter of a scale-free random graph”. *Combinatorica*, 24(1) (2004), pp. 5–34 (cit. on p. 84).

[206] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Vol. 8. Cambridge: Cambridge University Press, 1994 (cit. on p. 84).

[207] R. C. Oldfield. “The assessment and analysis of handedness: the Edinburgh inventory”. *Neuropsychologia*, 9(1) (1971), pp. 97–113 (cit. on p. 86).

[208] P. Wollstadt, M. Martínez-Zarzuela, R. Vicente, F. J. Díaz-Pernas, and M. Wibral. “Efficient transfer entropy analysis of non-stationary neural time series”. *PLoS ONE*, 9(7) (2014), e102833 (cit. on pp. 90, 129, 144, 145).

[209] A. Kulesa, M. Krzywinski, P. Blainey, and N. Altman. “Points of significance: sampling distributions and the bootstrap”. *Nature Methods*, 12 (2015), pp. 477–478 (cit. on p. 90).

[210] S. Laughlin and T. Sejnowski. “Communication in neuronal networks”. *Science*, 301(5641) (2003), pp. 1870–1874 (cit. on p. 92).

[211] R. Kötter and F. T. Sommer. “Global relationship between anatomical connectivity and activity propagation in the cerebral cortex.” *Philosophical Transactions of the Royal Society B: Biological Sciences*, 355(1393) (2000), pp. 127–134 (cit. on p. 92).

[212] D. B. Chklovskii and A. A. Koulakov. “Maps in the brain: what can we learn from them?” *Annual Review of Neuroscience*, 27 (2004), pp. 369–392 (cit. on p. 92).

[213] C. Cherniak. “Neural component placement.” *Trends in Neuroscience*, 18(12) (1995), pp. 522–527 (cit. on p. 92).

[214] K. J. Friston, L. Harrison, and W. Penny. “Dynamic causal modelling”. *Neuroimage*, 19(4) (2003), pp. 1273–1302 (cit. on pp. 93, 149, 176).

[215] O. David, S. J. Kiebel, L. M. Harrison, et al. “Dynamic causal modeling of evoked responses in EEG and MEG.” *Neuroimage*, 30(4) (2006), pp. 1255–72 (cit. on pp. 93, 176).

[216] S. J. Kiebel, O. David, and K. J. Friston. “Dynamic causal modelling of evoked responses in EEG/MEG with lead field parameterization”. *Neuroimage*, 30(4) (2006), pp. 1273–1284 (cit. on p. 93).

[217] D. Marinazzo, M. Pellicoro, and S. Stramaglia. “Causal information approach to partial conditioning in multivariate data sets”. *Computational and Mathematical Methods in Medicine*, 2012 (2012) (cit. on pp. 93, 96, 97, 174, 175).

[218] S. Stramaglia, J. M. Cortes, and D. Marinazzo. “Synergy and redundancy in the Granger causal analysis of dynamical networks”. *New Journal of Physics*, 16 (2014), p. 105003 (cit. on pp. 96, 175).

[219] S. Dehaene and J.-P. Changeux. “Experimental and theoretical approaches to conscious processing”. *Neuron*, 70(2) (2011), pp. 200–227 (cit. on p. 100).

[220] G. Tononi. “An information integration theory of consciousness”. *BMC Neuroscience*, 5(42) (2004) (cit. on p. 100).

[221] A. G. Hudetz. “Suppressing consciousness: mechanisms of general anesthesia”. *Seminars in Anesthesia*, 25(4) (2006), pp. 196–204 (cit. on pp. 100, 120, 147).

[222] M. T. Alkire, A. G. Hudetz, and G. Tononi. “Consciousness and anesthesia”. *Science*, 322(5903) (2008), pp. 876–880 (cit. on pp. 100, 117, 147).

[223] M. D. Krasowski and N. L. Harrison. “General anaesthetic actions on ligand-gated ion channels”. *Cellular and Molecular Life Sciences*, 55(10) (1999), pp. 1278–1303 (cit. on pp. 101, 119, 211).

[224] P. Wollstadt, K. K. Sellers, A. Hutt, F. Fröhlich, and M. Wibral. “Anesthesia-related changes in information transfer may be caused by reduction in local information generation”. In: *Engineering in Medicine and Biology Society (EMBC), 2015 37$^{th}$ Annual International Conference of the IEEE*. IEEE. 2015, pp. 4045–4048 (cit. on p. 102).

[225] I. Nemenman, F. Shafee, and W. Bialek. “Entropy and inference, revisited”. In: *Advances in Neural Information Processing Systems 14*. Cambridge, MA: MIT Press, 2002, pp. 471–478 (cit. on pp. 103, 105, 130, 163, 164, 211).

[226] I. Nemenman, F. Shafee, and R. R. van Steveninck. “Entropy and information in neural spike trains: progress on the sampling problem”. *Physical Reviews E*, 69 (2004), p. 056111 (cit. on pp. 103, 105, 130, 163, 164, 211).

[227] E. Archer, I. M. Park, and J. W. Pillow. “Bayesian entropy estimation for countable discrete distributions”. *Journal of Machine Learning Research*, 15 (2014), pp. 2833–2868 (cit. on pp. 107, 130).

[228] L. Barnett and A. K. Seth. “Behaviour of Granger causality under filtering: theoretical invariance and practical application”. *Journal of Neuroscience Methods*, 201(2) (2011), pp. 404–419 (cit. on p. 112).

[229] E. Florin, J. Gross, J. Pfeifer, G. R. Fink, and L. Timmermann. “The effect of filtering on Granger causality based multivariate causality measures”. *Neuroimage*, 50(2) (2010), pp. 577–588 (cit. on p. 112).

[230] P. L. Purdon, E. T. Pierce, E. A. Mukamel, et al. “Electroencephalogram signatures of loss and recovery of consciousness from propofol”. *Proceedings of the National Academy of Sciences*, 110(12) (2013), E1142–E1151 (cit. on p. 117).

[231] D. J. Felleman and D. C. Van Essen. “Distributed hierarchical processing in the primate cerebral cortex”. *Cerebral Cortex*, 1(1) (1991), pp. 1–47 (cit. on p. 119).

[232] P.-A. Salin and J. Bullier. “Corticocortical connections in the visual system: structure and function”. *Physiological Reviews*, 75(1) (1995), pp. 107–154 (cit. on p. 119).

[233] R. Tomioka, K. Okamoto, T. Furuta, et al. “Demonstration of long-range GABAergic connections distributed throughout the mouse neocortex”. *European Journal of Neuroscience*, 21(6) (2005), pp. 1587–1600 (cit. on p. 119).

