

Unifying pairwise interactions in complex dynamics

Received: 1 July 2022

Accepted: 14 August 2023

Published online: 25 September 2023

 Check for updates

Oliver M. Cliff^{1,2}, Annie G. Bryant^{1,2}, Joseph T. Lizier^{2,3}, Naotsugu Tsuchiya^{4,5,6} & Ben D. Fulcher^{1,2}✉

Scientists have developed hundreds of techniques to measure the interactions between pairs of processes in complex systems, but these computational methods—from contemporaneous correlation coefficients to causal inference methods—define and formulate interactions differently, using distinct quantitative theories that remain largely disconnected. Here we introduce a large assembled library of 237 statistics of pairwise interactions, and assess their behavior on 1,053 multivariate time series from a wide range of real-world and model-generated systems. Our analysis highlights commonalities between disparate mathematical formulations of interactions, providing a unified picture of a rich interdisciplinary literature. Using three real-world case studies, we then show that simultaneously leveraging diverse methods can uncover those most suitable for addressing a given problem, facilitating interpretable understanding of the quantitative formulation of pairwise dependencies that drive successful performance. Our results and accompanying software enable comprehensive analysis of time-series interactions by drawing on decades of diverse methodological contributions.

A fundamental question in science is how complex dynamics can be characterized by measuring the interactions within a distributed system. To address this question, many approaches have been developed to measure different types of pairwise interactions from dynamical data. For example, in neuroimaging, functional connections between pairs of brain regions are quantified through statistical correlations, which mark changes in human behaviors¹ and differ in neurological diseases². In Earth system science, pairwise causal models have been used to infer mechanistic drivers of natural processes, from the influence of sea-surface temperature on sardine and anchovy populations³, to the atmospheric drivers of air circulation⁴. Furthermore, economic analysts have studied the co-integration of paired non-stationary time series—including stock-market indices and their associated future contracts—to infer a statistically significant coupling for building econometric models⁵.

As illustrated schematically in Fig. 1a, the common goal of these studies is to extract meaningful pairwise relationships from multivariate time series (MTS): sets of observations taken regularly over time⁶. In the age of big data, the scientific problems that are studied in diverse disciplinary contexts—from genomics to astronomy⁷—require novel ways to extract information from MTS data; however, despite the myriad ways to quantify a statistical dependency between two time series, it remains common practice to manually select a single method with minimal comparisons against alternatives. For instance, Pearson correlation remains the most commonly used tool for measuring pairwise relationships in neuroimaging⁸ and Earth system science⁴, despite rather restrictive (and often unsatisfied) assumptions that the data are serially independent and normally distributed⁹. Fortunately, many more sophisticated and powerful algorithms have been developed, including those where dependencies are lagged in time (for example,

¹School of Physics, The University of Sydney, Camperdown, New South Wales, Australia. ²Centre for Complex Systems, The University of Sydney, Camperdown, New South Wales, Australia. ³School of Computer Science, The University of Sydney, Camperdown, New South Wales, Australia. ⁴Turner Institute for Brain and Mental Health & School of Psychological Sciences, Faculty of Medicine, Nursing, and Health Sciences, Monash University, Melbourne, Victoria, Australia. ⁵Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology (NICT), Suita-shi, Japan. ⁶Advanced Telecommunications Research Computational Neuroscience Laboratories, Seika-cho, Japan. ✉e-mail: ben.fulcher@sydney.edu.au

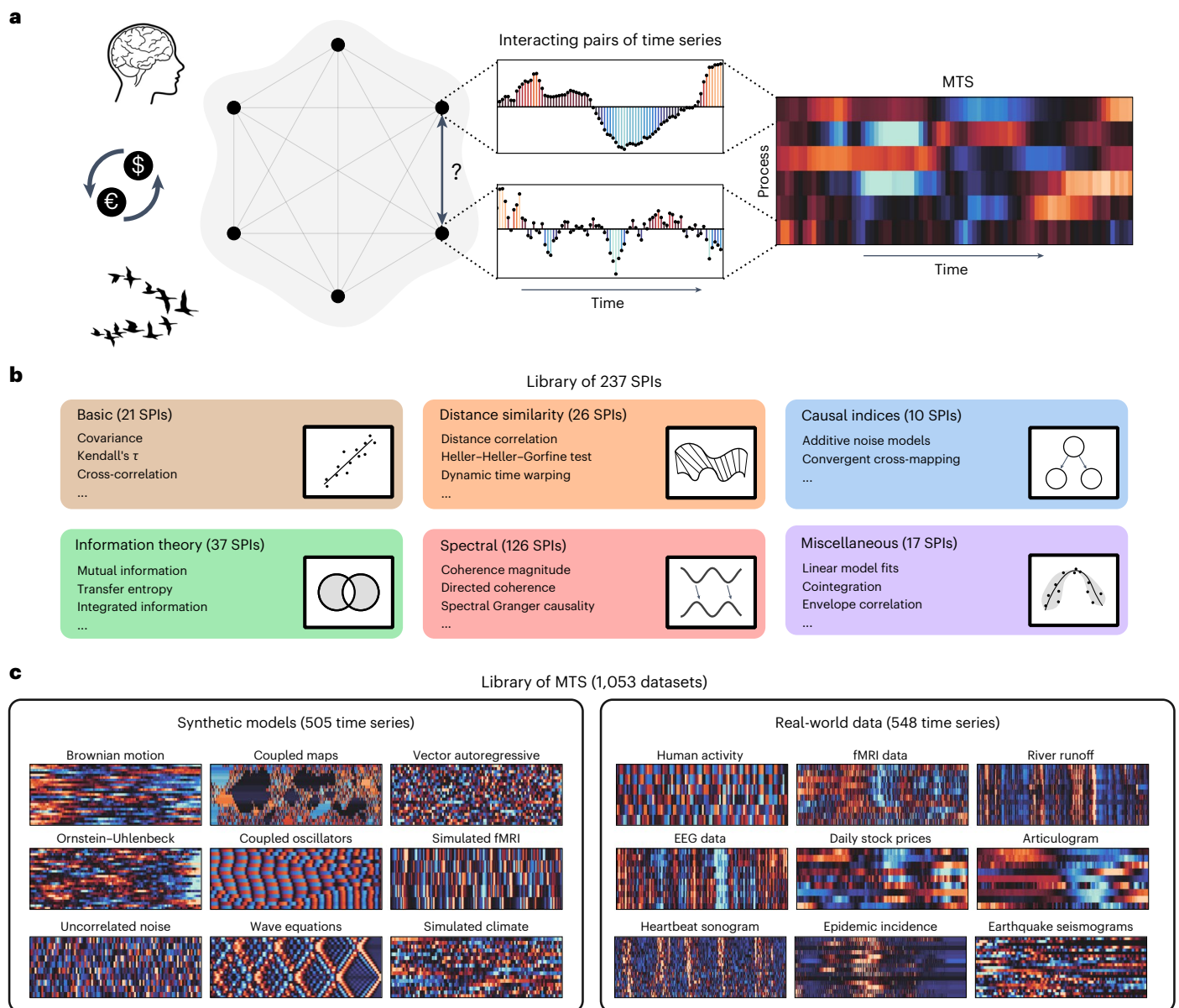


Fig. 1 | The behavior of a scientific library of 237 SPIs was evaluated using a collection of 1,053 MTS. a, In many disciplines, including neuroscience, economics and biology, scientists analyze interactions between pairs of processes in a MTS. For illustration, two time series are plotted, using colour to indicate time-series values on a scale from low (light blue), to the mean (black), to high (light orange). An MTS can be visualized as a heat map using the same coloring, in which individual processes (rows) evolve over time (columns). **b**, The 237 SPIs analyzed here derive from six broad disciplinary categories:

'basic', 'distance similarity', 'causal indices', 'information theory', 'spectral' and 'miscellaneous', as described in detail in Supplementary Note 1; selected examples of SPIs within each category are listed here, along with a schematic depiction. **c**, Selected examples from our diverse collection of 1,053 MTS—which are used to evaluate the empirical behavior of the SPIs—are plotted as process $\times$ time heat maps (using the same colors as in **a**). Our library includes data generated from synthetic models (left) and measured from real-world systems (right), as described in detail in Supplementary Note 2.

cross-correlation⁶), may be misaligned (for instance, dynamic time warping¹⁰), or where the knowledge of one variable improves the predictability of another (for example, Granger causality¹¹).

In this Article, we represent algorithms that measure interactions between pairs of time series as real-valued summary statistics, that is, statistics of pairwise interactions (SPIs). Figure 1b illustrates the diverse theoretical tools and types of interactions covered in the scientific literature on SPIs, from covariance (the foundation of statistics and machine learning) to convergent cross-mapping³ (developed to infer causal effect in complex ecosystems); however, as the theory underlying statistical interactions between pairs of time series has been developed largely independently across disciplinary contexts,

these methods remain disconnected from one another. In this work, we unify this wealth of interdisciplinary scientific knowledge, empirically connecting previously disjoint methodological traditions to yield a unified set of tools for quantifying interactions in complex dynamics.

As diverse scientific methods for quantifying pairwise interactions have never been compared at scale, there remain many unanswered questions. Are all of these methods capturing unique information on the interactions occurring within a time-varying system? Is there synergy between complementary approaches that, when combined, tells us more about the underlying system than any single method can? Or, do interesting redundancies exist between techniques used across disciplines that hint at a common theoretical underpinning

that has not previously been identified? Following previous highly comparative studies of univariate time series¹² and graphs¹³, this work addresses these questions by simultaneously evaluating hundreds of different SPIs directly from data. Our empirical approach to this problem involves first assembling a comprehensive annotated library of computational methods for quantifying pairwise interactions from data (summarized in Fig. 1b), and then analyzing their behavior across a large and diverse library of MTS (Fig. 1c).

