ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

DATA SCIENCE AND COMPUTATION

Ciclo 35

Settore Concorsuale: 09/G2 - BIOINGEGNERIA

Settore Scientifico Disciplinare: ING-INF/06 - BIOINGEGNERIA ELETTRONICA E INFORMATICA

UNCOVERING THE ENCODING AND TRANSMISSION OF BEHAVIOURALLY RELEVANT INFORMATION IN NEURAL ACTIVTY

Presentata da: Marco Celotto

Coordinatore Dottorato
Daniele Bonacorsi

Supervisore
Stefano Vittorio Tiziano Panzeri

Co-supervisore
Andrea Cavalli

Esame finale anno 2024

Uncovering the encoding and transmission of behaviourally relevant information in neural activity

Marco Celotto

Abstract

Most cognitive functions require the encoding and routing of information across distributed networks of brain regions. Information propagation is typically attributed to physical connections existing between brain regions, and contributes to the formation of spatially correlated activity patterns, known as functional connectivity. While structural connectivity provides the anatomical foundation for neural interactions, the exact manner in which it shapes functional connectivity is complex and not yet fully understood. Additionally, traditional measures of directed functional connectivity only capture the overall correlation between neural activity, and provide no insight on the content of transmitted information, limiting their ability in understanding neural computations underlying the distributed processing of behaviorally-relevant variables.

In this work, we first study the relationship between structural and functional connectivity in simulated recurrent spiking neural networks with spike timing dependent plasticity. We use established measures of time-lagged correlation and overall information propagation to infer the temporal evolution of synaptic weights, showing that measures of dynamic functional connectivity can be used to reliably reconstruct the evolution of structural properties of the network.

Then, we extend current methods of directed causal communication between brain areas, by deriving an information-theoretic measure of Feature-specific Information Transfer (FIT) quantifying the amount, content and direction of information flow. We test FIT on simulated data, showing its key properties and advantages over traditional measures of overall propagated information. We show applications of FIT to several neural datasets obtained with different recording methods (magneto- and electro-encephalography, spiking activity, local field potentials) during various cognitive functions, ranging from sensory perception to decision making and motor learning. Overall, these analyses demonstrate the ability of FIT to advance the investigation of communication between brain regions, uncovering the previously unaddressed content of directed information flow.

Acknowledgments

I would like to express my gratitude to my supervisor, Stefano Panzeri. Your invaluable feedback and guidance not only helped shape my scientific perspective but also inspired me to surpass my own expectations.

I would like to thanks Stefan Lemke for collaborating extensively on the paper that forms Chapter 4, along with Stefano Panzeri. Additionally, I would like to thanks Stefan for his scientific enthusiasm. Working with you felt less like a task and more like a shared passion.

I extend my gratitude to all the collaborators I've had the privilege to work with over these years, including, but not limited to, Loren Koçillari, Hamed Nili, Alessandro Toso, Daniel Chicharro, Alejandro Tlaie Boria, Hame Park, Vito De Feo, Jan Bím, Roberto Maffulli, Tobias Donner, Ileana Hanganu-Opatz, Andrea Brovelli, Michele Zoccali, Malte Bieler, Ilaria Zanchi, Mariangela Panniello, and Karunesh Ganguly. Science thrives on a synergistic effort, and much of my progress would not have been achievable without your technical expertise and personal support.

A special thanks to all my friends and colleagues at the Department of Excellence for Neural Information Processing, DonnerLab and Cognition and Action Lab at UKE. Being a part of such an incredible community made the everyday challenges of pursuing a PhD far more manageable and fun.

I would like to thanks my mom, my dad, and my whole family for always giving me so much, without asking for anything back. Your love over these years pushed me to always give my best.

I want to express my profound appreciation to all my lifetime friends. Your unwavering support, stimulating discussions, and boundless love have been pivotal in my journey. The strength from our bonds has been my bedrock of resilience, and I carry you all with me wherever I go.

Lastly, I would like to thanks Alessandra: of all the experiences and memories from my PhD years, you are the best one.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 The brain as an information processing machine

The mammalian brain is possibly the most complex [1, 2] and efficient [3, 4] information processing machine in nature. It constantly orchestrates a huge variety of operations, underlying our thoughts, actions, emotions, and perceptions [5–7]. These operations are supported by different degrees of structural and functional complexity, ranging from controlling simple reflexes [8] using relatively simple neural circuits, to higher cognitive processes like decision-making and problem-solving [9, 10], involving the integrated activity of neural populations in several brain areas [11, 12]. The brain can be broadly divided into two main regions: the cerebral cortex and the subcortical structures [13]. The cerebral cortex, is the exterior part of the brain and is thought to be responsible for many of our higher-order cognitive functions, such as sensory perception [14], voluntary and goal-directed motor control [15, 16], and decision making [17]. Beneath the cortex lie the subcortical structures, including the thalamus, basal ganglia, and brainstem, among others [13]. These areas play crucial roles such as regulating vital functions [18], driving automatic movements [19, 20], process early-stage sensory information [21–23], and releasing neurotransmitters [24–26].

Neurons are the primary functional units of the brain. Central to their function is the ability to generate action potentials, also known as spikes. Spikes are brief electrical impulses generated when the summation of excitatory inputs to a neuron surpasses a certain threshold, resulting in a nonlinear depolarization of the neuron's membrane potential [13, 27]. Once generated, the action potential *propagates* down the neuron's axon, ultimately leading to the release of neurotransmitters into synapses, influencing neighboring neurons. When large groups of neurons fire action potentials in a synchronized manner, the resultant electric activity summates, producing detectable macroscopic electrical signals [28].

At the core of its operations, the brain encodes environmental stimuli into neural activity [14, 29], transfers information through its densely connected networks [30,

31], and uses such information to generate actions or drive learning [9]. However, neural activity is very variable [32] as it results from the integrated dynamics of many stochastic units (e.g. ion channels) and due to the concurrent processing of many cognitive and regulatory functions [33]. To partially reduce such variability when trying to understand neural information processing, it is common to study neural activity collected during environmentally controlled cognitive tasks.

At the microscopic scale, neural systems comprise billions of neurons interconnected via synapses [13], while from a macroscopic perspective the brain can be conceptualized as a system of highly interconnected brain areas [1] functionally specialized yet adapting their function depending on the ongoing computations [10, 29]. The dense connectivity in neural circuitry and brain areas led researchers to conceptualize structural (i.e. physical links between regions) and functional (i.e. correlations in neural activity) connections in the brain as forming networks of neurons or areas [1]. Crucially, both types of connections are dynamic: they evolve over time due to the temporal plasticity of physical connections [34, 35] and the varying cognitive and operational states governing functional interactions [36–38]. However, the relationship between structural and functional connectivity remains elusive [39–42]. In Chapter two, we use measures of time-lagged correlation and causal information transfer to study whether, in simulated recurrent spiking neural networks with spike timing-dependent plasticity [43], it is possible to infer the temporal evolution of structural properties of the network from measures of dynamic functional connectivity [44].

Our capacity to explore neural information processing magnifies as technological advancements, notably in throughput-intensive neural-data-acquisition techniques [45–47] and computational resources available to process data [48, 49], ease the recording and analysis of activity simultaneously recorded from several neurons or brain areas. It is essential to accompany such technological developments with new versatile mathematical tools that can be used to analyze distributed computations involving the encoding and transmission of task-relevant variables across several regions. In Chapter three and four, we derive and validate on several simulated and real neural datasets [50, 51] a new measure, termed Feature-specific Information Transfer (FIT). FIT advances traditional measures of causal communication between the activity of simultaneously recorded brain regions, by isolating the components of information transmitted about specific variables of interest.

1.2 Recording neural activity

Nowadays experimental techniques allow the recording of neural activity across many spatial and temporal scales.

Recordings with individual cell resolution, allowed researchers to study fundamental properties of neural information processing, such as the role of individual spikes (in the scale of milliseconds) in encoding sensory information [52, 53] and the

emergence of population-codes arising from the tiling of information across single cells in space and time [17, 54, 55]

The superposition of the coortinated action potentials of many neuron generate frequency-rich voltage fluctuations in the extracellular medium [28]. These lower dimensional signals capture the aggregate activity of hundreds up to billions of neurons, depending on the recording technique. Over the decades, recordings of these aggregate signals shed light on the multiple functions of oscillations in the brain, from providing a common reference relative to which spiking activity can carry information [56], to the study of information carried [11, 57, 58] and transmitted [30, 31, 59] by neural activity in specific frequency-bands.

1.2.1 Invasive recording techniques

Invasive recording techniques, such as electrophysiology and calcium imaging, allow measuring the spiking activity of individual cells and populations of neurons with high temporal resolution (milliseconds to tens of milliseconds).

Electrophysiology uses electrodes to measure voltage changes in the extracellular medium, capturing a range of signals, from low-frequency fluctuations to high-frequency spikes. Broadband recording can be high-passed filtered and processed with spike-sorting algorithms [60] to isolate the spike trains from up to hundred of individual neurons [45]. In Chapter two, we study communication between simulated spike trains obtained from the activity of recurrent spiking neural networks [44]. In Chapter four, we include analyses of spike train data recorded from motor cortex and dorsolateral striatum of rats learning a reach-to-grasp task [51].

Broadband activity obtained from electrophysiological recordings can also be used to measure aggregate (or mass) signals from neural populations. The two most common aggregate electrophysiological signals are multi-unit activity (MUA) and local-field potentials (LFPs) [28]. MUA is the combined spiking activity of multiple neurons nearby the recording site. It is derived by high-pass filtering the raw extracellular recording to capture the fast voltage transients associated with action potentials of neurons. LFPs reflect the slower changes in the extracellular voltage that arise primarily from the combined synaptic activity within a local population of neurons [61] - typically on a larger spatial scale than the one captured by MUA. LFPs are obtained by low-pass filtering the raw extracellular voltage recording.

In Chapter three, we show applications of FIT to MUA data recorded from the thalamocortical system of mice, to prove the specificity of sensory modality processed by different thalamocortical pathways [50]. In Chapter four, we include extensive analyses of LFP recordings from motor cortex and dorsolateral striatum of rats learning of a reach-to-grasp task [51].

Another common invasive imaging technique is calcium imaging [17, 62], in which calcium-sensitive fluorescent indicators are used to optically measure the activity of cells [63]. This technique allows recording from large-scale populations of neurons [46], and is compatible with optogenetic techniques for the manipulation of neural

activity, giving the possibility to simultaneously record and perturb neural circuits [9]. Additionally, calcium imaging has been key in discovering the active information processing properties of glial cells in the brain, such as astrocytes [64]. Since these cells do not generate electric activity, their rich calcium dynamics have been overlooked for decades, and recent studies performed with calcium imaging started unraveling their contribution in different aspects of neural computations [65–67].

1.2.2 Non-invasive recording techniques

Non-invasive neuroimaging techniques provide a window into the brain's activity without the need for surgical interventions, minimizing the associated risks. Most common techniques include Electroencephalography (EEG), Magnetoencephalography (MEG) and functional Magnetic Resonance Imaging (fMRI).

EEG measures the electrical activity generated by the synchronous activity of populations of neurons, recorded using electrodes placed on the scalp [28]. It provides high temporal resolution, on the order of milliseconds, making it particularly suitable for studying the dynamics of neural oscillations and event-related potentials. However, its spatial resolution is limited.

In Chapter three we present application of FIT to EEG recordings, to study stimulus-feature specific information transfer across hemispheres in the human brain during a face detection task [50] .

MEG measures the magnetic fields produced by neural electrical activity using highly sensitive magnetometers. Like EEG, MEG offers high temporal resolution, but has typically higher spatial resolution. Additionally, magnetic fields are less distorted by the skull and scalp compared to electric ones [28]. However, MEG requires specialized facilities due to its high sensitivity to external magnetic noise.

In chapter three, we show applications of FIT to MEG recordings, quantifying the stimulus- and choice-specific information transfer in a network of visual regions of human participants during a perceptual decision-making task [50].

Functional Magnetic Resonance Imaging (fMRI) measures changes in blood oxygenation levels, which is an indirect indicator of neural activity. It offers high spatial resolution, allowing researchers to discern activity in different cortical and subcortical regions. However, its temporal resolution is lower than EEG or MEG, capturing changes over seconds rather than milliseconds.

1.3 Information theory in neuroscience

The brain continuously performs a huge set of operations. Therefore, even in controlled experimental scenarios, neural activity is highly variable across experimental trials [32]. Variability in neural recordings can be divided as coming from two main, conceptually distinct sources. The first source of variability is the intrinsic ongoing spontaneous activity that is not related to the feature of interest (determined by endogenous factors including the ongoing operational state of neural populations [29],

attention [68], thirst [69], or the processing of other relevant variables). The second is the measurement noise arising from the recording technology, biological sources like muscle activity, and external interferences. This second type of variability can be reduced using appropriate signal processing techniques.

Given such large variability, most common techniques to study the relationship between neural activity and external world variables rely on probabilistic approaches. Such techniques include generalized linear models (GLMs) [17, 70], linear discriminant analysis, support vector machines [71], and more sophisticated deep-learning-based decoders [72]. While these different modelling techniques proved to be key in understanding how external world variables are encoded in neural activity, they are conceptually limited by the assumptions (e.g., on the type of noise and the nature of interactions) they take when describing neural activity. A very natural and almost assumption-free probabilistic framework to study neural computations is information theory [73], which revolves around a very general definition of variability and relationship between stochastic variables.

1.3.1 A brief history of information theory in neuroscience

First quantitative works posing the basis to current conceptualization of the brain as an information processing system go back to early 1940s. In 1943 Pitts and McCulloch [74] proposed neurons as binary threshold units capable of logical computations, introducing the first mathematical model imitating the functionality biological neurons. A few years later, in 1948 C. Shannon [75] provided a mathematical definition of the elusive concept of information and introduced a framework (information theory) which proved to be essential in designing artificial (and understanding biological) computing systems [76]. Upon the introduction of information theory, McCulloch [77] promptly recognized Shannon's work and applied it shortly after [78] to set theoretical limits to the amount of information that can be transmitted through a synapse, showing that neurons can in principle transmit large amount of information. This line of research lead to the first attempts to use information theory to characterize information encoding in real neural systems [79].

As it was succinctly expressed by D. Perkel and T. Bullock in 1968: "*Whereas the heart pumps blood and the lungs effect gas exchange, ..., the nervous system processes information*" [80]. As suggested by its name, information theory offers a very natural yet rigorous framework to study information processing. Unsurprisingly then, over the course of the following decades, information theory has been successfully applied to study the encoding and transmission of information across several spatial and temporal scales in individual neurons [81], neural populations [54, 82, 83], and large brain networks [31, 84, 85].

1.3.2 Information Encoding

At the center of information theory lies the concept of *entropy* [75]. Entropy of a random variable X is a quantity that can be computed from the probability

distribution $P(X)$ and measures the overall variability (or uncertainty) of X [73]. In our context, the variable X could be the number of spikes emitted by a neuron in a given time window [81, 86]. By quantifying the intersection between the entropy of neural response X and the one of an external world feature S (such as a feature of a sensory stimulus), *mutual information* measures the amount of neural variability that is explained, at a single-trial level, by the variability of the feature. This approach is conceptually similar to quantifying the information encoded by X about feature S as the explained variance coefficient R^2 of a Pearson correlation between X and S, or the deviance explained about X by the predictor S using a GLM [17, 70]. However, mutual information quantifies relationship in term of the full probability distribution of X and S, and is therefore a non-parametric measure sensitive to both linear and non-linear interactions between variables [87, 88].

Moreover, the information theoretic framework is intrinsically fit for studying the relationship between source variables (e.g. two neurons X_1 and X_2) in carrying information about a target (such as a feature of a sensory stimulus S). Indeed, given that mutual information is additive for independent variables [73], if X_1 and X_2 independently encode S, then the joint mutual information carried by the two neurons is equal to the sum of the information carried individually by the two cells [87, 89]. If the sum of the two individual information is larger than the joint information, then the two neurons carry redundant (or shared) information about S. If, instead, the sum of the individual information is smaller than the joint information, the two neurons carry synergistic information about S, i.e. a component of information that is not present in either neuron alone but is available when observing both neurons together.[1]

1.3.3 Information Transmission and Causality

Common neural computations involve the transmission of activity and information across networks of neurons or brain areas. Indeed, even in regular sensorimotor processing, sensory variables are first encoded in subcortical areas (typically in the thalamus), they are transmitted to primary sensory cortical areas, then further flow toward downstream areas, where information is integrated with contextual and previously acquired knowledge, and finally be transformed into appropriate actions [90]. To perform this type of processing, areas have to communicate to each other, sending and receiving information about specific variables at each stage of the computation.

[1] To better explain redundancy and synergy, we make two examples. In the first example X_2 is a copy of X_1. Therefore, the joint feature information carried by X_1 and X_2 equals the information encoded individually by each variable. The joint information is half the sum of the two individual information, meaning that all of the information encoded in the response a neuron in redundant with the other one [87]. In the second example the feature S modulates the correlation between X_1 and X_2, without affecting the average response nor the noise magnitude of any neuron. In this example, X_1 and X_2 individually carry zero feature information but the joint information is larger than zero. This exemplifies synergistic information encoding provided by feature-dependent single-trial correlations [89].

Importantly, this communication always requires some time due to physical limits in the speed of propagation of electrical signals along axons [13]. Therefore, finding significant relationships in the time-lagged activity between a putative sender and a putative receiver area is mandatory to infer directed communication. However, time-lagged correlation in the activity is not enough to infer directed *causal* communication [91]. For the stronger claim of causality, it is required that the past activity of the sender explains the variability of receiver's present activity beyond the past of the receiver itself. Otherwise, current activity in the receiver could not be attributed communication with the sender. This principle of causal communication is know as *Wiener-Granger causality principle* [92, 93]. The application of Wiener-Granger causality principle to probabilistic models, typically in the form of Vectorial Autoregressive Models, has led to the definition of the measure of Granger causality [94]. Given its parametric nature, Granger Causality has been mainly applied to study communication between mass signals (whose noise distributions are approximately normal) simultaneously recorded in two or more areas [59, 95]. However, extensions of these methods also exist to non-normal data, and Granger causality has also been used to study communication in networks of neurons [83, 96].

Transfer Entropy (TE) encapsulates Wiener-Granger causality in non-parametric, information theoretic terms, and is equivalent to Granger casuality for linear Gaussian systems [97]. Transfer Entropy has also been widely applied to neuroscience [31, 85, 98], proving to be a reliable tool to infer brain communication, also being sensitive to nonlinear interactions [84]. However, TE is only sensitive to the propagation of overall information by neural activity, lacking the ability of specifically select feature-related components of transmitted information. To conceptually advance the investigation of information transfer in the brain, it is fundamental to overcome this limitation by developing methods quantifying how much of the transmitted information is about specific features of interest.

In Chapter two we use time-lagged correlations and transfer entropy to reconstruct structural connectivity properties from the activity of a recurrent neural network and infer the temporal evolution of the synaptic weights from measures of dynamic functional connectivity [44].

In Chapter three [50] and four we extensively apply TE to simulated and real neural data, showing its strengths and limitations in estimating the overall propagation of information between brain regions.

1.3.4 Partial Information Decomposition

Multivariate information theory has, so far, been crucial in understanding neural information processing. Many of most prominent application of information theory in neuroscience benefited from the rigorous framework offered by information theory when studying the relationship between several neurons or regions in encoding external world features [52, 54]. Over the years, theoretical developments

of multivariate Shannon information theory allowed researchers to study important properties of population coding, such as the role of correlations in enhancing or limiting the amount of information encoded at the population level [89, 99].

However, classical multivariate information theory has two main issues. The first one is that it can only quantify the net effect of synergistic and redundant information encoding [100], lacking the ability to provide a separate quantification of these conceptually different terms [76, 101]. The second is that multivariate information quantities defined within Shannon framework have issues generalizing to more than two source variables , providing unintuitive results when used to quantify redundancy or synergy between three or more variables [102].

In 2010, Williams and Beer [100] introduced a mathematical framework, called Partial Information Decomposition (PID), to quantify how information that two or more source variables carry about a target variable is distributed among the sources. PID braks down the joint mutual information carried by source variables about a target into non-negative components representing shared (redundant) encoding between the sources, unique encoding by some of the sources, or synergistic encoding in the combination of different sources [100, 103]. In the case of two source variables (e.g., two neurons) encoding a feature, PID decomposes the joint feature information carried by the two neurons into a piece of shared information carried by both neurons, two pieces of unique information (carried by either one of the two neurons, but not by the other), and a piece of synergistic information that is only available when observing the two neurons simultaneously.

By providing a separate quantification of redundancy and synergy, and a framework that naturally generalizes to three or more source variables, PID opened the venue to a new set of scientific questions that can be addressed about computations in neural [104–106], other biological [107], and arificial [108, 109] processing systems.

While PID is still a very active research field [110], with the community still discussing the fundamental properties of PID measures [111, 112] and structures [105, 113], in the recent years it has already provided methodological advances in the analysis of neural data. These include providing a rigorous information-theoretic definition of neural codes in decision making [83, 104] (i.e. quantify the amount of sensory information encoded in neural activity that is readout to guide decisions), separating redundant and synergistic functional connectivity between pairs of neurons or brain regions [41, 101, 114], and quantifying the shared, unique and synergitic representation of multimodal sensory stimuli in the brain [115, 116].

In Chapter three and four we use PID to define and validate on simulated and real neural data a new nonparametric measure of Feature-specific Information Transfer, that can capture - within the overall propagated information between areas - the amount of communication about a specific variable of interest. By merging the Wiener-Granger causality principle with content-specificity, FIT can improve our understanding of how brain regions communicate, uncovering previously unaddressed feature-specific information flow.

Chapter 2

Inferring the Temporal Evolution of Synaptic Weights from Dynamic Functional Connectivity

The content of this chapter was published and awarded as the best conference paper at the 15th International Conference on Brain Informatics, held in Padua, Italy, in July 15-17 2022 [117]. An extended and refined version of the paper was published, under invitation, in the Brain Informatics journal [44].

2.1 Introduction

Neurons in biological networks are sparsely connected by directed, plastic synapses, with communication delays that can vary across different pairs of cells [1, 118, 119]. The patterns of synaptic connectivity have a profound influence on the computations and functions of neural circuits [120–122]. Importantly, such synaptic connectivity is not static. The strength of each synapse can change over different time scales—ranging from milliseconds to days—due to processes including synaptic potentiation and depression [34]. Such changes in synaptic weights are thought to be neural-activity dependent and driven by local Hebbian mechanisms of plasticity such as spike timing-dependent plasticity (STDP). In these mechanisms, the potentiation and depression of synaptic weights depends on the precise temporal relationship between pre- and post-synaptic spikes [43].

It is challenging to directly measure time changes of synaptic weights in vivo. One possible approach to study in vivo changes in synaptic strength is to simultaneously record the spiking activity of several neurons within a network and estimate changes in their functional connectivity with the statistical analysis of simultaneous recordings. Though the relationship between fixed structural connectivity and "static" time-averaged functional connectivity (FC), in which FC is computed over long time intervals, has been studied extensively [39, 40, 123], how changes in synaptic and functional connectivity relate at different time scales remains unclear.

Understanding the relationship between changes in synaptic and functional connectivity is relevant to a range of neuroscientific questions, such as the role of sleep in synaptic homeostasis and memory formation. Several theories and experimental findings posit that non-REM sleep is accompanied by profound changes in anatomical synaptic connectivity, including the general down-scaling of synaptic connectivity related to homeostasis [124–126] as well as context-specific upscaling in synaptic connectivity, such as sleep-dependent dendritic spine formation after motor learning [127]. The anatomical and theoretical evidence for changes in synaptic strength in sleep have been accompanied by evidence for changes in FC, as observed across the motor network during motor learning [128, 129]. It remains challenging to relate the evidence for structural and functional changes during sleep [130, 131], as robust methods to relate dynamic functional connectivity (DFC) to the underlying temporal evolution of synaptic connectivity are not yet established.

Neural network models are a powerful tool to relate structural and functional connectivity, as the former is known because it is put into the model's equation by the modeler, and the latter can be computed by activity generated by the model [39, 132]. Previous studies have utilized network models of Izhikevich neurons [118] to investigate the relationship between FC measures and synaptic connectivity because these models are generated by simple equations that can produce firing patterns resembling several types of cortical neurons in vivo [133, 134]. These studies highlighted that static bivariate FC measures, such as cross-covariance and transfer entropy, provide robust estimates of the underlying fixed structural synaptic connectivity in simulated networks. However, they did not examine the temporal evolution of functional and synaptic connectivity within spiking networks incorporating STDP.

Here, we relate the temporal evolution of synaptic connectivity to DFC in a neural network model. We examined the performance of several different DFC methods in estimating the temporal dynamics of synaptic weights (termed dynamic synaptic connectivity or DSC) from up to 90 min of spiking activity in simulated spiking networks whose synaptic strength changed over time due to STDP. We first determined the performance of static FC measures in inferring fixed structural properties of the simulated networks (such as presence or absence of pairwise synaptic connectivity and the associated communication delays). We then applied these measures with a sliding time window approach to compute DFC and quantify its relationship with DSC. We found cross- covariance outperformed other DFC measures in capturing the evolution of synaptic weights over time. We also established how to use the information obtained from the static, time-averaged analysis of the network, to enhance the estimate of DSC from DFC.

