Data-driven causal analysis of observational time series: a synthesis

Alex E. Yuan[1,2*] and Wenying Shou[1*]

August 3, 2020

1 Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America, 2 Molecular and Cellular Biology PhD program, University of Washington, Seattle, Washington, United States of America

*** alexericyuan@gmail.com (AEY); wenying.shou@gmail.com (WS)**

Abstract

Complex systems such as microbial communities play key roles in global processes and human life, yet are often challenging to understand. Although mechanistic knowledge in biology is generally rooted in manipulative experiments, perturbing these systems can encounter practical and ethical barriers. Thus, extensive attempts have been made to infer causal knowledge by analyzing observations of taxon abundance over time. When, and to what extent, does this strategy yield genuine insight? Unfortunately, the literature of causal inference can be formidable and controversial, as it draws from divergent fields such as philosophy, statistics, econometrics, and chaos theory. Most benchmarking papers focus on performance details of causal inference approaches, rather than fundamental issues such as the the underlying assumptions and their reasons, conceptual distinctions, and universal limitations. Here, we provide a synthesis of popular causal inference approaches including pairwise correlation and Reichenbach's common cause principle, Granger causality, and state space reconstruction. We find that each of these requires that certain properties of the data do not change with time (e.g. "IID", "stationarity", "reverting dynamics"). We provide new ways of visualizing key concepts, point out important issues that have been under-emphasized, and in some cases describe novel pathologies of causal inference methods. Although our synthesis is motivated by microbial communities, all arguments apply to other types of dynamic systems. We strive to balance precision with accessibility, and hope that our synthesis will motivate future development on causal inference approaches. To facilitate communication to a broad audience, we have made an accompanying video walkthrough (https://youtu.be/TZvEk3jXQfY).

Introduction

Microbial communities play key roles in diverse environments, from the human body, to soil, to industrial food fermentation. Microbial communities are often highly complex, with many species engaging in diverse interactions such as the release and consumption of multiple chemical compounds. Thus, it is challenging to understand or control microbial communities [1, 2].

Ideally, biologists acquire mechanistic knowledge from manipulative experiments. However, manipulative experiments can be infeasible or inappropriate: Natural ecosystems may not present enough replicates for comprehensive manipulative experiments, and perturbations can be impractical at large scales and may have unanticipated negative consequences. Laboratory experiments, which are imperfect approximations of natural ecosystems, can also be difficult when species are "unculturable". Instead, we have an abundance of observational population dynamics data (i.e. species population size over time without intentional perturbations).

If a mechanistic understanding of the ecosystem is available, then we can use this understanding to write a set of equations, fit equations to the data, and obtain model parameters which may describe which species

interact and how strongly. In microbiology, this approach commonly uses the Lotka-Volterra equations, which assume that the growth rate of one species varies linearly with the population sizes of other species and environmental factors [3, 4, 5]. One can additionally include assumptions about the strength and number of interactions using statistical techniques known as Bayesian priors [5] and regularization penalties [3, 6]. Unfortunately, this approach is inadequate for multiple reasons [7, 8, 9], and it is often impossible to assess the appropriateness of a mechanistic model for a given ecosystem. As there are multiple thoughtful explorations of the many considerations related to mechanistic modeling [6, 10], we will not focus on these topics here.

Without resorting to mechanistic modeling, how might we formulate likely hypotheses on how one species affects another species? In this essay, we explore current data-driven approaches for inferring causal relationships between species from observational longitudinal data. We focus on absolute abundance data, since analyzing compositional or relative abundance data is known to be fraught with problems and can potentially be avoided [11, 12, 13, 14].

The literature of causal inference can be impenetrable, even for those with quantitative training. Causal inference approaches have originated from an extremely diverse set of communities including philosophy, statistics, econometrics, and chaos theory, each employing a different lexicon of specialized jargon. Moreover, as we will see, causality has multiple definitions. Some definitions have stringent requirements that cannot be easily checked, while others have more relaxed requirements but offer results of questionable value [15]. Indeed, research using causal inference from time series has been highly controversial, with an abundance of "letters to the editor", often followed by impassioned back and forths [16, 17, 18]. We suspect that these controversies reflect disagreement and confusion about when certain causality inference approaches are valid, and how to interpret their results under different contexts. Although benchmarking papers exist [19, 20, 21, 22, 23, 24], many focus on demonstrating successes or failures of particular methods under a particular set of conditions rather than on conceptual issues.

This essay is targeted towards biologists and medical scientists who wish to delve deeper into causal inference and how it may (or may not) be applied to biological questions. We strive to balance precision and readability, and provide extended discussions of key mathematical notions in the Appendices. We will first define causality in a way that suits the need of life scientists. Then, we will focus on three commonly-used approaches that have been applied to microbial communities [25, 26, 27, 28]: pairwise correlation and Reichenbach's common cause principle, Granger causality, and state space reconstruction. For each approach, we ask a series of questions: What is the rationale behind this approach? What flavor of causality does it describe? How does it relate to other approaches? When do theoretical guarantees, if any, hold? When can this approach fail, and why? Can known failure modes be comprehensively categorized and tested for? That is, can we run our data through a series of diagnostic tests that allow us to identify an inference method that is likely predictive?

Causality

Perturbation causality: Definition and properties

Multiple related but nonidentical definitions of causality have been put forward for dynamical systems [29, 30, 31]. To avoid confusion, here we define causality as "perturbation causality" (e.g. [32, 33]; [30] Proposition 6.13 & Section 10.1): Variable X causes variable Y if some externally applied perturbation of X at a point in time results in a perturbation in Y at a future (or current) time (Fig 1A). We focus on a binary notion of causality (i.e. X either does or does not cause Y). Perturbation causality is not transitive ([30] Example 6.34): For example in Fig 1B, X perturbation causes Y and Y perturbation causes W, but X does not perturbation cause W because X's positive effect on W through Y is canceled by a negative effect through another path. Fig 1C illustrates a notion of direct versus indirect causes. Perturbation causality is distinct from direct causality: In Fig 1B, X directly causes W but does not perturbation cause W; in Fig 1C, U does not directly cause Y but does perturbation cause Y. We will use hollow arrows as in Fig 1B (bottom panel) to denote perturbation causality throughout this article.

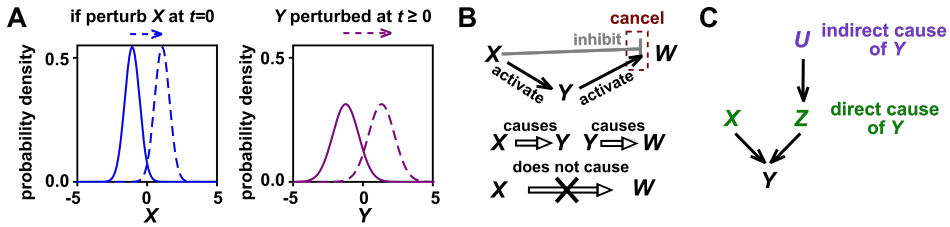

Figure 1: Perturbation causality. (**A**) If a perturbation in X causes future (or current) Y to change, then X perturbation causes Y. This definition does not require that *any* perturbation in X will perturb Y. For example, if the effect of X on Y has saturated, then a further increase in X will not affect Y. To embody probabilistic thinking (e.g. drunk driving increases the chance of accidents)[32], X and Y are depicted as probability distributions instead of single values (a "probability density" plot is basically a histogram). Perturbing X can perturb the current value of Y if, for example, X and Y are identical or if X and Y are linked by conservation law (e.g. conservation of energy). When sampling frequency is limited, perturbing X can also appear to perturb the current value of Y. For example, when stool samples are available only once a day, the "current" state can last many hours during which perturbing X can perturb "current" Y. (**B**) Perturbation causality is not transitive. In this example X perturbation causes Y and Y perturbation causes W, but X does not perturbation cause W due to a cancellation. (**C**) Direct versus indirect causes. The direct causes of Y are the minimal set of variables such that once the entire set is fixed, Y is not affected by any other variable. Here, three players (X, Z, and U) activate Y. Of these, X and Z are direct causes of Y, since if we fix both X and Z, then Y becomes independent of U. In contrast, the set $\{U, Z\}$ is not the set of direct causes of Y since Y is always affected by X. The direct causes of a variable are also known as its Markovian parents [32, 34]. Note that direct and indirect causes should be considered within the scope of included variables. For example, suppose that yeast releases acetate, which inhibits bacteria. If acetate is not in our scope, then yeast directly causes bacteria. Conversely, if acetate is included in our scope, then acetate directly causes bacteria while yeast indirectly causes bacteria.

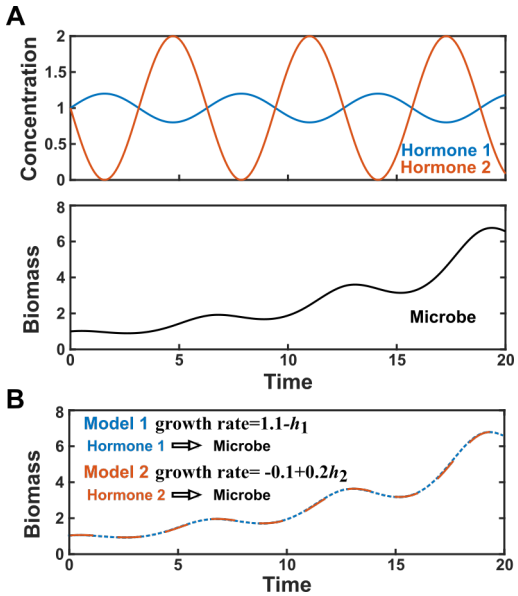

Figure 2: An example where perturbation causality cannot be inferred from observational data. (**A**) Observational data from two hormones and one microbe. (**B**) The microbial biomass data in (A) are consistent with both Model 1 (blue) where Hormone 1 (h_1) decreases the growth rate of Microbe, and Model 2 (orange) where Hormone 2 (h_2) increases the growth rate of Microbe. In other words, parameters of the model *growth rate* $= a + bh_1 + ch_2$ are not identifiable from the data.

The "nonidentifiability" problem of causal inference

In some cases, inferring causality from observational data is fundamentally limited by the "nonidentifiability" problem. We say that a causal network is nonidentifiable if data are not sufficient to fully specify the network. Imagine the ground truth is that X causes Y, and that the dynamics of Z is identical to that of X. Then without doing a perturbation experiment on X or Z, we cannot identify whether X or Z or both causes Y. More broadly, this problem can occur if $X(t)$ is a deterministic function of $Z(t)$ (Fig 2). Without performing perturbations, no approach can resolve a nonidentifiability problem.

Below, we will describe three commonly-used causal inference approaches. As we will see, the problem of nonidentifiability surfaces in different circumstances.

Pairwise correlation and Reichenbach's common cause principle

It is well-known that "correlation is not causality". In practice however, correlation approaches have often been used to hypothesize causal relationships [35, 36, 37]. We say that X and Y are correlated if an increase in X systematically accompanies an increase in Y (or a decrease in the case of anticorrelation).. A wide variety of formulae exist to quantify this notion. In the microbiome literature, authors often use "correlation" to refer to Pearson correlation or Spearman (rank) correlation.

When might we use correlation to infer causality? As we will see, although Reichenbach's common cause principle connects correlation to causality for certain types of data, inferring causality from temporal correlations is generally fraught. For example, when predicting which viruses infected which microbes in an *in silico* community, correlation approaches usually performed similarly to random guessing [22]. More broadly, two species might have correlated abundances because their dynamics share a common biotic or environmental driver [38, 4]. If this were the only problem, then correlation approaches would at least give us some information: a correlation between two species would indicate that these species belong to the same network of interacting biotic and abiotic factors. Unfortunately, even this cannot be guaranteed. As we discuss below, entirely independent temporal processes can also appear highly correlated [39, 38, 10].

Causally unrelated temporal processes can appear significantly correlated

Figure 3 provides three examples of causally unrelated temporal processes that nevertheless appear correlated. In the first example, two microbial cultures in exponential phase independently grown in different tubes have correlated densities because of their shared growth law (Fig 3A). In the second example involving island populations controlled by migrations, population sizes of two causally unrelated island populations usually appear significantly correlated according to currently popular statistical techniques (Fig 3B; see also [39, 10]). The third example considers island populations whose sizes are driven by immigration and cell death (Fig 3B). This final example demonstrates that independent processes can appear correlated even if they tend to return to an equilibrium (see also [40]).

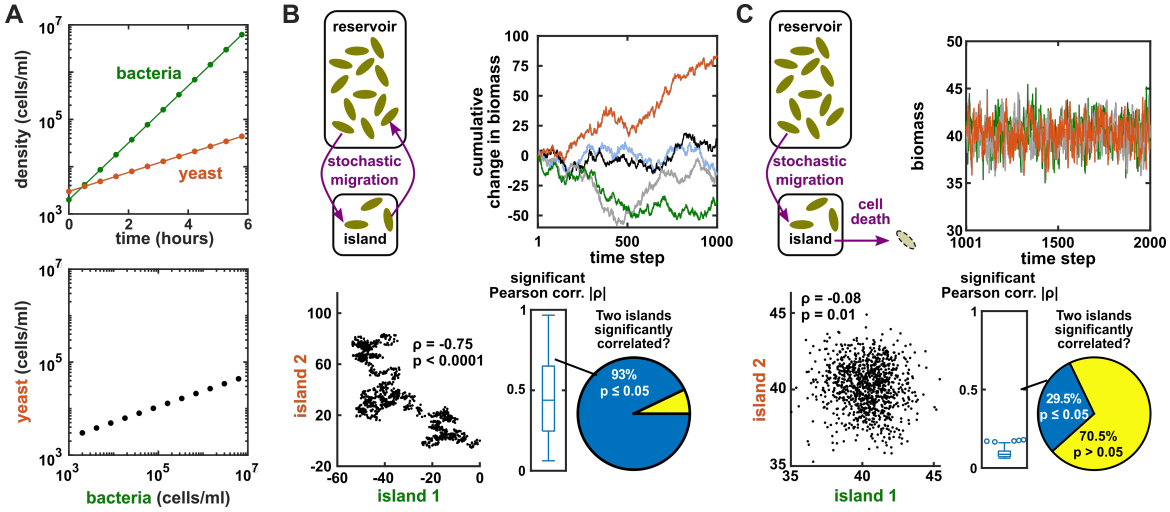

Figure 3: Two completely independent temporal processes can appear significantly correlated. (**A**) Top: Densities of yeast and bacteria cultures growing exponentially in two different tubes. Bottom: The densities of bacteria and yeast are correlated. (**B**) Consider a population (the "island") whose size is driven by individuals stochastically migrating to and from a "reservoir" (a random walk). At each time step, the net change in island biomass is drawn from a standard normal distribution (mean = 0; standard deviation = 1 biomass unit). Cumulative changes in biomass over 1000 time steps are plotted for five independent islands. We simulated 1000 pairs of independent islands and correlated the biomass of one island with that of the other island. In > 90% of these simulations, the Pearson correlation was significant (0.05 level, permutation test). For example, the blue and black trajectories are positively correlated, while the orange and green trajectories are negatively correlated. The magnitudes of significant correlations are shown in a boxplot. (**C**) Consider a population which receives cells through migration and loses cells to death and whose size has reached an equilibrium. Specifically, at each time step, 25% of the biomass is lost to death, and the biomass gained through migration is randomly chosen from a normal distribution with a mean of 10 biomass units and standard deviation of 1 unit. We simulated this process for 1000 pairs of independent islands, starting with an initial biomass of 40, for 2000 time steps, and correlated the final 1000 steps of one island with the final 1000 steps of the other island. In 29.5% of these simulations, the correlation was significant (0.05 level, permutation test). We discuss more correct methods of calculating p values for longitudinal data toward the end of this article. Circles in the boxplot are outliers (the boxplot of B has no outliers).

Time series typically violate the IID requirement of statistical correlation

Even if correlation does not always imply causality, might some forms of correlation suggest causality [41]? To answer this question, we need two concepts: random variables and IID (independent and identically distributed) data.

A random variable is, intuitively, a variable whose values (i.e. experimental measurements) depend on outcomes of a random phenomenon according to a particular probability distribution. Reichenbach's common cause principle (i.e. Principle 1.1 in [30]) connects random variables to causality in an intuitive

way. It states that if X and Y are two random variables, and if X and Y have a statistical dependence (such as a correlation; Appendix 1.1), then one or more of three statements is true: X perturbation causes Y, Y perturbation causes X, or a third variable Z perturbation causes both X and Y. Although on its own, this principle is an instance of nonidentifiability, it does narrow down the scope of causal hypotheses. The common cause principle can be proven ([30] Proposition 6.28) for systems that can be correctly described by a structural causal model (Appendix 1.3). However, the common cause principle contains crucial data requirements that often go overlooked.

Data requirements of Reichenbach's principle are not about relationships between random variables (capital letters $X, Y, Z, \cdots$), but rather about relationships between measurements of a single random variable (e.g. lower-case letters $x_1, x_2, \cdots x_n$). Specifically, all measurements $x_1, x_2, \cdots x_n$ must be independent and identically distributed (IID). That is, they must (1) follow the same probability distribution (since they are measurements of the same random phenomenon) and (2) be independent of one another (so that testing the dependence between two random variables X and Y makes mathematical sense) [41]. We call a dataset an "IID dataset" if for every random variable, each measurement is IID with respect to other measurements of the same random variable. IID relationships can also be discussed between two random variables (Appendix 1.2).

As an example of an IID dataset, consider a population of replicate mice where we observe that fecal biomass measurements of Microbe 1 and Microbe 2 are correlated, and we want to know whether we can expect a causal explanation. We can attempt to satisfy the requirement of independence by ensuring that none of the mice in this study interacts with any of the other mice in this study (e.g. sampling from different cages). We can attempt to satisfy the requirement of identical distributions by imposing similar environmental and genetic conditions on each mouse so that any differences in the biology of our replicate mice are random. Moreover, suppose that we measure Microbe 1 load in each of several mice where female mice tend to have a higher load than male mice. Our measurements can still be IID if we measure Microbe 1 from both male and female mice, as long as we randomly choose among male and female mice instead of preassigning a male to female ratio (Fig 4). In typical situations, we cannot guarantee IID with mathematical certainty (e.g. we cannot guarantee that no two human subjects interact), so the researcher must decide whether the IID assumption is adequate.

Most data with a temporal component are not IID and are thus not appropriate for the common cause principle. The data in Fig 3A are from a deterministic system, and therefore the concept of IID does not apply. Data points in Fig 3B are not identically distributed, since the variance increases with time. Data points within a time series in Fig 3B-C are not independent, since measurements at one time depend on measurements from previous times. We thus must be cautious about inferring causal relationships from correlations of longitudinal species density measurements. One needs to be vigilant when correlation assumes aliases such as "connections", "species-species associations", or any pairwise association method intended for IID data. The highly cited local similarity analysis [42] was originally one such method, although recent versions account for temporal dependence [43, 44]. As we discuss near the end of this article, falsely inferred causality arising from non-IID data can sometimes be mitigated by artfully chosen surrogate data tests, although this is not foolproof. Artifactual correlation may also arise from cross-sectional datasets when we pool measurements from independent but asynchronous replicates, unless these replicates have reached an equilibrium (Appendix 1.4).

In the upcoming sections, we describe two approaches that are designed for time-series: Granger causality for stochastic systems and state space reconstruction (SSR) for largely deterministic systems. A discussion on determinstic versus stochastic systems is provided in Appendix 1.8.

Granger causality

Granger causality is a statistical notion based on the idea that the history of a causer is useful in predicting the future of its causee [15]. We will first define Granger causality, then discuss practical considerations for testing Granger causality, and finally point out where Granger causality can fail as a heuristic for perturbation causality.

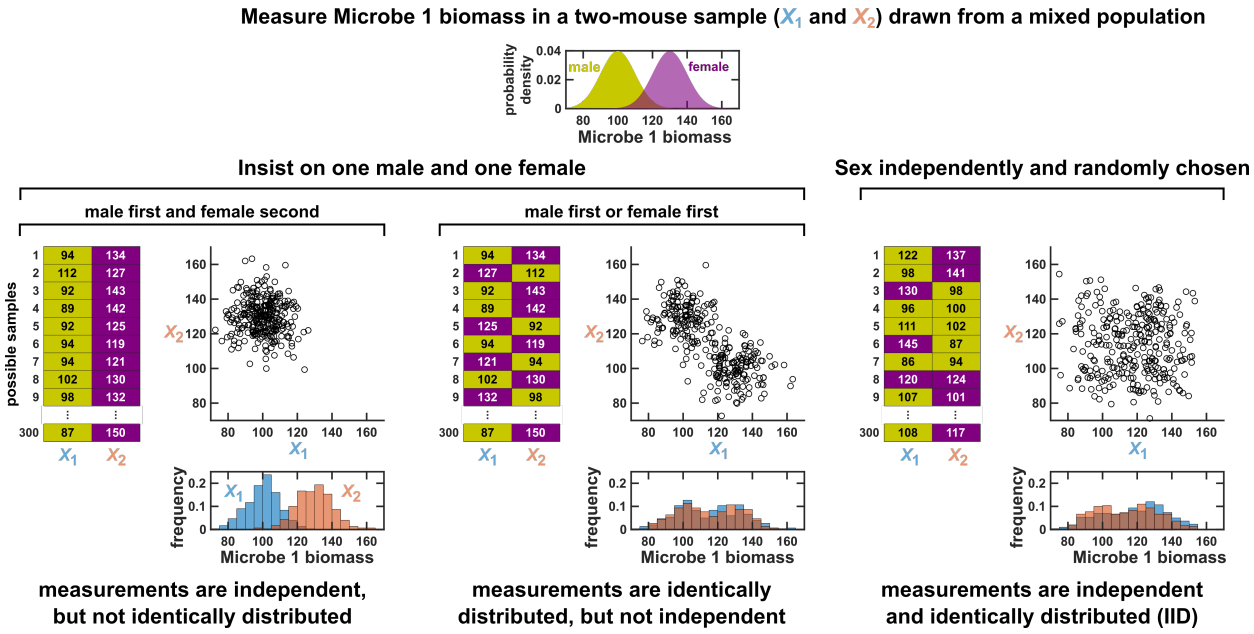

Figure 4: **Measurements taken from a mixed population may still be IID, as long as sampling is independent and random.** Consider a study in which we measure Microbe 1 biomass from a mixed population of low-load male mice and high-load female mice. For simplicity, suppose that each sample comprises measurements of two mice. To see whether a sample could be an IID dataset, we imagine drawing many possible versions of that sample, and ask whether our first measurement X_1 and second measurement X_2 are identically distributed and independent across possible samples. We could draw this sample in 3 different ways (3 sets of charts). On the left, we take our first measurement from a male and second measurement from a female. In this case, our two measurements are independent, but not identically distributed, so our sample is not an IID dataset. In the middle, we choose one male and one female per sample, but choose the first measurement randomly from a male or female. Now, our measurements are identically distributed (save for sampling error) but not independent (so also not IID). On the right, the sex of each measurement is randomly and independently chosen, and thus a sample could have two measurements from the same sex. In this case our sample is an IID dataset.