[234] J. Hohwy. *The Predictive Mind*. New York, NY: Oxford University Press, 2013 (cit. on pp. 119, 120).

[235] J. Hawkins and S. Blakeslee. *On Intelligence*. New York, NY: Henry Holt and Company, 2007 (cit. on pp. 119, 154, 179).

[236] A. M. Bastos, W. M. Usrey, R. A. Adams, et al. “Canonical microcircuits for predictive coding”. *Neuron*, 76(4) (2012), pp. 695–711 (cit. on pp. 120, 154, 155).

[237] G. A. Mashour. “Top-down mechanisms of anesthetic-induced unconsciousness”. *Frontiers in Systems Neuroscience*, 8(115) (2014) (cit. on pp. 120, 147).

[238] A. Brodski, G.-F. Paasch, S. Helbling, and M. Wibral. “The faces of predictive coding”. *Journal of Neuroscience*, 35(24) (2015), pp. 8997–9006 (cit. on pp. 120, 134, 156).

[239] R. E. Spinney, M. Prokopenko, and J. T. Lizier. “Transfer entropy in continuous time, with applications to jump and neural spiking processes”. *arXiv preprint: arXiv:1610.08192v1* (2016) (cit. on p. 121).

[240] F. Fröhlich and D. A. McCormick. “Endogenous electric fields may guide neocortical network activity”. *Neuron*, 67(1) (2010), pp. 129–143 (cit. on p. 122).

[241] T. Tao. “Szemerédi’s regularity lemma revisited”. *arXiv preprint, arXiv:math/0504472* (2005) (cit. on p. 122).

[242] F. Roux, M. Wibral, H. M. Mohr, W. Singer, and P. J. Uhlhaas. “Gamma-band activity in human prefrontal cortex codes for the number of relevant items maintained in working memory”. *Journal of Neuroscience*, 32(36) (2012), pp. 12411–12420 (cit. on p. 122).

[243] K. K. Sellers, D. V. Bennett, A. Hutt, and F. Fröhlich. “Anesthesia differentially modulates spontaneous network dynamics by cortical area and layer”. *Journal of Neurophysiology*, 110(12) (2013), pp. 2739–2751 (cit. on p. 123).

[244] K. K. Sellers, D. V. Bennett, A. Hutt, J. Williams, and F. Fröhlich. “Awake versus anesthetized: layer-specific sensory processing in visual cortex and functional connectivity between cortical areas”. *Journal of Clinical Neurophysiology*, 113 (2015), pp. 3798–3815 (cit. on pp. 123, 138).

[245] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. New York, NY: Wiley, 2006 (cit. on pp. 124, 126).

[246] M. Staniek and K. Lehnertz. “Symbolic transfer entropy”. *Physical Review Letters*, 100 (2008), p. 158101 (cit. on p. 128).

[247] S. Frenzel and B. Pompe. “Partial mutual information for coupling analysis of multivariate time series.” *Physical Review Letters*, 99(20) (2007), p. 204101 (cit. on pp. 128, 129).

[248] D. H. Wolpert and D. R. Wolf. “Estimating functions of probability distributions from a finite set of samples”. *Physical Review E*, 52(6) (1995), p. 6841 (cit. on p. 130).

[249] M. Anderson and C. T. Braak. “Permutation tests for multi-factorial analysis of variance”. *Journal of Statistical Computation and Simulation*, 73(2) (2003), pp. 85–113 (cit. on p. 134).

[250] J. Suckling and E. Bullmore. “Permutation tests for factorially designed neuroimaging experiments”. *Human Brain Mapping*, 22(3) (2004), pp. 193–205 (cit. on p. 134).

[251] S. Helbling. “Advances in MEG methods and their applications to investigate auditory perception”. PhD thesis. Goethe-University, Frankfurt, Germany, 2015 (cit. on p. 134).

[252] E. Aarts, M. Verhage, J. V. Veenvliet, C. V. Dolan, and S. van der Sluis. “A solution to dependency: using multilevel analysis to accommodate nested data”. *Nature Neuroscience*, 17(4) (2014), pp. 491–496 (cit. on pp. 135, 138).

[253] L. Fahrmeir, T. Kneib, and S. Lang. *Regression*. Berlin, Heidelberg: Springer, 2007 (cit. on p. 135).

[254] R Development Core Team. “R: a language and environment for statistical computing”. *R Foundation for Statistical Computing* (2008). `http://www.R-project.org`. Accessed 27 December 2016 (cit. on p. 136).

[255] D. Bates, M. Mächler, B. Bolker, and S. Walker. “Fitting linear mixed-effects models using lme4”. *Journal of Statistical Software*, 67(1) (2015), pp. 1–48 (cit. on p. 136).

[256] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily. “Random effects structure for confirmatory hypothesis testing: keep it maximal”. *Journal of Memory and Language*, 68(3) (2013), pp. 255–278 (cit. on p. 136).

[257] M. Wibral, R. Vicente, and J. T. Lizier, eds. *Directed Information Measures in Neuroscience*. Berlin, Heidelberg: Springer, 2014 (cit. on pp. 147, 160).

[258] M. Boly, M. Massimini, M. I. Garrido, et al. “Brain connectivity in disorders of consciousness”. *Brain Connectivity*, 2(1) (2012), pp. 1–10 (cit. on p. 147).

[259] U. Lee, S. Blain-Moraes, and G. A. Mashour. “Assessing levels of consciousness with symbolic analysis”. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 373(2034) (2015), p. 20140117 (cit. on pp. 147, 211).

[260] H. Hinrichs, T. Noesselt, and H.-J. Heinze. “Directed information flow—A model free measure to analyze causal interactions in event related EEG-MEG-experiments”. *Human Brain Mapping*, 29(2) (2008), pp. 193–206 (cit. on p. 147).

[261] D. Gencaga, K. H. Knuth, and W. B. Rossow. “A recipe for the estimation of information flow in a dynamical system”. *Entropy*, 17(1) (2015), pp. 438–470 (cit. on pp. 148, 161).

[262] R. G. James, N. Barnett, and J. P. Crutchfield. “Information flows? A critique of transfer entropies”. *Physical Review Letters*, 116 (2016), p. 238701 (cit. on pp. 148, 172).

[263] N. Perret and G. Longo. “Reductionist perspectives and the notion of information”. *Progress in Biophysics and Molecular Biology* (2016), pp. 3–7 (cit. on p. 148).

