

Article

# Minimising the Kullback–Leibler Divergence for Model Selection in Distributed Nonlinear Systems

Oliver M. Cliff <sup>1,2,\*</sup> , Mikhail Prokopenko <sup>2</sup>  and Robert Fitch <sup>1,3</sup> 

<sup>1</sup> Australian Centre for Field Robotics, The University of Sydney, Sydney NSW 2006, Australia; rfitch@uts.edu.au

<sup>2</sup> Complex Systems Research Group, The University of Sydney, Sydney NSW 2006, Australia; mikhail.prokopenko@sydney.edu.au

<sup>3</sup> Centre for Autonomous Systems, University of Technology Sydney, Ultimo NSW 2007, Australia

\* Correspondence: o.cliff@acfr.usyd.edu.au; Tel.: +61-2-9351-3040

Received: 21 December 2017; Accepted: 18 January 2018; Published: 23 January 2018

**Abstract:** The Kullback–Leibler (KL) divergence is a fundamental measure of information geometry that is used in a variety of contexts in artificial intelligence. We show that, when system dynamics are given by distributed nonlinear systems, this measure can be decomposed as a function of two information-theoretic measures, transfer entropy and stochastic interaction. More specifically, these measures are applicable when selecting a candidate model for a distributed system, where individual subsystems are coupled via latent variables and observed through a filter. We represent this model as a directed acyclic graph (DAG) that characterises the unidirectional coupling between subsystems. Standard approaches to structure learning are not applicable in this framework due to the hidden variables; however, we can exploit the properties of certain dynamical systems to formulate exact methods based on differential topology. We approach the problem by using reconstruction theorems to derive an analytical expression for the KL divergence of a candidate DAG from the observed dataset. Using this result, we present a scoring function based on transfer entropy to be used as a subroutine in a structure learning algorithm. We then demonstrate its use in recovering the structure of coupled Lorenz and Rössler systems.

**Keywords:** Kullback–Leibler divergence; model selection; information theory; transfer entropy; stochastic interaction; nonlinear systems; complex networks; state space reconstruction

## 1. Introduction

Distributed information processing systems are commonly studied in complex systems and machine learning research. We are interested in inferring data-driven models of such systems, specifically in the case where each subsystem can be viewed as a nonlinear dynamical system. In this context, the Kullback–Leibler (KL) divergence is commonly used to measure the quality of a statistical model [1–3]. When a model is compared with fully observed data, computing the KL divergence can be straightforward. However, in the case of spatially distributed dynamical systems, where individual subsystems are coupled via latent variables and observed through a filter, the presence of hidden variables renders typical approaches unusable. We derive the KL divergence in such systems as a function of two information-theoretic measures using methods from differential topology.

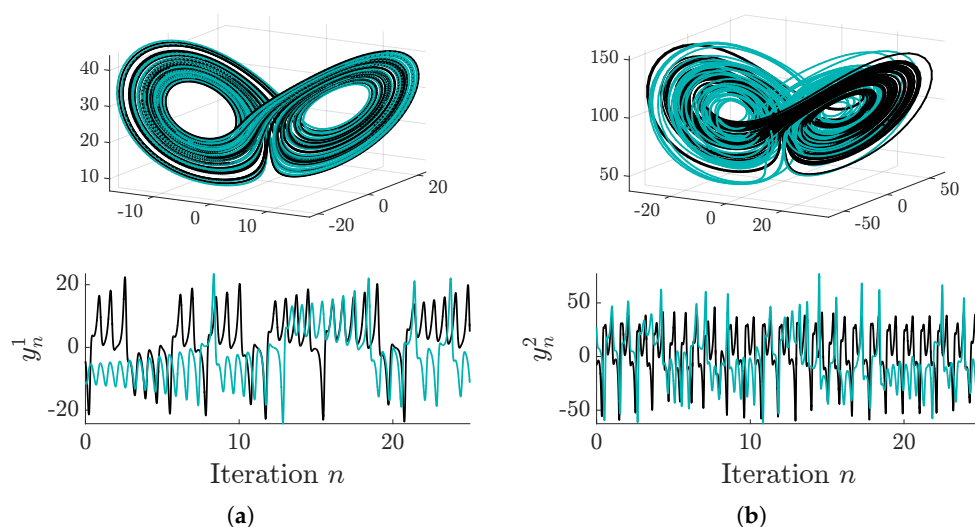
The *model selection* problem has applications in a wide variety of areas due to its usefulness in performing efficient inference and understanding the underlying phenomena being studied. Dynamical systems are an expressive model characterised by a map that describes their evolution over time and a read-out function through which we observe the latent state. Our research focuses on the more general case of a multivariate system, where a set of these subsystems are distributed and unidirectionally coupled to one another. The problem of inferring this coupling is an important

multidisciplinary study in fields such as ecology [4], neuroscience [5,6], multi-agent systems [7–9], and various others that focus on artificial and biological networks [10].

We represent such a spatially distributed system as a probabilistic graphical model termed a *synchronous graph dynamical system (GDS)* [11,12], whose structure is given by a directed acyclic graph (DAG). Model selection in this context is the problem of inferring directed relationships between hidden variables from an observed dataset, also known as *structure learning*. A main challenge in structure learning for DAGs is the case where variables are unobserved. Exact methods are known for fully observable systems (i.e., Bayesian networks (BNs)) [13]; however, these are not applicable in the more expressive case when the state variables in dynamical systems are latent. The main focus of this paper is to analytically derive a measure for comparing a candidate graph to the underlying graph that generated a measured dataset. Such a measure can then be used to solve the two subproblems that comprise structure learning, *evaluation* and *identification* [14], and hence find the optimal model that explains the data.

For the evaluation problem, it is desirable to select the *simplest* model that incorporates all statistical knowledge. This concept is commonly expressed via information theory, where an established technique is to evaluate the encoding length of the data, given the model [1,15,16]. The simplest model should aim to minimise code length [2], and therefore we can simplify our problem to that of minimising KL divergence for the synchronous GDS. Using this measure, we find a factorised distribution (given by the graph structure) that is closest to the complete (unfactorised) distribution. We first analytically derive an expression for this divergence, and build on this result to present a scoring function for evaluating candidate graphs based on a dataset.

The main result of this paper is an exact decomposition of the KL divergence for synchronous GDSs. We show that this measure can be decomposed as the difference between two well-known information-theoretic measures, stochastic interaction [17,18] and collective transfer entropy [19]. We establish this result by first representing discrete-time multivariate dynamical systems as dynamic Bayesian networks (DBNs) [20]. In this form, both the complete and factorised distributions cannot be directly computed due to the hidden system state. Thus, we draw on state space reconstruction methods from differential topology to reformulate the KL divergence in terms of computable distributions. Using this expression, we show that the maximum transfer entropy graph is the most likely to have generated the data. This is experimentally validated using toy examples of a Lorenz–Rössler system and a network of coupled Lorenz attractors (Figure 1) of up to four nodes. These results support the conjecture that transfer entropy can be used to infer effective connectivity in complex networks.



**Figure 1.** Trajectory of a pair of coupled Lorenz systems. *Top row:* original state of the subsystems. *Bottom row:* time-series measurements of the subsystems. In each figure, the black lines represent an uncoupled simulation ( $\lambda = 0$ ), and teal lines illustrate a simulation where the first (leftmost) subsystem was coupled to the second ( $\lambda = 10$ ). (a)  $\sigma = 10, \beta = 8/3, \rho = 28$ ; (b)  $\sigma = 10, \beta = 8/3, \rho = 90$ .

## 2. Related Work

Networks of coupled dynamical systems have been introduced under a variety of terms, such as complex networks [10], distributed dynamical systems [6] and master–slave configurations [21]. The defining feature of these networks is that the dynamics of each subsystem are given by a set of either discrete-time maps or first-order ordinary differential equations (ODEs). In this paper, we use the discrete-time formulation, where a map can be obtained numerically by integrating ODEs or recording observations at discrete-time intervals [22].

An important precursor to network reconstruction is inferring causality and coupling strength between complex nonlinear systems. Causal inference is intractable when the experimenter can not intervene with the dataset [23], and so we focus our attention on methods that determine conditional independence (coupling) rather than causality. In seminal work, Granger [24] proposed *Granger causality* for quantifying the predictability of one variable from another; however, a key requirement of this measure is linearity of the system, implying subsystems are separable [4]. Schreiber [25] extended these ideas and introduced *transfer entropy* using the concept of finite-order Markov processes to quantify the information transfer between coupled nonlinear systems. Transfer entropy and Granger causality are equivalent for linearly-coupled Gaussian systems (e.g., Kalman models) [26]; however, there are clear distinctions between the concepts of information transfer and causal effect [27]. Although transfer entropy has received criticism over spuriously identifying causality [28–30], we are concerned with statistical modelling and not causality of the underlying process.

Recently, a number of measures have been proposed to infer coupling between distributed dynamical systems based on reconstruction theorems. Sugihara et al. [4] proposed convergent cross-mapping that involves collecting a history of observed data from one subsystem and uses this to predict the outcome of another subsystem. This history is the delay reconstruction map described by Takens' Delay Embedding Theorem [31]. Similarly, Schumacher et al. [6] used the Bundle Delay Embedding Theorem [32,33] to infer causality and perform inference via Gaussian processes. Although the algorithms presented in these papers can infer driving subsystems in a spatially distributed dynamical system, the results obtained differ from ours as inference is not considered for an entire network structure, nor is a formal derivation presented. Contrasting this, we recently derived an information criterion for learning the structure of distributed dynamical systems [12]. However, the criterion proposed required parametric modelling of the probability distributions, and thus a detailed understanding of the physical phenomena being studied. In this paper, we extend this framework by first showing that KL divergence can be decomposed as information-theoretically useful measures, and then arriving at a similar result but employing non-parametric density estimation techniques to allow for no assumptions about the underlying distributions.

It is important to distinguish our approach from dynamic causal modelling (DCM), which attempts to infer the parameters of explicit dynamic models that cause (generate) data. In DCM, the set of potential models is specified a priori (typically in the form of ODEs) and then scored via marginal likelihood or evidence. The parameters of these models include *effective connectivity* such that their posterior estimates can be used to infer coupling among distributed dynamical systems [34]. As a consequence, these approaches can be used to recover networks that reveal the effective structure of observed systems [35,36]. In contrast, our approach does not require an explicitly specified model because the scoring function can be computed directly from the data. However, it does assume an implicit model in the form of a DAG where the subsystem processes are generated by generic functions.

Unlike effective connectivity, which is defined in relation to a (dynamic causal) model, the concept of *functional connectivity* refers to recovering statistical dependencies [37]. Consequently, statistical measures such as Granger causality and transfer entropy are typically used to identify functional, rather than effective structure. For example, transfer entropy has been used previously to infer networks in numerous fields, e.g., computational neuroscience [5,38], multi-agent systems [8], financial markets [39], supply-chain networks [40], and biology [41]. However, most of these results build on the work of Schreiber [25] by assuming the system is composed of finite-order Markov chains and

thus there is a dearth of work that provides formal derivations for the use of this measure for inferring effective connectivity. Our work allows us to compute scoring functions directly from multivariate time series (as in functional connectivity), yet still assumes an implicit model (albeit with weaker assumptions on the model than those considered in inferring effective connectivity).

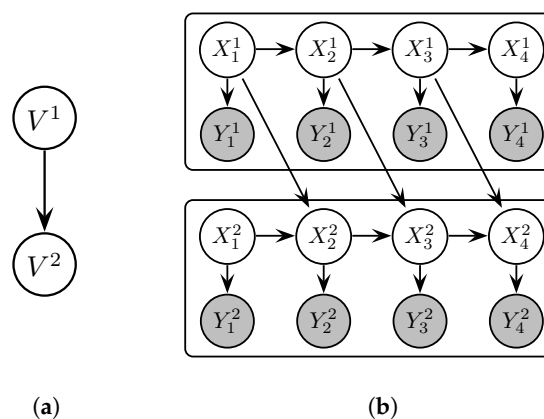
### 3. Background

#### 3.1. Notation

We use the convention that  $(\cdot)$  denotes a sequence,  $\{\cdot\}$  a set, and  $\langle \cdot \rangle$  a vector. In this work, we consider a collection of stationary stochastic temporal processes  $\mathbf{Z}$ . Each process  $Z^i$  comprises a sequence of random variables  $(Z_1^i, \dots, Z_N^i)$  with realisation  $(z_1^i, \dots, z_N^i)$  for countable time indices  $n \in \mathbb{N}$ . Given these processes, we can compute probability distributions of each variable by counting relative frequencies or by density estimation techniques [42,43]. We use bold to denote the set of all variables, e.g.,  $\mathbf{z}_n = \{z_n^1, \dots, z_n^M\}$  is the collection of  $M$  realisations at index  $n$ . Furthermore, unless otherwise stated,  $X_n^i$  is a latent (hidden) variable,  $Y_n^i$  is an observed variable, and  $Z_n^i$  is an arbitrary variable; thus,  $\mathbf{Z}_n = \{X_n, Y_n\}$  is the set of all hidden and observed variables at temporal index  $n$ . Given a graphical model  $G$ , the  $p^i$  parents of variable  $Z_{n+1}^i$  are given by the parent set  $\Pi_G(Z_{n+1}^i) = \{Z_n^{ij}\}_j = \{Z_n^{i1}, \dots, Z_n^{ip^i}\}$ . Finally, let the superscript  $z_n^{i,(k)} = \langle z_n^i, z_{n-1}^i, \dots, z_{n-k+1}^i \rangle$  denote the vector of  $k$  previous values taken by variable  $Z_n^i$ .