As shown in Fig. 1b, our annotated library of SPIs was organized into six broad categories on the basis of their underlying theory: ‘basic’, ‘distance similarity’, ‘causal indices’, ‘information theory’, ‘spectral’ and ‘miscellaneous’ (see ‘Library of 237 SPIs’ section in the Methods for more details, and Supplementary Note 1 for a full list of SPIs with descriptions and references). To understand the behavior of these SPIs on data, we constructed a library of 1,053 diverse MTS, curated with the aim of capturing the main classes of systems and dynamics that are studied across scientific disciplines, including synchronization, spatiotemporal chaos, wave propagation, criticality and phase transitions. As depicted in Fig. 1c, our MTS library contains 505 synthetic MTS generated from a range of mathematical models, including coupled maps, ordinary differential equations and partial differential equations, and 548 diverse real-world MTS assembled from public databases such as geophysical, medical and financial data (see ‘Library of 1,053 MTS’ section in the Methods for further details, and Supplementary Note 2 for a full list of included MTS).

Results

Organizing pairwise interactions by their empirical behavior

Having assembled diverse libraries of methods (237 SPIs) and data (1,053 MTS), we aimed to analyze how similarly the different SPIs behave on the data. To achieve this, we developed an empirical similarity index, R ($0 \leq R \leq 1$), that captures the relationship between any two SPIs by comparing their output when applied to all 195,112 pairwise interactions present across all 1,053 MTS. This index is derived from the average absolute Spearman correlation between a pair of SPIs when applied to all pairs of processes in all datasets (see ‘Quantifying similarity between SPIs using the empirical similarity index’ section of the Methods for details). The minimum value, $R = 0$, indicates a pair of maximally distinct SPIs (with uncorrelated behavior on all datasets), whereas the maximum, $R = 1$, indicates a pair of SPIs that are perfectly correlated on all datasets (that is, they behave as simple monotonic transformations of one another); thus, a pair of SPIs with a high R reflects broadly similar behavior across MTS containing very different types of structures, and therefore acts as a suitable candidate index of empirical similarity.

Using the dissimilarity measure, $D = 1 - R$, we organized all 237 SPIs using hierarchical clustering, yielding the dendrogram shown in Fig. 2b. This presents a data-driven, structured representation of a diverse literature that allows us to probe and interpret relationships between scientific methods at multiple levels. We focus our analysis here on a 14-module resolution (with modules labeled M1–M14), which captures important methodological connections between groups of SPIs with similar behavior on data. As summarized in Fig. 2a, these fourteen modules group common conceptual and theoretical approaches to measuring interactions between pairs of time series, demonstrating the ability of our empirical approach to meaningfully organize the interdisciplinary literature.

In addition to grouping similar types of methods into modules, we found that different high-level conceptual formulations of dynamical interactions were recapitulated in the relationships between modules. For example, modules M3–M6 contain distinct types of SPIs (including Granger causality¹¹, directed information¹⁴ and integrated information^{15,16}), all of which capture statistical dependencies between two time series by considering the context of their past. This idea that observable interactions are predicated (or, from a statistical standpoint, conditioned) on the history of a process was first proposed by

the Wiener–Granger theories of causality and feedback¹⁷, specifically by measuring how one time series might improve the self-predictability of another. Our results group SPIs on the basis of this underlying theoretical formulation, due to their characteristic behavior on data. Other types of SPIs (that do not predicate on the self-predictability of a process) also display distinctive behavior, including measures of contemporaneous relationships (the correlation coefficients of M14), dependencies that account for temporal lags (the coherence measures of M12), or temporal dilation and shifts (dynamic time warping and related methods in M10).

Ten of the fourteen modules are homogeneous, containing methods that are derived from similar underlying theories (as indicated by the color of the category labels in Fig. 2bii). Of these ten homogeneous modules, six of them (M4, M7, M8, M9, M11 and M13) comprise SPIs for measuring one specific type of pairwise interaction, differing in either their: specific algorithmic implementation (for example, M13 contains both the Engle–Granger⁵ and the Johansen¹⁸ tests for measuring co-integration); extracted summary statistics (for example, M8 contains both the mean and maximum of the wavelet-based phase lag index¹⁹); or parameter settings (for example, M4 contains SPIs that use five different estimation techniques for transfer entropy²⁰). The remaining four homogeneous modules (M1, M2, M6 and M12) comprise methods with very similar theoretical underpinnings; for example, M12 contains many SPIs for measuring undirected interactions via Fourier transformations, such as the magnitude and the imaginary part of the coherence²¹. Of particular interest are the four heterogeneous modules (M3, M5, M10 and M14), which mix SPIs from different literature categories, revealing interesting connections between different theoretical bases for quantifying pairwise dependencies between time series. Although M3 contains a mix of SPIs based on information theory (six labeled ‘information theory’ measures and one, information-geometric conditional independence, labeled as a ‘causal index’), the remaining three modules establish interesting connections between the behavior of seemingly disparate SPIs on MTS data. Three networks that are derived from these modules are plotted in Fig. 2c–e and are investigated in detail below.

As illustrated in Fig. 2c, module M5 contains a mix of two different types of methods: the first includes five linear estimators for integrated information (such as geometric integrated information, Φ_G (ref. 22), ϕ^* (ref. 16) and stochastic interaction²³), whereas the second includes 16 estimators for Granger causality in both the time and frequency domains²⁴. Although Granger causality and integrated information theory were developed in very different contexts (for example, Granger’s investigations into causality between economic time series in 1969¹¹, versus Tononi’s recent integrated information theory (Φ) of consciousness^{15,16,22}), our analysis reveals that all SPIs in this module nevertheless behave similarly on data (with an empirical similarity index, $R = 0.52$, which lies in the 95th percentile for all R values; see Supplementary Fig. 1c). Recent results have indeed shown that Granger causality can be formulated as the information-theoretic measure, transfer entropy²⁵, and can thus be grouped under the same information-geometric framework as integrated information theory^{22,26}. However, it was not known whether these information-theoretic SPIs behave similarly in practice and, as such, their relationship was not widely recognized. Module M5 thus demonstrates an important confirmation of our empirical approach in being able to recapitulate emerging theory and unify scientific tools for understanding interacting processes.

Module M10, shown in Fig. 2d, highlights striking connections between three conceptually distinct types of methods: (1) dynamic time warping (DTW), which was developed in the data-mining community to quantify the similarity between two (potentially shifted and dilated) audio signals¹⁰; (2) cross-spectral phase-based measures—the maximum phase coherence²¹, and the mean and maximum phase-locking value²⁷—which were developed to examine frequency-specific

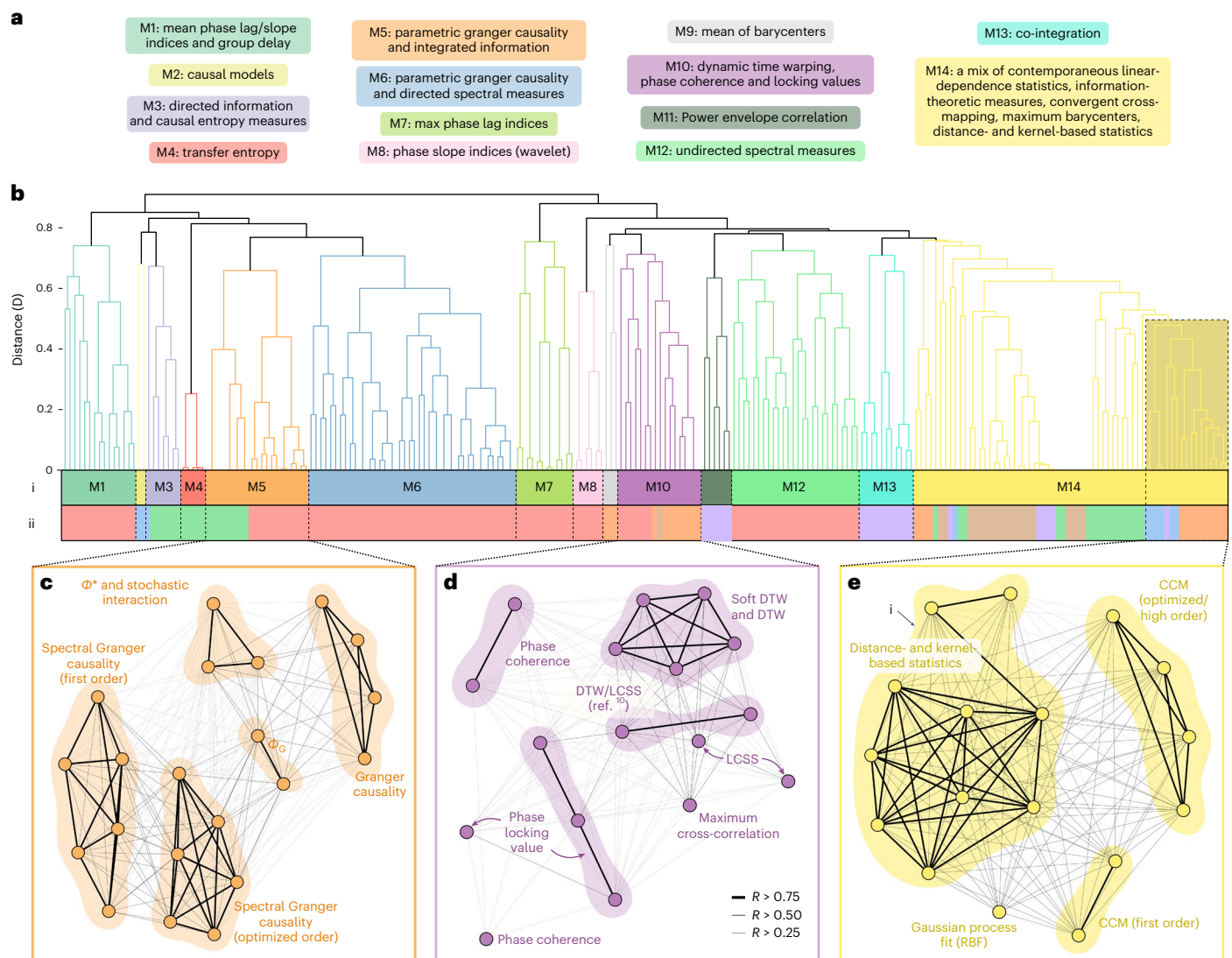


Fig. 2 | Statistics for measuring pairwise interactions between time series can be organized into 14 modules on the basis of their behavior on over 1,000 MTS, providing an intuitive, data-driven organization of interdisciplinary scientific literature. a, A brief summary of the main types of methods in each of the modules (see Supplementary Note 1 for a full list with descriptions). **b**, The dendrogram used to infer the modules, produced by hierarchical clustering using a dissimilarity measure, $D = 1 - R$, that is based on the empirical similarity index, R (see Methods). SPIs are colored according to: (i) their module label (upper row) and (ii) their literature categorization (lower row), displayed using the color scheme defined in Fig. 1b. A high-resolution version of this dendrogram,