[264] J. S. Chan, M. Wibral, P. Wollstadt, et al. “Predict age through multisensory integration”. *In preparation* (2016) (cit. on pp. 149, 176, 178).

[265] N. Wiener. “The theory of prediction.” In: *Modern Mathematics for the Engineer*. Ed. by E. F. Beckmann. New York, NY: McGraw-Hill, 1956 (cit. on pp. 149, 161).

[266] D. Materassi. “Norbert Wiener’s legacy in the study and inference of causation”. In: *2014 IEEE Conference on Norbert Wiener in the 21st Century (21CW)*. IEEE. 2014, pp. 1–6 (cit. on p. 149).

[267] C. W. J. Granger. “Investigating causal relations by econometric models and cross-spectral methods”. *Econometrica: Journal of the Econometric Society* (1969), pp. 424–438 (cit. on p. 149).

[268] M. Lungarella, K. Ishiguro, Y. Kuniyoshi, and N. Otsu. “Methods for quantifying the causal structure of bivariate time series”. *International Journal of Bifurcation and Chaos*, 17(03) (2007), pp. 903–921 (cit. on pp. 150, 160).

[269] G. Van Dijck, J. Van Vaerenbergh, and M. M. Van Hulle. “Information theoretic derivations for causality detection: application to human gait”. In: *Proceedings of the 17th International Conference on Artificial Neural Networks vol:17*. Springer. 2007, pp. 159–168 (cit. on p. 150).

[270] J. Pearl. *Causality: models, reasoning, and inference*. Cambridge: Cambridge University Press, 2000 (cit. on p. 150).

[271] P. A. Valdes-Sosa, A. Roebroeck, J. Daunizeau, and K. Friston. “Effective connectivity: Influence, causality and biophysical modeling.” *Neuroimage*, 58(2) (2011), pp. 339–361 (cit. on p. 150).

[272] D. Battaglia. “Function follows dynamics: state-dependency of directed functional influence”. In: *Directed information measures in neuroscience*. Ed. by M. Wibral, R. Vicente, and J. T. Lizier. Berlin, Heidelberg: Springer, 2014, pp. 111–136 (cit. on p. 150).

[273] D. Zipser, B. Kehoe, G. Littlewort, and J. Fuster. “A spiking network model of short-term active memory.” *The Journal of Neuroscience*, 13(8) (1993), pp. 3406–3420 (cit. on p. 151).

[274] S. Panzeri, R. Senatore, M. A. Montemurro, and R. S. Petersen. “Correcting for the sampling bias problem in spike train information measures.” *Journal of Neurophysiology*, 98(3) (2007), pp. 1064–72 (cit. on pp. 153, 162, 163).

[275] M. Wibral, V. Priesemann, J. W. Kay, J. T. Lizier, and W. A. Phillips. “Partial information decomposition as a unified approach to the specification of neural goal functions”. *Brain and Cognition* (2015) (cit. on p. 153).

[276] A. Fairhall, E. Shea-Brown, and A. Barreiro. “Information theoretic approaches to understanding circuit function”. *Current Opinion in Neurobiology*, 22(4) (2012), pp. 653–659 (cit. on p. 153).

[277] J. T. Lizier, S. Pritam, and M. Prokopenko. “Information dynamics in small-world boolean networks.” *Artif Life*, 17(4) (2011), pp. 293–314 (cit. on p. 154).

[278] M. Wibral, D. Rathbun, W. M. Usrey, A. M. Bastos, and P. Wollstadt. “One man’s prediction is another man’s error - Quantifying predictive coding at the retino-geniculate synapse independent of the observer’s assumptions”. *Neuroscience Meeting Planner. Washington, DC: Society for Neuroscience, 2015* (2015). Program No. 543.16/DD22. Accessed 09.09.2016 (cit. on pp. 154, 156, 157, 168).

[279] J. Hohwy. “The hypothesis testing brain: some philosophical applications”. In: *ASCS09: Proceedings of the 9th Conference of the Australasian Society for Cognitive Science*. Ed. by W. Christensen, E. Schier, and J. Sutton. Sydney: Macquarie Centre for Cognitive Science, 2010, pp. 135–144 (cit. on pp. 154, 179).

[280] D. Mumford. “On the computational architecture of the neocortex”. *Biological Cybernetics*, 66 (1992), pp. 241–251 (cit. on p. 155).

[281] R. P. N. Rao and D. H. Ballard. “Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects.” *Nature Neuroscience*, 2(1) (1999), pp. 79–87 (cit. on p. 155).

[282] J. Hohwy, A. Roepstorff, and K. J. Friston. “Predictive coding explains binocular rivalry: an epistemological review”. *Cognition*, 108(3) (2008), pp. 687–701 (cit. on p. 155).

[283] R. Desimone and J. Duncan. “Neural Mechanisms of selective visual attention”. *Annual Review of Neuroscience*, 18 (1995), pp. 193–222 (cit. on p. 155).

[284] S. Grossberg. “How does a brain build a cognitive code”. *Psychological Review*, 87 (1980), pp. 1–51 (cit. on p. 155).

[285] S. Grossberg. “Adaptive Resonance Theory: How a brain learns to consciously attend, learn, and recognize a changing world”. *Neural Networks*, 37 (2013), pp. 1–47 (cit. on p. 155).

[286] M. W. Spratling. “Reconciling predictive coding and biased competition models of cortical function.” *Frontiers in Computational Neuroscience*, 2 (2008), p. 4 (cit. on p. 155).

[287] K. Kveraga, A. S. Ghuman, and M. Bar. “Top-down predictions in the cognitive brain”. *Brain and Cognition*, 65(2) (2007), pp. 145–168 (cit. on p. 155).

[288] C. Summerfield and T. Egner. “Expectation (and attention) in visual cognition”. *Trends in Cognitive Sciences*, 13(9) (2009), pp. 403–409 (cit. on p. 155).

[289] C. Summerfield and F. P. de Lange. “Expectation in perceptual decision making: neural and computational mechanisms”. *Nature Reviews Neuroscience* (2014), pp. 1–12 (cit. on p. 155).

[290] S. O. Murray, P. Schrater, and D. Kersten. “Perceptual grouping and the interactions between visual cortical areas”. *Neural Networks*, 17(5-6) (2004), pp. 695–705 (cit. on p. 155).

[291] L. De-Wit, B. Machilsen, and T. Putzeys. “Predictive coding and the neural response to predictable stimuli.” *The Journal of Neuroscience*, 30(26) (2010), pp. 8702–8703 (cit. on p. 156).