2.2 Simulated spiking network and inference pipeline

To investigate the relationship between DSC and DFC, we simulated a spiking neural network in which the strength of synaptic weights changed over time according to an STDP rule. We then compared the performance of different functional connectivity measures in estimating both the ground truth structure of the network (i.e. which pairs of neurons were connected, their communication lag, and the type of synapse), and how the strengths of the synaptic weights changed over time (Fig. 2.1). We simulated a spiking network of N = 100 neurons in which the dynamics of each neuron was described using the Izhikevich neuron model [135]. In this model, the voltage v of each neuron is described by two coupled differential equations:

$$\begin{gathered} v' = 0.04v^2 + 5v + 140 - u + I_{syn}, \\ u' = a(bv - u), \\ \text{if } v(t) = 30mV \text{ then } v \leftarrow c \text{ and } u \leftarrow u + d \end{gathered} \tag{2.1}$$

where u is a recovery variable, prime symbols ($'$) denote time derivatives, I_{syn} is the total synaptic input to the neuron and (a, b, c, d) is a set of parameters controlling the firing behavior. Depending on the set of parameters, the Izhikevich model can reproduce several firing patterns observed in cortical neurons. As in the original Izhikevich cortical network model [118], we set $(a, b, c, d) = (0.02, 0.2, -65, 8)$ to simulate excitatory regular spiking neurons, and $(a, b, c, d) = (0.1, 0.2, -65, 2)$ for inhibitory fast spiking neurons. The term I_{syn} is a sum of the voltages generated by the firing of the presynaptic neurons plus an external input term. The external input term consisted of a voltage of 20 mV added to a randomly selected neuron in each simulation time step, as in Ref. [118]. The synaptic voltages were set to an initial value of 6 mV for excitatory synapses and -5 mV for inhibitory synapses, as in Ref. [118].

The structure of the network was set by Izhikevich [118] to mimic the connectivity of a real population of cortical neurons (Fig. 2.1A). 80% of neurons in the network were excitatory and 20% were inhibitory. Excitatory neurons were randomly connected to 10 neurons which could be either excitatory or inhibitory (800 excitatory synapses in total). Each excitatory synapse had a random communication delay (δ) whose value was uniformly distributed between 1 and $20ms$ and was constant over time. Inhibitory neurons were randomly connected to 10 excitatory neurons (200 inhibitory synapses), therefore no inhibitory-to-inhibitory (I-I) connections were present in the network. The lack of I-I synapses caused the average firing rate of excitatory neurons (5.11 ± 0.03 Hz) to be lower than the one of inhibitory neurons (8.23 ± 0.04 Hz). Inhibitory connections had a communication delay of $1ms$. The simulation ran with $1ms$ temporal precision for a duration decided by the user. During the simulation, the strength of excitatory synapses - which were all initialized to the same, positive, value - changed dynamically due to an STDP

rule: when a presynaptic neuron i fired before a postsynaptic neuron j the strength of the synapse from i to j (w_{ij}) was strengthened, on the other hand when j fired before i w_{ij} got weaker (Fig. 2.1B). The decay time of the STDP rule was $\tau = 20ms$ and synaptic weights were updated every $1s$ with a memory factor which made the weights change, on average, over the timescale of 1-2 minutes (obtained measuring the synaptic weights autocorrelation, not shown).

We used different measures to compute the static and dynamic functional connectivity of the network from the spiking activity (Fig. 2.1C). Such measures were all directional and allowed the computation of the strength of communication for different delays (δ). When computing static functional connectivity, we used data from the whole simulated recording to compute a single connectivity value for each pair of neurons (i, j). We computed all connectivity measures with δ ranging from 1 to $50ms$ then, for each pair, we determined the static functional connectivity (w_{ij}) as the maximum connectivity value across delays. We selected the communication delay (δ_{ij}) as the lag that maximized the functional connectivity. We did not compute FC measures at delay values equal to zero, since spike propagation and synaptic transmission requires time to occur. Even if zero-lag correlations between real neurons have been reported [136], such forms of FC are most likely due to the presence of coordinated activity driven by other areas, and not due to the presence of synapses between neuron pairs. Calling $f_{ij}(\delta)$ the generic measure of functional connectivity, then $w_{ij} = max_\delta(f_{ij}(\delta))$ and $\delta_{ij} = argmax_\delta(f_{ij}(\delta))$. By taking the top percentile of connectivity values for each measure we obtained sparse static networks (Fig. 2.1D). If the measure f was signed we could also infer whether a synapse was excitatory or inhibitory. Then, we used a sliding window approach to compute, for each measure, the DFC of all the synapses that were inferred as present (Fig. 2.1E). We exploited the static measures of communication delay between pairs to compute delay-consistent DFC and then evaluated the performance of the different measures in recovering the ground-truth dynamics of synaptic weights.

2.3 Inferring the presence of synapses

We tested the performance of different measures of functional connectivity in estimating the presence of synapses from spiking activity. Two of these measures were based on Pearson correlation, which is commonly used to estimate the connectivity between pairs of neurons [128, 133, 137]. The first method was normalized cross-correlation ($XCorr$):

$$XCorr_{ij}(\delta) = \frac{E[i_{t-\delta} j_t]}{\sigma_i \sigma_j} \tag{2.2}$$

Where i_t and $j_{t'}$ are the binary values of the spike trains from neurons i and j at times t and t', and the expected value was computed across time. σ_i and σ_j are standard deviations of the spike trains of neurons i and j, respectively.

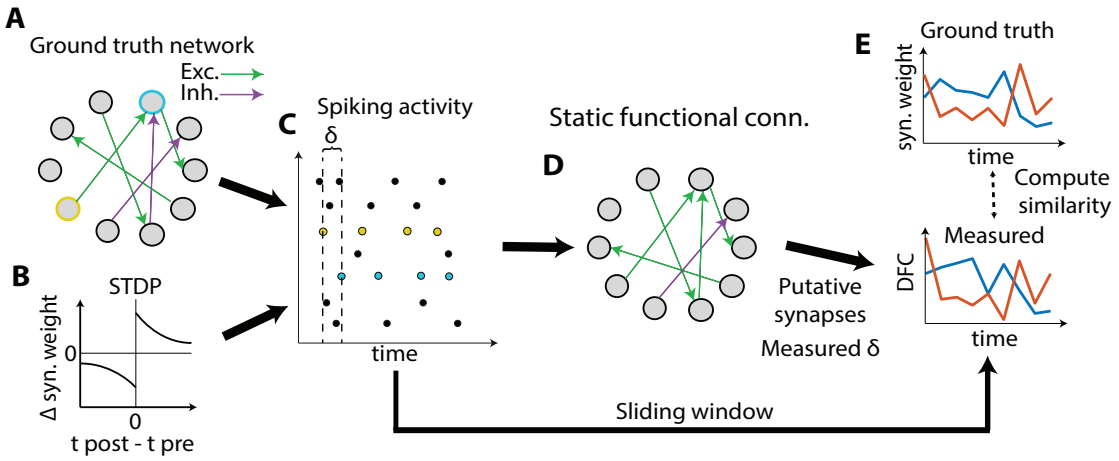

Figure 2.1: Graphical depiction of the inference pipeline. A) Structural connectivity of the simulated network for N=10 neurons. Synaptic weights could be either excitatory (green) or inhibitory (purple). Excitatory connections had randomly distributed communication delays. B) The strength of the synaptic weights changed over time due to STDP. C) Structural and biophysical properties of the network determined the spiking activity of the neural population. D) Static functional connectivity was measured from spiking activity. E) Dynamic functional connectivity was measured from activity, also leveraging on the inferred static connectivity of the network.

The second method was the normalized cross-covariance ($XCov$), which is insensitive to correlations in the average firing rate due to subtraction of the average activity value from the spike trains before computing the correlation:

$$XCov_{ij}(\delta) = \frac{E[(i_{t-\delta} - \bar{i})(j_t - \bar{j})]}{\sigma_i \sigma_j} \tag{2.3}$$

Here $\bar{i}$ and $\bar{j}$ are the average firing rates of neurons i and j, respectively.

Additionally, we computed the functional connectivity using two variants of the information-theoretic measure of information transfer known as transfer entropy [138, 139], a measure that has been successfully used to characterize time-dependent changes in recurrent connectivity between mass signals [31]. Transfer entropy has the theoretical advantage - with respect to correlation measures - of being assumption-free in terms of the joint probability distribution of the lagged activity of neuron i and j. This also means that transfer entropy does not assume that the interactions between neurons are linear. Additionally, this measure respects the Wiener-Granger causality principle of causal communication by conditioning the information between the past of the emitter and the present of the receiver neuron on the past activity of the receiver neuron. Our first implementation of transfer entropy uses single time-points statistics to build the probability distribution of lagged neural activity. We refer to this implementation as TE:

$$TE_{ij}(\delta) = I(i_{t-\delta}; j_t | j_{t-1}) = \sum p(i_{t-\delta}, j_t, j_{t-1}) \log_2 \frac{p(j_t | i_{t-\delta}, j_{t-1})}{p(j_t | j_{t-1})} \tag{2.4}$$

Where $p(i_{t-\delta}, j_t, j_{t-1})$ is the joint probability distribution of the present state of the receiver neuron j_t, its past lagged by one time step j_{t-1} and the past state of the emitter neuron lagged by δ time steps $i_{t-\delta}$. The sum occurs over all the $(i_{t-\delta}, j_t, j_{t-1})$ triplets of events in the probability space. The probability distribution is sampled across time. The lag of the receiver past is set to -1 since it has been proven to be theoretically optimal for determining real communication delays [140].

The second implementation of transfer entropy uses multidimensional pasts of the emitter and the receiver to consider the possible relevance of time windows longer than $1ms$ when transmitting information. According to [133] we refer to this measure as Higher Order Transfer Entropy ($HOTE$):

$$HOTE_{ij}(\delta) = I(i^{(k)}_{t-\delta}; j_t | j^{(l)}_{t-1}) = \sum p(i^{(k)}_{t-\delta}, j_t, j^{(l)}_{t-1}) \log_2 \frac{p(j_t | i^{(k)}_{t-\delta}, j^{(l)}_{t-1})}{p(j_t | j^{(l)}_{t-1})} \tag{2.5}$$

Where k and l are the dimensions of the past activity of the emitter and the receiver neuron i and j, respectively. For the analysis reported in this paper we set $k = l = 5ms$.

We computed these four functional connectivity measures between all pairs of neurons in the network and estimated the communication strength and delay for each pair as described in the previous section. We then evaluated the performance of the different metrics in determining the presence or absence of synapses between pairs of neurons, varying the threshold probability of connectivity strength incrementally from 0 to 1 in steps of 0.01. Since the two classes of present and absent synapses were unbalanced (only 10% of all the possible synapses were present in the network) we used precision-recall (PR) curves to study the performance in this classification task [141] (Fig. 2.2A). Calling TP, FP and FN the number of true positive, false positive and false negative inferred synapses, respectively, we have that $precision = \frac{TP}{TP+FP}$ and $recall = \frac{TP}{TP+FN}$. Therefore, if for a given measure the two distributions of present and absent links were perfectly separable, we would get that for $recall = 1$ also $precision = 1$. On the other hand, a random classifier would always have a precision equal to the ratio of synapses present in the model (10%, dashed line in Fig. 2.2A) for each recall value.

After 90 minutes of simulation, XCov, TE and HOTE all performed well in the classification task, having a PR curve whose shape approached the optimal one. Among these three measures, XCov showed the best PR curve and TE the worst one. XCorr, on the other hand, performed poorly, with a PR curve far from optimal. The area under the precision-recall curve (AUPR) is a useful metric to summarize the goodness of a PR curve; a perfect classifier has an AUPR equal to one. We computed how AUPR scales with simulation length for different measures. This analysis confirmed that XCov and HOTE were the best metrics in evaluating which

links were present for long recordings, while HOTE worked better than XCov and TE for recording shorter than 10 minutes (Fig. 2.2B). We measured how the precision of the different measures scaled with the simulation time for the top 10th and top 5th percentile of inferred synapses. For the top 10th percentile (i.e. 1000 inferred synapses, which equals the ground truth number of connections) we found that the maximum precision in the classification was obtained with XCov, which topped at 92% for 90 minutes of simulated recording (Fig. 2.2C top). With a more conservative threshold of the top 5th percentile of connections (i.e. half of the true total number), we captured the top 500 real connections after 30 minutes of simulation (Fig. 2.2C bottom) for all measures but XCorr. To investigate why XCorr performance was so poor when compared to the other measures, we computed the fraction of links inferred by each measure in the four subgroups of excitatory-to-excitatory (E-E), excitatory-to-inhibitory (E-I), inhibitory-to-excitatory (I-E) and inhibitory-to-inhibitory (I-I) synapses (Fig. 2.2D). XCov performed best in determining the correct fraction of synapses belonging to each group, while XCorr overestimated the number of I-I connections and underestimated the number of E-E connections. This behavior of XCorr is observed due to the differences in average firing rate between inhibitory and excitatory neurons, with a higher firing rate for inhibitory neurons, as XCorr is sensitive to the correlation between average firing rates. Given the poor performance of XCorr in estimating the presence of synapses, we discarded it in the following analyses.

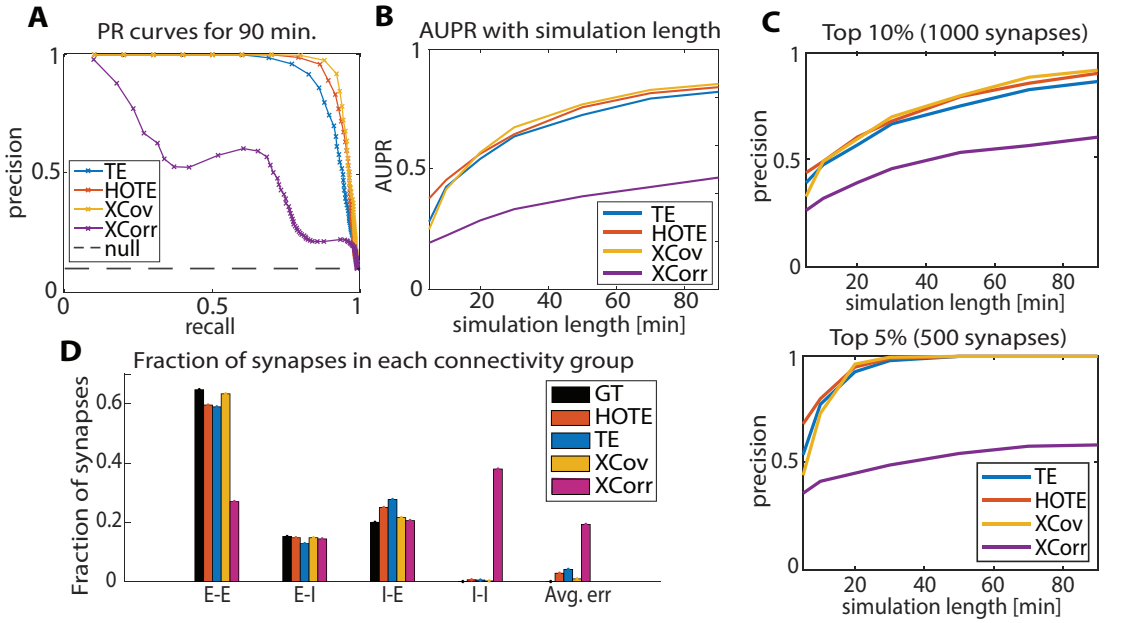

Figure 2.2: Performance of functional connectivity measures in estimating structural connectivity. A) Precision-recall (PR) curves computed from 90 minutes of simulated activity for TE, HOTE, XCov and XCorr. Each point is one percentile of the distribution of functional connectivity values across pairs. B) AUPR trend with simulation length (length ranges from 5 to 90 minutes). C) Comparison of precision in identifying connected pairs with simulation lengths, for top 10th (1000 pairs) and top 5th (500 pairs) percentiles of each measure's distribution. D) Fraction of pairs belonging to each group of synapses, from 90 minute simulation and using the top 10th percentile of connections. GT = ground truth.

2.4 Inferring synapse type and communication delay

We studied how, for each ground truth synapse, different functional connectivity measures performed in inferring whether the synapse was excitatory or inhibitory, and the value of the communication delay of that pair of neurons.

We could not use information-theoretic measures to infer whether synapses were excitatory or inhibitory as these measures are only positively defined. Therefore, we only examined the performance on this excitatory/inhibitory classification for XCov. We classified a connection as excitatory and inhibitory based on XCov value, with positive correlation values assigned as excitatory connections and negative correlation values as inhibitory connections. After 90 minutes of recording XCov could reliably separate excitatory and inhibitory synapses (Fig. 2.3A). We found that the performance of the classifier increased with recording time for both the excitatory and the inhibitory class (Fig. 2.3B).

We also compared how functional connectivity measures performed in inferring

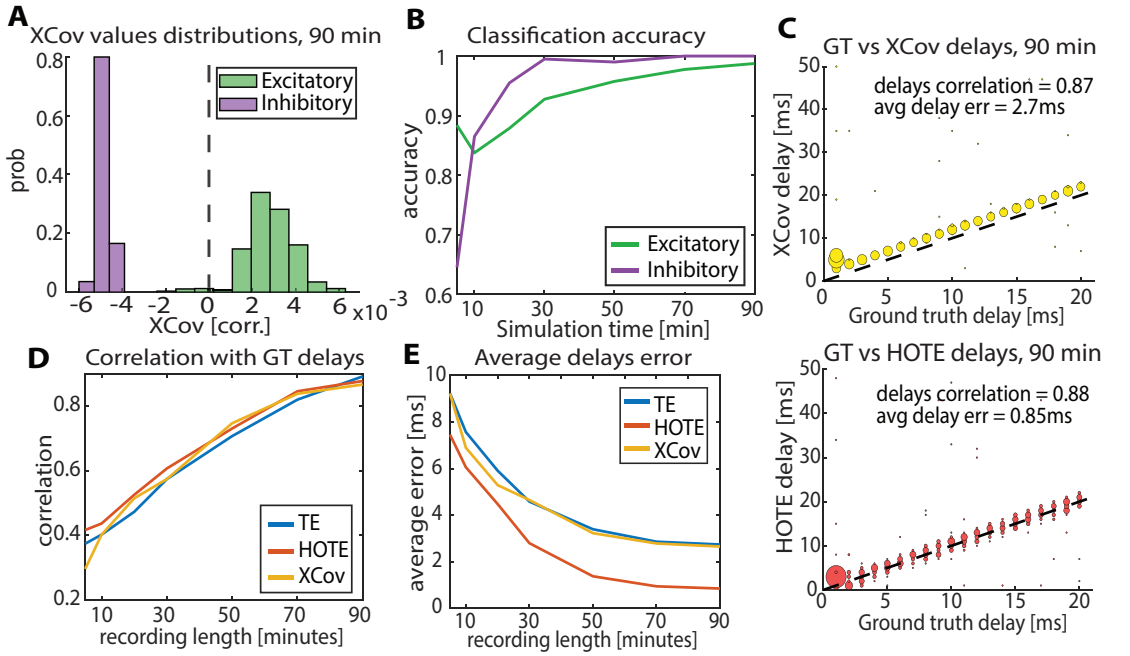

Figure 2.3: Performance of the measures in estimating connection type and delays. A) Distributions of functional connectivity values measured using XCov for excitatory (green) and inhibitory (purple) cells. B) Performance of a classifier in identifying excitatory and inhibitory synapses, the decision boundary of the classifier was set to $XCov = 0$. C) Scatter plots of real and estimated delays across cell pairs using XCov (top) and HOTE (bottom). The size of the markers is proportional to the number of pairs having that specific combination of ground truth and estimated delay. The dashed line is the identity line $x = y$. Black dots far from the identity line correspond to pairs of measured and real delays that occurred only once. D) Correlation between ground truth and estimated delays with simulation length. E) Average error in delay estimation with simulation length.

ground truth communication delays. After 90 minutes of simulation, all measures estimated delays with a correlation across synapses that was above 0.85 (see Fig. 2.3C for the relationship between the ground truth delays and those estimated using XCov - on the top - and using HOTE - on the bottom). The trend of the correlation between ground truth and estimated delays with simulation lengths was approximately linear in the explored range (Fig. 2.3D). Nonetheless, HOTE estimated the delays more precisely than XCov and TE. After 90 minutes of simulation HOTE had an average delay error below $1ms$, while XCov and TE showed a systematic error in the delay estimation of approximately $2ms$ (see Fig. 2.3E and Fig. 2.3C).

2.5 Relationship between dynamic functional connectivity and the temporal evolution of synaptic weights

Lastly, we investigated how the ground truth evolution of the synaptic weights, that is the DSC, related to the measured DFC. We computed DFC using a sliding window approach. We first selected a size for the sliding window T and then shifted it through the simulated recording in steps of length T. We computed DFC only for pairs of neurons that were putatively connected, which we selected as the top 5th percentile of links for each measure after 90 minutes of simulation (Fig. 2.1C), and only at the communication delay that we measured for each pair (Fig. 2.3C). Moreover, we computed DFC only for excitatory synapses since the inhibitory ones had a constant synaptic weight in the simulated network. We calculated the across-time correlation between DFC and DSC for all synapses to quantify the performance of each functional connectivity measure in estimating the DSC. To do this, we averaged the DSC over windows of width T, so that the number of DSC and DFC samples over time were matched.

In Fig. 2.4A we show the DSC (top left), the DFC computed using TE (top right), HOTE (bottom left) and XCov (bottom right) for three example synapses and $T = 10min$. It is visible that, while all measures work reasonably well in tracking how the strength of the gray and the green synapses change over time, TE and HOTE fail in quantifying the temporal evolution of the brown synapse. We found that, on average, DFC computed via XCov correlates with DSC better than the DFC computed via TE or HOTE (Fig. 2.4B). In particular, while DFC computed via TE and HOTE had a high temporal correlation with DSC (above 0.7) for the majority of synapses, their distributions showed a large tail of synapses whose correlation between DSC and DFC was distributed around zero (such as the brown one in Fig. 2.4A). For XCov, the number of synapses whose DSC was poorly estimated decreased rapidly with the correlation strength, and the average correlation was 0.82 (Fig. 2.4B, right). Therefore, the DFC computed using XCov outperformed the one obtained from TE and HOTE in inferring the simulated changes of the synaptic weights over time.

Finally, we studied how the across-time correlation between DSC and DFC depends on the width of the sliding window T. The correlation between DFC and DSC increased with the window size, reaching a plateau around $T = 5min$ (Fig. 2.4C, left). Below $T = 5min$ the correlation dropped due to the limited sample size used to compute DFC, manifesting a tradeoff between the temporal precision of the DFC measures (T) and their performance in estimating DSC. We repeated the same analysis without keeping the delay consistent when computing DFC but simply taking the maximum connectivity value across delays (between 1 and 50ms) for each window (Fig. 2.4C, middle). When not keeping the delay consistent with the previously measured one, the correlation between DSC and DFC dropped sub-

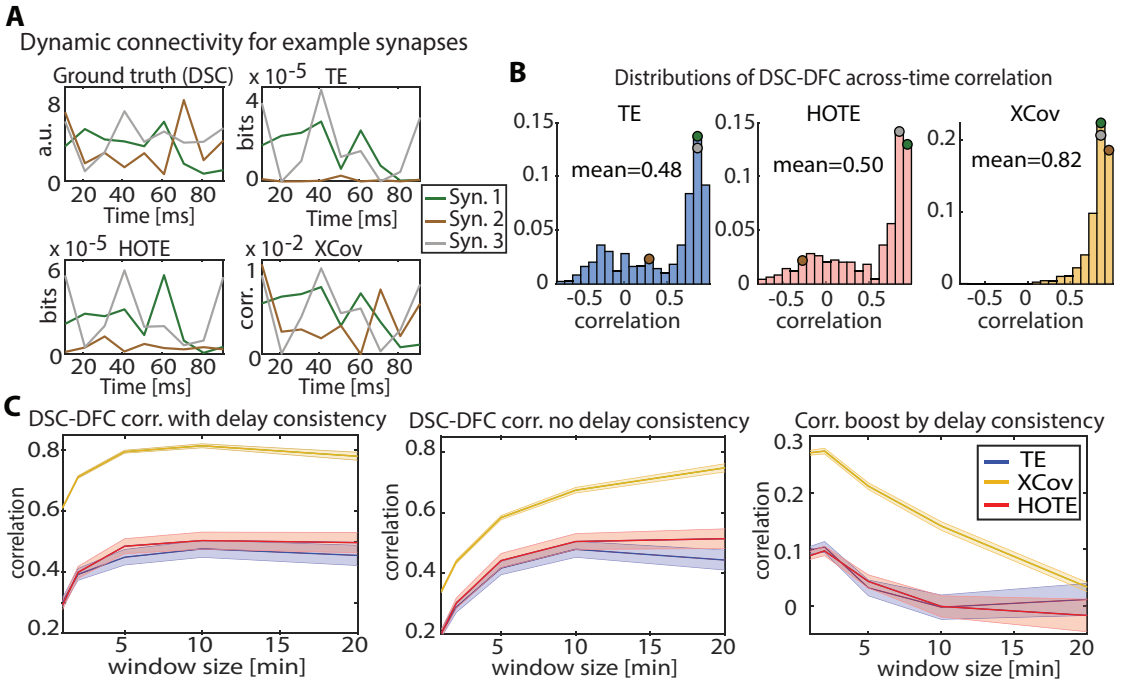

Figure 2.4: Relationship between dynamic structural and functional connectivity. A) Dynamic connectivity for 3 example synapses, $T = 10$ minutes. Top left: ground truth dynamics of synaptic weights (DSC). Top right: Transfer entropy DFC. Bottom left: HOTE DFC. Bottom right: Cross-covariance DFC. B) Distribution of the across-time correlation coefficients between DSC and DFC, $T = 10$ minutes. Left: Transfer entropy. Middle: HOTE. Right: Cross-covariance. Colored dots show where the synapses in panel A are in the correlation distributions. C) Average correlation between DSC and DFC over time for different sizes of the moving window. Shaded areas are SEM across synapses. Left: DFC keeping delay consistency (i.e. measures computed only at previously estimated delay); Middle: DFC without delay consistency; Right: Boost in correlation between DFC and DSC when keeping delay consistency (difference between left and middle panels).

stantially. For sizes of the sliding window lower than $T = 5$ minutes, the advantage of keeping a consistent delay was particularly evident, with a boost in the correlation between DSC and DFC larger than 0.2 (Fig. 2.4C, right). This showed a clear benefit in leveraging estimates of delay derived from entire simulated recordings when inferring DSC from DFC.