Granger causality: Intuition and definitions

Granger's precise definition of causality [29] is not practical for real world applications (e.g. we need to know all relevant and non-redundant information in the universe). Thus, multiple practical definitions of Granger causality have been proposed, including several by Granger himself [15]. The ensuing multiplicity of definitions can cause confusion. In neuroscience, it has been argued that Granger causality statistics can be meaningfully interpreted in their own right (instead of as a heuristic for perturbation causality) [45], although this idea has been questioned [46]. In Box 1, we summarize the intuition for Granger causality and state an operational definition most commonly used in microbiome research (linear Granger causality) [26, 28, 47, 24], as well as a more general definition used in fields such as economics and neuroscience (general Granger causality) [48, 49, 50, 51, 52].

Box 1: Granger causality

1. Intuition for Granger causality: Time series X Granger-causes time series Y if a dataset containing all relevant historical information predicts future Y better than a similar dataset that excludes the history of X.

2. Linear Granger causality: Let $X_t, Y_t, \cdots$ be series of random variables indexed by time t. Under linear Granger causality, X_t Granger-causes time series Y_t if including the history of X in a linear autoregressive model (Eq 1) allows for a better prediction of future Y than not including the history of X. By "linear autoregressive model", we mean that the future value of variable Y is modeled as a linear combination of historical values of X and Y and all other included variables "$\cdots$":

$$Y_{t+1} = c + \sum_{k=0}^{n} (\alpha_k X_{t-k} + \beta_k Y_{t-k} + \cdots) + \varepsilon_t \tag{1}$$

Here, c is a constant, $k = 0, 1, \cdots, n$ are the time lags included in the model, α and β are coefficients representing the strength of contributions from the respective terms. ε_t is IID and represents process noise (e.g. chance events such as death or migration) whose effects on population size propagate over time (Fig 9A). To test whether X Granger-causes Y, we fit our data to the above equation (without the process noise term ε_t which now shows up in residuals), and test whether the α_k coefficeints differ significantly from 0.

3. General Granger causality[15]: Let X_t Y_t, and Z_t be series of random variables indexed by time t. X Granger-causes Y with respect to the information set $\{X_t, Y_t, Z_t\}$ if:

$$f(Y_t|\{X_k, Y_k, Z_k \text{ for all } k < t\}) \neq f(Y_t|\{Y_k, Z_k \text{ for all } k < t\}) \tag{2}$$

where $f(Y_t|\mathcal{S})$ is the probability distribution of Y_t conditional on $\mathcal{S}$. Unlike linear Granger causality, general Granger causality does not make assumptions about the underlying process, and is thus considered "model free". Note that Z_k in Eq. 2 may be multivariate and thus plays the same role as "$\cdots$" in Eq. 1.

Both linear and general Granger causality require all time series to be stationary. In statistics, a stationary time series is one whose statistical properties do not change over time (see Appendix 1.5 for details). For linear Granger causality, although we can fit Eq. 1 to nonstationary data, standard statistical tests of the null hypothesis (i.e. that all α_k terms in Eq. 1 are 0; e.g. [53]) are invalid when data are nonstationary [54]. Indeed when applied to nonstationary time series[55, 56] such as random walks (Fig 5Bii, [57]), statistical tests produce false positive results more often than they should (i.e. the p-values are miscalibrated) . For general Granger causality, one would ideally conduct many independent trials, and then estimate and compare the two distributions in Eq. 2 across time. Indeed, if one could accomplish this, then Theorem 10.3 of [30] states additional theoretical conditions under which Granger causality is equivalent to direct causality. However, obtaining multiple independent trials is often not feasible, and Granger causality is most commonly used with many time points from a single trial. This is only valid if a time series is strongly stationary. Briefly, this means that the distribution f in Eq. 2 is independent of time so that each distribution can be estimated from historical data of a single trial.

Granger causality also requires that the dynamics be stochastic. This is because in deterministic systems, causer history may already be "encoded" in causee history [58, 59, 60], and thus does not improve prediction

of causee (Fig 5, Bi).

Practical considerations for testing Granger causality

Granger causality requires that data are stochastic and stationary. Microbial dynamics can generally be assumed to be at least partially stochastic since they are affected by many random factors. In fact, even a completely deterministic system can appear stochastic if it has many unobserved variables (Appendix 1.8).

Testing for stationarity is problematic when we only have a single time series, because stationarity is by definition a property of a population of time series (Fig 15). Therefore, any test for stationarity of a single time series must make assumptions. For example some studies (e.g. [61, 26]) used the ADF test for stationarity, which assumes a model similar to Eq. 1 (with only Y), but where the process noise terms ϵ_t may be correlated across time [62]. If a time series is nonstationary, one can sometimes "make it stationary" by computing the rate of change (differencing) or subtracting a deterministic trend (detrending) [26], although naive detrending procedures do not always work well [63].

In microbiome research, linear Granger causality seems to be more popular than general Granger causality [26, 28, 47, 24]. Tests for linear Granger causality are straightforward to interpret, computationally inexpensive, and available in multiple free and well-documented software packages [64, 53]. However, the linear autoregressive model is only a convenient statistical approximation, and linear Granger causality has been criticized as encouraging model-based inferences before assessing model validity [65]. Still, we can assess the plausibility of linear autoregression (Eq. 1) after fitting to data (e.g. by checking whether the model residuals ε_t are uncorrelated across time [66], as is assumed by Eq. 1).

To test for general Granger causality, it is common to compute a statistic known as transfer entropy: X Granger causes Y if and only if the transfer entropy from X to Y is nonzero ([52], Appendix 1.6). Tests for general Granger causality using transfer entropy are available for two-variable systems (e.g. the package RTransferEntropy [67]) and for multivariable systems (IDTxl [68]). Significance tests are based on a type of null model called surrogate data, which we discuss near the end of this article.

How do Granger causality and perturbation causality differ?

Granger causality tests heuristically for direct perturbation causality, and excludes indirect perturbation causality. To see this, consider a time series with 3 observed variables X, Y, and Z where Z of tomorrow is caused by Y of today, and Y of today is caused by X of yesterday (Fig 5A). Thus, once we know Y of today, knowledge of X of yesterday is no longer useful for predicting Z of tomorrow. Therefore, while X perturbation causes Z indirectly through Y, we would nonetheless say that X does not Granger cause Z. In this example, we can also see that Granger causality is not transitive: X Granger causes Y and Y Granger causes Z, but X does not Granger cause Z (Fig 5A).

As a heuristic for perturbation causality, Granger causality suffers several failure modes. As alluded to above, deterministic systems and nonstationary data can create problems for Granger causality (Fig 5B i and ii). Moreover, an unobserved common driver can induce Granger causality without the corresponding perturbation causality. For example, X drives Y with a lag of 1 and drives Z with a lag of 2. If X is not observed, then Y appears to Granger cause Z because Y receives the same "information" before Z and thus, historical data of Y helps in predicting Z (Fig 5 Biii). Thus, any application of Granger causality in an open system is technically "nonidentifiabile" because one can rarely rule out the possibility of a common cause. Additionally, if sampling frequency is too low, then perturbation causality may not be detected by Granger causality [29]. Suppose that X varies frequently with time and drives Y after a short time lag, but X and Y are sampled infrequently (Fig 5Biv). Then, past values of X will not appear to be useful in predicting current Y.

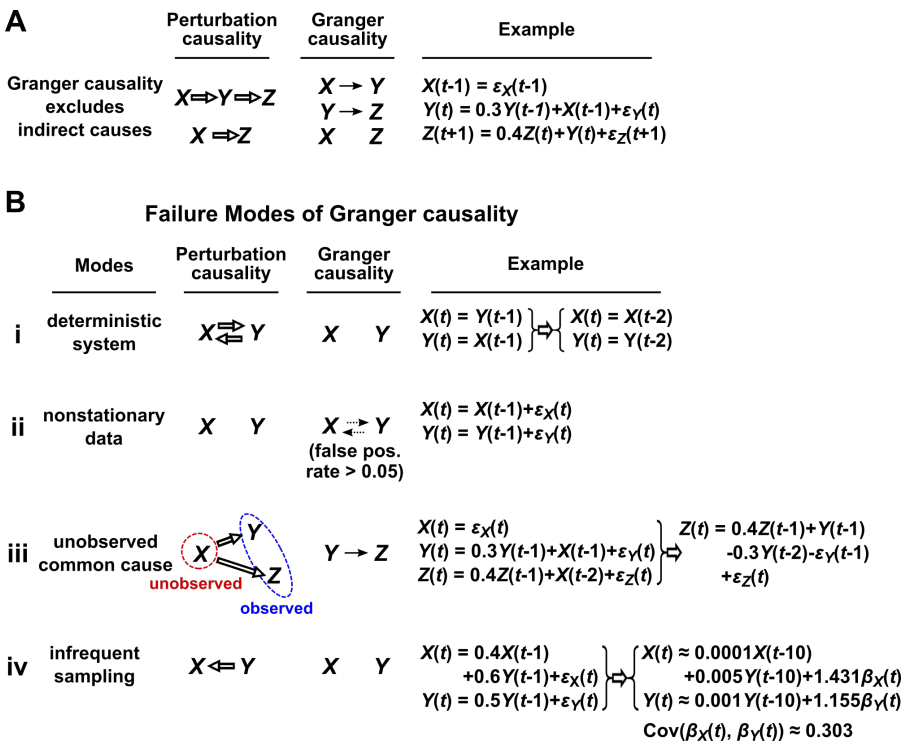

Figure 5: Perturbation causality versus Granger causality. (**A**) Granger causality is not designed to uncover indirect perturbation causes. Although X perturbation causes Z, X does not Granger cause Z because with the history of Y available, the history of X no longer adds value for predicting Z. (**B**) Failure modes of Granger causality when inferring perturbation causality. ϵ_X, ϵ_Y, ϵ_Z, β_X, and β_Y are normal random variables with mean of 0 and variance of 1 and represent process noise. ϵ_X, ϵ_Y, and ϵ_Z are uncorrelated with one another. Coefficients are chosen to ensure stationarity, except in (**ii**). (**i**) False negative due to lack of stochasticity. X and Y mutually and deterministically cause one another through a copy operation [30]: $X(t)$ copies $Y(t-1)$ and vice versa. In this case, since $X(t-2)$ already contains sufficient information to know $X(t)$ exactly, the history of Y cannot improve prediction of X, and so Y does not Granger cause X. By symmetry, X does not Granger cause Y. (**ii**) False positives appear more often than expected if data are not stationary [55, 56] (dotted arrows). The dynamics are identical to those in Fig 3B. (**iii**) False positive due to unobserved common cause. X causes Y with a short delay, and causes Z with a long delay. We only observe Y and Z. Since the history of Y helps to predict Z, Y Granger causes Z, resulting in a false positive. (**iv**) Infrequent sampling induces false negatives by reducing the correlation between current value of causee and previous value(s) of causer. The quantitative relationship is derived in Appendix 1.7. Note that in the equation of X, the ratio of signal (the coefficient of causer Y) to noise (the coefficient of process noise β or ε) is much smaller when sampling is every ten time steps (right) compared to every one time step (left).

State space reconstruction (SSR)

In contrast to Granger causality, which relies entirely on stochastic process noise (Fig B5), the "state space reconstruction" (SSR) approach is intended for systems that are primarily deterministic. SSR has been widely proposed as an alternative to model-based inference for detecting interactions within microbial ecosystems [69, 70, 27, 71, 24, 10]. Intuitively, the SSR approach is based on the idea that if X drives Y, then a record of the X's influence will be present in the history of Y. In this section, we will first walk through an example of SSR working successfully in a toy deterministic system, using a new way to visualize potential causal relationships. We will then point out that SSR lacks theoretical guarantees when used to infer perturbation causality. To illustrate this insufficiency, we will describe some failure modes and associated countermeasures which reduce or diagnose, but do not eliminate, failures.

Visualizing SSR causal inference

We will walk through an example of SSR correctly inferring causality, using our visualization method. Consider the deterministic system whose equations are given in Figure 6A, and time series are shown in 6B. In this system, Z is perturbation caused by Y, but not W. We can take the current value $Z(t)$ and two past values $Z(t-\tau)$ and $Z(t-2\tau)$ (red dots) to form a delay vector (Fig 6B). This delay vector $[Z(t), Z(t-\tau), Z(t-2\tau)]$ can be represented as a single point in the 3-dimensional Z delay space (Fig 6C, red dot). We then shade each point of the Z delay space according to the contemporaneous value of Y or $Y(t)$. Since each point of the Z delay space corresponds to one and only one $Y(t)$ value, we call this a "delay map" from Z to Y. Notice that the $Y(t)$ gradient in this plot looks gradual in the sense that if two time points are nearby in the delay space of Z, then their corresponding $Y(t)$ shades are also similar. This property is called "continuity" (Fig 18). We provide additional details on continuity, and the more stringent criterion of smoothness, in Appendix 1.11. Overall, there is a continuous map from the Z delay space to Y, or more concisely, a "continuous delay map" from Z to Y. A similar continuous delay map also exists from Z to causer X. On the other hand, if we shade the delay space of Z by W (which does not cause Z), the relationship is not continuous (Fig 6D). Thus, a continuous map seems to exist from the delay space of a causee to the current values of a causer. This phenomenon is mostly insensitive to the choice of τ and the delay vector length (although a delay vector that is too short will cause problems; see Appendix 1.9). We note that there is a striking contrast between Granger causality and SSR: Whereas in Granger causality the history of a causer can predict the future of its causee, in SSR the history of a causee can estimate the current value of its causer.

The SSR approach [31, 72, 60] essentially relies on this heuristic assumption that A causes B if and only if there is a continuous delay map from B to A. While it has been argued that continuous delay maps can be interpreted in information theoretic terms [31], here we focus on SSR methods purely as heuristics for inferring perturbation causality. As a side note, some authors [72] look for the smoothness (Fig 18) of a delay map as evidence of causality, rather than the related (but weaker) continuity property as we do here. As we discuss in Appendix 1.10, these two criteria arise from slightly different theoretical assumptions. For brevity we focus only on continuity, which is advocated by [73] and is easier to detect with our shading visualization.

SSR failure modes

When can we be confident that SSR will successfully reveal perturbation causality? Ideally, we would like to find a set of conditions where we could show theoretically that "X perturbation causes Y if and only if there exists a continuous delay map from Y to X". Unfortunately, we are not aware of any general conditions, even if a system is purely deterministic. It is commonly thought that the use of SSR to detect causality is justified by theorems from Takens [58] and Sauer et al. [59]. However, these theorems state the conditions under which continuous delay maps are likely (but not quite guaranteed) to arise. They do not show (or even suggest) that the existence of a continuous delay map implies a causal relationship. For the interested reader, we visually explain Takens's theorem and briefly discuss its extensions in Appendix 1.10.

Fig 7 illustrates some failure modes where a causal relationship and a continuous delay map do not coincide. The top row of Fig 7 shows a failure mode which we call "nonreverting continuous dynamics". This example consists of two unrelated time series: a wavy linear increase and a parabolic trajectory. Despite being unrelated, we can find continuous delay maps between them. This is because there is (i) a continuous map from the delay vector $[X(t), X(t-\tau)]$ to t ("nonreverting X"), and (ii) a continuous map from t to Z ("continuous Z"), and thus there is a continuous delay map from X to Z ("nonreverting continuous dynamics"). Nonreverting continuous dynamics may lead one to infer causality where there is none. In simpler language, given the current value and sufficient history of X, we can accurately determine the current time, which allows accurate estimation of the current Z. Swapping X and Z in the above argument explains why we see a continuous delay map in the other direction. The problem of nonreverting continuous dynamics in SSR is similar to the non-stationarity problem in Granger causality, although they are distinct. For example, reverting dynamics may be nonstationary (Fig 15 bottom row).

The second row of Fig 7 shows a failure mode where a continuous delay map appears in both directions even though the causal relationship is only unidirectional. This has been called "strong forcing" or "synchrony" [60], because intuitively, it may arise when one variable impacts another variable very strongly so that the dynamics of the two variables are synchronized. This is an instance of nonidentifiability because synchrony

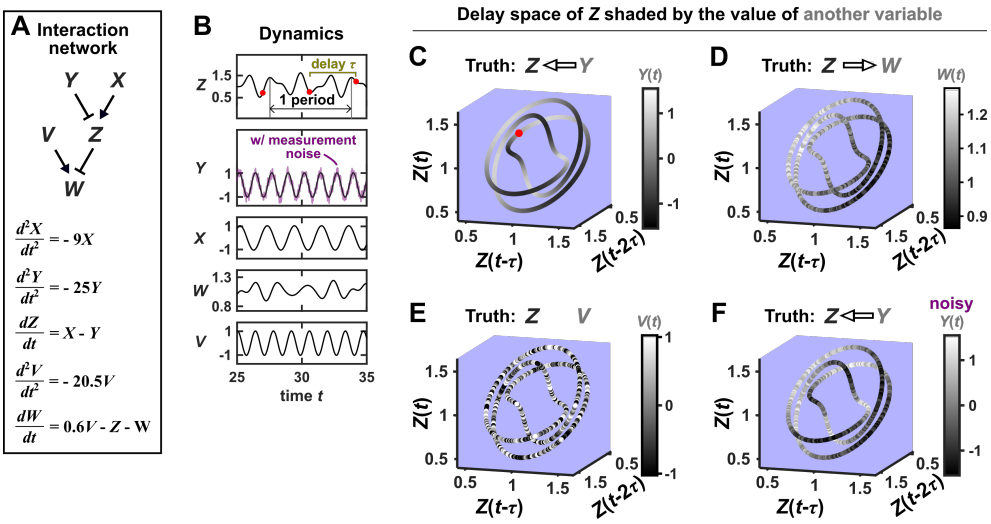

Figure 6: SSR causal inference detects continuous delay maps from noiseless systems, and becomes more difficult with noise. (**A**) A 5-variable toy system. (**B**) Time series. The delay vector $[Z(t), Z(t-\tau), Z(t-2\tau)]$ can be represented as a single point in the 3-dimensional Z delay space (**C**, red dot). (**C**) We then shade each point in the Z delay space by its corresponding contemporaneous value of $Y(t)$. The shading is continuous (i.e. gradual shade transitions), consistent with Y causing Z. (**D**) When we repeat this procedure, but now shade the Z delay space by $W(t)$, the shading is bumpy, consistent with W not causing Z. However, there is still a global pattern in **D** (right side darker than left side) due to Z's effect on W. (**E**) Shading the delay space of Z by the causally unrelated V is not only bumpy, but also yields no global pattern. Note that sometimes a global pattern can still be observed (replace 20.5 in the equation for V by the square of a rational number) or even a continuous pattern (replace 20.5 by the square of an integer; see the failure mode in Fig 7 row 3). (**F**) Dynamics as in (**C**), but now with noisy measurements of Y (purple in **B**). The shading is no longer gradual. Thus with noisy data, causal inference is difficult (compare **D** and **F**).

prevents us from resolving the direction of causality. A strategy (the "prediction lag" test; Fig 8C right panel) has been proposed to resolve the direction of causality in this scenario [74]. However, we report here that this test suffers a range of failure modes including unclear results with periodic drivers, false negatives, and false positives (Appendix 1.12). Nevertheless, this test dramatically improved the performance of SSR-based causality detection in a simulation model of disease transmission [19].

In the third row of Fig 7, the oscillation period of X is an integer multiple of that of Y (e.g. X has a period of a week and Z has a period of a day). A continuous delay map exists from X to Z, but not from Z to X, even if X and Z are causally unrelated.

The bottom row of Fig 7 shows a case in which SSR gives a false negative error. Here, X has a causal influence on Z, but this influence is not detected. This is because, as we mentioned earlier, satisfying conditions in Takens's theorem makes continuous mapping likely but not guaranteed (Appendix 1.10). Here, Z is an example of this lack of guarantee [75]. However, systems with this type of "pathological symmetry" may be rare in real biological settings.