including each of the SPIs (leaf nodes) that form the modules, is presented in Supplementary Fig. 2 (and the assignment of SPIs to modules is in Supplementary Data 1). **c–e**, Three selected modules that include a mixture of SPIs developed in different disciplinary contexts are shown as network plots in M5 (**c**), M10 (**d**) and a subset of M14 (**e**). In these plots, SPIs are represented as nodes, three different edge weight thresholds are displayed (corresponding to $R > 0.25$, $R > 0.5$ and $R > 0.75$), and connected components with high similarities, $R > 0.75$, are shaded yellow. DTW, dynamic time warping; LCSS, longest continuous subsequence; CCM, convergent cross-mapping; RBF, radial basis function.

synchronization in neuroimaging data²¹; and (3) the maximum cross-correlation⁶, a classical statistical technique for correlating two time series at different temporal lags. All of these SPIs capture time-lagged interactions between two processes, but in slightly different ways: the maximum cross-correlation finds the highest fixed-lag match, DTW extends this idea by optimizing the distortion of the time axis to best match potentially misaligned time series, and the cross-spectral measures account for time lags in terms of phase differences. Module M10 thus reveals previously unreported connections between diverse approaches to capturing associations between pairs of potentially unaligned time series, indicating a common conceptual basis for methods developed and applied across disciplines—whether they are measuring synchronization between neuroelectric recordings or recognizing speech from audio signals.

Finally, we discuss module M14, which groups 66 SPIs from all literature categories except for ‘spectral’ (see Fig. 1b for categories). This module recapitulates some theoretical relationships that have already been established, such as the equivalence between linear-Gaussian mutual information and absolute correlation²⁸ (with a maximum similarity of $R = 1$). To highlight some previously unreported relationships, we focus on a demonstrative submodule (shown in Fig. 2e) comprising 17 SPIs from the ‘causal indices’, ‘distance similarity’ and ‘miscellaneous’ literature categories. We first note the tight cluster of SPIs, labeled ‘i’ in Fig. 2e, that were developed independently in two different domains: distance correlation-based methods²⁹ from the statistics community, and kernel-based methods from the machine-learning community. This cluster first highlights a recent finding that distance correlation and the Hilbert–Schmidt independence criteria

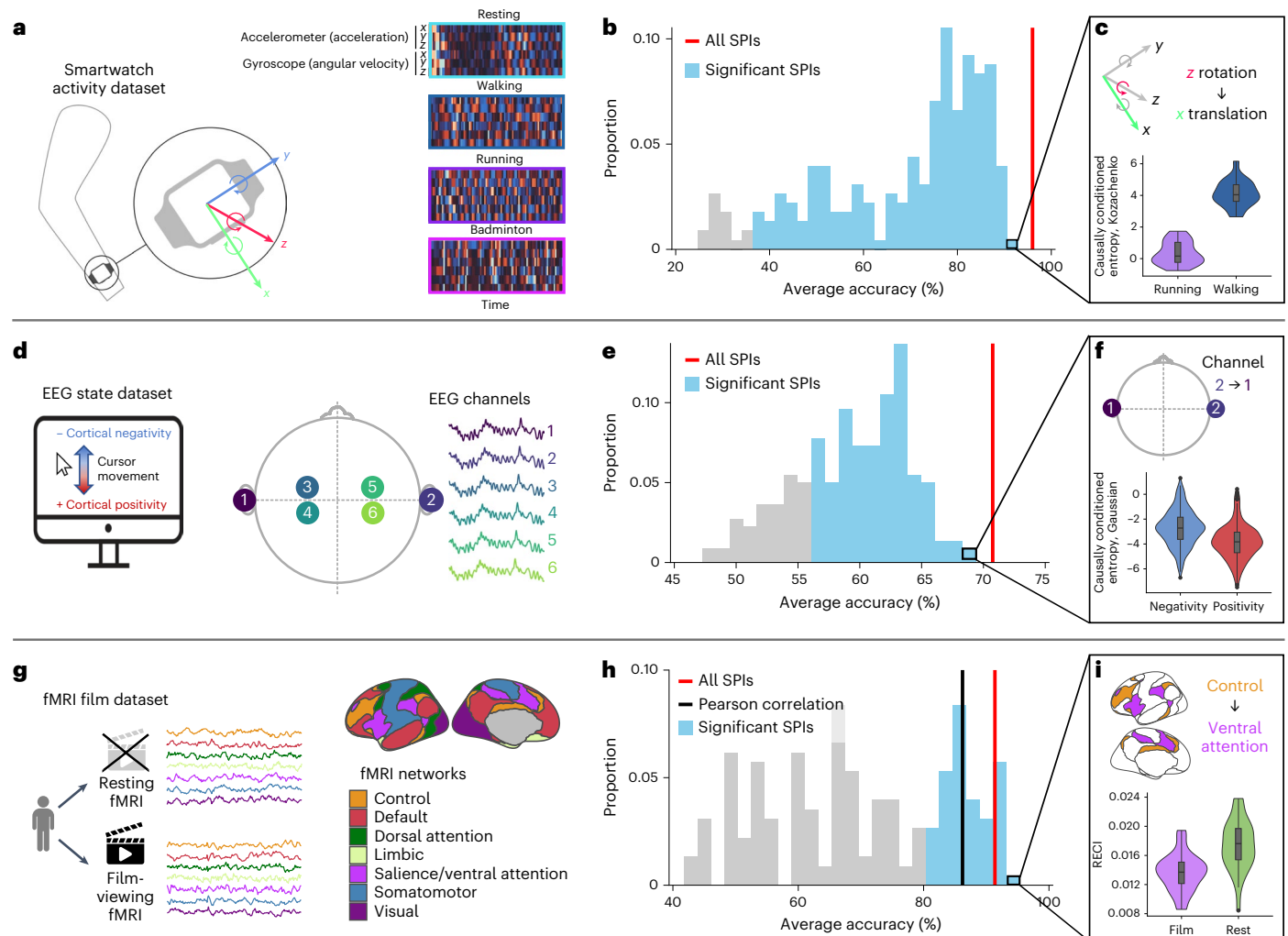


Fig. 3 | A comprehensive library of SPIs can be used to accurately classify and understand differences in human movement and neural activity datasets.

a, In the smartwatch activity dataset, we aimed to determine which of four activities a participant is performing ('resting', 'walking', 'running' or 'badminton') from smartwatch accelerometer recordings. An example MTS from each of the four classes is shown as a process × time heat map (colored as in Fig. 1a). **b**, Distribution of average classification accuracy (over train–test resamples) across 228 SPIs, where all pairwise interactions are used as the basis for classification for each SPI. The 213 SPIs with significant accuracy are shaded blue (one-sided permutation test, Bonferroni-corrected $P < 0.05$). Concatenating feature vectors from all SPIs into a single classification model yields an average accuracy of 96% (red). **c**, Violin plots showing the distribution of the top-performing SPI (causally conditioned entropy with a Kozachenko–Leonenko density estimator) between the wrist's x-axis translation and z-axis rotation, indicating a stronger interaction during walking (mean $\pm$ s.d. = 4.1 ± 0.8 , $N = 20$) than running (0.4 ± 0.7 , $N = 20$). In each violin plot, the annotated box plot shows the box centre and extent as

the median, 25th and 75th percentiles; whiskers extend for $1.5 \times$ the interquartile range and outliers are shown as black dots. **d**, In the EEG state dataset, we aimed to classify positive versus negative cortical activity states using data from six EEG channels, for which the electrode locations and example time series are depicted. **e**, Distribution of average classification accuracy across all 219 SPIs, including the 165 significant SPIs, and a combination of all SPIs (71% accuracy). **f**, The top-performing SPI (causally conditioned entropy with a Gaussian density estimator) is visualized from EEG channel 2 to 1 in cortical negativity (-2.7 ± 1.4 , $N = 282$) versus positivity (-3.8 ± 1.4 , $N = 279$). **g**, In the fMRI film dataset, we aimed to classify rest versus film-viewing using fMRI time series from seven brain networks. **h**, Distribution of classification accuracy across all 227 SPIs, including the 67 significant SPIs (shaded blue) and annotating the performance of Pearson correlation coefficient (86%, black) and the combined set of all 227 SPIs (91%, red) for comparison. **i**, The top-performing SPI (regression error-based causal inference, RECI) is shown from the control to ventral attention networks in film-viewing (0.014 ± 0.003 , $N = 29$) and rest (0.018 ± 0.004 , $N = 29$) conditions.

