Identifying key drivers in a stochastic dynamical system through estimation of transfer entropy between univariate and multivariate time series

Julian Lee [*]

Department of Bioinformatics and Life Science, Soongsil University, 06978 Seoul, Korea

(Received 1 October 2024; revised 11 January 2025; accepted 21 January 2025; published 13 February 2025)

Transfer entropy (TE) is a widely used tool for quantifying causal relationships in stochastic dynamical systems. Traditionally, TE and its conditional variants are applied pairwise between dynamic variables to infer these relationships. However, identifying key drivers in such systems requires a measure of the causal influence exerted by each component on the entire system. I propose using outgoing transfer entropy (OutTE), the transfer entropy from a given variable to the collection of remaining variables, to quantify the causal influence of the variable on the rest of the system. Conversely, the incoming transfer entropy (InTE) is also defined to quantify the causal influence received by a component from the rest of the system. Since OutTE and InTE involve transfer entropy between univariate and multivariate time series, naive estimation methods can result in significant errors, especially when the number of variables is large relative to the number of samples. To address this, I introduce a novel estimation scheme that computes outgoing and incoming TE only between significantly interacting partners. The feasibility and effectiveness of this approach are demonstrated using synthetic data and real oral microbiota data. The method successfully identifies the bacterial species known to be key players in the bacterial community, highlighting its potential for uncovering causal drivers in complex systems.

DOI: 10.1103/PhysRevE.111.024308

I. INTRODUCTION

Causal inference in interacting dynamic systems is a critical area of study that aims to understand the cause-and-effect relationships within complex systems, such as those found in neuroscience, economics, ecology, and biology [1–22]. The goal is to distinguish true causal interactions from mere correlations, which is essential for predicting system behavior and designing effective interventions. Over the past several decades, frameworks like Granger causality [1] and transfer entropy [2] have been developed to address these challenges. In particular, transfer entropy (TE) [2] is a model-free, information-theoretic measure of causal influence between two time series. TE uniquely detects nonlinear and asymmetric interactions, making it ideal for complex systems where traditional linear methods fall short. The concept of TE has been extended to conditional transfer entropy, also known as causation entropy, to quantify the direct causal relationship between a pair of components in a dynamic system consisting of many interacting components [12–21]. Although there are nuances in interpreting transfer entropy and its conditional variant as information flow or causal influence quantification [23], these measures have been widely used to uncover causal relationships in complex systems, including neural networks [24–26], social networks [27], and gene regulatory networks [28,29].

In all these applications, the causal relationships between pairs of components in the system were estimated. Influential components in the system were identified within the framework of such pairwise relationships. A directed binary network was constructed, where each variable was treated as a node, and a directed edge was generated between nodes with statistically significant values of transfer entropy or conditional transfer entropy. Nodes with a number of outgoing links significantly larger than the average were then identified as the most influential nodes. For instance, key regulatory genes in gene regulatory networks have been identified in this manner by analyzing single-cell RNA sequencing data [28,29]. However, constructing a binary network disregards the actual values of transfer entropy or its conditional variant. It is possible that a component exerts a strong influence on only a few other components, which in turn influence others in a hierarchical manner. Key components in such a hierarchical structure cannot be identified by simply counting the number of outgoing links in a directed network without considering their weights. A straightforward approach to incorporating weights when computing the outward influence of a node in a weighted network is to sum the weights of the outgoing edges. Yet, as elaborated later, transfer entropy and its conditional variant are not additive quantities. From an information-theoretic perspective, the causal influence of a component on the rest of the system is better quantified by the transfer entropy from the corresponding variable to the rest of the system, which I will refer to as “outgoing TE (OutTE).” Here, the system involves two variables: a univariate source variable and a multivariate target variable. Components with OutTE values significantly higher than the average can be identified as key components that exert a strong influence on the rest of the system. Similarly, “incoming TE (InTE),” defined as the transfer entropy from the rest of the system to a given component, quantifies the causal influence exerted on

[*] Contact author: jul@ssu.ac.kr

the component by the rest of the system. As will be elaborated further, naive estimates of OutTE or InTE can lead to significant estimation errors, especially when the number of variables is comparable to or larger than the length of the time series. In this work I will introduce the novel estimation method for the OutTE and InTE, where estimation is performed only between significantly interacting partners.

The outline of the paper is as follows. In Sec. II, I will review the concept of TE and conditional TE, and introduce OutTE and InTE. In Sec. III, I will illustrate the estimation problems of OutTE and InTE using synthetic data from simple models, showing how estimation errors grow as the number of unrelated variables increases. In Sec. IV I will propose a novel estimation method for OutTE and InTE which overcomes the estimation problem. In Sec. V I will apply my method to microbiota data, showing that the method successfully identifies the bacterial species known to be key players in the bacterial community, where traditional network centrality measures fail. Section VI concludes the paper.

II. OUTGOING AND INCOMING TRANSFER ENTROPY

Information theory offers rigorous foundation for causal inference by quantifying the information shared between a pair of variables X and Y through mutual information $I(X,Y)$ defined as

$$I(X,Y) \equiv \left\langle \log_2 \left(\frac{P(X,Y)}{P(X)P(Y)} \right) \right\rangle \tag{1}$$

where $P(X)$ and $P(Y)$ are the marginal probability distributions for the random variables X and Y, $P(X,Y)$ are their joint probability distribution, and $\langle\rangle$ denotes the expectation value. In systems with more than two variables, one often seeks the direct correlation between X and Y after accounting for other variables. Denoting all variables other than X and Y by Z, the conditional mutual information $I(X,Y/Z)$ is defined as

$$I(X,Y/Z) \equiv \left\langle \log_2 \left(\frac{P(X,Y/Z)}{P(X/Z)P(Y/Z)} \right) \right\rangle, \tag{2}$$

which quantifies the direct correlation between X and Y.

In dynamic systems, given two time series $X(t)$ and $Y(t)$, where the integer t is an index for a discretized time, one might initially attempt to quantity the causal influence of X on Y using $I[X_-, Y(t)]$, where $X_- \equiv [X(t-1), X(t-2), \dots X(t-L)]$ represents the past history of X, where L is the maximum time lag to be considered. However, as previously mentioned, $I[X_-, Y(t)]$ only quantifies the shared information between the history of X and the present state of Y, not the causal influence of X on Y. It is possible that Y actually causes X, which could still result in a nonzero value of $I[X_-, Y(t)]$. Therefore, to quantify the true causal influence of X on Y, we must correct for the effect due to the history of Y, using the measure

$$T_{X\rightarrow Y} \equiv I[X_-, Y(t)/Y_-]. \tag{3}$$

This is known as transfer entropy (TE) from X to Y [2]. Transfer entropy can also be expressed as:

$$T_{X\rightarrow Y} = H[Y(t)/Y_-] - H[Y(t)/X_-, Y_-], \tag{4}$$

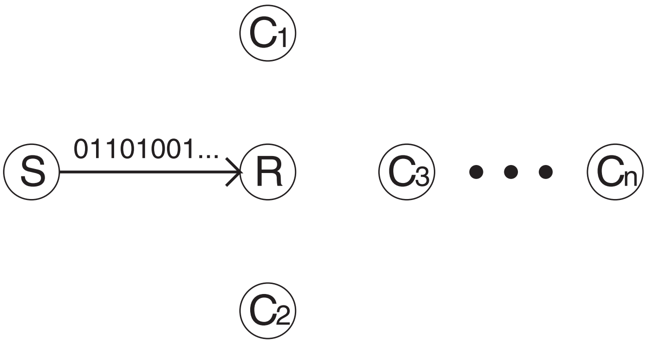

FIG. 1. The SR model with a sender variable S and a receiver variable R. The variables $C_1, \dots C_n$ represent confounding variables that are independent of the dynamics of S and R.

where

$$\begin{aligned} H[Y(t)/Y_-] &\equiv \langle -\log_2 P[Y(t)/Y_-]\rangle \\ H[Y(t)/X_-, Y_-] &\equiv \langle -\log_2 P[Y(t)/X_-, Y_-]\rangle. \end{aligned} \tag{5}$$

Here, the conditional entropies $H[Y(t)/Y_-]$ and $H[Y(t)/X_-, Y_-]$ represent the uncertainties of $Y(t)$ after observing the history of Y and after observing the histories of both X and Y, respectively. Thus, $T_{X\rightarrow Y}$ can be interpreted as the reduction in the uncertainty of $Y(t)$ after observing the history of X, given that we already know the history of Y. It can be proved that a conditional entropy is non-negative [30], and consequently,

$$H[Y(t)/Y_-] \geqslant 0, \quad H[Y(t)/X_-, Y_-] \geqslant 0.$$

It can also be shown that

$$H[Y(t)/Y_-] \geqslant H[Y(t)/X_-, Y_-],$$

so that $TE_{X\rightarrow Y} \geqslant 0$. Specifically, this implies that if $Y(t)$ is completely determined by Y_- such that $H[Y(t)/Y_-] = 0$, then $T_{X\rightarrow Y} = 0$. The intuitive interpretation is clear: If there is no uncertainty on $Y(t)$ after we observe Y_-, then obviously there is no additional uncertainty to be removed by knowing X_-.

In systems with more than two time series, we are often interested in the direct causal influence of X on Y after accounting for indirect effects due to other variables. Denoting the multivariate time series of variables other than X and Y by Z, the direct causal influence of X on Y is quantified by the measure [12–15]:

$$\begin{aligned} T_{(X\rightarrow Y)/Z} &\equiv I[X_-, Y(t)/Y_-, Z_-] \\ &= H[Y(t)/Y_-, Z] - H[Y(t)/X_-, Y_-, Z]. \end{aligned} \tag{6}$$

This measure is called multivariate transfer entropy, conditional transfer entropy, causation entropy, or full conditional mutual information [12–21].