#### 3.2. Representing Distributed Dynamical Systems as Probabilistic Graphical Models

We are interested in modelling discrete-time multivariate dynamical systems, where the state is a vector of real numbers given by a point  $x_n$  lying on a compact  $d$ -dimensional manifold  $\mathcal{M}$ . A map  $f : \mathcal{M} \rightarrow \mathcal{M}$  describes the temporal evolution of the state at any given time, such that the state at the next time index  $x_{n+1} = f(x_n)$ . Furthermore, in many practical scenarios, we do not have access to  $x_n$  directly, and can instead observe it through a *measurement function*  $\psi : \mathcal{M} \rightarrow \mathbb{R}^M$  that yields a scalar representation  $y_n = \psi(x_n)$  of the latent state [22,44]. We assume the multivariate system can be factorised and modelled as a DAG with spatially distributed dynamical subsystems, termed a synchronous GDS (see Figure 2a). This definition is restated from [12] as follows.



**Figure 2.** Representation of (a) the synchronous GDS with two vertices ( $V^1$  and  $V^2$ ), and (b) the rolled-out DBN of the equivalent structure. Subsystems  $V^1$  and  $V^2$  are coupled by virtue of the edge  $X_n^1 \rightarrow X_{n+1}^2$ .

**Definition 1** (Synchronous GDS). A synchronous GDS  $(G, \mathbf{x}_n, \mathbf{y}_n, \{f^i\}, \{\psi^i\})$  is a tuple that consists of: a finite, directed graph  $G = (\mathcal{V}, \mathcal{E})$  with edge-set  $\mathcal{E} = \{E^i\}$  and  $M$  vertices comprising the vertex set  $\mathcal{V} = \{V^i\}$ ; a multivariate state  $\mathbf{x}_n = \langle x_n^i \rangle$ , composed of states for each vertex  $V^i$  confined to a  $d^i$ -dimensional manifold  $x_n^i \in \mathcal{M}^i$ ; an  $M$ -variate observation  $\mathbf{y}_n = \langle y_n^i \rangle$ , composed of scalar observations for each vertex  $y_n^i \in \mathbb{R}$ ; a set of

local maps  $\{f^i\}$  of the form  $f^i : \mathcal{M} \rightarrow \mathcal{M}^i$ , which update synchronously and induce a global map  $f : \mathcal{M} \rightarrow \mathcal{M}$ ; and a set of local observation functions  $\{\psi^1, \psi^2, \dots, \psi^M\}$  of the form  $\psi^i : \mathcal{M}^i \rightarrow \mathbb{R}$ .

The global dynamics and observations can therefore be described by the set of local functions [12]:

$$x_{n+1}^i = f^i(x_n^i, \langle x_n^{ij} \rangle_j) + v_{fi}, \tag{1}$$

$$y_{n+1}^i = \psi^i(x_{n+1}^i) + v_{\psi^i}, \tag{2}$$

where  $v_{fi}$  and  $v_{\psi^i}$  are additive noise terms. The subsystem dynamics (1) are a function of the subsystem state  $x_n^i$  and the subsystem parents' state  $\langle x_n^{ij} \rangle_j$  at the previous time index, i.e.,  $f^i : (\mathcal{M}^i \times_j \mathcal{M}^{ij}) \rightarrow \mathcal{M}^i$ . However, the observation  $y_{n+1}^i$  is a function of the subsystem state alone, i.e.,  $\psi^i : \mathcal{M}^i \rightarrow \mathbb{R}$ . We assume that the maps  $\{f^i\}$  and  $\{\psi^i\}$ , as well as the graph  $G$ , are time-invariant.

The discrete-time mapping for the dynamics (1) and measurement functions (2) can be modelled as a DBN in order to facilitate structure learning of the graph [12] (see Figure 2b). DBNs are a probabilistic graphical model that represent probability distributions over trajectories of random variables  $(Z_1, Z_2, \dots)$  using a prior BN and a *two-time-slice BN (2TBN)* [45]. To model the maps, however, we need only to consider the 2TBN  $B = (G, \Theta_G)$ , which can model a first-order Markov process  $p_B(z_{n+1} | z_n)$  graphically via a DAG  $G$  and a set of conditional probability distribution (CPD) parameters  $\Theta_G$  [45]. Given a set of stochastic processes  $(Z_1, Z_2, \dots, Z_N)$ , the realisation of which constitutes the sample path  $(z_1, z_2, \dots, z_N)$ , the 2TBN distribution is given by  $p_B(z_{n+1} | z_n) = \prod_i \Pr(z_{n+1}^i | \pi_G(Z_{n+1}^i))$ , where  $\pi_G(Z_{n+1}^i)$  denotes the (index-ordered) set of realisations  $\{z_o^j : Z_o^j \in \Pi_G(Z_{n+1}^i)\}$ .

To model the synchronous GDS as a DBN, we associate each subsystem vertex  $V^i$  with a state variable  $X_n^i$  and an observation variable  $Y_n^i$ . The parents of subsystem  $V^i$  are denoted  $\Pi_G(V^i)$  [12]. From the dynamics (1), variables in the set  $\Pi_G(X_{n+1}^i)$  come strictly from the preceding time slice, and additionally, from the measurement function (2),  $\Pi_G(Y_{n+1}^i) = X_{n+1}^i$ . Thus, we can build the edge set  $\mathcal{E}$  in the GDS by means of the edges in the DBN [12], i.e., given an edge  $X_n^i \rightarrow X_{n+1}^j$  of the DBN, the equivalent edge  $V^i \rightarrow V^j$  exists for the GDS. The distributions for the dynamics (1) and observation (2) maps of  $M$  arbitrary subsystems can therefore be factorised according to the DBN structure such that [12]

$$p_B(z_{n+1} | z_n) = \prod_{i=1}^M \Pr(x_{n+1}^i | x_n^i, \langle x_n^{ij} \rangle_j) \cdot \Pr(y_{n+1}^i | x_{n+1}^i). \tag{3}$$

The goal of learning nonlinear dynamical networks thus becomes that of inferring the parent set  $\Pi_G(X_n^i)$  for each latent variable  $X_n^i$ .

Finally, recall that the parents of each observation are constrained such that  $\Pi_G(Y_{n+1}^i) = X_{n+1}^i$ . As a consequence, we use the shorthand notation  $y_n^{ij}$  to denote the observation of the  $j$ -th parent of the  $i$ -th subsystem at time  $n$  (and the same for  $x_n^{ij}$ ).

### 3.3. Network Scoring Functions

A number of exact and approximate DBN structure learning algorithms exist that are based on Bayesian statistics and information theory. We have shown in prior work how to compute the log-likelihood function for synchronous GDSs. In this section, we will briefly summarise the problem of structure learning for DBNs, focusing on the factorised distribution (3).

The *score and search* paradigm [46] is a common method for recovering graphical models from data. Given a dataset  $D = (y_1, y_2, \dots, y_N)$ , the objective is to find a DAG  $G^*$  such that

$$G^* = \arg \max_{G \in \mathcal{G}} g(B : D), \tag{4}$$



where  $g(B:D)$  is a scoring function measuring the degree of fitness of a candidate DAG  $G$  to the data set  $D$ , and  $\mathcal{G}$  is the set of all DAGs. Finding the optimal graph  $G^*$  in Equation (4) requires solutions to the two subproblems that comprise structure learning: the *evaluation* problem and the *identification* problem [14]. The main problem we focus on in this paper is the evaluation problem, i.e., determining a score that quantifies the quality of a graph, given data. Later, we will address the identification problem by discussing the attributes of this scoring function in efficiently finding the optimal graph structure.

In prior work, we developed a score based on the posterior probability of the network structure  $G$ , given data  $D$ . That is, we considered maximising the expected log-likelihood [12]

$$\ell(\hat{\Theta}_G : D) = \mathbf{E} [\log \Pr(D | G, \hat{\Theta}_G)] = \mathbf{E} \left[ \log \left( p_B(\mathbf{z}_{n+1} | \mathbf{z}_n) \right) \right], \tag{5}$$

where the expectation  $\mathbf{E}[Z] = \int_{-\infty}^{\infty} z \Pr(z) dz$ . It was shown that state space reconstruction techniques (see Appendix A) can be used to compute the log-likelihood of Equation (3) as a difference of conditional entropy terms [12]. In the same work, we illustrated that the log-likelihood ratio of a candidate DAG  $G$  to the empty network  $G_{\emptyset}$  is given by collective transfer entropy (see Appendix B), i.e.,

$$\ell(\hat{\Theta}_G : D) - \ell(\hat{\Theta}_{G_{\emptyset}} : D) = N \cdot \sum_{i=1}^M T_{(Y^{ij})_j \rightarrow Y^i}. \tag{6}$$

For the nested log-likelihoods above, the statistics of  $2(\ell(\hat{\Theta}_G : D) - \ell(\hat{\Theta}_{G_{\emptyset}} : D))$  asymptotically follow the  $\chi_q^2$ -distribution, where  $q$  is the difference between the number of parameters of each model [47,48]. We will draw on this log-likelihood decomposition in later sections for statistical significance testing.

#### 4. Computing Conditional KL Divergence

In this section, we present our main result, which is an analytical expression of KL divergence that facilitates structure learning in distributed nonlinear systems. We begin by considering the problem of finding an optimal DBN structure as searching for a parsimonious *factorised distribution*  $p_B$  that best represents the complete digraph distribution  $p_{K_M}$ . That is,  $p_{K_M}$  is the joint distribution yielded by assuming no factorisation (the complete graph  $K_M$ ) and thus no information loss. The distribution is expressed as:

$$p_{K_M}(\mathbf{z}_{n+1} | \mathbf{z}_n^{(n)}) = \Pr \left( \{z_{n+1}^1, \dots, z_{n+1}^M\} | \{z_n^1, \dots, z_n^M\}, \{z_{n-1}^1, \dots, z_{n-1}^M\}, \{z_1^1, \dots, z_1^M\} \right). \tag{7}$$

We quantify the similarity of the factorised distribution  $p_B$  to this joint distribution via KL divergence. In prior work, De Campos [3] derived the *MIT* scoring function for BNs by this approach and it was later used for DBN structure learning with complete data [49]. We extend the analysis to DBNs with latent variables, i.e., we compare the joint and factorised distributions of time slices, given the entire history,

$$\begin{aligned} D_{\text{KL}} [p_{K_M} \parallel p_B] &= D_{\text{KL}} \left[ p_{K_M}(\mathbf{z}_{n+1} | \mathbf{z}_n^{(n)}) \parallel p_B(\mathbf{z}_{n+1} | \mathbf{z}_n^{(n)}) \right] \\ &= \sum_{\mathbf{z}_n^{(n)}} \Pr(\mathbf{z}_n^{(n)}) \sum_{\mathbf{z}_{n+1}} \Pr(\mathbf{z}_{n+1} | \mathbf{z}_n^{(n)}) \log \frac{\Pr(\mathbf{z}_{n+1} | \mathbf{z}_n^{(n)})}{p_B(\mathbf{z}_{n+1} | \mathbf{z}_n^{(n)})} \\ &= \mathbf{E} \left[ \log \frac{\Pr(\mathbf{z}_{n+1} | \mathbf{z}_n^{(n)})}{p_B(\mathbf{z}_{n+1} | \mathbf{z}_n)} \right]. \end{aligned} \tag{8}$$

Substituting the synchronous GDS model (3) into Equation (8), we get

$$D_{\text{KL}} [p_{K_M} \parallel p_B] = \mathbf{E} \left[ \log \frac{\Pr(z_{n+1} | z_n^{(n)})}{\prod_{i=1}^M \Pr(x_{n+1}^i | x_n^i, \langle x_n^{ij} \rangle_j) \cdot \Pr(y_{n+1}^i | x_{n+1}^i)} \right]. \tag{9}$$

However, Equation (9) comprises maximum likelihood distributions with unobserved (latent) states  $x_n$ . It is common in model selection to decompose the KL divergence as

$$D_{\text{KL}} [p_{K_M} \parallel p_B] = \mathbf{E} \left[ \log \left( \Pr(z_{n+1} | z_n^{(n)}) \right) \right] - \mathbf{E} \left[ \log \left( p_B(z_{n+1} | z_n) \right) \right], \tag{10}$$

where the second term is simply the log-likelihood (5). In this form,  $p_{K_M}$  is often identical for all models considered and, in practice, it suffices to ignore this term and thus avoid the problem of computing distributions of latent variables. The resulting simpler expression can be viewed as log-likelihood maximisation (as in our previous work outlined in Section 3.3). However, as we show in this section,  $p_{K_M}$  is not equivalent for all models unless certain parameters of the dynamical systems are known. Hence, for now, we cannot ignore the first term of Equation (10) and we instead propose an alternative decomposition of KL divergence that comprises only observed variables.

#### 4.1. A Tractable Expression via Embedding Theory

In order to compute the distributions in (9), we use the Bundle Delay Embedding Theorem [32,33] to reformulate the factorised distribution (denominator), and the Delay Embedding Theorem for Multivariate Observation Functions [50] for the joint distribution (numerator). We describe these theorems in detail in Appendix A, along with the technical assumptions required for  $(f, \psi)$ . Although the following theorems assume a diffeomorphism, we also discuss application of the theory towards inferring the structure of endomorphisms (e.g., coupled map lattices [51]) in the same appendix.

The first step is to reproduce a prior result for computing the factorised distribution (denominator) in Equation (9). First, the embedding

$$y_n^{i,(\kappa^i)} = \langle y_n^i, y_{n-\tau^i}^i, \dots, y_{n-(\kappa^i-1)\tau^i}^i \rangle, \tag{11}$$

where  $\tau^i$  is the (strictly positive) lag, and  $\kappa^i$  is the embedding dimension of the  $i$ -th subsystem (the *embedding parameters*). Note that, although we can take either the future or past delay embedding (11) for diffeomorphisms, we explicitly consider a *history* of values to account for both endomorphisms and diffeomorphisms. Moreover, an important assumption of our approach is that the structure (enforced by coupling between subsystems) is a DAG; this comes from the Bundle Delay Embedding Theorem [32,33] (see Lemma 1 of [12] for more detail). Our previous result is expressed as follows.