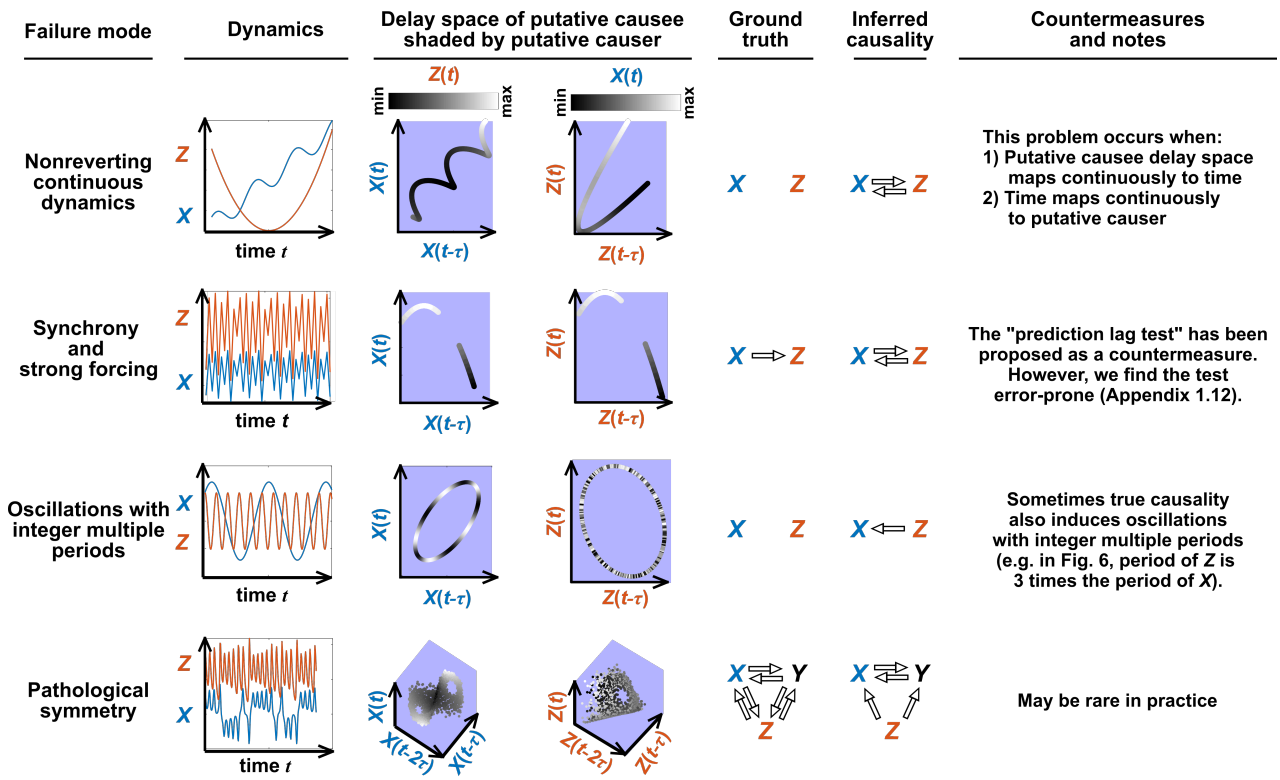

Figure 7: Failure modes associated with state space reconstruction. **Top row**: A smooth delay map can occur between two causally unrelated time series when we have nonreverting continuous dynamics. That is, a continuous map can be found from the delay space of one variable to time and from time to another variable. **Second row**: X drives Z in a manner such that their dynamics are "synchronzied", and consequently, we find a continuous delay map also from X to Z even though Z does not drive X. Note that the extent of synchronization is not always apparent from inspecting equations (e.g. Fig 12 of [21]) or dynamics (Fig 19). See rows 1 and 5 in Fig 19 for examples of strong forcing. **Third row**: X oscillates at a period that is 5 times the oscillatory period of Y. There is a continuous delay map from X to Z even through X and Z are causally unrelated. **Bottom row**: In the classic chaotic Lorenz attractor, the Z variable is perturbation caused by X and Y, but we do not see a continuous map from the delay space of Z to either X (shown) or Y (not shown). This is due to a symmetry in the system that causes the dynamics of X and Y to fold over on themselves when viewed from the Z axis ("Background definitions for causation in dynamic systems" in Supplementary Information of [60]).

Convergent cross mapping: Detecting SSR causal signals from real data

Although a continuous delay map is neither necessary nor sufficient for a causal relationship even in fully deterministic systems, one might still attempt to use it to hypothesize causal relationships. In this case, one must test for continuity. If we have infinite and noiseless data, then continuity is a natural notion and can be mathematically formalized (Fig 18). However, real data are finite and noisy. Thus, if we observe a delay map from a putative causee to a putative causer that is not continuous, we need to decide whether our delay map is discontinuous only because of noise and finite sampling, or instead because the putative causal relationship is not there (Fig 6, compare D and F).

Several methods can be used to detect SSR causal signals by detecting approximate continuity [73] or related properties [31, 72, 60]. The most popular is convergent cross mapping (CCM) [60], which is based on "cross map skill". Cross map skill quantifies how well a causer can be predicted from delay vectors of its causee (Fig 8A), conceptually similar to checking for gradual transitions when shading the causee delay space by causer values (Fig 6). Cross map skill is thought to detect local smoothness (and thus local continuity) of the delay map (Appendix 1.10 and [72]). Four criteria have been proposed to infer causality [60, 74, 19] (Fig 8B): First, the cross map skill must be positive. That is, when using the causee delay space to predict the causer, the Pearson correlation coefficient between true values and predictions must be positive. Second, the cross map skill must be significant. That is, the cross map skill must be greater than the "null" cross map skill when the putative causer is replaced with"surrogate data" (a type of null model discussed toward the end of this article). Third, the cross map skill must increase with an increasing amount of training data. Last, cross map skill must be greater when predicting past values of causer than when predicting future values of causer (the "prediction lag" test; [74, 19]). This test was proposed to overcome the strong forcing failure mode (the second row of Fig 7). However, see Appendix 1.12 for problems that we have found using this criterion. In practice, many CCM analyses use only a subset of these four criteria [60, 76, 77, 78]. Other approaches to detect various aspects of continuous delay maps have also been proposed [31, 73, 72]. We do not know of a systematic comparison of these alternatives.

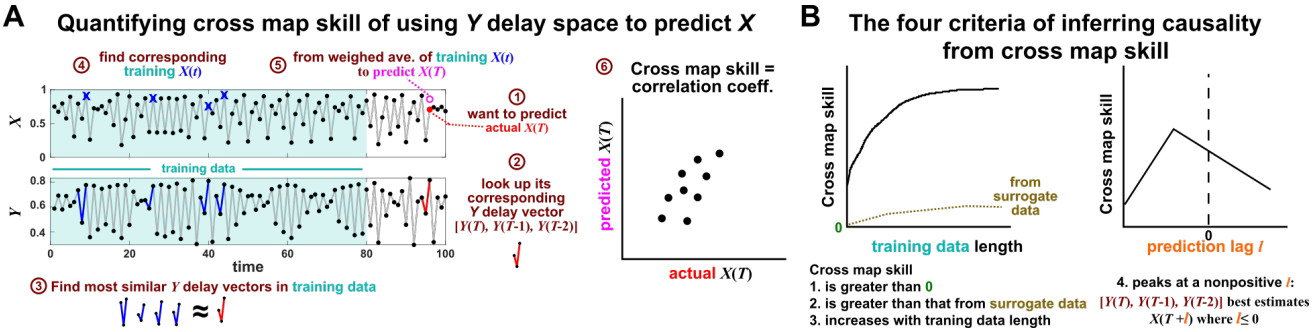

Figure 8: Illustration of the convergent cross mapping (CCM) procedure for testing whether X causes Y. (**A**) Quantifying cross map skill. Consider $X(T)$ denoted by the red dot ("actual $X(T)$" in **1**), which we want to predict from Y delay vectors. We first look up the contemporaneous Y delay vector ($[Y(T), Y(T-1), Y(T-2)]$) (red dynamics in **2**), and identify times within our training data when delay vectors of Y were the most similar to our red delay vector (blue segments **3**). We then look up their contemporaneous values of X (**4**, blue crosses), and use their weighted average to predict $X(T)$ (**5**, open magenta circle). We repeat this procedure for many choices of $X(T)$ and calculate the Pearson correlation between the actual values of $X(T)$ and our predicted values of $X(T)$ (**6**). This correlation is called the "cross map skill". While Sugihara et al. [60] note that other measures of cross map skill, such as mean squared error, may also be used, this article follows the convention of [60] in using Pearson correlation to measure cross map skill. (**B**) Four criteria for inferring causality from the cross map skill. Data points in **A** are marked by dots and connecting lines serve as visual aids.

Granger causality and SSR performance are sensitive to environmental drivers, process noise, and measurement noise

In this section, we test how well Granger causal inference and SSR/CCM perform with one toy system. We constructed a toy ecological system whose true causal structure is known, and applied a linear Granger causality test (using the MVGC package) and CCM (using the rEDM package) to test how well we can infer causal relationships.

We simulated a two-species community in which one species (S_1) perturbation causes the other species (S_2) but S_2 has no influence on S_1 (Fig 9B). Additionally, S_1 is influenced by an unobserved periodic external driver and S_2 either is (Fig 9D) or is not (Fig 9E) influenced by its own (also unobserved) periodic external driver. Species dynamics were modeled by the Lotka-Volterra equations, which are valid when, for example, S_1 affects S_2 via a reusable chemical [9]. We added process noise to model the stochastic and/or many-variable nature of typical ecosystems (Appendix 1.8) and added measurement noise to model measurement uncertainty. Process noise, such as those resulting from stochastic migration or death, propagates to future time steps (Fig 9A). In contrast, measurement noise does not propagate over time, and includes instrument noise as well as ecological processes during sampling (e.g. cell births within a stool sample as the stool passes through the intestine). Unlike in linear Granger causality, there is no standard and automatic test procedure for CCM causality criteria [19, 24]. We therefore tested for CCM criteria using two different procedures (Fig 9 legend and Methods).

Although Granger causality and CCM can perform well when requirements are met, both are highly sensitive to the levels of process and measurement noise (see the fraction of green in pie charts of Fig 9D and E; see also [79]). Unfortunately, where a system lies in the spectrum of process versus measurement noise is often unknown and we are not aware of any test that reliably distinguishes between process noise and measurement noise. Both Granger causality and CCM are also sensitive to details of the ecosystem (whether or not S_2 has its own external driver; compare Fig 9D with E). CCM is additionally sensitive to test procedure details (Fig 9 D and E, olive brackets).

Model-free causality tests are not assumption-free

When inferring causality, it is important to ask whether the signal for causality could have been generated even in the absence of a real causal relationship. That is, we need to calculate the significance of our inferred causal relationship. If our test relies on a model such as when testing for linear Granger causality, we can compute a p-value with a standard parametric test. However, if our heuristic for causality is model-free as with pairwise correlation, general Granger causality, or many SSR methods, standard parametric statistical tests generally do not make sense. Instead, it is common to create a synthetic control dataset known as "surrogate data". Surrogate data retain certain aspects of the original dataset, but would be unlikely if there was a causal relationship. We then repeat the causality test on surrogate data to see whether they give a signal that is as strong as the data. If so, then our evidence of a causal relationship is likely suspect. Many methods exist for generating surrogate data, some more appropriate than others.

Ideally, our surrogate data should preserve important aspects of the data that do not depend on a causal relationship, while removing aspects of the data that do depend on a causal relationship. Since we do not know which aspect of our data depends on a causal relationship, we need to make assumptions. A popular approach for generating surrogate data in the microbiome field is to shuffle experimental measurements [42]. This procedure, known as a *permutation test*, preserves the probability mass function of each time series (i.e. the histogram of population size). However, the permutaiton test is often problematic for temporal data because it destroys the temporal dependence in the data. In particular, early (and highly cited) versions of the popular local similarity analysis [42, 82] used a permutation test to generate surrogate data, which can create false positive errors, even for simple stationary systems [43, 44]. More recent versions have incorporated alternative tests that account for temporal dependence [43, 44]. As another example, the *stationary bootstrap* resamples data in blocks, keeping nearby data points together [83, 52]. Yet another procedure, known as the *random phase test*, first uses the Fourier transform to represent a time series as a sum of sine waves, then randomly shifts each of the component sine waves in time, and finally sums the phase-shifted components [84] (Fig 21). This procedure preserves the power spectrum (frequency components)

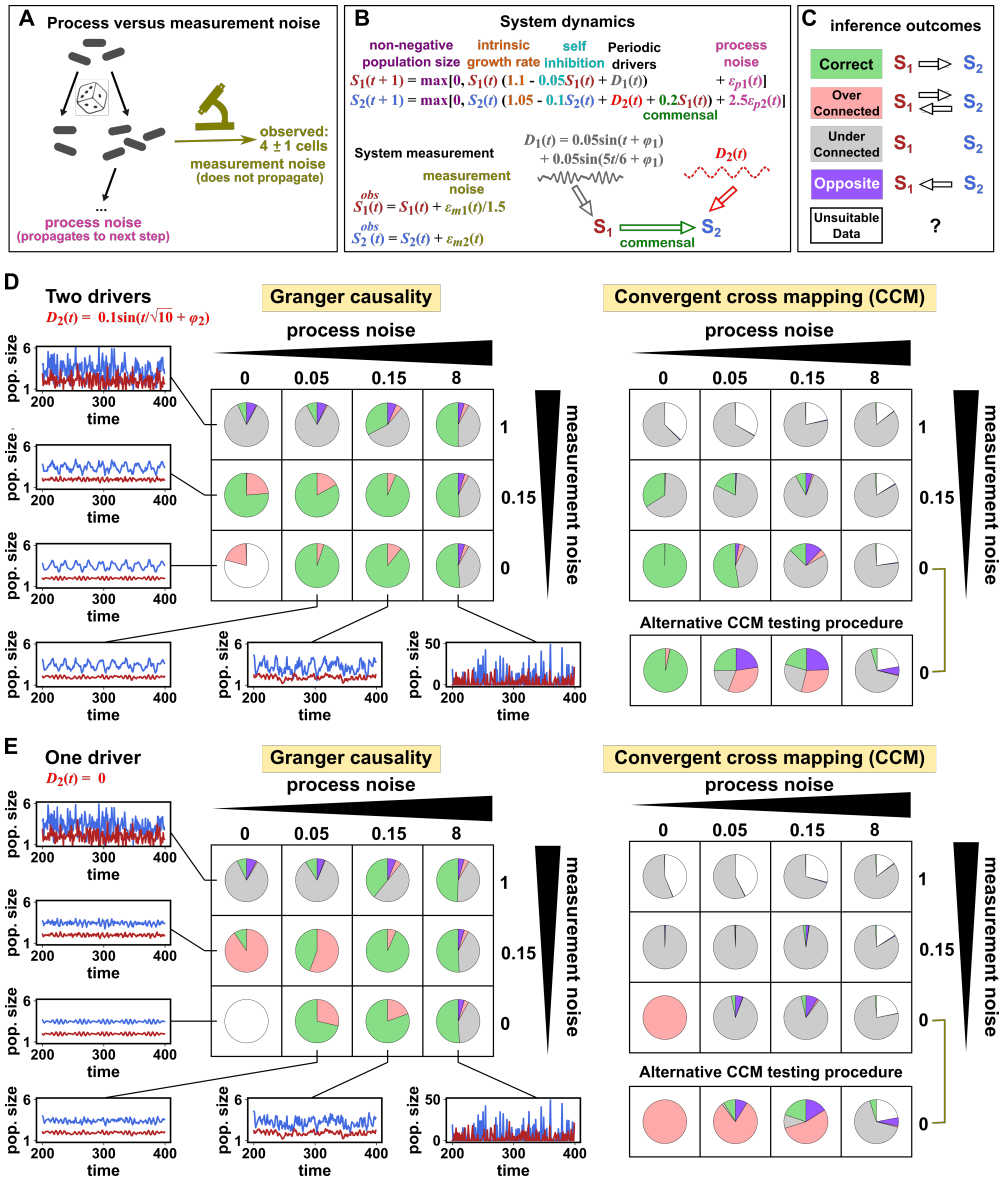

Figure 9: Performance of Granger causality and convergent cross mapping (an SSR method) in a toy model with noise. (**A**) The effect of process noise but not measurement noise propagates to samples taken at subsequent time points. (**B**) We simulated a two-species commensal community with one (**E**) or two (**D**) external drivers. The process noise terms $\epsilon_{p1}(t)$ and $\epsilon_{p2}(t)$, as well as the measurement noise terms $\epsilon_{m1}(t)$ and $\epsilon_{m2}(t)$, are IID normal random variables with a mean of zero and a standard deviation whose value we vary. (**C**) Five possible outcomes of causal inference. (**D**, **E**) Community dynamics and causal inference outcomes. We varied the level (i.e. standard deviation) of process noise and measurement noise. Each pie chart shows the distribution of inference outcomes from 1000 independent replicates. Granger causality: In the deterministic setting (lower left corner), the MVGC tool correctly rejects the data as inappropriate. With modest (but not high) process noise and no measurement noise, Granger causality frequently infers the correct causal structure. However, adding an intermediate level of measurement noise can dramatically increase the false positive rate (salmon color). MVGC's checks for data suitability may not flag pathological measurement noise [20]. Smoothing data before Granger causal inference can partially alleviate problems due to measurement noise [80]. CCM: The performance of CCM is sensitive to test procedure and ecological details (one versus two drivers), and quickly deteriorates with measurement or process noise. Note that when S_2 does not have its own external driver (**E**), CCM cannot infer causality in the deterministic case due to the synchronization of S_1 and S_2. In both the main and alternative CCM procedures, distributions of cross map skill (ρ) were computed by bootstrapping to test for criterion 1 (positive ρ), and random phase surrogate data were used to test criterion 2 (significance of ρ). Because the dynamics are periodic, criterion 4 (prediction lag test) was not used (Fig 19). The two procedures differ only in how they test criterion 3 (increases with more training data): the main procedure uses bootstrap testing following [19]while the alternative procedure uses a Kendall's τ as suggested by [81].

instead of the probability mass function of a time series, and thus preserves some temporal features. The random phase test thus assumes that the power spectrum of each time series does not depend on the presence of a causal relationship. In Fig 10, we repeat the analysis of Fig 3C using both the permutation test and the random phase test of [84]. The false positive rate (blue wedge) in random phase test is reduced to 5%, consistent with the p-value threshold of 5%. Thus, due to different underlying assumptions, different methods of generating surrogate data can yield drastically different results.

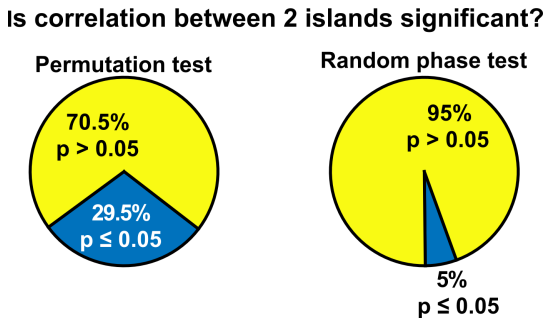

Figure 10: **Appropriate surrogate data can sometimes mitigate false positive errors**. Shown here is the same system shown in Fig 3C in which a population (the island) receives cells through immigration and loses cells to death, and population size has reached an equilibrium. We correlate the biomass of two independent islands against one another (using Pearson correlation) and test whether this correlation is significant at the 0.05 level. We run this numerical experiment 1000 times and show the fraction of times we find significance as pie charts. The permutation test, which randomly shuffles one of the time series, finds a spurious correlation in 29.5% of runs. In contrast, the random phase test [84] finds a spurious correlation in only 5% of cases, consistent with the significance cutoff of 0.05.

Closing thoughts

Inferring perturbation causality from observational time series data is tantalizing but challenging. Although nonidentifiability remains a fundamental problem, proper causal inference can reduce the hypothesis space. Pairwise correlation approaches that assume IID data frequently produce erroneous results when applied to time series data. Granger causality and state space reconstruction are two causal inference approaches designed with time series data in mind. Whereas Granger causality is designed for stochastic systems, SSR is designed for largely deterministic systems. Both approaches have predictive power in certain situations, but they are far from perfect (Fig 9). Both can break down in similar ways (Table 1). Whereas Granger causality tests are not valid for nonstationary data, state space reconstruction has its own kind of problem with "nonstationarity", which we call "nonreverting continuous dynamics". For both, failure modes cannot be adequately tested for in a data-driven way, as far as we know, especially given the difficulty of quantifying process versus measurement noise.

Properties and failure modes	Granger causality	State space reconstruction
Does it detect indirect perturbation causality?	No (Fig 5A)	Yes
Is it (Granger causality or SSR) Transitive?	No (Fig 5A)	Yes
Process noise?	Requires moderate process noise	Assumes no (or low) process noise
"Nonstationarity" failure mode	Nonstationarity (Fig 5Bii)	Nonreverting continuous dynamics (Fig 7)

Table 1: Comparison of Granger causality and state space reconstruction approaches to causal inference from time series data. Note that indirectness and transitivity (top two rows) are distinct: perturbation causality can be both nontransitive (Fig1B) and indirect (Fig1C).

From this synthesis, the outlook for mechanism-free causal inference methods appears rather discouraging. One way forward might be to quantify causal signals by scoring possible causal links (e.g. using a Granger causality p-value or cross map skill) and then look for the strongest (or most likely) interactions [72, 85], instead of assuming that 95% of significant inferred interactions are "real". Alternatively, one might relax the directed nature of causal inference, and instead draw an analogy to Reichenbach's principle (e.g. if X Granger or SSR causes Y, then X causes Y, or Y causes X, or Z causes both) [86]. Thus strong causal signals, while clearly imperfect, may provide evidence to be used along with other lines of evidence.

The move toward causal inference from observational time series is a response to practical barriers to performing manipulative experiments. Given the difficulties associated with the methods we have discussed, an attractive alternative way to deal with the precious nature of manipulative experiments would be to get away with as few perturbations as possible. Thus systematic ways to determine which perturbations are most informative in revealing causality [87, 88] constitute an immensely valuable research direction.

Acknowledgment

Ideally, writers of this essay should be Renaissance scholars with mastery of divergent fields. However, human knowledge has expanded drastically since the Renaissance and we are not modern da Vincis. We therefore sought feedback and advice from domain experts: Bree Cummins (Montana State University) and Tim Sauer (George Mason University) discussed topology with us; Sean Gibbons (Institute for Systems Biology) and Nathan Kutz (University of Washington) critiqued our manuscript. We thank David Fredricks and Sujatha Srinivasan (Fred Hutch) for discussions that inspired this effort, and members of the Shou group for helpful comments.