(HSIC, a kernel-based method)³⁰ are equivalent when computed using certain distance kernels³¹; our results suggest that similar theoretical connections can be established between the other SPIs of the cluster (including the Heller–Heller–Gorfine test³² and multiscale graph correlation²⁹). Second, we find that the distance- and kernel-based statistics display strikingly similar behavior as common implementations of the convergent cross-mapping (CCM) algorithm, which was originally developed for inferring causality in complex ecosystems³. Convergent cross-mapping aims to measure the causal effect of one time series on another by the ability of the second to reconstruct the first using a nearest-neighbor approach. The fact that these methods

behave so similarly on MTS data indicates that the well-studied techniques of phase-space reconstruction (used in the CCM algorithm) have a correspondence to the nonlinear kernel-estimation techniques from the statistics and machine learning communities. These observed connections also have important practical ramifications; for example, our results suggest candidate proxy algorithms to substitute for the computationally expensive CCM, which could yield major computational efficiencies and enable new applications on larger datasets.

In summary, our empirical approach to unifying a large and disjoint literature of methods for characterizing pairwise interactions in time series has allowed us to: (1) capture the commonalities in

the conceptualizations of dynamical interactions (such as whether methods predicate a statistical dependence on self-predictability) and (2) understand empirical and theoretical connections between diverse methods, with implications for shaping theory and practice.

Leveraging diverse methods to address scientific problems

Our results above illustrate the rich diversity of scientific methods for quantifying pairwise interactions. This diversity suggests that, when quantifying pairwise interactions for a given application, there is potential to make comparisons across the scientific literature of SPIs to: (1) select the best-performing SPI in an unbiased, data-driven way; and (2) leverage a synergistic combination of multiple complementary SPIs to better capture complex underlying interactions in MTS. Here we provide a simple demonstration of this strategy (referred to as ‘highly comparative’ due to the broad methodological comparison involved^{12,13}) on three MTS classification problems using three open datasets: (1) the smartwatch activity dataset (Fig. 3a), where the aim is to classify one of four behavioral states (walking, running, resting, or playing badminton) from six-sensor smartwatch MTS (comprising a three-axis accelerometer and three-axis gyroscope)³³; (2) the electroencephalogram (EEG) state dataset (Fig. 3d), where the aim is to distinguish positive versus negative slow cortical potentials from single-subject EEG data (originally used to move a cursor up or down on a computer screen)^{34,35}; and (3) the functional magnetic resonance imaging (fMRI) film dataset (Fig. 3g), where the aim is to classify resting and film-watching conditions from fMRI data³⁶ (see ‘Classification case studies’ section in the Methods for additional details on each dataset). To investigate the performance of different SPIs on these tasks, we represented each MTS as a set of features corresponding to all pairwise interactions between its constituent processes and compared their classification performance using a linear support vector machine (SVM) with cross-validation (see ‘Classification’ section in the Methods for details).

We observed a wide range of SPI performances in each case study, ranging from null performance up to high and statistically significant performance: 27–92% accuracy on the smartwatch activity dataset (Fig. 3b), 48–69% accuracy on the EEG state dataset (Fig. 3e) and 41–95% accuracy on the fMRI film dataset (Fig. 3h). Many SPIs displayed significant classification performance on each dataset (permutation test, Bonferroni-corrected $P < 0.05$): 213 SPIs for the smartwatch activity dataset, 165 SPIs for the EEG state dataset, and 67 SPIs for the fMRI film dataset. This wide range of observed SPI performance on all three datasets demonstrates the crucial importance of selecting an SPI that is able to capture the relevant types of interactions underlying a given dataset.

To understand the types of interactions that characterize the labeled classes of MTS, we analyzed and interpreted the highest-performing SPIs on each dataset (see Supplementary Data 2 for the full results). To provide a simple demonstration of this process, here we focus on the top-performing individual SPI in each case study. In the smartwatch activity dataset, the top-performing SPI was causally conditioned entropy (CCE) using a Kozachenko–Leonenko estimator (92% accuracy; SPI label ‘cce_kozachenko’; see Supplementary Note 1.4.7 for details). Its high performance is driven in part by its ability to capture the coupling from the wrist’s z -rotation to its x -translation, which differs strongly between running and walking, as shown in Fig. 3c. This tells us that wrist rotation in the z -direction is more informative of subsequent elbow movement in the x -direction (as the elbow moves side to side) in walking than in running.

The top-performing SPI for the EEG state dataset was, as above, a causally conditioned entropy, but using a Gaussian density estimator (69%, SPI label ‘cce_gaussian’; see Supplementary Note 1.4.7 for details). Its performance was driven in part by its ability to capture the increased coupling from EEG channel 2 to 1 (from near the right ear to near the left ear) in cortical negativity versus positivity states, as shown in Fig. 3f. Other statistics designed to capture directed information flow (in a way

that includes instantaneous interactions in the presence of feedback) also performed well on this task, including directed information with a Gaussian density estimator (69% accuracy, SPI label ‘di_gaussian’; see Supplementary Note 1.4.8). Although past applications of SPIs derived from directed information theory to neuroimaging data are limited^{37,38}, our highly comparative analysis suggests them as high-performing methods for measuring EEG coupling alongside other novel candidates for further investigation, including the direct directed transfer function evaluated over high frequencies (67%, SPI label ‘ddtf_multi-taper_mean_fs-1_fmin-0-25_fmax-0-5’; see Supplementary Note 1.5.9) and the Hilbert–Schmidt Independence Criterion (66%, SPI label ‘hsic’; see Supplementary Note 1.2.5). Of the classical methods for quantifying EEG connectivity, some are recapitulated as high performers by our data-driven analysis, including mean directed coherence across various frequency bands³⁹ (all 67% accuracy; see Supplementary Note 1.5.10), whereas others exhibited surprisingly low accuracy, such as algorithmic variants of partial directed coherence⁴⁰ (between 55% and 64%).

Finally, for the fMRI film dataset, the top-performing SPI was regression error-based causal inference (95% accuracy; SPI label ‘reci’; see Supplementary Note 1.3.4). Its high performance is driven in part by its ability to capture the stronger coupling from the control network to the ventral attention network during film-watching compared to rest (Fig. 3i). The dominant way of measuring coupling in whole-brain fMRI is to use the Pearson correlation coefficient^{8,41} (annotated in Fig. 3h), which exhibits strong and statistically significant classification accuracy on this problem (86%); however, our data-driven approach highlights 30 alternative SPIs with higher performance (88% to 95%; compare with Supplementary Data 2). These high-performing methods include alternative types of covariance (for example, minimum covariance determinant, 93% accuracy, SPI label ‘cov_MinCovDet’; see Supplementary Note 1.1.1) and precision estimates (for example, using Ledoit–Wolf shrinkage, 91%, SPI label ‘prec_LedoitWolf’; see Supplementary Note 1.1.2) that better deal with non-Gaussian bivariate distributions. Other high performers include information-theoretic SPIs (such as conditional entropy, joint entropy and mutual information using Gaussian density estimators, all of which exhibited 91% accuracy; see Supplementary Note 1.4.1–1.4.3) and directed SPIs which distinguish asymmetric coupling (for example, the top-performing regression error-based causal inference, ‘reci’). Compared with the typically subjective process of selecting an appropriate method to analyze a given dataset, the highly comparative approach demonstrated here highlights the most useful scientific methods automatically, facilitating interpretable understanding of the conceptual formulations of pairwise dependencies that drive successful performance.

As different types of systems involve different types of interactions between measured processes, we expected different SPIs to perform well across the three datasets. Indeed, we found that an SPI with high performance on one problem does not imply its high performance on other problems. In particular, some SPIs performed well on only a single dataset; for example, dynamic time warping with an Itakura parallelogram (SPI label ‘dtw_constraint-itakura’; see Supplementary Note 1.2.7), which was a top performer on the fMRI film dataset (91%), but showed weaker performance on the smartwatch activity dataset (78%) and null performance on the EEG state dataset (53%). Yet some SPIs did perform strongly across all three datasets, such as the cross distance correlation (SPI label ‘dcorr_maxlag-10’; see Supplementary Note 1.2.2), which ranked among the top 10 SPIs for all three problems (90%, 66% and 91%, for the smartwatch activity, EEG state and fMRI film datasets, respectively). Moreover, different algorithmic variants of causally conditioned entropy were top performers for the smartwatch activity (‘cce_kozachenko’, 92%), EEG state (‘cce_gaussian’, 69%) and fMRI film (‘cce_kernel_W-0.5’, 90%) datasets (although we note a strong dependence of the density estimation approach on CCE performance). We also found that grouping SPIs based on the fourteen data-driven modules (identified by their similarity of behavior on data, Fig. 2)

better captured their relative performance on these tasks than the six literature categories (from Fig. 1b), as shown in Supplementary Fig. 3, suggesting our modular representation as a useful one for understanding differential SPI performance on a given task.

Relative to investigating individual SPIs one at a time, we finally aimed to investigate the value of drawing on multiple SPIs simultaneously. We developed a combined representation of the pairwise dependence structures captured by all SPIs, allowing us to simultaneously represent each MTS using a large and diverse set of pairwise dependency structures (through feature concatenation, as described in the ‘Classification’ section of the Methods). Although this approach represents each MTS in a much higher-dimensional space than the individual SPI representation analyzed above (with associated challenges for robust classifier fitting), we expected it to outperform the individual best SPI on datasets involving multiple types of interactions, such that simultaneously leveraging multiple SPIs provides complementary and useful information about class differences. Relative to the top-performing individual SPI, our combined SPI approach improved classification accuracy on the smartwatch activity dataset (to 96%) and the EEG state dataset (to 71%), shown as vertical red lines in Fig. 3b,e. On the fMRI film dataset, it yielded slightly lower performance (91%) than the top individual SPI, ‘reci’ (95%), suggesting that the associated interactions are well-captured by a single, well-chosen SPI on this dataset (that is, multiple SPIs do not provide an advantage sufficient to overcome the challenges of fitting a classifier in a higher-dimensional space). By simultaneously drawing on a wide range of SPIs, the simple statistical approach demonstrated here can quantify multiple complementary types of interactions from MTS data (and is likely to yield improved accuracy through optimization; see ‘Discussion’ section).

