Investigating Information Flows in Spiking Neural Networks With High Fidelity

David Shorten

A thesis submitted in fulfilment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Engineering
The University of Sydney

2022

To Mom, Pops, Gogo, Granny, Granddad and Grandpa

ABSTRACT

The brains of many organisms are capable of a wide variety of complex computations. This capability must be undergirded by a more general purpose *computational capacity*. The exact nature of this capacity, how it is distributed across the brains of organisms and how it arises throughout the course of development is an open topic of scientific investigation.

Individual neurons are widely considered to be the fundamental computational units of brains. Moreover, the finest scale at which large-scale recordings of brain activity can be performed is the spiking activity of neurons and our ability to perform these recordings over large numbers of neurons and with fine spatial resolution is increasing rapidly. This makes the spiking activity of individual neurons a highly attractive data modality on which to study neural computation.

The framework of information dynamics has proven to be a successful approach towards interrogating the capacity for general purpose computation. It does this by revealing the atomic information processing operations of information storage, transfer and modification. Unfortunately, the study of information flows and other information processing operations from the spiking activity of neurons has been severely hindered by the lack of effective tools for estimating these quantities on this data modality. This thesis remedies this situation by presenting an estimator for information flows, as measured by Transfer Entropy (TE), that operates in continuous time on event-based data such as spike trains. Unlike the previous approach to the estimation of this quantity, which discretised the process into time bins, this estimator operates on the raw inter-spike intervals. It is demonstrated to be far superior to the previous discrete-time approach in terms of consistency, rate of convergence and bias. Most importantly, unlike the discrete-time approach, which requires a hard tradeoff between capturing fine temporal precision or history effects occurring over reasonable time intervals, this estimator can capture history effects occurring over relatively large intervals without any loss of temporal precision.

This estimator is applied to developing dissociated cultures of cortical rat neurons, therefore providing the first high-fidelity study of information flows on spiking data. It is found that the spatial structure of the flows locks in to a significant extent. at the point of their emergence and that certain nodes occupy specialised computational roles as either transmitters, receivers or mediators of information flow. Moreover, these roles are also found to lock in early.

In order to fully understand the structure of neural information flows, however, we are required to go beyond pairwise interactions, and indeed multivariate information flows have become an important tool in the inference of effective networks from neuroscience data. These are directed networks where each node is connected to a minimal set of sources which maximally reduce the uncertainty in its present state. However, the application of multivariate information flows to the inference of effective networks from spiking data has been hampered by the above-mentioned issues

with preexisting estimation techniques. Here, a greedy algorithm which iteratively builds a set of parents for each target node using multivariate transfer entropies, and which has already been well validated in the context of traditional discretely sampled time series, is adapted to use in conjunction with the newly-developed estimator for event-based data. The combination of the greedy algorithm and continuous-time estimator is then validated on simulated examples for which the ground truth is known.

The new capabilities in the estimation of information flows and the inference of effective networks on event-based data presented in this work represent a very substantial step forward in our ability to perform these analyses on the ever growing set of high resolution, large scale recordings of interacting neurons. As such, this work promises to enable substantial quantitative insights in the future regarding how neurons interact, how they process information, and how this changes under different conditions such as disease.

Declaration of Authorship

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any other degree. I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

David Shorten, March 2022.

AUTHORSHIP ATTRIBUTION STATEMENT

Parts of this thesis have appeared in the following publications and submitted manuscripts during the candidature for this degree.

Chapter 3 of this thesis is published as:

> D. P. Shorten, R. E. Spinney, and J. T. Lizier, "Estimating transfer entropy in continuous time between neural spike trains or other event-based data," *PLoS Computational Biology*, vol. 17, no. 4, e1008054, 2021.

I contributed to the conceptualization of the work, designed the methodology, wrote the software, wrote the first draft of the paper and edited the paper.

Chapter 4 of this thesis is published as:

> D. P. Shorten, V. Priesemann, M. Wibral, and J. T. Lizier, "Early lock-in of structured and specialised information flows during neural development," *eLife*, vol. 11, e74651, Mar. 2022, ISSN: 2050-084X. DOI: 10.7554/eLife.74651. [Online]. Available: https://doi.org/10.7554/eLife.74651.

This article is distributed under the terms of a Creative Commons Attribution License that permits unrestricted use and redistribution provided that the original author and source are credited. I designed the research, analyzed the data, wrote the first draft of the paper and edited the paper.

Chapter 5 of this thesis will be published as:

> D. P. Shorten, V. Priesemann, M. Wibral, and J. T. Lizier, "Inferring effective networks of spiking neurons using a continuous-time estimator of transfer entropy," unpublished, N.D.

I contributed to the conceptualization of the work, wrote the software, wrote the first draft of the paper and edited the paper.

David Shorten, March 2022.

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

A/Prof. Joseph Lizier, March 2022.

Acknowledgements

The work that is presented here is to a large degree a product of the institutions that I have been a part of and the communities and individuals that I have been lucky enough to interact with. They have provided essential input into and support of this work.

On an institutional level, I would like to thank the South African government and the University of Cape Town for providing extensive funding towards the earlier parts of my tertiary education. None of this would have been possible without the foundational education that you provided. I would also like to thank the University of Sydney for funding the research presented in this thesis as well as providing a fantastic environment in which to study. This research was also funded by the Australian Research Council DECRA grant DE160100630 and The University of Sydney Research Accelerator (SOAR) Fellowship program.

None of the work in this thesis would have been possible without the patient and dedicated supervision of my primary supervisor Joseph Lizier. Joe, I am deeply appreciative of the amount of time, effort and expertise you have dedicated towards my thesis. I have also appreciated how you have been able to balance serious science with a sense of humour within the research group. I will always have particularly fond memories of our group Zoom meetings during the Covid lock-downs, which were able to lighten up this challenging period.

Richard Spinney played a vital role in the development of the continuous-time TE estimator presented in this thesis. Our discussions on point processes and the bias and consistency of estimators were invaluable both in terms of deriving the result and helping me develop my understanding of the topic.

The Complex Systems Research group at The University of Sydney has been a fantastic environment in which to conduct this research. My thanks go to Mikhail for leading the group as well as the many group members I have been lucky enough to interact with during my time there. As well as our great discussions on science and politics, we were able to have great fun decorating cones, testing the blending properties of various materials and verifying the performance of Suzukis.

Finally, my thanks goes to my family. Your love and support is what allowed me to begin my journey in Australia and your emails, letters, Zooms and Skypes were essential in me completing it.

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Information Flow in Spiking Neural Networks

It is evident that brains possess the ability to perform advanced computations in a highly distributed manner [1], [2]. This ability for computation requires an intrinsic information processing capacity (see Section 2.2 for a formal definition of this capacity and related terms). However, there remain numerous unanswered questions pertaining to the exact nature of this capacity. How is it distributed over neural systems? Do different brain regions specialise in different information processing operations? At what stage in development does it emerge? Are there specific patterns or relationships in its emergence?

It is worth asking at what scale this computational ability should be investigated. The computations performed by brains are carried out by their neurons acting in a distributed, coordinated fashion [3], [4]. These cells can communicate with each other through a number of mechanisms, including through chemical signalling. However, it is widely acknowledged that the dominant form of communication between neurons is through changes in the electrical potential on the membrane of the cells (that is, changes in the difference in the electrical potential between the interior and exterior of the cell) [3], [5]. Although recordings of neural systems are performed at a variety of spatial scales, the most fine-grained recordings that are capable of recording from multiple neurons are measuring this membrane potential (or some correlate of it) [6], [7]. As this data allows us to interrogate neural systems at the level of the individual computational units, it is of great interest to the neuroscience community. The membrane potentials of neurons are characterised by highly-pronounced near-instantaneous spikes. Each spike is commonly referred to as an action potential [3]. It is generally accepted that the primary method of electrical communication between neurons is through this spiking activity, where the change in potential in the pre-synaptic neuron induces a change in potential in the post-synaptic neuron after a short synaptic delay[3], [5].

As such, the initial processing of membrane potential recordings usually involves the extraction of the times of these spikes [8]. This pre-processed data is usually referred to as a spike train. The spike times captured in these spike trains are considered to represent the sensory information the brain receives [9], [10], to encode the communication from one brain region to another [11], [12] as well as capture the spontaneous dynamics of networks [13]. As spike trains contain this information and do so at the finest spatial scale readily available, they represent an incredibly important data modality within neuroscience, particularly for the inference of information flows in order to reveal computations at this fine scale.

Information dynamics [14], [15] is a framework grounded in Shannon's information theory [16]–[18], that has proven remarkably successful at revealing how, when and where information is intrinsically processed in the interactions of activity in complex systems. It does this by first considering the uncertainty of the current state of a given system component. This uncertainty can be rigorously defined and measured using the Shannon entropy [16]–[18]. We can then study the reductions in this uncertainty that are provided by knowing the histories of various system components. Specifically, the information storage of a given component is measured by the reduction in uncertainty provided by knowledge of its own history [15]. Similarly, the information transfer from a source to a target, as defined by the Transfer Entropy (TE) [19], is measured as the reduction in the uncertainty of the target's present state provided by the knowledge of the source's history, conditioned on the target's own history. See Section 2.2 for formal definitions of these quantities. Information modification can be measured by decomposing the information transfer into synergistic, unique and redundant components [20], [21]. However, as its definition and measurement is still a topic of ongoing research, it is not a focus of this thesis.

The information dynamics framework has already been successfully used to reveal the computational properties of a variety of systems. A number of studies [22]–[24] have focused on the changes in information dynamics as networks move from ordered to chaotic dynamics through a critical transition. This work concluded that there are computational advantages to the network dynamics being situated at the critical transition by increasing the system's capacity for information storage and transfer. More specifically, they found that TE and active information storage were either both maximised at or near this critical point or that an optimal tradeoff between these two quantities was found in this region.

Other work [25], [26] has demonstrated that information dynamics measures can be used as a useful early warning of changes in the regime of network dynamics. These measures were shown to be a good indicator of cascading failures in energy networks [25]. In oscillator networks [26], it was found that the TE was able to indicate synchrony substantially earlier than domain-specific measures, such as the order parameter. They were also able to elucidate some of the computational mechanisms [23] involved in the approach of the network towards synchronization.

Given the success of information dynamics at revealing these computational structures [27], in such a variety of systems, it is natural to query whether it could be similarly exploited to uncover the computational properties of biological neural systems, particularly from spike-train data. Indeed, information dynamics, in particular the information flow as measured by TE, has already been widely applied to neuroscientific data. Among these applications include the interrogation of the complex, dynamic, structure of information transfer revealed by calcium imaging [28], fMRI [29], [30], MEG [31] and EEG [32]–[35], the role of reduced information storage in autism spectrum disorders [36], brain-heart information flows [37], the relationship between changes in storage and transfer in gain-mediated phase transitions [38], changes in information modification in developing neural networks [39], and the role of information storage in representing visual stimuli [40].

One particular application, for which there has been a surge of recent interest, is the inference of connectivity from neuroscience data [41]–[43] from sources such as EEG [35], fMRI [30], calcium imaging [28] and electrode arrays [44]. These networks usually fall into one of two categories: *functional* or *effective*. In functional networks [45], [46], an edge is placed between two given system components

based on some pairwise measure of their statistical dependence. In effective networks [47], [48], by contrast, the goal is to find a minimal set of parents that can explain the activity of a target component. Information flow, as measured by TE, has become an important measure for inferring these networks [32], [49], [50]. Moreover, the use of multivariate TE, when paired with a greedy inference algorithm, has been thoroughly validated for use with standard estimators applied to regularly-sampled time series [49]. There are numerous advantages to this approach, including that it can capture nonlinear effects, unlike the commonly used correlation measures such as the Pearson correlation. Moreover, given its grounding within the information dynamics framework, networks inferred using TE are capable of illuminating the computational signature [38] of the neural system. This grounding in information theory also provides a natural interpretation of the resulting networks. Specifically, for every incoming edge for a given target node, the source nodes of these incoming edges represent the set of nodes whose states' maximally reduce the uncertainty of the given target node's state updates.

1.2 Challenges and Objectives

Given the importance of spike train data and the demonstrated utility of both the information dynamics framework and network inference, as described in Section 1.1, the application of these techniques to this data type holds great promise. Indeed, there have been a number of previous studies which have used TE to interrogate spike trains [39], [44], [51]–[58]. These studies primarily focused on inferring directed functional networks, finding that they exhibit a highly non-random structure [55], including rich-club topologies [54]. Other work [39], [58] has focussed on how the components of information can be decomposed into unique, redundant and synergystic components, as well as how some flows can be localised on certain time scales [44]. See Section 2.7.2 for more detail on this previous work.

However, despite the valuable insights provided by this existing literature, the application of TE to this data modality have been hindered by the available estimation techniques. Previous applications of TE to spike train data have made use of a discrete-time estimator. This estimator operates by dividing the process into bins of width Δt. Each bin is then given a binary value corresponding to whether or not there was any spiking activity in that bin (it could also be assigned a natural number corresponding to the number of spikes that occurred in the bin). A simple plugin estimator (see Section 2.3.1) is then applied to this discretised data. This estimation strategy does, however, suffer from a number of drawbacks which have impeded such work:

1. As time discretisation is a lossy transformation, the resulting estimator is not consistent. That is, in general it does not converge to the true value of the TE in the limit of infinite data (see Section 3.2.2).

2. The estimated TE values exhibit a strong dependence on the size of bin chosen [44].

3. Any estimator is going to suffer from the curse of dimensionality. This greatly limits the number of bins that can be used in the history embeddings which aim to capture the statistical relationship between process histories and current state. This implies that the only way that the discrete-time estimator can capture history effects occurring over larger time scales is to increase the bin size. However, this will decrease the estimator's ability to capture effects occurring over fine time scales [44].

4. Relating to point 3, as effective use of the discrete-time estimator requires the use of many bins in the history embeddings, it has trouble handling multiple conditional processes as the dimensionality rises too rapidly as additional processes are added. This limits its use in the context of the inference of effective networks, as this requires considering information contributions collectively or conditionally from multiple processes [49].

5. The discrete-time estimator converges very slowly with the amount of available data (see Section 3.2.2).

6. The discrete-time estimator is usually paired with a time-shift method for surrogate generation in order to test for statistically significant non-zero TE values. Examples can be found where this method yields incredibly high false-positive rates (see Section 3.2.3).

The first goal of this thesis is to overcome these problems that have been present in previous studies. A promising approach for doing so lies in the recently-developed continuous-time framework for information dynamics [59], [60]. This framework might facilitate the development of an estimator which will allow us to estimate TE, without sacrificing time precision, whilst still being capable of capturing history effects that occur over long periods of time. The core desired features of this estimator are:

1. No loss of time precision.

2. Consistency — it must converge to the true value of the TE in the limit of infinite data.

3. The ability to represent reasonably long histories efficiently, that is, with few dimensions.

4. An associated surrogate generation scheme which will allow for accurate testing for non-zero TE.

This thesis also sets out to apply such an estimator to spike-train recordings from biological data and thus reveal neural information flows at high fidelity for the first time ("high-fidelity" describes how we will be performing estimation with fine temporal precision and capturing long-range history effects). Moreover, all previous applications of TE to neuroscience data have studied recordings from mature animals or cultures. This thesis sets out to fill this gap in the research literature by studying the changes in the information dynamics of a developing neural cell culture.

Developing an estimator for TE which is able to represent history embeddings efficiently will make it possible to use TE for the inference of effective networks from spiking data, using the greedy algorithm already validated on regularly-sampled time series, as discussed above [49]. We therefore intend to validate this capability with the new estimator that we develop and therefore open the door to the inference of effective connectivity using information flows for this data modality.

1.3 Contributions of this Thesis

This thesis develops a novel estimator for TE on event-based data (such as spike trains) that operates in continuous time (without time binning), then applies this estimator to data from neural recordings in order to understand how patterns of information flow emerge during neural development, and

validates its use in the inference of effective networks. As such, the core contribution of this thesis is the first study of information flows on spike trains with high fidelity. The overall contribution is split across the three articles that I contributed during my candidature.

The first article, presented in Chapter 3, *Estimating transfer entropy in continuous time between neural spike trains or other event-based data* [61], presents a novel estimator for TE on event-based data which operates in continuous time. This estimator is demonstrated to able to circumvent the challenges presented by the discrete-time estimator that were discussed in Section 1.2 and further satisfies the desired properties listed in there. A vital property of this new estimator is that, unlike the discrete-time estimator, it is provably consistent. Moreover, as it operates in continuous time, it does not suffer from any loss of time precision and its results are not dependent on the choice of a bin width. Fundamental to its operation is that it makes use of inter-spike intervals (the time between successive spikes) to represent history embeddings. This allows it to represent reasonably long histories, with few dimensions and no loss of time precision, thus opening the door to its use in the inference of effective networks. Validation experiments were performed which demonstrated that it has substantially lower bias than the discrete-time approach and converged orders of magnitude faster. This paper also presents an adaptation of a recently-proposed local permutation test [62] for conditional mutual information for use in conjunction with the presented estimator. This method of generating surrogates was validated to generate surrogates which conformed to the null hypothesis of zero TE. Further, it was demonstrated on simulated data that, in some cases, it was capable of achieving far lower false-positive rates for conditionally independent processes than the traditional approach of time-shifted surrogates.

The second article, presented in Chapter 4, *Early lock-in of structured and specialised information flows during neural development* [63], applies this estimator to recordings from developing cultures of dissociated cortical rat neurons [64]. As such, it is the first high-fidelity study of information flows in spike-train data. Moreover, as this paper estimates the information flow at various different points in the development of the cultures, it is the first work to track changes in information flows longitudinally during neural development. It was found in this work that the information flow structure exhibits a marked early lock-in phenomenon, whereby the information flow of the mature network is highly correlated with the flows early in development. It was also revealed that nodes that burst in the middle of the burst propagation occupy a unique computational role as the mediators of information flow, confirming prior speculation about the role of these nodes.

A unique benefit of the high-fidelity estimator presented in the first article is its highly-efficient representation of history embeddings. This allows for successful estimation, even in the case of large conditioning sets, therefore opening up the possibility of performing the inference of effective networks. The third article, presented in Chapter 5, *Inferring networks of spiking neurons using a continuous-time estimator of TE*, validates the use of the continuous-time estimator when used in conjunction with a (slightly modified) existing approach to effective network inference using TE [49], [65]. It is the first article to demonstrate the use of TE in the inference of effective networks from spike trains. This article validates the presented approach on simulated networks of spiking neurons for which the ground truth structural connectivity is known. The continuous-time TE approach is demonstrated to have excellent precision and recall for networks in a wide variety of dynamical regimes. It achieves this performance on relatively large networks consisting of 50 nodes and 5 sources

per target.

This new-found ability to estimate TE on spike trains with high-fidelity opens up exciting research opportunities. As discussed in Chapter 6, this partially involves the application of the estimator to recordings of spiking neurons collected using more modern techniques. Such techniques can allow for the collection of continuous long-term recordings, which allows for spike sorting [66]. Alternative techniques allow for recordings to be performed at incredibly high spatial resolution [67]. Applying the new estimator to such large scale, high resolution recordings will deliver a much higher fidelity understanding of the nature of neural information flows, and how they relate to higher level function, disease, and neural computation.

1.4 References

[1] A. J. Maren, C. T. Harston, and R. M. Pap, *Handbook of neural computing applications*. Academic Press, 2014.

[2] S. Navlakha and Z. Bar-Joseph, "Distributed information processing in biological and computational systems," *Communications of the ACM*, vol. 58, no. 1, pp. 94–102, 2014.

[3] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. Siegelbaum, A. J. Hudspeth, and S. Mack, *Principles of neural science*. McGraw-hill New York, 2000, vol. 4.

[4] T. Trappenberg, *Fundamentals of computational neuroscience*. OUP Oxford, 2009.

[5] M. Bear, B. Connors, and M. A. Paradiso, *Neuroscience: Exploring the Brain, Enhanced Edition: Exploring the Brain*. Jones & Bartlett Learning, 2020.

[6] N. A. Steinmetz, C. Aydin, A. Lebedeva, *et al.*, "Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings," *Science*, vol. 372, no. 6539, 2021.

[7] L. Cong, Z. Wang, Y. Chai, *et al.*, "Rapid whole brain imaging of neural activity in freely behaving larval zebrafish (danio rerio)," *Elife*, vol. 6, e28158, 2017.

[8] P. Dayan and L. F. Abbott, *Theoretical neuroscience: computational and mathematical modeling of neural systems*. Computational Neuroscience Series, 2001.

[9] F. Rieke, D. Warland, R. d. R. Van Steveninck, and W. Bialek, *Spikes: exploring the neural code*. 1999.

[10] E. N. Brown, R. E. Kass, and P. P. Mitra, "Multiple neural spike train data analysis: State-of-the-art and future challenges," *Nature Neuroscience*, vol. 7, no. 5, pp. 456–461, 2004.

[11] S. L. Keeley, D. M. Zoltowski, M. C. Aoi, and J. W. Pillow, "Modeling statistical dependencies in multi-region spike train data," *Current Opinion in Neurobiology*, vol. 65, pp. 194–202, 2020.

[12] N. A. Steinmetz, P. Zatka-Haas, M. Carandini, and K. D. Harris, "Distributed coding of choice, action and engagement across the mouse brain," *Nature*, vol. 576, no. 7786, pp. 266–273, 2019.

[13] M. B. Feller, "Spontaneous correlated activity in developing neural circuits," *Neuron*, vol. 22, no. 4, pp. 653–656, 1999.

[14] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, "A framework for the local information dynamics of distributed computation in complex systems," in *Guided Self-Organization: Inception*, Springer, 2014, pp. 115–158.

[15] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, "Local measures of information storage in complex distributed computation," *Information Sciences*, vol. 208, pp. 39–54, 2012.

[16] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[17] D. J. MacKay, *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

[18] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[19] T. Schreiber, "Measuring information transfer," *Physical Review Letters*, vol. 85, no. 2, p. 461, 2000.

[20] C. Finn and J. T. Lizier, "Pointwise partial information decomposition using the specificity and ambiguity lattices," *Entropy*, vol. 20, no. 4, p. 297, 2018.

[21] P. L. Williams and R. D. Beer, "Nonnegative decomposition of multivariate information," *arXiv preprint arXiv:1004.2515*, 2010.

[22] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, "The information dynamics of phase transitions in random boolean networks.," in *ALIFE*, 2008, pp. 374–381.

[23] J. Boedecker, O. Obst, J. T. Lizier, N. M. Mayer, and M. Asada, "Information processing in echo state networks at the edge of chaos," *Theory in Biosciences*, vol. 131, no. 3, pp. 205–213, 2012.

[24] J. T. Lizier, S. Pritam, and M. Prokopenko, "Information dynamics in small-world boolean networks," *Artificial Life*, vol. 17, no. 4, pp. 293–314, 2011.

[25] J. T. Lizier, M. Prokopenko, and D. J. Cornforth, "The information dynamics of cascading failures in energy networks," in *Proceedings of the European Conference on Complex Systems (ECCS), Warwick, UK*, Citeseer, 2009, p. 54.

[26] R. V. Ceguerra, J. T. Lizier, and A. Y. Zomaya, "Information storage and transfer in the synchronization process in locally-connected networks," in *2011 IEEE Symposium on Artificial Life (ALIFE)*, IEEE, 2011, pp. 54–61.

[27] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, "Coherent information structure in complex computation," *Theory in Biosciences*, vol. 131, no. 3, pp. 193–203, 2012.

[28] J. G. Orlandi, O. Stetter, J. Soriano, T. Geisel, and D. Battaglia, "Transfer entropy reconstruction and labeling of neuronal connections from simulated calcium imaging," *PloS One*, vol. 9, no. 6, e98842, 2014.

[29] V. Maki-Marttunen, I. Diez, J. M. Cortes, D. R. Chialvo, and M. Villarreal, "Disruption of transfer entropy and inter-hemispheric brain functional connectivity in patients with disorder of consciousness," *Frontiers in Neuroinformatics*, vol. 7, p. 24, 2013.

[30] J. T. Lizier, J. Heinzle, A. Horstmann, J.-D. Haynes, and M. Prokopenko, "Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fmri connectivity," *Journal of Computational Neuroscience*, vol. 30, no. 1, pp. 85–107, 2011.

[31] M. Wibral, B. Rahm, M. Rieder, M. Lindner, R. Vicente, and J. Kaiser, "Transfer entropy in magnetoencephalographic data: Quantifying information flow in cortical and cerebellar networks," *Progress in Biophysics and Molecular Biology*, vol. 105, no. 1-2, pp. 80–97, 2011.

[32] M. H. I. Shovon, N. Nandagopal, R. Vijayalakshmi, J. T. Du, and B. Cocks, "Directed connectivity analysis of functional brain networks during cognitive activity using transfer entropy," *Neural Processing Letters*, vol. 45, no. 3, pp. 807–824, 2017.

[33] C.-S. Huang, N. R. Pal, C.-H. Chuang, and C.-T. Lin, "Identifying changes in eeg information transfer during drowsy driving by transfer entropy," *Frontiers in human neuroscience*, vol. 9, p. 570, 2015.

[34] S. Stramaglia, G.-R. Wu, M. Pellicoro, and D. Marinazzo, "Expanding the transfer entropy to identify information circuits in complex systems," *Physical Review E*, vol. 86, no. 6, p. 066 211, 2012.

[35] D. Marinazzo, O. Gosseries, M. Boly, *et al.*, "Directed information transfer in scalp electroencephalographic recordings: Insights on disorders of consciousness," *Clinical EEG and Neuroscience*, vol. 45, no. 1, pp. 33–39, 2014.

[36] A. Brodski-Guerniero, M. J. Naumer, V. Moliadze, *et al.*, "Predictable information in neural signals during resting state is reduced in autism spectrum disorder," *Human brain mapping*, vol. 39, no. 8, pp. 3227–3240, 2018.

[37] L. Faes, G. Nollo, F. Jurysta, and D. Marinazzo, "Information dynamics of brain–heart physiological networks during sleep," *New Journal of Physics*, vol. 16, no. 10, p. 105 005, 2014.

[38] M. Li, Y. Han, M. J. Aburn, *et al.*, "Transitions in information processing dynamics at the whole-brain network level are driven by alterations in neural gain," *PLoS Computational Biology*, vol. 15, no. 10, e1006957, 2019.

[39] M. Wibral, C. Finn, P. Wollstadt, J. T. Lizier, and V. Priesemann, "Quantifying information modification in developing neural networks via partial information decomposition," *Entropy*, vol. 19, no. 9, p. 494, 2017.

[40] M. Wibral, J. Lizier, S. Vögler, V. Priesemann, and R. Galuske, "Local active information storage as a tool to understand distributed neural information processing," *Frontiers in Neuroinformatics*, vol. 8, p. 1, 2014.

[41] O. Sporns, *Networks of the Brain*. MIT press, 2010.

[42] J. D. Medaglia, M.-E. Lynall, and D. S. Bassett, "Cognitive network neuroscience," *Journal of Cognitive Neuroscience*, vol. 27, no. 8, pp. 1471–1491, 2015.

[43] R. F. Betzel, "Network neuroscience and the connectomics revolution," in *Connectomic Deep Brain Stimulation*, Elsevier, 2022, pp. 25–58.

[44] N. Timme, S. Ito, M. Myroshnychenko, *et al.*, "Multiplex networks of cortical and hippocampal neurons revealed at different timescales," *PloS One*, vol. 9, no. 12, e115764, 2014.

[45] P. Uhlhaas, G. Pipa, B. Lima, *et al.*, "Neural synchrony in cortical networks: History, concept and current status," *Frontiers in Integrative Neuroscience*, vol. 3, p. 17, 2009.

[46] P. J. Uhlhaas and W. Singer, "Neural synchrony in brain disorders: Relevance for cognitive dysfunctions and pathophysiology," *Neuron*, vol. 52, no. 1, pp. 155–168, 2006.

[47] D. S. Bassett and O. Sporns, "Network neuroscience," *Nature Neuroscience*, vol. 20, no. 3, pp. 353–364, 2017.

[48] F. de Vico Fallani, J. Richiardi, M. Chavez, and S. Achard, "Graph analysis of functional brain networks: Practical issues in translational neuroscience," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1653, p. 20 130 521, 2014.

[49] L. Novelli, P. Wollstadt, P. Mediano, M. Wibral, and J. T. Lizier, "Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing," *Network Neuroscience*, vol. 3, no. 3, pp. 827–847, 2019.

[50] L. Novelli and J. T. Lizier, "Inferring network properties from time series using transfer entropy and mutual information: Validation of multivariate versus bivariate approaches," *Network Neuroscience*, vol. 5, no. 2, pp. 373–404, 2021.

[51] B. Gourévitch and J. J. Eggermont, "Evaluating information transfer between auditory cortical neurons," *Journal of Neurophysiology*, vol. 97, no. 3, pp. 2533–2543, 2007.

[52] M. Garofalo, T. Nieus, P. Massobrio, and S. Martinoia, "Evaluation of the performance of information theory-based methods and cross-correlation to estimate the functional connectivity in cortical networks," *PloS One*, vol. 4, no. 8, e6482, 2009.

[53] S. Ito, M. E. Hansen, R. Heiland, A. Lumsdaine, A. M. Litke, and J. M. Beggs, "Extending transfer entropy improves identification of effective connectivity in a spiking cortical network model," *PLoS One*, vol. 6, no. 11, e27431, 2011.

[54] S. Nigam, M. Shimono, S. Ito, *et al.*, "Rich-club organization in effective connectivity among cortical neurons," *Journal of Neuroscience*, vol. 36, no. 3, pp. 670–684, 2016.

[55] M. Shimono and J. M. Beggs, "Functional clusters, hubs, and communities in the cortical microconnectome," *Cerebral Cortex*, vol. 25, no. 10, pp. 3743–3757, 2015.

[56] E. Matsuda, T. Mita, J. Hubert, *et al.*, "Multiple time scales observed in spontaneously evolved neurons on high-density cmos electrode array," in *Artificial Life Conference Proceedings 13*, MIT Press, 2013, pp. 1075–1082.

[57] M. Kajiwara, R. Nomura, F. Goetze, *et al.*, "Inhibitory neurons exhibit high controlling ability in the cortical microconnectome," *PLoS Computational Biology*, vol. 17, no. 4, e1008846, 2021.

[58] N. M. Timme, S. Ito, M. Myroshnychenko, *et al.*, "High-degree neurons feed cortical computations," *PLoS Computational Biology*, vol. 12, no. 5, e1004858, 2016.

[59] R. E. Spinney and J. T. Lizier, "Characterizing information-theoretic storage and transfer in continuous time processes," *Physical Review E*, vol. 98, no. 1, p. 012 314, 2018.

[60] R. E. Spinney, M. Prokopenko, and J. T. Lizier, "Transfer entropy in continuous time, with applications to jump and neural spiking processes," *Physical Review E*, vol. 95, no. 3, p. 032 319, 2017.

[61] D. P. Shorten, R. E. Spinney, and J. T. Lizier, "Estimating transfer entropy in continuous time between neural spike trains or other event-based data," *PLoS Computational Biology*, vol. 17, no. 4, e1008054, 2021.

[62] J. Runge, "Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2018, pp. 938–947.

[63] D. P. Shorten, V. Priesemann, M. Wibral, and J. T. Lizier, "Early lock-in of structured and specialised information flows during neural development," *eLife*, vol. 11, e74651, Mar. 2022, ISSN: 2050-084X. DOI: 10.7554/eLife.74651. [Online]. Available: https://doi.org/10.7554/eLife.74651.

[64] D. A. Wagenaar, J. Pine, and S. M. Potter, "An extremely rich repertoire of bursting patterns during the development of cortical cultures," *BMC Neuroscience*, vol. 7, no. 1, pp. 1–18, 2006.

[65] J. Sun, D. Taylor, and E. M. Bollt, "Causal network inference by optimal causation entropy," *SIAM Journal on Applied Dynamical Systems*, vol. 14, no. 1, pp. 73–106, 2015.

[66] J. Kreutzer, L. Ylä-Outinen, A.-J. Mäki, M. Ristola, S. Narkilahti, and P. Kallio, "Cell culture chamber with gas supply for prolonged recording of human neuronal cells on microelectrode array," *Journal of Neuroscience Methods*, vol. 280, pp. 27–35, 2017.

[67] X. Yuan, M. Schröter, M. E. J. Obien, *et al.*, "Versatile live-cell activity analysis platform for characterization of neuronal dynamics at single-cell and network level," *Nature Communications*, vol. 11, no. 1, pp. 1–14, 2020.

CHAPTER 2

BACKGROUND

The first three sections of this chapter provide an overview of some fundamental background knowledge that is necessary for what follows in the remainder of the thesis. As this thesis requires a working knowledge of information theory, Section 2.1 provides an introduction to the necessary quantities in this field. Section 2.2 then builds on these fundamentals by introducing the core quantities within the information dynamics framework that will be used in this thesis, including TE as a measure of information flow. As a key focus of this thesis is the derivation and application of a novel information-theoretic estimator, Section 2.3 provides a thorough background on the k-nearest-neighbours class of information-theoretic estimators before Section 2.4 explains the common methods of surrogate generation that are paired with these estimators in order to test for values that are statistically significant against a null hypothesis of zero directed relationship.

The remaining sections aim to provide more general background knowledge in order to assist the reader in situating this work within the wider literature. As Chapter 4 infers functional networks and Chapter 5 infers effective networks from spike-train data, Section 2.5 provides some context on the difference between these two types of network inference in the neuroscientific context. Section 2.6 then describes how information theory more generally has been applied to spike train data, before Section 2.7.1 and Section 2.7.2 discuss aspects of the application of TE to this data type. Finally, Section 2.7.3 briefly describes developments deriving a continuous-time framework for information dynamics, which will be used in the derivation of the TE estimator for spiking data presented in this thesis.

2.1 Information Theory

2.1.1 Fundamental Quantities

This section introduces the core information-theoretic quantities that will be relevant to this thesis. Table 2.1 contains a glossary of notation for quick reference.

2.1.1.1 Entropy

The fundamental quantity of information theory, as developed by Claude Shannon [5], is the entropy [1], [2]. The entropy of a discrete random variable X with n outcomes and probability distribution p is

Symbol	Description	Reference
$H_p(X)$	The entropy of the random variable X, distributed according to $p(X)$	[1], [2]
$H_p(X \mid Y)$	The conditional entropy of the random variable X, conditioned on Y, distributed according to $p(X, Y)$	[1], [2]
$H_{p,q}(X)$	The cross entropy between the distributions $p(X)$ and $q(X)$	[1], [2]
$I(X;Y)$	The mutual information between the random variables X and Y, distributed according to $p(X, Y)$	[1], [2]
$I(X;Y \mid Z)$	The conditional mutual information between the random variables X and Y, given Z, distributed according to $p(X, Y, Z)$	[1], [2]
$D_{KL}(p \,\|\| q)$	The Kullback–Leibler (KL) divergence between the distributions $p(X)$ and $q(X)$	[1], [2]
$\dot{\mathbf{S}}_X$	The Active Information Storage (AIS) rate of the random process X	[3]
$\dot{\mathbf{T}}_{Y \to X}$	The Transfer Entropy (TE) rate from the source process Y to the target process X.	[4]
$\dot{\mathbf{T}}_{Y \to X \mid Z}$	The conditional Transfer Entropy rate from the source process Y to the target process X, conditioned on the third process Z.	[4]

Table 2.1: List of notation used in this section.

defined as:

$$H_p(X) \equiv -\sum_{i=1}^{n} p(x_i) \log p(x_i). \tag{2.1}$$

Entropy is a measure of the uncertainty in a variable. If we consider flipping a coin, for instance, if the coin is fully biased (guaranteed to give either heads or tails) then the entropy will be zero. There is no uncertainty. Conversely, the entropy will be maximised if the coin is completely unbiased (there is a 0.5 chance of heads or tails). This is the case where we are most uncertain about the outcome of the coin flip.

Entropy is usually measured in either bits or nats. The former corresponds to using a base two logarithm in (2.1) and the latter corresponds to using the natural logarithm. In our coin flipping example, the maximum entropy, occurring when the coin is completely unbiased, is 1 bit.

2.1.1.2 Conditional Entropy

We can also define a conditional entropy, which measures the uncertainty in a variable X given that we know the outcome of another variable Y. If X has n outcomes and Y has m then the conditional

entropy is:

$$H_p(X \mid Y) \equiv -\sum_{i=1,j=1}^{n,m} p(x_i, y_j) \log p(x_i \mid y_j). \tag{2.2}$$

2.1.1.3 Mutual Information

The conditional entropy allows us to then define the central quantity of information theory, the mutual information I, as the reduction in uncertainty in a variable X that comes from knowing a second variable Y. It is, therefore, the difference between the entropy of X and the conditional entropy of X given Y:

$$\begin{aligned} I(X;Y) &\equiv H_p(X) - H_p(X \mid Y) \\ &= \sum_{i=1,j=1}^{n,m} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}. \end{aligned} \tag{2.3}$$

Mutual information is a measure of dependence between X and Y. It is zero if and only if X and Y are independent. Unlike other correlation measures, such as the usual Pearson correlation [6], it is capable of capturing any nonlinear relationship between X and Y. It is also symmetric, that is $I(X;Y) = I(Y;X)$.

2.1.1.4 Conditional Mutual Information

We can also consider the conditional mutual information, which is the reduction in uncertainty in a variable X that comes from knowing a second variable Y, given that we know the outcome of a third variable Z. This can be defined as the difference in the information provided by the joint variable (Y, Z) and the variable Z on its own:

$$\begin{aligned} I(X;Y \mid Z) \equiv& I(X;Y,Z) - I(X;Z) \\ &= \sum_{i=1,j=1,k=1}^{n,m,l} p(x_i, y_j, z_j) \log \frac{p(x_i, y_j \mid z_k)}{p(x_i \mid z_k)p(y_j \mid z_k)}. \end{aligned} \tag{2.4}$$

2.1.1.5 Cross Entropy

This thesis makes use of the Kullback-Leibler divergence [1], [2] in a number of places, and so we will introduce it in the next subsection. Doing so will, however, require us to first define the cross entropy. Suppose that we now have one variable X, with n outcomes, and two probability distributions p and q defined on X. The cross entropy of q with respect to p is then:

$$H_{p,q}(X) \equiv -\sum_{i=1}^{n} p(x_i) \log q(x_i). \tag{2.5}$$

Note that this measure is not symmetric: $H_{p,q}(X)$ is not usually equal to $H_{q,p}(X)$. One possible way of interpreting cross entropy is as the average uncertainty we perceive about X, if we believe that it is distributed as q, whereas it is actually distributed as p.

2.1.1.6 Kullback-Leibler Divergence

The Kullback-Leibler (KL) divergence [1], [2] is a measure of how different two probability distributions are from one another. If we have a variable X, again with n outcomes, and two probability distributions p and q defined on X. The KL divergence is defined as:

$$\begin{aligned} D_{KL}(p \,||\, q) &\equiv H_{p,q}(X) - H_p(X) \\ &= \sum_{i=1}^{n} p(x_i) \log \frac{p(x_i)}{q(x_i)}. \end{aligned} \tag{2.6}$$

This measure is not symmetric and so is not a distance metric. Relating this back to the above interpretation of cross entropy, we can think of the KL divergence as how much our uncertainty about X is increased if we believe that it is distributed as q, as opposed to the correct distribution p. It is worth noting that the mutual information defined in Section 2.1.1.3 is the KL divergence between the joint distribution $p(X, Y)$ and the product of the marginal distributions $p(X)p(Y)$.

2.1.2 Defining Quantities for Continuous Variables

The previous section discussed and defined the various information-theoretic quantities in terms of discrete variables. This was done for ease of exposition. However, in many cases, including in much of this thesis, we are interested in variables that can take on continuous values. It is necessary, therefore, to provide and briefly discuss the definition of the relevant information-theoretic quantities on continuous-valued variables. We will begin with the definitions of the differential entropy and cross entropy, before using these quantities to define the continuous KL-divergence and mutual information.

2.1.2.1 Differential Entropy and Cross Entropy

For a continuous random variable X, with probability density function $p(x)$, the differential entropy [1], [2] is defined as:

$$\begin{aligned} H_p(X) &\equiv - \mathbb{E}_p \left[\log p(x)\right] \\ &= - \int_{-\infty}^{\infty} p(x) \log p(x) dx. \end{aligned} \tag{2.7}$$

Note that, unlike the entropy, the differential entropy can be negative. It is also possible to find density functions p for which the differential entropy is not finite.

If we have two continuous random variables X and Y, with a joint probability density function $p(x, y)$ we can then also easily define the differential conditional entropy:

$$\begin{aligned} H_p(X \,|\, Y) &\equiv - \mathbb{E}_p \left[\log p(x \,|\, y)\right] \\ &= - \int_{-\infty}^{\infty} p(x, y) \log p(x \,|\, y) dx. \end{aligned} \tag{2.8}$$

If we have two density functions p and q, then the cross entropy of q with respect to p is defined as:

$$\begin{aligned} H_{p,q}(X) &\equiv -\mathbb{E}_p\left[\log q(x)\right] \\ &= -\int_{-\infty}^{\infty} p(x)\log q(x)dx. \end{aligned} \tag{2.9}$$

2.1.2.2 Other Quantities

Once we have defined the differential entropy and cross entropy, we can define other quantities in terms of them. This is similar to how we defined the various information theoretic quantities for discrete variables in terms of the discrete entropy and cross entropy. Along these lines, the mutual information is then:

$$\begin{aligned} I_p(X;Y) &\equiv H_p(X) - H_p(X\,|\,Y) \\ &= \mathbb{E}_p\left[\log\frac{p(x,y)}{p(x)p(y)}\right] \\ &= \int_{-\infty}^{\infty} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}dxdy. \end{aligned} \tag{2.10}$$

$$\tag{2.11}$$

Similarly, if we have a continuous-valued random variable X and two probability distributions p and q, the KL divergence can be defined as:

$$\begin{aligned} D_{KL}(p\,||\,q) &\equiv H_{p,q}(X) - H_p(X) \\ &= \mathbb{E}_p\left[\log\frac{p(x)}{q(x)}\right] \\ &= \int_{-\infty}^{\infty} p(x)\log\frac{p(x)}{q(x)}dx. \end{aligned} \tag{2.12}$$

$$\tag{2.13}$$

Note that, although the differential entropy and cross entropy can be negative, the KL divergence defined on continuous variables is guaranteed to be greater than or equal to zero. As the mutual information is a KL divergence, it has this same property. This implies that it has a consistent interpretation for continuous variables, as the information held commonly between two variables. This is as opposed to the differential entropies themselves, which can be negative. As such, differentiable entropies should only be viewed as relative quantities which can be compared with other differential entropies.

2.2 Information Dynamics

Now that we have introduced the various important information-theoretic quantities, we can discuss how they are applied to time-series data within the framework of information dynamics. A common intuition in the study of complex systems is that such systems perform *computation* or *information processing* [7]. In order for economies to react to new demands [8] or organisms to respond to their

environment [9], some sort of computation is required. Moreover, the literature in complex systems is replete with references to computational operations such as the storage or transmission of information [10]. *Information dynamics* [11]–[13] is an emerging set of techniques for quantifying this intuition. It provides rigorous measures for the computational primitives which a complex system performing computation is composed of. These measures can then be applied to such systems in order to reveal their computational dynamics. They have already been applied to a diverse range of systems, including: brain imaging data [14]–[16], schooling behaviour in fish [17], random boolean networks [18] and cellular automata [11].

Information dynamics models the state of a system component X as being computed from its past as well as the past of covarying components. It characterises the computation performed by X by its *predictive capacity*, $\dot{\mathbf{C}}_X$, which is the reduction in uncertainty in the state of X gained by knowing its history as well as the history of its covariates. Without loss of generality, we can assume a single covarying component Y and write:

$$\dot{\mathbf{C}}_X \equiv \tfrac{1}{\Delta t}\mathbf{I}\left(X_t\,;\,\mathbf{X}_{<t},\mathbf{Y}_{<t}\right) \tag{2.14}$$

Here, $\mathbf{I}$ refers to the mutual information (see Section 2.1.1.3). $\mathbf{Y}_{<t}$ and $\mathbf{X}_{<t}$ are history embedding vectors of the source and target process. Usually, they will consist of the k and l preceding values of the time series, that is $\mathbf{x}_{<t} = [x_{t-k}, x_{t-k+1}, \ldots, x_{t-1}]$ and $\mathbf{y}_{<t} = [y_{t-l}, y_{t-l+1}, \ldots, y_{t-1}]$. Δt is the sampling interval of the time series. We normalise by it so that we have a quantity that is independent of the choice of sampling rate [19].

The reduction in uncertainty about the current state of X can be decomposed into the uncertainty reduction stemming solely from the knowledge of the history of X and the uncertainty reduction from the history of Y, given that the history of X is known [20]. That is:

$$\dot{\mathbf{C}}_X = \tfrac{1}{\Delta t}\mathbf{I}\left(X_t\,;\,\mathbf{X}_{<t}\right) + \tfrac{1}{\Delta t}\mathbf{I}\left(X_t\,;\,\mathbf{Y}_{<t}|\mathbf{X}_{<t}\right) \tag{2.15}$$

The two terms on the right of (2.15) are given the names *active information storage* ($\dot{\mathbf{S}}_X$) and *transfer entropy* ($\dot{\mathbf{T}}_{Y\to X}$), respectively.

Active information storage [3], is a measure of the information stored by an individual system component. It is defined as the mutual information between the current state of a component and its history.

$$\dot{\mathbf{S}}_X \equiv \tfrac{1}{\Delta t}\mathbf{I}\left(X_t\,;\,\mathbf{X}_{<t}\right) \tag{2.16}$$

$$= \frac{1}{\Delta t}\sum_{t=1}^{N} p(x_t, \mathbf{x}_{<t}) \log_2\left(\frac{p(x_t|\mathbf{x}_{<t})}{p(x_t)}\right) \tag{2.17}$$

On the other hand, transfer entropy [4] is a measure of information flow between system components. It is defined as the conditional mutual information between the history of the source and the current state of the target, where the conditioning is done on the history of the target.

$$\dot{\mathbf{T}}_{Y\to X} \equiv \tfrac{1}{\Delta t}\mathbf{I}\left(X_t\,;\,\mathbf{Y}_{<t}|\mathbf{X}_{<t}\right) \tag{2.18}$$

$$= \frac{1}{\Delta t}\sum_{t=1}^{N} p(x_t, \mathbf{y}_{<t}, \mathbf{x}_{<t}) \log_2\left(\frac{p(x_t|\mathbf{y}_{<t}, \mathbf{x}_{<t})}{p(x_t|\mathbf{x}_{<t})}\right) \tag{2.19}$$

It is possible to use the chain rule for mutual information to decompose the TE into terms from each of the individual source embedding components. If $\mathbf{y}_{<t} = \{y_{t-l}, y_{t-l+1}, \ldots, y_{t-1}\}$, then we have:

$$\dot{\mathbf{T}}_{Y\to X} = \tfrac{1}{\Delta t}\mathbf{I}\left(X_t\,;\, Y_{t-1}|\mathbf{X}_{<t}\right) + \tfrac{1}{\Delta t}\mathbf{I}\left(X_t\,;\, Y_{t-2}|\mathbf{X}_{<t}, Y_{t-1}\right) + \ldots + \tfrac{1}{\Delta t}\mathbf{I}\left(X_t\,;\, Y_{t-l}|\mathbf{X}_{<t}, \mathbf{Y}_{t-l+1:t-1}\right) \quad (2.20)$$

The first term contains the information flow from the most recent source embedding element, the second contains the flow from the second element, conditioned on the first and so forth. Each term can, therefore, be considered a conditional TE. We can define the conditional TE more generally in terms of other processes Z. This gives:

$$\dot{\mathbf{T}}_{Y\to X\,|\,Z} \equiv \tfrac{1}{\Delta t}\mathbf{I}\left(X_t\,;\, \mathbf{Y}_{<t}|\mathbf{X}_{<t}, \mathbf{Z}_{<t}\right) \quad (2.21)$$

$$= \frac{1}{\Delta t}\sum_{t=1}^{N} p(x_t, \mathbf{y}_{<t}, \mathbf{x}_{<t}, \mathbf{z}_{<t}) \log_2 \left(\frac{p(x_t|\mathbf{y}_{<t}, \mathbf{x}_{<t}, \mathbf{z}_{<t})}{p(x_t|\mathbf{x}_{<t}, \mathbf{z}_{<t})}\right) \quad (2.22)$$

2.3 Estimation of Information-Theoretic Quantities

The information-theoretic quantities that have been discussed so far have all been defined in terms of probability distributions. However, if we want to use these quantities to study real-world systems using empirical data, we will not generally have access to the underlying probability distributions. Instead, we have access to samples drawn from these distributions. We then need to estimate the quantities from the samples. This section will discuss techniques for doing this. We will begin with the simpler case of discrete data in Section 2.3.1, before introducing the problem of performing this same task on continuous data in Section 2.3.2. Sections 2.3.3 through 2.3.7 will then introduce and discuss the very popular class of *k*-nearest-neighbour estimators of information theoretic quantities for continuous data.

The various information-theoretic quantities introduced in Section 2.1.1 are all functionals $T(p)$ of probability distribution p (or functionals of two probability distributions). The aim of estimation is to construct estimators $\widehat{T}(\mathbf{X})$ which estimate the true underlying value of $T(p)$ from the samples drawn from p, denoted by $\mathbf{X}$.

The nature of estimation means that our estimates $\widehat{T}(\mathbf{X})$ may have a *bias* with respect to the true value $T(p)$, and a *variance*, as a function of the size of the data being provided to the estimator. The bias is a measure of the degree to which the estimator systematically deviates from the true value of the quantity being estimated, for finite data size. It is expressed as

$$\text{bias}(\widehat{T}(\mathbf{X})) = \mathbb{E}\left[\widehat{T}(\mathbf{X})\right] - T(p). \quad (2.23)$$

The variance of an estimator is a measure of the degree to which it provides different estimates for distinct, finite, samples from the same process. It is expressed as

$$\text{variance}(\widehat{T}(\mathbf{X})) = \mathbb{E}\left[\widehat{T}(\mathbf{X})^2\right] - \mathbb{E}\left[\widehat{T}(\mathbf{X})\right]^2. \quad (2.24)$$

Another important property is *consistency*, which refers to whether, in the limit of infinite data points, the estimator converges to the true value. That is, if n is the number of data points, then an estimator

is consistent if and only if

$$\lim_{n\to\infty} \widehat{T}(\mathbf{X}) = T(p). \tag{2.25}$$

Bias, variance and consistency are important metrics to be considered when evaluating the efficacy of an estimator and we will make frequent reference to them in this thesis.

2.3.1 Discrete Data

Estimation of information-theoretic quantities on discrete data can be done using the straightforward plugin estimator [4], [21]. A more general definition of this estimator, from which each specific estimator can easily be derived, is that if our quantity is a functional $T(p)$ of the probability distribution p, then the plugin estimator is:

$$\widehat{T}_{\text{plugin}} = T\left(\widehat{p}_{\text{freq}}(x_i)\right). \tag{2.26}$$

$\widehat{p}_{\text{freq}}$ is the frequency-based estimator of the probability mass function. That is, $\widehat{p}_{\text{freq}}(x_i) = \frac{n_i}{N}$, where n_i is the number of occurrences of x_i and N is the number of samples. In cases where we have a quantity defined in terms of two distributions p and q (such as cross entropy and KL-divergence), (2.26) can be extended in a straightforward manner:

$$\widehat{T}_{\text{plugin}} = T\left(\widehat{p}_{\text{freq}}(x_i), \widehat{q}_{\text{freq}}(x_i)\right). \tag{2.27}$$

We can then substitute in the specific functionals from Section 2.1.1 in order to get our required estimator of each quantity. As an example, the plugin estimator for mutual information (Section 2.1.1.3) is:

$$\widehat{I}_{\text{plugin}}(X\,;\,Y) = \sum_{i=1}^{n} \widehat{p}_{\text{freq}}(x_i, y_i) \log \frac{\widehat{p}_{\text{freq}}(x_i, y_i)}{\widehat{p}_{\text{freq}}(x_i)\widehat{p}_{\text{freq}}(y_i)}. \tag{2.28}$$

2.3.2 Continuous Data

The simplest way to estimate information-theoretic quantities for continuous data is to discretise the data [4]. That is, we define 'bins' over the domain of the variable and treat all values occurring in a given bin as having the same single discrete value. We can then use the plugin estimators discussed in the previous section. This approach does, however, have a number of problems, most significantly that the resulting estimators are not usually consistent — they are not guaranteed to converge to the true underlying value in the limit of infinite data. In fact, they often converge to a value very far from the true value. See Section 3.2.2 for examples of this behaviour and also refer to [4], [19], [22].

There are a number of different approaches to performing estimation without discretisation. There has, for instance, been much recent progress on parametric information-theoretic estimators [23]. However, such estimators will always inject modelling assumptions into the estimation process. Even in the case that large, general, parametric models are used — as in [24] — there are no known methods of determining whether such a model is capturing all dependencies present within the data.

In comparison, nonparametric estimators make less explicit model assumptions regarding the probability distributions. Early approaches included the use of kernels for the estimation of the probability densities [25], however this has the disadvantage of operating at a fixed kernel 'resolution'. An improvement was achieved by the successful, widely applied, class of nonparametric estimators

making use of k-nearest-neighbour statistics [26]–[29], which dynamically adjust their resolution given the local density of points. Crucially, there are consistency proofs [28], [30] for kNN estimators, meaning that these methods are known to converge to the true values in the limit of infinite data size. These estimators operate by decomposing the information quantity of interest into a sum of differential entropy terms H_*. Each entropy term is subsequently implicitly or explicitly estimated by estimating the probability densities $p(x_i)$ at all the points in the sample by finding the distances to the kth nearest-neighbours of the points x_i. The average of the logarithms of these densities is found and is adjusted by bias correction terms. In some instances, most notably the Kraskov-Stögbauer-Grassberger (KSG) estimator for mutual information [27], many of the terms in each entropy estimate cancel and so each entropy is only implicitly estimated. There are consistency proofs for many of the estimators within this class [26], [28], [31], and they have been shown to have favourable bias properties [30].

These bias and consistency properties are highly desirable. It is for these reasons that this class of estimator has become ubiquitous within the application of information theory to empirical data [26]–[29]. Moreover, this has resulted in a wide literature discussing features of these estimators as well as potential improvements [28], [30]. It was, therefore, decided that the novel estimator for TE on event-based data, operating on the continuous-valued ISIs (or inter-event times), presented in Chapter 3 would be of this class. As such, the remainder of this section will focus solely on k-nearest-neighbour estimators. We will begin by presenting the use of this technique for the estimation of differential entropy, and then build up the other estimators as combinations of this core estimator.

2.3.3 The Kozachenko-Leonenko Estimator of Differential Entropy

Following [26], assume that we want to estimate the differential entropy of the distribution $\mu(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^d$. μ is unknown, but we have a set S_X of N_X points drawn from μ. If we want to use these points to estimate the differential entropy H we need to construct estimates of the form

$$\widehat{H}(S_X) = -\frac{1}{N_X}\sum_{i=1}^{N_X} \widehat{\ln \mu(\mathbf{x}_i)} \tag{2.29}$$

where $\widehat{\ln \mu(\mathbf{x}_i)}$ is an estimate of the logarithm of the true density. We do this by finding the k nearest-neighbours of $\mathbf{x}_i$ under some norm L. We can then use the distance to this kth neighbour in order to estimate the probability density at $\mathbf{x}_i$ as well as its logarithm. Let $\epsilon\left(k, \mathbf{x}_i, S_X\right)$ be the distance to the kth nearest-neighbour of $\mathbf{x}_i$ in the set S_X under the norm L. Further, let p_i^μ be the probability mass of the ϵ-ball surrounding $\mathbf{x}_i$. We make the assumption that $\mu(\mathbf{x}_i)$ is constant within the ϵ-ball to arrive at $p_i^\mu = \frac{k}{N_X - 1} = c_{d,L}\epsilon\left(k, \mathbf{x}_i, S_X\right)^d \mu(\mathbf{x}_i)$ where $c_{d,L}$ is the volume of the d-dimensional unit ball under the norm L. This relationship can be used to construct a simple estimator of the differential entropy:

$$\widehat{H}(S_X) = -\frac{1}{N_X}\sum_{i=1}^{N_X} \ln \frac{k}{\left(N_X - 1\right) c_{d,L}\epsilon\left(k, \mathbf{x}_i, S_X\right)^d}. \tag{2.30}$$

We then add a bias-correction term $\ln k - \psi(k)$. $\psi(x) = \Gamma^{-1}(x) d\Gamma(x)/dx$ is the digamma function and $\Gamma(x)$ the gamma function. This yields $\widehat{H}_{\mathrm{KL}}$, the Kozachenko-Leonenko [1] [26] estimator of differential

[1] One should be vigilant of the potential confusion that can stem from Kozachenko-Leonenko and Kullback-Leibler sharing the same abbreviation.

entropy:

$$\widehat{H}_{\mathrm{KL}}(S_X) = -\psi(k) + \ln(N_X - 1) + \ln c_{d,L} + \frac{d}{N_X}\sum_{i=1}^{N_X} \ln \epsilon\left(k, \mathbf{x}_i, S_X\right). \tag{2.31}$$

This estimator has been shown to be consistent [26], [32].

2.3.4 The KSG Estimator of Differential Mutual Information

A widely-employed and successful approach to estimating other information-theoretic quantities is to decompose them into their constituent entropy terms and then implicitly or explicitly estimate each of these using the Kozachenko-Leonenko estimator [27], [31]. Applying this to mutual information, suppose we have an unknown joint probability density $\mu(\mathbf{x}, \mathbf{y})$, and a set of samples S_{XY} (with $s_{XY,i} = (\mathbf{x}_i, \mathbf{y}_i)$). We can then construct the following simple estimator of mutual information:

$$\widehat{H}_{\mathrm{simple}}(S_{XY}) = \widehat{H}_{\mathrm{KL}}(S_X) + \widehat{H}_{\mathrm{KL}}(S_Y) - \widehat{H}_{\mathrm{KL}}(S_{XY}). \tag{2.32}$$

It has been generally observed that, when constructing estimators for information-theoretic quantities by combining nearest-neighbour estimators for entropy terms, the bias can be reduced by sharing k-nearest-neighbour distances across the estimators [27], [30], [31]. In the context of estimating the mutual information, this is done by first performing a k_{joint} nearest-neighbour search in the joint sample space S_{XY} around each point pair $(\mathbf{x}_i, \mathbf{y}_i)$, finding a distance ε_i. We then find the number of samples $k_{X,i}$ that fall within the distance ε_i of $\mathbf{x}_i$ and the number of samples $k_{Y,i}$ that fall within the distance ε_i of $\mathbf{y}_i$ (unlike k, $k_{X,i}$ and $k_{Y,i}$ include the point itself). The use of this same distance across all searches means that, if we expand (2.32) according to the definition of $\widehat{H}_{\mathrm{KL}}$ in (2.31), many of the terms (including all distances) will cancel. If we do this, we arrive at the Kraskov-Stögbauer-Grassberger estimator of mutual information:

$$\widehat{I}_{\mathrm{KSG}}(S_{XY}) = \psi(k) + \psi(N) - \langle \psi(k_X) + \psi(k_Y) \rangle \tag{2.33}$$

Kraskov et. al. also present a second version of this estimator where the distances in the marginal spaces are trimmed to the furthermost point that fell within ε_i [27].

2.3.5 KL Divergence Estimators

As the KL divergence is one of the fundamental information-theoretic quantities (Section 2.1.1.6), it is important to consider how to go about estimate it. Moreover, in the construction of the TE estimator for event-based data presented in this thesis, we have to estimate two KL divergence terms (see Section 3.4.1). As KL divergence involves an entropy and a cross entropy term, if we would like to estimate it by combining entropy estimators, then we need to adapt the entropy estimator discussed in Section 2.3.3 in order to be able to handle cross entropies.

2.3.5.1 Estimating Cross Entropy

Following [32], we would like to estimate the cross entropy between two (unknown) probability distributions $\mu(\mathbf{x})$ and $\beta(\mathbf{x})$. Although we do not have access to the probability distributions, suppose that, we have a set S_X of N_X points drawn from μ and a set S_Y of N_Y points drawn from β. Using

similar arguments to above (Section 2.3.3), we use $\epsilon\left(k, \mathbf{x}_i, S_Y\right)/2$ to denote the distance from the ith element of X to its kth nearest neighbour in S_Y. We then make the assumption that $\beta(\mathbf{x}_i)$ is constant within the ϵ-ball, and we have $p_i^{\beta} = \frac{k}{N_Y} = c_{d,L}\epsilon\left(k, \mathbf{x}_i, S_Y\right)^d \beta(\mathbf{x}_i)$. We can then construct a straightforward estimator of the cross entropy

$$\widehat{H}_{\beta}(S_X, S_Y) = -\frac{1}{N_X}\sum_{i=1}^{N_X} \ln \frac{k}{N_Y c_{d,L}\epsilon\left(k, \mathbf{x}_i, S_Y\right)^d}. \tag{2.34}$$

Again, we add the bias-correction term $\ln k - \psi(k)$ to arrive at an estimator of the cross entropy.

$$\widehat{H}_{\beta,\mathrm{KL}}(S_X, S_Y) = -\psi(k) + \ln N_Y + \ln c_{d,L} + \frac{d}{N_X}\sum_{i=1}^{N_X} \ln \epsilon\left(k, \mathbf{x}_i, S_Y\right). \tag{2.35}$$

This estimator has been shown to be consistent [32].

It is worth briefly noting the core difference between estimating entropy and cross entropy using kNN estimators. An entropy estimator takes a set S_X and, for each $\mathbf{x}_i \in S_X$, performs a k nearest-neighbour search *in the same set* S_X. An estimator of cross entropy takes two sets, S_X and S_Y and, for each $\mathbf{x}_i \in S_X$, performs a k nearest-neighbour search *in the other set* S_Y (since it is the samples of Y that are used to estimate the probability density function evaluated inside the integral of the cross entropy).

2.3.5.2 Combining Entropy and Cross Entropy Estimators

Suppose we have two unknown probability densities $\mu(\mathbf{x})$ and $\beta(\mathbf{x})$ along with a set S_X of N_X points drawn from μ and a set S_Y of N_Y points drawn from β. We can then combine the Kozachenko-Leonenko estimators for entropy and cross entropy in order to arrive at an estimator for the KL-divergence [31].

$$\begin{aligned}\widehat{D}_{\mathrm{KL}}(S_X, S_Y) =& \widehat{H}_{\beta,\mathrm{KL}}(S_X, S_Y) - \widehat{H}_{\mathrm{KL}}(S_X) \\ =& \ln N_Y - \ln(N_X - 1) + \frac{d}{N_X}\sum_{i=1}^{N_X}\left[\ln \epsilon\left(k, \mathbf{x}_i, S_Y\right) - \ln \epsilon\left(k, \mathbf{x}_i, S_X\right)\right].\end{aligned} \tag{2.36}$$

Wang et. al. [31] showed that this estimator is consistent. As with the Kraskov estimator for mutual information (Section 2.3.4), the bias of this estimator can be reduced by sharing the same distance across the nearest-neighbour searches. In the mutual information case, we are guaranteed that the distance to the kth point in the joint space will always be greater than or equal to this distance in each of the marginal spaces. This then provides us with the distance to kth nearest neighbour in the joint space as the natural choice for what the shared distance will be. For KL-divergence, however, there is no such natural choice of distance. Instead, we perform our k nearest-neighbour search in each space as before and choose the largest of the two resulting distances. We then have to perform a distance search in the space which had the smaller distance, in order to find the number of points that fall within the shared distance. This results in the following estimator [31]:

$$\begin{aligned}\widehat{D}_{\mathrm{KL}}(S_X, S_Y) =& \widehat{H}_{\beta,\mathrm{KL}}(S_X, S_Y) - \widehat{H}_{\mathrm{KL}}(S_X) \\ =& \ln N_Y - \ln(N_X - 1) + \frac{d}{N_X}\sum_{i=1}^{N_X}\left[\psi(k_{X,i}) - \psi(k_{Y,i})\right],\end{aligned} \tag{2.37}$$

where $k_{X,i}$ is the number of points found in S_X within the shared distance of $\mathbf{x}_i$ and $k_{Y,i}$ is the number of points found in S_Y within the shared distance of $\mathbf{x}_i$. One of $k_{X,i}$ or $k_{Y,i}$ will always be equal to the value of k used for the initial searches. This estimator was shown to be consistent and have lower bias than the estimator which operates without distance sharing [31].

2.3.6 Estimating Conditional Mutual Information

In order to estimate conditional mutual information [33], [34], we proceed as we did for mutual information (Section 2.3.4), by first decomposing the mutual information into its constituent entropy terms, applying the Kozachenko-Leonenko [26] (Section 2.3.3) and then reducing bias by sharing distances. For the first step, we arrive at:

$$\widehat{I}_{\text{KSG, cond}}(S_{XYZ}) = \widehat{H}_{\text{KL}}(S_{XZ}) + \widehat{H}_{\text{KL}}(S_{YZ}) - \widehat{H}_{\text{KL}}(S_{XYZ}) - \widehat{H}_{\text{KL}}(S_Z) \tag{2.38}$$

We then set the distance to the k nearest neighbour found in the joint space (S_{XYZ}) as the search distance with which to perform nearest-neighbour searches in the other sets. This yields the estimator:

$$\widehat{I}_{\text{KSG, cond}}(S_{XYZ}) = \psi(k) - \langle \psi(k_{X,Z}) + \psi(k_{Y,Z}) - \psi(k_Z) \rangle, \tag{2.39}$$

Where $k_{X,Z,i}$, $k_{Y,Z,i}$ and $k_{Y,i}$ are the numbers of points found within the distance ε_i of the point $(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)$ within the associated spaces (including points on the boundary of this ε-ball), and ε_i is the distance to the kth nearest neighbour of this point in the joint space.

2.3.7 TE Estimation

As TE is just a conditional mutual information (see equation (2.18)), it can be estimated using (2.40). Rewriting this in terms of the variables used in the definition of TE [34], we have:

$$\widehat{\mathbf{T}}_{Y \to X,\text{KSG}} = \psi(k) - \langle \psi(k_{X,X_{<t}}) + \psi(k_{Y_{<t},X_{<t}}) - \psi(k_{X_{<t}}) \rangle, \tag{2.40}$$

2.4 Significance Testing

Often we are particularly interested in whether an estimated information-theoretic value is consistent with zero or not. For instance, when inferring networks, a non-zero value is taken to signify the existence of an edge. In the case of mutual information, a non-zero mutual information (Section 2.1.1.3) indicates a statistical dependence between two variables. In the case of KL-divergence (Section 2.1.1.6), a non-zero value indicates that the two distributions are not identical.

However, as discussed in Section 2.3, all estimators have some variance. That is, for finite data we do not expect them to produce estimates that are exactly equal to the true underlying value. We therefore require a method for determining whether the estimated value would be likely to occur, even if the true underlying value was zero, thus enabling us to determine if the estimated quantity is statistically different from zero.

As there are no results for the sampling distribution of many of the estimators we use, including the k-nearest-neighbour estimators [35], we turn to permutation (surrogate) methods in order to

provide an approximation for the distribution of estimates under the null hypothesis that the variables from which we have sampled are (conditionally) independent and therefore that the value of the quantity is zero. The idea behind these methods is to permute the data on which the estimation was performed and then perform estimation on this permuted data for many different permutations. The algorithm which we use should create surrogates which are identically distributed to the original data if and only if the null hypothesis holds [36]. We can then use the estimated values on this permuted data in order to approximate the distribution of estimates under the assumption of the null hypothesis of the value being zero. We can also use these surrogate values to estimate the probability that the estimated value, or one more extreme, would occur under the assumption of the null hypothesis (that is, estimate a p value). We can then choose to reject the null hypothesis for a sufficiently low p value.

2.4.1 Shuffling Permutation for Mutual Information

An instructive example of a permutation that can be used for significance testing is simple shuffling of the data, which can be applied to data on which mutual information is being estimated. As the mutual information between X and Y is zero if and only X and Y are independent, testing for non-zero mutual information is a test for statistical dependence. As such, we are testing against the null hypothesis that X and Y are independent ($X \perp\!\!\!\perp Y$) or, equivalently, that the joint probability distribution of X and Y factorises as $p(X, Y) = p(X)p(Y)$. It is straightforward to construct surrogate pairs $(\check{x}, \check{y})$ that conform to this null hypothesis. We start with the original pairs (x, y) and resample the y values across pairs. This shuffling process will maintain the marginal distributions $p(X)$ and $p(Y)$, and the same number of samples, but will destroy any relationship between X and Y, yielding the required factorisation for the null hypothesis. If X and Y are already independent, then this shuffling will have no effect on the joint distribution. Note that this may be problematic when there are serial or dynamic correlations in time series samples. Using the Theiler window is known to correct for this with nearest-neighbour information-theoretic estimators [37].

2.4.2 Local Permutation for Conditional Mutual Information

It is worthwhile asking how one would go about extending the shuffling method discussed in Section 2.4.1 to conditional mutual information. For conditional mutual information, we are testing whether X and Y are statistically dependent given Z. This is equivalent to asking whether the factorisation $p(X, Y \mid Z) = p(X \mid Z)p(Y \mid Z)$ holds.

Deriving a permutation algorithm for this situation is relatively straightforward in the case of discrete data. We can simply look at each z_i separately, and shuffle the y values among all the triplets (x, y, z) that have $z = z_i$.

Recent work by Runge [35] has derived a suitable permutation scheme for conditional mutual information on continuous variables. The core step of this algorithm is to conduct a nearest-neighbour search based on the conditional z values, finding a set of k points that have similar z values to a given triplet. We can then exchange the y value of the given triplet with a randomly chosen y value in the nearest-neighbour set.

2.5 Connectivity Inference Using Information Theory

Network inference has become a popular tool for summarising high-dimensional time-series data in an accessible and visual manner, particularly within neuroscience [38]. It aims to reduce the often very large number of data points down to a single network representation. Each node in this network usually corresponds to a variable in the original system, such as beamformed sources in MEG data [15]. An edge between nodes indicates a statistical relationship between them.

As information theoretic quantities can measure the relationships between variables, they are well suited to the task of network inference. In particular, mutual information (Section 2.1.1.3), and derived quantities such as transfer entropy (Section 2.2), are measures of statistical dependence and so are widely used for network inference.

2.5.1 Functional Networks

Broadly speaking, there are two approaches to performing network inference. The first of these infers what are normally referred to as *functional* networks [38], [39]. These networks are inferred by considering each pair of variables in isolation. As such, this approach will only capture the pairwise relationships between variables. However, one advantage of using TE for the inference of functional networks is that, unlike more common approaches such as simple Pearson correlation [38], the functional networks will be *directed*. Moreover, TE can capture non-linear effects which are missed by the Pearson correlation [4].

On an algorithmic level, the inference of functional networks using TE is fairly straightforward. One simply considers every unique ordered pair of variables, with one member of this pair acting as the target and the other as the source. The TE is then estimated between this source-target pair and, if it is significantly different from zero, a directed edge is place connecting the source to the target.

2.5.2 Effective Networks

Often in multivariate systems, many if not all the variables will be highly correlated. This is particularly true within neuroscience where synchrony is commonplace [40], [41]. When using a functional network approach, this will result in an edge between nearly all nodes. This is not a result which reveals much structure about the data (assuming that we are only interested in the presence of edges, rather than their weights). In very many of these cases, however, even though a variable might have a pairwise dependence with most or all other variables in the system, these dependencies might be removed by conditioning on another variable in the system. Specifically, there will be some smaller subset of variables which a given variable is statistically dependent on for which it is conditionally independent of all other variables in the system.

This is the motivation behind the other main approach to network inference, namely the inference of *effective* networks [42], [43]. For these networks, we only place an edge between a target node and a source, if the source is part of some minimal set that captures all the dependence between the target and the population. That is, in the case of TE-based network inference, we are looking for a minimal set of parents (source variables) whose history can maximally reduce our uncertainty about the current state of each target node.

Different information-theoretic measures can be incorporated into each approach. For instance, functional networks could be inferred using TE or mutual information, depending on whether we wanted to only capture instantaneous effects, or whether we wanted to capture directional and dynamic dependence. Effective networks, on the other hand, can be inferred using conditional mutual information or multivariate TE.

Algorithmically, the inference of effective networks using TE is a fair bit more complicated than the functional case. An approach which has proven effective for the inference of the minimal set of parents is to iteratively add sources to each target in a greedy fashion [44]–[48]. Specifically, for each target process, we select the source with the strongest information flow (without any conditioning). We then select the next source as the component with the highest information flow when conditioned on the first source and add this new source to the conditioning set. We continue adding sources to the conditioning set in this fashion until we are unable to find a source with a statistically significant non-zero information flow. The process then finishes with a pruning step, where it is verified that each source still has a non-zero information flow when conditioned on all other sources in the set.

2.6 Application of Information Theory to Spiking Neural Networks

Before looking at how TE and the information dynamics framework has been applied to spike trains, it is worth reviewing how information theory, more generally, has been applied to this domain as well as how information-theoretic methods relate to other types of analysis commonly performed on spike trains. There is a very large literature surrounding the application of information theory to other types of neuroscience data, however, a review of this work is beyond the scope of this short literature review. The reader is referred to the existing reviews of this subject matter [49]–[52].

Research on the application of information theory to spike trains is deeply intertwined with the concepts of encoders and decoders [50], [53]. Given that authors in these studies make frequent reference to 'information transmission' [54]–[56] (transfer entropy is often referred to as 'information transfer'), that our proposed method for estimating transfer entropy makes use of an encoder and that certain methodologies used have significant similarities to our proposed approach, it is worth paying particular attention to how these concepts are related.

In order to model the mechanism by which a neuron encodes information, an encoding function is often fit to data. This could be a function which maps entire stimulus histories onto entire spike trains. However, it is usually a function which maps only the stimulus history, or the stimulus history and the spiking history of the neuron, onto an instantaneous spike rate [57]. Such a function is often referred to as a *tuning curve* or *response function*. It could potentially take any form, but is frequently a generalised linear model (GLM) [58]. It is worth noting that the estimates of $\lambda_{x|y}$, which will be required for estimating the transfer entropy in continuous time using equation (2.42) are tuning curves (see Chapter 3 for the full derivation and evaluation of this estimator). Once this function has been estimated, calculating the likelihood of the spike train - $P(r|s,\theta)$ - given the stimulus s and tuning curve θ is a simple matter. Specifically:

$$P(r|s,\theta) = \left(\prod_{i=1}^{N} \theta\left(s(t_i)\right)\right) e^{\int_0^t \theta(s(t))dt} \tag{2.41}$$

where the product is over the N spikes of the spike train.