[292] A. Alink, C. M. Schwiedrzik, A. Kohler, W. Singer, and L. Muckli. “Stimulus predictability reduces responses in primary visual cortex”. *The Journal of Neuroscience*, 30(8) (2010), pp. 2960–2966 (cit. on p. 156).

[293] M. Wibral, J. T. Lizier, and V. Priesemann. “How to measure local active information storage in neural systems”. *2014 8th Conference of the European Study Group on Cardiovascular Oscillations (ESGCO)*, (Esgco) (2014), pp. 131–132 (cit. on pp. 157, 158).

[294] S. Deneve. “Bayesian spiking neurons I: inference”. *Neural computation*, 20(1) (2008), pp. 91–117 (cit. on p. 157).

[295] D. Rich, F. Cazettes, Y. Wang, J. L. Peña, and B. J. Fischer. “Neural representation of probabilities for Bayesian inference”. *Journal of Computational Neuroscience*, 38(2) (2015), pp. 315–323 (cit. on p. 157).

[296] A. M. Bastos and J.-M. Schoffelen. “A tutorial review of functional connectivity analysis methods and their interpretational pitfalls”. *Frontiers in Systems Neuroscience*, 9 (2015), pp. 1–23 (cit. on p. 159).

[297] E. Pereda, R. Quiroga, and J. Bhattacharya. “Nonlinear multivariate analysis of neurophysiological signals”. *Progress in Neurobiology*, 77(1-2) (2005), pp. 1–37 (cit. on p. 159).

[298] V. Sakkalis. “Review of advanced techniques for the estimation of brain connectivity measured with EEG/MEG”. *Computers in Biology and Medicine*, 41(12) (2011), pp. 1110–1117 (cit. on p. 159).

[299] R. E. Greenblatt, M. E. Pflieger, and A. E. Ossadtchi. “Connectivity measures applied to human brain electrophysiological data”. *Journal of Neuroscience Methods*, 207(1) (2012), pp. 1–16 (cit. on p. 159).

[300] J. Jeong, J. C. Gore, and B. S. Peterson. “Mutual information analysis of the EEG in patients with Alzheimer’s disease”. *Clinical Neurophysiology*, 112(5) (2001), pp. 827–835 (cit. on p. 159).

[301] S. H. Na, S.-H. Jin, S. Y. Kim, and B.-J. Ham. “EEG in schizophrenic patients: mutual information analysis”. *Clinical Neurophysiology*, 113(12) (2002), pp. 1954–1960 (cit. on p. 159).

[302] J. F. Alonso, J. Poza, M. Á. Mañanas, et al. “MEG connectivity analysis in patients with Alzheimer’s disease using cross mutual information and spectral coherence”. *Annals of Biomedical Engineering*, 39(1) (2011), pp. 524–536 (cit. on p. 159).

[303] R. Kleeman. “Information flow in ensemble weather predictions”. *Bulletin of the American Meteorological Society*, 88(4) (2007), pp. 491–492 (cit. on p. 159).

[304] M. J. Kaminski and K. J. Blinowska. “A new method of the description of the information flow in the brain structures”. *Biological Cybernetics*, 65(3) (1991), pp. 203–210 (cit. on p. 160).

[305] K. Sameshima and L. A. Baccalá. “Using partial directed coherence to describe neuronal ensemble interactions”. *Journal of Neuroscience Methods*, 94(1) (1999), pp. 93–103 (cit. on p. 160).

[306] G. Sugihara, R. May, H. Ye, et al. “Detecting causality in complex ecosystems”. *Science*, 338(6106) (2012), pp. 496–500 (cit. on p. 160).

[307] J. Schumacher, T. Wunderle, P. Fries, F. Jäkel, and G. Pipa. “A statistical framework to infer delay and direction of information flow frommeasurements of complex systems”. *Neural Computation*, 27 (2015), pp. 1555–1608 (cit. on p. 160).

[308] J. Bhattacharya, E. Pereda, and H. Petsche. “Effective detection of coupling in short and noisy bivariate data”. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 33(1) (2003), pp. 85–95 (cit. on p. 160).

[309] J. Garland, R. G. James, and E. Bradley. “Leveraging information storage to select forecastoptimal parameters for delaycoordinate reconstructions”. *Physical Reviews E*, 93(2) (2016), p. 022221 (cit. on p. 161).

[310] A. Antos and I. Kontoyiannis. “Convergence properties of functional estimates for discrete distributions”. *Random Structures and Algorithms*, 19(3-4) (2001), pp. 163–193 (cit. on p. 162).

[311] L. Paninski. “Estimation of Entropy and Mutual Information”. *Neural Computation*, 15(6) (2003), pp. 1191–1253 (cit. on p. 162).

[312] A. Khadem, G. A. Hossein-Zadeh, and A. Khorrami. “Long-range reduced predictive information transfers of autistic youths in EEG sensor-space during face processing”. *Brain Topography*, 29(2) (2016), pp. 283–295 (cit. on p. 163).

[313] A. Montalto, L. Faes, and D. Marinazzo. “MuTE : a MATLAB toolbox to compare established and novel estimators of the multivariate transfer entropy”. *PLoS ONE*, 9(10) (2014), pp. 1–18 (cit. on pp. 163, 174).

[314] S. Gao, G. V. Steeg, and A. Galstyan. “Efficient estimation of mutual information for strongly dependent variables”. In: *18th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2015, pp. 277–286 (cit. on pp. 163, 164).

[315] S. Haufe and A. Ewald. “A simulation framework for benchmarking EEG-based brain connectivity estimation methodologies”. *Brain Topography (in press)* (2016) (cit. on p. 164).

[316] A. Brodski-Guerniero, G.-F. Paasch, P. Wollstadt, et al. “Activating task relevant prior knowledge increases active information storage in content specific brain areas”. *bioRxiv preprint: bioRxiv 089300* (2016) (cit. on pp. 165, 179).

[317] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. “When is "nearest neighbor" meaningful?” In: *Lecture Notes in Computer Science*. Vol. 1540. 1999, pp. 217–235 (cit. on p. 166).

[318] C. M. Bishop. *Pattern Recognition and Machine Learning*. New York, NY: Springer, 2006 (cit. on p. 166).

[319] L. Faes, G. Nollo, S. Erla, et al. “Detecting nonlinear causal interactions between dynamical systems by non-uniform embedding of multiple time series.” *Conf Proc IEEE Eng Med Biol Soc*, 2010 (2010), pp. 102–105 (cit. on p. 166).

[320] P. Wollstadt, M. Wibral, J. T. Lizier, et al. “IDTxl”. *GitHub repository* (2016). https://github.com/pwollstadt/IDTxl. Accessed 14 November 2016 (cit. on p. 174).

