# Dynamical independence: Discovering emergent macroscopic processes in complex dynamical systems

L. Barnett[1,*] and A. K. Seth[1,2]

[1]*Sussex Centre for Consciousness Science, Department of Informatics, University of Sussex, Falmer, Brighton BN1 9QJ, United Kingdom*
[2]*Canadian Institute for Advanced Research, Program on Brain, Mind, and Consciousness, Toronto, Ontario M5G 1M1, Canada*

(Received 29 June 2022; revised 18 January 2023; accepted 15 June 2023; published 17 July 2023)

We introduce a notion of emergence for macroscopic variables associated with highly multivariate microscopic dynamical processes. *Dynamical independence* instantiates the intuition of an emergent macroscopic process as one possessing the characteristics of a dynamical system "in its own right," with its own dynamical laws distinct from those of the underlying microscopic dynamics. We quantify (departure from) dynamical independence by a transformation-invariant Shannon information-based measure of *dynamical dependence*. We emphasize the data-driven *discovery* of dynamically independent macroscopic variables, and introduce the idea of a multiscale "emergence portrait" for complex systems. We show how dynamical dependence may be computed explicitly for linear systems in both time and frequency domains, facilitating discovery of emergent phenomena across spatiotemporal scales, and outline application of the linear operationalization to inference of emergence portraits for neural systems from neurophysiological time-series data. We discuss dynamical independence for discrete- and continuous-time deterministic dynamics, with potential application to Hamiltonian mechanics and classical complex systems such as flocking and cellular automata.

DOI: 10.1103/PhysRevE.108.014304

## I. INTRODUCTION

When we observe a murmuration of starlings twisting, stretching, and wheeling, it is hard to escape the impression that we are witnessing an individuated dynamical entity quite distinct from the thousands of individual birds which we know to constitute the flock. The singular dynamics of the murmuration as a whole, it seems, "emerges" from the collective behavior of its constituents [1]. Analogously, the gliders and particles observed in some cellular automata appear to emerge as distinct and distinctive dynamical entities from the collective interactions between cells [2]. In both cases, these emergent phenomena reveal dynamical structure at coarser "macroscopic" scales than the "microscopic" scale of interaction between individual components of the system—structure which is not readily apparent from the microscopic perspective. Frequently, dynamical interactions at the microscopic level are reasonably simple and/or well-understood; yet an appropriate macroscopic perspective reveals dynamics that do not flow transparently from the micro-level interactions, and, furthermore, appear to be governed by laws quite distinct from the microscopic dynamics. For a class of complex systems, it seems, emergence proffers a window into inherent parsimonious structure at characteristic spatiotemporal scales.

In both of the above examples emergent structure "jumps out at us" visually. But this need not be the case, and in general may not be the case. For example, directly observing the population activity of large numbers of cortical neurons (e.g., via calcium or optogenetic imaging [3]) may not reveal any visually obvious macroscopic patterning (besides phenomena such as widespread synchrony), even though this activity underlies complex organism-level cognition and behavior. Even in flocking starlings, while distal visual observation clearly reveals emergent macroscopic structure, might there be additional emergent structure that would only be apparent from a very different—possibly nonvisual, or otherwise nonintuitive—perspective?

In this paper we address two key questions regarding emergent properties in complex dynamical systems:

$$\text{How may we \textit{characterize} those perspectives which reveal emergent dynamical structure?} \tag{1a}$$

$$\text{Knowing the microscopic dynamics, how may we \textit{find} these revealing perspectives?} \tag{1b}$$

By providing principled data-driven methods for identifying emergent structure across spatiotemporal scales, we hope to enable new insights into many complex systems from brains to ecologies to societies.

### A. Emergence and dynamical independence

Emergence is broadly understood as a gross (macroscopic) property of a system of interacting elements, which is not a property of the individual (microscopic) elements themselves. A distinction is commonly drawn between "strong"

[*] l.c.barnett@sussex.ac.uk

and "weak" emergence [4,5]. A strongly emergent macroscopic property (i) is in principle not deducible from its microscopic components, and (ii) has irreducible causal power over these components [6]. This flavour of emergence appears to reject mechanistic explanations altogether, and raises awkward metaphysical issues about causality, such as how to resolve competition between competing micro- and macrolevel "downward" causes [7]. By contrast, Bedau [4, p. 375] characterizes emergent phenomena as "somehow constituted by, and generated from, underlying processes," while at the same time "somehow autonomous from underlying processes" (*ibid.*). He goes on to define a process as weakly emergent iff it "can be derived from [micro-level dynamics and] external conditions but only by simulation." Weakly emergent properties are therefore ontologically reducible to their microscopic causes, though they remain epistemically opaque from these causes.

We propose a new notion and measure of emergence inspired by Bedau's formulation of weak emergence. *Dynamical independence* shares with weak emergence the aim to capture the sense in which a flock of starlings seems to have a "life of its own" distinct from the micro-level interactions among individual birds, even though there is no mystery that the flock is in fact constituted by the birds. Following Bedau's formulation, a dynamically independent macroscopic process is "ontologically reducible" to its microscopic causes, and downward (physical) causality is precluded. However, dynamically independent macroscopic processes may display varying degrees of "epistemic opacity" from their microscopic causes, loosening the constraint that (weak) emergence relations can only be understood through exhaustive simulation (cf. Seth [8]).

Dynamical independence is defined for macroscopic dynamical phenomena associated with a microscopic dynamical system—macroscopic variables—which *supervene* [9,10] on the microscopic. Here, supervenience of macro on micro is operationalized in a predictive sense: that macroscopic variables convey no information about their own evolution in time beyond that conveyed by the microscopic dynamics. The paradigmatic example of such macroscopic variables, and one which we mostly confine ourselves to in this study, is represented by the *coarse-graining* of the microscopic system by aggregation of microscopic components at some characteristic scale. Dynamical independence, like supervenience, is framed in predictive terms: a macroscopic variable is defined to be dynamically independent if, even while supervenient on the microscopic process, knowledge of the microscopic process adds nothing to prediction of the macroscopic process beyond the extent to which the macroscopic process already self-predicts (Sec. II C). In the language of Bertschinger *et al.* [11], the macroscopic process is "informationally closed" with respect to the microscopic process [12]—see also Sec. V A. Note that this does not imply that a dynamically independent process need self-predict *well*, if indeed at all; see Sec. V C.

To bolster intuition, consider a large group of particles, such as a galaxy of stars. The system state is described by the ensemble of position and momentum vectors of the individual stars in some inertial coordinate system, and the dynamics by (to a degree sufficient for illustrating the point) Newtonian gravitation. We may construct a low-dimensional coarse-grained macroscopic variable by taking the average position and total momentum (and, if we like, also the angular momentum and total energy) of the stars in the galaxy. Elementary physics tells us that this macroscopic variable in fact self-predicts perfectly without any knowledge of the detailed microscopic state; it has a "life of its own," perfectly understandable without recourse to the microscopic level. Yet an arbitrarily concocted coarse-graining—i.e., an arbitrary mapping of the microscopic state space to a lower-dimensional space—will almost certainly *not* have this property [13]: indeed, the vast majority of possible coarse-grainings do not define dynamically independent processes.

Dynamical independence is defined over the range of scales from microscopic, through mesoscopic to macroscopic. It is expressed, and quantified, solely in terms of Shannon (conditional) mutual information [14], and as such is fully *transformation-invariant*; that is, for a physical process it yields the same quantitative answers no matter how the process is measured. Under some circumstances, it may also be defined in the frequency domain, thus enabling analysis of emergence across temporal scales. It applies in principle to a broad range of dynamical systems, continuous and discrete in time and/or state, deterministic and stochastic. Examples of interest include neural systems, econometric processes, Hamiltonian dynamics, cellular automata, flocking, and evolutionary processes. Dynamical independence is also easily extended to accommodate a dynamically coupled *environment* (see Sec. V B); in this study, for clarity we consider only "closed" microscopic systems with no environmental coupling.

As previously indicated, our specific aims are (1a) to *quantify* the degree of dynamical independence of a macroscopic variable, and (1b) given the microlevel dynamics, to *discover* dynamically independent macroscopic variables. In the current study we address these aims primarily for stochastic processes in discrete time, and analyze in detail the important and nontrivial case of stationary linear systems.

The article is organized as follows: in Sec. II we set out our approach. We present the formal underpinnings of dynamical systems, macroscopic variables and coarse-graining, the information-theoretic operationalization of dynamical independence, its quantification and its properties. We declare an *ansatz* on a practical approach to our primary objectives (1). In Sec. III we specialize to linear stochastic systems in discrete time and analyze in depth how our *ansatz* may be realized for such systems; in particular, we detail how dynamically independent macroscopic variables for linear systems may be discovered via numerical optimization. In Sec. IV we discuss approaches to dynamical independence for deterministic and continuous-time systems. In Sec. V we summarize our findings, discuss related approaches in the literature, and examine some potential applications in neuroscience.

## II. APPROACH

### A. Dynamical systems

We describe a dynamical system by a sequence of variables $X_t$ taking values in some state space $\mathcal{X}$ at times indexed by $t$ (when considered as a whole, we sometimes drop the time index and write just $X$). In full generality, the state space might be discrete or real-valued, and possibly endowed with further structure (e.g., topological, metric, linear, etc.). Typically, a microscopic state $x \in \mathcal{X}$ will represent the high-dimensional ensemble state of a large number of atomic elements, e.g., birds in a flock, molecules in a gas, cells in a cellular automaton, neurons in a neural system, etc. The sequential time index may be discrete or continuous. The dynamical law, governing how $X_t$ evolves over time, specifies the system state at time $t$ given the history of states at prior times $t' < t$; this specification may be deterministic or probabilistic. In this study, we largely confine our attention to discrete-time stochastic processes, where the $X_t$, $t \in \mathbb{Z}$, are jointly distributed random variables. The distribution of $X_t$ is thus contingent on previously instantiated historical states; that is, on the set $x_t^- = \{x_{t'} : t' < t\}$ given that $X_{t'} = x_{t'}$ for $t' < t$ (throughout, we use a superscript dash to denote a set of prior states). In Sec. IV, we discuss extension of dynamical independence to deterministic and/or continuous-time systems.

### B. Macroscopic variables and coarse-graining

Given a microscopic dynamical system $X_t$, we associate an emergent phenomenon explicitly with some "macroscopic variable" associated with the microscopic system. Intuitively, we may think of a macroscopic variable as a gross perspective on the system, a "way of looking at it" (cf. Sec. I), or a particular mode of description of the system [15]. We operationalize this idea in terms of a process $Y_t$ that in some sense aggregates microscopic states in $\mathcal{X}$ into common states in a lower-dimension or cardinality state space $\mathcal{Y}$, with a consequent loss of information. (We do not rule out that aggregation may occur over time as well as state.) The dimension or cardinality of $\mathcal{Y}$ defines the *scale*, or "granularity," of the macroscopic variable.

The supervenience of macroscopic variables on the microscopic dynamics (Sec. I A) is operationalized in predictive, information-theoretic terms: We assume that a macroscopic variable conveys no information about its own future beyond that conveyed by the microscopic history. Explicitly, we demand the condition

$$\mathbf{I}(Y_t : Y_t^- \mid X_t^-) \equiv 0, \tag{2}$$

where $\mathbf{I}(\cdots)$ denotes Shannon (conditional) mutual information [14,16]. A canonical example of a macroscopic variable in the above sense is one of the form $Y_t = f(X_t)$, where $f : \mathcal{X} \to \mathcal{Y}$ is a deterministic, surjective mapping from the microscopic onto the lower dimensional/cardinality macroscopic state space (here aggregation is over states, but not over time [17]). The relation (2) is then trivially satisfied. We refer to this as *coarse-graining*, to be taken in the broad sense of dimensionality reduction. Coarse-graining partitions the state space, "lumping together" microstates in the preimage $f^{-1}(y) \subseteq \mathcal{X}$ of macrostates $y \in \mathcal{Y}$, with a concomitant loss of information: many microstates correspond to the same macrostate.

If the state space $\mathcal{X}$ is endowed with some structure (e.g., topological, metric, smooth, linear, etc.), then we generally restrict attention to structure-preserving mappings (morphisms). In particular, we restrict coarse-grainings $f$ to epimorphisms (surjective structure-preserving mappings) [18]. There is a natural equivalence relation among coarse-grainings: given $f : \mathcal{X} \to \mathcal{Y}$, $f' : \mathcal{X} \to \mathcal{Y}'$, we write

$$f' \sim f \iff \exists \text{ an isomorphism } \psi : \mathcal{Y} \to \mathcal{Y} \\ \text{such that } f' = \psi \circ f. \tag{3}$$

$f$ and $f'$ then lump together the same subsets of microstates. When we talk of a coarse-graining, we implicitly intend an equivalence class $\{f\}$ of mappings $\mathcal{X} \to \mathcal{Y}$ under the equivalence relation (3). In the remainder of this article we restrict attention to coarse-grained macroscopic variables.

### C. Dynamical independence

As we have noted, not *every* coarse-graining $f : \mathcal{X} \to \mathcal{Y}$ will yield a macroscopic variable $Y_t = f(X_t)$ which we would be inclined to describe as emergent (cf. the galaxy example in Sec. I A). Quite the contrary; for a complex microscopic system comprising many interacting components, we may expect that an arbitrary coarse-grained macroscopic variable will fail to behave as a dynamical entity in its own right, with its own distinctive law of evolution in time. When applied to the coarse-grained variable, the response to the question: *What will it do next?* will be: *Well, without knowing the* microscopic*history, we really can't be sure*; unsurprising, perhaps, as coarse-graining, by construction, loses information.

By contrast, for an *emergent* macroscopic variable, despite the loss of information incurred by coarse-graining, the macroscopic dynamics are parsimonious in the following specific sense: knowledge of the microscopic history adds nothing to the capacity of the macroscopic variable to self-predict. Dynamical independence formalizes this parsimony as follows:

$$\begin{aligned}&\text{Given jointly stochastic processes } (X, Y), Y \text{ is}\\&\text{dynamically independent of } X \text{ iff, conditional}\\&\text{on its own history, } Y \text{ is independent of the}\\&\text{history of } X.\end{aligned} \tag{4}$$

In information-theoretic terms, Eq. (4) holds (at time $t$) precisely when $\mathbf{I}(Y_t : X_t^- \mid Y_t^-)$ vanishes identically. We recognize this quantity as the *transfer entropy* (TE) [19–22] from $X$ to $Y$ at time $t$:

$$\mathbf{T}_t(X \to Y) = \mathbf{I}(Y_t : X_t^- \mid Y_t^-) \tag{5a}$$

$$= \mathbf{H}(Y_t \mid Y_t^-) - \mathbf{H}(Y_t \mid X_t^-, Y_t^-) \tag{5b}$$

$$= \mathbf{H}(Y_t \mid Y_t^-) - \mathbf{H}(Y_t \mid X_t^-); \tag{5c}$$

[the last equality follows from Eq. (2)] and state formally:

$$Y \text{ is dynamically independent of } X \text{ at time } t \\ \iff \mathbf{T}_t(X \to Y) \equiv 0. \tag{6}$$

The definition (6) establishes an information-theoretic *condition* [23] for dynamical independence of $Y$ with respect to $X$; we further propose the transfer entropy $\mathbf{T}_t(X \to Y)$ itself, the *dynamical dependence* of $Y$ on $X$, as a quantitative, nonnegative measure of the extent to which $Y$ *departs* from dynamical independence with respect to $X$ at time $t$ [24].

Dynamical (in)dependence is naturally interpreted in predictive terms: the *un* predictability of the process $Y$ at time $t$ given its own history is naturally quantified by the entropy rate $\mathbf{H}(Y_t \mid Y_t^-)$. We may contrast this with the unpredictability $\mathbf{H}(Y_t \mid X_t^-, Y_t^-)$ of $Y$ given not only its own history, but also the history of $X$. Thus, dynamical dependence quantifies the extent to which $X$ predicts $Y$ over-and-above the extent to which $Y$ already self-predicts.

The dynamical dependence $\mathbf{T}_t(X \to Y)$ will in general be time-varying, except when all processes are strongly stationary. For the remainder of this article we restrict ourselves to the stationary case and drop the time index subscript. We note at this stage that in the case that the processes $X, Y$ are deterministic, mutual information is not well-defined, and dynamical independence must be framed differently. We discuss deterministic systems in Sec. IV.