Discussion

Our empirical organization of SPIs (Fig. 2b) identifies connections between a rich literature of diverse methods for quantifying interactions between time series, and thereby highlights fruitful directions for future research to consolidate, extend and develop new theory. These results were obtained using a similarity index, R , that averages across a large and diverse range of 1,053 MTS but—as the behavior of an SPI depends on the types of interactions present in a dataset—the resulting relationships between SPIs depend on the MTS we chose to include. As a simple illustrative example, consider the case that our MTS library only contained data generated by lag-1 vector autoregressive models, VAR(1). Then the unique behavior of more sophisticated SPIs (such as those capturing nonlinear dependence or causal coupling on longer time lags) would not be observed, as data containing those types of interactions would be absent. Consequently, the variety of behaviors exhibited by our library of SPIs would be reduced, and the modular structure evident in the dendrogram of Fig. 2b would become far more homogeneous. In constructing the MTS data library used in this work, we thus aimed to be as even-handed and comprehensive as possible in sampling from diverse systems, which has been sufficient to yield a meaningful and useful representation of the interdisciplinary literature on SPIs. However, as any such finite sample is incomplete and relies on subjective decisions in its construction, future work may explore the dataset-dependent similarity of SPI pairs in detail to construct more nuanced organizations of the literature. This would provide new understanding of how different mechanisms (and hence empirical dependency structures) play out in different classes of complex systems. Future work may also revisit the simple methodological choices made here, including our decisions to: (1) quantify the similarity of a pair of SPIs using a single number (R); (2) represent the resulting relationships in one dimension (as a dendrogram); and (3) analyze the resulting dendrogram at a single selected resolution (a 14-module decomposition). For example, more complex unsupervised methods, such as overlapping community-detection methods^{42,43}, could reveal interesting new relationships between SPIs at different resolution scales.

As different time-varying systems contain different statistical relationships between their elements (for example, those that capture instantaneous or time-delayed responses, linear or nonlinear interactions, are conditioned on the past, allow for non-constant delays, or infer directional coupling), our results demonstrate that a highly comparative approach can be used to detect the most informative types of SPIs for capturing the relevant types of interactions underlying a given dataset. Unlike many machine-learning approaches to MTS classification, which can be challenging to interpret³³, this approach connects scientists to interpretable theory that shapes understanding of the most important types of pairwise interactions in a dataset, following recent undertakings to unify interdisciplinary literatures on summary statistics (‘features’) for univariate time series^{12,44,45} and complex networks¹³. This broad, comprehensive methodological comparison stands in contrast to a more conventional approach in which the data analyst manually selects a method—a practice that leaves open the possibility that alternative methods may provide clearer interpretation, better performance, or computational efficiencies. The three classification case studies analyzed here provide a simple demonstration of the procedure, automatically highlighting high-performing SPIs from across the literature and providing interpretable understanding of the relevant interactions in each dataset. We observed a wide range of performance in all cases, highlighting the importance of careful selection of SPIs for a given application. In the EEG state dataset, for example, our analysis flagged high-performing SPIs consistent with common methodological practice in EEG analysis (thus recapitulating existing domain knowledge), and others with surprisingly poor performance, as well as flagging novel high-performing SPIs as promising candidates for methodological innovation in the field. Future work investigating which types of SPIs are best suited to which types of problems will yield new insights into the interactions underlying different complex systems, and is likely to uncover additional novel applications of SPIs to new problems.

Although the ability to compare across the full library of SPIs is powerful, it comes at a computational cost, particularly for larger MTS datasets in which each MTS is sampled over many time points and processes. The high redundancy between many groups of SPIs (Fig. 2), together with the preliminary indications from case studies that some SPIs may exhibit generally high performance, suggest the potential for major efficiency gains by constructing a high-performing and minimally redundant reduced set of SPIs. Beyond a simple approach of selecting a representative SPI from each of the 14 modules, a more systematic approach could compare SPI performance across a wide range of representative MTS tasks, for example, following recent work reducing a large pool of thousands of interdependent univariate time-series features to just 22 representatives⁴⁶. There is also much scope for statistical optimization in applying any given set of SPIs to real-world problems, beyond the simple choices made here. In particular, performance could be optimized by tackling the challenges of high-dimensionality and strong interdependence of SPIs using dimensionality reduction, regularization or feature-selection methods to find small but highly explanatory combinations of complementary SPIs tailored to a given dataset⁴⁷. Given the methodological freedoms involved in highly comparative time-series analysis, pre-registration should be considered for transparency and to avoid bias⁴⁸ (for example, ref. 49).

Beyond the applications to MTS classification presented here, the ability to represent the coupling structure in a MTS using a large and diverse set of pairwise interactions could form the foundation for tackling myriad statistical inference problems, including regression, clustering, anomaly detection, and causal network inference. For example, current methods for inferring networks of causal interactions from MTS data^{50,51} range from heuristic approaches based on thresholding a set of pairwise dependencies defined by a given SPI through to full statistical inference^{50,52–54}. Rather than manually selecting an SPI for this purpose, the ability to compare diverse SPIs provides the flexibility to

capture different types of underlying interactions that could form the basis of improved network-inference algorithms. Such an extension may also go beyond point estimates of pairwise dependence strengths (considered here) towards assessing statistical significance through comparison to an appropriate null model^{9,51,55}. Future work could also explore comprehensive statistical representations of MTS using more diverse properties than just the set of pairwise interactions, by also incorporating properties of univariate dynamics of individual system components⁴⁵, system-level structure in the full set of pairwise interactions represented as a network¹³ and higher-order interactions⁵⁶.

The flexible and extendable software accompanying this work, *pyspi*⁵⁷ (see ‘Code Availability’ section), facilitates the application of our highly comparative approach to quantifying pairwise interactions to myriad applications. Through careful framing of systems-level inference⁵⁵, it has the potential to highlight unexpected and powerful methodological approaches to quantifying interaction patterns in time-varying systems. The data library of over 1,000 MTS provided with this work⁵⁸ (see ‘Data Availability’ section) is another resource that will allow researchers to characterize the behavior of their computational methods on a comprehensive range of real-world and simulated systems, addressing issues associated with only testing new complex systems methods on idealized datasets⁵⁵. In summary, this work demonstrates the utility of an empirical approach to unifying diverse complex dynamical systems and their methods of analysis, providing insights and tools for scientific discovery.

Methods

Comprehensive scientific libraries of methods and data

Library of 237 SPIs. This work takes an empirical approach to understanding the relationships between diverse scientific methods for quantifying pairwise interactions between time series, formulated as an SPI. The first step in this approach was to construct a comprehensive, annotated library of SPIs. Here we introduce a library of 237 SPIs, organized into six broad categories based on their underlying theory: ‘basic’ (such as co-variance, Kendall’s τ (ref. 59) and cross-correlation); ‘distance similarity’ (for example, distance correlations²⁹, kernel-based independence tests^{30,32} and dynamic time warping¹⁰); ‘causal indices’ (including additive noise models and convergent cross-mapping³); ‘information theory’ (such as Granger causality^{11,25} (note that Granger causality is included in the information theory category as it is equivalent to transfer entropy evaluated with a Gaussian estimator²⁵), transfer entropy²⁰ and integrated information¹⁶); ‘spectral’ (derived from Fourier or wavelet transformations, for example, coherence magnitude, phase-locking value²⁷ and spectral Granger causality⁶⁰); and ‘miscellaneous’ (for example, co-integration⁵ and model fits). A full list of SPIs, along with descriptions and references, is presented in Supplementary Note 1.

Library of 1,053 MTS. To understand how each SPI behaves on different types of multivariate time-series data, we constructed a library of 1,053 diverse model-generated and real-world MTS. Our aim in assembling these data was to capture the main classes of systems and dynamics that are studied across scientific disciplines, including synchronization, spatiotemporal chaos, wave propagation, criticality and phase transitions. Our library contains 505 synthetic MTS generated from mathematical models, including: uncorrelated and correlated noise (for example, Cauchy and normally distributed noise and Brownian motion); coupled maps (such as vector autoregression⁶ and coupled map lattices⁶¹); coupled ordinary differential equations (including Kuramoto oscillators⁶², Hodgkin–Huxley and Wilson–Cowan networks); and partial differential equations (namely, wave equations, in which processes are embedded in physical space). It also contains 548 diverse real-world MTS from public databases across: geophysics (such as earthquake seismograms and atmospheric processes); medicine (for example, heartbeat sonograms, fMRI data and EEGs);

physiology (including accelerometer and gyroscope readings for sports and basic motions); and finance (for instance, exchange rates and stock prices), among others. Each MTS comprises between 5 and 40 processes, and between 100 and 2,000 observations—characteristics that were designed to match many real-world datasets. Across all MTS in our library, we have a total of 195,112 pairwise interactions that we used to evaluate the SPIs. Descriptions of the full set of MTS analyzed here are presented in Supplementary Note 2.