2.6 Discussion

We studied how different measures of static and dynamic functional connectivity measured from simulated spiking activity of a recurrent neural network can be used to infer the fixed and time-varying properties of synapses within the network. This question is relevant as in vivo experiments typically rely on recording spiking activity

or other functional measures (such as field potentials) to examine network structure using FC. To infer how changes in FC relate to changes in the underlying synaptic structure of the network requires an understanding of the relationship between the static and dynamic FC measures and the fixed and dynamic synaptic properties of the network. We addressed the problem of inferring synaptic weights and their temporal evolution at the level of simulated recordings with single-neuron cellular resolution. As such, our approach differs from and complements other studies of DFC at the level of mass neural activity [37, 142], which lack the ability to resolve interactions between pairs of individual neurons.

We found that among the considered static FC measures, XCov and HOTE outperformed other measures in inferring the presence of synapses. Using cross-covariance as a static FC measure could also reliably classify excitatory and inhibitory synapses, while HOTE was the best measure to estimate ground-truth communication delay between neurons. Cross-covariance performed best in inferring DSC, with an across-time correlation above 0.8 between DFC and DSC for sliding window sizes larger than 10 min.

We also found that, when computing DFC, keeping the communication delay consistent with the one obtained from the static network analysis increased the correspondence between DFC and DSC, especially for sliding windows shorter than 10 min. This benefit is likely to arise from the fact that, in situations like those simulated here in which the communication delay is a fixed structural property of the neuron pair over the considered time scales, estimating the delay from long time windows increases the precision of its detection without missing out on capturing possible changes of this parameter. This specifically holds under the assumption that communication delays are constant in the recording period as is the case of our spiking network.

Reliable methods to infer structural properties of neural networks are relevant to several open questions in system neuroscience, ranging from investigating the relationship between structural connectivity and computational properties of neural populations to understanding the physiological mechanisms that control the up- and down-scaling of FC, e.g., how the dynamics of synaptic weights relate to changes in functional connectivity during sleep. Another relevant potential application of such methods concerns the inference of STDP rules from recordings of spiking activity. Many studies support the idea that several STDP rules might coexist in different cells or brain areas [143, 144]. Nonetheless, such theories are complicated to test in vivo due to lack of statistical methodologies to estimate how synaptic weights evolve after STPD-triggering events. The methods presented in this work could potentially be used to infer STDP rules governing network plasticity from in vivo recordings, by estimating how synaptic weights change after the occur- rence of pre- and post-synaptic spikes with precise temporal relationships. Moreover, previous works used recurrent neural networks with short term synaptic plasticity to investigate the role of plasticity in working memory [145, 146], showing that short term plasticity facilitates robust memory maintenance. By leveraging the methodologies developed

in the present study to infer the ongoing evolution of synaptic weights from real neural populations recorded during tasks involving working memory, it would be possible to provide further empirical validation of these theoretical models.

The present study has limitations that we plan to address in future works. First of all, it will be important to validate DFC measures on more biologically realistic simulated neural networks with global oscillations, correlated inputs to neurons or global network covariations (which induce FC not related to direct synaptic connections between the neurons [55, 147]), and more heterogeneity in the firing rates and in the average synaptic weights over time. Such effects could act as confounders of the relationship between DFC and DSC or could require refined null hypotheses based on permutation tests to assess the presence of synapses. Furthermore, DFC does not depend solely on temporal changes in structural connectivity. Factors influencing the dynamics of FC, potentially on different timescales, range from the endogenous state of the network to changes in environment [148, 149]. In this work, we started investigating the relationship between the dynamics of FC and the evolution of synaptic weight in a simple recurrent neural network where such factors are absent. Future work involving more complex simulations will be required to disentangle the concurrent contributions of changes in synaptic weights and changes in network or environmental states on DFC. In the model we also assumed that communication delays between neurons are fixed and no synapses are formed or eliminated over time. The former assumes that the main parameters determining the conductance velocity of action potentials (e.g., axons diameters and myelination) are approximately constant over time scales of a few hours. Experimental finding suggest that this assumption is reasonable, especially in adult mice where the formation of new myelin occurs in the range of weeks [150]. The latter assumption is more delicate since in mice it has been shown that, especially during sleep, dendritic spines can be formed and eliminated within hours [127]. It will be important to investigate how much we can relax these hypotheses while still exploiting the knowledge obtained from static FC measures. Moreover, we plan to test the performance of other bivariate (e.g., Granger Causality) and multivariate measures for estimating DSC. These measures include using Granger Causality estimates based on Generalized Linear Models [83, 96, 151] and maximum entropy models [54, 152]. Such multivariate measures could be useful to alleviate the effect of confounders such as common inputs. Another open question that is relevant for the application of methods developed in this work to real neural data, is how to best select the threshold to infer synaptic presence from static FC measures, when no ground truth is known. While an empirical approach would be leveraging physiological knowledge about the average connectivity degree in the recorded area, this method would be limited by the high heterogeneity of connectivity properties across real neurons and subpopulations [153].

Lastly, it will be crucial to apply such methods to data collected from real neural populations and validate, in the first place, the performance of inferring fixed structural connectivity properties from static FC (Figs. 2.2,2.3). A first way to validate

the method proposed here is to verify if the static connectivity networks obtained from two long (e.g., ¿90 min) independent recordings of the same population converge to the same inferred synapses and delays. A second possible validation of the static part of our methodology would be to apply the FC measures to a long recording of a population whose fixed structural properties were reconstructed post-mortem using, e.g., electron microscopy [121, 154]. Such methods typically identify the synapses of neurons whose functional activity was recorded with two-photon calcium imaging rather than with electrophysiology. Given the lower signal-to- noise ratio and temporal resolution of calcium imaging recordings [62], it would be important to first extend and then validate in simulations our proposed methodology to simulated two-photon imaging recordings, rather than simulated electrophysiological recordings as done here.

In conclusion, here we laid down foundations for relating dynamic functional connectivity to the temporal evolution of synaptic weights in spiking neural networks. The results obtained here provide a benchmark for further improving methodologies that infer DSC from DFC.

Chapter 3

An information-theoretic quantification of the content of communication between brain regions

The content of this Chapter was accepted for publication at the 37th Advances in Neural Information Processing Systems (NeurIPS) conference, and is currently in press [50].

3.1 Introduction

Cognitive functions, such as perception and action, emerge from the processing and routing of information across brain regions [17, 55, 155–157]. Methods to study within-brain communication [93, 158, 159] are often based on the Wiener-Granger causality principle, which identifies propagation of information between simultaneously recorded brain regions as the ability to predict the current activity of a putative receiving region from the past activity of a putative sending region, discounting the self-prediction from the past activity of the receiving region [91, 92]. While early measures implementing this principle, such as Granger causality [159], capture only linear interactions, successive information theoretic measures (the closely-related Directed Information [160] and Transfer Entropy [138]) are capable of capturing both linear and nonlinear time-lagged interactions between brain regions [31, 84, 161]. While using such measures has advanced our understanding of brain communication [30, 31, 158, 162–166], they are designed to capture only the overall information propagated across regions, and are insensitive to the content of information flow. Assessing the content of information flow, not only its presence, would be invaluable to understand how complex brain functions, involving distributed processing and flow of different types of information, arise.

Here, we leverage recent progress in Partial Information Decomposition (PID; [100, 110]) to develop a new non-negative measure (Feature-specific Information Transfer; FIT) that quantifies the directed flow of information about a specific feature of interest between neural populations (Fig. 3.1A). The PID decomposes the total information that a set of source variables encodes about a specific target variable into components representing shared (redundant) encoding between the variables, unique encoding by some of the variables, or synergistic encoding in the combination of different variables. FIT isolates features-specific information flowing from one region to another by identifying the part of the feature information encoded in the current activity of the receiving region that is shared (redundant) with information present in the past activity of the sending region (because a piece of transmitted information is first found in the sender and then in the receiver) and that is new and unique with respect to the information encoded in the past activity of the receiver (because information already encoded would not have come from the sender).

We first mathematically derive a definition of FIT based on PID. We then use it to demonstrate, on simulated data, that it is specifically sensitive to the flow of information about specific features, correctly discarding feature-unrelated transmission. We then demonstrate that FIT is able to track the feature-specific content and direction of information flow using three different types of simultaneous multi-region brain recordings (electroencephalography - EEG, magnetoencephalography - MEG, and spiking activity). We also address how introducing appropriate null hypotheses and defining conditioned versions of FIT can deal with potential confounding factors, such as the time-lagged encoding of information in two regions without actual communication between them.

3.2 Defining and Computing Feature-specific Information Transfer (FIT)

We consider two time-series of neural activity X and Y simultaneously recorded from two brain regions over several experimental trials. X and Y might carry information about a feature S varying from trial to trial, e.g. a feature of a sensory stimulus or a certain action. The activity measured in each region, X and Y, may be any type of brain signal, e.g. the spiking activity of single or multiple neurons, or the aggregate activity of neural populations, such as EEG or MEG. We call Y_{pres} the activity of Y at the present time point t , and X_{past} and Y_{past} the past activity of X and Y respectively (Fig. 3.1). Established information theoretic measures such as TE [138] use the Wiener-Granger principle to quantify the overall information propagated from a putative sender X to a putative receiver Y as the mutual information I between the receiver's present neural activity Y_{pres} and the sender's past activity X_{past}, conditioned on the receiver's past activity Y_{past} (Fig. 3.1):

$$TE(X \to Y) = I(X_{past}; Y_{pres} | Y_{past}) \tag{3.1}$$

(see Section 3.8.1.1 for how TE depends on probabilities of past and present activity). TE captures the overall information propagated activity across regions but lacks the ability to isolate information flow about specific external variables. To overcome this limitation, here we define FIT, which quantifies the flow of information specifically about a feature S from a putative sending area X to a putative receiving area Y (Fig. 3.1A). We define FIT using the PID [100]. PID decomposes the joint mutual information $I(S;\underline{X})$, that a set of N source variables $\underline{X} = (X_1, X_2, ..., X_N)$ carries about a target variable S, into non-negative components called information atoms (see SM1.2). For $N = 2$, PID breaks down the joint mutual information $I(S; X_1, X_2)$ into four atoms: the Shared (or redundant) Information $SI(S : X_1, X_2)$ that both X_1 and X_2 encode about S; the two pieces of Unique Information about S, $UI(S : X_1 \setminus X_2)$ and $UI(S : X_2 \setminus X_1)$, provided by one source variable but not by the other ; and the Complementary (synergistic) information about S ,$CI(S : X_1, X_2)$, encoded in the combination of X_1 and X_2 . Several measures have been proposed to quantify information atoms [100, 103, 111, 112]. Here we use the measure I_{min} originally defined in [100], which guarantees non-negative values for information atoms for any N (see Section 3.8.1.2).

Using I_{min}, the Shared Information that X_1 and X_2 carry about S is defined as follows:

$$SI(S : X_1, X_2) = \sum_{s \in S} p(s) \min_{X_i \in \{X_1, X_2\}} I(S = s; X_i) \tag{3.2}$$

where $I(S = s; X_i)$ is the specific information that source X_i carries about a specific outcome of the target variable $s \in S$, and is defined as:

$$I(S = s; X_i) = \sum_{x_i \in X_i} p(x_i|s) \Big[\log \frac{p(s|x_i)}{p(s)}\Big] \tag{3.3}$$

Intuitively, the shared information computed as in eq. 3.2 quantifies redundancy as the similarity between X_1 and X_2 in discriminating individual values of the feature S. In the general case of N source variables, information atoms are hierarchically ordered in a lattice structure, and I_{min} can be used to quantify any atom in the decomposition (including the Unique and Complementary information atoms introduced above for the case $N = 2$; see Section 3.8.1.2).

We wanted FIT to measure the directed flow of information about S between X and Y, rather than the overall propagation of information measured by TE (Fig. 3.1A). We thus isolated the information about a feature S in the past of the sender X that Y receives at time t. Because of the Wiener-Granger causality principle, such information should not have been present in the past activity of the receiver Y. Therefore, we performed the PID in the space of four variables S, X_{past}, Y_{past}, and Y_{pres} to compute information atoms that combine Shared, Unique and Complementary Information carried by three sources about one target [100]. One natural candidate atom to measure FIT is the information about S that X_{past} shares with Y_{pres} and is unique with respect to Y_{past}: $SUI(S : X_{past}, Y_{pres} \setminus Y_{past})$ (Fig. 3.1B; Fig. 3.S1B). This atom is defined as the difference between the shared information

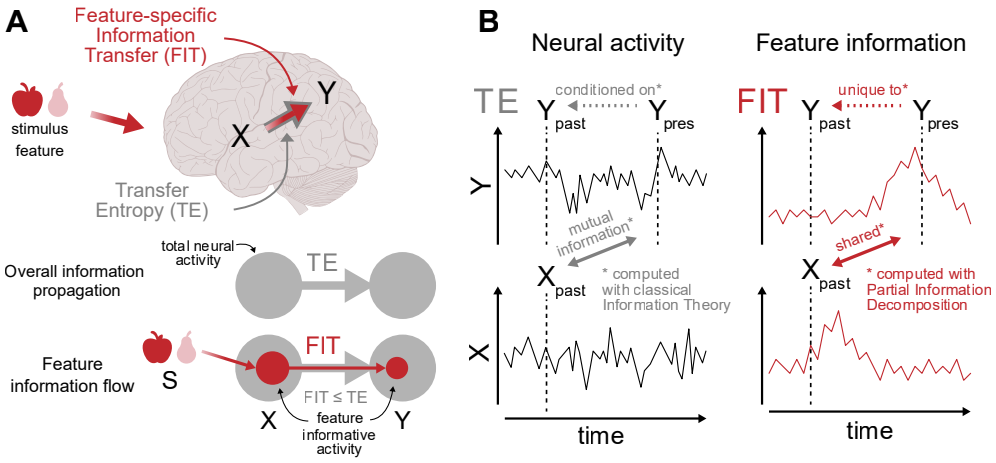

Figure 3.1: Sketch of FIT and TE. (A) TE is the established information-theoretic measure to quantify the overall information propagated between two simultaneously recorded brain regions X (sender) and Y (receiver). FIT measures the information flowing from X to Y about the stimulus feature S. B) TE and FIT incorporate content-unspecific and content-specific versions of the Wiener-Granger causality principle. TE is the mutual information between the past activity of X and the present activity of Y conditioned on the past of Y. FIT is the feature information in the present of Y shared with the past information of X and unique with respect to the past information of Y.

that the two source variables X_{past} and Y_{pres} carry about S, and the shared information that the three source variables X_{past}, Y_{pres} and Y_{past} carry about S. Redundancy can only decrease when adding more sources. Hence by removing the information that is also redundant with Y_{past}, $SUI(S : X_{past}, Y_{pres} \setminus Y_{past})$ quantifies a non-negative component of shared information between X_{past} and Y_{pres} about S that is unique with respect to Y_{past}. Importantly, using unique information to remove the feature information in Y_{past} is more conservative than conditioning on Y_{past} as in TE (Fig. 3.1B) [167] . $SUI(S : X_{past}, Y_{pres} \setminus Y_{past})$ intuitively captures what we are interested in, and satisfies two desirable mathematical properties: it is upper bounded by the feature information encoded in the past of X ($I(S; X_{past})$) and in the present of Y ($I(S; Y_{pres})$). This is because the PID defines redundancy between source variables as sub-components of the joint information carried by each of the sources (see Section 3.8.1.3). However, the information atom $SUI(S : X_{past}, Y_{pres} \setminus Y_{past})$ has two undesirable properties. The first is that its value can exceed the total amount of information propagated from X to Y (TE). This can happen since the unique information in the PID decomposition is a component of the conditional mutual information about the target. However, the target in $SUI(S : X_{past}, Y_{pres} \setminus Y_{past})$ is the feature S, which means that this atom is not constrained to be smaller than the TE, which is independent of S (see eq. 3.1 and SM1.3.4). This property is undesirable because the overall propagation of activity (Fig. 3.1A, bottom) must be an upper bound to the information transmitted about a specific feature. The second is that by construction (see Section 3.8.1.3) this atom depends on X_{past}, Y_{pres}, S only

through the pairwise marginal distributions $P(X_{past}, S)$ and $P(Y_{pres}, S)$, but not through the marginal distribution $P(X_{past}, Y_{pres})$, which implies that this atom by itself cannot identify confounding scenarios where both sender and receiver encode feature information at different times with no transmission taking place (see Section 3.8.1.3).

To address these limitations, following [113] we considered the alternative PID taking S, Y_{past}, and X_{past} as source variables and Y_{pres} as a target. In this second PID (Fig. 3.S1B), the atom that intuitively relates to FIT is $SUI(Y_{pres} : X_{past}, S \setminus Y_{past})$, the information about Y_{pres} that X_{past} shares with S that is unique with respect to Y_{past}. While being intuitively similar to $SUI(S : X_{past}, Y_{pres} \setminus Y_{past})$, $SUI(Y_{pres} : X_{past}, S \setminus Y_{past})$ has Y_{pres} as target variable and hence is upper bounded by TE (but not by $I(S; X_{past})$) and depends on the pairwise marginal distribution $P(X_{past}, Y_{pres})$ (see Section 3.8.1.3). Thus, this second atom has useful properties that complement those of the first atom. Importantly, Shannon information quantities impose constraints that relate PID atoms across decompositions with different targets. We [113] demonstrated that, for PID with $N = 2$ sources, these constraints reveal the existence of finer information components shared between similar atoms of different decompositions. Here, we extended this approach (see Section 3.8.1.3) to $N = 3$ sources and demonstrated that the second atom is the only one in the second PID that has a pairwise algebraic relationship with the first atom, indicating that these atoms share a common, finer information component. Therefore we defined FIT by selecting this finer common component by taking the minimum between these two atoms:

$$FIT = \min[SUI(S : X_{past}, Y_{pres} \setminus Y_{past}), SUI(Y_{pres} : X_{past}, S \setminus Y_{past})] \qquad (3.4)$$

With this definition, FIT is upper bounded by $I(S; X_{past})$, by $I(S; Y_{pres})$ and by $TE(X \to Y)$. That FIT satisfies such bounds is essential to interpret it as transmitted information. If FIT could be larger than the feature information encoded by sender X or receiver Y, or than the total information transmitted ($TE(X \to Y)$), then FIT could not be interpreted as feature information transmitted from X to Y. Additionally, FIT depends on the joint distribution $P(S, X_{past}, Y_{pres})$ through all the pairwise marginals $P(S, X_{past})$, $P(S, Y_{pres})$, and $P(X_{past}, Y_{pres})$, implying that it can rule out, using appropriate permutation tests, false-communication scenarios in which X and Y encode the stimulus independently with a temporal lag, without any within-trial transmission (see Section 3.8.1.3).

Note that the definition of FIT holds when defining present and/or past activity as multidimensional variables, potentially spanning several time points. However, use of multidimensional neural responses requires significantly more data for accurate computation of information. For this reason, following [31, 168, 169], in all computations of TE and FIT we computed the present of Y at a single time point t and the past of X and of Y at individual time points lagged by a delay δ: $X_{past} = X_{t-\delta}$ and $Y_{past} = Y_{t-\delta}$. Note also that in all calculations of FIT and TE, we estimated probabilities from empirical occurrences after discretizing both features

and neural activities. SM1.5 reports details of the procedure and Table S1 the number of bins used for each analysis. Simulations of accuracy of these estimates as function of the data size are reported in Section 3.8.2.5 and Fig. 3.S6.

3.3 Validation of FIT on simulated data

To test the ability of FIT to measure feature-specific information flow between brain regions, we performed simulations in scenarios of feature-specific and feature-unrelated information transfer.

We performed (Fig. 3.2A-B) a simulation (details in SM2.1) in which the encoded and transmitted stimulus feature S was a stimulus-intensity integer value (1 to 4) . The activity of the sender X was a two-dimensional variable with one stimulus-feature-informative X_{stim} and one stimulus-uninformative component X_{noise}. The stimulus-feature-informative dimension had temporally-localized stimulus-dependent activity from 200 to 250ms and had multiplicative Gaussian noise (similar results were found with additive noise, see Section 3.8.2.1 and Fig. 3.S3). The stimulus-unrelated component was, at any time point, a zero-mean Gaussian noise. The activity of the receiver Y was the weighted sum of X_{stim} and X_{noise} with a delay δ, plus Gaussian noise . The delay δ was chosen randomly in each simulation repetition in the range 40-60ms. Here and in all further simulations, we averaged information values across simulation repetitions we determined their significance via non-parametric permutation tests. For TE, we permuted X across all trials to test for the presence of significant within-trial transmission between X and Y [31, 170]. For FIT, we conducted two different permutation tests: one for the presence of stimulus information in X and Y (shuffling S across trials), and another for the contribution of within-trial correlations between X and Y to the transmission of S (shuffling X across trials at fixed stimulus). We set the threshold for FIT significance as the 99th percentile of the element-wise maximum between the two permuted distributions (see Section 3.8.1.7).

We investigated how FIT and TE from X to Y depended on the amount of stimulus-feature-related transmission (increased by increasing w_{stim}) and of -unrelated transmission (increased by increasing w_{noise}). We report values at the first time point in which information in Y was received from X, but similar results hold for later time points. Both FIT and TE increased when increasing w_{stim} (Fig. 3.2A). However, TE increased with w_{noise} (Fig. 3.2A), as expected from a measure that captures the overall information propagation. In contrast, FIT decreased when increasing w_{noise}, indicating that FIT specifically captures the flow of information about the considered feature.

We then investigated how well TE and FIT temporally localize the stimulus-feature-related information transmitted from X to Y (Fig. 3.2B). We simulated a case in which stimulus-feature-related information was transmitted from X to Y only in a specific window ($[240, 310]ms$) and computed FIT and TE at each time point (see Section 3.8.2.1 for details) . FIT was significant only in the time window in

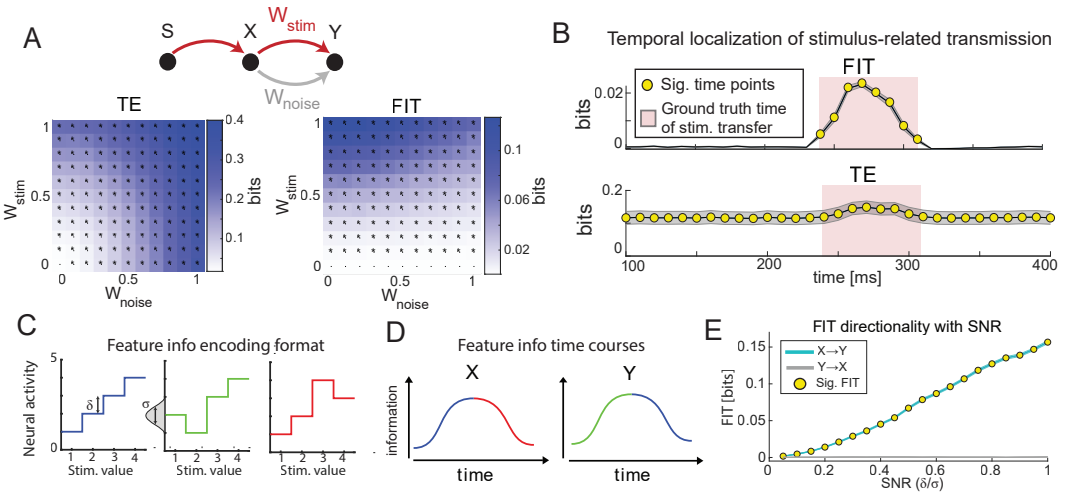

Figure 3.2: Testing FIT on simulated data. A) FIT and TE as function of stimulus-feature-related (w_{stim}) and -unrelated (w_{noise}) transmission strength. * indicate significant values ($p < 0.01$, permutation test) for the considered parameter set. B) Dynamics of FIT and TE in a simulation with time-localized stimulus-feature-information transmission. Red area shows the window of stimulus-feature-related information transfer. Results plot mean (lines) and SEM (shaded area) across 50 simulations (2000 trials each). C) Different neural encoding functions used for the simulations in panels D,E. D) Sketch of simultaneous stimulus feature information profiles. Different types of information content are color-coded. E) FIT in the $X \to Y$ (cyan) and $Y \to X$ (grey) directions as a function of SNR (plot in log scale). Results plot mean (lines) and SEM (shaded area) across 100 simulations (2000 trials each). Yellow dots in panels B and E show points with significant FIT (p ¡ 0.01, permutation test).

which Y received the stimulus information from X. In contrast, TE was significant at any time point, reflecting that noise was transmitted from X to Y throughout the whole simulation time.

Importantly, FIT can detect feature-specific information flow even when information is encoded in the sender and receiver with an overlapping timecourse. To illustrate this, we simulated the activity of two regions X and Y encoding an integer stimulus feature S with the same amount of information at each instant of time (Fig. 3.2D), but with feature specific transmission taking place only from X to Y. Because FIT could correctly detect that the format of information representation of S in the present of Y was equal to that of the past of X but different to that of the past of Y (Fig. 3.2D), it could correctly detect that feature information flows from X to Y (Fig. 3.2E).

We also performed simulations to investigate whether the non-parametric permutation test described above can correctly rule out as non-significant feature-specific transmission the scenario in which X and Y independently encode S without actual communication occurring between them. We simulated a scenario in which feature information was encoded with a temporal lag in X and Y, with no transmission

from X to Y. We found that the resulting values were always non-significant (see Section 3.8.2.6 and Fig. 3.S7C) when tested against a surrogate null-hypothesis distribution (pairing X and Y in randomly permuted trials with the same feature) that destroy the within-trial communication between X and Y without changing the feature information encoding in X and Y (see Section 3.8.1.7). Importantly, this null hypothesis also ruled out false communication scenarios where the measured FIT and TE were only due to the presence of instantaneous mixing of sources (see Section 3.8.2.7).