**Lemma 1** (Cliff et al. [12]). *Given an observed dataset  $D$ , where  $y_n \in \mathbb{R}^M$ , generated by a directed and acyclic synchronous GDS  $(G, x_n, y_n, \{f^i\}, \{\psi^i\})$ , the 2TBN distribution can be written as*

$$\prod_{i=1}^M \Pr(x_{n+1}^i | x_n^i, \langle x_n^{ij} \rangle_j) \cdot \Pr(y_{n+1}^i | x_{n+1}^i) = \frac{\prod_{i=1}^M \Pr(y_{n+1}^i | y_n^{i,(\kappa^i)}, \langle y_n^{ij,(\kappa^i)} \rangle_j)}{\Pr(x_n | \langle y_n^{i,(\kappa^i)} \rangle)}. \tag{12}$$

Next, we present a method for computing the joint distribution (numerator) in Lemma 3. For convenience, Lemma 2 restates part of the delay embedding theorem in [50] in terms of subsystems of a synchronous GDS and establishes existence of a map  $\mathbf{G}$  for predicting future observations from a history of observations.

**Lemma 2.** *Consider a diffeomorphism  $f : \mathcal{M} \rightarrow \mathcal{M}$  on a  $d$ -dimensional manifold  $\mathcal{M}$ , where the multivariate state  $x_n$  consists of  $M$  subsystem states  $\langle x_n^1, \dots, x_n^M \rangle$ . Each subsystem state  $x_n^i$  is confined to a submanifold*

$\mathcal{M}^i \subseteq \mathcal{M}$  of dimension  $d^i \leq d$ , where  $\sum_i d^i = d$ . The multivariate observation is given, for some map  $\mathbf{G}$ , by  $\mathbf{y}_{n+1} = \mathbf{G}(\langle \mathbf{y}_n^{i,(\kappa^i)} \rangle)$ .

**Proof.** The proof restates part of the proof of Theorem 2 of Deyle and Sugihara [50] in terms of subsystems. Given  $M$  inhomogeneous observation functions  $\{\psi^i\}$ , the following map

$$\Phi_{f,\psi}(\mathbf{x}) = \langle \Phi_{f^1,\psi^1}(\mathbf{x}), \Phi_{f^2,\psi^2}(\mathbf{x}), \dots, \Phi_{f^M,\psi^M}(\mathbf{x}) \rangle \tag{13}$$

is an embedding where each subsystem (local) map  $\Phi_{f^i,\psi^i} : \mathcal{M} \rightarrow \mathbb{R}^{\kappa^i}$ , smoothly (at least  $\mathbb{C}^2$ ), and, at time index  $n$  is described by

$$\begin{aligned} \Phi_{f^i,\psi^i}(\mathbf{x}_n) &= \langle \psi^i(\mathbf{x}_n), \psi^i(\mathbf{x}_{n-\tau}), \dots, \psi^i(\mathbf{x}_{n-(k-1)\tau}) \rangle \\ &= \mathbf{y}_n^{i,(\kappa^i)}, \end{aligned} \tag{14}$$

where  $\sum_i \kappa^i = 2d + 1$  [50]. Note that, from (13) and (14), we have the global map

$$\Phi_{f,\psi}(\mathbf{x}_n) = \langle \mathbf{y}_n^{i,(\kappa^i)} \rangle = \langle \mathbf{y}_n^{1,(\kappa^1)}, \dots, \mathbf{y}_n^{m,(\kappa^M)} \rangle.$$

Now, since  $\Phi_{f,\psi}$  is an embedding, it follows that the map  $\mathbf{F} = \Phi_{f,\psi} \circ f \circ \Phi_{f,\psi}^{-1}$  is well defined and a diffeomorphism between two observation sequences  $\mathbf{F} : \mathbb{R}^{2d+1} \rightarrow \mathbb{R}^{2d+1}$ , i.e.,

$$\begin{aligned} \langle \mathbf{y}_{n+1}^{i,(\kappa^i)} \rangle &= \Phi_{f,\psi}(\mathbf{x}_{n+1}) = \Phi_{f,\psi}(f(\mathbf{x}_n)) \\ &= \Phi_{f,\psi}\left(f\left(\Phi_{f,\psi}^{-1}\left(\langle \mathbf{y}_n^{i,(\kappa^i)} \rangle\right)\right)\right) = \mathbf{F}(\langle \mathbf{y}_n^{i,(\kappa^i)} \rangle). \end{aligned}$$

The last  $2d + 1$  components of  $\mathbf{F}$  are trivial, i.e., the set  $\langle \mathbf{y}_n^{i,(\kappa^i)} \rangle$  is observed; denote the first  $M$  components by  $\mathbf{G} : \Phi_{f,\psi} \rightarrow \mathbb{R}^M$ , and then we have  $\mathbf{y}_{n+1} = \mathbf{G}(\langle \mathbf{y}_n^{i,(\kappa^i)} \rangle)$ .  $\square$

We now use the result of Lemma 2 to obtain a computable form of the KL divergence.

**Lemma 3.** Consider a discrete-time multivariate dynamical system with generic  $(f, \psi)$  modelled as a directed and acyclic synchronous GDS  $(G, \mathbf{x}_n, \mathbf{y}_n, \{f^i\}, \{\psi^i\})$  with  $M$  subsystems. The KL divergence of a candidate graph  $G$  from the observed dataset  $D$  can be computed from tractable probability distributions:

$$D_{\text{KL}} [p_{K_M} \parallel p_B] = \mathbf{E} \left[ \log \frac{\Pr(\mathbf{y}_{n+1} \mid \langle \mathbf{y}_n^{i,(\kappa^i)} \rangle)}{\prod_{i=1}^M \Pr(\mathbf{y}_{n+1}^i \mid \mathbf{y}_n^{i,(\kappa^i)}, \langle \mathbf{y}_n^{ij,(\kappa^{ij})} \rangle_j)} \right]. \tag{15}$$

**Proof.** Lemma 1, we can substitute (12) into (9), and express the KL divergence  $D_{\text{KL}} [p_{K_M} \parallel p_B]$  as

$$D_{\text{KL}} [p_{K_M} \parallel p_B] = \mathbf{E} \left[ \log \left( \Pr(\mathbf{z}_{n+1} \mid \mathbf{z}_n^{(n)}) \cdot \frac{\Pr(\mathbf{x}_n \mid \langle \mathbf{y}_n^{i,(\kappa^i)} \rangle)}{\prod_{i=1}^M \Pr(\mathbf{y}_{n+1}^i \mid \mathbf{y}_n^{i,(\kappa^i)}, \langle \mathbf{y}_n^{ij,(\kappa^{ij})} \rangle_j)} \right) \right]. \tag{16}$$

We now focus on  $p_{K_M}(\mathbf{z}_{n+1} \mid \mathbf{z}_n^{(n)})$ . Using the chain rule,

$$p_{K_M}(\mathbf{z}_{n+1} \mid \mathbf{z}_n^{(n)}) = \Pr(\mathbf{x}_{n+1} \mid \mathbf{z}_n^{(n)}) \cdot \Pr(\mathbf{y}_{n+1} \mid \mathbf{x}_{n+1}, \mathbf{z}_n^{(n)}).$$

Given the Markov property of the dynamics (1) and observation (2) maps, we get

$$p_{K_M}(\mathbf{z}_{n+1} \mid \mathbf{z}_n^{(n)}) = \Pr(\mathbf{X}_{n+1} = f(\mathbf{x}_n) \mid \mathbf{x}_n) \cdot \Pr(\mathbf{Y}_{n+1} = \psi(\mathbf{x}_{n+1}) \mid \mathbf{x}_{n+1}). \tag{17}$$



Now, recall from Lemma 2 that global equations for the entire system state  $\mathbf{x}_n$  and observation  $\mathbf{y}_n$  are

$$\mathbf{x}_{n+1} = f(\mathbf{x}_n) + \mathbf{v}_f = f\left(\Phi_{f,\psi}^{-1}(\langle \mathbf{y}_n^{i,(\kappa^i)} \rangle)\right) + \mathbf{v}_f, \tag{18}$$

$$\mathbf{y}_{n+1} = \psi(\mathbf{x}_{n+1}) + \mathbf{v}_\psi = \mathbf{G}(\langle \mathbf{y}_n^{i,(\kappa^i)} \rangle) + \mathbf{v}_\psi. \tag{19}$$

Given the assumption of i.i.d noise on the function  $f$ , from (18), we express the probability of the dynamics  $\mathbf{x}_{n+1}$ , given by the embedding, as

$$\begin{aligned} \Pr\left(\mathbf{x}_{n+1} \mid \langle \mathbf{y}_n^{i,(\kappa^i)} \rangle\right) &= \Pr\left(\mathbf{X}_{n+1} = f\left(\Phi_{f,\psi}^{-1}\left(\langle \mathbf{y}_n^{i,(\kappa^i)} \rangle\right)\right) \mid \langle \mathbf{y}_n^{i,(\kappa^i)} \rangle\right) \\ &= \Pr\left(\mathbf{X}_n = \Phi_{f,\psi}^{-1}\left(\langle \mathbf{y}_n^{i,(\kappa^i)} \rangle\right) \mid \langle \mathbf{y}_n^{i,(\kappa^i)} \rangle\right) \cdot \Pr\left(\mathbf{X}_{n+1} = f(\mathbf{x}_n) \mid \mathbf{x}_n\right). \end{aligned} \tag{20}$$

By assumption, the observation noise is i.i.d or dependent only on the state  $\mathbf{x}_{n+1}$ , and thus the probability of observing  $\mathbf{y}_{n+1}$ , from (19) is

$$\begin{aligned} \Pr\left(\mathbf{y}_{n+1} \mid \langle \mathbf{y}_n^{i,(\kappa^i)} \rangle\right) &= \Pr\left(\mathbf{Y}_{n+1} = \mathbf{G}(\langle \mathbf{y}_n^{i,(\kappa^i)} \rangle) \mid \langle \mathbf{y}_n^{i,(\kappa^i)} \rangle\right) \\ &= \Pr\left(\mathbf{X}_{n+1} = f\left(\Phi_{f,\psi}^{-1}\left(\langle \mathbf{y}_n^{i,(\kappa^i)} \rangle\right)\right) \mid \langle \mathbf{y}_n^{i,(\kappa^i)} \rangle\right) \\ &\quad \times \Pr\left(\mathbf{Y}_{n+1} = \psi(\mathbf{x}_{n+1}) \mid \mathbf{x}_{n+1}\right). \end{aligned} \tag{21}$$

By (20) and (21), we have that

$$\Pr(\mathbf{x}_{n+1} \mid \mathbf{x}_n) \cdot \Pr(\mathbf{y}_{n+1} \mid \mathbf{x}_{n+1}) = \frac{\Pr(\mathbf{y}_{n+1} \mid \langle \mathbf{y}_n^{i,(\kappa^i)} \rangle)}{\Pr(\mathbf{x}_n \mid \langle \mathbf{y}_n^{i,(\kappa^i)} \rangle)}. \tag{22}$$

Substituting Equation (22) into (17) gives

$$p_{K_M}(\mathbf{z}_{n+1} \mid \mathbf{z}_n^{(n)}) = \frac{\Pr(\mathbf{y}_{n+1} \mid \langle \mathbf{y}_n^{i,(\kappa^i)} \rangle)}{\Pr(\mathbf{x}_n \mid \langle \mathbf{y}_n^{i,(\kappa^i)} \rangle)}. \tag{23}$$

Finally, substituting (23) back into (16) yields the statement of the theorem.  $\square$

Given all variables in (15) are observed, it is now straightforward to compute KL divergence; however, as we will see, it is more convenient to express (15) as a function of known information-theoretic measures.

#### 4.2. Information-Theoretic Interpretation

The main theorem of this paper states KL divergence in terms of transfer entropy and stochastic interaction. These information-theoretic concepts are defined in Appendix B for convenience.

**Theorem 4.** Consider a discrete-time multivariate dynamical system with generic  $(f, \psi)$  represented as a directed and acyclic synchronous GDS  $(G, \mathbf{x}_n, \mathbf{y}_n, \{f^i\}, \{\psi^i\})$  with  $M$  subsystems. The KL divergence  $D_{\text{KL}} [p_{K_M} \parallel p_B]$  of a candidate graph  $G$  from the observed dataset  $D$  can be expressed as the difference between stochastic interaction (A9) and collective transfer entropy (A8), i.e.,

$$D_{\text{KL}} [p_{K_M} \parallel p_B] = S_Y - \sum_{i=1}^M T_{\{Y^{ij}\}_j \rightarrow Y^i}. \tag{24}$$

**Proof.** We can reformulate the KL divergence in (15) as

$$\begin{aligned}
 D_{\text{KL}} [p_{K_M} \parallel p_B] &= \mathbf{E} \left[ \log \left( \Pr(y_{n+1} \mid \langle y_n^{i,(\kappa^i)} \rangle) \right) \right] - \mathbf{E} \left[ \log \left( \prod_{i=1}^M \Pr(y_{n+1}^i \mid y_n^{i,(\kappa^i)}, \langle y_n^{ij,(\kappa^{ij})} \rangle_j) \right) \right] \\
 &= -H(Y_{n+1} \mid \{Y_n^{(\kappa^i)}\}) + \sum_{i=1}^M H(Y_{n+1}^i \mid Y_n^{i,(\kappa^i)}, \{Y_n^{ij,(\kappa^{ij})}\}_j) \\
 &= -H(Y_{n+1} \mid \{Y_n^{(\kappa^i)}\}) + \sum_{i=1}^M H(Y_{n+1}^i \mid Y_n^{i,(\kappa^i)}) \\
 &\quad + \sum_{i=1}^M \left( H(Y_{n+1}^i \mid Y_n^{i,(\kappa^i)}, \{Y_n^{ij,(\kappa^{ij})}\}_j) - H(Y_{n+1}^i \mid Y_n^{i,(\kappa^i)}) \right).
 \end{aligned}
 \tag{25}$$

Substituting in the definitions of transfer entropy (A8) and stochastic interaction (A9) completes the proof.  $\square$

To conclude this section, we present the following corollary showing that, when we assume a maximum or fixed embedding dimension  $\kappa^i$  and time delay  $\tau^i$ , it suffices to maximise the collective transfer entropy alone in order to minimise KL divergence for a synchronous GDS.