As a simple example to illustrate the concept of transfer entropy, consider a system consisting of two variables: a sender variable S and a recipient variable R. I will refer to this system as the SR model (Fig. 1). The variable $S(t)$ randomly takes the value of either 0 or 1 with equal probability. $R(0)$ is also random, but for $t > 0$, $R(t)$ is fully determined by $S(t-1)$ and $R(t-1)$ according to the rule $R(t) = S(t-1) \oplus R(t-1)$, where $\oplus$ denotes the exclusive

OR operation. The dynamics here is Markovian, so we have $S_- = S(t-1)$ and $R_- = R(t-1)$. Without knowledge of the past of S, the dynamics of the recipient variable appears entirely random, leading to $H[R(t)/R_-] = 1$ bit. Furthermore, since $R(t)$ is fully determined by $S(t-1)$ and $R(t-1)$, we have $H[R(t)/R_-, S_-] = 0$, resulting in $T_{S\to R} = 1$ bit. In contrast, since $H[S(t)/S_-] = H[S(t)/R_-, S_-] = 1$ bit, it follows that $T_{R\to S} = 0$.

The focus of this work is on identifying key variables that exert significant influence on the rest of a stochastic dynamical system consisting of many interacting components. To achieve this, we compute OutTE for each variable X, $\mathrm{OutTE}(X) \equiv T_{X\to\mathrm{rest}}$, where "rest" denotes the collection of all variables except X. In this context, the "rest" is a multivariate target, whereas the source variable X is univariate. Variables with exceptionally large OutTE values are identified as key influencers of the system's overall dynamics. Conversely, the incoming transfer entropy (InTE) is defined as $\mathrm{InTE}(X) \equiv T_{\mathrm{rest}\to X}$, quantifying the causal influence exerted on X by the rest of the system.

It is important to note that neither $\mathrm{OutTE}(X)$ nor $\mathrm{InTE}(X)$ can be decomposed into the sum of bivariate (conditional) transfer entropies, unless the dynamics of remaining variables are completely decoupled from one another. For example, suppose we have three variables S, R_1, and R_2, and $R_1(t) = R_2(t)$ at all times. Let us assume that the dynamics of S and R_1 (R_2) is the same as that of the SR model described above. We then have

$$\mathrm{OutTE}(S) = T_{S\to(R_1,R_2)} = T_{S\to R_1} = T_{S\to R_2} = 1.$$

Also, we have

$$\begin{aligned} T_{S\to R_1/R_2} &= I[S(t-1), R_1(t)/R_1(t-1), R_2(t-1)] \\ &= I[S(t-1), R_1(t)/R_1(t-1)] = T_{S\to R_1}, \\ T_{S\to R_2/R_1} &= I[S(t-1), R_2(t)/R_1(t-1), R_2(t-1)] \\ &= I[S(t-1), R_2(t)/R_2(t-1)] = T_{S\to R_2}. \end{aligned} \tag{7}$$

Therefore, $\mathrm{TE}_{S\to R_1} + \mathrm{TE}_{S\to R_2} = \mathrm{TE}_{S\to R_1/R_2} + \mathrm{TE}_{S\to R_2/R_2} = 2$ bits, leading to

$$\begin{aligned} &\mathrm{OutTE}(S) \neq T_{S\to R_1} + T_{S\to R_2}, \\ &\mathrm{OutTE}(S) \neq T_{S\to R_1/R_2} + T_{S\to R_2/R_1}. \end{aligned} \tag{8}$$

As another example, suppose now we have two copies of sender variables $S_1(t) = S_2(t)$, and the receiver variable $R(t)$. Then we have

$$\mathrm{InTE}(R) = T_{(S_1,S_1)\to R} = T_{S_1\to R} = T_{S_2\to R} = 1.$$

Since $H[R(t)/R(t-1), S_1(t-1)] = H[R(t)/R(t-1), S_2(t-1)] = H[R(t)/R(t-1), S_1(t-1), S_2(t-1)]$, we have

$$\begin{aligned} T_{S_1\to R/S_2} &= H[R(t)/R(t-1), S_2(t-1)] \\ &\quad - H[R(t)/R(t-1), S_1(t-1), S_2(t-1)] = 0, \\ T_{S_2\to R_2/S_1} &= H[R(t)/R(t-1), S_1(t-1)] \\ &\quad - H[R(t)/R(t-1), S_1(t-1), S_2(t-1)] = 0. \end{aligned} \tag{9}$$

Therefore, we have

$$\begin{aligned} &\mathrm{InTE}(R) \neq T_{S_1\to R} + T_{S_2\to R}, \\ &\mathrm{InTE}(R) \neq T_{S_1\to R_1/S_2} + T_{S_2\to R_2/S_1}. \end{aligned} \tag{10}$$

III. ESTIMATION PROBLEM OF OUTGOING AND INCOMING TRANSFER ENTROPY

In practice, the true values of OutTE and InTE are not available and must be *estimated* from data, which can introduce estimation errors. Recall that $T_{X\to Y}$, $H[Y(t)/Y_-]$, and $H[Y(t)/X_-, Y_-]$ are all non-negative quantities. Assume $T_{X\to Y} > 0$, so $H[Y(t)/Y_-] > 0$. Also, let X be univariate and Y multivariate with dimension D_Y. On one hand, the estimator $\hat{H}[Y(t)/Y_-]$ tends to underestimate the true value $H[Y(t)/Y_-]$ when $D_Y \gtrsim N_{\mathrm{samp}}$, where N_{samp} is the number of samples, due to insufficient observation of events (Appendix). From the fact that $\widehat{H}[Y(t)/Y_-] \geqslant \widehat{H}[Y(t)/X_-, Y_-]$, we see that if the underestimation is so severe that $\hat{H}[Y(t)/Y_-] \simeq 0$, then $\hat{H}[Y(t)/X_-, Y_-] \simeq 0$, leading to $\hat{T}_{X\to Y} = \hat{H}[Y(t)/Y_-] - \hat{H}[Y(t)/X_-, Y_-] \simeq 0$ even if $T_{X\to Y} > 0$. That is, $\hat{T}_{X\to Y}$ underestimates the true value of $T_{X\to Y}$. On the other hand, the estimator $\hat{H}[X(t)/X_-, Y_-]$ tends to underestimate the true value for $D_Y \gtrsim N_{\mathrm{samp}}$, while $\hat{H}[X(t)/X_-]$ remains unaffected by D_Y. Consequently, $\hat{T}_{Y\to X} = \hat{H}[X(t)/X_-] - \hat{H}[X(t)/X_-, Y_-]$ overestimates the true value of $T_{Y\to X}$. Representing the collection of variables in the system other than X as Y, we find that the estimators $\widehat{\mathrm{OutTE}}(X)$ and $\widehat{\mathrm{InTE}}(X)$ tend to underestimate and overestimate their true values, $\mathrm{OutTE}(X)$ and $\mathrm{InTE}(X)$, respectively, if the true values are positive.

For the special cases where $\mathrm{OutTE}(X)$ and $\mathrm{InTE}(X)$ are zero, underestimation cannot occur since these quantities are non-negative. In such cases, overestimation is possible, but $\widehat{\mathrm{OutTE}}(X)$ will vanish if the number of samples is sufficiently small compared to the number of variables. However, our primary goal is to identify a few variables with the highest values of $\widehat{\mathrm{OutTE}}$ and $\widehat{\mathrm{InTE}}$. Therefore, cases where the true values vanish are of limited relevance in real data applications.

To illustrate the issue of underestimation of OutTE, consider the SR model introduced in the previous section (Fig. 1). Suppose we have four observed transitions of variables (S, R):

$$\begin{aligned} &(0,0)\to(0,0), \quad (0,1)\to(1,1), \quad (1,0)\to(1,1), \\ &(1,1)\to(1,0). \end{aligned} \tag{11}$$

In this case, the empirical conditional probability distributions $\hat{P}[R(t)/R_-]$ and $\hat{P}[R(t)/S_-, R_-]$ coincide with the true conditional probability distributions $P[R(t)/R_-]$ and $P[R(t)/S_-, R_-]$, respectively, so $\hat{T}_{S\to R} = T_{S\to R} = 1$ bit.

Next, suppose we add some variables $C_1, C_2, \dots C_n$, each of which takes random value of either 0 or 1 and is entirely disconnected from the rest of the system. I will refer to these as "confounding variables" (see Fig. 1). The true values of $\mathrm{OutTE}(S)$ and $\mathrm{OutTE}(R)$ remain unaffected, as the dynamics of $C_1, C_2, \dots C_n$ are independent of S and R. However, they can severely impact the *estimated* values of $\mathrm{OutTE}(S)$ and $\mathrm{OutTE}(R)$. Let us assume just one confounding variable C is added to the example in Eq. (11) so that the four observed

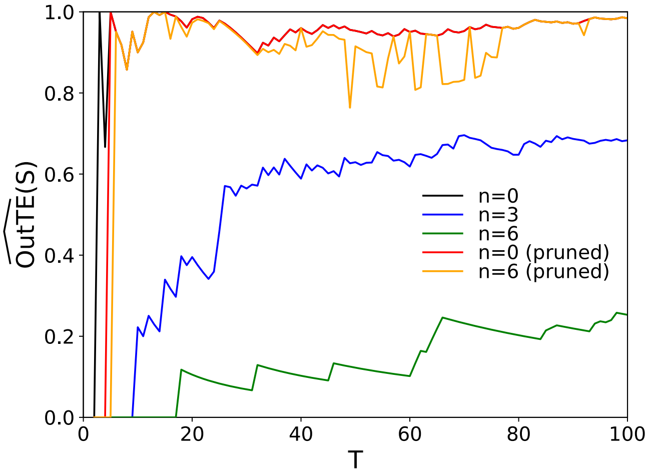

FIG. 2. $\widehat{\mathrm{OutTE}}(S)$ in the SR model as a function of T, without pruning (black, blue, and green lines for $n = 0, 3, 6$, respectively) and with pruning (red and orange lines for $n = 0$ and 6, respectively).