[321] M. Lindner, R. Vicente, V. Priesemann, et al. “TRENTOOL3”. *GitHub repository* (2015). https://github.com/trentool/TRENTOOL3. Accessed 14 November 2016 (cit. on pp. 176, 178, 179).

[322] K. E. Stephan, W. D. Penny, R. J. Moran, et al. “Ten simple rules for dynamic causal modeling.” *Neuroimage*, 49(4) (2010), pp. 3099–3109 (cit. on pp. 176–178).

[323] J. Daunizeau, O. David, and K. E. Stephan. “Dynamic causal modelling: a critical review of the biophysical and statistical foundations”. *Neuroimage*, 58(2) (2011), pp. 312–322 (cit. on p. 176).

[324] K. J. Friston, B. Li, J. Daunizeau, and K. E. Stephan. “Network discovery with DCM”. *Neuroimage*, 56(3) (2011), pp. 1202–1221 (cit. on p. 176).

[325] O. David, L. Harrison, and K. J. Friston. “Modelling event-related responses in the brain”. *Neuroimage*, 25(3) (2005), pp. 756–770 (cit. on p. 177).

[326] G. Lohmann, K. Erfurth, K. Müller, and R. Turner. “Critical comments on dynamic causal modelling”. *Neuroimage*, 59(3) (2012), pp. 2322–2329 (cit. on p. 177).

[327] K. J. Friston, R. Moran, and A. K. Seth. “Analysing connectivity with Granger causality and dynamic causal modelling”. *Current Opinion in Neurobiology*, 23(2) (2013), pp. 172–178 (cit. on p. 178).

[328] M. L. Seghier and K. J. Friston. “Network discovery with large DCMs”. *Neuroimage*, 68 (2013), pp. 181–191 (cit. on p. 178).

# List of Figures

# List of Tables

# Deutsche Zusammenfassung

## Einleitung

In der Neurowissenschaftlichen Forschung spricht man häufig davon, dass neuronale Systeme „Berechnungen“ durchführen oder „Information verarbeiten“, um komplexes Verhalten zu ermöglichen. Diesen Aussagen liegt die Annahme zugrunde, dass neuronale Systeme ihre Umwelt in Form physikalischer Variablen repräsentieren (z.B. als Membranpotential oder Pheromonkonzentration) [1], und die in den Variablen gespeicherte Information nutzen und verarbeiten, um durch Verhalten mit der Umwelt zu interagieren [1]. Das Verstehen dieser Informationsverarbeitung ist damit wichtig, um zu verstehen, *wie* Verhalten aus physiologischen Prozessen entsteht.

Die Informationsverarbeitung in neuronalen Systemen wird in der bestehenden Forschung jedoch kaum *direkt* untersucht. Stattdessen wird häufig versucht, aus der Beobachtung von Verhalten oder von physiologischen Prozessen Rückschlüsse auf die zugrundeliegende Informationsverarbeitung zu ziehen [1–4]. Diese indirekte Untersuchung ist oft wenig erfolgreich–selbst dann nicht, wenn detaillierte Beschreibungen von Verhalten und Physiologie vorliegen [2]. So sind beispielsweise die neuronale Architektur sowie viele neuronale Prozesse des Modellorganismus C. elegans vollständig beschrieben [36–39]; es ist jedoch weiterhin nicht möglich auf Grundlage dieses Wissens um die physiologische Implementierung, Verhalten oder Lernen des Organismus vorherzusagen. Als eine mögliche Ursache für den fehlenden Übertrag zwischen Implementierung und Verhalten, wurde vorgeschlagen, dass es sich bei Informationsverarbeitung, Verhalten und Implementierung um drei abgegrenzte Erklärungsebenen eines Systems handelt [2, 40]. Dabei sind diese Ebenen voneinander unabhängig, sodass sich Phänomene auf einzelnen Ebenen kaum gegenseitig beschränken: beispielsweise lässt sich das selbe Verhalten durch eine Vielzahl möglicher, informationsverarbeitender Algorithmen erklären, sodass sich der Algortihmus aus dem Verhalten nicht eindeutig ableiten lässt [40]. Da unter dieser Annahme eindeutige Schlüsse von einer Erklärungsebene auf die andere nicht möglich sind, muss die Informationsverarbeitung als eigenständige Erklärungsebene explizit untersucht werden [2].

Die Informationsverarbeitung und ihre formale Beschreibung ist für klassische informationsverarbeitende Systeme wie Computer oder abstrakte Rechnermodelle klar definiert. Dagegen fehlt eine entsprechende Definition für biologische Systeme [6, 7] wie z.B. Schwärme [45, 46], Genregulationsnetzwerke [49, 50] oder neuronale Systeme [51, 52]. Im Gegensatz zu traditionellen, informationsverarbeitenden Systemen sind biologische Systeme verteilte Systeme, welche Information hoch-parallel und stochastisch repräsentieren und verarbeitet, um globales Verhalten zu generieren [6]. Dabei ist es für einen externen Betrachter nicht notwendigerweise ersichtlich, welche (Teil-)Funktion die lokale Informationsverarbeitung in einzelnen Subsystem erfüllt. Weiterhin ist es möglich, dass sich die lokale Informationsverarbeitung über die Zeit verändert. Es müssen daher neue Werkzeuge zur Analyse dieser verteilten und dynamischen Informationsverarbeitung in biologischen Systemen gefunden werden, die über die Beschreibung von Implementierung und Funktion hinaus gehen und erklären, wie diese Art der Informationsverarbeitung zu globalem Verhalten führt.