Quantifying similarity between SPIs using the empirical similarity index

Our main aim in this work was to assess the similarity of behavior of any two SPIs across a diverse range of MTS. In this section we describe the empirical similarity index, R , used to capture the similarity between the behavior of all pairs of SPIs across our diverse library of 1,053 MTS. This index measures the monotonic relationship between a pair of SPIs across a large number of datasets that have different numbers of processes (and thus a different number of pairwise interactions). For an M -variate time series, $\mathbf{z} = (z_1, \dots, z_M)$, we first consider the $M \times M$ matrix of pairwise interactions (MPI), constructed by evaluating a given SPI for all pairs of M processes within the MTS. The scalar values of the MPI, $S = (s_{ij}) \in \mathbb{R}^{M \times M}$, where the (i, j) entries of the matrix, $s_{ij} = s(z_i, z_j)$, denotes an SPI evaluated on the i th and the j th time series, z_i and z_j , respectively. In general, undirected statistics (for which $s_{ij} = s_{ji}$, such as Kendall’s τ) yield symmetric MPIs, whereas directed statistics (such as transfer entropy, which have $s_{ij} \neq s_{ji}$) yield asymmetric MPIs. Furthermore, some SPIs are signed (for example, correlation coefficients are within $[-1, 1]$), whereas others are unsigned (for example, the distance correlation is within $[0, 1]$). To compare all methods appropriately, we converted signed SPIs to their absolute value, such that all statistics increase with the strength of dependency between z_i and z_j (for example, we analyze the magnitude of the covariance rather than its sign). Some example MPIs are shown in Supplementary Fig. 1a. Having computed MPIs for all SPIs on all 1,053 MTS, we computed the empirical similarity index between each pair of SPIs via a two-step process (depicted in Supplementary Fig. 1b). First, we defined the similarity of a pair of SPIs, k and l , on a given dataset, d . We did this by computing MPIs for k and l (yielding S_{kd} and S_{ld} , respectively) and then computing the absolute value of the Spearman’s rank correlation coefficient between each of their (off-diagonal) entries, $|r_{d,kl}|$. The resulting correlation value, $|r_{d,kl}|$, thus captures the strength of a monotonic relationship between the output of the two SPIs on the dataset d . After repeating this computation for all MTS in our library, we calculated the empirical similarity index, R , as the average of $|r_{d,kl}|$ across all datasets. Although R provides a useful scalar summary of the similarity between a pair of SPIs, it is important to note that some pairs of SPIs have quite a wide distribution of scores, $|r_{d,kl}|$, across datasets, indicating that they yield highly correlated outputs on some MTS (high $|r_{d,kl}|$) but not on others (low $|r_{d,kl}|$). See Supplementary Note 3.1 for more details, including a detailed visual breakdown of $|r_{d,kl}|$ distributions for some selected pairs of SPIs. We found a wide range of similarity indices across pairs of SPIs, from $R \approx 0$ to $R \approx 1$, illustrated by the cumulative distribution function in Supplementary Fig. 1c.

Classification case studies

Datasets. Smartwatch activity dataset. The smartwatch activity dataset is derived from the BasicMotions problem in the University of East Anglia (UEA) MTS classification repository³⁴. Each MTS includes six sensors—a three-axis accelerometer and a three-axis gyroscope—recorded for 10 s at 10 Hz, yielding 1,000 time points. There are 20 MTS in each of the four classes (resting, walking, running or playing badminton) for a total of 80 MTS in the dataset. This dataset was recently analyzed in a large MTS classification challenge³³ in which the data was split 50%–50% into training and test subsets with 30 stratified repeats. The baseline classifier (based on dynamic time warping) achieved 95.25% accuracy; the best algorithm (HIVE–COTE) achieved 100% accuracy³³.

EEG state dataset. The EEG state dataset corresponds to the SelfRegulationSCP1 problem in the UEA MTS classification repository³⁴ and was originally published in work by Birbaumer and colleagues³⁵. Electrical activity was measured from six EEG channels in one participant as they were instructed to move a cursor up or down on a computer screen by generating negative or positive slow cortical potentials, respectively. The physical placement of these EEG electrodes is depicted in Fig. 3d. Cortical activity (measured in microvolts) was recorded for 3.5 s at 256 Hz, yielding 896 time points in each channel per MTS, with 282 negativity trials and 279 positivity trials. This dataset was analyzed in work by Ruiz et al.³³, assessing accuracy using 30 stratified train–test splits of 268 training and 293 test samples. In the original comparison, the baseline DTW algorithm achieved an accuracy of 81.81% and the best algorithm (TapNet) achieved 95.68% (ref. 33).

fMRI film dataset. The fMRI film dataset is derived from a functional connectome fingerprinting study examining individual signatures of cortical activity in $N = 29$ individuals at rest or while watching a film³⁶. In this dataset, blood-oxygen-level-dependent signals were recorded in 114 parcellated cortical regions with a repetition time of 813 ms (a sampling rate of 1.23 Hz) for either 1,200 frames (resting), or between 952 and 1,000 frames (film-watching). Each trial type (rest versus film-watching) was repeated four times per participant. Here we analyzed pre-processed data obtained from work by Betzel and co-workers⁶³. For consistency and simplicity, we examined the first rest and first film-watching session per participant; across all participants and trials, only the first 947 frames contain real data, and thus we restricted our analysis to this time range. To retain a comparable number of processes as the first two classification case studies, we averaged blood-oxygen-level-dependent signals from the 114 original brain regions into the seven functional networks from Yeo et al.⁶⁴, as depicted in Fig. 3g. We compared the performance of the Pearson correlation coefficient, used to construct functional connectivity matrices in the original publication (ref. 36) to our library of SPIs.

Classification. For all three case studies, our simple approach to SPI-based classification involved computing the MPIs (6×6 for the smartwatch activity and EEG state datasets, and 7×7 for the fMRI film dataset) for each z-scored MTS and repeating for each SPI. We then used the elements of these matrices as features for a linear SVM classifier. Note that we used the most recent version of pyspi (v.0.4.0) to compute SPIs for the classification case studies, which included some improved implementations of some SPIs. Features were extracted from each MPI differently for directed and undirected SPIs (see Supplementary Data 1): we used the upper triangular entries as features for undirected SPIs (for which the corresponding MPI is symmetric), whereas we used all non-diagonal elements as features for directed SPIs. As a preprocessing step, for each case study, we removed any SPI that had invalid entries (due to numerical issues) in any of the MPIs, or gave constant results across all MTS (see Supplementary Data 3 for a list of omitted SPIs). This yielded a set of 228 SPIs for the smartwatch activity problem, 219 SPIs for the EEG state problem and 227 SPIs for the fMRI film problem. For the analysis involving combining all SPIs into a single classifier, this yielded a total of 4,755 features for the smartwatch activity dataset, 4,659 features for the EEG state dataset and 6,743 features for the fMRI film dataset.

The linear SVM was implemented using default settings from Python's scikit-learn⁶⁵ as part of a classification pipeline that involved z-score feature normalization (fitted on training data and applied to unseen test data). The very simple methodological choices made here allowed us to focus on demonstrating the key conceptual types of analyses made possible by drawing on a diverse set of SPIs, aiding transparency while acknowledging that more complicated statistical methodologies are likely to improve the classification performance quoted here. We implemented 30 class-stratified train–test splits for

cross-validation with the same proportions implemented in ref. 33 using the StratifiedShuffleSplit function for the smartwatch activity and EEG state datasets. As there are $N = 29$ individuals in the fMRI film dataset, we implemented leave-one-individual-out cross-validation, such that each classifier was trained with the rest and film scans of $N = 28$ participants, and tested on the rest and film fMRI scans of the left-out participant.

We measured classification performance using total accuracy for all three case studies. Statistical significance was estimated using a permutation testing approach whereby 100 null models were fitted (using randomly shuffled class labels) and evaluated using the same cross-validation classification procedure described above. The observed classification performance for each SPI was then compared with the pooled null distribution of all SPIs, yielding P -values that were then adjusted for multiple comparisons by controlling the family-wise error at 0.05 using the method of Bonferroni⁶⁶. The performance metric for the union of all SPIs was similarly compared with its corresponding null permutation distribution to yield a single P -value per classification problem.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The full database of 1,053 diverse real-world and simulated MTS analyzed here is available on the Zenodo repository at: <https://doi.org/10.5281/zenodo.7118947> (ref. 58). This resource could be used to test scientific methods on a diverse sample of MTS. Time-series data used in the three case studies are from open sources, as described in the 'Classification case studies' section in the Methods. Source data are provided with this paper.

Code availability

Accompanying this paper is an extendable Python-based package called pyspi⁵⁷, which includes implementations of all 237 SPIs. This software allows users to compare the behavior of an interdisciplinary literature on methods for quantifying interactions between pairs of time series. Furthermore, code and instructions for reproducing the results and figures presented in this work can be found on the Zenodo repository at: <https://doi.org/10.5281/zenodo.8027702> (ref. 67).

References

- Bassett, D. S. & Sporns, O. Network neuroscience. *Nat. Neurosci.* **20**, 353–364 (2017).
- Buckner, R. L. et al. Molecular, structural, and functional characterization of Alzheimer's disease: evidence for a relationship between default activity, amyloid, and memory. *J. Neurosci.* **25**, 7709 (2005).
- Sugihara, G. et al. Detecting causality in complex ecosystems. *Science* **338**, 496–500 (2012).
- Runge, J. et al. Inferring causation from time series in Earth system sciences. *Nat. Commun.* **10**, 2553 (2019).
- Engle, R. F. & Granger, C. W. Co-integration and error correction: representation, estimation, and testing. *Econometrica* **55**, 251–276 (1987).
- Reinsel, G. C. *Elements of Multivariate Time Series Analysis* (Springer, 2003).
- Stephens, Z. D. *PLoS Biol.* **13**, e1002195 (2015).
- van den Heuvel, M. P. & Hulshoff Pol, H. E. Exploring the brain network: a review on resting-state fMRI functional connectivity. *Eur. Neuropsychopharmacol.* **20**, 519 (2010).
- Cliff, O. M., Novelli, L., Fulcher, B. D., Shine, J. M. & Lizier, J. T. Assessing the significance of directed and multivariate measures of linear dependence between time series. *Phys. Rev. Res.* **3**, 013145 (2021).