Finally, we addressed how to remove the confounding effect of transmission of feature information to Y not from X but from a third region Z. In Granger causality or TE analyses, this is addressed conditioning the measures on Z [94, 171]. In an analogous way, we developed a conditioned version of FIT, called cFIT (see see Section 3.8.1.4), which measures the feature information transmitted from X to Y that is unique with respect to the past activity of a third region Z. We tested its performance in simulations in which both X and Z transmitted feature information to Y and found that cFIT reliably estimated the unique contribution of X in transmitting feature information to Y, beyond what was transmitted by Z (see Section 3.8.2.6 and Fig. 3.S7D).

3.4 Analysis of real neural data

We assessed how well FIT detects direction and specificity of information transfer in real neural data.

3.4.1 Flow of stimulus and choice information across the human visual system

We analyzed a previously published dataset ([11], see also SM3.1 for details) of source-level MEG data recorded while human participants performed a visual decision-making task. At the beginning of each trial, a reference stimulus was presented (contrast 50%), followed by a test stimulus that consisted of a sequence of 10 visual samples with variable contrasts (Fig. 3.3A). After the test stimulus sequence, participants reported their choice of whether the average contrast of the samples was greater or smaller than the reference contrast. The previous study on these data ([11]) analysed the encoding of stimulus and choice signals in individual areas but did not study information transfer. We focused on gamma-band activity (defined as the instantaneous power of the 40-75Hz frequency band), because it is the most prominent band for visual information encoding [172–174] and information propagation [31, 59] in the visual system. Previous work has demonstrated that gamma-band transmission is stronger in the feedforward (from lower to higher in the visual cortical hierarchy) than in the feedback (from higher to lower in the visual cortical hierarchy) direction [59, 164, 175], suggesting that gamma is a privileged frequency

channel for transmitting feedforward information. However, these previous studies did not determine the content of the information being transmitted.

To address this question, we quantified FIT in a network of three visual cortical areas (Fig. 3.3B) that we selected because they encoded high amounts of stimulus information and because they were sufficiently far apart ($\geq$ 2.8cm) to minimize leakage in source reconstruction (see Fig. 3.S9) [11, 176]. The areas, listed in order of position, from lower to higher, in the cortical hierarchy were: primary visual cortex (V1), area V3A (carrying maximal stimulus information in the dorsal stream visual cortex), and area LO3 (carrying high stimulus information in the MT+ complex). Because participants made errors (behavioral performance was 75% correct), in each trial the stimulus presented could differ from the participant's choice. We thus assessed the content of the information flow by computing FIT about either the sensory stimulus (FIT_S; using as feature the mean contrast across all 10 visual samples) or the reported choice (FIT_C), in each instant of time in the $[-100, 500]$ms peri-stimulus time window (because stimulus information was higher in the first 500ms post-stimulus, see Section 3.8.3.1 and Fig. 3.S9) and across a range of putative inter-area delays δ. In Fig. 3.3C we show the resulting FIT_S time-delay information maps for the example pair of regions V1 and V3A. A cluster-permutation analysis [177, 178] revealed significant feedforward stimulus-specific information transmission from V1 to V3A (but no significant feedback from V3A to V1) localized 200-400ms after the stimulus onset, with an inter-area communication delay between 65 and 250ms (see Section 3.8.3.1 and Fig. 3.S9).

We compared properties of overall information propagation (computed with TE) and feature-specific information flow (computed with FIT) across all pairs of areas within the considered visual cortical network. To determine the prevalent content of information flow in the network, we compared the amount of FIT_S and FIT_C transmitted in the feedforward and in the feedback directions (Fig. 3.3D). Gamma-band transmitted more information about the stimulus than about choice (i.e. FIT_S ¿. FIT_C) in both the feedforward ($p < 10^{-3}$ two-tailed paired t-test) and in the feedback ($p < 0.01$ two-tailed paired t-test) direction, with a larger difference for the feedforward direction. This result is supported by simulations where we show that, in presence of multiple simultaneously transmitted features, FIT ranks correctly the features about which most information is transmitted (see Section 3.8.2.3 and Fig. 3.S4). Thus, we focused on stimulus-specific information flow in the following FIT analyses. We then studied the leading direction of information flow. Both the total amount of information propagation (TE) and the stimulus-specific information flow (FIT_S) were larger in the feedforward than in the feedback direction (Fig. 3.3E), but with a larger effect for FIT_S ($p < 10^{-6}$ two-tailed paired t-test) compared to TE ($p < 0.05$ two-tailed paired t-test). Together, these results show that gamma-band activity in the visual system carries principally information about the stimulus (rather than choice) and propagates it more feedforward than feedback.

We next assessed the behavioral relevance of the feedforward stimulus information transmitted by the gamma band. A previous study showed that the over-

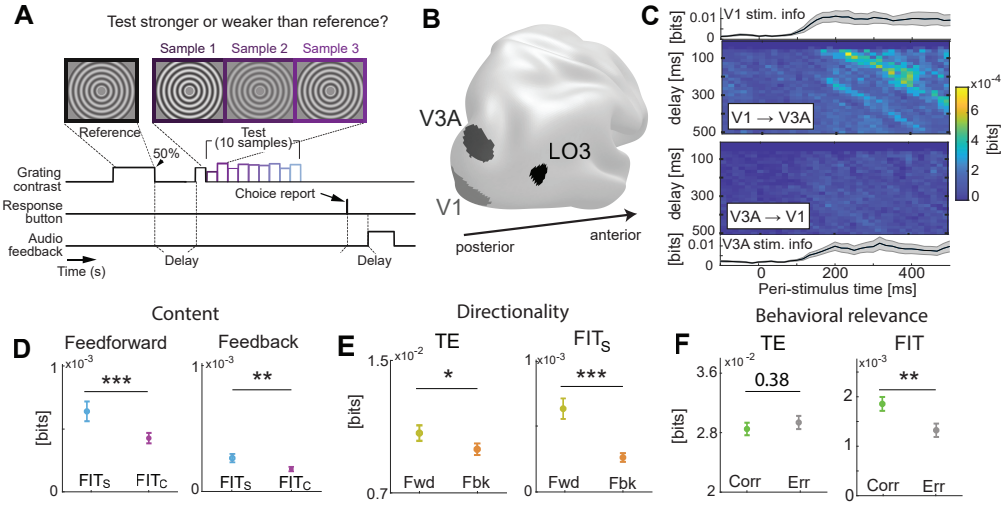

Figure 3.3: Information flow across the human visual hierarchy with MEG. A) Sketch of task B) Cortical surface map of the location of the four considered visual regions. C) Temporal profiles of stimulus information and time-delay stimulus FIT maps for an example regions pair (V1, V3A). Top to bottom: stimulus information in V1; time-delay FIT map in the feedforward (V1 $\rightarrow$ V3A) direction; time-delay FIT map in the feedback (V3A $\rightarrow$ V1) direction; stimulus information in V3A. D) Comparison between FIT about stimulus (FIT_S) and FIT about choice (FIT_C) in the visual network in the feedforward (left) and feedback (right) directions. E) Comparison between feedforward and feedback transmission in the network for TE (left) and stimulus FIT (right). F) Same as F but for feedforward transmission on correct vs error trials. In all panels, lines and image plots show averages and errorbars SEM across participants, experimental sessions and regions pairs (in case of FIT and TE) or regions (in case of mutual information). *: p<0.05, **: p<0.01, ***: p<0.001. Information-theoretic quantities were computed from the gamma band ([40-75]Hz) power of source-level MEG, first computed separately for the left and right hemisphere and then averaged.

all (feature-unspecific) strength of feedforward gamma band information propagation negatively correlates with reaction times, indicating that stronger feedforward gamma actvity propagation favours faster decisions [95]. However, no study has addressed whether stimulus information transmitted forward in the gamma band helps accuracy of decision making. We addressed this question by comparing how FIT_S varied between trials when participants made a correct or incorrect choice (Fig. 3.3F). We matched the number of correct and error trials to avoid data size confounds [179]. FIT_S in the feedforward direction was significantly lower in error than in correct trials (Fig. 3.3F, right; $p < 0.001$ two-tailed paired t-test), while TE did not vary between correct and error trials (Fig. 3.3F, left; $p = 0.24$ two-tailed paired t-test). Feedback information transmission (both in terms of overall transmission, TE, and stimulus specific information flow, FIT_S), did not vary between correct and incorrect trials. This indicates that the feedforward flow of stimulus information, rather than the overall information propagation, is key for forming

correct choices based on sensory evidence.

These results provide the new discovery that the gamma band transmits feedfoward stimulus information of behavioral relevance, and highlight the power of FIT in revealing the content and direction of information flow between brain areas.

3.4.2 Eye-specific interhemispheric information flow during face detection

We next tested the ability of FIT to detect feature-specific information flow between brain hemispheres. We analyzed a published EEG dataset recorded from human participants detecting the presence of either a face or a random texture from an image covered by a bubble mask randomly generated in each trial ([180]; see Section SM3.2.1 for details). Previous analysis of these data [181] showed that eye visibility in the image (defined as the proportion of image pixels in the eye region visible through the mask) is the most relevant image feature for successful face discrimination. It then showed that eye-specific information appears first at ∼120ms post-image presentation in the Occipito-Temporal (OT) region of the hemisphere contralateral with respect to the position of the eye, and then appears ∼20-40ms later in the ipsilateral OT region (Fig. 3.4A). However, this study did not determine if the eye information in the ipsilater hemisphere is received from the contralateral hemisphere. To address this issue, we computed FIT transmission of eye-specific information between the Left OT (LOT) and Right OT (ROT) regions (using the electrodes within these regions that had most information as in [181], see Section 3.8.3.2). Left Eye (LE) FIT from the contra- to the ipsi-lateral OT (ROT to LOT; Fig. 3.4C) peaked between 150 to 190ms after image onset with transfer delays of 20-80ms (Fig. 3.4B). Right eye (RE) FIT the contra- to the ipsi-lateral OT (LOT to ROT) peaked with similar times and delays. Both contra-to-ipsilateral LE and RE had statistically significant FIT peaks in the time-delay maps (cluster-permutation analysis, p ¡ 0.01; see Section 3.8.3.2 and Fig. 3.S10). Thus, FIT determined the communication window for contralateral flow of eye-specific information with high precision.

To gain further insight about the directionality and feature-specificity of the information flow across hemispheres, we compared FIT and TE across transfer directions and/or eye-specific visibility conditions (Fig. 3.4D, middle and right). Right-to-left LE FIT was significantly larger than left-to-right LE FIT ($p < 0.001$ two-tailed paired t-test) or right-to-left RE FIT ($p < 0.01$ two-tailed paired t-test). Left-to-right RE FIT was significantly larger than right-to-left RE FIT ($p < 0.05$ two-tailed paired t-test) or left-to-right LE FIT ($p < 0.05$ two-tailed paired t-test). In contrast, we found no significant difference between directions for the overall propagated information (TE), Fig. 3.4D, left). Thus, the use of FIT revealed a temporally localized flow of eye information across hemispheres that was feature-selective (i.e. about mainly the contralateral eye) and direction-specific (contra-to-ipsilateral), without direction specificity in the overall information propagation (TE) across hemispheres.

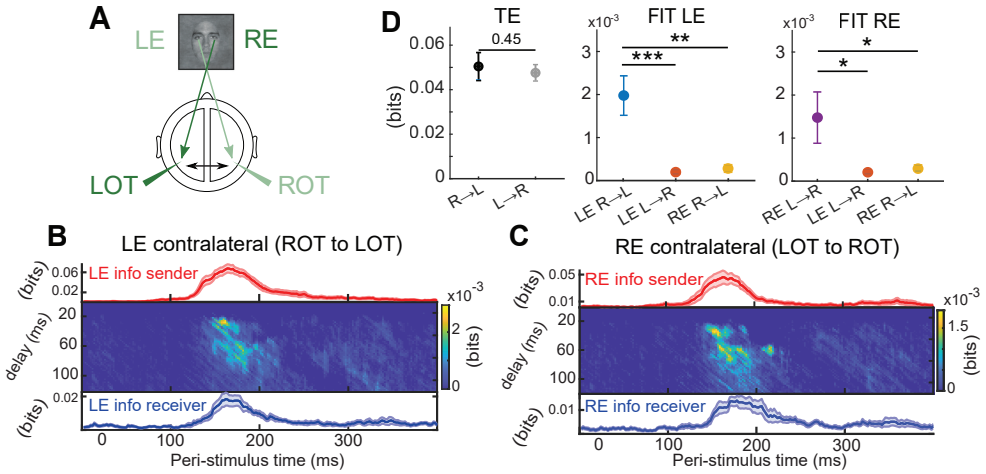

Figure 3.4: Inter-hemispheric eye-specific information flow during face detection using human EEG. (A) Schematic of the putative information flow. LOT (ROT) denote Left (Right) occipito-temporal regions. LE (RE) denote the Left (Right) Eye visibility feature (B) Information (lines) carried by the EEG in each region, and FIT (image plot) about LE across regions. (C) Same as B for RE. (D) Contra- to ipsi-lateral vs ipsi- to contra-lateral transfers for TE and FIT for both LE and RE. Dots and images plot averages and errorbars plot SEM across participants (N=15).

Finally, to more tightly localize the origin of eye-specific contralateral information flow, we asked whether the contralateral OT electrodes selected in our analyses were the sole senders of inter-hemispheric eye-specific information. We used the conditioned version of FIT, cFIT, to compute the amount of transfer of eye information from the contra- to the ipsi-lateral OT after removing the effect of eye-specific information possibly routed through alternative sending locations (see Section 3.8.1.4). We found (Fig. 3.S12A) that the effect we measured with FIT was robust even when conservatively removing with cFIT the information that could have been routed through other locations.

3.4.3 Stimulus-specific information flow in a thalamocortical network

We finally used FIT to measure the feature- and direction-specificity of information flow in the thalamocortical somatosensory and visual pathways. We analysed a published dataset in which multi-unit spiking activity was simultaneously recorded in anaesthetized rats from the primary visual and somatosensory cortices, and from first-order visual and somatosensory thalamic nuclei ([23], see Section 3.8.3.3 for details), during either unimodal visual, unimodal tactile, or bimodal (visual and tactile) stimulation. This analysis tests FIT on another major type of brain recordings (spiking activity). Moreover, due to the wealth of knowledge about the thalamocortical network [22, 182], we can validate FIT against the highly-credible predictions

that information about basic sensory features flows from thalamus to cortex, and that somatosensory and visual pathways primarily transmit tactile and visual information, respectively. Using FIT, we found (see 3.8.3.3 and Fig. 3.S11) that sensory information flowed primarily from thalamus to cortex, rather than from cortex to thalamus. We also found that the feedforward somatosensory pathway transmits more information about tactile- than about visually-discriminative features, and that the feedforward visual pathway transmits more information about visually- than tactile-discriminative features. Importantly, TE was similarly strong in both directions, and when considering tactile- or visually-discriminative features. This confirms the power of FIT for uncovering stimulus-specific information transfer, and indicates a partial dissociation between overall information propagation and neural transfer of specific information.

3.5 Comparison with previously published measures

We finally examine how FIT differs from alternative methods for identifying components of the flow of information about specific features. We focus on measures that implement the Wiener-Granger discounting of the information present in the past activity of the sender. Other methods, that do not implement this (and thus just correlate past information of the sender with present information of the receiver), erroneously identify information already encoded in the past activity of the receiver as information transmitted from a sender (see Section 3.8.4.3).

A possible simple proxy for identifying feature-specific information flow is quantifying how the total amount of transmitted information (TE) is modulated by the feature [31]. For the case of two feature values, this amounts to the difference of TE computed for each individual value. We show in Section 3.8.4.1, using simulations, that this measure can fail in capturing feature-related information flow even in simple scenarios of feature information transmission. Additionally, when tested on MEG data, it could not assess the directionality of information transmission within brain networks (see Section 3.8.4.1).

A previous study [183] defined a measure, Directed Feature Information (DFI), which computes feature-specific information redundant between the present activity of the receiver and the past activity of the sender, conditioned on the past activity of the receiver. However, DFI used a measure of redundancy that conflated the effects of redundancy and synergy (see See 3.8.1.5). Because of this, DFI is, both on real and on simulated data, often negative and thus not interpretable as measure of information flow (see Section 3.8.4.2). In contrast, FIT is non-negative and uses PID to consider only redundant information between sender and receiver, as appropriate to identify transmission of information. Moreover, because DFI discounts only past activity of the sender rather than its feature-specific information, it was less precise and less conservative or sensitive in localizing direction and timing of feature-specific

information flow (as shown in Section 3.8.4.2 and Fig. 3.S15 with simulated and real data).

Finally, a study defined feature-specific information using PID in the space of four variables S, X_{past}, Y_{past}, and Y_{pres} [108]. However, this measure was not upper bounded by either feature information encoded in the past of the sending region or the total information flowing between regions.

3.6 Discussion

We developed and validated FIT, an information theoretic measure of the feature-specific information transfer between a sender X and a receiver Y. FIT combines the PID concepts of redundancy and uniqueness of information [100] with the Wiener-Granger causality principle [93] to isolate, within the overall transmitted information (TE), the flow of information specifically related to a feature S.

The strengths and limitations of FIT as a neural data analysis tool stem from those of information theory for studying neural information processing. Information theory has led to major advances to neural coding because of its ability to capture linear and non-linear interactions at all orders making little assumptions [86, 184]. This is important because deviations from linearity and order of interactions vary in often unknown ways between brain areas, stimulus types and recording modalities [14, 185, 186]. Using such a general formalism avoids potentially biasing results with wrong assumptions. However, the price to pay for the fact that information theory includes full probability distributions is that it is data hungry. While our definitions of FIT and cFIT are straightforwardly valid for multivariate analyses including conditioning on the information of multiple regions [171] (as in cFIT) or obtaining more conservative estimates of information transmission on which information in the receiver Y is requested to be unique with respect to the information of the sender and receiver at multiple past time points [84], for data sampling reasons in practice in real data these analyses are confined to conditioning to one region or a single past time points[31, 168, 169]. In future work, we aim to make FIT applicable to analyses of multiple regions or time points coupling it with advanced non-parametric [187] methods to robustly estimate its multivariate probability distributions.

The generality of our approach lends itself to further developments. Importantly, we defined FIT directly at the level of PID atoms. This means that, although here we implemented FIT using the original definition of redundancy in PID [100] because it has the advantage of being non-negative for all information atoms, FIT can be easily implemented also using other PID redundancy measures [103, 111, 112, 188, 189] with complementary advantages and disadvantages (see Section 3.8.1.2). Additionally, and even though the surrogate permutation test we developed to assess FIT significance provided reasonable results on real data and worked work well also with artefacts due to instantaneous mixing of sources (see See 3.8.2.7), further research is needed to generate more refined surrogate data generation techniques to rule out more conservatively false feature-specific communication scenarios [190].

To demonstrate the properties of FIT, we performed numerical simulations in different communication scenarios and compared FIT against TE (Fig. 3.2). These simulations confirmed that TE effectively detected the overall propagation of information, but it did not detect the flow of feature-specific information. FIT, in contrast, reliably detected feature- and direction-specific information flow with high temporal sensitivity. We confirmed the utility of FIT in applications to neural data. In three brain datasets spanning the range of electrophysiological recordings (spiking activity, MEG and EEG), FIT credibly determined the directionality and feature specificity of information flow. Importantly, in most of these datasets this happened in the absence of variations in the overall flow of activity between the same brain regions (measured with TE). The partial dissociation between overall activity flow and feature-specific flow found consistently in simulations and data has important implications. First, it highlights the need of introducing a specific measure of feature information transfer such as FIT, as it resolves question unaddressed by content-unspecific measures. Second, it establishes that measuring feature-specific components of information flow between brain regions is critical to go beyond the measurement of overall neural activity propagation and uncover aspects of cross-area communication relevant for ongoing behavior. Much effort in neuroscience has been devoted to the identification of feature specific information pathways in the brain, such as the diverging processing of different visual features by the ventral and the dorsal visual streams [191, 192]. FIT is well-suited to further test, in vivo, current theories about information processing hierarchies in the brain, as well as uncover new fundamental principles in how brain regions communicate at several spatial and temporal scales.

3.7 Acknowledgements

This research has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3), from the NIH Brain Initiative (grants U19 NS107464, R01 NS109961, R01 NS108410), and from the Simons Foundation (SFARI Human Cognitive and Behavioral Science grant 982347). We thank G. Bondanelli and G.M. Lorenz for useful discussion and feedback on the manuscript.

3.8 Supplementary Material

3.8.1 Definitions, derivations, and properties of the information theoretic quantities

In this section, we first define the basic information theoretic quantities that we use in the paper. We next introduce basic concepts of the PID theory needed for our

derivations. We then use these concepts to derive a mathematical definition of FIT and then prove some of its key mathematical properties.

3.8.1.1 Definition of the Shannon information quantities used in this study

In the following we describe and provide analytical expression for the quantities of Shannon's information theory that we used in this paper. These are the mutual information between the stimulus and the neural activity, and the Transfer Entropy (TE), which are estimated in terms of the probabilities of activities of neural signals X and Y and of stimulus features S.

The mutual information $I(S;X_{pres})$ between the stimulus feature S and the neural activity X_{pres} of X at the present time is a non-parametric measure that quantifies the full single-trial statistical relationship between S and X_{pres}. It captures the effect of all linear and nonlinear interactions between these variables. It is defined as follows:

$$I(S;X_{pres}) = \sum_{s,x_{pres}} p(s,x_{pres}) \log \frac{p(s,x_{pres})}{p(s)p(x_{pres})} \tag{3.S5}$$

where $p(s,x_{pres})$ is the joint probability, sampled across experimental trials, of observing in a given trial the joint occurrence of stimulus feature value $s \in S$, and activity $x_{pres} \in X_{pres}$. The sum spans all possible events. $I(S;X_{pres})$ is non-negative, and it is zero if and only if S and X_{pres} are independent. Similar expressions and properties hold for the information $I(S;X_{past})$, $I(S;Y_{past})$ between the stimulus and the past activity X_{past}, Y_{past} of X and Y, respectively.

TE [138] is an information theoretic measure that utilizes the Wiener-Granger principle to quantify the overall propagation of information by neural activity from a putative sender X to a putative receiver Y as the mutual information between the present neural activity of the receiver Y_{pres} and the past activity of the sender X_{past}, conditioned upon the past activity of the receiver Y_{past}. The expression of TE as a function of the joint probability distribution $P(X_{past},Y_{pres},Y_{past})$ is as follows:

$$\begin{aligned} TE(X \to Y) &= I(X_{past};Y_{pres}|Y_{past}) = \\ &= \sum_{x_{past},y_{pres},y_{past}} p(x_{past},y_{pres},y_{past}) \log \frac{p(x_{past},y_{pres}|y_{past})}{p(x_{past}|y_{past})p(y_{pres}|y_{past})} \end{aligned} \tag{3.S6}$$

where $p(x_{past},y_{pres},y_{past})$ is the joint probability, sampled across experimental trials, of observing the joint occurrence of $x_{past} \in X_{past}$, $y_{pres} \in Y_{pres}$, and $y_{past} \in Y_{past}$, and the sum spans all possible events. Importantly, TE does not depend on the stimulus feature S and thus cannot tell how much of the overall information being transmitted from X to Y is about S or about other factors unrelated to S.

3.8.1.2 Elements of PID theory

PID was introduced first in Ref [100] and is a very active field of research [110]. To make our paper self-standing, here we briefly summarize the basic concepts of PID that are most needed for our reasoning and derivations.

In the general case of N source variables $\underline{X} = (X_1, \ldots, X_N)$, PID dissects the joint mutual information that the source variables jointly carry about a target variable T, $I(\underline{X}; T)$, into non-overlapping pieces of redundant, unique, and synergistic information. Let $A_1, \ldots, A_M$ be all the non-empty and potentially overlapping subsets of $\underline{X}$, that we call *sources* in the following. PID considers the collections of sources $\alpha \in P_1(P_1(\underline{X}))$, where $P_1(\underline{X})$ denotes the set of all non-empty subsets of $\underline{X}$. That is, a collection α corresponds to a non-empty subset of *sources*, namely to a non-empty subset of non-empty subsets of source variables. In the following, for brevity, we will call *collections* the collections of sources. Collections α are indicated using a bracketed notation (e.g., $\alpha = \{X_1X_2\}\{X_1X_3\}$ represents the collection of the two overlapping sources $\{X_1X_2\}$ and $\{X_1X_3\}$). Importantly, pieces of unique and synergistic information can be defined and computed algorithmically once the redundant information is identified and computed. Thus, in what follows we focus principally on defining and computing redundancies. For each α, PID defines the amount of information about T that is redundant between all sources in the collection: $I_\cap(T; \alpha)$. Conceptually, the redundancy of any collection α for which a source $A_i \in \alpha$ is a subset of another source $A_j \in \alpha$ $(i \neq j)$ should be equal to the redundancy of the same collection after removing the superset A_j [100]. Therefore, the collections of interest to compute $I_\cap(T; \alpha)$ are only those for which no source is a superset of any other, and hence removing any source $A_i \in \alpha$ could potentially reduce the redundancy. These collections form a domain called $\mathcal{A}(\underline{X})$:

$$\mathcal{A}(\underline{X}) = \{\alpha \in P_1(P_1(\underline{X})) : \forall A_i, A_j \in \alpha, A_i \nsubseteq A_j\} \tag{3.S7}$$

It is possible to define a partial order over the collections of $\mathcal{A}(\underline{X})$. A collection α precedes another collection β if for each source B in β it exists a source A in α that is a subset of B, formally:

$$\forall \alpha, \beta \in \mathcal{A}(\underline{X})(\alpha \preceq \beta \Leftrightarrow \forall B \in \beta \, \exists \, A \in \alpha \,|\, A \subseteq B) \tag{3.S8}$$

Applying the order relationship in eq. 3.S8 to the elements of $\mathcal{A}(\underline{X})$ produces redundancy lattices, in which a collection that succeeds α provides at least as much redundant information about T as α [100] (see Fig. 3.S1A,B for the lattices for $N = 2$ and $N = 3$ source variables). PID allows quantifying the amount of redundant information $I_\partial(T; \alpha)$ that a specific collection α contributes to the joint mutual information about T, and that is not already redundant in any collections preceding α (in the following, we will call $I_\partial(T; \alpha)$ the *information atom* provided by collection α). $I_\partial(T; \alpha)$ is implicitly defined by the following relationship [100]:

$$I_\cap(T; \alpha) = \sum_{\beta \preceq \alpha} I_\partial(T; \beta) \tag{3.S9}$$

Due to the so-called self-redundancy axiom of the PID theory [100], if an individual source A_i appears in collection α, the redundancy computed on collection α is equal to the mutual information between all source variables in A_i and the target variable T:

$$I_\cap(T;\alpha) = I_\cap(T;\{A_i\}) = I(T;A_i) \tag{3.S10}$$

By combining eqs. 3.S9 and 3.S10 we can write Shannon information theoretic quantities as the sum of partial information atoms:

$$I(T;A_i) = \sum_{\beta \preceq A_i} I_\partial(T;\beta) \tag{3.S11}$$

Eq. 3.S11 will be fundamental to provide upper bounds for FIT in terms of Shannon information quantities. When applied to the trivariate system (S, X_1, X_2), taking S as target and (X_1, X_2) as source variables, eq. 3.S11 provides the decomposition of the joint mutual information $I(S;X,Y)$ that we discussed in the main text:

$$I(S;X_1,X_2) = I_\partial(S;\{X_1\}\{X_2\}) + I_\partial(S;\{X_1\}) + I_\partial(S;\{X_2\}) + I_\partial(S;\{X_1X_2\}) \tag{3.S12}$$

where in the main text we called $I_\partial(S;\{X_1\}\{X_2\}) = SI(S:X_1,X_2)$, $I_\partial(S;\{X_1\}) = UI(S:X_1 \setminus X_2)$, $I_\partial(S;\{X_2\}) = UI(S:X_2 \setminus X_1)$, $I_\partial(S;\{X_1X_2\}) = CI(S:X_1,X_2)$ to improve clarity for readers not familiar with PID. SI is shorthand for Shared (that is, redundant) Information; UI is short-hand for Unique information; CI is shorthand for Complementary (that is, synergistic) information.