A common task in computational neuroscience is the fitting of decoding functions [57]. These functions map a segment of a spike train (which we will refer to as a 'word') to either a most likely stimulus, or to a distribution of stimuli. There is a tight relationship between this decoder and the mutual information between the response word (spike-train segment) and the stimulus ($I(R\,;\,S)$). As the mutual information is the reduction in uncertainty about S that arises from knowing R, it places a bound on how effective the decoder can be. Indeed, if $I(R\,;\,S) = 0$, the decoder will not be able to perform better than random chance. Further, decoders are a commonly used method for providing a lower bound on the mutual information $I(R\,;\,S)$ [50], [59]–[61]. For a given a decoder D, we know that $I(D(R)\,;\,S) < I(R\,;\,S)$ due to the data processing inequality. Estimating mutual information in stimulus space can be substantially easier than in spike-train space, partially because the experimenter can specify the stimulus so as to facilitate its calculation. Specifically, the experimenter can constrain the variation in stimulus to only a few discrete states, rendering estimation task comparatively easy.

The relationship between encoders, decoders and mutual information is made clear when we consider that a popular class of decoding algorithms use Bayesian inference [62]–[65]. That is, they will first estimate a tuning curve, calculate the likelihood of the response $P(r|s,\theta)$ and then use Bayes' rule to calculate the posterior probability $P(s|r)$. The most likely stimulus can be calculated by finding $s^p = \texttt{argmax}_s(P(s|r))$. The results of this calculation, over all stimulus-response pairs, can be used to estimate the quantity $I(D(R)\,;\,S)$, and thus place a lower bound on the mutual information [53], [66]. In order to emphasise the similarities between mutual information and decoding, it is worth quoting Quiroga and Panzeri [50] "The complementarities between decoding and information theory are explicit when considering Bayesian decoders, because in this case both decoding and information theory are just two different computations over the posterior probability $P(s|r)$"

It is worth highlighting the large similarities between this method of calculating mutual information and the approach for calculating transfer entropy which is detailed in Chapter 3. The Bayesian approach to mutual information estimation first estimates an encoding function θ for the spike train. The likelihoods of response words r given this encoder and a stimulus s - $P(r|s,\theta)$ - are then easily calculated. For each stimulus-response pair the most likely stimulus s^p is predicted by finding the s^p that maximises the posterior probability, $P(s|r)$. The mutual information is then the amount by which the knowledge of s^p reduces the uncertainty about s. The proposed approach for calculating transfer entropy begins in the same manner by estimating the encoding function θ. However, it proceeds to directly calculate the reduction in the uncertainty of the *response* that is provided by knowing the stimulus and the encoding function. Further differences are that the encoding function is only estimated implicitly and that this is done non-parametrically. By comparing distances to sampled history embeddings, the estimator compares the probability of a given history being observed from spiking or non-spiking points in time. This allows it to estimate the reduction in uncertainty provided by the source history, however this could easily be adapted to estimate the history-dependent spike rate. Indeed, this rate is being implicitly estimated.

2.7 Information Dynamics on Spike Trains

2.7.1 Discrete-Time Estimation of TE on Spike Trains

Past applications of TE to spike trains [22], [67]–[80] (see Section 2.7.2 for discussion of the contributions of this work), have made use of a discrete-time estimator of TE. In this estimation scheme, the time series is divided into small bins of width Δt. Each bin is then assigned a binary value which denotes the presence or absence of spikes in the bin. Alternatively, each bin could be assigned a natural number corresponding to the number of spikes that fell within the bin. A choice is made as to the number of time bins, l and m, to include in the source and target history embeddings $\mathbf{y}_{<t}$ and $\mathbf{x}_{<t}$. For each time bin, we can then create a triplet of the current state x_t, the target past $\mathbf{x}_{<t}$, and the source past $\mathbf{y}_{<t}$. We can then apply a simple plugin estimator (Section 2.3.1) for conditional mutual information on discrete data. More specifically, for a given combination $\mathbf{x}_{<t}$ and $\mathbf{y}_{<t}$, the probability of the target's value in the current bin conditioned on these histories, $p\left(x_t \mid \mathbf{x}_{<t}, \mathbf{y}_{<t}\right)$, can be directly estimated by counting its frequency of occurence. The probability of the target's value in the current bin conditioned on only the target history, $p\left(x_t \mid \mathbf{x}_{<t}\right)$, can be estimated in the same fashion. From these estimates the TE can be calculated in a straightforward manner via its definition (equation (2.18)).

There are two large disadvantages to this approach [19]. As time discretisation is a lossy transformation, it will result in an inaccurate estimate of the TE. Thus, any estimator based on time discretisation is not consistent (it is not guaranteed to converge to the true value of the TE in the limit of infinite data, see Section 2.3). Secondly, whilst the loss of resolution of the discretization will reduce with decreasing bin size Δt, this requires larger dimensionality in the history embeddings to capture correlations over similar time intervals. This increase in dimension will result in an exponential increase in the state space size being sampled to estimate $p\left(x_t \mid \mathbf{x}_{<t}, \mathbf{y}_{<t}\right)$, and therefore the data requirements. In practice, the problems of dimensionality are sufficiently severe that most authors tend to use very few bins in their history embeddings. Indeed, nearly all previous applications of the discrete-time TE estimator to spiking data from cell cultures used only a single bin in their history embeddings (the exceptions being [74], which used 1 and 3 bins for the target and source, and [81], which used up to 5 bins). The bin widths used in those studies were 40 µs [73], 50 µs [81], 0.3 ms [82], and 1 ms [74], [83], [84]. Some studies chose to examine the TE values produced by multiple different bin widths, specifically: 0.6 ms and 100 ms [85], 1.6 ms and 3.5 ms [70] and 10 different widths ranging from 1 ms to 750 ms [22].

In instances where a variety of bin widths were used, the rationale is clear in that the authors are trying to interrogate the differences in information flows occurring on different time scales. It is of more interest to investigate the rationale behind single bin widths that were used. In some cases, no rationale is given. Reference [82] used a bin width of 0.3 ms, as this was close to the mean inter-spike interval. Reference [84] used a width of 1 ms due to this being a common discretisation scale for a variety of analyses on spike trains.

In the cases where narrow (< 5 ms) bins were used, only a very narrow slice of history is being considered in the estimation of the history-conditional spike rate. This is problematic, as correlations in spike trains exhist over distances of (at a minimum) hundreds of milliseconds [86], [87]. Conversely, in the instances where broad (> 5 ms) bins were used, relationships occurring on fine time scales will be completely missed. This is significant given that it is established that correlations at the millisecond

and sub-millisecond scale play a role in neural function [88]–[91]. This highlights the hard trade-off that is encountered when performing the estimation of TE on spike trains in discrete time. One can either capture fine temporal details or effects occurring over larger spans of time. However, it is not possible to do both simultaneously.

2.7.2 Application of TE to Biological Neural Networks

There have been a number of studies which have applied the discrete-time TE estimator for spike trains described in Section 2.7.1 to biological recordings of neural populations. This work primarily made use of recordings from neural cell cultures [22], [67], [70], [73], [83]–[85]. These studies primarily focused on inferring the directed functional networks implied by the estimated TE values between pairs of nodes and analysing various features of these information flow networks. These studies found interesting results, such as Shimono and Beggs [83], who found that these networks exhibited a highly non-random structure and contained a long-tailed degree distribution. This work was expanded by Nigam et. al. [73], where it was found that the functional networks contained a rich-club topology. Conversely, Timme et. al. [22] found that the hubs of these networks were localised to certain time scales. Other work [67], [70] has instead focussed on how the components of information flows in cell cultures can be decomposed into unique, redundant and synergystic components.

2.7.3 Information Dynamics on Spike Trains in Continuous Time

Recent work [19], [20] has derived a continuous-time formalism for information dynamics. This work holds the promise of allowing for the estimation of quantities like TE on spike trains in continuous-time, thus overcoming some of the drawbacks of the discrete-time estimator described in Section 2.7.1. Indeed, this potential is realised in this thesis in the estimator derived in Chapter 3. Here, we briefly review this prior work.

It was found that the transfer entropy rate (the information flow per unit time) between two spike trains is:

$$\dot{\mathbf{T}}_{Y\to X} = \lim_{T\to\infty} \frac{1}{T} \sum_{i=1}^{N_X} \ln \frac{\lambda_{x|y}\left[x_{<t_i}, y_{<t_i}\right]}{\lambda_x\left[x_{<t_i}\right]} \tag{2.42}$$

Here, the spike train is observed over a period of time T, during which N_X spikes occurred. $\lambda_{x|y}\left[x_{<t_i}, y_{<t_i}\right]$ is the instantaneous conditional spike rate of the neuron X. The conditioning is performed over the history of the neurons X and Y, where this history is represented by the k most recent inter-spike intervals of X and the l most recent inter-spike intervals of Y. Similarly, $\lambda_x\left[x_{<t_i}\right]$ is the instantaneous conditional spike rate of X, where the conditioning is performed only on the history of X. The contributions to the TE are, therefore, only coming from the spikes (events). As shown in [19], the contributions from the quiescent periods cancel and average to zero - thus making no contribution to the TE.

It was further found that the active information storage diverges in continuous time. It was, therefore, decomposed into a diverging and a non-diverging component. That is:

$$\mathbf{A}_X = \mathbf{I}_X + \dot{\mathbf{M}}_X \Delta t + \mathcal{O}\left(\Delta t^2\right) \tag{2.43}$$

$\mathbf{I}_X$ refers to the *instantaneous predictive capacity*, which captures the uncertainty reduction due to the

path regularity of the process. This quantity diverges for many processes, but is zero for spike trains. $\dot{\mathbf{M}}_X$ refers to the *active memory utilisation rate*, which is the uncertainty reduction due to correlations other than path regularity.

$$\dot{\mathbf{M}}_{Y\to X} = \lim_{T\to\infty} \frac{1}{T} \sum_{i=1}^{N_X} \ln \frac{\lambda_x \left[x_{<t_i}\right]}{\langle \lambda_x \rangle} \tag{2.44}$$

Here, $\langle \lambda_x \rangle$ refers to the average spike rate of X over the entire length of the spike train.

2.8 References

[1] D. J. MacKay, *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

[2] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012.

[3] R. V. Ceguerra, J. T. Lizier, and A. Y. Zomaya, "Information storage and transfer in the synchronization process in locally-connected networks," in *2011 IEEE Symposium on Artificial Life (ALIFE)*, IEEE, 2011, pp. 54–61.

[4] T. Bossomaier, L. Barnett, M. Harré, and J. T. Lizier, "An introduction to transfer entropy," *Cham: Springer International Publishing*, vol. 65, 2016.

[5] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[6] L. Wasserman, *All of statistics: a concise course in statistical inference*. Springer, 2004, vol. 26.

[7] D. P. Feldman, C. S. McTague, and J. P. Crutchfield, "The organization of intrinsic computation: Complexity-entropy diagrams and the diversity of natural information processing," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 18, no. 4, p. 043 106, 2008.

[8] L. Tesfatsion, "Agent-based computational economics: Modeling economies as complex adaptive systems," *Information Sciences*, vol. 149, no. 4, pp. 262–268, 2003.

[9] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.

[10] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, "Local measures of information storage in complex distributed computation," *Information Sciences*, vol. 208, pp. 39–54, 2012.

[11] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, "Local information transfer as a spatiotemporal filter for complex systems," *Physical Review E*, vol. 77, no. 2, p. 026 110, 2008.

[12] J. T. Lizier, "Jidt: An information-theoretic toolkit for studying the dynamics of complex systems," *Frontiers in Robotics and AI*, vol. 1, p. 11, 2014.

[13] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, "Information modification and particle collisions in distributed computation," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 20, no. 3, p. 037 109, 2010.

[14] A. Brodski-Guerniero, G.-F. Paasch, P. Wollstadt, I. Özdemir, J. T. Lizier, and M. Wibral, "Information-theoretic evidence for predictive coding in the face-processing system," *Journal of Neuroscience*, vol. 37, no. 34, pp. 8273–8283, 2017.

[15] A. Brodski-Guerniero, M. J. Naumer, V. Moliadze, *et al.*, "Predictable information in neural signals during resting state is reduced in autism spectrum disorder," *Human brain mapping*, vol. 39, no. 8, pp. 3227–3240, 2018.

[16] M. Wibral, B. Rahm, M. Rieder, M. Lindner, R. Vicente, and J. Kaiser, "Transfer entropy in magnetoencephalographic data: Quantifying information flow in cortical and cerebellar networks," *Progress in Biophysics and Molecular Biology*, vol. 105, no. 1-2, pp. 80–97, 2011.

[17] E. Crosato, L. Jiang, V. Lecheval, *et al.*, "Informative and misinformative interactions in a school of fish," *Swarm Intelligence*, vol. 12, no. 4, pp. 283–305, 2018.

[18] J. T. Lizier, S. Pritam, and M. Prokopenko, "Information dynamics in small-world boolean networks," *Artificial Life*, vol. 17, no. 4, pp. 293–314, 2011.

[19] R. E. Spinney, M. Prokopenko, and J. T. Lizier, "Transfer entropy in continuous time, with applications to jump and neural spiking processes," *Physical Review E*, vol. 95, no. 3, p. 032 319, 2017.

[20] R. E. Spinney and J. T. Lizier, "Characterizing information-theoretic storage and transfer in continuous time processes," *Physical Review E*, vol. 98, no. 1, p. 012 314, 2018.

[21] L. Wasserman, *All of nonparametric statistics*. Springer Science & Business Media, 2006.

[22] N. Timme, S. Ito, M. Myroshnychenko, *et al.*, "Multiplex networks of cortical and hippocampal neurons revealed at different timescales," *PloS One*, vol. 9, no. 12, e115764, 2014.

[23] B. Poole, S. Ozair, A. van den Oord, A. A. Alemi, and G. Tucker, "On variational lower bounds of mutual information," in *NeurIPS Workshop on Bayesian Deep Learning*, 2018.

[24] M. I. Belghazi, A. Baratin, S. Rajeshwar, *et al.*, "Mutual information neural estimation," in *International Conference on Machine Learning*, PMLR, 2018, pp. 531–540.

[25] J. Beirlant, E. J. Dudewicz, L. Györfi, and E. C. Van der Meulen, "Nonparametric entropy estimation: An overview," *International Journal of Mathematical and Statistical Sciences*, vol. 6, no. 1, pp. 17–39, 1997.

[26] L. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Problemy Peredachi Informatsii*, vol. 23, no. 2, pp. 9–16, 1987.

[27] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E*, vol. 69, no. 6, p. 066 138, 2004.

[28] T. B. Berrett, R. J. Samworth, M. Yuan, *et al.*, "Efficient multivariate entropy estimation via k-nearest neighbour distances," *The Annals of Statistics*, vol. 47, no. 1, pp. 288–318, 2019.

[29] K. R. Moon, K. Sricharan, K. Greenewald, and A. O. Hero, "Ensemble estimation of information divergence," *Entropy*, vol. 20, no. 8, p. 560, 2018.

[30] W. Gao, S. Oh, and P. Viswanath, "Demystifying fixed k-nearest neighbor information estimators," *IEEE Transactions on Information Theory*, vol. 64, no. 8, pp. 5629–5661, 2018.

[31] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation for multidimensional densities via k-nearest-neighbor distances," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2392–2405, 2009.

[32] N. Leonenko, L. Pronzato, and S. Vippal, "A class of Rényi information estimators for multidimensional densities," *The Annals of Statistics*, vol. 36.5, pp. 2153–2182, 2008.

[33] S. Frenzel and B. Pompe, "Partial mutual information for coupling analysis of multivariate time series," *Physical Review Letters*, vol. 99, no. 20, p. 204 101, 2007.

[34] G. Gómez-Herrero, W. Wu, K. Rutanen, M. C. Soriano, G. Pipa, and R. Vicente, "Assessing coupling dynamics from an ensemble of time series," *Entropy*, vol. 17, no. 4, pp. 1958–1970, 2015.

[35] J. Runge, "Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2018, pp. 938–947.

[36] G. Doran, K. Muandet, K. Zhang, and B. Schölkopf, "A permutation-based kernel conditional independence test," in *30th Conference on Uncertainty in Artificial Intelligence (UAI 2014)*, AUAI Press, 2014, pp. 132–141.

[37] J. Nichols, M. Seaver, S. Trickey, M. Todd, C. Olson, and L. Overbey, "Detecting nonlinearity in structural systems using the transfer entropy," *Physical Review E*, vol. 72, no. 4, p. 046 217, 2005.

[38] O. Sporns, *Networks of the Brain*. MIT press, 2010.

[39] A. A. Fingelkurts, A. A. Fingelkurts, and S. Kähkönen, "Functional connectivity in the brain—is it an elusive concept?" *Neuroscience & Biobehavioral Reviews*, vol. 28, no. 8, pp. 827–836, 2005.

[40] P. Uhlhaas, G. Pipa, B. Lima, *et al.*, "Neural synchrony in cortical networks: History, concept and current status," *Frontiers in Integrative Neuroscience*, vol. 3, p. 17, 2009.

[41] P. J. Uhlhaas and W. Singer, "Neural synchrony in brain disorders: Relevance for cognitive dysfunctions and pathophysiology," *Neuron*, vol. 52, no. 1, pp. 155–168, 2006.

[42] D. S. Bassett and O. Sporns, "Network neuroscience," *Nature Neuroscience*, vol. 20, no. 3, pp. 353–364, 2017.

[43] F. de Vico Fallani, J. Richiardi, M. Chavez, and S. Achard, "Graph analysis of functional brain networks: Practical issues in translational neuroscience," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1653, p. 20 130 521, 2014.

[44] L. Novelli, P. Wollstadt, P. Mediano, M. Wibral, and J. T. Lizier, "Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing," *Network Neuroscience*, vol. 3, no. 3, pp. 827–847, 2019.

[45] J. Lizier and M. Rubinov, "Multivariate construction of effective computational networks from observational data," 2012.

[46] L. Faes, G. Nollo, and A. Porta, "Information-based detection of nonlinear granger causality in multivariate processes via a nonuniform embedding technique," *Physical Review E*, vol. 83, no. 5, p. 051 112, 2011.

[47] J. Sun, D. Taylor, and E. M. Bollt, "Causal network inference by optimal causation entropy," *SIAM Journal on Applied Dynamical Systems*, vol. 14, no. 1, pp. 73–106, 2015.

[48] I. Vlachos and D. Kugiumtzis, "Nonuniform state-space reconstruction and coupling detection," *Physical Review E*, vol. 82, no. 1, p. 016 207, 2010.

[49] A. Borst and F. E. Theunissen, "Information theory and neural coding," *Nature Neuroscience*, vol. 2, no. 11, p. 947, 1999.

[50] R. Q. Quiroga and S. Panzeri, "Extracting information from neuronal populations: Information theory and decoding approaches," *Nature Reviews Neuroscience*, vol. 10, no. 3, p. 173, 2009.

[51] M. Wibral, R. Vicente, and J. T. Lizier, *Directed information measures in neuroscience*. Springer, 2014.

[52] N. M. Timme and C. Lapish, "A tutorial for information theory in neuroscience," *eNeuro*, vol. 5, no. 3, 2018.

[53] L. Paninski, J. Pillow, and J. Lewi, "Statistical models for neural encoding, decoding, and optimal stimulus design," *Progress in Brain Research*, vol. 165, pp. 493–507, 2007.

[54] N. Brenner, W. Bialek, and R. d. R. Van Steveninck, "Adaptive rescaling maximizes information transmission," *Neuron*, vol. 26, no. 3, pp. 695–702, 2000.

[55] T. Toyoizumi, J.-P. Pfister, K. Aihara, and W. Gerstner, "Generalized bienenstock–cooper–munro rule for spiking neurons that maximizes information transmission," *Proceedings of the National Academy of Sciences*, vol. 102, no. 14, pp. 5239–5244, 2005.

[56] J. Gjorgjieva, R. A. Mease, W. J. Moody, and A. L. Fairhall, "Intrinsic neuronal properties switch the mode of information transmission in networks," *PLoS Computational Biology*, vol. 10, no. 12, e1003962, 2014.

[57] C. R. Holdgraf, J. W. Rieger, C. Micheli, S. Martin, R. T. Knight, and F. E. Theunissen, "Encoding and decoding models in cognitive electrophysiology," *Frontiers in Systems Neuroscience*, vol. 11, p. 61, 2017.

[58] J. Aljadeff, B. J. Lansdell, A. L. Fairhall, and D. Kleinfeld, "Analysis of neuronal spike trains, deconstructed," *Neuron*, vol. 91, no. 2, pp. 221–259, 2016.

[59] W. Bialek, F. Rieke, R. D. R. Van Steveninck, and D. Warland, "Reading a neural code," *Science*, vol. 252, no. 5014, pp. 1854–1857, 1991.

[60] F. Rieke, D. Warland, R. d. R. van Steveninck, and W. Bialek, *Spikes: Exploring the neural code*, 1999.

[61] L. Paninski, "Estimation of entropy and mutual information," *Neural Computation*, vol. 15, no. 6, pp. 1191–1253, 2003.

[62] S. Koyama and L. Paninski, "Efficient computation of the maximum a posteriori path and parameter estimation in integrate-and-fire and more general state-space models," *Journal of Computational Neuroscience*, vol. 29, no. 1-2, pp. 89–105, 2010.

[63] M. Paradiso, "A theory for the use of visual orientation information which exploits the columnar structure of striate cortex," *Biological cybernetics*, vol. 58, no. 1, pp. 35–49, 1988.

[64] T. D. Sanger, "Probability density estimation for the interpretation of neural population codes," *Journal of Neurophysiology*, vol. 76, no. 4, pp. 2790–2793, 1996.

[65] M. W. Oram, P. Földiák, D. I. Perrett, and F. Sengpiel, "Theideal homunculus': Decoding neural population signals," *Trends in Neurosciences*, vol. 21, no. 6, pp. 259–265, 1998.

[66] R. Barbieri, L. M. Frank, D. P. Nguyen, *et al.*, "Dynamic analyses of information encoding in neural ensembles," *Neural computation*, vol. 16, no. 2, pp. 277–307, 2004.

[67] M. Wibral, C. Finn, P. Wollstadt, J. T. Lizier, and V. Priesemann, "Quantifying information modification in developing neural networks via partial information decomposition," *Entropy*, vol. 19, no. 9, p. 494, 2017.

[68] J.-P. Thivierge, "Scale-free and economical features of functional connectivity in neuronal networks," *Physical Review E*, vol. 90, no. 2, p. 022 721, 2014.

[69] J. J. Harris, E. Engl, D. Attwell, and R. B. Jolivet, "Energy-efficient information transfer at thalamocortical synapses," *PLoS Computational Biology*, vol. 15, no. 8, e1007226, 2019.

[70] N. M. Timme, S. Ito, M. Myroshnychenko, *et al.*, "High-degree neurons feed cortical computations," *PLoS Computational Biology*, vol. 12, no. 5, e1004858, 2016.

[71] K. E. Schroeder, Z. T. Irwin, M. Gaidica, *et al.*, "Disruption of corticocortical information transfer during ketamine anesthesia in the primate brain," *Neuroimage*, vol. 134, pp. 459–465, 2016.

[72] R. Kobayashi and K. Kitano, "Impact of network topology on inference of synaptic connectivity from multi-neuronal spike data simulated by a large-scale cortical network model," *Journal of Computational Neuroscience*, vol. 35, no. 1, pp. 109–124, 2013.

[73] S. Nigam, M. Shimono, S. Ito, *et al.*, "Rich-club organization in effective connectivity among cortical neurons," *Journal of Neuroscience*, vol. 36, no. 3, pp. 670–684, 2016.

[74] S. Ito, M. E. Hansen, R. Heiland, A. Lumsdaine, A. M. Litke, and J. M. Beggs, "Extending transfer entropy improves identification of effective connectivity in a spiking cortical network model," *PLoS One*, vol. 6, no. 11, e27431, 2011.

[75] S. A. Neymotin, K. M. Jacobs, A. A. Fenton, and W. W. Lytton, "Synaptic information transfer in computer models of neocortical columns," *Journal of Computational Neuroscience*, vol. 30, no. 1, pp. 69–84, 2011.

[76] B. Gourévitch and J. J. Eggermont, "Evaluating information transfer between auditory cortical neurons," *Journal of Neurophysiology*, vol. 97, no. 3, pp. 2533–2543, 2007.

[77] A. Buehlmann and G. Deco, "Optimal information transfer in the cortex through synchronization," *PLoS Computational Biology*, vol. 6, no. 9, e1000934, 2010.

[78] O. Stetter, D. Battaglia, J. Soriano, and T. Geisel, "Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals," *PLoS Computational Biology*, vol. 8, no. 8, e1002653, 2012.

[79] J. G. Orlandi, O. Stetter, J. Soriano, T. Geisel, and D. Battaglia, "Transfer entropy reconstruction and labeling of neuronal connections from simulated calcium imaging," *PloS One*, vol. 9, no. 6, e98842, 2014.

[80] I. M. de Abril, J. Yoshimoto, and K. Doya, "Connectivity inference from neural recording data: Challenges, mathematical bases and research directions," *Neural Networks*, vol. 102, pp. 120–137, 2018.

[81] P. C. Antonello, T. F. Varley, J. Beggs, M. Porcionatto, O. Sporns, and J. Faber, "Self-organization of in vitro neuronal assemblies drives to complex network topology," *bioRxiv*, 2021.

[82] M. Garofalo, T. Nieus, P. Massobrio, and S. Martinoia, "Evaluation of the performance of information theory-based methods and cross-correlation to estimate the functional connectivity in cortical networks," *PloS One*, vol. 4, no. 8, e6482, 2009.

[83] M. Shimono and J. M. Beggs, "Functional clusters, hubs, and communities in the cortical microconnectome," *Cerebral Cortex*, vol. 25, no. 10, pp. 3743–3757, 2015.

[84] M. Kajiwara, R. Nomura, F. Goetze, *et al.*, "Inhibitory neurons exhibit high controlling ability in the cortical microconnectome," *PLoS Computational Biology*, vol. 17, no. 4, e1008846, 2021.

[85] E. Matsuda, T. Mita, J. Hubert, *et al.*, "Multiple time scales observed in spontaneously evolved neurons on high-density cmos electrode array," in *Artificial Life Conference Proceedings 13*, MIT Press, 2013, pp. 1075–1082.

[86] J. W. Aldridge and S. Gilman, "The temporal structure of spike trains in the primate basal ganglia: Afferent regulation of bursting demonstrated with precentral cerebral cortical ablation," *Brain Research*, vol. 543, no. 1, pp. 123–138, 1991.

[87] L. Rudelt, D. G. Marx, M. Wibral, and V. Priesemann, "Embedding optimization reveals long-lasting history dependence in neural spiking activity," *PLOS Computational Biology*, vol. 17, no. 6, e1008927, 2021.

[88] I. Nemenman, G. D. Lewen, W. Bialek, and R. R. D. R. Van Steveninck, "Neural coding of natural stimuli: Information at sub-millisecond resolution," *PLoS Computational Biology*, vol. 4, no. 3, e1000025, 2008.

[89] C. Kayser, N. K. Logothetis, and S. Panzeri, "Millisecond encoding precision of auditory cortex neurons," *Proceedings of the National Academy of Sciences*, vol. 107, no. 39, pp. 16 976–16 981, 2010.

[90] S. J. Sober, S. Sponberg, I. Nemenman, and L. H. Ting, "Millisecond spike timing codes for motor control," *Trends in Neurosciences*, vol. 41, no. 10, pp. 644–648, 2018.

[91] J. A. Garcia-Lazaro, L. A. Belliveau, and N. A. Lesica, "Independent population coding of speech with sub-millisecond precision," *Journal of Neuroscience*, vol. 33, no. 49, pp. 19 362–19 372, 2013.

CHAPTER 3

IMPROVING THE ESTIMATION OF TE ON SPIKE TRAINS

As discussed in Chapter 1, despite the fact that we know that brains perform a dizzying array of advanced computations, there is still much work to be done in revealing the information dynamics that undergird this ability. Investigations into the information processing operations of brains would ideally be performed on the finest scale for which we have abundant recordings, which is the spiking activity of individual neurons. Moreover, individual neurons are considered the fundamental computational units of the brain [1].

This thesis pays particularly close attention to the information transfer component of computation, as measured by the Transfer Entropy (TE). There have already been a number of studies which have applied TE to the spike times of neurons. However, although this work provided many valuable contributions, interpretability of results was limited by the traditional approach for estimating TE on spike trains.

The traditional approach towards estimating TE involved an initial step of time discretisation, whereby the process is divided into bins of a fixed width. The process is then transformed into a binary sequence, with the binary values signifying the presence or absence of spikes in bins. The TE is then estimated using a simple plugin estimator [2], [3]. See Section 2.7.1 for more details on this traditional method of estimation.

There are numerous problems with this method of estimation. Chief among these is that the estimator is not consistent, that is, it does not usually converge to the true value of the TE in the limit of infinite data. Moreover, it requires a hard tradeoff between the length of history dependence that can be captured and the temporal precision. Small time bins allow for better time precision, but reduce the length of the history effects that can be studied. Large time bins have the opposite effect. It is not possible for this estimator to perform both simultaneously (for any realistically sized dataset). Moreover, if we would like to estimate the total information flow, it is not sufficient to sum up estimates performed at different time scales, as this would ignore any synergistic effects between them.

In order to make possible the high-fidelity study of information flows on spike trains that we will perform in subsequent chapters, this chapter presents a novel estimator of TE on spike trains. This estimator makes use of a recently-developed continuous-time formalism for information dynamics [4], [5] in order to operate without time-discretisation by performing estimation using the inter-spike intervals. The resulting estimator is provably consistent (section 3.4.1). Further, the use of the inter-spike intervals for the history embeddings provides an efficient representation of the history of the

spike train, allowing for relatively long histories to be examined, with no loss of precision and minimal use of extra dimensions.

This new estimator makes possible, for the first time, the high fidelity study of information flows from spike-train data. This then allows us to perform the study of information flows in developing neural cell cultures that we do in Chapter 4. The efficient use of dimension in the history embeddings of this new estimator allows for the use of larger sets of conditioning processes. This, in turn, makes it possible for TE to be used in the inference of effective networks from spike times. We demonstrate this capability in Chapter 5.

[1] F. López-Muñoz, J. Boya, and C. Alamo, "Neuron theory, the cornerstone of neuroscience, on the centenary of the nobel prize award to santiago ramón y cajal," *Brain Research Bulletin*, vol. 70, no. 4-6, pp. 391–405, 2006.

[2] L. Wasserman, *All of nonparametric statistics*. Springer Science & Business Media, 2006.

[3] T. Bossomaier, L. Barnett, M. Harré, and J. T. Lizier, "An introduction to transfer entropy," *Cham: Springer International Publishing*, vol. 65, 2016.

[4] R. E. Spinney and J. T. Lizier, "Characterizing information-theoretic storage and transfer in continuous time processes," *Physical Review E*, vol. 98, no. 1, p. 012 314, 2018.

[5] R. E. Spinney, M. Prokopenko, and J. T. Lizier, "Transfer entropy in continuous time, with applications to jump and neural spiking processes," *Physical Review E*, vol. 95, no. 3, p. 032 319, 2017.

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Estimating Transfer Entropy in Continuous Time Between Neural Spike Trains or Other Event-Based Data

David P. Shorten[1]*, **Richard E. Spinney**[1,2], **Joseph T. Lizier**[1]

1 Complex Systems Research Group and Centre for Complex Systems, Faculty of Engineering, The University of Sydney, Sydney, Australia, **2** School of Physics and EMBL Australia Node Single Molecule Science, School of Medical Sciences, The University of New South Wales, Sydney, Australia

* david.shorten@sydney.edu.au

OPEN ACCESS

Citation: Shorten DP, Spinney RE, Lizier JT (2021) Estimating Transfer Entropy in Continuous Time Between Neural Spike Trains or Other Event-Based Data. PLoS Comput Biol 17(4): e1008054. https://doi.org/10.1371/journal.pcbi.1008054

Editor: Daniele Marinazzo, Ghent University, BELGIUM

Received: June 10, 2020

Accepted: February 19, 2021

Published: April 19, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1008054

Data Availability Statement: The code for generating the datasets is available in a public repository: https://github.com/dpshorten/CoTETE_experiments.

Abstract

Transfer entropy (TE) is a widely used measure of directed information flows in a number of domains including neuroscience. Many real-world time series for which we are interested in information flows come in the form of (near) instantaneous events occurring over time. Examples include the spiking of biological neurons, trades on stock markets and posts to social media, amongst myriad other systems involving events in continuous time throughout the natural and social sciences. However, there exist severe limitations to the current approach to TE estimation on such event-based data via discretising the time series into time bins: it is not consistent, has high bias, converges slowly and cannot simultaneously capture relationships that occur with very fine time precision as well as those that occur over long time intervals. Building on recent work which derived a theoretical framework for TE in continuous time, we present an estimation framework for TE on event-based data and develop a *k*-nearest-neighbours estimator within this framework. This estimator is provably consistent, has favourable bias properties and converges orders of magnitude more quickly than the current state-of-the-art in discrete-time estimation on synthetic examples. We demonstrate failures of the traditionally-used source-time-shift method for null surrogate generation. In order to overcome these failures, we develop a local permutation scheme for generating surrogate time series conforming to the appropriate null hypothesis in order to test for the statistical significance of the TE and, as such, test for the conditional independence between the history of one point process and the updates of another. Our approach is shown to be capable of correctly rejecting or accepting the null hypothesis of conditional independence even in the presence of strong pairwise time-directed correlations. This capacity to accurately test for conditional independence is further demonstrated on models of a spiking neural circuit inspired by the pyloric circuit of the crustacean stomatogastric ganglion, succeeding where previous related estimators have failed.

PLOS COMPUTATIONAL BIOLOGY

Estimating Transfer Entropy in Continuous Time Between Neural Spike Trains

Funding: JL was supported through the Australian Research Council DECRA grant DE160100630 - https://www.arc.gov.au/grants/discovery-program/discovery-early-career-researcher-award-decra and The University of Sydney Research Accelerator (SOAR) Fellowship program - https://sydney.edu.au/research/our-researchers/sydney-research-accelerator-fellows.html. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Transfer Entropy (TE) is an information-theoretic measure commonly used in neuroscience to measure the directed statistical dependence between a source and a target time series, possibly also conditioned on other processes. Along with measuring information flows, it is used for the inference of directed functional and effective networks from time series data. The currently-used technique for estimating TE on neural spike trains first time-discretises the data and then applies a straightforward plug-in information-theoretic estimation procedure. This approach has numerous drawbacks: it has high bias, cannot capture relationships occurring on both fine and large timescales simultaneously, converges very slowly as more data is obtained, and indeed does not even converge to the correct value for any practical non-vanishing discretisation scale. We present a new estimator for TE which operates in continuous time and demonstrate, via application to synthetic examples, that it addresses these problems and can reliably differentiate statistically significant flows from (conditionally) independent spike trains. Further, we also apply it to more biologically-realistic spike trains obtained from a biophysical model inspired by the pyloric circuit of the crustacean stomatogastric ganglion; our correct inference of directed conditional dependence and independence between neurons here provides an important validation for our approach where similar methods have previously failed.

This is a *PLOS Computational Biology* Methods paper.

Introduction

In analysing time series data from complex dynamical systems, such as in neuroscience, it is often useful to have a notion of information flow. We intuitively describe the activities of brains in terms of such information flows: for instance, information from the visual world must flow to the visual cortex where it will be encoded [1]. Further, information coded in the motor cortex must flow to muscles where it will be enacted [2].

Transfer entropy (TE) [3, 4] has become a widely accepted measure of such flows. It is defined as the mutual information between the past of a source time-series process and the present state of a target process, conditioned on the past of the target. More specifically (in discrete time), the transfer entropy rate [5] is:

$$\dot{\mathbf{T}}_{Y\to X} = \frac{1}{\Delta t} I(X_t\,;\,\mathbf{Y}_{<t}\,|\,\mathbf{X}_{<t}) = \frac{1}{\tau}\sum_{t=1}^{N_T} \ln \frac{p(x_t\,|\,\mathbf{x}_{<t},\mathbf{y}_{<t})}{p(x_t\,|\,\mathbf{x}_{<t})}. \tag{1}$$

Here the information flow is being measured from a source process Y to a target X, $I(\cdot;\cdot|\cdot)$ is the conditional mutual information [6], $p(\cdot|\cdot)$ is a conditional probability, x_t is the current state of the target, $\mathbf{x}_{<t}$ is the history of the target, $\mathbf{y}_{<t}$ is the history of the source, Δt is the interval between time samples (in units of time), τ is the length of the time series and $N_T = \tau/\Delta t$ is the number of time samples. The histories $\mathbf{x}_{<t}$ and $\mathbf{y}_{<t}$ are usually captured via embedding vectors, e.g. $\mathbf{x}_{<t} = \mathbf{x}_{t-m:t-1} = \{x_{t-m}, x_{t-m+1}, \ldots, x_{t-1}\}$. The average here is taken over time, as opposed to possible states and histories (both formulations are equivalent under the assumptions of stationarity and ergodicity). Recent work [5] has highlighted the importance of normalising the TE by the width of the time bins, as above, such that it becomes a *rate*, in order to ensure convergence in the limit of small time bin size.

It is also possible to condition the TE on additional processes [4]. Given additional processes $\mathscr{Z} = \{Z_1, Z_2, \ldots, Z_{n_{\mathscr{Z}}}\}$ with histories $\mathscr{Z}_{<t} = \{\mathbf{Z}_{1,<t}, \mathbf{Z}_{2,<t}, \ldots, \mathbf{Z}_{n_{\mathscr{Z}},<t}\}$, we can write the conditional TE rate as

$$\dot{\mathbf{T}}_{Y \to X|\mathscr{Z}} = \frac{1}{\Delta t} I(X_t ; \mathbf{Y}_{<t} \,|\, \mathbf{X}_{<t}, \mathscr{Z}_{<t}). \tag{2}$$

When combined with a suitable statistical significance test, the TE (and conditional TE) can be used to show that the present state of X is conditionally independent of the past of Y–when conditioned on the past of X (and on the conditional processes $\mathscr{Z}$). Of course, we refer to conditional independence in the statistical sense (i.e. $p(x_t \,|\, \mathbf{x}_{<t}, \boldsymbol{\varkappa}_{<t}, \mathbf{y}_{<t}) = p(x_t \,|\, \mathbf{x}_{<t}, \boldsymbol{\varkappa}_{<t})$) rather than the causal sense. Such a conditional independence test can be used as a component in a network inference algorithm and, as such, TE is widely used for inferring directed functional and effective network models [7, 8, 9, 10, 11, 12] (and see [4, Sec. 7.2] for a review).

TE has enjoyed widespread application in neuroscience in particular [13, 14]. Uses have included the functional/effective network inference as mentioned above, as well as the measurement of the direction and magnitude of information flows [15, 16] and the determination of transmission delays [17]. Such applications have been performed using data from multiple diverse sources such as MEG [18, 19], EEG [20], fMRI [21], electrode arrays [22], calcium imaging [9] and simulations [23].

Previous applications of TE to spike trains [22, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34] and other types of event-based data [35], including for the purpose of network inference [9, 36, 37], have relied on time discretisation. As shown in Fig 1, the time series is divided into small bins of width Δt. The value of a sample for each bin could then be assigned a binary value—denoting the presence or absence of events (spikes) in the bin—or a natural number denoting the number of events (spikes) that fell within the bin (the experiments in this paper use the former). A choice is made as to the number of time bins, l and m, to include in the source and target history embeddings $\mathbf{y}_{<t}$ and $\mathbf{x}_{<t}$. This results in a finite number of possible history embeddings. For a given combination $\mathbf{x}_{<t}$ and $\mathbf{y}_{<t}$, the probability of the target's value in the current bin conditioned on these histories, $p(x_t|\mathbf{x}_{<t}, \mathbf{y}_{<t})$, can be directly estimated using the plugin (histogram) [38] estimator. The probability of the target's value in the current bin conditioned on only the target history, $p(x_t|\mathbf{x}_{<t})$, can be estimated in the same fashion. From these estimates the TE can be calculated in a straightforward manner via Eq (1). See Results for a description of the application of the discrete time TE estimator to synthetic examples including spiking events from simulations of model neurons.

There are two large disadvantages to this approach [5]. If the process is genuinely occurring in discrete time, then the estimation procedure just described is consistent. That is, it is guaranteed to converge to the true value of the TE in the limit of infinite data. However, if we are considering a fundamentally continuous-time process (with full measurement precision), such as a neuron's action potential, then the lossy transformation of time discretisation ($\Delta t > 0$) will result in an inaccurate estimate of the TE. Thus, in these cases, any estimator based on time discretisation is not consistent. Secondly, whilst the loss of resolution of the discretization will reduce with decreasing bin size Δt, this requires larger dimensionality in the history embeddings to capture correlations over similar time intervals. This increase in dimension will result in an exponential increase in the state space size being sampled to estimate $p(x_t|\mathbf{x}_{<t}, \mathbf{y}_{<t})$, and therefore the data requirements. However, some recordings of the activities of neurons are done with low time precision. For example, recordings from calcium imaging experiments usually use a sampling rate of around 1 to 10 Hz [39]. In such cases, we could use bin sizes on the order of the experimental precision and still capture a reasonable history

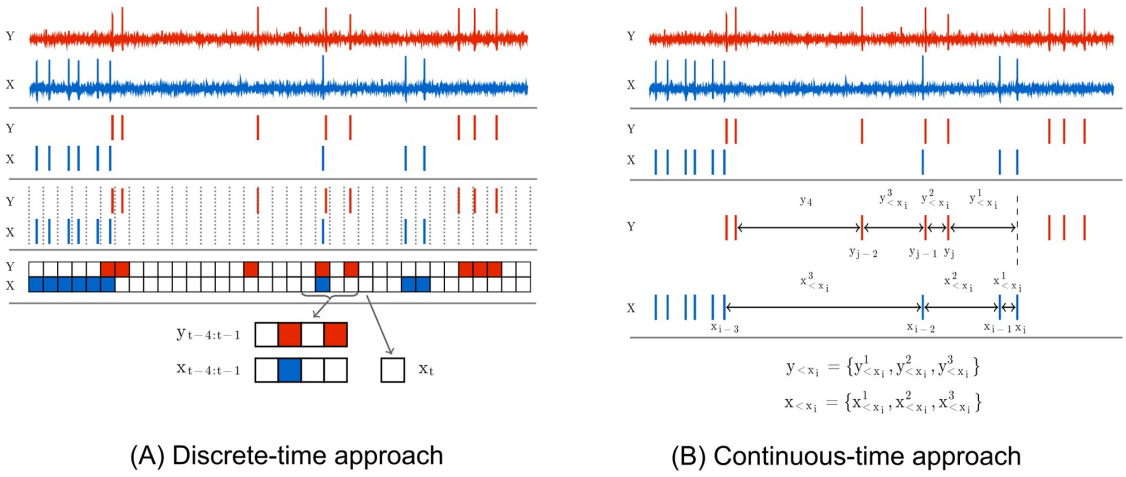


Fig 1. Diagrams highlighting the differences in the embeddings used by the discrete and continuous-time estimators. The discrete-time estimator (A) divides the time series into time bins. A binary value is assigned to each bin denoting the presence or absence of a (spiking) event–alternatively, this could be a natural number to represent the occurrence of multiple events. The process is thus recast as a sequence of binary values and the history embeddings ($\mathbf{x}_{t-4:t-1}$ and $\mathbf{y}_{t-4:t-1}$) for each point are binary vectors. The probability of an event occurring in a bin, conditioned on its associated history embeddings, is estimated via the plugin (histogram) [38] estimator. Conversely, the continuous-time estimator (B) performs no time binning. History embeddings $\mathbf{x}_{<x_i}$ and $\mathbf{y}_{<x_i}$ for events or $\mathbf{x}_{<u_i}$ and $\mathbf{y}_{<u_i}$ for arbitrary points in time (not shown in this figure, see Fig 10) are constructed from the raw interspike intervals. This approach estimates the TE by comparing the probabilities of the history embeddings of the target processes' history as well as the joint history of the target and source processes at both the (spiking) events and arbitrary points in time.

https://doi.org/10.1371/journal.pcbi.1008054.g001

length with history embeddings composed of only a small number of bins. This might keep the size of the history state space small enough that we can collect an adequate number of samples for each history permutation with the available data. In such cases, we might expect the discrete-time approach to perform as well as can be expected given the limitations imposed by the apparatus. On the other hand, data from microelectrode arrays can be sampled at rates over 70 kHz [40]. When using data collected with this high temporal precision, if we use bin sizes corresponding to the sampling rate, we will be forced to use incredibly short history embeddings in order to avoid the size of the history state space growing to a point where it can no longer be sampled.

In practice then, if the data has been collected with fine temporal precision, the application of transfer entropy to event-based data such as spike trains has often required a trade-off between fully resolving interactions that occur with fine time precision and capturing correlations that occur across long time intervals. There is substantial evidence that spike correlations at the millisecond and sub-millisecond scale play a role in encoding visual stimuli [41, 42], motor control [43] and speech [44]. On the other hand, correlations in spike trains exist over lengths of hundreds of milliseconds [45]. A discrete-time TE estimator cannot capture both of these classes of effects simultaneously, and remains heavily dependent on the value of Δt [31].

Recent work by Spinney et al. [5] derived a continuous-time formalism for TE. It was demonstrated that, for stationary point processes such as spike trains, the pairwise TE rate is given by:

$$\dot{\mathrm{T}}_{Y\to X} = \lim_{\tau\to\infty} \frac{1}{\tau} \sum_{i=1}^{N_X} \ln \frac{\lambda_{x|\mathbf{x}_{<t},\mathbf{y}_{<t}}[\mathbf{x}_{<x_i}, \mathbf{y}_{<x_i}]}{\lambda_{x|\mathbf{x}_{<t}}[\mathbf{x}_{<x_i}]}. \tag{3}$$

Here N_X is the number of events in the target process and τ is the length in time of this process

whilst $\lambda_{x|\mathbf{x}_{<t},\mathbf{y}_{<t}}[\mathbf{x}_{<x_i}, \mathbf{y}_{<x_i}]$ is the instantaneous firing rate of the target conditioned on the histories of the target $\mathbf{x}_{<x_i}$ and source $\mathbf{y}_{<x_i}$ at the time points x_i of the events in the target process. $\lambda_{x|\mathbf{x}_{<t}}[\mathbf{x}_{<x_i}]$ is the instantaneous firing rate of the target conditioned on its history alone, ignoring the history of the source. Note that $\lambda_{x|\mathbf{x}_{<t},\mathbf{y}_{<t}}$ and $\lambda_{x|\mathbf{x}_{<t}}$ are defined at all points in time and not only at target events. It is worth emphasizing that, in this context, the processes X and Y are series of the time points x_i and y_j of the events i and j in the target and source respectively. This is contrasted with Eq (1), where X and Y are time series of values at the sampled time points t_i. To avoid confusion we use the notation that the $y_j \in Y$ are the raw time points and $\mathbf{y}_{<x_i}$ is some representation of the history of Y observed at the time point x_i (see Methods).

Eq (3) can easily be adapted to the conditional case:

$$\dot{\mathbf{T}}_{Y\to X|\mathscr{Z}} = \lim_{\tau\to\infty} \frac{1}{\tau} \sum_{i=1}^{N_X} \ln \frac{\lambda_{x|\mathbf{x}_{<t},\mathbf{y}_{<t},\boldsymbol{\varkappa}_{<t}}[\mathbf{x}_{<x_i}, \mathbf{y}_{<x_i}, \boldsymbol{\varkappa}_{<x_i}]}{\lambda_{x|\mathbf{x}_{<t},\boldsymbol{\varkappa}_{<t}}[\mathbf{x}_{<x_i}, \boldsymbol{\varkappa}_{<x_i}]}. \tag{4}$$

Here $\lambda_{x|\mathbf{x}_{<t},\mathbf{y}_{<t},\boldsymbol{\varkappa}_{<t}}[\mathbf{x}_{<x_i}, \mathbf{y}_{<x_i}, \boldsymbol{\varkappa}_{<x_i}]$ is the instantaneous firing rate of the target conditioned on the histories of the target $\mathbf{x}_{<x_i}$, source $\mathbf{y}_{<x_i}$ and other possible conditioning processes $\boldsymbol{\varkappa}_{<x_i} = \{\mathbf{z}_{1,<x_i}, \mathbf{z}_{2,<x_i}, \ldots, \mathbf{z}_{n_{\mathscr{Z}},<x_i}\}$. $\lambda_{x|\mathbf{x}_{<t},\boldsymbol{\varkappa}_{<t}}[\mathbf{x}_{<x_i}, \boldsymbol{\varkappa}_{<x_i}]$ is the instantaneous firing rate of the of the target conditioned on the histories of the target and the additional conditioning processes, ignoring the history of the source.

Crucially, it was demonstrated by Spinney et al., and later shown more rigorously by Cooper and Edgar [46], that if the discrete-time formalism of the TE (in Eq (1)) could be properly estimated as $\lim_{\Delta t \to 0}$, then it would converge to the same value as the continuous-time formalism. This is due to the contributions to the TE from the times between target events vanishing in expectation. Yet there are two important distinctions in the continuous-time formalism which hold promise to address the consistency issues of the discrete-time formalism. Firstly, the basis in continuous time allows us to efficiently represent the history embeddings by inter-event intervals, suggesting the possibility of jointly capturing subtleties in both short and long time-scale effects that has evaded discrete-time approaches. See Fig 1 for a diagrammatic representation of these history embeddings, contrasted with the traditional way of constructing histories for the discrete-time estimator. Secondly, note the important distinction that the sums in Eqs (3) and (4) are taken over the N_X (spiking) events in the target during the time-series over interval τ; this contrasts to a sum over all time-steps in the discrete-time formalism. An estimation strategy based on Eqs (3) and (4) would only be required to calculate quantities at events, ignoring the inter-event interval time where the neuron is quiescent. This implies a potential computational advantage, as well as eliminating one source of estimation variability.

These factors all point to the advantages of estimating TE for event-based data using the continuous-time formalism in Eqs (3) and (4). This paper presents an empirical approach to performing such estimation. The estimator (presented in Methods) operates by considering the probability densities of the history embeddings observed at events and contrasts these with the probability densities of those embeddings being observed at other (randomly sampled) points. This approach is distinct in utilising a novel Bayesian inversion on Eq (4) in order to operate on these probability densities of the history embeddings, rather than making a more difficult direct estimation of spike rates. Furthermore, this allows us to utilise k-Nearest-Neighbour (kNN) estimators for the entropy terms based on these probability densities. These estimators have known advantages of consistency, data efficiency, low sensitivity to parameters and known bias corrections. By combining these entropy estimators, and making use of established bias reduction techniques for combinations of kNN estimators, we arrive at our proposed estimator. The resulting estimator is consistent (see Methods) and is demonstrated on

synthetic examples in Results to be substantially superior to estimators based on time discretisation across a number of metrics.

To conclude that there exists non-zero TE (and thus establish conditional dependence) between two processes a suitable hypothesis test is required. This is usually done by creating a surrogate population of processes (or samples of histories) which conform to the null hypothesis of zero TE, or in other words, directed conditional independence of the target spikes from the source. The algorithm which we use should create surrogates which are identically distributed to the original processes (or history samples) if and only if the null hypothesis holds [47]. The historically used method for generating these surrogates was to either shuffle the original source samples or to shift the source process in time. However, this results in surrogates which conform to an incorrect null hypothesis–that the transitions in the target are completely independent of the source histories. That is, they conform to the factorisation $p(X_t, \mathbf{Y}_{<t} \mid \mathbf{X}_{<t}, \boldsymbol{\mathcal{Z}}_{<t}) = p(X_t \mid \mathbf{X}_{<t}, \boldsymbol{\mathcal{Z}}_{<t}) p(\mathbf{Y}_{<t})$. In cases where there is a pairwise correlation between the present state of the target and the history of the source, but they are nonetheless conditionally independent, shuffling or time shifting will create surrogates that are not identically distributed to the original history samples. This is despite the fact that the null hypothesis holds. This can result in the estimate of the TE on the original processes being statistically different from those on the surrogate population, leading to the incorrect inference of non-zero TE.

As shown in Results, this can lead to incredibly high false positive rates for conditional dependence in certain settings such as the presence of strong common driver effects. Therefore, in order to have a suitable significance test for use in conjunction with the proposed estimator, we also present (in Methods) an adaptation of a recently proposed local permutation method [48] to our specific case. This adapted scheme produces surrogates which conform to the correct null hypothesis of conditional independence of the present of the target and the source history, given the histories of the target and further conditioning processes. This is the condition that $p(X_t, \mathbf{Y}_{<t} \mid \mathbf{X}_{<t}, \boldsymbol{\mathcal{Z}}_{<t}) = p(X_t \mid \mathbf{X}_{<t}, \boldsymbol{\mathcal{Z}}_{<t}) p(\mathbf{Y}_{<t} \mid \mathbf{X}_{<t}, \boldsymbol{\mathcal{Z}}_{<t})$.

It is easy to intuit that the second factorisation is correct by rewriting the discrete-time TE (Eq (2)) as:

$$\dot{\mathbf{T}}_{Y \to X} = \frac{1}{\tau} \sum_{t=1}^{N_T} \ln \frac{p(x_t, \mathbf{y}_{<t} \mid \mathbf{x}_{<t}, \boldsymbol{\mathcal{z}}_{<t})}{p(x_t \mid \mathbf{x}_{<t}, \boldsymbol{\mathcal{z}}_{<t}) p(\mathbf{y}_{<t} \mid \mathbf{x}_{<t}, \boldsymbol{\mathcal{z}}_{<t})}. \tag{5}$$

That is, transfer entropy can be readily interpreted as a measure of the difference between the distributions $p(X_t, \mathbf{Y}_{<t} \mid \mathbf{X}_{<t}, \boldsymbol{\mathcal{Z}}_{<t})$ and $p(X_t \mid \mathbf{X}_{<t}, \boldsymbol{\mathcal{Z}}_{<t}) p(\mathbf{Y}_{<t} \mid \mathbf{X}_{<t}, \boldsymbol{\mathcal{Z}}_{<t})$.

We show in Results that the combination of the proposed estimator and surrogate generation method is capable of correctly distinguishing between zero and non-zero information flow in difficult cases, such as where the history of the source has a strong pairwise correlation with the occurrence of events in the target, but is nevertheless conditionally independent. The combination of the current state-of-the-art in discrete-time estimation and a traditional method of surrogate generation is shown to be incapable of making this distinction.

Similarly, we demonstrate that the proposed combination is capable of correctly distinguishing between conditional dependence and independence relationships in data taken from a simple circuit of biophysical model neurons inspired by the crustacean stomatogastric ganglion [49]. Despite the presence of strong pairwise correlations, the success of our estimator here contrasts not only with known failure of a related Granger causality estimator, but also our demonstration that the discrete-time estimator is incapable of correctly performing this task.

Our results provide strong impetus for the application of our proposed techniques to investigate information flows in spike-train data recorded from biological neurons. Furthermore, we underline the importance of our correct identification of conditional dependence and independence in these experiments. Whilst functional/effective network inference algorithms using TE estimators such as ours are *not* expected to align with structural networks in general, they would be expected to do so under certain idealised assumptions (e.g. full observability, large sample size, etc., as outlined in Methods 16) implemented in these experiments. As recently discussed by Novelli and Lizier [50], and specifically for spiking neural networks by Das and Fiete [51], inference aligning with underlying structure under such conditions is a crucial validation that the effective network models they infer at scale are readily interpretable. As such, the demonstration of the efficacy of our proposed approach to detecting conditional dependence in small networks here implies that it holds promise for larger scale effective network inference once paired with a suitable (conditional-independence-based) network inference algorithm (e.g. IDTxl as described in [7, 52]).

Results

The first two subsections here present the results of the continuous-time estimator applied to two different synthetic examples for which the ground truth value of the TE is known. The first example considers independent processes where $\dot{\mathbf{T}}_{Y\to X} = 0$, whilst the second examines coupled point processes with a known, non-zero $\dot{\mathbf{T}}_{Y\to X}$. The continuous-time estimator's performance is also contrasted with that of the discrete-time estimator. The emphasis of these sections is on properties of the estimators in isolation, as opposed to when combined with a statistical test. As such, we focus on the estimators' bias, variance and consistency (see Methods).

The third, fourth and fifth subsections present the results of the combination of the continuous-time estimator and the local permutation surrogate generation scheme applied to two examples: the first two synthetic and the last a biologically plausible model of neural activity. The comparison of the estimates to a population of surrogates produces p-values for the null hypothesis of zero TE. Rejection of this null hypothesis and the resulting conclusion of non-zero TE implies a directed statistical dependence. The results are compared to the known connectivity of the studied systems. Whilst we do not expect directed statistical dependence to have a one-to-one correspondence with structural connectivity in general, these experiments are designed under ideal conditions such that they would. This provides important test cases for detection of conditional dependence and independence. These p-values could be translated into other metrics such as ROC curves and false-positive rates, but we choose to instead visualise the distributions of the p-values themselves. The combination of the discrete-time estimator along with a traditional method for surrogate generation (time shifts) is also applied to these examples for comparison.

No TE between independent homogeneous poisson processes

The simplest processes on which we can attempt to validate the estimator are independent homogeneous Poisson processes, where the true value of the TE between such processes is zero.

Pairs of independent homogeneous Poisson processes were generated, each with rate $\bar{\lambda} = 1$, and contiguous sequences of $N_X \in \{1 \times 10^2, 1 \times 10^3, 1 \times 10^4, 1 \times 10^5\}$ target events were selected. For the continuous-time estimator, the parameter N_U for the number of placed sample points was varied (see Methods) to check the sensitivity of estimates to this parameter. At

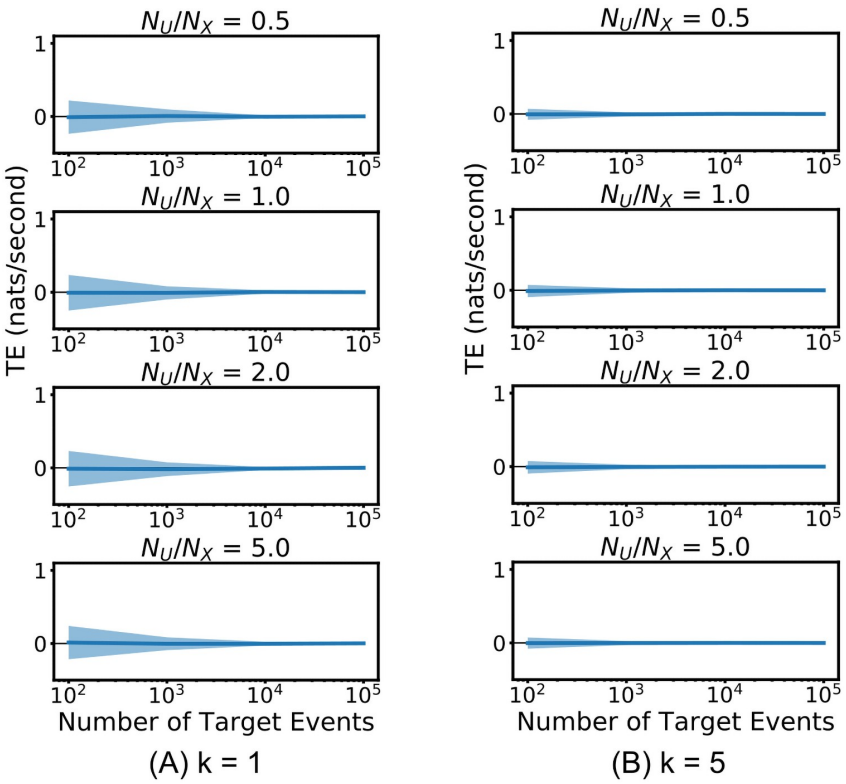


Fig 2. Evaluation of the continuous-time estimator on independent homogeneous Poisson processes. The solid line shows the average TE rate across multiple runs and the shaded area spans from one standard deviation below the mean to one standard deviation above it. Plots are shown for two different values of k nearest neighbours, and four different values of the ratio of the number of sample points to the number of events N_U/N_X (See Methods).

https://doi.org/10.1371/journal.pcbi.1008054.g002

each of these numbers of target events N_X, the averages are taken across 1000, 100, 20 and 20 tested process pairs respectively.

Fig 2 shows the results of these runs for the continuous-time estimator, using various parameter settings. In all cases, the Manhattan (ℓ_1) norm is used as the distance metric and the embedding lengths are set to $l_X = l_Y = 1$ spike. See Methods for a description of these parameters. See also [52] for a discussion on how to set these embedding lengths. For this example, the set of conditioning processes $\mathscr{Z}$ is empty. S1 Fig shows results with longer history embeddings.

The plots show that the continuous-time estimator converges to the true value of the TE (equal to 0). This is a numerical confirmation of its consistency for independent processes. Moreover, it exhibits very low bias (as compared to the discrete-time estimator, Fig 3) for all values of the k nearest neighbours and N_U/N_X parameters. The variance is relatively large for $k = 1$, although it dramatically improves for $k = 5$—this reflects known results for variance of this class of estimators as a function of k, where generally k above 1 is recommended [53].

Fig 3 shows the result of the discrete-time estimator applied to the same independent homogeneous Poisson processes for two different combinations of the source and target history embedding lengths, l and m time bins, and four different bin sizes Δt (see S2 Fig for

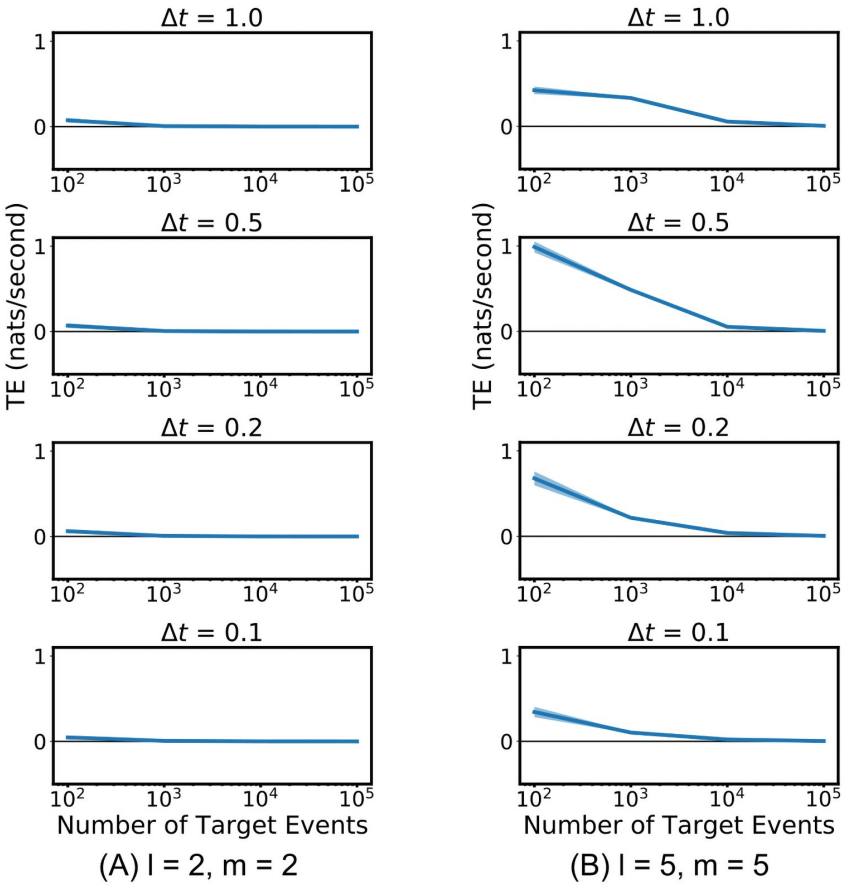


Fig 3. Result of the discrete-time estimator applied to independent homogeneous Poisson processes. The solid line shows the average TE rate across multiple runs and the shaded area spans from one standard deviation below the mean to one standard deviation above it. Plots are shown for four different values of the bin width Δt as well as different source and target embedding lengths, l and m.

https://doi.org/10.1371/journal.pcbi.1008054.g003

different choices of l and m). At each of the numbers of target events N_X, the averages are taken across 1000, 100, 100 and 100 tested process pairs respectively. The variance of this estimator on this process is low and comparable to the continuous-time estimator, however the bias is very large and positive for short processes. The bias of both estimators could be reduced by subtracting the mean of the estimates over a population of surrogates (see the following subsection for an example of this being done with the continuous-time approach). We do observe the discrete-time estimator converging to zero (the true value of the TE) as we increase the available data. This would suggest that it might be consistent on this specific example. However, we will shortly encounter an example where this is not the case.

Consistent TE between unidirectionally coupled processes

The estimators were also tested on an example of unidirectionally coupled spiking processes with a known value of TE (previously presented as example B in [5]). Here, the source process Y is a homogoneous Poisson process. The target process X is produced as a conditional point

process where the instantaneous rate is a function of the time since the most recent source event. More specifically:

$$\lambda_{y|\mathbf{x}_{<t},\mathbf{y}_{<t}}[\mathbf{x}_{<t},\mathbf{y}_{<t}] = \bar{\lambda}_y$$

$$\lambda_{x|\mathbf{x}_{<t},\mathbf{y}_{<t}}[\mathbf{x}_{<t},\mathbf{y}_{<t}] = \lambda_x\left[t_y^1\right] = \begin{cases} \lambda_x^{\text{base}} & t_y^1 > t_{\text{cut}} \\ \lambda_x^{\text{base}} + m\exp\left[-\frac{1}{2\sigma^2}\left(t_y^1 - \frac{t_{\text{cut}}}{2}\right)^2\right] & \\ \quad - m\exp\left[-\frac{1}{2\sigma^2}\left(-\frac{t_{\text{cut}}}{2}\right)^2\right] & t_y^1 \leq t_{\text{cut}}. \end{cases}$$

Here, t_y^1 is the time since the most recent source event. As a function of t_y^1, the target spike rate $\lambda_{x|\mathbf{x}_{<t},\mathbf{y}_{<t}}[\mathbf{x}_{<t},\mathbf{y}_{<t}]$ rises from a baseline λ_x^{base} at $t_y^1 = 0$ to a peak at $t_y^1 = t_{\text{cut}}/2$, before falling back to the baseline λ_x^{base} from $t_y^1 = t_{\text{cut}}$ onwards (see Fig 4A). We simulated this process using the parameter values $\bar{\lambda}_y = 0.5$, $m = 5$, $t_{\text{cut}} = 1$, $\lambda_x^{\text{base}} = 0.5$ and $\sigma^2 = 0.01$. This simulation was performed using a thinning algorithm [54]. Specifically, we first generated the source process at rate $\bar{\lambda}_y$. We then generate the target as a homogeneous Poisson process with rate λ_h such that $\lambda_h > \lambda_x[t_y^1]$ for all values of t_y^1. We then went back through all the events in this process and removed each event with probability $1 - \lambda_x[t_y^1]/\lambda_h$. As with the previous example, once a pair of processes had been generated, a contiguous sequence of N_X target events was selected. Tests were conducted for the values of $N_X \in \{1 \times 10^2, 1 \times 10^3, 1 \times 10^4, 1 \times 10^5, 1 \times 10^6\}$. For the continuous-time estimator, the number of placed sample points N_U was set equal to N_X (see Methods). At each N_X, the averages are taken over 1000, 100, 20, 20 and 20 tested process pairs respectively.

Spinney et al. [5] present a numerical method for calculating the TE for this process, based on known conditional firing rates in the system under stationary conditions. For the parameter values used here the true value of the TE is 0.5076 ± 0.001.

Given that we know that the dependence of the target on the source is fully determined by the distance to the most recent event in the source, we used a source embedding length of $l_Y = 1$. The estimators were run with three different values of the target embedding length $l_X \in \{1, 2, 3\}$ (see Methods). For this example, the set of conditioning processes $\mathscr{Z}$ is empty.

Fig 4B shows the results of the continuous-time estimator applied to the simulated data. We used the value of $k = 4$ and the Manhattan (ℓ_1) norm. The results displayed are as expected in that for a short target history embedding length of $l_X = 1$ spike, the estimator converges to a slight over-estimate of the TE. The overestimate at shorter target history embedding lengths l_X can be explained in that perfect estimates of the $\sum_{i=1}^{N_X} \ln \lambda_{x|\mathbf{x}_{<t}}[\mathbf{x}_{<t}]$ component require full knowledge of the target past within the previous $t_{\text{cut}} = 1$ time unit; shorter values of l_X don't cover this period in many cases, leaving this rate underestimated and therefore the TE overestimated. For longer values of $l_X \in \{2, 3\}$ we see that they converge closely to the true value of the TE. This is a further numerical confirmation of the consistency of the continuous-time estimator. See S1 Fig for plots with a different value of l_Y.

Fig 4C shows the results of the discrete-time estimator applied to the same process, run for three different values of the bin width $\Delta t \in \{1, 0.5, 0.2, 0.1\}$ time units. The number of bins included in the history embeddings was chosen such that they extended one time unit back (the known length of the history dependence). Smaller bin sizes could not be used as this leads to undersampling of the possible history permutations, resulting in far inferior performance. The plots are a clear demonstration that the discrete-time estimator is very biased and not consistent. At a bin size of $\Delta t = 0.2$ it converges to a value around half the true TE. Moreover, its

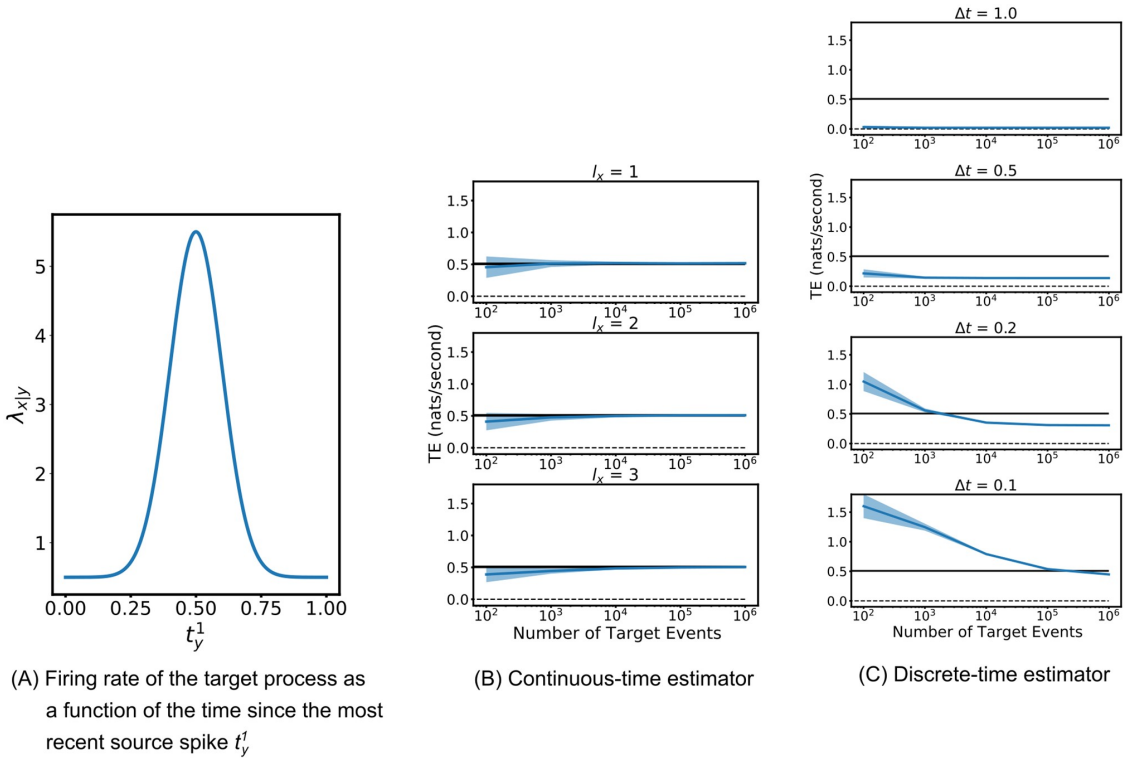


Fig 4. The discrete-time and continuous-time estimators were run on coupled point processes for which the ground-truth value of the TE is known. (A) shows the firing rate of the target process as a function of the history of the source. (B) and (C) show the estimates of the TE provided by the two estimators. The solid blue line shows the average TE rate across multiple runs and the shaded area spans from one standard deviation below the mean to one standard deviation above it. The black line shows the true value of the TE. For the continuous-time estimator the parameter values of $N_U/N_X = 1$ and $k = 4$ were used along with the ℓ_1 (Manhattan) norm. Plots are shown for three different values of the length of the target history component l_X. For the discrete-time estimator, plots are shown for four different values of the bin width Δt. The source and target history embedding lengths are chosen such that they extend back one time unit (the known length of the history dependence).

https://doi.org/10.1371/journal.pcbi.1008054.g004

convergence is incredibly slow. At the bin size of $\Delta t = 0.1$ it would appear to not have converged even after 1 million target events, and indeed it is not even converging to the true value of the TE. The significance of the performance improvement by our estimator is explored further in Discussion.