**Corollary 1.** Fix an embedding dimension  $\kappa^i$  and time delay  $\tau^i$  for each subsystem  $V^i \in \mathcal{V}$ . Then, the graph  $G$  that minimises the KL divergence  $D_{\text{KL}} [p_{K_M} \parallel p_B]$  is equivalent to the graph that maximises transfer entropy, i.e.,

$$\arg \min_{G \in \mathcal{G}} D_{\text{KL}} [p_{K_M} \parallel p_B] = \arg \max_{G \in \mathcal{G}} \sum_{i=1}^M T_{\{Y^{ij}\}_j \rightarrow Y^i}.
 \tag{26}$$

**Proof.** The first term of (24) is constant, given a constant vertex set  $\mathcal{V}$ , time delay  $\tau$  and embedding dimension  $\kappa$  and is thus unaffected by the parent set  $\Pi_G(V^i)$  of a variable. As a result,  $S_Y$  does not depend on the graph  $G$  being considered, and, therefore, we only need to consider transfer entropy when optimising KL divergence (24).  $\square$

As mentioned above, Corollary 1 is, in practice, equivalent to the maximum log-likelihood (5) and log-likelihood ratio (6) approaches. However, the statement only holds for constant embedding parameters. In the general case, where these parameters are unknown, one requires Theorem 4 to perform structure learning. Given this result, we can now confidently derive scoring functions from Corollary 1.

### 5. Application to Structure Learning

We now employ the results above in selecting a synchronous GDS that best fits data generated by a multivariate dynamical system. The most natural way to find an optimal model based on Theorem 4 is to minimise KL divergence. Here, we assume constant embedding parameters and use Corollary 1 to present the *transfer entropy score* and discuss some attributes of this score. We then use this scoring function as a subroutine for learning the structure of coupled Lorenz and Rössler attractors.

From Corollary 1, a naive scoring function can be defined as

$$g_{\text{TE}}(B : D) = \sum_{i=1}^M T_{\{Y^{ij}\}_j \rightarrow Y^i}.
 \tag{27}$$

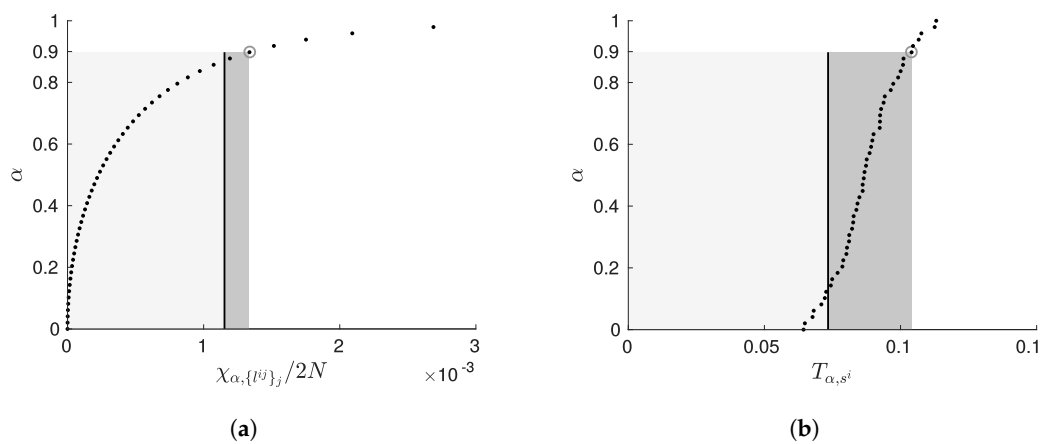
Given parameterised probability distributions, this score is insufficient, since the sum of transfer entropy in (27) is non-decreasing when including more parents in the graph [38]. Thus, we use statistical significance tests in our scoring functions to mitigate this issue.

### 5.1. Penalising Transfer Entropy by Independence Tests

Building on the maximum likelihood score (27), we propose using independence tests to define two new scores of practical value. Here, we draw on the result of de Campos [3], who derived a scoring function for BN structure learning based on conditional mutual information and statistical significance tests, called *MIT*. The central idea is to use collective transfer entropy  $T_{\langle Y^{ij} \rangle_j \rightarrow Y^i}$  to measure the degree of interaction between each subsystem  $V^i$  and its parent subsystems  $\Pi_G(V^i)$ , but also to penalise this term with a value based on significance testing. As with the *MIT* score, this gives a principled way to re-scale the transfer entropy when including more edges in the graph.

To develop our scores, we form a *null hypothesis*  $H_0$  that there is no interaction  $T_{\langle Y^{ij} \rangle_j \rightarrow Y^i}$ , and then compute a test statistic to penalise the measured transfer entropy. To compute the test statistic, it is necessary to consider the measurement distribution in the case where the hypothesis is true. Unfortunately, this distribution is only analytically tractable in the case of discrete and linear-Gaussian systems, where  $2NT_{\langle Y^{ij} \rangle_j \rightarrow Y^i}$  is known to asymptotically approach the  $\chi^2$ -distribution [48]. Since this distribution is a function of the parents of  $Y^i$ , we let it be described by the function  $\chi^2(\{I^{ij}\}_j)$ . Now, given this distribution, we can fix some *confidence level*  $\alpha$  and determine the value  $\chi_{\alpha, \{I^{ij}\}_j}$  such that  $p(\chi^2(\{I^{ij}\}_j) \leq \chi_{\alpha, \{I^{ij}\}_j})$ . This represents a conditional independence test: if  $2NT_{\langle Y^{ij} \rangle_j \rightarrow Y^i} \leq \chi_{\alpha, \{I^{ij}\}_j}$  then we accept the hypothesis of conditional independence between  $Y^i$  and  $\langle Y^{ij} \rangle_j$ ; otherwise, we reject it. We express this idea as the *TEA* score:

$$g_{TEA}(B : D) = \sum_{i=1}^M \left( 2NT_{\langle Y^{ij} \rangle_j \rightarrow Y^i} - \chi_{\alpha, \{I^{ij}\}_j} \right). \tag{28}$$



**Figure 3.** Distributions of the (a) *TEA* penalty function (28) and the (b) *TEE* penalty function (28). Both distributions were generated by observing the outcome of 1000 samples from two Gaussian variables with a correlation of 0.05. The figures illustrate: the distribution as a set of 100 sampled points (black dots); the area considered independent (grey regions); the measured transfer entropy (black line); and the difference between measurement and penalty term (dark grey region). Both tests use a value of  $\alpha = 0.9$  ( $p$ -value of 0.1). The distribution in (a) was estimated by assuming variables were linearly-coupled Gaussians, and the distribution in (b) was computed via a kernel box method (computed by the Java Information Dynamics Toolkit (JIDT), see [52] for details).

In general, we only have access to *continuous* measurements of dynamical systems, and so are limited by the discrete or linear-Gaussian assumption. We can, however, use *surrogate* measurements  $T_{\langle Y^{ij} \rangle_j^s \rightarrow Y^i}$  to empirically compute the distribution under the assumption of  $H_0$  [52]. This same technique has been used by [38] to derive a greedy structure learning algorithm for effective network analysis. Here,  $\langle Y^{ij} \rangle_j^s$  are surrogate sets of variables for  $\langle Y^{ij} \rangle_j$ , which have the same statistical

properties as  $\langle Y^{ij} \rangle_j$ , but the correlation between  $\langle Y^{ij} \rangle_j^s$  and  $Y^i$  is removed. Let the distribution of these surrogate measurements be represented by some general function  $T(s^i)$  where, for the discrete and linear-Gaussian systems, we could compute  $T(s^i)$  analytically as an independent set of  $\chi^2$ -distributions  $\chi^2(\{I^{ij}\}_j)$ . When no analytic distribution is known, we use a resampling method (i.e., permutation or bootstrapping), creating a large number of surrogate time-series pairs  $\{\langle Y^{ij} \rangle_j^s, Y^i\}$  by shuffling (for permutations, or redrawing for bootstrapping) the samples of  $Y^i$  and computing a population of  $T_{\langle Y^{ij} \rangle_j^s \rightarrow Y^i}$ . As with the TEA score, we fix some confidence level  $\alpha$  and determine the value  $T_{\alpha, s^i}$ , such that  $p(T(s^i) \leq T_{\alpha, s^i}) = \alpha$ . This results in the TEE scoring function as

$$g_{TEE}(B : D) = \sum_{i=1}^M \left( T_{\langle Y^{ij} \rangle_j \rightarrow Y^i} - T_{\alpha, s^i} \right). \tag{29}$$

We can obtain the value  $T_{\alpha, s^i}$  by (1) drawing  $S$  samples  $T_{\langle Y^{ij} \rangle_j^s \rightarrow Y^i}$  from the distribution  $T(s^i)$  (by permutation or bootstrapping), (2) fixing  $\alpha \in \{0, 1/S, 2/S, \dots, 1\}$ , and then (3) taking  $T_{\alpha, s^i}$  such that

$$\alpha = \frac{1}{S} \sum_{T_{\langle Y^{ij} \rangle_j \rightarrow Y^i}} \mathbb{1}_{T_{\langle Y^{ij} \rangle_j^s \rightarrow Y^i} \leq T_{\alpha, s^i}}.$$

We can alternatively limit the number of surrogates  $S$  to  $\lceil \alpha / (1 - \alpha) \rceil$  and take the maximum as  $T_{\alpha, s^i}$  [22]; however, taking a larger number of surrogates will improve the validity of the distribution  $T(s^i)$ .

Both the analytical (TEA) and empirical (TEE) scoring functions are illustrated in Figure 3. Note that the approach of significance testing is functionally equivalent to considering the log-likelihood ratio in (6), where, as stated, nested log-likelihoods (and thus transfer entropy) follows the above  $\chi^2$ -distribution [48].

### 5.2. Implementation Details and Algorithm Analysis

The two main implementation challenges that arise when performing structure learning are: (1) computing the score for every candidate network and (2) obtaining a sufficient number of samples to recover the network. The main contributions of this work are theoretical justifications for measures already in use and, fortunately, algorithmic performance has already been addressed extensively using various heuristics. Here, we present an exact, exhaustive implementation for the purpose of validating our theoretical contributions.

First, for computing collective transfer entropy for the score (29), we require CPDs to be estimated from data. Given these CPDs, collective transfer entropy (A8) decomposes as a sum of  $p$  conditional transfer entropy (A7) terms, where  $p = |\{Y^{ij}\}_j|$  is the size of the parent set (see Appendix B for details). Since most observations of dynamical systems are expected to be continuous, we employ a non-parametric, nearest-neighbour based approach to density estimation called the Kraskov–Stögbauer–Grassberger (KSG) estimator [43]. For any arbitrary decomposition of collective transfer entropy (i.e., any ordering of the parent set), this density estimation can be computed in time  $O(\kappa(p + 1)KN^{\kappa(p+1)} \log(N))$ , where  $K$  is the number of nearest neighbours for each observation in a dataset of size  $N$ , and  $\kappa$  is the embedding dimension [52]. We upper bound this as  $O(\kappa MKN^{\kappa M} \log(N))$  since the maximum  $p$  is  $M - 1$ .

Now, the above density estimation was described for an arbitrary ordering of the parent set. In the case of parametric (discrete or linear-Gaussian) density estimation, every permutation of the parent set yields equivalent results, with potentially different  $\chi_{\alpha, \{I^{ij}\}_j}$  values for each permutation [3]; however, this is not the case for non-parametric density estimation techniques, e.g., the KSG estimator. Hence, as a conservative estimate of the score, we compute all  $p!$  permutations of the parent set and take the minimum collective transfer entropy. In order to obtain the surrogate distribution, we require  $S$

uncorrelated samples of the density. Since the surrogate distributions decompose in a similar manner, the score for a candidate network can be computed in time  $O(S \cdot M! \cdot \kappa MKN^{\kappa M} \log(N))$ , where, again, we have upper bounded  $p!$  as  $M!$ .

Using this approach, we can now compute the score (29), and thus the optimal graph  $G^*$  can be found using any search procedure over DAGs. Exhaustive search, where all DAGs are enumerated, is typically intractable because the search space is super-exponential in the number of variables (about  $2^{O(M^2)}$ ), and so heuristics are often applied for efficiency. We restrict our attention to a relatively small network (a maximum of  $M = 4$  nodes) and thus we are able employ the dynamic programming (DP) approach of Silander and Myllymaki [53] to search through the space of all DAGs efficiently. This approach requires first computing the scores for all local parent sets, i.e.,  $2^M$  scores. Once each score is calculated, the DP algorithm runs in time  $o(M \cdot 2^{M-1})$  and the entire search procedure run in time  $O(M \cdot 2^{M-1} + 2^M \cdot S \cdot M! \cdot \kappa MKN^{\kappa M} \log(N))$ . As a consequence, the time complexity of the exhaustive algorithm is dominated by computing the  $2^M$  scores and, in smaller networks, most of the time is spent on density estimation for surrogate distributions.

Finally, the problem of inferring optimal embedding parameters is well studied in the literature. In our experimental evaluation, we set the embedding dimension to the maximum, i.e.,  $\kappa = 2d + 1$ , where  $d$  is the dimensionality of the entire latent state space (e.g., if  $M = 3$  and  $d^i = 3$  for each subsystem, then  $\kappa = 2 \sum_i d^i + 1 = 19$ ). However, determining these parameters would give more insight into the system and reduce the number of samples required for inference. There are numerous criteria for optimising these parameters (e.g., [54]); most notably, the work of [55] suggests an information-theoretic approach that could be integrated into the scoring function (29) to search over the embedding parameters and DAG space simultaneously.

## 6. Experimental Validation

The dynamics (1) and observation (2) maps can be obtained by either differential equations, discrete-time maps, or real-world measurements. To validate our approach, we use the toy example of distributed flows, whereby the dynamics of each node are given by either the Lorenz [56] or the Rössler system of ODEs [57]. The discrete-time measurements are obtained by integrating these ODEs over constant intervals. In this section, we formally introduce this model, study the effect of changing the parameters of a coupled Lorenz–Rössler system, and finally apply our scoring function to learn the structure of up to four coupled Lorenz attractors with arbitrary graph topology. To compute the scores, we use the Java Information Dynamics Toolkit (JIDT) [52], which includes both the KSG estimator and methods for generating the surrogate distributions.