Thus far we covered elements of PID theory that hold for a generic redundancy measure $I_\cap$, but did not discuss how to compute $I_\cap(T;\alpha)$ for a specific collection α. Several measures of redundant information have been proposed [103, 111, 112, 193], in this work we use the original measure I_{min} from Williams and Beer, as it has the fundamental property of being non-negative for any information atoms for any number N of source variables (not only for $N = 2$) (for a proof, see Appendix D of Ref [100]). The redundant information I_{min} for a collection α is defined as follows:

$$I_{min}(T;\alpha) = \sum_{(t \in T)} (p(t) \min_{A_i \in \alpha} I(T=t;A_i) \tag{3.S13}$$

where $I(T=t;A_i)$ is the specific information that source A_i carries about a specific outcome of the target variable $t \in T$, and is defined as:

$$I(T=t;A) = \sum_a p(a|t)\Big[\log \frac{p(t|a)}{p(t)}\Big] \tag{3.S14}$$

Intuitively, I_{min} quantifies redundancy as the as the overlap in the distributions of specific information across individual values of target variable. This corresponds to quantifying the degree to which all sources in collection α are similarly discriminative about individual values of the target. We decided to use I_{min} because of its advantages in terms of being defined for an arbitrary number of source variables N

(something that is needed because FIT is defined in terms of $N = 3$ source variables and cFIT in terms of $N = 4$ source variables) and being non-negative for all atoms (which is important to guarantee that FIT is interpretable as a measure of information transmission). Importantly, similarly to other redundancy measures [112], I_{min} satisfies the *pairwise marginals* property, meaning that $I_{min}(T; \alpha)$ only depends on the pairwise marginals distributions $p(T, A_i)$ between the target T and each source $A_i \in \alpha$.

Alternative redundancy measures proposed so far are either not straightforward to generalize beyond $N = 2$ source variables [103, 193] or can provide negative information atoms [111]. However, these alternative measures have complementary advantages with respect to I_{min}, such as satisfying the identity property $I_\cap(X, Y; (X, Y)) = I(X; Y)$ which guarantees that, in a system made of two independent variables, the two variables cannot carry redundant information about the whole system. Despite not satisfying this property, the I_{min} measure has been applied to study information processing in simulated neural networks [108], providing insightful and interpretable results.

3.8.1.3 Derivation of FIT

In this subsection, we derive the definition of FIT. In the main text, we used the notation $SUI(S : X_{past}, Y_{pres} \setminus Y_{past})$ to denote the atom of information that is shared by variables X_{past} and Y_{pres} about target S but is unique with respect to a third variable Y_{past}. Using the bracketed notation introduced in Section 3.8.1.2 to denote information atoms, $SUI(S : X_{past}, Y_{pres} \setminus Y_{past})$ corresponds to $I_\partial(S; \{X_{past}\}\{Y_{pres}\})$. From eq. 3.S9, this atom is the difference between the information that X_{past} and Y_{pres} share about S minus the information that X_{past}, Y_{past} and Y_{pres} share about S (see also eq. 3.S18). In the PID literature, information redundant in set of sources about a target that is not redundant with information from another set, has been termed *shared unique* information [194, 195]. Therefore, using the bracketed notation to denote the two atoms of shared unique information $SUI(S : X_{past}, Y_{pres} \setminus Y_{past})$ and $SUI(Y_{pres} : X_{past}, S \setminus Y_{past})$, we can write the definition of FIT as:

$$FIT = min[I_\partial(S; \{X_{past}\}\{Y_{pres}\}), I_\partial(Y_{pres}; \{X_{past}\}\{S\})] \tag{3.S15}$$

We first discuss the mathematical properties of $I_\partial(S; \{X_{past}\}\{Y_{pres}\})$, the *first atom* appearing in FIT definition. We then discuss the complementary mathematical properties met by $(I_\partial(Y_{pres}; \{X_{past}\}\{S\}))$, the *second atom* appearing in FIT definition. To keep our reasoning as general as possible, we discuss properties of the atoms that are valid when the atom is computed using any of the redundancy measures that satisfy the pairwise marginal property (which include the I_{min} redundancy measure that we implemented here). We then demonstrate that a specific pairwise algebraic relationship exists between these two atoms. This relationship is derived from the Shannon information theoretic quantities that relate atoms in the two decompositions. Importantly, this relationship uncovers the presence of a more refined

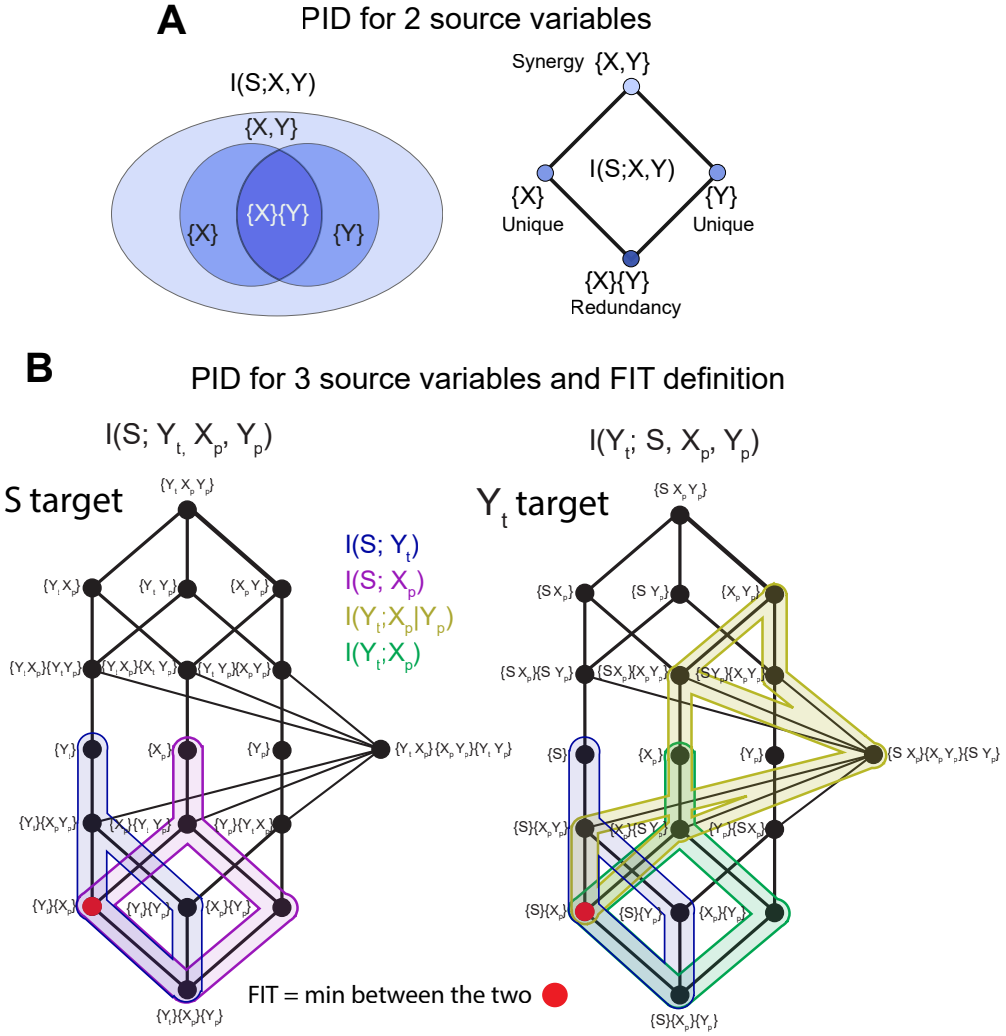

Figure 3.S1: Schematic of the concepts of PID. A) The information $I(S;X,Y)$ that two source variables X, Y carry about a target variable S can be decomposed into four PID atoms. Left: a set-theoretic diagram of the decomposition. Shared information $\{X\}\{Y\}$, darkest shade of blue; unique information $\{X\}$ and $\{Y\}$, lighter shade of blue; synergistic information $\{X,Y\}$, lightest blue. Right: the same decomposition plotted as lattice. A link between two regions symbolizes the ordering relationship of eq. 3.S8. B) FIT is defined on two PID lattices with three sources and one target. Left: The PID lattice with S as target and $(X_{past}, Y_{past}, Y_{pres})$ as sources. Right: the PID with Y_{pres} as target and (S, X_{past}, Y_{past}) as sources. FIT is the minimum between the two atoms highlighted in red. Classical Shannon information theoretic quantities are mapped on the two lattices with different colors (i.e. the sum of all the atoms bounded by a given color is equal to a classical information-theoretic quantity). $I(S;Y_{pres})$ is mapped using blue, $I(S;X_{past})$ using purple, $TE(X \to Y)$ using yellow, and $I(X_{past};Y_{pres})$ using green. The p and t subscript in the Figure is a shorthand for *past* and *pres* respectively.

information component that is shared between the two atoms. Finally, we discuss how taking the minimum between these two atoms ensures that FIT fulfill simultaneously a series of fundamental properties, including being upper bounded at the same time by the feature information encoded in the past activity of the sender X, $I(S;X_{past})$, and in the present activity of the receiver Y, $I(S;Y_{pres})$, and by the total information flowing from X to Y, namely $TE(X \to Y)$.

Properties of the first atom in the FIT definition Our intuitive definition is that FIT should be the information shared between the past activity of a sender region X_{past} and the present activity of a receiver region Y_{pres} about S that is unique with respect to the past activity of the receiver Y_{past}. Thus, within PID of the $(S, X_{past}, Y_{pres}, Y_{past})$ system the most natural candidate is the first atom in eq. 4.S5 ($I_\partial(S;\{X_{past}\}\{Y_{pres}\})$) coming from the decompisition taking $(X_{past}, Y_{past}, Y_{pres})$ as source variable and S as target variable. Using eq. 3.S11, we show that the two Shannon information quantities $I(S;X_{past})$ and $I(S;Y_{pres})$ (i.e., the feature information encoded in the past values of the sender X and of the receiver Y, respectively) set an upper bound on $I_\partial(S;\{X_{past}\}\{Y_{pres}\})$. Indeed, $I(S;X_{past})$ and $I(S;Y_{pres})$ can be written as the sum of information atoms appearing on the lattice having S as target:

$$\begin{aligned} I(S;X_{past}) = I_\partial(S;\{X_{past}\}\{Y_{pres}\}\{Y_{past}\}) + I_\partial(S;\{X_{past}\}\{Y_{past}\}) \\ +I_\partial(S;\{X_{past}\}\{Y_{pres}\}) + I_\partial(S;\{X_{past}\}\{Y_{past}Y_{pres}\}) \\ +I_\partial(S;\{X_{past}\}) \geq I_\partial(S;\{X_{past}\}\{Y_{pres}\}) \end{aligned} \tag{3.S16}$$

$$\begin{aligned} I(S;Y_{pres}) = I_\partial(S;\{X_{past}\}\{Y_{pres}\}\{Y_{past}\}) + I_\partial(S;\{Y_{pres}\}\{Y_{past}\}) \\ +I_\partial(S;\{X_{past}\}\{Y_{pres}\}) + I_\partial(S;\{Y_{pres}\}\{Y_{past}X_{past}\}) \\ +I_\partial(S;\{Y_{pres}\}) \geq I_\partial(S;\{X_{past}\}\{Y_{pres}\}) \end{aligned} \tag{3.S17}$$

which proves that $I_\partial(S;\{X_{past}\}\{Y_{pres}\})$ is upper bounded by both quantities (see Fig. 3.S1B for a graphical depiction of $I(S;Y_{pres})$, in blue, and $I(S;X_{past})$, in purple, upper bound the first atom). However, eq. 3.S11 does not establish any relationship between $I_\partial(S;\{X_{past}\}\{Y_{pres}\})$ and Shannon information between the source variables of the decomposition, including $TE(X \to Y)$. Therefore, the value of the first atom can exceed the total amount of information transmitted from X to Y $TE(X \to Y)$.

Next we prove that, when computed using a redundancy measure that satisfies the *pairwise marginals* property (see Section 3.8.1.2), $I_\partial(S;\{X_{past}\}\{Y_{pres}\})$ only depends on the probability distribution
$P(S, X_{past}, Y_{pres})$ through the pairwise marginal distributions $P(S, X_{past})$ and $P(S, Y_{pres})$, and does not depend explicitly on $P(X_{past}, Y_{pres})$. Indeed, using eq. 3.S9 we can express $I_\partial(S;\{X_{past}\}\{Y_{pres}\})$ as the difference between the redundancy about S computed on collection $\{X_{past}\}\{Y_{pres}\}$ minus the redundancy computed on collection $\{X_{past}\}\{Y_{pres}\}\{Y_{past}\}$:

$$I_\partial(S;\{X_{past}\}\{Y_{pres}\}) = I_\cap(S;\{X_{past}\}\{Y_{pres}\}) - I_\cap(S;\{X_{past}\}\{Y_{pres}\}\{Y_{past}\}) \tag{3.S18}$$

If $I_\cap$ satisfies the pairwise marginals property, then the right-hand side of eq. 3.S18 only depends on the full probability distribution $P(S, X_{past}, Y_{past}, Y_{pres})$ through the pairwise marginal distributions between the target S and the individual sources $P(S, X_{past})$, $P(S, Y_{past})$, and $P(S, Y_{pres})$, but not through the pairwise marginals between the sources, including $P(Y_{pres}, X_{past})$. This implies that if we partially disrupt the dependency structure of our data to create surrogate data, where the individual dependencies of X and of Y on S are preserved (i.e., the pairwise marginals $P(S, X_{past})$ and $P(S, Y_{pres})$ do not change) and the within-trial correlations at a fixed stimulus are disrupted (i.e., the conditional distribution $P(X_{pres}, Y_{past}|S)$ changes), this atom will retain the same value it had in the original data. Therefore, this atom alone cannot rule out confounding scenarios where X and Y encode S independently with a temporal lag, with no information transfer at fixed stimulus value.

Properties of the second atom in the FIT definition Atoms satisfying mathematical properties that are complementary to the ones of $I_\partial(S; \{X_{past}\}\{Y_{pres}\})$ exist on the decomposition with Y_t as target. On this decomposition one atom that intuitively captures feature-specific information flow is $I_\partial(Y_{pres}; \{X_{past}\}\{S\})$, i.e. the information that the past activity of the sender X and the feature S share about the present activity of the receiver Y that is unique with respect to the past activity of the receiver Y. We first prove that this atom is upper bounded by the value of $TE(X \to Y)$, and then that this atom depends on $P(Y_{pres}, X_{past})$.

To prove that the value of $TE(X \to Y)$ sets an upper bound to $I_\partial(Y_{pres}; \{X_{past}\}\{S\})$, we first use the information-theoretic chain rule [73], to write the conditioned mutual information in eq. 3.S6 as the difference between the joint mutual information that X_{past} and Y_{past} carry about Y_{pres} minus the mutual information between Y_{pres} and Y_{past}. Then, we use eq. 3.S11 to write the two information quantities as the sum of non-negative information atoms, including $I_\partial(Y_{pres}; \{X_{past}\}\{S\})$:

$$\begin{aligned} TE(X \to Y) &= I(Y_{pres}; X_{past}, Y_{past}) - I(Y_{pres}; Y_{past}) = \\ &= I_\partial(Y_{pres}; \{S\}\{X_{past}\}) + I_\partial(Y_{pres}; \{X_{past}\}\{SY_{past}\}) \\ &\quad + I_\partial(Y_{pres}; \{S\}\{X_{past}Y_{past}\}) + I_\partial(Y_{pres}; \{X_{past}\}) \\ &\quad + I_\partial(Y_{pres}; \{X_{past}S\}\{Y_{past}S\}\{X_{past}Y_{past}\}) \\ &\quad + I_\partial(Y_{pres}; \{X_{past}S\}\{X_{past}Y_{past}\}) + I_\partial(Y_{pres}; \{Y_{past}S\}\{X_{past}Y_{past}\}) \\ &\quad + I_\partial(Y_{pres}; \{X_{past}Y_{past}\}) \geq I_\partial(Y_{pres}; \{S\}\{X_{past}\}) \end{aligned} \tag{3.S19}$$

which proves that $I_\partial(Y_{pres}; \{X_{past}\}\{S\})$ is upper bounded by $TE(X \to Y)$ (see Fig. 3.S1B for a graphical depiction of the mapping of $TE(X \to Y)$, in yellow, on the lattice to which this second atom belongs).

Similarly to the first atom, this second atom is also upper bounded by $I(S; Y_{pres})$ (not proven, but see the blue quantity in Fig. 3.S1B for a graphical depiction of this property), however it is not upper bounded by Shannon information quantities between the source variables of the decomposition with Y_{pres} as target, and in particular

by $I(S; X_{past})$. This is important because it proves that neither the second atom alone satisfies all the properties that we require for a measure of feature-specific information transfer.

We then prove that the second atom depends on $P(Y_{pres}, X_{past})$, a property which makes it suited to rule out confounding scenarios where X and Y independently encode the S but no communication occurs between the two. To do so, we use eq. 3.S9 to write the second atom as the difference between two redundancy terms:

$$I_\partial(Y_{pres}; \{X_{past}\}\{S\}) = I_\cap(Y_{pres}; \{X_{past}\}\{S\}) - I_\cap(Y_{pres}; \{X_{past}\}\{S\}\{Y_{past}\}) \tag{3.S20}$$

If $I_\cap$ satisfies the pairwise marginals property, then the right-hand side. of eq. 3.S18 depends on the full probability distribution $P(S, X_{past}, Y_{past}, Y_{pres})$ through the marginal distributions between the target Y_{pres} and the individual sources $P(Y_{pres}, X_{past})$, $P(Y_{pres}, S)$, and $P(Y_{pres}, Y_{past})$. This implies that if we partially disrupt the dependency structure of our data and create surrogate data where the individual dependencies of X and of Y on S are preserved (i.e., the pairwise marginals $P(S, X_{past})$ and $P(S, Y_{pres})$ do not change) and the within-trial correlations at a fixed stimulus are disrupted (i.e., the conditional distribution $P(X_{pres}, Y_{past}|S)$ changes), the value of the referenced atom may differ from its original value. This change occurs because this operation generally disrupts $P(X_{pres}, Y_{past})$. Therefore, this atom can rule out confounding scenarios where X and Y encode S independently with a temporal lag, with no information transfer at fixed stimulus value.

The two atoms in the FIT definition are related by Shannon Information theoretic quantities This Section is structured as follows. First, we present some basic findings from Ref. [113] where, the authors showed that atoms from different decompositions are algebraically constrained by Shannon's information-theoretic quantities and used these constraints to identify, specifically for a trivariate system, a reduced set of finer information components which could describe all atoms across different decompositions. Next, we express the algebraic constraints between two decompositions as a homogeneous linear system of equations. We demonstrate that the reduced set of information components derived for two decompositions in Ref. [113] can be obtained as solutions to this homogeneous system. Finally, we derive the analogous homogeneous system in the case of four variables. One solution to this system relates specifically the two atoms appearing in the FIT definition.

Ref [113] showed that atoms belonging to different decompositions are algebraically constrained by information-theoretic quantities. These constraints derive from fundamental axioms of PID theory, specifically the fact that in the system $(X_1, \ldots, X_N)$, we can use eq. 3.S11 to express the mutual information between two variables X_i and X_j (conditioned on up to $N-2$ other variables) as the sum of information atoms from both the decomposition with X_i and the decomposition with X_j as target variable. For the trivariate system (S, X, Y), Shannon information quantities impose two linear constraints between the 4 atoms of information having S as target and the 4 atoms of information having Y as target (all atoms are:

$I_\partial(S;\{X\}\{Y\})$, $I_\partial(S;\{Y\})$, $I_\partial(S;\{X\})$, $I_\partial(S;\{XY\})$, $I_\partial(Y;\{X\}\{S\})$, $I_\partial(Y;\{S\})$, $I_\partial(Y;\{X\})$, $I_\partial(Y;\{XS\})$):

$$\begin{aligned} I(S;Y) &= I_\partial(S;\{X\}\{Y\}) + I_\partial(S;\{Y\}) = I_\partial(Y;\{X\}\{S\}) + I_\partial(Y;\{S\}) \\ I(S;Y|X) &= I_\partial(S;\{XY\}) + I_\partial(S;\{Y\}) = I_\partial(Y;\{XS\}) + I_\partial(Y;\{S\}) \end{aligned} \tag{3.S21}$$

Combining the two equations in the system of eqs. 3.S21 reveals an equality among the differences in the amount of information carried by pairs of similar atoms (the two redundancies $I_\partial(S;\{X\}\{Y\})$ and $I_\partial(Y;\{X\}\{S\})$, the two synergies $I_\partial(S;\{XY\})$ and $I_\partial(Y;\{XS\})$, and the two unique information $I_\partial(Y;\{S\})$ and $I_\partial(S;\{Y\})$):

$$\begin{aligned} &I_\partial(S;\{X\}\{Y\}) - I_\partial(Y;\{X\}\{S\}) \\ &= I_\partial(Y;\{S\}) - I_\partial(S;\{Y\}) = I_\partial(S;\{XY\}) - I_\partial(Y;\{XS\}) \end{aligned} \tag{3.S22}$$

Therefore, 6 atoms of the 8 atoms belonging to the two decompositions (those appearing in eq 3.S21) are not independent, while $I_\partial(S;\{X\})$ and $I_\partial(Y;\{X\})$ are independent from all other atoms. In Ref. [113] the authors showed that, due to the two constraints of eq. 3.S21, the 8 atoms can be described by 6 finer independent information components (that they called information *subatoms*). In Ref. [113] they quantify these 6 subatoms as follows: three subatoms are the minimum between pairs of similar atoms belonging to the two decomposition (i.e., the two redundancies $I_\partial(S;\{X\}\{Y\})$ and $I_\partial(Y;\{X\}\{S\})$, the two synergies $I_\partial(S;\{XY\})$ and $I_\partial(Y;\{XS\})$ and the two unique information $I_\partial(S;\{Y\})$ and $I_\partial(Y;\{S\})$); one subatom is equal to the difference between the maximum and the minimum in each of the above pairs (which is equal for the three pairs, see eq. 3.S22); two subatoms are equal to the unconstrained atoms not appearing in eq. 3.S21 ($I_\partial(S;\{X\})$ and $I_\partial(Y;\{X\})$).