Identifying conditional independence despite strong pairwise correlations

The existence of a set of conditioning processes under which the present of the target component is conditionally independent of the past of the source implies that, under certain assumptions, there is no causal connection from the source to the target [55, 56, 57] (see Methods for details on the assumptions we use to conclude the ground truth of dependence/independence in the examples we use here). More importantly, TE can be used to test for such conditional independence (see Methods), thus motivating its use in directed functional (using pairwise TE) and effective (using multivariate TE) network inference. A large challenge faced in testing for conditional independence is correctly identifying "spurious" correlations, whereby conditionally independent components might have a strong pairwise correlation. This problem is particularly pronounced when investigating the spiking activity of biological neurons, as

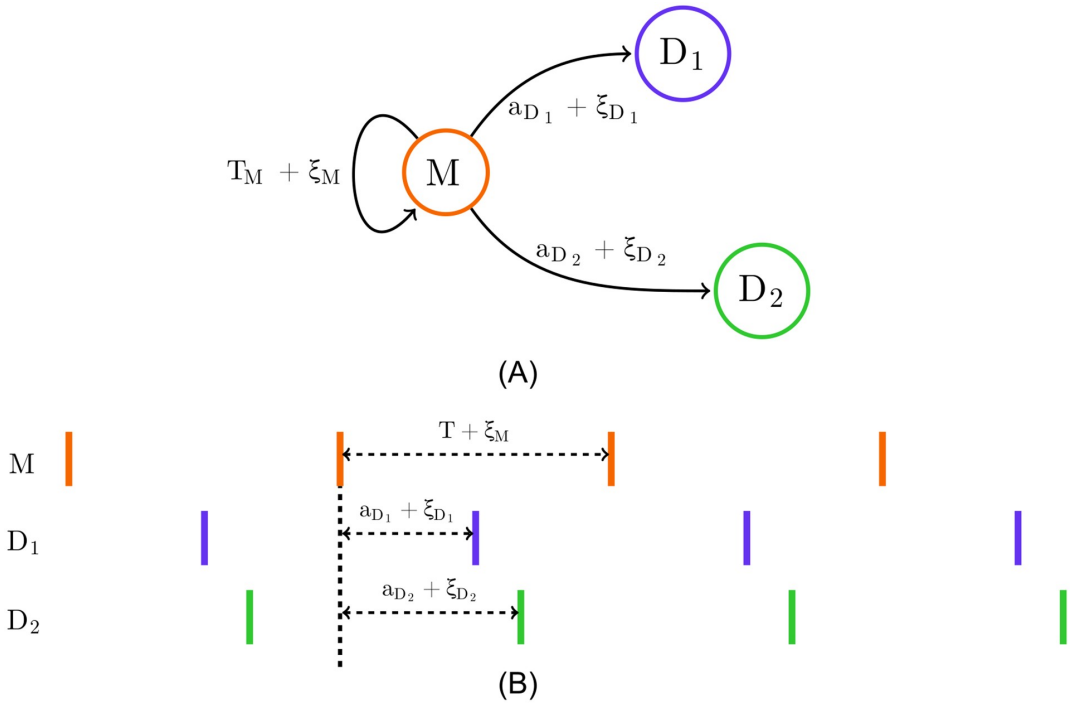


Fig 5. Diagram of the noisy copy process. Events in the mother process M occur periodically with intervals $T + \xi_M$ (ξ_M and ξ_{D_i} are noise terms). Events in the daughter processes D_1 and D_2 occur after each event in the mother process, at a distance of $a_{D_i} + \xi_{D_i}$ (with $a_{D_1} < a_{D_2}$). (A) shows a graph of the dependencies with the labels on the edges representing delays. (B) shows a diagram of a representative spike raster.

https://doi.org/10.1371/journal.pcbi.1008054.g005

populations of which often exhibit highly correlated behaviour through various forms of synchrony [58, 59, 60] or common drivers [61, 62]. In this subsection, we demonstrate that the combination of the presented estimator and surrogate generation scheme is particularly adept at identifying conditional independence in the face of strong pairwise correlations on a synthetic example. Moreover, the combination of the traditional discrete-time estimator and surrogate generation techniques are demonstrated to be ineffective on this task.

The chosen synthetic example in this subsection models a common driver effect, where an apparent directed coupling between a source and target is only due to a common parent. In such cases, despite a strong induced correlation between the source history and the occurrence of an event in the target, we expect to measure zero information flow when conditioning on the common driver. Our system here consists of a quasi-periodic 'mother' process M (the common driver) and two 'daughter' processes, D_1 and D_2 (see Fig 5A for a diagram of the process). The mother process contains events occurring at intervals of $T + \xi_M$, with the daughter processes being noisy copies with each event shifted by an amount $a_{D_i} + \xi_{D_i}$ (ξ_M and ξ_{D_i} are noise terms). We also choose that $a_{D_1} < a_{D_2}$; so long as the difference between these a_{D_i} values is large compared to the size of the noise terms, this will ensure that the events in D_1 precede those in D_2. When conditioning on the mother process, the TE from the first daughter to the second, $\dot{\mathbf{T}}_{D_1 \to D_2 | M}$, should be 0. However, accurately detecting this is difficult, as the history of source daughter process D_1 is strongly correlated with the occurrence of events in the second

daughter process D_2—the events in D_1 will precede those in D_2 by the constant amount $a_{D_2} - a_{D_1}$ plus a small noise term $\xi_{D_2} - \xi_{D_1}$.

Due to the noise in the system, this level of correlation will gradually break down if we translate the source daughter process relative to the others. This allows us to do two things. Firstly, we can get an idea of the bias of the estimator on conditionally independent processes for different levels of pairwise correlation between the history of the source and events in the target. Secondly, we can evaluate different schemes of generating surrogate TE distributions as a function of this correlation. We would expect that, for well-generated surrogates which reflect the relationships to the conditional process, the TE estimates on these conditionally independent processes will closely match the surrogate distribution.

We simulated this process using the parameter values of $T = 1.0$, $a_{D_1} = 0.25$, $a_{D_2} = 0.5$, $\xi_{D_1} \sim \mathcal{N}(0, \sigma_D^2)$ and $\xi_{D_2} \sim \mathcal{N}(0, \sigma_D^2)$ where $\sigma_D = 0.05$. ξ_M was distributed as a left-truncated normal distribution, with mean 0 and standard deviation $\sigma_M = 0.05$, with a left truncation point of $-T + \varepsilon$, where $\varepsilon = 1 \times 10^{-6}$, ensuring that $T + \xi_M > 0$. Once the process had been simulated, the source process D_1 was translated by an amount ω. We used values of ω between -10T and 10T, at intervals of 0.13T. For each such ω, the process was simulated 200 times. For each simulation, the TE was estimated on the original process with the translation ω in the first daughter as well as on a surrogate generated according to our proposed local permutation scheme (see Methods for a detailed description). The parameter values of $k_{\text{perm}} = 10$ and $N_{U,\text{surrogate}} = N_X$ were used. For comparison, we also generated surrogates according to the traditional source time-shift method, where this shift was distributed randomly uniform between 200 and 300 time units. A contiguous region of 50 000 target events was extracted and the estimation was performed on this data. The continuous-time estimator used the parameter values of $l_X = l_Y = l_{Z_1} = 1$, $k = 10$, $N_U = N_X$ and the Manhattan (ℓ_1) norm.

The results in Fig 6A demonstrate that the null distribution of TE values produced by the the local permutation surrogate generation scheme closely matches the distribution of TE values produced by the continuous-time estimator applied to the original data. Whilst the raw TE estimates retain a slight negative bias (explored further in Discussion), we can generate a bias-corrected TE with the surrogate mean subtracted from the original estimate (giving an "effective transfer entropy" [63]). This bias-corrected TE as displayed in Fig 6B is consistent with zero because of the close match between our estimated value and surrogates, which is the desired result in this scenario. On the other hand, the TE values estimated on the surrogates generated by the traditional time-shift method are substantially lower than those estimated on the original process (Fig 6A); comparison to these would produce very high false positive rates for significant directed statistical relationships (see the values of TE bias-corrected to these surrogates, which are not consistent with 0, in Fig 6B). This is most pronounced for high levels of pairwise source-target correlation (with translations ω near zero). The reason behind this difference in the two approaches is easy to intuit. The traditional time-shift method destroys all relationship between the history of the source and the occurrence of events in the target. This means that we are comparing estimates of the TE on the original processes (where there is a strong pairwise correlation between the history of the source and the occurrence of target events) with estimates of the TE on fully independent surrogate processes. Specifically, in discrete time, the joint distribution of the present state of the target and the source history, conditioned on the other histories decomposes as $p(X_t, \mathbf{Y}_{<t} \mid \mathbf{X}_{<t}, \boldsymbol{\mathcal{Z}}_{<t}) = p(X_t \mid \mathbf{X}_{<t}, \boldsymbol{\mathcal{Z}}_{<t}) p(\mathbf{Y}_{<t})$ when using a naive shift method.

By contrast, the proposed local permutation scheme produces surrogates where, although the history of the source and the occurrence of events in the target are conditionally independent, the relationship between the history of the source and the mediating variable,

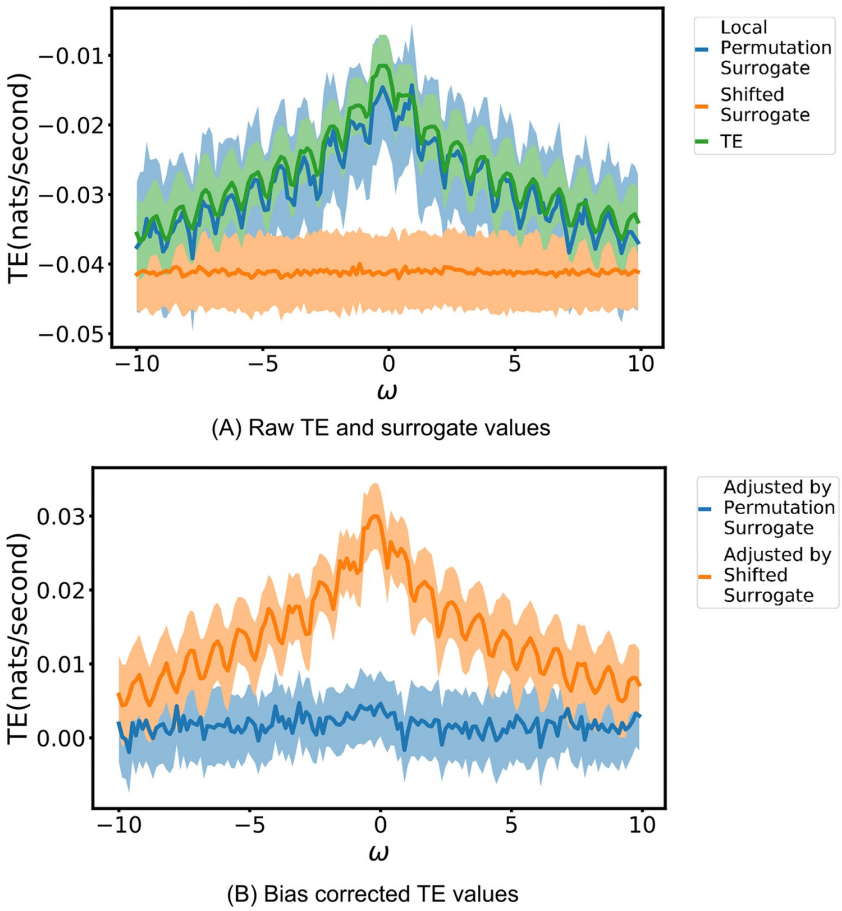


Fig 6. Results of the continuous-time estimator run on a noisy copy process $\dot{T}_{D_1 \to D_2 | M}$, where conditioning on a strong common driver M should lead to zero information flow being inferred. The translation ω of the source, relative to the target and common driver, controls the strength of the correlation between the source and target (maximal at zero translation). For each translation, the estimator is run on both the original process as well as embeddings generated via two surrogate generation methods: our proposed local permutation method and a traditional source time-shift method. The solid lines show the average TE rate across multiple runs and the shaded areas span from one standard deviation below the mean to one standard deviation above it. The bias of the estimator changes with the translation ω, and we expect the estimates to be consistent with appropriately generated surrogates reflecting the same strong common driver effect. This is the case for our local permutation surrogates, as shown in (A). This leads to the correct bias-corrected TE value of 0, as shown in (B).

https://doi.org/10.1371/journal.pcbi.1008054.g006

which in this case is the history of the mother process, is maintained. That is, the scheme produces surrogates where (working in the discrete-time formalism for now) the joint distribution of the present of the target and the source history, conditioned on the other histories decomposes appropriately as $p(X_t, \mathbf{Y}_{<t} \,|\, \mathbf{X}_{<t}, \mathscr{Z}_{<t}) = p(X_t \,|\, \mathbf{X}_{<t}, \mathscr{Z}_{<t}) p(\mathbf{Y}_{<t} \,|\, \mathbf{X}_{<t}, \mathscr{Z}_{<t})$. See Methods for the analogous decomposition within the continuous-time event-based TE framework.

We then confirm that the proposed scheme is able to correctly distinguish between cases where an information flow does or does not exist. To do so, we applied it to measure $\dot{\mathbf{T}}_{M \to D_2 | D_1}$

in the above system, where we would expect to see non-zero information flow from the common driver or mother to one daughter process, conditioned on the other. The setup used was identical to above however focussing on a translation of $\omega = 0$, and for completeness, two different levels of noise in the daughter processes were used: $\sigma_D = 0.05$ and $\sigma_D = 0.075$. The translation of $\omega = 0$ was chosen as, in the cases of zero information flows ($\dot{\mathbf{T}}_{D_1 \to D_2|M}$), the pairwise source-target correlations will be at their highest, increasing the difficulty of correctly identifying these zero flows.

We recorded the p values produced by the combination of the proposed continuous-time estimator and the local permutation surrogate generation scheme when testing for conditional information flow where it is expected to be non-zero through $\dot{\mathbf{T}}_{M \to D_2|D_1}$, in addition to where there is expected to be zero flow through $\dot{\mathbf{T}}_{D_1 \to D_2|M}$. These flows were measured in 10 runs each and the distributions of the resulting p values are shown in Fig 7. We observe that our proposed combination assigns a p value of zero in every instance of $\dot{\mathbf{T}}_{M \to D_2|D_1}$ as expected; whilst for $\dot{\mathbf{T}}_{D_1 \to D_2|M}$ it assigns p values in a broad distribution above zero, meaning the estimates are consistent with the null distribution as expected.

We also applied the combination of the discrete-time estimator and the traditional time-shift method of surrogate generation to this same task of distinguishing between zero and non-zero conditional information flows. We used time bins of width $\Delta t = 0.05$ and history lengths of 7 bins for the target, source and conditional histories. In order to increase the length of history being considered, while keeping the length of the history embeddings constant, application of the discrete-time estimator often makes use of the fact that the present state of the target might be conditionally independent of the most recent source history due to, for instance, transmission delays. In order to exploit this property of the processes, a lag parameter is determined. This lag parameter is a number of time bins to skip between the target present bin and the start of the source history embedding. We followed the current best practice in determining this lag parameter [17]. That is, before calculating the conditional TE from the source to the target, we determined the optimal lag between the conditional history and the target by calculating the pairwise TE between the conditioning process and the target for all lags between 0 and 10. The lag which produced the maximum such TE was used. We then

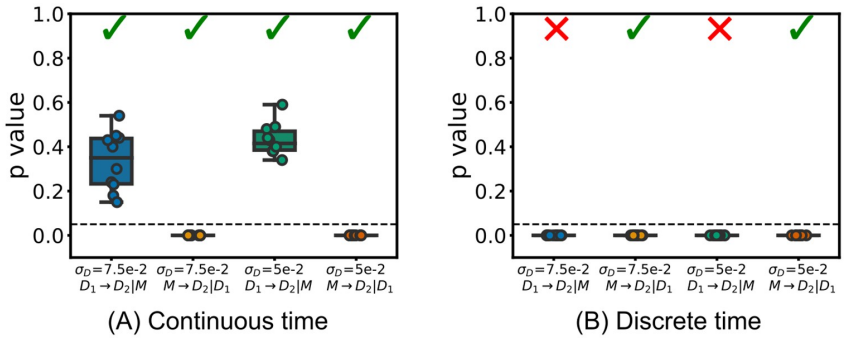


Fig 7. The p-values obtained when using continuous and discrete-time estimators to infer non-zero information flow in the noisy copy process. The estimators are applied to both $\dot{\mathbf{T}}_{D_1 \to D_2|M}$ (expected to have zero flow) and $\dot{\mathbf{T}}_{M \to D_2|D_1}$ (expected to have non-zero flow and therefore be indicated as statistically significant). Only the results from the continuous-time estimator match these expectations. Ticks represent the particular combination of estimator and surrogate generation scheme making the correct inference in the majority of cases when a cutoff value of $p = 0.05$ is used. The dotted line shows $p = 0.05$.

https://doi.org/10.1371/journal.pcbi.1008054.g007

calculated the conditional TE between the source and the target, using this determined lag for the conditioning process, for all lags to the source process between 0 and 10. The TE was then determined to be the maximum TE estimated over all these different lags applied to the source process. This procedure was applied when estimating the TE on the original process as well as on each separate surrogate. The results of this procedure are also displayed in Fig 7. Here we see that the combination of the discrete-time estimator and the traditional time-shift method of surrogate generation assigns a p value indistinguishable from zero to all individual runs of both $\dot{\mathbf{T}}_{M\to D_2|D_1}$ and $\dot{\mathbf{T}}_{D_1\to D_2|M}$. This result–contradicting the expectation that $\dot{\mathbf{T}}_{D_1\to D_2|M}$ is consistent with zero–suggests that this benchmark approach has an incredibly high false positive rate here.

Finally, we investigated whether the poor performance of the traditional combination of the discrete-time estimator and source time-shift surrogate generation scheme was entirely due to the surrogate generation scheme, or at least partially due to time discretisation. To do so, we reran the experiments for the discrete-time estimator shown in Fig 7B, but replaced the time-shift surrogate generation scheme for an approach which is equivalent to our local permutation scheme, but operates on categorical variables (such as binary numbers). This is a pre-existing conditional-permutation-based surrogate generation technique [64]. The results were identical to those shown in Fig 7B for which the time-shift method of surrogate generation (the usual approach for TE analysis) was used. This suggests that time discretisation plays a substantial role in the failure of the traditional approach on this example. That is, good performance here also requires estimation in continuous time.

Scaling of conditional independence testing in higher dimensions

The previous subsection demonstrated the ability of the proposed continuous-time TE estimator and local permutation surrogate generation scheme to perform conditional independence tests despite strong pairwise correlations. The results and analysis there demonstrated how the distribution of the TE values over the surrogates was able to match those over the original time series in cases of zero TE, resulting in a broad distribution of p values between 0 and 1. It was further demonstrated that the distribution of p values obtained from cases with a non-zero TE was clustered around 0, thus providing us with an effective test between zero and non-zero TE. As argued in Introduction and Methods, this is equivalent to a test for conditional independence.

One of the main applications of conditional independence tests is as a component in network inference algorithms [7, 50, 65]. In such cases, the number of processes included in the conditioning set can be as large as one less than the degree of the node. The previous subsection performed a detailed analysis of the distribution of TE values of the original time series, TE values of the surrogate time series as well as the resulting p values in a case where there is a single process in the conditioning set. It was also demonstrated that the inference of non-zero TE could be performed successfully in this case. In this subsection, we study the scaling of the inference of non-zero TE with the size of the conditioning set. As such, we provide a demonstration of the suitability of the combination of the proposed estimator and surrogate generation scheme as a component in a conditional-independence based network inference algorithm.

We generate synthetic data on which to test this scaling. The simulated example consists of a single Leaky-Integrate-and-Fire (LIF) [66] neuron and a set of stimuli to it. See Methods for a full description of this model. The LIF neuron has parameters $V_0 = -65$ mv, $V_{\text{reset}} = -75$ mv, $V_{\text{threshold}} = -45$ mv, a time constant of $\tau = 10$ ms and a hard refractory period of 5 ms.

Each stimulus is a separately generated inhomogeneous Poisson process, with an added refractory period of 5 ms. All the stimuli have a common rate. This rate is constant across windows of 0.5*s* and is generated uniformly randomly between 0 Hz and 40 Hz. As in the above example of unidirectionally coupled process pairs, the stimuli are generated using a thinning algorithm. The process is first generated as a homogeneous Poisson process with rate $R > 40$ Hz. Spikes are excluded with probability $1 - r_i/R$, where r_i is the common rate of the window of the spike. All spikes within the refractory period of the previous spike are also excluded. The stimuli are divided into a set of background processes B, with $|B| \in \{6, 12, 18\}$, and a source Y. One third of the stimuli in the background set are inhibitory and remainder are excitatory. The strength of the connection V_{connect} associated with each stimulus was adjusted by hand such that the average firing rate of the target LIF neuron was around 20 Hz when only the stimuli in the background set were connected to the target (that is, the extra source stimulus was unconnected). The resulting connection strengths used are 18 mV, 13 mV and 10 mV for each of the three sizes of the background set, respectively. All connections have a fixed delay of 2 ms. The source stimulus Y is set to be either inhibitory, excitatory or is otherwise unconnected to the target LIF neuron.

In the case where the source neuron is unconnected, when conditioning on all the processes in the background set, the TE between the source and the target LIF neuron is zero. In the cases where it is connected in either an inhibitory or excitatory manner, the TE will be non-zero. This follows from the assumptions made explicit in Methods relating conditional independence and dependence to network structure. We tested the ability of both estimator and surrogate generation scheme combinations to correctly infer zero or non-zero TE.

For the continuous-time estimator and local permutation surrogate generation scheme we used the parameter values of $l_X = l_Y = l_{Z_i} = 1$, $k = 5$, $k_{\text{perm}} = 10$, $N_U/N_X = 1$ and $N_{U,\text{surrogate}}/N_X = 10$. The discrete-time estimator used the same history embedding length for the source, target and conditioning processes. This was set at 3, 2 or 1 bins for each of the conditioning set sizes (6, 12 or 18), respectively. These embedding lengths were chosen so as to keep the total number of bins used across the target, source and conditioning processes below 25. Using more than 25 bins resulted in the space of possible history permutations growing too large, leading to undersampling and far inferior performance. The bin width Δt was set at 8 ms, 11 ms and 22 ms for each of these three embedding lengths. These bin widths were chosen so that the history would extend back a distance of at least twice the time constant of the LIF target neuron, plus the transmission delay from the stimuli.

For both combinations, 100 surrogates and a threshold of $p = 0.05$ for the inference of non-zero TE were used. Tests were conducted for the number of target spikes $N_X \in \{100, 500, 1000, 2000, 5000, 10000\}$. For each data set size, both approaches were tested on 30 independent simulations for each setting of Y as either inhibitory, excitatory or unconnected.

Fig 8 shows the results of running the two approaches on the simulated data for different data set sizes. The combination of the discrete-time estimator and the traditional time-shift surrogate generation scheme is found to be inadequate. For data set sizes of $N_X \geq 1000$, we see that this approach assigns non-zero TE to all 30 runs of every connection class (excitatory, inhibitory or absent) at each size, despite the fact that the 30 runs on absent connections correspond to cases of zero TE. Moreover, in the case of absent connections, the direction of convergence is in the wrong direction—this approach performs worse as we provide more data. This is likely due to this scheme's poor ability to identify conditional independence in the presence of pairwise correlations, as we have already seen in the previous subsection. In the instances where the source is not connected to the target, its spiking activity will still be correlated with that of the target, due to it sharing a common rate with the background processes.

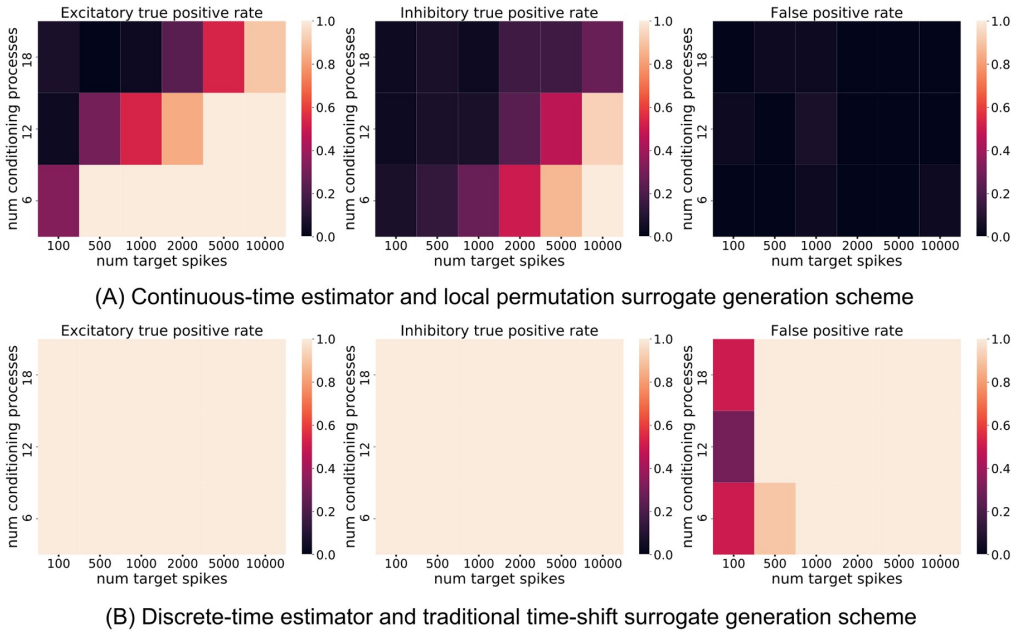


Fig 8. The scaling of the combination of the continuous-time estimator and the local permutation surrogate generation scheme on correctly identifying conditional independence relationships with increasing dimension and data size (A). This is compared with the performance of the combination of the discrete-time estimator with the traditional time-shift surrogate generation procedure (B). The *y* axis represents the number of background processes being conditioned on. Above a certain moderate threshold of data size, the discrete-time approach infers a non-zero TE in all of the runs, including those where the source was in fact not connected to the target. This renders it impractical for this task.

https://doi.org/10.1371/journal.pcbi.1008054.g008

S4 Fig displays the same results as Fig 8, but in the simpler case where all stimuli have a constant rate of 20 Hz. In this case, with the pairwise correlations removed, we see that the discrete-time estimator is capable of more consistently correctly identifying cases of zero TE, although it still displays a substantially inflated false positive rate compared to the expected value of 0.05. Moreover, it is worth emphasising that this is an unrealistic scenario as it is assuming completely independent sources, whereas the activity of biological neurons are known to exhibit a wide variety of correlations in their activities [58, 59, 60].

Returning to Fig 8A, in the cases where the conditioning set contains 6 or 12 processes, the combination of the continuous-time estimator and local permutation surrogate generation scheme is able to correctly identify zero versus non-zero TE provided that it has access to around 10000 target spikes. In the case where the conditioning set contains 18 processes, it is capable of correctly identifying non-zero TE for excitatory connections as well as correctly identifying zero TE in the case of an unconnected source. In all combinations of numbers of spikes and number of conditionals, our method is able to control the false positive rate at the prescribed level. This is crucial: in the context of network inference applied to neuroscientific data, false positives are considered more detrimental than false negatives [67]. This is due to such false positives often existing between communities and thus resulting in substantial errors in the inferred topology. With that said, the true positive rate is below 50% for inhibitory sources, though it is observed to rise with an increase in the number of target spikes being considered. Importantly, were a greedy approach to effective network inference to be used, as in

[7, 50], whereby edges are iteratively added to the conditioning set based on their TE value, then the majority of conditional independence tests will be performed at a dimension well below the degree of the node. In order to measure the performance of our proposed approach at the start of this process (where no sources have yet been selected and conditioned on), S5 Fig displays the same results as Fig 8 but where the background processes are not included in the conditioning set. Here we see higher true positive rates at lower numbers of spikes, with the inhibitory connections being easily identified. This implies that, when used as a component in such a greedy algorithm, our approach will be able to identify the principal sources whilst controlling the false-positive rate, although it may miss some true sources in higher dimensions.

Finally, we investigated whether the poor performance of the traditional combination of the discrete-time estimator and source time-shift surrogate generation scheme was entirely due to the surrogate generation scheme. That is, could it be rescued by using a better surrogate generation technique? We therefore repeated the discrete-time experiments shown in Fig 8, S4 and S5 Figs, but replaced the time-shift surrogate generation scheme for an approach which is equivalent to our local permutation scheme, but operates on categorical variables (such as binary numbers). This is an established conditional-permutation based surrogate generation scheme [64]. The results of these runs are displayed in S6 Fig. We observe qualitatively similar results for the use of these two surrogate generation techniques. The only substantial difference is that the conditional-permutation based scheme has lower true positive rates for inhibitory connections when less data is available under all setups. This implies that the poor performance of the traditional approach is largely due to time-discretisation. Once again, we see that good performance here requires estimation in continuous time.

Testing for conditional independence on the simulated pyloric circuit of the crustacean stomatogastric ganglion

The pyloric circuit of the crustacean stomatogastric ganglion has received significant attention in terms of statistical modelling and has been proposed as a benchmark circuit on which to test spike-based connectivity inference techniques [68, 69]. Such modelling attempts have faced substantial difficulties. For instance, it has been shown that Granger causality is unable to infer the connectivity of this network [68] (Granger causality and TE are equivalent for linear dynamics with Gaussian noise [70]). We demonstrate here that our proposed approach is able to correctly infer the conditional dependence and independence relationships in this circuit (which, as per the previous examples, are expected to match connectivity under the conditions of this experiment, see Methods).

The crustacean stomatogastric ganglion [49, 71, 72] has received substantial research attention as a simple model circuit. The fact that its full connectivity is known is of great use for modelling and statistical analysis. The pyloric circuit is a partially independent component of the greater circuit and consists of an Anterior Burster (AB) neuron, two Pyloric Driver (PD) neurons, a Lateral Pyloric (LP) neuron and multiple Pyloric (PY) neurons. As the AB neuron is electrically coupled to the PD neurons and the PY neurons are identical, for the purposes of modelling, the circuit is usually represented by a single AB/PD complex, and single LP and PY neurons [68, 69, 73, 74].

The AB/PD complex undergoes self-sustained rhythmic bursting. It inhibits the LP and PY neurons through slow cholinergic and fast glutamatergic synapses. These neurons then burst on rebound from this inhibition. The PY and LP neurons also inhibit one another through fast glutamatergic synapses and the LP neuron similarly inhibits the AB/PD complex.

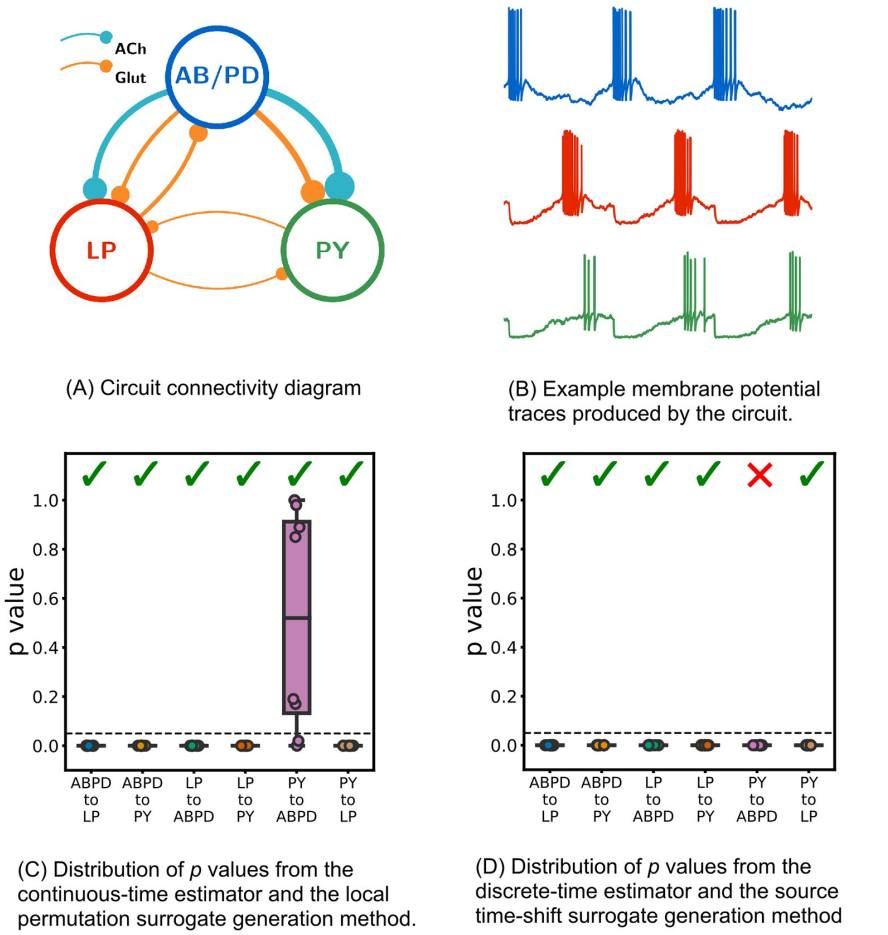


Fig 9. Results of both estimator and surrogate generation combinations being applied to data from simulations of a biophysical model of a neural circuit inspired by the pyloric circuit of the crustacean stomatogastric ganglion. The circuit, shown in (A), is fully connected apart from the missing connection between the PY neuron and the AB/PD complex, and generates membrane potential traces which are bursty and highly-periodic with cross-correlated activity. The distribution of p values from the combination of the continuous-time estimator and local permutation surrogate generation scheme are shown in (C). They demonstrate that this combination is capable of correctly identifying the conditional dependence and independence relationships in this circuit in all runs, apart from two false negatives. By contrast, the distribution of p values produced by the combination of the discrete-time estimator and the traditional source time-shift surrogate generation method shown in (D) mis-specified the relationship from the PY to the ABPD in every run. Ticks represent the particular combination of estimator and surrogate generation scheme making the correct inference of dependence or independence in the majority of cases when a cutoff value of $p = 0.05$ is used.

https://doi.org/10.1371/journal.pcbi.1008054.g009

Fig 9 shows sample membrane potential traces from simulations of this circuit as well as a connectivity diagram. Despite its small size, inference of the relationships between neurons is challenging [68, 69] due to the fact that it is highly periodic. Although there is no structural connection from the PY to the ABPD neuron (implying conditional independence due to full observability and the causal Markov assumption), there is a strong, time-directed, correlation between their activity—the PY neuron always bursts shortly before the ABPD. Recognising

that this is a spurious correlation, and that the AB/PD complex is thus conditionally independent of the history of the PY neuron, requires fully resolving the influence of the AB/PD's history on itself as well as that of the LP on the AB/PD. To further complicate matters, the dependence implied by the connection between the LP and ABPD neurons (along with the contraposition of our assumption of faithfulness) is very challenging to detect. The AB/PD complex will continue bursting regardless of any input from the LP. Correctly inferring this dependence requires detecting the subtle changes in the timing of AB/PD bursts that result from the activity of the LP.

Previous work on statistical modelling of the pyloric circuit has used both *in vitro* and *in silico* data [68, 69]. We ran simulations of biophysical models inspired by this network, similar to those used in [68] (see S1 Text). Attempts were then made to identify the conditional dependence/independence relationships in the network by detecting non-zero conditional information flow from the spiking event times produced by the simulations. This was done by estimating the TE from the source to the target, conditioned on the activity of the third remaining neuron for every source-target pair. Both the combination of the proposed continuous-time estimator and local permutation surrogate generation scheme and the combination of the discrete-time estimator and source time-shift surrogate generation scheme were applied to this task. As the dynamics of the network are fully captured in the three neurons of the network (we have full observability), and due to the causal Markov assumption, in the case where there is no causal directed connection from a source to a target, the target's present will be conditionally independent of the source's past. By the contraposition of the faithfulness assumption, in the presence of a connection the target's current state will be dependent on the source's past (see Methods).

Both combinations were applied to nine independent simulations (ten simulations were instantiated but one was discarded due to early termination from a numerical instability) of the network and the number of target events $N_X = 2 \times 10^4$ was used. For the continuous-time estimator the parameter values of $l_X = l_Y = l_{Z_1} = 3$, $k = 10$, $N_U = N_X$, $N_{U,\text{surrogate}} = 5N_X$ and $k_{\text{perm}} = 10$ were used along with the Manhattan (ℓ_1) norm (see Methods). The discrete-time estimator made use of a bin size of $\Delta t = 0.05s$ and history embedding lengths of seven bins for each of the source, target and conditioning processes. Searches were performed to determine the optimum embedding lag for both the source and conditioning histories (as above) with a maximum search value of 20 bins being used. We designed the search procedure to include times up to the inter-burst interval (around 1 time unit), which placed an effective lower bound on the width of the time bins (as bin sizes below $\Delta t = 0.05s$ resulted in impractically large search spaces). For both estimators, p values were inferred from 100 independently generated surrogates (see Methods). The source time-shift surrogate generation scheme used time shifts distributed uniformly randomly between 200 and 400 time units.

Fig 9C and 9D show the distributions of p values resulting from the application of both estimator and surrogate generation scheme combinations. The continuous-time estimator and local permutation surrogate generation scheme were able to correctly infer the dependence/independence relationships in the network in the majority of cases (indicated by p-values approaching 0 for the true positives, and spread throughout [0, 1] for the true negatives). On the other hand, the discrete-time estimator and source time-shift surrogate generation scheme produced an erroneous inference on every run: a dependence between the PY neuron and the AB/PD complex. S7 and S8 Figs contain plots showing runs of the continuous-time estimator using different values of the parameters l_X, l_Y, l_{Z_1} and N_X. The results are qualitatively very

similar to those presented in Fig 9C, showing that, on this example, our methodology is robust to these parameter choices.

As in the previous subsections, we investigated whether the poor performance of the traditional combination of the discrete-time estimator and source time-shift surrogate generation scheme was entirely due to the surrogate generation scheme, or at least partially due to time discretisation. To do so, we reran the experiments for the discrete-time estimator shown in Fig 9D, but replaced the time-shift surrogate generation scheme for an approach which is equivalent to our local permutation scheme, but operates on categorical variables (such as binary numbers). As previously, this is a pre-existing conditional-permutation-based surrogate generation method [64]. The results were identical to those shown in Fig 9D for which the time-shift method of surrogate generation (the usual approach for TE analysis) was used. This suggests that time discretisation plays a substantial role in the failure of the traditional approach on this example. Mirroring our previous findings, we observe that good performance here requires estimation in continuous time.

On this particular example, the inference of all connections using the continuous-time approach took 13 minutes and 6 seconds when using 20 cores of an Intel Xeon E5-2670. The discrete-time approach took around 37 minutes and 4 seconds when running on the same hardware. We would, however, point out that the computational requirements for both methods are highly sensitive to their parameters. The discrete-time approach will be particularly sensitive to Δt and the number of lag settings searched over. The continuous-time approach is particularly sensitive to the embedding lengths.

Discussion

Despite transfer entropy being a popular tool within neuroscience and other domains of enquiry [7, 8, 9, 13, 14, 15, 16, 17, 18, 19], it has received more limited application to event-based data such as spike trains. This is at least partially due to current estimation techniques requiring the process to be recast as a discrete-time phenomenon. The resulting discrete-time estimation task has been beset by difficulties including a lack of consistency, high bias, slow convergence and an inability to capture effects which occur over fine and large time scales simultaneously.

This paper has built on recent work presenting a continuous-time formalism for TE [5] in order to derive an estimation framework for TE on event-based data in continuous time. This framework has the unique advantage of only estimating quantities at events in the target process alongside efficient representation of the data as inter-spike intervals, providing a significant computational advantage. Instead of comparing spike rates conditioned on specific histories at each target spiking event, we use a Bayesian inversion to instead make the empirically easier comparison of probabilities of histories at target events versus anywhere else along the target process. This comparison, using KL divergences, is made using k-NN techniques, which brings desirable properties such as efficiency for the estimator. This estimator is provably consistent. Moreover, as it operates on inter-event intervals, it is capable of capturing relationships which occur with fine time precision along with those that occur over longer time distances.

The estimator was first evaluated on two simple examples for which the ground truth is known: pairs of independent Poisson processes (first subsection of Results) as well as pairs of processes unidirectionally coupled through a simple functional relationship (second subsection of Results). The current state-of-the-art in discrete-time estimation was also applied to these processes. It was found that the continuous-time estimator had substantially lower bias than the discrete-time estimator, converged orders of magnitude faster (in terms of the

number of sample spikes required), and was relatively insensitive to parameter selections. Moreover, these examples provided numerical confirmation of the consistency of the continuous-time estimator, and further demonstration that the discrete-time estimator is not consistent. The latter simple example highlighted the magnitude of the shortcomings of the discrete-time estimator. In the authors' experience, spike-train datasets which contain 1 million spiking events for a single neuron are vanishingly rare. However, even in the unlikely circumstance that the discrete-time estimator is presented with a dataset of this size, as in the second subsection of Results, it could not accurately estimate the TE for a simple one-way relationship between only two neurons. Moreover, this example neatly demonstrates a known [31], notable problem with the use of the discrete-time estimator, which is that it provides wildly different estimates for different values of Δt. Whilst the underlying theory [5] suggests that in principle taking the discrete time TE rate as $\Delta t \to 0$ converges with the continuous time formalism, the use of smaller Δt values leads to issues in undersampling and inability to represent patterns on long time scales. In real-world applications, where the ground truth is unknown, there is no principled method for choosing which resulting TE value from the various bin sizes to use.

One of the principal use-cases of TE is the inference of non-zero information flow. As the TE is estimated from finite data, we require a manner of determining the statistical significance of the estimated values. Traditional methods of surrogate generation for TE either shift the source in time, or shuffle the source embeddings. However, whilst this retains the relationship of the target to its past and other conditionals, it completely destroys the relationship between the source and any conditioning processes, which can lead to very high false positive rates as detailed in the third subsection of Results and Methods. We developed a local permutation scheme, based on [48], for use in conjunction with this estimator which is able to maintain the relationship of the source history embeddings with the history embeddings of the target and conditioning processes. The combination of the proposed estimator and this surrogate generation scheme were applied to an example where the history of the source and the occurrence of events in the target are highly correlated, but conditionally independent given their common driver (third subsection of Results). The established time-shift method for surrogate generation produced a null distribution of TE values substantially below that estimated on the original data, incorrectly implying non-zero information flow. Conversely, the proposed local permutation method produced a null distribution which closely tracked the estimates on the original data. The proposed combination was also shown to be able to correctly distinguish between cases of zero and non-zero information flow. When applied to the same example, the combination of the discrete-time estimator and the traditional method of time-shifted surrogates inferred the existence of information flow in all cases, even when no such flow was present. The scaling of these results with the size of the conditioning set was investigated in the fourth subsection of Results. Here, in a highly simplified model of the input-output relationships of a neuron, it was demonstrated that the proposed method could correctly identify conditional dependence vs. independence in cases of up to 12 conditioning processes with access to 10^4 target spikes. Moreover, it maintained robustness to pairwise correlations despite conditional independence. Again, the traditional combination of discrete-time estimator and time shifted surrogates was found to be lacking.

Finally, our proposed approach was applied to inferring the dependence/independence relationships in a more biologically faithful example in the fifth subsection of Results. For this purpose, we made use of models inspired by the pyloric circuit of the crustacean stomatogastric ganglion. The full observability and large noise provided by this model allowed us to conclude that the conditional dependence/independence relationships would match the underlying connectivity of the model, thus providing us with a ground truth against which to

test our approach. Statistical modelling of this network is challenging due to its highly periodic dynamics. For instance, attempts to use Granger causality, using a more established estimator, to infer its connectivity have been unsuccessful [68]; furthermore, we showed that the discrete-time binary-valued TE estimator (with time-shifted surrogates) also could not successfully infer the independence and dependence relationships in the network. It is worth highlighting in this context that Granger causality and TE are equivalent for linear dynamics with Gaussian noise [70]. Given that discrete-time TE (capable of capturing nonlinear relationships) failed on this network, we suspect that the reason for the earlier failures of Granger causality applied to this network were due, at least in part, to time binning and not entirely due to its inability to find nonlinear relationships. Despite these challenges, our combination of continuous-time estimator and surrogate generation scheme was able to correctly infer the relationships implied by the pyloric network. This provides an important validation of the efficacy of our presented approach on a challenging example of representative biological spiking data.

This work represents a substantial step forward in the estimation of information flows from event-based data. To the best of the authors' knowledge it is the first consistent estimator of TE for event-based data. That is, it is the first estimator which is known to converge to the true value of the TE in the limit of infinite data, let alone to provide efficient estimates with finite data. As demonstrated in the first and second subsections of Results it has substantially favourable bias and convergence properties as compared to the discrete-time estimator. The fact that this estimator uses raw inter-event intervals as its history representation allows it to efficiently capture relevant information from the past of the source, target and conditional processes. This allows it to simultaneously measure relationships that occur both with very fine time scales as well as those that occur over long intervals. This was highlighted in the fifth subsection of Results, where it was shown that our proposed approach is able to correctly infer the conditional dependence/independence relationships implied by a model inspired by the pyloric circuit of the crustacean stomatogastric ganglion. The inference of these relationships requires capturing subtle changes in spike timing. However, its bursty nature means that there are long intervals of no spiking activity. This is contrasted with the poor performance of the discrete-time estimator on this same task, as above. The use of the discrete-time estimator requires a hard trade-off in the choice of bin size: small bins will be able to capture relationships that occur over finer timescales but will result in an estimator that is blind to history effects existing over large intervals. Conversely, whilst larger bins might be capable of capturing these relationships occurring over larger intervals, the estimator will be blind to effects occurring with fine temporal precision.

Further, real-world data is of course sampled at some limited resolution; this means that any estimator cannot detect TE in the underlying process associated with smaller time scales than available in the data, though the consistency property of our continuous-time estimator means that it will converge to the TE value of the process at the available resolution. Of course, as per our Introduction, where temporal resolution in recordings is very poor (such as in calcium imaging experiments) the aforementioned trade-offs for the discrete-time estimator are likely to be less problematic and the advantages of the continuous-time estimator less pronounced.

To the best of our knowledge, this work showcases the first use of a surrogate generation scheme for statistical significance estimates which correctly handles strong source-conditional relationships for event-based data. This has crucial practical benefit in that it greatly reduces the occurrence of false positives in cases where the history of a source is strongly correlated with the present of the target, but conditionally independent.

We make note of the fact that inspection of some plots, notably Fig 6 shows that, in some cases, the estimator can exhibit small though not insignificant bias. Indeed, similar biases can readily be demonstrated with the standard KSG estimator for transfer entropy on continuous

variables in discrete time, in similar circumstances where a strong source-target relationship is fully explained by a conditional process. The reason for the small remaining bias is that while the underlying assumption of the nearest neighbour estimators is of a uniform probability density within the range of the k nearest neighbours, strong conditional relationships tend to result in correlations remaining between the variables within this range. For the common use-case of inferring non-zero information flows this small remaining bias will not be an issue as the proposed method for surrogate generation is capable of producing null distributions with very similar bias properties. Furthermore, such bias can be removed from an estimate by subtracting the mean of the surrogate distribution (as shown via the effective transfer entropy [63] in the third subsection of Results). However, it is foreseeable that certain scenarios might benefit from an estimator with lower bias, without having to resort to generating surrogates. In such cases it will likely prove beneficial to explore the combination of various existing bias reduction techniques for k-NN estimators with the approach proposed here. These include performing a whitening transformation on the data [75], transforming each marginal distribution to uniform or exploring alternative approaches to sharing radii across entropy terms (see Methods). The authors believe that the most probable cause of the observed bias in the case of strong pairwise correlations is that these correlations cause the assumption of local uniformity (see Methods) to be violated. Gao, Ver Steeg and Galstyan [76] have proposed a method for reducing the bias of k-NN information theoretic estimators which specifically addresses cases where local uniformity does not apply. The application of this technique to our estimator holds promise for addressing this remaining bias.

We foresee that one of the more useful applications of the conditional independence test that the combination of estimator and surrogate generation scheme provides will be network inference. Strictly speaking, statistical methods such as these produce effective network models which are not generally expected to provide precise matches to underlying structural connectivity. Under certain idealised circumstances though, as implemented in our experiments (see Methods), the two can be expected to match, and this provided for the important validation that our methods detect directed conditional independence where it exists in these small networks. The extent to which our method can be validated in this manner on larger more latent-confounded networks, and more importantly the extent to which the network models it infers correlate with underlying structure outside of such idealised conditions including faithfulness (see Methods), remain open questions. This is an intended focus of future work. Indeed, the inference of the connectivity of spiking neural networks from their activity is an active area of research [77, 78] which includes recently proposed continuous-time approaches [79, 80]. However, any conditional independence test will suffer from the curse of dimensionality. This means that performing effective network inference requires pairing the conditional independence test with a suitable (conditional-independence-based) network inference algorithm which reduces the dimensionality of the tests. Fortunately, a variety of such algorithms exist [65] (see Runge [81] for a methodology for reducing the dimensionality outside of network inference). In particular, the greedy algorithm [7, 50], which has already been validated for use in combination with TE (for different types of dynamics on larger networks), holds particular promise. Further, it was recently shown by Das and Fiete [51] that popular existing approaches to the inference of spiking neural networks, such as generalised linear models and maximum entropy-based reverse Ising inference, had very high false-positive rates in instances where the activity of unconnected neurons was highly correlated. Given our focus on demonstrating that our conditional independence test is highly robust to strong pairwise correlations despite conditional independence, we believe that the work presented in this paper holds great promise towards making progress on this important issue.

Finally, it is worth pointing out that, as well as presenting a specific estimator and surrogate generation algorithm, this paper is also presenting an approach to testing for time-directed statistical dependence in spike trains much more generally. Any estimator of KL divergence can be plugged into our framework by being applied to estimate the two KL divergence terms appearing in Eq (10). Moreover, a different surrogate generation scheme could be used, so long as it factorises the distribution of histories as specified in Eq (20) (see Methods). There has been substantial recent progress towards the efficient estimation of divergences [82, 83] in high dimension, pointing to the future promise of this work being applied in the context of network inference.

Methods

There are a variety of approaches available for estimating information theoretic quantities from continuous-valued data [84]; here we focus on methods for generating estimates $\hat{\dot{\mathbf{T}}}_{Y\to X|\mathcal{Z}}$ of a true underlying (conditional) transfer entropy $\dot{\mathbf{T}}_{Y\to X|\mathcal{Z}}$.

The nature of estimation means that our estimates $\hat{\dot{\mathbf{T}}}_{Y\to X|\mathcal{Z}}$ may have a *bias* with respect to the true value $\dot{\mathbf{T}}_{Y\to X|\mathcal{Z}}$, and a *variance*, as a function of some metric n of the size of the data being provided to the estimator (we use the number of spikes, or events, in the target process). The bias is a measure of the degree to which the estimator systematically deviates from the true value of the quantity being estimated, for finite data size. It is expressed as $\text{bias}(\hat{\dot{\mathbf{T}}}_{Y\to X|\mathcal{Z}}) = \mathbb{E}[\hat{\dot{\mathbf{T}}}_{Y\to X|\mathcal{Z}}] - \dot{\mathbf{T}}_{Y\to X|\mathcal{Z}}$. The variance of an estimator is a measure of the degree to which it provides different estimates for distinct, finite, samples from the same process. It is expressed as $\text{variance}(\hat{\dot{\mathbf{T}}}_{Y\to X|\mathcal{Z}}) = \mathbb{E}[\hat{\dot{\mathbf{T}}}^2_{Y\to X|\mathcal{Z}}] - \mathbb{E}[\hat{\dot{\mathbf{T}}}_{Y\to X|\mathcal{Z}}]^2$. Another important property is *consistency*, which refers to whether, in the limit of infinite data points, the estimator converges to the true value. That is, an estimator is consistent if and only if $\lim_{n\to\infty} \hat{\dot{\mathbf{T}}}_{Y\to X|\mathcal{Z}} = \dot{\mathbf{T}}_{Y\to X|\mathcal{Z}}$.

The first half of this methods section is concerned with the derivation of a consistent estimator of TE which operates in continuous time. In order to be able to test for non-zero information flow given finite data, we require a surrogate generation scheme to use in conjunction with the estimator. Such a surrogate generation scheme should produce surrogate history samples that conform to the null hypothesis of zero information flow. The second half of this section will focus on a scheme for generating these surrogates.

The presented estimator and surrogate generation scheme have been implemented in a software package which is freely available online (see the Implementation subsection).

Continuous-time estimator for transfer entropy between spike trains

In the following subsections, we describe the algorithm for our estimator $\hat{\dot{\mathbf{T}}}_{Y\to X|\mathcal{Z}}$ for the transfer entropy between spike trains. We first outline our choice of a *k*NN type estimator, due to the desirable consistency and bias properties of this class of estimator. In order to use such an estimator type, we then describe a Bayesian inversion we apply to the definition of transfer entropy for spiking processes, which allows us to operate on probability densities of histories of the processes, rather than directly on spike rates. This results in a sum of differential entropies to which *k*NN estimator techniques can be applied. The evaluation of these entropy terms using *k*NN estimators requires a method for sampling history embeddings, which is presented before attention is turned to a technique for combining the separate *k*NN estimators in a manner that will reduce the bias of the final estimate.

Consideration of estimator type. Although there has been much recent progress on parametric information-theoretic estimators [85], such estimators will always inject modelling assumptions into the estimation process. Even in the case that large, general, parametric models are used—as in [82]—there are no known methods of determining whether such a model is capturing all dependencies present within the data.

In comparison, nonparametric estimators make less explicit model assumptions regarding the probability distributions. Early approaches included the use of kernels for the estimation of the probability densities [86], however this has the disadvantage of operating at a fixed kernel 'resolution'. An improvement was achieved by the successful, widely applied, class of nonparametric estimators making use of k-nearest-neighbour statistics [53, 87, 88, 89], which dynamically adjust their resolution given the local density of points. Crucially, there are consistency proofs [88, 90] for kNN estimators, meaning that these methods are known to converge to the true values in the limit of infinite data size. These estimators operate by decomposing the information quantity of interest into a sum of differential entropy terms H^*. Each entropy term is subsequently estimated by estimating the probability densities $p(x_i)$ at all the points in the sample by finding the distances to the kth nearest neighbours of the points x_i. The average of the logarithms of these densities is found and is adjusted by bias correction terms. In some instances, most notably the Kraskov-Stögbauer-Grassberger (KSG) estimator for mutual information [53], many of the terms in each entropy estimate cancel and so each entropy is only implicitly estimated.

Such bias and consistency properties are highly desirable–given the efficacy of kNN estimators, it would be advantageous to be able to make use of such techniques in order to estimate the transfer entropy of point processes in continuous time. However the continuous time formulations in Eqs (3) and (4) contain no entropy terms, being written in terms of *rates* as opposed to probability densities. Moreover, the estimators for each differential entropy term H^* in a standard kNN approach operate on sets of points in $\mathbb{R}^d$, and it is unclear how to sample points so as to get an unbiased estimate of the rate.

The following subsection is concerned with deriving an expression for continuous-time transfer entropy on spike trains as a sum of H^* terms, in order to define a kNN type estimator.

Formulating continuous-time TE as a sum of differential entropies. Consider two point processes X and Y represented by sets of real numbers, where each element represents the time of an event. That is, $X \in \mathbb{R}^{N_X}$ and $Y \in \mathbb{R}^{N_Y}$. Further, consider the set of extra conditioning point processes $\mathscr{Z} = \{Z_1, Z_2, \ldots, Z_{n_{\mathscr{Z}}}\}$, $Z_i \in \mathbb{R}^{N_{Z_i}}$. We can define a *counting process* $\mathbf{N}_X(t)$ on X. $\mathbf{N}_X(t)$ is a natural number representing the 'state' of the process. This state is incremented by one at the occurrence of an event. The instantaneous firing rate of the target is then $\lambda_X(t) = \lim_{\Delta t \to 0} p(\mathbf{N}_X(T + \Delta t) - \mathbf{N}_X(t) = 1)/\Delta t$. Using this expression, Eq (4) can then be rewritten as

$$\dot{\mathbf{T}}_{Y \to X|\mathscr{Z}} = \bar{\lambda}_X \lim_{\Delta t \to 0} \mathbb{E}_{P_X} \left[\ln \frac{p_U(\mathbf{N}_X(x + \Delta t) - \mathbf{N}_X(x) = 1 \mid \mathbf{x}_{<x}, \mathbf{y}_{<x}, \boldsymbol{\mathscr{z}}_{<x})}{p_U(\mathbf{N}_X(x + \Delta t) - \mathbf{N}_X(x) = 1 \mid \mathbf{x}_{<x}, \boldsymbol{\mathscr{z}}_{<x})} \right]. \tag{6}$$

Here, $\bar{\lambda}_X$ is the average, unconditional, firing rate of the target process, that is $\bar{\lambda}_X = \lim_{N_X, \tau \to \infty} N_X/\tau$. In practice this is estimated through a trivial bias free estimate e.g. $\hat{\bar{\lambda}}_X = (N_X - 1)/\tau$ with $\tau = x_{N_X} - x_1$. $\mathbf{x}_{<x} \in \mathbf{X}_{<X}$, $\mathbf{y}_{<x} \in \mathbf{Y}_{<X}$ and $\boldsymbol{\mathscr{z}}_{<x} = \{\mathbf{z}_{1,<x}, \mathbf{z}_{2,<x}, \ldots, \mathbf{z}_{n_{\mathscr{Z}},<x}\} \in \mathscr{Z}_{<X}$ are the histories of the target, source and conditioning processes, respectively, at time x. The probability density p_U is taken to represent the probability density at any arbitrary point in the target process, unconditional of events in any of the

processes. Conversely, p_X is taken to represent the probability density of observing a quantity at target events. The expectation $\mathbb{E}_{P_X}$ is taken over this distribution. That is $\mathbb{E}_{P_X}[f(Y)] = \int_Y f(y)p_X(y)dy$.

By applying Bayes' rule we can make a Bayesian inversion to arrive at:

$$\dot{\mathbf{T}}_{Y\to X|\mathscr{Z}} = \bar{\lambda}_X \lim_{\Delta t\to 0} \mathbb{E}_{P_X}\left[\ln \frac{p_U(\mathbf{x}_{<x}, \mathbf{y}_{<x}, \boldsymbol{\varkappa}_{<x} \mid \mathbf{N}_X(x+\Delta t) - \mathbf{N}_X(x) = 1)}{p_U(\mathbf{x}_{<x}, \boldsymbol{\varkappa}_{<x} \mid \mathbf{N}_X(x+\Delta t) - \mathbf{N}_X(x) = 1)} \times \frac{p_U(\mathbf{x}_{<x}, \boldsymbol{\varkappa}_{<x})}{p_U(\mathbf{x}_{<x}, \mathbf{y}_{<x}, \boldsymbol{\varkappa}_{<x})}\right]. \tag{7}$$

Eq (7) can be written as

$$\dot{\mathbf{T}}_{Y\to X|\mathscr{Z}} = \bar{\lambda}_X \mathbb{E}_{P_X}\left[\ln \frac{p_X(\mathbf{x}_{<x}, \mathbf{y}_{<x}, \boldsymbol{\varkappa}_{<x})}{p_X(\mathbf{x}_{<x}, \boldsymbol{\varkappa}_{<x})} + \ln \frac{p_U(\mathbf{x}_{<x}, \boldsymbol{\varkappa}_{<x})}{p_U(\mathbf{x}_{<x}, \mathbf{y}_{<x}, \boldsymbol{\varkappa}_{<x})}\right]. \tag{8}$$

Eq (8) can be written as a sum of differential entropy and cross entropy terms

$$\dot{\mathbf{T}}_{Y\to X|\mathscr{Z}} = \bar{\lambda}_X \quad [-H(\mathbf{X}_{<X}, \mathbf{Y}_{<X}, \mathscr{Z}_{<X}) + H(\mathbf{X}_{<X}, \mathscr{Z}_{<X}) + H_{P_U}(\mathbf{X}_{<X}, \mathbf{Y}_{<X}, \mathscr{Z}_{<X}) - H_{P_U}(\mathbf{X}_{<X}, \mathscr{Z}_{<X})]. \tag{9}$$

Here, H refers to an entropy term and H_{P_U} refers to a cross entropy term. More specifically,

$$H(\mathbf{X}_{<X}, \mathscr{Z}_{<X}) = -\int p_X(\mathbf{x}_{<x}, \boldsymbol{\varkappa}_{<x}) \ln p_X(\mathbf{x}_{<x}, \boldsymbol{\varkappa}_{<x}) d\mathbf{x}_{<x} d\boldsymbol{\varkappa}_{<x}$$

and

$$H_{P_U}(\mathbf{X}_{<X}, \mathscr{Z}_{<X}) = -\int p_X(\mathbf{x}_{<x}, \boldsymbol{\varkappa}_{<x}) \ln p_U(\mathbf{x}_{<x}, \boldsymbol{\varkappa}_{<x}) d\mathbf{x}_{<x} d\boldsymbol{\varkappa}_{<x}.$$

It is worth noting in passing that Eq (8) can also be written as a difference of Kullback-Leibler divergences:

$$\dot{\mathbf{T}}_{Y\to X|\mathscr{Z}} = \bar{\lambda}_X [D_{KL}(P_X(\mathbf{X}_{<X}, \mathbf{Y}_{<X}, \mathscr{Z}_{<X}) || P_U(\mathbf{X}_{<X}, \mathbf{Y}_{<X}, \mathscr{Z}_{<X})) - D_{KL}(P_X(\mathbf{X}_{<X}, \mathscr{Z}_{<X}) || P_U(\mathbf{X}_{<X}, \mathscr{Z}_{<X}))]. \tag{10}$$

The expressions in Eqs (9) and (10) represent a general framework for estimating the TE between point processes in continuous time. Any estimator of differential entropy $\hat{H}$ which can be adapted to the estimation of cross entropies can be plugged into Eq (9) in order to estimate the TE. Similarly, any estimator of the KL divergence can be plugged into Eq (10).

Constructing *k*NN estimators for differential entropies and cross entropies. Following similar steps to the derivations in [53, 75, 90], assume that we have an (unknown) probability distribution $\mu(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^d$. Note that here $\mathbf{X}$ is a general random variable (not necessarily a point process). We also have a set X of N_X points drawn from μ. In order to estimate the differential entropy H we need to construct estimates of the form

$$\hat{H}(X) = -\frac{1}{N_X}\sum_{i=1}^{N_X} \widehat{\ln \mu(\mathbf{x}_i)} \tag{11}$$

where $\widehat{\ln \mu(\mathbf{x}_i)}$ is an estimate of the logarithm of the true density. Denote by $\epsilon(k, \mathbf{x}_i, X)$ the distance to the kth nearest neighbour of $\mathbf{x}_i$ in the set X under some norm L. Further, let p_i^μ be the probability mass of the ϵ-ball surrounding $\mathbf{x}_i$. If we make the assumption that $\mu(\mathbf{x}_i)$ is constant within the ϵ-ball, we have $p_i^\mu = \frac{k}{N_X - 1} = c_{d,L}\epsilon(k, \mathbf{x}_i, X)^d \mu(\mathbf{x}_i)$ where $c_{d,L}$ is the volume of the

d-dimensional unit ball under the norm L. Using this relationship, we can construct a simple estimator of the differential entropy:

$$\hat{H}(X) = -\frac{1}{N_X}\sum_{i=1}^{N_X} \ln \frac{k}{(N_X - 1)c_{d,L}\epsilon(k, \mathbf{x}_i, X)^d}. \tag{12}$$

We then add the bias-correction term $\ln k - \psi(k)$. $\psi(x) = \Gamma^{-1}(x)d\Gamma(x)/dx$ is the digamma function and $\Gamma(x)$ the gamma function. This yields $\hat{H}_{\mathrm{KL}}$, the Kozachenko-Leonenko [87] estimator of differential entropy:

$$\hat{H}_{\mathrm{KL}}(X) = \quad -\psi(k) + \ln(N_X - 1) + \ln c_{d,L} + \frac{d}{N_X}\sum_{i=1}^{N_X} \ln \epsilon(k, \mathbf{x}_i, X). \tag{13}$$

This estimator has been shown to be consistent [87, 91].

Assume that we now have two (unknown) probability distributions $\mu(\mathbf{x})$ and $\beta(\mathbf{x})$. We have a set X of N_X points drawn from μ and a set Y of N_Y points drawn from β. Using similar arguments to above, we denote by $\epsilon(k, \mathbf{x}_i, Y)$ the distance from the ith element of X to its kth nearest neighbour in Y. We then make the assumption that $\beta(\mathbf{x}_i)$ is constant within the ϵ-ball, and we have $p_i^\beta = \frac{k}{N_Y} = c_{d,L}\epsilon(k, \mathbf{x}_i, Y)^d \beta(\mathbf{x}_i)$. We can then construct a naive estimator of the cross entropy

$$\hat{H}_\beta(X) = -\frac{1}{N_X}\sum_{i=1}^{N_X} \ln \frac{k}{N_Y c_{d,L}\epsilon(k, \mathbf{x}_i, Y)^d}. \tag{14}$$

Again, we add the bias-correction term $\ln k - \psi(k)$ to arrive at an estimator of the cross entropy.

$$\hat{H}_{\beta,\mathrm{KL}}(X) = \quad -\psi(k) + \ln N_Y + \ln c_{d,L} + \frac{d}{N_X}\sum_{i=1}^{N_X} \ln \epsilon(k, \mathbf{x}_i, Y). \tag{15}$$

This estimator has been shown to be consistent [91].

Attention should be brought to the fundamental difference between estimating entropies and cross entropies using kNN estimators. An entropy estimator takes a set X and, for each $x_i \in X$, performs a nearest neighbour search *in the same set* X. An estimator of cross entropy takes two sets, X and Y and, for each $x_i \in X$, performs a nearest neighbour search *in the other set* Y.

We will be interested in applying these estimators to the entropy and cross entropy terms in Eq (9). For instance, we could use $\hat{H}_{\beta,\mathrm{KL}}(X)$ to estimate $H_{P_U}(\mathbf{X}_{<x}, \boldsymbol{\mathcal{Z}}_{<x})$, where we have that $\mu = p_X(\mathbf{x}_{<x}, \boldsymbol{\varkappa}_{<x})$ and $\beta = p_U(\mathbf{x}_{<x}, \boldsymbol{\varkappa}_{<x})$. This will be covered in more detail in a later subsection, after we first consider how to represent the history embeddings $\mathbf{x}_{<x}, \mathbf{y}_{<x}, \boldsymbol{\varkappa}_{<x}$ as well as sample them from their distributions.

Selection and representation of sample histories for entropy estimation. Inspection of Eqs (8) and (9) informs us that we will need to be able to estimate four distinct differential entropy terms and, implicitly, the associated probability densities:

1. The probability density of the target, source and conditioning histories at target events $p_X(\mathbf{x}_{<x}, \mathbf{y}_{<x}, \boldsymbol{\varkappa}_{<x})$.
2. The probability density of the target, and conditioning histories at target events $p_X(\mathbf{x}_{<x}, \boldsymbol{\varkappa}_{<x})$.

3. The probability density of the target, source and conditioning histories independent of target activity $p_U(\mathbf{x}_{<x}, \mathbf{y}_{<x}, \boldsymbol{\varkappa}_{<x})$.
4. The probability density of the target and conditioning histories independent of target activity $p_U(\mathbf{x}_{<x}, \boldsymbol{\varkappa}_{<x})$.

Estimation of these probability densities will require an associated set of samples for a kNN estimator to operate on. These samples for $\mathbf{x}_{<x}, \mathbf{y}_{<x}, \boldsymbol{\varkappa}_{<x}$ will logically be representated as history embeddings from the raw event times of the target $X \in \mathbb{R}^{N_X}$, source $Y \in \mathbb{R}^{N_Y}$ and conditioning $\boldsymbol{\mathcal{Z}} = \{Z_1, Z_2, \ldots, Z_{n_{\boldsymbol{\mathcal{Z}}}}\}, Z_i \in \mathbb{R}^{N_{Z_i}}$ processes. It is assumed that these sets are indexed in ascending order (from the first event to the last). The length of the history embeddings (in terms of how many previous spikes are referred to) must be restricted in order to avoid the difficulties associated with the estimation of probability densities in high dimensions. The lengths of the history embeddings along each process are specified by the parameters $l_X, l_Y, l_{Z_1}, \ldots, l_{Z_{n_{\boldsymbol{\mathcal{Z}}}}}$.

We label the sets of samples as $J_{<X} = \{\mathbf{j}_{<x_i}\}_{i=1}^{N_X}$, $C_{<X} = \{\mathbf{c}_{<x_i}\}_{i=1}^{N_X}$, $J_{<U} = \{\mathbf{j}_{<u_i}\}_{i=1}^{N_U}$, and $C_{<U} = \{\mathbf{c}_{<u_i}\}_{i=1}^{N_U}$, for each probability density $p_X(\mathbf{x}_{<x}, \mathbf{y}_{<x}, \boldsymbol{\varkappa}_{<x})$, $p_X(\mathbf{x}_{<x}, \boldsymbol{\varkappa}_{<x})$, $p_U(\mathbf{x}_{<x}, \mathbf{y}_{<x}, \boldsymbol{\varkappa}_{<x})$, and $p_U(\mathbf{x}_{<x}, \boldsymbol{\varkappa}_{<x})$ respectively (J for 'joint' and C for 'conditioning', i.e. without the source).

For the two sets of joint embeddings $J_{<^*}$ (where $^* \in \{X, U\}$) each $\mathbf{j}_{<*_i} \in J_{<*}$ is made up of target, source and conditioning components. That is, $\mathbf{j}_{<*_i} = \{\mathbf{x}_{<*_i}, \mathbf{y}_{<*_i}, \boldsymbol{\varkappa}_{<*_i}\}$ where $\boldsymbol{\varkappa}_{<*_i} = \{\mathbf{z}_{1,<*_i}, \mathbf{z}_{2,<*_i}, \ldots, \mathbf{z}_{n_{\boldsymbol{\mathcal{Z}}},<*_i}\}$. Similarly, for the two sets of conditioning embeddings $C_{<^*}$ (where $^* \in \{X, U\}$) each $\mathbf{c}_{<*_i} \in C_{<*}$ is made up of target, and conditioning components. That is, $\mathbf{c}_{<*_i} = \{\mathbf{x}_{<*_i}, \boldsymbol{\varkappa}_{<*_i}\}$.

Each set of embeddings $J_{<^*}$ is constructed from a set of observation points $T \in \mathbb{R}^{N_T}$. Each individual embedding $\mathbf{j}_{<*_i}$ is constructed at one such observation t_i. We denote by $\text{pred}(t_i, P)$, the index of the most recent event in the process P to occur before the observation point t_i.

The values of $\mathbf{x}_{<*_i} = \{x^1_{<*_i}, x^2_{<*_i}, \ldots, x^{l_X}_{<*_i}\} \in X_{<*}$ are set as follows:

$$x^k_{<*_i} := \begin{cases} t_i - x_{\text{pred}(t_i,X)} & k = 1 \\ x_{\text{pred}(t_i,X)-k+2} - x_{\text{pred}(t_i,X)-k+1} & k \neq 1. \end{cases} \tag{16}$$

Here, the $t_i \in T$ are the raw observation points and the $x_j \in X$ are the raw event times in the process X. The first element of $\mathbf{x}_{<*_i}$ is then the interval between the observation time and the most recent target event time $x_{\text{pred}(t_i, X)}$. The second element of $\mathbf{x}_{<*_i}$ is the inter-event interval between this most recent event time and the next most recent event time and so forth. The values of $\mathbf{y}_{<*_i} = \{y^1_{<*_i}, y^2_{<*_i}, \ldots, y^{l_X}_{<*_i}\} \in Y_{<*}$ and $\mathbf{z}_{<*_i} = \{z^1_{<*_i}, z^2_{<*_i}, \ldots, z^{l_X}_{<*_i}\} \in \boldsymbol{\mathcal{Z}}_{<*}$ are set in the same manner.

The set of samples $J_{<X} = \{\mathbf{j}_{<x_i}\}_{i=1}^{N_X} \subseteq \mathbb{R}^{l_X+l_Y+\sum l_{Z_j}}$ for $p_X(\mathbf{x}_{<x}, \mathbf{y}_{<x}, \boldsymbol{\varkappa}_{<x})$ is constructed using this scheme, with the set of observation points T being simply set as the N_X event times x_j of the target process X. As such, $J_{<X} = X_{<X} \times Y_{<X} \times \boldsymbol{\mathcal{Z}}_{<X}$.

In contrast, while the set of samples $J_{<U} = \{\mathbf{j}_{<u_i}\}_{i=1}^{N_U} \subseteq \mathbb{R}^{l_X+l_Y+\sum l_{Z_j}}$ for $p_U(\mathbf{x}_{<x}, \mathbf{y}_{<x}, \boldsymbol{\varkappa}_{<x})$ is also constructed using this scheme, the set of observation points T is set as $U \subseteq \mathbb{R}^{N_U}$. U is composed of sample time points placed independently of the occurrence of events in the target process. These N_U sample points between the first and last events of the target process X can either be placed randomly or at fixed intervals. In the experiments presented in this paper they were placed at fixed intervals. Importantly, note that N_U is not necessarily equal to N_X, with their ratio N_U/N_X a parameter for the estimator which is investigated in our Results. We also

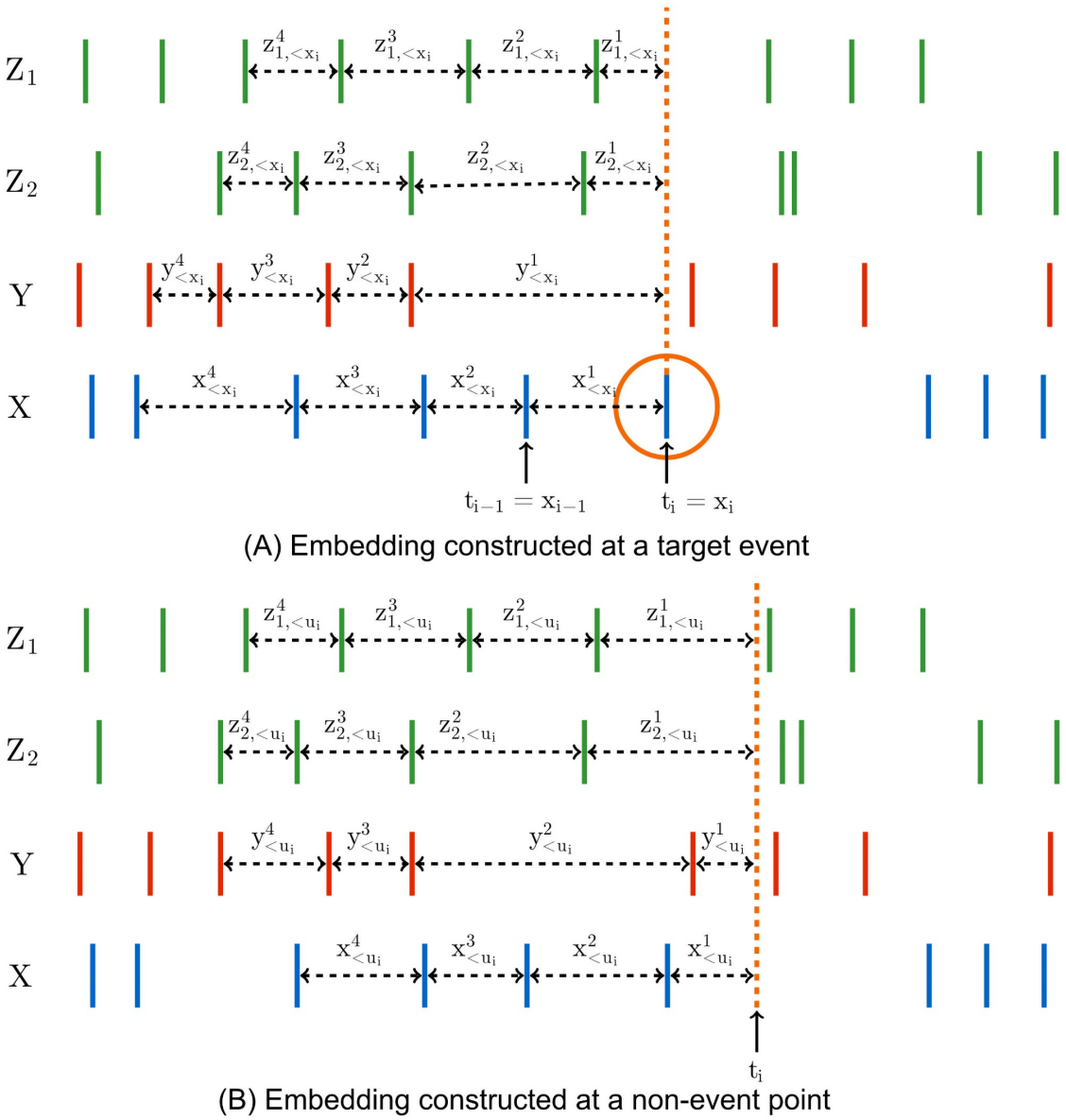


Fig 10. Examples of history embeddings. (A) shows an example of a joint embedding constructed at a target event ($\mathbf{j}_{<x_i} \in J_{<X}$). (B) shows an example of a joint embedding constructed at a sample event ($\mathbf{j}_{<u_i} \in J_{<U}$).

https://doi.org/10.1371/journal.pcbi.1008054.g010

have that $J_{<U} = X_{<U} \times Y_{<U} \times \mathcal{Z}_{<U}$. Fig 10 shows diagramatic examples of an embedded sample from $J_{<X}$ as well as one from $J_{<U}$. Notice the distinction that for $J_{<X}$, the $x^1_{<x_i}$ in the embeddings $\mathbf{x}_{<x_i}$ are specifically an interspike interval from the current spike at $t_i = x_i$ back to the previous spike, which is not the case for $J_{<U}$.