Methods

Methodological details for Fig 3

For panel B, we simulated the random walk system

$$X(t+1) = X(t) + \epsilon(t)$$

where $\epsilon(t)$ terms were drawn independently from a normal distribution with mean of 0 and standard deviation of 1. We simulated this system from the initial condition of $X(1) = 0$ through 999 subsequent steps. For panel C, we simulated the autoregressive system

$$X(t+1) = 0.75X(t) + 10 + \epsilon(t)$$

where $\epsilon(t)$ terms were again drawn independently from a normal distribution with mean of 0 and standard deviation of 1. We simulated this system from the initial condition of $X(1) = 40$ for 1999 subsequent steps. We only used the final 1000 steps for computing the correlation between two time series.

To compute the significance of the Pearson correlation between two time series, we used a permutation test. For a pair of time series $[X_1(1), X_1(2), ..., X_1(1000)], [X_2(1), X_2(2), ..., X_2(1000)]$, we first computed the Pearson correlation $\hat{\rho}$ between the two time series. We then shuffled the X_2 values and recomputed the Pearson correlation as $\tilde{\rho}$. We computed this shuffled correlation 10^4 times to get a null distribution $[\tilde{\rho}_1, \tilde{\rho}_2, ..., \tilde{\rho}_{10^4}]$. We computed the p value as the fraction of shuffled correlations $\tilde{\rho}$ whose magnitude was greater than or equal to the magnitude of the original correlation $\hat{\rho}$.

Methodological details for Fig 4

The original subpopulation distributions are normal distributions with standard deviation of 10 and mean of 100 (male) or 130 (female). Each plot shows 300 samples randomly drawn from the appropriate mixture distribution.

Methodological details for Fig 6

The system of equations was numerically integrated using the ode45 method in matlab from $t = 0$ to $t = 200$ in time steps of 0.03, and plotted in the delay space Z with $\tau = 3.6$. The initial condition for all state variables (V, W, X, Y, Z, $\frac{dX}{dt}$, $\frac{dY}{dt}$, and $\frac{dV}{dt}$) was 1. For panel F, measurement noise was added to $Y(t)$. Specifically, noisy data were generated as:

$$Y^{obs}(t) \sim \text{Unif}\left(Y(t) - 3^{1/2}\left(0.15\Delta_Y\right), Y(t) + 3^{1/2}\left(0.15\Delta_Y\right)\right)$$

where $\text{Unif}(a, b)$ is a uniform random variable bounded by a and b, and Δ_Y is the difference between the maximum and minimum values of $Y(t)$ between $t = 0$ and $t = 200$. These noise parameters are chosen so that $Y^{obs}(t)$ is centered at $Y(t)$ and has a standard deviation of $0.15\Delta_Y$.

Methodological details for Fig 7

The dynamics in the top row of Fig 7 were generated from the equations:

$$X(t) = \sin(t) + 0.5t$$
$$Z(t) = 0.1(t - 10)^2$$

This continuous-time system was discretized from $t = 1$ to $t = 20$ on an evenly spaced grid of 400 data points for visualizing delay spaces where the time delay is $\tau = 50$.

The dynamics in the second row of Fig 7 were generated from the equations:

$$X(t + 1) = X(t)(3.61 - 3.61X(t))$$
$$Z(t + 1) = Z(t)(3.61 - 3.61X(t))$$

with initial conditions of $X(1) = 0.4$ and $Z(1) = 0.7$. For this system, $\tau = 1$ and 2000 time points were used to make the plots of delay spaces.

The dynamics in the third row of Fig 7 were generated from the equations:

$$\frac{dX^2}{dt} = -X(t)$$
$$\frac{dZ^2}{dt} = -25Z(t)$$

with initial conditions of $X(1) = X'(1) = Z(1) = Z'(1) = 1$. For this system, $\tau = 0.9$ was used for delay spaces. This continuous-time system was numerically integrated using the ode45 method in Matlab from $t = 0$ to $t = 13.998$ on a grid of 4667 evenly-spaced time points for plotting dynamics, and time points $t = 0.003$ through $t = 7.698$ were used for visualizing delay spaces.

The dynamics in the bottom row of Fig 7 were generated from the classic Lorenz attractor equations:

$$\frac{dX}{dt} = -10X(t) + 10Y(t)$$
$$\frac{dY}{dt} = 28X(t) - Y(t) - X(t)Z(t)$$
$$\frac{dZ}{dt} = -\frac{8}{3}Z(t) + X(t)Y(t)$$

with initial conditions of $X(0) = Y(0) = Z(0) = 1$. A delay of $\tau = 0.14$ was used to make delay spaces. This continuous-time system was numerically integrated using the ode45 method in Matlab from $t = 0$ to $t = 399.98$ on an evenly spaced grid of 5715 data points for visualizing delay spaces.

Methodological details for Fig 9

Ground truth model and data generation

We used the ground truth model:

$$S_1(t+1) = \max\left(0, S_1(t)\left(1.1 - 0.05S_1(t) + D_1(t)\right) + \epsilon_{p1}(t)\right)$$
$$S_2(t+1) = \max\left(0, S_2(t)\left(1.05 - 0.1S_2(t) + D_2(t) + 0.3S_1(t) + 2.5\epsilon_{p2}(t)\right)\right)$$

$S_1(t)$ and $S_2(t)$ represent the population sizes of species 1 and 2 at time t. $D_1(t)$ and $D_2(t)$ are the values of periodic drivers at time t. Specifically, in both the two-driver and one-driver cases:

$$D_1(t) = 0.05\sin\left(\frac{5t}{6} + \phi_1\right) + 0.05\sin\left(t + \phi_1\right)$$

In the two-driver case:

$$D_2(t) = 0.1\sin\left(\frac{t}{\sqrt{10}} + \phi_2\right)$$

Conversely, in the one-driver case $D_2(t) = 0$. The process noise terms $\epsilon_{p1}(t)$ and $\epsilon_{p2}(t)$ are both IID normal random variables with mean of 0 and with shared standard deviation σ_p. Specifically, for any pair of times $t_1 \neq t_2$, $\epsilon_{p1}(t_1)$ and $\epsilon_{p1}(t_2)$ are independent, and similarly for ϵ_{p2}. Also, all values $\epsilon_{p1}(1), \epsilon_{p1}(2), \cdots$ are independent of all values $\epsilon_{p2}(1), \epsilon_{p2}(2), \cdots$. At the beginning of each replicate simulation, the phases ϕ_1 and ϕ_2 are independently assigned a random number from a uniform distribution between 0 and 2π, and do not change with time.

To generate data without measurement noise, we simulated this system for $t = 1, 2, ..., 400$ with the initial conditions $S_1(1) = 2$; $S_2(1) = 4.5$. We used the final 200 time points for inference to help ensure that the system had reached equilibrium behavior.

We also introduced additive measurement noise to simulate instrument uncertainty:

$$S_1^{obs}(t) = S_1(t) + \epsilon_{m1}(t)/1.5$$
$$S_2^{obs}(t) = S_2(t) + \epsilon_{m2}(t)$$

where S_1^{obs} and S_2^{obs} represent the observed values (i.e. noisy measurments) of S_1 and S_2. $\epsilon_{m1}(t)$ and $\epsilon_{m2}(t)$ are also IID normal random variables with mean of 0 and standard deviation σ_m. The tables in Fig 9D are generated by varying σ_p from 0 to 8 and varying σ_m from 0 to 1.

Causal inference using Granger causality and CCM

For each combination of σ_m and σ_p (the standard deviation of measurement noise and process noise, respectively), we generated 1000 time series for S_1 and S_2 as described above. For each replicate pair of time series, we used Granger causality and CCM to infer whether S_1 causes S_2 (it does) and whether S_2 causes S_1 (it does not).

Granger causality inference

We used the multivariate Granger causality Matlab package (MVGC, [53]). We used the following settings:

- regmode = 'OLS' (We fit the autoregressive model by the ordinary least squares method).
- icregmode = 'LWR' (We determined the information criterion using the LWR algorithm. This is the default setting).
- morder = 'AIC' (We used Akaike information criterion to determine the number of lags in the autoregressive model).
- momax = 50 (We used a maximum of 50 lags in the autoregressive model).
- tstat = '' (We used Granger's F-test for statistical significance. This is the default setting).

We inferred the presence of a causal link if the p-value was less than or equal to 0.05. We inferred no causal link otherwise. When σ_m and σ_p were both 0, the MVGC package (correctly) exited with an error on most trials. We reported this as "unsuitable data" in Fig 9D.

When σ_m and σ_p are both 0,the inferred spectral radius of the stochastic process is close to 1, the MVGC routines can be prohibitively slow (i.e. when running 1000 trials, the program would hang at an early stage for hours). In this case, the authors note that switching from the package's default single-regression mode to an alternative dual-regression mode may improve runtime [53]. We thus switched to the dual-regression mode when the spectral radius was between 0.9999 and 1 (a spectral radius of 1 or more causes an error). This fix had no effect on benchmark results as long as at least one of σ_m and σ_p was not 0.

Convergent cross mapping

Convergent cross mapping looks for a delay map from X to Y. That is, CCM looks for a map from $[X(t), X(t-\tau), X(t-2\tau), ..., X(t-(E-1)\tau)]$ to $Y(t)$. Thus in order to apply CCM one needs to choose the delay τ and the vector length (dimension of the delay space) E. E and τ should ideally be "generic" in the sense of Takens's theorem: we want to avoid line-crossing (such as the symbol "∞") in the delay space, because otherwise, Φ^{-1} in Fig 17 does not exist. There are different ways to do this, but no method is obviously best ([19, 31]). Following [19] and [60] we chose τ and E to maximize univariate one-step-ahead forecast of the putative causee X. That is, for $X(n)$, we try to predict $X(n+1)$ using the simplex projection method by finding delay vectors in the training data of X that are most similar to $[X(n), X(n-\tau), X(n-2\tau), ..., X(n-(E-1)\tau)]$, and take weighed average of their X values 1 step in the future (i.e. Fig 8A where $X = Y$ and the prediction lag is 1). If the delay space has a line crossing, then at the cross-point, a one-step-ahead forecast may have more than one possible outcome and thus perform poorly. In more detail, we made one-step-ahead forecasts within the time range 201-400 (we did not use time range 1-200 to avoid transient dynamics). As per the field standard, we used leave-one-out cross-validation to do simplex projection. That is, when making a forecast for a time t, we used all times within 201-400 other than t as training data. We performed a grid search, varying τ from 1 to 6 and varying E from 1 to 6. We then used the combination of τ and E that maximized the forecast skill (the Pearson correlation between forecasts and true values) for subsequent CCM analysis. Additionally, following [60], if the optimal combination of τ and E failed to give a significant forecast skill, we did not report CCM results for that trial and reported the trial as "unsuitable data". To test for significance of univariate one-step-ahead forecast skill, we used a naive bootstrap approach. We computed the forecast skill with a training library composed of randomly chosen delay vectors (sampling with replacement: some vectors may not be sampled while other may be sample more than once) from the original training data using the 'random_libs' setting in the rEDM (version 1.5) ccm method. The training library size (the number of delay vectors in the library) was chosen to be the total number of training vectors ($= 200 - 1 - (E-1)\tau$, where 200 is the number of training data points and the -1 comes from one-step-ahead forecasting). We calculated the forecast skill with 1000 such libraries and considered the forecast skill "significant" if at least least 95% of the 1000 libraries gave a forecast skill greater than 0. Basically, this procedure asks whether a positive forecast skill is robust to small changes in the dataset, and rejects a forecast skill that is not robust as insignificant.

Having chosen τ and E, we used three CCM criteria to infer causality (criteria 1-3 in Fig 8). We did not use the fourth criterion (the prediction lag test) since its interpretation is unclear for periodic systems (Fig 19). For all three criteria, we used the same cross-validation setting that we used to choose τ and E. The first CCM criterion is that cross map skill is greater than 0. To test for this criterion we again used a naive bootstrap approach similar to our above test for univariate forecast skill. We computed the cross map skill with a training library composed of randomly chosen delay vectors sampled with replacement from the original training data time points. The training library size (the number of delay vectors available for the library) was again chosen to be the total number of distinct delay vectors ($200 - (E-1)\tau$). We calculated the cross map skill with 1000 such libraries and considered the cross map skill "significant" if at least least 95% of the 1000 libraries gave a cross map skill greater than 0.

The second CCM criterion is that the cross map skill from causee to causer with real data must be greater than the cross map skill when the putative causer is replaced with surrogate data. To test this criterion, we first computed cross map skill using the same training and testing time points as before to obtain a single cross map skill value. We then repeatedly (1000 times) computed cross map skill in the same way, but now

with the putative causer time series replaced with random phase surrogate data. Random phase surrogate data were generated by Ebisuzaki's method as implemented in the rEDM function make_surrogate_data. We computed the significance p-value as the fraction of cross map skill values obtained with surrogate time series that were equal to or greater than the cross map skill value obtained with the original putative causer time series. A putative causal link passed this criterion if the p-value was less than or equal to 0.05.

The third CCM criterion is that cross map skill increases with more training data. Following [19], we again used a naive bootstrap approach to test for this criterion. Specifically, we computed the cross map skill with a training library composed of randomly chosen delay vectors sampled with replacement from the original training data time points. We used either a large library with $200 - (E-1)\tau$ available training vectors as used previously, or a small library with 15 training vectors. For each of 1000 bootstrap trials, we compared the cross map skill from a randomly chosen small library to the cross map skill from a randomly chosen large library. We said that the cross map skill increased with training data if the cross map skill of the large library was greater than that of the small library in at least 95% of the 1000 bootstrap trials.

For "alternative" CCM testing, we only changed how the third CCM criterion (cross map skill increases with more training data) were tested. Here, instead of using the bootstrap test of [19], we tested the third CCM criterion using Kendall's τ test as suggested in [81]. To do this, we varied the library size from a minimum of 15 vectors to the the maximum library size $(200 - (E-1)\tau - 1)$, in increments of 3 vectors. For each library size, we computed cross map skill using 50 libraries randomly sampled without replacement (e.g. the 50 libraries would be identical at the maximal library size). We then computed the median cross map skill for each library size. Finally we ran a 1-tailed Kendall's τ test for a positive association between library size and median cross map skill. We used the function stats.kendalltau from the Python package SciPy to compute a 2-tailed p-value, and then divided this p-value by 2 to get a 1-tailed p-value. We said that cross map skill increased with training data if the τ statistic was positive and the 1-tailed p-value was ≤ 0.05.

Methodological details for Fig 10

The data-generating process and permutation test are equivalent to those in Fig 3C. Random phase surrogate data were generated by Ebisuzaki's random phase method [84] as implemented in the rEDM (version 1.5) function make_surrogate_data.

Methodological details for Fig 17

To generate data for panels C-J, the system of panel A was numerically integrated using the ode45 method in matlab with a time step of 0.005 and with the intial condition that X, Y, Z, $\frac{dX}{dt}$, $\frac{dY}{dt}$ were all set to 1 at $t = 0$. Panels C , E, F, G, I, and J (non-inset) show data from a single period. For panel H the system was integrated for about 5 periods to more clearly visualize the lack of a continuous delay map. For the panel J inset the system was integrated for about 12 periods to better see the separated legs of the curve upon zooming in. Panels C, D, F, H, and J were colored $\mathrm{mod}(t, T)$. That is, they were colored by the remainder of t (time) after dividing by T (here $T = 2\pi$). $\tau = 3.6$ was used for all delay spaces.

Methodological details for Fig 19

Top row: For this system, we used the initial conditions $X(1) = 0.2, Y(1) = 0.4$ and composed delay vectors of length $E = 2$ with a delay of $\tau = 2$. We visualized the delay space using data from time points 501-2000. We used points 801-1000 for training data and points 1001-2000 for testing cross map predictions.

Second row: For this system, we used the initial conditions $X(1) = 0.4, Y(1) = 0.2$ and the delay vector parameters $(E = 2, \tau = 1)$. We visualized the delay space using data from time points 501-2000. We used points 801-1000 for training data and points 1001-2000 for testing cross map predictions.

Third row: For this system, we used the initial conditions $X(1) = 0.2, Y(1) = 0.4$ and the delay vector parameters $(E = 3, \tau = 2)$. We visualized the delay space using data from time points 501-2000 (time points 1-6 $\times 10^5$ for the zoomed-in inset). We used points 801-1000 for training data and points 1001-2000 for testing cross map predictions.

Fourth row: For this system, we used the initial conditions $W(0) = Y(0) = 0$ and $X(0) = \dot{W}(0) = \dot{Y}(0) = 1$. We numerically integrated this system using ode45 in Matlab with a time step of 0.1. We visualized the delay space using data from $t = 50.1$ through $t = 200$ (time indices 501-2000). We used the delay vector parameters ($E = 3, \tau = 7.2$). We used data from $t = 70.1$ through $t = 100$ (time indices 701-1000) for training data and data from $t = 100.1$ through $t = 200$ (time indices 1001-2000) for testing cross map predictions.

Fifth row: For this system, we used the initial conditions $X(1) = 0.2, Y(1) = 0$ and composed delay vectors of length $E = 2$ with a delay of $\tau = 2$. We visualized the delay space using data from time points 501-2000. We used points 801-1000 for training data and points 1001-2000 for testing cross map predictions.

Sixth row: For this system, the "initial" conditions specified the first 3 timepoints since we included a lag of 3. Thus, for $k = 1, 2, 3$, $W(k) = 0.2$, $X(k) = 0.4$, and $Y(k) = 0.3$. We composed delay vectors of length $E = 3$ with a delay of $\tau = 1$. We visualized the delay space using data from time points 501-2000. We used points 801-1000 for training data and points 1001-2000 for testing cross map predictions.

For convergent cross mapping, we used the same τ and E as for visualizing delay spaces. The training data size is the number of delay vectors in the training library. For the plots in the fourth column, we chose 300 random libraries of training delay vectors with variable training data size, and used the standard prediction lag of 0. Delay vectors were chosen without replacement. Note that at large training data size, some or all of the 300 random libraries can be identical. Each dot in these CCM plots represents the average forecast skill over all 300 randomly-chosen libraries. Error bars represent the 95% confidence interval as calculated by the bias-corrected and accelerated bootstrap (1000 bootstraps) as implemented in Matlab's bootci function. Error bars are the same color as the dots and so are not visible when they fit inside the dots.

In all rows, the cross map skill for the putative causer Y was greater than for at least 95% of random phase surrogate time series (purple dot). The 5% cutoff value was computed for the maximum library size (156 for row 4 and ~ 200 for all other rows) by running the CCM procedure after replacing the putative causer Y with 500 random phase surrogate time series generated using the rEDM function make_surrogate_data.

For the plots in the fifth column we used the full library contained within the training data window (156 delay vectors for row 4 and ~ 200 for all other rows) and varied the prediction lag. That is, we did not use random libraries for these plots.

Methodological details for Fig 20

To generate randomized parameter sets, we randomly selected R_X, R_Y, A_{XY} and A_{YX} from uniform distributions. We also randomly selected the initial conditions $X(1)$ and $Y(1)$ from uniform distributions. To make systems in the "friendly" parameter regime, we drew R_X and R_Y independently from the range $3.7 - 3.9$, we drew A_{XY} and A_{YX} independently from the range $0.05 - 0.1$, and we drew $X(1)$ and $Y(1)$ independently from the range $0.01 - 0.99$. These are the same parameters used in the randomized numerical simulations of [74]. Next, to make systems in the "pathological" parameter regime, we drew R_X from the range $3.7 - 3.9$, we drew R_Y from the range $3.1 - 3.3$, we drew A_{XY} and A_{YX} independently from the range $0.15 - 0.2$, and we drew $X(1)$ and $Y(1)$ independently from the range $0.01 - 0.99$. For both parameter regimes we randomly chose 500 sets of parameters and ran the system for 3000 time points. Occasionally a randomly chosen system would leave the basin of attraction and reach large values, represented on the computer as positive or negative infinity, or "not a number". When this occurred, we discarded the data and resampled parameters.

To apply CCM on each system, we generated a training library of delay vectors of X by randomly selecting 200 vectors from among time points 100-2000. We then evaluated cross map skill from delay vectors of X to values of Y at points 2001-3000. Following [74], we used delay vectors of length $E = 2$ and a delay duration of $\tau = 1$. We evaluated cross map skill with a prediction horizon of -8 through 8.

References

[1] S. Widder, R. J. Allen, T. Pfeiffer, T. P. Curtis, C. Wiuf, W. T. Sloan, O. X. Cordero, S. P. Brown, B. Momeni, W. Shou, *et al.*, "Challenges in microbial ecology: building predictive understanding of community function and dynamics," *The ISME journal*, vol. 10, no. 11, pp. 2557–2568, 2016.

[2] S. R. Lindemann, H. C. Bernstein, H.-S. Song, J. K. Fredrickson, M. W. Fields, W. Shou, D. R. Johnson, and A. S. Beliaev, "Engineering microbial consortia for controllable outputs," *The ISME journal*, vol. 10, no. 9, pp. 2077–2084, 2016.

[3] R. R. Stein, V. Bucci, N. C. Toussaint, C. G. Buffie, G. Ratsch, E. G. Pamer, C. Sander, and J. B. Xavier, "Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota," *PLoS computational biology*, vol. 9, no. 12, p. e1003388, 2013.

[4] C. K. Fisher and P. Mehta, "Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression," *PLOS ONE*, vol. 9, pp. 1–10, 07 2014.

[5] V. Bucci, B. Tzen, N. Li, M. Simmons, T. Tanoue, E. Bogart, L. Deng, V. Yeliseyev, M. L. Delaney, Q. Liu, B. Olle, R. R. Stein, K. Honda, L. Bry, and G. K. Gerber, "Mdsine: Microbial dynamical systems inference engine for microbiome time-series analyses," *Genome Biology*, vol. 17, p. 121, Jun 2016.

[6] H.-T. Cao, T. E. Gibson, A. Bashan, and Y.-Y. Liu, "Inferring human microbial dynamics from temporal metagenomics data: Pitfalls and lessons," *BioEssays*, vol. 39, no. 2, 2017.