### 6.1. Distributed Lorenz and Rössler Attractors

For validating our scoring function, we study coupled Lorenz and Rössler attractors. The Lorenz attractor exhibits chaotic solutions for certain parameter values and has been used to describe numerous phenomena of practical interest [56,58,59]. Each Lorenz system comprises three components ( $d^i = 3$ ), which we denote  $x = \langle u, v, w \rangle$ ; the state dynamics are given by:

$$\dot{x} = g(x) = \begin{cases} \dot{u} = \sigma(v - u), \\ \dot{v} = u(\rho - w) - v, \\ \dot{w} = uv - \beta w, \end{cases} \quad (30)$$

with free parameters  $\{\sigma, \rho, \beta\}$ . Similarly, the Rössler attractor has state dynamics given by:

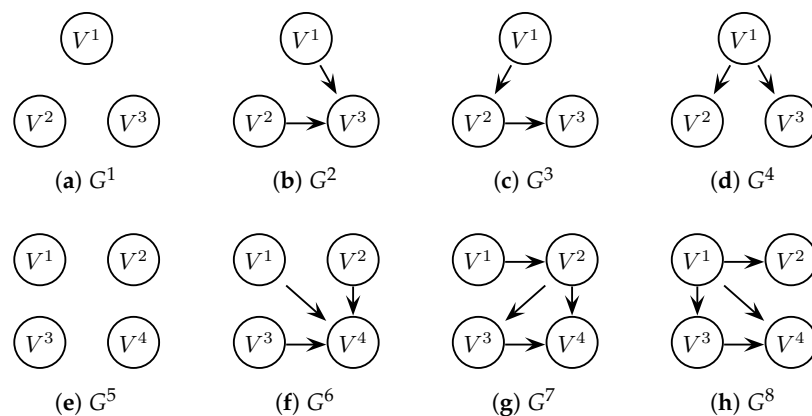
$$\dot{x} = g(x) = \begin{cases} \dot{u} = -y - z, \\ \dot{v} = x + ay, \\ \dot{w} = b + z(x - c), \end{cases} \quad (31)$$

with free parameters  $\{a, b, c\}$  [57].

In the distributed case, the components of each state vector  $x_t^i$  are also driven by components of another subsystem. A number of different schemes have been proposed for coupling these variables, e.g., using the product [21,60] and the difference [61,62] of components. Our model uses the latter approach of linear differencing between one or more subsystem variables to couple the network. Let  $\lambda$  denote the coupling strength,  $C$  denote a three-dimensional vector of binary values, and  $A$  denote an adjacency (coupling) matrix (i.e., an  $M \times M$  matrix of zeros with  $A_{ij} = 1$  iff  $V^i \in \Pi_G(V^j)$ ). Then, the state equations for  $M$  spatially distributed systems can be expressed as

$$\dot{x}_t^i = g^i(x_t^i) + v_f + \lambda C \sum_{j=1}^M A_{ij}(x_t^j - x_t^i), \tag{32}$$

where  $g^i(\cdot)$  represents the  $i$ -th chaotic attractor and  $v_f$  is additive noise. In our simulations, we use  $\lambda = 2$ ,  $C = \langle 1, 0, 0 \rangle$  (each subsystem is coupled via variable  $u$ ), and the adjacency matrices shown in Figure 4. In our experiments, we use common parameters for both attractors, i.e.,  $\sigma = 10, \beta = 8/3, \rho = 28$  and  $a = 0.1, b = 0.1, c = 14$ . For the observation  $y_t^i$ , it is common to use one component of the state as the read-out function [4,32,33]; we therefore let  $y_t^i = u_t^i + v_\psi$ . The noise terms are normally distributed with  $v_f \sim \mathcal{N}(0, \sigma_f)$  and  $v_\psi \sim \mathcal{N}(0, \sigma_\psi)$ . Figure 1 illustrates example trajectories of Lorenz–Lorenz attractors coupled via this model.



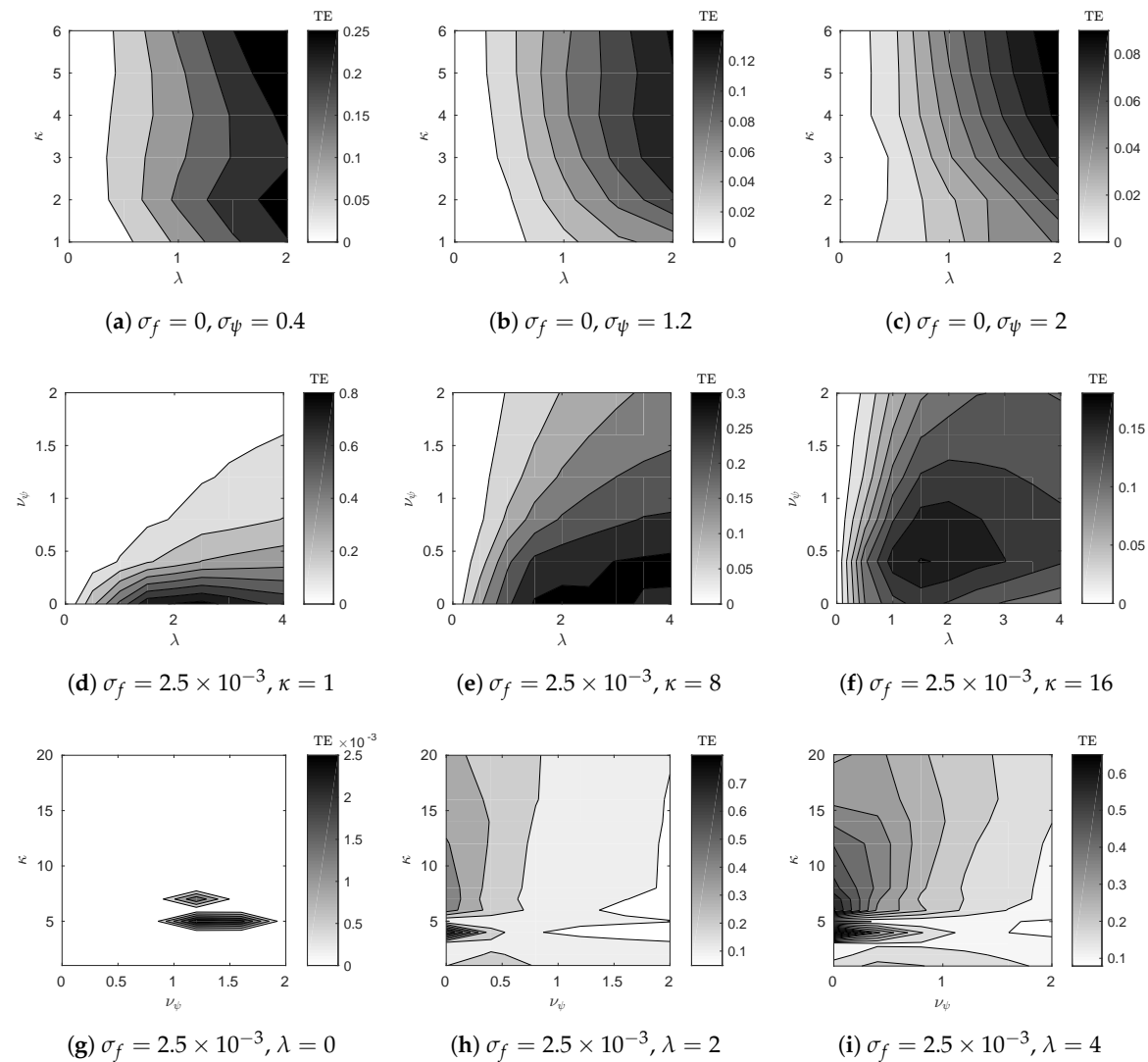
**Figure 4.** The network topologies used in this paper. The top row (a–d) are four arbitrary networks with three nodes ( $M = 3$ ) and the bottom row (e–h) are four arbitrary networks with four nodes ( $M = 4$ ).

### 6.2. Case Study: Coupled Lorenz–Rössler System

In order to characterise the effect of coupling on our score, we begin our evaluation by measuring the transfer entropy of a coupled Lorenz–Rössler attractor. In this setup,  $M = 2$ ,  $\Pi_G(V^1) = \emptyset$ , and  $\Pi_G(V^2) = V^1$ ,  $g^1(x)$  was given by (30), and  $g^2(x)$  was given by (31). The transfer entropy was computed with a finite sample size of  $N = 100,000$ .

Figure 5 shows the transfer entropy as a function of numerous parameters. In particular, the figure illustrates the effect of varying the coupling strength  $\lambda$ , embedding dimension  $\kappa$ , dynamics noise  $\sigma_f$ , and observation noise  $\sigma_\psi$ . As expected, increasing  $\lambda$ , or reducing either noise  $\sigma$ , increases the transfer entropy. The embedding dimension, however, increases to a set point, remains approximately constant, and then decreases. The  $\kappa$ -value above which transfer entropy remains constant illustrates the embedding dimension at which the dynamics are reconstructed; the decrease in transfer entropy after this point, however, is likely due to the finite sample size used for density estimation.





**Figure 5.** Transfer entropy as a function of the parameters of a coupled Lorenz–Rössler system. These components are: coupling strength  $\lambda$  and embedding dimension  $\kappa$  in the top row (a–c); coupling strength  $\lambda$  and observation noise  $\sigma_\psi$  in the middle row (d–f); and observation noise  $\sigma_\psi$  and embedding dimension  $\kappa$  in the bottom row (g–i).

There are two interesting features in Figure 5 due to the dynamical systems studied. First, in the bottom row (Figure 5g–i), there is a bifurcation around  $\kappa = 6$ . The theoretical embedding dimension for this system is  $\kappa = 2(d^1 + d^2) + 1 = 7$ , and, in this case, for  $\kappa < 6$ , the embedding does not suffice to reconstruct the dynamics. Second, in Figure 5i, the transfer entropy decreases after about  $\lambda = 2$ . This appears to be the case of synchrony due to strong coupling, where the dynamics of the forced variable become subordinate to the forcing [4], thus reducing the information transferred between the two subsystems.

### 6.3. Case Study: Network of Lorenz Attractors

In this section, we evaluate the score (27) in learning the structure of distributed dynamical systems. We will look at systems of three and four nodes of coupled Lorenz subsystems with arbitrary topologies. Unfortunately, significantly higher number of nodes become computationally expensive due to an increased embedding dimension  $\kappa$ , number of data points  $N$ , and number of permutations required to calculate the collective transfer entropy. To evaluate the performance of the score (27), the dynamics noise is constant  $\sigma_f = 0.01$ , whereas the observation noise  $\sigma_\psi$  and the number



**Table 2.**  $F_1$ -scores for four-node ( $M = 4$ ) networks. We present the classification summary for the three arbitrary topologies of coupled Lorenz systems represented by Figure 4f–h (network  $G^5$  has no edges and thus an undefined  $F_1$ -score). The  $p$ -value of the  $TEE$  score is given in the top row of each table, with  $\infty$  signifying using no significance testing, i.e., score (27).

Graph	$N$	$p = \infty$		$p = 0.01$		$p = 0.001$		$p = 0.0001$	
		$\sigma_\psi = 1$	$\sigma_\psi = 10$	$\sigma_\psi = 1$	$\sigma_\psi = 10$	$\sigma_\psi = 1$	$\sigma_\psi = 10$	$\sigma_\psi = 1$	$\sigma_\psi = 10$
$G^6$	5 K	0.57	0.5	0.57	0.29	0.57	0.29	0.57	-
	25 K	0.75	0.33	0.75	0.33	0.75	0.29	0.75	0.33
	100 K	1	0.33	1	0.57	1	0.4	1	0.33
$G^7$	5 K	1	0.25	1	0.29	0.75	0.25	0.75	0.57
	25 K	1	0.5	1	0.86	1	0.86	1	0.5
	100 K	1	0.86	1	0.86	1	0.86	1	0.86
$G^8$	5 K	1	0.25	1	0.57	1	0.75	1	0.25
	25 K	1	0.86	1	0.86	1	0.86	1	0.86
	100 K	1	0.86	1	0.86	1	0.57	1	0.86

Interestingly, the statistical significance testing does not have a strong effect on the results. It is unclear if this is due to the use of the non-parametric density estimators, which, in effect, are parsimonious in nature since transfer entropy will likely reduce when conditioning on more variables with a fixed samples size. One challenging case is the empty networks  $G^1$  and  $G^5$ ; this is shown in Appendix C, where the fallout is rarely 0 for any of the  $p$ -values or sample sizes (although a large number of observations  $N = 100$  K appears to reduce spurious edges). It would be expected that significance testing on these networks would outperform the naive score (27) given that a non-zero bias is introduced for a finite number of observations. Further investigation is required to understand why the null case fails.

## 7. Discussion and Future Work

We have presented a principled method to compute the KL divergence for model selection in distributed dynamical systems based on concepts from differential topology. The results presented in Figure 5 and Tables 1 and 2 illustrate that this approach is suitable for recovering synchronous GDSs from data. Further, KL divergence is related to model encoding, which is a fundamental measure used in complex systems analysis. Our result, therefore, has potential implications for other areas of research. For example, the notion of equivalence classes in BN structure learning [63] should lend insight into the area of effective network analysis [35,36].

More specifically, the approach proposed here complements explicit Bayesian identification and comparison of state space models. In DCM, and more generally in approximate Bayesian inference, models are identified in terms of their parameters via an optimisation of an approximate posterior density over model parameters with respect to a variational (free energy) bound on log evidence [64]. After these parameters have been identified, this bound can be used directly for model comparison and selection. Interestingly, free energy is derived from the KL divergence between the approximate and true posterior and thus automatically penalises more complex models; however, in Equation (8), these distributions are inverted. In future work, it would be interesting to explore the relationship between transfer entropy and the variational free energy bound. Specifically, computing an evidence bound directly from the transfer entropy may allow us to avoid the significance testing described in Section 5 and instead use an approximation to evidence for structure learning.