As a conditional Shannon mutual information [14], the transfer entropy $\mathbf{T}(X \to Y)$ is *nonparametric* in the sense that it is invariant with respect to reparametrization of source and target variables by isomorphisms of the respective state spaces [21]. Thus, if $\varphi$ is an isomorphism of $\mathcal{X}$ and $\psi$ an isomorphism of $\mathcal{Y}$, then

$$\mathbf{T}[\varphi(X) \to \psi(Y)] \equiv \mathbf{T}(X \to Y). \tag{7}$$

In particular, dynamical (in)dependence respects the equivalence relation (3) for coarse-grainings. This means that (at least for coarse-grained macroscopic variables) transfer entropy from macro to micro vanishes trivially; i.e., $\mathbf{T}(Y \to X) \equiv 0$, which we may interpret as the nonexistence of "downward causation" in this formalism.

To guarantee *transitivity* of dynamical independence (see below), we introduce a mild technical restriction on admissible coarse-grainings to those $f : \mathcal{X} \to \mathcal{Y}$ with the following property:

$$\begin{aligned}&\exists \text{ an epimorphism } u : \mathcal{X} \to \mathcal{U} \text{ such that } \varphi \\ &\quad = f \times u : \mathcal{X} \to \mathcal{Y} \times \mathcal{U} \text{ is an isomorphism.}\end{aligned} \tag{8}$$

Intuitively, this means that there is a complementary mapping $u$ which, along with $f$ itself, defines a nonsingular transformation of the system. For example, if $f$ is a projection of the real Euclidean space $\mathbb{R}^n$ onto the first $m < n$ coordinates, $u$ could be taken as the complementary projection of $\mathbb{R}^n$ onto the remaining $n - m$ coordinates (cf. Sec. III). Trivially, Eq. (8) respects the equivalence relation (3). The restriction holds universally for some important classes of structured dynamical systems, e.g., the linear systems analyzed in Sec. III, and also in general for discrete-state systems; otherwise, it might be relaxed to obtain at least locally in the state space [25]. We assume Eq. (8) for all coarse-grainings from now on.

Given property (8), we may apply the transformation $\varphi = f \times u$ and exploit the dynamical dependence invariance (7) to obtain an equivalent system $X \sim (Y, U)$, $U = u(X)$ in which the coarse-graining $Y$ becomes a projection of $\mathcal{X} = \mathcal{Y} \times \mathcal{U}$ onto $\mathcal{Y}$, and dynamical dependence is given by

$$\mathbf{T}(X \to Y) = \mathbf{T}(Y, U \to Y) = \mathbf{T}(U \to Y). \tag{9}$$

Assuming Eq. (8) for all coarse-grainings and using Eq. (9) we may show that dynamical independence is transitive:

$$\begin{aligned}&\text{If } Y = f(X) \text{ is dynamically independent of } X \\ &\text{and } Z = g(Y) \text{ is dynamically independent of } Y, \\ &\text{then } Z = (g \circ f)(X) \text{ is dynamically independent of } X.\end{aligned} \tag{10}$$

We provide a formal proof in Appendix A. We have thus a partial ordering on the set of coarse-grained dynamically independent macroscopic variables, under which they may potentially be hierarchically nested at increasingly coarse scales.

At this point, we note the intimate relationship between transfer entropy and *Wiener-Granger causality* (GC) [26–29]. Specifically, for discrete-time, continuous-state processes, GC and TE are equivalent under Gaussian assumptions [30] (these may be relaxed to include a somewhat broader class of distributions; see Ref. [31]). More generally, for Markovian processes GC and TE are asymptotically equivalent under an interpretation of Granger causality as a log-likelihood ratio [32]. For a broad class of processes, GC has distinct advantages over TE in terms of analytic tractability, sample estimation and statistical inference, in both parametric [33,34] and nonparametric [35] scenarios. In Appendix B we provide a concise recap of (unconditional) Granger causality following the classical formulation of Geweke [29].

Systems featuring emergent properties are typically large ensembles of dynamically interacting elements; that is, system states $x \in \mathcal{X}$ are of the form $(x_1, \dots, x_n) \in \mathcal{X}_1 \times \dots \times \mathcal{X}_n$ with $x_i$ the state of the $i$th element. Dynamical independence for such systems may be interpreted in terms of the *causal graph* of the system [33,36] (here, "causality" is in the Wiener-Granger predictive sense; see above). This means that the causal graph encapsulates information transfer between system elements, which facilitates the construction of systems with prescribed dynamical-independence structure. Given a coarse-graining $y_k = f_k(x_1, \dots, x_n)$, $k = 1, \dots, m$ at scale $m$, using Eq. (8) it is not hard to show that we may always transform the system so that $y_k = x_k$ for $k = 1, \dots, m$; that is, under a change of coordinates the coarse-graining becomes a projection onto the subspace defined by the first $m$ dimensions of the microscopic state space. The dynamical dependence is then given by

$$\mathbf{T}(X \to Y) = \mathbf{T}(X_{m+1}, \dots, X_n \to X_1, \dots, X_m), \tag{11}$$

and we may show [37] that, under such a transformation,

$$\begin{aligned}&\mathbf{T}(X \to Y) = 0 \iff \mathbf{G}_{ij}(X) = 0 \\ &\text{for } i = 1, \dots, m, \quad j = m+1, \dots, n,\end{aligned} \tag{12}$$

where

$$\begin{aligned}&\mathbf{G}_{ij}(X) = \mathbf{T}(X_j \to X_i \mid X_{[ij]}), \\ &\qquad i, j = 1, \dots, n, \quad i \neq j,\end{aligned} \tag{13}$$

is the causal graph of the system $X$ (here the subscript "$[ij]$" denotes omission of the $i$ and $j$ components of $X$). According to Eq. (12) we may characterize dynamically independent macroscopic variables for ensemble systems as those coarse-grainings which are transformable into projections onto a subgraph of the causal graph with no incoming information transfer from the rest of the system. However, given two or more dynamically independent macroscopic variables (at the same or different scales), in general we cannot expect to find a transformation under which all of those variables *simultaneously* become projections onto causal subgraphs. Nonetheless, Eq. (12) is useful for constructing dynamical systems with prespecified dynamically independent macroscopic variables.

For many complex dynamical systems, there will be no fully dynamically independent macroscopic variables at some particular (or perhaps at any) scale; i.e., no macroscopic variables for which Eq. (6) holds *exactly* [38]. There may, however, be macroscopic variables for which the dynamical dependence (5) is small; in an empirical scenario, for instance, "small" might be taken to mean statistically insignificant at some given significance level, or in a Bayesian sense of evidence for no difference from zero. We take the view that even "near-dynamical independence" yields useful structural insights into emergence, and adopt the *ansatz*:

The *maximally* dynamically independent
macroscopic variables at a given scale
(i.e., those which minimize dynamical dependence)
characterize emergence at that scale. (14a)
The collection of maximally dynamically independent
macroscopic variables at *all* scales, along with
their degree of dynamical dependence, affords a multiscale
portrait of the emergence structure of the system. (14b)

The question—central to this study—now arises: given a complex dynamical system (specified, perhaps, as a model derived from empirical data), how, in practice, are we to *find* the maximally dynamically independent macroscopic variables across scales? On the face of it, this would seem to be a daunting task: those macroscopic variables which maximize dynamical independence will be needles in the haystack of all possible macroscopic variables at a given scale—a search-space which, for all but the most trivial systems, will generally be vast, and itself complex. We proceed, though, to analysis of a useful (and nontrivial) class of systems for which, perhaps surprisingly, this task turns out to be tractable.

## III. LINEAR SYSTEMS

In this section we consider linear discrete-time continuous-state systems, and later specialize to linear state-space (SS) systems. Linear models are commonly deployed in a variety of real-world scenarios, especially for econometric and neuroscientific time-series analysis. Reasons for their popularity include parsimony (linear models will frequently have fewer parameters than alternative nonlinear models [39]), simplicity of estimation, and mathematical tractability (see below for further discussion). Our aim here is to describe the space of linear macroscopic variables, calculate dynamical dependence for such variables and, in line with our *ansatz*, outline practical methods for minimization of dynamical dependence over the macroscopic variable space.

Our starting point is that the microscopic system may be modeled as a wide-sense stationary, purely nondeterministic, stable and invertible, zero-mean vector [40] stochastic process $X_t = [X_{1t}\ X_{2t}\ \dots\ X_{nt}]^\mathsf{T}$, $t \in \mathbb{Z}$, defined on the vector space $\mathbb{R}^n$. Wide-sense stationarity guarantees that the process has a unique stable and causal vector moving-average (VMA) representation [41]

$$X_t = \boldsymbol{\varepsilon}_t + \sum_{k=1}^{\infty} B_k \boldsymbol{\varepsilon}_{t-k} \quad \text{or} \quad X_t = H(z) \cdot \boldsymbol{\varepsilon}_t, \tag{15}$$

where $\boldsymbol{\varepsilon}_t$ is a white-noise innovations process with positive-definite covariance matrix $\Sigma = \mathbf{E}[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^\mathsf{T}]$, and

$$H(z) = I + \sum_{k=1}^{\infty} B_k z^k \tag{16}$$

the transfer function, with $z$ the back-shift operator (in the frequency domain, $z = e^{-i\omega}$, where $\omega$ is angular frequency in radians). The VMA coefficient matrices $B_k$ are square-summable, and all roots of $H(z)$ lie strictly outside the unit disk $|z| \leqslant 1$ in the complex plane. The cross-power spectral density (CPSD) matrix for the process exists at almost all frequencies, and is given by [42]

$$S(\omega) = H(\omega)\Sigma H^*(\omega), \tag{17}$$

where "*" denotes conjugate transpose [43]. Conversely, if a CPSD $S(\omega)$ is given, then under certain regularity conditions [44, Def. 1.1] unique $H(z)$ and $\Sigma$ exist that satisfy Eq. (17), and there are stable algorithms by which this *spectral factorization* may be computed numerically [35,44]. A standard result [45, Theorem 4.2] states that

$$\frac{1}{2\pi} \int_0^{2\pi} \log |S(\omega)|\, d\omega = \log |\Sigma|. \tag{18}$$

Invertibility requires that the process $X_t$ also have a stable vector-autoregressive (VAR) representation

$$X_t = \sum_{k=1}^{\infty} A_k X_{t-k} + \boldsymbol{\varepsilon}_t \quad \text{or} \quad H(z)^{-1} \cdot X_t = \boldsymbol{\varepsilon}_t. \tag{19}$$

The VAR coefficients $A_k$ are again square-summable, and all poles (as well as roots) of $H(z)$ lie strictly outside the unit disk; stability requires that the *spectral radius* $\rho = \max\{|z| : |H(z^{-1})| = 0\}$ of the process is <1. Following Geweke [29, Eq. (2.4)] we assume the slightly stronger condition that $S(\omega)$ be uniformly bounded away from zero almost everywhere. This guarantees invertibility and, importantly, that any *subprocess* of $X_t$ is also invertible.

Note that at this stage we do not assume that the innovations $\boldsymbol{\varepsilon}_t$ are (multivariate) Gaussian. Note too, that even though we describe the system as "linear," *this does not necessarily exclude processes with nonlinear generative mechanisms*—we just require that the conditions listed above are met. Wold's Decomposition Theorem [46] guarantees a VMA form (15) provided that the process is wide-sense stationary and purely nondeterministic; if in addition the Geweke spectral condition [29, Eq. (2.4)] holds, the VAR form (19) also exists, and all our conditions are satisfied. Thus, our analysis here also covers a large class of stationary "nonlinear" systems, with the caveat that for a given nonlinear generative model, the VMA/VAR representations will generally be infinite-order, and as such may not represent parsimonious models for the system.

Since we are in the linear domain, we restrict ourselves to *linear* coarse-grained macroscopic variables [47] (cf. Sec. II B). A surjective linear mapping $L : \mathbb{R}^n \to \mathbb{R}^m$, $0 < m < n$, corresponds to a full-rank $m \times n$ matrix, and the coarse-graining equivalence relation (3) identifies $L, L'$ iff there is a nonsingular linear transformation $\Psi$ of $\mathbb{R}^m$ such that $L' = \Psi L$. Note that since $L$ is full-rank, $Y_t = LX_t$ is purely nondeterministic, and by the Geweke spectral condition it is also invertible. Considering the rows of $L$ as basis vectors for an $m$-dimensional linear subspace (hyperplane) of $\mathbb{R}^n$, a linear transformation simply specifies a change of basis for the subspace. Thus, we may identify the set of linear coarse-grainings with the *Grassmannian manifold* $\mathcal{G}_m(n)$ of $m$-dimensional hyperplanes in $\mathbb{R}^n$.

The Grassmannian [48] is a compact smooth manifold of dimension $m(n-m)$. It is also a nonsingular *algebraic variety* (the set of solutions of a system of polynomial equations over the real numbers), a *homogeneous space* (it "looks the same at any point"), and an *isotropic space* (it "looks the same in all directions"). We have $\mathcal{G}_m(n) = \mathcal{O}(n)/[\mathcal{O}(m) \times \mathcal{O}(n-m)]$, where $\mathcal{O}(n)$ is the Lie group of real orthogonal matrices; this quotient structure induces a canonical Riemannian metric on $\mathcal{G}_m(n)$ [49], and there is a natural definition of *principal angles* between linear subspaces of Euclidean spaces, via which various invariant (noncanonical) metrics may be defined [50].

By transformation-invariance of dynamical dependence, we may assume without loss of generality that the row-vectors of $L$ form an orthonormal basis; i.e.,

$$LL^{\mathsf{T}} = I. \tag{20}$$

The manifold of linear mappings satisfying (20) is known as the *Stiefel manifold* $\mathcal{V}_m(n) \equiv \mathcal{O}(n)/\mathcal{O}(n-m)$, which, like the Grassmannian is a compact, homogeneous and isotropic algebraic variety, with dimension $nm - \frac{1}{2}m(m+1)$. Under the Euclidean inner product, every $m$-dimensional subspace of $\mathbb{R}^n$ has a unique orthogonal complement of dimension $n-m$ [51], which establishes an isometry of $\mathcal{G}_m(n)$ with $\mathcal{G}_{n-m}(n)$; for instance, in $\mathbb{R}^3$, every line through the origin has a unique orthogonal plane through the origin, and viceversa. The condition (8) is thus automatically satisfied for linear coarse-grainings; specifically, given $L$ satisfying (20), we may always find a surjective linear mapping $M : \mathbb{R}^n \to \mathbb{R}^{n-m}$ where the row-vectors of the $(n-m) \times n$ matrix $M$ form an orthonormal basis for the orthogonal complement of the subspace spanned by the row-vectors of $L$. The transformation $\Phi : \mathbb{R}^n \to \mathbb{R}^n$

$$\Phi = \begin{bmatrix} L \\ M \end{bmatrix} \tag{21}$$

is then nonsingular and orthonormal; i.e., $\Phi\Phi^{\mathsf{T}} = \Phi^{\mathsf{T}}\Phi = I$.

Given a linear mapping $L$, our task is to calculate the dynamical dependence $\mathbf{T}(X \to Y)$ for the coarse-grained macroscopic variable $Y_t = LX_t$. In the context of linear systems, it is convenient to switch from transfer entropy to Granger causality (Sec. II C and Appendix B), and we write the GC dynamical dependence as $\mathbf{F}(X \to Y)$. In case the innovations $\boldsymbol{\varepsilon}_t$ in Eqs. (15) and (19) are multivariate-normal, as noted previously the equivalence of TE and GC is exact [30]; else we may either consider the GC approach as an approximation to "actual" dynamical dependence, or, if we wish, consider dynamical dependence framed in terms of GC rather than TE as a linear prediction-based measure in its own right. We note that key properties of dynamical (in)dependence including transformation invariance (7), the existence of complementary mappings (8) [cf. Eq. (21)], transitivity (10), and relationship to the (Granger-)causal graph (12) carry over straightforwardly to the GC case.