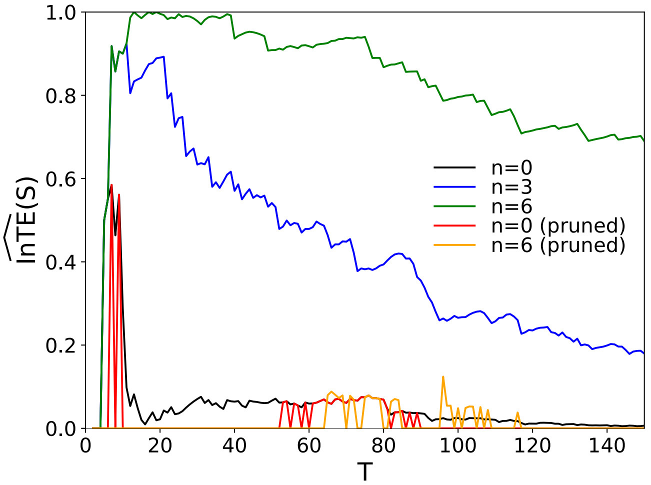

FIG. 4. $\widehat{\mathrm{InTE}}(S)$ in the SR model as a function of T, without pruning (black, blue, and green lines for $n = 0, 3, 6$, respectively) and with pruning (red and orange lines for $n = 0$ and 6, respectively).

transitions of the variables (S, R, C) become

$$\begin{aligned} &(0,0,1) \to (0,0,1), \quad (0,1,0) \to (1,1,1), \\ &(1,0,0) \to (1,1,0), \quad (1,1,1) \to (1,0,0). \end{aligned} \tag{12}$$

Now the dynamics of (R, C) *estimated* by the data above is completely deterministic. Specifically, the only four observed transitions for (R, C) are

$$\begin{aligned} &(0,1) \to (0,1), \quad (1,0) \to (1,1), \quad (0,0) \to (1,0), \\ &(1,1) \to (0,0), \end{aligned} \tag{13}$$

which are all unique. This leads to $\hat{H}[R(t), C(t)]/[R_-, C_-] = \hat{H}[R(t), C(t)]/[R_-, C_-, S_-] = 0$, resulting in $\widehat{\mathrm{OutTE}}(S) = \hat{T}_{S\to(R,C)} = 0$. This represents a drastic underestimation compared to the true value of $\mathrm{OutTE}(S) = 1$. As mentioned earlier, such an artifact occurs when the number of variables is comparable to or larger than the number of samples. In such situations, most observed transitions become rare events, typically appearing only once in the data, leading to the underestimation of the values of conditional entropy (Appendix). Examples of overestimation will be demonstrated with simulations below.

A synthetic time series of length 300 were generated using the probability distribution of the SR model, and the estimators $\widehat{\mathrm{OutTE}}$ and $\widehat{\mathrm{InTE}}$ were computed from the empirical distribution using partial series of length T. I assumed Markovian dynamics from the start, so the number of observed transitions is $T - 1$. The values of $\widehat{\mathrm{OutTE}}$ and $\widehat{\mathrm{InTE}}$ were computed using the JIDT toolkit [31], where empirical distributions were estimated by counting the frequencies of events, such as a variable taking a specific value or a transition occurring between a given pair of values. The graphs of $\widehat{\mathrm{OutTE}}(S)$, $\widehat{\mathrm{OutTE}}(R)$, $\widehat{\mathrm{InTE}}(S)$, and $\widehat{\mathrm{InTE}}(R)$ are shown as functions of T in Figs. 2–5, respectively, for different numbers of confounding variables: $n = 0$ (black line),

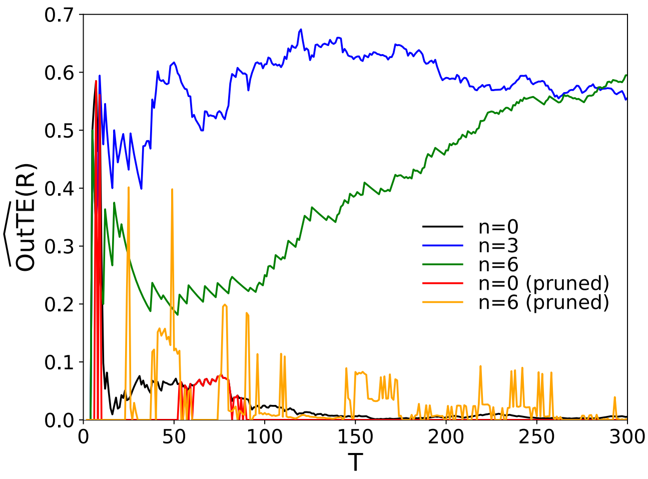

FIG. 3. $\widehat{\mathrm{OutTE}}(R)$ in the SR model as a function of T, without pruning (black, blue, and green lines for $n = 0, 3, 6$, respectively) and with pruning (red and orange lines for $n = 0$ and 6, respectively).

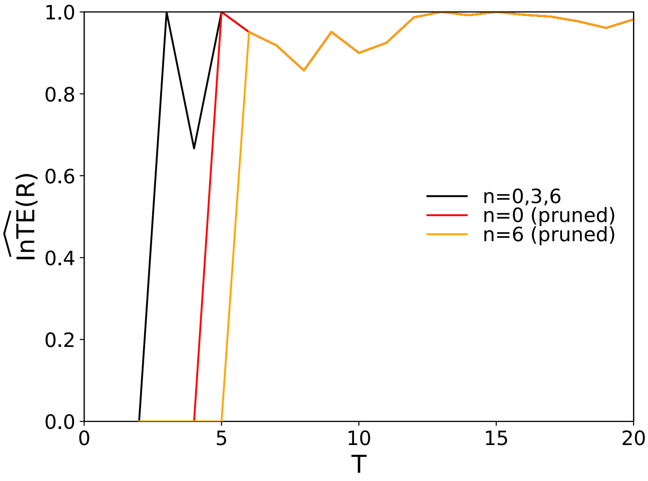

FIG. 5. $\widehat{\mathrm{InTE}}(R)$ in the SR model as a function of T, without pruning (black line for $n = 0, 3, 6$) and with pruning (red and orange lines for $n = 0$ and 6, respectively).

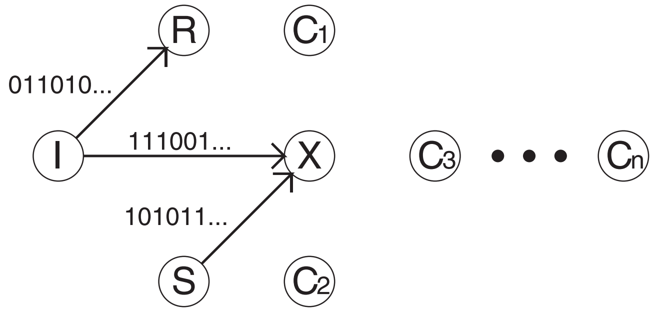

FIG. 6. The ISXR model with a source variable I, a sender variable S, a sink variable X, and a receiver variable R. The variables $C_1, \dots C_n$ represent confounding variables that are independent of the dynamics of I, S, X, and R.

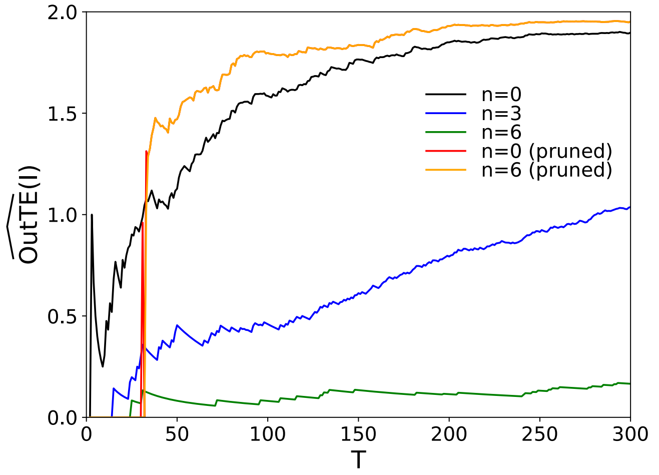

FIG. 7. $\widehat{\mathrm{OutTE}}(I)$ in the ISXR model as a function of T, for $n = 0, 3, 6$ without pruning (black, blue, and green lines for $n = 0, 3, 6$, respectively) and with pruning (red and orange lines for $n = 0$ and 6, respectively).

$n = 3$ (blue line), and $n = 6$ (green line). We observe that $\widehat{\mathrm{OutTE}}(S)$ is lower than the true value of $\mathrm{OutTE}(S) = 1$ bit for small sample sizes, but begins to converge to the true value around $T \simeq 10$ for $n = 0$ (Fig. 2). As the confounding variables are added, the underestimation becomes more severe for a given T, and convergence becomes slower. In the case of $\mathrm{OutTE}(R)$, whose true value is zero, we now encounter an overestimation problem in the presence of confounding variables (Fig. 3). We also observe that the estimation error increases with T over the range shown in the figure, and that the estimation with $n = 6$ is better for small T but becomes worse for $T \gtrsim 300$. This occurs because, for small T and large n, $\hat{H}[S(t), C_1(t), \dots, C_n(t)]/[S(t-1), C_1(t-1), \cdots, C_n(t-1)] \simeq 0$ and $\hat{H}[S(t), C_1(t), \dots, C_n(t)]/[R(t-1), S(t-1), C_1(t-1), \dots, C_n(t-1)] \simeq 0$, due to the underestimation problem, resulting in $\widehat{\mathrm{OutTE}}(R) \simeq 0$. Thus, the small value of $\widehat{\mathrm{OutTE}}(R)$ for small T and large n is an artifact of the limited sample size rather than an accurate estimation.