Einen möglicher Ansatz zur Analyse dieser biologischen Informationsverarbeitung beschreiben Mitchell et al. [7]: Die Autoren schlagen vor, Informationsverarbeitung nicht als „eine ‚sinnvolle' Transformation einer Eingabe in eine Ausgabe" [7] oder die Möglichkeit universeller Programmierbarkeit zu interpretieren, sondern die sogenannte *intrinsische Informationsverarbeitung* eines Systems zu untersuchen. Die Analyse dieser intrinsischen Informationsverarbeitung versucht, Berechnungen durch die Quantifizierung basaler, informationsverarbeitender Operationen zu beschreiben. Diese Operationen sind der Transfer, die Speicherung sowie die Modifikation von Information. Die Implementierung dieser drei Funktionen wurde als notwendige Voraussetzung für universelle Informationsverarbeitung in beliebigen Systemen beschrieben [5]. Es wurde weiterhin vorgeschlagen, über die Quantifizierung der Operationen die stattfindende Informationsverarbeitung einzugrenzen und zu beschreiben [8, 11]. Eine konkrete Umsetzung dieses Ansatzes wurde von Lizier [9] vorgestellt, welcher Maße aus der Informationstheorie nutzte, um die vorgestellten Operationen in verteilt rechnenden Systemen zu untersuchen. U.a. wurde die erfolgreiche Anwendung dieser Maße in der Analyse Zellulärer Automaten, einem Modellsystem für verteilte Informationsverarbeitung, demonstriert [11]. Die Informationstheorie ist hier eine natürliche Wahl für Maße der Informationsverarbeitung, da sie eine mathematisch vollständige Beschreibung der intuitiven Vorstellung von Information und ihrer Kommunikation bietet [10]. Weiterhin erlaubt die Informationstheorie eine Quantifizierung von Information, die unabhängig von der *Semantik*, d.h. der Bedeutung für einen Beobachter, ist. Somit ist die Informationstheorie besonders für die Anwendung in biologischen (Sub-)Systemen geeignet, in denen die *Funktion* lokaler Informationsverarbeitung unklar ist.

Die von Lizier [9] vorgeschlagenen Maße stellen den ersten umfassenden Ansatz zur direkten Untersuchung von Informationsverarbeitung in beliebigen, verteilten Systemen dar und sind damit ein vielversprechender Ansatz zur Untersuchung der Informationsverarbeitung in neuronalen Systemen [8]. Bisher hat insbesondere das vorgeschlagene Maß für den Informationstransfer, die sogenannte Transferentropie (TE) [12], Beachtung in den Neurowissenschaften gefunden (z.B. [14–22]). Die TE kann als Maß für den Transfer von Information zwischen einem Quell-

und einem Zielprozess verstanden werden. TE quantifiziert, inwiefern sich die Vorhersagbarkeit des Zielprozesses verbessert, wenn nicht nur die Vergangenheit dieses Prozesses zur Vorhersage genutzt wird, sondern auch die Vergangenheit eines zweiten Quellprozesses. Die TE ist ein modellfreies Maß, welches auch nicht-lineare Zusammenhänge erfasst. Die TE ist damit ein attraktives und vielfach verwendetes Maß für Informationstransfer in neuronalen Systemen. Jedoch existieren trotz dieser Vorzüge und der Popularität der TE noch immer eine Vielzahl praktischer Probleme in ihrer Schätzung und Interpretation (z.B. [13, 23]).

Ein Problem liegt in der großen Menge benötigter Daten, um TE zu schätzen. Dieses Problem verstärkt sich in den Neurowissenschaften, da beobachtete Prozesse häufig nicht-stationär sind (d.h. die den einzelnen Zufallsvariablen zugrunde liegenden Wahrscheinlichkeitsverteilungen sind nicht invariant über die Zeit), sodass die benötigten Datenmengen nicht aus Beobachtungen über die Zeit gewonnen werden können. Ein weiteres Problem ergibt sich aus der multivariaten Natur neurowissenschaftlicher Datensätze und Fragestellungen: Da multivariate Ansätze zur Schätzung der TE zu rechenaufwändig sind, wird TE oft nur bivariat zwischen Variablenpaaren geschätzt. Dieses Vorgehen kann zu fehlerhaften Ergebnissen führen [25, 28, 29]. Ein drittes Problem entsteht bei der Anwendung der TE wenn Schätzparameter nicht optimiert und durch heuristische Werte ersetzt werden (z.B. [32, 33]). Dies kann zur Unter- oder Überschätzung des tatsächlichen Informationstransfers führen [13]. Zuletzt entstehen noch immer Probleme bei der Interpretation der TE: häufig wird TE als Maß für *kausale* Interaktion, d.h. ein Maß für einen physikalischen oder mechanistischen Zusammenhang zwischen beobachteten Prozessen, interpretiert (z.B. [89–91]). Es wurde jedoch gezeigt, dass diese Interpretation nicht korrekt ist [68, 92] und somit zu Fehlschlüssen auf Grundlage der beobachteten Daten führen kann.

Die vorliegende Arbeit adressiert die vorgestellten Probleme in drei Studien: In der ersten Studie wird eine effiziente Implementierung zur Schätzung der TE aus nicht-stationären Daten vorgestellt. In der zweiten Studie wird eine Korrektur für Fehler in der Schätzung bivariater TE aufgrund multivariater Effekte vorgeschlagen. In der dritten Studie werden Möglichkeiten zur Optimierung von Schätzparametern diskutiert und es wird gezeigt, welche Konsequenzen die Wahl nicht-optimaler Schätzparametern hat. Weiterhin wird diskutiert, wie die Interpretation der TE als Maß für kausale Zusammenhängen zu inkorrekten Schlüssen führen kann.

## Studie 1

Die erste Studie beschreibt eine effiziente Implementierung zur Schätzung von TE aus nicht-stationären, neuronalen Daten.

Die Schätzung von TE erfordert eine erhebliche Menge an Daten, um die dem Maß zugrunde liegenden und typischerweise unbekannten Wahrscheinlichkeitsdichtefunktionen zu schätzen. In den Neurowissenschaften werden diese Beobachtungen häufig über die Zeit gewonnen. Dieses Vorgehen setzt jedoch voraus, dass die zugrunde liegenden Prozesse stationär, d.h. statistische Eigenschaften über die Zeit invariant sind. Diese Annahme ist für neuronale Daten häufig nicht zutreffend. Um auch im Fall nicht-stationärer Daten TE robust schätzen zu können, besteht die Möglichkeit, Daten über physischen oder zeitlichen Kopien der beobachteten Prozesse (sog. *Ensembles*) zusammenzufassen [24]. Ein Ensemble von zeitlichen Kopien liegt in neurowissenschaftlichen Daten typischerweise in Form von „Trials", d.h. einer Vielzahl von beobachteten Wiederholungen der experimentellen Aufgabe, vor. Trials sind „zyklostationär", d.h. sie sind Beobachtungen wiederkehrender, zeitlicher Kopien des selben Prozesses. Somit können Beobachtungen über das Ensemble der Trials zusammengefasst werde, um eine ausreichende Datenmenge für informationstheoretische Schätzungen zu erhalten. Damit kann TE bei einer genügend großen Anzahl vorliegender Trials für beliebig kleine Zeitfenster, d.h. *zeitaufgelöst*, geschätzt werden (wobei bei hinreichend kleinen Zeitfenstern von Stationarität ausgegangen werden kann).