### A. Dynamical dependence for linear systems

With $Y_t = LX_t$ and $\Phi$ as in Eq. (21), by transformation invariance we have $\mathbf{F}(X \to Y) = \mathbf{F}(\Phi X \to Y) = \mathbf{F}(MX \to LX)$, where the last equality holds since, given $LX_t^-$, the all-variable history $\Phi X_t^-$ yields no additional predictive information about $LX_t$ beyond that contained in $MX_t^-$. Now from Eq. (19) the $LX$ component of the innovations for the full process $\Phi X_t$ is $L\boldsymbol{\varepsilon}_t$, with covariance matrix $L\Sigma L^{\mathsf{T}}$. Thus, the dynamical dependence is (Appendix B)

$$\mathbf{F}(X \to Y) = \log \frac{|\Sigma^{\mathsf{R}}|}{|L\Sigma L^{\mathsf{T}}|}, \tag{22}$$

where $\Sigma^{\mathsf{R}}$ is the innovations covariance matrix for the reduced process $LX_t$. Again using transformation invariance, we note that transformation of $\mathbb{R}^n$ by the inverse of the left Cholesky factor of $\Sigma$ decorrelates and normalizes the innovations $\boldsymbol{\varepsilon}_t$ of $X_t$ so that $\Sigma = I$, and from Eq. (20) we have $L\Sigma L^{\mathsf{T}} = I$ in the transformed system. Thus, from now on, without loss of generality we assume $\Sigma = I$, and the dynamical dependence becomes simply

$$\mathbf{F}(X \to Y) = \log |\Sigma^{\mathsf{R}}|. \tag{23}$$

It remains to calculate $\Sigma^{\mathsf{R}}$.

Below (Secs. III B and III C) we shall show how the general problem of calculation of $\Sigma^{\mathsf{R}}$ may be achieved in a parametric scenario if $X_t$ satisfies a state-space or finite-order VAR model, via solution of a discrete algebraic Riccati equation (DARE). An alternative nonparametric approach utilizes the CPSD (17). From Eq. (15) it follows that the transfer function for the transformed process $\Phi X_t$ is $\Phi H(z)\Phi^{\mathsf{T}}$ and from Eq. (17) that the CPSD of $\Phi X_t$ is $\Phi S(\omega)\Phi^{\mathsf{T}}$. The CPSD of the macroscopic process $Y_t = LX_t$ is thus $LS(\omega)L^{\mathsf{T}}$, and From Eqs. (18) and (23) we have immediately

$$\mathbf{F}(X \to Y) = \frac{1}{2\pi} \int_0^{2\pi} \log |LS(\omega)L^{\mathsf{T}}| \, d\omega. \tag{24}$$

Empirically, the CPSD may be estimated by standard methods (such as averaged periodogram, multi-taper or wavelets), or else derived analytically from an estimated parametric model.

Like $\mathbf{F}(X \to Y)$, the spectral GC $\mathbf{f}(X \to Y; \omega)$ (B3) is invariant under nonsingular linear transformation of source or target variable at all frequencies [52], and furnishes a frequency decomposition for time-domain GC (B4). To calculate $\mathbf{f}(X \to Y; \omega)$, we again apply the orthonormal transformation (21) and calculate $\mathbf{f}(X \to Y; \omega) = \mathbf{f}(MX \to LX; \omega)$. It is then not hard to show that, under the assumed conditions $\Sigma = I$ and $LL^{\mathsf{T}} = I$, and noting that $L^{\mathsf{T}}L + M^{\mathsf{T}}M = I$, we have the *spectral dynamical dependence*

$$\mathbf{f}(X \to Y; \omega) = \log \frac{|LH(\omega)H^*(\omega)L^{\mathsf{T}}|}{|LH(\omega)L^{\mathsf{T}}LH^*(\omega)L^{\mathsf{T}}|}. \tag{25}$$

As per Eq. (B5), we may define the *band-limited dynamical dependence* as

$$\mathsf{F}(X \to Y; \omega_1, \omega_2) = \frac{1}{\omega_2 - \omega_1} \int_{\omega_1}^{\omega_2} \mathsf{f}(X \to Y; \omega)\, d\omega, \quad (26)$$

and we have the spectral decomposition for dynamical dependence (B4)

$$\mathsf{F}(X \to Y) = \frac{1}{2\pi} \int_0^{2\pi} \mathsf{f}(X \to Y; \omega)\, d\omega. \quad (27)$$

Significantly, band-limited dynamical dependence facilitate analysis of emergence across *temporal* as well as spatial scales.

### B. Linear state-space systems

We now specialize to the class of linear state-space systems, under the restrictions listed at the beginning of Sec. III. Here $X_t$ may be represented by a model of the form

$$W_{t+1} = AW_t + U_t \quad \text{state-transition equation}, \quad (28a)$$

$$X_t = CW_t + V_t \quad \text{observation equation}, \quad (28b)$$

where the state process $W_t = [W_{1t}\ W_{2t}\ \ldots\ W_{rt}]^\mathsf{T}$, $t \in \mathbb{Z}$, is defined on $\mathbb{R}^r$, $U_t, V_t$ are zero-mean jointly multivariate white noises, $C$ is the observation matrix, and $A$ the state transition matrix. Note the specialized use of the term "state space" in the linear systems vocabulary: The state variable $W_t$ is to be considered a notional unobserved process, or simply as a mathematical construct for expressing the dynamics of the observed process $X_t$, which as before stands as the microscopic variable.

The parameters of the model (26) are $(A, C, Q, R, S)$, where

$$\begin{bmatrix} Q & S \\ S^\mathsf{T} & R \end{bmatrix} \equiv \mathbf{E}\left[\begin{bmatrix} U_t \\ V_t \end{bmatrix} \begin{bmatrix} U_t^\mathsf{T} & V_t^\mathsf{T} \end{bmatrix}\right] \quad (29)$$

is the joint noise covariance matrix (the purely nondeterministic assumption implies that $R$ is positive-definite). Stationarity requires that the transition equation (28a) satisfy the stability condition $\max\{|\lambda| : \lambda \in \operatorname{eig}(A)\} < 1$. A process $X_t$ satisfying a stable, invertible SS model (26) also satisfies a stable, invertible vector-autoregressive moving-average (VARMA) model; conversely, any stable, invertible VARMA process satisfies a stable, invertible SS model of the form (26) [53].

To facilitate calculation of dynamical dependence (see below), it is useful to transform the SS model (26) to *innovations form*:

$$Z_{t+1} = AZ_t + K\boldsymbol{\varepsilon}_t, \quad (30a)$$

$$X_t = CZ_t + \quad \boldsymbol{\varepsilon}_t, \quad (30b)$$

with new state variable $Z_t = \mathbf{E}[W_t \,|\, X_t^-]$, white-noise innovations process $\boldsymbol{\varepsilon}_t = X_t - \mathbf{E}[X_t \,|\, X_t^-]$ with covariance matrix $\Sigma = \mathbf{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\mathsf{T}]$, and Kalman gain matrix $K$ [note that the $\boldsymbol{\varepsilon}_t$ are the same innovations as in the VMA and VAR representations (15), (19)]. The transfer function and its inverse for the innovations-form state-space (ISS) model (28) are given by [34]

$$H(z) = I + C(1 - Az)^{-1}Kz, \quad (31a)$$

$$H(z)^{-1} = I - C(1 - Bz)^{-1}Kz, \quad (31b)$$

respectively, where $B = A - KC$. The invertibility condition is thus $\max\{|\lambda| : \lambda \in \operatorname{eig}(B)\} < 1$. Eq. (31a) facilitates explicit calculation of the CPSD (17).

A general-form SS (26) may be converted to an ISS (28) by solving the associated DARE [53,54]:

$$P = APA^\mathsf{T} + Q - K\Sigma K^\mathsf{T}, \quad (32a)$$

$$\Sigma = CPC^\mathsf{T} + R, \quad (32b)$$

$$K = (APC^\mathsf{T} + S)\Sigma^{-1}, \quad (32c)$$

which under our assumptions has a unique stabilising solution for $P$. Stable computational algorithms exist for solving DAREs numerically [54], with efficient implementations available in popular software systems such as Matlab and Python.

From Eq. (26) it is clear that a macroscopic process $Y_t = LX_t$ will be of the same form; that is, the class of state-space systems is closed under full-rank linear mappings. Given an ISS model (28), we see that $Y_t$ satisfies the SS model (no longer in innovations form),

$$Z_{t+1} = AZ_t \quad + K\boldsymbol{\varepsilon}_t, \quad (33a)$$

$$Y_t = LCZ_t + L\boldsymbol{\varepsilon}_t. \quad (33b)$$

Again assuming without loss of generality that $\Sigma = I$, and using $LL^\mathsf{T} = I$, the SS model (31) has parameters $(A, LC, KK^\mathsf{T}, I, KL^\mathsf{T})$. To bring it into ISS form we solve the corresponding "reduced" DARE

$$P^\mathsf{R} = AP^\mathsf{R}A^\mathsf{T} + KK^\mathsf{T} - K^\mathsf{R}\Sigma^\mathsf{R}K^{\mathsf{RT}}, \quad (34a)$$

$$\Sigma^\mathsf{R} = LCP^\mathsf{R}C^\mathsf{T}L^\mathsf{T} + I, \quad (34b)$$

$$K^\mathsf{R} = (AP^\mathsf{R}C^\mathsf{T} + K)L^\mathsf{T}[\Sigma^\mathsf{R}]^{-1}, \quad (34c)$$

and Eq. (34b) furnishes the reduced innovations covariance matrix $\Sigma^\mathsf{R}$ required for calculation of the dynamical dependence (23).

The spectral dynamical dependence (25) may be calculated directly using the expression (31a) for the transfer function $H(z)$ without the need to solve a DARE.

### C. Finite-order VAR systems

Finite-order autoregressive systems are an important special case of state-space systems. We suppose that a VAR($p$) model, $p < \infty$, is given by

$$X_t = \sum_{k=1}^{p} A_k X_{t-k} + \boldsymbol{\varepsilon}_t. \quad (35)$$

The model parameters are the $n \times n$ coefficients matrices $A_1, \ldots, A_p$ and innovations covariance matrix $\Sigma = \mathbf{E}[\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t^\mathsf{T}]$. The transfer function is given by

$$H(z) = \left(I - \sum_{k=1}^{p} A_k z^k\right)^{-1}, \quad (36)$$

from which the CPSD (17) may be calculated. We may specify an equivalent ISS model (28) by taking $A$ to be the $pn \times pn$ "companion matrix" [53]

$$A = \begin{bmatrix} A_1 & A_2 & \dots & A_{p-1} & A_p \\ I & 0 & \dots & 0 & 0 \\ 0 & I & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & I & 0 \end{bmatrix}, \tag{37}$$

$C$ the $n \times pn$ matrix

$$C = [A_1 \quad A_2 \quad \dots \quad A_p], \tag{38}$$

and $K$ the $pn \times n$ matrix

$$K = [I \quad 0 \quad \dots \quad 0]^{\mathsf{T}}. \tag{39}$$

The VAR($p$) is stable iff the companion matrix (37) is stable.

For calculation of time-domain dynamical dependence, the reduced covariance matrix $\Sigma^{\mathsf{R}}$ may be obtained from solution of the DARE (32) corresponding to the derived ISS model as in Sec. III B [55]. Spectral dynamical dependence may again be calculated from the transfer function (25) using the transfer function (36).

### D. Perfect dynamical independence

In the parametric scenario of state-space or VAR models, we now examine generic conditions under which perfectly dynamically independent macroscopic variable can be expected to exist, where by "generic" we mean that the conditions hold everywhere except on a measure-zero subset of the model parameter space.

For an ISS model, applying the transformation (21) and using the condition given in Barnett and Seth [34, Eq. (17) and $ff$] it is easily found that

$$\mathbf{F}(X \to Y) \equiv 0 \iff LCA^kKM^{\mathsf{T}} \equiv 0$$
$$\text{for} \quad k = 0, \dots, r-1, \tag{40}$$

where $r$ is the dimension of the state space. Equation (40) constitutes $rm(n-m)$ multivariate-quadratic equations for the $m(n-m)$ free variables which parametrize the Grassmannian $\mathcal{G}_m(n)$. For $r = 1$, we would thus expect solutions yielding dynamically independent $Y_t = LX_t$ at all scales $0 < m < n$; however, as the equations are quadratic, some of these solutions may not be real. For $r > 1$, except on a measure-zero subset of the $(A, C, K)$ ISS parameter space, there will be solutions if $LC \equiv 0$ or $KM^{\mathsf{T}} \equiv 0$ (or both). The former comprises $rm$ linear equations, and the latter $r(n-m)$ equations, for the $m(n-m)$ Grassmannian parameters. Therefore, we expect generic solutions to Eq. (40) if $r < n$ and either $m \leqslant n-r$ or $m \geqslant r$ (or $r \leqslant m \leqslant n-r$, in which case $2r \leqslant n$ is required). Generically, for $r \geqslant n$ there will be no perfectly dynamically independent macroscopic variables. We note that $r < n$ corresponds to "simple" models with few spectral peaks; nonetheless, anecdotally it is not uncommon to estimate parsimonious model orders $< n$ for highly multivariate data, especially for limited time-series data.

In the generic VAR($p$) case, it is again easy to establish that the condition for vanishing dynamical dependence is

$$\mathbf{F}(X \to Y) \equiv 0 \iff LA_kM^{\mathsf{T}} \equiv 0$$
$$\text{for} \quad k = 1, \dots, p, \tag{41}$$

which constitutes $pm(n-m)$ multivariate-quadratic equations for the $m(n-m)$ Grassmannian parameters. Generically, for $p = 1$ we should again expect to find dynamically independent macroscopic variables at all scales $0 < m < n$, while for $p > 1$ we do not expect to find any dynamically independent macroscopic variables, except on a measure-zero subset of the VAR($p$) parameter space $(A_1, \dots, A_p)$.

Regarding spectral dynamical independence, we note that $\mathbf{f}(X \to Y; \omega)$ is an *analytic* function of $\omega$. Thus, by a standard property of analytic functions, if band-limited dynamical dependence (26) vanishes on any particular finite interval $[\omega_1, \omega_2]$ then it is zero everywhere, so that by Eq. (27) the time-domain dynamical dependence must also vanish identically.

### E. Statistical inference

Given empirical time-series data, a state-space or VAR model may be estimated via standard maximum-likelihood techniques, such as ordinary least squares [OLS; 56] for VAR estimation, or a subspace method [57] for state-space estimation. The dynamical dependence, as a Granger causality sample statistic, may then in principle be tested for significance at some prespecified level, and dynamical independence of a coarse-graining $Y_t = LX_t$ inferred by failure to reject the appropriate null hypothesis of zero dynamical dependence (40) or (41).

In the state-space case (Sec. III B), the asymptotic null sampling distribution of the maximum-likelihood estimator of the dynamical dependence (23) remains unknown, and surrogate/resampling methods are appropriate. For VAR modeling the maximum-likelihood estimator for Eq. (23) is a "single-regression" Granger causality estimator, for which an asymptotic generalized $\chi^2$ sampling distribution, in both time and (band-limited) frequency domains, has recently been obtained by Gutknecht and Barnett [58]. Alternatively, a likelihood-ratio, Wald or $F$-test [59] might be performed for the null hypothesis (41).

For the nonparametric case (24), the CPSD may be estimated at a given frequency resolution $d\omega$ by a standard method, and the integral approximated by numerical quadrature. Here again, the sampling distribution of the resulting dynamical dependence estimator is (to the authors' knowledge) unknown, and surrogate/resampling methods must be used.

### F. Maximizing dynamical independence

At this stage, we have effectively answered question (1a) of Sec. I for linear systems; that is, we have *characterized* dynamically independent macroscopic variables for linear systems. We now address question (1b): given a (microscopic) linear dynamical system, how shall we *find* dynamically independent—or, following our *ansatz* (14), at least *maximally* dynamically independent—macroscopic variables?

Whether or not perfectly dynamically independent macroscopic variables exist at any given scale, we thus seek to minimise the dynamical dependence $\mathbf{F}(X \to Y)$ over the Grassmannian manifold of linear coarse-grainings (i.e., over $L$ for $Y_t = LX_t$). The band-limited dynamical dependence (26) may also in principle be minimized at a given scale to yield maximally dynamically independent coarse-grainings associated with a given frequency range at that scale; we leave this for future research.

Solving the minimization of dynamical dependence over the Grassmannian analytically appears, at this stage, intractable in both nonparametric and parametric (state-space or VAR) cases; we thus proceed to numerical optimization. Optimization of an objective function—in our case the dynamical dependence—over the (non-Euclidean) Grassmannian manifold is nontrivial, and several approaches have been proposed in the literature, involving various parametrizations of the Grassmannian [60]. In our case, the natural parametrization at scale $m$ is via the corresponding Stiefel manifold, i.e., over $m \times n$ matrices $L$ with $LL^\mathsf{T} = I$; we largely follow the treatment in Edelman *et al.* [49], and briefly discuss alternative parametrizations in Appendix C. Note that for fixed macroscopic dimension $m$ the search space scales quadratically with microscopic dimension $n$. We have found that the number of local suboptima increases with $n$, although the rate of increase is unclear.