Multivariate extensions to transfer entropy are known to eliminate redundant pairwise relationships and take into account the influence of confounding relationships in a network (i.e., synergistic effects) [65,66]. In this work, we have shown that this intuition holds for distributed dynamical systems when confined to a DAG topology. We conjecture that these methods are also applicable when cyclic dependencies exist within a graph, given any generic observation can be used

in reconstructing the dynamics [50]; however, the methods presented are more likely to reveal *one* source in the cycle, rather than all information sources due to redundancy.

There are a number of extensions that should be considered for further practical implementations of this algorithm. Currently, we assume that the dimensionality of each subsystem is known, and thus we can bound the embedding dimension  $\kappa$  for recovering the hidden structure. However, this is generally infeasible in practice and a more general algorithm would infer the embedding dimension and time delay for an unknown system. Fortunately, there are numerous techniques to recover these parameters [54,55]. Furthermore, evaluating the quality of large graphs is infeasible with our current approach. However, our exact algorithm illustrates the feasibility of state space reconstruction in recovering a graph in practice. In the future, we aim to leverage the structure learning literature on reducing the search space and approximating scoring functions to produce more efficient algorithms.

Finally, the theoretical results of this work supplements understanding in fields where transfer entropy is commonly employed. Point processes are being increasingly viewed as models for a variety of information processing systems, e.g., as spiking neural trains [67] and adversaries in robotic patrolling models [68]. It was recently shown how transfer entropy can be computed for continuous time point processes such as these [67], allowing for efficient use of our analytical scoring function  $g_{TEA}$  in a number of contexts. Another intriguing line of research is the physical and thermodynamic interpretation of transfer entropy [69], particularly its relationship to the arrow of time [70]; this relationship between endomorphisms as discussed here and time asymmetry of thermodynamics should be explored further.

**Acknowledgments:** This work was supported in part by the Australian Centre for Field Robotics; the New South Wales Government; and the Faculty of Engineering & Information Technologies, The University of Sydney, under the Faculty Research Cluster Program. Special thanks go to Jürgen Jost, Michael Small, Joseph Lizier, and Wolfram Martens for their useful discussions.

**Author Contributions:** O.C, M.P. and R.F. conceived and designed the experiments; O.C. performed the experiments; O.C. and M.P. analyzed the data; O.C., M.P., and R.F. wrote the paper. All authors have read and approved the final manuscript.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A. Embedding Theory

We refer here to embedding theory as the study of inferring the (hidden) state  $x_n \in \mathcal{M}$  of a dynamical system from a sequence of observations  $y_n \in \mathbb{R}$ . This section will cover reconstruction theorems that define the conditions under which we can use delay embeddings for recovering the original dynamics  $f$  from this observed time series.

In differential topology, an *embedding* refers to a smooth map  $\Phi : \mathcal{M} \rightarrow \mathcal{N}$  between manifolds  $\mathcal{M}$  and  $\mathcal{N}$  if it maps  $\mathcal{M}$  diffeomorphically onto its image. In Takens' seminal work on turbulent flow [31], he proposed a map  $\Phi_{f,\psi} : \mathcal{M} \rightarrow \mathbb{R}^k$ , that is composed of delayed observations, can be used to reconstruct the dynamics for typical  $(f, \psi)$ . That is, fix some  $\kappa$  (the *embedding dimension*) and  $\tau$  (the *time delay*), the *delay embedding map*, given by

$$\Phi_{f,\psi}(x_n) = y_n^{(\kappa)} = \langle y_n, y_{n+\tau}, y_{n+2\tau}, \dots, y_{n+(\kappa-1)\tau} \rangle, \quad (\text{A1})$$

is an embedding. More formally, denote  $\Phi_{f,\psi} \in \mathcal{D}^r(\mathcal{M}, \mathbb{R}^k)$  as the space of  $C^r$ -diffeomorphisms on  $\mathcal{M}$  and  $C^r(\mathcal{M}, \mathbb{R})$  as the space of  $C^r$ -functions on  $\mathcal{M}$ , then the theorem can be expressed as follows.

**Theorem A1** (Delay Embedding Theorem for Diffeomorphisms [31]). *Let  $\mathcal{M}$  be a compact manifold of dimension  $d \geq 1$ . If  $\kappa \geq 2d + 1$  and  $r \geq 1$ , then there exists an open and dense set  $(f, \psi) \in \mathcal{D}^r(\mathcal{M}, \mathcal{M}) \times C^r(\mathcal{M}, \mathbb{R})$  for which the map  $\Phi_{f,\psi}$  is an embedding of  $\mathcal{M}$  into  $\mathbb{R}^k$ .*

The implication of Theorem A1 is that, for typical  $(f, \psi)$ , the image  $\Phi_{f,\psi}(\mathcal{M})$  of  $\mathcal{M}$  under the delay embedding map  $\Phi_{f,\psi}$  is completely equivalent to  $\mathcal{M}$  itself, apart from the smooth invertible

change of coordinates given by the mapping  $\Phi_{f,\psi}$ . An important consequence of this result is that we can define a map  $\mathbf{F} = \Phi_{f,\psi} \circ f \circ \Phi_{f,\psi}^{-1}$  on  $\Phi_{f,\psi}$ , such that  $y_{n+1}^{(\kappa)} = \mathbf{F}(y_n^{(\kappa)})$  [44]. The bound for the open and dense set referred to in Theorem A1 is given by a number of technical assumptions. Denote  $(Df)_x$  as the derivative of function  $f$  at a point  $x$  in the domain of  $f$ . The set of periodic points  $A$  of  $f$  with period less than  $\tau$  has finitely many points. In addition, the eigenvalues of  $(Df)_x$  at each  $x$  in a compact neighbourhood  $A$  are distinct and not equal to 1.

Theorem A1 was established for diffeomorphisms  $\mathcal{D}^r$ ; by definition, the dynamics are thus invertible in time. Thus, the time delay  $\tau$  in (A1) can be either positive (delay lags) or negative (delay leads). Takens later proved a similar result for endomorphisms, i.e., non-invertible maps that restricts the time delay to a negative integer. Denote by  $\mathcal{E}(\mathcal{M}, \mathcal{M})$  the set of the space of  $C^r$ -endomorphisms on  $\mathcal{M}$ , then the reconstruction theorem for endomorphisms can be expressed as the following.

**Theorem A2** (Delay Embedding Theorem for Endomorphisms [71]). *Let  $\mathcal{M}$  be a compact  $m$  dimensional manifold. If  $\kappa \geq 2d + 1$  and  $r \geq 1$ , then there exists an open and dense set  $(f, \psi) \in \mathcal{D}^r(\mathcal{M}, \mathcal{M}) \times C^r(\mathcal{M}, \mathbb{R})$  for which there is a map  $\pi_\kappa : \mathcal{X}_\kappa \rightarrow \mathcal{M}$  with  $\pi_\kappa \Phi_{f,\psi} = f^{\kappa-1}$ . Moreover, the map  $\pi_\kappa$  has bounded expansion or is Lipschitz continuous.*

As a result of Theorem A2, a sequence of  $\kappa$  successive measurements from a system determines the system state at the end of the sequence of measurements [71]. That is, there exists an endomorphism  $\mathbf{F} = \Phi_{f,\psi} \circ f \circ \Phi_{f,\psi}^{-1}$  to predict the next observation if one takes a negative time (lead) delay  $\tau$  in (A1).

In this work, we consider two important generalisations of the Delay Embedding Theorem A1. Both of these theorems follow similar proofs to the original and have thus been derived for diffeomorphisms, not endomorphisms. However, encouraging empirical results in [6] support the conjecture that they can both be generalised to the case of endomorphisms by taking a negative time delay, as is done in Theorem A2 above. This would allow for not only distributed flows that are used in our work, but endomorphic maps, e.g., the well-studied coupled map lattice structure [51].

The first generalisation is by Stark et al. [44] and deals with a skew-product system. That is,  $f$  is now forced by some second, independent system  $g : \mathcal{N} \rightarrow \mathcal{N}$ . The dynamical system on  $\mathcal{M} \times \mathcal{N}$  is thus given by the set of equations

$$x_{n+1} = f(x_n, \omega_n), \quad \omega_{n+1} = g(\omega_n). \tag{A2}$$

In this case, the delay map is written as

$$\Phi_{f,g,\psi}(x, \omega) = \langle y_n, y_{n+\tau}, y_{n+2\tau}, \dots, y_{n+(\kappa-1)\tau} \rangle, \tag{A3}$$

and the theorem can be expressed as follows.

**Theorem A3** (Bundle Delay Embedding Theorem [44]). *Let  $\mathcal{M}$  and  $\mathcal{N}$  be compact manifolds of dimension  $d \geq 1$  and  $e$ , respectively. Suppose that  $\kappa \geq 2(d + e) + 1$  and the periodic orbits of period  $\leq d$  of  $g \in \mathcal{D}^r(\mathcal{N})$  are isolated and have distinct eigenvalues. Then, for  $r \geq 1$ , there exists an open and dense set of  $(f, \psi) \in \mathcal{D}^r(\mathcal{M} \times \mathcal{N}, \mathcal{M}) \times C^r(\mathcal{M}, \mathbb{R})$  for which the map  $\Phi_{f,g,\psi}$  is an embedding of  $\mathcal{M} \times \mathcal{N}$  into  $\mathbb{R}^\kappa$ .*

Finally, all theorems up until now have assumed a single read-out function for the system in question. Recently, Sugihara et al. [4] showed that multivariate mappings also form an embedding, with minor changes to the technical assumptions underlying Takens' original theorem. That is, given  $M \leq 2d + 1$  different observation functions, the delay map can be written as

$$\Phi_{f,(\psi^i)}(x) = \langle \Phi_{f,\psi^1}(x), \Phi_{f,\psi^2}(x), \dots, \Phi_{f,\psi^M}(x) \rangle, \tag{A4}$$

where each delay map  $\Phi_{f,\psi^i}$  is as per (A1) for individual embedding dimension  $\kappa^i \leq \kappa$ . The theorem can then be stated as follows.

**Theorem A4** (Delay Embedding Theorem for Multivariate Observation Functions [50]). *Let  $\mathcal{M}$  be a compact manifold of dimension  $d \geq 1$ . Consider a diffeomorphism  $f \in \mathcal{D}^r(\mathcal{M}, \mathcal{M})$  and a set of at most  $2d + 1$  observation functions  $\langle \psi^i \rangle$  where each  $\psi^i \in C^r(\mathcal{M}, \mathbb{R})$  and  $r \geq 2$ . If  $\sum_i \kappa^i \geq 2d + 1$ , then, for generic  $(f, \langle \psi^i \rangle)$ , the map  $\Phi_{f, \langle \psi^i \rangle}$  is an embedding.*

### Appendix B. Information Theory

In this section, we introduce some key concepts of information theory: conditional entropy; conditional and collective transfer entropy; and stochastic interaction.

Consider two arbitrary random variables  $X$  and  $Y$ ; the conditional entropy  $H(X | Y)$  represents the uncertainty of  $X$  after taking into account the outcomes of another random variable  $Y$  by the equation

$$H(X | Y) = - \sum_{x,y} \Pr(x,y) \log \Pr(x | y) = \mathbf{E} [\Pr(x | y)]. \tag{A5}$$

Transfer entropy detects the directed exchange of information between random processes by marginalising out common history and static correlations between variables; it is thus considered a measure of information transfer within a system [25]. Let the processes  $X$  and  $Y$  have associated embedding dimensions  $\kappa^X$  and  $\kappa^Y$ . The transfer entropy of  $X$  to  $Y$  is given in terms of conditional entropy:

$$T_{X \rightarrow Y} = H(Y_{n+1} | Y_n^{(\kappa^Y)}) - H(Y_{n+1} | X_n^{i,(\kappa^X)}, Y_n^{(\kappa^Y)}). \tag{A6}$$

Now, given a third process  $Z$  with embedding dimension  $\kappa^Z$ , we can compute the information transfer of  $X$  to  $Y$  in the context of  $Z$  as:

$$T_{X \rightarrow Y | Z} = H(Y_{n+1} | Y_n^{(\kappa^Y)}, Z_n^{i,(\kappa^Z)}) - H(Y_{n+1} | X_n^{i,(\kappa^X)}, Y_n^{(\kappa^Y)}, Z_n^{i,(\kappa^Z)}). \tag{A7}$$

The collective transfer entropy computes the information transfer between a set of  $M$  source processes and a single destination process [19]. Consider the set  $\mathbf{Y} = \{Y^i\}$  of source processes. We can compute the collective transfer entropy from  $\mathbf{Y}$  to the destination process  $X$  as a function of conditional entropy (A5) terms:

$$T_{\mathbf{Y} \rightarrow X} = T_{Y^1 \rightarrow X} + \sum_{i=1}^M T_{Y^i \rightarrow X | \{Y^1, \dots, Y^{i-1}\}}, \tag{A8}$$

where the ordering of the source processes are arbitrary.

Stochastic interaction measures the complexity of dynamical systems by quantifying the excess of information processed, in time, by the system beyond the information processed by each of the nodes [17,18,72,73]. Using the same notation, stochastic interaction of the collection of processes  $\mathbf{Y}$  is

$$S_{\mathbf{Y}} = -H(\mathbf{Y}_{n+1} | \{Y_n^{i,(\kappa^i)}\}) + \sum_{i=1}^M H(Y_{n+1}^i | Y_n^{i,(\kappa^i)}). \tag{A9}$$

The standard definition assumes a first-order Markov process [17,18]; In (A9), we generalise stochastic interaction to arbitrary  $\kappa$ -order Markov chains.