For $\mathrm{InTE}(S)$, whose true value is zero, we also encounter an overestimation problem, but the artifact observed for $\widehat{\mathrm{OutTE}}(R)$ does not appear here because the target is univariate. The overestimation becomes progressively worse as the confounding variables are added (Fig. 4). Finally, the estimate of $\mathrm{InTE}(R)$, whose true value is one bit, remains unaffected by the presence of the confounding variables, as shown by the black line in Fig. 5. Note that according to the definition of the transfer entropy in Eq. (4), the estimate $\widehat{\mathrm{InTE}}(R)$ of $\mathrm{InTE}(R)$ can be written as

$$\widehat{\mathrm{InTE}}(R) \equiv \hat{T}_{\mathrm{rest}\to R} = \hat{T}_{(S,C_1,\dots,C_N)\to R} = \hat{H}[R(t)/R_-] - \hat{H}[R(t)/R_-, S_-, C_{1-}, \dots, C_{N-}]. \tag{14}$$

Since the first term $\hat{H}[R(t)/R_-]$ in Eq. (14) does not depend on the confounding variables, the only possible dependence on the confounding variables originates from the second term $\hat{H}[R(t)/R_-, S_-, C_{1-}, \dots, C_{N-}]$. However, in this model, the values of $R(t-1)$ and $S(t-1)$ uniquely determine the value of $R(t)$, which also holds true for the observed transitions. The addition of confounding variables does not affect this determinism. Therefore, $\hat{H}[R(t)/R_-, S_-, C_{1-}, \dots, C_{N-}] = 0$, regardless of the number of confounding variables, leading to $\hat{H}[R(t)/R_-, S_-, C_{1-}, \dots, C_{N-}] = 0$, which is independent of the confounding variables. This property is specific to this model. In general, $\widehat{\mathrm{InTE}}(R)$ tends to overestimate the true value, as mentioned earlier in this section.

I conducted a similar analysis on a system containing a source variable I, a sender S, a sink variable X, and a receiver variable R, as shown in Fig. 6. I will refer to this system as the ISXR model. In this model the variable I sends two independent bits of information to two nodes X and R, and S sends one bit of information to X. As a result, X receives one bit of information from both I and S (total two bits), while R receives one bit of information from I alone. The estimates $\widehat{\mathrm{OutTE}}(I)$, $\widehat{\mathrm{OutTE}}(S)$, $\widehat{\mathrm{OutTE}}(X)$, $\widehat{\mathrm{OutTE}}(R)$, $\widehat{\mathrm{InTE}}(I)$, $\widehat{\mathrm{InTE}}(S)$, $\widehat{\mathrm{InTE}}(X)$, and $\widehat{\mathrm{InTE}}(R)$ are shown in Figs. 7–14 for different numbers of confounding variables ($n = 0, 3, 6$), represented by black, blue, and green lines, respectively.

The results are qualitatively similar to those of the SR model. Although $\widehat{\mathrm{OutTE}}(I)$ and $\widehat{\mathrm{OutTE}}(S)$ converge towards their true values (2 bits for I and 1 bit for S) as T increases,

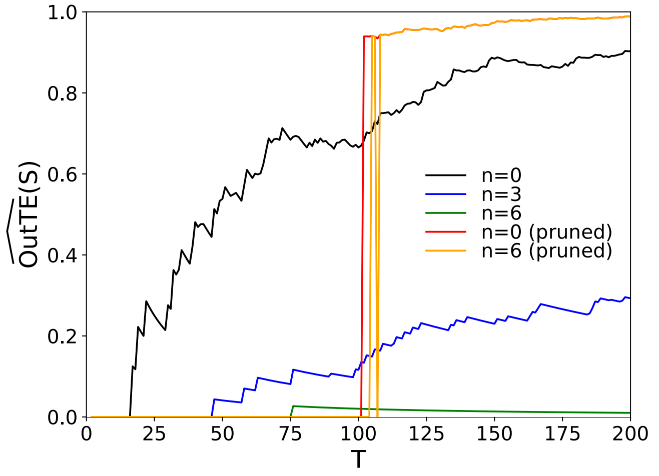

FIG. 8. $\widehat{\mathrm{OutTE}}(S)$ in the ISXR model as a function of T, without pruning (black, blue, and green lines for $n = 0, 3, 6$, respectively) and with pruning (red and orange lines for $n = 0$ and 6, respectively).

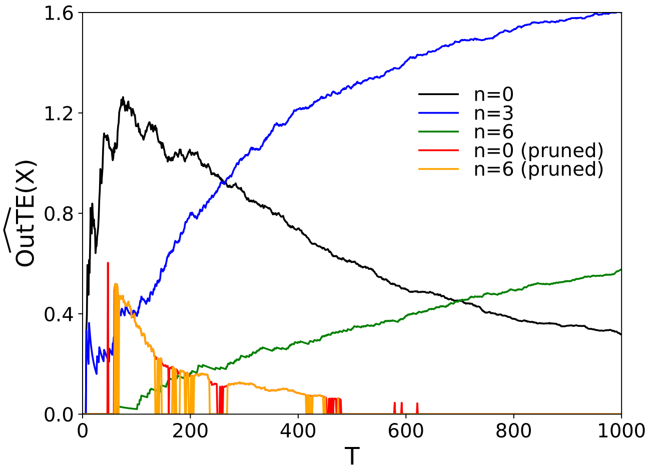

FIG. 9. $\widehat{\mathrm{OutTE}}(X)$ in the ISXR model as a function of T, without pruning (black, blue, and green lines for $n = 0, 3, 6$, respectively) and with pruning (red and orange lines for $n = 0$ and 6, respectively).

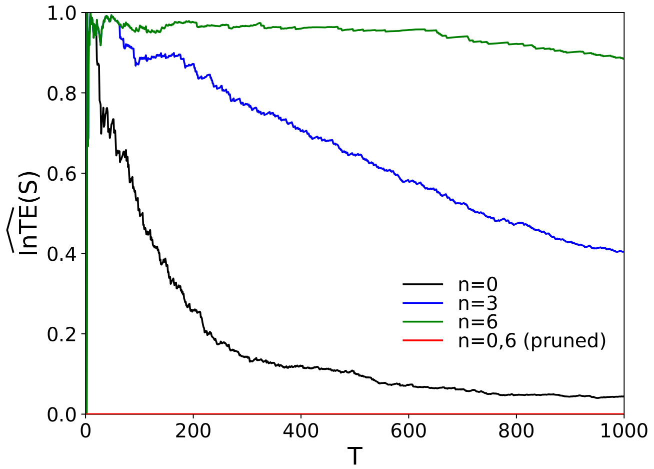

FIG. 12. $\widehat{\mathrm{InTE}}(S)$ in the ISXR model as a function of T, without pruning (black, blue, and green lines for $n = 0, 3, 6$, respectively) and with pruning (red line for $n = 0$ and 6).

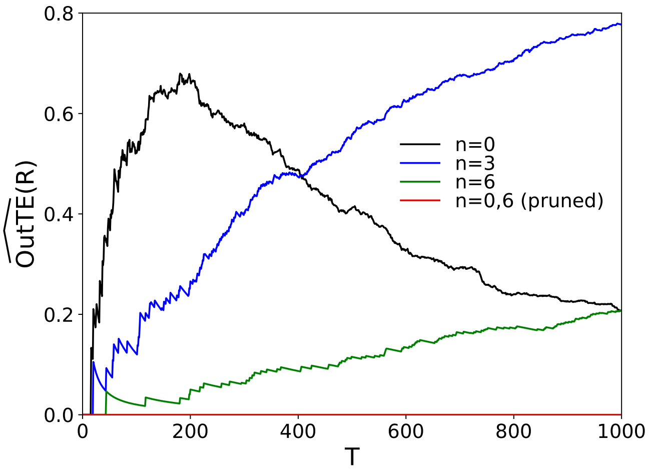

FIG. 10. $\widehat{\mathrm{OutTE}}(R)$ in the ISXR model as a function of T, without pruning (black, blue, and green lines for $n = 0, 3, 6$, respectively) and with pruning (red line for $n = 0$ and 6).

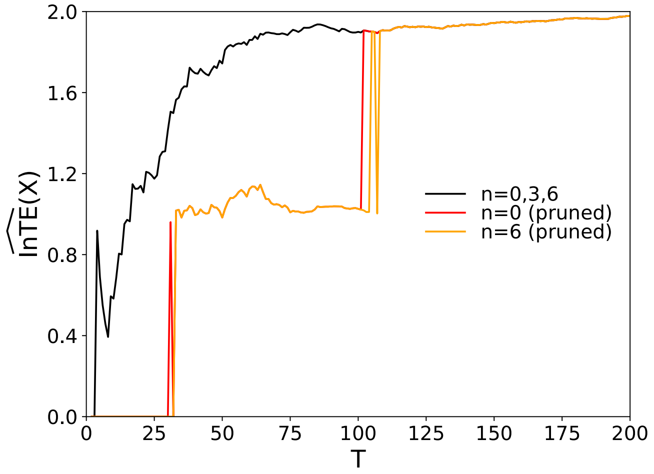

FIG. 13. $\widehat{\mathrm{InTE}}(X)$ in the ISXR model as a function of T, without pruning (black line for $n = 0, 3, 6$) and with pruning (red and orange lines for $n = 0$ and 6, respectively).

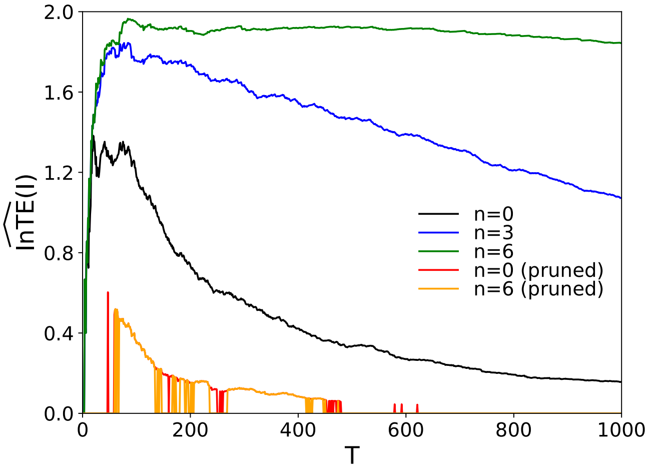

FIG. 11. $\widehat{\mathrm{InTE}}(I)$ in the ISXR model as a function of T, without pruning (black, blue, and green lines for $n = 0, 3, 6$, respectively) and with pruning (red and orange lines for $n = 0$ and 6, respectively).