A novel perspective on the relationships between the amounts of information carried by specific sets of atoms from different decompositions is to conceptualize the eight atoms as forming an eight-dimensional vector space, $\mathcal{V}$. We can represent a generic column vector in $\mathcal{V}$ as $\underline{v}$ and the 2×8 matrix of constraints imposed by Shannon information quantities relating the atoms of the two decompositions as $\underline{\underline{B}}$. With these definitions, we can express the system of eqs. 3.S21 as a homogeneous linear system:

$$\underline{\underline{B}}\,\underline{v} = \underline{0} \tag{3.S23}$$

Specifically, the coefficients of $\underline{\underline{B}}$ are obtained by taking the difference between the middle- and the right-term in the two eqs. 3.S21. Ordering the dimensions of $\mathcal{V}$ as ($I_\partial(S;\{X\}\{Y\})$, $I_\partial(S;\{Y\})$, $I_\partial(S;\{X\})$, $I_\partial(S;\{XY\})$, $I_\partial(Y;\{X\}\{S\})$, $I_\partial(Y;\{S\})$, $I_\partial(Y;\{X\})$, $I_\partial(Y;\{XS\})$), $\underline{\underline{B}}$ has the following form:

$$\underline{\underline{B}} = \begin{bmatrix} 1 & 1 & 0 & 0 & -1 & -1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & -1 & 0 & -1 \end{bmatrix} \tag{3.S24}$$

It can easily be verified that, for instance, $I_\partial(S;\{X\}\{Y\}) = I_\partial(Y;\{X\}\{S\})$, is a solution of the homogeneous system in eq. 3.S23 for the matrix $\underline{\underline{B}}$ defined as in eq. 3.S24. Consider a matrix multiplication between $\underline{\underline{B}}$ and the vector $\underline{v}_{SI}$, whose only non-zero components are $I_\partial(S;\{X\}\{Y\})$ and $I_\partial(Y;\{X\}\{S\})$ (i.e., $\underline{v}_{SI} = (I_\partial(S;\{X\}\{Y\}), 0, 0, 0, I_\partial(Y;\{X\}\{S\}), 0, 0, 0)$). This matrix multiplication is equivalent to multiplying the coefficients in columns 1 and 5 of $\underline{\underline{B}}$ by the two atoms, respectively, and doing the element-wise sum of the two resulting vectors. This sum is zero if and only if $I_\partial(S;\{X\}\{Y\}) = I_\partial(Y;\{X\}\{S\})$. Put simply, columns of $\underline{\underline{B}}$ with element-wise opposite coefficients correspond to pairs (or triplets) of atoms that form a solution of eq. 3.S23 when they have equal value. As a result, the following are all nontrivial solutions of the homogeneous system in eq. 3.S23, or equivalently, they belong to the null space of $\underline{\underline{B}}$: $I_\partial(S;\{X\}\{Y\}) = I_\partial(Y;\{X\}\{S\})$, $I_\partial(S;\{Y\}) = I_\partial(Y;\{S\})$, $I_\partial(S;\{XY\}) = I_\partial(Y;\{XS\})$, $I_\partial(S;\{X\})$, $I_\partial(Y;\{X\})$, and $I_\partial(S;\{X\}\{Y\}) = I_\partial(S;\{XY\}) = I_\partial(Y;\{S\})$. These solutions uncover specific relationships between pairs or triplets of atoms across different decompositions. Importantly, considering that the eight atoms are not independent and can be represented by six finer, independent quantities, these solutions lend support to the notion that these finer components of information are shared among the atoms linked by a single solution. Remarkably, the atoms identified as related by a solution precisely correspond to the six subatoms previously defined in Ref [113].

Similar to the case of $N = 2$ source variables, for $N = 3$ source variables, there are 36 information atoms (18 per lattice) that belong to two decompositions with different targets. Shannon information quantities impose constraints relating these 36 atoms, implying the existence of finer information components (or subatoms) that can describe the two decompositions even when there are $N = 3$ source variables. Our goal here is not to uncover the complete set of components that describe all atoms belonging to the two decompositions. Rather, we aim to demonstrate that a specific algebraic relationship, similar to the ones discussed above, exists between the two atoms present in the FIT definition. To do this, we generalize the homogeneous linear system in eq. 3.S23 to the four-variable case $(S, X_{past}, Y_{past}, Y_{pres})$. In this scenario, the two decompositions that have S and Y_{pres} as their respective targets (represented by the two lattices in Fig. 3.S1B) are constrained by the following four Shannon information quantities:

$$\begin{gathered} I(S;Y_{pres}) \\ I(S;Y_{pres}|X_{past}) \\ I(S;Y_{pres}|Y_{past}) \\ I(S;Y_{pres}|Y_{past},X_{past}) \end{gathered} \tag{3.S25}$$

Similarly to eq. 3.S21, we can use eq. 3.S11 to express the 4 quantities in eq. 3.S25 as sums of atoms either belonging to the decomposition with S as target, or the one with Y_{pres} as target. As an example, in Fig. 3.S1B, we demonstrate that $I(S;Y_{pres})$ is the sum of atoms belonging to both decompositions, which together

consist of 36 atoms (18 per decomposition). In general, each quantity in eq. 3.S25 imposes constraints between two sets of many atoms from the two decompositions.

To study numerically the solutions of 3.S23 for these 36 atoms, we wrote a MATLAB script named *FIT_nullB.m*. This script computes the four quantities in eq. 3.S25 as the sum of atoms from either the decomposition with S or Y_{pres} as target. It then constructs the 4×36 matrix $\underline{\underline{B}}$ (Fig. 3.S2) in a similar way to how we derived the 2×8 matrix in eq. 3.S24 from the eqs. in 3.S21. From Fig. 3.S2, it is clear that in this four-variable case, some atoms, such as $I_\partial(Y_{pres}; \{X_{past}\}\{Y_{past}\})$, are not constrained by $\underline{\underline{B}}$ and can vary independently (analogously to how for $N = 2$ the two terms $I_\partial(S; \{X\})$ and $I_\partial(Y; \{X\})$) were unconstrained). However, there are also pairs of atoms that are not independent when considered individually, but that are specifically related by eq. 3.S23 (i.e. the equality between the two atoms in the pair is a solution of eq. 3.S23). A notable example of these pairwise solutions is made by the pair of atoms appearing in the FIT definition: $I_\partial(S; \{X_{past}\}\{Y_{pres}\}) = I_\partial(Y_{pres}; \{X_{past}\}\{S\})$. This relationship can be easily verified from Fig. 3.S2, where the first and the second atom are highlighted in red and white, respectively. Indeed, drawing from the intuition developed in the $N = 2$ source variables case, these two atoms belong to columns of $\underline{\underline{B}}$ with element-wise opposite coefficients. This solution (Fig. 3.S2) reveals a specific pairwise relationship between the two atoms appearing in the FIT definition and supports the existence of a finer component of information shared by these two atoms. It is actually apparent from the plot in Fig. 3.S2 that this is the only pairwise relationship involving any of the two atoms. We quantify this finer component of information by taking the minimum between the two related atoms.

Proofs and summary of the main mathematical properties of FIT Here we prove that FIT defined as in eq. 4.S5 satisfies the two following properties:

1. FIT is simultaneously upper bounded by $I(S; X_{past})$, $I(S; Y_{pres})$, and $TE(X \to Y)$.

2. FIT depends on $P(S, X_{past}, Y_{pres})$ through all the pairwise marginal distributions $P(S, X_{past})$, $P(S, Y_{pres})$, and $P(X_{past}, Y_{pres})$. Thus, FIT can rule out confounding scenarios where X and Y independently encode S with a temporal lag in absence of within-trial correlations between X and Y at fixed stimulus.

To prove that FIT is simultaneously upper bounded by $I(S; X_{past})$, $I(S; Y_{pres})$, and $TE(X \to Y)$, it is sufficient to note that FIT is simultaneously upper bounded by all quantities that set an upper bound to the two atoms appearing in its definition. This can be seen from eqs. 3.S16, 3.S17, and 3.S19, which show:

$$\begin{aligned} FIT &\leq I_\partial(S; \{X_{past}\}\{Y_{pres}\}) \leq I(S; X_{past}) \\ FIT &\leq I_\partial(S; \{X_{past}\}\{Y_{pres}\}) \leq I(S; Y_{pres}) \\ FIT &\leq I_\partial(Y_{pres}; \{X_{past}\}\{S\}) \leq TE(X \to Y) \end{aligned} \tag{3.S26}$$

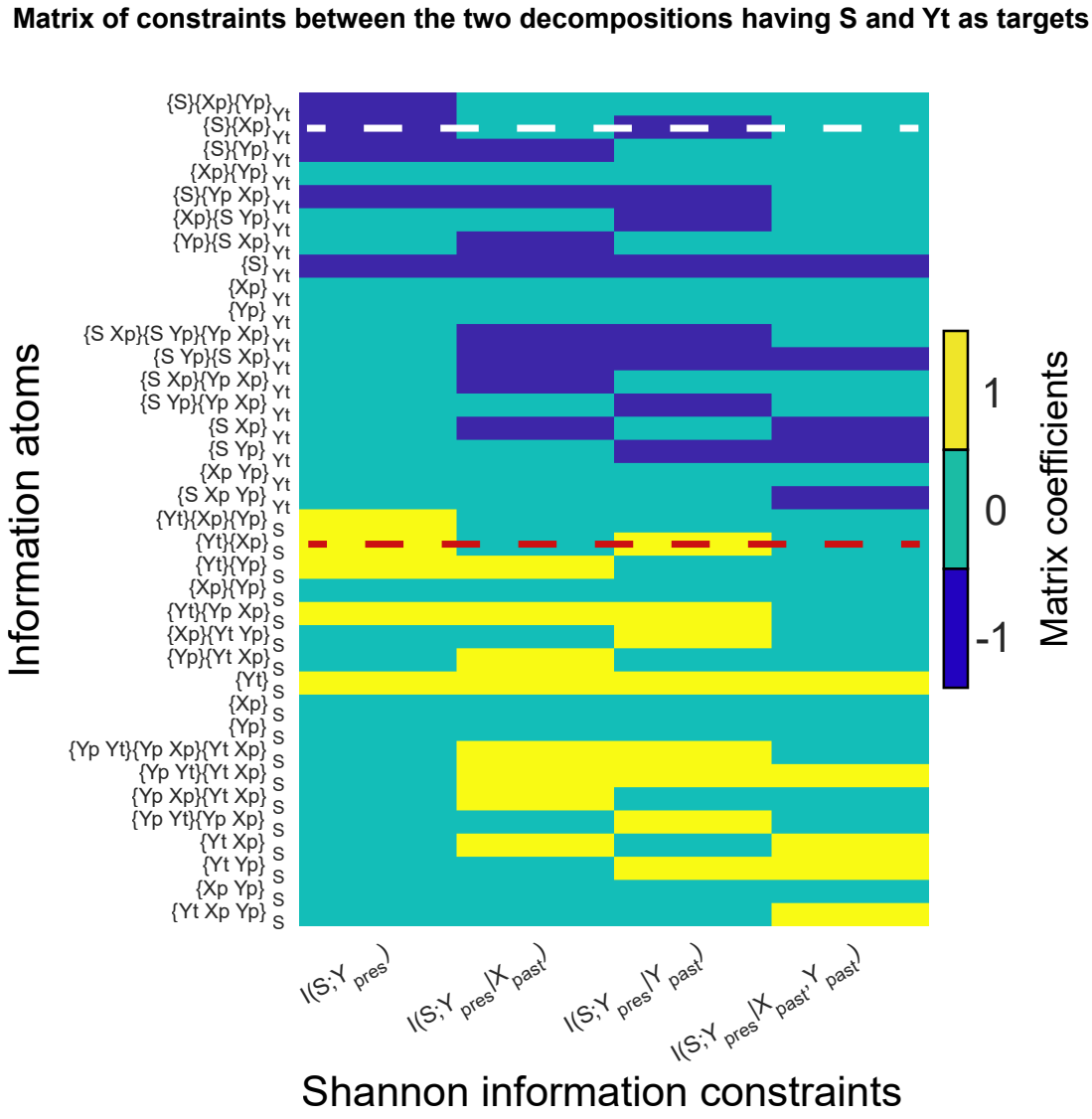

Figure 3.S2: Matrix of constraints imposed by Shannon information-theoretic quantities relating the PID having $(Y_{pres}, X_{past}, Y_{past})$ as source variables and S as target variable to the PID having (S, X_{past}, Y_{past}) as source variables and Y_{pres} as target variable. For brevity, we used the notation $Y_t = Y_{pres}, X_p = X_{past}, Y_p = Y_{past}$ and denoted each atom (y axis) directly with the collection it is computed on, with a subscript indicating the target variable of the decomposition (e.g. $\{X_p\}\{Y_t\}_S = I_\partial(S; \{X_{past}\}\{Y_{pres}\})$). For better visibility, we plotted the transpose of the 4×36 matrix appearing in eq. 3.S23. The red dashed line highlights the first atom appearing in FIT definition $I_\partial(S; \{X_{past}\}\{Y_{pres}\})$, the white line highlights the second atom in FIT definition $I_\partial(Y_{pres}; \{X_{past}\}\{S\})$. Importantly, only atom highlighted in red has coefficients that are opposite to the ones of the atom highlighted in white.

A particularly important consequence of the upper bound set by $TE(X \to Y)$ is that if X and Y are independent, then $FIT = 0$. Indeed, if X is independent of Y,

then X_{past} is independent of (Y_{past}, Y_{pres}), and therefore $I(X_{past}; Y_{past}, Y_{pres}) = 0$. By applying the information theoretic chain rule [73] to $I(X_{past}; Y_{past}, Y_{pres})$, we obtain:

$$\begin{aligned} I(X_{past}; Y_{past}, Y_{pres}) &= I(X_{past}; Y_{pres}|Y_{past}) + I(X_{past}; Y_{past}) \\ &\geq I(X_{past}; Y_{pres}|Y_{past}) \geq FIT \end{aligned} \tag{3.S27}$$

Proving that, if X and Y are independent, then $FIT = 0$. Another important point is that none of the 36 atoms belonging to either decomposition with S or decomposition with Y_{pres} as target satisfies the first property. Indeed eq. 3.S11 does not establish any relationship between atoms of the decomposition with S as target and Shannon information between the sources of the decomposition, including $TE(X \to Y)$ (see Fig. 3.S1B, left), nor between atoms of the decomposition with Y_{pres} as target and Shannon information between the sources of the decomposition, including $I(S; X_{past})$ (see Fig. 3.S1B, right). Therefore it is necessary to simultaneously consider atoms belonging to different decompositions to obtain a quantity that satisfies the first property.

To prove that FIT depends on $P(S, X_{past}, Y_{pres})$ through all the pairwise marginal distributions $P(S, X_{past})$, $P(S, Y_{pres})$, and $P(X_{past}, Y_{pres})$ we also leveraged on the simultaneous dependence of FIT on $I_\partial(S; \{X_{past}\}\{Y_{pres}\})$ and on $I_\partial(Y_{pres}; \{X_{past}\}\{S\})$. $I_\partial(S; \{X_{past}\}\{Y_{pres}\})$ and $I_\partial(Y_{pres}; \{X_{past}\}\{S\})$ depend of $P(S, X_{pres}, Y_{past})$ through the marginals $P(S, X_{pres})$ and $P(S, Y_{past})$, and $P(Y_{pres}, X_{pres})$ and $P(S, Y_{past})$, respectively. Therefore:

$$FIT = f(P(S, X_{pres}, Y_{past})) = f(P(S, X_{past}), P(S, Y_{pres}), P(X_{past}, Y_{pres})) \tag{3.S28}$$

This implies that if we partially disrupt the dependency structure of our data and create surrogate data where the individual dependencies of X and of Y on S are preserved (i.e., the pairwise marginals $P(S, X_{past})$ and $P(S, Y_{pres})$ do not change) and the within-trial correlations at a fixed stimulus are disrupted (i.e., the conditional distribution $P(X_{pres}, Y_{past}|S)$ changes), the value of the FIT can differ from its original value. This change occurs because this operation generally disrupts $P(X_{pres}, Y_{past})$. Therefore, FIT can rule out confounding scenarios where X and Y encode S independently with a temporal lag, with no information transfer at fixed stimulus value.

3.8.1.4 The conditional feature specific information cFIT

Here we discuss the definition and the properties of the conditioned version of FIT, termed cFIT.

Definition and derivation of cFIT We defined a conditioned version of FIT, to remove from the feature information transmitted from X to Y (that in this Section we term FIT_X) the information potentially routed through the past activity of a third region Z (Z can, in principle, also be the multivariate activity of a set

of regions). To do so, we identified subcomponents of the two atoms in the FIT definition that quantified pieces of information that were also shared with the past of Z, and removed them from FIT.

In this subsection, we will be working with atoms computed on collections belonging to decompositions with $N = 3$ and $N = 4$ source variables. To avoid any confusion, we will explicitly denote the number of source variables of each collection in the following discussion. For example, we will use the notation $\{X_{past}\}\{Y_{pres}\}_{(3)}$ to indicate the collection on which the first atom in the FIT definition is computed. This collection refers to the PID with $N = 3$ source variables $(X_{past}, Y_{pres}, Y_{past})$ and target variable S.

Previous studies showed that, using eq. 3.S11 , atoms on the PID with target T and N source variables $\underline{X}_N = (X_1, \ldots, X_N)$ can be written as the sum of finer atoms belonging to the PID with same target and an additional source variable $\underline{X}_{N+1} = (X_1, \ldots, X_N, X_{N+1})$[196]. Importantly, collections α present in the PID with N source variables, denoted as $\alpha_{(N)}$, also exist in the PID with $N + 1$ source variables, denoted as $\alpha_{(N+1)}$, since $\underline{X}_N \subset \underline{X}_{N+1}$. However, the opposite does not necessarily hold.

The atom $I_\partial(T; \alpha_{(N)})$ is the sum of atoms $I_\partial(T; \beta_{(N+1)})$, where $\beta_{(N+1)}$ simultaneously precedes $\alpha_{(N+1)}$ in the PID with $N+1$ source variables (as per the ordering relationship of eq. 3.S8, in which precedence includes equality), but does not precede any collections $\gamma_{(N+1)}$ such that $\gamma_{(N)}$ precedes $\alpha_{(N)}$ [196]. For example, the collection $\{X_{past}\}\{Y_{pres}\}_{(3)}$, present in the PID with $N = 3$ source variables $(X_{past}, Y_{past}, Y_{pres})$ and target S, is preceded only by the collection $\{X_{past}\}\{Y_{pres}\}\{Y_{past}\}_{(3)}$ (i.e., the information that all source variables $(X_{past}, Y_{past}, Y_{pres})$ share about S) and by itself (see Fig.3.S1B, left). When adding Z_{past} to the set of source variables, the collection $\{X_{past}\}\{Y_{pres}\}_{(4)}$ is preceded by four collections (additionally to itself): $\{X_{past}\}\{Y_{pres}\}\{Y_{past}\}_{(4)}$, $\{X_{past}\}\{Y_{pres}\}\{Y_{past}\}\{Z_{past}\}_{(4)}$, $\{X_{past}\}\{Y_{pres}\}\{Z_{past}\}_{(4)}$, and $\{X_{past}\}\{Y_{pres}\}\{Z_{past}Y_{past}\}_{(4)}$ (see Fig.3.S7A, right). Collection $\{X_{past}\}\{Y_{pres}\}\{Y_{past}\}_{(4)})$, which was not preceded by any collection apart from itself in the PID with $N = 3$ variables, is preceded also by $\{X_{past}\}\{Y_{pres}\}\{Y_{past}\}\{Z_{past}\}_{(4)})$. Therefore:

$$\begin{aligned} &I_\partial(S; \{X_{past}\}\{Y_{pres}\}\{Y_{past}\}_{(3)}) \\ &= I_\partial(S; \{X_{past}\}\{Y_{pres}\}\{Y_{past}\}_{(4)}) + I_\partial(S; \{X_{past}\}\{Y_{pres}\}\{Y_{past}\}\{Z_{past}\}_{(4)}) \end{aligned} \tag{3.S29}$$

which intuitively means that, when considering also the past of a third region Z, the information that $(X_{past}, Y_{past}, Y_{pres})$ share about S breaks down into a component that is also shared with Z_{past} and a component that is unique with respect to Z_{past}.

The other two collections $\{X_{past}\}\{Y_{pres}\}\{Z_{past}\}_{(4)})$ and $\{X_{past}\}\{Y_{pres}\}\{Z_{past}Y_{past}\}_{(4)})$ precede
$\{X_{past}\}\{Y_{pres}\}_{(4)})$ but do not precede $\{X_{past}\}\{Y_{pres}\}\{Y_{past}\}\}_{(4)})$ (Fig.3.S7A). Therefore:

$$\begin{aligned} &I_\partial(S; \{X_{past}\}\{Y_{pres}\}_{(3)}) = I_\partial(S; \{X_{past}\}\{Y_{pres}\}_{(4)}) \\ &+ I_\partial(S; \{X_{past}\}\{Y_{pres}\}\{Y_{past}Z_{past}\}_{(4)}) + I_\partial(S; \{X_{past}\}\{Y_{pres}\}\{Z_{past}\}_{(4)}) \end{aligned} \tag{3.S30}$$

which shows how the first atom in FIT definition (eq. 4.S5) breaks down into three components in the PID with $(X_{past}, Y_{pres}, Y_{past}, Z_{past})$ as source variables and S as target variable (one component that is unique with respect to Y_{past} but shared with Z_{past}, one that is unique with respect to both Y_{past} and Z_{past} but shared with $\{Y_{past}Z_{past}\}$, and one that is also unique with respect to $\{Y_{past}Z_{past}\}$). One of these atoms is the information that X_{past}, Y_{pres}, and Z_{past} share about S, i.e. the component of the first FIT atom that is also shared with Z_{past}: $I_\partial(S; \{X_{past}\}\{Y_{pres}\}\{Z_{past}\}_{(4)})$.

Similarly, the second atom appearing in the FIT definition, is the sum of finer atoms belonging to the PID with $(X_{past}, S, Y_{past}, Z_{past})$ as source variables and Y_{pres} as target variable:

$$\begin{aligned} I_\partial(Y_{pres}; \{X_{past}\}\{S\}_{(3)}) = I_\partial(Y_{pres}; \{X_{past}\}\{S\}_{(4)}) + \\ + I_\partial(Y_{pres}; \{X_{past}\}\{S\}\{Y_{past}Z_{past}\}_{(4)}) + I_\partial(Y_{pres}; \{X_{past}\}\{S\}\{Z_{past}\}_{(4)}) \end{aligned} \tag{3.S31}$$

One of these atoms is the information that X_{past}, S, and Z_{past} share about Y_{pres}, i.e. the component of the second FIT atom that is also shared with Z_{past}: $I_\partial(Y_{pres}; \{X_{past}\}\{S\}\{Z_{past}\}_{(4)})$.

To remove from FIT the information that is also shared with Z_{past} we defined the conditioned FIT (cFIT) from X to Y conditioned to Z as:

$$\begin{aligned} cFIT_{X|Z} = min[I_\partial(S; \{X_{past}\}\{Y_{pres}\}_{(3)}), I_\partial(Y_{pres}; \{X_{past}\}\{S\}_{(3)})] + \\ - min[I_\partial(S; \{X_{past}\}\{Y_{pres}\}\{Z_{past}\}_{(4)}), I_\partial(Y_{pres}; \{X_{past}\}\{S\}\{Z_{past}\}_{(4)})] \end{aligned} \tag{3.S32}$$

Therefore $cFIT_{X|Z}$ is equal to FIT from X to Y (cf. eq. 4.S5) minus a term that is the minimum between two similar information atoms (both quantifying intuitively the feature information about S that both the past of X and the past of Z share with the present of Y, but is unique with respect to the past of Y) on the two PID having S and having Y_{pres} as target variables, respectively.

Properties of cFIT In this Section we prove two properties of cFIT, under the assumption that we compute PID atoms using a redundancy measure (such as I_{min}) that is non-negative for each atom. The first property we prove (i) is that $cFIT_{X|Z}$ is upper bounded by FIT_X and is lower bounded by the maximum between 0 and $FIT_X - FIT_Z$ (where we denote as FIT_X the feature information transmitted from X to Y and FIT_Z the one transmitted from Z to Y). The second property that we prove (ii) is that if $S \to Z_{past} \to Y_{pres}$ is a Markov chain (i.e. $P(S; Y_{pres}|Z_{past}) = P(S|Z_{past})P(Y_{pres}|Z_{past})$) then $cFIT_{X|Z} = 0$. This second property is important because it means that if the present of Y received all its feature information from the past of a recorded region Z, then there is no residual FIT through X once any contribution from Z is eliminated.

We start by proving property (i). From eq. 3.S32, since we subtract from FIT the minimum between two non-negative quantities, it immediately follows that $cFIT_{X|Z} \leq FIT_X$. This proves that $cFIT_{X|Z}$ is upper bounded by FIT_X. Then, since from eqs. 3.S30 and 3.S31 we have that

$$\begin{aligned} I_\partial(S; \{X_{past}\}\{Y_{pres}\}_{(3)}) \geq I_\partial(S; \{X_{past}\}\{Z_{past}\}\{S\}_{(4)}) \\ I_\partial(Y_{pres}; \{X_{past}\}\{Y_{pres}\}_{(3)}) \geq I_\partial(Y_{pres}; \{X_{past}\}\{Z_{past}\}\{S\}_{(4)}) \end{aligned} \tag{3.S33}$$

from which it follows that:

$$\begin{aligned} &min[I_\partial(S;\{X_{past}\}\{Z_{past}\}\{Y_{pres}\}_{(4)}), I_\partial(Y_{pres};\{X_{past}\}\{Z_{past}\}\{S\}_{(4)})] \\ &\leq min[I_\partial(S;\{X_{past}\}\{Y_{pres}\}_{(3)}), I_\partial(Y_{pres};\{X_{past}\}\{S\}_{(3)})] \end{aligned} \tag{3.S34}$$

Eq. 3.S34 shows that the term that we subtract from the right-hand side in eq. 3.S32 is lower or equal to the first one, proving that $cFIT_{X|Z} \geq 0$.