10. Sakoe, H. & Chiba, S. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Signal Process* **26**, 43 (1978).
11. Granger, C. W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 424 (1969).
12. Fulcher, B. D., Little, M. A. & Jones, N. S. Highly comparative time-series analysis: the empirical structure of time series and their methods. *J. R. Soc. Interface* **10**, 20130048 (2013).
13. Peach, R. L. et al. HCGA: Highly comparative graph analysis for network phenotyping. *Patterns* **2**, 100227 (2021).
14. Massey, J. Causality, feedback and directed information. In *International Symposium on Information Theory and its Applications (ISITA-90)* 303–305 (IEEE, 1990).
15. Tononi, G. An information integration theory of consciousness. *BMC Neurosci.* **5**, 42 (2004).
16. Oizumi, M., Amari, S.-i, Yanagawa, T., Fujii, N. & Tsuchiya, N. Measuring integrated information from the decoding perspective. *PLoS Comput. Biol.* **12**, e1004654 (2016).
17. Wiener, N. in *Modern Mathematics for Engineers* (ed. Beckenback, E.) (McGraw-Hill, 1956).
18. Pesaran, M. H. & Shin, Y. An autoregressive distributed lag modelling approach to cointegration analysis. In *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium* (ed. Strøm, S.) (Cambridge Univ. Press, 1999).
19. Nolte, G. et al. Robustly estimating the flow direction of information in complex physical systems. *Phys. Rev. Lett.* **100**, 234101 (2008).
20. Schreiber, T. Measuring information transfer. *Phys. Rev. Lett.* **85**, 461 (2000).
21. Bastos, A. M. & Schoffelen, J.-M. A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Front. Syst. Neurosci.* <https://doi.org/10.3389/fnsys.2015.00175> (2016).
22. Oizumi, M., Tsuchiya, N. & Amari, S.-I. Unified framework for information integration based on information geometry. *Proc. Natl Acad. Sci. USA* **113**, 14817 (2016).
23. Ay, N. Information geometry on complexity and stochastic interaction. *Entropy* **17**, 2432 (2015).
24. Geweke, J. Measurement of linear dependence and feedback between multiple time series. *J. Am. Stat. Assoc.* **77**, 304 (1982).
25. Barnett, L., Barrett, A. B. & Seth, A. K. Granger causality and transfer entropy are equivalent for Gaussian variables. *Phys. Rev. Lett.* **103**, 238701 (2009).
26. Cliff, O. M., Prokopenko, M. & Fitch, R. Minimising the Kullback–Leibler divergence for model selection in distributed nonlinear systems. *Entropy* **20**, 51 (2018).
27. Lachaux, J.-P., Rodriguez, E., Martinerie, J. & Varela, F. J. Measuring phase synchrony in brain signals. *Hum Brain Mapp* **8**, 194–208 (1999).
28. MacKay, D. J. in *Information Theory, Inference and Learning Algorithms* (Cambridge Univ. Press, 2003).
29. Shen, C., Priebe, C. E. & Vogelstein, J. T. From distance correlation to multiscale graph correlation. *J. Am. Stat. Assoc.* **115**, 280 (2020).
30. Gretton, A. et al. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems* (eds Roweis, S. et al.) Vol. 20 (Curran Associates, 2008).
31. Sejdinovic, D., Sriperumbudur, B., Gretton, A. & Fukumizu, K. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Annals Stat.* **41**, 2263–2291 (2013).
32. Heller, R., Heller, Y. & Gorfine, M. A consistent multivariate test of association based on ranks of distances. *Biometrika* **100**, 503–510 (2013).
33. Ruiz, A. P., Flynn, M., Large, J., Middlehurst, M. & Bagnall, A. The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Min. Knowl. Discov.* **35**, 401–449 (2021).
34. Bagnall, A. et al. The UEA multivariate time series classification archive. Preprint at <https://arxiv.org/abs/1811.00075> (2018).
35. Birbaumer, N. et al. A spelling device for the paralysed. *Nature* **398**, 297–298 (1999).
36. Byrge, L. & Kennedy, D. P. Accurate prediction of individual subject identity and task, but not autism diagnosis, from functional connectomes. *Hum. Brain Mapp.* **41**, 2249 (2020).
37. Liu, Y. & Aviyente, S. Quantification of effective connectivity in the brain using a measure of directed information. *Comput. Math Methods Med.* **2012**, 635103 (2012).
38. Mehta, K. & Kliewer, J. Directional and causal information flow in EEG for assessing perceived audio quality. *IEEE Trans. Mol. Biol. Multiscale Commun.* **3**, 150–165 (2017).
39. Wang, G. & Takigawa, M. Directed coherence as a measure of interhemispheric correlation of EEG. *Int J. Psychophysiol.* **13**, 119–128 (1992).
40. Schelter, B. et al. Testing for directed influences among neural signals using partial directed coherence. *J. Neurosci. Methods* **152**, 210–219 (2006).
41. Smith, S. M. et al. Network modelling methods for FMRI. *NeuroImage* **54**, 875–891 (2011).
42. Evans, T. & Lambiotte, R. Line graphs, link partitions, and overlapping communities. *Phys. Rev. E* **80**, 016105 (2009).
43. Ahn, Y.-Y., Bagrow, J. P. & Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature* **466**, 761–764 (2010).
44. Fulcher, B. D. Feature-based time-series analysis. In *Feature Engineering for Machine Learning and Data Analytics* 87–116 (CRC, 2018).
45. Fulcher, B. D. & Jones, N. S. *hctsa*: a computational framework for automated time-series phenotyping using massive feature extraction. *Cell Syst.* **5**, 527–531.e3 (2017).
46. Lubba, C. H. et al. *catch22*: CAnonical Time-series CHaracteristics. *Data Min. Knowl. Discov.* **33**, 1821–1852 (2019).
47. Phinyomark, A. et al. Navigating features: A topologically informed chart of electromyographic features space. *J. R. Soc.* **14**, 20170734 (2017).
48. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T. The preregistration revolution. *Proc. Natl Acad. Sci. USA* **115**, 2600 (2018).
49. Leung, A. et al. Towards blinded classification of loss of consciousness: distinguishing wakefulness from general anesthesia and sleep in flies using a massive library of univariate time series analyses. Preprint at <https://psyarxiv.com/rmsv8/> (2023).
50. Peixoto, T. P. Network reconstruction and community detection from dynamics. *Phys. Rev. Lett.* **123**, 128301 (2019).
51. Banerjee, A., Chandra, S. & Ott, E. Network inference from short, noisy, low time-resolution, partial measurements: application to *C. elegans* neuronal calcium dynamics. *Proc Natl Acad. Sci. USA* **120**, e2216030120 (2023).
52. Hoffmann, T., Peel, L., Lambiotte, R. & Jones, N. S. Community detection in networks without observing edges. *Sci Adv* **6**, eaav1478 (2020).
53. McCabe, S. et al. Netrd: A library for network reconstruction and graph distances. *J. Open Source Softw.* **6**, 2990 (2021).
54. Wollstadt, P. et al. IDTxL: The Information Dynamics Toolkit xl: a Python package for the efficient analysis of multivariate information dynamics in networks. *J. Open Source Softw.* **4**, 1081 (2019).
55. Peel, L., Peixoto, T. P. & De Domenico, M. Statistical inference links data and theory in network science. *Nat. Commun.* **13**, 6794 (2022).
56. Battiston, F. et al. The physics of higher-order interactions in complex systems. *Nat. Phys.* **17**, 1093–1098 (2021).
57. Cliff, O. M. *DynamicsAndNeuralSystems/pyspi* (Zenodo, 2023); <https://doi.org/10.5281/zenodo.8223340>

58. Cliff, O. M. *Library of Multivariate Time Series* (Zenodo, 2022); <https://doi.org/10.5281/zenodo.7118947>
59. Kendall, M. G. A new measure of rank correlation. *Biometrika* **30**, 81 (1938).
60. Dhamala, M., Rangarajan, G. & Ding, M. Estimating Granger causality from Fourier and wavelet transforms of time series data. *Phys. Rev. Lett.* **100**, 018701 (2008).
61. Kaneko, K. & Tsuda, I. *Complex systems: Chaos and Beyond* (Springer, 2011).
62. Strogatz, S. H. From Kuramoto to Crawford: exploring the onset of synchronization in populations of coupled oscillators. *Physica D* **143**, 1–20 (2000).
63. Betzel, R., Kennedy, D. & Byrge, L. *Resting-State and Movie-Watching Data* (Figshare, 2020); <https://doi.org/10.6084/m9.figshare.12971162.v2>
64. Yeo, B. T. et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* **106**, 1125–1165 (2011).
65. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825 (2011).
66. Bonferroni, C. *Teoria Statistica Delle Classi e Calcolo Delle Probabilita* Vol. 8, 3 (Seeber, 1936).
67. Bryant, A. G. *DynamicsAndNeuralSystems/pyspi_paper_classification: Documentation ft. readme and code comments* (Zenodo, 2023); <https://doi.org/10.5281/zenodo.8027702>

Acknowledgements

O.M.C., N.T. and B.D.F. were supported by NHMRC Ideas Grant 1183280. N.T. was supported by Japan Society for the Promotion of Science, Grant-in-Aid for Transformative Research Areas (20H05710, 23H04830, 23H04829). A.G.B. was supported by an Australian Government Research Training Program Scholarship, an American Australian Association Graduate Education Fund Scholarship, and the University of Sydney Physics Foundation. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the paper. High-performance computing facilities provided by the School of Physics, the University of Sydney contributed to our results.