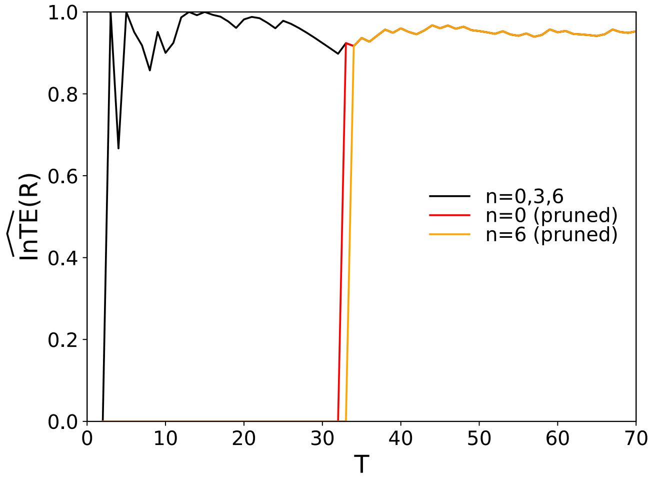

FIG. 14. $\widehat{\mathrm{InTE}}(R)$ in the ISXR model as a function of T, without pruning (black line for $n = 0, 3, 6$) and with pruning (red and orange lines for $n = 0$ and 6, respectively).

the convergence slows down for larger n. For OutTE(X) and OutTE(R), the estimates are larger than their true values (which are zero). However, this overestimation is mitigated for large n and small T, due to the same artifact as in the case of $\widehat{\text{OutTE}}(R)$ in the SR model. The overestimation of InTE(I) and InTE(S), whose true values are zero, worsens with increasing n, much like the overestimation of InTE(S) in the SR model. Finally, $\widehat{\text{InTE}}(X)$ and $\widehat{\text{InTE}}(R)$ remain unaffected by the presence of confounding variables for the same reason as in $\widehat{\text{InTE}}(R)$ in the SR model.

IV. THE METHOD FOR ACCURATE ESTIMATION OF OUTGOING AND INCOMING TRANSFER ENTROPY

In the previous examples, the estimation errors were due to the proliferation of confounding variables. If we had eliminated unrelated variables beforehand, the estimation error would have been reduced. Therefore, instead of estimating OutTE and InTE between a variable and the collection of all the other variables, we first take a pruning step where only variables causally related to the variable of interest are selected. That is, we construct a directed binary network in which each variable is represented by a node, and an edge $X \to Y$ exists if and only if $\widehat{T}_{(X\to Y)/Z}$ is statistically significant, where Z denotes the collection of all variables except X and Y. The construction of such a network using the exact computation of $\widehat{T}_{(X\to Y)/Z}$, along with rigorous statistical tests, is computationally costly, and various approximation schemes have been proposed to construct the causal network [13,15,20,32–34]. Here, I used the method developed in Ref. [20], where the set of source variables with statistically meaningful causal influence is constructed step by step for each target variable. To elaborate, for a given target variable Y, let S denote the set of candidate source variables, initialized as an empty set. Assuming Markovian dynamics, where $X_- = X(t-1)$ for any variable X, the procedure is as follows:

(1) Forward selection of source variables:

(a) First, select the candidate source variable X_1 that maximizes the conditional mutual information $I[X_1(t-1), Y(t)/Y(t-1)]$ and test its statistical significance.

(b) If X_1 is statistically significant, include it in S. Then select X_2 such that $I[X_2(t-1), Y(t)/Y(t-1), X_1(t-1)]$ is maximal and test its significance.

(c) Repeat this process until no further statistically significant source variables are found.

(2) Backward elimination of redundant variables:

(a) Among the candidate source variables $X_1, \ldots, X_n \in S$, identify the variable $\tilde{X}_1$ that minimizes the conditional mutual information $I(\tilde{X}_1(t-1), Y(t)/Y(t-1), X_1(t-1), \cdots, \xcancel{\tilde{X}_1(t-1)}\cdots, X_n(t-1)]$.

(b) Test the statistical significance of this value. If it is not significant, exclude $\tilde{X}_1$ from S and repeat the process with $\tilde{X}_2$, $\tilde{X}_3$, and so on.

(c) Stop when the variable with the minimal conditional mutual information in S is statistically significant or when S becomes empty.

(3) Final test and edge construction:

(a) If S is nonempty, test the statistical significance of the transfer entropy from the collection of variables in S to the target variable Y.

(b) If significant, S is finalized as the set of source variables for Y. Otherwise, no source variables are identified for Y.

(c) Draw directed edges from each variable in S to the target variable Y.

This process is repeated for each variable in the system, taking it as the target and identifying its source variables. The resulting directed edges form a causal network. This method has been implemented as a publicly available PYTHON package called IDTxl [35] and has been used for constructing a causal network from the brain record data comprising 100 variables and 10 000 samples [20].

By computing OutTE and InTE only between the node of interest and its connected nodes in the directed binary network derived from the causal inference algorithm described above, estimation errors are significantly reduced, as shown in Figs. 2–5 for the SR model and Figs. 7–14 for the ISXR model for $n = 0$ (red lines) and $n = 6$ (orange lines). We find that the estimation error with pruning does not significantly increase, even as the number of confounding variables rises from 0 to 6.

In the SR model, the pruned estimate $\widehat{\text{OutTE}}(S)$ quickly converges to the true value of 1 bit around $T \simeq 10$ (Fig. 2) regardless of n, overcoming the underestimation problem. Pruning also reduces the overestimation problem of $\widehat{\text{OutTE}}(R)$, even for $n = 6$, where $\widehat{\text{OutTE}}(R) < 0.1$ for $T \geqslant 112$ (Fig. 3). The same applies to $\widehat{\text{InTE}}(S)$, where $\widehat{\text{InTE}}(S) < 0.1$ for $T \geqslant 97$ even for $n = 6$ (Fig. 4). In the case of $\widehat{\text{InTE}}(R)$ where confounding variables are not problematic, pruning is not useful; in fact, it increases estimation error for very small sample sizes. However, this error quickly diminishes once T reaches 6 (Fig. 5), so the damage is minimal.

In the ISXR model, the pruned estimate $\widehat{\text{OutTE}}(I)$ is zero up to $T \simeq 30$ because the network construction algorithm could not detect the outgoing edges from the node I. However, the estimate quickly converges to the true value of 2 bits for $T \gtrsim 30$, regardless of n (Fig. 7). In this range, the pruned estimate of OutTE(I) outperforms the unpruned estimate, even for $n = 0$, as variable S, which is unrelated to I in terms of OutTE(I), acts as a confounding variable for I. The same behavior is observed for $\widehat{\text{OutTE}}(S)$, where the estimate with pruning converges to the true value of 1 bit at around $T \simeq 100$. Again, for $T \gtrsim 100$, pruning proves beneficial even for $n = 0$, because it eliminates the confounding effect from the variable I (Fig. 8). Pruning also reduces the overestimation problem of $\widehat{\text{OutTE}}(X)$ (Fig. 9) by removing the confounding effect from the variable R as well as those from the variables C_i $(i = 1, \ldots n)$. The effect of the pruning is even more drastic in the case of the estimate $\widehat{\text{OutTE}}(R)$, since it makes the estimate coincide with the true value of zero, as the network algorithm does not detect any meaningful connection from the variable R to any other variable for any sample size (Fig. 10). The difference in the behavior of the network construction algorithm for OutTE(R) compared to OutTE(X) suggests that

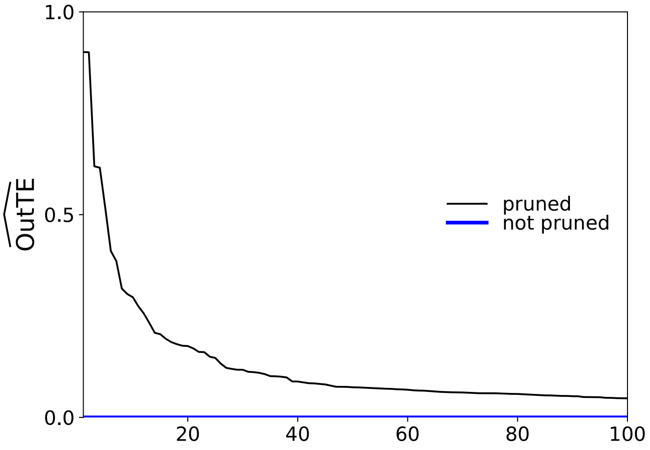

FIG. 15. $\widehat{\text{OutTE}}$ of oral microbiota, obtained with pruning, sorted in descending order of their values (black line) shown for top 100 bacteria, compared with estimates obtained without pruning (blue line).

the presence of many incoming edges for X, which should not affect the true value of $\text{OutTE}(X)$, somehow confuses the network construction algorithm at small sample sizes. This might result in the introduction of some false outgoing edges, leading to small but nonzero values of $\widehat{\text{OutTE}}(X)$.

Regarding InTE, pruning resolves the overestimation problem of $\text{InTE}(I)$, where all of the S, R, and X variables act as additional confounding variables (Fig. 11). For $\widehat{\text{InTE}}(S)$, the pruned estimate coincides with the true value of 0 bits (Fig. 12). The difference between the effect of pruning on $\widehat{\text{InTE}}(I)$ and $\widehat{\text{InTE}}(S)$ is similar to that between $\widehat{\text{OutTE}}(X)$ and $\widehat{\text{OutTE}}(R)$. Pruning does not benefit $\widehat{\text{InTE}}(X)$ or $\widehat{\text{InTE}}(R)$, as confounding variables are not an issue. However, introducing pruning does no harm for $T \gtrsim 100$ for $\widehat{\text{InTE}}(X)$ and $T \gtrsim 30$ for $\widehat{\text{InTE}}(R)$. The situation is similar to that of $\widehat{\text{InTE}}(R)$ in the SR model.