Finally, we prove that $cFIT_{X|Z} \geq FIT_X - FIT_Z$. We do so by proving that the term we subtract from FIT_X in the definition of $cFIT_{X|Z}$ (eq. 3.S32) is smaller than FIT_Z. FIT_Z is defined on the two decompositions having $(Y_{pres}, Y_{past}, Z_{past})$ as sources and S as target variable, and the one having (S, Y_{past}, Z_{past}) as sources and Y_{pres} as target variable:

$$FIT_Z = min[I_\partial(S;\{Z_{past}\}\{Y_{pres}\}_{(3)}), I_\partial(Y_{pres};\{Z_{past}\}\{S\}_{(3)})] \tag{3.S35}$$

Similarly to eqs. 3.S30 and 3.S31, the two atoms in 3.S35 break down into the sum of finer information atoms - when adding variable X_{past} to the respective sets of source variables (in Fig.3.S7A we show a graphical depiction of the decomposition of the first atom in FIT_Z definition, depicted in light blue):

$$\begin{aligned} &I_\partial(S;\{Z_{past}\}\{Y_{pres}\}_{(3)}) = I_\partial(S;\{Z_{past}\}\{Y_{pres}\}_{(4)}) \\ &+I_\partial(S;\{Z_{past}\}\{Y_{pres}\}\{Y_{past}X_{past}\}_{(4)}) + I_\partial(S;\{Z_{past}\}\{Y_{pres}\}\{X_{past}\}_{(4)}) \end{aligned} \tag{3.S36}$$

$$\begin{aligned} &I_\partial(Y_{pres};\{X_{past}\}\{S\}_{(3)}) = I_\partial(Y_{pres};\{X_{past}\}\{S\}_{(4)}) \\ &+I_\partial(Y_{pres};\{X_{past}\}\{S\}\{Y_{past}Z_{past}\}_{(4)}) + I_\partial(Y_{pres};\{X_{past}\}\{S\}\{Z_{past}\}_{(4)}) \end{aligned} \tag{3.S37}$$

From eqs. 3.S36 and 3.S37 it follows that $I_\partial(S;\{Z_{past}\}\{Y_{pres}\}_{(3)}) \geq I_\partial(S;\{Z_{past}\}\{X_{past}\}\{Y_{pres}\}_{(4)})$ (i.e. the information the the past of Z and the present of Y share about S is larger than the information that they both share also with the past of X about S) and $I_\partial(Y_{pres};\{Z_{past}\}\{S\}\}_{(3)}) \geq I_\partial(Y_{pres};\{Z_{past}\}\{X_{past}\}\{S\}\}_{(4)})$. Thus:

$$FIT_Z \geq min[I_\partial(S;\{Z_{past}\}\{Y_{pres}\}\{X_{past}\}_{(4)}), I_\partial(Y_{pres};\{X_{past}\}\{S\}\{Z_{past}\}_{(4)})] \tag{3.S38}$$

The above proves that $cFIT_{X|Z} \geq FIT_X - FIT_Z$. This is important because it assures that the component that we subtract from FIT_X when removing from it any contribution potentially due to Z_{past} cannot exceed the feature information transmitted from Z to Y (it we only remove the FIT_Z that is shared with FIT_X). To summarize, we proved that $cFIT_{X|Z} \leq FIT_X$, that $cFIT_{X|Z} \geq 0$ and that $cFIT_{X|Z} \geq FIT_X - FIT_Z$, meaning that $cFIT_{X|Z}$ is upper bounded by FIT_X and is lower bounded by $max[0, FIT_X - FIT_Z]$.

We now prove property (ii): if $S \rightarrow Z_{past} \rightarrow Y_{pres}$ is a Markov chain (i.e. $I(S;Y_{pres}|Z_{past}) = 0$) then $cFIT_{X|Z} = 0$). If $S \rightarrow Z_{past} \rightarrow Y_{pres}$ is a Markov chain, that is $P(S;Y_{pres}|Z_{past}) = P(S|Z_{past})P(Y_{pres}|Z_{past})$, then $I(S;Y_{pres}|Z_{past}) = 0$ [73]. Using the information-theoretic chain rule [73] we can write:

$$\begin{aligned} I(S;Y_{pres}|Z_{past}) &= I(S;Y_{pres},Z_{past}) - I(S;Z_{past}) \\ &= I(Y_{pres};S,Z_{past}) - I(Y_{pres};Z_{past}) \end{aligned} \tag{3.S39}$$

Therefore $I(S; Y_{pres}|Z_{past}) = 0$ implies $I(S; Y_{pres}, Z_{past}) = I(S; Z_{past})$ (meaning that all PID atoms that are a subpart of $I(S; Y_{pres}, Z_{past})$, but not of $I(S; Z_{past})$, are zero) and also $I(Y_{pres}; S, Z_{past}) = I(Y_{pres}; Z_{past})$ (meaning that all PID atoms that are a subpart of $I(Y_{pres}; S, Z_{past})$, but not of $I(Y_{pres}; Z_{past})$, are zero). In particular, in eqs. 3.S30 all atoms are computed on collections preceding (according to eq. 3.S8) collection $\{Y_{pres}Z_{past}\}$, meaning that, due to eq. 3.S11, they are all a subcomponent of $I(S; Y_{pres}, Z_{past})$. However, among these atoms, only the collection in $I_\partial(S; \{X_{past}\}\{Y_{pres}\}\{Z_{past}\}_{(4)})$ precedes collection $\{Z_{past}\}$ on this decomposition and, therefore, is a subcomponent of $I(S; Z_{past})$. Since in our case $I(S; Y_{pres}, Z_{past}) = I(S; Z_{past})$, the other two atoms on the right-hand side of 3.S30 are zero. Thus, if $S \to Z_{past} \to Y_{pres}$ is a Markov chain, the following identity holds for the first atom in FIT definition $I_\partial(S; \{X_{past}\}\{Y_{pres}\}_{(3)}) = I_\partial(S; \{X_{past}\}\{Y_{pres}\}\{Z_{past}\}_{(4)})$. Similarly, in eqs. 3.S31 all atoms are computed on collections preceding collection $\{SZ_{past}\}$ (meaning that, they are a subcomponent of $I(Y_{pres}; S, Z_{past})$). However, among these atoms, only $I_\partial(Y_{pres}; \{X_{past}\}\{S\}\{Z_{past}\}_{(4)})$ precedes collection $\{Z_{past}\}$ on this decomposition and, therefore, is a subcomponent of $I(Y_{pres}; Z_{past})$. Since in our case $I(Y_{pres}; S, Z_{past}) = I(Y_{pres}; Z_{past})$, the other two atoms on the r.h.s. of 3.S30 are zero. Thus, if $S \to Z_{past} \to Y_{pres}$ is a Markov chain, the following identity holds for the second atom in FIT definition $I_\partial(Y_{pres}; \{X_{past}\}\{S\}_{(3)}) = I_\partial(Y_{pres}; \{X_{past}\}\{S\}\{Z_{past}\}_{(4)})$. Altogether, we found that, if $S \to Z_{past} \to Y_{pres}$ is a Markov chain, the two atoms appearing in FIT definition (eq. 4.S5) are exactly equal to the two atoms between which we minimize to remove the effect of Z from FIT_X in eq. 3.S32, proving that in this scenario $cFIT_{X|Z} = 0$.

3.8.1.5 PID decomposition of DFI

We next use PID to examine a previously introduced measure of the information about a specific stimulus feature S flowing from X to Y, called Directed Feature Information (DFI) [183].

This measure was defined by reasoning to first consider TE between X and Y as a measure of the overall information transmitted from X to Y and the to subtract out from it the information that is not due to changes in the value of the stimulus feature. The latter was estimated as $TE(X \to Y|S)$, the value of TE conditioned on the stimulus feature, that is the expected value of the TE when it is conditioned on the value of a particular stimulus feature. The reasoning of [183] is that the conditioning removes information not related to variations of the stimulus feature, and that thus $TE(X \to Y|S)$ quantifies the amount of information transferred from X to Y that is not related to the variations in the stimulus feature. With this reasoning, the authors of [183] defined the DFI to measure stimulus-feature specific information transfer by subtracting out from the total information their estimate of the one that is not related to variations in stimulus features [183]:

$$DFI(X \to Y) = TE(X \to Y) - TE(X \to Y|S) \tag{3.S40}$$

The authors of Ref [183] showed that DFI is equivalent to the difference between the sum of the information about S that each of X_{past} and Y_{pres} individually carry, minus the information about S jointly carried by X_{past} and Y_{pres}, with all the information quantities conditioned on Y_{past}:

$$DFI(X \to Y) = I(S; X_{past}|Y_{past}) + I(S; Y_{pres}|Y_{past}) - I(S; X_{past}, Y_{pres}|Y_{past}) \quad \text{(3.S41)}$$

The difference between information individually carried and information jointly carried is often referred to as co-information [100, 103]. This measure of co-information has been used in the literature as a measure of the net effect of redundancy and synergy and it indicates prevalent redundancy when positive and prevalent synergy when negative [87, 89]. In this rewriting, DFI has some similarities with FIT, in that it uses a measure of redundancy (although conflating synergy and redundancy) between stimulus information in the past of X and in the present of Y, as well as a discounting, by conditioning, of the past activity of Y.

Previous work on PID has shown that co-information can be expressed as the difference between two non-negative pieces of information which properly quantify synergy and redundancy [100, 103]. Therefore a simple difference between DFI and FIT is that DFI possibly also includes terms of synergy between X_{past} and Y_{pres} than should not be included in a definition of transmission of feature information from X to Y. Moreover, given that DFI conditions on the past activity of Y rather requiring uniqueness with respect to the past feature information of Y (as in FIT), it does not isolate information in the present activity of Y that has not been present before in Y. To understand better the consequences of these facts in terms of the difference between DFI and FIT, we reformulated DFI as a sum of the partial information terms from the PID, as follows:

$$\begin{aligned} DFI_{X\to Y} &= I(S; X_{past}|Y_{past}) + I(S; Y_{pres}|Y_{past}) \\ &\quad - I(S; Y_{pres}, Y_{past}, X_{past}) + I(S; Y_{past}) \\ &= I_\partial(S; \{X_{past}, Y_{past}\}\{Y_{pres}, Y_{past}\}) \\ &\quad + I_\partial(S; \{X_{past}, Y_{past}\}\{Y_{pres}, Y_{past}\}\{X_{past}, Y_{pres}\}) + \\ &\quad + I_\partial(S; \{X_{past}\}\{Y_{pres}, Y_{past}\}) + I_\partial(S; \{Y_{pres}\}\{X_{past}, Y_{past}\}) \\ &\quad + I_\partial(S; \{X_{past}\}\{Y_{pres}\}) - I_\partial(S; \{X_{past}, Y_{pres}\}) - I_\partial(S; \{X_{past}, Y_{past}, Y_{pres}\}) \end{aligned} \quad \text{(3.S42)}$$

Note that in the above expression all terms involving pieces of redundant information are positive and those only involving synergistic information are negative. Thus this decomposition of DFI demonstrates that it is the linear combination of (mostly) redundant information terms appearing with a positive sign and synergistic information terms appearing with a negative sign. This explains why, as a result of not separating redundancy from synergy, DFI can be negative and difficult to interpret as information about a stimulus feature transmitted from X to Y.

The fact that DFI can become negative also shows that using $TE(X \to Y|S)$ to remove from the total transmitted information the one not about the stimulus feature S (as done in DFI, see [183]) is incorrect. This is because $TE(X \to Y|S)$

does not really quantify the information from X to Y which is not about S, as conceptualized in Ref [183]. It actually quantifies the information transmitted on average within each feature condition. This can overestimate the information from X to Y which is not about S. In simple terms, when using data from the same feature conditions, some information sent from X to Y in this subset of data could be about the specific value of the feature in the considered set of trials. When the strength of communication about S between X and Y varies from one feature value to another, this overestimation may become even more severe, because in this case additional information about S is encoded synergistically within the network of X ad Y by their feature-dependent relationship [89].

3.8.1.6 Numerical computation of FIT and other information quantities

FIT and all other information theoretic quantities were computed from both simulated and real data by plugging into the corresponding equations the numerical evaluation of the response probabilities from the data. We computed by the response probabilities discretizing neural activity into a number R of equipopulated bins [170] and then computing empirically the frequency of occurrence of each binned response across all available trials.

In Table 3.S1 we summarize the number of bins we used to discretize neural activity for each figure in the paper. In Section 3.8.2.5 we study the accuracy of the FIT and TE estimates with the number of available trials and we show that the estimates of FIT and TE are accurate and unbiased for the number of bins and number of trials used for all analyses. However, in the code we provide to compute FIT and TE, we also implemented limited-sampling bias correction routines that can be used to obtain more accurate estimates when data are more scarce (see Section 3.8.2.5).

Number of bins	2	3	4
Figures	Fig.3 Fig.S8 Fig.S9 Fig.S11 Fig.S12B Fig.S13B Fig.S14D-E	Fig.2A,B,E Fig.S3 Fig.S4 Fig.S5 Fig.S7 Fig.S13A Fig.S14A-B	Fig.2C Fig.4 Fig.S10 Fig.S12A

Table 3.S1: Number of bins used to discretize neural activity for information-theoretic analyses of simulated and real data, for each main text and SM figure

3.8.1.7 Permutation-based non-parametric null hypotheses for FIT and TE

To test for significance of the information theoretic quantities, we used non-parametric permutation tests, described below.

To test for the significance of mutual information values about the feature of interest, we used established non-paramametric procedures [83, 101, 178]. We constructed surrogate datasets in which we destroyed any feature information by randomly permuting across trials the values of S, and then we recomputed information on the surrogate data to obtain a null-hypothesis distribution of null information values.

To test for the significance of FIT, we developed a permutation test in which we created surrogate data in which we preserve the feature information in the past of X and the present of Y while destroying the communication of this information between X and Y. We shuffled X within trials with the same value of the feature S, destroying any within-trial statistical relationship between the activity of X and the activity of Y at fixed values of S, and recomputed FIT on the surrogate data. This data shuffling preserves the marginal distributions between the feature and the past activity of X and between the stimulus and the present activity of Y, thereby preserving the information about the stimulus that each carries. However, it destroys the within-trial statistical relationship at fixed stimulus between X and Y that would be present if X sends stimulus information to Y. Because FIT depends on $P(X_{past}, Y_{pres})$ (see mathematical proof in Section 3.8.1.3), the values of FIT on the permuted data will be smaller than the ones on the original data whenever there is direct within-trial communication of stimulus information between X and Y, but will be similar to the value of the original data when there is no such direct within-trial communication. As shown by numerical simulations (see Section 3.8.2.6) the so generated null hypothesis distribution of FIT values when X and Y encode but do not communicate stimulus feature information is more conservative (see Fig. 3.S7C) than the simpler one that would be obtained by a permutation test destroying all information about S in X and Y by randomly permuting S across all trials, as for the mutual information quantities above. (This permutation test would implement the idea that is no stimulus-feature information is present, it cannot be transmitted). However, in limiting cases in which the stimulus information in the neural data is absent, we found it numerically better to perform this second random permutation of the label of S across trials (because more possible independent data permutation are available in this second permutation, which therefore may have some advantages in the case of zero or negligible stimulus information, see Fig. 3.S7E). To reduce the probability of false positives in such cases of no information present in the network, we computed and then intersected the two above describe possible permuted distributions by taking the element-wise maximum between the two distributions, and obtained a null distribution for FIT. (In practice, in real data and simulations with stimulus information present, the maximum of the two permuted values coincided in all simulations with the maximum of the first permutation. This is exemplified

in Fig. 3.S7C, in which for higher value of the parameter W_{ZY} some information about the stimulus is created in both X and Y in absence of communication between X and Y, the null hypothesis for FIT taking the maximum between the first and second permutation has values not only larger than the FIT value measures in the simulation, but also much larger values then the ones based on only shuffling S. In simulations with null information, the maximum value of the permuted data in each simulation could instead belong to either permutation.)

An identical procedure was applied to test for the significance of DFI. For TE, since by design the measure captures the total amount of information flowing from X to Y, we permuted the neural activity of the sender X across all trials.

Since in all simulations and real data analyses we wanted to test for the significance of the information values averaged either across simulations or participants, we computed the average over simulations and participants of the information values in each realization of the random permutation and we used this distribution of null hypothesis of averaged values for testing the significance of the averaged information value. (To compute the null-hypothesis distribution, we generated 500 different realizations of the permuted average information for each test we conducted.)

In some analyses (e.g. Fig. 3C and 4B,C) we had to identify the cluster of post-stimulus times and transmission delays for which FIT or TE were significantly different from zero (shown, e.g., in Figs. 3.S9G, 3.S10A,B). We individuated these clusters of points in the time-delay space using a cluster-based permutation test [177, 178] using as null hypothesis values those obtained from the permutation test described above. We computed the cluster forming threshold as the 99th percentile of information values in the surrogate data. We created information clusters in the original and shuffled datasets by summing together all adjacent information values above the cluster forming threshold. We then determined a null distribution for information clusters using the maximum cluster value from each shuffled dataset. Finally, we assigned significance to clusters in the original dataset if their value was larger than the 99th percentile of the clusters null distribution (p ¡ 0.01).

3.8.2 Details of simulations and and additional analyses of simulated data

3.8.2.1 Simulations of FIT and TE as a function of signal and noise transmission

This section pertains to the description of Fig. 2A-B of the main text.

The goal of the first simulation (whose results are reported in Fig. 2A) was to evaluate the dependence of FIT and TE on stimulus-feature-related and -unrelated transmission. The goal of the the second simulation (whose results are reported Fig. 2B) was to test the ability of FIT and TE to localize in time the stimulus-feature-related information transmission. The setting of both simulations was identical and is described in the following.

We simulated 500ms of activity of activity of X and Y, in time steps of 10ms. The sending region X encoded a stimulus S over time and transmitted stimulus-feature-related and -unrelated activity to the receiver Y with a given temporal lag δ. The stimulus feature S being encoded by X and transmitted to Y was an integer (between 1 and 4) drawn independently and uniformly in each trial (500 trials per stimulus for each of the 50 simulations). The activity of the sender was a two-dimensional variable with one feature-informative X_{stim} and one feature-uninformative component X_{noise}. The stimulus-feature-informative dimension had a temporally-localized feature-dependent bump in the activity (from 200 to 250ms) and multiplicative Gaussian noise

$$X(t)_{stim} = S(t)(1 + \mathcal{N}(0, \sigma_{stim})) \tag{3.S43}$$

where $S(t)$ was a function equal to the value of the the stimulus $s \in [1, 4]$ during the time window $[200, 250]ms$ and was zero outside of this window. The presence of noise in $X(t)_{stim}$ was needed to test for the impact of within-trial encoding of S in X on the within-trial encoding of S in Y, at fixed values of S (i.e., when X encodes S incorrectly, also Y encodes S in a similar way). If $X(t)_{stim}$ encoded the stimulus perfectly (no noise in $X(t)_{stim}$, therefore $X(t)_{stim} = S$ for $t \in [200, 250]ms$), it would be impossible to determine whether Y is receiving stimulus information from X or directly from S. We choose the noise in the stimulus-feature-informative dimension to be multiplicative because it made it a more challenging scenario for FIT. In fact using multiplicative noise X developed a stimulus-dependent noise in the encoding of S. The stimulus-dependent noise in the encoding of S leads to stimulus-dependent within-trial correlations between X and Y, which potentially induces synergies in the encoding of S in X and Y [89]. Since FIT computes information transmission by identifying a component of redundant information between the past of X and the present of Y, using simulations that have both redundancy between X and Y induced by information transmission and synergy between X and Y induced by stimulus-dependent amount of noise encoded and transmitted (using this kind of multiplicative noise), makes it potentially harder for a measure of feature-specific information transmission to separate out the redundant information that was transmitted. In fact, we will see that measures that do not separate well redundancy and synergy, such as DFI, will suffer under such conditions (leading to negative values of transmitted information (Fig. 3.S15A), whereas FIT seems to work well even under this condition because it uses PID to only include redundant time-lagged information about S in X and Y, discarding synergy. (However, we found similar results for TE and FIT by replacing the multiplicative noise with an additive noise in Eq. 3.S43 (Fig. 3.S3)). The stimulus-feature-unrelated component was, at any time point, a zero-mean Gaussian noise $X(t)_{noise} = \mathcal{N}(0, \sigma)$. The activity of the receiver Y was the weighted sum of X_{stim} and X_{noise} with a delay δ, plus a Gaussian noise: $Y(t) = W_{stim}X_{stim}(t - \delta) + W_{noise}X_{noise}(t - \delta) + \mathcal{N}(0, \sigma)$. The delay δ was chosen randomly in each repetition from a uniform distribution in the range between $40ms$ and $60ms$, in steps of $10ms$. Therefore, across repetitions of the simulation, Y received information from X only in the time window $[240, 310]ms$. In all simu-

lations we set a standard deviation $\sigma = 2$ for the additive Gaussian noise in X_{noise} and Y, and a standard deviation $\sigma_{stim} = \frac{\sigma}{5} = 0.4$ for the multiplicative Gaussian noise in X_{stim}.

In the first simulation, we computed FIT and TE at the first time instant in which Y received information from X ($t = 200ms + \delta$), and at the ground truth delay δ, for all combinations of W_{stim} and W_{noise} in the range between 0 and 1, in steps of 0.1. In the second simulation we set $W_{stim} = 0.5$ and $W_{noise} = 1$ and computed FIT and TE at all time points, in a rage of communication delays between 0 and 100ms, and averaged their values over delays to obtain temporal profiles of transmitted information.

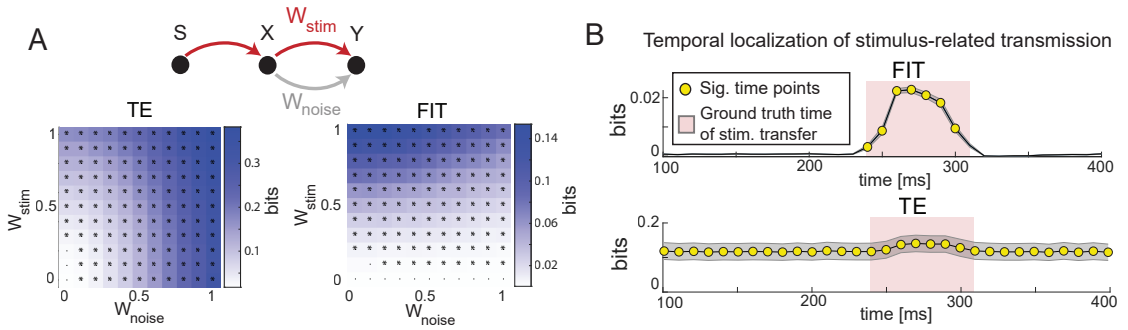

Figure 3.S3: Further tests of FIT on simulated data with additive noise in X. This figure is similar to Fig. 2 of the main text except that the results are now obtained with additive (rather than multiplicative) noise in X A) FIT and TE as function of stimulus-feature-related (W_{stim}) and -unrelated (W_{noise}) transmission strength. * indicate significant values (p ¡ 0.01, permutation test) for the considered parameter set. B) Dynamics of FIT and TE in a simulation with time-localized stimulus-feature-information transmission. The red area shows the window of stimulus-feature-related information transfer. Yellow dots show time points with significant information (p ¡ 0.01, permutation test). Results plot mean (lines) and SEM (shaded area) across 50 simulations (2000 trials each).

3.8.2.2 FIT can detect feature specific information flow even with overlapping time courses of stimulus information

This section pertains to the description of Fig. 2C-E of the main text.

One often used method to infer hierarchical flow of information across ares is to consider the timing of neural activation or of stimulus selectivity of activity across brain regions [197]. However, this method is neither necessary nor sufficient to determine real communication. On the one hand time lagged information selectivity between two regions may arise in absence of communication for example if the two regions received a partly shared input signal with a different delay. On the other hand, as we will exemplify in this section, real features-specific communication between two brain regions could take place even without detectable differences in timing of information across the considered regions. The purpose of this subsection

is to illustrate that this can happen and also to show that in such case FIT has power to discriminate between cases in which feature-specific information flow does or does not take place. We will show that this is because FIT can assess that in cases of real communication the format of information encoding is the same in the past activity of the sender and in the present activity of the receiver.

We simulated a scenario where a sending regions X encodes and transmits to a receiving region Y information about an integer stimulus-feature S (ranging between 1 and 4). Importantly, the past of Y and the present of X carry the same amount of feature information of the past of X and the present of Y, but encode the information with different formats.

The *feature encoding format* of each region is determined by the encoding function $f(S)$ controlling the average response of each region to individual stimulus values. We simulated responses with three different encoding functions:

$$\begin{aligned} f_1(S) &= 1 + \delta[0, 1, 2, 3] \\ f_2(S) &= 1 + \delta[1, 0, 2, 3] \\ f_3(S) &= 1 + \delta[0, 1, 3, 2] \end{aligned} \tag{3.S44}$$

where δ is a parameter controlling the separation of the average responses to different stimuli, and therefore the amount of feature information carried by each region at a specific time point. We set $\delta = 1$ in all simulations. The three encoding functions $f_1(S)$, $f_2(S)$, $f_3(S)$ are depicted in Fig.2C in blue, green and red, respectively. The encoding function determined the feature values that each region preferentially encodes at each specific time point. Specifically, due to the presence of additive Gaussian noise, regions were most informative (according to Eq. 3.S14) about stimulus values for which the response was either minimum (i.e. equal to 1 in eq. 3.S44) or maximum (i.e. equal to $1 + 3\delta$ in eq. 3.S44). Indeed, activity distributions in response to these stimulus values were less overlapped with activity in response to other stimuli. For example, regions encoding the stimulus as $f_1(S)$ would carry high specific information about stimulus values 1 and 4, and low specific information about stimulus values 2 and 3. On the other hand, regions encoding the stimulus as $f_2(S)$ would carry high specific information about stimulus values 2 and 4, and low specific information about stimulus values 1 and 3. Therefore, since the I_min measure quantifies redundancy as the overlap in the distributions of specific information across individual values of target variable, the responses of two regions X and Y would be maximally redundant if they encoded the feature with the same encoding format (e.g., $f_X(S) = f_Y(S) == f_1(S)$), partially redundant if they both carried high specific information about one stimulus value (e.g., $f_X(S) == f_1(S)$ and $f_Y(S) == f_2(S)$) or minimally redundant if they carried high specific information about different pairs of stimulus values (e.g., $f_X(S) == f_2(S)$ and $f_Y(S) == f_3(S)$).

In our simulation, X and Y activity at different time points was described by the following set of equations:

$$\begin{aligned} X_{past} &= f_1(S) + E_1 \\ Y_{pres} &= X_{past} \\ Y_{past} &= f_2(S) + E_2 \\ X_{pres} &= f_3(S) + E_3 \end{aligned} \tag{3.S45}$$

where E_1, E_2, and E_3 are additive Gaussian noise with standard deviation equal to σ. Importantly, real feature transfer only occurs in the $X \to Y$ direction, causing the past of the sender and the present of the receiver to encode the feature with the same format. Since σ was equal for all noise terms in eq. 3.S45, both X and Y carried the same amount of stimulus-feature information in the past and in the present, removing any contribution to FIT due to time-lagged information levels in the sender and the receiver region. We measured FIT in the two directions ($X \to Y$ and $Y \to X$) for different levels of noise σ. By changing σ we controlled the $SNR = \frac{\delta}{\sigma}$ in both past and preset activity of X and Y by changing σ. We repeated the simulation 100 times for each SNR value ranging between 0.05 and 1 with a precision of 0.05. We measured the FIT significance in the two direction using the permutation test described in section 3.8.1.7.