[7] S. H. Levine, "Competitive interactions in ecosystems," *The American Naturalist*, vol. 110, no. 976, pp. 903–910, 1976.

[8] J. T. Wootton, "Indirect effects in complex ecosystems: recent progress and future challenges," *Journal of Sea Research*, vol. 48, no. 2, pp. 157–172, 2002.

[9] B. Momeni, L. Xie, and W. Shou, "Lotka-volterra pairwise modeling fails to capture diverse pairwise microbial interactions," *Elife*, vol. 6, p. e25051, 2017.

[10] A. R. Coenen, S. K. Hu, E. Luo, D. Muratore, and J. S. Weitz, "A primer for microbiome time-series analysis," *Frontiers in Genetics*, vol. 11, 2020.

[11] A. Tkacz, M. Hortala, and P. S. Poole, "Absolute quantitation of microbiota abundance in environmental samples," *Microbiome*, vol. 6, no. 1, p. 110, 2018.

[12] T. S. Schmidt, J. Raes, and P. Bork, "The human gut microbiome: from association to modulation," *Cell*, vol. 172, no. 6, pp. 1198–1215, 2018.

[13] L. Qian, H. Song, and W. Cai, "Determination of bifidobacterium and lactobacillus in breast milk of healthy women by digital pcr," *Beneficial microbes*, vol. 7, no. 4, pp. 559–569, 2016.

[14] F. Bonk, D. Popp, H. Harms, and F. Centler, "Pcr-based quantification of taxa-specific abundances in microbial communities: quantifying and avoiding common pitfalls," *Journal of microbiological methods*, vol. 153, pp. 139–147, 2018.

[15] C. W. Granger, "Testing for causality: a personal viewpoint," *Journal of Economic Dynamics and control*, vol. 2, pp. 329–352, 1980.

[16] L. Barnett, A. B. Barrett, and A. K. Seth, "Misunderstandings regarding the application of granger causality in neuroscience," *Proceedings of the National Academy of Sciences*, vol. 115, no. 29, pp. E6676–E6677, 2018.

[17] E. B. Baskerville and S. Cobey, "Does influenza drive absolute humidity?," *Proceedings of the National Academy of Sciences*, vol. 114, no. 12, pp. E2270–E2271, 2017.

[18] H. Alrasheed, R. Jin, and J. S. Weitz, "Caution in inferring viral strategies from abundance correlations in marine metagenomes," *Nature communications*, vol. 10, no. 1, p. 501, 2019.

[19] S. Cobey and E. B. Baskerville, "Limits to causal inference with state-space reconstruction for infectious disease," *PloS one*, vol. 11, no. 12, p. e0169050, 2016.

[20] B. Lusch, P. D. Maia, and J. N. Kutz, "Inferring connectivity in networked dynamical systems: Challenges using granger causality," *Physical Review E*, vol. 94, no. 3, p. 032220, 2016.

[21] D. Mønster, R. Fusaroli, K. Tylén, A. Roepstorff, and J. F. Sherson, "Causal inference from noisy time-series data-testing the convergent cross-mapping algorithm in the presence of noise and external influence," *Future Generation Computer Systems*, vol. 73, pp. 52–62, 2017.

[22] A. R. Coenen and J. S. Weitz, "Limitations of correlation-based inference in complex virus-microbe communities," *mSystems*, vol. 3, no. 4, pp. e00084–18, 2018.

[23] A. Krakovská, J. Jakubík, M. Chvosteková, D. Coufal, N. Jajcay, and M. Paluš, "Comparison of six methods for the detection of causality in a bivariate time series," *Physical Review E*, vol. 97, no. 4, p. 042207, 2018.

[24] F. Barraquand, C. Picoche, M. Detto, and F. Hartig, "Inferring species interactions using granger causality and convergent cross mapping," *arXiv preprint arXiv:1909.00731*, 2019.

[25] D. Vandeputte, G. Kathagen, K. D'hoe, S. Vieira-Silva, M. Valles-Colomer, J. Sabino, J. Wang, R. Y. Tito, L. De Commer, Y. Darzi, *et al.*, "Quantitative microbiome profiling links gut community variation to microbial load," *Nature*, vol. 551, no. 7681, p. 507, 2017.

[26] S. M. Gibbons, S. M. Kearney, C. S. Smillie, and E. J. Alm, "Two dynamic regimes in the human gut microbiome," *PLoS computational biology*, vol. 13, no. 2, p. e1005364, 2017.

[27] E. Margolis, A. Oot, and D. N. Fredricks, "Human Microbiome Dynamics: Causality Detection With Convergent Cross Mapping," *Open Forum Infectious Diseases*, vol. 3, 10 2016. 2224.

[28] K. Mainali, S. Bewick, B. Vecchio-Pagan, D. Karig, and W. F. Fagan, "Detecting interaction networks in the human microbiome with conditional granger causality," *PLoS computational biology*, vol. 15, no. 5, p. e1007037, 2019.

[29] C. W. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica: Journal of the Econometric Society*, pp. 424–438, 1969.

[30] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

[31] D. Harnack, E. Laminski, M. Schünemann, and K. R. Pawelzik, "Topological causality in dynamical systems," *Physical review letters*, vol. 119, no. 9, p. 098301, 2017.

[32] J. Pearl, *Causality*. Cambridge university press, 2009.

[33] J. Woodward, "Causation and manipulability," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, winter 2016 ed., 2016.

[34] D. M. Hausman and J. Woodward, "Independence, invariance and the causal markov condition," *The British journal for the philosophy of science*, vol. 50, no. 4, pp. 521–583, 1999.

[35] T. Pollet, L. Berdjeb, C. Garnier, G. Durrieu, C. Le Poupon, B. Misson, and J.-F. Briand, "Prokaryotic community successions and interactions in marine biofilms: the key role of flavobacteriia," *FEMS microbiology ecology*, vol. 94, no. 6, p. fiy083, 2018.

[36] C. Li, K. M. K. Lim, K. R. Chng, and N. Nagarajan, "Predicting microbial interactions through computational approaches," *Methods*, vol. 102, pp. 12–19, 2016.

[37] F. Ju and T. Zhang, "Bacterial assembly and temporal dynamics in activated sludge of a full-scale municipal wastewater treatment plant," *The ISME journal*, vol. 9, no. 3, pp. 683–695, 2015.

[38] A. Carr, C. Diener, N. S. Baliga, and S. M. Gibbons, "Use and abuse of correlation analyses in microbial ecology," *The ISME journal*, p. 1, 2019.

[39] C. Granger and P. Newbold, "Spurious regressions in econometrics," *Journal of Econometrics*, vol. 2, no. 2, pp. 111–120, 1974.

[40] C. W. Granger IV, N. Hyung, and Y. Jeon, "Spurious regressions with stationary series," *Applied Economics*, vol. 33, no. 7, pp. 899–904, 2001.

[41] C. Hitchcock and M. Rédei, "Reichenbach's common cause principle," in *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.), Metaphysics Research Lab, Stanford University, spring 2020 ed., 2020.

[42] Q. Ruan, D. Dutta, M. S. Schwalbach, J. A. Steele, J. A. Fuhrman, and F. Sun, "Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors," *Bioinformatics*, vol. 22, no. 20, pp. 2532–2538, 2006.

[43] F. Zhang, A. Shan, and Y. Luan, "A novel method to accurately calculate statistical significance of local similarity analysis for high-throughput time series," *Statistical applications in genetics and molecular biology*, vol. 17, no. 6, 2018.

[44] F. Zhang, F. Sun, and Y. Luan, "Statistical significance approximation for local similarity analysis of dependent time series data," *BMC bioinformatics*, vol. 20, no. 1, p. 53, 2019.

[45] L. Barnett, A. B. Barrett, and A. K. Seth, "Solved problems for granger causality in neuroscience: A response to stokes and purdon," *NeuroImage*, vol. 178, pp. 744–748, 2018.

[46] P. A. Stokes and P. L. Purdon, "Reply to barnett et al.: Regarding interpretation of granger causality analyses," *Proceedings of the National Academy of Sciences*, vol. 115, no. 29, pp. E6678–E6679, 2018.

[47] D. Ai, X. Li, G. Liu, X. Liang, and L. C. Xia, "Constructing the microbial association network from large-scale time series data using granger causality," *Genes*, vol. 10, no. 3, p. 216, 2019.

[48] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, "Transfer entropy—a model-free measure of effective connectivity for the neurosciences," *Journal of computational neuroscience*, vol. 30, no. 1, pp. 45–67, 2011.

[49] F. Roux, M. Wibral, W. Singer, J. Aru, and P. J. Uhlhaas, "The phase of thalamic alpha activity modulates cortical gamma-band activity: evidence from resting-state meg recordings," *Journal of Neuroscience*, vol. 33, no. 45, pp. 17827–17835, 2013.

[50] C. Diks and V. Panchenko, "A new statistic and practical guidelines for nonparametric granger causality testing," *Journal of Economic Dynamics and Control*, vol. 30, no. 9-10, pp. 1647–1669, 2006.

[51] S. D. Bekiros and C. G. Diks, "The nonlinear dynamic relationship of exchange rates: Parametric and nonparametric causality testing," *Journal of macroeconomics*, vol. 30, no. 4, pp. 1641–1650, 2008.

[52] A. Papana, C. Kyrtsou, D. Kugiumtzis, and C. Diks, "Assessment of resampling methods for causality testing: A note on the us inflation behavior," *PloS one*, vol. 12, no. 7, p. e0180852, 2017.

[53] L. Barnett and A. K. Seth, "The mvgc multivariate granger causality toolbox: a new approach to granger-causal inference," *Journal of neuroscience methods*, vol. 223, pp. 50–68, 2014.

[54] J. Geweke, "Measurement of linear dependence and feedback between multiple time series," *Journal of the American statistical association*, vol. 77, no. 378, pp. 304–313, 1982.

[55] L. E. Ohanian, "The spurious effects of unit roots on vector autoregressions: A monte carlo study," *Journal of Econometrics*, vol. 39, no. 3, pp. 251–266, 1988.

[56] H. Y. Toda and P. C. Phillips, "The spurious effect of unit roots on vector autoregressions: an analytical study," *Journal of Econometrics*, vol. 59, no. 3, pp. 229–255, 1993.

[57] Z. He and K. Maekawa, "On spurious granger causality," *Economics Letters*, vol. 73, no. 3, pp. 307–313, 2001.

[58] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical systems and turbulence, Warwick 1980*, pp. 366–381, Springer, 1981.

[59] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," *Journal of statistical physics*, vol. 65, no. 3-4, pp. 579–616, 1991.

[60] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch, "Detecting causality in complex ecosystems," *Science*, vol. 338, no. 6106, pp. 496–500, 2012.

[61] 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 biology*, vol. 15, no. 7, p. R89, 2014.

[62] S. E. Said and D. A. Dickey, "Testing for unit roots in autoregressive-moving average models of unknown order," *Biometrika*, vol. 71, no. 3, pp. 599–607, 1984.

[63] L. E. Ohanian, "A note on spurious inference in a linearly detrended vector autoregression," *The Review of Economics and Statistics*, pp. 568–571, 1991.

[64] S. Seabold and J. Perktold, "statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

[65] P. A. Stokes and P. L. Purdon, "A study of problems encountered in granger causality analysis from a neuroscience perspective," *Proceedings of the National Academy of Sciences*, vol. 114, no. 34, pp. E7063–E7072, 2017.

[66] E. L. Feige and D. K. Pearce, "The casual causal relationship between money and income: Some caveats for time series analysis," *The Review of Economics and Statistics*, pp. 521–533, 1979.

[67] S. Behrendt, T. Dimpfl, F. J. Peter, and D. J. Zimmermann, "Rtransferentropy—quantifying information flow between different time series using effective transfer entropy," *SoftwareX*, vol. 10, p. 100265, 2019.

[68] P. Wollstadt, J. Lizier, R. Vicente, C. Finn, M. Martinez-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," *Journal of Open Source Software*, vol. 4, no. 34, p. 1081, 2019.

[69] G. K. Gerber, "The dynamic microbiome," *FEBS letters*, vol. 588, no. 22, pp. 4131–4139, 2014.

[70] K. Faust, L. Lahti, D. Gonze, W. M. De Vos, and J. Raes, "Metagenomics meets time series analysis: unraveling microbial community dynamics," *Current opinion in microbiology*, vol. 25, pp. 56–66, 2015.

[71] J. S. Biteen, P. C. Blainey, Z. G. Cardon, M. Chun, G. M. Church, P. C. Dorrestein, S. E. Fraser, J. A. Gilbert, J. K. Jansson, R. Knight, *et al.*, "Tools for the microbiome: nano and beyond," 2016.

[72] H. Ma, K. Aihara, and L. Chen, "Detecting causality from nonlinear dynamics with short-term time series," *Scientific reports*, vol. 4, p. 7464, 2014.

[73] B. Cummins, T. Gedeon, and K. Spendlove, "On the efficacy of state space reconstruction methods in determining causality," *SIAM Journal on Applied Dynamical Systems*, vol. 14, no. 1, pp. 335–381, 2015.

[74] H. Ye, E. R. Deyle, L. J. Gilarranz, and G. Sugihara, "Distinguishing time-delayed causal interactions using convergent cross mapping," *Scientific reports*, vol. 5, p. 14750, 2015.

[75] E. R. Deyle and G. Sugihara, "Generalized theorems for nonlinear state space reconstruction," *PLOS ONE*, vol. 6, pp. 1–8, 03 2011.

[76] E. Brookshire and T. Weaver, "Long-term decline in grassland productivity driven by increasing dryness," *Nature communications*, vol. 6, no. 1, pp. 1–7, 2015.

[77] K. L. Cramer, A. O'Dea, T. R. Clark, J.-x. Zhao, and R. D. Norris, "Prehistorical and historical declines in caribbean coral reef accretion rates driven by loss of parrotfish," *Nature communications*, vol. 8, no. 1, pp. 1–8, 2017.

[78] Y. Wang, J. Yang, Y. Chen, P. De Maeyer, Z. Li, and W. Duan, "Detecting the causal effect of soil moisture on precipitation using convergent cross mapping," *Scientific reports*, vol. 8, no. 1, pp. 1–8, 2018.

[79] P. Newbold, "Feedback induced by measurement errors," *International Economic Review*, pp. 787–791, 1978.

[80] H. Nalatore, M. Ding, and G. Rangarajan, "Mitigating the effects of measurement noise on granger causality," *Physical Review E*, vol. 75, no. 3, p. 031123, 2007.

[81] C.-W. Chang, M. Ushio, and C.-h. Hsieh, "Empirical dynamic modeling for beginners," *Ecological Research*, vol. 32, no. 6, pp. 785–796, 2017.

[82] L. C. Xia, J. A. Steele, J. A. Cram, Z. G. Cardon, S. L. Simmons, J. J. Vallino, J. A. Fuhrman, and F. Sun, "Extended local similarity analysis (elsa) of microbial community and other time series data with replicates," in *BMC systems biology*, vol. 5, p. S15, BioMed Central, 2011.

[83] D. N. Politis and J. P. Romano, "The stationary bootstrap," *Journal of the American Statistical association*, vol. 89, no. 428, pp. 1303–1313, 1994.

[84] W. Ebisuzaki, "A method to estimate the statistical significance of a correlation when the data are serially correlated," *Journal of Climate*, vol. 10, no. 9, pp. 2147–2153, 1997.

[85] D. Janzing, D. Balduzzi, M. Grosse-Wentrup, B. Schölkopf, *et al.*, "Quantifying causal influences," *The Annals of Statistics*, vol. 41, no. 5, pp. 2324–2358, 2013.

[86] G. Sugihara, E. R. Deyle, and H. Ye, "Reply to baskerville and cobey: Misconceptions about causation with synchrony and seasonal drivers," *Proceedings of the National Academy of Sciences*, vol. 114, no. 12, pp. E2272–E2274, 2017.

[87] J. Bongard and H. Lipson, "Automated reverse engineering of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 104, no. 24, pp. 9943–9948, 2007.

[88] G. Stepaniants, B. W. Brunton, and J. N. Kutz, "Inferring causal networks of dynamical systems through transient dynamics and perturbation," 2020.

[89] K. D. Baksi, B. K. Kuntal, and S. S. Mande, "'time': a web application for obtaining insights into microbial ecology using longitudinal microbiome data," *Frontiers in microbiology*, vol. 9, p. 36, 2018.

[90] J. D. Silverman, L. Shenhav, E. Halperin, S. Mukherjee, and L. A. David, "Statistical considerations in the design and analysis of longitudinal microbiome studies," *bioRxiv*, 2018.

[91] I. Florescu, *Probability and stochastic processes*. John Wiley & Sons, 2014.

[92] S. M. Ross, *Stochastic processes*, vol. 2. Wiley New York, 1996.

[93] P. Chodrow, "Divergence, entropy, information: An opinionated introduction to information theory," *arXiv preprint arXiv:1708.07459*, 2017.

[94] C. T. Perretti, S. B. Munch, and G. Sugihara, "Model-free forecasting outperforms the correct mechanistic model for simulated and experimental data," *Proceedings of the National Academy of Sciences*, vol. 110, no. 13, pp. 5253–5257, 2013.

[95] J. Huke, "Embedding nonlinear dynamical systems: A guide to takens' theorem," 2006.

1 Appendices

1.1 Dependence between random variables

Let's start with intuition. Consider the sizes of cells in a microbial population. We can use (random) variables X_1 and X_2 to record the volume and the length of a cell, and then repeat measurements for many cells (trials) to get data (realizations). If larger cells tend to be longer, then the volume and the length random variables are correlated and thus dependent. Mathematically, two random variables are dependent if they are not independent. A rigorous definition of independence is presented in Fig 11 B.

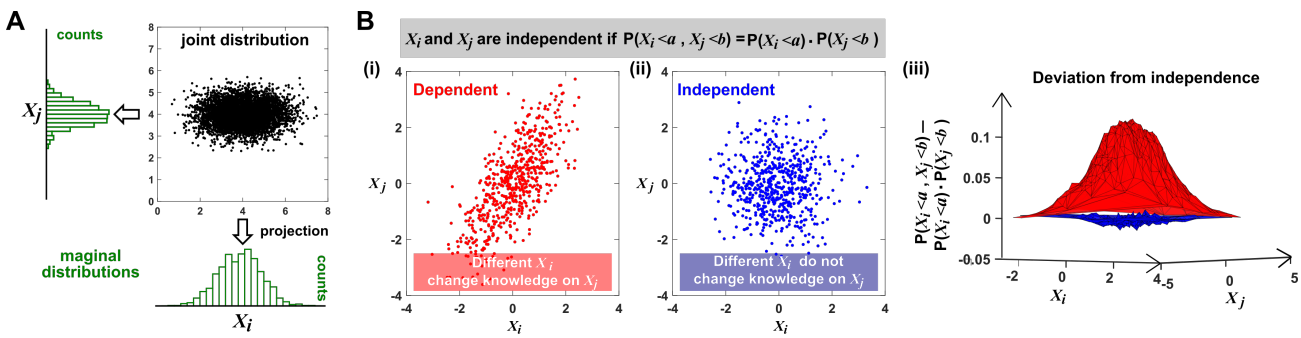

Figure 11: Joint distribution, marginal distributions, and dependence of two random variables. (**A**) A scatter-plot of data associated with random variables X_i and X_j represents a "joint distribution" (black). Histograms for data associated with X_i and for data associated with X_j represent "marginal distributions" (green). Strictly speaking, joint and marginal distributions must be normalized so that probabilities (here represented as "counts") sum to 1. Graphically, marginal distributions are projections of the joint distribution on the axes. Two random variables are identically distributed if their marginal distributions are identical. (**B**) Independence between two random variables. Gray box: the mathematical definition of independence, where "P" means probability. Visually, if two random variables are independent, then different values of one random variable will not change our knowledge about another random variable. In (**i**), X_j increased as X_i increased (different X_i led to different knowledge on X_j), and thus, X_i and X_j are not independent (i.e. they are dependent). In (**ii**), X_i and X_j were repeatedly drawn from two normal distributions. Thus, the two random variables are independent. One might argue that when X_i values become extreme, X_j values tend to land in the middle. However, this is a visual artifact caused by fewer data points at the more extreme X_i values. If we had plotted histograms of X_j at various X_i values, we would see that X_j is always normally distributed with the same mean and variance. (**iii**) Indeed, when we plotted the difference between the observed probability $P(X_i < a, X_j < b)$ and the probability expected from X_i and X_j being independent $P(X_i < a) \cdot P(X_j < b)$, (ii) showed a near-zero difference (blue), while (i) showed deviation from zero (red). This is consistent with X_i and X_j being independent in (ii) but not in (i).

Note that when measuring a random variable, sampling is with replacement, or can be thought of as from an infinite population. For example during dice rolling, if the first trial registers 1, then the second trial can register 1 as well. Otherwise, if sampling was done *without* replacement, then the second trial would not register 1, which means that the outcome of the second trial would depend on the outcome of the first trial. This would violate the requirement that realizations of a random variable should be independent. As another example, imagine that we raised 20 mice, and sacrificed 4 mice per time point for 5 time points. Although at each time point sampling was done without replacement, the original 20 mice can be thought of as being sampled from an infinite population of possible mice. Thus any pair of mice in this experiment can be considered independent.

1.2 Independent and identically distributed (IID) random variables

Two random variables are IID if they have the same probability distribution and are independent. Fig 11 above illustrates how to check whether two random variables have identical distributions and are independent.

In Fig 12 we give examples of random variables that are (or are not) identically distributed, and that are (or are not) independent. Note that two dependent random variables can be correlated (Fig 12 3rd column), or not (Fig 12 4th column). A common quantitative measurement of linear correlation known as Pearson correlation is nicely illustrated at https://en.wikipedia.org/wiki/Pearson_correlation_coefficient.