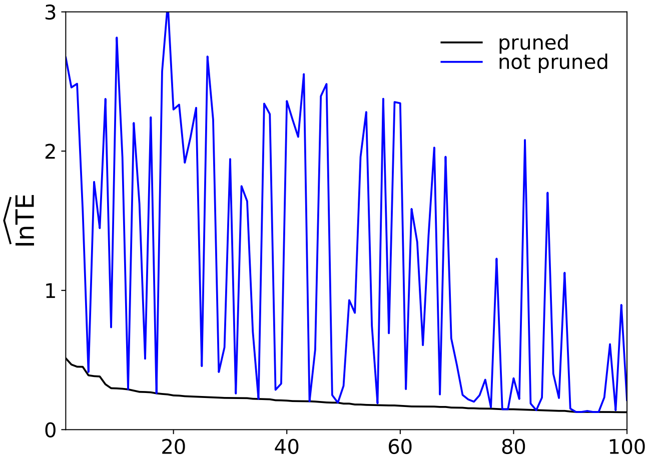

FIG. 16. $\widehat{\text{InTE}}$ of oral microbiota, obtained with pruning, sorted in descending order of their values (black line) shown for top 100 bacteria, compared with estimates obtained without pruning (blue line).

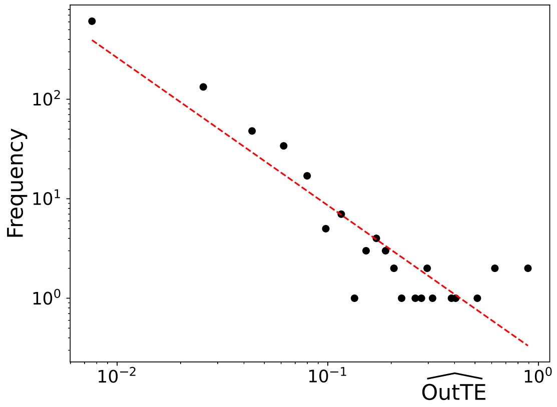

FIG. 17. Histogram of $\widehat{\text{OutTE}}$ (black dots) of oral microbiota, along with a power-law fit (red dashed line).

V. APPLICATION TO MICROBIOTA DATA

We now apply the current method to microbiota data obtained from a saliva sample observed over 226 days, consisting of 879 variables [36]. Most of the variables represent operational taxonomic units (OTUs), the lowest taxonomic levels for bacteria, but some of the bacteria could only be resolved at higher taxonomic levels.

For reduced computational costs, we assume Markovian dynamics so $X_{-} = X(t-1)$ for any variable X. The causal network was constructed using the IDTxl package [35], and transfer entropy was computed with the JIDT toolkit [31]. The values of $\widehat{\text{OutTE}}$ and $\widehat{\text{InTE}}$ for all variables are compared, with and without pruning, in Figs. 15 and 16, respectively, where the variables are sorted in descending order of the pruned estimates, and the top 100 are shown. We observe that $\widehat{\text{OutTE}} = 0$ for all variables without pruning, due to severe underestimation, while the pruned estimates range between 0 and 0.9 (Fig. 15). In the case of $\widehat{\text{InTE}}$, the estimated values are nonzero even without pruning. However, while the estimates obtained with pruning range between 0 and 0.5, those obtained without pruning can be as high as 3 and are unrelated to those obtained with pruning (Fig. 16), suggesting large estimation errors without pruning. Also, compared to $\widehat{\text{OutTE}}$, where

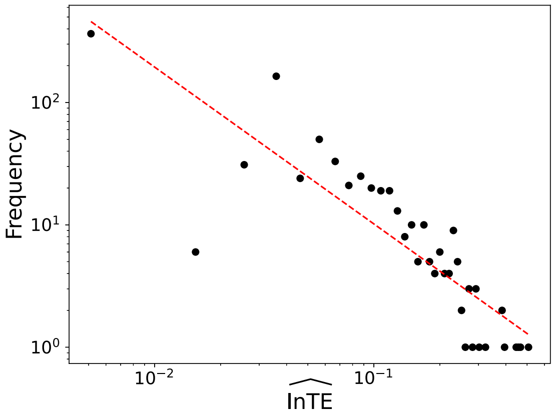

FIG. 18. Histogram of $\widehat{\text{INTE}}$ (black dots) of oral microbiota, along with a power-law fit (red dashed line).

TABLE I. Top five microbiota with highest values of $\widehat{\text{OutTE}}$ and $\widehat{\text{InTE}}$.

Rank	$\widehat{\text{OutTE}}$		$\widehat{\text{InTE}}$	
1	Corynebacterium Durum	0.90	Gemellales (order)[a]	0.51
2	Fusobacterium (genus)[b]	0.90	Oribacterium (genus)[c]	0.47
3	Prevotella melaninogenica	0.62	Rothia mucilaginosa	0.45
4	Unknown	0.62	SR1 (phylum)[d]	0.45
5	Coriobacteriaceae (family)[e]	0.52	Flavobacterium succinicans	0.39

[a] Families other than Gemellaceae.

[b] Includes all the OTUs.

[c] Includes all the OTUs.

[d] Includes all the classes.

[e] Genera other than Adlercreutzia, Atopobium, Collinsella, ggerthella, Slackia, and Rubrobacter.

a few variables have very high values, such peaks are less pronounced for $\widehat{\text{InTE}}$, where no values exceed 0.5. This difference in behavior is also evident in the histograms, which were constructed using 50 bins to count the frequency of occurrence (Figs. 17 and 18). In network analysis, the power-law distribution of the number of links is often investigated, as it indicates the presence of hubs, nodes with disproportionately high connectivity compared to others in the network [37–39]. To explore whether $\widehat{\text{OutTE}}$ ($\widehat{\text{InTE}}$), the information theoretic analog of the number of outgoing (incoming) edges in a directed binary network, exhibits such behavior, I examined their histograms. The histogram of $\widehat{\text{OutTE}}$ indeed follows a power-law behavior, suggesting the presence of nodes acting as information sources (Fig. 17). In contrast, the histogram of $\widehat{\text{InTE}}$ fits the power law less well (Fig. 18).

The top five bacteria ranked by $\widehat{\text{OutTE}}$ and $\widehat{\text{InTE}}$ are shown in the Table I. We find that the OTU *Corynebacterium durum* and the genus *Fusobacterium* stand out in terms of high $\widehat{\text{OutTE}}$ values. *C. durum* is indeed known to play a crucial role in the community of oral microbiota. As a gram-positive bacterium, it is a prolific biofilm and extracellular matrix producer [40], and a decrease in this bacterium is associated with a disease [41]. The genus *Fusobacterium*, particularly the OTU *Fusobacterium nucleatum*, is also well known as a key player in the community of oral bacteria [42,43]. As a gram-negative bacterium, *F. nucleatum* is a major coaggregation bridge organism linking early and late colonizers in dental biofilm [44] and plays a role in carcinogenesis [45,46]. Note that the fourth rank, denoted as "unknown," represents a group of bacteria unidentified at any taxonomic level. Relatively high OutTE of this group may simply result from its heterogeneity, as it likely comprises a diverse mixture of bacteria. No particular bacterium stands out in terms of $\widehat{\text{InTE}}$ values, as noted above. Moreover, since InTE measures how much a variable acts as an information sink rather than an information source, it is unclear whether such downstream species can be easily identified experimentally. Therefore, the biological relevance of the top microbiota in terms of $\widehat{\text{InTE}}$ is less evident compared to those ranked by $\widehat{\text{OutTE}}$.

Since we constructed a directed binary network from the causal inference in the pruning step, we can compare the information-theoretic measures OutTE and InTE with traditional network centrality measures. Here I consider three measures: The number of outgoing edges, referred to as the out-degree; the number of incoming edges, referred to as the in-degree; and the number of shortest paths passing through a node, known as the betweenness centrality. For weighted directed networks, the out-degree and in-degree are typically generalized by summing the weights of the out-going and incoming edges, respectively. In the network obtained from the causal inference on a stochastic dynamical system, the transfer entropy or its conditional version can be assigned to each directed edge. However, as explained in Sec. II, there is no theoretical justification for summing these values.

The correlation between the rankings obtained from various centrality measures are shown in Table II. Positive correlations are observed between rankings obtained from conceptually similar centrality measures, $\widehat{\text{OutTE}}$ ($\widehat{\text{InTE}}$), sum

TABLE II. Correlations between rankings obtained from various centrality measures. The labels bv$\widehat{\text{OutTE}}$ and bv$\widehat{\text{InTE}}$ denote the sum of bivariate $\widehat{\text{OutTE}}$s and sum of bivariate $\widehat{\text{InTE}}$s, repectively.

	$\widehat{\text{OutTE}}$	bv$\widehat{\text{OutTE}}$	Out-degree	$\widehat{\text{InTE}}$	bv$\widehat{\text{InTE}}$	In-degree	Betweenness
$\widehat{\text{OutTE}}$	1.000	0.734	0.560	0.109	0.101	0.052	0.542
bv$\widehat{\text{OutTE}}$		1.000	0.865	−0.059	−0.036	0.000	0.742
Out-degree			1.000	−0.120	−0.103	0.047	0.826
$\widehat{\text{InTE}}$				1.000	0.936	0.513	0.135
bv$\widehat{\text{InTE}}$					1.000	0.578	0.168
In-degree						1.000	0.330
Betweenness							1.000

TABLE III. Top five microbiota with highest values of out-degree, in-degree, and betweenness.

Rank	Out-degree		In-degree		Betweenness	
1	Ellin6067 (order)[a]	57	Oxalobacter (genus)[b]	28	258ds10 (order)[c]	0.052
2	Rhodobacter (genus)[d]	51	Cellvibrio (genus)[e]	28	Blautia producta	0.050
3	Alphaproteobacteria (class)[f]	50	Aliivibrio fischeri	25	Ellin6067 (order)[g]	0.044
4	Nelumbo nucifera	49	Chitinophagaceae (family)[h]	24	Bacillaceae (family)[i]	0.039
5	Chromatiales (order)[j]	46	258ds10 (order)[k]	24	Volucribacter psittacicida	0.037

[a] Includes all the families.

[b] OTUs other formigenes.

[c] Includes all the families.

[d] Includes all the OTUs.

[e] Includes all the OTUs.