Author contributions

N.T. and B.D.F. conceived of the project. O.M.C. developed the software and large data repositories and performed the main analysis. A.G.B. performed the classification case study analyses. B.D.F. supervised the project with input from J.T.L. and N.T. O.M.C. and B.D.F. wrote the paper, with input from all other co-authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s43588-023-00519-x>.

Correspondence and requests for materials should be addressed to Ben D. Fulcher.

Peer review information *Nature Computational Science* thanks the anonymous reviewers for their contribution to the peer review of this work. Primary Handling Editor: Jie Pan, in collaboration with the *Nature Computational Science* team. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | | |
|-------------------------------------|--|
| n/a | Confirmed |
| ☐ | ☒ The exact sample size (<i>n</i>) for each experimental group/condition, given as a discrete number and unit of measurement |
| ☐ | ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| ☐ | ☒ The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| ☐ | ☒ A description of all covariates tested |
| ☐ | ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| ☐ | ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| ☐ | ☒ For null hypothesis testing, the test statistic (e.g. <i>F</i> , <i>t</i> , <i>r</i>) with confidence intervals, effect sizes, degrees of freedom and <i>P</i> value noted
<i>Give P values as exact values whenever suitable.</i> |
| ☒ | ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| ☒ | ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| ☐ | ☒ Estimates of effect sizes (e.g. Cohen's <i>d</i> , Pearson's <i>r</i>), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection	Data analyzed in this study were generated via numerical simulation, as well as data downloaded from publicly available sources. See Supplementary Section 2 for full details of the datasets simulated and analyzed.
Data analysis	The pyspi software package was developed by us; it accompanies this work. Classification analyses were performed using pyspi v0.4.0: https://github.com/DynamicsAndNeuralSystems/pyspi (https://doi.org/10.5281/zenodo.8223340). A docker image for pyspi is available here: https://hub.docker.com/r/arianguyen/pyspi Analysis and visualizations were done using a combination of R (v4.3.0) and python (v3.9.7). Full code and data to reproduce the presented analyses are in https://github.com/DynamicsAndNeuralSystems/pyspi_paper_code (https://doi.org/10.5281/zenodo.8027702). Python package versions are as follows: pandas (v1.5.2), numpy (v1.22.4), scikit-learn (v0.24.1), pyarrow (v10.0.1), scipy (v1.7.3), statsmodels (v0.12.1), pyyaml (v5.4), tqdm (v4.50.2), nitime (v0.9), hyppo (v0.2.1), pyEDM (v1.9.3.0), jpype1 (v1.2.0), sktime (v0.8.0), dill (v0.3.2), spectral-connectivity (v0.2.4.dev0), torch (v1.10.0), cdt (v0.5.23), oct2py (v5.2.0), tslearn (v0.5.2), mne (v0.23.0), seaborn (v0.11.0). R package versions are as follows: tidyverse (v2.0.0), glue (v1.6.2), reticulate (v1.28), icesTAF (v4.2.0), cowplot (v1.1.1), plyr (v1.8.8).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Data analyzed in this study were both generated via numerical simulation and assembled from publicly available sources. See Supplementary Section 2 for full details and sources. The full set of multivariate time-series datasets analyzed are available in the following Zenodo repository: <https://zenodo.org/record/7118947>.

Human research participants

Policy information about [studies involving human research participants and Sex and Gender in Research](#).

Reporting on sex and gender	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Our interdisciplinary library of 1053 diverse multivariate time series aimed to be as comprehensive as possible. This large collection of MTS was curated with the aim of capturing the main classes of systems and dynamics that are studied across scientific disciplines, including synchronization, spatiotemporal chaos, wave propagation, criticality, and phase transitions. Our aim was to encompass a sufficiently diverse sample of the types of data analyzed in the scientific literature as comprehensively as possible within practical constraints, such that the empirical similarity index evaluated in this work constitutes a meaningful measure of the behavioral similarity of two statistics of pairwise interaction (SPIs).
Data exclusions	We included MTS that contained between 5–40 processes and between 100–2000 observations. We set this range both for practical purposes (avoiding large calculation times required of larger datasets), and because these characteristics match those of real-world datasets that are commonly analyzed.
Replication	This study aims to understand the relationship between different algorithms for quantifying statistical dependence between a pair of time series. We note in the Discussion that, since the behavior of an SPI depends on the types of interactions present in a dataset, the organization we obtain depends on the data included in this set of MTS. Please see the relevant paragraph in Discussion for additional discussion of this point, including why we might expect different structures to be found when analyzing data obtained from different types of systems. Cross-validation and permutation-based statistics were applied to the classification case studies, as described in Methods: 'Classification'.
Randomization	Randomization is not applicable to this study as it does not involve an experimental design that involves measuring data from participants.
Blinding	Blinding is not applicable to this study as it does not involve an experimental design that involves measuring data from participants.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).
Research sample	State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.
Sampling strategy	Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.
Data collection	Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.
Timing	Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.
Data exclusions	If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.
Non-participation	State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.
Randomization	If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.
Research sample	Describe the research sample (e.g. a group of tagged <i>Passer domesticus</i> , all <i>Stenocereus thurberi</i> within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.
Sampling strategy	Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.
Data collection	Describe the data collection procedure, including who recorded the data and how.
Timing and spatial scale	Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken
Data exclusions	If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.
Reproducibility	Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.
Randomization	Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.
Blinding	Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.

Did the study involve field work? ☐ Yes ☒ No

Field work, collection and transport

Field conditions	<i>Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).</i>
Location	<i>State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).</i>
Access & import/export	<i>Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).</i>
Disturbance	<i>Describe any disturbance caused by the study and how it was minimized.</i>

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	<i>Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.</i>
Validation	<i>Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.</i>

Eukaryotic cell lines

Policy information about [cell lines and Sex and Gender in Research](#)

Cell line source(s)	<i>State the source of each cell line used and the sex of all primary cell lines and cells derived from human participants or vertebrate models.</i>
Authentication	<i>Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.</i>
Mycoplasma contamination	<i>Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.</i>
Commonly misidentified lines (See ICLAC register)	<i>Name any commonly misidentified cell lines used in the study and provide a rationale for their use.</i>

Palaeontology and Archaeology

Specimen provenance	<i>Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information). Permits should encompass collection and, where applicable, export.</i>
Specimen deposition	<i>Indicate where the specimens have been deposited to permit free access by other researchers.</i>

Dating methods

If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other research organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research, and [Sex and Gender in Research](#)

Laboratory animals

For laboratory animals, report species, strain and age OR state that the study did not involve laboratory animals.

Wild animals

Provide details on animals observed in or captured in the field; report species and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Reporting on sex

Indicate if findings apply to only one sex; describe whether sex was considered in study design, methods used for assigning sex. Provide data disaggregated for sex where this information has been collected in the source data as appropriate; provide overall numbers in this Reporting Summary. Please state if this information has not been collected. Report sex-based analyses where performed, justify reasons for lack of sex-based analysis.

Field-collected samples

For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about [clinical studies](#)

All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration

Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.

Study protocol

Note where the full trial protocol can be accessed OR if not available, explain why.

Data collection

Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.

Outcomes

Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.

Dual use research of concern

Policy information about [dual use research of concern](#)

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No | Yes

- ☒ ☐ Public health
☒ ☐ National security
☒ ☐ Crops and/or livestock
☒ ☐ Ecosystems
☒ ☐ Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No Yes

- ☒ ☐ Demonstrate how to render a vaccine ineffective
- ☒ ☐ Confer resistance to therapeutically useful antibiotics or antiviral agents
- ☒ ☐ Enhance the virulence of a pathogen or render a nonpathogen virulent
- ☒ ☐ Increase transmissibility of a pathogen
- ☒ ☐ Alter the host range of a pathogen
- ☒ ☐ Enable evasion of diagnostic/detection modalities
- ☒ ☐ Enable the weaponization of a biological agent or toxin
- ☒ ☐ Any other potentially harmful combination of experiments and agents

ChIP-seq

Data deposition

- ☐ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- ☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links

May remain private before publication.

For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.

Files in database submission

Provide a list of all files available in the database submission.

Genome browser session

(e.g. [UCSC](#))

Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.

Methodology

Replicates

Describe the experimental replicates, specifying number, type and replicate agreement.

Sequencing depth

Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.

Antibodies

Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.

Peak calling parameters

Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.

Data quality

Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.

Software

Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

- ☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☐ All plots are contour plots with outliers or pseudocolor plots.
- ☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.

Instrument

Identify the instrument used for data collection, specifying make and model number.

Software	<i>Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.</i>
Cell population abundance	<i>Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.</i>
Gating strategy	<i>Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.</i>

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	<i>Indicate task or resting state; event-related or block design.</i>
Design specifications	<i>Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.</i>
Behavioral performance measures	<i>State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).</i>

Acquisition

Imaging type(s)	<i>Specify: functional, structural, diffusion, perfusion.</i>
Field strength	<i>Specify in Tesla</i>
Sequence & imaging parameters	<i>Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.</i>
Area of acquisition	<i>State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.</i>
Diffusion MRI	☐ Used ☐ Not used

Preprocessing

Preprocessing software	<i>Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).</i>
Normalization	<i>If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.</i>
Normalization template	<i>Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.</i>
Noise and artifact removal	<i>Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).</i>
Volume censoring	<i>Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.</i>

Statistical modeling & inference

Model type and settings	<i>Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).</i>
Effect(s) tested	<i>Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.</i>
Specify type of analysis:	☐ Whole brain ☐ ROI-based ☐ Both
Statistic type for inference (See Eklund et al. 2016)	<i>Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.</i>
Correction	<i>Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).</i>

Models & analysis

n/a	Involvement in the study
☐	☒ Functional and/or effective connectivity
☒	☐ Graph analysis
☒	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

We compared over 230 measures of functional connectivity in the "fMRI film" case study.

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.