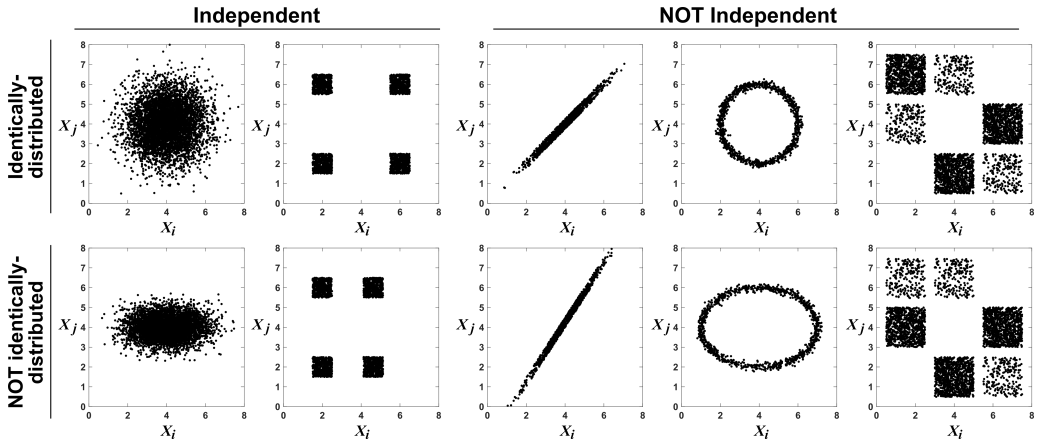

Figure 12: **Examples of random variables that are identically distributed or not identically distributed, and independent or not independent.** In the top row, X_i and X_j are identically distributed (projections of the scatter plot on both axes would have the same shape). Note that in the top row of the rightmost column, the scatter plot is not symmetric along the diagonal line, yet projections on both axes yield identical marginal distributions: three segments of equal densities. Thus, the two random variables are identically distributed. In the bottom row, X_i and X_j are not identically distributed. In the leftmost two columns, the two random variables are independent (for more details about independence, see Fig 11B). In the last three columns, the two random variables are dependent: different X_i values alter our knowledge of X_j.

A sample of values drawn from a mixed population can still be IID, as long as sample members are chosen randomly and independently (Fig 4).

1.3 Structural causal models

A structural causal model [30] is a popular way of thinking about causal relationships in statistics. Let's first consider the following example:

$$
\begin{aligned}
X_1 &:= \epsilon_1 \\
X_2 &:= \tan(X_1 + \epsilon_2) \\
X_3 &:= X_1 + X_2 + \epsilon_3
\end{aligned}
$$

In this example, X_1, X_2, and X_3 are random variables whose randomness is introduced by random process noise terms ϵ_1, ϵ_2, and ϵ_3, respectively. By using the := notation (instead of =) for assignments, we mean that each line in the above example is not only a mathematical statement of equality, but additionally has a directional causal meaning. $A := B$ means that B directly causes A, but not the reverse. In the above example, if one were to externally change the value of X_2, the perturbation would have an effect on X_3 (since X_2 shows up on the right side of the assignment for X_3), but not on X_1 (even though X_1 shows up on the right side of the assignment for X_2).

More generally, a structural causal model consists of a set of random variables X_1, X_2,...,X_n that are related by a set of assignments denoted by ":=":

$$X_j \coloneqq f_j(\mathcal{P}_j, \epsilon_j)$$
$$j = 1, ..., n$$

Here, $\mathcal{P}_j$ is interpreted as the subset of variables $X_1, ..., X_n$ that directly cause X_j. ϵ_j is a noise term that encodes randomness. That is, X_j is entirely determined by $\mathcal{P}_j$ and ϵ_j. Crucially, all of the noise terms $\epsilon_1, ..., \epsilon_n$ must be jointly independent ($P(\epsilon_1, ..., \epsilon_n) = P(\epsilon_1)...P(\epsilon_n)$). Note that the set of assignments induces a network if we draw a directed link from any direct causer to its direct causee. We require that this network have no cycles (colloquially, X_j must not cause itself either directly or indirectly).

Strictly speaking, these assumptions are sufficient for proving the common cause principle ([30] Proposition 6.28). We now remark on the requirement of an acyclic network. Note that our definition of causality (perturbation causality) in the main text does not exclude acyclic networks. For example, in the classic Lotka-Volterra system, the predator population size has a causal influence on the prey population size and vice versa. However, we can represent this system as an acyclic causal system by increasing our time resolution: The prey time t influences the predator at time $t + 1$ and the predator at time t influences the prey at time $t + 1$. Thus, by separating different time points into different variables, we can represent a cyclic dynamic system as an acyclic system.

1.4 Simpsons paradox: Temporal effects can induce artifactual correlations even between random variables

Here, we provide an example where temporal effects induce artifactual correlation between two random variables. We explore this example using the causal framework outlined in Appendix 1.3 and show that the example technically does not violate Reichenbach's common cause principle.

We consider an experiment on a population of animals. We inoculate the gut of each animal with two microbial taxa (X and Y) and then after either one or two time units we simultaneously measure the biomass of each taxon. These two taxa have no interaction and no common drivers (i.e. they are causally unrelated). However, each taxon tends to increases in biomass over time as it colonizes the animal gut. This is represented mathematically in Fig 13A. Here, X_1 and X_2 are the biomass of taxon X at time 1 and 2 respectively, and X_{obs} is the biomass of X at the time of measurement (randomly and independently chosen from X_1 and X_2). Y_1, Y_2, and Y_{obs} are similar but for taxon Y. C represents our random choice of when to measure an animal. The biomass measurements of our two taxa are correlated (Fig 13C), despite the fact that these taxa have no interaction or common driver. The X_{obs} values in Fig 13C are IID since each was generated by an independent and random draw from the same process (similar to Fig 4), and note that the Y_{obs} values are similarly IID.

What explains this correlation? The key is that X_{obs} is a distinct variable from X_1 and X_2 (and similar for Y_{obs}). X_1 and X_2 are causally independent of Y_1 and Y_2, but X_{obs} shares the common cause of C with Y_{obs} (Fig 13**B**). In other words, our choice of when to measure each animal is the causal explanation for the observed correlation. We can see this by noting that if we only consider those animals which were measured on time 1 (13**C**, olive dots) the correlation vanishes (and similarly with time 2). Thus, while Reichenbach's common cause principle technically does not fail in this instance, the observed correlation remains mostly unilluminating about the ecological relationship between X and Y. The emergence of a new trend upon combining different groups of data is known as Simpson's paradox.

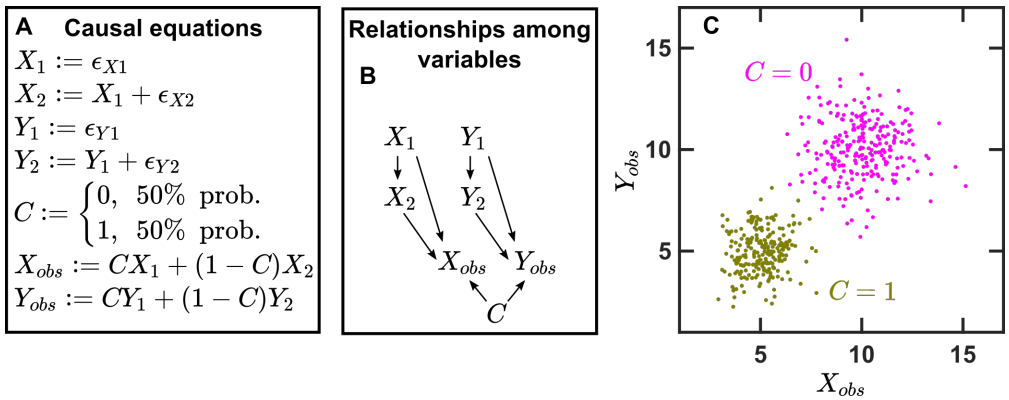

Figure 13: Even IID data may generate an artifactual correlation. (**A**) Equations describing a system where an animal gut is inoculated with two taxa and then measured after either one or two time units. X_1, X_2, and X_{obs} are the biomass of taxon X one time unit after inoculation, two time units after inoculation, and at the time of measurement respectively. Y_1, Y_2, and Y_{obs} are the same, but for taxon Y. C is our choice of whether to take measurements at time 1 or 2. The noise terms $\epsilon_{X1}, \epsilon_{X2}, \epsilon_{Y1}, \epsilon_{Y2}$ are all independent of one another and are all normal random variables with mean of 5 and standard deviation of 1. Thus X_1 and X_2 are independent of Y_1 and Y_2. As described in Appendix 1.3, the ":=" symbol denotes a directional causal assignment (e.g. $X_2 := X_1 + \epsilon_{X2}$ means that externally changing X_1 would affect X_2, but changing X_2 would not affect X_1). (**B**) A diagrammatic representation of the equations in Fig 13**A** where an arrow is drawn from variable A to variable B if A appears in the causal equation that determines B. (**C**) A scatterplot showing the correlation between X_{obs} and Y_{obs}. Points are colored according to their corresponding value of C (i.e. the time in which they were measured).

1.5 Stationarity

Intuitively speaking, stationarity means that the statistical properties of a time series do not change over time. In this context, a time series is represented by a temporally ordered series of random variables known as a *stochastic process*. Stationarity is a requirement for Granger causality tests [26, 89, 47, 90], which seek to learn Granger causal relationships between dynamic variables. Despite the importance of stationarity, a clear explanation of this concept is not always given. Here we describe definitions of stationarity.

There are two notions of stationarity: strong (strict) stationarity ([91], p. 636) and weak (wide-sense) stationarity ([91]p. 637). In strict stationarity, the probability distribution functions of random variables and joint distributions between any number of temporally adjacent random variables do not change over time.

A stochastic process with random variables $X_1, X_2, X_3, \ldots$ is strongly (or strictly) stationary if

$$F(X_1, X_2, \ldots, X_n) = F(X_{\tau+1}, X_{\tau+2}, \ldots, X_{\tau+n})$$

for all $\tau = 1, 2, ...$ and all $n = 1, 2, ...$, where F denotes a cumulative joint distribution function.

A process composed of a series of IID random variables (Appendix 1.2) is strictly stationary (top row of Fig 15). Ross [92] (p. 396) gives several other examples of stationary processes. Strict stationarity is very difficult to test for in practice [91]. Below we discuss weak stationarity.

A stochastic process X_t is weakly (or wide-sense) stationary if:

1. $\mathbb{E}[\mathrm{X_t}]$ (the ensemble mean) does not depend on t
2. $\mathrm{Var}[X_t]$ is finite and does not depend on t
3. For all choices of h, $\mathrm{Cov}(X_t, X_{t+h})$ does not depend on t

As an example similar to Fig 3C, consider a population whose dynamics are governed by death and stochastic migration:

$$X_t = (1-a)X_{t-1} + c + \epsilon_t \tag{3}$$

Here, X_t is the population size at time t, a is the probability of death during the time interval of 1, c is the average number of individuals migrating into the population during the time interval of 1, and ϵ_t is a random variable with a mean of zero which represents temporal fluctuations in the number of migrants. Suppose that we observed the dynamics of 10 populations governed by Eq. 3 such that the populations all have the same parameters, but are independent (Fig14A). Then, at each time point t, we will have some distribution of values of X_t. In fact, if we have not just 10, but 1,200 replicates, we can see that the distribution of values of X_t does not appear to depend on time (Fig 14B, top). Furthermore, the covariance between X_t and X_{t+1} does not appear to depend on time either (Fig 14B, bottom).

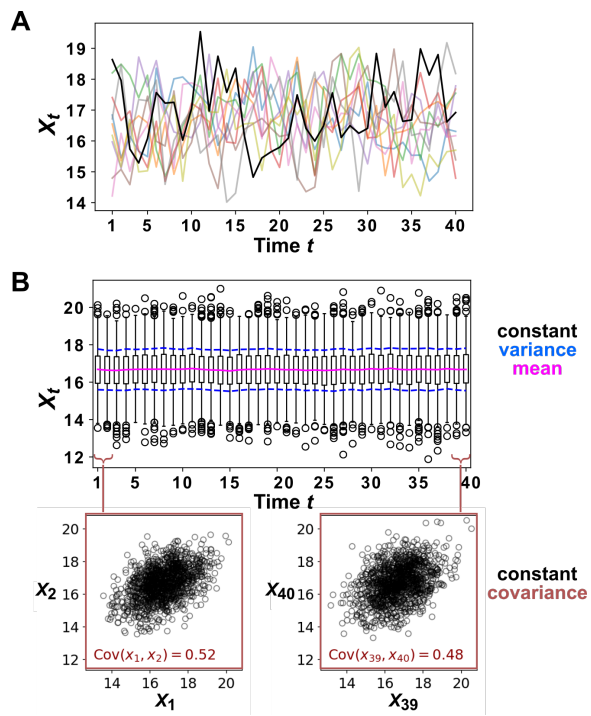

Figure 14: Example of a stationary process. (**A**) Ten replicate runs of the stochastic process described in Eq. 3 with parameter choices $a = 0.6$, $c = 10$, and ϵ is a normal random variable with mean of zero and standard deviation of 1. (**B**) The distribution X_t values shown for $1,200$ replicates runs of the same stochastic process as in (A). The mean of X_t is given as a solid red line and the mean $\pm$ standard deviation of X_t is shown as a dashed blue line. Bottom: X_t is plotted against X_{t+1} for two values of t.

Whether a time series obtained from some process is stationary depends on other time series obtained from the same process (Fig 15). The middle row of Fig 15 shows a system that can be readily shown to be weakly stationary:

$$\mathbb{E}[X_2(t)] = \int_0^{2\pi} \cos\left(\frac{t}{3} + \theta_2\right) d\theta_2 = 0$$

$$\mathrm{Var}[X_2(t)] = \int_0^{2\pi} \left(\cos\left(\frac{t}{3} + \theta_2\right)\right)^2 d\theta_2 = \pi$$

$$\mathrm{Cov}(X_2(t), X_2(t+h)) = \int_0^{2\pi} \cos\left(\frac{t}{3} + \theta_2\right) \cos\left(\frac{t+h}{3} + \theta_2\right) d\theta_2 = \pi\cos\left(\frac{h}{3}\right)$$

The time series of the bottom row of Fig 15 has the same dynamics as that of the middle row, but has

a more constrained set of phases. We can see that this time series is clearly nonstationary since its mean changes over time.

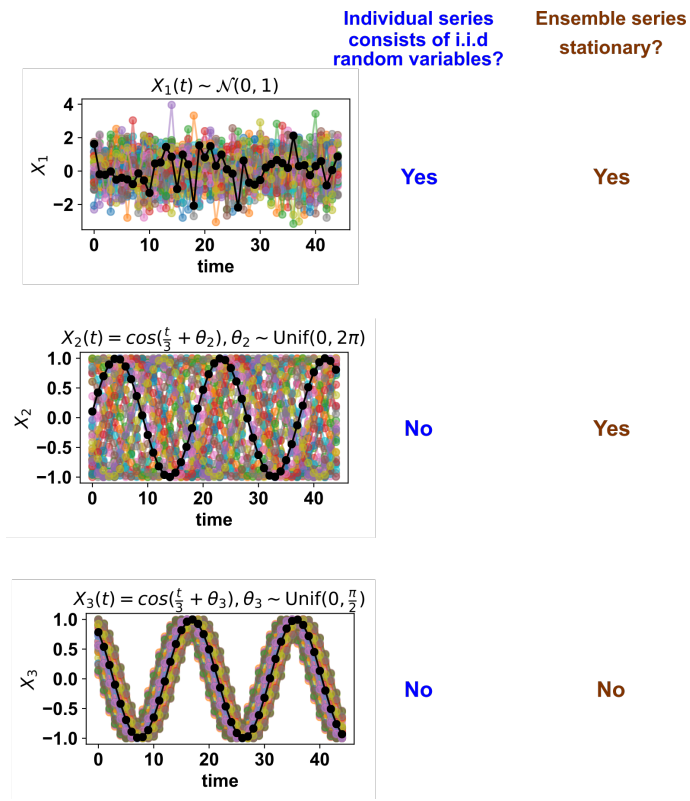

Figure 15: Whether a single time series is stationary or not depends on what other replicates look like. The top panel shows IID standard normal noise. The middle and bottom panel both show sinusoidal curves. The time series in the middle panel, but not the bottom panel, is weakly stationary. More examples on this topic can be found in the accompanying video.

1.6 General Granger causality and transfer entropy

Here we prove the claim from the main text that "X Granger causes Y (in the sense of Box 1, Definition 3) if and only if the transfer entropy from X to Y is nonzero". We begin by reviewing definitions and introducing a convenient notation, and then prove both directions of the statement. Our proof covers the discrete case, where state variables can be binned into discrete values.

We first restate the definition of general Granger causality here, using slightly different notation to facilitate comparison with transfer entropy. Let X Y, and Z be time series composed of random variables indexed by time (i.e. X_t is the value of X at time t). Let the bolded $\boldsymbol{X_t}$ be the current and historical values of X (i.e. $\boldsymbol{X_t} = \{X_t, X_{t-1}, X_{t-2}, \cdots\}$). Then X Granger-causes Y with respect to this information set $\{X_t, Y_t, Z_t\}$ if:

$$f(Y_{t+1}|\boldsymbol{X_t}, \boldsymbol{Y_t}, \boldsymbol{Z_t}) \neq f(Y_{t+1}|\boldsymbol{Y_t}, \boldsymbol{Z_t})$$

where $f(Y_{t+1}|\mathcal{S})$ is the probability distribution of Y_{t+1} conditional on $\mathcal{S}$.

We next give the multivariate definition of Transfer etropy. This is called "partial transfer entropy" in [52] because it is the *part* of the distribution of Y's future value that is explained by the history of X, but not by other variables (i.e. Z). The (partial) transfer entropy PTE from X to Y is defined as:

$$PTE_{X \to Y|Z} = \sum_{\boldsymbol{X_t}} \sum_{\boldsymbol{Y_t}} \sum_{\boldsymbol{Z_t}} \sum_{Y_{t+1}} f(Y_{t+1}, \boldsymbol{X_t}, \boldsymbol{Y_t}, \boldsymbol{Z_t}) \log \left(\frac{f\left(Y_{t+1}|\boldsymbol{X_t}, \boldsymbol{Y_t}, \boldsymbol{Z_t}\right)}{f\left(Y_{t+1}|\boldsymbol{Y_t}, \boldsymbol{Z_t}\right)} \right)$$

To see that a lack of Granger causality from X to Y implies that there is no partial transfer entropy from X to Y, simply assume that X does not Granger cause Y:

$$f(Y_{t+1}|\boldsymbol{X_t}, \boldsymbol{Y_t}, \boldsymbol{Z_t}) = f(Y_{t+1}|\boldsymbol{Y_t}, \boldsymbol{Z_t})$$

Then,

$$\begin{aligned} PTE_{X\to Y|Z} &= \sum_{\boldsymbol{X_t}}\sum_{\boldsymbol{Y_t}}\sum_{\boldsymbol{Z_t}}\sum_{Y_{t+1}} f(Y_{t+1}, \boldsymbol{X_t}, \boldsymbol{Y_t}, \boldsymbol{Z_t})\log(1) \\ &= 0 \end{aligned}$$

To see that zero partial transfer entropy from X to Y implies that X does not Granger cause Y, we can set $PTE_{X\to Y|Z} = 0$:

$$PTE_{X\to Y|Z} = \sum_{\boldsymbol{X_t}}\sum_{\boldsymbol{Y_t}}\sum_{\boldsymbol{Z_t}}\sum_{Y_{t+1}} f(Y_{t+1}, \boldsymbol{X_t}, \boldsymbol{Y_t}, \boldsymbol{Z_t})\log\left(\frac{f\left(Y_{t+1}|\boldsymbol{X_t}, \boldsymbol{Y_t}, \boldsymbol{Z_t}\right)}{f\left(Y_{t+1}|\boldsymbol{Y_t}, \boldsymbol{Z_t}\right)}\right) = 0$$

By dividing both sides by $f(\boldsymbol{X_t}, \boldsymbol{Y_t}, \boldsymbol{Z_t})$ we have:

$$\sum_{\boldsymbol{X_t}}\sum_{\boldsymbol{Y_t}}\sum_{\boldsymbol{Z_t}}\sum_{Y_{t+1}} f\left(Y_{t+1}|\boldsymbol{X_t}, \boldsymbol{Y_t}, \boldsymbol{Z_t}\right)\log\left(\frac{f\left(Y_{t+1}|\boldsymbol{X_t}, \boldsymbol{Y_t}, \boldsymbol{Z_t}\right)}{f\left(Y_{t+1}|\boldsymbol{Y_t}, \boldsymbol{Z_t}\right)}\right) = 0 \tag{4}$$

At this point, note that the Gibbs inequality (e.g. Theorem 2 in [93]) states that if $p(a)$ and $q(a)$ are two discrete probability distributions defined over the same values (e.g. $p(a)$ and $q(a)$ are probability distributions of shoe sizes a purchsed by males and females defined between 1 and 15) then:

$$\sum_a p(a)\log\left(\frac{p(a)}{q(a)}\right) \geq 0 \tag{5}$$

where the equality holds if and only if $p = q$. Since $f\left(Y_{t+1}|\boldsymbol{X_t}, \boldsymbol{Y_t}, \boldsymbol{Z_t}\right)$ and $f\left(Y_{t+1}|\boldsymbol{X_t}, \boldsymbol{Y_t}\right)$ are both probability distributions over the same values (Y_{t+1}), Eq. 5 tells us that for all choices of $\boldsymbol{X_t}, \boldsymbol{Y_t}, \boldsymbol{Z_t}$,

$$\sum_{Y_{t+1}} f\left(Y_{t+1}|\boldsymbol{X_t}, \boldsymbol{Y_t}, \boldsymbol{Z_t}\right)\log\left(\frac{f\left(Y_{t+1}|\boldsymbol{X_t}, \boldsymbol{Y_t}, \boldsymbol{Z_t}\right)}{f\left(Y_{t+1}|\boldsymbol{Y_t}, \boldsymbol{Z_t}\right)}\right) \geq 0 \tag{6}$$

The lefthand side of Eq. 4 is a sum of terms shown in the lefthand side of Eq. 6, and if multiple nonnegative terms sum to zero, then each term must be zero. Thus:

$$\sum_{Y_{t+1}} f\left(Y_{t+1}|\boldsymbol{X_t}, \boldsymbol{Y_t}, \boldsymbol{Z_t}\right)\log\left(\frac{f\left(Y_{t+1}|\boldsymbol{X_t}, \boldsymbol{Y_t}, \boldsymbol{Z_t}\right)}{f\left(Y_{t+1}|\boldsymbol{Y_t}, \boldsymbol{Z_t}\right)}\right) = 0 \tag{7}$$

Eq. 7 is the equality case of a Gibbs inequality, which implies that $f\left(Y_{t+1}|\boldsymbol{X_t}, \boldsymbol{Y_t}, \boldsymbol{Z_t}\right) = f\left(Y_{t+1}|\boldsymbol{Y_t}, \boldsymbol{Z_t}\right)$. This shows that X does not Granger cause Y, which completes the proof.