Ein technischer Nachteil des Zusammenfassens von Beobachtungen über Trials ist, dass für individuelle Trials keine Einzelschätzungen mehr vorliegen, welche für effiziente statistische Tests der geschätzten TE genutzt werden können. Ein statistischer Test der geschätzten TE ist notwendig, um den systematischen Fehler des TE-Schätzers zu korrigieren (welcher nicht analytisch korrigierbar ist) [85]. Hierzu wird der geschätzte TE-Wert auf Signifikanz gegenüber einer Nullverteilung getestet. Liegen Schätzungen für einzelne Trials vor, kann diese Nullverteilung effizient konstruiert werden [86]. Fehlen diese Einzelschätzungen jedoch, muss eine Nullverteilung generiert werden, indem TE wiederholt aus Daten geschätzt wird, in denen Beobachtungen des Ziel-Prozesses in einem Trial mit Beobachtungen eines zufällig gewählten weiteren Trials vertauscht werden. Durch dieses Permutieren der Beobachtungen wird der Informationstransfer zwischen den Prozessen zerstört, während die statistischen Eigenschaften der beobachteten Zeitserien erhalten bleiben. Durch wiederholtes Permutieren und Schätzen von TE entsteht eine Nullverteilung von TE-Schätzwerten bei nicht vorhandenem Informationstransfer, gegen die die TE-Schätzung aus den originalen Daten auf Signifikanz getestet werden kann.

Diese Art der Generierung einer Nullverteilung erfordert das wiederholte Schätzen von TE bis die gewünschte Verteilungsgröße erreicht ist. Hierdurch multipliziert sich der Berechnungsaufwand der Schätzung und Testung der TE mit der Größe der Nullverteilung. Aufgrund dieses erhöhten Berechnungsaufwands erreichen bisherige Implementierungen von Schätzern keine praktikablen Laufzeiten wenn sie für die Ensemble-Schätzung von TE eingesetzt werden. Die vorliegende Studie adressiert dieses Problem, indem sie eine neue Implementierung der zentralen Algorithmen der TE-Schätzung vorschlägt: diese Algorithmen sind Nächste-Nachbarn-Suchen [83], für die eine hoch-parallele Implementierung für Grafikprozessoren (GPUs) verwendet wird. Diese Implementierung erlaubt es, den erhöhten Rechenaufwand durch Parallelisierung der Suchen über Datensätze (bspw. original und mehrere Hundert permutierte Datensätze) zu bewältigen.

Diese effiziente Implementierung der Ensemble-Methode ermöglicht somit die Schätzung informationstheoretischer Maße aus nicht-stationären Daten, wie sie häufig in den Neurowissenschaften vorkommen. Dies erlaubt auch das Schätzen von zeitaufgelöster TE in beliebig kleinen Zeitfenstern bei ausreichender Größe des Ensembles. Es wird eine Referenz-Implementierung des vorgeschlagenen Verfahrens vorgestellt, sowie die Anwendung der Implementierung zur Schätzung von TE aus magnetoenzephalographischen Daten.

## Studie 2

Die zweite Studie beschreibt eine Methode zur post-hoc Korrektur von multivariaten Effekten in bivariaten TE-Schätzungen.

Neurowissenschaftliche Fragestellungen befassen sich häufig mit einer Vielzahl beobachteter Variablen (z.B. simultane Beobachtungen mehrerer Neurone oder kortikaler Areale). Typischerweise wird der Informationstransfer zwischen diesen Variablen durch iteratives Schätzen bivariater TE (von *einem* Quell- zu *einem* Ziel-Prozess) rekonstruiert, da eine vollständig multivariate Schätzung von TE ein NP-hartes Problem ist [25–27]. Dieses Vorgehen lässt multivariate Interaktionen außer Acht und kann zu fehlerhaften Ergebnissen führen; besonders gravierend sind hierbei falsch-positive Ergebnisse, d.h., das Auftreten signifikanter TE, ohne dass ein tatsächlicher Informationstransfer vorliegt [25, 28, 29]. Diese Fehler entstehen dann z.B. dann, wenn zwei Prozesse gleichzeitig Information von einer gemeinsamen Quelle erhalten: hier kann fälschlicherweise TE zwischen den Prozessen gefunden werden, wenn der Input aus der gemeinsamen Quelle zu einer Korrelation in den beobachteten Zeitserien führt.

Die vorliegende Studie stellt einen approximativen Ansatz zur Identifizierung potentieller falsch-positiver Ergebnisse in bivariaten TE-Schätzungen aus multivariaten Datensätzen vor. Der Ansatz geht davon aus, dass falsch-positive Ergebnisse durch charakteristische Graph-Motive im Netzwerk der bivariaten Interaktionen identifiziert werden können; die Motive zeichnen sich insbesondere durch Muster in der Latenz $u$ der Informationstransfers zwischen Prozess-Paaren aus. Die bivariat geschätzte TE wird deshalb als Graph repräsentiert, der mit den rekonstruierten

Werten für $u$ gewichtet wird. Es wird ein Algorithmus vorgestellt, welcher den so konstruierten Graph nach charakteristischen Motiven durchsucht. In den identifizierten Motiven werden potenziell falsche Verbindungen markiert, welche aus dem Graph entfernt werden können, um eine konservativere Approximation des tatsächlichen Informationstransfer-Netzwerks zu erhalten.

Der vorgestellte Ansatz erlaubt die approximative Rekonstruktion multivariaten Informationstransfers in multivariaten Datensätzen. Der beschriebene Graph-Algorithmus kann auf jedes Netzwerk gerichteter, bivariater Zusammenhangsmaße angewendet werden, dessen Verbindungen mit Interaktions-Latenzen gewichtet wurden. Die Anwendbarkeit des Algorithmus auf simulierte und magnetoenzephalographische Datensätze wird demonstriert, außerdem werden mögliche Anwendungsszenarien in den Neurowissenschaften diskutiert.

## Studie 3

Die dritte Studie untersucht, inwieweit sich lokale Informationsverarbeitung im Quell- und Ziel-Prozess der TE auf den Informationstransfer zwischen den Prozessen auswirkt. Es wird diskutiert, wie sich eine ungünstige Wahl der Schätzparameter auf die geschätzte TE auswirken und wie die Interpretation der TE als Kausalitätsmaß zu Fehlschlüssen bei der Interpretation von Ergebnissen führen kann.