[f] Orders other than BD7-3, Caulobacterales, Ellin329, RF32, Rhizobiales, Rhodobacterales, hodospirillales, Rickettsiales, and Sphingomonadales.

[g] Includes all the families.

[h] Genera other than Chitinophaga, Flavihumibacter, Flavisolibacter, Niabella, Sediminibacterium, and Segetibacter.

[i] Genera other than Anoxybacillus, Bacillus, Geobacillus, and Natronobacillus.

[j] Families other than Chromatiaceae.

[k] Includes all the families.

of bivariate $\widehat{\text{OutTE}}$s (sum of bivariate $\widehat{\text{InTE}}$s), Out-degree (In-degree), as expected. However, as shown in Table III, where the top five microbiota are selected based on centrality measures of the directed binary network, measures that do not account for the magnitude of causal influence fail to identify key oral microbiota such as *C. durum* and *Fusobacterium*.

In fact, the out-degrees of both *C. Durum* and *Fusobacterium* are only 3, suggesting that they exert significant causal influence on the rest of the system hierarchically through a few neighboring nodes. In contrast, all the top five nodes with highest out-degrees have values of $\widehat{\text{OutTE}}$ values less than 0.065. This behavior is highlighted in Fig. 19, which shows a portion of the network including *C. Durum*, *Fusobacterium*, and *Rhodobacter* (with an out-degree of 51). In the figure, node sizes are proportional their $\widehat{\text{OutTE}}$ values.

As shown in Table IV, which lists the top five bacteria with highest values of the sum of bivariate TEs, the sum of outgoing bivariate TEs outperforms the centrality measures of the directed binary network by identifying *C. durum* as the fourth-ranking species. This is likely because it accounts for the weight in the network. However, its performance is inferior to the more rigorous information theoretic measure OutTE, which ranks *C. durum* and *Fusobacterium* as the top two bacterial groups.

Note that if the out-degree or the in-degree of a node in the causal inference network is too large, the effect of pruning is reduced, and underestimation of OutTE or overestimation of InTE may still occur. This suggests that when selecting nodes with high values of $\widehat{\text{OutTE}}$ and $\widehat{\text{InTE}}$, we may encounter false negatives for the former and false positives for the latter. For the ten bacterial groups listed in Table I, all degrees are at most five, except for *Flavobacterium succinicans*, which is ranked as fifth in terms of InTE and has an in-degree of 20. This degree is still small compared to the $N_{\text{obs}} = 225$. As shown in Table III, the largest value of out-degree is only 57, which is also smaller than $N_{\text{obs}} = 225$. Although there is a possibility that the values of OutTE are underestimated for nodes with large out-degrees, no bacteria known to play an important role in the oral bacterial community appears in the list of five nodes with the largest out-degrees (Table III).

VI. CONCLUSION

In this study I introduced outgoing transfer entropy (OutTE) and incoming transfer entropy (InTE) as novel measures for quantifying the causal influence of each component

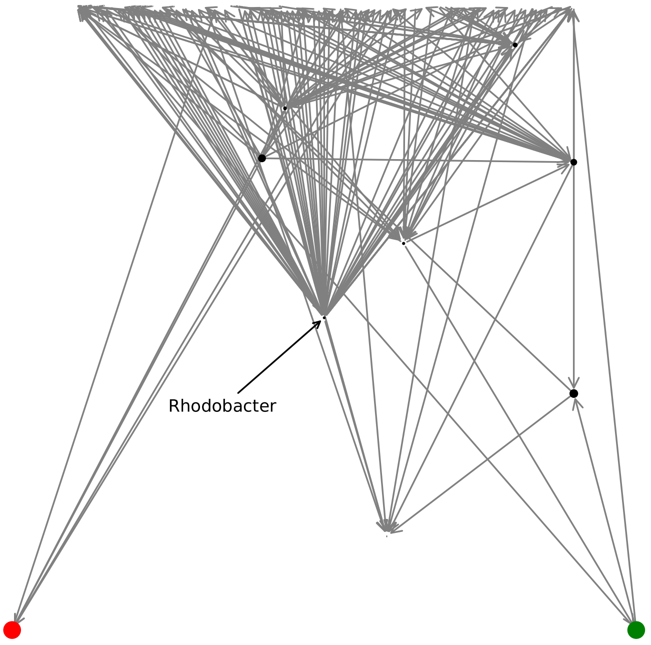

FIG. 19. A subgraph containing two nodes with highest values of $\widehat{\text{OutTE}}$, *C. Durum* (red) and *Fusobacterium* (green), along with *Rhodbacter* whose out-degree is 51. The size of the nodes are proportional to the values of $\widehat{\text{OutTE}}$. The nodes outside this subgraph are placed above the figure. The figure was drawn using NeworkX [47] and Matplotlib [48].

TABLE IV. Top five microbiota with highest values of sum of bivariate $\widehat{\text{OutTE}}$s and sum of bivariate $\widehat{\text{InTE}}$s.

Rank	Sum of bivariate $\widehat{\text{OutTE}}$s		Sum of bivariate $\widehat{\text{InTE}}$s	
1	Porphyromonas (genus)[a]	1.79	Flavobacterium succinicans	0.52
2	Nelumbo nucifera	1.69	Gemellales (order)[b]	0.51
3	Rhodobacter (genus)[c]	1.50	Rothia mucilaginosa	0.50
4	Corynebacterium durum	0.96	Rikenellaceae (family[d]	0.45
5	Ellin6067 (order)[e]	0.94	Marinomonas (genus)[f]	0.43

[a] OTUs other than endodontalis.
[b] Families other than Gemellaceae.
[c] Includes all the OTUs.
[d] Genera other than Alistipes and Rikenella.
[e] Includes all the families.
[f] Includes all the OTUs.

on the entire system and vice versa in stochastic dynamical systems. OutTE and InTE are analogous to out-degree and in-degree in network analysis but uniquely quantify the causal influence based on information theory. Variables with exceptionally high values of OutTE can be considered as key drivers of the dynamical system.

To address estimation errors, particularly when the number of variables approaches or exceeds the number of samples, I proposed an estimation method enhanced by pruning. A directed binary network was first constructed using causal inference to prune unrelated variables. Through simulations on synthetic data, I demonstrated the effectiveness of this pruning-enhanced method in accurately estimating OutTE and InTE values. These results suggest that the method reliably highlight key drivers in a dynamical system. The application of this method to microbiota data from human saliva further validated its utility, successfully identifying *Corynebacterium durum* and *Fusobacterium*, known to play critical roles in the oral bacterial community. These two bacterial groups would have not been identified using traditional centrality measures, such as out-degree or the sum of bivariate outgoing transfer entropies in the causal interaction network.

For some time series data, the estimation of OutTE after pruning may fail to detect certain key drivers of the system if their out-degrees remain comparable to the sample size. In such cases, examining nodes with high out-degrees or large values of the sum of bivariate $\widehat{\text{OutTE}}$s can serve as complementary approaches to identify missing drivers, albeit without theoretical rigor.

Constructing a binary network of meaningful causal relationships is a crucial first step for pruning in this approach, for which a publicly available state-of-the-art tool has been used [20,35]. Selecting parameters and options that best match the underlying data distribution is essential for optimal efficacy. For the oral microbiota data, assuming Markovian dynamics still yielded good results. In the absence of prior knowledge about the distribution, trial and error may be required. The development of faster and more accurate causal network reconstruction algorithms would further enhance the application of the estimation method presented here. Additionally, applying this method to larger and more diverse datasets could validate its broader utility and uncover new insights into the structure and function of complex systems.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea, funded by the Ministry of Science and ICT (NRF-2020R1A2C1005956).

APPENDIX: UNDERESTIMATION OF CONDITIONAL ENTROPY FOR SMALL SAMPLE SIZES

Conditional entropy $H(Y/X)$ measures the amount of uncertainty remaining in a random variable Y given that the value of another random variable X is known. The true conditional entropy is defined as

$$H(Y/X) = -\sum_x P(X=x)\sum_y P(Y=y/X=x) \times \log P(Y=y/X=x).$$

When estimating conditional entropy from a finite sample, the empirical conditional entropy $\hat{H}(Y/X)$ is calculated by substituting the true probabilities with their empirical estimates based on the observed data:

$$\hat{H}(Y/X) = -\sum_x \hat{P}(X=x)\sum_y \hat{P}(X=y/X=x) \times \log \hat{P}(X=y/X=x),$$

where empirical probabilities are calculated as

$$\hat{P}(X=x) \equiv \frac{N(x)}{N}, \quad \hat{P}(Y=y/X=x) \equiv \frac{N(x,y)}{N(x)},$$

with $N(x,y)$ being the number of times the pair (x,y) occurs, and $N(x)$ the number of times x occurs, and N the total number of observations. In practice, the estimated conditional entropy $\hat{H}(Y/X)$ tends to be smaller than the true entropy $H(Y/X)$ when the sample size is small. This underestimation occurs due to insufficient observations, which leads to a biased estimation of probabilities.

The primary reason for the underestimation of conditional entropy lies in the concavity of the logarithm function and the bias introduced by finite sample sizes. Note that Jensen's inequality can be written in the form

$$\langle f(X)\rangle \leqslant f(\langle X\rangle) \tag{A1}$$

for any random variable X and a concave function $f(x)$. The logarithm function $\log(x)$ is concave, which means that for any random variable Z,

$$\langle \log Z\rangle \leqslant \log\langle Z\rangle.$$

In the context of conditional entropy, this inequality implies that the expected value of the logarithm of the estimated probability is less than the logarithm of the true probability. When probabilities are estimated from a small sample, the estimates $\hat{P}(Y = y/X = x)$ are typically more concentrated around zero for rare events, leading to a lower average entropy. That is, for small sample sizes, many possible pairs (x, y) may not be observed at all, leading to $\hat{P}(Y = y/X = x) = 0$ for these pairs. Since $\hat{P}(Y = y/X = x)\log\hat{P}(Y = y/X = x)$ is defined as $\lim_{p\to 0} p\log p = 0$ when $\hat{P}(Y = y/X = x) = 0$, this results in a lower estimated entropy. In contrast, the true probability $P(Y = y/X = x)$ might be nonzero, leading to a nonzero contribution to the true entropy $H(Y/X)$.