The Stiefel parametrization is redundant; an element of $\mathcal{G}_m(n)$ is represented by an equivalence class of matrices $\{L\}$. This redundancy would seem to rule out population-based optimization methods such as cross-entropy optimization [61]. Our exploratory investigations indeed indicated that such algorithms generally fail to converge, apparently because the population diffuses along equicost surfaces in $\mathcal{V}_m(n)$. Simplex methods [62], which are generally better at locating global optima, also fared poorly, although the reasons are less clear. We have had success, however, with gradient descent methods—see below.

Regarding computational efficiency, in the parametric scenario we may apply a useful preoptimization technique. From Eqs. (40) and (41), respectively, $\mathbf{F}(X \to Y)$ vanishes precisely where the "proxy objective function"

$$\mathbf{F}^*(X \to Y) = \sum_k \|LQ_kM^\mathsf{T}\|^2 \tag{42}$$

vanishes, where $Q_k = CA^kK$, $k = 0, \dots, r-1$ in the state-space case, and $Q_k = A_k$, $k = 1, \dots, p$ in the VAR case. As before $M$ spans the orthogonal complement of $L$, and $\|U\|^2 = \text{trace}[UU^\mathsf{T}]$ is the squared Frobenius matrix norm. While $\mathbf{F}^*(X \to Y)$ will not in general vary monotonically with $\mathbf{F}(X \to Y)$, simulations indicate strongly that subspaces $L$ which locally minimise $\mathbf{F}(X \to Y)$ will lie in regions of $\mathcal{G}_m(n)$ with near-locally minimal $\mathbf{F}^*(X \to Y)$. As a quadratic function of $L$, $\mathbf{F}^*(X \to Y)$, is considerably less computationally expensive to calculate than $\mathbf{F}(X \to Y)$ (note that the sequence $Q_k$ may be precalculated), and we have found that pre-optimising $\mathbf{F}^*(X \to Y)$ leads to significantly accelerated optimization of $\mathbf{F}(X \to Y)$, especially for high-dimensional and/or high-complexity models. The same techniques may be used to optimize $\mathbf{F}^*(X \to Y)$ as $\mathbf{F}(X \to Y)$.

For optimization of $\mathbf{F}(X \to Y)$, it will frequently be more computationally efficient to use the spectral integral form (24) rather than the parametric forms via solution of the DARE (34). A good heuristic choice for the frequency resolution is $d\omega \approx \frac{1}{2}\nu \log\rho/\log\varepsilon$, where $\rho$ is the spectral radius of the process (Sec. III), $\nu$ the sampling frequency, and $\varepsilon$ the machine floating-point epsilon [63]. We have found that numerical quadrature is generally computationally cheaper than (and of comparable accuracy to) the corresponding DARE-based computation provided that $\rho$ is not too close to 1, and for fixed $\rho$ scales better with system size.

Gradient descent techniques may be classified into those for which (i) the gradient is unknown or uncomputable, (ii) those for which just the gradient is exploited, (iii) those for which both gradient and Hessian are exploited. In Appendix D we calculate the gradient of $\mathbf{F}(X \to Y)$ for the spectral form (24) and also the gradient of the proxy form $\mathbf{F}^*(X \to Y)$ (42) under the Stiefel parametrization with the canonical metric on $\mathcal{G}_m(n)$. The Hessians may likewise be calculated analytically, but due to their computational complexity do not appear to offer much advantage over gradient-only techniques. We found that using Stiefel parametrization (non-Hessian) gradient descent with pre-optimization and multiple random initial values of $L$ is effective at identification of most local suboptima over a range of microscopic and macroscopic dimensions. See Appendix E for benchmark results.

### G. Interpretation of results

At each macroscopic scale $m$, the results of dynamical independence optimization for a microscopic linear system $X_t$ of dimension $n$ will be a set of linear projections $\{L : \mathbb{R}^n \to \mathbb{R}^m\}$. For any such $L$, the macroscopic process $Y_t = LX_t$ minimizes, at least locally, the dynamical dependence $\mathbf{F}(X \to Y)$ over all projections onto $\mathbb{R}^m$, and may be viewed as a (local, approximate) macroscopic "dynamical process in its own right," capturing a particular macroscopic perspective on the microscopic dynamics $X_t$. Each such $L$ defines a specific $m$-dimensional hyperplane—the space in which the macroscopic dynamics play out—projected from the $n$-dimensional space in which the microscopic dynamics play out.

To bolster our intuition, consider the simple case of a two-dimensional macroscopic variable in a three-dimensional microscopic space. In some system of coordinates, the projection $L$ will be of the form:

$$y_1 = L_{11}x_1 + L_{12}x_2 + L_{13}x_3, \tag{43}$$

$$y_2 = L_{21}x_1 + L_{22}x_2 + L_{23}x_3. \tag{44}$$

Now suppose, say, that $L_{13} = L_{23} = 0$. Then the macroscopic process $Y_t$ resides entirely on the $(x_1, x_2)$ plane in $\mathbb{R}^3$. So, for example, if $X_t$ represents channels of neural activity recorded from three brain regions, then the corresponding macroscopic variable only implicates the first two of those regions. We can restate the condition $L_{13} = L_{23} = 0$ as the vanishing of the *angle* between the plane defined by $L$ and the $(x_1, x_2)$ plane, or alternatively, as the vanishing of the angles between the plane and each of the $x_1$ and $x_2$ axes. If those angles are "small," then we might say that the macroscopic plane is "close" to the $(x_2, x_2)$ plane. Thus, in our toy neural example, hyperplane angles give a handle on the extent to which specific brain regions participate in the dynamics of a given macroscopic variable.

Unlike planes (or lines) in three-dimensional space, in higher dimensions there is no single unique angle between hyperplanes. Rather, for hyperplanes of dimensions $m_1 \leqslant m_2$ in $n$-dimensional Euclidean space there are $m_1$ unique *principal angles* $0 \leqslant \theta_1 \leqslant \ldots \leqslant \theta_{m_1} \leqslant \pi/2$, and $\sqrt{\theta_1^2 + \cdots + \theta_{m_1}^2}$ defines a transformation-invariant [64] measure of distance between hyperplanes [50]; if this distance is zero then the hyperplanes are colinear, while if it attains its maximum value of $(\pi/2)\sqrt{m_1}$ the hyperplanes are orthogonal. Note, though, that unlike for lower dimensions, if $n > 3$ and $1 < m < n-1$ the distances (in this case the first principal angles $\theta_1$) between a given $m$-dimensional hyperplane and each of the $x_i$ axes, $i = 1, \ldots, n$, do not uniquely determine the hyperplane; in higher dimensions there is more "wiggle room." For a finer-grained analysis, we might in addition consider the distances between an $m$-dimensional hyperplane and each of the $\binom{n}{m}$ coordinate hyperplanes spanned by combinations $x_{i_1}, \ldots, x_{i_m}$, $1 \leqslant i_1 < \ldots < i_m \leqslant n$, of the coordinate axes; collectively, these uniquely (over-)determine the given hyperplane.

While in our neural scenario a principal-angles analysis along the above lines yields insight into the degree to which specific brain regions participate in the dynamics of a given macroscopic variable, it has little to say about the internal dynamics of the macroscopic process. We may, via Eq. (33), calculate the dynamics of a (nearly) dynamically independent macroscopic variable explicitly as a linear system in its own right and, e.g., calculate its causal graph. However, the causal graph is not transformation-invariant. An interesting (and seemingly difficult) question for future research, is whether an invariant "canonical form" for the causal graph of a given linear system might be derivable through an appropriate linear transformation.

This concludes our analysis of characterization, discovery and interpretation of dynamically independent macroscopic variables in linear systems. In forthcoming work (in preparation), we demonstrate the techniques presented here in an empirical application: inference of emergent macroscopic dynamics from steady-state human MEG data recorded in placebo and under the influence of LSD, psilocybin, and subanaesthetic ketamine [65].

## IV. DETERMINISTIC AND CONTINUOUS-TIME DYNAMICS

Although the main thrust of this article concerns dynamical independence for discrete-time stochastic systems, and in particular discrete-time linear systems (Sec. III), many systems commonly associated with emergent phenomena feature deterministic and/or continuous-time dynamics. For stochastic processes in continuous time, although the basic information-theoretic formulation of Sec. II C carries through, transfer entropy is more nuanced and considerably more complex, both analytically and to estimate empirically [66], even in the linear case [67]. For deterministic systems the question immediately arises as to how the information-theoretic framework might apply, since Shannon information is indeterminate for nonstochastic variables. Below we preview some suggested approaches to this challenge.

### A. Discrete-state deterministic dynamics

In discrete time, many discrete-state systems of interest, such as cellular automata, flocking models and chaotic systems, dynamics take the deterministic Markovian form

$$x_{t+1} = \xi(x_t), \tag{45}$$

with finite or countably infinite microscopic state space $\mathcal{X}$ and state transition function $\xi : \mathcal{X} \to \mathcal{X}$. Thus, given some initial condition $x_0$, we have $x_t = \xi^t(x_0)$, $t \geqslant 0$ where $\xi^t$ denotes $t$ iterations of the mapping $\xi$.

Our preferred strategy for this case is to consider the dynamics (45) over an ensemble of realizations with *stochastic initial conditions*. Here a random variable $X_0$ on $\mathcal{X}$ is introduced to represent the statistical distribution of initial ($t = 0$) microscopic states, yielding the microscopic stochastic process $X_t = \xi^t(X_0)$, $t \geqslant 0$, on $\mathcal{X}$. Given a coarse-graining $f : \mathcal{X} \to \mathcal{Y}$, dynamical independence for the macroscopic variable $Y_t = f(X_t)$ may then be analyzed along the lines of the general discrete-time stochastic case (Sec. II C). In practice, choice of the initial distribution may be based on general principles (e.g., maximum-entropy), or on domain-specific *a priori* considerations. Then, since $Y_t$ depends deterministically on $X_t^- = \{X_0, X_1, \ldots, X_{t-1}\}$, the dynamical dependence (5) of $Y$ on $X$ at time $t \geqslant 0$ is given simply by

$$\mathbf{T}_t(X \to Y) = \mathbf{H}(Y_t|Y_t^-) = \mathbf{H}(Y_{t+1}^-) - \mathbf{H}(Y_t^-). \tag{46}$$

Given a probability mass function $p(x_0)$ for $X_0$, a general expression for the entropy $\mathbf{H}(Y_{t+1}^-) = \mathbf{H}(Y_0, Y_1, \ldots, Y_t)$, and thence $\mathbf{T}_t(X \to Y)$, may be calculated.

The initial stochastic conditions approach may also be extended to discrete-state systems in continuous time (e.g., *point processes* [68]).

### B. Continuous-state deterministic dynamics

When the state space $\mathcal{X}$ is *continuous*-valued, however, the above construction fails. In this case conditional entropies are replaced by conditional *differential* entropies [14], and it transpires that for deterministic systems with initial stochastic conditions both the macroscopic variable supervenience condition (2) and the dynamical dependence (5) are undefined (in both discrete and continuous time). Specifically, certain conditional differential entropies which appear in these equations diverge to $-\infty$ [69]. These divergences, furthermore, appear to be inherent; it is not hard to show, for example, that in linear continuous-state deterministic systems of the form $X_t = AX_{t-1}$, $Y_t = LX_t$ (cf. Sec. III A) with Gaussian initial conditions, they cannot be "finessed away" by the addition of intrinsic noise (on the process $X_t$) and/or extrinsic noise (on the process $Y_t$). In both cases, the relevant differential entropies again diverge as the noise magnitude tends to zero.

An important case of deterministic dynamics in continuous time is that of *flows*, defined as the group action $\xi : \mathcal{X} \times \mathbb{R} \to \mathcal{X}$ of the additive group $\mathbb{R}$ on a set $\mathcal{X}$, such that

$$\xi(\boldsymbol{x}, 0) = \boldsymbol{x}, \tag{47a}$$

$$\xi(\xi(\boldsymbol{x}, s), t) = \xi(\boldsymbol{x}, s + t), \tag{47b}$$

for $\boldsymbol{x} \in \mathcal{X}$, $s, t \in \mathbb{R}$. If $\mathcal{X}$ is a differentiable manifold, then the flow is *smooth* if the function $\xi$ is differentiable, and for any fixed $t$ the function $\boldsymbol{x} \mapsto \xi(\boldsymbol{x}, t)$ is a diffeomorphism. If $\boldsymbol{x} \mapsto \xi(\boldsymbol{x}, t)$ is only a diffeomorphism on a strict subset of $\mathcal{X} \times \mathbb{R}$, then $\xi$ is said to define a *local* flow; from now on, we use the term "flow" to include local flows. On Euclidean space $\mathcal{X} = \mathbb{R}^n$, smooth flows are essentially equivalent to first-order autonomous ordinary differential equations (ODEs); the *trajectory* $\boldsymbol{x}(t) = \xi(\boldsymbol{x}_0, t)$ is the unique solution of the autonomous ODE [70] $\dot{\boldsymbol{x}}(t) = g(\boldsymbol{x})$ with initial condition $\boldsymbol{x}(0) = \boldsymbol{x}_0$, where $g(\boldsymbol{x}) = \dot{\xi}(\boldsymbol{x}, 0)$. Many classical dynamical systems, such as Hamiltonian mechanics, flocking, and chaotic dynamical systems, are expressed as ODEs and may thus be considered as flows.

For flows, stochastic initial conditions founder on the problem of diverging differential entropies as described above. As an alternative formulation, we reconsider the original expression (4) of dynamical independence. There, independence is interpreted in a statistical sense; here we propose a "functional" interpretation more aligned with dynamical systems theory: given a smooth flow $\xi : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ at the microscopic level, a differentiable coarse-graining function $f : \mathbb{R}^n \to \mathbb{R}^m$, $0 < m < n$, is considered to define a dynamically independent macroscopic variable for $\xi$ iff there is a flow $\eta : \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^m$ on the macroscopic space such that

$$f(\xi(\boldsymbol{x}, t)) = \eta(f(\boldsymbol{x}), t), \tag{48}$$

for all $\boldsymbol{x} \in \mathbb{R}^n$ and $t \in \mathbb{R}$; i.e., the following diagram commutes:

$$\begin{tikzcd}
\mathbb{R}^n \times \mathbb{R} \arrow[r, "\xi"] \arrow[d, "f \times \mathbf{1}"'] & \mathbb{R}^n \arrow[d, "f"] \\
\mathbb{R}^m \times \mathbb{R} \arrow[r, "\eta"'] & \mathbb{R}^m
\end{tikzcd}$$

where $\mathbf{1}$ denotes the identity function on $\mathbb{R}$ [71]. In terms of ODEs, this is equivalent to the existence of an autonomous ODE [72] $\dot{\boldsymbol{y}}(t) = h(\boldsymbol{y})$ on $\mathbb{R}^m$ such that for any trajectory $\boldsymbol{x}(t)$ of $\xi$, $\boldsymbol{y}(t) = f(\boldsymbol{x}(t))$ is a trajectory of $\eta$. Thus, in the spirit of Eq. (4), a dynamically independent macroscopic system is self-determining: given an initial condition $\boldsymbol{y}_0 \in \mathbb{R}^m$, the coarse-grained macroscopic system determines its evolution in time without reference to the microlevel dynamics. Dynamical independence in this sense is invariant with respect to smooth coordinate transformations of both the microscopic space $\mathbb{R}^n$ and the macroscopic space $\mathbb{R}^m$; dynamical independence may thus be extended to flows on *differentiable manifolds* via overlapping coordinate charts [73].

In preliminary work (in preparation) we derive necessary and sufficient condition for dynamical independence in the above sense, and show that dynamically independent coarse-grainings $f : \mathbb{R}^n \to \mathbb{R}^m$ are built from *invariants* (conserved quantities) of the flow, along with a "timelike" scalar function. Thus, dynamical independence in the functional sense essentially reduces to the classical problem of invariants of flows on differentiable manifolds [74]; cf. the example of a Newtonian galaxy at the beginning of Sec. II C. By the celebrated First Theorem of Noether [75], for Lagrangian systems invariants are associated with symmetries of the Lagrangian action. Thus, for such systems Noether's Theorem characterizes the dynamically independent macroscopic variables. However, by no means all systems of interest fall into this class. (The central role of symmetry in Noether's Theorem, though, seems worth bearing in mind.)