Die Studie untersucht die TE zwischen zwei lokalen Feldpotentialen, aufgezeichnet in primären visuellen und präfrontalen kortikalen Arealen zweier Frettchen. Die Aufnahmen wurden simultan und unter verschiedenen Anästhesie-Konzentrationen durchgeführt. Zusätzlich zur TE wurde die lokale Informationsverarbeitung in beiden Arealen, d.h. dem Quell- und dem Ziel-Prozess der TE, gemessen, indem die aktive Informationsspeicherung (der sogenannte „active information storage" oder AIS [30]) und die Signalentropie geschätzt wurden. Bisherige Studien fanden Belege für eine Reduktion der TE bei höheren Anästhesie-Konzentrationen (z.B. [31–35]). Die Reduktion von TE wird hierbei oft durch eine Veränderung der zugrundeliegenden physiologischen Kopplung der Prozesse erklärt. Die vorliegende Studie legt dagegen eine alternative Erklärung für eine Abnahme der TE auf Grundlage der gemessenen, lokalen Informationsverarbeitung nahe: unter höheren Anästhesie-Konzentrationen wurde neben einer Abnahme der TE eine Abnahme der Signalentropie gefunden. Die Signalentropie im Quell-Prozess der TE ist eine mathematische, obere Schranke für die TE; die Reduktion der Signalentropie kann somit eine Reduktion der TE verursachen. Die Studie zeigt durch diese Untersuchung der lokalen Informationsverarbeitung eine potentielle alternative Ursache für die Reduktion von Informationstransfer unter Anästhesie auf.

Die Studie diskutiert weiterhin aktuelle Empfehlungen für die optimale Schätzung von TE und weist auf methodische Problem existierender Anästhesie-Forschung hin: Zum einen kann die Nutzung approximativer Schätzer (der sogenannte „symbolic transer entropy“ [109]) dazu führen, dass Informationstransfer nicht gefunden wird. Die vorliegende Studie nutzt deshalb zwei alternative TE-Schätzer, welche unter den derzeit verfügbaren Schätzern die günstigsten Fehlereigenschaften aufweisen: ein Nächste-Nachbarn-basiertes Verfahren nach Kraskov et al. [83] und ein Bayes-Schätzer nach Nemenman et al. [225] und Nemenman et al. [226]. Beide Verfahren kamen zu qualitativ gleichen Ergebnissen, welche auch existierenden Befunden zur Veränderungen von TE unter Anästhesie entsprechen [31–35] (welche zum Teil mit approximativen Methoden gewonnen wurden [32–35, 259]). Zum anderen wird die Wahl optimaler Schätzparameter diskutiert, insbesondere wie die Verwendung nicht-optimaler Parameter zu einer Über- oder Unterschätzung der TE führt, wodurch u.a. die dominante Richtung des Informationstransfers zwischen zwei bidirektional gekoppelten Prozessen nicht identifiziert werden kann.

Zusammenfassend zeigen unsere Ergebnisse, dass eine verringerte Informations*übertragung* unter Anästhesie durch eine verringerte Informations*produktion*, vor allem im Quell-Prozess der TE, verursacht werden kann. Die Veränderung in der TE wäre damit unabhängig von Veränderungen in der physiologischen Verbindung zwischen den beteiligten kortikalen Arealen. Dieser Befund wird gestützt durch das Wissen, dass das verwendete Anästhetikum Isofluran die Signalübertragung über weitreichende axonale Verbindungen, wie sie in der Studie untersucht wurden, kaum beeinflusst [223]. Die Studie zeigt damit, wie die (implizite) Interpretation von TE als kausales Maß, d.h. als Maß für den physikalischen Mechanismus, der Informationstransfer ermöglicht, zu fehlerhaften Interpretationen führen kann. Weiterhin unterstützt die Studie existierende Befunde zur Reduktion der TE unter Anästhesie und diskutiert aktuelle Empfehlungen für die präzise Schätzung von TE aus neuronalen Daten.

## Diskussion

Zusammenfassend stellt die vorliegende Arbeit zwei – besonders für die neurowissenschaftliche Anwendung relevante – methodische Ergänzungen für die Schätzung von TE vor: zum einen eine effiziente Implementierung für die zeitaufgelöste Schätzung von TE aus nicht-stationären Daten, zum anderen eine post-hoc Korrektur für multivariate Effekte in bivariat geschätzter TE. Weiterhin wird das empfohlene Vorgehen bei der Schätzung von TE aus experimentellen Daten, insbesondere im Hinblick auf die Wahl des Schätzers und der Optimierung von Schätzparametern, diskutiert. Zuletzt wird demonstriert, wie die Interpretation von TE als Maß physikalischer Interaktion zu Fehlschlüssen bei der Interpretation von Ergebnissen führen kann.

Es werden damit zwei wichtige methodische Probleme in der Schätzung der TE aus neurowissenschaftlichen Daten gelöst, sowie Voraussetzung für die korrekte Schätzung und Interpretation von TE diskutiert. Die vorgestellte Methode zur Schätzung aus nicht-stationären Daten sowie einige Aspekte der methodischen Empfehlungen haben weiterhin Gültigkeit für die Schätzung anderer informationstheoretischer Maße, wie z.B. des AIS als Maß für Informationsspeicherung. Für das offene Problem einer vollständig multivariaten Schätzung von TE und die offene Frage nach einem verbesserten Maß der Informationsmodifikation werden der aktuelle Stand der Forschung sowie mögliche zukünftige Entwicklungen vorgestellt.

Unter Berücksichtigung der diskutierten methodischen Erweiterungen und Empfehlungen, ist die TE ein etabliertes Maß für den Informationstransfer in neuronalen und anderen informationsverarbeitenden Systemen: zentral sind hier vor allem die Verwendung etablierter Schätzer, die Optimierung der Schätzparameter sowie eine entsprechende Planung der Datenerhebung. Wird diesen methodischen Empfehlungen gefolgt, bietet die TE ein robustes Maß des Informationstransfers, mit dem auch schwache, nicht-lineare Zusammenhänge, wie sie häufig in neuronalen Systemen vorkommen, erfasst werden können. Ein weiterer Vorzug der TE gegenüber anderen Zusammenhangsmaßen ist ihr Ursprung in der Informationstheorie, wodurch sie mit anderen informationstheoretischen Größen in Verbindung gesetzt werden kann; so ermöglicht die TE insbesondere in Verbindung mit Maßen der Informationsspeicherung und -modifikation [9] eine umfassende Analyse neuronaler Informationsverarbeitung, die eingebettet ist in eine exakte mathematische Definition von Information und ihrer Kommunikation.

# Colophon

This thesis was typeset with LaTeX 2$_{\varepsilon}$. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at `http://cleanthesis.der-ric.de/`.