1.7 Infrequent sampling

In this section, we consider the systems shown in Fig 5Biv and show how one can derive the infrequent sampling case from the frequent sampling case. In Fig 5Biv we show the system

$$\begin{aligned} X_t &= 0.4X_{t-1} + 0.6Y_{t-1} + \epsilon_{X,t-1} \\ Y_t &= 0.5Y_{t-1} + \epsilon_{Y,t-1} \end{aligned} \tag{8}$$

where $\epsilon_{X,t-1}$ and $\epsilon_{Y,t-1}$ are independent normal random variables with 0 mean and standard deviation of 1. We state that if one samples this system infrequently, one arrives at the following system

$$
\begin{aligned}
X_t &\approx 0.0001X_{t-10} + 0.005Y_{t-10} + 1.431\beta_{X,t} \\
Y_t &\approx 0.001Y_{t-10} + 1.155\beta_{Y,t} \\
\mathrm{Cov}(\beta_{X,t}, \beta_{Y,t}) &\approx 0.303
\end{aligned} \tag{9}
$$

where $\beta_{X,t}$ and $\beta_{Y,t}$ are normal random variables with 0 mean and standard deviation of 1. This demonstrates that increasing the sampling frequency can reduce the signal-to-noise ratio (effect of causer : effect of process noise). In this section we show how one can derive Eq. 9 from Eq. 8.

We begin by rewriting Eq. 8 in matrix form and putting $t+1$ rather than t on the left hand side:

$$
\begin{pmatrix} X_{t+1} \\ Y_{t+1} \end{pmatrix} = \begin{pmatrix} 0.4 & 0.6 \\ 0 & 0.5 \end{pmatrix} \begin{pmatrix} X_t \\ Y_t \end{pmatrix} + \begin{pmatrix} \epsilon_{X,t} \\ \epsilon_{Y,t} \end{pmatrix}
$$

In general, this looks like:

$$
\boldsymbol{U}_{t+1} = \boldsymbol{A}\boldsymbol{U}_t + \boldsymbol{\epsilon}_t \tag{10}
$$

Here, for a system of m state variables, $\boldsymbol{U}_t$ is a $m \times 1$ vector describing variables of the system at time t, $\boldsymbol{A}$ is an $m \times m$ matrix describing how the system evolves, and $\boldsymbol{\epsilon}_t$ is an $m \times 1$ random variable that introduces process noise to the system. Eq. 10 looks one step ahead. We can look two steps ahead by plugging Eq. 10 into itself.

$$
\begin{aligned}
\boldsymbol{U}_{t+2} &= \boldsymbol{A}\boldsymbol{U}_{t+1} + \boldsymbol{\epsilon}_{t+1} \\
&= \boldsymbol{A}^2\boldsymbol{U}_t + \boldsymbol{A}\boldsymbol{\epsilon}_t + \boldsymbol{\epsilon}_{t+1}
\end{aligned}
$$

More generally, we can look n steps ahead (as if we were sampling every n steps rather than every step) according to

$$
\boldsymbol{U}_{t+n} = \boldsymbol{A}^n\boldsymbol{U}_t + \sum_{j=1}^{n} \boldsymbol{A}^{n-j}\boldsymbol{\epsilon}_{t+j-1} \tag{11}
$$

A proof of Eq. 11 is provided at the end of this section. Let us now assume that $\boldsymbol{\epsilon}_t$ $(t = 1, 2, ...)$ are IID multivarite normal distributions with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{C}$. For example, if ϵ is the IID standard normal distribution, then $\boldsymbol{\mu}$ is the $m \times 1$ zero vector and $\boldsymbol{C}$ is the $m \times m$ identity matrix whose diagonal entries are 1s and whose off-diagonal entries are 0s. Then, we can rewrite Eq. 11 as:

$$
\boldsymbol{U}_{t+n} = \boldsymbol{A}^n\boldsymbol{U}_t + \boldsymbol{\eta}_{n,t} \tag{12}
$$

where $\boldsymbol{\eta}_{n,t}$ is a multivariate normal variable with mean $\boldsymbol{\mu}_n$ and covariance $\boldsymbol{C}_n$ given by:

$$
\begin{aligned}
\boldsymbol{\mu}_n &= \left(\sum_{j=1}^{n} \boldsymbol{A}^{n-j} \right) \boldsymbol{\mu} \\
\boldsymbol{C}_n &= \sum_{j=1}^{n} \boldsymbol{A}^{n-j}\boldsymbol{C}\left(\boldsymbol{A}^{n-j}\right)^\top
\end{aligned} \tag{13}
$$

where $M^\top$ is the transpose of M. The top line of Eq. 13 follows from the linearity of expectation. The bottom line of Eq. 13 follows from the covariance rules for linear transformation (Eq. 14) or sum (Eq. 15) of vector-valued random variables.

If we plug the relevant parameters for Eq. 8 (i.e. $A = \begin{pmatrix} 0.4 & 0.6 \\ 0 & 0.5 \end{pmatrix}$, $\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\boldsymbol{C} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$), and $n = 10$, then $\boldsymbol{C}_n \approx \begin{pmatrix} 2.048 & 0.500 \\ 0.500 & 1.333 \end{pmatrix}$ and Eq. 12 becomes Eq. 9. The diagonal entries of $\boldsymbol{C}_n$ are variances and the off-diagonal entries are identical and represent the covariance between the two elements of $\boldsymbol{\eta}_{n,t}$, the

process noises of X and Y accumulated over the $n = 10$ steps since t. Finally, we rescale the noise terms in 9 by letting

$$\begin{bmatrix} \sqrt{2.048}\beta_{X,t} \\ \sqrt{1.333}\beta_{Y,t} \end{bmatrix} = \boldsymbol{\eta}_{n,t}.$$

This rescaling gives us noise terms (β) in Eq. 9 with unit variance, putting them on the same scale as the ϵ noise terms in Eq. 8. Note that now, the covariance of $\beta_{X,t}$ and $\beta_{Y,t}$ is:

$$\begin{aligned}
\mathrm{Cov}(\beta_{X,t}, \beta_{Y,t}) &\approx \mathrm{Cov}(\frac{\boldsymbol{\eta}_{n,t}(1)}{\sqrt{2.048}}, \frac{\boldsymbol{\eta}_{n,t}(2)}{\sqrt{1.333}}) \\
&\approx \mathrm{E}\left[\frac{1}{\sqrt{2.048}}\left(\boldsymbol{\eta}_{n,t}(1) - \mathrm{E}\left[\boldsymbol{\eta}_{\mathrm{n,t}}(1)\right]\right)\frac{1}{\sqrt{1.333}}\left(\boldsymbol{\eta}_{n,t}(2) - \mathrm{E}\left[\boldsymbol{\eta}_{\mathrm{n,t}}(2)\right]\right)\right] \\
&\approx \frac{\mathrm{Cov}(\boldsymbol{\eta}_{\mathrm{n,t}}(1), \boldsymbol{\eta}_{\mathrm{n,t}}(2))}{\sqrt{2.048}\sqrt{1.333}} \\
&\approx \frac{0.500}{\sqrt{2.048}\sqrt{1.333}} \\
&\approx 0.303
\end{aligned}$$

Covariance of a linear transformation of a vector-valued random variable

Suppose $\boldsymbol{X}$ is an $m \times 1$ vector of random variables whose mean vector is the $m \times 1$ vector $\boldsymbol{\mu_X}$ and whose covariance is the $m \times m$ matrix $\boldsymbol{C_X}$. Let $\boldsymbol{Y} = \boldsymbol{AX}$ where $\boldsymbol{A}$ is an $m \times m$ matrix. Thus, $\boldsymbol{Y}$ is an $m \times 1$ vector. Denote the mean and covariance of $\boldsymbol{Y}$ by $\boldsymbol{\mu_Y}$ and $\boldsymbol{C_Y}$. Then:

$$\boldsymbol{C_Y} = \boldsymbol{AC_XA}^\top \tag{14}$$

To see this, first note that $\boldsymbol{\mu_Y} = \boldsymbol{A\mu_X}$. Then we can derive Eq. 14 directly from the definition of the covariance matrix.

$$\begin{aligned}
\boldsymbol{C_Y} &= \mathrm{E}\left[\left(\boldsymbol{Y} - \boldsymbol{\mu_Y}\right)\left(\boldsymbol{Y} - \boldsymbol{\mu_Y}\right)^\top\right] \\
&= \mathrm{E}\left[\left(\boldsymbol{AX} - \boldsymbol{A\mu_X}\right)\left(\boldsymbol{AX} - \boldsymbol{A\mu_X}\right)^\top\right] \\
&= \mathrm{E}\left[\boldsymbol{A}\left(\boldsymbol{X} - \boldsymbol{\mu_X}\right)\left(\boldsymbol{X} - \boldsymbol{\mu_X}\right)^\top \boldsymbol{A}^\top\right] \\
&= \boldsymbol{A}\left(\mathrm{E}\left[\left(\boldsymbol{X} - \boldsymbol{\mu_X}\right)\left(\boldsymbol{X} - \boldsymbol{\mu_X}\right)^\top\right]\right)\boldsymbol{A}^\top \\
&= \boldsymbol{AC_XA}^\top
\end{aligned}$$

Covariance of a sum of independent vector-valued random variables

Suppose that $\boldsymbol{X}$ is an $m \times 1$ random variables with covariance matrix $\boldsymbol{C_X}$ and that $\boldsymbol{Y}$ is an $m \times 1$ random variables with covariance matrix $\boldsymbol{C_Y}$. Suppose further that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent of each other. That is, each element of $\boldsymbol{X}$ is independent of each element of $\boldsymbol{Y}$. Let $\boldsymbol{C_{X+Y}}$ be the covariance matrix of $\boldsymbol{X} + \boldsymbol{Y}$. Then it is a fact that:

$$\boldsymbol{C_{X+Y}} = \boldsymbol{C_X} + \boldsymbol{C_Y} \tag{15}$$

We can see that Eq. 15 is true by showing that it holds for each element of the matrix $\boldsymbol{C_{X+Y}}$. Specifically,

we will apply the definition of covariance to the element of $\boldsymbol{C_{X+Y}}$ at the ith row and the jth column:

$$
\begin{aligned}
(\boldsymbol{C_{X+Y}})_{ij} &= \mathrm{Cov}(X_i + Y_i, X_j + Y_j) \\
&= \mathrm{E}\left[\left((X_i - \mathrm{E}\left[X_i\right]) + (Y_i - \mathrm{E}\left[Y_i\right])\right)\left((X_j - \mathrm{E}\left[X_j\right]) + (Y_j - \mathrm{E}\left[Y_j\right])\right)\right] \\
&= \mathrm{E}[(X_i - \mathrm{E}\left[X_i\right])\left(X_j - \mathrm{E}\left[X_j\right]\right) + (X_i - \mathrm{E}\left[X_i\right])\left(Y_j - \mathrm{E}\left[Y_j\right]\right) + \\
&\qquad \left(Y_i - \mathrm{E}\left[Y_i\right]\right)\left(X_j - \mathrm{E}\left[X_j\right]\right) + \left(Y_i - \mathrm{E}\left[Y_i\right]\right)\left(Y_j - \mathrm{E}\left[Y_j\right]\right)] \\
&= \mathrm{Cov}(X_i, X_j) + \mathrm{Cov}(X_i, Y_j) + \mathrm{Cov}(Y_i, X_j) + \mathrm{Cov}(Y_i, Y_j) \\
&= \mathrm{Cov}(X_i, X_j) + \mathrm{Cov}(Y_i, Y_j) \\
&= (\boldsymbol{C_X})_{ij} + (\boldsymbol{C_Y})_{ij}
\end{aligned}
$$

where the second to last line uses the fact that $\mathrm{Cov}(Y_i, X_j) = \mathrm{Cov}(X_i, Y_j) = 0$ since $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent of each other.

Proof of Eq. 11

We prove Eq. 11 by induction. For the base case, let $n = 1$. Then Eq. 11 reduces to Eq. 10. For the inductive step, we assume Eq. holds for $n = k$ and wish to show that it holds for $n = k + 1$. Thus,

$$
\begin{aligned}
\boldsymbol{U}_{t+k+1} &= \boldsymbol{A}(\boldsymbol{U}_{t+k}) + \boldsymbol{\epsilon}_{t+k} \\
&= \boldsymbol{A}\left(\boldsymbol{A}^k \boldsymbol{U}_t + \sum_{j=1}^{k} \boldsymbol{A}^{k-j} \boldsymbol{\epsilon}_{t+j-1}\right) + \boldsymbol{\epsilon}_{t+k} \\
&= \boldsymbol{A}^{k+1} \boldsymbol{U}_t + \sum_{j=1}^{k} \boldsymbol{A}^{k-j+1} \boldsymbol{\epsilon}_{t+j-1} + \boldsymbol{\epsilon}_{t+k} \\
&= \boldsymbol{A}^{k+1} \boldsymbol{U}_t + \left(\boldsymbol{A}^k \boldsymbol{\epsilon}_t + \boldsymbol{A}^{k-1} \boldsymbol{\epsilon}_{t+1} + \cdots + \boldsymbol{A} \boldsymbol{\epsilon}_{t+k-1}\right) + \boldsymbol{\epsilon}_{t+k} \\
&= \boldsymbol{A}^{k+1} \boldsymbol{U}_t + \sum_{j=1}^{k+1} \boldsymbol{A}^{k+1-j} \boldsymbol{\epsilon}_{t+j-1}
\end{aligned}
$$

The final line shows that we have shown that Eq. 11 holds for $n = k + 1$, completing the inductive step and hence the proof.

1.8 Deterministic processes with many variables may appear stochastic

A deterministic time series from a system with many variables can be approximated as stochastic. This is illustrated below in Fig 16. When we track the trajectory of a particle in a box with 99 other particles (16 bottom row), the observed trajectory appears random, even though the governing equations of motion are deterministic. In particular, the motion of our particle over each time step can be approximated as random. This sets the apparent randomness of many-variable dynamics apart from that of a different phenomenon called chaos. In chaotic dynamics, each time step need not be random, but small changes in initial conditions lead to large changes in later states.

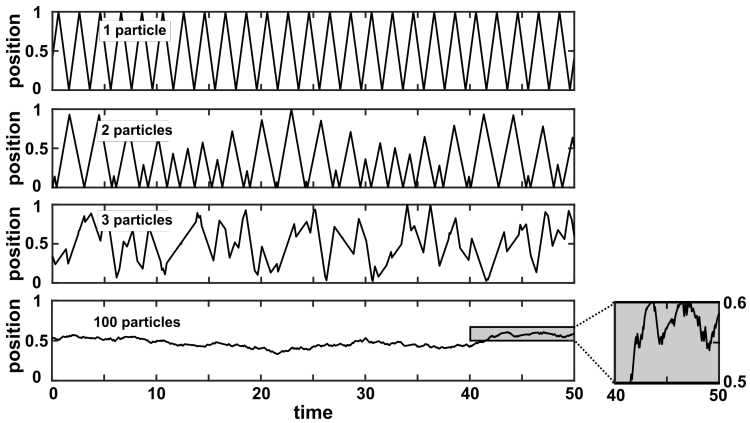

Figure 16: A many-variable deterministic system can be approximated as a stochastic system. The position of a particle in a system of particles bouncing in a 1-dimensional box is plotted over time. In each simulation, particles with radius 0 bounced in a box with walls of infinite mass placed at positions 0 and 1. All particles have mass 1 and are initialized at a random position between 0 and 1 according to a uniform distribution. Initial velocities are chosen in the following way: The initial velocity of each particle in a box is first randomly chosen from between -1 and 1 according to a uniform distribution. Then, all initial velocities in a given box are multiplied by the same constant to ensure that the total kinetic energy of each box is 0.5. Kinetic energy is conserved throughout the simulation. The simulation then follows the particles as they experience momentum-conserving collisions with one another and with the walls.

1.9 Considerations for selecting delay vector parameters for SSR

To construct delay vectors for SSR, one must choose the delay vector length E and the time delay τ. How does one choose E and τ? In general, detecting a continuous delay map requires that E be high enough so that no two parts of the delay space cross. For example, using $E = 2$ (instead of $E = 3$) to make Fig 6C would have projected the delay space onto 2 dimensions. This would introduce line crossings, which would in turn produce artifactual discontinuities in the shading. SSR is less sensitive to τ, although it is possible to mask a continuous delay map by choosing a "bad" τ. For example, consider what would happen to Fig 6C if we set τ to the period of Z. Since the delay vector is $[Z(t), Z(t-\tau), Z(t-2\tau)]$, setting set τ to the period of Z would force all 3 elements of the delay vector to always be equal. In geometric terms, this would compress the delay space onto a line, destroying the continuous delay map. However, bad choices of τ such as this are rare. In practice, a variety of methods are available for systematically choosing E and τ [19, 31]. Appendix 1.10 discusses theoretical requirements for E and τ in greater detail.

1.10 Historical notes on the basis of SSR

SSR is about the existence and properties of delay maps. Here the term "Map" is used interchangeably with "function": a map from X to Y sends every point in X to one and only one point in Y. Different authors have focused on different aspects of delay maps when analyzing SSR, possibly introducing confusion over which aspect is the "correct" criterion for casuality. In particular, while CCM is thought to look for local smoothness of the map from the delay space of a causee to the values of a causer ([72]), [73] proposes that a method to detect continuity of the map is more consistent with the underlying theory. So is SSR about smoothness, or continuity? As we will see, Takens's theorem guarantees smoothness (a stronger criterion) but for a restricted set of cases, whereas the theorem of Sauer et al. essentially guarantees continuity (a weaker condition) but for a broader set of cases including fractals.

Takens's celebrated 1981 paper [58] is arguably the first major theoretical result underpinning a variety of data-driven methods for both causality detection (as discussed in this article) and forecasting (e.g. [94]). Theorem 1 of [58] is reproduced below, except that we have changed the names of some variables:

Takens's theorem (theorem 1 of [58]): Let M be a compact manifold of dimension m. For pairs (ϕ, f), $\phi : M \to M$ a smooth diffeomorphism and $f : M \to \mathbb{R}$ a smooth function, it is a generic property that the map $\Phi_{(\phi,f)} : M \to \mathbb{R}^{2m+1}$, defined by

$$\Phi_{(\phi,f)}(p) = (f(p), f(\phi(p)), ..., f(\phi^{2m}(p))$$

is an embedding; by "smooth" we mean at least C^2.

We will attempt to illustrate Takens's theorem using the example in Fig 17. This system is described by a five-dimensional space of $[X, Y, Z, dX/dt, dY/dt]$ at various times (Fig 17A). However, once we have chosen the five initial conditions, X and Y uniquely determine dX/dt and dY/dt. Thus we visualize the state space in just three dimensions (X, Y, Z) (Fig 17C), and color the trajectory with time (a colored ring in Fig 17D similar to a clock to reflect the periodic nature of system dynamics). This trajectory is the manifold M in Takens's theorem and is 1-dimensional ($m = 1$) since it is a loop. We can think of ϕ as a function that maps a point p on the manifold M at current time t to the point q at a previous time $t - \tau$, and therefore $\phi^2(p)$ would operate ϕ twice and map p at current time t to the point r at $t - 2\tau$ (olive in Fig 17B). The term "diffeomorphism" in the theorem means that both this function ϕ and its inverse function (the map from past to present) are smooth (Fig 18). f can be viewed as an "observation" function that maps each point on the manifold to a single real number (e.g. the corresponding value of Z in Fig 17E; i.e $f(p) = Z$). Takens's theorem then asks us to consider a function Φ that maps a point $p(t)$ on our state space manifold (Fig 17E) to the "delay space" $[Z(t), Z(t-\tau), Z(t-2\tau)]$ (Fig 17F). This choice of delay space comes from three earlier choices: First, we consider delayed values of Z since Z is our observation function f; second, since $m = 1$, the delay space should be of dimension 3 ($= 2m + 1$); third, the delay τ comes from our diffeomorphism ϕ. Then, Takens's theorem states that for "most" (technically, "generic") choices of f and ϕ, Φ is an embedding. This means that Φ is diffeomorphic to its image, i.e. the curve in delay space will map smoothly to the state space manifold and vice versa [95].

Indeed from Fig 17 (C-F), we can see that when the observation function $f = Z(t)$, Φ from the state space to the delay space is a map. This is because each dot in the state space corresponds to a single time color (i.e. a point within a period), and each time color corresponds to a single dot in delay space, and thus, each point in the state space corresponds to a single point in the delay space. Moreover, Φ is continuous because the maps from state space to the time ring and from the time ring to delay space are both continuous. Similarly, the inverse of Φ from delay space to state space is also a continuous map. Moreover, we know from solutions to our equations that both maps are additionally smooth. Thus, we have verified Takens's theorem for this observation function.