The functional approach has one drawback: unlike the transfer entropy measure (5) in the discrete-time stochastic case, it lacks an information-theoretic interpretation, and does not yield up an obvious candidate measure for dynamical dependence, let alone a transformation-invariant one (we are currently investigating whether such a measure may exist); there is thus, as it stands, no ready notion of "near-dynamical independence." It might also be argued that the functional approach is unsuited to many-body scenarios as considered in statistical mechanics, where macroscopic variables are commonly built on probabilistic mesolevel descriptions of the system, possibly via a *repeated randomness* assumption [76–78]. In such cases, dynamical independence might more appropriately be applied with respect to the stochastic mesolevel.

## V. DISCUSSION

In this paper we introduce a notion of emergence of macroscopic dynamical structure in highly multivariate microscopic dynamical systems, which we term *dynamical independence*. Complementary to treatments of emergence which are largely concerned with part-whole and synergistic relationships between system components (see Sec. V A below), dynamical independence instantiates the intuition of an emergent process at a macroscopic scale as one which evolves over time according to its own dynamical laws, distinct from and independently of the dynamical laws operating at the microscopic level. More specifically, while prescribed by the microscopic process, a dynamically independent macroscopic process is, conditional on its own history, independent of the history of the microscopic process. Dynamical independence is quantified by a Shannon information-based (and hence transformation-invariant) measure of dynamical dependence. Importantly, dynamical independence may be conditional on a codistributed externally demarcated process, thus accommodating systems which feature input-output interactions with a dynamic environment.

Critical to any theory of emergence over a potential range of spatiotemporal scales, is how we should construe a "macroscopic variable." Here, we try to keep this question as open as possible, with one key constraint: a macroscopic variable may be any process co-distributed with the microscopic process which, in predictive terms, does not "bring anything new to the table" beyond the microscopic: a macroscopic variable is prescribed by the microscopic process in the sense that it does not self-predict beyond the prediction afforded by the microscopic (and environmental) dynamics. We might thus conclude that if a macroscopic process appears to emerge as a process in its own right—with a "life of its own"—this apparent autonomy is in the eye of a beholder blind to the microlevel dynamics. Emergence, in our approach, is therefore best thought of as being associated with particular "ways of looking" at a system.

A key aspect of our approach is an emphasis on *discovery* of emergent macroscopic variables—"ways of looking" at the system—given a microlevel description. Although specific problem domains may present "natural" and/or intuitively appealing prospective emergent macroscopic variables (which may be tested for degree of emergence by our dynamical dependence measure), this is by no means always the case. For neural systems, for example, it is in general far from clear how to identify candidate emergent processes. Groups of neurons firing in synchrony might intuitively suggest an emergent variable, but there may be many more (and more subtle) patterns of neural activity that may count as emergent without being intuitively apparent to an observer. Our approach addresses this issue by formalising and operationalising the discovery process, through consideration of the full space of all admissible macroscopic variables; discovery of emergent variables then becomes a search/optimization problem across this space. We introduce an *ansatz* that proposes the results of this search, across all scales, as an informative account of the emergence structure of the given system—an "emergence portrait." Parametric modeling, furthermore, opens up the possibility of *data-driven* discovery of emergent variables. We developed our approach by presenting an explicit calculation of dynamical dependence, along with a detailed account of the search/optimization process, for the important class of linear systems, suitable for wide deployment across a range of domains, including neural systems.

### A. Related approaches

A difference between our approach and many related approaches in the literature is our emphasis on discovery of emergent phenomena. Other differences concern the role of the environment (Sec. V B), and in particular the system/environment distinction [79]. Dynamical independence furthermore, in contrast to some treatments of emergence, does not overtly address part-whole (mereological) relationships within the microscopic system.

Our notion of dynamical independence is closely related to the property of *informational closure*, introduced by Bertschinger *et al.* [11] to address the system-environment distinction, individuality and autonomy. Essentially, a dynamical system $X_t$ is informationally closed with respect to a coupled dynamical environment $E_t$ if there is zero information flow from environment to system; i.e., if the transfer entropy $\mathbf{T}(E \to X)$ vanishes. Closer to dynamical independence, Pfante *et al.* [80] (see also Refs. [81,82]) effectively consider the microscopic process as the "environment" for a coarse-grained macroscopic process. Although the motivation appears to be similar, they only consider discrete-state systems, single time-step histories, and Markovian dynamics (which are preserved by coarse-graining in the case of informational closure). Dynamical independence is thus a significant generalization of this approach.

To mitigate against "trivial" information closure where the system is simply decoupled from the environment, Bertschinger *et al.* [11] deem an information closure to be nontrivial, if, in addition to low environment → system information flow, the system also shares information with the environment to a significant degree. They accordingly introduce a measure of *nontrivial informational closure* (NTIC). We remark that NTIC seems less than relevant to supervenient macroscopic variables (Sec. II B), since here the microscopic "environment" and macroscopic process are inherently coupled. Chang *et al.* [83] apply NTIC specifically to the case of coarse-grained macroscopic variables in the context of an environment [84]. Their definition of a *C-process* requires that the macroscopic variable $Y$ be (i) dynamically independent of the system-environment "universe" $(X, E)$, and (ii) NTIC with respect to the environment $E$. Note that condition (i) is not equivalent to dynamical independence of $Y$ with respect to the system in the context of the environment (53). While they clearly recognize the importance, in empirical scenarios, of "[finding] appropriate coarse-graining functions which map microscopic processes to macroscopic C-processes," they do not offer any concrete proposals on how this might be achieved.

Another relevant construct is *G-emergence* [Granger emergence; 8]. Seth [8] first operationalizes the "self-causation" or "self-determination" of a variable $Y$ with respect to an external (multivariate) variable $Z$ as *G-autonomy* [85]

$$\begin{aligned} ga(Y \mid Z) &= \mathbf{I}(Y_t : Y_t^- \mid Z_t^-) \\ &= \mathbf{H}(Y_t \mid Z_t^-) - \mathbf{H}(Y_t \mid Z_t^-, Y_t^-), \end{aligned} \tag{49}$$

which measures the degree to which inclusion of its *own* past enhances prediction of $Y_t$ by the past of the external variable $Z_t$. Given a microscopic process $X_t$ and a macroscopic process $Y_t$, the G-emergence of $Y$ from $X$ is then specified as [86]

$$ge(Y \mid X) = ga(Y \mid X) + \mathbf{T}(X \to Y). \tag{50}$$

This expression operationalizes the notion that an emergent macroscopic process is, in a predictive sense, at once autonomous from, but also dependent on, the microscopic process—again recalling the conceptual definition of "weak emergence" from Ref. [4]. G-emergence differs from dynamical independence in two main respects. First, it requires that a macroscopic variable be nontrivially self-predictive. Second, it includes a micro-to-macro term to assure, in an ad hoc way, that they are related, in contrast to the principled approach to coarse-graining taken by dynamical independence.

We recognize immediately the second (transfer entropy) term in Eq. (50)—designed to ensure that the macro and the micro are related—as our dynamical dependence (5) in the absence of a coupled environment, although for G-emergence it "pulls in the opposite direction", in the sense that increasing $\mathbf{T}(X \to Y)$ increases G-emergence, but *de*creases dynamical independence. Note also, though, that our supervenience requirement (2) on macroscopic variables—which holds in particular for coarse-grained variables—actually stipulates (in the absence of an environment) that the G-autonomy contribution $ga(Y \mid X)$ in (50) vanishes identically, thus leaving G-emergence as precisely our dynamical *dependence* rather than *in*dependence, for the situations we consider for dynamical independence.

A recent approach with both parallels and differences to ours is that of Rosas *et al.* [87] (see also Mediano *et al.* [88]). In contrast to our approach, their concern is explicitly with mereological causal relationships, such as *downward causation*, what they term *causal decoupling* and, in particular, *causal emergence*. The latter is quantified as the unique predictive capacity of a supervenient feature over the microscopic system, beyond the predictive capacity of (parts of) the microscopic system. This is almost the obverse of dynamical independence, which hinges on prediction of the *macroscopic* rather than the microscopic process. Supervenience for "features" as defined by Rosas *et al.* [87], it should be noted, does not generally correspond to our notion of supervenience for macroscopic variables. In contrast to our supervenience condition (2), the comparable condition in Ref. [87, Sec. II] is, in our notation

$$\mathbf{I}(X_t : Y_t^- \mid X_t^-) \equiv 0. \tag{51}$$

Although coarse-grained variables trivially satisfy both Eqs. (2) and (51), the latter again speaks to prediction of the microscopic, rather than macroscopic variable.

To express causal emergence in information-theoretic terms, Rosas *et al.* [87] make use of a *partial information decomposition* [PID; 89,90]. One challenge for this approach is a lack of consensus on what a "canonical" PID might look like. Further, current PID candidates tend to be computationally intractable (but see, e.g., Ref. [91]) and scale poorly with system size and macroscopic scale. In addition, the proposed measures are frequently framed in terms of discrete-valued (often finite) systems, and it is often unclear how they might be realized—or they become counterintuitive and/or exhibit discontinuous behavior—when extended to continuous-valued variables [92]. Connected with the last point, many (though not all; e.g., Ref. [91]) lack the transformation invariance of Shannon information [93,94]. In recognition of the computational burden attached to PIDs, Rosas *et al.* [87] define Shannon information-based "large system approximations" for their measures, although it is unclear to what extent these reflect the intent of the respective PID formulations.

Closer in spirit to our approach is the theory of emergent brain macrostates propounded by Allefeld *et al.* [15]. Along similar lines to dynamical independence, they consider dynamics for macroscopic systems which are in a sense "self-contained" with respect to the microscopic dynamics; however, unlike our more general information-theoretic approach, they associate such dynamics with a (1st-order, discrete) Markov property: "...the Markov-property criterion distinguishes descriptive levels at which the system exhibits a self-contained dynamics ('eigendynamics'), independent of details present at other levels." Emergent macroscopic processes are then identified with coarse-grainings which preserve the Markov property ["Markov partitions" [95]].

Since low-level neural processes, and indeed neurophysiological recordings of these processes, do not naturally take the form of first-order discrete-valued Markov processes, Allefeld *et al.* [15] devise a discrete approximation scheme [96]. They then seek Markovian coarse-grainings of the discretized Markov model in the form of metastable macrostates [97] and (putatively emergent) dynamics that transition between such macrostates at slow timescales compared to the underlying microscopic dynamics. This latter idea, more closely aligned with a thermodynamical perspective on coarse-graining [98,99], seems worthy of further investigation in regard to dynamical independence [100].

Along comparable lines, Strasberg *et al.* [77], in an attempt to reconcile the repeated randomness assumption in statistical mechanics [76] with microscopic reversibility, introduce a concept which they term "microstate independence" which holds, roughly, if all microstates compatible with the observed (coarse-grained) macrostate give rise to the same dynamics. Although a precise definition is not given, it seems that microstate independence should imply dynamical independence. Although it cannot hold strictly for Hamiltonian system due to time reversibility at the microstate level, on the basis of a detailed mathematical argument and some plausible heuristic assumptions it is claimed as in some sense generic in the case where the macroscopic observable is "slow" and "coarse," and the dynamics thus approximately Markovian on a coarse timescale.

Hoel *et al.* [101] formulate a notion of causal emergence based on *effective information* [102]. Here, although macro is supervenient on micro, a coarse-grained macroscopic variable is deemed emergent to the extent that it leads to a gain in effective information. Effective information is calculated by comparing the distribution of prior states that could have caused a given current state (the "causal distribution"), with the uniform distribution over the full repertoire of possible prior states. The KL-divergence of the causal distribution with respect to the uniform distribution is then averaged over the distribution of current states. The procedure is motivated by the Pearlian approach [103] which identifies causation with the effects of counterfactual interventions (perturbations) on the system; the uniform (maximum entropy) distribution then stands as an injection of random perturbations. A drawback of effective information, however, is that it assumes the *existence* of a uniform distribution of states, thus ruling out a large class of (in particular, continuous-state) physical systems, for which the uniform distribution does not exist; and even if it exists, it is not clear that the effective information will be transformation-invariant. It may also be argued that a uniform distribution over prior states is in any case a purely notional, unphysical construct, and that its deployment consequently fails to reflect causation "as it actually happens"—that is, as stochastic dynamics play out over time. In a related approach, Friston *et al.* [104] present a recursive partitioning of neuronal states based on effective connectivity graphs [105] and Markov blankets [106], which they associate with emergent intrinsic brain networks at hierarchical spatiotemporal scales [107].

Millidge [108] presents a mathematical theory of abstraction which shares some commonalities with theories of emergence. An abstraction is considered as a set of "summaries" of a system which are sufficient to answer a specified set of "queries" regarding the time evolution of the system. Like macroscopic variables (in the broad sense), abstractions discard information about the system's detailed dynamics—in this case such information as turns out to be irrelevant to the specific queries. It is proposed that the irrelevant information be considered via the *maximum-entropy principle* [109], whereby uncertainty about detailed system behavior is maximized within the constraint of retention of the ability to answer the queries. Like dynamically independent macrovariables, abstractions might be considered to have a "life of their own" insofar as they retain sufficient information to predict their own behaviors at a macroscopic level. In common with our approach, Millidge [108] places an emphasis on *data-driven discovery* of abstractions, by minimising their "leakiness"—that is, their departure from accurate prediction of the associated macrophenomena (cf. dynamical dependence). In contrast to dynamical independence, abstractions might be said to be driven by the agenda of the observer (in the form of specific queries), rather than, as in our case, unconstrained and intrinsic to the dynamical structure of the microsystem.

Finally, our approach is also clearly related to the general idea of dimensionality reduction in information theory, machine learning, and beyond. Importantly, dynamical independence defines a very specific basis for dimensionality reduction, one which flows explicitly from the dynamics of the underlying microscopic system. This might be contrasted, for example, with principal components analysis (PCA), which is essentially determined by correlations within a dataset. In case the data derives from a dynamical process (e.g., econometric data, neuroimaging data, etc.), these correlations are *contemporaneous*, and as such fail to reflect in full the temporal dynamics of the generative process.

### B. Dynamically coupled environment

Suppose that the microscopic process $X_t$ is jointly distributed with an *environmental process* $E_t$. We may easily extend our formalism by simply conditioning on the history of the environmental process; thus the condition (2) for supervenience of a macroscopic variable $Y_t$ becomes

$$\mathbf{I}(Y_t : Y_t^- \mid X_t^-, E_t^-) \equiv 0, \tag{52}$$

and dynamical dependence (5) of $Y_t$ on $X_t$ in the context of the environment $E_t$ becomes

$$\mathbf{T}_t(X \to Y \mid E) = \mathbf{I}(Y_t : X_t^- \mid Y_t^-, E_t^-). \tag{53}$$

As regards the linear operationalization (Sec. III A), this may also be readily extended by conditioning the Granger causalities [110] on the coupled environmental process. We remark, however, that Eqs. (24) and (25) no longer obtain as they stand. In this case dynamical dependence in time and frequency domains may still be computed parametrically (Secs. III B and III C), or nonparametrically via spectral factorization (17) of the joint CPSD of $(X, E)$ [35,44]; we leave this for a future study.

### C. Relationship with autonomy

One might be tempted to describe a macroscopic process $Y$ that is dynamically independent with respect to the microscopic process $X$ as being "autonomous of $X$." We avoid this usage, though, because conventionally the term autonomy carries two distinct connotations [111]: an autonomous process should not only be independent of external "driving" processes, but should also *self-determine* its evolution over time [8]. As remarked in Sec. I A, a dynamically independent macroscopic variable need not fulfill the self-determination criterion; dynamical independence does not equate to autonomy (cf. Sec. V A, Granger autonomy/emergence). In the extreme case, a dynamically independent macroscopic variable might in fact be completely random, as in the following trivial VAR(1) example:

$$X_{1,t} = aX_{1,t-1} + bX_{2,t-1} + \varepsilon_{1,t}, \tag{54a}$$

$$X_{2,t} = \varepsilon_{2,t}, \tag{54b}$$

where $\varepsilon_{1,t}, \varepsilon_{2,t}$ are uncorrelated white noises. Here the macroscopic (coarse-grained) white noise $Y_t = X_{2,t}$ is clearly dynamically independent of the microscopic process $X_t$. Note, however, that a completely random macroscopic variable is not necessarily dynamically independent: if we replace (54b) with

$$X_{2,t} = \varepsilon_{2,t} + c\varepsilon_{1,t-1}, \tag{55}$$

then, while $Y_t = X_{2,t}$ is still a white noise, it is no longer dynamically independent of $X_t$ [112].