The set of samples $C_{<X} \subseteq \mathbb{R}^{l_X+\sum l_{Z_j}}$ for $p_X(\mathbf{x}_{<x}, \boldsymbol{\varkappa}_{<x})$ and $C_{<U} \subseteq \mathbb{R}^{l_X+\sum l_{Z_j}}$ for $p_U(\mathbf{x}_{<x}, \boldsymbol{\varkappa}_{<x})$ are constructed in a similar manner to their associated sets $J_{<X}$ and $J_{<U}$, however, the source embeddings $\mathbf{y}_{<*_i}$ are discarded. We will also have that $C_{<X} = X_{<X} \times \boldsymbol{\mathcal{Z}}_{<X}$ and $C_{<U} = X_{<U} \times \boldsymbol{\mathcal{Z}}_{<U}$.

Note that, as $J_{<X} = X_{<X} \times Y_{<X} \times \boldsymbol{\mathcal{Z}}_{<X}$ and $C_{<X} = X_{<X} \times \boldsymbol{\mathcal{Z}}_{<X}$, these two sets are closely related. Specifically, the i-th element of $C_{<X}$ will be identical to the i-th element of $J_{<X}$, apart from missing the source component $\mathbf{y}_{<x_i}$. Further, as the same set U is used for both $C_{<U}$ and $J_{<U}$, we will have that the i-th element of $C_{<U}$ will be identical to the i-th element of $J_{<U}$, apart from missing the source component $\mathbf{y}_{<u_i}$.

Combining $\hat{H}_*$ estimators for $\hat{\hat{T}}_{Y\to X|\boldsymbol{\mathcal{Z}}}$. With sets of samples and their embedded representation determined as per the previous subsection, we are now ready to estimate each of the four $\hat{H}_*$ terms in Eq (9). Here we consider how to combine the entropy and cross entropy estimators of these terms (Eqs (13) and (15)) into a single estimator.

We could simply estimate each H^* term in Eq (9) using $\hat{H}_{\mathrm{KL}}$ as specified in Eq 13 and $\hat{H}_{p_U,\mathrm{KL}}$ as specified in Eq (15), with the same number k of nearest neighbours in each of the four estimators and at each sample in the set for each estimator. Following the convention introduced in [90] we shall refer to this as a 4KL estimator of transfer entropy (the '4' refers to the 4 kNN searches and the 'KL' to Kozachenko-Leonenko):

$$\begin{aligned}\hat{\mathbf{T}}_{Y\to X|\boldsymbol{\mathcal{Z}},\mathrm{4KL}} = \frac{\bar{\lambda}_X}{N_X}\sum_{i=1}^{N_X}\Big\{ & l_J[-\ln\epsilon(k,\mathbf{j}_{<x_i},J_{<X}) + \ln\epsilon(k,\mathbf{j}_{<x_i},J_{<U})] \\ & + l_C[\ln\epsilon(k,\mathbf{c}_{<x_i},C_{<X}) - \ln\epsilon(k,\mathbf{c}_{<x_i},C_{<U})]\Big\}.\end{aligned} \tag{17}$$

Here, $l_J = (l_X + l_Y + \sum_{j=1}^{n_{\boldsymbol{\mathcal{Z}}}} l_{Z_j})$ is the dimension of the joint samples and $l_C = (l_X + \sum_{j=1}^{n_{\boldsymbol{\mathcal{Z}}}} l_{Z_j})$ is the dimension of the conditional-only samples. Note that the $\ln(N_X - 1) - \psi(k)$ terms cancel between the $J_{<X}$ and $C_{<X}$ terms (also for $\ln(N_U) - \psi(k)$ between the $J_{<U}$ and $C_{<U}$ terms), whilst the $\ln c_{d,L}$ terms cancel between $J_{<X}$ and $J_{<U}$ as well as between $C_{<X}$ and $C_{<U}$. It is crucial also to notice that all terms are averaged over N_X samples taken at target events (the cross-entropies which evaluate probability densities using $J_{<U}$ and $C_{<U}$ still evaluate those densities on the samples $\mathbf{j}_{<x_i} \in J_{<X}$ and $\mathbf{c}_{<x_i} \in C_{<X}$, following the definition in Eq (15)), regardless of whether $N_U = N_X$.

It is, however, not only possible to use a different k at every sample, but desirable when the k are chosen judiciously (as detailed below). We shall refer to this as the *generalised* kNN estimator:

$$\begin{aligned}\hat{\mathbf{T}}_{Y\to X|\boldsymbol{\mathcal{Z}},\mathrm{generalised}} = \frac{\bar{\lambda}_X}{N_X}\sum_{i=1}^{N_X}\Big\{ & \psi\left(k_{J_{<X},i}\right) - \psi\left(k_{J_{<U},i}\right) - \psi\left(k_{C_{<X},i}\right) + \psi\left(k_{C_{<U},i}\right) \\ & + l_J[-\ln\epsilon(k_{J_{<X},i},\mathbf{j}_{<x_i},J_{<X}) + \ln\epsilon(k_{J_{<U},i},\mathbf{j}_{<x_i},J_{<U})] \\ & + l_C[\ln\epsilon(k_{C_{<X},i},\mathbf{c}_{<x_i},C_{<X}) - \ln\epsilon(k_{C_{<U},i},\mathbf{c}_{<x_i},C_{<U})]\Big\}.\end{aligned} \tag{18}$$

Here $k_{A,i}$ is the number of neighbours used for the ith sample in set A for the corresponding entropy estimator for that set of samples. By theorems 3 and 4 of [75] this estimator (and, by implication, the 4KL estimator) is consistent. Application of the generalised estimator requires a scheme for choosing the $k_{A,i}$ at each sample. Work on constructing H^* kNN estimators for mutual information [53] and KL divergence [75] has found advantages in having certain H^* terms share the same or similar radii, e.g. resulting in lower overall bias due to components of

biases of individual H^* terms cancelling. Given that we have four H^* terms, there are a number of approaches we could take to sharing radii.

Our algorithm, which we refer to as the CT estimator of TE—$\hat{\mathbf{T}}_{Y\to X|\mathscr{Z},\mathrm{CT}}$—is specified in detail in Box 1. Our algorithm applies the approach proposed in [75] (referred to as the 'bias improved' estimator in that work) to each of the Kullback-Leibler divergence terms separately. In broad strokes, whereas Eq (17) uses the same k for each nearest-neighbour search, this estimator uses the same *radius* for each of the two nearest-neighbour searches relating to a given KL divergence term. In practice, this requires first performing searches with a fixed k in order to determine the radius to use. As such, we start with a fixed parameter k_{global}, which will be

Box 1: Algorithm for the CT TE estimator

input : /* The joint history embeddings at the target events and at the sampled points */
$J_{<X} = \{\mathbf{j}_{<x_i}\}_{i=1}^{N_X}$; $J_{<U} = \{\mathbf{j}_{<u_i}\}_{i=1}^{N_U}$
/* The conditioning history embeddings at the target events and at the sampled points */
$C_{<X} = \{\mathbf{c}_{<x_i}\}_{i=1}^{N_X}$; $C_{<U} = \{\mathbf{c}_{<u_i}\}_{i=1}^{N_U}$
/* The average firing rate of the target process */
$\bar{\lambda}_X$
/* The dimension of each element in the conditioning and joint sets */
l_C; l_J
/* The minimum number of nearest neighbours to consider in any set */
k_{global}

output: $\hat{\mathbf{T}}_{Y\to X|\mathscr{Z},\mathrm{CT}}$.

$\hat{\mathbf{T}}_{Y\to X|\mathscr{Z},\mathrm{CT}} \leftarrow 0$
for $i \leftarrow 1$ **to** N_X **do**
  /* Find the radii associated with the history embeddings constructed at events. */
  $\xi\left(k_{\mathrm{global}}, \mathbf{j}_{<x_i}, J_{<X}\right) \leftarrow$ `findDistanceToKthNearestNeighbour` $\left(k_{\mathrm{global}}, \mathbf{j}_{<x_i}, J_{<X}\right)$
  $\xi\left(k_{\mathrm{global}}, \mathbf{j}_{<x_i}, J_{<U}\right) \leftarrow$ `findDistanceToKthNearestNeighbour` $\left(k_{\mathrm{global}}, \mathbf{j}_{<x_i}, J_{<U}\right)$
  $r_{\mathrm{joint},i} \leftarrow \max\left\{\xi\left(k_{\mathrm{global}}, \mathbf{j}_{<x_i}, J_{<X}\right), \xi\left(k_{\mathrm{global}}, \mathbf{j}_{<x_i}, J_{<U}\right)\right\}$
  $k_{J_{<X},i} \leftarrow$ `findNumberOfNeighboursInRadius` $\left(r_{\mathrm{joint},i}, \mathbf{j}_{<x_i}, J_{<X}\right)$
  $\epsilon\left(k_{J_{<X},i}, \mathbf{j}_{<x_i}, J_{<X}\right) \leftarrow 2 *$ `findDistanceToKthNearestNeighbour` $\left(k_{J_{<X},i}, \mathbf{j}_{<x_i}, J_{<X}\right)$
  $k_{J_{<U},i} \leftarrow$ `findNumberOfNeighboursInRadius` $\left(r_{\mathrm{joint},i}, \mathbf{j}_{<x_i}, J_{<U}\right)$
  $\epsilon\left(k_{J_{<U},i}, \mathbf{j}_{<x_i}, J_{<U}\right) \leftarrow 2 *$ `findDistanceToKthNearestNeighbour` $\left(k_{J_{<U},i}, \mathbf{j}_{<x_i}, J_{<U}\right)$
  /* Find the radii associated with the embeddings constructed at randomly sampled points. */
  $\xi\left(k_{\mathrm{global}}, \mathbf{c}_{<x_i}, C_{<X}\right) \leftarrow$ `findDistanceToKthNearestNeighbour` $\left(k_{\mathrm{global}}, \mathbf{c}_{<x_i}, C_{<X}\right)$
  $\xi\left(k_{\mathrm{global}}, \mathbf{c}_{<x_i}, C_{<U}\right) \leftarrow$ `findDistanceToKthNearestNeighbour` $\left(k_{\mathrm{global}}, \mathbf{c}_{<x_i}, C_{<U}\right)$
  $r_{\mathrm{conditioning},i} \leftarrow \max\left\{\xi\left(k_{\mathrm{global}}, \mathbf{c}_{<x_i}, C_{<X}\right), \xi\left(k_{\mathrm{global}}, \mathbf{c}_{<x_i}, C_{<U}\right)\right\}$
  $k_{C_{<X},i} \leftarrow$ `findNumberOfNeighboursInRadius` $\left(r_{\mathrm{conditioning},i}, \mathbf{c}_{<x_i}, C_{<X}\right)$
  $\epsilon\left(k_{C_{<X},i}, \mathbf{c}_{<x_i}, C_{<X}\right) \leftarrow 2 *$ `findDistanceToKthNearestNeighbour` $\left(k_{C_{<X},i}, \mathbf{c}_{<x_i}, C_{<X}\right)$
  $k_{C_{<U},i} \leftarrow$ `findNumberOfNeighboursInRadius` $\left(r_{\mathrm{conditioning},i}, \mathbf{c}_{<x_i}, C_{<U}\right)$
  $\epsilon\left(k_{C_{<U},i}, \mathbf{c}_{<x_i}, C_{<U}\right) \leftarrow 2 *$ `findDistanceToKthNearestNeighbour` $\left(k_{C_{<U},i}, \mathbf{c}_{<x_i}, C_{<U}\right)$
  /* We now combine these quantities into the contribution to the TE from the given target event. */
  $\hat{\mathbf{T}}_{Y\to X|\mathscr{Z},\mathrm{CT}} \leftarrow \hat{\mathbf{T}}_{Y\to X|\mathscr{Z},\mathrm{CT}} + \psi\left(k_{J_{<X},i}\right) - \psi\left(k_{J_{<U},i}\right) - \psi\left(k_{C_{<X},i}\right) + \psi\left(k_{C_{<U},i}\right)$
  $+ l_J\left[-\ln\epsilon\left(k_{J_{<X},i}, \mathbf{j}_{<x_i}, J_{<X}\right) + \ln\epsilon\left(k_{J_{<U},i}, \mathbf{j}_{<x_i}, J_{<U}\right)\right]$
  $+ l_C\left[\ln\epsilon\left(k_{C_{<X},i}, \mathbf{c}_{<x_i}, C_{<X}\right) - \ln\epsilon\left(k_{C_{<U},i}, \mathbf{c}_{<x_i}, C_{<U}\right)\right]$
end
$\hat{\mathbf{T}}_{Y\to X|\mathscr{Z},\mathrm{CT}} \leftarrow \frac{\bar{\lambda}_X}{N_X}\hat{\mathbf{T}}_{Y\to X|\mathscr{Z},\mathrm{CT}}$

the minimum number of nearest neighbours in any search space. For each joint sample at a target event, that is, each $\mathbf{j}_{<x_i}$ in $J_{<X}$, we perform a k_{global}NN search in this same set $J_{<X}$ and record the distance to the k_{global}-th nearest neighbour (line 3 of Box 1). We perform a similar k_{global}NN search for $\mathbf{j}_{<x_i}$ in the set of joint samples independent of target activity $J_{<U}$, again recording the distance to the k_{global}-th nearest neighbour (line 4). We define a search radius as the maximum of these two distances (line 5). We then find the number of points in $J_{<X}$ that fall within this radius of $\mathbf{j}_{<x_i}$ and set $k_{J<X}, i$ as this number (line 6). We also find twice the distance to the $k_{J<X}, i$-th nearest neighbour, which is the term $\epsilon(k_{J_{<X},i}, \mathbf{j}_{<x_i}, J_{<X})$ in Eq (18) (line 7). Similarly, we find the number of points in $J_{<U}$ that fall within the search radius of $\mathbf{j}_{<x_i}$ and set $k_{J<U}, i$ as this number (line 8). We find twice the distance to the $k_{J<U}, i$-th nearest neighbour, which is the term $\epsilon(k_{J_{<U},i}, \mathbf{j}_{<x_i}, J_{<U})$ (line 9).

In the majority of cases, only one of these two ϵ terms will be exactly twice the search radius, and its associated $k_{A,i}$ will equal k_{global}. In such cases, the other ϵ will be less than twice the search radius and its associated $k_{A,i}$ will be greater than or equal to k_{global}.

The same set of steps is followed for each conditioning history embedding that was constructed at an event in the target process, that is, each $\mathbf{c}_{<x_i}$ in $C_{<X}$, over the sets $C_{<X}$ and $C_{<U}$ (lines 10 through 16 of Box 1).

The values that we have found for $k_{J_{<X},i}$, $k_{J_{<U},i}$, $k_{C_{<X},i}$, $k_{C_{<U},i}$, $\epsilon(k_{J_{<X},i}, \mathbf{j}_{<x_i}, J_{<x_i})$, $\epsilon(k_{J_{<U},i}, \mathbf{j}_{<x_i}, J_{<U})$, $\epsilon(k_{C_{<X},i}, \mathbf{c}_{<x_i}, C_{<x_i})$ and $\epsilon(k_{C_{<U},i}, \mathbf{c}_{<x_i}, C_{<U})$ can be plugged into Eq (18) (lines 17 and 19 of Box 1).

Handling dynamic correlations. The derivation of the kNN estimators for entropy and cross entropy given above assumes that the points are independent [53]. However, nearby interspike intervals might be autocorrelated (e.g. during bursts), and indeed our method for constructing history embeddings (see Selection and Representation of Sample Histories for Entropy Estimation) will incorporate the same interspike intervals at different positions in consecutive samples. This contradicts the assumption of independence. In order to satisfy the assumption of independence when counting neighbours, conventional neighbour counting estimators can be made to ignore matches within a dynamic or serial correlation exclusion window (a.k.a. Theiler windows [92, 93]).

For our estimator, we maintain a record of the start and end times of each history embedding, providing us with an exclusion window. The start time is recorded as the time of the first event that formed part of an interval which was included in the sample. This event could come from the embedding of any of the processes from which the sample was constructed. The end of the window is the observation point from which the sample is constructed. When performing nearest neighbour and radius searches (lines lines 3, 4, 6, 7, 8, 9, 10, 11, 13, 14, 15 and 16 of Box 1 and line 6 of Box 2), any sample whose exclusion window overlaps with the exclusion window of the original data point around which the search is taking place is ignored. Subtleties concerning dynamic correlation exclusion for surrogate calculations are considered in the next subsection.

Local permutation method for surrogate generation

A common use of this estimator would be to ascertain whether there is a non-zero conditional information flow between two components of a system. When using TE for directed functional network inference, this is the criteria we use to determine the presence or absence of a connection. Given that we are estimating the TE from finite samples, we require a statistical test in order to determine the significance of the measured TE value. Unfortunately, analytic results do not exist for the sampling distribution of kNN estimators of information theoretic

Box 2: Algorithm for the local permutation method for surrogate generation.

Input : /* The joint history embeddings at the target events */
$J_{<X} = \{\mathbf{j}_{<x_i}\}_{i=1}^{N_X} = \{\mathbf{x}_{<x_i}, \mathbf{y}_{<x_i}, \boldsymbol{\varkappa}_{<x_i}\}_{i=1}^{N_X}$
/* The joint history embeddings at the sampled points */
$J_{<U,\text{surr}} = \{\mathbf{j}_{<u_i}\}_{i=1}^{N_{U,\text{surr}}} = \{\mathbf{x}_{<u_i}, \mathbf{y}_{<u_i}, \boldsymbol{\varkappa}_{<u_i}\}_{i=1}^{N_{U,\text{surr}}}$
k_{perm}

Output : $J_{<X,\text{surr}}$

/* Set to keep a record of the used indices in the independently sampled embeddings. */
$\mathcal{U} \leftarrow \emptyset$ /* Initialise this set to be empty */
$J_{<X,\text{surrogate}} \leftarrow \emptyset$ /* Initialise the surrogate embeddings as empty */
$I \leftarrow \{i\}_{i=1}^{N_X}$ /* Initialise the indices to iterate over */
/* Shuffle the indices to ensure that different samples are assigned duplicate source componenents each time surrogate sample sets are generated. */
$I \leftarrow$ `shuffle`(I)
for $i \in I$ **do**
  /* Search for the nearest neighbours in the set of embeddings at sampled points;ignoring the source components. The function findIndicesOfNearestNeighbours(k;a;B) finds the indices of the k nearest neighbours of the point a in the set B. */
  $\mathcal{N} \leftarrow$ `findIndicesOfNearestNeighbours` $\left(k_{\text{perm}}, \{\mathbf{x}_{<x_i}, \boldsymbol{\varkappa}_{<x_i}\}, \{\mathbf{x}_{<u_j}, \boldsymbol{\varkappa}_{<u_j}\}_{j=1}^{N_{U,\text{surr}}}\right)$
  /* Create a set of candidate indices by removing those already used. */
  $\mathcal{E} \leftarrow \mathcal{N} \setminus (\mathcal{N} \cap \mathcal{U})$
  if $\|\mathcal{E}\| > 0$ **then**
    $h \leftarrow$ `chooseRandomElement` $(\mathcal{E})$
  end
  else
    $h \leftarrow$ `chooseRandomElement` $(\mathcal{N})$
  end
  /* Append the set of surrogate samples with an embedding composed of the original target and conditioning components (at index i);but with the source component swapped for that at index h of the independently sampled embeddings. */
  $J_{<X,\text{surrogate}} \leftarrow J_{<X,\text{surrogate}} \cup \{\mathbf{x}_{<x_i}, \mathbf{y}_{<u_h}, \boldsymbol{\varkappa}_{<x_i}\}$
  /* Add the index that we chose to the set keeping track of used indices */
  $\mathcal{U} \leftarrow \mathcal{U} \cup h$
end

quantities [48]. This necessitates a scheme for generating surrogate samples from which the null distribution can be empirically constructed.

It is instructive to first consider the more general case of testing for non-zero mutual information. As the mutual information between X and Y is zero if and only X and Y are independent, testing for non-zero mutual information is a test for statistical dependence. As such, we are testing against the null hypothesis that X and Y are independent ($X \perp\!\!\!\perp Y$) or, equivalently, that the joint probability distribution of X and Y factorises as $p(X, Y) = p(X)p(Y)$. It is straightforward to construct surrogate pairs $(\check{x}, \check{y})$ that conform to this null hypothesis. We start with the original pairs (x, y) and resample the y values across pairs, commonly by shuffling (in conjunction with handling dynamic correlations, as per Implementation). This shuffling process will maintain the marginal distributions $p(X)$ and $p(Y)$, and the same number of samples, but will destroy any relationship between X and Y, yielding the required factorisation for the null

hypothesis. One shuffling process produces one set of surrogate samples; estimates of mutual information on populations of such surrogate sample sets yields a null distribution for the mutual information.

As transfer entropy is a conditional mutual information ($I(X_t ; \mathbf{Y}_{<t} \mid \mathbf{X}_{<t}, \mathcal{Z}_{<t})$), we are testing against the null hypothesis that the current state of the target X_t is conditionally independent of the history of the source $\mathbf{Y}_{<t}$ ($X_t \perp\!\!\!\perp \mathbf{Y}_{<t} \mid \mathbf{X}_{<t}, \mathcal{Z}_{<t}$). That is, the null hypothesis states that the joint distribution factorises as: $p(X_t, \mathbf{Y}_{<t} \mid \mathbf{X}_{<t}, \mathcal{Z}_{<t}) = p(X_t \mid \mathbf{X}_{<t}, \mathcal{Z}_{<t}) p(\mathbf{Y}_{<t} \mid \mathbf{X}_{<t}, \mathcal{Z}_{<t})$.

Historically, the generation of surrogates for TE has been done by either shuffling source history embeddings or by shifting the source time series (see discussions in e.g. [4, 94]). These approaches lead to various problems. These problems stem from the fact that they destroy any relationship between the source history ($\mathbf{Y}_{<t}$) and both the target ($\mathbf{X}_{<t}$) and conditioning ($\mathcal{Z}_{<t}$) histories. As such, they are testing against the null hypothesis that the joint distribution factorises as: $p(X_t, \mathbf{Y}_{<t} \mid \mathbf{X}_{<t}, \mathcal{Z}_{<t}) = p(X_t \mid \mathbf{X}_{<t}, \mathcal{Z}_{<t}) p(\mathbf{Y}_{<t})$ [48]. The problems associated with this factorisation become particularly pronounced when we are considering a system whereby the conditioning processes $\mathcal{Z}_{<t}$ drive both the current state of the target X_t as well as the history of the source $\mathbf{Y}_{<t}$. This can lead to $\mathbf{Y}_{<t}$ being highly correlated with X_t, but conditionally independent. This is the classic case of a "spurious correlation" between $\mathbf{Y}_{<t}$ and X_t being mediated through the "confounding variable" $\mathcal{Z}_{<t}$. If, in such a case, we use time shifted or shuffled source surrogates to test for the significance of the TE, we will be comparing the TE measured when X_t and $\mathbf{Y}_{<t}$ are highly correlated (albeit potentially conditionally independent) with surrogates where they are independent. This subtle difference in the formulation of the null may result in a high false positive rate in a test for conditional independence. An analysis of such a system is presented in the third subsection of Results. Alternately, if we can generate surrogates where the joint probability distribution factorises correctly and the relationship between $\mathbf{Y}_{<t}$ and the histories $\mathbf{X}_{<t}$ and $\mathcal{Z}_{<t}$ is maintained, then $\mathbf{Y}_{<t}$ will maintain much of its correlation with X_t through the mediating variables $\mathcal{Z}_{<t}$ and $\mathbf{X}_{<t}$. We would anticipate conditional independence tests using surrogates generated under this properly formed null to have a false positive rate closer to what we expect.

Generating surrogates for testing for conditional dependence is relatively straightforward in the case of discrete-valued conditioning variables. If we are testing for dependence between X and Y given Z, then, for each unique value of Z, we can shuffle the associated values of Y. This maintains the distributions $p(X|Z)$ and $p(Y|Z)$ whilst, for any given value of Z, the relationship between the associated X and Y values is destroyed.

The problem is more challenging when Z can take on continuous values. However, recent work by Runge [48], demonstrated the efficacy of a local permutation technique. In this approach, to generate one surrogate sample set, we separately generate a surrogate sample $(x, \tilde{y}, z)$ for each sample (x, y, z) in the original set. We find the k_{perm} nearest neighbours of z in Z: one of these neighbours, z', is chosen at random, and y is swapped with the associated y' to produce the surrogate sample (x, y', z). In order to reduce the occurrence of duplicate y values, a set $\mathcal{U}$ of used indices is maintained. After finding the k_{perm} nearest neighbours, those that have already been used are removed from the candidate set. If this results in an empty candidate set, one of the original k_{perm} candidates are chosen at random. Otherwise, this choice is made from the reduced set. As before, a surrogate conditional mutual information is estimated for every surrogate sample set, and a population of such surrogate estimates provides the null distribution.

This approach needs to be adapted slightly in order to be applied to our particular case, because we have implictly removed the target variable (whether or not the target is spiking)

from our samples via the novel Bayesian inversion. We can rewrite Eq (8) as:

$$\dot{\mathbf{T}}_{Y\to X|\mathscr{Z}} = \bar{\lambda}_X \mathbb{E}_X\left[\ln \frac{p_X(\mathbf{x}_{<x}, \mathbf{y}_{<x}, \boldsymbol{\varkappa}_{<x})}{p_X(\mathbf{x}_{<x}, \boldsymbol{\varkappa}_{<x}) p_U(\mathbf{y}_{<x} \mid \mathbf{x}_{<x}, \boldsymbol{\varkappa}_{<x})}\right]. \tag{19}$$

This makes it clear that we are testing whether the following factorisation holds:

$$p_X(\mathbf{x}_{<x}, \mathbf{y}_{<x}, \boldsymbol{\varkappa}_{<x}) = p_X(\mathbf{x}_{<x}, \boldsymbol{\varkappa}_{<x}) p_U(\mathbf{y}_{<x} \mid \mathbf{x}_{<x}, \boldsymbol{\varkappa}_{<x}) \tag{20}$$

(recall the difference between probability densities at target events p_X and those not conditioned at target events p_U). In order to create surrogates $J_{<X,\text{surr}}$ that conform to this null distribution, we resample a new set from our original data in a way that maintains the relationship between the source histories and the histories of the target and conditioning processes, but decouples (only) the source histories from target events. (As above, simply shuffling the source histories across $J_{<X}$ or shifting the source events does not properly maintain the relationship of the source to the target and conditioning histories). The procedure to achieve this is detailed in Box 2. We start with the samples at target events $J_{<X} = \{\mathbf{x}_{<x_i}, \mathbf{y}_{<x_i}, \boldsymbol{\varkappa}_{<x_i}\}_{i=1}^{N_X}$ and resample the source components $\mathbf{y}_{<x_i}$ as follows. We first construct a new set $J_{<U,\text{surr}} = \{\mathbf{x}_{<u_i}, \mathbf{y}_{<u_i}, \boldsymbol{\varkappa}_{<u_i}\}_{i=1}^{N_{U,\text{surr}}}$ from the set U_{surr} of $N_{U,\text{surr}}$ points sampled independently of events in the target. This set is constructed in the same manner as $J_{<U}$, although we might choose to change the number of sample points ($N_{U,\text{surr}} \neq N_U$) at which the embeddings are constructed, or whether the points are placed randomly or at fixed intervals. For each original sample $\mathbf{j}_{<x_i}$ from $J_{<X}$, we then find the nearest neighbours $\{\mathbf{x}_{<u_i}, \boldsymbol{\varkappa}_{<u_i}\}_{i=1}^{k_{\text{perm}}}$ of $\{\mathbf{x}_{<x_i}, \boldsymbol{\varkappa}_{<x_i}\}$ in $J_{<U,\text{surr}}$ (line 9 of Box 12), select $\mathbf{y}_{<u_j}$ randomly from amongst the k_{perm} nearest neighbours (line 6 or Box 2), and add a sample $\{\mathbf{x}_{<x_i}, \mathbf{y}_{<u_j}, \boldsymbol{\varkappa}_{<x_i}\}$ to $J_{<X,\text{ surr}}$ (line 14). The construction of such a sample is also displayed in Fig 11. Similar to Runge [48], we also keep a record of used indices in order

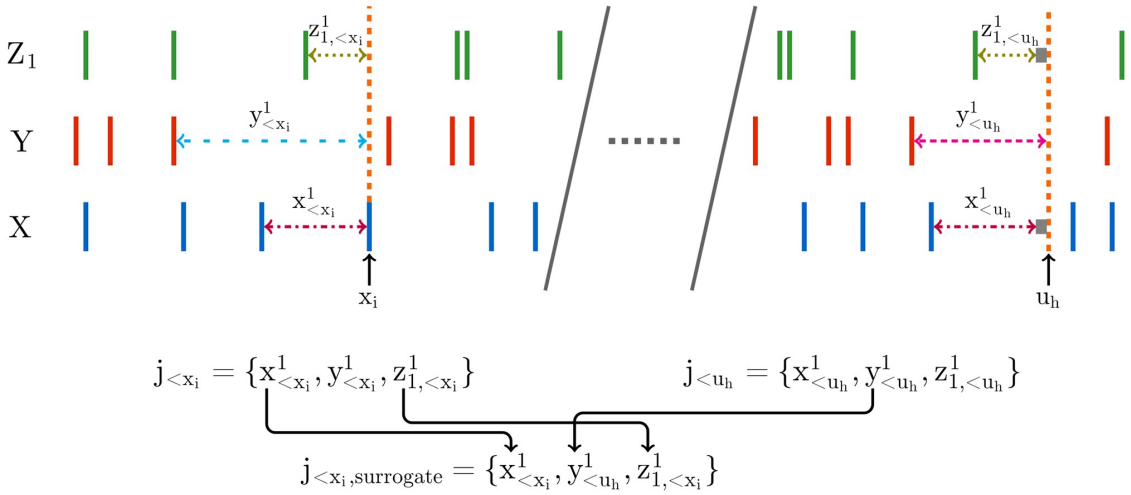


Fig 11. Diagrammatic representation of the local permutation surrogate generation scheme. For our chosen sample $\mathbf{j}_{<x_i}$ we find a $\mathbf{j}_{<u_h} \in J_{<U,\text{surr}}$ where we have that the $\mathbf{x}_{<x_i}$ component of $\mathbf{j}_{<x_i}$ is similar to the $\mathbf{x}_{<u_h}$ component of $\mathbf{j}_{<u_h}$ and $\boldsymbol{\varkappa}_{<x_i}$ component of $\mathbf{j}_{<x_i}$ is similar to the $\boldsymbol{\varkappa}_{<u_h}$ component of $\mathbf{j}_{<u_h}$. We then form a single surrogate sample by combining the $\mathbf{x}_{<x_i}$ and $\boldsymbol{\varkappa}_{<x_i}$ components of $\mathbf{j}_{<x_i}$ with the $\mathbf{y}_{<u_h}$ component of $\mathbf{j}_{<u_h}$. Corresponding colours of the dotted interval lines indicates corresponding length. The grey boxes indicate a small delta.

https://doi.org/10.1371/journal.pcbi.1008054.g011

to reduce the incidence of duplicate $\mathbf{y}_{<u_j}$ (line 15). For each redrawn surrogate sample set $J_{<X,\,\mathrm{surr}}$ a surrogate conditional mutual information is estimated (utilising the same $J_{<U}$ selected independently of the target events as was used for the original TE estimate) following the algorithm outlined earlier; the population of such surrogate estimates provides the null distribution as before.

The p values are calculated by constructing $N_{\mathrm{surrogates}}$ surrogates by the algorithm just described. The TE is estimated on these surrogates and compared to the TE estimated on the original embeddings. The p value is then the number of estimates on surrogate embeddings which were larger than the estimate on the original data divided by the total number of surrogates.

Finally, we note an additional subtlety for dynamic correlation exclusion for the surrogate calculations. Samples in the surrogate calculations will have had their history components originating from two different time windows. One will be from the construction of the original sample and the other from the sample with which the source component was swapped. A record is kept of both these exclusion windows and, during neighbour searches, points are excluded if their exclusion windows intersect either of the exclusion windows of the surrogate history embedding.

Implementation

The algorithms shown in Boxes 2 and 1 as well as all experiments were implemented in the Julia language. The implementation of the algorithms is freely available at the following repository: github.com/dpshorten/CoTETE.jl. Scripts to run the experiments in the paper can be found here: github.com/dpshorten/CoTETE_experiments. Implementations of kNN information-theoretic estimators have commonly made use of KD-trees to speed up the nearest neighbour searches [94]. A popular Julia nearest neighbours library (NearestNeighbors.jl, available from github.com/KristofferC/NearestNeighbors.jl) was modified such that checks for dynamic exclusion windows (see Handling Dynamic Correlations) were performed during the KD-tree searches when considering adding neighbours to the candidate set.

Assumptions used to conclude conditional independence or dependence

We summarise here the conditions and assumptions that allow us to draw conclusions about conditional independence relationships from the structure in a model. Although these relationships are obvious in some of our examples (see Results), they are less so in others. If we consider, for now, the discrete-time case, then for a sufficiently small Δt there will be no instantaneous effects. This implies that the causal relationships in these models can be represented by a Directed Acyclic Graph (DAG); specifically a Dynamic Bayesian Network with multiple time slices and connections only going forward in time (see [95]). In order to conclude that connected nodes will be statistically dependent we need to use the contraposition of the faithfulness assumption [55, 56, 57]. This assumption states that, if two nodes are conditionally independent, given some conditioning set S, then they are d-separated [55] given the same set. This in turn implies that if there exists some set of conditioning processes by which a node is conditionally independent of another, then there is no direct causal link between these nodes. It is worth asking how reasonable the faithfulness assumption is. After all, particularly for the case of deterministic dynamics, it is easy to construct examples whereby the present state of each of a pair of processes is determined by the history of the other process, but where the present state of each process is conditionally independent of the history of the other [96, 97] (e.g. one can have zero TE when a real causal connection exists, for instance, the system $x_t = y_{t-1}$, $y_t = x_{t-1}$, $x_1 = 0$ and $y_1 = 1$). Such examples violate

faithfulness. However, determinism is not a realistic assumption for biological systems or their models. Moreover, it can be shown that almost all discrete probability distributions (such as those of spike trains) satisfy faithfulness. Indeed, the set of discrete probability distributions that violate this assumption has measure zero [98]. In order to determine that the present state of a process is independent of an unconnected source, when conditioning on its direct causal parents, we need to assume sufficiency and the causal Markov condition [55, 56, 57]. Sufficiency assumes that we have observed all relevant variables (which is easy to meet if we are defining the model). The causal Markov condition states that d-separation implies conditional independence. Conditioning on all the direct causal parents of a variable provides us d-separation. In summary then, under these conditions the directed structural connections designed in our models are expected to have a one-to-one correspondence with directed conditional dependence (or independence, in their absence), when appropriately conditioned on other nodes. Correctly differentiating conditional dependence and independence then, in alignment with the underlying structural connections in these models, provides an important validation of the correctness of the estimators.

Specification of leaky-integrate-and-fire model

What follows is a specification of the Leaky-Integrate-and-Fire (LIF) model which we used in the Results subsection Scaling of Conditional Independence Testing in Higher Dimensions. The membrane potential evolves according to:

$$\tau \frac{dV}{dt} = V_0 - V. \tag{21}$$

When V crosses the threshold $V_{\text{threshold}}$, the timestamp of crossing is recorded as a spike. V is then set to V_{reset} and the evolution of the membrane potential is subsequently paused for the duration of the hard refractory period. In the case of excitatory connections, when a presynaptic spike occurs, V is instantaneously increased by the connection strength of the synapse (specified in millivolts) at the delay specified by the connections delay parameter. Inhibitory connects behave in the same manner, but lead to a decrease in V. We use the initial condition $V(t = 0) = V_0$.

Supporting information

S1 Fig. Longer embeddings on homogeneous. The results of an identical experimental setup to those displayed in Fig 2, but with history embedding lengths of $l_X = l_Y = 3$.
(TIFF)

S2 Fig. Different embeddings on discrete homogeneous. The results of an identical experimental setup to those displayed in Fig 3, but where the history embedding lengths (l and m) were set to cover the distance of an average interspike interval. Specifically, these lengths were 1, 2, 5 and 10, corresponding to the Δt values of 1.0, 0.5, 0.2 and 0.1.
(TIFF)

S3 Fig. Different embeddings on continuous coupled. The results of an identical experimental setup to those displayed in Fig 4B, but where the history embedding length of the source is set to $l_Y = 3$.
(TIFF)

S4 Fig. Conditional independence scaling at constant rate. The results of an identical experimental setup to those displayed in Fig 8, but with a constant rate of 20 Hz in all the stimuli. This removes the correlation between the unconnected source and the firing of the target. The

top row shows results of the continuous-time approach, the bottom shows results of the discrete-time approach.
(TIFF)

S5 Fig. Independence test with no conditioning. The results of an identical experimental setup to those displayed in Fig 8, but where the background processes are not included in the conditioning set (the conditioning set is left empty). This represents the nature of the inference task at the early stage of a greedy network inference algorithm being applied to a node. We see that the continuous-time estimator performs well on inhibitory connections in this case. Due to the change in dimension, different source and target embedding lengths (l and m) as well as bin widths Δt were used for the discrete-time estimator. These were set at $l = m = 12$ and $\Delta t = 2$ms. The top row shows results of the continuous-time approach, the bottom shows results of the discrete-time approach.
(TIFF)

S6 Fig. Conditional independence testing with the discrete-time estimator and permutation-based surrogates. The results of identical experimental setups to those displayed in the bottom rows of Fig 8, S4 and S5 Figs. As the bottom rows of all of these figures show the results of the discrete-time estimator, the plots in this figure similarly all display the results of runs of the discrete-time estimator. However, where the other plots make use of the source time-shift method for surrogate generation (as is traditionally used in conjunction with TE estimators), these plots make use of a standard conditional-permutation-based surrogate generation scheme for categorical variables [64]. The top row of this figure corresponds to the bottom row of Fig 8, the middle row corresponds to the middle row of S4 Fig and the bottom row corresponds to the bottom row of S5 Fig.
(TIFF)

S7 Fig. Pyloric STG continuous different embedding lengths. The results of an identical experimental setup to those displayed in Fig 9C, but where different embeddings lengths (l_X, l_Y and l_{Z_1}) are used. The left plot shows $l_X = l_Y = l_{Z_1} = 2$ and the right plot shows $l_X = l_Y = l_{Z_1} = 4$.
(TIFF)

S8 Fig. Pyloric STG continuous different dataset sizes. The results of an identical experimental setup to those displayed in Fig 9C, but where different numbers of target spikes N_X are used. The left plot shows $N_X = 1 \times 4$ and the right plot shows $N_X = 3.5 \times 4$.
(TIFF)

S1 Text. Description of the biophysical neural network model insipred by the Pyloric STG.
(PDF)

Acknowledgments

We would like to thank Mike Li for performing preliminary benchmarking of the performance of the discrete-time estimator on point processes.

Author Contributions

Conceptualization: David P. Shorten, Richard E. Spinney, Joseph T. Lizier.

Funding acquisition: Joseph T. Lizier.

Methodology: David P. Shorten, Richard E. Spinney, Joseph T. Lizier.

Software: David P. Shorten.

Supervision: Richard E. Spinney, Joseph T. Lizier.

Writing – original draft: David P. Shorten.

Writing – review & editing: David P. Shorten, Richard E. Spinney, Joseph T. Lizier.

References

1. Olshausen BA, Field DJ. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature. 1996; 381(6583):607. https://doi.org/10.1038/381607a0
2. Georgopoulos AP, Ashe J, Smyrnis N, Taira M. The motor cortex and the coding of force. Science. 1992; 256(5064):1692–1695. https://doi.org/10.1126/science.256.5064.1692
3. Schreiber T. Measuring information transfer. Physical Review Letters. 2000; 85(2):461. https://doi.org/10.1103/PhysRevLett.85.461
4. Bossomaier T, Barnett L, Harré M, Lizier JT. An introduction to transfer entropy. Cham: Springer International Publishing. 2016; p. 65–95.
5. Spinney RE, Prokopenko M, Lizier JT. Transfer entropy in continuous time, with applications to jump and neural spiking processes. Physical Review E. 2017; 95(3):032319. https://doi.org/10.1103/PhysRevE.95.032319
6. MacKay DJ. Information theory, inference and learning algorithms. Cambridge University Press; 2003.
7. Novelli L, Wollstadt P, Mediano P, Wibral M, Lizier JT. Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing. Network Neuroscience. 2019; 3(3):827–847. https://doi.org/10.1162/netn_a_00092
8. Honey CJ, Kötter R, Breakspear M, Sporns O. Network structure of cerebral cortex shapes functional connectivity on multiple time scales. Proceedings of the National Academy of Sciences. 2007; 104(24):10240–10245. https://doi.org/10.1073/pnas.0701519104
9. Stetter O, Battaglia D, Soriano J, Geisel T. Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals. PLoS Computational Biology. 2012; 8(8):e1002653. https://doi.org/10.1371/journal.pcbi.1002653
10. Sun J, Bollt EM. Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings. Physica D: Nonlinear Phenomena. 2014; 267:49–57. https://doi.org/10.1016/j.physd.2013.07.001
11. Faes L, Nollo G, Porta A. Information-based detection of nonlinear Granger causality in multivariate processes via a nonuniform embedding technique. Physical Review E. 2011; 83(5):051112. https://doi.org/10.1103/PhysRevE.83.051112
12. Stramaglia S, Wu GR, Pellicoro M, Marinazzo D. Expanding the transfer entropy to identify information circuits in complex systems. Physical Review E. 2012; 86(6):066211. https://doi.org/10.1103/PhysRevE.86.066211
13. Wibral M, Vicente R, Lizier JT. Directed information measures in neuroscience. Springer; 2014.
14. Timme NM, Lapish C. A tutorial for information theory in neuroscience. eNeuro. 2018; 5(3). https://doi.org/10.1523/ENEURO.0052-18.2018 PMID: 30211307
15. Palmigiano A, Geisel T, Wolf F, Battaglia D. Flexible information routing by transient synchrony. Nature Neuroscience. 2017; 20(7):1014. https://doi.org/10.1038/nn.4569
16. Lungarella M, Sporns O. Mapping information flow in sensorimotor networks. PLoS Computational Biology. 2006; 2(10):e144. https://doi.org/10.1371/journal.pcbi.0020144
17. Wibral M, Pampu N, Priesemann V, Siebenhühner F, Seiwert H, Lindner M, et al. Measuring information-transfer delays. PLoS One. 2013; 8(2):e55809. https://doi.org/10.1371/journal.pone.0055809 PMID: 23468850
18. Wibral M, Rahm B, Rieder M, Lindner M, Vicente R, Kaiser J. Transfer entropy in magnetoencephalographic data: quantifying information flow in cortical and cerebellar networks. Progress in biophysics and molecular biology. 2011; 105(1-2):80–97. https://doi.org/10.1016/j.pbiomolbio.2010.11.006
19. Brodski-Guerniero A, Naumer MJ, Moliadze V, Chan J, Althen H, Ferreira-Santos F, et al. Predictable information in neural signals during resting state is reduced in autism spectrum disorder. Human Brain Mapping. 2018; 39(8):3227–3240. https://doi.org/10.1002/hbm.24072 PMID: 29617056

20. Vakorin VA, Kovacevic N, McIntosh AR. Exploring transient transfer entropy based on a group-wise ICA decomposition of EEG data. Neuroimage. 2010; 49(2):1593–1600. https://doi.org/10.1016/j.neuroimage.2009.08.027

21. Lizier JT, Heinzle J, Horstmann A, Haynes JD, Prokopenko M. Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity. Journal of Computational Neuroscience. 2011; 30(1):85–107. https://doi.org/10.1007/s10827-010-0271-2

22. Wibral M, Finn C, Wollstadt P, Lizier J, Priesemann V. Quantifying information modification in developing neural networks via partial information decomposition. Entropy. 2017; 19(9):494. https://doi.org/10.3390/e19090494

23. Li M, Han Y, Aburn MJ, Breakspear M, Poldrack RA, Shine JM, et al. Transitions in brain-network level information processing dynamics are driven by alterations in neural gain. PloS Computational Biology. 2019; 15:e1006957. https://doi.org/10.1371/journal.pcbi.1006957 PMID: 31613882

24. Thivierge JP. Scale-free and economical features of functional connectivity in neuronal networks. Physical Review E. 2014; 90(2):022721. https://doi.org/10.1103/PhysRevE.90.022721

25. Harris JJ, Engl E, Attwell D, Jolivet RB. Energy-efficient information transfer at thalamocortical synapses. PLoS Computational Biology. 2019; 15(8):e1007226. https://doi.org/10.1371/journal.pcbi.1007226

26. Timme NM, Ito S, Myroshnychenko M, Nigam S, Shimono M, Yeh FC, et al. High-degree neurons feed cortical computations. PLoS Computational Biology. 2016; 12(5):e1004858. https://doi.org/10.1371/journal.pcbi.1004858 PMID: 27159884

27. Timme N, Ito S, Myroshnychenko M, Yeh FC, Hiolski E, Litke AM, et al. Multiplex networks of cortical and hippocampal neurons revealed at different timescales. BMC Neuroscience. 2014; 15(1):P212. https://doi.org/10.1371/journal.pone.0115764 PMID: 25536059

28. Schroeder KE, Irwin ZT, Gaidica M, Bentley JN, Patil PG, Mashour GA, et al. Disruption of corticocortical information transfer during ketamine anesthesia in the primate brain. Neuroimage. 2016; 134:459–465. https://doi.org/10.1016/j.neuroimage.2016.04.039 PMID: 27095309

29. Kobayashi R, Kitano K. Impact of network topology on inference of synaptic connectivity from multi-neuronal spike data simulated by a large-scale cortical network model. Journal of Computational Neuroscience. 2013; 35(1):109–124. https://doi.org/10.1007/s10827-013-0443-y

30. Nigam S, Shimono M, Ito S, Yeh FC, Timme N, Myroshnychenko M, et al. Rich-club organization in effective connectivity among cortical neurons. Journal of Neuroscience. 2016; 36(3):670–684. https://doi.org/10.1523/JNEUROSCI.2177-15.2016 PMID: 26791200

31. Ito S, Hansen ME, Heiland R, Lumsdaine A, Litke AM, Beggs JM. Extending transfer entropy improves identification of effective connectivity in a spiking cortical network model. PLoS One. 2011; 6(11):e27431. https://doi.org/10.1371/journal.pone.0027431

32. Neymotin SA, Jacobs KM, Fenton AA, Lytton WW. Synaptic information transfer in computer models of neocortical columns. Journal of Computational Neuroscience. 2011; 30(1):69–84. https://doi.org/10.1007/s10827-010-0253-4

33. Gourévitch B, Eggermont JJ. Evaluating information transfer between auditory cortical neurons. Journal of neurophysiology. 2007; 97(3):2533–2543. https://doi.org/10.1152/jn.01106.2006

34. Buehlmann A, Deco G. Optimal information transfer in the cortex through synchronization. PLoS Computational Biology. 2010; 6(9):e1000934. https://doi.org/10.1371/journal.pcbi.1000934

35. Ver Steeg G, Galstyan A. Information transfer in social media. In: Proceedings of the 21st International Conference on World Wide Web; 2012. p. 509–518.

36. Orlandi JG, Stetter O, Soriano J, Geisel T, Battaglia D. Transfer entropy reconstruction and labeling of neuronal connections from simulated calcium imaging. PLoS One. 2014; 9(6):e98842. https://doi.org/10.1371/journal.pone.0098842

37. de Abril IM, Yoshimoto J, Doya K. Connectivity inference from neural recording data: Challenges, mathematical bases and research directions. Neural Networks. 2018; 102:120–137. https://doi.org/10.1016/j.neunet.2018.02.016

38. Wasserman L. All of nonparametric statistics. Springer Science & Business Media; 2006.

39. Weisenburger S, Vaziri A. A guide to emerging technologies for large-scale and whole-brain optical imaging of neuronal activity. Annual Review of Neuroscience. 2018; 41:431–452. https://doi.org/10.1146/annurev-neuro-072116-031458

40. Obien MEJ, Deligkaris K, Bullmann T, Bakkum DJ, Frey U. Revealing neuronal function through microelectrode array recordings. Frontiers in Neuroscience. 2015; 8:423. https://doi.org/10.3389/fnins.2014.00423

41. Nemenman I, Lewen GD, Bialek W, Van Steveninck RRDR. Neural coding of natural stimuli: information at sub-millisecond resolution. PLoS Computational Biology. 2008; 4(3):e1000025. https://doi.org/10.1371/journal.pcbi.1000025

42. Kayser C, Logothetis NK, Panzeri S. Millisecond encoding precision of auditory cortex neurons. Proceedings of the National Academy of Sciences. 2010; 107(39):16976–16981. https://doi.org/10.1073/pnas.1012656107

43. Sober SJ, Sponberg S, Nemenman I, Ting LH. Millisecond spike timing codes for motor control. Trends in Neurosciences. 2018; 41(10):644–648. https://doi.org/10.1016/j.tins.2018.08.010

44. Garcia-Lazaro JA, Belliveau LA, Lesica NA. Independent population coding of speech with sub-millisecond precision. Journal of Neuroscience. 2013; 33(49):19362–19372. https://doi.org/10.1523/JNEUROSCI.3711-13.2013

45. Aldridge JW, Gilman S. The temporal structure of spike trains in the primate basal ganglia: afferent regulation of bursting demonstrated with precentral cerebral cortical ablation. Brain Research. 1991; 543 (1):123–138. https://doi.org/10.1016/0006-8993(91)91055-6

46. Cooper JN, Edgar CD. Transfer Entropy in Continuous Time. arXiv preprint arXiv:190506406. 2019;.

47. Doran G, Muandet K, Zhang K, Schölkopf B. A Permutation-Based Kernel Conditional Independence Test. In: 30th Conference on Uncertainty in Artificial Intelligence (UAI 2014). AUAI Press; 2014. p. 132–141.

48. Runge J. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In: International Conference on Artificial Intelligence and Statistics. PMLR; 2018. p. 938–947.

49. Marder E, Bucher D. Understanding circuit dynamics using the stomatogastric nervous system of lobsters and crabs. Annu Rev Physiol. 2007; 69:291–316. https://doi.org/10.1146/annurev.physiol.69.031905.161516

50. Novelli L, Lizier JT. Inferring network properties from time series using transfer entropy and mutual information: validation of multivariate versus bivariate approaches. Network Neuroscience. 2020;(Just Accepted):1–52. https://doi.org/10.1162/netn_a_00178

51. Das A, Fiete IR. Systematic errors in connectivity inferred from activity in strongly recurrent networks. Nature Neuroscience. 2020; p. 1–11.

52. Wollstadt P, Lizier J, Vicente R, Finn C, Martinez-Zarzuela M, Mediano P, et al. IDTxl: The Information Dynamics Toolkit xl: a Python package for the efficient analysis of multivariate information dynamics in networks. The Journal of Open Source Software. 2019; 4(34):1081. https://doi.org/10.21105/joss.01081

53. Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Physical review E. 2004; 69 (6):066138. https://doi.org/10.1103/PhysRevE.69.066138

54. Lewis PW, Shedler GS. Simulation of nonhomogeneous Poisson processes by thinning. Naval Research Logistics Quarterly. 1979; 26(3):403–413. https://doi.org/10.1002/nav.3800260304

55. Spirtes P, Glymour CN, Scheines R, Heckerman D. Causation, prediction, and search. MIT Press; 2000.

56. Peters J, Janzing D, Schölkopf B. Elements of causal inference: foundations and learning algorithms. MIT Press; 2017.

57. Runge J. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. Chaos: An Interdisciplinary Journal of Nonlinear Science. 2018; 28(7):075310. https://doi.org/10.1063/1.5025050

58. Buzsaki G. Rhythms of the Brain. Oxford University Press; 2006.

59. Riehle A, Grün S, Diesmann M, Aertsen A. Spike synchronization and rate modulation differentially involved in motor cortical function. Science. 1997; 278(5345):1950–1953. https://doi.org/10.1126/science.278.5345.1950

60. Maeda E, Robinson H, Kawana A. The mechanisms of generation and propagation of synchronized bursting in developing networks of cortical neurons. Journal of Neuroscience. 1995; 15(10):6834–6845. https://doi.org/10.1523/JNEUROSCI.15-10-06834.1995

61. Litwin-Kumar A, Oswald AMM, Urban NN, Doiron B. Balanced synaptic input shapes the correlation between neural spike trains. PLoS Computational Biology. 2011; 7(12). https://doi.org/10.1371/journal.pcbi.1002305 PMID: 22215995

62. Trong PK, Rieke F. Origin of correlated activity between parasol retinal ganglion cells. Nature Neuroscience. 2008; 11(11):1343. https://doi.org/10.1038/nn.2199

63. Marschinski R, Kantz H. Analysing the information flow between financial time series. The European Physical Journal B. 2002; 30(2):275–281. https://doi.org/10.1140/epjb/e2002-00379-2

64. Tsamardinos I, Borboudakis G. Permutation testing improves Bayesian network learning. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer; 2010. p. 322–337.

65. Glymour C, Zhang K, Spirtes P. Review of causal discovery methods based on graphical models. Frontiers in Genetics. 2019; 10:524. https://doi.org/10.3389/fgene.2019.00524

66. Burkitt AN. A review of the integrate-and-fire neuron model: I. Homogeneous synaptic input. Biological Cybernetics. 2006; 95(1):1–19. https://doi.org/10.1007/s00422-006-0068-6

67. Zalesky A, Fornito A, Cocchi L, Gollo LL, van den Heuvel MP, Breakspear M. Connectome sensitivity or specificity: which is more important? Neuroimage. 2016; 142:407–420.

68. Kispersky T, Gutierrez GJ, Marder E. Functional connectivity in a rhythmic inhibitory circuit using Granger causality. Neural Systems & Circuits. 2011; 1(1):9. https://doi.org/10.1186/2042-1001-1-9

69. Gerhard F, Kispersky T, Gutierrez GJ, Marder E, Kramer M, Eden U. Successful reconstruction of a physiological circuit with known connectivity from spiking activity alone. PLoS Computational Biology. 2013; 9(7):e1003138. https://doi.org/10.1371/journal.pcbi.1003138

70. Barnett L, Barrett AB, Seth AK. Granger causality and transfer entropy are equivalent for Gaussian variables. Physical Review Letters. 2009; 103(23):238701. https://doi.org/10.1103/PhysRevLett.103.238701

71. Selverston AI. Dynamic biological networks: the stomatogastric nervous system. MIT Press; 1992.

72. Marder E, Bucher D. Central pattern generators and the control of rhythmic movements. Current Biology. 2001; 11(23):R986–R996. https://doi.org/10.1016/S0960-9822(01)00581-4

73. Prinz AA, Bucher D, Marder E. Similar network activity from disparate circuit parameters. Nature Neuroscience. 2004; 7(12):1345. https://doi.org/10.1038/nn1352

74. O'Leary T, Williams AH, Franci A, Marder E. Cell types, network homeostasis, and pathological compensation from a biologically plausible ion channel expression model. Neuron. 2014; 82(4):809–821. https://doi.org/10.1016/j.neuron.2014.04.002

75. Wang Q, Kulkarni SR, Verdú S. Divergence estimation for multidimensional densities via *k*-nearest-neighbor distances. IEEE Transactions on Information Theory. 2009; 55(5):2392–2405. https://doi.org/10.1109/TIT.2009.2016060

76. Gao S, Ver Steeg G, Galstyan A. Efficient estimation of mutual information for strongly dependent variables. In: Artificial Intelligence and Statistics; 2015. p. 277–286.

77. Zaytsev YV, Morrison A, Deger M. Reconstruction of recurrent synaptic connectivity of thousands of neurons from simulated spiking activity. Journal of Computational Neuroscience. 2015; 39(1):77–103. https://doi.org/10.1007/s10827-015-0565-5

78. Ladenbauer J, McKenzie S, English DF, Hagens O, Ostojic S. Inferring and validating mechanistic models of neural microcircuits based on spike-train data. Nature Communications. 2019; 10(1):1–17. https://doi.org/10.1038/s41467-019-12572-0

79. Casadiego J, Maoutsa D, Timme M. Inferring network connectivity from event timing patterns. Physical Review Letters. 2018; 121(5):054101. https://doi.org/10.1103/PhysRevLett.121.054101

80. Rosenblum M, et al. Reconstructing networks of pulse-coupled oscillators from spike trains. Physical Review E. 2017; 96(1):012209. https://doi.org/10.1103/PhysRevE.96.012209 PMID: 29347231

81. Runge J, Heitzig J, Petoukhov V, Kurths J. Escaping the curse of dimensionality in estimating multivariate transfer entropy. Physical Review Letters. 2012; 108(25):258701. https://doi.org/10.1103/PhysRevLett.108.258701

82. Belghazi MI, Baratin A, Rajeshwar S, Ozair S, Bengio Y, Courville A, et al. Mutual information neural estimation. In: International Conference on Machine Learning. PMLR; 2018. p. 531–540.

83. Noshad M, Zeng Y, Hero AO. Scalable mutual information estimation using dependence graphs. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2019. p. 2962–2966.

84. Khan S, Bandyopadhyay S, Ganguly AR, Saigal S, Erickson DJ III, Protopopescu V, et al. Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. Physical Review E. 2007; 76(2):026209. https://doi.org/10.1103/PhysRevE.76.026209

85. Poole B, Ozair S, van den Oord A, Alemi AA, Tucker G. On variational lower bounds of mutual information. In: NeurIPS Workshop on Bayesian Deep Learning; 2018.

86. Beirlant J, Dudewicz EJ, Györfi L, Van der Meulen EC. Nonparametric entropy estimation: An overview. International Journal of Mathematical and Statistical Sciences. 1997; 6(1):17–39.

87. Kozachenko L, Leonenko NN. Sample estimate of the entropy of a random vector. Problemy Peredachi Informatsii. 1987; 23(2):9–16.

88. Berrett TB, Samworth RJ, Yuan M, et al. Efficient multivariate entropy estimation via *k*-nearest neighbour distances. The Annals of Statistics. 2019; 47(1):288–318. https://doi.org/10.1214/18-AOS1688

89. Moon K, Sricharan K, Greenewald K, Hero A. Ensemble estimation of information divergence. Entropy. 2018; 20(8):560. https://doi.org/10.3390/e20080560

90. Gao W, Oh S, Viswanath P. demystifying fixed *k*-nearest neighbor information estimators. IEEE Transactions on Information Theory. 2018; 64(8):5629–5661. https://doi.org/10.1109/TIT.2018.2807481

91. Leonenko N, Pronzato L, Vippal S. A class of Rényi information estimators for multidimensional densities. The Annals of Statistics. 2008; 36.5:2153–2182. https://doi.org/10.1214/07-AOS539

92. Theiler J. Estimating fractal dimension. Journal of the Optical Society of America A. 1990; 7(6):1055–1073. https://doi.org/10.1364/JOSAA.7.001055

93. Kantz H, Schreiber T. Nonlinear time series analysis. vol. 7. Cambridge University Press; 2004.

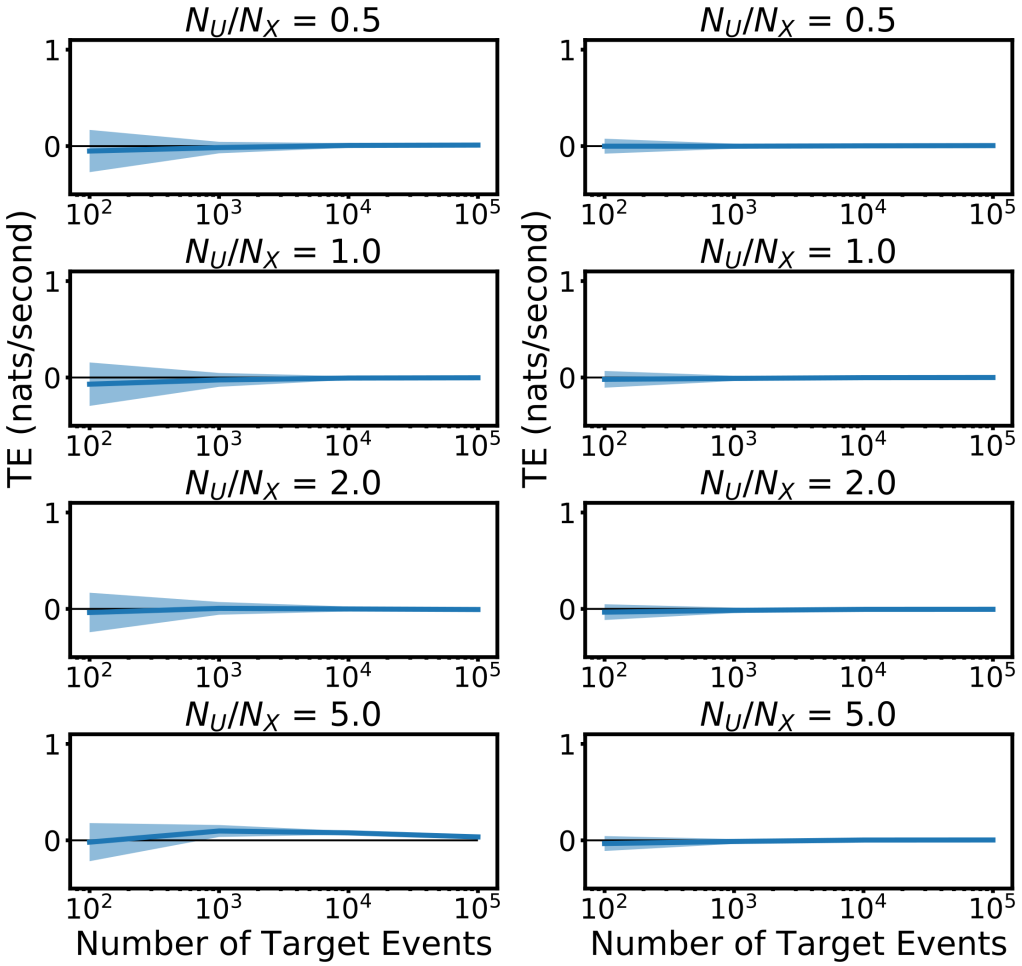
94. Lizier JT. JIDT: an information-theoretic toolkit for studying the dynamics of complex systems. Frontiers in Robotics and AI. 2014; 1:11. https://doi.org/10.3389/frobt.2014.00011

95. Cliff OM, Prokopenko M, Fitch R. An Information Criterion for Inferring Coupling of Distributed Dynamical Systems. Frontiers in Robotics and AI. 2016; 3:71. https://doi.org/10.3389/frobt.2016.00071

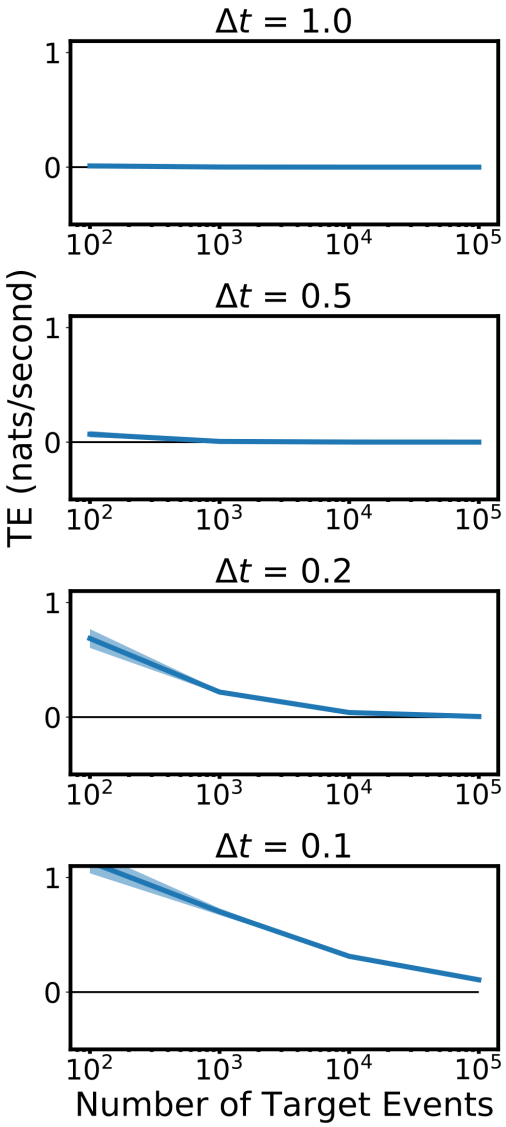
96. Ay N, Polani D. Information flows in causal networks. Advances in Complex Systems. 2008; 11(01):17–41. https://doi.org/10.1142/S0219525908001465

97. Janzing D, Balduzzi D, Grosse-Wentrup M, Schölkopf B, et al. Quantifying causal influences. The Annals of Statistics. 2013; 41(5):2324–2358. https://doi.org/10.1214/13-AOS1145

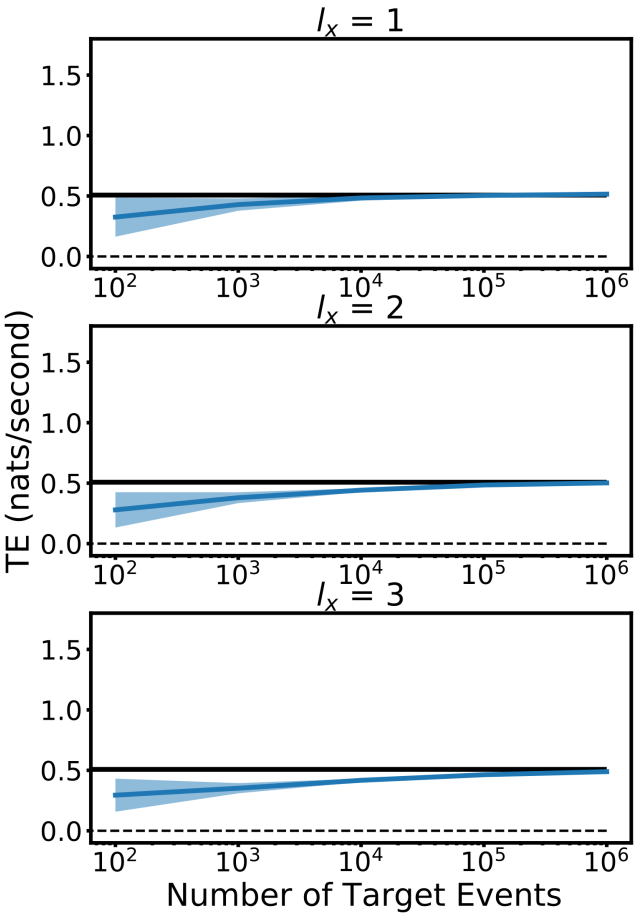
98. Meek C. Strong completeness and faithfulness in bayesian networks. In: Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence. UAI'95. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1995. p. 411–418.