Strikingly, if the observation function is Y, we will no longer have a continuous map from the delay space of Y to the state space. This is visualized as "bumpy coloring" in Fig 17H. In fact, we cannot even map the delay space to the time ring or the state space: (p, q, r) and (p', q', r') occupy the same point in the Y delay space, yet correspond to different times within a period (Fig 17B)) and thus they correspond to different locations in the state space. In summary, we cannot map the Y delay space to state space. Takens's theorem took care of this pathology using the word "generic". That is, Y is not considered a generic observation function here. On the other hand, if we use an observation function based on 95% Y mixed with 5% Z, we get an embedding from the state space and the delay space (Fig 17I-J). This is essentially what the term "generic" means in the context of topology: Although some observation functions do not give you an embedding, these "bad" observation functions can be tweaked just a little bit to become "good" ones. Similarly, some choices of ϕ do not work (i.e. $\tau = T$ for this system), but these are exceptions (see [95] Theorem 2 for what makes a ϕ "generic").

Roughly speaking, Sugihara and colleagues essentially used the values of one variable to shade the delay space of another variable, and used the local smoothness of the map to infer causality. In the example of Fig 6 in the main text, shading delay space of Z with Y generates a continuous (and smooth) pattern, consistent with Y causing Z. On the other hand, shading delay space of Z with W shows a bumpy pattern, consistent with W not causing Z.

Sauer and colleagues [59] later extended this work by proving a similar result that is in some ways more general. Theorem 2.5 in [59] is similar to Theorem 1 in [58], but now the system does not need to evolve on a manifold, but can evolve on a broader set of spaces, including fractals. Additionally, [59] replaces the concept of "generic" with a different notion ("prevalence"), which is closer to a probabilistic statement. Cummins

et al. then proved an important corollary to Sauer's result (corollary 4.3 in [73]), which showed that even under the more general conditions of [59], the map from the delay space to the state space is continuous.

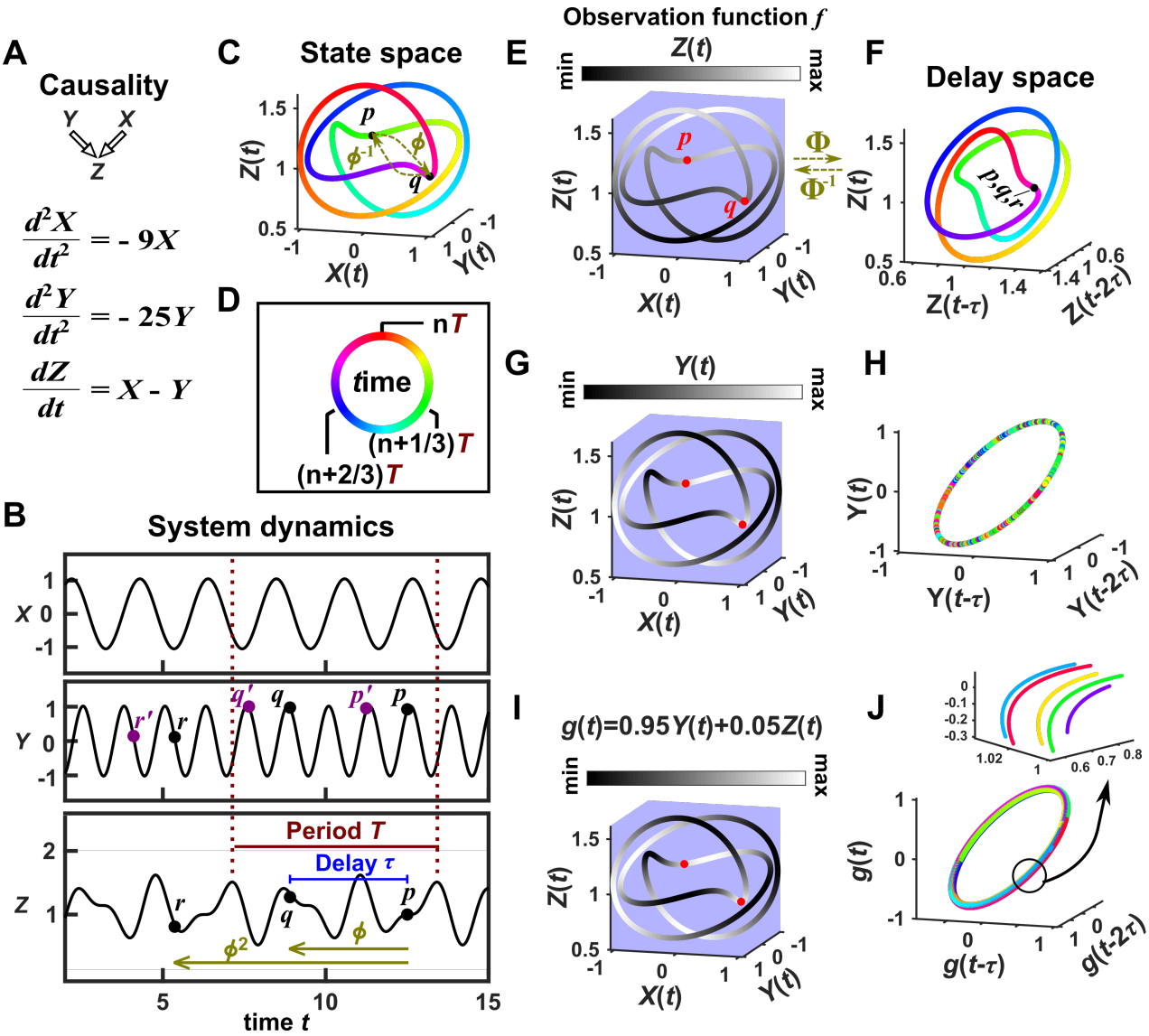

Figure 17: Illustration of Takens's theorem. (**A**) We consider a 3-variable toy system in which X and Y causally influence Z, but Z does not influence X or Y. (**B**) Time series of the three variables. (**C**) We can plot time series data in the state space M. Takens's theorem requires that ϕ, a function that maps a point p at current time t to the point q at a previous time $t-\tau$, and its inverse ϕ^{-1} (from past to current) are both smooth (C^2: the first and second derivatives of the function exist and are continuous for all time). To mark time progression, we color each point along the trajectory with its corresponding time value where time is represented as a color ring similar to a clock to reflect the periodic nature of system dynamics (**D**). (**E, G, I**) Shading the state space with the observation function (f in Takens's theorem) marked above. (**F, H, J**) Delay space based on the observation function, colored with time. The map Φ in Takens's theorem maps, for example, point p in E to point pqr in F. The theorem states that for "generic" observation functions, this map Φ and its inverse Φ^{-1} are both smooth (differentiable). In this example $\tau = 3.6$. In J, multiple colors in a region are due to one period wrapping around the delay space multiple times (inset), but the color shading transition is gradual (similar to F).

1.11 Difficulty of evaluating the continuity or smoothness of a function with finite or noisy data

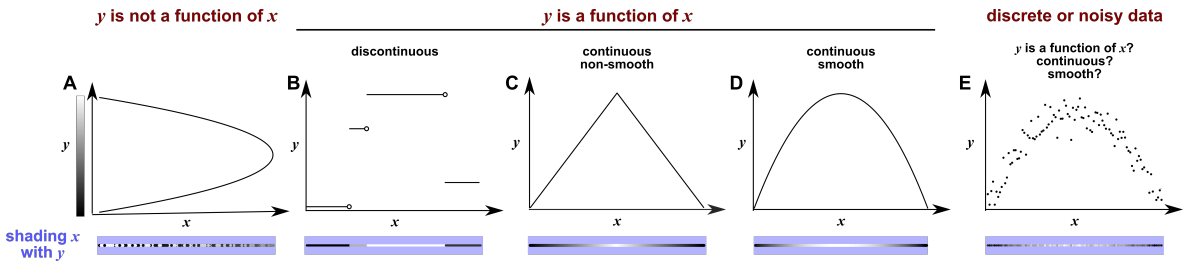

Figure 18: Difficulty of evaluating the continuity or smoothness of a function with finite or noisy data. (**A**) y is not a function of x because a single $x$value can correspond to more than one y value. Here, when we shade x with y value, we randomly choose the upper or the lower y value, leading to bumpy shading, similar to what we might expect to occur in the real world. (**B**) y is a discontinuous function of x. This is because at any "breakpoint" (circle) between two adjacent segments, the limit taken from the left-hand side is unequal to the limit taken from the right-hand side. Shading x with y generates a "bumpy" pattern. (**C**) y is a continuous and nonsmooth function of x. The function is nonsmooth because at the maximum point, the slope taken from the left-hand side is unequal to the slope taken from the right-hand side. It is a continuous function, and shading x with y generates a gradual pattern. (**D**) y is a continuous and smooth function of x. At every point, the slope taken from the left-hand side is equal to the slope taken from the right-hand side (the slope at the maximal point being zero). Shading x with y generates a gradual pattern. (**E**) With finite and noisy data, often shading x with y generates a bumpy pattern. However, it is unclear where y is a function of x, and if yes, whether the function is continuous (and perhaps even smooth).

1.12 The prediction lag test suffers serious failure modes

State space reconstruction methods suffer false positive errors in the presence of "strong forcing" [60]. This occurs when "the dependence of the dynamics of the forced variable on its own state is no longer significant" [60]. Ye et al. proposed a test in an effort to solve this problem [74]. Their procedure relies on finding mappings from the delay vector $[X(t), X(t-\tau), X(t-2\tau)...X(t-(\text{E-1})\tau)]$ to $Y(t+l)$, where E is the delay vector length, τ is the time lag, and l is a key variable known as the "prediction lag". They then examine how the cross map skill (Fig 8B) varies with the prediction lag. According to the authors, if the cross map skill is maximized at a positive prediction lag ($l > 0$), then the putative causality is spurious and arose from, for example, strong unidirectional forcing. On the other hand, if the highest quality mapping occurs at a non-positive prediction lag ($l \leq 0$), then we have further evidence that the detected causality is real and not spurious.

We find that while this test correctly distinguishes between real and spurious causal signals at some times, at other times it does not. In Fig 19, we explore the behavior of this test. Within each row of Fig 19, we examine a different system and ask whether Y causes X according to: (1) the ground truth model, (2) our visual continuity test, (3) a CCM cross map skill test (without the prediction lag test), and (4) the prediction lag test.

In rows 1 and 2 of Fig 19, the prediction lag test performs well, overturning the results of the visual continuity and CCM tests when apparent causality is spurious (row 1), and agreeing with the continuity and CCM tests (row 2) when apparent causality is real (modified from [74] Eq. 2). However in row 3, the prediction lag test dismisses a true causal link as spurious. Moreover, when we apply the prediction lag test to a system with a periodic putative driver (Fig 19 row 4), we find that cross map skill is a periodic function of the prediction lag. While this result is precisely what we would expect mathematically, its causal interpretation is unclear. The fifth row of Fig 19 takes the idea of strong forcing to the extreme, so that $Y(t+1)$ is a function of $X(t)$, but not $Y(t)$. Here the prediction lag test gives a false positive error. In the bottom row, X and Y do not interact, but are both driven by a common cause W with different lags. Specifically, $W(t)$ exerts a direct effect on $Y(t+1)$ and on $X(t+3)$. Thus, Y receives the same information

as X, but at an earlier time, analogous to Fig 5iii. Consistent with this, delay vectors of X predict past values of Y better than future values of Y. Thus, the prediction lag test produces a false positive error.

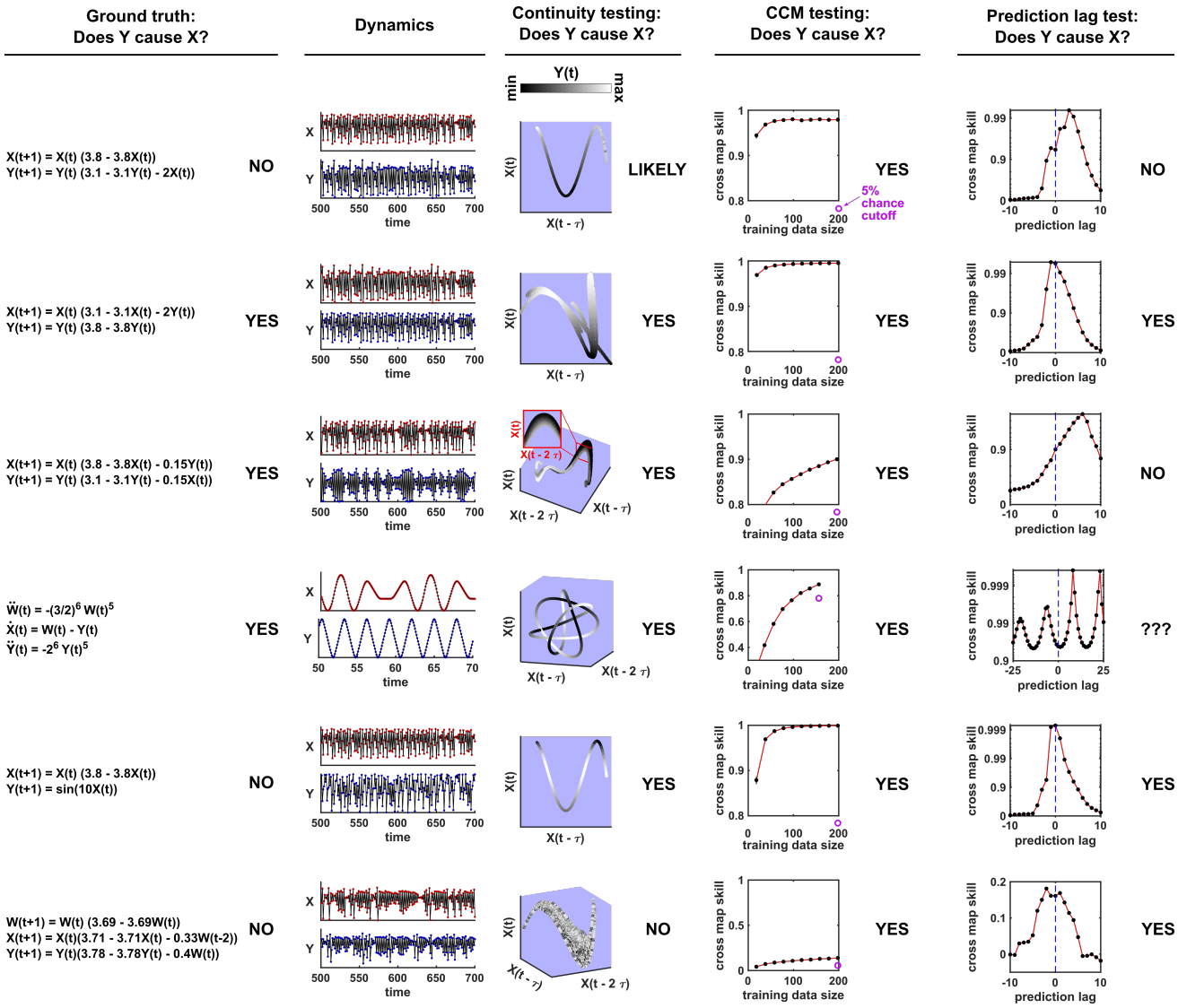

Figure 19: Comparison of visual continuity testing, cross map skill testing, and prediction lag testing in assessing six putative causal links. Each row represents a two-variable or three-variable system where Y may or may not causally influence X. The leftmost column shows the equations and ground truth causality. The second column shows a sample of X and Y dynamics. Red and blue dots represent X and Y values, respectively; black lines connecting the dots serve as a visual aid. The third column shows visual continuity testing and causal interpretation. We write "likely" in the top row because the map from X delay space to Y appears to have some small bumps on the right side of the plot. The fourth column shows cross map skill testing (without the prediction lag test) and causal interpretation. Black dots show cross map skill. Open purple dots show the 5% chance cutoff at the maximum library size according to random phase surrogate data testing (see Methods), or are placed below the horizontal axis if the 5% chance cutoff is below the plot. In all systems Y appears to cause X according to cross map skill testing since cross map skill is positive, increases with training data size, and is significant according to the surrogate data test. The rightmost column shows the prediction lag test and causal interpretation.

Ye et al. [74] applied their test to 500 systems with the same form as in the third row of 19 but with

randomly chosen parameters. They found that within the parameter range they sampled, false negative errors as in Fig 19 do occur, but such errors are rare. However, we repeated a randomized numerical experiment from [74] for both the original parameter range of [74] (Fig 20B, "friendly" parameter regime) and a second parameter range of the same volume in parameter space (Fig 20B, "pathological" parameter regime). In this pathological parameter regime, false negative errors occur in the overwhelming majority of cases.

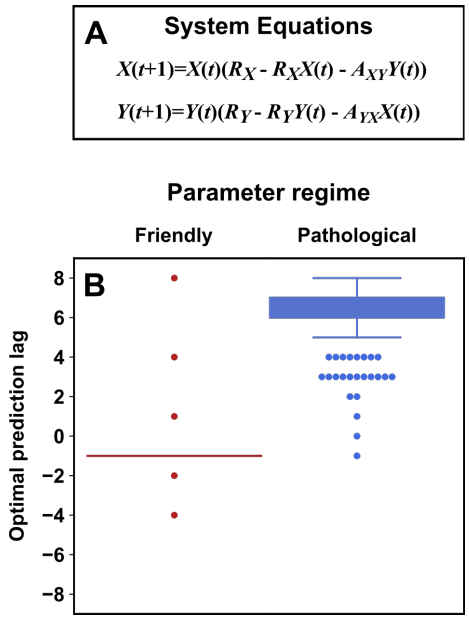

Figure 20: Parameters within a "pathological" regime almost always cause the prediction lag test of [74] to erroneously reject a true causal link. (**A**) System equations. For both "friendly" and "pathological" regimes, initial conditions $X(1)$ and $Y(1)$ were independently and randomly drawn from the range $0.01-0.99$, and R_X was randomly drawn from the range $3.7-3.9$. R_Y was randomly drawn from the range $3.7-3.9$ ("friendly") or $3.1-3.3$ ("pathological"). A_{XY} and A_{YX} were independently (and randomly) drawn from the range $0.05-0.1$ ("friendly") or $0.15-0.2$ ("pathological"). (**B**) Boxplots show the optimal prediction lag when using delay vectors made from X to predict values of Y in 500 systems with randomly selected parameters. In the ground truth model for this system, Y exerts a causal influence on X. In the "friendly" parameter regime explored in [74], the optimal prediction horizon is negative, correctly indicating that Y does indeed cause X. In the "pathological" regime, the optimal prediction horizon is positive, and so the rule of [74] would wrongly lead us to conclude that Y does not cause X. In the friendly regime the "box" is shown as a line because the vast majority of trials had the same optimal prediction lag of -1.

1.13 Intuition for random phase test

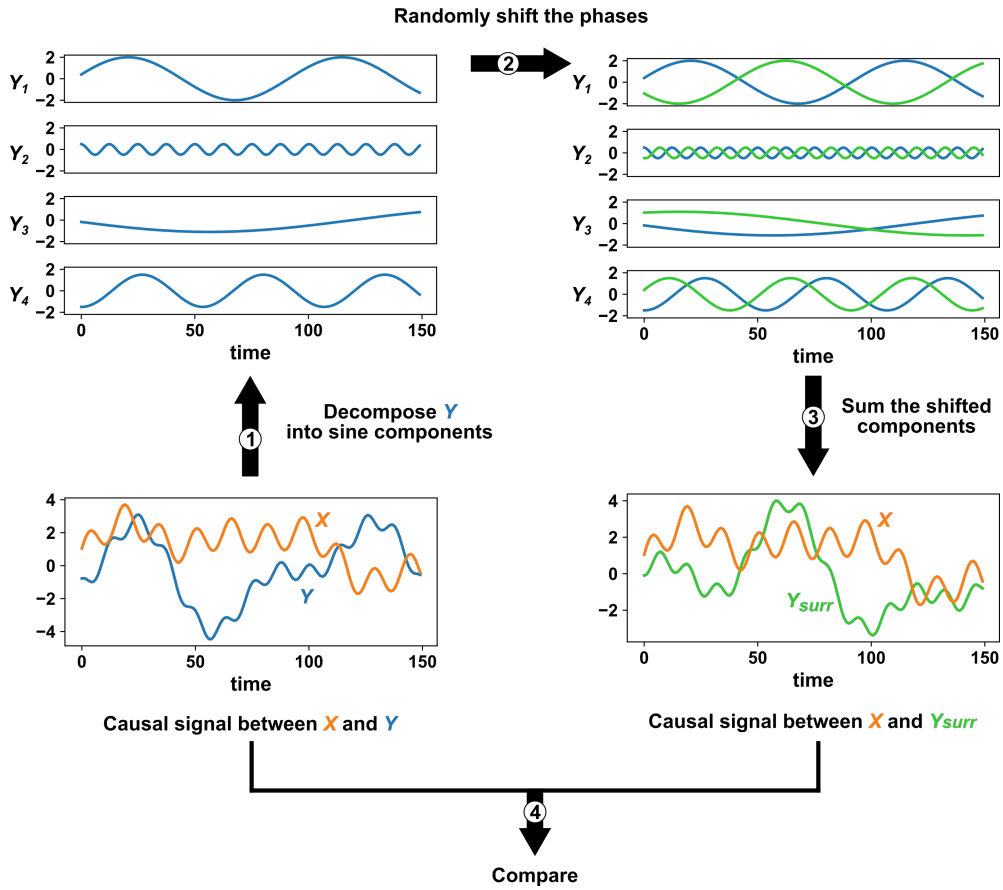

Figure 21: Intuition for random phase surrogate data methods. Given two time series X and Y (lower left), we can compute a statistic that may indicate a causal relationship ("causal signal"). Surrogate data methods test the significance of this causal signal by comparing it to causal signals generated when one of the time series (Y) is replaced by a similar time series (Y_{surr}) which is assumed to be causally independent of X. Random phase surrogate data methods generate Y_{surr} by representing Y as a sum of sine waves (upper left), randomly shifting the phases of the component sine waves (upper right green), and summing up the shifted sine waves (lower right green). A p-value is calculated as the fraction of (X, Y_{surr}) time series that produce a causal signal that is at least as strong as the causal signal from the original (X, Y) time series.