For example, consider a simple case where X and Y are binary variables, and the true conditional probabilities are

$$P(Y = 0/X = 0) = 1/2, \quad P(Y = 1/X = 0) = 1/2,$$

$$P(Y = 0/X = 1) = 1/2, \quad P(Y = 1/X = 1) = 1/2.$$

Suppose we have just two observations, with $(X, Y) = (0, 0)$ and $(X, Y) = (1, 1)$. The empirical conditional probabilities are

$$\hat{P}(Y = 0/X = 0) = 1, \quad \hat{P}(Y = 1/X = 0) = 0,$$

$$\hat{P}(Y = 0/X = 1) = 0, \quad \hat{P}(Y = 1/X = 1) = 1.$$

The empirical conditional entropy $\hat{H}(Y/X) = 0$ calculated from these estimates is less than the true conditional entropy $H(Y/X) = 1$ calculated from the true probabilities, illustrating the underestimation bias.

[1] C. W. J. Granger, Investigating causal relations by econometric models and cross-spectral methods, Econometrica **37**, 424 (1969).

[2] T. Schreiber, Measuring information transfer, Phys. Rev. Lett. **85**, 461 (2000).

[3] J. Pearl, *Causality: Models, Reasoning, and Inference* (Cambridge University Press, Cambridge, 2000).

[4] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search* (The MIT Press, Boston, MA, 2000).

[5] M. Staniek and K. Lehnertz, Symbolic transfer entropy, Phys. Rev. Lett. **100**, 158101 (2008).

[6] M. Vejmelka and M. Palus, Inferring the directionality of coupling with conditional mutual information, Phys. Rev. E **77**, 026214 (2008).

[7] N. Ay and D. Polani, Information flows in causal networks, Adv. Complex Syst. **11**, 17 (2008).

[8] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, Transfer entropy—A model-free measure of effective connectivity for the neurosciences, J. Comput. Neurosci. **30**, 45 (2011).

[9] M. Wibral, N. Pampu, V. Priesemann, F. Siebenhühner, H. Seiwert, M. Lindner, J. T. Lizier, and R. Vicente, Measuring information-transfer delays, PLoS ONE **8**, e55809 (2013).

[10] Y. M. Song, J. Jeong, A. A. V. de los Reyes, D. Lim, C.-H. Cho, J. W. Yeom, T. Lee, J.-B. Lee, H.-J. Lee, and J. K. Kim, Causal dynamics of sleep, circadian rhythm, and mood symptoms in patients with major depression and bipolar disorder: Insights from longitudinal wearable device data, eBioMedicine **103**, 105094 (2024).

[11] S. H. Park, S. Ha, and J. K. Kim, A general model-based causal inference method overcomes the curse of synchrony and indirect effect, Nat. Commun. **14**, 4287 (2023).

[12] J. Runge, J. Heitzig, N. Marwan, and J. Kurths, Quantifying causal coupling strength: A lag-specific measure for multivariate time series related to transfer entropy, Phys. Rev. E **86**, 061121 (2012).

[13] J. Runge, J. Heitzig, V. Petoukhov, and J. Kurths, Escaping the curse of dimensionality in estimating multivariate transfer entropy, Phys. Rev. Lett. **108**, 258701 (2012).

[14] J. Sun and E. Bollt, Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings, Physica D **267**, 49 (2014).

[15] J. Sun, D. Taylor, and E. Bollt, Causal network inference by optimal causation entropy, SIAM J. Appl. Dyn. Syst. **14**, 73 (2015).

[16] J. Runge, R. V. Donner, and J. Kurths, Optimal model-free prediction from multivariate time series, Phys. Rev. E **91**, 052909 (2015).

[17] J. Runge, V. Petoukhov, J. F. Donges, J. Hlinka, N. Jajcay, M. Vejmelka, D. Hartman, N. Marwan, M. Palus, and J. Kurths, Identifying causal gateways and mediators in complex spatio-temporal systems, Nat. Commun. **6**, 8502 (2015).

[18] J. Runge, Causal network reconstruction from time series: From theoretical assumptions to practical estimation, Chaos **28**, 075310 (2018).

[19] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí, E. H. van Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang *et al.*, Inferring causation from time series in earth system sciences, Nat. Commun. **10**, 2553 (2019).

[20] L. Novelli, P. Wollstadt, P. Mediano, M. Wibral, and J. T. Lizier, Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing, Network Neurosci. **3**, 827 (2019).

[21] L. Novelli and J. T. Lizier, Inferring network properties from time series using transfer entropy and mutual information: Validation of multivariate versus bivariate approaches, Network Neurosci. **5**, 373 (2021).

[22] L. Novelli, F. M. Atay, J. Jost, and J. T. Lizier, Deriving pairwise transfer entropy from network structure and motifs, Proc. R. Soc. A **476**, 20190779 (2020).

[23] R. G. James, N. Barnett, and J. P. Crutchfield, Information flows? A critique of transfer entropies, Phys. Rev. Lett. **116**, 238701 (2016).

[24] J. G. Orlandi, O. Stetter, J. Soriano, T. Geisel, and D. Battaglia, Transfer entropy reconstruction and labeling of neuronal

connections from simulated calcium imaging, PLoS ONE **9**, e98842 (2014).

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

[26] R. E. Spinney, M. Prokopenko, and J. T. Lizier, Transfer entropy in continuous time, with applications to jump and neural spiking processes, Phys. Rev. E **95**, 032319 (2017).

[27] M. Kim, D. Newth, and P. Christen, Macro-level information transfer in social media: Reflections of crowd phenomena, Neurocomputing **172**, 84 (2016).

[28] J. Kim, S. T. Jakobsen, K. N. Natarajan, and K.-J. Won, Tenet: Gene network reconstruction using transfer entropy reveals key regulatory factors from single cell transcriptomic data, Nucl. Acids Res. **49**, e1 (2021).

[29] G. Weng, J. Kim, K. N. Natarajan, and K.-J. Won, scmTE: Multivariate transfer entropy builds interpretable compact gene regulatory networks by reducing false predictions, bioRxiv doi: 10.1101/2022.11.08.515579.

[30] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, Inc., New York, 1991).

[31] J. T. Lizier, Jidt: An information-theoretic toolkit for studying the dynamics of complex systems, Front. Robot. AI **1**, 11 (2014).

[32] I. Vlachos and D. Kugiumtzis, Nonuniform state-space reconstruction and coupling detection, Phys. Rev. E **82**, 016207 (2010).

[33] L. Faes, G. Nollo, and A. Porta, Information-based detection of nonlinear Granger causality in multivariate processes via a nonuniform embedding technique, Phys. Rev. E **83**, 051112 (2011).

[34] A. Montalto, L. Faes, and D. Marinazzo, MuTE: A MATLAB toolbox to compare established and novel estimators of the multivariate transfer entropy, PLoS ONE **9**, e109462 (2014).

[35] P. Wollstadt, J. T. Lizier, R. Vicente, C. Finn, M. Martínez-Zarzuela, P. Mediano, L. Novelli, and M. Wibral, IDTXL: The information dynamics toolkit XL: A PYTHON package for the efficient analysis of multivariate information dynamics in networks, Open Source Software **4**, 1081 (2019).

[36] L. A. David, A. C. Materna, J. Friedman, M. I. Campos-Baptista, M. C. Blackburn, A. Perrotta, S. E. Erdman, and E. J. Alm, Host lifestyle affects human microbiota on daily timescales, Genome Biol. **15**, R89 (2014).

[37] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, Science **286**, 509 (1999).

[38] M. E. J. Newman, Power laws, Pareto distributions and Zipf's law, Contemp. Phys. **46**, 323 (2005).

[39] A. Clauset, C. R. Shalizi, and M. E. J. Newman, Power-law distributions in empirical data, SIAM Rev. **51**, 661 (2009).

[40] J. Kreth, E. Helliwell, P. Treerat, and J. Merritt, Molecular commensalism: How oral corynebacteria and their extracellular membrane vesicles shape microbiome interactions, Front. Oral. Health **5**, 1410786 (2024).

[41] R. Francavilla, D. Ercolini, M. Piccolo, L. Vannini, S. Siragusa, F. D. Filippis, I. D. Pasquale, R. D. Cagno, M. D. Toma, G. Gozzi, D. I. Serrazanetti, M. D. Angelis, and M. Gobbetti, Salivary microbiota and metabolome associated with celiac disease, Appl. Environ. Microbiol. **80**, 3416 (2014).

[42] C. A. Brennan and W. S. Garrett, Fusobacterium nucleatum—Symbiont, opportunist and oncobacterium, Nat. Rev. Microbiol. **17**, 156 (2019).

[43] S. Groeger, Y. Zhou, S. Ruf, and J. Meyle, Pathogenic mechanisms of fusobacterium nucleatum on oral epithelial cells, Front. Oral. Health **3**, 831607 (2022).

[44] P. E. Kolenbrander, R. J. Palmer, Jr., S. Periasamy, and N. S. Jakubovics, Oral multispecies biofilm development and the key role of cell-cell distance, Nat. Rev. Microbiol. **8**, 471 (2010).

[45] P. Pignatelli, F. Nuccio, A. Piattelli, and M. Curia, The role of fusobacterium nucleatum in oral and colorectal carcinogenesis, Microorganisms **11**, 2358 (2023).

[46] P. Gholizadeh, H. Eslami, and H. S. Kafil, Carcinogenesis mechanisms of fusobacterium nucleatum, Biomed. Pharmacother. **89**, 918 (2017).

[47] A. A. Hagberg, D. A. Schult, and P. J. Swart, Exploring network structure, dynamics, and function using networkX, in *Proceedings of the 7th Python in Science Conference (SciPy2008)*, edited by G. Varoquaux, T. Vaught, and J. Millman (Pasadena, CA, 2008), p. 11.

[48] J. D. Hunter, Matplotlib: A 2D graphics environment, Comput. Sci. Eng. **9**, 90 (2007).