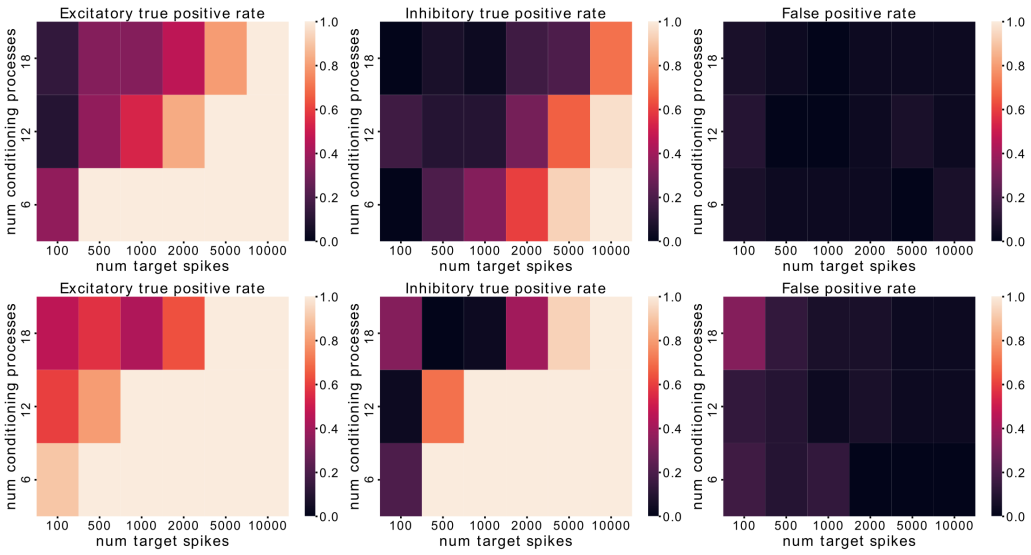
Supplementary figure 1



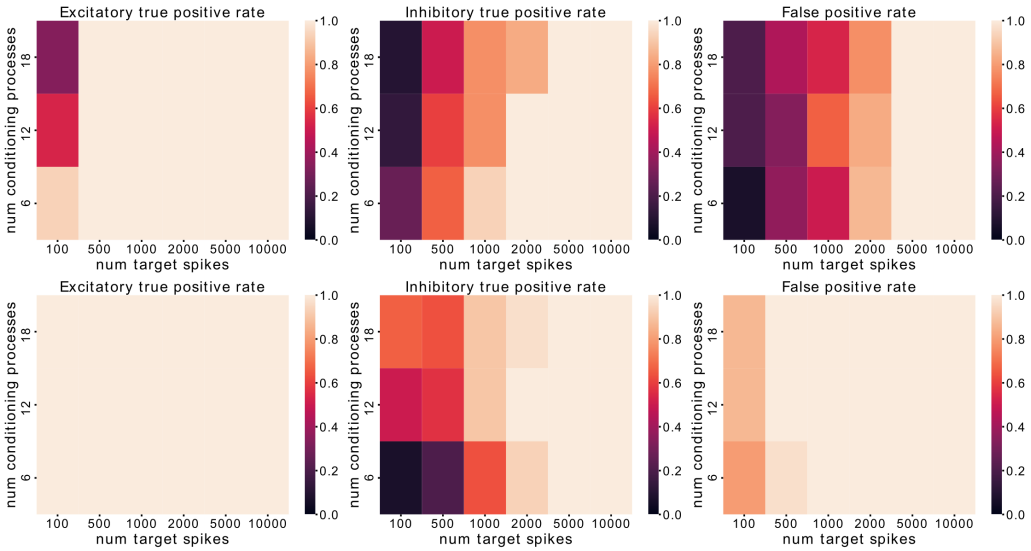
Supplementary figure 2



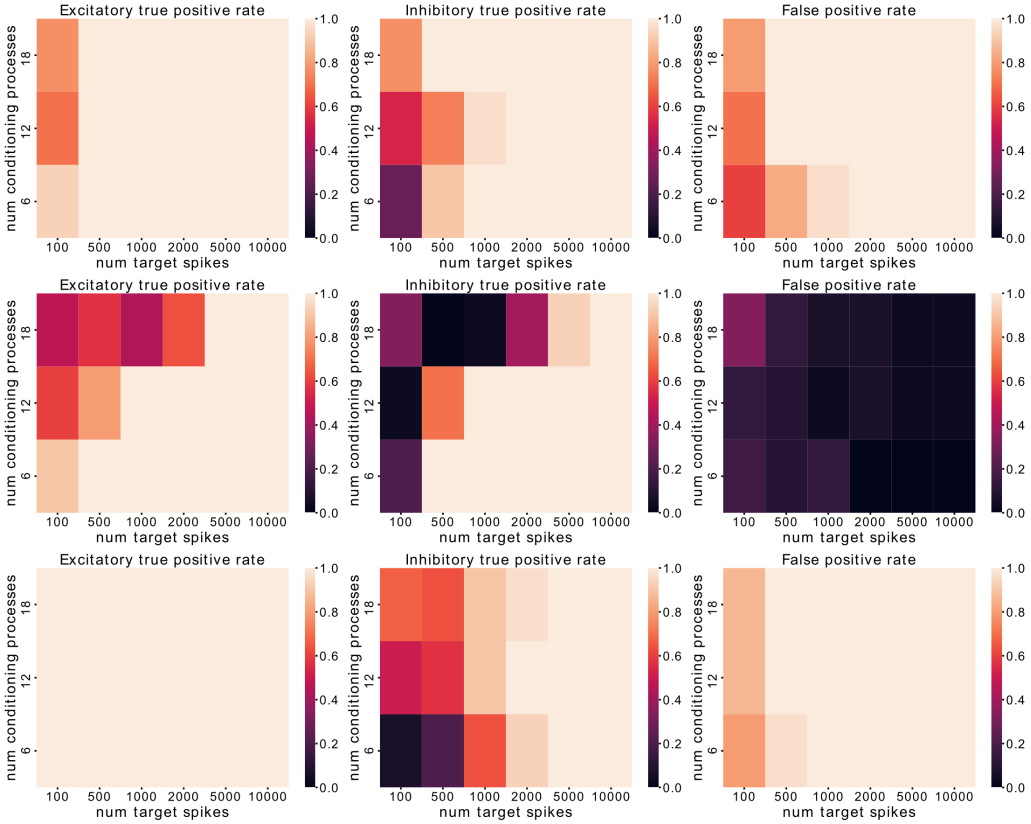
Supplementary figure 3



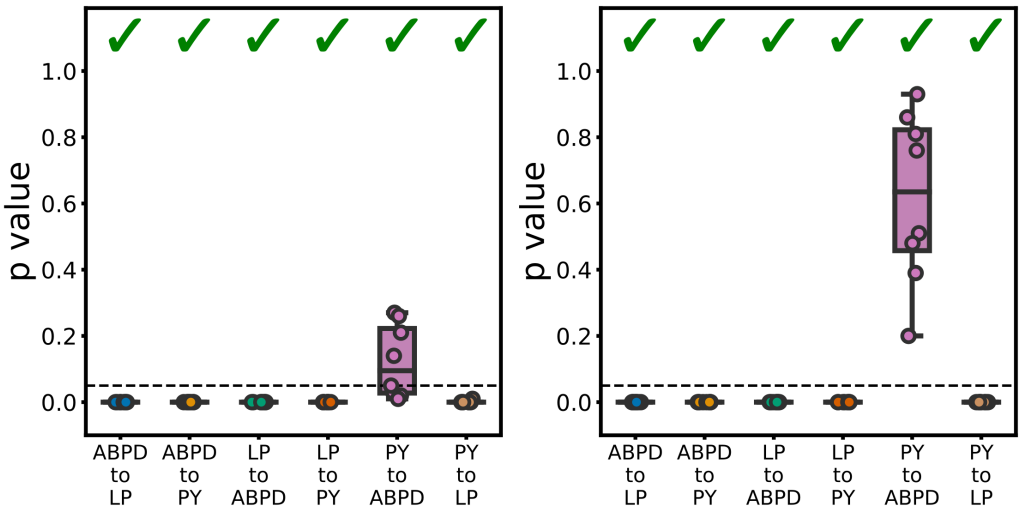
Supplementary figure 4



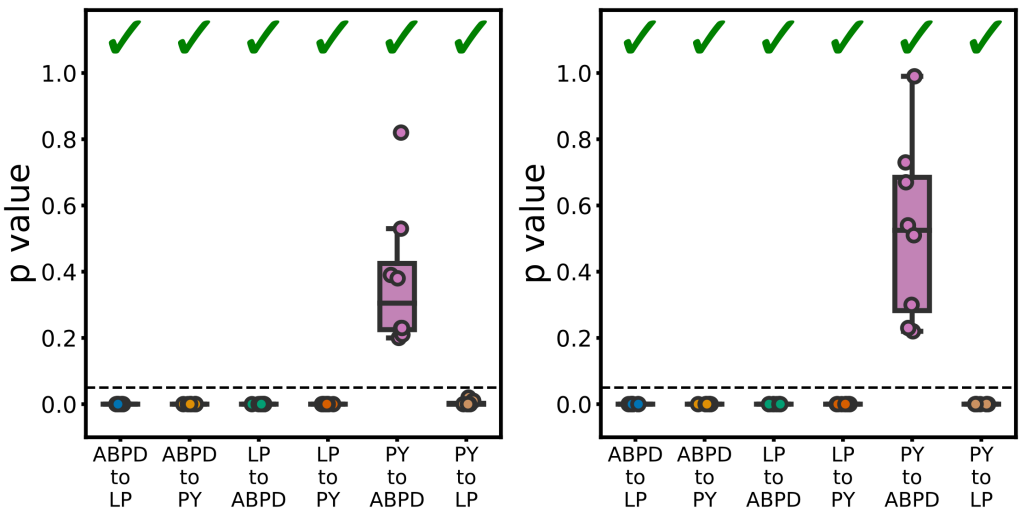
Supplementary figure 5



Supplementary figure 6



Supplementary figure 7



Supplementary figure 8

SUPPLEMENTARY TEXT S9 FOR
ESTIMATING TRANSFER ENTROPY IN CONTINUOUS TIME BETWEEN NEURAL SPIKE TRAINS OR OTHER EVENT-BASED DATA

David P. Shorten, Richard E. Spinney and Joseph T. Lizier

Specification of Biophysical Network Models Inspired by the Pyloric Circuit of the Crustacean Stomatogastric Ganglion

Current	E	p	q	m_∞	h_∞
I_{Na}	50	3	1	$\frac{1}{1+\exp\left(\frac{V+25.5}{-5.29}\right)}$	$\frac{1}{1+\exp\left(\frac{V+48.9}{5.18}\right)}$
I_{CaT}		3	1	$\frac{1}{1+\exp\left(\frac{V+27.1}{-7.2}\right)}$	$\frac{1}{1+\exp\left(\frac{V+32.1}{5.5}\right)}$
I_{CaS}		3	1	$\frac{1}{1+\exp\left(\frac{V+33}{-8.1}\right)}$	$\frac{1}{1+\exp\left(\frac{V+60}{6.2}\right)}$
I_A	-80	3	1	$\frac{1}{1+\exp\left(\frac{V+27.2}{-8.7}\right)}$	$\frac{1}{1+\exp\left(\frac{V+56.9}{4.9}\right)}$
I_{KCa}	-80	4	0	$\frac{[\text{Ca}]}{[\text{Ca}]+3}\frac{1}{1+\exp\left(\frac{V+28.3}{-12.6}\right)}$	
I_{Kd}	-80	4	0	$\frac{1}{1+\exp\left(\frac{V+12.3}{-11.8}\right)}$	
I_H	-20	1	0	$\frac{1}{1+\exp\left(\frac{V+75}{5.5}\right)}$	

Table 1: Parameters and functions used in the conductance based model.

We devised a simulation approach which follows very closely that presented in [1, 2, 3]. The only significant deviation is the addition of a noise term.

We modelled the neurons of the pyloric circuit using a conductance-based model. The membrane potential (V) evolves according to

$$C\frac{dV}{dt} = -\sum_i I_i - \sum_s I_s + \xi. \tag{1}$$

$C = 0.628\,\text{nF}$ is the membrane conductance. ξ is a noise term. Each current is specified by

$$I_i = g_i m_i^{q_i} h_i^{p_i}(V - E_i).$$

E_i is the reversal potential and its values are listed in table 1. The reversal potentials associated with the calcium channels are not listed as these are calculated according to the

Current	τ_m	τ_h
I_{Na}	$2.64 - \frac{2.52}{1+\exp\left(\frac{V+120}{-25}\right)}$	$\frac{1.34}{1+\exp\left(\frac{V+62.9}{-10}\right)}\left[1.5 - \frac{42.6}{1+\exp\left(\frac{V+34.9}{3.6}\right)}\right]$
I_{CaT}	$43.4 - \frac{42.6}{1+\exp\left(\frac{V+68.1}{-20.5}\right)}$	$210 - \frac{179.6}{1+\exp\left(\frac{V+55}{-16.9}\right)}$
I_{CaS}	$2.8 + \frac{14}{\exp\left(\frac{V+27}{10}\right)+\exp\left(\frac{V+70}{-13}\right)}$	$120 + \frac{300}{\exp\left(\frac{V+55}{9}\right)+\exp\left(\frac{V+65}{-16}\right)}$
I_A	$23.2 - \frac{20.8}{1+\exp\left(\frac{V+32.9}{-15.2}\right)}$	$77.2 - \frac{58.4}{1+\exp\left(\frac{V+38.9}{-26.5}\right)}$
I_{KCa}		$180.6 - \frac{150.2}{1+\exp\left(\frac{V+46}{-22.7}\right)}$
I_{Kd}		$14.4 - \frac{12.8}{1+\exp\left(\frac{V+28.3}{-19.2}\right)}$
I_H		$\frac{2}{\exp\left(\frac{V+169.7}{-11.6}\right)+\exp\left(\frac{V-26.7}{14.3}\right)}$

Table 2: Parameters and functions used in the conductance based model.

Nernst equation. Specifically, $E_{\mathrm{Ca}} = \frac{RT}{2F}\log_{10}\left(\frac{[\mathrm{Ca}^{2+}]_{\mathrm{ext}}}{[\mathrm{Ca}^{2+}]}\right)$ where $R = 8.314\,\mathrm{J\,K^{-1}\,mol^{-1}}$ is the universal gas constant, $T = 293.3\,\mathrm{K}$ is the temperature and $F = 96\,485.332\,12\,\mathrm{C\,mol^{-1}}$ is Faraday's constant. $[\mathrm{Ca}^{2+}]_{\mathrm{ext}} = 3\,\mathrm{mM}$ is the extracellular Ca^{2+} concentration and $[\mathrm{Ca}^{2+}]$ is the intracellular Ca^{2+} concentration. The intracellular Ca^{2+} concentration evolves according to

$$\tau_{\mathrm{Ca}}\frac{d[\mathrm{Ca}^{2+}]}{dt} = -f\left(I_{\mathrm{CaT}} + I_{\mathrm{CaS}}\right) - [\mathrm{Ca}^{2+}] + [\mathrm{Ca}^{2+}]_0.$$

$[\mathrm{Ca}^{2+}]_0 = 0.05\,\mu\mathrm{M}$ is the steady-state Ca^{2+} concentration, $f = 14.96\,\mu\mathrm{M\,nA^{-1}}$ and $\tau_{\mathrm{Ca}} = 200\,\mathrm{ms}$

The values of q_i and p_i are listed in table 1. The activation variables m_i evolve according to

$$\tau_{m_i}\frac{dm_i}{dt} = m_{\infty,i} - m_i.$$

The inactivation variables h_i evolve according to

$$\tau_{h_i}\frac{dh_i}{dt} = h_{\infty,i} - h_i.$$

$m_{\infty,i}$ and $h_{\infty,i}$ are given in table 1 and τ_{m_i} and τ_{h_i} are given in table 2.

The synaptic currents are specified by

$$I_s = g_s a(V - E_s).$$

E_s is the synaptic reversal potential. It was set to $-70\,\mathrm{mV}$ for glutamatergic synapses and $-80\,\mathrm{mV}$ for cholinergic synapses.

The activation variable a_s evolves according to

$$\tau_{a_s}\frac{da_s}{dt} = a_{\infty,s} - a_s$$

where

$$\tau_{a_s} = \frac{1 - a_s}{k_-}.$$

In the previous work which we are following [1, 2, 3], $a_{\infty,s}$ was a function of the presynaptic membrane potential. The functional form was such that the majority of current flow across the synapse was in the vicinity of spikes. However, the fact that all the influence over the synapse is not contained in the spikes opens the possibility for causal sufficiency not being met. As the principal purpose of this model in the context of this paper is to produce examples in which we have a known ground-truth of conditional dependence/independence we made a slight modification to the model to ensure that this condition was met. We specified that $a_{\infty,s}$ would evolve according to

$$\tau_{a_{\infty,s}}\frac{da_{\infty,s}}{dt} = 0 - a_{\infty,s}$$

where $\tau_{a_{\infty,s}} = 25\,\mathrm{ms}$. On the occurrence of a presynaptic spike $a_{\infty,s}$ was set to 0.99.

$\Delta = 5\,\mathrm{mV}$ provides the slope of the activation curve. $V_{\mathrm{th}} = -35\,\mathrm{mV}$ is the half activation potential of the synapse. V_{pre} is the membrane potential of the presynaptic neuron. k_- is the rate constant for the transmitter-receptor dissociation rate. For the glutamatergic synapses we used $k_- = 0.025\,\mathrm{ms}$ and for the cholinergic synapses we used $k_- = 0.01\,\mathrm{ms}$.

Simulations of the pyloric rhythm can be run with fixed maximum conductance values g_i and g_s, as in [2]. However, it was found that these models were less robust to the addition of a noise term. It was, therefore, decided to use adaptive conductances as described in [3]. Each conductance evolved according to:

$$\tau_g\frac{dg_j}{dt} = m_j - g_j$$

where

$$\tau_{m_j}\frac{dm_j}{dt} = [\mathrm{Ca}^{2+}] - \mathrm{Ca}_{\mathrm{tgt}}.$$

$\tau_g = 100\,\mathrm{ms}$ was common across all channels. The time constants τ_{m_j} are listed in table 3. The time constants provided in [3] were not used as these produce a pyloric rhythm with unrealistically short period. Instead, the approach presented in [3] for arriving at conductance time constants from desired conductance values was used. The conductance values in [2] were used as these desired values. Specifically, we used the values presented in table 2 in [2] for AB/PD 1, LP 2 and PY 1. The time constant τ_{m_j} associated with g_j was set as $\tau_{m_j} = c/g_j$. c is a constant with units of seconds that was adjusted by hand so that the activity converged in a reasonable amount of time and the time constants were of the same order of magnitude as those provided in [3]. As in [3], the leak conductances were fixed. They were set at 0 for the AB/PD neuron, and $0.0628\,\mu\mathrm{S}$ for the PY neuron and $0.1256\,\mu\mathrm{S}$ for the LP neuron.

Conductance	AB/PD	LP	PY
g_{Na}	0.25	1	1
g_{CaT}	40	1e15	40
g_{CaS}	16.67	25	40
g_{A}	2	5	2
g_{KCa}	10	20	1e15
g_{Kd}	1	4	0.8
g_{H}	1×10^4	2×10^3	2×10^3

Table 3: The conductance time constants τ_{m_j}. All values are in seconds.

LP $\rightarrow$ AB/PD, glutamatergic	2×10^4
AB/PD $\rightarrow$ LP, cholinergic	500
AB/PD $\rightarrow$ LP, glutamatergic	1×10^4
PY $\rightarrow$ LP, glutamatergic	1×10^4
AB/PD $\rightarrow$ PY, cholinergic	5×10^3
AB/PD $\rightarrow$ PY, glutamatergic	250
LP $\rightarrow$ PY, glutamatergic	1×10^6

Table 4: The conductance time constants τ_{m_j} for the synapses. All values are in seconds.

The TE approach to network inference will not work in a fully deterministic system. As such, noise was added to the system. There are a number of techniques for adding noise to conductance-based models [4]. The simplest such technique is to add noise to the currents in (1). Although this is not a biophysically realistic method, it has been shown to produce resulting behaviours which closely match those produced by more realistic techniques [4, 5]. As such, we decided to make use of this procedure in our simulations. The associated noise term is shown in (1). The noise was generated using an AR(1) process

$$\xi_t = \theta \xi_{t-1} + \epsilon_t. \tag{2}$$

We used the parameter value of $\theta = 0.005$. ϵ_t was distributed normally with mean 0 and standard deviation 9×10^{-9}

A simulation timestep of $\Delta t = 0.005\,\mathrm{ms}$ was used.

References

[1] Prinz AA, Billimoria CP, Marder E. Alternative to hand-tuning conductance-based models: construction and analysis of databases of model neurons. Journal of Neurophysiology. 2003;90(6):3998–4015.

[2] Prinz AA, Bucher D, Marder E. Similar network activity from disparate circuit parameters. Nature Neuroscience. 2004;7(12):1345.

[3] O'Leary T, Williams AH, Franci A, Marder E. Cell types, network homeostasis, and pathological compensation from a biologically plausible ion channel expression model. Neuron. 2014;82(4):809–821.

[4] Goldwyn JH, Shea-Brown E. The what and where of adding channel noise to the Hodgkin-Huxley equations. PLoS Computational Biology. 2011;7(11).

[5] Rowat P. Interspike interval statistics in the stochastic Hodgkin-Huxley model: Coexistence of gamma frequency bursts and highly irregular firing. Neural Computation. 2007;19(5):1215–1250.

Chapter 4

Early Lockin of Information Flows

Chapter 3 derived a new estimator for TE on spike trains which makes possible the study of information flow on this data with high fidelity. Specifically, TE can be estimated without any loss of time precision while still considering relatively long histories. Moreover, this new estimator allows for much greater confidence in the resulting estimates, due to its consistency property.

In this chapter, we make use of this new estimator to perform the first high-fidelity study of information flows on biological spike train recordings. We do so by applying it to recordings of the spike times of developing cultures of dissociated cortical rat neurons. These cultures were recorded on different days *in vitro*. This allows us to contrast the information flows on different days of development, producing the first study of how information flows change over the course of development in biological neural systems.

We study the emergence of distributed computation (that is, computation occurring over a network, as opposed to a Von Neumann architecture [1]) during the development of these cultures, where the development of the networks is being guided by the spontaneous activity of the neurons [2]. We are able to uncover a number of fascinating aspects of this emergence. Firstly, we find that there is a remarkable early lock-in phenomenon of the information flows. Flows between nodes early in development are highly correlated with flows later in development, indicating that the nature of the information transmission in the networks is determined early in development. We further find that certain nodes, specifically those that burst in the middle of the burst propagation, occupy the specialised role of the mediators of information flow, having a balance between both transmitting and receiving information. Moreover, these computational roles are also locked in early in development.

We also studied the information flows in simulated networks developing according to an STDP learning rule in order to confirm a putative mechanism for the observed phenomena. It was found that the changes in the information flows in these networks closely mirrored those in the biological cell cultures. Specifically, we observed that information flows locked in early and that middle bursters were occupying the specialised computational role of mediators of information flow.

In this chapter, information flows were studied for all pairs across the network. That is, a given source-target pair was considered in isolation. In Chapter 5, we will perform network inference using multivariate information flows. That is, we will consider extra conditioning processes other than just the source and target when inferring their relationship. This will allow us to arrive at minimal sets of sources which explain the activity of the target. It is important to note that, in both chapters, spike-sorting is not performed, and so we are considering multi-unit activity [3].

[1] F. Kuhn, N. Lynch, and R. Oshman, "Distributed computation in dynamic networks," in *Proceedings of the forty-second ACM symposium on Theory of computing*, 2010, pp. 513–522.

[2] L. A. Kirkby, G. S. Sack, A. Firl, and M. B. Feller, "A role for correlated spontaneous activity in the assembly of neural circuits," *Neuron*, vol. 80, no. 5, pp. 1129–1144, 2013.

[3] M. S. Schroeter, P. Charlesworth, M. G. Kitzbichler, O. Paulsen, and E. T. Bullmore, "Emergence of rich-club topology and coordinated dynamics in development of hippocampal functional networks in vitro," *Journal of Neuroscience*, vol. 35, no. 14, pp. 5459–5470, 2015.

RESEARCH ARTICLE

Early lock-in of structured and specialised information flows during neural development

David P Shorten[1]*, Viola Priesemann[2], Michael Wibral[3], Joseph T Lizier[1]

[1]Centre for Complex Systems, Faculty of Engineering, The University of Sydney, Sydney, Australia; [2]Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany; [3]Campus Institute for Dynamics of Biological Networks, Georg August University, Göttingen, Germany

Abstract The brains of many organisms are capable of complicated distributed computation underpinned by a highly advanced information processing capacity. Although substantial progress has been made towards characterising the information flow component of this capacity in mature brains, there is a distinct lack of work characterising its emergence during neural development. This lack of progress has been largely driven by the lack of effective estimators of information processing operations for spiking data. Here, we leverage recent advances in this estimation task in order to quantify the changes in transfer entropy during development. We do so by studying the changes in the intrinsic dynamics of the spontaneous activity of developing dissociated neural cell cultures. We find that the quantity of information flowing across these networks undergoes a dramatic increase across development. Moreover, the spatial structure of these flows exhibits a tendency to lock-in at the point when they arise. We also characterise the flow of information during the crucial periods of population bursts. We find that, during these bursts, nodes tend to undertake specialised computational roles as either transmitters, mediators, or receivers of information, with these roles tending to align with their average spike ordering. Further, we find that these roles are regularly locked-in when the information flows are established. Finally, we compare these results to information flows in a model network developing according to a spike-timing-dependent plasticity learning rule. Similar temporal patterns in the development of information flows were observed in these networks, hinting at the broader generality of these phenomena.

Editor's evaluation

This work analyses how meaningful connections develop in the nervous system. The authors study the dissociated neuronal cultures and find that the information processing connections develop after 5–10 days. The direction of the information flow is influenced by neuronal bursting properties: the early bursting neurons emerge as sources and late bursting neurons become sinks in the information flow.

***For correspondence:** david.shorten@sydney.edu.au

Competing interest: The authors declare that no competing interests exist.

Funding: See page 26

Preprinted: 30 June 2021
Received: 12 October 2021
Accepted: 13 March 2022
Published: 14 March 2022

Reviewing Editor: Tatyana O Sharpee, Salk Institute for Biological Studies, United States

Introduction

Throughout development, how do brains gain the ability to perform advanced computation? Given that the distributed computations carried out by brains require an intrinsic information processing capacity, it is of utmost importance to decipher the nature of the emergence of this capacity during development.

For brains to engage in the computations required for specific tasks, they require a general-purpose computational *capacity*. This capacity is often studied within the framework of information dynamics, where it is decomposed into the atomic operations of information storage, transfer, and modification (***Lizier et al., 2014***; ***Lizier, 2013***). We are particularly interested in the information flow component, which is measured using the transfer entropy (TE) (***Schreiber, 2000***; ***Bossomaier et al., 2016***). There exists a substantial body of work examining the structure and role of computational capacity in terms of these operations in mature brains. This includes the complex, dynamic, structure of information transfer revealed by calcium imaging (***Orlandi et al., 2014***), fMRI (***Mäki-Marttunen et al., 2013***; ***Lizier et al., 2011***), MEG (***Wibral et al., 2011***), and EEG (***Shovon et al., 2016***; ***Huang et al., 2015***; ***Stramaglia et al., 2012***; ***Marinazzo et al., 2014a***), and the role of information storage in representing visual stimuli (***Wibral et al., 2014a***), among others.

Given the established role of information flows in enabling the computations carried out by mature brains, we aim to study how they self-organise during neural development. There are a number of requirements for such a study. Firstly, it needs to be performed at a fine spatial scale (close to the order of individual neurons) to capture the details of development. It also needs to be conducted longitudinally in order to track changes over developmental timescales. Finally, the estimation of the information flow as measured by TE needs to be performed with a technique which is both accurate and able to capture the subtleties of computations performed on both fine and large timescales simultaneously.

Considering the first requirement of fine spatial scale, cell cultures plated over multi-electrode arrays (MEAs) allow us to record from individual neurons in a single network, providing us with this fine spatial resolution. There have been a number of previous studies examining information flows in neural cell cultures, for example, ***Nigam et al., 2016***; ***Shimono and Beggs, 2015***; ***Matsuda et al., 2013***; ***Timme et al., 2014***; ***Kajiwara et al., 2021***; ***Timme et al., 2016***; ***Wibral et al., 2017***. Such work has focussed on the directed functional networks implied by the estimated TE values between pairs of nodes, which has revealed interesting features of the information flow structure. See section 'Previous application of the discrete-time estimator' for a more detailed description of this previous work.

However, moving to our second requirement of a longitudinal study, these studies have almost exclusively examined only single points in neural development since nearly all of them examined recordings from slice cultures of mature networks. By contrast, we aim to study the information flows longitudinally by estimating them at different stages in development. Using recordings from developing cultures of dissociated neurons (***Wagenaar et al., 2006b***) makes this possible.

In terms of our third requirement of accurate and high-fidelity estimation of TE, we note that all previous studies of information flows in neural cell cultures made use of the traditional discrete-time estimator of TE. As recently demonstrated (***Shorten et al., 2021***), the use of this estimator is problematic as it can only capture effects occurring on a single timescale. In contrast, a novel continuous-time TE estimator (***Shorten et al., 2021***) captures effects on multiple scales, avoiding time binning, is data efficient, and consistent. See section 'Transfer entropy estimation' for a more detailed discussion of the differences between the continuous-time and discrete-time estimators.

In this article, we thus examine the development of neural information flows for the first time, addressing the above requirements by applying the continuous-time TE estimator to recordings of developing dissociated cultures. We find that the amount of information flowing over these cultures undergoes a dramatic increase throughout development and that the patterns of these flows tend to be established when the flows arise. During bursting periods, we find that nodes often engage in specialised computational roles as either transmitters, receivers, or mediators of information flow. Moreover, these roles usually correspond with the node's mean position in the burst propagation, with middle bursters tending to be information mediators. This provides positive evidence for the pre-existing conjecture that nodes in the middle of the burst propagation play the vital computational role of 'brokers of neuronal communication' (***Schroeter et al., 2015***). Intriguingly, the designation of computational roles (transmitter, receiver, or mediator) appears to also be determined early when the information flows are established. Finally, in order to investigate the generality of these phenomena, as well as a putative mechanism for their emergence, we study the dynamics of information flow in a model network developing according to a spike-timing-dependent plasticity (STDP) (***Caporale and Dan, 2008***) update rule. We find that the abovementioned phenomena are present in this model system, hinting at the broader generality of such patterns of information flow in neural development.

Results

Data from overnight recordings of developing cultures of dissociated cortical rat neurons at various stages of development (designated by days in vitro [DIV]) was analysed. These recordings are part of an open, freely available, dataset (***Wagenaar et al., 2006b***; ***Network, 2021***). See 'Materials and methods' (section 'Cell culture data') for a summary of the setup that produced the recordings. We studied all cultures for which there were overnight recordings. We restricted our analysis to these overnight recordings as they provided sufficient data for the estimation of TE. In what follows, we refer to the cultures by the same naming convention used in the open dataset: 1-1 through 1-5 and 2-1 through 2-6. The majority of cultures have recordings at three different time points, three have recordings at four points (1-3, 2-2, and 2-5) and one has only two recordings (2-1). The days on which these recordings took place vary between the 4th and 33rd DIV. By contrasting the TE values estimated at these different recording days, we are able to obtain snapshots of the emergence of these culture's computational capacity.

In the analysis that follows, for space considerations, we show plots for four representative cultures: 1-1, 1-3, 2-2, and 2-5. The latter three were chosen as they were the only cultures with four recording days. Culture 1-1 was selected from the group with three recording days, having the latest final recording day that was no more than a week later than the penultimate recording day. Plots for the remaining cultures are shown in Appendix 3. We also display certain summary statistics for the results of all cultures in the main text. Culture 1-5 is anomalous (among the recordings studied in this work) in that on its final day it has ceased to burst regularly, as stated in Figure 3A of ***Wagenaar et al., 2006b***. This leads to its results being substantially different from those of the other cultures, as will be presented below.

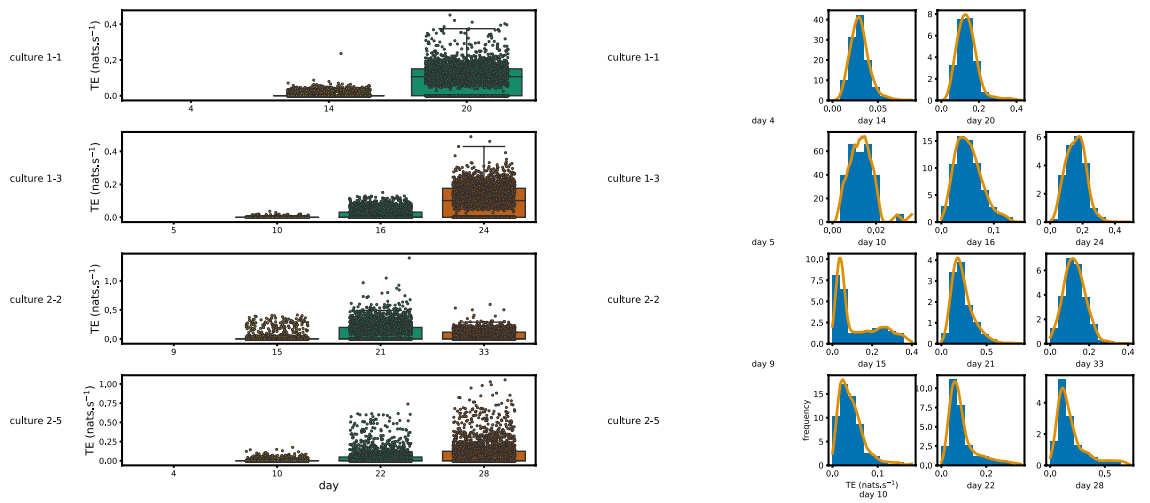
The TE between all pairs of electrodes was estimated using a recently introduced continuous-time estimator (***Shorten et al., 2021***; see section 'Transfer entropy estimation'). This produces a directed functional network at each recording day, and we aim to analyse how the connections in this network change over development time. Spike sorting was not performed because we would not be able to match the resulting neural units across different recordings and could not then fulfil our aim of contrasting the information flow between specific source-target pairs at different recording days. As such, the activity on each node in the directed functional networks we study is multi-unit activity (MUA) (***Schroeter et al., 2015***) formed of the spikes from all neurons detected by a given electrode, with connections representing information flows in the MUA. For more details on data preprocessing as well as the parameters used with the estimator, see 'Materials and methods'.

The dramatic increase in the flow of information during development

We first investigate how the amount of information flowing between the nodes changes over the lifespan of the cultures. ***Table 1*** shows the mean TE between all source-target pairs (***Appendix 3—table 1*** shows these values for the additional cultures). We observe that this mean value increases monotonically with the number of DIV, with only a single exception in the main cultures (a slight drop in the mean TE between days 21 and 33 of culture 2-2). We can make the same observation for the additional cultures, where the only drop is caused by day 5 of culture 2-4, a day still very early in development which had no significant TE values. We performed a two-sided Student's *t*-test

Table 1. Mean transfer entropy (TE) in nats per second between every source-target pair for each recording studied.

Culture 1-1	Day 4	Day 14	Day 20	
	0	0.060	0.097	
Culture 1-3	Day 5	Day 10	Day 16	Day 24
	0	2×10^{-4}	0.017	0.098
Culture 2-2	Day 9	Day 15	Day 21	Day 33
	0	0.015	0.11	0.057
Culture 2-5	Day 4	Day 10	Day 22	Day 28
	0	2×10^{-3}	0.037	0.082



(a) Scatters and boxplots of TE values. (b) Histograms and kernel density estimates of TE values.

Figure 1. Plots of the distributions of estimated transfer entropy (TE) values in the recordings analysed in this study. (**a**) Scatters of the TE values are overlaid on box plots. The box plots show the quartiles and the median (values greater than 10 SDs from the mean have been removed from both the box and scatter plots as outliers). (**b**) Density estimates of the nonzero (statistically significant) TE distribution on top of a histogram. The densities are estimated using a Gaussian kernel. The histogram bin width and kernel histogram are both 10% of the data range. Recordings with fewer than 10 statistically significant TE values are excluded.

for the difference in the mean between all pairs of recordings for each culture. All such differences (increases and decreases) were found to be statistically significant at $p<0.01$ with Bonferroni correction for multiple comparisons.

Overall, the magnitude of the increase in the mean TE is substantial. All the first recordings for the main cultures had a mean estimated TE of $0\,\text{nats.s}^{-1}$ (with no statistically significant transfer entropies measured as per section 'The emergence of functional information flow networks'). By contrast, all recordings beyond 20 DIV had a mean TE greater than $0.037\,\text{nats.s}^{-1}$.

Figure 1 shows scatter plots of the TE values in each recording laid over box-and-whisker plots (***Appendix 3—figure 1*** shows equivalent plots for the additional cultures). The large increase over time in the amount of information flowing over the networks is clearly visible in these plots. However, it is interesting to note that certain source-target pairs do have large information flows between them on early recording days even while the average remains very low.

Figure 1b shows histograms of the TE values estimated in each recording along with probability densities estimated using a Gaussian kernel (***Appendix 3—figure 1*** shows these for the additional runs). The distributions only include the nonzero (statistically significant) estimated TE values. Some of these distributions do, qualitatively, appear to be log-normal, in particular for later DIV. Moreover, previous studies have placed an emphasis on the observation of log-normal distributions of TE values in in vitro cultures of neurons (***Shimono and Beggs, 2015***; ***Nigam et al., 2016***). As such, we quantitatively analysed the distribution of the nonzero (statistically significant) estimated TE values in each individual recording. However, contrary to expectations, we found that these values were not well described by a log-normal distribution. It is worth noting that previous work which analysed the distribution of TE values in networks of spiking neurons was performed on organotypic cultures, as opposed to the dissociated cultures studied in this work. See Appendix 1 for further details and discussion.

The emergence of functional information flow networks

By considering each electrode as a node in a network, we can construct directed functional networks of information flow by assigning a directed edge between each source-target pair of electrodes with a statistically significant information flow. This results in weighted directed networks, the weight being provided by the TE value. Diagrams of these networks for the main cultures are shown in ***Figure 2***, and in ***Appendix 3—figure 3*** for the additional cultures. Note that, in all subsequent analysis presented in this article, a TE value of zero was assigned to all cases where the TE was not statistically significant.

We are able to notice a number of interesting spatiotemporal patterns in these diagrams. Firstly, the density (number of edges) of the networks increases over time. This is quantified in ***Table 2***, which shows the number of source-target pairs of electrodes for which a statistically significant nonzero TE value was estimated. In all the main cultures studied, the number of such pairs (and, therefore, the network density) increased by orders of magnitude over the life of the culture. For instance, in all four cultures, no statistically significant TE values are estimated on the first recording day. However, over 1000 source-target pairs have significant TE values between them on the final day of recording for each culture. ***Appendix 3—table 2*** shows the same values for the additional cultures. With a few exceptions (such as the abovementioned anomalous culture 1-5), the same relationship is observed. Note that the final recording day for culture 2-4 is relatively early (day 11), and so the low number of significant edges on this day is consistent with the other cultures.

We are, therefore, observing the networks moving from a state where no nodes are exchanging information, to one in which information is being transferred between a substantial proportion of the pairs of nodes (≈30% density of possible directed connections in most networks). Put another way, the functional networks are emerging from an unconnected state to a highly connected state containing the information flow structure that underpins the computational capacity of the network. This helps to explain the overall increase in information flow across the network reported in section 'The dramatic increase in the flow of information during development'.

We observe that the information flow (both incoming and outgoing) is spread somewhat evenly over the networks – in the sense that in the later, highly connected, recordings there are very few areas with neither incoming nor outgoing flow (the one notable exception to this is culture 2-6). A number of clear hubs (with particularly high either outgoing or incoming information flow) do stand out against this strong background information flow however. The strongest such hubs (with many high-TE edges) are all information sinks: they have low outgoing information flow, but receive high flow from a number of other nodes.

One can observe many instances in these diagrams where nodes have either very high incoming flow and very low outgoing flow, or very low incoming flow and very high outgoing flow. That is, they are taking on the roles of source (information-transmitting) hubs or target (information-receiving) hubs. Notable instances of information-receiving hubs include node 49 of day 16 of culture 1-3, node 42 of day 22 of culture 2-5, and node 5 of day 15 of culture 2-2 (see ***Figure 2*** for the node numbers used here). Notable examples of information transmitting hubs include node 28 of day 10 of culture 1-3 and nodes 18, 19, 22, and 30 of day 22 of culture 2-5. The specialist computational roles that nodes can take on will be studied in more detail quantitatively in section 'Information flows quantify computational role of burst position', with a particular focus on how this relates to the burst propagation.

It is possible to observe some notable instances whereby the information processing properties of a node remain remarkably similar across recording days. For example, nodes 55, 50, and 39 of culture 2-2 are outgoing hubs (with almost no incoming TE) on all four recording days. This offers us a tantalising hint that the information processing structure of these networks might be locked-in early in development, being reinforced as time progresses. Section 'Early lock-in of information flows' performs a quantitative analysis of this hypothesis.

Early lock-in of information flows

In the previous subsection, analysis of the directed functional networks of information flow suggested that the structure of the information processing capacity of the developing networks might be determined early in development and reinforced during the subsequent neuronal maturation.

In order to quantitatively investigate this hypothesis, we examine the relationships in the information flow from a given source to a given target between different recording days. That is, we are probing whether the amount of information flowing between a source and a target on an early day

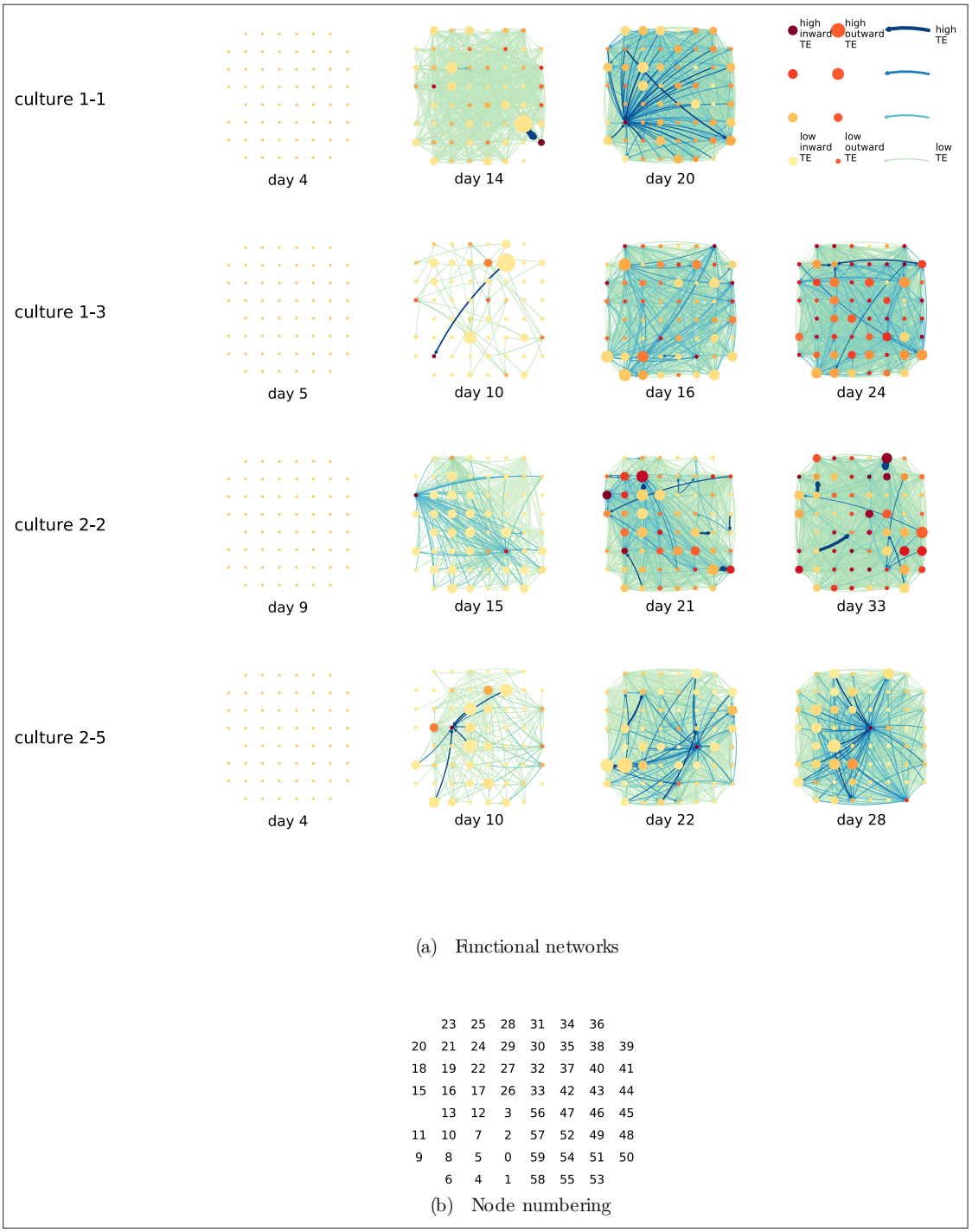


Figure 2. Functional networks overlaid on the spatial layout of the electrodes. (**a**) The directed functional networks implied by the estimated transfer entropy (TE) values. Each node represents an electrode in the original experimental setup. The nodes are spatially laid out according to their position in the recording array. An edge is present between nodes if there is a statistically significant information flow between them. The edge weight and colour are indicative of the amount of information flowing between electrodes (see the legend). The scaling of this weight and colour is done relative to the

Figure 2 continued on next page

Figure 2 continued

mean and variance of the information flow in each recording separately. The size and colour of the nodes are assigned relative to the total outgoing and incoming information flow on the node, respectively. As with the edge colour and size, this is done relative to the distribution of these values in each recording separately. (**b**) The spatial layout of the nodes. The numbering is identical to that used in the documentation of the open dataset studied in this work (***Wagenaar et al., 2006b***; ***Network, 2021***).

of development will be correlated with the amount flowing on a later day of development. This is equivalent to studying the correlation in the weights of the edges of the functional networks across different recording days. ***Figure 3*** shows scatter plots between the TE values estimated between each source-target pair on earlier and later days. Note that, in every case where the null hypothesis of zero TE could not be rejected (the TE was not statistically significant), a value of zero was used. Days with fewer than 10 nonzero values were excluded from the analysis as they could not lead to meaningful insights. By observing the pair scatters in ***Figure 3a–d*** (equivalent plots for the additional cultures are shown in ***Appendix 3—figure 4***), we see that, in many pairs of days, there appears to be a substantial correlation between the TE values on the edges across days. This is particularly pronounced for cultures 1-3 and 1-1, though visual assessment of the trend is complicated by the many zero values (where TE is not significant), gaps in the distribution and outliers. As such, ***Figure 3a–d*** also display the Spearman rank-order correlation (ρ) for each early-late pair of days for each culture. This correlation is positive and statistically significant at the $p<0.01$ level (after Bonferroni correction for multiple comparisons) between all the final pairs of recording days in the analysed cultures (including those in the additional cultures). ***Table 3*** summarises the proportions of pairs of recordings (including the additional cultures) which had significant positive Spearman correlations between the TE values on the edges across days. Whether we focus on either the final pairs of recordings or also include pairs that occur after day 15 (by which time the information flow networks have emerged), either all or all but one of the pairs of recordings exhibit such correlations. Moreover, the probability of this number of correlations arising by chance is found to be very low. This represents a strong tendency for the relatively strong information flows between a given source and target on later days to be associated with the relatively strong information flow between the same source and target on an earlier day of development. ***Figure 3*** displays these Spearman correlations between the early and late TE between source-target pairs visually. We notice a trend, whereby the correlation of the TE values seems to be higher between closer days (sample point being closer to the diagonal) and where those days are later in the development of the cultures (sample points being further to the right).

We also investigated the manner in which a node's tendency to be an information source hub might be bound once information flows are established. ***Figure 4*** shows scatter plots between the outgoing TE of each node (averaged across all targets) on different days of development along with the associated Spearman correlations (***Appendix 3—figure 5*** shows these plots for the additional cultures). By observing the scatter plots, it is easy to see that there is a strong positive relationship between the outgoing information flow from a given node on an earlier day of development and the outgoing

Table 2. The number of source-target pairs of electrodes with a statistically significant transfer entropy (TE) value between them for each recording studied.
This corresponds to the number of possible edges in the functional networks shown in ***Figure 2***. As the electrode arrays used to record the data had 59 electrodes, the total number of unique ordered pairs of electrodes (and, therefore, the number of possible edges) is 3422.

Culture 1-1	Day 4	Day 14	Day 20	
	0	607	2166	
Culture 1-3	Day 5	Day 10	Day 16	Day 24
	0	44	999	1902
Culture 2-2	Day 9	Day 15	Day 21	Day 33
	0	371	1409	1386
Culture 2-5	Day 4	Day 10	Day 22	Day 28
	0	185	975	1263

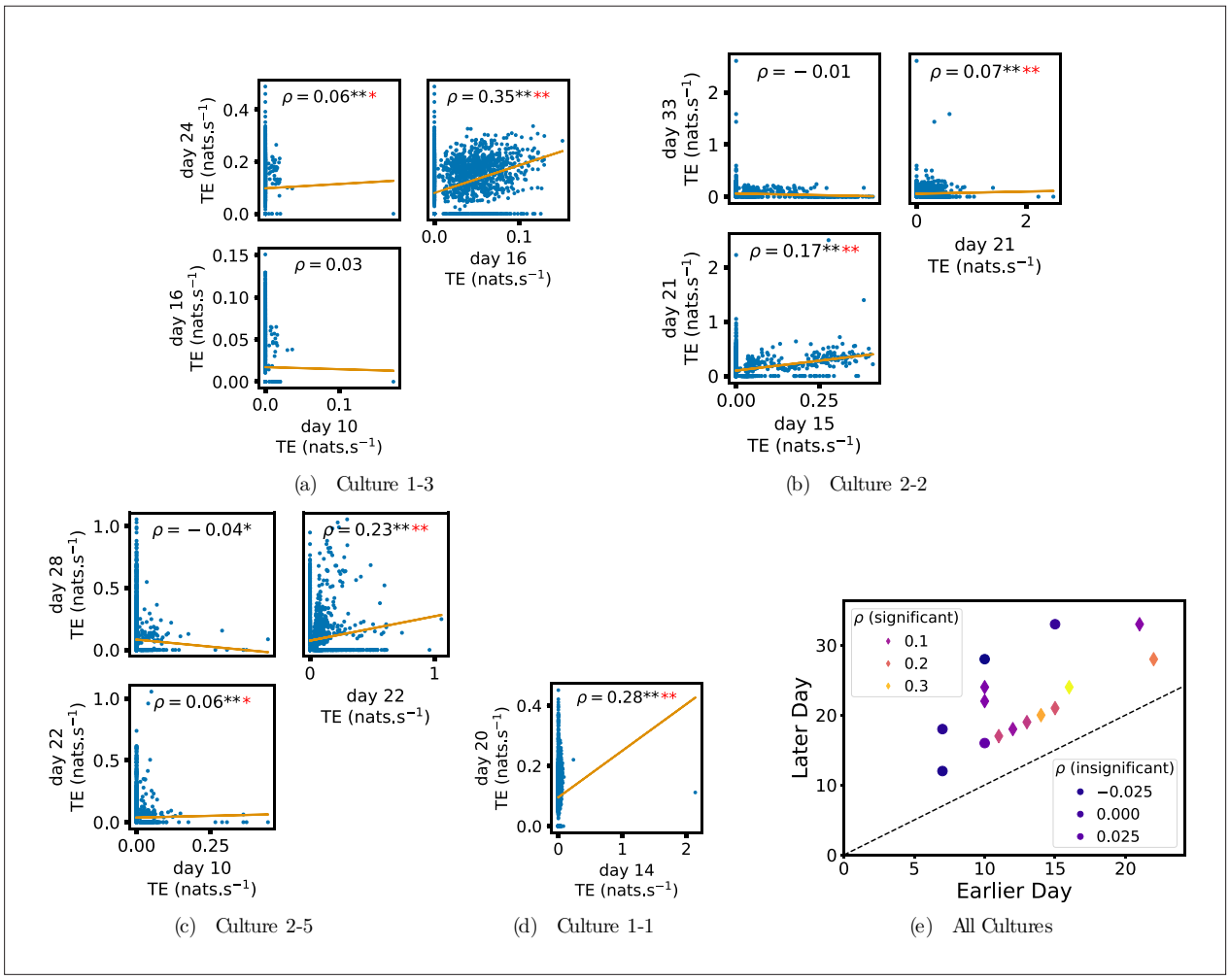


Figure 3. Plots investigating the relationship between the information flow on a given source-target pair over different days of development. (**a**–**d**) show scatter plots between all pairs of days for each culture (excluding days with less than 10 significant transfer entropy [TE] values). Specifically, in each scatter plot, the x value of a given point is the TE on the associated edge on an earlier day and the y value of that same point is the TE on the same edge but on a later day. The days in question are shown on the bottom and sides of the grids of scatter plots. The orange line shows the ordinary least-squares regression. The Spearman correlation (ρ) between the TE values on the two days is displayed in each plot. Values of ρ significant at the 0.05 level are designated with an asterisk and those significant at the 0.01 level are designated with a double asterisk. Red asterisks are used to denote significance after performing a Bonferroni correction for multiple comparisons. (**e**) shows all recording day pairs for all cultures (where the pairs are always from the same culture) and the associated Spearman correlation between the TE on the edges across this pair of recording days. Diamonds indicate significance at $p<0.05$, with Bonferroni correction.

flow from that same node on a later day (when we restrict ourselves to focusing on pairs occurring after a substantial number of statistically significant information flows have been established). This is not surprising, given the correlation we already established for TE on individual pairs, but does not automatically follow from that. As with the TE on the edges, ***Table 3*** summarises the proportions of pairs of recordings (including the additional cultures) which had significant positive Spearman correlations between the outgoing TE from each node across days. We see that, whether we focus on just the last recordings of each culture or also include those after day 15, a substantial majority of pairs of recordings exhibit such correlations and that such a majority would be very unlikely to arise by chance. Some of these correlations are particularly strong, and indeed stronger than that observed on the TEs

Table 3. Summary of significance values for the lock-in results.
Each table cell shows the number of relationships that were found to be significant at the $p<0.05$ level, out of the total number of relationships tested. Note that relationships were only tested in cases where both recordings in the pair had at least 10 significant transfer entropy (TE) values. A hypothesis test is conducted against the null hypothesis that the original p-values that produced these results were uniformly distributed between 0 and 1 (giving a 0.05 chance of a significant result). * indicates that this probability is less than 0.05. ** indicates that the probability of the observed number of significant results has probability less than 0.001 under this null hypothesis, with a Bonferroni correction for multiple comparisons. The first row summarises the significance values for the value of the TE on the edges, shown in ***Figure 3*** and ***Appendix 3—figure 4***. The second and third rows summarise the significance values for the mean outward and inward TE on each node, shown in ***Figure 4*** and ***Appendix 3—figure 5*** as well as ***Appendix 2—figure 1*** and ***Appendix 3—figure 6***, respectively. The final row summarises the significance values for the ratio of inward to outward burst-local TE, shown in ***Figure 6*** and ***Appendix 3—figure 9***. The columns which restrict the analysis to bursting recordings exclude the final recording of culture 1-5, as this culture ceased bursting, after having previously been bursty (***Wagenaar et al., 2006b***).

	All cultures, final pair	All cultures, final pair or post day 15	Bursting, final pair	Bursting, final pair or post day 15
Edge	7/7**	8/9**	6/6**	7/8**
Out	5/7**	6/9**	5/6**	6/8**
In	3/7*	3/9*	3/6*	3/8*
Burst-local ratios	5/7**	6/9**	5/6**	6/8**

of individual node pairs. For instance, between days 16 and 24 of culture 1-3 we have that $\rho = 0.62$ and between days 14 and 20 of culture 1-1 we have that $\rho = 0.71$. ***Figure 4*** visualises all Spearman correlations between the early and late total outgoing TE of a given node. As per the TEs for individual node pairs, the correlation is higher between closer days and where those days are later in the development of the cultures.

Appendix 2—figure 1 shows similar plots to ***Figure 4***, but for the average inward TE on each node (with ***Appendix 3—figure 6*** showing plots for the additional cultures). We observe three cases of significant correlations in this value between early and late days, indicating a weaker yet still statistically significant (***Table 3***) propensity for the average inward TE to also lock-in.

Of course, some pairs involving earlier days of some cultures (such as day 10 of cultures 1-3 and 2-5) do not exhibit such lock-in tendencies. However, as displayed in ***Table 2***, there are very few significant information flows at this early stage of development (44 and 185, respectively). This represents a point in development perhaps too early for any substantial information flow networks to have emerged.

In summary, the data suggests that, in these developing neural cell cultures, the structure of the information flows is to a large degree locked-in early in development, around the point at which the information dynamics emerge. There is a strong tendency for properties of these flows on later days to be correlated with those same properties on earlier days. Specifically, we have looked at the flows between source-target pairs, the average outgoing flow from a source, and the average incoming flow to a target. The values of these variables on later DIV were found, in the majority of cases, to be positively correlated with the same values on earlier DIV. Further, there were no cases where a statistically significant negative correlation was found.

Information flows quantify computational role of burst position

Developing cultures of dissociated neurons have a tendency to self-organise so as to produce population bursts or avalanches (***Wagenaar et al., 2006b***; ***Pasquale et al., 2008***). Such spike avalanches are not only a feature of cell cultures, being a ubiquitous feature of in vivo neural recordings (***Priesemann et al., 2014***; ***Priesemann et al., 2013***; ***Priesemann et al., 2009***). There is a wide body of work discussing the potential computational importance of such periods of neuronal activity (***Lisman, 1997***; ***Krahe and Gabbiani, 2004***; ***Shew et al., 2011***; ***Kinouchi and Copelli, 2006***; ***Haldeman and Beggs, 2005***; ***Rubinov et al., 2011***; ***Cramer et al., 2020***). It has been observed that cultures often follow one

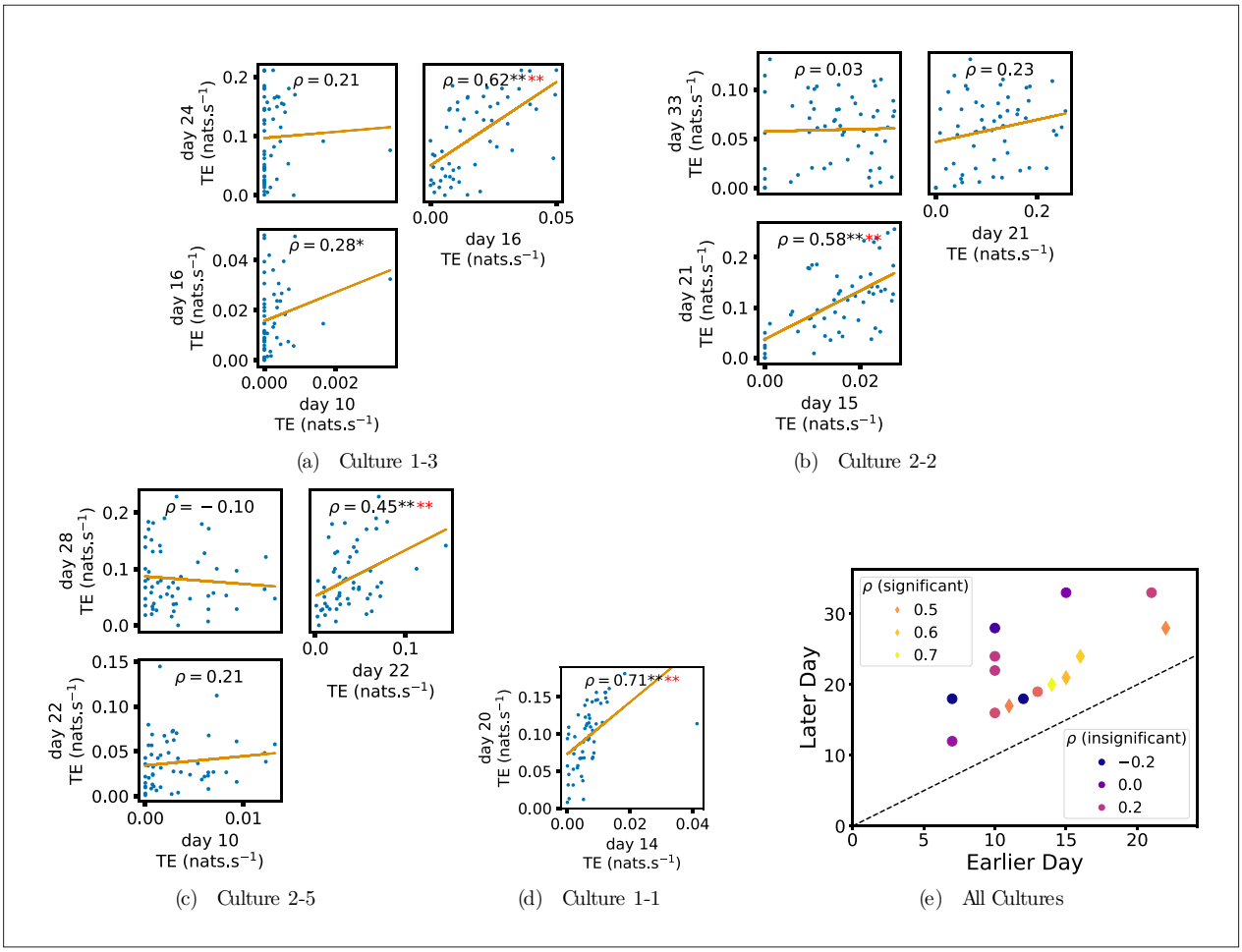


Figure 4. Plots investigating the relationship between the outward information flow from a given node over different days of development. (**a–d**) show scatter plots between all pairs of days for each culture (excluding days with less than 10 significant transfer entropy [TE] values). Specifically, in each scatter plot, the x value of a given point is the average outgoing TE from the associated node on an earlier day and the y value of that same point is the total outgoing TE from the same node but on a later day. The days in question are shown on the bottom and sides of the grids of scatter plots. The orange line shows the ordinary least-squares regression. The Spearman correlation (ρ) between the outgoing TE values on the two days is displayed in each plot. Values of ρ significant at the 0.05 level are designated with an asterisk and those significant at the 0.01 level are designated with a double asterisk. Red asterisks are used to denote significance after performing a Bonferroni correction for multiple comparisons. (**e**) shows all recording day pairs for all cultures (where the pairs are always from the same culture) and the associated Spearman correlation between the outward TEs of nodes across this pair of recording days. Diamonds indicate significance at $p<0.05$, with Bonferroni correction.

or more ordered patterns of burst propagation (***Maeda et al., 1995***), with some nodes exhibiting a tendency to burst towards the beginning of these patterns and others towards their end (***Schroeter et al., 2015***). More recent work has proposed that the nodes which tend to burst at different points in these progressions play different computational roles (***Schroeter et al., 2015***). This work has placed special importance on those nodes which usually burst during the middle of the burst progression, conjecturing that they act as the 'brokers of neuronal communication'.

The framework of information dynamics is uniquely poised to illuminate the computational dynamics during population bursting as well as the different roles that might be played by various nodes during these bursts. This is due to its ability to analyse information processing *locally in time* (***Lizier, 2013***; ***Lizier et al., 2008***; ***Lizier, 2014***; ***Wibral et al., 2014b***), as well as directionally between information sources and targets via the asymmetry of TE. This allows us to isolate the information

processing taking place during population bursting activity. We can then determine the information processing roles undertaken by the different nodes and examine how this relates to their average position in the burst propagation.

We analyse the information flowing between nodes during population bursts by estimating the *burst-local* TE between nodes in each recording (i.e. averaging the TE rates only during bursting periods, using probability distribution functions estimated over the whole recordings; see section 'Estimation of burst-local TE'). We also measure the mean position of each node within bursts (with earlier bursting nodes having a lower numerical position; see section 'Analysis of population bursts'). Note that, although there is variability of the burst position across bursts, certain nodes have much lower or higher mean burst positions, indicating a strong tendency to burst earlier or later within the propagation.

Figure 5a and b show plots of the mean burst position plotted against the total inward (***Figure 5a***) and outward (***Figure 5b***) burst-local TE of each node. ***Appendix 3—figure 7*** shows these plots for the additional cultures. Plots are only shown for days where there were at least 10 statistically significant burst-local TE values. The Spearman correlation (ρ) between these variables is also displayed on the plots.

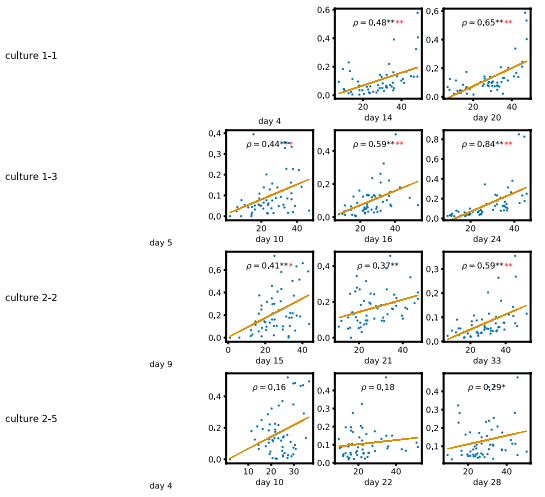
We see from ***Figure 5*** that, particularly for the final recording days, in most cases there is a positive correlation between the mean burst position of the node and the total inward burst-local TE. In some cases, this correlation is particularly strong. For instance, on the 24th DIV of culture 1-3, we observe a Spearman correlation of $\rho = 0.84$. In other words, we observe that nodes which tend to burst later have higher incoming information flows. ***Table 4*** summarises the proportions of recordings (including the additional cultures) which had significant positive Spearman correlations between the mean burst position of the node and the total inward burst-local TE. By focussing on recordings that have reached a state of established information dynamics, by either selecting all final recordings or all that were performed post day 15, we see that in all cases a clear majority of cases had a statistically significant positive Spearman correlation. Moreover, the probability of this number of correlations arising by chance is found to be very low. These relationships suggest that there is a tendency for the late bursters to occupy the specialised computational role of information receivers.

Conversely, as shown in ***Figure 5***, there is a tendency for the mean burst position of the nodes to be negatively correlated with the outward burst-local TE. Again, this correlation is particularly strong in many cases. For example, on the 24th DIV of culture 1-3, we observe a Spearman correlation of $\rho = -0.80$. In other words, we observe that nodes which tend to burst earlier have higher outward information flows. ***Table 4*** summarises the proportions of recordings (including the additional cultures) which had significant negative Spearman correlations between the mean burst position of the node and the total outward burst-local TE. We see that a clear majority of either all final recording days or all recordings performed post day 15 had a statistically significant negative Spearman correlation. Moreover, the probability of this number of correlations arising by chance is found to be very low. These relationships suggest that there is a tendency for the early bursters to occupy the specialised computational role of information receivers.

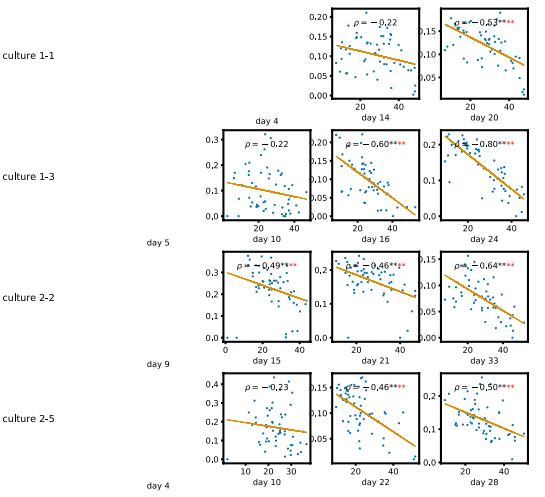
Figure 5 plots the total incoming burst-local TE on each node against the total outgoing burst-local TE, with points coloured according to the node's mean burst position (***Appendix 3—figure 8*** shows these plots for the additional cultures). We see a very clear pattern in these plots, which is remarkably clear on some later recording days: nodes which often fire at the beginning of the burst progression have high outgoing information flows with lower incoming flows, whereas those which tend to sit at the end of the progression have high incoming flows with lower outgoing flows. By contrast, those nodes which, on average, occupy the middle of the burst progression have a balance between outgoing and incoming information transfer. These nodes within the middle of the burst propagation are, therefore, occupying the suggested role of mediators of information flow.

Early lock-in of specialised computational roles

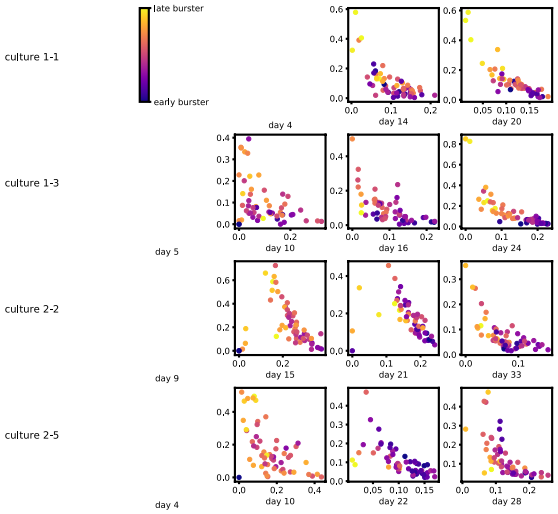
Given that we have seen in section 'Information flows quantify computational role of burst position' that nodes tend to occupy specialised computational roles based on their average position in the burst propagation and that we have seen in section 'Early lock-in of information flows' that information processing properties can lock-in early in development, it is worth asking whether the specialised



(a) Burst position vs TE in



(b) Burst position vs TE out



(c) Burst position relative to outgoing and incoming TE

Figure 5. The relationship between the amount of incoming and outgoing local (in burst) transfer entropy (TE) on a given node and its average burst position. (**a**) and (**b**) show the burst position of each node on the x axis of each plot, plotted against either the total incoming (**a**) or outgoing (**b**) TE on the node. The Spearman correlation (ρ) between the mean burst position and the incoming or outgoing TE values is displayed in each plot. Values of ρ significant at the 0.05 level are designated with an asterisk and those significant at the 0.01 level are designated with a double asterisk. Red asterisks are used to denote significance after performing a Bonferroni correction for multiple comparisons. (**e**) plots the outgoing TE on the x axis and the incoming TE on the y axis with the points coloured according to the mean burst position of the node: late bursters are coloured yellow and early bursters are purple.

Table 4. Summary of significance values for the results relating to computational roles.
Each table cell shows the number of relationships that were found to be significant at the $p<0.05$ level, out of the total number of relationships tested. A hypothesis test is conducted against the null hypothesis that the original p-values that produced these results were uniformly distributed between 0 and 1 (giving a 0.05 chance of a significant result). ** indicates that the probability of the observed number of significant results has probability less than 0.001 under this null hypothesis, with a Bonferroni correction for multiple comparisons. The first row summarises the significance values for the relationships between inward burst-local TE and burst position, as shown in ***Figure 5*** and ***Appendix 3—figure 7***. The second row summarises the significance values for the relationships between outward burst-local TE and burst position, as shown in ***Figure 5*** and ***Appendix 3—figure 7***.

	All cultures, final day	All cultures post day 15	Bursting, final day	Bursting, post day 15
In	7/11**	10/15**	7/10**	10/14**
Out	7/11**	11/15**	7/10**	11/14**

computational roles that nodes occupy during population bursts lock-in during the earlier stages of neuronal development.

In order to investigate this question, we quantified the computational role occupied by a node by measuring the proportion of its total incoming and outgoing burst-local TE that was made up by its outgoing burst-local TE. Scatters of these proportions between earlier and later DIV are plotted in ***Figure 6*** for the main cultures (the additional cultures are shown in ***Appendix 3—figure 9***). They also display the Spearman rank-order correlations (ρ) between the ratios on different days. We see that, in many cases, there are strong, significant, correlations in this ratio between earlier and later DIV. ***Table 3*** summarises the proportions of pairs of recordings (including the additional cultures) which had significant positive Spearman correlations between this ratio on each node across days. We see that, whether we focus on just the last recordings of each culture or also include those after day 15, a clear majority of pairs of recordings exhibit such correlations and that such a majority would be very unlikely to arise by chance. ***Figure 6*** visualises all these Spearman correlations between the early and late day pairs.

These results suggest that, if a node is an information transmitter during population bursts at the point at which the information flows are established, it has a tendency to maintain this specialised role later in development. Similarly, being an information receiver earlier in development increases the probability that the node will occupy this same role later.

Information flows in an STDP model of development

In order to investigate the generality of the phenomena revealed in this article, we reimplemented a model network (***Khoshkhou and Montakhab, 2019***) of Izhikevich neurons (***Izhikevich, 2003***) developing according to an STDP (***Caporale and Dan, 2008***) update rule as described in section 'Network of Izhikevich neurons'. For the low value of the synaptic time constant which we used (see section 'Network of Izhikevich neurons'), these networks developed from a state where each neuron underwent independent tonic spiking at a regular firing rate to one in which the dynamics were dominated by periodic population bursts (***Zeraati et al., 2021***; ***Khoshkhou and Montakhab, 2019***). It is worth noting that these population bursts are significantly more regular than those in the biological data used in this article. Small modifications were made to the original model in order that the development occurred over a greater length of time. The greater length of development allowed us to extract time windows which were short relative to the timescale of development (resulting in the dynamics being approximately stationary in these windows) yet still long enough to sample enough spikes for reliable TE rate estimation. The windows which we used resulted in a median of 5170 spikes per neuron per window compared with a median of 17,399 spikes per electrode in the biological data. See section 'Network of Izhikevich neurons' for more details on the modifications made. A single simulation was run. The dynamics of the model are very consistent across independent runs. Three windows were extracted, extending between the simulation timepoints of 200 and 250 seconds, 400 and 450 seconds, and 500 and 550 seconds. These time windows were labelled 'early', 'mid', and

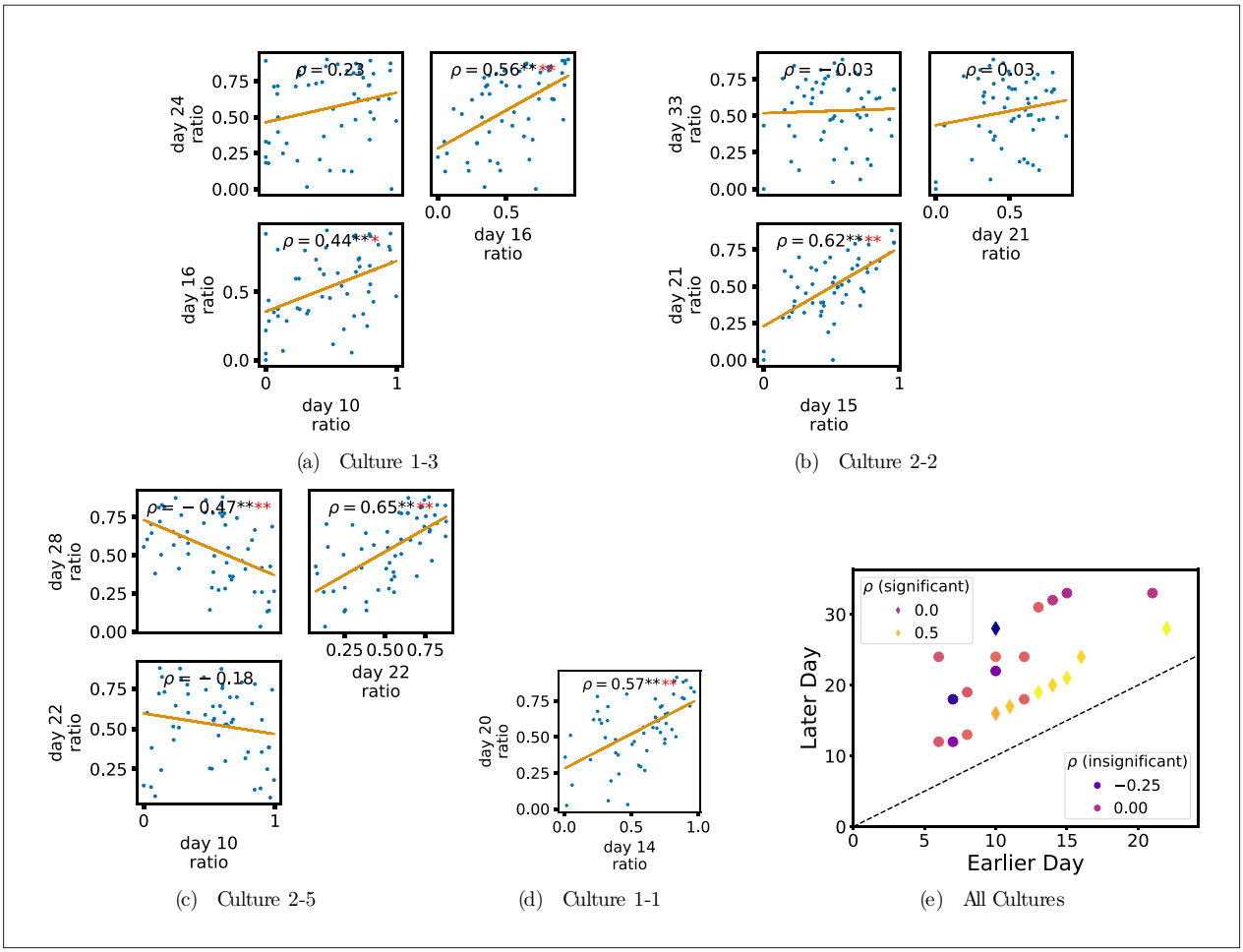


Figure 6. Plots investigating the relationship between the ratio of outward to total burst-local information flow from a given node over different days of development. (**a–d**) show scatter plots between all pairs of days for each culture (excluding days with less than 10 significant burst-local transfer entropy [TE] values). Specifically, in each scatter plot, the x value of a given point is the ratio of total outgoing burst-local TE on the associated node to the total burst-local TE on the same node on one day and the y value of that same point is this same ratio on the same node but on a different day. The days in question are shown on the bottom and sides of the grids of scatter plots. The orange line shows the ordinary least-squares regression. The Spearman correlation (ρ) between the TE values on the two days is displayed in each plot. Values of ρ significant at the 0.05 level are designated with an asterisk and those significant at the 0.01 level are designated with a double asterisk. Red asterisks are used to denote significance after performing a Bonferroni correction for multiple comparisons. (**e**) shows all recording day pairs for all cultures (where the pairs are always from the same culture) and the associated Spearman correlation between the outward TE of the nodes across this pair of recording days. Diamonds indicate significance at $p<0.05$, with Bonferroni correction.

'late', respectively. The early window was chosen such that it had a nonzero number of significant TE values, but such that this number was of the same (order of magnitude in) proportion as observed in the first recording days of the cell cultures (refer to ***Table 2***). The mid period was set at the point where population bursting begun to emerge, and the late period was set at the point where all neurons were bursting approximately synchronously in a pronounced manner.

TE values between all pairs of model neurons were estimated, as described in section 'Transfer entropy estimation'. These estimates were then subjected to the same statistical analysis as the cell culture data, the results of which are presented in the preceding subsections of this section. The plots of this analysis are displayed in ***Figures 7 and 8***.