### Appendix C. Extended Results

Here, we present the extended results of Tables 1 and 2. That is, we give the precision, recall, fallout, and  $F_1$ -scores for the eight networks of Lorenz attractors shown in Figure 4. These results are given for a number of different sample sizes to illustrate the sample complexity of this problem:



$N = 5000$  (Tables A1 and A2),  $N = 10,000$  (Tables A3 and A4),  $N = 25,000$  (Tables A5 and A6),  $N = 50,000$  (Tables A7 and A8), and  $N = 100,000$  (Tables A9 and A10). Each table has results for various  $p$ -values (with a  $p$ -value of  $\infty$  denoting the maximum likelihood score (27)), as well as two different observation noise variances,  $\sigma_\psi = 1$  and  $\sigma_\psi = 10$ .

**Table A1.** Classification results for three-node ( $M = 3$ ) networks for  $N = 5000$  samples. We present the precision (P), recall (R), fallout (F), and  $F_1$ -score for the eight arbitrary topologies of coupled Lorenz systems represented by Figure 4.

Graph	$p$ -Value	$\infty$		0.01		0.001		0.0001	
		$\sigma_\psi$	1	10	1	10	1	10	1
$G^1$	$R$	-	-	-	-	-	-	-	-
	$F$	0.33	0.22	0.33	0.22	0.22	0.33	0.33	0.22
	$P$	0	0	0	0	0	0	0	0
	$F_1$	-	-	-	-	-	-	-	-
$G^2$	$R$	1	0.5	1	0.5	1	0.5	1	0.5
	$F$	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
	$P$	0.67	0.5	0.67	0.5	0.67	0.5	0.67	0.5
	$F_1$	0.8	0.5	0.8	0.5	0.8	0.5	0.8	0.5
$G^3$	$R$	1	0.5	1	1	1	1	1	0.5
	$F$	0	0	0	0	0	0	0	0
	$P$	1	1	1	1	1	1	1	1
	$F_1$	1	0.67	1	1	1	1	1	0.67
$G^4$	$R$	1	0	1	1	1	0.5	1	0
	$F$	0.14	0.43	0.14	0.14	0.14	0.14	0.14	0.43
	$P$	0.67	0	0.67	0.67	0.67	0.5	0.67	0
	$F_1$	0.8	-	0.8	0.8	0.8	0.5	0.8	-

**Table A2.** Classification results for four-node ( $M = 4$ ) networks for  $N = 5000$  samples. We present the precision (P), recall (R), fallout (F), and  $F_1$ -score for the eight arbitrary topologies of coupled Lorenz systems represented by Figure 4.

Graph	$p$ -Value	$\infty$		0.01		0.001		0.0001	
		$\sigma_\psi$	1	10	1	10	1	10	1
$G^5$	$R$	-	-	-	-	-	-	-	-
	$F$	0.31	0.25	0.31	0.19	0.31	0.25	0.31	0.19
	$P$	0	0	0	0	0	0	0	0
	$F_1$	-	-	-	-	-	-	-	-
$G^6$	$R$	0.67	0.67	0.67	0.33	0.67	0.33	0.67	0
	$F$	0.15	0.23	0.15	0.23	0.15	0.23	0.15	0.31
	$P$	0.5	0.4	0.5	0.25	0.5	0.25	0.5	0
	$F_1$	0.57	0.5	0.57	0.29	0.57	0.29	0.57	-
$G^7$	$R$	1	0.25	1	0.25	0.75	0.25	0.75	0.5
	$F$	0	0.25	0	0.17	0.083	0.25	0.083	0.083
	$P$	1	0.25	1	0.33	0.75	0.25	0.75	0.67
	$F_1$	1	0.25	1	0.29	0.75	0.25	0.75	0.57
$G^8$	$R$	1	0.25	1	0.5	1	0.75	1	0.25
	$F$	0	0.25	0	0.083	0	0.083	0	0.25
	$P$	1	0.25	1	0.67	1	0.75	1	0.25
	$F_1$	1	0.25	1	0.57	1	0.75	1	0.25

**Table A3.** Classification results for three-node ( $M = 3$ ) networks for  $N = 10,000$  samples. We present the precision (P), recall (R), fallout (F), and  $F_1$ -score for the eight arbitrary topologies of coupled Lorenz systems represented by Figure 4.

Graph	$p$ -Value	$\infty$		0.01		0.001		0.0001	
		1	10	1	10	1	10	1	10
$G^1$	$\sigma_\psi$	-	-	-	-	-	-	-	-
	R	-	-	-	-	-	-	-	-
	F	0.22	0.11	0.22	0.11	0.22	0.22	0.22	0.11
	P	0	0	0	0	0	0	0	0
$G^2$	$F_1$	-	-	-	-	-	-	-	-
	R	1	0.5	1	0.5	1	0.5	1	0.5
	F	0	0.14	0	0.14	0	0.14	0	0.14
	P	1	0.5	1	0.5	1	0.5	1	0.5
$G^3$	$F_1$	1	0.5	1	0.5	1	0.5	1	0.5
	R	1	0.5	1	1	1	0	1	0.5
	F	0	0.14	0	0	0	0.29	0	0.14
	P	1	0.5	1	1	1	0	1	0.5
$G^4$	$F_1$	1	0.5	1	1	1	-	1	0.5
	R	1	1	1	0.5	1	0.5	1	1
	F	0.14	0.14	0	0	0.14	0.14	0.14	0.14
	P	0.67	0.67	1	1	0.67	0.5	0.67	0.67
$G^5$	$F_1$	0.8	0.8	1	0.67	0.8	0.5	0.8	0.8

**Table A4.** Classification results for four-node ( $M = 4$ ) networks for  $N = 10,000$  samples. We present the precision (P), recall (R), fallout (F), and  $F_1$ -score for the eight arbitrary topologies of coupled Lorenz systems represented by Figure 4.

Graph	$p$ -Value	$\infty$		0.01		0.001		0.0001	
		1	10	1	10	1	10	1	10
$G^5$	$\sigma_\psi$	-	-	-	-	-	-	-	-
	R	-	-	-	-	-	-	-	-
	F	0.31	0.25	0.31	0.19	0.31	0.19	0.31	0.25
	P	0	0	0	0	0	0	0	0
$G^6$	$F_1$	-	-	-	-	-	-	-	-
	R	0.67	0.33	0.67	0	1	1	0.67	0.33
	F	0.15	0.15	0.15	0.15	0.15	0.15	0.15	0.15
	P	0.5	0.33	0.5	0	0.6	0.6	0.5	0.33
$G^7$	$F_1$	0.57	0.33	0.57	-	0.75	0.75	0.57	0.33
	R	0.75	0.5	1	0.5	1	0.25	0.75	0.5
	F	0.083	0.083	0	0.083	0	0.17	0.083	0.083
	P	0.75	0.67	1	0.67	1	0.33	0.75	0.67
$G^8$	$F_1$	0.75	0.57	1	0.57	1	0.29	0.75	0.57
	R	1	0.25	1	0.25	1	0	1	0.25
	F	0	0.17	0	0.17	0	0.25	0	0.17
	P	1	0.33	1	0.33	1	0	1	0.33
$G^9$	$F_1$	1	0.29	1	0.29	1	-	1	0.29

**Table A5.** Classification results for three-node ( $M = 3$ ) networks for  $N = 25,000$  samples. We present the precision (P), recall (R), fallout (F), and  $F_1$ -score for the eight arbitrary topologies of coupled Lorenz systems represented by Figure 4.

Graph	$p$ -Value	$\infty$		0.01		0.001		0.0001	
		1	10	1	10	1	10	1	10
$G^1$	$\sigma_\psi$	-	-	-	-	-	-	-	-
	R	-	-	-	-	-	-	-	-
	F	0.22	0.11	0.22	0.11	0.22	0.22	0.22	0.11
	P	0	0	0	0	0	0	0	0
$G^2$	$F_1$	-	-	-	-	-	-	-	-
	R	1	1	1	0.5	1	0.5	1	1
	F	0	0.14	0	0.14	0	0.14	0	0.14
	P	1	0.67	1	0.5	1	0.5	1	0.67
$G^3$	$F_1$	1	0.8	1	0.5	1	0.5	1	0.8
	R	1	1	1	0.5	1	1	1	1
	F	0	0	0	0.14	0	0	0	0
	P	1	1	1	0.5	1	1	1	1
$G^4$	$F_1$	1	1	1	0.5	1	1	1	1
	R	1	1	1	1	1	0.5	1	1
	F	0	0	0	0	0	0.14	0	0
	P	1	1	1	1	1	0.5	1	1
$G^5$	$F_1$	1	1	1	1	1	0.5	1	1

**Table A6.** Classification results for four-node ( $M = 4$ ) networks for  $N = 25,000$  samples. We present the precision (P), recall (R), fallout (F), and  $F_1$ -score for the eight arbitrary topologies of coupled Lorenz systems represented by Figure 4.

Graph	$p$ -Value	$\infty$		0.01		0.001		0.0001	
		1	10	1	10	1	10	1	10
$G^5$	$\sigma_\psi$	-	-	-	-	-	-	-	-
	R	-	-	-	-	-	-	-	-
	F	0.31	0.19	0.31	0.19	0.31	0.19	0.31	0.19
	P	0	0	0	0	0	0	0	0
$G^6$	$F_1$	-	-	-	-	-	-	-	-
	R	1	0.33	1	0.33	1	0.33	1	0.33
	F	0.15	0.15	0.15	0.15	0.15	0.23	0.15	0.15
	P	0.6	0.33	0.6	0.33	0.6	0.25	0.6	0.33
$G^7$	$F_1$	0.75	0.33	0.75	0.33	0.75	0.29	0.75	0.33
	R	1	0.5	1	0.75	1	0.75	1	0.5
	F	0	0.17	0	0	0	0	0	0.17
	P	1	0.5	1	1	1	1	1	0.5
$G^8$	$F_1$	1	0.5	1	0.86	1	0.86	1	0.5
	R	1	0.75	1	0.75	1	0.75	1	0.75
	F	0	0	0	0	0	0	0	0
	P	1	1	1	1	1	1	1	1
$G^9$	$F_1$	1	0.86	1	0.86	1	0.86	1	0.86

**Table A7.** Classification results for three-node ( $M = 3$ ) networks with  $N = 50,000$  samples. We present the precision (P), recall (R), fallout (F), and  $F_1$ -score for the eight arbitrary topologies of coupled Lorenz systems represented by Figure 4.

Graph	$p$ -Value	$\infty$		0.01		0.001		0.0001	
		1	10	1	10	1	10	1	10
$G^1$	$\sigma_\psi$								
	R	-	-	-	-	-	-	-	-
	F	0	0.11	0	0	0	0.11	0	0.22
	P	-	0	-	-	-	0	-	0
$G^2$	$F_1$	-	-	-	-	-	-	-	-
	R	1	0.5	1	0.5	1	0.5	1	0.5
	F	0	0.14	0	0.14	0	0.14	0	0.14
	P	1	0.5	1	0.5	1	0.5	1	0.5
$G^3$	$F_1$	1	0.5	1	0.5	1	0.5	1	0.5
	R	1	1	1	0.5	1	1	1	1
	F	0	0.14	0	0.14	0	0.14	0	0
	P	1	0.67	1	0.5	1	0.67	1	1
$G^4$	$F_1$	1	0.8	1	0.5	1	0.8	1	1
	R	1	0.5	1	1	1	0.5	1	1
	F	0	0.14	0	0	0	0.14	0	0
	P	1	0.5	1	1	1	0.5	1	1
$G^5$	$F_1$	1	0.5	1	1	1	0.5	1	1

**Table A8.** Classification results for four-node ( $M = 4$ ) networks with  $N = 50,000$  samples. We present the precision (P), recall (R), fallout (F), and  $F_1$ -score for the eight arbitrary topologies of coupled Lorenz systems represented by Figure 4.

Graph	$p$ -Value	$\infty$		0.01		0.001		0.0001	
		1	10	1	10	1	10	1	10
$G^5$	$\sigma_\psi$								
	R	-	-	-	-	-	-	-	-
	F	0.19	0.062	0.19	0.19	0.19	0.12	0.19	0.12
	P	0	0	0	0	0	0	0	0
$G^6$	$F_1$	-	-	-	-	-	-	-	-
	R	1	0.33	1	0	1	0.33	1	0.33
	F	0	0.15	0	0	0	0.23	0.15	0.15
	P	1	0.33	1	-	1	0.25	0.6	0.33
$G^7$	$F_1$	1	0.33	1	-	1	0.29	0.75	0.33
	R	1	0.75	1	0.5	1	0.5	1	0.75
	F	0	0	0	0.17	0	0.083	0	0
	P	1	1	1	0.5	1	0.67	1	1
$G^8$	$F_1$	1	0.86	1	0.5	1	0.57	1	0.86
	R	1	0.75	1	0.75	1	0.75	1	0.75
	F	0	0	0	0	0	0	0	0
	P	1	1	1	1	1	1	1	1
$G^8$	$F_1$	1	0.86	1	0.86	1	0.86	1	0.86

**Table A9.** Classification results for three-node ( $M = 3$ ) networks with  $N = 100,000$  samples. We present the precision (P), recall (R), fallout (F), and  $F_1$ -score for the eight arbitrary topologies of coupled Lorenz systems represented by Figure 4.

Graph	$p$ -Value	$\infty$		0.01		0.001		0.0001	
		1	10	1	10	1	10	1	10
$G^1$	$\sigma_\psi$	-	-	-	-	-	-	-	-
	R	-	-	-	-	-	-	-	-
	F	0	0.22	0	0.11	0	0.22	0	0.11
	P	-	0	-	0	-	0	-	0
$G^2$	$F_1$	-	-	-	-	-	-	-	-
	R	1	0.5	1	1	1	1	1	1
	F	0	0.14	0	0	0	0	0	0.14
	P	1	0.5	1	1	1	1	1	0.67
$G^3$	$F_1$	1	0.5	1	1	1	1	1	0.8
	R	1	1	1	1	1	1	1	1
	F	0	0	0	0	0	0	0	0
	P	1	1	1	1	1	1	1	1
$G^4$	$F_1$	1	1	1	1	1	1	1	1
	R	1	1	1	1	1	1	1	1
	F	0	0	0	0	0	0	0	0
	P	1	1	1	1	1	1	1	1
$G^4$	$F_1$	1	1	1	1	1	1	1	1

**Table A10.** Classification results for four-node ( $M = 4$ ) networks with  $N = 100,000$  samples. We present the precision (P), recall (R), fallout (F), and  $F_1$ -score for the eight arbitrary topologies of coupled Lorenz systems represented by Figure 4.