We consider this as a positive feature of our definition of dynamical independence: As per our *ansatz* (1), if we discover that our microscopic system features a completely random macroscopic variable at some scale, this tells us something useful about the system. We might even, via Eq. (8), choose to "factor out" this embedded randomness to better reveal significant causal structure.

### D. Discovery of emergent macroscopic processes in neural systems

Notwithstanding that the generative mechanisms underlying neural processes may be highly nonlinear, linear modeling is routinely deployed for the functional analysis of neural systems via neurophysiological recordings (indeed, correlation statistics are associated with linear regression; see also discussion in Sec. III). Granger causality based on VAR (and more recently state-space) modeling in particular is a popular technique for inference of directed functional connectivity [34,36] from EEG, MEG, and iEEG data [113]. The techniques described in Sec. III F may thus be applied directly to estimated state-space models for such data, to infer the emergence portrait of neural systems. Issues of scale remain significant, but with sufficient computing power do not appear to be intractable.

While it may be tempting to draw analogies between dynamically independent macrovariables in neural systems and network-level constructs prominent in neuroimaging analyses, e.g., default-mode networks [114], this would be misleading; a dynamically independent macrovariable is not a static "network," but rather a macroscale dynamical entity in its own right, emerging from interactions on the "microscopic" scale (in this case, the scale set by neural recording channels associated with "small" brain regions). A fascinating question for future empirical research, is whether specific emergent (dynamically independent) macrovariables might be associated with ("neural correlates" of) large-scale neural phenomena, such as behaviors, cognition, and specific states of, or disorders of, consciousness.

Whether in the analysis of neural systems or in other application domains, the underlying intuition behind any study of emergence for real-world systems is that identifying emergent structure is likely to advance our understanding of the physical phenomena in question. While reasonable, this conclusion is not a given. Whether emergent dynamical structures turn out to be functionally relevant for explaining a particular system's behavior will most often be an empirical question.

## ACKNOWLEDGMENTS

This research was supported by the European Research Council (Advanced Investigator Grant No. 101019254) and by the Dr. Mortimer and Theresa Sackler Foundation. The authors thank Nadine Spychala, Adam Barrett, Javier Galadí, Fernando Rosas, Pedro Mediano, and Boki Milinkovic for useful discussions and two anonymous reviewers for constructive comments.

## APPENDIX A: PROOF OF TRANSITIVITY OF DYNAMICAL INDEPENDENCE

From property (8) we construct isomorphisms

$$f \times u : \mathcal{X} \to \mathcal{Y} \times \mathcal{U}, \tag{A1a}$$

$$g \times v : \mathcal{Y} \to \mathcal{Z} \times \mathcal{V}, \tag{A1b}$$

as in the diagram below

$$\begin{tikzcd}
\mathcal{X} \arrow[r, "f"] \arrow[dr, "u"] & \mathcal{Y} \arrow[r, "g"] \arrow[dr, "v"] & \mathcal{Z} \\
& \mathcal{U} & \mathcal{V}
\end{tikzcd}$$

so that

$$(g \circ f) \times (v \circ f) \times u : \mathcal{X} \to \mathcal{Z} \times \mathcal{V} \times \mathcal{U} \tag{A2}$$

is an isomorphism. Setting $U_t = u(X_t)$, $V_t = v(Y_t) = v(f(X_t))$, under this isomorphism, we have $X_t \sim (Z_t, V_t, U_t)$ with $Y_t \sim (Z_t, V_t)$, and by (9)

$$\mathbf{T}(X \to Y \mid E_t) = \mathbf{T}(U \to Z, V \mid E_t), \tag{A3a}$$

$$\mathbf{T}(Y \to Z \mid E_t) = \mathbf{T}(V \to Z \mid E_t), \tag{A3b}$$

$$\mathbf{T}(X \to Z \mid E_t) = \mathbf{T}(V, U \to Z \mid E_t). \tag{A3c}$$

The dynamical independence of $Y_t$ from $X_t$, and of $Z_t$ from $Y_t$ then become [cf. (5)]

$$\mathbf{H}(Z_t, V_t \mid Z_t^-, V_t^-, E_t^-) = \mathbf{H}(Z_t, V_t \mid Z_t^-, V_t^-, U_t^-, E_t^-), \tag{A4a}$$

$$\mathbf{H}(Z_t \mid Z_t^-, E_t^-) = \mathbf{H}(Z_t \mid Z_t^-, V_t^-, E_t^-), \tag{A4b}$$

respectively, while

$$\mathbf{T}(X \to Z \mid E_t) = \mathbf{H}(Z_t \mid Z_t^-, E_t^-) - \mathbf{H}(Z_t \mid Z_t^-, V_t^-, U_t^-, E_t^-). \tag{A5}$$

Now by Eq. (A4a), conditional on $(Z_t^-, V_t^-, E_t^-)$, the joint variable $(Z_t, V_t)$ is independent of $U_t^-$. Thus, again conditional on $(Z_t^-, V_t^-, E_t^-)$, the marginal $Z_t$ is itself independent of $U_t^-$. We thus have [115]

$$\mathbf{H}(Z_t \mid Z_t^-, V_t^-, E_t^-) = \mathbf{H}(Z_t \mid Z_t^-, V_t^-, U_t^-, E_t^-), \tag{A6}$$

which, together with Eqs. (A4b) and (A5), yields $\mathbf{T}(X \to Z \mid E_t) = 0$ and Eq. (10) holds as required. ■

## APPENDIX B: GRANGER CAUSALITY

We begin by noting that the optimal linear prediction of $X_t$ in the least-squares sense is given by the conditional expectation $\mathbf{E}[X_t \mid X_t^-] = \sum_{k=1}^{\infty} A_k X_{t-k}$ (19). The residual prediction errors are then just the innovations $\boldsymbol{\varepsilon}_t = X_t - \mathbf{E}[X_t \mid X_t^-]$, and in the formulation of Geweke [29] the magnitude of the prediction error is quantified by the generalized variance [116,117] $|\Sigma|$, where $\Sigma = \mathbf{E}[\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^{\mathsf{T}}]$ is the error covariance matrix.

Suppose now that the vector process $X_t$ is partitioned into two subprocesses $X_t = [X_{1t}^{\mathsf{T}}\ X_{2t}^{\mathsf{T}}]^{\mathsf{T}}$. To specify the Granger causality $\mathbf{F}(X_2 \to X_1)$ we compare the prediction error $|\Sigma_{11}|$ of $X_{1t}$ predicted on the joint past $X_t^-$ of both itself and $X_{2t}$, with the prediction error $|\Sigma_{11}^{\mathsf{R}}|$ of $X_{1t}$ predicted only on its own past $X_{1t}^-$; here the superscript "$^{\mathsf{R}}$" refers to the "restricted" VAR representation,

$$X_{1t} = \sum_{k=1}^{\infty} A_k^{\mathsf{R}} X_{1,t-k} + \boldsymbol{\varepsilon}_{1t}^{\mathsf{R}}, \tag{B1}$$

and $\Sigma_{11}^{\mathsf{R}} = \mathbf{E}[\boldsymbol{\varepsilon}_{1t}^{\mathsf{R}} \boldsymbol{\varepsilon}_{1t}^{\mathsf{RT}}]$ is the corresponding error covariance matrix. Geweke [29] then defines the Granger causality as the log-ratio of generalized variances,

$$\mathbf{F}(X_2 \to X_1) = \log \frac{|\Sigma_{11}^{\mathsf{R}}|}{|\Sigma_{11}|}, \tag{B2}$$

which quantifies the degree to which the history of $X_{2t}$ enhances prediction of $X_{1t}$ beyond the degree to which $X_{1t}$ is predicted by its own history alone. We note that if the innovations are Gaussian, then the generalized variance $|\Sigma|$ is proportional to the likelihood function for $X_t$, so that (B2) is a log-likelihood ratio, which under ergodic assumptions is asymptotically equivalent to the conditional entropy $\mathbf{H}(X_t \mid X_t^-)$ [32]; this circumscribes the relationship between Granger causality and transfer entropy [cf. (5b)]. Under the classical "large-sample theory" [118–120], the log-likelihood ratio as a sample statistic furnishes asymptotic $F$ and $\chi^2$ tests for statistical inference on GC (but see also Gutknecht and Barnett [58]).

Unlike transfer entropy, Granger causality may also be defined in the frequency domain: the spectral GC from $X_{2t}$ to $X_{1t}$ at angular frequency $\omega \in [0, 2\pi]$ is given by [29,34]

$$\mathbf{f}(X_2 \to X_1; \omega) = \log \frac{|S_{11}(\omega)|}{|S_{11}(\omega) - H_{12}(\omega)\Sigma_{22|1}H_{12}^*(\omega)|}. \tag{B3}$$

Here $S(\omega)$ is the CPSD matrix (17), $H(z)$ the transfer function (15), and $\Sigma_{22|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ a partial covariance matrix. Spectral GC averages across the broadband frequency range $[0, 2\pi]$ to yield time-domain GC [29]:

$$\mathbf{F}(X_2 \to X_1) = \frac{1}{2\pi} \int_0^{2\pi} \mathbf{f}(X_2 \to X_1; \omega)\, d\omega. \tag{B4}$$

Given a frequency band $[\omega_1, \omega_2] \subseteq [0, 2\pi]$, we may define the "band-limited" (time-domain) Granger causality [52]

$$\mathbf{F}(X_2 \to X_1; \omega_1, \omega_2) = \frac{1}{\omega_2 - \omega_1} \int_{\omega_1}^{\omega_2} \mathbf{f}(X_2 \to X_1; \omega)\, d\omega, \tag{B5}$$

which may be interpreted as the information transfer from $X_2$ to $X_1$ associated with frequencies $\omega_1 \leqslant \omega \leqslant \omega_2$.

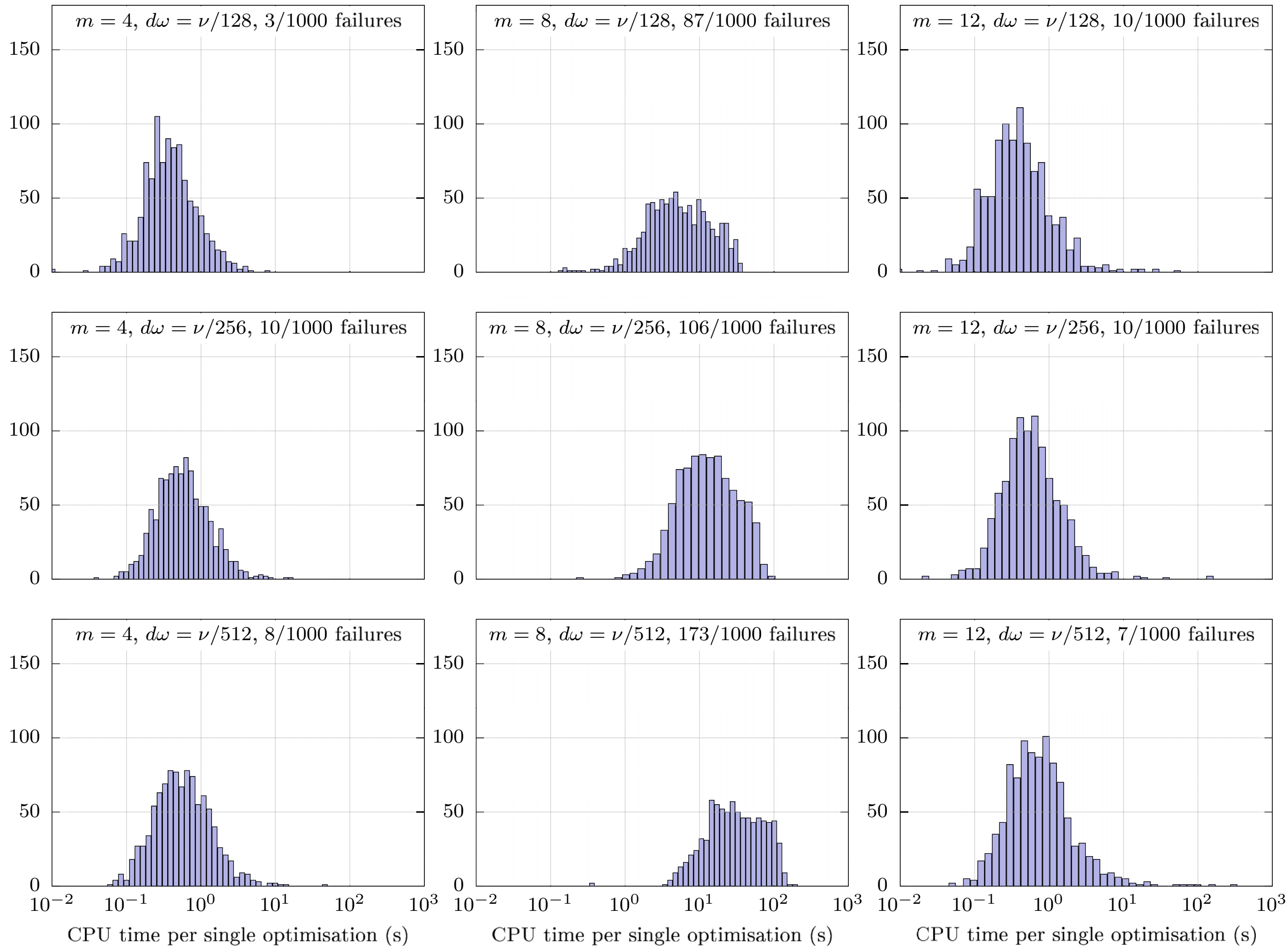

FIG. 1. Distributions of CPU time per dynamical dependence optimization run, for state-space systems of dimension $n = 16$ (state dimension $r = 11$), for macroscopic dimension $m = 4$, 8, and 12 (columns), and frequency resolution $d\omega = \nu/128$, $\nu/256$, and $\nu/512$ (rows), where $\nu$ is the Nyqvist sampling frequency. Histograms were compiled from 1000 independent runs; "failures" denote the number of runs that failed to converge to a local minimum within the specified number of gradient-descent steps: 100 000 for pre-optimization and 10 000 for full optimization. See main text for details.

## APPENDIX C: ALTERNATIVE PARAMETRIZATIONS FOR OPTIMIZATION ON THE GRASSMANNIAN MANIFOLD

An alternative approach is to use local coordinate charts for the $\mathcal{G}_m(n)$. Any full-rank $m \times n$ matrix $L$ can be represented as

$$L = \Psi[I_{m\times m}\ M]\Pi, \tag{C1}$$

where $\Pi$ is an $n \times n$ permutation matrix (i.e., a row or column permutation of $I_{n\times n}$), $\Psi$ an $m \times m$ nonsingular transformation matrix, and $M$ is $m \times (n-m)$ full-rank. For given $\Pi$, $\mathcal{G}_m(n)$ is then locally and diffeomorphically mapped by the $m \times (n-m)$ full-rank matrices $M$ [121]. But note that for given $\Pi, M$, while there is no redundancy in the (injective) mapping $M : \mathbb{R}^{m(n-m)} \to \mathcal{G}_m(n)$, the space of such $M$ is unbounded and doesn't cover the entire Grassmannian, which again makes numerical optimization awkward.

A partial resolution is provided by a surprising mathematical result due to Knuth [122] (see also Ref. [123]), which states roughly that given any fixed $\delta > 1$, for any full-rank $m \times n$ matrix $L_0$ there is a neighbourhood of $L_0$, a permutation matrix $\Pi$, and a transformation $\Psi$ such that for any $L$ in the neighbourhood of $L_0$, all elements of $M$ satisfying Eq. (C1) are bounded to lie in $[-\delta, \delta]$. That is, in the local neighbourhood of any subspace in the Grassmannian, we can always find a suitable permutation matrix $\Pi$ such that Eq. (C1) effectively parametrizes the neighbourhood by a *bounded* submanifold of $\mathbb{R}^{m(n-m)}$. During the course of an optimization process, then, if the current local search (over $M$) drifts outside its $\delta$ bounds, we can always find a *new* bounded local parametrization of the search neighbourhood "on the fly." Finding a suitable new $\Pi$ is, however, not straightforward, and calculating the requisite $\Psi$ for Eq. (C1) is computationally expensive [124]. Nor is this scheme particularly convenient for population-based optimization algorithms, which will generally require keeping track of different permutation matrices for different subpopulations, and—worse—for some algorithms (such as CE optimization), it seems that the procedure can only work if the entire current population resides in a single $(\Psi, \Pi)$-chart.