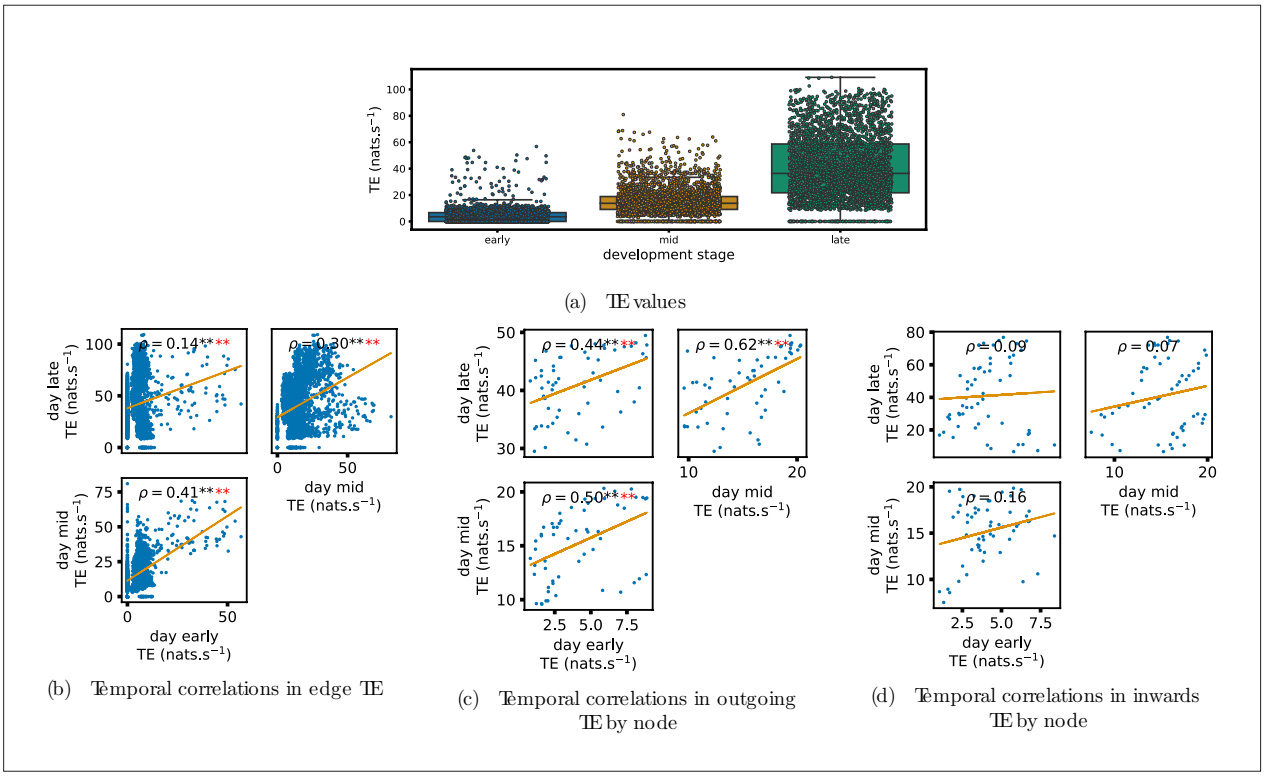
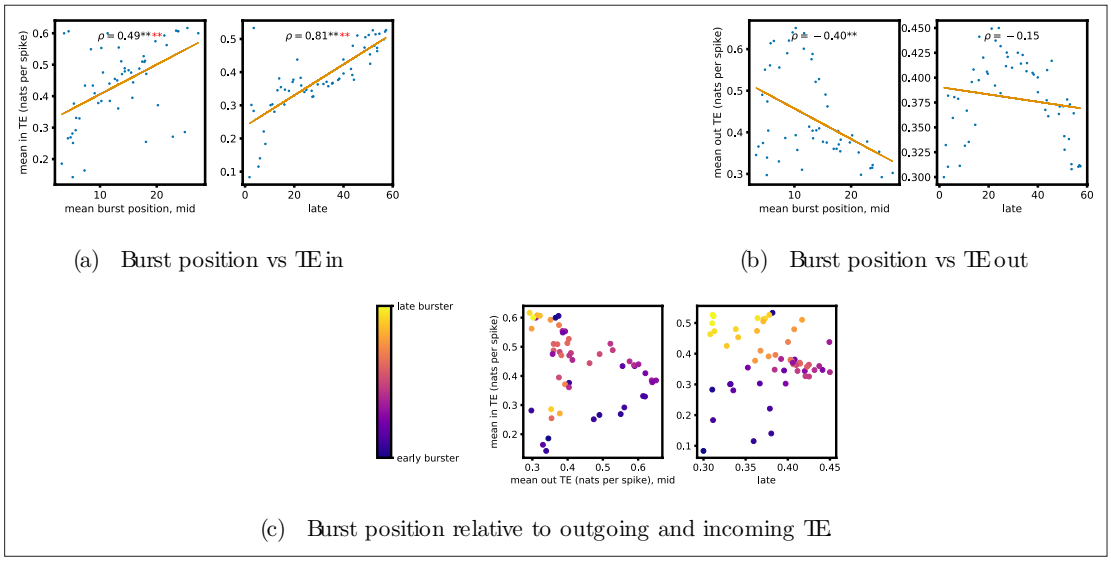


Figure 7. Equivalent plots to those shown in ***Figures 1, 3 and 4*** and ***Appendix 2—figure 1***, but for the simulated spiking network developing under spike-timing-dependent plasticity (STDP). (**a**) shows scatters of the transfer entropy (TE) values overlaid on box plots. The box plots show the quartiles and the median (values greater than 10 SDs from the mean have been removed from both the box and scatter plots as outliers). It corresponds to ***Figure 1***. (**b –d**) show scatter plots investigating the relationship between TE values (or derived summary statistics) over different stages of development. Specifically, in each scatter plot, the x value of a given point is a TE value or derived statistic at an earlier simulation stage and the y value of that same point is a TE value (or derived statistic) on the corresponding edge or node, but later in the simulation. The orange line shows the ordinary least-squares regression. The Spearman correlation (ρ) between the TE values on the two days is displayed in each plot. Values of ρ significant at the 0.05 level are designated with an asterisk and those significant at the 0.01 level are designated with a double asterisk. Red asterisks are used to denote significance after performing a Bonferroni correction for multiple comparisons. (**b**) corresponds to the scatter plots in ***Figure 3***, (**c**) corresponds to the scatter plots in ***Figure 4***, and (**d**) corresponds to the scatter plots in ***Appendix 2—figure 1***.

Scatters and box plots of the TE values estimated in each developmental window are shown in ***Figure 7***. We observe a large, monotonic, increase in these values over development. This mirrors the finding in cell cultures, as described in section 'The dramatic increase in the flow of information during development'.

We also observe a similar lock-in phenomenon of information processing as was found in the cell cultures (described in section 'Early lock-in of information flows'). ***Figure 7a–d*** show the correlation in information flow between different stages of development. Specifically, ***Figure 7a*** plots the correlation in TE values between each ordered pair of neurons between early and later windows. ***Figure 7d*** plots this same correlation, but for the total incoming TE on each neuron, and ***Figure 7c*** does this for the total outgoing TE. In all six of the plots for the relationships between the TE on each edge and for the total outgoing TE, we observe a substantial statistically significant positive correlation between values on earlier and later days (significant at the $p<0.01$ level, with Bonferroni correction). We observe smaller positive correlations in these values for the total incoming TE on each node, although these correlations are not significant, which aligns somewhat with the weaker effect observed for incoming TE in the cultures. As with the cell cultures, some of the observed correlations are particularly strong, such as the Spearman correlation of $\rho = 0.62$ between the total outgoing TE on each node in the mid window and this same value in the late window. This implies that the spatial structure of the



(a) Burst position vs TE in (b) Burst position vs TE out

(c) Burst position relative to outgoing and incoming TE

Figure 8. Equivalent plots to those shown in ***Figure 5***, but for the simulated spiking network developing under spike-timing-dependent plasticity (STDP). Plots show the relationship between the amount of incoming and outgoing local (in burst) transfer entropy (TE) on a given node and its average burst position. (**a**) and (**b**) show the burst position of each node on the x axis of each plot, plotted against either (**a**) the total incoming or (**b**) outgoing TE on the node. The Spearman correlation (ρ) between the mean burst position and the incoming or outgoing TE values is displayed in each plot. Values of ρ significant at the 0.05 level are designated with an asterisk and those significant at the 0.01 level are designated with a double asterisk. Red asterisks are used to denote significance after performing a Bonferroni correction for multiple comparisons. (**c**) plots the outgoing TE on the x axis and the incoming TE on the y axis with the points coloured according to the mean burst position of the node: late bursters are coloured yellow and early bursters are purple.

information flow has a tendency to be determined in the earlier stages of development, after which they are locked-in – in a similar fashion to what was observed in the biological experiments in earlier sections.

We also performed the same analysis on computational roles as presented in section 'Information flows quantify computational role of burst position'. This analysis, the results of which are presented in ***Figure 8***, only looked at the mid and late windows. The early window was ignored due to its lack of bursting activity. In the mid recording window, we observe a strong relationship between the mean burst position of the neuron and its computational role. ***Figure 8*** shows that there is a significant (at the $p<0.01$ level) positive correlation between the mean burst position of a neuron and its total incoming burst-local TE (see section 'Estimation of burst-local TE' for more details on the burst-local TE). There is also a negative correlation between the mean burst position and the total outgoing burst-local TE, as shown in ***Figure 8***. However, this relationship is not significant after Bonferroni correction. These same figures also display these relationships for the late window. Here, we observe the same directions of relationships. This relationship is incredibly strong between incoming TE and burst position ($\rho = 0.81$). The relationship between the outgoing TE and burst position is still negative, although it is not as strong as in the mid window and is no longer significant. Inspection of ***Figure 8*** reveals that there is still a very clear negative relationship between burst position and outgoing TE after a burst position of about 20. Indeed, if we condition on the burst position being greater than 20, then we find a Spearman correlation of $\rho = 0.80$, which is significant at the $p<0.01$ with Bonferroni correction. Inspection of the spike rasters of these simulations suggests that the anomalous results for the earliest bursters may be due to their spiking a very substantial distance ahead of the rest of the population in these simulations. These differing burst dynamics mean that the earliest bursters are then less able to reduce the uncertainty in the spike times of the majority of the neurons which begin spiking significantly later.

This implies that, midway through development, we are observing the same specialisation into computational roles based on burst position as was observed in the cell cultures: early bursters display

a tendency to be information transmitters, late bursters operate as receivers, and middle bursters exhibit a balance of the two. Later on in development, we do, however, observe a slight departure from the roles we observed in the cell cultures. The computational roles are shifted further down the burst propagation and the earliest bursters here are less strongly driving the rest of the population.

It is worth noting that the estimated TE values in the model are substantially higher than in the biological dataset. The median estimated TE in the late window of the model was around $40\,\mathrm{nats.s}^{-1}$ (***Figure 7***). Conversely, it was less than $0.2\,\mathrm{nats.s}^{-1}$ for every last recording day of the cell cultures (***Figure 1***). This is due to the much higher spike rate of the model implying that the dynamics are operating on different timescales. Indeed, if we compare the magnitude of the burst-local TE – which is measured in nats per spike (see section 'Estimation of burst-local TE') – between the model and the biological data (***Figures 8 and 5***, respectively), we find values of similar magnitude.

In summary, in a network model of Izhikevich neurons developing according to STDP towards a state of population bursts, we observe a similar developmental information processing phenomena as in the cell cultures. Namely, the amount of information flowing across the network increases dramatically, the spatial structure of this flow locks in early, and the neurons take on specialised computational roles based on their burst position.

Discussion

Biological neural networks are imbued with an incredible capacity for computation, which is deployed in a flexible manner in order to achieve required tasks. Despite the importance of this capacity to the function of organisms, how it emerges during development has remained largely a mystery. Information dynamics (***Lizier, 2013***; ***Lizier et al., 2014***; ***Lizier et al., 2008***; ***Lizier et al., 2010***; ***Lizier et al., 2012***) provides a framework for studying such computational capacity by measuring the degree to which the fundamental information processing operations of information storage, transfer, and modification occur within an observed system.

Previous work on the information flow component of computational capacity in neural cell cultures (***Nigam et al., 2016***; ***Shimono and Beggs, 2015***; ***Matsuda et al., 2013***; ***Timme et al., 2014***; ***Kajiwara et al., 2021***; ***Timme et al., 2016***; ***Wibral et al., 2017***) has focussed on the static structure of information flow networks at single points in time. This has mostly taken the form of elucidating properties of the functional networks implied by the information flows. However, such work leaves open questions concerning how these structures are formed. We address this gap here.

An initial goal in addressing how computational capacity emerges was to determine when the information flow component arrived. It is plausible that this capacity could have been present shortly after plating or that it could have arrived suddenly at a later point in maturation. What we see, however, is that the capacity for information transmission is either not present or only minimally present in the early DIV. This can be seen by looking at the very low mean TE values in the first column of ***Table 1***. However, over the course of development we see that the TE values increase progressively, reaching values orders of magnitude larger. This implies that information transmission is a capacity which is developed enormously during neuronal development and that its gain is spread consistently throughout the observed period.

The information processing operations of a system tend to be distributed over it in a heterogeneous fashion. For example, it has been found in models of whole-brain networks (***Li et al., 2019***; ***Marinazzo et al., 2012***; ***Marinazzo et al., 2014b***), abstract network models (***Ceguerra et al., 2011***; ***Novelli et al., 2020***; ***Goodman and Porfiri, 2020***), and even energy networks (***Lizier et al., 2009***) that nodes with high indegrees tend to also have high outgoing information flows. Section 'The emergence of functional information flow networks' examined the emergent information flow networks, formed by connecting nodes with a statistically significant TE value between them. In accordance with this previous work – and indeed the large variation in shared, unique and synergistic information flow components observed on the same dataset (albeit with the discrete-time estimator) (***Wibral et al., 2017***) – these networks exhibited a high degree of heterogeneity. Notably, as shown in ***Figure 2a***, they have prominent hubs of inward flow (sinks) along with less pronounced hubs of outgoing flow (sources). Moreover, along with heterogeneity within individual networks, large structural differences are easily observed between the different networks shown in ***Figure 2a***.

Keeping with our goal of uncovering how features of mature information flow networks self-organise, we examined how this heterogeneity at both the intra-network and inter-network levels

emerged. It was found in section 'Early lock-in of information flows' that the key features of the information flow structure are locked-in early in development, around the point at which the information flows emerge. This effect was identified for the outgoing TE from each node, for example, where we found strong correlations over the different days of development. It is worth further noting that this lock-in phenomenon occurs remarkably early in development. Specifically, in very many cases, we observe strong correlations between quantities estimated on the first recording days with nonzero TE and the same quantities estimated on later days. This early lock-in provides us with a mechanism for how the high heterogeneity exhibited in the inflow and outflow hubs emerges. Small differences between networks on early DIV will be magnified on subsequent days. This leads to the high levels of inter-network heterogeneity that we observe. A similar phenomenon has been observed with STDP, which can lead to symmetry breaking in network structure (***Gilson et al., 2009***; ***Kunkel et al., 2011***), whereby small fluctuations in early development can set the trajectory of the synaptic weights on a specific path with a strong history dependence. In order to confirm a hypothesis that this observed lock-in of information flows could be induced by STDP, in section 'Information flows in an STDP model of development' we studied the information dynamics of a model network of Izhikevich neurons developing according to an STDP (***Caporale and Dan, 2008***) update rule from a state of independent tonic firing to population bursting. The lock-in of key features of the information flow structure was evident over the period where the network developed from independent firing to approximately synchronous bursting (i.e., bursts occurring at approximately the same point in time). This indicates a plausible mechanism for our observations and suggests a broader generality of these phenomena.

An interesting difference between the results for the model and the biological data is that the lock-in effect was stronger for outward TE as well as the TE on the edges than in the biological data. The reasons for this difference require further investigation; however, it might be due to the multi-unit nature of the biological data. A possible direction of future work is to modify the model such that the activity on neurons is sub-sampled and aggregated in order to model the effect of placing electrodes in a culture. However, this will detract from the simplicity of the model in its current form.

It has been hypothesised that different neural units take on specialised computational roles (***Schroeter et al., 2015***; ***Frost and Goebel, 2012***; ***Cohen and D'Esposito, 2016***). In section 'Information flows quantify computational role of burst position', we investigated the information flows occurring during the critical bursting periods of the cultures' dynamics. Specifically, we studied the burst-local TE in order to measure the information being transferred between nodes during these periods. The plots shown in ***Figure 5*** show a clear tendency for the nodes to take on specialised computational roles as development progresses. Moreover, these computational roles were tightly coupled to the position the node tended to burst within in the burst propagation. Nodes that tended to initiate the bursts had a tendency to have high outgoing information transfer combined with low incoming information flow, implying their role as information transmitters. The opposite relationship is observed for typically late bursters, indicating their role as information receivers. By contrast, nodes that usually burst during the middle of the progression have a balance between outward and inward flows. This indicates that they are the crucial links between the transmitters and receivers of information. Neurons bursting in the middle of the burst progression of dissociated cell cultures have received special attention in past work using undirected measures, where it was conjectured that they act as the 'brokers of neuronal communication' (***Schroeter et al., 2015***). In this work, we have provided novel supporting evidence for this conjecture by specifically identifying the *directed* information flows into and out of these nodes. Moreover, in section 'Information flows in an STDP model of development', we observed that this same specialisation of neurons into computational roles based on average burst position occurred in a model network of Izhikevich neurons which had developed via an STDP learning rule to a state of population bursting. This suggests that this phenomenon might exist more generally than the specific cell cultures studied. It is also worth noting that some of these relationships, notably those shown in ***Figure 7b and d***, are much stronger than what was observed in the cell culture. It is likely that this is due to the fact that in the model we estimated TE between individual model neurons, whereas in the cultures we estimated TE between the MUA on each electrode. A possible direction for future work will be to study how the estimated information flow changes when we aggregate the spikes from multiple model neurons into simulated MUA.

It is worth reflecting on the fact that the observed correlations between burst-local information transfer and average burst position will not occur in all neuronal populations. For instance, in

populations with strictly periodic bursts, each node's behaviour will be well explained by its own history, resulting in very low burst-local TE's, regardless of burst position. Furthermore, neuronal populations develop bursty dynamics to different extents and such quantities (let alone their correlation) are simply less meaningful in the absence of burstiness (e.g. for the final day of culture 1-5 as stated in Figure 3A of ***Wagenaar et al., 2006b***).

Returning once more to our focus on investigating the emergence of information flows, we have demonstrated, in section 'Early lock-in of specialised computational roles', that these specialist computational roles have a tendency to lock-in early. There we looked at the ratio of outgoing burst-local TE to the total burst-local TE on each node. It was found that there is a strong tendency for this ratio to be correlated between early and late days of development. This suggests that the computational role that a node performs during population bursts is determined to a large degree early in development.

Insights into development aside, a fundamental technical difference between the work presented here and previous studies of TE in neural cultures is that here we use a recently developed continuous-time estimator of TE (***Shorten et al., 2021***). This estimator was demonstrated to have far higher accuracy in estimating information flows than the traditional discrete-time estimator. The principal challenge which is faced when using the discrete-time estimator is that the curse of dimensionality limits the number of previous time bins that can be used to estimate the history-dependent spike rates. All applications of this estimator to spiking data from cell cultures of which the authors are aware (***Nigam et al., 2016***; ***Shimono and Beggs, 2015***; ***Matsuda et al., 2013***; ***Timme et al., 2014***; ***Kajiwara et al., 2021***; ***Timme et al., 2016***) made use of only a single previous bin in the estimation of these rates. This makes it impossible to simultaneously achieve high time precision and capture the dependence of the spike rate on spikes occurring further back in time. Conversely, by operating on the interspike intervals, the continuous-time estimator can capture the dependence of the spike rate on events occurring relatively far back in time, while maintaining the time precision of the raw data. Looking at a specific representative example, our target history embeddings made use of the previous four interspike intervals (section 'Selection of embedding lengths'). For the recording on day 24 of culture 1-3, the mean interspike interval was 0.71 s. This implies that the target history embeddings on average extended over a period of 2.84 s. The raw data was collected with a sampling rate of $25\,\mathrm{kHz}$ (***Wagenaar et al., 2006b***). In order to lose no time precision, the discrete-time estimator would thus have to use bins of 40 s, and then in order to extend over 2.84 s, the target history embeddings would therefore need to consist of around 70,000 bins which is empirically not possible to sample well.

It is worth noting that, as we were performing a longitudinal analysis where each studied recording was separated by days or weeks, we did not perform spike sorting as we would have been unable to match the different units on an electrode across different recordings. We would then not have been able to compare the TE values on a given unit over the course of development. Instead, we analysed the spikes on each electrode without sorting. As such, this work studies MUA (***Schroeter et al., 2015***). Spike sorting applied to data collected from a near-identical recording setup found an average of four neurons per electrode (***Wagenaar et al., 2006a***). This situates this work at a spatial scale slightly larger than spike-sorted neural data, but still orders of magnitude finer than fMRI, EEG, or MEG (***Bassett and Sporns, 2017***).

Functional and effective networks arising from the publicly available dataset used in this article (***Wagenaar et al., 2006b***) have been studied by other authors. For instance, it was shown that the functional networks were small world (***Downes et al., 2012***) and that more connected nodes exhibited stronger signatures of nonlinearity (***Minati et al., 2019***). Work has also been conducted analysing the burst propagation of these cultures, finding that there are 'leader' nodes which consistently burst before the rest of the population (***Eckmann et al., 2008***). These are the information transmitters that we have observed in this work here.

An exciting direction for future work will be to move beyond directed functional networks to examine the information flow provided by higher-order multivariate TEs in an effective network structure (***Novelli and Lizier, 2021***; ***Novelli et al., 2019***). The networks inferred by such higher-order TEs are able to better reflect the networks' underlying structural features (***Novelli and Lizier, 2021***). As was the case with bivariate TEs prior to this work, there is an absence of work investigating how the networks of multivariate information flow emerge during neural development. Moreover, moving to higher-order measures will allow us to more fully characterise the multifaceted specialised computational roles undertaken by neurons.

Table 5. File numbers used for each culture on each day.
These correspond to the file numbering used in the freely available dataset used in this study, provided by ***Wagenaar et al., 2006b***; ***Network, 2021***.

Culture 1-1	Day 4	Day 14	Day 20	
	2	2	2	
Culture 1-2	Day 6	Day 11	Day 17	
	2	2	2	
Culture 1-3	Day 5	Day 10	Day 16	Day 24
	2	2	2	2
Culture 1-4	Day 8	Day 13	Day 19	
	2	2	2	
Culture 1-5	Day 7	Day 12	Day 18	
	2	2	2	
Culture 2-1	Day 14	Day 32		
	2	2		
Culture 2-2	Day 9	Day 15	Day 21	Day 33
	2	2	2	2
Culture 2-3	Day 6	Day 12	Day 24	
	2	2	2	
Culture 2-4	Day 3	Day 5	Day 11	
	1	1	1	
Culture 2-5	Day 4	Day 10	Day 22	Day 28
	1	1	2	1
Culture 2-6	Day 7	Day 13	Day 31	
	1	1	1	

Materials and methods

Cell culture data

The spike train recordings used in this study were collected by ***Wagenaar et al., 2006b*** and are freely available online (***Network, 2021***). The details of the methodology used in these recordings can be found in the original publication (***Wagenaar et al., 2006b***). A short summary of their methodology follows.

Dissociated cultures of rat cortical neurons had their activity recorded. This was achieved by plating 8 × 8 MEAs, operating at a sampling frequency of 25 kHz with neurons obtained from the cortices of rat embryos. The spacing between the electrodes was 200 m centre-to-centre. The MEAs did not have electrodes on their corners and one electrode was used as ground, resulting in recordings from 59 electrodes. In all recordings, electrodes with less than 100 spikes were removed from the analysis. This resulted in electrodes 37 and 43 (see ***Figure 2*** for the position of these electrodes) being removed from every recording as no spikes were recorded on them. The spatial layout of the electrodes is available from the website associated with the dataset (***Network, 2021***), allowing us to overlay the functional networks onto this spatial layout as is done in ***Figure 2a***.

30 min recordings were conducted on most days, starting from 3 to 4 DIV. The end point of recording varied between 25 and 39 DIV. Longer overnight recordings were also conducted on some cultures at sparser intervals. As the accurate estimation of information-theoretic quantities requires substantial amounts of data (***Shorten et al., 2021***; ***Kraskov et al., 2004***), in this work we make use

of these longer overnight recordings. These recordings were split into multiple files. The specific files used, along with the names of the cultures and days of the recordings, are listed in ***Table 5***.

The original study plated the electrodes with varying densities of cortical cells. However, overnight recordings were only performed on the 'dense' cultures, plated with a density of $2500\,\text{cells/L}$.

The original study performed threshold-based spike detection by determining that a spike was present in the case of an upward or downward excursion beyond 4.5 times the estimated RMS noise of the recorded potential on a given electrode. The analysis presented in this article makes use of these detected spike times. No spike sorting was performed, and, as such, we are studying MUA (***Schroeter et al., 2015***).

Network of Izhikevich neurons

The model spiking network used to generate the data analysed in section 'Information flows in an STDP model of development' is identical to that presented in ***Khoshkhou and Montakhab, 2019***, with a few minor alterations. This model consists of Izhikevich neurons (***Izhikevich, 2003***) developing according to an STDP (***Caporale and Dan, 2008***) update rule. At the beginning of the simulation, each neuron performs independent tonic spiking; however, the network develops towards population bursts.

The specific model settings used were based on those used to produce Figure 5A in ***Khoshkhou and Montakhab, 2019***. That is, the proportion of inhibitory neurons (α) and the synapse time delay (τ_{ij}) were both set to 0. The first change made was to use 59 neurons, as opposed to the 500 used in ***Khoshkhou and Montakhab, 2019***, in order to correspond to the number of electrodes used in the cell culture recordings. The maximum connection strength (g_{max}) was also increased from 0.6 to 10 in order to compensate for this reduction in the network size.

The only remaining change was made in order to slow the rate of development of the population. The reasoning behind this was to allow for the extraction of windows which were much shorter than the timescale of development, resulting in the dynamics within these windows being approximately stationary (and including enough samples for estimation of the TE rates). Specifically, this change was to greatly reduce the values of the maximum synaptic potentiation and depression (A_+ and A_-). These values were reduced from 5×10^{-2} to 4×10^{-4}.

Data preprocessing

As the data was sampled at 25 kHz, uniform noise distributed between –20 s and 20 s was added to each spike time. This is to prevent the TE estimator from exploiting the fact that, in the raw data, interspike intervals are always an integer multiple of 40 s.

Transfer entropy estimation

The (bivariate) TE (***Schreiber, 2000***; ***Bossomaier et al., 2016***) was estimated between each pair of electrodes in each of the recordings listed in ***Table 5***. TE is the mutual information between the past state of a source process and the present state of a target process, conditioned on the past state of the target. More specifically (in discrete time), the TE rate is

$$\dot{\mathbf{T}}_{Y\to X} = \frac{1}{\Delta t} I\left(X_t\,;\,\mathbf{Y}_{<t}\,\middle|\,\mathbf{X}_{<t}\right) = \frac{1}{\tau}\sum_{t=1}^{N_T} \ln \frac{p\left(x_t\,\middle|\,\mathbf{x}_{<t},\mathbf{y}_{<t}\right)}{p\left(x_t\,\middle|\,\mathbf{x}_{<t}\right)}. \tag{1}$$

The TE above is being measured from a source Y to a target X, $I(\cdot\,;\,\cdot|\cdot)$ is the conditional mutual information (***MacKay and Kay, 2003***), x_t is the current state of the target, $\mathbf{x}_{<t}$ is the history of the target, $\mathbf{y}_{<t}$ is the history of the source, Δt is the bin width (in time units), τ is the length of the processes, and $N_T = \tau/\Delta t$ is the number of time samples (bins). The histories $\mathbf{x}_{<t}$ and $\mathbf{y}_{<t}$ are usually captured via embedding vectors, for example, $\mathbf{x}_{<t} = \mathbf{x}_{t-m:t-1} = \{x_{t-m}, x_{t-m+1}, \ldots, x_{t-1}\}$.

Previous application of the discrete-time estimator

Previous applications of TE to spiking data from neural cell cultures (***Nigam et al., 2016***; ***Shimono and Beggs, 2015***; ***Matsuda et al., 2013***; ***Timme et al., 2014***; ***Kajiwara et al., 2021***; ***Timme et al., 2016***; ***Wibral et al., 2017***) made use of this discrete-time formulation of TE. This work was primarily focussed on the directed functional networks implied by the estimated TE values between pairs of

nodes which has revealed interesting features of the information flow structure. ***Shimono and Beggs, 2015*** found that these networks exhibited a highly non-random structure and contained a long-tailed degree distribution. This work was expanded by ***Nigam et al., 2016***, where it was found that the functional networks contained a rich-club topology. Conversely, ***Timme et al., 2014*** found that the hubs of these networks were localised to certain timescales. Other work (***Timme et al., 2016***; ***Wibral et al., 2017***) has instead focussed on how the components of information flows in cell cultures can be decomposed into unique, redundant, and synergistic components.

Continuous-time estimation

It has, relatively recently, been shown that, for event-based data such as spike trains, in the limit of small bin size, the TE is given by the following expression (***Spinney and Lizier, 2018***):

$$\dot{\mathbf{T}}_{Y\to X} = \lim_{\tau\to\infty} \frac{1}{\tau} \sum_{i=1}^{N_X} \ln \frac{\lambda_{x|\mathbf{x}_{<t},\mathbf{y}_{<t}}\left[\mathbf{x}_{<x_i}, \mathbf{y}_{<x_i}\right]}{\lambda_{x|\mathbf{x}_{<t}}\left[\mathbf{x}_{<x_i}\right]}. \tag{2}$$

Here, $\lambda_{x|\mathbf{x}_{<t},\mathbf{y}_{<t}}\left[\mathbf{x}_{<x_i}, \mathbf{y}_{<x_i}\right]$ is the instantaneous firing rate of the target conditioned on the histories of the target $\mathbf{x}_{<x_i}$ and source $\mathbf{y}_{<x_i}$ at the time points x_i of the spike events in the target process. $\lambda_{x|\mathbf{x}_{<t}}\left[\mathbf{x}_{<x_i}\right]$ is the instantaneous firing rate of the target conditioned on its history alone, ignoring the history of the source. It is important to note that the sum is being taken over the N_X spikes of the target, thereby evaluating log ratios of the expected spike rates of the target given source and target histories versus target histories alone, *when* the target does spike. As this expression allows us to ignore the 'empty space' between events, it presented clear potential for allowing for more efficient estimation of TE on spike trains.

This potential was recently realised in a new continuous-time estimator of TE presented in ***Shorten et al., 2021*** (and utilised in ***Mijatovic et al., 2021***), and all TE estimates in this article were performed using this new estimator. In ***Shorten et al., 2021*** it is demonstrated that this continuous-time estimator is far superior to the traditional discrete-time approach to TE estimation on spike trains. For a start, unlike the discrete-time estimator, it is consistent. That is, in the limit of infinite data, it will converge to the true value of the TE. It was also shown to have much preferable bias and convergence properties. Most significantly, perhaps, this new estimator utilises the interspike intervals to efficiently represent the history embeddings $\mathbf{x}_{<x_i}$ and $\mathbf{y}_{<x_i}$ in estimating the relevant conditional spike rates in (***Lizier, 2013***). This then allows for the application of the highly effective nearest-neighbour family of information-theoretic estimators (***Kozachenko and Leonenko, 1987***; ***Kraskov et al., 2004***), which bring estimation efficiency, bias correction, and together with their application to interspike intervals enable capture of long timescale dependencies.

This is in contrast to the traditional discrete-time estimator which uses the presence or absence of spikes in time bins as its history embeddings (it sometimes also uses the number of spikes occurring in a bin). In order to avoid the dimensionality of the estimation problem becoming sufficiently large so as to render estimation infeasible, only a small number of bins can be used in these embeddings. Indeed, to the best of the authors' knowledge, all previous applications of the discrete-time TE estimator to spiking data from cell cultures used only a single bin in their history embeddings. The bin widths used in those studies were 40 s (***Nigam et al., 2016***), 0.3 ms (***Garofalo et al., 2009***), and 1 ms (***Shimono and Beggs, 2015***; ***Kajiwara et al., 2020***). Some studies chose to examine the TE values produced by multiple different bin widths, specifically, 0.6 ms and 100 ms (***Matsuda et al., 2013***), 1.6 ms and 3.5 ms (***Timme et al., 2016***), and 10 different widths ranging from 1 ms to 750 ms (***Timme et al., 2014***). Specifically those studies demonstrated the unfortunate high sensitivity of the discrete-time TE estimator to the bin width parameter. In the instances where narrow (<5 ms) bins were used, only a very narrow slice of history is being considered in the estimation of the history-conditional spike rate. This is problematic as it is known that correlations in spike trains exist over distances of (at least) hundreds of milliseconds (***Aldridge and Gilman, 1991***; ***Rudelt et al., 2021***). Conversely, in the instances where broad (>5 ms) bins were used, relationships occurring on fine timescales will be completely missed. This is significant given that it is established that correlations at the millisecond and sub-millisecond scale play a role in neural function (***Nemenman et al., 2008***; ***Kayser et al., 2010***; ***Sober et al., 2018***; ***Garcia-Lazaro et al., 2013***). In other words, previous applications of TE to electrophysiological data from cell cultures either captured some correlations occurring with fine temporal

Table 6. The parameter values used in the continuous-time transfer entropy (TE) estimator.
A complete description of these parameters, along with analysis and discussion of their effects, can be found in ***Shorten et al., 2021***.

Parameter	Description	Value
N_X	Number of spikes in the target spike train	Varied (see text)
l_X	Number of interspike intervals in target history embeddings	4
l_Y	Number of interspike intervals in source history embeddings	2
k_{global}	Number of nearest neighbours to find in the initial search	10
k_{perm}	Number of nearest neighbours to consider during surrogate generation	10
N_U	Number of random samples of histories at non-spiking points in time	$50N_X$
$N_{U,\text{surrogates}}$	Number of random samples of histories at non-spiking points in time used for surrogate generation	$10N_X$
$N_{\text{surrogates}}$	Number of surrogates to generate for each node pair	100

precision or they captured relationships occurring over larger intervals, but never both simultaneously. This can be contrasted with the interspike interval history representation used in this study. To take a concrete example, for the recording on day 24 of culture 1-3, the average interspike interval was 0.71 s. This implies that the target history embeddings (composed of four interspike intervals) on average extended over a period of 2.84 s and the source history embeddings (composed of two interspike intervals) on average extended over a period of 1.42 s. This is despite the fact that our history representations retain the precision of the raw data (40 s) and the ability to measure relationships on this scale where they are relevant (via the underlying nearest-neighbour estimators).

The parameters used with this estimator are shown in ***Table 6***. The values of k_{global} and k_{perm} were chosen because, in previous work (***Shorten et al., 2021***), similar values were found to facilitate stable performance of the estimator. The high values of N_U and $N_{U,\text{surrogates}}$ were chosen so that histories during bursting periods could be adequately sampled. These two parameters refer to sample points placed randomly in the spike train, at which history embeddings are sampled. As the periods of bursting comprise a relatively small fraction of the total recording time, many samples need to be placed in order to achieve a good sample of histories potentially observed during these periods. The choice of embedding lengths is discussed in the section 'Selection of embedding lengths', and the choice of $N_{\text{surrogates}}$ is discussed in the section 'Significance testing of TE values'.

Instead of selecting a single number of target spikes N_X to include in the analysis, we chose to include all the spikes that occurred within the first hour of recording time. The reason for doing this was that the spike rates varied by orders of magnitude between the electrodes. This meant that fixing the number of target spikes would result in the source spikes being severely undersampled in cases where the target spike rate was much higher than the source spike rate. When using 1 hr of recording time, among the main cultures the smallest number of spikes per electrode was 481, the maximum was 69,627, and the median was 17,399.

Selection of embedding lengths

The target embedding lengths were determined by adapting the technique (***Erten et al., 2017***; ***Novelli et al., 2019***) extending (***Garland et al., 2016***) of maximising the bias-corrected active information storage (AIS) (***Lizier et al., 2012***) over different target embedding lengths for a given target. Our adaptations sought to select a consensus embedding parameter for all targets on all trials to avoid different bias properties due to different parameters across targets and trials, in a similar fashion to ***Hansen et al., 2021***. As such, our approach determines a target embedding length l_X which maximises the *average* bias-corrected AIS across all electrodes using one representative recording

Table 7. Summary statistics for the active information storage (AIS) values estimated at different target embedding lengths l_X.
These were estimated across all electrodes of a representative recording (day 23 of culture 1-3). The p-values shown in the fourth column are associated with the null hypothesis that the mean AIS at the given l_X is equal to the mean AIS at $l_X - 1$.

l_X	Mean AIS	SD	p-Value
1	7.73	4.71	–
2	8.27	4.97	3.0×10^{-19}
3	8.41	5.08	5.8×10^{-8}
4	8.44	5.11	2.7×10^{-4}
5	8.43	5.12	0.85

(selected as day 23 of culture 1-3). To estimate AIS within the continuous-time framework (***Spinney and Lizier, 2018***) for this purpose, we estimated the difference between the second KL divergence of Equation 10 of ***Shorten et al., 2021*** and the mean firing rate of the target. These estimates contain inherent bias correction as per the TE estimator itself. Moreover, the mean of surrogate values was subtracted to further reduce the bias. The embedding length l_X was continuously increased so long as each subsequent embedding produced a statistically significant (at the $p<0.05$ level) increase in the average AIS across the electrodes. The resulting mean AIS values (along with standard deviations) and p-values are shown in ***Table 7***. We found that every increase in l_X up to 4 produced a statistically significant increase in the mean AIS. The increase from 4 to 5 produced a non-significant decrease in the mean AIS and so l_X was set to 4.

With the target embedding length determined, we set about similarly determining a consensus source embedding length l_Y by estimating the TE between all directed electrode pairs on the same representative recording for different values of l_Y. Each estimate also had the mean of the surrogate population subtracted to reduce its bias (see section 'Significance testing of TE values').

The embedding length was continuously increased so long as each subsequent embedding produced a statistically significant (at the $p<0.05$ level) increase in the average TE across all electrode pairs. The resulting mean TE values (along with standard deviations) and p-values are shown in ***Table 8***. We found that increasing l_Y from 1 to 2 produced a statistically significant increase in the mean TE. However, increasing l_Y from 2 to 3 produced a non-significant decrease in the mean TE. As such, we set l_Y to 2.

Significance testing of TE values

In constructing the directed functional networks displayed in ***Figure 2a***, we tested whether the estimated TE between each source-target pair was statistically different from the distribution of TEs under the null hypothesis of conditional independence of the target from the source (i.e. TE consistent with zero). Significance testing for TE in this way is performed by constructing a population of surrogate time-series or history embeddings that conform to the null hypothesis of zero TE (***Novelli et al., 2019***; ***Wollstadt et al., 2019***; ***Novelli and Lizier, 2021***). We then estimate the TE on each of

Table 8. Summary statistics for the transfer entropy (TE) values estimated at different source embedding lengths l_Y.
These were estimated between all electrodes of a representative recording (day 23 of culture 1–-3). The p-values shown in the fourth column are associated with the null hypothesis that the mean TE at the given l_Y is equal to the mean TE at $l_Y - 1$.

l_Y	Mean TE	SD	p-Value
1	0.031	0.043	–
2	0.058	0.056	0.0
3	0.057	0.069	0.84

these surrogates to generate a null distribution of TE. Specifically, we generate the surrogates and compute their TEs according to the method associated with the continuous-time spiking TE estimator (***Shorten et al., 2021***) and using the parameters shown in ***Table 6***. One small change was made to that surrogate generation method: instead of laying out the $N_{U,\text{surrogates}}$ sample points randomly uniformly, we placed each one at an existing target spike, with the addition of uniform noise on the interval $[-240\,\text{ms},\ 240\,\text{ms}]$. This was to ensure that these points adequately sampled the incredibly dense burst regions.

With the surrogate TE distribution constructed, the resulting p-value for our TE estimate can be computed by counting the proportion of these surrogate TEs that are greater than or equal to the original estimate. Here, we seek to compare significance against a threshold of $\alpha < 0.01$. We chose this lower threshold as false positives are generally considered more damaging than false negatives when applying network inference to neuroscientific data (***Zalesky et al., 2016***). We also applied a Bonferroni correction (***Miller, 2012***) to all the significance tests done on a given recording. Given that there are 59 electrodes in the recordings, 3422 tests were performed in each recording. This meant that, once the Bonferroni correction was included, the significance threshold dropped to $p<2.9 \times 10^{-6}$. Comparing against such a low significance threshold would require an infeasible number of surrogates for the many pairs within each recording, if computing the p-value by counting as above. Instead, we assume that the null TE distribution is Gaussian and compute the p-value for our TE estimate using the CDF of the Gaussian distribution fitted from 100 surrogates (e.g. as per ***Lizier et al., 2011***). Specifically, the p-value reports the probability that a TE estimate on history embeddings conforming to the null hypothesis of zero TE being greater than or equal to our original estimated TE value. If this p-value is below the threshold, then the null hypothesis is rejected and we conclude that there is a statistically significant information flow between the electrodes.

Analysis of population bursts

A common family of methods for extracting periods of bursting activity from spike-train recordings examines the length of adjacent interspike intervals. The period spanned by these intervals is designated a burst if some summary statistic of the intervals (e.g. their sum or maximum) is below a certain threshold (***Kaneoke and Vitek, 1996***; ***Wagenaar et al., 2005***; ***Wagenaar et al., 2006b***; ***Selinger et al., 2007***; ***Bakkum et al., 2013***). In order to detect single-electrode as well as population-wide bursts, we implement such an approach here.

We first determine the start and end points of the bursts of each individual electrode. The locations of the population bursts were subsequently determined using the results of this per-electrode analysis.

The method for determining the times during which an individual electrode was bursting proceeded as follows: the spikes were moved through sequentially. If the interval between a given spike and the second most recent historic spike for that electrode was less than α, then, if the electrode was not already in a burst, it was deemed to have a burst starting at the kth most recent historic spike. A burst was taken to continue until an interspike interval greater than $a * \alpha$ was encountered. If such an interval was encountered, then the end of the burst was designated as the timestamp of the earlier of the two spikes forming the interval.

The starts and ends of population bursts were similarly determined by moving through the time series in a sequential fashion. If the population was not already designated to be in a burst, but the number of electrodes currently bursting was greater than the threshold β, then a burst start position was set at the point this threshold was crossed. Conversely, if the electrode was already designated to be in a burst and the number of individual electrodes currently bursting dropped below the threshold γ ($\gamma < \beta$), then a burst stop position was set at the point this threshold was crossed.

In this article, we always made use of the parameters $k = 2, \alpha = \frac{1}{2\bar{\lambda}}, a = 3, \beta = 15$ and $\gamma = 10$, where $\bar{\lambda}$ is the average spike rate. These parameters were chosen by trial-and-error combined with visual inspection of the resulting inferred burst positions. The results of this scheme showed low sensitivity to the choice of these parameters.

For the simulated network dynamics, we used the parameters $k = 1, \alpha = \frac{1}{2\bar{\lambda}}, a = 1.5, \beta = 2$ and $\gamma = 1$. These parameters were found to better suit the stereotyped dynamics of the simulated networks.

Estimation of burst-local TE

The information dynamics framework provides us with the unique ability to analyse information processing locally in time (***Lizier et al., 2008***; ***Lizier, 2013***; ***Lizier, 2014***). We make use of that ability here to allow us to specifically examine the information flows during the important period of population bursts. The TE estimator which we are employing here (***Shorten et al., 2021***) sums contributions from each spike in the target spike train. It then divides this total by the time length of the target spike train that is being examined. In order to estimate the burst-local TE, we simply sum the contributions from the target spikes where those spikes occurred during a population burst. We then normalise by the number of such spikes, providing us with a burst-local TE estimate in units of nats per spike, instead of nats per second. Note that the burst-local TE is different to the approach of ***Stetter et al., 2012***, who extracted the bursting activity prior to any analysis, rendering a TE conditioned on bursting occurring. Specifically, in contrast to the burst-local TE, in their work the non-spiking activity is ignored for the purposes of estimating the log densities.

Code availability

Scripts for performing the analysis in this article can be found at bitbucket.org/dpshorten/cell_cultures (***Shorten, 2022***; copy archived at swh:1:rev:8ee5e519da5cb90590865e9a692b96ad7e68a69e).

Acknowledgements

JL was supported through the Australian Research Council DECRA grant DE160100630 (https://www.arc.gov.au/grants/discovery-program/discovery-early-career-researcher-award-decra) and The University of Sydney Research Accelerator (SOAR) prize program (https://www.sydney.edu.au/research/our-researchers/sydney-research-accelerator-prizes.html). The authors acknowledge the technical assistance provided by the Sydney Informatics Hub, a Core Research Facility of the University of Sydney. In particular, the analysis presented in this work made use of the Artemis HPC cluster.

Additional information

Funding

Funder	Grant reference number	Author
Australian Research Council	DE160100630	Joseph T Lizier
University of Sydney	SOAR Fellowship	Joseph T Lizier
Deutsche Forschungsgemeinschaft	SFB 1528	Viola Priesemann

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

David P Shorten, Conceptualization, Investigation, Methodology, Software, Visualization, Writing - original draft, Writing - review and editing; Viola Priesemann, Michael Wibral, Conceptualization, Writing - review and editing; Joseph T Lizier, Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing - review and editing

Author ORCIDs

David P Shorten http://orcid.org/0000-0003-2412-4705
Viola Priesemann http://orcid.org/0000-0001-8905-5873
Michael Wibral http://orcid.org/0000-0001-8010-5862
Joseph T Lizier http://orcid.org/0000-0002-9910-8972

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.74651.sa1
Author response https://doi.org/10.7554/eLife.74651.sa2

Additional files

Supplementary files

- Transparent reporting form

Data availability

This work made use of a publicly available dataset which can be found at: http://neurodatasharing.bme.gatech.edu/development-data/html/index.html. Analysis scripts are available at: https://bitbucket.org/dpshorten/cell_cultures (copy archived at swh:1:rev:8ee5e519da5cb90590865e9a692b96ad7e68a69e).

The following previously published dataset was used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Pine J, Potter S	2006	Network activity of developing cortical cultures in vitro	http://neurodatasharing.bme.gatech.edu/development-data/html/index.html	neurodatasharing, development-data

References

Aldridge JW, Gilman S. 1991. The temporal structure of spike trains in the primate basal ganglia: afferent regulation of bursting demonstrated with precentral cerebral cortical ablation. *Brain Research* **543**:123–138. DOI: https://doi.org/10.1016/0006-8993(91)91055-6, PMID: 2054667

Bakkum DJ, Radivojevic M, Frey U, Franke F, Hierlemann A, Takahashi H. 2013. Parameters for burst detection. *Frontiers in Computational Neuroscience* **7**:193. DOI: https://doi.org/10.3389/fncom.2013.00193, PMID: 24567714

Bassett DS, Sporns O. 2017. Network neuroscience. *Nature Neuroscience* **20**:353–364. DOI: https://doi.org/10.1038/nn.4502, PMID: 28230844

Bossomaier T, Barnett L, Harré M, Lizier JT. 2016. An Introduction to Transfer Entropy. Cham: Springer International Publishing. DOI: https://doi.org/10.1007/978-3-319-43222-9

Caporale N, Dan Y. 2008. Spike timing–dependent plasticity: a hebbian learning rule, Annu. *Annual Review of Neuroscience* **31**:25–46. DOI: https://doi.org/10.1146/annurev.neuro.31.060407.125639, PMID: 18275283

Ceguerra RV, Lizier JT, Zomaya AY. 2011. Information storage and transfer in the synchronization process in locally-connected networks. 2011 IEEE Symposium On Artificial Life - Part Of 17273 - 2011 Ssci. Paris, France. DOI: https://doi.org/10.1109/ALIFE.2011.5954653

Cohen JR, D'Esposito M. 2016. The segregation and integration of distinct brain networks and their relationship to cognition. *The Journal of Neuroscience* **36**:12083–12094. DOI: https://doi.org/10.1523/JNEUROSCI.2965-15.2016, PMID: 27903719

Cramer B, Stöckel D, Kreft M, Wibral M, Schemmel J, Meier K, Priesemann V. 2020. Control of criticality and computation in spiking neuromorphic networks with plasticity. *Nature Communications* **11**:2853. DOI: https://doi.org/10.1038/s41467-020-16548-3, PMID: 32503982

Downes JH, Hammond MW, Xydas D, Spencer MC, Becerra VM, Warwick K, Whalley BJ, Nasuto SJ. 2012. Emergence of a small-world functional network in cultured neurons. *PLOS Computational Biology* **8**:e1002522. DOI: https://doi.org/10.1371/journal.pcbi.1002522, PMID: 22615555

Eckmann JP, Jacobi S, Marom S, Moses E, Zbinden C. 2008. Leader neurons in population bursts of 2D living neural networks. *New Journal of Physics* **10**:015011. DOI: https://doi.org/10.1088/1367-2630/10/1/015011

Erten EY, Lizier JT, Piraveenan M, Prokopenko M. 2017. Criticality and information dynamics in epidemiological models. *Entropy* **19**:194. DOI: https://doi.org/10.3390/e19050194

Frost MA, Goebel R. 2012. Measuring structural–functional correspondence: spatial variability of specialised brain regions after macro-anatomical alignment. *NeuroImage* **59**:1369–1381. DOI: https://doi.org/10.1016/j.neuroimage.2011.08.035, PMID: 21875671

Garcia-Lazaro JA, Belliveau LAC, Lesica NA. 2013. Independent population coding of speech with sub-millisecond precision. *The Journal of Neuroscience* **33**:19362–19372. DOI: https://doi.org/10.1523/JNEUROSCI.3711-13.2013, PMID: 24305831

Garland J, James RG, Bradley E. 2016. Leveraging information storage to select forecast-optimal parameters for delay-coordinate reconstructions. *Physical Review. E* **93**:022221. DOI: https://doi.org/10.1103/PhysRevE.93.022221, PMID: 26986345

Garofalo M, Nieus T, Massobrio P, Martinoia S. 2009. Evaluation of the performance of information theory-based methods and cross-correlation to estimate the functional connectivity in cortical networks. *PLOS ONE* **4**:e6482. DOI: https://doi.org/10.1371/journal.pone.0006482, PMID: 19652720

Gibbons JD, Chakraborti S. 2020. . *Nonparametric Statistical Inference*. 6th edition. Boca Raton: CRC press. DOI: https://doi.org/10.1201/9781315110479

Gilson M, Burkitt AN, Grayden DB, Thomas DA, van Hemmen JL. 2009. Emergence of network structure due to spike-timing-dependent plasticity in recurrent neuronal networks. *Biological Cybernetics* **101**:427–444. DOI: https://doi.org/10.1007/s00422-009-0346-1, PMID: 19937070

Goodman R, Porfiri M. 2020. Topological features determining the error in the inference of networks using transfer entropy. *Mathematics in Engineering* **2**:34–54. DOI: https://doi.org/10.3934/mine.2020003

Haldeman C, Beggs JM. 2005. Critical branching captures activity in living neural networks and maximizes the number of metastable states. *Physical Review Letters* **94**:058101. DOI: https://doi.org/10.1103/PhysRevLett.94.058101, PMID: 15783702

Hansen MJ, Burns AL, Monk CT, Schutz C, Lizier JT, Ramnarine I, Ward AJW, Krause J. 2021. The effect of predation risk on group behaviour and information flow during repeated collective decisions. *Animal Behaviour* **173**:215–239. DOI: https://doi.org/10.1016/j.anbehav.2021.01.005

Huang CS, Pal NR, Chuang CH, Lin CT. 2015. Identifying changes in eeg information transfer during drowsy driving by transfer entropy. *Frontiers in Human Neuroscience* **9**:570. DOI: https://doi.org/10.3389/fnhum.2015.00570, PMID: 26557069

Izhikevich EM. 2003. Simple model of spiking neurons. *IEEE Transactions on Neural Networks* **14**:1569–1572. DOI: https://doi.org/10.1109/TNN.2003.820440, PMID: 18244602

Kajiwara M, Nomura R, Goetze F, Akutsu T, Shimono M. 2020. Inhibitory Neurons Are a Central Controlling Regulator in the Effective Cortical Microconnectome. *bioRxiv*. DOI: https://doi.org/10.1101/2020.02.18.954016

Kajiwara M, Nomura R, Goetze F, Kawabata M, Isomura Y, Akutsu T, Shimono M. 2021. Inhibitory neurons exhibit high controlling ability in the cortical microconnectome. *PLOS Computational Biology* **17**:e1008846. DOI: https://doi.org/10.1371/journal.pcbi.1008846, PMID: 33831009

Kaneoke Y, Vitek JL. 1996. Burst and oscillation as disparate neuronal properties. *Journal of Neuroscience Methods* **68**:211–223. DOI: https://doi.org/10.1016/0165-0270(96)00081-7, PMID: 8912194

Kayser C, Logothetis NK, Panzeri S. 2010. Millisecond encoding precision of auditory cortex neurons. *PNAS* **107**:16976–16981. DOI: https://doi.org/10.1073/pnas.1012656107, PMID: 20837521

Khoshkhou M, Montakhab A. 2019. Spike-timing-dependent plasticity with axonal delay tunes networks of izhikevich neurons to the edge of synchronization transition with scale-free avalanches. *Frontiers in Systems Neuroscience* **13**:73. DOI: https://doi.org/10.3389/fnsys.2019.00073, PMID: 31866836

Kinouchi O, Copelli M. 2006. Optimal dynamical range of excitable networks at criticality. *Nature Physics* **2**:348–351. DOI: https://doi.org/10.1038/nphys289

Kozachenko L, Leonenko NN. 1987. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii* **23**:9–16.

Krahe R, Gabbiani F. 2004. Burst firing in sensory systems. *Nature Reviews. Neuroscience* **5**:13–23. DOI: https://doi.org/10.1038/nrn1296, PMID: 14661065

Kraskov A, Stögbauer H, Grassberger P. 2004. Estimating mutual information. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* **69**:066138. DOI: https://doi.org/10.1103/PhysRevE.69.066138, PMID: 15244698

Kunkel S, Diesmann M, Morrison A. 2011. Limits to the development of feed-forward structures in large recurrent neuronal networks. *Frontiers in Computational Neuroscience* **4**:160. DOI: https://doi.org/10.3389/fncom.2010.00160, PMID: 21415913

Li M, Han Y, Aburn MJ, Breakspear M, Poldrack RA, Shine JM, Lizier JT. 2019. Transitions in information processing dynamics at the whole-brain network level are driven by alterations in neural gain. *PLOS Computational Biology* **15**:e1006957. DOI: https://doi.org/10.1371/journal.pcbi.1006957, PMID: 31613882

Lisman JE. 1997. Bursts as a unit of neural information: making unreliable synapses reliable. *Trends in Neurosciences* **20**:38–43. DOI: https://doi.org/10.1016/S0166-2236(96)10070-9, PMID: 9004418

Lizier JT, Prokopenko M, Zomaya AY. 2008. Local information transfer as a spatiotemporal filter for complex systems. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* **77**:026110. DOI: https://doi.org/10.1103/PhysRevE.77.026110, PMID: 18352093

Lizier JT, Prokopenko M, Cornforth DJ. 2009. The information dynamics of cascading failures in energy networks. Proceedings of the European Conference on Complex Systems (ECCS). .

Lizier JT, Prokopenko M, Zomaya AY. 2010. Information modification and particle collisions in distributed computation. *Chaos (Woodbury, N.Y.)* **20**:037109. DOI: https://doi.org/10.1063/1.3486801, PMID: 20887075

Lizier JT, Heinzle J, Horstmann A, Haynes JD, Prokopenko M. 2011. Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fmri connectivity. *Journal of Computational Neuroscience* **30**:85–107. DOI: https://doi.org/10.1007/s10827-010-0271-2, PMID: 20799057

Lizier JT, Prokopenko M, Zomaya AY. 2012. Local measures of information storage in complex distributed computation. *Information Sciences* **208**:39–54. DOI: https://doi.org/10.1016/j.ins.2012.04.016

Lizier JT. 2013. The Local Information Dynamics of Distributed Computation in Complex Systems. Berlin, Heidelberg: Springer. DOI: https://doi.org/10.1007/978-3-642-32952-4

Lizier JT. 2014. Measuring the dynamics of information processing on a local scale in time and space. Wibral M, Vicente R, Lizier JT (Eds). *Directed Information Measures in Neuroscience*. Berlin/Heidelberg: Springer. p. 161–193. DOI: https://doi.org/10.1007/978-3-642-54474-3_7

Lizier JT, Prokopenko M, Zomaya AY. 2014. A framework for the local information dynamics of distributed computation in complex systems. Prokopenko M (Ed). *Guided Self-Organization: Inception*. Springer. p. 115–158. DOI: https://doi.org/10.1007/978-3-642-53734-9_5

MacKay DJ, Kay DJM. 2003. Information Theory, Inference and Learning Algorithms. Cambridge university press.

Maeda E, Robinson HP, Kawana A. 1995. The mechanisms of generation and propagation of synchronized bursting in developing networks of cortical neurons. *The Journal of Neuroscience* **15**:6834–6845. DOI: https://doi.org/10.1523/JNEUROSCI.15-10-06834.1995, PMID: 7472441

Mäki-Marttunen V, Diez I, Cortes JM, Chialvo DR, Villarreal M. 2013. Disruption of transfer entropy and inter-hemispheric brain functional connectivity in patients with disorder of consciousness. *Frontiers in Neuroinformatics* **7**:24. DOI: https://doi.org/10.3389/fninf.2013.00024, PMID: 24312048

Marinazzo D, Wu G, Pellicoro M, Angelini L, Stramaglia S. 2012. Information flow in networks and the law of diminishing marginal returns: evidence from modeling and human electroencephalographic recordings. *PLOS ONE* **7**:e45026. DOI: https://doi.org/10.1371/journal.pone.0045026, PMID: 23028745

Marinazzo D, Gosseries O, Boly M, Ledoux D, Rosanova M, Massimini M, Noirhomme Q, Laureys S. 2014a. Directed information transfer in scalp electroencephalographic recordings: insights on disorders of consciousness. *Clinical EEG and Neuroscience* **45**:33–39. DOI: https://doi.org/10.1177/1550059413510703, PMID: 24403318

Marinazzo D, Pellicoro M, Wu G, Angelini L, Cortés JM, Stramaglia S. 2014b. Information transfer and criticality in the ising model on the human connectome. *PLOS ONE* **9**:e93616. DOI: https://doi.org/10.1371/journal.pone.0093616, PMID: 24705627

Matsuda E, Mita T, Hubert J, Oka M, Bakkum D, Frey U, Takahashi H, Ikegami T, University of Tokyo. 2013. Multiple time scales observed in spontaneously evolved neurons on high-density cmos electrode array. European Conference on Artificial Life 2013. 1075–1082. DOI: https://doi.org/10.7551/978-0-262-31709-2-ch161

Mijatovic G, Antonacci Y, Loncar-Turukalo T, Minati L, Faes L. 2021. An information-theoretic framework to measure the dynamic interaction between neural spike trains. *IEEE Transactions on Bio-Medical Engineering* **68**:3471–3481. DOI: https://doi.org/10.1109/TBME.2021.3073833, PMID: 33872139

Miller RG. 2012. Simultaneous Statistical Inference. Springer Science & Business Media.

Minati L, Ito H, Perinelli A, Ricci L, Faes L, Yoshimura N, Koike Y, Frasca M. 2019. Connectivity influences on nonlinear dynamics in weakly-synchronized networks: Insights from rössler systems, electronic chaotic oscillators, model and biological neurons. *IEEE Access* **7**:174793–174821. DOI: https://doi.org/10.1109/ACCESS.2019.2957014

Nemenman I, Lewen GD, Bialek W, de Ruyter van Steveninck RR. 2008. Neural coding of natural stimuli: information at sub-millisecond resolution. *PLOS Computational Biology* **4**:e1000025. DOI: https://doi.org/10.1371/journal.pcbi.1000025, PMID: 18369423

Network. 2021. Network activity of developing cortical cultures in vitro. http://neurodatasharing.bme.gatech.edu/development-data/html/index.html [Accessed January 3, 2021].

Nigam S, Shimono M, Ito S, Yeh F-C, Timme N, Myroshnychenko M, Lapish CC, Tosi Z, Hottowy P, Smith WC, Masmanidis SC, Litke AM, Sporns O, Beggs JM. 2016. Rich-club organization in effective connectivity among cortical neurons. *The Journal of Neuroscience* **36**:670–684. DOI: https://doi.org/10.1523/JNEUROSCI.2177-15.2016, PMID: 26791200

Novelli L, Wollstadt P, Mediano P, Wibral M, Lizier JT. 2019. Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing. *Network Neuroscience (Cambridge, Mass.)* **3**:827–847. DOI: https://doi.org/10.1162/netn_a_00092, PMID: 31410382

Novelli L, Atay FM, Jost J, Lizier JT. 2020. Deriving pairwise transfer entropy from network structure and motifs. *Proceedings. Mathematical, Physical, and Engineering Sciences* **476**:20190779. DOI: https://doi.org/10.1098/rspa.2019.0779, PMID: 32398937

Novelli L, Lizier JT. 2021. Inferring network properties from time series using transfer entropy and mutual information: Validation of multivariate versus bivariate approaches. *Network Neuroscience (Cambridge, Mass.)* **5**:373–404. DOI: https://doi.org/10.1162/netn_a_00178, PMID: 34189370

Orlandi JG, Stetter O, Soriano J, Geisel T, Battaglia D. 2014. Transfer entropy reconstruction and labeling of neuronal connections from simulated calcium imaging. *PLOS ONE* **9**:e98842. DOI: https://doi.org/10.1371/journal.pone.0098842, PMID: 24905689

Pasquale V, Massobrio P, Bologna LL, Chiappalone M, Martinoia S. 2008. Self-organization and neuronal avalanches in networks of dissociated cortical neurons. *Neuroscience* **153**:1354–1369. DOI: https://doi.org/10.1016/j.neuroscience.2008.03.050, PMID: 18448256

Priesemann V, Munk MHJ, Wibral M. 2009. Subsampling effects in neuronal avalanche distributions recorded in vivo. *BMC Neuroscience* **10**:40. DOI: https://doi.org/10.1186/1471-2202-10-40, PMID: 19400967

Priesemann V, Valderrama M, Wibral M, Le Van Quyen M. 2013. Neuronal avalanches differ from wakefulness to deep sleep–evidence from intracranial depth recordings in humans. *PLOS Computational Biology* **9**:e1002985. DOI: https://doi.org/10.1371/journal.pcbi.1002985, PMID: 23555220

Priesemann V, Wibral M, Valderrama M, Pröpper R, Le Van Quyen M, Geisel T, Triesch J, Nikolić D, Munk MHJ. 2014. Spike avalanches in vivo suggest a driven, slightly subcritical brain state. *Frontiers in Systems Neuroscience* **8**:108. DOI: https://doi.org/10.3389/fnsys.2014.00108, PMID: 25009473

Rubinov M, Sporns O, Thivierge JP, Breakspear M. 2011. Neurobiologically realistic determinants of self-organized criticality in networks of spiking neurons. *PLOS Computational Biology* **7**:e1002038. DOI: https://doi.org/10.1371/journal.pcbi.1002038, PMID: 21673863

Rudelt L, González Marx D, Wibral M, Priesemann V. 2021. Embedding optimization reveals long-lasting history dependence in neural spiking activity. *PLOS Computational Biology* **17**:e1008927. DOI: https://doi.org/10.1371/journal.pcbi.1008927, PMID: 34061837

Schreiber T. 2000. Measuring information transfer. *Physical Review Letters* **85**:461–464. DOI: https://doi.org/10.1103/PhysRevLett.85.461, PMID: 10991308

Schroeter MS, Charlesworth P, Kitzbichler MG, Paulsen O, Bullmore ET. 2015. Emergence of rich-club topology and coordinated dynamics in development of hippocampal functional networks in vitro. *Journal of Neuroscience* **35**:5459–5470. DOI: https://doi.org/10.1523/JNEUROSCI.4259-14.2015

Selinger JV, Kulagina NV, O'Shaughnessy TJ, Ma W, Pancrazio JJ. 2007. Methods for characterizing interspike intervals and identifying bursts in neuronal activity. *Journal of Neuroscience Methods* **162**:64–71. DOI: https://doi.org/10.1016/j.jneumeth.2006.12.003, PMID: 17258322

Shapiro SS, Wilk MB. 1965. An analysis of variance test for normality (complete samples)). *Biometrika* **52**:591. DOI: https://doi.org/10.2307/2333709

Shew WL, Yang H, Yu S, Roy R, Plenz D. 2011. Information capacity and transmission are maximized in balanced cortical networks with neuronal avalanches. *The Journal of Neuroscience* **31**:55–63. DOI: https://doi.org/10.1523/JNEUROSCI.4637-10.2011, PMID: 21209189

Shimono M, Beggs JM. 2015. Functional clusters, hubs, and communities in the cortical microconnectome. *Cerebral Cortex (New York, N.Y)* **25**:3743–3757. DOI: https://doi.org/10.1093/cercor/bhu252, PMID: 25336598

Shorten DP, Spinney RE, Lizier JT, Marinazzo D. 2021. Estimating transfer entropy in continuous time between neural spike trains or other event-based data. *PLOS Computational Biology* **17**:e1008054. DOI: https://doi.org/10.1371/journal.pcbi.1008054

Shorten D. 2022. cell_cultures. Software Heritage. https://archive.softwareheritage.org/swh:1:dir:c37a7f336c6ead1deb9bf62f0b5202014a51384c;origin=https://bitbucket.org/dpshorten/cell_cultures;visit=swh:1:snp:9afa324ac020f1c169cae2c8f220a29e1ff11375;anchor=swh:1:rev:8ee5e519da5cb90590865e9a692b96ad7e68a69e

Shovon MHI, Nandagopal N, Vijayalakshmi R, Du JT, Cocks B. 2016. Directed connectivity analysis of functional brain networks during cognitive activity using transfer entropy. *Neural Processing Letters* **45**:807–824. DOI: https://doi.org/10.1007/s11063-016-9506-1

Sober SJ, Sponberg S, Nemenman I, Ting LH. 2018. Millisecond spike timing codes for motor control. *Trends in Neurosciences* **41**:644–648. DOI: https://doi.org/10.1016/j.tins.2018.08.010

Spinney RE, Lizier JT. 2018. Characterizing information-theoretic storage and transfer in continuous time processes. *Physical Review. E* **98**:012314. DOI: https://doi.org/10.1103/PhysRevE.98.012314, PMID: 30110808

Stetter O, Battaglia D, Soriano J, Geisel T. 2012. Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals. *PLOS Computational Biology* **8**:e1002653. DOI: https://doi.org/10.1371/journal.pcbi.1002653, PMID: 22927808

Stramaglia S, Wu GR, Pellicoro M, Marinazzo D. 2012. Expanding the transfer entropy to identify information circuits in complex systems. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* **86**:066211. DOI: https://doi.org/10.1103/PhysRevE.86.066211, PMID: 23368028

Timme N, Ito S, Myroshnychenko M, Yeh FC, Hiolski E, Hottowy P, Beggs JM. 2014. Multiplex networks of cortical and hippocampal neurons revealed at different timescales. *PLOS ONE* **9**:e115764. DOI: https://doi.org/10.1371/journal.pone.0115764, PMID: 25536059

Timme NM, Ito S, Myroshnychenko M, Nigam S, Shimono M, Yeh F-C, Hottowy P, Litke AM, Beggs JM, Pillow JW. 2016. High-degree neurons feed cortical computations. *PLOS Computational Biology* **12**:e1004858. DOI: https://doi.org/10.1371/journal.pcbi.1004858

Wagenaar D, DeMarse TB, Potter SM. 2005. Meabench: A toolset for multi-electrode data acquisition and on-line analysis. 2nd International IEEE EMBS Conference on Neural Engineering, 2005. 518–521. DOI: https://doi.org/10.1109/CNE.2005.1419673

Wagenaar DA, Nadasdy Z, Potter SM. 2006a. Persistent dynamic attractors in activity patterns of cultured neuronal networks. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* **73**:051907. DOI: https://doi.org/10.1103/PhysRevE.73.051907, PMID: 16802967

Wagenaar DA, Pine J, Potter SM. 2006b. An extremely rich repertoire of bursting patterns during the development of cortical cultures. *BMC Neuroscience* **7**:11. DOI: https://doi.org/10.1186/1471-2202-7-11, PMID: 16464257

Wibral M, Rahm B, Rieder M, Lindner M, Vicente R, Kaiser J. 2011. Transfer entropy in magnetoencephalographic data: quantifying information flow in cortical and cerebellar networks. *Progress in Biophysics and Molecular Biology* **105**:80–97. DOI: https://doi.org/10.1016/j.pbiomolbio.2010.11.006, PMID: 21115029

Wibral M, Lizier JT, Vögler S, Priesemann V, Galuske R. 2014a. Local active information storage as a tool to understand distributed neural information processing. *Frontiers in Neuroinformatics* **8**:1. DOI: https://doi.org/10.3389/fninf.2014.00001, PMID: 24501593

Wibral M, Vicente R, Lizier JT. 2014b. Directed Information Measures in Neuroscience. Berlin, Heidelberg: Springer. DOI: https://doi.org/10.1007/978-3-642-54474-3

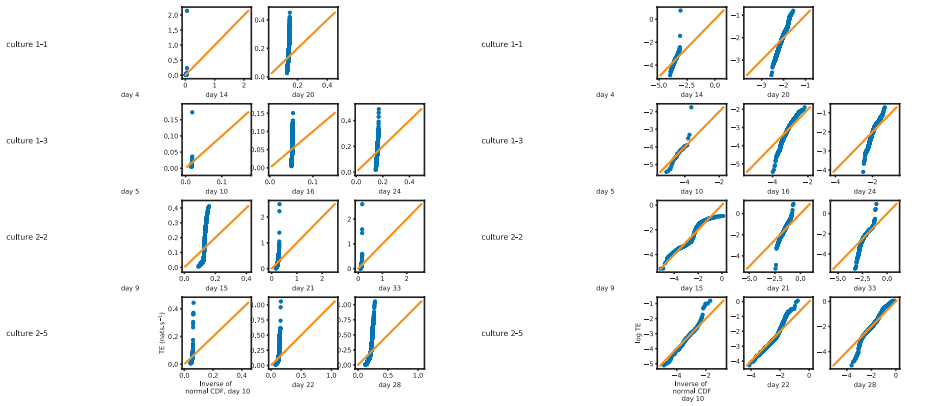
Wibral M, Finn C, Wollstadt P, Lizier JT, Priesemann V. 2017. Quantifying information modification in developing neural networks via partial information decomposition. *Entropy* **19**:494. DOI: https://doi.org/10.3390/e19090494

Wollstadt P, Lizier JT, Vicente R, Finn C, Martinez-Zarzuela M, Mediano P, Novelli L, Wibral M. 2019. Idtxl: The information dynamics toolkit xl: a python package for the efficient analysis of multivariate information dynamics in networks. *Journal of Open Source Software* **4**:1081. DOI: https://doi.org/10.21105/joss.01081, PMID: 30043789

Zalesky A, Fornito A, Cocchi L, Gollo LL, van den Heuvel MP, Breakspear M. 2016. Connectome sensitivity or specificity: which is more important *NeuroImage* **142**:407–420. DOI: https://doi.org/10.1016/j.neuroimage.2016.06.035, PMID: 27364472

Zeraati R, Priesemann V, Levina A. 2021. Self-organization toward criticality by synaptic plasticity. *Frontiers in Physics* **9**:103. DOI: https://doi.org/10.3389/fphy.2021.619661

Appendix 1



(a) QQ plots of TE values against the normal distribution. (b) QQ plots of log TE values against the normal distribution.

Appendix 1—figure 1. Quantile-quantile (QQ) plots (***Gibbons and Chakraborti, 2020***) of the nonzero estimated transfer entropy (TE) values against normal (**a**) and log-normal (**b**) distributions, respectively. The y axis shows estimated TE values (or their logarithm), whereas the x axis shows the value of the normal distribution at the same quantile. The solid orange line shows the line $y = x$. If the data is drawn from the distribution against which it is being plotted, then the blue marks will sit along this line. We observe that the distributions of TE values deviate substantially from both normal and log-normal distributions in all recordings analysed.

Appendix 1—table 1. p-Values for the Shapiro–Wilk test (***Shapiro and Wilk, 1965***) of normality for the distribution of transfer entropy (TE) values estimated in each recording.
Only the statistically significant TE values are included in these tests. Recordings for which there were no statistically significant values estimated are left blank. These p-values represent the probability that the associated test statistic is more extreme than that calculated on the estimated TE values, under the null hypothesis that these values are normally distributed. For any reasonable choice of p cutoff value, the null hypothesis is rejected in all recordings.

Culture 1-1	Day 4	Day 14	Day 20	
	–	9.8×10^{-45}	4.2×10^{-35}	
Culture 1-3	Day 5	Day 10	Day 16	Day 24
	4.4×10^{-13}	4.7×10^{-24}	2.7×10^{-36}	1.0×10^{-36}
Culture 2-2	Day 9	Day 15	Day 21	Day 33
	7.9×10^{-6}	1.6×10^{-28}	2.6×10^{-35}	9.5×10^{-38}
Culture 2-5	Day 4	Day 10	Day 22	Day 28
	–	7.5×10^{-10}	2.4×10^{-28}	3.7×10^{-29}

Appendix 1—table 2. p-Values for the Shapiro–Wilk test (***Shapiro and Wilk, 1965***) of normality for the distribution of transfer entropy (TE) values estimated in each recording.
Only the statistically significant TE values are included in these tests. Recordings for which there were no statistically significant values estimated are left blank. These p-values represent the probability that the associated test statistic is more extreme than that calculated on the estimated TE values, under the null hypothesis that these values are normally distributed. For any reasonable choice of p cutoff value, the null hypothesis is rejected in all recordings.