Graph	$p$ -Value	$\infty$		0.01		0.001		0.0001	
		1	10	1	10	1	10	1	10
$G^5$	$\sigma_\psi$	-	-	-	-	-	-	-	-
	R	-	-	-	-	-	-	-	-
	F	0.19	0.062	0.19	0.062	0.19	0.19	0.19	0.12
	P	0	0	0	0	0	0	0	0
$G^6$	$F_1$	-	-	-	-	-	-	-	-
	R	1	0.33	1	0.67	1	0.33	1	0.33
	F	0	0.15	0	0.15	0	0.077	0	0.15
	P	1	0.33	1	0.5	1	0.5	1	0.33
$G^7$	$F_1$	1	0.33	1	0.57	1	0.4	1	0.33
	R	1	-	1	-	1	-	1	-
	F	0	-	0	-	0	-	0	-
	P	1	-	1	-	1	-	1	-
$G^8$	$F_1$	1	-	1	-	1	-	1	-
	R	1	0.75	1	0.75	1	0.5	1	0.75
	F	0	0	0	0	0	0.083	0	0
	P	1	1	1	1	1	0.67	1	1
$G^8$	$F_1$	1	0.86	1	0.86	1	0.57	1	0.86

## References

1. Akaike, H. Information theory and an extension of the maximum likelihood principle. In Proceedings of the Second International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, 2–8 September 1971; pp. 267–281.
2. Lam, W.; Bacchus, F. Learning Bayesian belief networks: An approach based on the MDL principle. *Comput. Intell.* **1994**, *10*, 269–293.
3. de Campos, L.M. A Scoring Function for Learning Bayesian Networks Based on Mutual Information and Conditional Independence Tests. *J. Mach. Learn. Res.* **2006**, *7*, 2149–2187.
4. Sugihara, G.; May, R.; Ye, H.; Hsieh, C.H.; Deyle, E.; Fogarty, M.; Munch, S. Detecting causality in complex ecosystems. *Science* **2012**, *338*, 496–500.
5. Vicente, R.; Wibral, M.; Lindner, M.; Pipa, G. Transfer entropy—A model-free measure of effective connectivity for the neurosciences. *J. Comput. Neurosci.* **2011**, *30*, 45–67.
6. Schumacher, J.; Wunderle, T.; Fries, P.; Jäkel, F.; Pipa, G. A statistical framework to infer delay and direction of information flow from measurements of complex systems. *Neural Comput.* **2015**, *27*, 1555–1608.
7. Best, G.; Cliff, O.M.; Patten, T.; Mettu, R.R.; Fitch, R. Decentralised Monte Carlo Tree Search for Active Perception. In Proceedings of the International Workshop on the Algorithmic Foundations of Robotics (WAFR), San Francisco, CA, USA, 18–20 December 2016.
8. Cliff, O.M.; Lizier, J.T.; Wang, X.R.; Wang, P.; Obst, O.; Prokopenko, M. Delayed Spatio-Temporal Interactions and Coherent Structure in Multi-Agent Team Dynamics. *Art. Life* **2017**, *23*, 34–57.
9. Best, G.; Forrai, M.; Mettu, R.R.; Fitch, R. Planning-aware communication for decentralised multi-robot coordination. In Proceedings of the International Conference on Robotics and Automation, Brisbane, Australia, 21 May 2018.
10. Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M.; Hwang, D.U. Complex networks: Structure and dynamics. *Phys. Rep.* **2006**, *424*, 175–308.
11. Mortveit, H.; Reidys, C. *An Introduction to Sequential Dynamical Systems*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2007.
12. Cliff, O.M.; Prokopenko, M.; Fitch, R. An Information Criterion for Inferring Coupling in Distributed Dynamical Systems. *Front. Robot. AI* **2016**, *3*, doi:10.3389/frobt.2016.00071.
13. Daly, R.; Shen, Q.; Aitken, J.S. Learning Bayesian networks: Approaches and issues. *Knowl. Eng. Rev.* **2011**, *26*, 99–157.
14. Chickering, D.M. Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res.* **2002**, *2*, 445–498.
15. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464.
16. Rissanen, J. Modeling by shortest data description. *Automatica* **1978**, *14*, 465–471.
17. Ay, N.; Wennekers, T. Temporal infomax leads to almost deterministic dynamical systems. *Neurocomputing* **2003**, *52*, 461–466.
18. Ay, N. Information geometry on complexity and stochastic interaction. *Entropy* **2015**, *17*, 2432–2458.
19. Lizier, J.T.; Prokopenko, M.; Zomaya, A.Y. Information modification and particle collisions in distributed computation. *Chaos* **2010**, *20*, 037109, doi:10.1063/1.3486801.
20. Murphy, K. *Dynamic Bayesian Networks: Representation, Inference and Learning*. Ph.D. Thesis, UC Berkeley, Berkeley, CA, USA, 2002.
21. Kocarev, L.; Parlitz, U. Generalized synchronization, predictability, and equivalence of unidirectionally coupled dynamical systems. *Phys. Rev. Lett.* **1996**, *76*, 1816–1819.
22. Kantz, H.; Schreiber, T. *Nonlinear Time Series Analysis*; Cambridge University Press: Cambridge, UK, 2004.
23. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*; Morgan Kaufmann: Burlington, MA, USA, 2014.
24. Granger, C.W.J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **1969**, *37*, 424–438.
25. Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **2000**, *85*, 461–464.
26. Barnett, L.; Barrett, A.B.; Seth, A.K. Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables. *Phys. Rev. Lett.* **2009**, *103*, e238701.



27. Lizier, J.T.; Prokopenko, M. Differentiating information transfer and causal effect. *Eur. Phys. J. B* **2010**, *73*, 605–615.
28. Smirnov, D.A. Spurious causalities with transfer entropy. *Phys. Rev. E* **2013**, *87*, 042917.
29. James, R.G.; Barnett, N.; Crutchfield, J.P. Information flows? A critique of transfer entropies. *Phys. Rev. Lett.* **2016**, *116*, 238701.
30. Liang, X.S. Information flow and causality as rigorous notions *ab initio*. *Phys. Rev. E* **2016**, *94*, 052201.
31. Takens, F. Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence; Lecture Notes in Math*; Springer: Berlin/Heidelberg, Germany, 1981; Volume 898, pp. 366–381.
32. Stark, J. Delay embeddings for forced systems. I. Deterministic forcing. *J. Nonlinear Sci.* **1999**, *9*, 255–332.
33. Stark, J.; Broomhead, D.S.; Davies, M.E.; Huke, J. Delay embeddings for forced systems. II. Stochastic forcing. *J. Nonlinear Sci.* **2003**, *13*, 519–577.
34. Valdes-Sosa, P.A.; Roebroek, A.; Daunizeau, J.; Friston, K. Effective connectivity: influence, causality and biophysical modeling. *Neuroimage* **2011**, *58*, 339–361.
35. Sporns, O.; Chialvo, D.R.; Kaiser, M.; Hilgetag, C.C. Organization, development and function of complex brain networks. *Trends Cogn. Sci.* **2004**, *8*, 418–425.
36. Park, H.J.; Friston, K. Structural and functional brain networks: From connections to cognition. *Science* **2013**, *342*, 1238411.
37. Friston, K.; Moran, R.; Seth, A.K. Analysing connectivity with Granger causality and dynamic causal modelling. *Curr. Opin. Neurobiol.* **2013**, *23*, 172–178.
38. Lizier, J.T.; Rubinov, M. *Multivariate Construction of Effective Computational Networks from Observational Data*; Preprint 25/2012; Max Planck Institute for Mathematics in the Sciences: Leipzig, Germany, 2012.
39. Sandoval, L. Structure of a global network of financial companies based on transfer entropy. *Entropy* **2014**, *16*, 4443–4482.
40. Rodewald, J.; Colombi, J.; Oyama, K.; Johnson, A. Using Information-theoretic Principles to Analyze and Evaluate Complex Adaptive Supply Network Architectures. *Procedia Comput. Sci.* **2015**, *61*, 147–152.
41. Crosato, E.; Jiang, L.; Lecheval, V.; Lizier, J.T.; Wang, X.R.; Tichit, P.; Theraulaz, G.; Prokopenko, M. Informative and misinformative interactions in a school of fish. *arXiv* **2017**, arXiv:1705.01213.
42. Kozachenko, L.; Friston, L.F.; Leonenko, N.N. Sample estimate of the entropy of a random vector. *Probl. Peredachi Inf.* **1987**, *23*, 9–16.
43. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138.
44. Stark, J.; Broomhead, D.S.; Davies, M.E.; Huke, J. Takens embedding theorems for forced and stochastic systems. *Nonlinear Anal. Theory Methods Appl.* **1997**, *30*, 5303–5314.
45. Friedman, N.; Murphy, K.; Russell, S. Learning the structure of dynamic probabilistic networks. In Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, Madison, WI, USA, 24–26 July 1998; pp. 139–147.
46. Koller, D.; Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*; MIT Press: Cambridge, MA, USA, 2009.
47. Wilks, S.S. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Stat.* **1938**, *9*, 60–62.
48. Barnett, L.; Bossomaier, T. Transfer entropy as a log-likelihood ratio. *Phys. Rev. Lett.* **2012**, *109*, 138105.
49. Vinh, N.X.; Chetty, M.; Coppel, R.; Wangikar, P.P. GlobalMIT: Learning globally optimal dynamic Bayesian network with the mutual information test criterion. *Bioinformatics* **2011**, *27*, 2765–2766.
50. Deyle, E.R.; Sugihara, G. Generalized theorems for nonlinear state space reconstruction. *PLoS ONE* **2011**, *6*, e18295.
51. Lloyd, A.L. The coupled logistic map: a simple model for the effects of spatial heterogeneity on population dynamics. *J. Theor. Biol.* **1995**, *173*, 217–230.
52. Lizier, J.T. JIDT: An information-theoretic toolkit for studying the dynamics of complex systems. *Front. Robot. AI* **2014**, *1*, doi:10.3389/frobt.2014.00011.
53. Silander, T.; Myllymaki, P. A simple approach for finding the globally optimal Bayesian network structure. In Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence, Cambridge, MA, USA, 13–16 July 2006; pp. 445–452.
54. Ragwitz, M.; Kantz, H. Markov models from data by simple nonlinear time series predictors in delay embedding spaces. *Phys. Rev. E* **2002**, *65*, 056201.

55. Small, M.; Tse, C.K. Optimal embedding parameters: A modelling paradigm. *Physica* **2004**, *194*, 283–296.
56. Lorenz, E.N. Deterministic nonperiodic flow. *J. Atmos. Sci.* **1963**, *20*, 130–141.
57. Rössler, O.E. An equation for continuous chaos. *Phys. Lett. A* **1976**, *57*, 397–398.
58. Haken, H. Analogy between higher instabilities in fluids and lasers. *Phys. Lett. A* **1975**, *53*, 77–78.
59. Cuomo, K.M.; Oppenheim, A.V. Circuit implementation of synchronized chaos with applications to communications. *Phys. Rev. Lett.* **1993**, *71*, 65–68.
60. He, R.; Vaidya, P.G. Analysis and synthesis of synchronous periodic and chaotic systems. *Phys. Rev. A* **1992**, *46*, 7387–7392.
61. Fujisaka, H.; Yamada, T. Stability theory of synchronized motion in coupled-oscillator systems. *Prog. Theor. Phys.* **1983**, *69*, 32–47.
62. Rulkov, N.F.; Sushchik, M.M.; Tsimring, L.S.; Abarbanel, H.D. Generalized synchronization of chaos in directionally coupled chaotic systems. *Phys. Rev. E* **1995**, *51*, 980–994.
63. Acid, S.; de Campos, L.M. Searching for Bayesian network structures in the space of restricted acyclic partially directed graphs. *J. Artif. Intell. Res.* **2003**, *18*, 445–490.
64. Friston, K.; Kilner, J.; Harrison, L. A free energy principle for the brain. *J. Physiol. Paris* **2006**, *100*, 70–87.
65. Williams, P.L.; Beer, R.D. Generalized measures of information transfer. *arXiv* **2011**, arXiv:1102.1507.
66. Vakorin, V.A.; Krakovska, O.A.; McIntosh, A.R. Confounding effects of indirect connections on causality estimation. *J. Neurosci. Methods* **2009**, *184*, 152–160.
67. Spinney, R.E.; Prokopenko, M.; Lizier, J.T. Transfer entropy in continuous time, with applications to jump and neural spiking processes. *Phys. Rev. E* **2017**, *95*, 032319.
68. Hefferan, B.; Cliff, O.M.; Fitch, R. Adversarial Patrolling with Reactive Point Processes. In Proceedings of the Australasian Conference on Robotics and Automation (ACRA), Brisbane, Australia, 5–7 December 2016.
69. Prokopenko, M.; Einav, I. Information thermodynamics of near-equilibrium computation. *Phys. Rev. E* **2015**, *91*, 062143.
70. Spinney, R.E.; Lizier, J.T.; Prokopenko, M. Transfer entropy in physical systems and the arrow of time. *Phys. Rev. E* **2016**, *94*, 022135.
71. Takens, F. The reconstruction theorem for endomorphisms. *Bull. Braz. Math. Soc.* **2002**, *33*, 231–262.
72. Ay, N.; Wennekers, T. Dynamical properties of strongly interacting Markov chains. *Neural Netw.* **2003**, *16*, 1483–1497.
73. Edlund, J.A.; Chaumont, N.; Hintze, A.; Koch, C.; Tononi, G.; Adami, C. Integrated information increases with fitness in the evolution of animats. *PLoS Comput. Biol.* **2011**, *7*, e1002236.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).