Other suggested approaches in the literature to optimization on Grassmannian manifolds include the "proxy matrix" algorithm of Nagananda *et al.* [125], and the "involution"

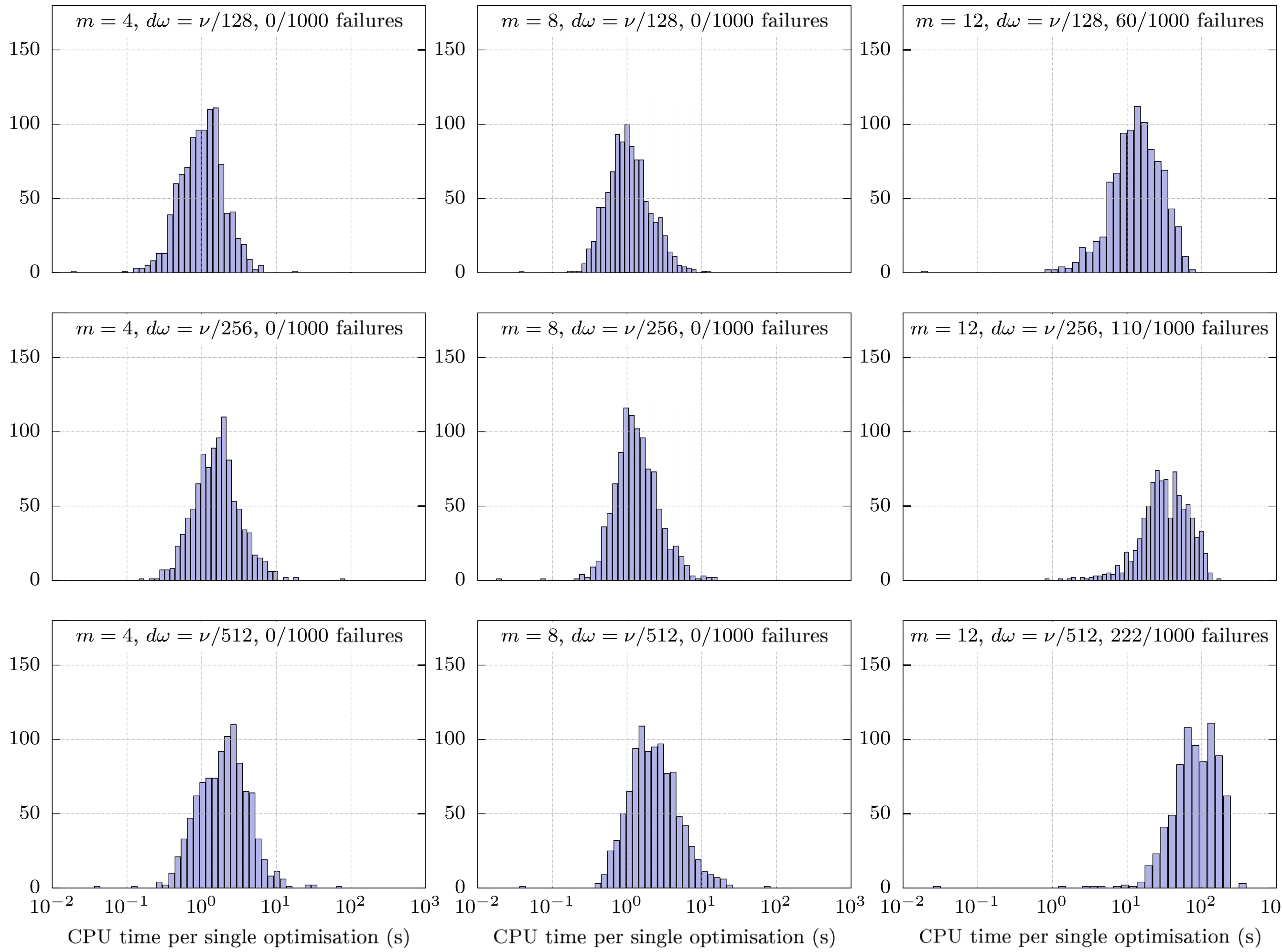

FIG. 2. Distributions of CPU time per dynamical dependence optimization run, for state-space systems of dimension $n = 32$ (state dimension $r = 21$). See Fig. 1 and main text for details.

parametrization of Lai *et al.* [60]; regarding the latter, though, it is not apparent how to efficiently convert the parametrized form of a Grassmannian element into a linear projection operator $L$ as required for calculation of dynamical dependence.

## APPENDIX D: GRADIENT OF DYNAMICAL DEPENDENCE ON THE GRASSMANNIAN

Let $X_t$ be a linear process (Sec. III), $L$ an $m \times n$ matrix with $LL^{\mathsf{T}} = I$, and $Y_t = LX_t$. Under the assumption of decorrelated innovations (i.e., $\Sigma = I$), the dynamical dependence $\mathbf{F} = \mathbf{F}(X \to Y)$ for the corresponding Grassmannian element $\{L\}$ is given by Eq. (24), and the gradient under the canonical metric for $\mathcal{G}_m(n)$ by [49, Sec. 2.5.2]

$$\nabla \mathbf{F} = \frac{\partial \mathbf{F}}{\partial L}(I - L^{\mathsf{T}}L), \tag{D1}$$

where $\frac{\partial \mathbf{F}}{\partial L}$ denotes the $m \times n$ matrix with entries $\frac{\partial \mathbf{F}}{\partial L_{\alpha i}}$, $\alpha = 1, \dots, m$, $i = 1, \dots, n$ [126]. From the formula for the derivative of a log-determinant (for compactness we drop the *omega* argument of $S$)

$$\frac{\partial}{\partial L_{\alpha i}} \log |LSL^{\mathsf{T}}| = \operatorname{trace}\left([LSL^{\mathsf{T}}]^{-1} \frac{\partial}{\partial L_{\alpha i}}[LSL^{\mathsf{T}}]\right) \tag{D2}$$

Using the Hermitian property of $S$, so that $S^{\mathsf{T}} = \bar{S}$ (complex conjugate), we may calculate that

$$\frac{\partial}{\partial L_{\alpha i}}[LSL^{\mathsf{T}}]_{\beta\gamma} = \delta_{\alpha\beta}[LS]_{\gamma i} + \delta_{\alpha\gamma}[L\bar{S}]_{\beta i}, \tag{D3}$$

so that Eq. (D2) yields

$$\frac{\partial}{\partial L} \log |LSL^{\mathsf{T}}| = 2\mathfrak{Re}\{[LSL^{\mathsf{T}}]^{-1}LS.\} \tag{D4}$$

We may check that multiplying Eq. (D4) on the right by $L^{\mathsf{T}}$ yields $2I$, so that by Eqs. (24) and (D1),

$$\nabla \mathbf{F} = 2\left(\frac{1}{2\pi}\int_0^{2\pi} \mathfrak{Re}\{[LS(\omega)L^{\mathsf{T}}]^{-1}LS(\omega)\}d\omega - L.\right) \tag{D5}$$

We may similarly calculate the gradient of the "proxy" dynamical dependence $\mathbf{F}^*(X \to Y)$ given by Eq. (42). Given an $n \times n$ matrix $Q$, using the identity $L^{\mathsf{T}}L + M^{\mathsf{T}}M = I$ and

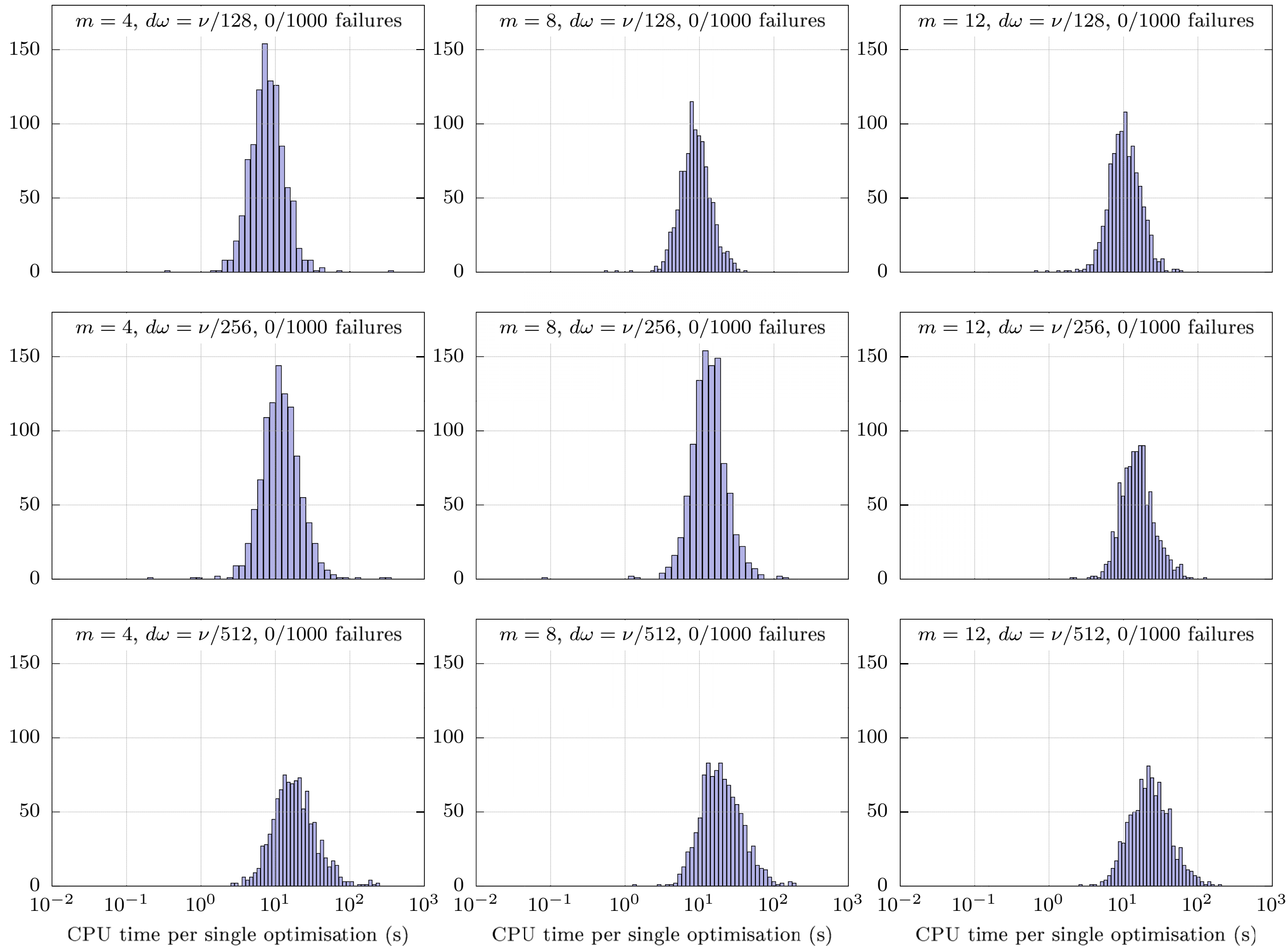

FIG. 3. Distributions of CPU time (log scale) per dynamical dependence optimization run, for state-space systems of dimension $n = 64$ (state dimension $r = 43$). See Fig. 1 and main text for details.

setting $P = L^{\mathsf{T}}L$, we may derive

$$\begin{aligned}\|LQM^{\mathsf{T}}\|^2 &= \text{trace}[LQM^{\mathsf{T}}MQ^{\mathsf{T}}L^{\mathsf{T}}]\\ &= \text{trace}[QQ^{\mathsf{T}}P] - \text{trace}[QPQ^{\mathsf{T}}P]. \qquad \text{(D6)}\end{aligned}$$

We may calculate

$$\frac{\partial}{\partial L}\text{trace}[QQ^{\mathsf{T}}P] = 2LQQ^{\mathsf{T}}, \qquad \text{(D7)}$$

and after some algebra that

$$\frac{\partial}{\partial L}\text{trace}[QPQ^{\mathsf{T}}P] = 2L(QPQ^{\mathsf{T}} + Q^{\mathsf{T}}PQ). \qquad \text{(D8)}$$

Thus, from Eq. (42) and the Grassmannian gradient formula [49], we have

$$\nabla\mathbf{F}^* = 2L\sum_k(Q_kQ_k^{\mathsf{T}} - Q_kL^{\mathsf{T}}LQ_k^{\mathsf{T}} - Q_k^{\mathsf{T}}L^{\mathsf{T}}LQ_k)(I - L^{\mathsf{T}}L). \qquad \text{(D9)}$$

## APPENDIX E: DYNAMICAL DEPENDENCE OPTIMIZATION FOR LINEAR SYSTEMS: BENCHMARK RESULTS

In benchmark tests, we used (non-Hessian) gradient descent with a naïve line-search strategy whereby the step size is increased by a constant factor if following the gradient results in smaller dynamical dependence, otherwise decreased by a constant factor. The algorithm terminates when either step size or gradient magnitude fall below given tolerances, dynamical dependence falls below a given tolerance [127], or the number of iterations exceeds a specified maximum. For state-space or VAR models, pre-optimization (Sec. III F) using the proxy dynamical dependence (42) followed by dynamical dependence optimization using the spectral form (24) allows for successful optimization of dynamical dependence up to system dimension $n \approx 100$ in minutes per restart on a standard multi-core workstation, provided the spectral radius is not too close to 1.

Figures 1–3 plot histograms of the distribution of CPU time (log scale), for randomly generated state-space systems of dimension 16, 32, and 64, respectively (state dimensions were $r \approx 2n/3$; see Sec. III B). Simulations, written in Matlab, were performed on a Xeon 12-core, 3.5 GHz workstation running Linux. At each state-space (microscopic) dimension and each macroscopic dimension, we performed 1000 optimization runs, comprising pre-optimization followed by full optimization. Each individual run was performed independently on a *different* randomly generated state-space model, with initial projection uniformly random on the Grassmannian. Runs terminated when the gradient descent step size or gradient magnitude fell below a tolerance set to $10^{-8}$. Runs timed-out ("failed") if the number of steps (100 000 for pre-optimization and 10 000 for full optimization) was exceeded. We see that CPU times per run are very roughly log-normally distributed, and increase with microscopic dimension, macroscopic dimension and frequency resolution.

Optimization times in our benchmark tests reported here are somewhat pessimistic: in practice we find that, for a given microscopic system and macroscopic dimension, multiple independent pre-optimization runs with random initial projections will converge to the *same* local optima [128]; gradient descent on the actual dynamical dependence then need only be performed starting at the unique preoptimized projections. Since restarts are independent and may be run concurrently, parallel high-performance computing offers considerable scope for improved efficiency. GPU computing potentially also offers significant efficiency improvements.

[1] A. Cavagna, S. M. Duarte Queirós, I. Giardina, F. Stefanini, and M. Viale, Proc. R. Soc. B **280**, 20122484 (2013).

[2] C. Bays, in *Encyclopedia of Complexity and Systems Science*, edited by R. A. Meyers (Springer, New York, NY, 2009), pp. 4240–4249.

[3] S. Weisenburger, F. Tejera, J. Demas, B. Chen, J. Manley, F. T. Sparks, F. Martínez Traub, T. Daigle, H. Zeng, A. Losonczy, and A. Vaziri, Cell **177**, 1050 (2019).

[4] M. A. Bedau, Noûs **31**, 375 (1997).

[5] D. J. Chalmers, in *The Re-emergence of Emergence: The Emergentist Hypothesis From Science to Religion*, edited by P. Davies and P. Clayton (Oxford University Press, Oxford, UK, 2006), pp. 244–256.

[6] M. A. Bedau, Principia **6**, 5 (2002).

[7] J. Kim, Synthese **151**, 547 (2006).

[8] A. K. Seth, Artificial Life **16**, 179 (2010).

[9] A property A is said to be supervenient on a property B if A "depends on" B, in the sense that a difference in the state of A implies a difference in the state of B.

[10] D. Davidson, in *Essays on Actions and Events* (Oxford University Press, New York, NY, 1980), Chap. 11.

[11] N. Bertschinger, E. Olbrich, N. Ay, and J. Jost, in *Explorations in the Complexity of Possible Life: Proceedings of the 7th German Workshop of Artificial Life* (IOS Press, Amsterdam, 2006), pp. 26–28.