Culture 1-1	Day 4	Day 14	Day 20	

Appendix 1—table 2 Continued on next page

Appendix 1—table 2 Continued

	–	0	3.3×10^{-33}	
Culture 1-2	Day 6	Day 11	Day 17	
	–	2.3×10^{-21}	3.8×10^{-43}	
Culture 1-3	Day 5	Day 10	Day 16	Day 24
	–	6.4×10^{-13}	3.3×10^{-14}	5.3×10^{-14}
Culture 1-4	Day 8	Day 13	Day 19	
	–	2.5×10^{-28}	2.1×10^{-31}	
Culture 1-5	Day 7	Day 12	Day 18	
	8.4×10^{-2}	1×10^{-35}	1.5×10^{-12}	
Culture 2-1	Day 14	Day 32		
	2.8×10^{-1}	0		
Culture 2-2	Day 9	Day 15	Day 21	Day 33
	–	1.4×10^{-19}	7.1×10^{-44}	0
Culture 2-3	Day 6	Day 12	Day 24	
	–	2.6×10^{-1}	1.4×10^{-45}	
Culture 2-4	Day 3	Day 5	Day 11	
	3.5×10^{-12}	–	1.0×10^{-2}	
Culture 2-5	Day 4	Day 10	Day 22	Day 28
	–	1.2×10^{-21}	3.5×10^{-40}	1.2×10^{-37}
Culture 2-6	Day 7	Day 13	Day 31	
	1.0×10^{-1}	–	1.9×10^{-19}	–

Appendix 1—table 3. p-Values for the Shapiro–Wilk test (***Shapiro and Wilk, 1965***) of log-normality for the distribution of transfer entropy (TE) values estimated in each recording.

Only the statistically significant TE values are included in these tests. Recordings for which there were no statistically significant values estimated are left blank. These p-values represent the probability that the associated test statistic is more extreme than that calculated on the logarithms of the estimated TE values, under the null hypothesis that these values are normally distributed. For any reasonable choice of p cutoff value, the null hypothesis is rejected in all recordings (apart from those with very few significant TE values). It is interesting to note that the p-values are often smaller on later days, despite the Q-Q plots in ***Appendix 1—figure 1***, suggesting the distribution is closer to log-normal. This is probably due to there being many more statistically significant TE values on these later days (see ***Table 2***).

Culture 1-1	Day 4	Day 14	Day 20	
	–	3.1×10^{-21}	2.9×10^{-10}	
Culture 1-2	Day 6	Day 11	Day 17	
	–	3.2×10^{-14}	1.6×10^{-22}	
Culture 1-3	Day 5	Day 10	Day 16	Day 24
	–	1.3×10^{-4}	3.0×10^{-13}	1.3×10^{-22}
Culture 1-4	Day 8	Day 13	Day 19	
	–	3.0×10^{-15}	2.4×10^{-7}	
Culture 1-5	Day 7	Day 12	Day 18	
	3.3×10^{-2}	7.8×10^{-24}	2.0×10^{-4}	

Appendix 1—table 3 Continued on next page

Appendix 1—table 3 Continued

Culture 2-1	Day 14	Day 32		
	9.7×10^{-1}	3.6×10^{-22}		
Culture 2-2	Day 9	Day 15	Day 21	Day 33
	–	1.8×10^{-12}	3.6×10^{-14}	5.8×10^{-29}
Culture 2-3	Day 6	Day 12	Day 24	
	–	6.1×10^{-2}	1.78×10^{-7}	
Culture 2-4	Day 3	Day 5	Day 11	
	9.8×10^{-13}	–	5.1×10^{-2}	
Culture 2-5	day 4	day 10	Day 22	Day 28
	-	1.2×10^{-3}	1.1×10^{-16}	2.4×10^{-14}
Culture 2-6	day 7	day 13	Day 31	
	7.4×10^{-1}	-	1.9×10^{-14}	–

Previous studies have placed an emphasis on the observation of log-normal distributions of TE values in in vitro cultures of neurons (***Shimono and Beggs, 2015***; ***Nigam et al., 2016***). As such, we analysed the distribution of the nonzero (statistically significant) estimated TE values in each individual recording.

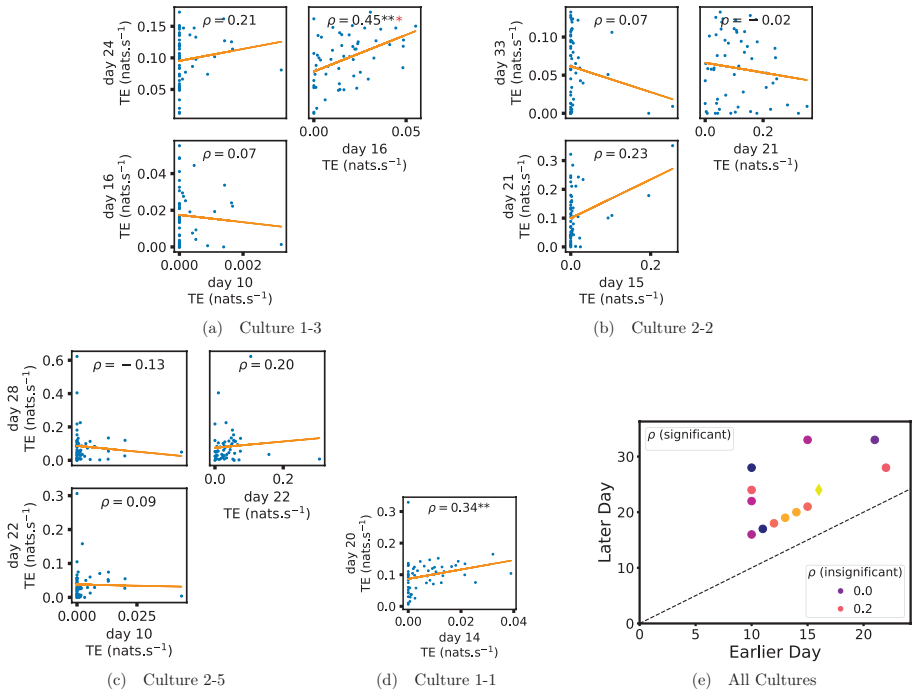
Figure 1 shows histograms as well as probability density functions estimated by a kernel density estimator (KDE) of the nonzero TE values for each recording. From these plots, we can see that the distributions of TE values exhibit a clear right (positive) skew. In order to ascertain how well the estimated TE values were described by a log-normal distribution, we constructed quantile-quantile (QQ) plots (***Gibbons and Chakraborti, 2020***) for the TE values against the log-normal distribution in ***Appendix 1—figure 1***. In all recordings, the plotted points deviate from the line $y = x$, indicating that the data is not well described by a log-normal distribution. However, this deviation appears only slight for some recordings, most notably days 22 and 28 of culture 2-5. We also perform Shapiro–Wilk tests (***Shapiro and Wilk, 1965***) for log-normality, the resulting p-values are displayed in ***Appendix 1—table 3***. The p-values for every recording are incredibly low, meaning that we reject the null hypothesis of a log-normal distribution in every case.

Given that the distributions of the TE values were not well described by a log-normal distribution, we investigated the alternative that they could be described by a normal distribution. ***Appendix 1—figure 1*** displays QQ plots (***Gibbons and Chakraborti, 2020***) for the TE values against the normal distribution. In all recordings, the plotted points deviate substantially from the line $y = x$, indicating that the data is poorly described by a normal distribution. We also perform Shapiro–Wilk tests (***Shapiro and Wilk, 1965***) for normality, the resulting p-values are displayed in ***Appendix 1—table 1***. The p-values for every recording are incredibly low, meaning that we reject the null hypothesis of a normal distribution in every case.

These results contrast with observation of log-normal distributions of TE values in in vitro cultures of neurons (***Shimono and Beggs, 2015***; ***Nigam et al., 2016***). The difference may be due to the use of continuous-time estimator here in contrast to the discrete-time estimator used in previous studies. This estimator is more faithful to capturing the true underlying TE for spike trains (as per ***Shorten et al., 2021***); however, it may be that the combination of the discrete-time estimator and use of only a single previous time-bin – in specifically *not* representing history dependence well – align more strongly with the component of the statistical relationship that follows a log-normal distribution. It is also possible that log-normal distributions of TE emerge later in development and are simply not yet present in the early developmental stages observed here (noting that the fit to a log-normal distribution seems to improve for later DIV in ***Appendix 1—figure 1***).

Appendix 2

Plots for early lock-in of incoming TE



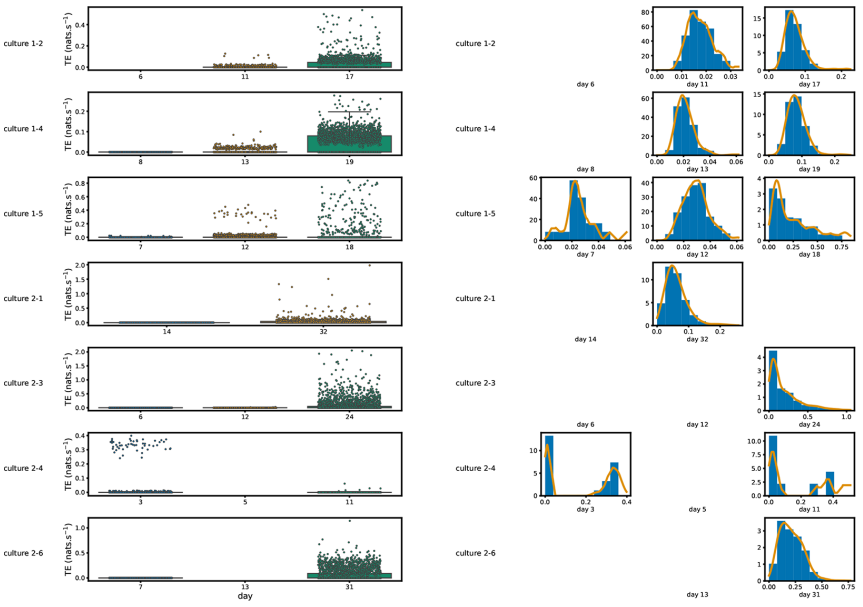
Appendix 2—figure 1. Plots investigating the relationship between the inward information flow from a given node over different days of development. (**a**–**d**) show scatter plots between all pairs of days for each culture (excluding days with zero significant transfer entropy [TE] values). Specifically, in each scatter plot, the x value of a given point is the average inward TE from the associated node on an earlier day and the y value of that same point is the total outgoing TE from the same node but on a later day. The days in question are shown on the bottom and sides of the grids of scatter plots. The orange line shows the ordinary least squares regression. The Spearman correlation (ρ) between the outgoing TE values on the two days is displayed in each plot. Values of ρ significant at the 0.05 level are designated with an asterisk and those significant at the 0.01 level are designated with a double asterisk. A Bonferroni correction for multiple comparisons was used. (**e**) shows all recording day pairs for all cultures (where the pairs are always from the same culture) and the associated Spearman correlation between the outward TEs of nodes across this pair of recording days. Diamonds indicate significance at $p<0.05$, with Bonferroni correction.

Appendix 3

Extra cultures

Appendix 3—table 1. Mean transfer entropy (TE) in nats per second between every source-target pair for the additional cultures.

Culture 1-2	Day 6	Day 11	Day 17
	0	6.6×10^{-4}	0.023
Culture 1-4	Day 8	Day 13	Day 19
	0	0.017	0.040
Culture 1-5	Day 7	Day 12	Day 18
	5.6×10^{-5}	6.6×10^{-3}	0.016
Culture 2-1	Day 14	Day 32	
	2.9×10^{-6}	0.028	
Culture 2-3	Day 6	Day 12	Day 24
	0	2.0×10^{-5}	0.075
Culture 2-4	Day 3	Day 5	Day 11
	6.4×10^{-3}	0	5.3×10^{-5}
Culture 2-6	Day 7	Day 13	Day 31
	0	0	0.061



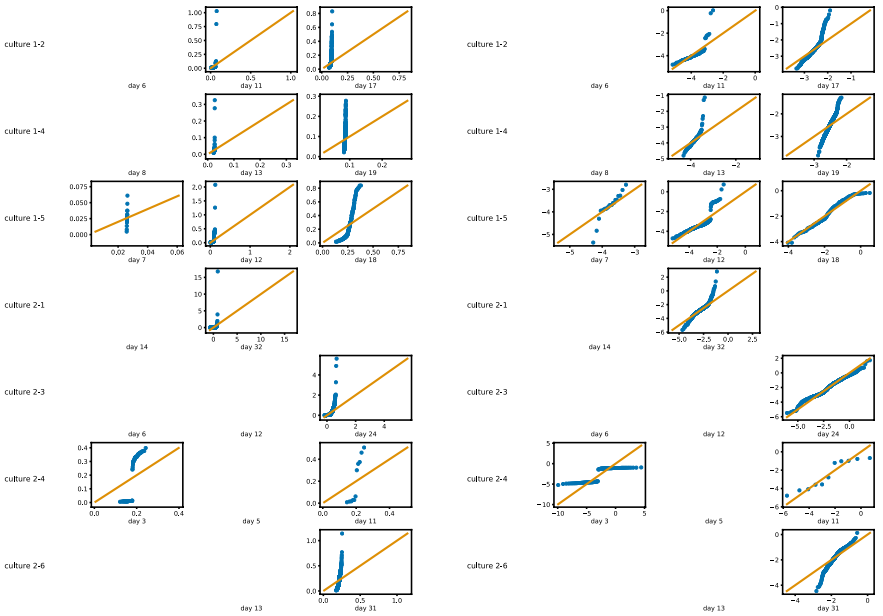
(a) Scatters and boxplots of TE values. (b) Histograms and kernel density estimates of TE values.

Appendix 3—figure 1. Identical plots to those shown in ***Figure 1***, but showing the cultures left out of that plot for space considerations. (**a**) Scatters of the TE values are overlaid on box plots. The box plots show the quartiles and the median (values greater than 10 standard deviationSDs from the mean have been removed from both the box and scatter plots as outliers). (**b**) Density estimates of the nonzero (statistically significant) TE distribution on top of a histogram. The densities are estimated using a Gaussian kernel. The histogram bin width and kernel

histogram are both 10% of the data range. Recordings with fewer than 10 statistically significant TE values are excluded.

Appendix 3—table 2. Displays the same information as ***Figure 2***, but for the additional cultures.

Culture 1-2	Day 6	Day 11	Day 17
	0	105	860
Culture 1-4	Day 8	Day 13	Day 19
	1	214	1457
Culture 1-5	Day 7	Day 12	Day 18
	21	375	195
Culture 2-1	Day 14	Day 32	
	5	1165	
Culture 2-3	Day 6	Day 12	Day 24
	2	9	1000
Culture 2-4	Day 3	Day 5	Day 11
	97	0	11
Culture 2-6	Day 7	Day 13	Day 31
	9	0	873

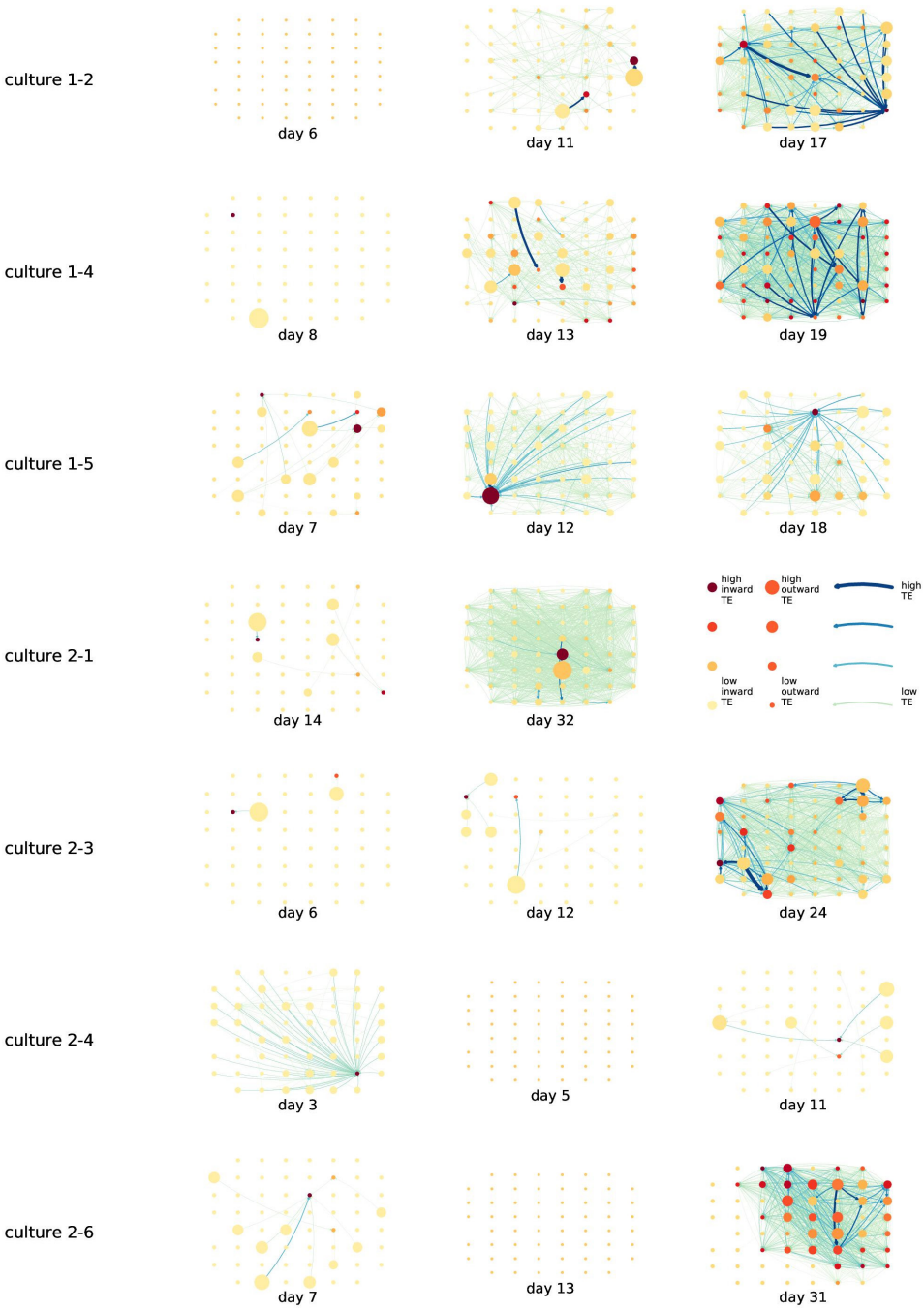


(a) QQ plots of TE values against the normal distribution. (b) QQ plots of log TE values against the normal distribution.

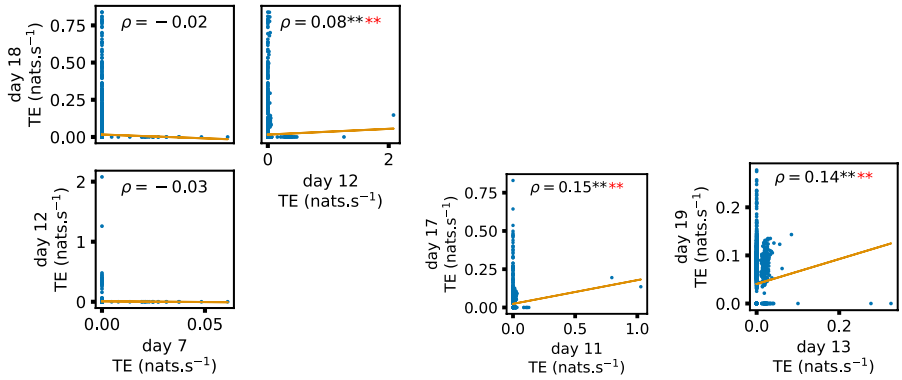
Appendix 3—figure 2. Identical plots to those shown in ***Appendix 2—figure 1***, but for the additional cultures. The y axis shows estimated TE values (or their logarithm), whereas the axis shows the value of the normal distribution at the same quantile. The solid orange line shows the line $y = x$. If the data is drawn from the

Appendix 3—figure 2 continued on next page

Appendix 3—figure 2 continued
distribution against which it is being plotted, then the blue marks will sit along this line. We observe that the distributions of TE values deviate substantially from both normal and log-normal distributions in all recordings analysed.

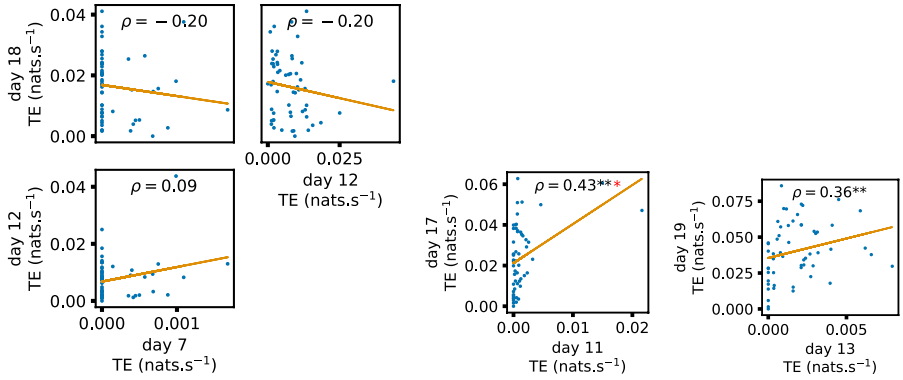


Appendix 3—figure 3. Identical plots to those in ***Figure 2***, but for the additional cultures.



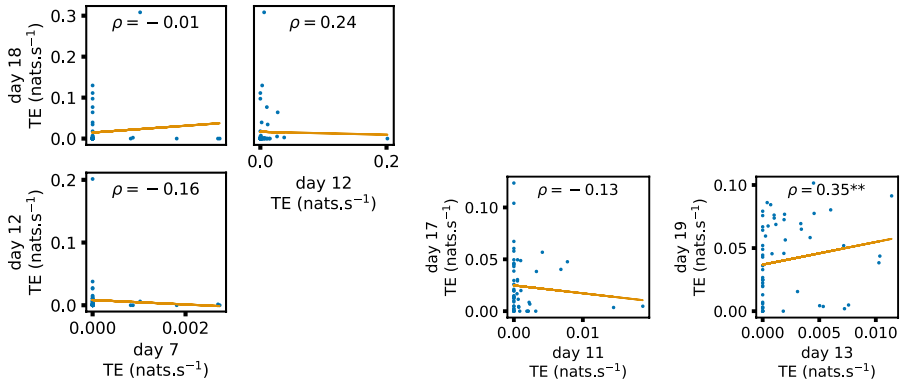
(a) Culture 1-5 (b) Culture 1-2 (c) Culture 1-4

Appendix 3—figure 4. Identical plots to those in ***Figure 3***, but for the additional cultures. (a) Contains plots for culture 1-5, (b) contains plots for culture 1-2 and (c) contains plots for culture 1-4.



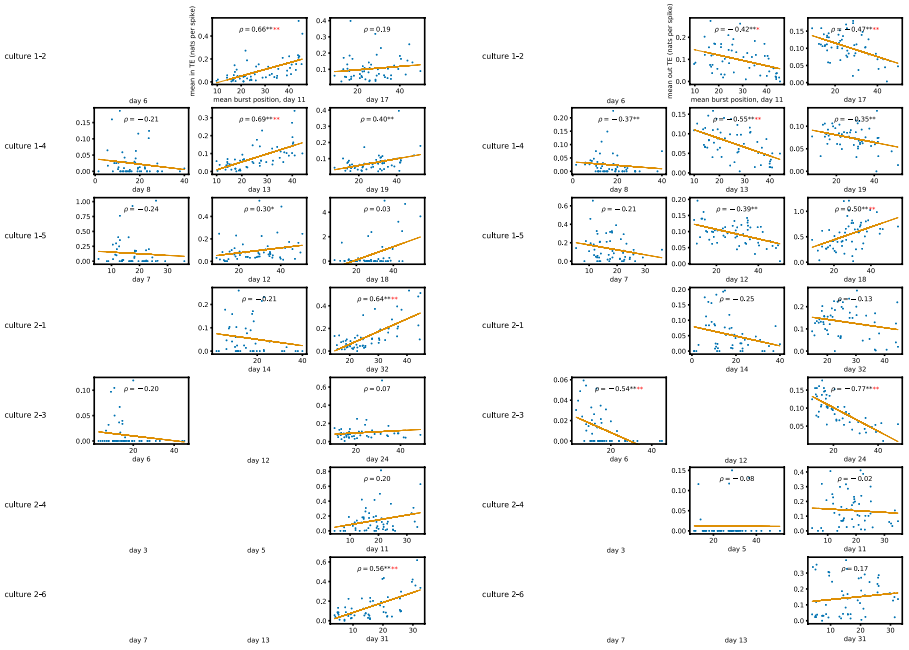
(a) Culture 1-5 (b) Culture 1-2 (c) Culture 1-4

Appendix 3—figure 5. Identical plots to those in ***Figure 4***, but for the additional cultures. (**a**) Contains plots for culture 1-5, (**b**) contains plots for culture 1-2 and (**c**) contains plots for culture 1-4.



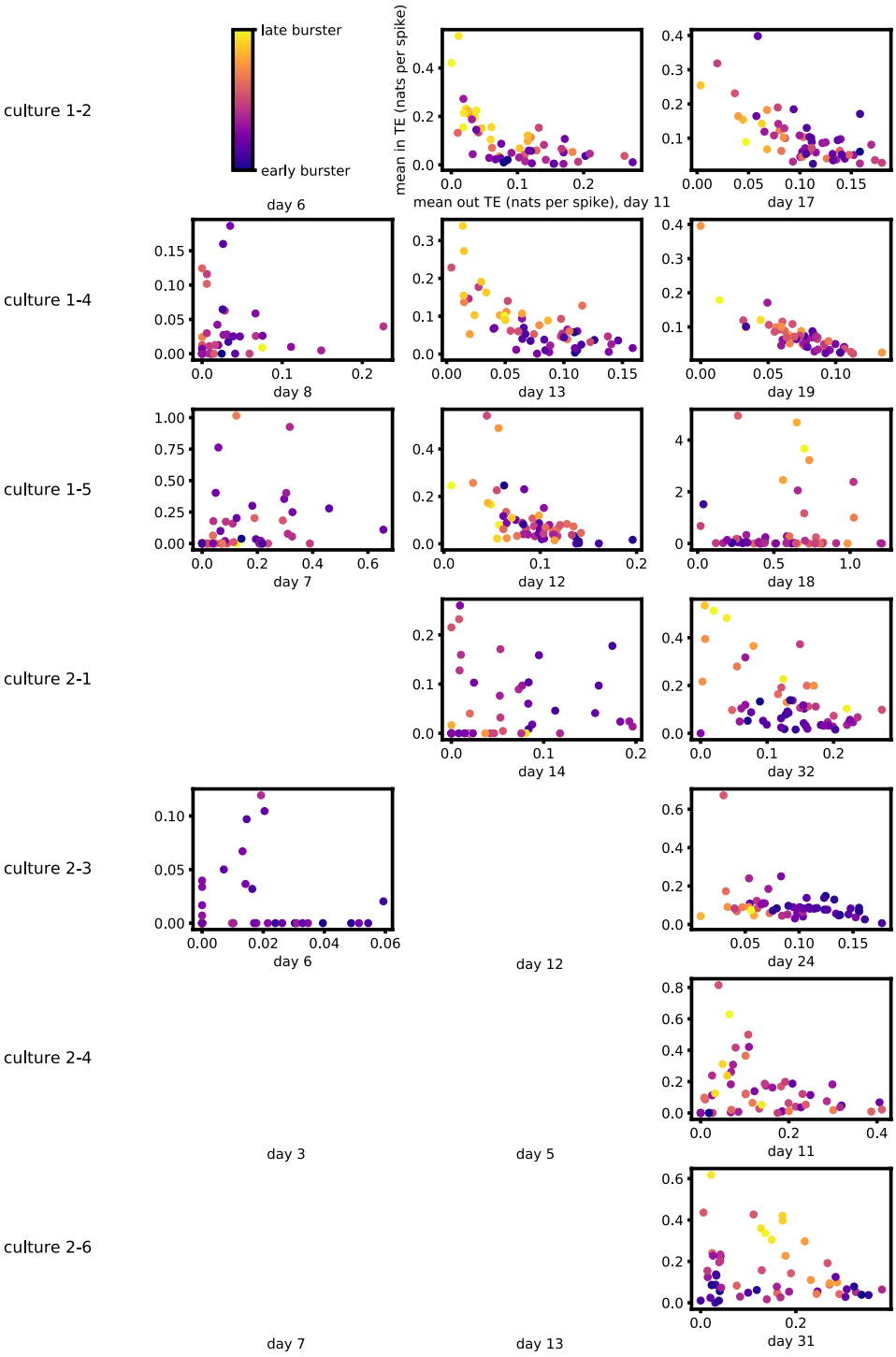
(a) Culture 1-5 (b) Culture 1-2 (c) Culture 1-4

Appendix 3—figure 6. Identical plots to those in ***Appendix 2—figure 1***, but for the additional cultures. (**a**) Contains plots for culture 1-5. (**b**) Contains plots for culture 1-2. (**c**) Contains plots for culture 1-4.

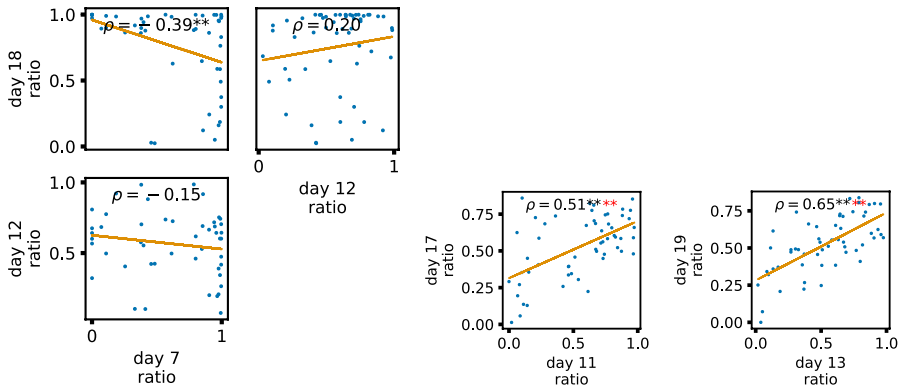


(a) Burst position vs TEin (b) Burst position vs TEout

Appendix 3—figure 7. Identical plots to those in ***Figure 5a and b***, but for the additional cultures. (**a**) Plots the mean burst position against the total incoming TE. (**b**) Plots the mean burst position against the total outgoing TE.



Appendix 3—figure 8. Identical plots to those in ***Figure 5c***, but for the additional cultures.



(a) Culture 1-5 (b) Culture 1-2 (c) Culture 1-4

Appendix 3—figure 9. Identical plots to those in ***Figure 6***, but for the additional cultures. (**a**) Contains plots for culture 1-5. (**b**) Contains plots for culture 1-2. (**c**) Contains plots for culture 1-4.

Chapter 5

Network Inference

TE is a particularly attractive technique to use for the inference of effective networks from the recordings of spiking neurons. This stems from the fact that the underlying estimation techniques which allow for the estimation of this quantity are non-parametric [1], [2]. This implies that they do not rely on any given model for the underlying system, but can rather detect any statistical relationship. Network inference of spiking neurons can be performed using biophysical models of how neurons spike [3], [4]. However, all such models are merely approximations of the actual behaviour of neurons [5]. It is unclear how the simplifying assumptions in these models affect the resulting network inference.

When inferring effective networks using TE, we aim to find a minimal set of source nodes which together provide the maximum reduction in the uncertainty of the target [6]–[8]. The inference of effective networks using TE requires an estimator that can reliably estimate conditional TE values even in the case of large sets of conditioning processes. The traditionally-used discrete-time estimator for TE on spike trains requires the use of multiple time bins in the history embedding of each process in order to be able to capture that history both over a reasonable span of time as well as with decent time precision. This makes it very challenging to use this estimator for conditional TE estimation with reasonably large conditioning sets, as the addition of multiple history embeddings, each composed of many individual bins, causes the dimensionality of the estimation task to become impractically large

By constructing history embeddings using the raw inter-spike intervals, the estimator presented in Chapter 3 is capable of capturing effects which occur over fairly large time intervals, with no loss of precision and using very few embedding dimensions. This opens up the possibility of using it for the inference of effective networks, as the dimensionality of the estimation task will be growing more slowly with the addition of each extra conditioning process, as compared to when the standard time-binning estimator is used. Moreover, Chapter 3 also presented an adaptation of a recently-proposed local permutation scheme for the generation of the surrogates necessary for performing significance tests. It was demonstrated that this method was much superior to the traditional time-shift method, especially when a target was strongly pairwise dependent, but conditionally independent of a source.

These substantial improvements in our ability to estimate pairwise and conditional TE on spike trains imply that it is now feasible to infer effective networks on them. This chapter validates this ability by utilising the novel estimator in conjunction with a slightly modified version of a pre-existing greedy TE-based effective network inference algorithm [6], [9]. The author is only aware of a single, very recent, piece of work which performs effective network inference on spiking data [10].

Moreover, this is the first study which validates effective network inference to spike-train data using TE by comparing the inferred networks against a known ground truth. We validate the estimator on simulated networks for which the ground-truth is known, achieving high accuracy at relatively low spike train lengths. We also inferred the effective networks from the spikes of neural cell cultures, demonstrating the use of this technique on biological data.

[1] T. Bossomaier, L. Barnett, M. Harré, and J. T. Lizier, "An introduction to transfer entropy," *Cham: Springer International Publishing*, vol. 65, 2016.

[2] M. Wibral, R. Vicente, and J. T. Lizier, *Directed information measures in neuroscience*. Springer, 2014.

[3] L. Paninski, J. Pillow, and J. Lewi, "Statistical models for neural encoding, decoding, and optimal stimulus design," *Progress in Brain Research*, vol. 165, pp. 493–507, 2007.

[4] I. M. de Abril, J. Yoshimoto, and K. Doya, "Connectivity inference from neural recording data: Challenges, mathematical bases and research directions," *Neural Networks*, vol. 102, pp. 120–137, 2018.

[5] E. R. Kandel, J. H. Schwartz, T. M. Jessell, S. Siegelbaum, A. J. Hudspeth, and S. Mack, *Principles of neural science*. McGraw-hill New York, 2000, vol. 4.

[6] L. Novelli, P. Wollstadt, P. Mediano, M. Wibral, and J. T. Lizier, "Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing," *Network Neuroscience*, vol. 3, no. 3, pp. 827–847, 2019.

[7] L. Novelli and J. T. Lizier, "Inferring network properties from time series using transfer entropy and mutual information: Validation of multivariate versus bivariate approaches," *Network Neuroscience*, vol. 5, no. 2, pp. 373–404, 2021.

[8] M. H. I. Shovon, N. Nandagopal, R. Vijayalakshmi, J. T. Du, and B. Cocks, "Directed connectivity analysis of functional brain networks during cognitive activity using transfer entropy," *Neural Processing Letters*, vol. 45, no. 3, pp. 807–824, 2017.

[9] J. Sun, D. Taylor, and E. M. Bollt, "Causal network inference by optimal causation entropy," *SIAM Journal on Applied Dynamical Systems*, vol. 14, no. 1, pp. 73–106, 2015.

[10] P. C. Antonello, T. F. Varley, J. Beggs, M. Porcionatto, O. Sporns, and J. Faber, "Self-organization of in vitro neuronal assemblies drives to complex network topology," *bioRxiv*, 2021.

Inferring effective networks of spiking neurons using a continuous-time estimator of transfer entropy

David P. Shorten,[1,*] Viola Priesemann,[2] Michael Wibral,[3] and Joseph T. Lizier[1,†]

[1] *Centre for Complex Systems, Faculty of Engineering, The University of Sydney, Sydney, Australia*

[2] *Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany*

[3] *Campus Institute for Dynamics of Biological Networks, Georg August University, Göttingen, Germany*

When analysing high-dimensional time-series datasets, the inference of effective networks has proven to be a valuable modelling technique. This technique produces networks where each target node is associated with a set of source nodes that are capable of providing explanatory power for its dynamics. Multivariate Transfer Entropy (TE) has proven to be a popular and effective tool for inferring these networks. Recently, a continuous-time estimator of TE for event-based data such as spike trains has been developed which, in more efficiently representing event data in terms of inter-event intervals, is significantly more capable of measuring multivariate interactions. The new estimator thus presents an opportunity to use TE for the inference of effective networks from spike trains, and we demonstrate in this paper for the first time its efficacy at this task. Using data generated from models of spiking neurons — for which the ground-truth connectivity is known — we demonstrate the accuracy of this approach in various dynamical regimes. We further show that it exhibits far superior performance to a pairwise TE-based approach at inference as well as a recently-proposed convolutional neural network approach. Moreover, comparison with Generalised Linear Models (GLMs), which are commonly applied to spike-train data, showed clear benefits, particularly in cases of high synchrony. Finally, we demonstrate its utility in gleaning insight from recordings of *in vitro* spiking neurons.

I. INTRODUCTION

For many of the complex systems that scientists are most interested in, our ability to record high-fidelity data from the numerous components of these systems is improving rapidly. For instance, the number of biological neurons that can be simultaneously recorded from is increasing exponentially, with a doubling rate of around six to seven years [1, 2], while the spatial resolution at which neural electrical activity can be recorded continues to increase dramatically[3]. The process of drawing scientific insight from this flood of data is, however, not always straightforward [4].

The inference of effective networks [5] from high-dimensional time-series data has become a popular and productive technique for reducing the complexity of this class of data. Such data sets often consist of millions of (or far more) individual data points [6]. The inference of effective networks aims to produce a minimal model of the data, by finding

* david.shorten@sydney.edu.au
† joseph.lizier@sydney.edu.au

the smallest set of system source components capable of explaining the activity of each target component [7]. As such, it compresses the large number of data points down to a single directed network diagram describing the relationship between components of the system, thus facilitating the interrogation of the data at hand.

There are different philosophical approaches to the inference of these networks. These include: uncovering causal relationships [8], inferring the coupling parameters in models faithful to the underlying system [9] or delineating the computational properties of the system by revealing information flows [7, 10]. When the latter approach is properly grounded in information theory, it provides the unique advantage of giving us networks that are readily interpretable in terms of the fundamental computational operations of information storage, transfer and modification [7, 11]. Moreover, as these measures can be estimated non-parametrically [12], they are not dependent on model assumptions and can capture any form of non-linear relationship.

As Transfer Entropy (TE) [13, 14] is a widely accepted measure of information transfer, it has become a popular method in the inference of effective networks [15]. When applying transfer entropy to network inference, we aim to establish a minimal set of parents whose activity is able to maximally explain the dynamic updates of each target node. This set is minimal in the sense that the addition of any extra parents will not further decrease our uncertainty about the state of the target. On the other hand, it provides maximal explainability in the sense that the removal of any parent will increase our uncertainty.

The challenge, then, is to infer this minimal set. An approach which has proven effective is to iteratively add sources to each target in a greedy fashion [7, 10, 16–18]. Specifically, for each target process, we select the source with the strongest information flow (without any conditioning). We then select the next source as the component with the highest information flow when conditioned on the first source and add this new source to the conditioning set. We continue adding sources to the conditioning set in this fashion until we are unable to find a source with a statistically significant non-zero information flow. The process then finishes with a pruning step, where it is verified that each source still has a non-zero information flow when conditioned on all other sources in the set. See Methods for more details.

In this work, we specifically focus on the inference of effective networks for event-based data. Such data is characterised by being represented by the timestamps of events (e.g.: the times of social-media posts or the times of stock-market trades), as opposed to regular samples from a continuously varying signal. This type of data is of particular importance in neuroscience as the activity of neurons is often summarised by the timestamps of their action potentials (spikes). There have been several previous studies which have proposed TE-based methods for inferring networks from the spike times of neurons and evaluated them against ground truth [19–22]. There have also been a number of studies which used TE to infer networks from *in vitro* [23–29] and *in vivo* [30] recordings of spiking activity. These networks were found to exhibit highly non-random structure [24], including rich-club topologies [23]. All of this work has estimated the TE in a pairwise fashion, that is, without conditioning on other recorded processes. Networks inferred based on pairwise statistics are often referred to as *functional* (as opposed to *effective*) networks [5], since they are reporting pairwise relationships rather than a minimal multivariate directed model of the dynamics.

The main obstacle that has prevented the inference of effective networks from spike trains using TE has been the manner in which the traditional method of estimating TE on spike train data causes a rapid increase in the dimensionality as we add conditioning processes [31]. This traditional method operates by first discretising the process into time bins. The TE is then estimated on the resulting binary sequences. The estimation of TE requires the use of

embedding vectors to represent histories for the target, source and conditioning processes in the relevant conditional probability distributions for the target process. In order for these embeddings to both extend over a reasonable period of time and also capture fine subtleties in event timings, each embedding vector needs to consist of multiple time bins. Capturing effects occurring on both fine and large time scales is necessary as it is known that correlations in spike trains exhist over distances of (at least) hundreds of milliseconds [32, 33]. Moreover, it is established that correlations at the millisecond and sub-millisecond scale play a role in neural function [34–37]. The use of multi-bin embedding vectors causes an explosion in the dimensionality of the state space over which probability distributions needs to be estimated as conditioning processes are added, rendering the estimation of TE with substantial sets of conditioning processes infeasible.

Recent work has developed a continuous-time estimator of TE for event-based data [31] which bypasses the embedding dimension problems of the discrete-time estimator. Specifically, as it uses inter-spike intervals to efficiently represent the history embeddings, it is capable of using embeddings that extend over relatively long periods of time (on the order of seconds [38]), with no loss of time precision. This makes the estimation of TE with significant conditioning sets feasible, thus allowing for the inference of effective networks.

In this paper, we bring the greedy network inference algorithm and the continuous-time TE estimator together for the first time. We validate the efficacy of this combination on synthetic examples, where the underlying causal network is recovered by the model. We further compare its efficacy against generalised linear models [39] and a convolutional neural network based approach [40], finding its performance to be highly competitive. We finally demonstrate its ability to uncover biological insight by inferring the effective networks of developing cell cultures of dissociated cortical rate neurons [41].

II. RESULTS

In this section we apply the greedy TE-based effective network inference algorithm [7] in conjunction with the continuous-time TE estimator for event-based data [31]. Please see Sec. IV A for details of the operation of the greedy algorithm, along with a description of a few minor changes that were made for the application to event-based data. Sec. IV C summarises the TE estimation approach used.

The first three subsections of this section focus on the inference of simulated spiking networks for which the ground truth connectivity is known. We must emphasize, however, that in general we do not expect the effective networks inferred by TE to align with the causal structure [7]. Whilst the effective networks always provide a useful model for interpreting the directed relationships in the system, it is only under certain specific conditions that we expect them to match the causal structure, most importantly full observability of the nodes involved in the dynamics, and under certain assumptions such as faithfulness and the causal Markov property [42, 43]. We evaluate the performance of the network inference scheme by comparing the inferred network to this ground truth under these idealised conditions, since this provides an important validation of the output of the inference when this match can be expected. In order to measure the accuracy of the inference scheme, we make use of the commonly-employed classification metrics of *recall* and *precision*. They are defined as:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

4

and

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{2}$$

Here, TP is the number of true positives, FP is the number of false positives and FN is the number of false negatives. In the context of network inference, recall can be interpreted as the proportion of true connections that were predicted by the algorithm. Precision, on the other hand, is the proportion of predicted edges that are true edges.

The final subsection of the results focuses on the application of the estimator to the spike times from recordings of cultures of dissociated rat cortical neurons. This provides a demonstration of the utility of this approach for extracting insights from biological data.

A. Inference at varying levels of synchrony

We constructed networks of Leaky-Integrate-and-Fire (LIF) [44] neurons with alpha synapses [45]. These networks were composed of 30 excitatory and 20 inhibitory neurons. Each neuron had exactly three excitatory and two inhibitory sources, where these sources were selected randomly from the respective sets. By varying the ratio of inhibitory to excitatory connection strength g, we could vary the level of synchrony within the networks. We ran simulations for three different levels of synchrony, which we refer to as "low" ($g = 3$), "medium" ($g = 1.5$) and "high" ($g = 1$). Varying the relative strength of inhibitory connections across the excitation-inhibition balance threshold is a known method for adjusting the degree of synchrony in these networks [46]. Please see Sec. IV E for full details on these network models. It is also worth noting that the level of synchrony present in these networks (even at "high" synchrony) is far lower than in the biological data examined in Sec. II D.

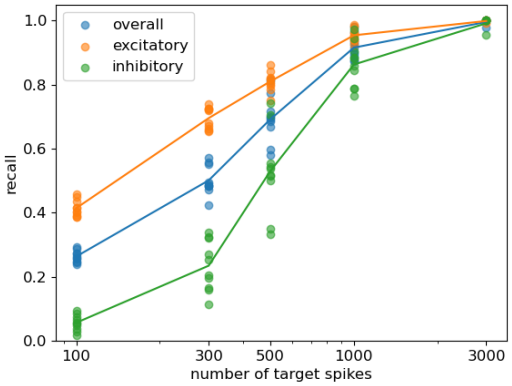
The combination of the greedy inference algorithm and the continuous-time estimator was applied to these networks, and the resulting inferred networks were compared against the ground truth for varying numbers of target spikes available to the estimator: 100, 300, 500, 1000, and 3000 (extra runs at 5000 target spikes were included for the high synchrony network, as the recall rose more slowly in this case). 10 independent simulations of the network model were run for each level of synchrony and the algorithm was applied to each run for each number of target spikes, although it was only applied to the first 5 simulations at 3000 spikes and the first 3 simulations at 5000 spikes due to the high computational requirements.

The precision and recall of the resulting inferences was calculated and is plotted in Fig. 1 for the different dynamical regimes and numbers of target spikes.

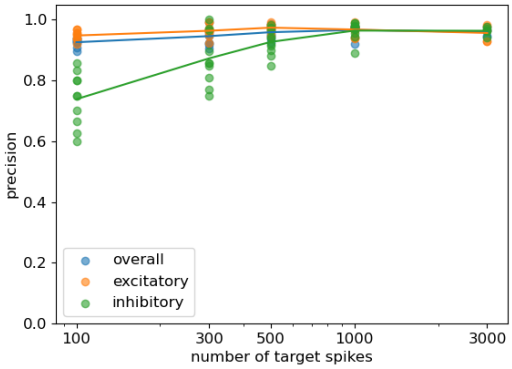
In the results shown in Fig. 1, we see that the algorithm exhibits high precision for all combinations of dynamical regime and number of target spikes. The precision only drops below 0.9 where the recall is very low, when few links are inferred. This demonstrates the high confidence with which it predicts links — a very low proportion of the predicted links turn out to be false positives – which can also be seen as a conservative approach.

In these plots, the recall begins low, but rises rapidly with the increase in the number of target spikes available. In the case of low network synchrony (Fig. 1a), we observe that, by 3000 target spikes, the recall has risen to nearly one. As the precision is also nearly one at this number of target spikes, the networks are being inferred nearly perfectly. Taken in conjunction with the apparent trends towards converging on perfect inference as the number of

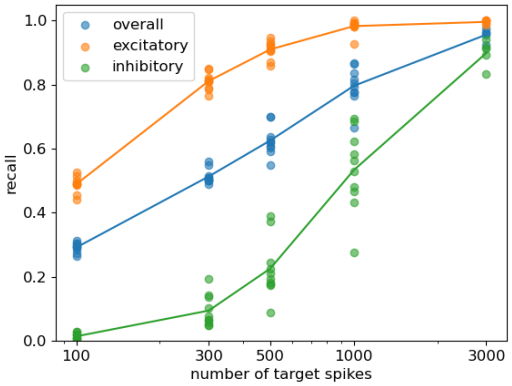
5



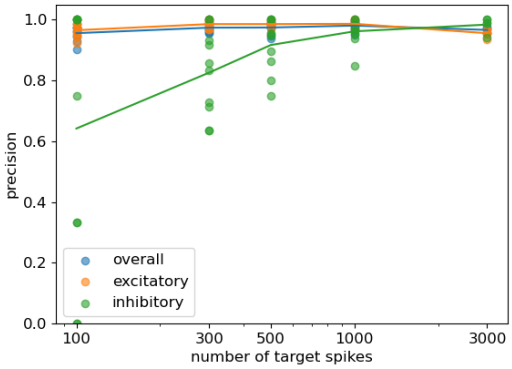
(a) Lower synchrony network, recall



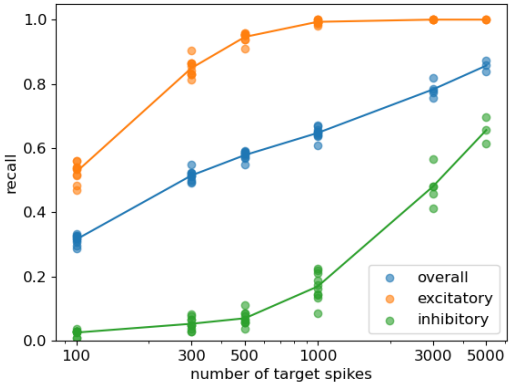
(b) Lower synchrony network, precision



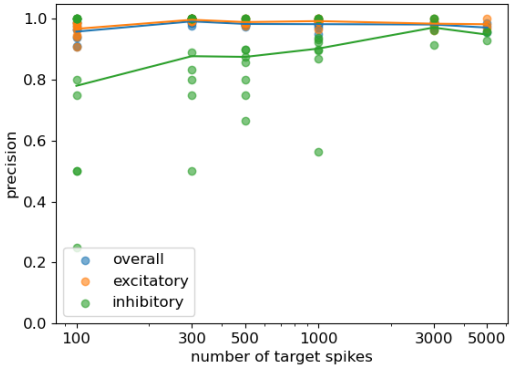
(c) Medium synchrony network, recall



(d) Medium synchrony network, precision



(e) High synchrony network, recall



(f) High synchrony network, precision

FIG. 1: Plots showing the resulting precision and recall from running the network-inference scheme on networks of LIF neurons composed of 30 excitatory neurons and 20 inhibitory neurons. The ratio of the inhibitory to excitatory connection strength was varied in order to change the level of synchrony in the network. Plots are shown for three different synchrony levels. Each plot contains points for each experiment the precision and recall for the inhibitory and excitatory sources separately as well as for their overall weighted average. The lines pass through the means of these points.

spikes increases for the other regimes, this suggests evidence for validation that the approach provides a consistent inference of the underlying network under the aforementioned idealised conditions.

By comparing Figs. 1a, 1c and 1e, we can observe that the achieved overall recall drops as the level of synchrony is increased. This is entirely driven by a drop in the recall on inhibitory connections. In fact, we observe a small increase in the recall on excitatory connections. This drop in recall is due to the increased complexity in the nature of the statistical relationship between the activity of a given target and an inhibitory source in the case of high synchrony. When the populations are highly synchronous, all of the cells spike close together, so the firing of an inhibitory source can become positively correlated with the firing of its target, when considering the purely pairwise relationship between the source and target. However, when conditioned on the target's excitatory sources (which becomes possible with more target spikes observed), for any given firing pattern of the excitatory sources, the firing of the inhibitory source is associated with a decrease in the probability of the firing of the target for that given pattern, allowing the inhibitory source to be identified. Crucially the precision remains high despite the spurious pairwise correlations that appear in the highly synchronous regime, due to the conditioning in the multivariate approach removing such redundant sources being included in the inference. Fig. 3f shows an ROC curve for the use of a purely pairwise approach on this same high-synchrony example for 1000 target spikes, which will be discussed in Sec. II C. We see that it cannot achieve a high true positive rate without a substantial increase in the false positive rate (and thus a decrease in the precision). This highlights the necessity of using a full multivariate approach when dealing with highly-synchronous neural populations.

In order to compare the performance of the proposed scheme with an existing network inference approach, we ran the recently-proposed CoNNECT [40] algorithm on the spike times from the simulations. This approach makes use of pre-trained convolutional neural networks and has been demonstrated to be competitive when compared with other existing network inference algorithms for spike trains. In order to perform this inference, we made use of the associated web-app provided by the authors [47]. The resulting precision and recall plots are shown in Fig. 8 of Appendix A. We consistently see that, for any given combination of number of target spikes and dynamical regime, the proposed approach is able to achieve both higher precision and higher recall. In particular, the precision of the results of the CoNNECT inference is particularly low, being largely in the region 0.1-0.2.

We also compared the performance of the proposed approach with a Generalised Linear Model (GLM), which is a popular approach for modelling spiking neural data [39, 48, 49], including for inferring connectivity [50, 51]. We closely followed previous work [50] which demonstrated the use of these models for connectivity inference, with a few minor differences, as specified in Sec. IV G. The resulting precision and recall plots are shown in Fig. 9 of Appendix A. The GLM approach exhibits markedly lower precision than the proposed TE-based approach, except for precision on inhibitory sources for very low numbers of available spikes in the target spike trains, where our more conservative approach is inferring very few links. Moreover, in the case of high synchrony, the precision of the proposed approach is far superior, whereas the very low precision of the GLM approach indicates that it is inferring almost half of all possible connections, rendering those inferred network models far less useful. For low numbers of target spikes, the GLM approach is able to achieve better recall. However, this is always at a cost of significantly higher precision, and any advantage in recall for the GLM approach disappears as the number of target spikes is increased above 1000 (except in the high-synchrony case). Interestingly, the better recall that our approach shows for excitatory versus inhibitory sources is reversed in the GLM approach.

B. Inference at varying levels of stimulus regularity

In Sec. II A neurons were always provided with independent Poisson stimulus. However, we can vary the properties of this stimulus in order to mimic different plausible inference scenarios. For instance, simulations constructed with a fully regular stimulus to the neurons provide us with an example of dynamics with no hidden sources of variability. That is, it is possible to perfectly predict the dynamics of the neuron based on its past and the past of those neurons connecting to it. Moreover, the system is completely deterministic. As we move towards the semi-regular and fully random Poisson stimuli, we are modelling increasing amounts of hidden activity or noise within the system, which prevents perfect predictability of the future state of the units. Full determinism is known to cause problems in the inference of networks from time series [42]. Conversely, very large amounts of noise can make it challenging to detect a comparatively weak relationship between two nodes. As such, it is important that both potentially problematic ends of this spectrum are tested.

We constructed model networks as in Sec. II A, however, this was done at a single level of ratio of inhibitory to excitatory connection strength. The same ratio ($g = 3$) used to produce the low synchrony runs was used. Instead, we varied the nature of the stimulus provided to each neuron, providing them with a regular, semi-regular or Poisson (fully random) stimulus. The regular stimulus was composed of spike times placed at a fixed interval. There was slight variation in this interval between neurons, in order to prevent the network settling into a simple, fixed, pattern. The semi-regular stimulus was similar to the regular stimulus, but with the addition of a small amount of Gaussian noise. Note that a few other minor simulation parameters had to be changed from the simulations used in Sec. II A in order to ensure numerical stability. See Sec. IV E for a full specification of these differences.

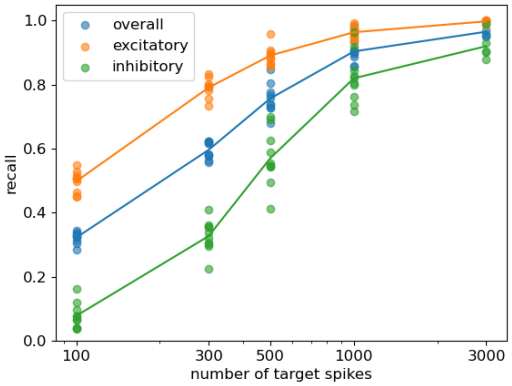
Fig. 2 shows plots of precision and recall, for different numbers of observed spikes in the target, for these different levels of stimulus regularity. We observe that the recall increases slightly when moving from the regular to the semi-regular stimulus, but then drops as we move to the Poisson stimulus. By contrast, the precision exhibited a slight increase with increasing irregularity. This is likely because the increasing irregularity reduces the correlations between true and false sources, thereby decreasing the likelihood of false positives. In all three cases, good performance is achieved at 3000 target spikes, with overall precision and recall both being around 0.9.

These results demonstrate that the proposed combination of estimator and network inference scheme is capable of successful inference at various levels of determinism or unobserved noise.

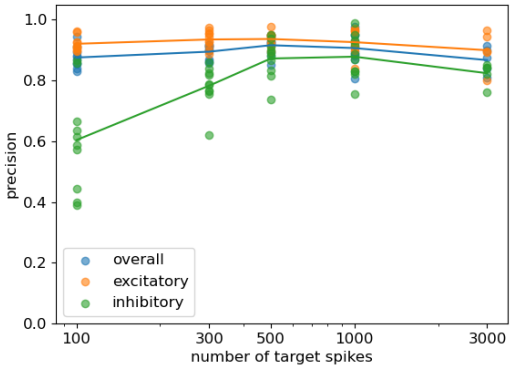
C. Comparing the greedy algorithm to pairwise inference

Recent work [52] has highlighted the improvements that can be gained when performing network inference using TE in its full multivariate sense, via the greedy algorithm, as opposed to the simpler pairwise approach. The pairwise approach operates by only checking for a statistically significant non-zero TE value between each source-target pair, without taking into account the other processes in the system. It is generally found that the full multivariate approach tends to exhibit much higher precision, as, among other reasons, it is able to distinguish true sources (which provide information about the target even when conditioned on all other system components) from spurious sources, which are merely correlated with the true sources but provide no additional information about the target when conditioning on these true sources.

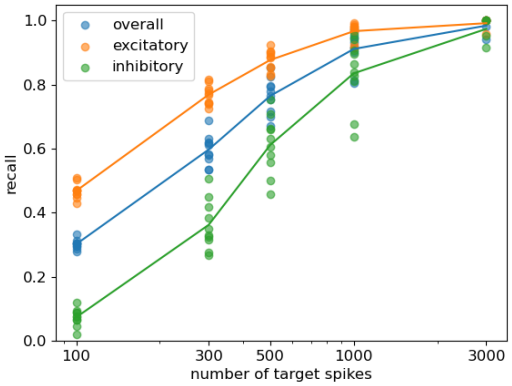
8



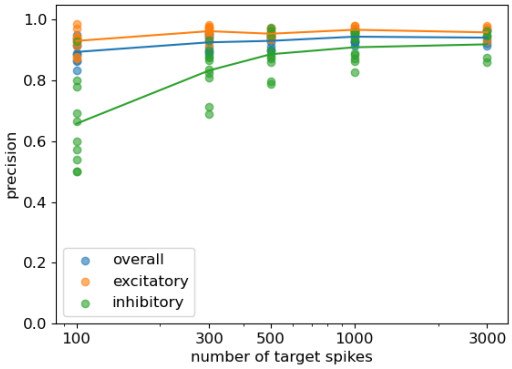
(a) Regular stimulus, recall



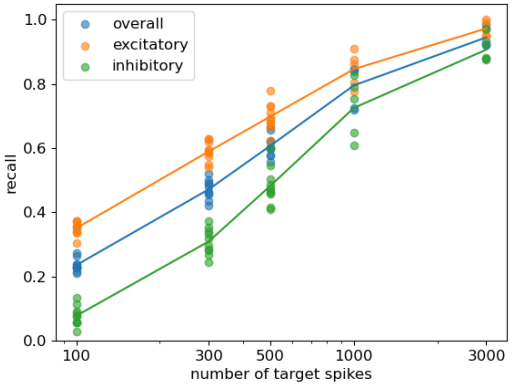
(b) Regular stimulus, precision



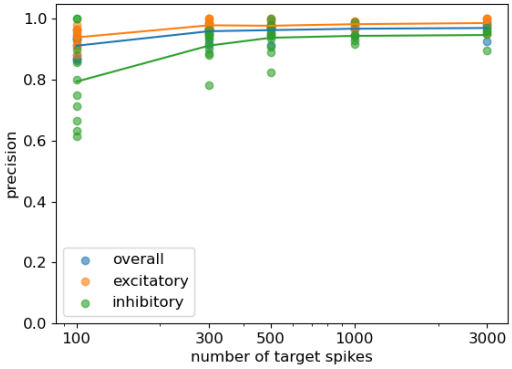
(c) Semi-regular stimulus, recall



(d) Semi-regular stimulus, precision



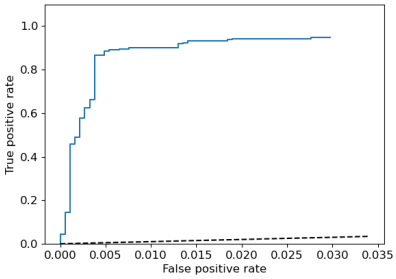
(e) Poisson stimulus, recall



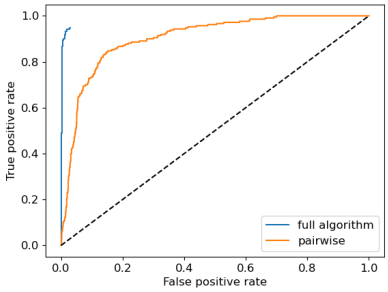
(f) Poisson stimulus, precision

FIG. 2: Plots showing the resulting precision and recall from running the network-inference scheme on networks of LIF neurons composed of 30 excitatory neurons and 20 inhibitory neurons. The regularity of the stimulus provided to each neuron was varied.

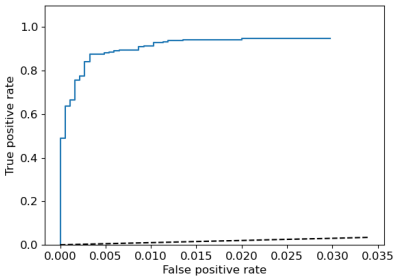
9



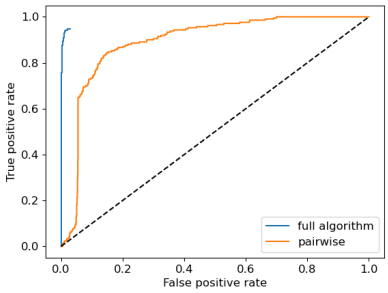
(a) Lower synchrony, full algorithm only



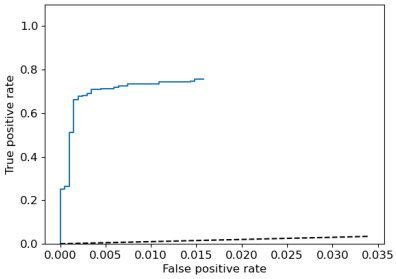
(b) Lower synchrony, pairwise inference and full algorithm



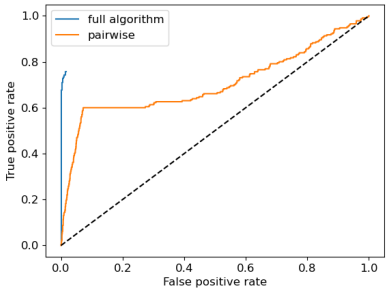
(c) Lower synchrony, full algorithm only, normal approximation



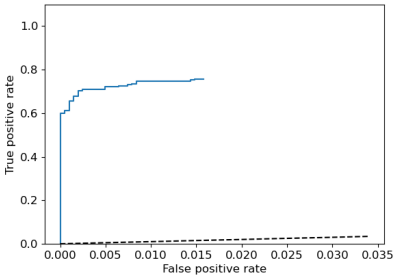
(d) Lower synchrony, pairwise inference and full algorithm, normal approximation



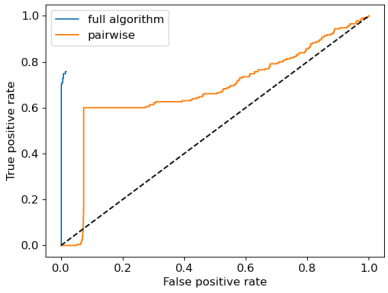
(e) High synchrony, full algorithm only



(f) High synchrony, pairwise inference and full algorithm



(g) High synchrony, full algorithm only, normal approximation



(h) High synchrony, pairwise inference and full algorithm, normal approximation

FIG. 3: ROC curves for both the presented network inference technique, as well as performing the inference using the continuous-time estimator in a simple pairwise fashion. Plots are shown for the higher synchrony and lower synchrony examples from Fig. 1, with 500 target spikes available to the algorithms. We also show plots for both the presented surrogate testing method as well as when using a normal approximation fitted to the surrogate population in order to estimate the p value.

The previous work [52] which compared multivariate TE and the greedy algorithm to the pairwise approach did so for standard time series of continuously varying signals sampled at a fixed interval. In this section, we verify that similar results hold when analysing event-based data using the continuous-time estimator of TE. Moreover, the analysis in this section will confirm the benefit of using the multivariate greedy approach, over the pairwise approach, when inferring networks from spike trains using TE.

We make use of the higher and lower synchrony simulations presented in Sec. II A, with 500 target spikes available. We applied a simple pairwise network-inference scheme to the resulting spike times from these simulations, which simply tested for statistically significant non-zero TE between each source-target pair. The resulting ROC curves are shown in Fig. 3. These ROC curves are created by sweeping through the α cutoff values (the threshold below which the p value must be for a link to be inferred) between 0 and 1 and recording the false-positive and true-positive rate observed at each α value. Calculating the p values for the statistical significance tests of non-zero TE was done by both counting the proportion of empirical surrogates (see Sec. IV B for a discussion of how these empirical surrogates are created) larger than the measured TE (Fig. 3f and Fig. 3b), as well as via a fitting a normal distribution to the surrogate values (Fig. 3h and Fig. 3d).

The purpose of also performing this normal approximation is that it allowed us to use much lower p value thresholds for a given number of surrogate calculations than is possible when evaluating p values by counting proportions of empirical surrogates, making it possible to efficiently gain more resolution on the far left of the ROC curves.

We also ran the full greedy algorithm with different α cutoff values between 0 and 0.75, providing it with 1000 target spikes (the pairwise approach made use of the same number of target spikes). The final pruning step (see Sec. IV A) was, however, excluded, as this allowed for for greater computational efficiency in a single bulk run. The resulting ROC curves are also plotted in Fig. 3. Note that these ROC curves will not reach the point where the true positive rate and false positive rate both equal one, as we are unable to inspect p values larger than 0.75.

By inspecting the ROC curves in Fig. 3a through Fig. 3d we can compare the performance of the two approaches on the networks with lower levels of correlation. The full multivariate approach is seen to very quickly arrive at a true positive rate of above 0.9, for very few false positives, which again underlines the effectiveness of this approach. In contrast, the true positive rate of the pairwise approach rises much more slowly; in other words it costs a substantially larger number of false positives to achieve the same true positive rate. Visually we see this in that the ROC curve of the multivariate approach is markedly above that of the pairwise apprach.

We see an even starker difference in the performance of these two approaches when we look at the results of inference run on the networks exhibiting higher synchrony (Fig. 3e through Fig. 3h). Here, we see that the true positive rate of the pairwise approach rises much more slowly than the multivariate approach as before, but its performance also saturates around a true positive rate of around 0.6 before false positives begin to strongly dominate further inference. In this regime the entire population has a tendency to be active together and also remain quiescent together. As the activity of all neurons are therefore correlated, the pairwise approach is unable to delineate which particular neurons are driving the activity of others. The multivariate approach is, by contrast, more robust to the higher synchrony and still able to achieve very high true positive rates at incredibly low false positive rates.

These results demonstrate the substantial advantages in using multivariate TE estimation in conjunction with the greedy algorithm as opposed to a pairwise (functional network) approach.

11

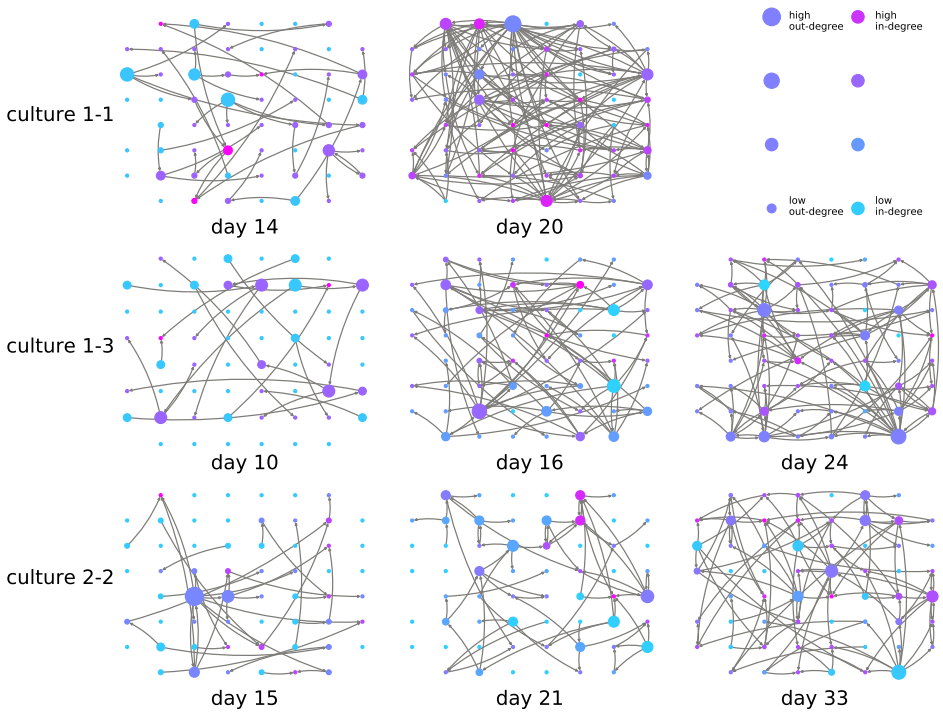
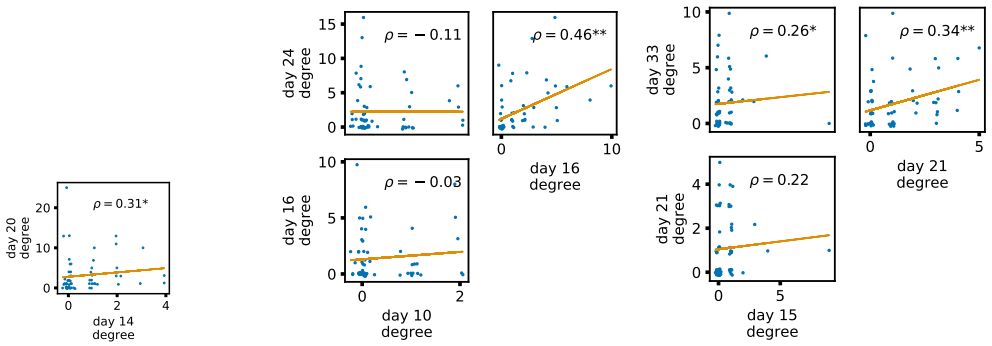


FIG. 4: Effective networks inferred using the presented approach between electrodes in developing cultures of dissociated cortical rat neurons. Each node in the network visualisations is placed in the same relative spatial location that the corresponding electrode occupied in the recording apparatus. Networks were inferred at different stages of development (days *in vitro*). The recordings used are part of an openly-available public dataset [41, 53]. Node colour and size is proportional to the in and out degrees (see the legend in the top right). The spacing between the electrodes is 200 µm centre to centre [41, 53].

D. Inference of the effective networks of developing cell cultures

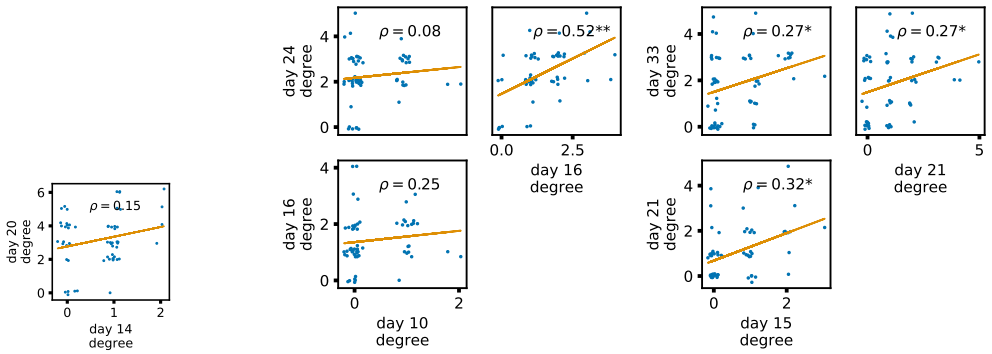
In order to demonstrate the utility of the application of this network inference scheme to biological data, we inferred the effective networks at various stages of development of cultures of dissociated cortical rat neurons. These recordings are part of a freely-available public dataset [41, 53]. See Sec. IV F for a summary of the nature of this dataset as well as details on how the network inference scheme was applied to it. In brief, cultures were allowed to develop over periods of around 30 days. On certain days, overnight recordings were performed. As these long overnight recordings contain sufficient numbers of spikes for effective application of information-theoretic estimators, they are eminently suitable for the application of our network inference approach. No spike sorting was performed, and so the networks are being inferred between the time series of the recording electrodes. This allows the nodes in the network to remain identifiable across different stages in development.

12



(a) Culture 1-1 (b) Culture 1-3 (c) Culture 2-2

FIG. 5: Plots showing the relationship between the out-degree of a given node over different days of development. Each group of plots shows scatter plots between all pairs of days for each culture analysed. Specifically, in each scatter plot, the x value of a given point is the out-degree of the associated node on an earlier day and the y value of that same point is the out-degree of the same node but on a later day. The days in question are shown on the bottom and sides of the grids of scatter plots. A small amount of Gaussian jitter ($\sigma = 0.1$) is added to the points to aid the visualisation of repeated values. The orange line shows the ordinary least squares regression. The Spearman correlation (ρ) between the out-degrees on the two days is displayed in each plot. Values of ρ significant at the 0.05 level are designated with an asterisk and those significant at the 0.01 level are designated with a double asterisk.

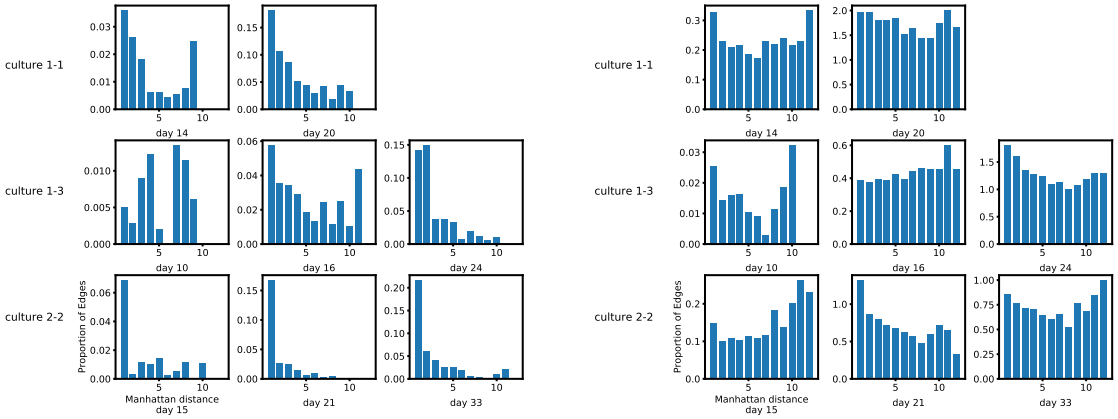


(a) Culture 1-1 (b) Culture 1-3 (c) Culture 2-2

FIG. 6: Plots showing the relationship between the in-degree of a given node over different days of development. Each group of plots shows scatter plots between all pairs of days for each culture analysed. Specifically, in each scatter plot, the x value of a given point is the in-degree of the associated node on an earlier day and the y value of that same point is the in-degree of the same node but on a later day. The days in question are shown on the bottom and sides of the grids of scatter plots. A small amount of Gaussian jitter ($\sigma = 0.1$) is added to the points to aid the visualisation of repeated values. The orange line shows the ordinary least squares regression. The Spearman correlation (ρ) between the in-degrees on the two days is displayed in each plot. Values of ρ significant at the 0.05 level are designated with an asterisk and those significant at the 0.01 level are designated with a double asterisk.

The results of applying the greedy algorithm along with the continuous-time estimator are displayed in Fig. 4. Note that the first recording days of each culture are not included in the figure as hardly any links (less than 10) were inferred in any of these recordings. We observe that effective networks with a rich and complex structure emerge, beginning to appear around the tenth day in vitro or so and quickly becoming more dense. This path of development correlates with the authors' previous investigation [38] of these recordings, with simpler directed functional networks

13



(a) Effective networks.

(b) Functional networks.

FIG. 7: Bar plots showing the proportion of possible edges that were inferred at different inter-node distances. (a) shows the proportions for the networks inferred using the presented greedy algorithm and, whose diagrams are displayed in Fig. 4. (b) shows the same proportions for the functional networks inferred on the same data in the authors' recent work [38]. The inference of such functional networks only considers pairwise relationships. The distances on the x axis are the Manhattan (cityblock) distances between electrodes. It is clear from the plots that, on this dataset, the effective network inference algorithm has a greater propensity to infer short distance links.

inferred using pairwise transfer entropy (via the same underlying estimator). The density of the networks inferred by the multivariate algorithm are lower than for the directed functional networks in [38], with the total number of edges in the last recording days declining from around 1000 to 2000 to around 100 to 200. This is common because the strongest action of the multivariate algorithm is to remove redundant sources [52].

Despite the difference in density, the effective network structures retain some of the interesting features observed for the directed functional networks in [38], such as containing pronounced inward and outward hubs (that is, nodes with particularly high in-degree or out-degree), and various features of the networks being locked in early in development. Specifically, in the directed functional networks in [38] characteristics such as the total inward or outward information flow for a given node exhibited high correlation between early and late days of development. Here, Fig. 5 shows scatter plots of the out-degrees of the inferred effective networks on earlier and later days of development. We see in these plots that, as with the functional networks, in all cases, there is a positive correlation between the out-degree on the earlier and later days of development. Moreover, there are no statistically significant negative correlations. The positive correlation for the out-degree across the last two recording days is statistically signfiicant for each culture, with some of these relationships, such as for culture 1-3, being particularly strong. Fig. 6 shows similar plots, but for the in-degrees of the nodes. Again, we see a positive correlation between the in-degree on earlier and later days in every case. These results indicate that features of the effective networks, representing the multivariate information flows here, are being locked in early in development. This is particularly interesting since these are more sparse network models than the directed functional networks in [38], suggesting that the lock-in effect is deeply ingrained in the system.

Fig. 7 displays the proportion of possible links that are inferred in the networks at various physical distances between

the nodes. Fig. 7a does so for the effective networks inferred in this work and Fig. 7b does so for the functional networks studied in the authors' previous work. These plots show that the effective networks inferred on this dataset exhibit a clear preference towards links between nodes that are physically close together. This preference appears to become stronger with developmental time. By contrast, the functional networks do not exhibit this preference.

III. DISCUSSION

In this work, we have validated the efficacy of the combination of an existing greedy multivariate TE-based network inference algorithm with a recently-introduced continuous-time estimator for TE on event-based data.

As the inference of networks from the spike times of neurons is a common goal within neuroscience, we expect this particular task to be a core application of the presented approach. Indeed, there is a significant body of existing work which validated [19–22] and applied [23–30, 54] TE to the task of inferring networks from the spike times of neurons. However, apart from a single very recent study [54], this previous work has always considered only the pairwise relationships between the activity on each node. As was demonstrated in Fig. 3, even when using the highly-effective continuous-time TE estimator, this approach suffers substantial drawbacks. Perhaps most notably, when the entire population is highly correlated, the pairwise approach is unable to distinguish a smaller subset of sources which can provide all the information about the target contained in the entire population, instead inferring large numbers of sources due to the redundant information they hold with the true sources.

Here, by contrast, we have presented a multivariate approach to inferring *effective* networks from neural spike trains using TE. Unlike the pairwise approach, this strategy infers a set of parents for a target collectively rather than individually for each source. In doing so, the multivariate strategy considers the activity of other nodes within the network when determining the directed relationship between any two nodes. Specifically, as the multivariate strategy iteratively or greedily adds new candidate sources to the parent set for a target, it requires that each source provide statistically significant non-zero TE, when conditioning on all other current parents for the target. This is in contrast to the pairwise approach, which only requires a non-zero TE value between the source and target, without taking other processes into account. That iterative conditioning, along with final associated pruning step, supports much more accurate inference because it eliminates redundant information from being spuriously attributed to other sources, and captures synergistic or collective interactions between multiple sources which jointly impact the target.

The term "accurate" here has specific meaning in the context in which we have evaluated the performance of the multivariate approach. Although we cannot and do not always expect the inferred effective networks to align with the underlying causal or structural network of the system being examined, under certain highly-specific idealised conditions (full observability etc., see Sec. I) we do indeed expect a minimal model explaining the dynamics of the variables (the effective network) to align with the causal structure in this way. As such, confirming such alignment – even though in highly-specific conditions – is an important validation of the performance of such an approach. Indeed, this validation is clear from our results, including in Fig. 1 and Fig. 2, and by the improved performance of the multivariate approach compared to pairwise, using the same underlying TE estimator, in Fig. 3. There, it was found that the multivariate approach was able to achieve a comparable true positive rate as the pairwise approach with a much lower corresponding false positive rate.

Maintaining a low false positive rate (or, equivalently, a high precision) is of utmost importance for network inference

in a neuroscientific context. Zalesky states that "False positives are at least twice as detrimental as false negatives" [55]. As demonstrated in all of our results, the presented approach errs on the conservative side and consistently maintains high precision (a low false-positive rate) even in the challenging cases where the activity on the nodes is highly correlated (Fig. 1f). This high precision is a result not only of the multivariate strategy, but also the local permutation surrogate generation method used for the significance testing of the individual TE estimates being different from zero. This method was developed in tandem with the recently-developed continuous-time TE estimator [31] that is used here and was demonstrated to have substantially lower false positive rates when compared with using traditionally-used approaches for surrogate generation in conjunction with this estimator. Our results also compared the proposed approach with two existing approaches for the inference of connectivity from spiking data (CoNNECT in Fig. 8 and GLM in Fig. 9), finding that it exhibited far superior precision in most cases. This is particularly important when we reflect on the goal of effective network inference being to provide a "minimal model" that can explain the dynamics.

Not only does this work present the first validation of TE for multivariate effective network inference on spike-train data, but it presents the first validation study of the using of the recently-developed continuous-time estimator for TE on event-based data such as spike trains in the context of network inference [31]. This estimator has been demonstrated to have many substantial advantages over the traditional discrete-time approach. These include consistency, lower bias and faster convergence. Of particular relevance to network inference, by representing the history embeddings of processes using inter-event intervals, it is able to represent histories of substantial length using few dimensions and without any loss of time precision. This efficient use of dimensions facilitates building conditioning sets of significant size.

These various benefits culminate in a technique that is highly effective in the inference of networks from event times. We have demonstrated its strong performance, with high precision, in dynamical regimes ranging from low to high network synchrony (Fig. 1) as well as with varying levels of unobserved noise sources in the system (Fig. 2). Furthermore, high quality inferences were made with relatively low numbers of target spikes. In some instances (eg: Fig. 1a and Fig. 1b), near perfect reconstruction was achieved with only 3000 events per target.

These results all point to the strong potential for deploying this methodology in the inference of networks from recordings of the spike times of biological neurons. Tantalising hints of the results that might be expected were provided in Sec. II D. Further such applications remain a focus of future work. It is of particular interest to note that, in these effective networks, there was a strong preference towards inferring edges between nodes that are spatially close together, especially when compared with the functional network approach. This is likely due to the fact that effective networks are known to conform closer to the underlying structural networks than those inferred using pairwise, functional, methods [52]. However, despite this change in the topology of the networks, it is worth noting that the lock-in of information flows early in development, which was previously observed in functional networks [38], remained in these effective networks.

IV. METHODS

A. Greedy Algorithm

The greedy network inference algorithm used here was proposed in a range of papers [10, 16–18], as summarised and studied in depth by Novelli et. al. [7], for traditional time series (a continuous-valued signal sampled at regular time intervals). We describe it here for completeness, and also to highlight some small changes that we made to adapt to the context of event-based data. The most significant change made is that only one inter-spike interval per source is considered as a candidate, and sequentially, from the most recent. This is as opposed to the original algorithm, where several lags from each source could be considered in the same selection round, and no ordering was imposed on the addition of these lagged samples.

The greedy algorithm is specified in Algorithm 1. We walk through its operation here, with reference to the line numbers in Algorithm 1.

We iterate over each process R_i in the set of processes $\mathcal{R}$ (line 1). These processes are the raw timestamps of events (spikes). Each process is being considered as a target, for which the sources need to be inferred. It is worth noting that the computations performed for each target are considered completely independent of one another. As such, it is easy to parallelise this algorithm across the different targets. Indeed, such parallelisation was performed for the experiments presented in this paper.

We initialise a data structure to keep track of the last interval added to the conditioning set for each source and the target itself (line 2). This algorithm makes the assumption that more recent inter-event intervals from a given source (or the target itself) always have more influence over the target than inter-event intervals further in the past. Based on this assumption, inter-event intervals for a given source are only considered as candidates once more recent intervals for that source have been added to the conditioning set. As above, this is distinct from the operation of the algorithm for traditional time-series.

Before considering candidate sources, we determine the number of target history intervals to condition on. We always include at least one (the most recent) such interval. We continue incrementing the total number of intervals that we are conditioning on until the next interval does not provide a statistically significant reduction in the uncertainty of the target (line 4). This reduction in uncertainty is measured by the conditional Active Information Storage (AIS) [11], which is the mutual information between the last target history interval being considered and the current state of the target, conditioned on the more recent target history intervals. Note that this quantity can be easily estimated using the continuous-time TE estimator by simply considering this last target interval as a source interval. The active information storage is estimated on the original spike train and on $N_{\text{surrogates}}$ surrogate processes (lines 5 and 6), constructed using the local permutation method described in Sec. IV D. The p value for the significance test is then the proportion of AIS estimates on the surrogates which are larger than the AIS estimate on the original process. If $p < p_{\text{cutoff}}$, then the null hypothesis of zero AIS is rejected and the number of target intervals being added is incremented.

Returning to the canidate sources then, we continuously iterate parent selection for the target until the candidate source interval with the highest TE is no longer statistically significant (line 13). For each of these iterations, we iterate over every process other than the target under consideration (line 16). For each such candidate source, we estimate

the TE between the most recent inter-event interval of that source that has not yet been added to the conditioning set and the target, conditioned on all intervals of all sources already added to the conditioning set (line 17). We then estimate the TE for $N_{\text{surrogates}}$ surrogate processes between the same source and target and conditioned on the same conditioning set (line 18). We also bias-correct the original and surrogate TE estimates by subtracting the mean value of the surrogates from each estimate (lines 19 and 20).

Once this has been performed for every candidate source interval, we select the interval which had the highest bias-corrected conditional TE (line 22). We then estimate the p value associated with the null hypothesis of the conditional TE from this source interval being zero (line 23). This is done using the maximum statistic test (see Sec. IV B).

The above process continues until the selected candidate source interval (with maximum bias-corrected conditional TE) fails the significance test (line 13). The algorithm then moves onto the final pruning step, where it is checked that every source retains a statistically significant conditional TE, once conditioning on every other process added to the conditioning set. This step is necessary as a source added early in the process might be providing information about the target that is fully redundant with that held by sources added later in the greedy building of the conditioning set. Such redundant sources need to be removed.

To perform the pruning, we continually try removing source intervals from the conditioning set one-by-one, until every final source interval in the set is found to be statistically significant. In a mirror image to how the candidate intervals, for a given source, are added iteratively from the most recent and then further back in time, they are removed in order from the last interval to the most recent. In each round of pruning, we iterate over all sources which had an interval added to the conditioning set (line 31). We then estimate the conditional TE and associated surrogates for the last remaining added interval for that source (lines 32 and 33). We then calculate the p value corresponding to the null hypothesis (line 34) of zero TE in the normal manner (that is, not using the maximum statistic test). After iterating over all sources in the conditioning set, we then find the source index with the maximum p value (line 36). If this p value is greater than the specified α cutoff, then we remove the last added interval for that source from the conditioning set (line 39).

B. Maximum Statistic Test

When considering adding sources to the conditioning set, we test the candidate source with the highest TE using the maximum statistic test (line 23 of Algorithm 1).

It is worth briefly describing the usual method for testing for non-zero TE using surrogates. We generate $N_{\text{surrogates}}$ surrogates, which conform to the null hypothesis of no temporal relationship (zero TE), using a given surrogate generation algorithm (see Sec. IV D for a description of the surrogate generation method used here). We then estimate the TE on each of these generated surrogate series. The proportion of these estimates which are greater than or equal to the estimate on the original data is then an estimate of the probability that we would observe a value greater than or equal to what we estimated on the original data, under the null hypothesis of zero TE (and therefore it is our p value).

Novelli et. al. [7] highlighted the fact that using this test as is, when adding sources to the conditioning set, would lead to high false-positive rates. This is effectively a multiple comparisons problem, in that the test is being performed

18

Algorithm 1: Greedy TE algorithm for network inference from event-based data.

input : /* Set of the event times of each process */
$\mathcal{R} = \{R_i\}_{i=1}^{N_{\text{proc}}}$
/* Cutoff significance value for adding sources */
p_{cutoff}
/* Number of surrogate estimates to perform per TE estimation */
$N_{\text{surrogates}}$

output: /* Set of the indices of the sources of each process */
$\mathcal{S} = \{S_i\}_{i=1}^{N_{\text{proc}}}$

/* Iterate over all target processes. */
for $i \leftarrow 1$ **to** N_{proc} **do**
  /* Variable to keep track of added sources, as well as the added target inter-event intervals. If c_k is zero, this implies the source is not added. Otherwise, the value of c_k indicates the number of added inter-event intervals. */
  $C \leftarrow \{0\}_{k=1}^{N_{\text{proc}}}$
  $p \leftarrow 0$
  /* Keep trying the next target interval, until the added information is no longer significant. */
  while $p < p_{\text{cutoff}}$ **do**
    /* Estimate the TE from the next interval of the target under consideration, conditional on all target intervals already added. */
    $a \leftarrow \texttt{estimateConditionalAIS}\,(i, c_i + 1, R,)$
    /* Estimate the associated surrogate values. */
    $\{a_{\text{surrogate},n}\}_{n=1}^{N_{\text{surrogates}}} \leftarrow \texttt{estimateConditionalAISOnSurrogates}\,(i, c_i + 1, R, N_{\text{surrogates}})$
    $p \leftarrow \texttt{calculatePVal}\left(t, \{\{t_{\text{surrogate},n}\}_{n=1}^{N_{\text{surrogates}}}\}\right)$
    /* If the null hypothesis is rejected, we increment the number of target intervals added */
    if $p < p_{\text{cutoff}}$ **then**
      $c_i \leftarrow c_i + 1$
    end
  end
  $p_{\max} \leftarrow 0$
  /* Keep looking for sources until the candidate with maximum TE is not significant. */
  while $p_{\max} < p_{\text{cutoff}}$ **do**
    /* Initialise TE and surrogate TE values to 0. */
    $\{t_j\}_{j=1}^{N_{\text{proc}}} \leftarrow \{0\}_{j=1}^{N_{\text{proc}}}$
    $\{\{t_{\text{surrogate},j,n}\}_{n=1}^{N_{\text{surrogates}}}\}_{j=1}^{N_{\text{proc}}} \leftarrow \{\{0\}_{n=1}^{N_{\text{surrogates}}}\}_{j=1}^{N_{\text{proc}}}$
    /* Iterate over all processes except the target under consideration. */
    for $j = 1 \ldots N_{\text{proc}}$ where $j \neq i$ **do**
      /* Estimate the TE from the next interval of the source under consideration, conditional on all source intervals already added. */
      $t_j \leftarrow \texttt{estimateConditionalTE}\,(i, j, c_j + 1, R, C)$
      /* Estimate the associated surrogate values. */
      $\{t_{\text{surrogate},j,n}\}_{n=1}^{N_{\text{surrogates}}} \leftarrow \texttt{estimateConditionalTEOnSurrogates}\,(i, j, c_j + 1, R, C, N_{\text{surrogates}})$
      /* Bias-correct both the original and surrogate TE estimates by subtracting the mean of the surrogates. */
      $t_j \leftarrow t_j - \texttt{mean}\left(\{t_{\text{surrogate},j,n}\}_{n=1}^{N_{\text{surrogates}}}\right)$
      $\{t_{\text{surrogate},j,n}\}_{n=1}^{N_{\text{surrogates}}} \leftarrow \{t_{\text{surrogate},j,n} - \texttt{mean}\left(\{t_{\text{surrogate},j,n}\}_{n=1}^{N_{\text{surrogates}}}\right)\}_{n=1}^{N_{\text{surrogates}}}$
    end
    /* Find which candidate source interval had the highest bias-corrected TE estimate. */
    $j_{\max} \leftarrow \texttt{findIndexOfMax}\left(\{t_j\}_{j=1}^{N_{\text{proc}}}\right)$
    /* Estimate the p value corresponding to the null hypothesis that this source had zero TE. */
    $p_{\max} \leftarrow \texttt{calculateMaxStatPVal}\left(t_{j_{\max}}, \{\{t_{\text{surrogate},j,n}\}_{n=1}^{N_{\text{surrogates}}}\}_{j=1}^{N_{\text{proc}}}\right)$
    /* If the null hypothesis is rejected, add the candidate source interval to the set of sources. */
    if $p_{\max} < p_{\text{cutoff}}$ **then**
      $c_{j_{\max}} \leftarrow c_{j_{\max}} + 1$
    end
  end
  /* We now check that all added source intervals are still significant, when conditioning on all other added sources. */
  $p_{\max} \leftarrow 0$
  /* Iterate until no checked source interval is not significant. */
  while $p_{\max} > p_{\text{cutoff}}$ **do**
    $\{p_j\}_{j=1}^{N_{\text{proc}}} \leftarrow \{1\}_{j=1}^{N_{\text{proc}}}$
    /* Iterate over all processes that had any intervals added. */
    for $j = 1$ **to** N_{proc} where $j \neq i$ and $c_j > 0$ **do**
      /* Estimate the TE from the last interval of the source under consideration, conditional on all other added source intervals. */
      $t \leftarrow \texttt{estimateConditionalTE}\,(i, j, c_j, R, C)$
      /* Estimate the associated surrogate values. */
      $\{t_{\text{surrogate},n}\}_{n=1}^{N_{\text{surrogates}}} \leftarrow \texttt{estimateConditionalTEOnSurrogates}\,(i, j, c_j, R, C, N_{\text{surrogates}})$
      /* Estimate the p value corresponding to the null hypothesis that this source had zero TE. */
      $p_j \leftarrow \texttt{calculatePVal}\left(t, \{t_{\text{surrogate},n}\}_{n=1}^{N_{\text{surrogates}}}\right)$
    end
    /* Find the index of the source whose final interval had the highest p value. */
    $j_{\max} \leftarrow \texttt{findIndexOfMax}\left(\{p_j\}_{j=1}^{N_{\text{proc}}}\right)$
    $p_{\max} \leftarrow p_{j_{\max}}$
    /* If the null hypothesis is not rejected for this source interval, remove the candidate source interval from the set of sources. */
    if $p_{\max} > p_{\text{cutoff}}$ **then**
      $c_{j_{\max}} \leftarrow c_{j_{\max}} - 1$
    end
  end
  /* Add the indices of the sources which had any intervals remaining in the conditional set to the final set of inferred sources for the given target. */
  $S_i \leftarrow \{j : j \in \{1, 2, \ldots, N_{\text{proc}}\} \text{ and } c_j > 0\}$
end

on the *maximum* estimated TE value from the set of candidate sources.

In order to compensate for this, we replicate the selection of the maximum candidate source in the significance testing step. Specifically, for each $i \in \{1, 2, \ldots, N_{\text{surrogates}}\}$, we compare the TE estimates on the i^{th} surrogate for each of the candidate sources. We select the maximum such estimate for each i. The population of surrogate values for the maximum statistic test is then made up of the resulting $N_{\text{surrogates}}$ maximum values. The test then proceeds as normal.

C. Transfer Entropy Estimation

It has, relatively recently, been shown that, for event-based data such as spike-trains, in the limit of small bin size, that the expected TE rate is given by the following expression [56]:

Parameter	Description	Value
N_X	Number of spikes in the target spike train	varied (see text)
k_{global}	Number of nearest neighbours to find in the initial search	4
k_{perm}	Number of nearest neighbours to consider during surrogate generation	20
N_U	Number of random samples of histories at non-spiking points in time	$20N_X$
$N_{U,\text{surrogates}}$	Number of random samples of histories at non-spiking points in time used for surrogate generation	$20N_X$
$N_{\text{surrogates}}$	Number of surrogates to generate for each node pair	100

TABLE I: The parameter values used in the continuous-time TE estimator when used for the inference of the simulated spiking networks. A complete description of these parameters, along with analysis and discussion of their effects can be found in [31].

$$\dot{\mathbf{T}}_{Y\to X} = \lim_{\tau\to\infty} \frac{1}{\tau} \sum_{i=1}^{N_X} \ln \frac{\lambda_{x|\mathbf{x}_{<t},\mathbf{y}_{<t}}\left[\mathbf{x}_{<x_i}, \mathbf{y}_{<x_i}\right]}{\lambda_{x|\mathbf{x}_{<t}}\left[\mathbf{x}_{<x_i}\right]}. \tag{3}$$

Here, $\lambda_{x|\mathbf{x}_{<t},\mathbf{y}_{<t}}\left[\mathbf{x}_{<x_i}, \mathbf{y}_{<x_i}\right]$ is the instantaneous firing rate of the target conditioned on the histories of the target $\mathbf{x}_{<x_i}$ and source $\mathbf{y}_{<x_i}$ at the time points x_i of the spike events in the target process. $\lambda_{x|\mathbf{x}_{<t}}\left[\mathbf{x}_{<x_i}\right]$ is the instantaneous firing rate of the target conditioned on its history alone, ignoring the history of the source. It is important to note that the sum is being taken over the N_X spikes of the target during the sampling period τ: thereby evaluating log ratios of the expected spike rates of the target given source and target histories versus target histories alone, *when* the target does spike. As this expression allows us to ignore the "empty space" between events, it presented clear potential for allowing for more efficient estimation of TE on spike trains.

This potential was recently realised in a new continuous-time estimator of TE presented in [31], and all TE estimates in this paper were performed using this new estimator. In [31] it is demonstrated that this continuous-time estimator is far superior to the traditional discrete-time approach to TE estimation on spike trains. For a start, unlike the

20

Parameter	Description	Value
N_X	Number of spikes in the target spike train	varied (see text)
k_{global}	Number of nearest neighbours to find in the initial search	10
k_{perm}	Number of nearest neighbours to consider during surrogate generation	20
N_U	Number of random samples of histories at non-spiking points in time	$20N_X$
$N_{U,\text{surrogates}}$	Number of random samples of histories at non-spiking points in time used for surrogate generation	$20N_X$
$N_{\text{surrogates}}$	Number of surrogates to generate for each node pair	100

TABLE II: The parameter values used in the continuous-time TE estimator when used for network inference on the *in vitro* spike recordings. A complete description of these parameters, along with analysis and discussion of their effects can be found in [31].

discrete-time estimator, it is consistent. That is, in the limit of infinite data, it will converge to the true value of the TE. It was also shown to have much preferable bias and convergence properties. Most significantly, perhaps, this new estimator utilises the inter-spike intervals to efficiently represent the history embeddings $\mathbf{x}_{<x_i}$ and $\mathbf{y}_{<x_i}$ in estimating the relevant conditional spike rates in (3).

This is in contrast with the traditional discrete-time estimator which uses the presence or absence of spikes in an array of time bins as its history embeddings (it sometimes also uses the number of spikes occurring in a bin). In order to avoid the dimensionality of the estimation problem becoming sufficiently large so as to render estimation infeasible, only a small number of bins can be used in these embeddings. To focus in on cell-culture data, previous applications of TE to this type of data have used a variety of bin sizes: 40 µs [23], 0.3 ms [19], and 1 ms [24, 27]. Some studies chose to examine the TE values produced by multiple different bin widths, specifically: 0.6 ms and 100 ms [25], 1.6 ms and 3.5 ms [28] and 10 different widths ranging from 1 ms to 750 ms [26]. And specifically, those studies demonstrated the unfortunate high sensitivity of the discrete-time TE estimator to the bin width parameter. Moreover, all of these studies have only used a single bin in the history embeddings. In the instances where narrow (< 5 ms) bins were used, only a very narrow slice of history is being considered in the estimation of the history-conditional spike rate. This is problematic, as it is known that correlations in spike trains exist over distances of (at least) hundreds of milliseconds [32, 33]. Conversely, in the instances where broad (> 5 ms) bins were used, relationships occurring on fine time scales will be completely missed. This is significant given that it is established that correlations at the millisecond and sub-millisecond scale play a role in neural function [34–37]. In other words, previous applications of transfer entropy to electrophysiological data from cell cultures either captured some correlations occurring with fine temporal precision or they captured relationships occurring over larger intervals, but never both simultaneously. This can be contrasted with the inter-spike interval history representation used by the continuous-time estimator. To take a concrete example, in the *in vitro* data we used, for the recording on day 24 of culture 1-3, the average interspike interval was 0.71 seconds. This implies that the history embeddings used are at least on average 0.71

seconds long, being longer than this in cases where multiple intervals are being used. This is despite the fact that our history representations retain the precision of the raw data (40 µs) and the ability to measure relationships on this scale where they are relevant (via the underlying nearest-neighbour estimators). Furthermore, the innovative representation of history embeddings as an array of inter-spike intervals allows for the application of the highly effective nearest-neighbour family of information-theoretic estimators [12, 57], which bring estimation efficiency and bias correction.

The challenges of using the discrete-time estimator only become more severe when one attempts to infer networks using conditional TEs. As there are now more processes being considered by the estimator (those in the conditioning set) the dimensionality of the estimation problem increases faster as we increase the embedding length. This places further pressure on keeping the number of bins in each embedding low, thus increasing the harshness in the tradeoff between history length and temporal precision. This is the likely reason behind the fact that almost all previous studies which evaluated the use of TE for the inference of spiking networks only made use of pairwise TE estimates [19–22]. This is as opposed to the multivariate conditional TE estimation used here, which takes into account the relationship of the target to other processes when considering its relationship to the given source.

The parameters used with this estimator for the simulated data are shown in Table I and those used for *in vitro* spike recordings are shown in Table II. The chief difference in the parameter values used in these situations is that, for the *in vitro* recordings a larger value of k_{global} (the number of nearest neighbours to consider in the initial searches) was employed (10 compared to 4). This was due to the observation that the estimates on the *in vitro* recordings exhibited much higher variance than those on the simulated data. It is a known property of nearest-neighbour information theoretic estimators that considering larger numbers of neighbours reduces their variance [12].

As in the authors' previous work applying the continuous-time TE estimator to *in vitro* spike recordings [38], a small change was made to the estimation procedure described in [31]. This was made in how random sample points were placed along the process both for the estimation of the TE and the generation of surrogates. Instead of laying out the N_U and $N_{U,\text{surrogates}}$ sample points randomly uniformly, we placed each one at an existing target spike, with the addition of uniform noise on the interval $[-80\,\text{ms}, 80\,\text{ms}]$. This was due to the fact that these recordings contain incredibly dense bursts. Such a sampling strategy is required in order to adequately sample these regions of intense activity.

An implementation of the estimator contained in the Java Information Dynamics Toolkit (JIDT) [58] software package was used in this study.

D. Surrogate Generation

Surrogate processes were generated by applying an adaptation of the permutation method of Runge [59] to the spiking TE estimator, as detailed in [31]. In brief, this method permutes the history embedding vectors to destroy the relationship between the source intervals and the existence or absence of spiking in the target. However, it retains the relationship between the source history embedding intervals and the embedding intervals from the target and conditioning processes.

22

497 **E. Spiking Network Simulation**

Parameter	Description	Value
N_{exc}	Number of excitatory neurons	30
N_{inh}	Number of inhibitory neurons	20
τ	Membrane time constant	20 ms
R_m	Membrane resistance	1 Ω
V_{reset}	Membrane reset potential	0 mV
V_0	Membrane resting potential	0 mV
$V_{\text{threshold}}$	Spike threshold potential	40 mV
τ_{syn}	Synaptic time constant	20 ms
g	Ratio of inhibitory to excitatory connection strength	$\{1, 1.5, 3\}$
$\bar{\alpha}_{\text{exc}}$	Excitatory connection strength	20 mA
n_{exc}	Number of excitatory sources per neuron	3
n_{inh}	Number of inhibitory sources per neurons	2
λ_{stim}	Rate of the Poisson stimuli	200 Hz
$\bar{\alpha}_{\text{stim}}$	Stimulus connection strength	6 mA

TABLE III: The parameter values used in the LIF network simulations at various levels of synchrony, presented in Sec. II A

Parameter	Description	Value
N_{exc}	Number of excitatory neurons	30
N_{inh}	Number of inhibitory neurons	20
τ	Membrane time constant	20 ms
R_m	Membrane resistance	1 Ω
V_{reset}	Membrane reset potential	0 mV
V_0	Membrane resting potential	0 mV
$V_{\text{threshold}}$	Spike threshold potential	40 mV
τ_{syn}	Synaptic time constant	20 ms
g	Ratio of inhibitory to excitatory connection strength	3
$\bar{\alpha}_{\text{exc}}$	Excitatory connection strength	17.5 mA
n_{exc}	Number of excitatory sources per neuron	3
n_{inh}	Number of inhibitory sources per neurons	2
$\bar{\alpha}_{\text{stim}}$	Stimulus connection strength	4 mA

TABLE IV: The parameter values used in the LIF network simulations at various levels of synchrony, presented in Sec. II B

498 All network simulations were conducted using Leaky-Integrate-and-Fire (LIF) model neurons [44]. In this model,

the membrane potential of the i^{th} evolves according to:

$$\tau \frac{dV_i}{dt} = (V_0 - V_i) + R_m I_{\text{syn},i}. \tag{4}$$

When V_i crosses the threshold $V_{\text{threshold}}$, the timestamp of crossing is recorded as a spike. V_i is then set to V_{reset} and the evolution of the membrane potential is subsequently paused for the duration of the hard refractory period. $I_{\text{syn},i}$ is the synaptic input current into neuron i. Neurons were connected using alpha synapses [60]. Each synapse connecting neuron j to neuron i evolves according to:

$$I_{i,j}(t) = a_{i,j}\bar{\alpha_j} \sum_{t_s \in S_{t,j}} \frac{t - t_s}{\tau_{\text{syn}}} \exp\left(-\frac{t - t_s}{\tau_{\text{syn}}}\right). \tag{5}$$

A is the connectivity matrix, with $a_{i,j} = 1$ indicating that neuron j is a pre-synaptic input to neuron i and $a_{i,j} = 0$ indicating otherwise. $\bar{\alpha_j}$ is the connection strength of the afferent connections from neuron j. All excitatory neurons share the same afferent connection strength $\bar{\alpha}_{\text{exc}}$. Inhibitory neurons, by contrast, have connection strength $g\bar{\alpha}_{\text{exc}}$. The sum is taken over the set of spike times in neuron j occurring before time t, $S_{t,j}$. The synaptic current for neuron i is then the sum of the currents from all other neurons in the network, that is, $I_{\text{syn},i} = \sum I_{i,j}$.

Each neuron was connected to exactly n_{exc} excitatory sources, chosen randomly from the set of excitatory neurons. Similarly, each neuron was connected to n_{inh} inhibitory sources, chosen randomly from the inhibitory sources. The specific parameter values used in the experiments described in Sec. II are shown in Table III and Table IV.

Each neuron also received an independent stimulus. In the experiments presented in Sec. II A, this source was an homogeneous Poisson point process. In the experiments presented in Sec. II B, it contained varying amounts of regularity. Specifically, in Sec. II B, each neuron received a stimulus with a spike rate λ drawn from a normal distribution with mean 500 hertz and standard deviation of 25 hertz. In the fully random case, the stimulus was generated as an homogeneous Poisson point process, with rate λ. In the fully regular case, spikes were placed at a fixed interval of $1/\lambda$. In the semi-regular case, the spikes were placed with this same fixed interval, but gaussian noise with mean 0 and standard deviation 0.5 millisecond was added to each spike time. The connectivity strength between each stimulus and its target was specified by the parameter $\bar{\alpha}_{\text{stim}}$.

F. Analysis of *in vitro* Data

We made use of the same dataset as in the authors' previous study [38] and analysed it in a very similar fashion. As such, the following section closely follows the discussion of this dataset in that previous work.

The spike train recordings used in this study were collected by Wagenaar et. al. [41] and are freely available online [53]. The details of the methodology used in these recordings can be found in the original publication [41]. A short summary of their methodology follows:

Dissociated cultures of rat cortical neurons had their activity recorded. This was achieved by plating 8x8 Multi-Electrode Arrays (MEAs), operating at a sampling frequency of 25 kHz with neurons obtained from the cortices of rat embryos. The spacing between the electrodes was 200 µm center-to-center. The MEAs did not have electrodes on their corners and one electrode was used as ground, resulting in recordings from 59 electrodes. In all recordings,

electrodes with less than 100 spikes were removed from the analysis. This resulted in electrodes 37 and 43 being removed from every recording as no spikes were recorded on them. The spatial layout of the electrodes is available from the website associated with the dataset [53], allowing us to overlay the inferred networks onto this spatial layout as is done in figure Fig. 4.

Recordings were conducted on most days, starting from 3-4 Days *In Vitro* (DIV). The end point of recording varied between 25 and 39 DIV. Longer overnight recordings were also conducted on some cultures at sparser intervals. In this work we make use of these longer overnight recordings. These recordings were split into multiple files. The specific files used, along with the names of the cultures and days of the recordings are listed in Table V. 30 Minute windows of spiking activity were extracted and used for network inference. Specifically, the number of target spikes N_X was set as the number of spikes that fell within this 30 minute window for the given target neuron.

The original study plated the electrodes with varying densities of cortical cells. However, overnight recordings were only performed on the 'dense' cultures, plated with a density of 2500 cells/µL.

The original study performed threshold-based spike detection by determining that a spike was present in the case of an upward or downward excursion beyond 4.5 times the estimated RMS noise of the recorded potential on a given electrode. The analysis presented in this paper makes use of these detected spike times. No spike sorting was performed and, as such, we are studying multi-unit activity (MUA) [61].

As the data was sampled at 25 kHz, uniform noise distributed between -20 µs and 20 µs was added to each spike time. This is to prevent the TE estimator from exploiting the fact that, in the raw data, inter-spike intervals are always an integer multiple of 40 µs.

Culture 1-1	day 4	day 14	day 20	
	2	2	2	
Culture 1-3	day 5	day 10	day 16	day 24
	2	2	2	2
Culture 2-2	day 9	day 15	day 21	day 33
	2	2	2	2

TABLE V: File numbers used for each culture on each day. These correspond to the file numbering used in the freely available dataset used in this study, provided by Wagenaar et. al.[41, 53]

G. GLM Model

The implementation of Generalised Linear Models (GLMs) of spiking activity followed that of Song et. al. [50] very closely. We briefly list the few minor differences.

For the B-spline basis functions, we excluded all knot locations beyond 100 ms. This was done due to the membrane potential decay time constant (τ) in the simulated models being set to 20 ms (see Table IV), implying that statistical relationships beyond 100 ms would be very unlikely.

Song et. al. [50] propose finding the penalty weight parameter λ using the Bayesian information criterion (BIC), by iteratively trialling various penalty weight values. Performing this for each target spike train would have been computationally prohibitive given the large networks and long simulation times used in this work. Instead, this step was performed on a few trial runs and a single value of $\lambda = 1{\times}10^{-3}$ was chosen as it closely approximated that chosen

by the BIC in all such trial runs.

We chose to designate the existence of a connection between a source and target when the GLM for the given target contained one or more non-zero weights assigned to the basis-splines associated with a given source.

Fitting of the GLM models was performed using the statsmodels [62] Python library.

CONTRIBUTIONS

David P. Shorten Conceptualisation, Wrote software, Analyzed data, Performed research, Wrote the paper, Edited the paper

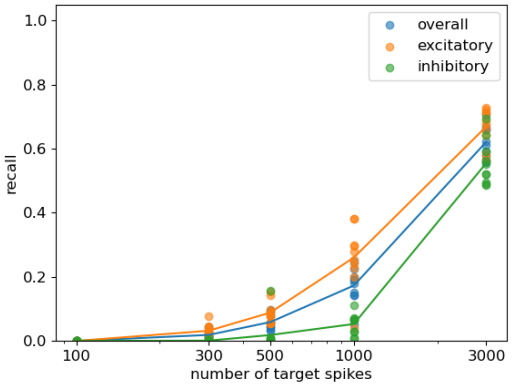
Michael Wibral Conceptualisation, Edited the paper

Viola Priesemann Conceptualisation, Edited the paper

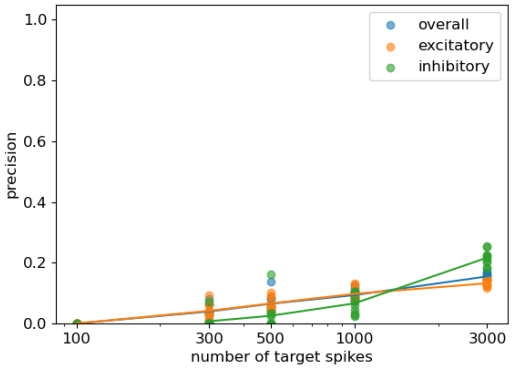
Joseph T. Lizier Conceptualisation, Wrote the paper, Edited the paper, Supervision, Funding Acquisition

Appendix A: Comparison with CoNNECT Algorithm and Generalised Linear Models

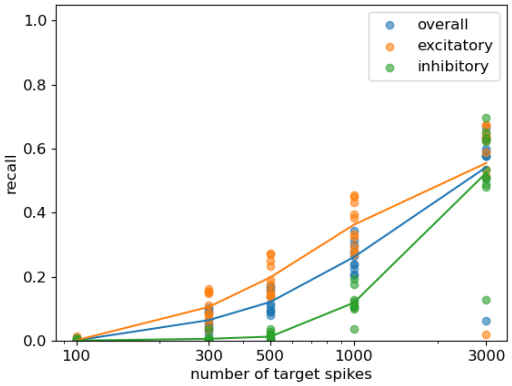
Plots identical to those in Fig. 1, but showing the results of applying competing spiking network inference techniques to the same data. Specifically, Fig. 8 applies the CoNNECT algorithm [40], which makes use of pretrained convolutional neural networks to classify the existence (or otherwise) of edges between spike trains. Fig. 9 shows the results of applying a GLM model of spiking activity [50] to the data, and basing the inference of connectivity on the existence of non-zero weights in this model.



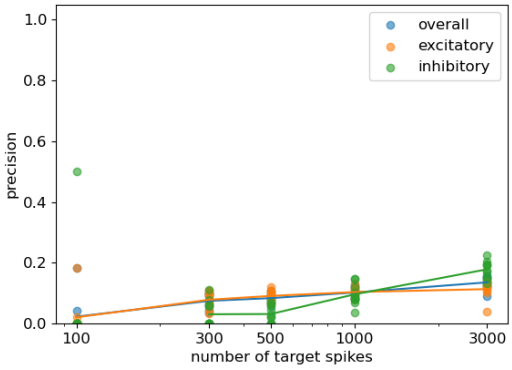
(a) Lower synchrony network, recall



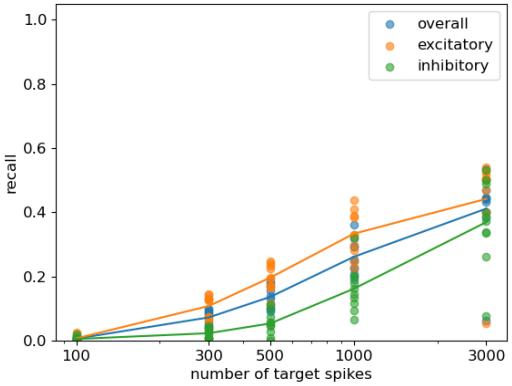
(b) Lower synchrony network, precision



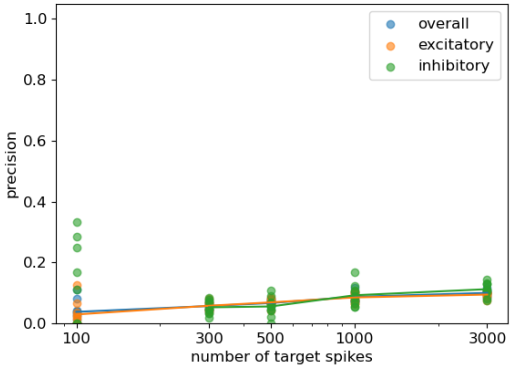
(c) Medium synchrony network, recall



(d) Medium synchrony network, precision



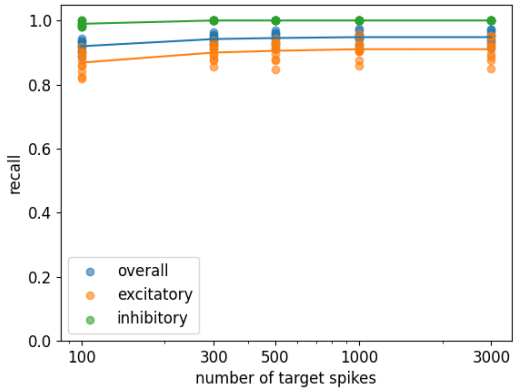
(e) High synchrony network, recall



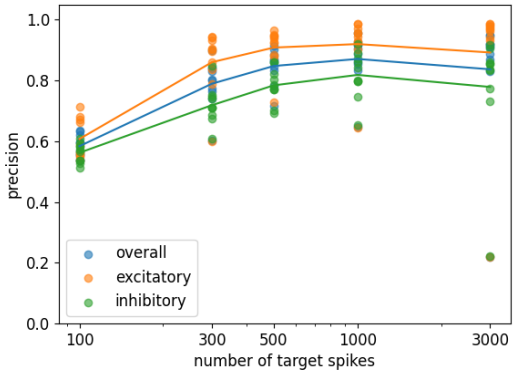
(f) High synchrony network, precision

FIG. 8: Plots showing the resulting precision and recall from running the CoNNECT algorithm [40] on networks of LIF neurons composed of 30 excitatory neurons and 20 inhibitory neurons. The ratio of the inhibitory to excitatory connection strength was varied in order to change the degree of synchrony in the network. Plots are shown for three different synchrony levels. Each plot contains points for the precision and recall for the inhibitory and excitatory neurons as well as the combined precision and recall. The lines pass through the means of these points.

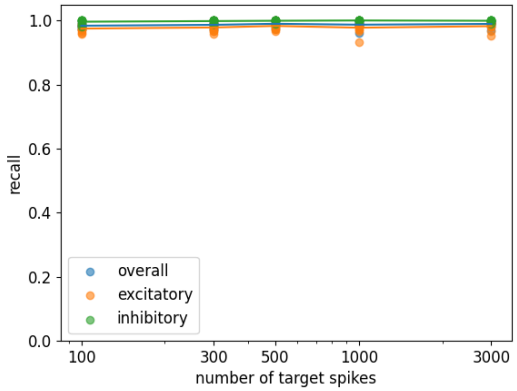
27



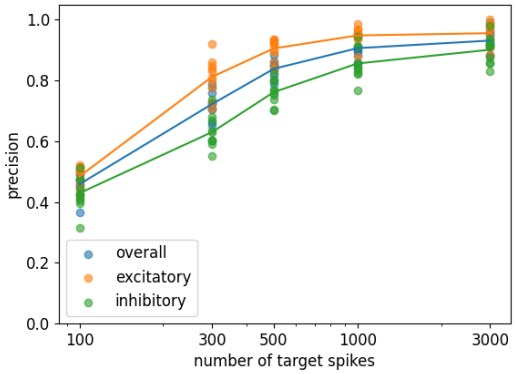
(a) Lower synchrony network, recall



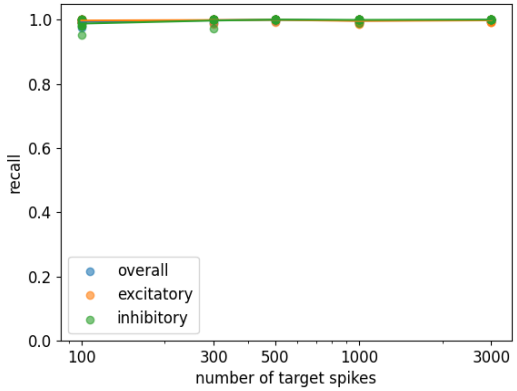
(b) Lower synchrony network, precision



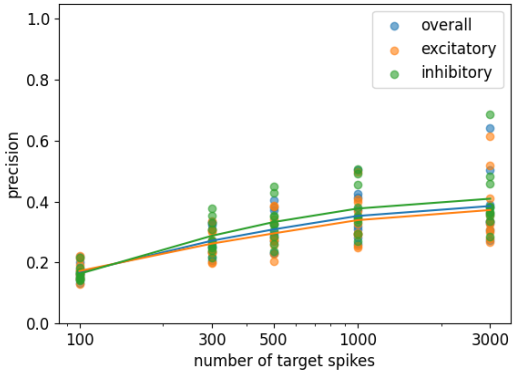
(c) Medium synchrony network, recall



(d) Medium synchrony network, precision



(e) High synchrony network, recall



(f) High synchrony network, precision

FIG. 9: Plots showing the resulting precision and recall from using a GLM model of spiking activity [50] to infer the connectivity of networks of LIF neurons composed of 30 excitatory neurons and 20 inhibitory neurons. The ratio of the inhibitory to excitatory connection strength was varied in order to change the degree of synchrony in the network. Plots are shown for three different synchrony levels. Each plot contains points for the precision and recall for the inhibitory and excitatory neurons as well as the combined precision and recall. The lines pass through the means of these points.

28

[1] I. H. Stevenson and K. P. Kording, How advances in neural recording affect data analysis, Nature Neuroscience **14**, 139 (2011).

[2] I. H. Stevenson, Tracking advances in neural recordings, `stevenson.lab.uconn.edu/scaling/#`, accessed: 2021-10-9.

[3] X. Yuan, M. Schröter, M. E. J. Obien, M. Fiscella, W. Gong, T. Kikuchi, A. Odawara, S. Noji, I. Suzuki, J. Takahashi, *et al.*, Versatile live-cell activity analysis platform for characterization of neuronal dynamics at single-cell and network level, Nature Communications **11**, 1 (2020).

[4] T. J. Sejnowski, P. S. Churchland, and J. A. Movshon, Putting big data to good use in neuroscience, Nature Neuroscience **17**, 1440 (2014).

[5] D. S. Bassett and O. Sporns, Network neuroscience, Nature Neuroscience **20**, 353 (2017).

[6] D. Bzdok and B. T. Yeo, Inference in the age of big data: Future perspectives on neuroscience, Neuroimage **155**, 549 (2017).

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

[8] S. Ryali, K. Supekar, T. Chen, and V. Menon, Multivariate dynamical systems models for estimating causal interactions in fMRI, Neuroimage **54**, 807 (2011).

[9] K. J. Friston, L. Harrison, and W. Penny, Dynamic causal modelling, Neuroimage **19**, 1273 (2003).

[10] J. Lizier and M. Rubinov, *Multivariate construction of effective computational networks from observational data*, Tech. Rep. 25 (Max-Planck-Institut für Mathematik in den Naturwissenschaften, 2012).

[11] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, Local measures of information storage in complex distributed computation, Information Sciences **208**, 39 (2012).

[12] A. Kraskov, H. Stögbauer, and P. Grassberger, Estimating mutual information, Physical Review E **69**, 066138 (2004).

[13] T. Schreiber, Measuring information transfer, Physical Review Letters **85**, 461 (2000).

[14] T. Bossomaier, L. Barnett, M. Harré, and J. T. Lizier, *An Introduction to Transfer Entropy* (Springer, 2016) pp. 65–95.

[15] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, Transfer entropy—a model-free measure of effective connectivity for the neurosciences, Journal of Computational Neuroscience **30**, 45 (2011).

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

[17] J. Sun, D. Taylor, and E. M. Bollt, Causal network inference by optimal causation entropy, SIAM Journal on Applied Dynamical Systems **14**, 73 (2015).

[18] I. Vlachos and D. Kugiumtzis, Nonuniform state-space reconstruction and coupling detection, Physical Review E **82**, 016207 (2010).

[19] M. Garofalo, T. Nieus, P. Massobrio, and S. Martinoia, Evaluation of the performance of information theory-based methods and cross-correlation to estimate the functional connectivity in cortical networks, PLoS One **4**, e6482 (2009).

[20] S. Ito, M. E. Hansen, R. Heiland, A. Lumsdaine, A. M. Litke, and J. M. Beggs, Extending transfer entropy improves identification of effective connectivity in a spiking cortical network model, PLoS One **6**, e27431 (2011).

[21] O. Stetter, D. Battaglia, J. Soriano, and T. Geisel, Model-free reconstruction of excitatory neuronal connectivity from calcium imaging signals, PLoS Computational Biology **8**, e1002653 (2012).

[22] J. G. Orlandi, O. Stetter, J. Soriano, T. Geisel, and D. Battaglia, Transfer entropy reconstruction and labeling of neuronal connections from simulated calcium imaging, PLoS One **9**, e98842 (2014).

615 [23] S. Nigam, M. Shimono, S. Ito, F.-C. Yeh, N. Timme, M. Myroshnychenko, C. C. Lapish, Z. Tosi, P. Hottowy, W. C. Smith,
616 *et al.*, Rich-club organization in effective connectivity among cortical neurons, Journal of Neuroscience **36**, 670 (2016).
617 [24] M. Shimono and J. M. Beggs, Functional clusters, hubs, and communities in the cortical microconnectome, Cerebral Cortex
618 **25**, 3743 (2015).
619 [25] E. Matsuda, T. Mita, J. Hubert, M. Oka, D. Bakkum, U. Frey, H. Takahashi, and T. Ikegami, Multiple time scales observed
620 in spontaneously evolved neurons on high-density CMOS electrode array, in *Artificial Life Conference Proceedings 13* (MIT
621 Press, 2013) pp. 1075–1082.
622 [26] N. Timme, S. Ito, M. Myroshnychenko, F.-C. Yeh, E. Hiolski, P. Hottowy, and J. M. Beggs, Multiplex networks of cortical
623 and hippocampal neurons revealed at different timescales, PLoS One **9**, e115764 (2014).
624 [27] M. Kajiwara, R. Nomura, F. Goetze, M. Kawabata, Y. Isomura, T. Akutsu, and M. Shimono, Inhibitory neurons exhibit
625 high controlling ability in the cortical microconnectome, PLoS Computational Biology **17**, e1008846 (2021).
626 [28] N. M. Timme, S. Ito, M. Myroshnychenko, S. Nigam, M. Shimono, F.-C. Yeh, P. Hottowy, A. M. Litke, and J. M. Beggs,
627 High-degree neurons feed cortical computations, PLoS Computational Biology **12**, e1004858 (2016).
628 [29] G. Mijatovic, Y. Antonacci, T. Loncar-Turukalo, L. Minati, and L. Faes, An information-theoretic framework to measure
629 the dynamic interaction between neural spike trains, IEEE Transactions on Biomedical Engineering **68**, 3471 (2021).
630 [30] B. Gourévitch and J. J. Eggermont, Evaluating information transfer between auditory cortical neurons, Journal of Neuro-
631 physiology **97**, 2533 (2007).
632 [31] D. P. Shorten, R. E. Spinney, and J. T. Lizier, Estimating transfer entropy in continuous time between neural spike trains
633 or other event-based data, PLoS Computational Biology **17**, e1008054 (2021).
634 [32] J. W. Aldridge and S. Gilman, The temporal structure of spike trains in the primate basal ganglia: afferent regulation of
635 bursting demonstrated with precentral cerebral cortical ablation, Brain Research **543**, 123 (1991).
636 [33] L. Rudelt, D. G. Marx, M. Wibral, and V. Priesemann, Embedding optimization reveals long-lasting history dependence
637 in neural spiking activity, PLoS Computational Biology **17**, e1008927 (2021).
638 [34] I. Nemenman, G. D. Lewen, W. Bialek, and R. R. D. R. Van Steveninck, Neural coding of natural stimuli: information at
639 sub-millisecond resolution, PLoS Computational Biology **4**, e1000025 (2008).
640 [35] C. Kayser, N. K. Logothetis, and S. Panzeri, Millisecond encoding precision of auditory cortex neurons, Proceedings of the
641 National Academy of Sciences **107**, 16976 (2010).
642 [36] S. J. Sober, S. Sponberg, I. Nemenman, and L. H. Ting, Millisecond spike timing codes for motor control, Trends in
643 Neurosciences **41**, 644 (2018).
644 [37] J. A. Garcia-Lazaro, L. A. Belliveau, and N. A. Lesica, Independent population coding of speech with sub-millisecond
645 precision, Journal of Neuroscience **33**, 19362 (2013).
646 [38] D. P. Shorten, V. Priesemann, M. Wibral, and J. T. Lizier, Early lock-in of structured and specialised information flows
647 during neural development, Elife **11**, e74651 (2022).
648 [39] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. Chichilnisky, and E. P. Simoncelli, Spatio-temporal correlations
649 and visual signalling in a complete neuronal population, Nature **454**, 995 (2008).
650 [40] D. Endo, R. Kobayashi, R. Bartolo, B. B. Averbeck, Y. Sugase-Miyamoto, K. Hayashi, K. Kawano, B. J. Richmond, and
651 S. Shinomoto, A convolutional neural network for estimating synaptic connectivity from spike trains, Scientific Reports
652 **11**, 1 (2021).
653 [41] D. A. Wagenaar, J. Pine, and S. M. Potter, An extremely rich repertoire of bursting patterns during the development of
654 cortical cultures, BMC Neuroscience **7**, 1 (2006).
655 [42] J. Runge, Causal network reconstruction from time series: From theoretical assumptions to practical estimation, Chaos:
656 An Interdisciplinary Journal of Nonlinear Science **28**, 075310 (2018).
657 [43] P. Spirtes, C. N. Glymour, R. Scheines, and D. Heckerman, *Causation, prediction, and search* (MIT Press, 2000).

[44] A. N. Burkitt, A review of the integrate-and-fire neuron model: I. homogeneous synaptic input, Biological Cybernetics **95**, 1 (2006).

[45] W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski, *Neuronal dynamics: From single neurons to networks and models of cognition* (Cambridge University Press, 2014).

[46] B. Kriener, H. Enger, T. Tetzlaff, H. E. Plesser, M.-O. Gewaltig, and G. T. Einevoll, Dynamics of self-sustained asynchronous-irregular activity in random networks of spiking neurons with strong synapses, Frontiers in Computational Neuroscience **8**, 136 (2014).

[47] Reconstructing neuronal circuitry from spike trains., `https://s-shinomoto.com/CONNECT/`, accessed: 2021-11-02.

[48] S. Linderman, C. H. Stock, and R. P. Adams, A framework for studying synaptic plasticity with neural spike train data, in *Advances in Neural Information Processing Systems*, Vol. 27, edited by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (Curran Associates, Inc., 2014).

[49] A. Calabrese, J. W. Schumacher, D. M. Schneider, L. Paninski, and S. M. Woolley, A generalized linear model for estimating spectrotemporal receptive fields from responses to natural sounds, PLoS One **6**, e16104 (2011).

[50] D. Song, H. Wang, C. Y. Tu, V. Z. Marmarelis, R. E. Hampson, S. A. Deadwyler, and T. W. Berger, Identification of sparse neural functional connectivity using penalized likelihood estimation and basis functions, Journal of Computational Neuroscience **35**, 335 (2013).

[51] S. Gerwinn, J. H. Macke, and M. Bethge, Bayesian inference for generalized linear models for spiking neurons, Frontiers in Computational Neuroscience **4**, 12 (2010).

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

[53] Network activity of developing cortical cultures *in vitro*, `http://neurodatasharing.bme.gatech.edu/development-data/html/index.html`, accessed: 2021-01-03.

[54] T. Varley, O. Sporns, H. Scherberger, and B. Dann, Information dynamics in neuronal networks of macaque cerebral cortex reflect cognitive state and behavior, bioRxiv , 2021.09.05.458983 (2021).

[55] A. Zalesky, A. Fornito, L. Cocchi, L. L. Gollo, M. P. van den Heuvel, and M. Breakspear, Connectome sensitivity or specificity: which is more important?, Neuroimage **142**, 407 (2016).

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

[57] L. Kozachenko and N. N. Leonenko, Sample estimate of the entropy of a random vector, Problemy Peredachi Informatsii **23**, 9 (1987).

[58] J. T. Lizier, Jidt: An information-theoretic toolkit for studying the dynamics of complex systems, Frontiers in Robotics and AI **1**, 11 (2014).

[59] J. Runge, Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information, in *International Conference on Artificial Intelligence and Statistics* (PMLR, 2018) pp. 938–947.

[60] A. Roth, M. C. van Rossum, *et al.*, Modeling synapses, Computational Modeling Methods for Neuroscientists **6**, 139 (2009).

[61] M. S. Schroeter, P. Charlesworth, M. G. Kitzbichler, O. Paulsen, and E. T. Bullmore, Emergence of rich-club topology and coordinated dynamics in development of hippocampal functional networks in vitro, Journal of Neuroscience **35**, 5459 (2015).

[62] S. Seabold and J. Perktold, Statsmodels: Econometric and statistical modeling with python, in *Proceedings of the 9th Python in Science Conference*, Vol. 57 (Austin, TX, 2010) p. 61.

CHAPTER 6

CONCLUSION

6.1 Summary of the Main Contributions

At a high level, this thesis presents the first high-fidelity study of information flows between the spiking activity of neurons. That is, it is the first time that the information flow between neurons has been studied without any loss of time precision whilst still considering history effects over reasonable time intervals.

Such a study was not possible until now due to the limitations of the traditional approach to estimating information flows (via TE) from event-based data (such as spike trains). This traditional approach operated by first discretising the process into bins of width Δt. The process was then cast as either a sequence of binary numbers or a sequence of natural numbers. For a binary sequence, each number represented the presence or absence of any spikes in the bin, whereas natural numbers would indicate the number of spikes that occurred within the bin. A straightforward plugin estimator was then applied to this data, which simply counts the frequency of values in the bins for each different unique history combination (see Section 2.3.1). There are a number of serious issues with this strategy for estimating TE. First and foremost, this estimation strategy is not consistent: it does not converge to the true value of the TE in the limit of infinite data. We also showed on some examples (Section 3.2.2) that it has very high bias and converges slowly. More importantly, it involves a fundamental tradeoff between being able to capture relationships occurring with fine temporal precision and those occurring over long periods of time. Using smaller bin sizes will reduce the loss of temporal precision, but also reduces the length of history that can be captured by the history embedding vectors. By contrast, larger bin sizes allow for longer histories to be represented, but reduce the temporal precision.

In Chapter 3, we presented a novel estimator of TE on event-based data (such as spike trains), which operates in continuous time and is able to overcome the above-mentioned challenges. This estimator is the first provably consistent estimator of TE on this data type. That is, it is the first estimator guaranteed to converge to the true value of the TE in the limit of infinite data. By operating on the raw inter-event intervals of the data, it does not lose any time precision. Moreover, for the usual spike density occuring in biological measurements, it can capture history effects occurring over fairly large intervals, with minimal use of dimensions. In Chapter 3, we also present an adaptation of a recently proposed local permutation method [1] for generating surrogate data, in order to perform statistical tests for non-zero TE. In Section 3.2.3 , we demonstrate that the traditionally-used time-shift method for surrogate generation can result in a very high false-positive rate when testing for non-zero

TE. The newly-proposed method, on the other hand, circumvents this issue by permuting the history embeddings correctly according to the null hypothesis of zero TE.

Given that we now have the ability to estimate TE with high-fidelity on event-based data, Chapter 4 makes use of this new ability to conduct the first ever high-fidelity study of information flows between the spiking activity of neurons. It does this using an openly-available dataset of recordings from developing cultures of dissociated cortical rat neurons [2]. Not only is this the first high-fidelity study of information flows between neurons, but it also represents the first study of information flows at different points in the development of neural cell cultures. Previous studies of information flows in neural cell cultures [3]–[9] studied recordings from mature cultures. By contrast, the dataset which we used contained recordings taken at multiple different days *in vitro*. This allowed us to contrast the information flows on earlier days with those on later days. We found that these flows exhibited an early lock in phenomenon, whereby the flows between nodes on earlier days of development were highly correlated with the flows on later days. We further found that nodes occupied specialised computational roles depending on their position in the burst propagation. Those nodes that tend to burst at the beginning of the propagation act as information transmitters and those that burst at the end of the propagation act as information receivers. By contrast, those that burst during the middle of the propagation perform a mixture of transmission and reception, occupying the vital role of the mediators of information flow. We also explored a plausible mechanism for these results by studying the information flows in simulated networks developing according to an STDP learning rule. We found that the changes in information flows exhibited remarkable similarities to those observed in the biological cell cultures, with early lockins and specialised computational roles both being present.

The networks inferred in Chapter 4 were *directed functional* networks. That is, each candidate source-target pair was considered in isolation. This resulted in very dense information flow networks. In Chapter 5, we move beyond this to consider multivariate information flows. That is, we are interested in whether there exists an information flow between a source and a target when conditioning on the activity of the rest of the neural population. Such an analysis aims to find the minimal set of sources for a given target whose histories will provide the maximal explanatory power of the activity of the target. The addition of further sources to the set will not reduce our uncertainty of the activity of the target, whereas removing sources from the set will increase our uncertainty. This is often referred to as the inference of an *effective* network. Although TE has been used for the inference of effective networks from other data modalities within neuroscience [10], [11], to date there has been minimal application of TE to the inference of effective networks from spike train data (the author is only aware of a single, very recent, contribution [12]). This is largely due to the limitations of the discrete-time estimator discussed above which prohibits the use of large conditioning sets and long embeddings. However, the new estimator proposed in Chapter 3 not only circumvents these issues, but also improves on the statistical testing for non-zero TE. As such, in Chapter 5, we take advantage of these exciting new advances to provide the first ever validation and application of TE to the task of inferring effective networks from spike train data. We utilise a (slightly-adapted) pre-existing greedy algorithm [10], [13] for the inference of effective networks using TE and then validate this approach on simulated spiking networks for which the ground truth is known. The approach is found to be able to achieve high accuracy with low data requirements. We also infer effective networks for the same dataset that was analysed in Chapter 4, providing a demonstration for its utility in deriving biological insights.

6.2 Directions for Future Research

6.2.1 Improving the Event-Based TE Estimator

One significant direction for future research is the improvement of the estimator presented in Chapter 3. In Section 3.4.1 , we showed how the TE on event-based data could be expressed as a sum of four KL-divergences. These KL-divergences are over the history embeddings of either the source, target and conditioning process or over just the history embeddings of the target and conditioning processes. These history embeddings are taken either at the events, or at randomly sampled points along the process. In the estimator presented in Chapter 3, the KL-divergences are estimated using relatively simple k-nearest neighbour estimators (with the addition of a radius sharing strategy in order to reduce bias). Although this scheme for estimating the divergences was quite adequate for the demonstration of the efficacy of this general approach for the estimation of TE on event-based data, there has been substantial recent research on the estimation of divergences [14]–[17]. By incorporating some of these results into the estimation of the divergence terms, we can expect to see reductions in bias as well as a better ability to scale to more dimensions. Better scaling over dimensions is particularly exciting, as it will allow for the handling of more conditioning processes in the network inference task.

One potential avenue for improvement involves advancements in the k-NN class of estimators for divergences. As just mentioned, in Chapter 3 we presented a strategy for estimating the necessary divergence terms using reasonably simple k-nearest neighbour estimators. There are a number of recent advances in this class of estimator that could be easily incorporated into the approach presented there. A particularly attractive possibility is the use of ensembles across various values of k [18]–[20], which has been shown to substantially reduce the bias of the estimators. The Kozachenko-Leonenko [21] class of k-NN estimators that we use assume that the probability density is constant in the ε-ball that surrounds the k closest neighbours. This assumption is often violated in the case high dimension relative to the number of data points. One avenue for ameliorating this issue is to use a shape which can more closely enclose the k points, and assume constant probability density within it. Suggested shapes include hyper-rectangles [22] and ellipsoids [23]. An alternate strategy is to relax the assumption of local uniformity, instead assuming that the probability distribution is locally Gaussian [24].

There has been significant recent research into methods for estimating divergences that do not rely on nearest-neighbour searches. An obvious extension to the estimation strategy presented in Chapter 3 would be to investigate the use of these newer techniques to estimate the two divergence terms required for the estimation of the TE. Notable examples of newer estimation approaches include: variational techniques [15]–[17], and the use of dependence graphs [14].

6.2.2 Further Applications

Another important direction for future work would be the further application of the novel estimator to spike train data. The study presented in Chapter 4, which analysed recordings of the spiking activity of neurons on different days of development, made use of an older dataset [2] (collected in 2006). As the recordings were collected sparsely (with many missing days between individual recordings), spike sorting was not practical. This is because it would not be possible to tell whether a

given sorted unit on one day was the same sorted unit on a different day. Some more recent work has performed continuous long-term recordings of neural cell cultures [25]. We could apply a spike-sorting algorithm capable of performing drift-tracking (that is, tracking the changes in the action-potential shapes over time, eg: [26]), to this data. This would provide us with spike-sorted data with consistent unit identities across development. We could then use this to study how information flows changed over the course of development between individual neural units, as opposed to between electrodes.

Another possibility would be to apply the presented estimation technique to modern cell-culture recordings collected at incredibly high spatial resolution. Modern high-density electrode arrays allow for recordings from cell cultures to be performed with far higher spatial resolution than the recordings used in Chapter 4 [27]. This then allows for incredibly accurate spike sorting, where we can be certain about the identities of the individual neurons associated with each spike. Given that this then provides us with full observability of the system, this type of data would be well suited to the application of the full effective network approach presented in Chapter 5.

It is worth bearing in mind that the development of neural networks in cell cultures does not perfectly mimic the development of neural tissue in animals [28], [29]. It is unclear how the differences between cell cultures and natural brain tissue will affect the development of information flows. As such, there is some uncertainty concerning to what degree the results presented in Chapter 4 are applicable to natural nervous systems. This implies that a clear focus for future application work will be applying the techniques presented in this thesis to recordings from live animals or experimental models that more faithfully mimic natural neural tissue [28].

The inference of effective networks has become an incredibly popular technique for analysing neuroscientific recordings [30]. However, its application to recordings of spiking neurons has been more limited. This is at least in part due to the limitations of previous information-theoretic estimation techniques when applied to event-based data. In particular, their inability to handle long-range history effects with high temporal fidelity has made them unsuitable for this task. The new continuous-time TE estimator which we have presented in this thesis, and validated in the context of network inference, circumvents these issues. It therefore opens up the possibility of much more widespread application of effective network analysis to spiking data within the neuroscience community.

6.3 References

[1] J. Runge, "Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information," in *International Conference on Artificial Intelligence and Statistics*, PMLR, 2018, pp. 938–947.

[2] D. A. Wagenaar, J. Pine, and S. M. Potter, "An extremely rich repertoire of bursting patterns during the development of cortical cultures," *BMC Neuroscience*, vol. 7, no. 1, pp. 1–18, 2006.

[3] S. Nigam, M. Shimono, S. Ito, *et al.*, "Rich-club organization in effective connectivity among cortical neurons," *Journal of Neuroscience*, vol. 36, no. 3, pp. 670–684, 2016.

[4] M. Shimono and J. M. Beggs, "Functional clusters, hubs, and communities in the cortical microconnectome," *Cerebral Cortex*, vol. 25, no. 10, pp. 3743–3757, 2015.

[5] E. Matsuda, T. Mita, J. Hubert, *et al.*, "Multiple time scales observed in spontaneously evolved neurons on high-density cmos electrode array," in *Artificial Life Conference Proceedings 13*, MIT Press, 2013, pp. 1075–1082.

[6] N. Timme, S. Ito, M. Myroshnychenko, *et al.*, "Multiplex networks of cortical and hippocampal neurons revealed at different timescales," *PloS One*, vol. 9, no. 12, e115764, 2014.

[7] M. Kajiwara, R. Nomura, F. Goetze, *et al.*, "Inhibitory neurons exhibit high controlling ability in the cortical microconnectome," *PLoS Computational Biology*, vol. 17, no. 4, e1008846, 2021.

[8] N. M. Timme, S. Ito, M. Myroshnychenko, *et al.*, "High-degree neurons feed cortical computations," *PLoS Computational Biology*, vol. 12, no. 5, e1004858, 2016.

[9] M. Wibral, C. Finn, P. Wollstadt, J. T. Lizier, and V. Priesemann, "Quantifying information modification in developing neural networks via partial information decomposition," *Entropy*, vol. 19, no. 9, p. 494, 2017.

[10] L. Novelli, P. Wollstadt, P. Mediano, M. Wibral, and J. T. Lizier, "Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing," *Network Neuroscience*, vol. 3, no. 3, pp. 827–847, 2019.

[11] L. Novelli and J. T. Lizier, "Inferring network properties from time series using transfer entropy and mutual information: Validation of multivariate versus bivariate approaches," *Network Neuroscience*, vol. 5, no. 2, pp. 373–404, 2021.

[12] P. C. Antonello, T. F. Varley, J. Beggs, M. Porcionatto, O. Sporns, and J. Faber, "Self-organization of in vitro neuronal assemblies drives to complex network topology," *bioRxiv*, 2021.

[13] J. Sun, D. Taylor, and E. M. Bollt, "Causal network inference by optimal causation entropy," *SIAM Journal on Applied Dynamical Systems*, vol. 14, no. 1, pp. 73–106, 2015.

[14] M. Noshad, Y. Zeng, and A. O. Hero, "Scalable mutual information estimation using dependence graphs," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 2962–2966.

[15] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, "On variational bounds of mutual information," in *International Conference on Machine Learning*, PMLR, 2019, pp. 5171–5180.

[16] J. Song and S. Ermon, "Understanding the limitations of variational mutual information estimators," *arXiv preprint arXiv:1910.06222*, 2019.

[17] M. I. Belghazi, A. Baratin, S. Rajeshwar, *et al.*, "Mutual information neural estimation," in *International Conference on Machine Learning*, PMLR, 2018, pp. 531–540.

[18] K. R. Moon, K. Sricharan, K. Greenewald, and A. O. Hero, "Ensemble estimation of information divergence," *Entropy*, vol. 20, no. 8, p. 560, 2018.

[19] K. R. Moon, K. Sricharan, K. Greenewald, and A. O. Hero, "Improving convergence of divergence functional ensemble estimators," in *2016 IEEE International Symposium on Information Theory (ISIT)*, IEEE, 2016, pp. 1133–1137.

[20] K. R. Moon and A. O. Hero, "Ensemble estimation of multivariate f-divergence," in *2014 IEEE International Symposium on Information Theory*, IEEE, 2014, pp. 356–360.

[21] L. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Problemy Peredachi Informatsii*, vol. 23, no. 2, pp. 9–16, 1987.

[22] S. Gao, G. Ver Steeg, and A. Galstyan, "Efficient estimation of mutual information for strongly dependent variables," in *Artificial Intelligence and Statistics*, PMLR, 2015, pp. 277–286.

[23] C. Lu and J. Peltonen, "Enhancing nearest neighbor based entropy estimator for high dimensional distributions via bootstrapping local ellipsoid," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 5013–5020.

[24] S. Gao, G. V. Steeg, and A. Galstyan, "Estimating mutual information by local gaussian approximation," *arXiv preprint arXiv:1508.00536*, 2015.

[25] J. Kreutzer, L. Ylä-Outinen, A.-J. Mäki, M. Ristola, S. Narkilahti, and P. Kallio, "Cell culture chamber with gas supply for prolonged recording of human neuronal cells on microelectrode array," *Journal of Neuroscience Methods*, vol. 280, pp. 27–35, 2017.

[26] N. A. Steinmetz, C. Aydin, A. Lebedeva, *et al.*, "Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings," *Science*, vol. 372, no. 6539, 2021.

[27] X. Yuan, M. Schröter, M. E. J. Obien, *et al.*, "Versatile live-cell activity analysis platform for characterization of neuronal dynamics at single-cell and network level," *Nature Communications*, vol. 11, no. 1, pp. 1–14, 2020.

[28] J. G. Roth, M. S. Huang, T. L. Li, *et al.*, "Advancing models of neural development with biomaterials," *Nature Reviews Neuroscience*, vol. 22, no. 10, pp. 593–615, 2021.

[29] E. Di Lullo and A. R. Kriegstein, "The use of brain organoids to investigate neural development and disease," *Nature Reviews Neuroscience*, vol. 18, no. 10, pp. 573–584, 2017.

[30] O. Sporns, *Networks of the Brain*. MIT press, 2010.