[12] Informational closure is usually framed as a property of a system with respect to its *environment*; but see, e.g., Pfante *et al.* [80].

[13] The macroscopic variables in this example are, of course, *conserved* (time-invariant) quantities, and as such trivially self-predicting; see Sec.IV B for further discussion.

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley-Interscience, New York, NY, 1991).

[15] C. Allefeld, H. Atmanspacher, and J. Wackermann, Chaos **19**, 015102 (2009).

[16] Throughout, if the state space is continuous, then mutual information is defined in terms of *differential* entropy [14].

[17] This might be extended to variables of the form $Y_t = f(X_t, X_t^-)$, which aggregate over time as well as state; the condition (2) still holds.

[18] We might frame our analysis more formally in the language of Category Theory [129].

[19] T. Schreiber, Phys. Rev. Lett. **85**, 461 (2000).

[20] M. Paluš, V. Komárek, Z. Hrnčíř, and K. Štěrbová, Phys. Rev. E **63**, 046211 (2001).

[21] A. Kaiser and T. Schreiber, Physica D **166**, 43 (2002).

[22] T. Bossomaier, L. Barnett, M. Harré, and J. T. Lizier, *An Introduction to Transfer Entropy: Information Flow in Complex Systems* (Springer International Publishing, Switzerland, 2016).

[23] In the language of Bertschinger *et al.* [11], $Y$ is "informationally closed" with respect to $X$; see Sec. V A.

[24] This is entirely analogous to statistical independence—a condition on the relationship between jointly distributed random variables—and mutual information, a quantitative measure of statistical dependence.

[25] For some systems (e.g., where the only structure is set-theoretic and the state space uncountable) (8) may require the Axiom of Choice [130].

[26] N. Wiener, in *Modern Mathematics for Engineers*, edited by E. F. Beckenbach (McGraw Hill, New York, NY, 1956), pp. 165–190.

[27] C. W. J. Granger, Inform. Control **6**, 28 (1963).

[28] C. W. J. Granger, Econometrica **37**, 424 (1969).

[29] J. Geweke, J. Am. Stat. Assoc. **77**, 304 (1982).

[30] L. Barnett, A. B. Barrett, and A. K. Seth, Phys. Rev. Lett. **103**, 238701 (2009).

[31] K. Hlaváčková-Schindler, Appl. Math. Sci. **5**, 3637 (2011).

[32] L. Barnett and T. Bossomaier, Phys. Rev. Lett. **109**, 138105 (2012).

[33] L. Barnett and A. K. Seth, J. Neurosci. Methods **223**, 50 (2014).

[34] L. Barnett and A. K. Seth, Phys. Rev. E **91**, 040101(R) (2015).

[35] M. Dhamala, G. Rangarajan, and M. Ding, Phys. Rev. Lett. **100**, 018701 (2008).

[36] A. K. Seth, A. B. Barrett, and L. Barnett, J. Neurosci. **35**, 3293 (2015).

[37] A proof may be constructed along the same lines as the proof in Appendix A.

[38] When we wish to stress that the condition (6) is satisfied exactly, we shall sometimes refer to "perfect" dynamical independence (cf. Sec. III D).

[39] This is particularly pertinent for systems where the signal-to-noise ratio is low (e.g., many econometric time series),

and model complexity is heavily penalized in terms of model estimation error.

[40] All vectors are considered to be *column* vectors, unless otherwise stated.

[41] Eq. (15) is the *Wold decomposition* [46]; the purely nondeterministic requirement states that the deterministic component of the decomposition is identically zero.

[42] P. Masani, in *Multivariate Analysis*, edited by P. R. Krishnaiah (Academic Press, New York, NY, 1966), pp. 351–382.

[43] By abuse of notation, we write $H(\omega)$ for $H(e^{-i\omega})$.

[44] G. T. Wilson, SIAM J. Appl. Math. **23**, 420 (1972).

[45] Y. A. Rozanov, *Stationary Random Processes* (Holden-Day, San Francisco, CA, 1967).

[46] J. Doob, *Stochastic Processes* (John Wiley, New York, NY, 1953).

[47] A further caveat is that restriction of coarse-graining to the linear domain neglects potentially parsimonious *non*linear macroscopic variables.

[48] S. Helgason, *Differential Geometry, Lie Groups, and Symmetric Spaces* (Academic Press, New York, NY, 1978).

[49] A. Edelman, T. A. Arias, and S. T. Smith, SIAM J. Matrix Anal. Appl. **20**, 303 (1998).

[50] Y.-C. Wong, Proc. Natl. Acad. Sci. USA **57**, 589 (1967).

[51] An orthonormal basis for the orthogonal complement may be found using a Singular Value Decomposition (SVD) of $L$.

[52] L. Barnett and A. K. Seth, J. Neurosci. Methods **201**, 404 (2011).

[53] E. J. Hannan and M. Deistler, *The Statistical Theory of Linear Systems* (SIAM, Philadelphia, PA, 2012).

[54] P. Lancaster and L. Rodman, *Algebraic Riccati Equations* (Oxford University Press, Oxford, UK, 1995).

[55] In Gutknecht and Barnett [58] it is shown that for a VAR-derived ISS, the DARE may be dimensionally reduced, resulting in improved computational efficiency.

[56] J. D. Hamilton, *Time Series Analysis* (Princeton University Press, Princeton, NJ, 1994).

[57] P. van Overschee and B. L. R. de Moor, *Subspace Identification for Linear Systems: Theory, Implementation, Applications* (Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996).

[58] A. J. Gutknecht and L. Barnett, Biometrika, asad009 (2023).

[59] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis* (Springer-Verlag, Berlin, 2005).

[60] Z. Lai, L.-H. Lim, and K. Ye, arXiv:2009.13502.

[61] Z. I. Botev, D. P. Kroese, R. Y. Rubinstein, and P. L'Ecuyer, in *Handbook of Statistics*, Handbook of Statistics, Vol. 31, edited by C. R. Rao and V. Govindaraju (Elsevier, Amsterdam, 2013), Chap. 3, pp. 35–59.

[62] J. A. Nelder and R. Mead, Comput. J. **7**, 308 (1965).

[63] The *autocorrelation* at lag $k$ of a VAR process decays $\propto \rho^k$, and we note that by the well-known Wiener-Khintchine theorem, the CPSD $S(e^{-i\omega})$ which appears under the integral in (24) is the discrete Fourier transform (DFT) of the autocorrelation sequence. Thus, according to the heuristic, any further precision in the quadrature conferred by finer spectral resolution is likely to be consumed by floating-point (relative) rounding error.

[64] Invariant, that is, under linear transformations of $\mathbb{R}^{m_1}$ and $\mathbb{R}^{m_2}$.

[65] S. D. Muthukumaraswamy, R. L. Carhart-Harris, R. J. Moran, M. J. Brookes, T. W. Williams, D. Errtizoe, B. Sessa, A. Papadopoulos, M. Bolstridge, K. D. Singh, A. Feilding, K. J. Friston, and D. J. Nutt, J. Neurosci. **33**, 15171 (2013).

[66] R. E. Spinney, M. Prokopenko, and J. T. Lizier, Phys. Rev. E **95**, 032319 (2017).

[67] L. Barnett and A. K. Seth, J. Neurosci. Methods **275**, 93 (2017).

[68] D. J. Daley and D. Vere-Jones, *An Introduction to the Theory of Point Processes* (Springer-Verlag, New York, NY, 1988), Vol. 1.

[69] Pfante *et al.* [81] analyse a coarse-graining of the well-known discrete-time "tent map", where the microscopic variable is continuous-state on [0,1] with a uniform stochastic initial distribution, and the macroscopic variable is discrete-state on $\{0, 1\}$. In this case, the relevant conditional entropies are well-defined.

[70] A dot indicates [partial] differentiation with respect to $t$.

[71] This bears comparison with Allefeld *et al.* [15], where the parsimony of macroscopic variables is associated with the preservation of a Markov property.

[72] Note that "autonomous" with reference to an ODE indicates that the equation has no explicit time dependence.

[73] S. Kobayashi and K. Nomizu, *Foundations of Differential Geometry, Volume 1* (John Wiley & Sons, New York, NY, 1996).

[74] A. Cohen, *An introduction to the Lie Theory of One-Parameter Groups, with Applications to the Solution of Differential Equations* (D. C. Heath & Co., New York, NY, 1911).

[75] V. I. Arnold, *Mathematical Methods of Classical Mechanics* (Springer-Verlag, New York, NY, 1978).

[76] N. G. van Kampen, J. Non-Equilib. Thermodyn. **11**, 327 (1985).

[77] P. Strasberg, A. Winter, J. Gemmer, and J. Wang, arXiv:2209.07977.

[78] The repeated randomness assumption guarantees classical dynamics.

[79] D. Krakauer, N. Bertschinger, E. Olbrich, J. C. Flack, and N. Ay, Theory Biosci. **139**, 209 (2020).

[80] O. Pfante, N. Bertschinger, E. Olbrich, N. Ay, and J. Jost, Adv. Complex Syst. **17**, 1450007 (2014).

[81] O. Pfante, E. Olbrich, N. Bertschinger, N. Ay, and J. Jost, Chaos **24**, 013136 (2014).

[82] N. Ay, N. Bertschinger, J. Jost, E. Olbrich, and J. Rauh, in *Complexity and Emergence : Lake Como School of Advanced Studies*, Springer proceedings in mathematics and statistics, Vol. 383, edited by S. Albeverio, E. Mastrogiacomo, E. R. Gianin, and S. Ugolini (Springer, Cham, 2022), pp. 87–105.

[83] A. Y. C. Chang, M. Biehl, Y. Yu, and R. Kanai, Front. Psychol. **11**, 1504 (2020).

[84] While Chang *et al.* [83] associate a C-process with a measure of consciousness, it is perhaps more generally (and less contentiously) construed as a notion of autonomy or emergence in complex systems.

[85] Seth [8] expresses G-autonomy and G-emergence in terms of linear prediction; here we present the information-theoretic analogues under Gaussian assumptions; cf. Sec. III below. In Seth's 2010 approach, the linear (Granger causality) formulation is important because his measure relies on a systematic *in*ability to capture the full dynamical behavior of a target system.

[86] Again, we have translated this into equivalent information-theoretic terms.

[87] F. E. Rosas, P. A. M. Mediano, H. J. Jensen, A. K. Seth, A. B. Barrett, R. L. Carhart-Harris, and D. Bor, PLoS Comput. Biol. **16**, e1008289 (2020).

[88] P. A. M. Mediano, F. E. Rosas, A. I. Luppi, H. J. Jensen, A. K. Seth, A. B. Barrett, R. L. Carhart-Harris, and D. Bor, Philos. Trans. R. Soc. A **380**, 20210246 (2022).

[89] P. L. Williams and R. D. Beer, arXiv:1004.2515.

[90] M. Wibral, V. Priesemann, J. W. Kay, J. T. Lizier, and W. A. Phillips, Brain Cogn. **112**, 25 (2017).

[91] A. Pakman, A. Nejatbakhsh, D. Gilboa, A. Makkeh, L. Mazzucato, M. Wibral, and E. Schneidman, Adv. Neural Inf. Process Syst. **34**, 20295 (2021).

[92] A. B. Barrett, Phys. Rev. E **91**, 052802 (2015).

[93] D. Chicharro, G. Pica, and S. Panzeri, Entropy **20**, 169 (2018).

[94] F. E. Rosas, P. A. M. Mediano, B. Rassouli, and A. B. Barrett, J. Phys. A: Math. Theor. **53**, 485001 (2020).

[95] R. L. Adler, Bull. Amer. Math. Soc. **35**, 1 (1998).

[96] It is unclear how neural processes, which typically feature signal propagation delays, feedback over a range of timescales and medium- to long-range memory, might in general be well-represented by 1st-order discrete-valued Markov processes at a fixed time increment.

[97] E. Olivieri and M. E. Vares, *Large Deviations and Metastability* (Cambridge University Press, Cambridge, UK, 2005).

[98] M. S. Green, J. Chem. Phys. **20**, 1281 (1952).

[99] K. Jeffery, R. Pollack, and C. Rovelli, Entropy **21**, 1211 (2019).

[100] We note that a Markovian coarse-grained macroscopic variable would automatically satisfy our criterion (4) for dynamical independence, although for continuous-state systems we do not discretise, so the respective emergence criteria are not directly comparable.

[101] E. P. Hoel, L. Albantakis, and G. Tononi, Proc. Natl. Acad. Sci. USA **110**, 19790 (2013).

[102] G. Tononi and O. Sporns, BMC Neurosci. **4**, 31 (2003).

[103] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed. (Cambridge University Press, Cambridge, UK, 2009).

[104] K. J. Friston, E. D. Fagerholm, T. S. Zarghami, T. Parr, I. Hipólito, L. Magrou, and A. Razi, Network Neurosci. **5**, 211 (2021).

[105] K. Friston, R. Moran, and A. K. Seth, Curr. Opin. Neurobiol. **23**, 172 (2013).

[106] J. Pearl, in *Quantified Representation of Uncertainty and Imprecision*, edited by P. Smets (Springer Netherlands, Dordrecht, 1998), pp. 367–389.

[107] Although described as a "renormalization group" approach, it is never adequately explained why (or indeed whether) the dimensional reductions associated with partitioning should lead to self-similar dynamics at increasingly coarse scales.

[108] B. Millidge, arXiv:2106.01826.

[109] E. T. Jaynes, in *Complex Systems—Operational Approaches in Neurobiology, Physics, and Computers*, edited by H. Haken (Springer, Berlin, 1985), pp. 254–269.

[110] J. Geweke, J. Am. Stat. Assoc. **79**, 907 (1984).

[111] N. Bertschinger, E. Olbrich, N. Ay, and J. Jost, Biosystems **91**, 331 (2008).

[112] $X_t$ is now VARMA(1,1) rather than VAR(1).

[113] Granger-causal analysis of fMRI BOLD data, however, remains controversial, due to confounds related to slow sampling rates [131] and potentially also to the hemodynamic response function [HRF; 132].

[114] M. E. Raichle, A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman, Proc. Natl. Acad. Sci. USA **98**, 676 (2001).

[115] Note that this conclusion holds for both discrete, and—with *differential* entropy—for continuous-valued state.

[116] S. S. Wilks, Biometrika **24**, 471 (1932).

[117] A. B. Barrett, L. Barnett, and A. K. Seth, Phys. Rev. E **81**, 041907 (2010).

[118] J. Neyman and E. S. Pearson, Philos. Trans. R. Soc. A **231**, 289 (1933).

[119] S. S. Wilks, Ann. Math. Statist. **9**, 60 (1938).

[120] A. Wald, Trans. Am. Math. Soc. **54**, 426 (1943).

[121] These charts comprise the "standard atlas" used to define the canonical differentiable manifold structure on the Grassmannian.

[122] D. E. Knuth, Linear Multilinear Algebra **17**, 1 (1985).

[123] K. Usevich and I. Markovsky, Automatica **50**, 1656 (2014).

[124] V. Mehrmann and F. Poloni, SIAM J. Matrix Anal. Appl. **33**, 780 (2012).

[125] N. Nagananda, B. Minnehan, and A. Savakis, arXiv:2104.08112.

[126] Note that our $L$ is transposed compared to the $Y$ in Edelman *et al.* [49].

[127] For iterative numerical minimization of dynamical dependence derived from empirical data via state-space or VAR modeling, a stopping criterion may be based on statistical inference (Sec. III E): iterated search may be terminated on failure to reject the appropriate null hypothesis of vanishing dynamical dependence at a predetermined significance level. As mentioned in Sec. III E, in lieu of known sampling distributions for the nonparametric and state-space Granger causality estimators, this is only likely to be practicable for VAR modeling.

[128] Identifying whether two linear mappings $L_1, L_2$ correspond to the same Grassmannian element is easily established by calculation of the hyperplane distance between $L_1$ and $L_2$ (Sec. III G).

[129] S. Mac Lane, *Categories for the Working Mathematician*, 2nd ed. (Springer Science+Business Media, New York, NY, 1978).

[130] J. L. Bell, in *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta (The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University, Stanford, CA. Online: https://plato.stanford.edu/entries/axiom-choice/, 2015).

[131] A. K. Seth, P. Chorley, and L. Barnett, NeuroImage **65**, 540 (2013).

[132] V. Solo, Neural Comput. **28**, 914 (2016).