We found that, because FIT could correctly detect that the format of information representation of S in the present of Y was equal to that of the past of X but different to that of the past of Y (Fig. 3.2D), and that feature information flowed from X to Y (Fig. 3.2E).

3.8.2.3 Simultaneous transfer of information about more than one feature

We performed a simulated study of how FIT performs when studying neural system that encode and transmit more than one feature (Fig. 3.S4).

We simulated two independent features (e.g. of a sensory stimulus) S1,S2 simultaneously encoded in a brain region X and transmitted to a brain region Y. In the simulation, S_1 is more strongly and encoded and transmitted than S_2. The equation for simulating the data are as follows:

$$X = S_1 + DS_2 + E_x \tag{3.S46}$$

where S_1, S_2 are independent binary variables (values equal to ± 1), E_x is Gaussian noise with standard deviation equal to 1, and Y equals X with a time lag, plus independent Gaussian noise with standard deviation equal to 1.

We simulated the system with different values of D (the strength of encoding and transmission of $S2$ relative to $S1$). We found (Fig. 3.S4) that FIT identifies correctly that both features are transmitted, and ranks correctly the features about which most information is transmitted. FIT also identified correctly the limiting case ($D = 1$) in which both features are encoded and transmitted with equal strength.

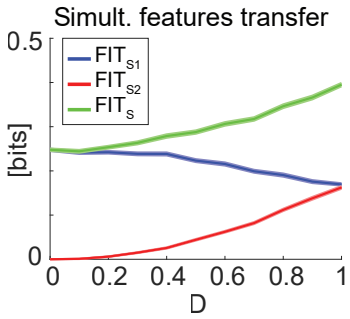

Figure 3.S4: Simulation of a system that encodes independently two features $S1$ and $S2$ in the activity of a brain region X and transmits them to another brain region Y. We compute FIT about each of the two stimulus features $S1$, $S2$ or their combination $S = (S_1, S_2)$ as a function of the parameter D describing the strength of encoding and transmission of S_2 with respect to S_1. We plot mean ± SEM over 50 simulations

3.8.2.4 Simulations of bidirectional transmission between X and Y

Here we describe the simulations whose results are presented in Fig. 3.S5.

To further investigate the ability of FIT to determine the direction of stimulus-feature information flow, we simulated a scenario with bidirectional (back and forth) communication between X and Y with stimulus-feature-related transfer from X and stimulus-feature-unrelated transfer from Y to X (Fig. 3.S5A).

In brief, both X and Y received information directly from a feature-information-sending region S. The region X received stimulus information from S early on (between 50 and 90 ms) and Y received stimulus information from S at a later time (between 110 and 150 ms). X sent its entire activity to Y (therefore communicating its stimulus information when it became available). Y instead only sent to Y a part of its activity that did not carry stimulus information. The details of how this was achieved are reported below.

We simulated 180ms of activity of X and Y, in steps of 1ms. The stimulus feature S being encoded and transmitted from a stimulus region S to X and Y was an integer (between 0 and 3) drawn independently and uniformly in each trial. The activity of X was one-dimensional. The activity of Y was two-dimensional. Both dimensions of Y (Y_+ and Y_-) were generated with a Poisson process whose mean was modulated over time by a Gaussian bump (whose amplitude was equal to the stimulus-feature value S) in the time window $[110, 150]$ms, plus an additive Gaussian noise and time-lagged readout of X activity (with a X to Y transmission delay $\delta_{xy} =$ 10ms). Importantly, Y_+ encoded the stimulus as a positive Gaussian bump and Y_- encoded the stimulus as a negative Gaussian bump. Therefore, the entire activity of Y, i.e. the sum of the two components, $Y_{noise} = Y_+ + Y_-$ carried no information about the stimulus, and the difference of the two components $Y_{stim} = Y_+ - Y_-$ carried all the stimulus information in Y. X was a Poisson process whose mean was positively modulated over time by a Gaussian bump - whose amplitude was modulated by

the stimulus - in the time window $[50, 90]$ms, plus an additive Gaussian noise and time-lagged readout of the entire activity of Y Y_{noise} (with a Y to X transmission delay $\delta_{yx} = 15$ms).

We measured FIT at each time step of the simulation over a range of communication delays and averaged the resulting information values over delays to obtain temporal profiles of information transmission. We found that FIT correctly captured the flow of information about S between X and Y that we put by design into these simulated data. FIT revealed that there was a significant stimulus-feature-related information transmission from X to Y, that was temporally localized in the actual "ground-truth" $[60, 100]ms$ window in which Y received stimulus information, and no significant stimulus-feature-information transmitted from Y to X (Fig. 3.S5B).

We also used these simulations to test the performance of the Directed Feature Information, DFI [183], using the same analysis pipeline described here for FIT. Results are discussed in Section 3.8.4.2.

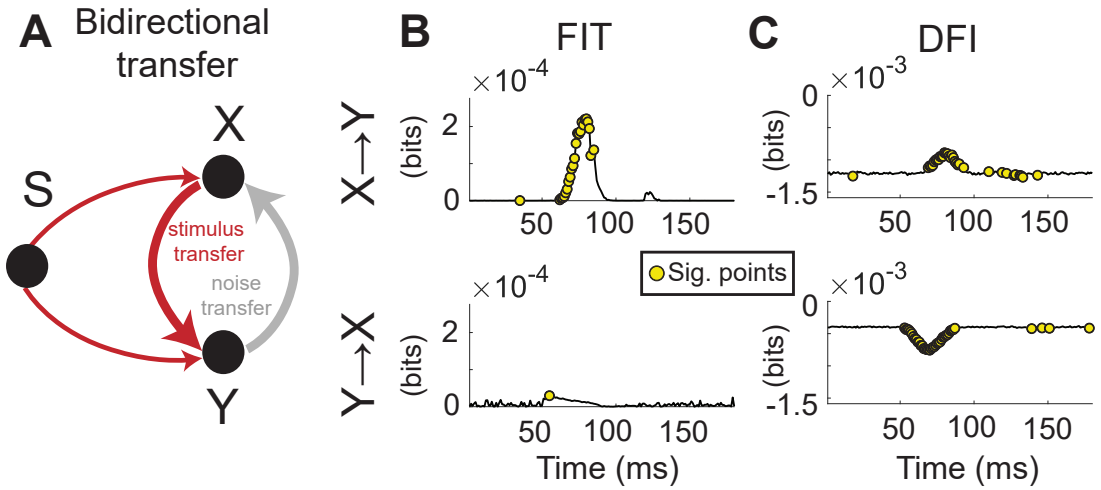

Figure 3.S5: Simulations of performance of FIT and other measures in a case of transfer of stimulus-feature-related and stimulus-unrelated information in different directions. A) Schematic of the simulation. A stimulus node S provides partly complementary information about a stimulus feature to X (in the 50-90ms interval) and to Y (in the 110-150ms interval). X transmits stimulus information to Y. Y instead has different components of activity and it projects to X only the component of its activity that is stimulus-unrelated. In other words, we have stimulus-feature-information transfer from X to Y and noise transfer from Y to X). B) Results of the analyses of this simulated activity using FIT (left panels) and DFI (right panels). Gray lines plot the value of these quantities Yellow dots plot time points in which the measure was significantly different from null (permutation test of Section 3.8.1.7; p<0.01)

.

3.8.2.5 Limited-sampling bias of FIT and TE

Information-theoretic quantities are know to suffer from a systematic error (called limited sampling bias) when the probabilities used to compute them are estimated from a limited number of experimental trials [179]. While the limited bias of Shannon information quantities such as TE have been studied well and has been shown to be inversely proportional to the number of available trials and directly proportional to the number of bins used to discretize the data [179, 198], those of FIT remain to be investigated.

Therefore here we simulated a simple scenario to study how FIT and TE scale with the number of trials available in the dataset and the number of bins of the discretized activity. In these simulations X encoded a binary feature S with additive Gaussian noise. Y was equal to X with a time lag of 1 plus independent Gaussian noise (standard deviation of noise = 0.5).

We found (Fig. 3.S6) that FIT behaved much better than TE with the data size and number of bins. The correct value of FIT, which can be estimated from large numbers of trials, was achieved already with smaller number of trials than for TE. We found that accurate calculations of FIT are possible with the number of trials available in empirical datasets (Fig. 3.S6); for comparison FIT calculations with real and simulated data in this paper were done with 2-4 discretization bins, see Table 3.S1). Our understanding is that the better scaling and sampling properties of FIT with respect to Shannon Information quantities arise because FIT considers a PID part of the total information which has lesser bias compared to other parts of the total information. Given that the PID atoms of FIT do not contain synergistic terms, this is in line with previous work [199] showing that synergistic components of information have much larger limited sampling bias, and that information quantities that do not include synergistic components have much better sampling properties than full multivariate Shannon information quantities. Thus, FIT can be computed from the datasets in which Shannon information measures typically applied to neural data. We applied a widely used bias correction technique, called the Quadratic Extrapolation [81, 179]. This method is based on subtracting the bias estimated from a second-order polynomial fitting of the dependence of the estimated quantity on sub-samples of the available data. We found (Fig. 3.S6) that this bias subtraction technique was helpful in further improving the estimate of information (reducing the limited sampling bias) in cases of very low numbers of trials available. This bias correction technique is made available in the software we provide for both FIT and TE.

3.8.2.6 Simulation tests of the significance of FIT and cFIT

In this Section we describe the simulations and results presented in Fig. 3.S7.

In this set of simulations, we first evaluated the effectiveness of the permutation-based non-parametric tests for FIT in a difficult scenario in which X and Y independently encode stimulus-feature information with a temporal lag, but no actual

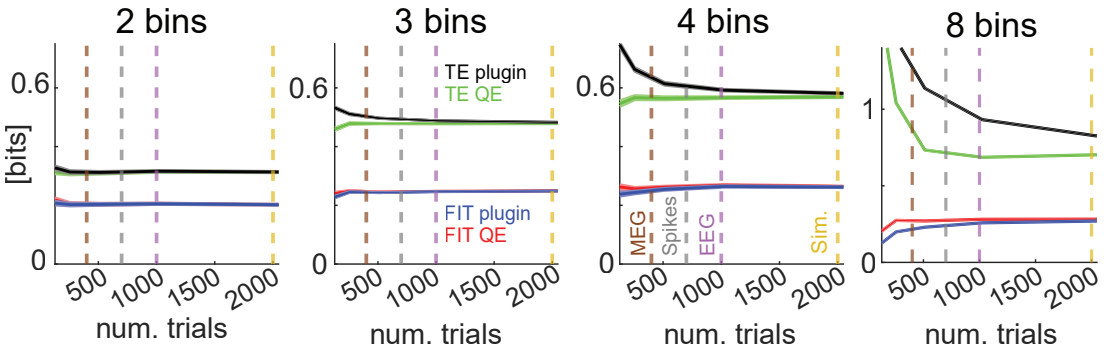

Figure 3.S6: FIT and TE as a function of the number of simulated trials used to compute them. Vertical lines: number of trials used for the analyses of real and simulated brain data in this paper. "Plugin": plugging empirical probability histograms into FIT, TE eqs. "QE": adding Quadratic Extrapolation bias correction. For all analyses in the rest of the paper we used 2-4 discretization bins (see Table 3.S1). Results in each panel are ploted as mean ± SEM over 100 simulations

.

communication occurs between them. We then evaluated the performance of cFIT in measuring the unique contribution of X in sending feature information to Y in presence of an alternative information route through a third node Z sending information to Y.

We addressed both questions using a the following simulation setup. We performed a simulation in which two senders (X and Z) both transmitted stimulus-feature information to Y (Fig. 3.S7B). X encoded the stimulus-feature linearly and Z non-linearly so that they carried partially different information about S. This is important because a good measure quantifying the unique contribution of X (but not Z) in the transmission of information to Y should capture that, even if the total amount of feature information transmitted from Z to Y is stronger than the one transmitted from X to Y, there can still be a different component of information that is uniquely transmitted by X and not by Z. We simulated 500ms of activity, in time steps of 10ms. The encoded and transmitted stimulus feature S was a stimulus-intensity integer value (0 to 3) drawn independently and uniformly in each trial (500 trials per stimulus). The activity of X had a temporally-localized square bump between 200 to 250ms whose amplitude depended linearly on S, with multiplicative Gaussian noise. Activity of Z had a temporally-localized square bump (from 200 to 250ms) encoded with a different format with respect to X, and multiplicative Gaussian noise. Specifically, X encoded the stimulus feature $S = (0, 1, 2, 3)$ with the encoding function while Z encoded S with the encoding function $S = (1, 0, 3, 2)$. In this way, X carries more specific information than Z (see eq 3.S14) about $S = 0, 3$ and Z carries more specific information than X about $S = 1, 2$, therefore both X and Z carry some unique information about S. Activity of Y was the weighted sum of X and Z with a temporal lag, plus additive Gaussian noise: $Y(t) = W_{xy}X(t-\delta) + W_{zy}Z(t-\delta) + \mathcal{N}(0, \sigma)$. The delay δ in the transmission

of information from X to Y was chosen in each repetition of the simulation randomly from a uniform distribution in the range between $40ms$ and $60ms$. We computed FIT and cFIT at the first time instant in which information in Y was received from X and Z ($t = 200ms + \delta$) using to define past activity the ground-truth delay δ actually used in that simulation. We set a standard deviation $\sigma = 2$ for the additive Gaussian noise in Y, and a standard deviation $\sigma_{stim} = \frac{\sigma}{5} = 0.4$ for the multiplicative Gaussian noise in X and Z.

Tests of significance of FIT accounting for the possible existence of encoded feature information in the absence of transfer of it across regions We first addressed the first question, that is how to deal with confounding scenarios where X and Y independently encode feature information with a temporal lag, but no actual communication occurs between them.

We studied how the FIT from X to Y depended on the strength of feature-related transmission from Z to Y W_{zy} when no stimulus-feature-information was transmitted from X to Y ($W_{xy} = 0$). We found that FIT from X to Y increased with W_{zy}, since X and Y carried redundant information about S with a temporal lag. However, FIT was always non-significant (Fig. 3.S7C) using the permutation test described in Section 3.8.1.7, since there were no within-trial correlations between the encoding of S in X and the time-lagged encoding of S in Y. This proves that, even if Z was not measured, the permutation test we provided for FIT can correctly rule out confounding scenarios where X and Y encode S with a temporal lag but with no actual communication occurring between X and Y (see Section 3.8.1.7 and 3.8.2.6).

Simulations testing cFIT in the presence of information transfer through an alternative route involving a third region Z We next addressed the second question, that is how to evaluate the unique contribution of X in sending feature information to Y in presence of an alternative information route through a third node Z sending information to Y. We studied how FIT from X to Y and cFIT from X to Y conditioned on the feature information in Z depended on the simultaneous transmission of feature information from X to Y and from Z to Y. To do this, we computed FIT and cFIT for all combinations of W_{xy} and W_{zy} in the range between 0 and 1, in steps of 0.1. We found that FIT grew both as a function of W_{xy} and of W_{zy} and was significant as soon as some information was transmitted from X to Y ($W_{xy} > 0$; Fig. 3.S7D, left). On the contrary, cFIT increased only as a function of W_{xy} and decreased with W_{zy}, correctly removing from the FIT from X to Y the feature information that was routed to Y through Z (Fig. 3.S7D, right). Crucially, cFIT did not simply subtract from FIT through X the FIT through Z, but it only removed the amount of feature information that was redundantly transmitted by X and by Z to Y. Indeed, since X and Z transmitted partially different feature information to Y, we have that cFIT was still significant for many combinations of

parameters where $W_{zy} > W_{xy}$ (Fig. 3.S7D, right) and, therefore, FIT through X was larger than FIT through Z (not shown).

3.8.2.7 Simulation studies of how FIT and TE are affected by the mixing of sources

In real electrophysiological recordings, it is possible that that separation and reconstruction of the underlying neural sources is imperfect, due to issues such as for example field spread or common referencing. As a result, electrophysiological recordings from different brain regions may contain, with different weights, a mixture of sources. It has been proposed that such source mixing may affect measures of communication between brain areas [200]. Here, we examine the effect of this source mixing in FIT and TE measures.

We simulated source mixing in different proportions in the sender X and the receiver Y. In our simulations we assumed that (as it is expected to be the case in real brain data) the mixing is instantaneous (i.e. sources are mixed with zero lag) and with a proportion of source sharing in X,Y that is stable across time.

We first simulated a source Z (informative about a stimulus feature S) shared between X and Y with a different proportion A:

$$\begin{aligned} X &= Z(s) + E_x \\ Y &= AZ(s) + E_y \end{aligned} \tag{3.S47}$$

with E_x, E_y independent Gaussian noise. We controlled the SNR of X and Y by changing A (which sets the relative level of stimulus-feature signal in X,Y) and fixing noise standard deviation to 1. On this model FIT and TE had spurious positive values (Fig. 3.S8B). We used the permutation test introduced in Section 3.8.1.7, testing for spurious values induced by X,Y covariations due to feature-signal sharing. We found that this test correctly ruled out as non-significant FIT and TE values generated only by source sharing with no real transmission (Fig. 3.S8B). Importantly, analysis of this model also showed that with instantaneous source mixing (and notably under the assumption that recording noise is constant over the time of the trial) the ratio between stimulus-feature info in X and Y is constant in time (Fig. 3.S8A). This gives a useful heuristic: while the finding that the feature information time courses of two individual regions that overlap in time cannot be used to rule in or out communication of information about the feature between the two areas (see 3.8.2.2 and Fig 2D-E), different timecourses of stimulus-feature info in X vs Y cannot be easily explained by instantaneous source mixing. We measured all real-data FIT in cases with a delay in stimulus-feature info latencies between X and Y (X to Y info latencies: MEG: 17-35ms between V1 and higher areas, Fig. 3.S9B. EEG: 25ms across hemispheres (Fig 4B). Spike data: 20ms from thalamus to cortex, Fig. 3.S11). Overall, these findings speaks against dominant mixing of a feature-informative source in our analyses.

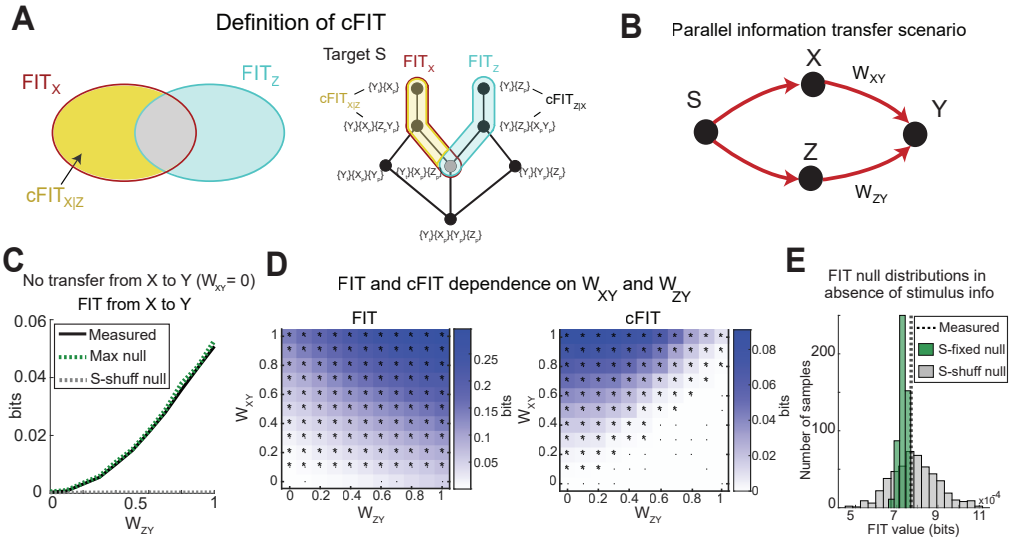

Figure 3.S7: Simulation tests of the significance of FIT and cFIT. A) Schematic of the cFIT definition. Left: intuitive definition with set-theoretic diagrams. Right: breakdown of FIT_X (in red) and FIT_Z (in light blue) into finer information atoms considered in the definition of $cFIT_{X|Z}$ (atom that can be part of $cFIT_{X|Z}$ are indicated in yellow). Only atoms of FIT_X, FIT_Z and $cFIT_{X|Z}$ belonging to the PID having S as target variable are shown. FIT_X is the feature information transmitted from X to Y, FIT_Z is the one transmitted from Z to Y, $cFIT_{X|Z}$ is the cFIT from X to Y conditioned on the stimulus-feature information of Z. B) Schematic of the scenario implemented in the simulations: both X and Z transmit feature information to Y. C) FIT dependence on the amount of stimulus-feature information transmitted from Z to Y (W_{ZY}) even when simulating a case ($W_{XY} = 0$) in which there was no within-trial transmission from X to Y. FIT grows with W_{ZY}, but its value is always non significant using the permutation null hypothesis described in Section 3.8.1.7. The dashed green line shows the 99th percentile of the FIT null hypothesis distribution described in Section 3.8.1.7. For comparison, the dashed gray line shows the 99th percentile of the null-hypothesis distribution that would have been obtained simply shuffling S across all trials. The fact that the latter remains so low across all values of W_{zy} highlights the need of using a shuffling procedure that preserves the stimulus-feature information in the individual nodes, as we did in this paper and described in Section 3.8.1.7 D) FIT and cFIT as function of feature-related transmission from X to Y (W_{XY}) and from Z to Y (W_{ZY}). * indicate significant values ($p < 0.01$, permutation test) for the considered parameter set. In Panels C,D, results plot mean across 50 simulations (2000 trials each). E) Example of shuffled distributions in a case in which there is not stimulus-feature information in Y. While the null hypothesis values of permuting X at fixed S give much more conservative and effective null hypothesis values when the analysed network has stimulus-feature information across the nodes (see panel C), in specific cases of no feature information in parts of the network it may be safer to consider also the permutation of S across all trials, as this has more available independent permutations form the data and thus gives wider distributions. The example is with the simulations performed in Fig. 2A, for the set of parameters ($W_{stim} = 0, W_{noise} = 0.6$).

Finally, we simulated the case with real FIT between two "pure" signals Z_1 and Z_2 that are unevenly mixed in the measured X,Y:

$$\begin{aligned} X &= Z_1 + AZ_2 + E_x \\ Y &= Z_2 + BZ_1 + E_y \end{aligned} \tag{3.S48}$$

Since adding a new feature-informative channel (Z_1 to Y and Z_2 to X) increases the stimulus-feature information in X,Y, we set the standard deviation of independent Gaussian noise E_x,E_y to equalize SNR of X and Y across the simulated parameters space. We found (Fig. 3.S8C) that mixing ($A, B > 0$) reduced FIT and TE compared to the pure case ($A = B = 0$). However, the correct direction of information transfer was always detected for all mixtures. Thus, FIT is reasonably conservative and robust to this mixing.

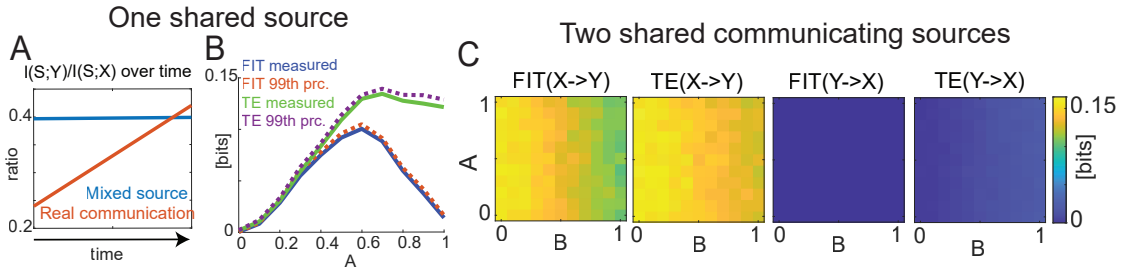

Figure 3.S8: Simulated tests of FIT and TE in the presence of source mixing. A-B: results of the null model with unequal sharing of one feature-informative source Z. A) Proportion of stimulus-feature info with only the null model (blue line, described in SM 3.8.2.7 and eq. 3.S47) and with real transmission of feature information $X \to Y$ added to it (red line). B) FIT and TE of null model (see 3.S47) fall below the permutation's test 99th perc significance for any proportion of sharing A. C) FIT and TE computed from a model with two stimulus-feature informative sources Z_1, Z_2 communicating $Z_1 \to Z_2$ and mixed into X and Y, vs mixing proportions A, B in X and Y (see 3.8.2.7 and eq. 3.S48). All panels plot results averaged over 50 simulations.

.

3.8.3 Details and further analyses of experimental data

3.8.3.1 MEG data

Behavioral task and MEG recordings We analyzed a publicly available MEG dataset [11], with source-reconstructed data available at [https://doi.org/ 10.6084/m9.figshare.12770366]. Full details (including details of approvals from Ethical Committees) are reported in the original publication and are briefly summarized here. The MEG data were recorded from the brain of participants while they performed a visual decision-making task. N=15 participants took part in the experiments, performing 4 experimental sessions each with on average 429 trials per session. At the beginning of each trial, a reference sample was shown

with a contrast of 50%. Each sample lasted 100ms and its contrast was drawn from a Gaussian distribution whose mean was randomly selected in each trial . At the end of each trial, participants reported whether the average contrast of the 10 sample gratings was higher or lower than the reference contrast. A staircase procedure was used to adjust the mean of the Gaussian distribution setting the average contrast of the 10 samples in each trial, by making trials harder (mean of the Gaussian closer to 50%) or easier (mean of the Gaussian further from 50%) depending on the behavioral performance of the participant until that moment in the experimental session. The staircase was set to obtain a behavioral performance of approximately 75% on each session. The Regions Of Interest (ROIs) in MEG source space used to identify signal from the considered brain areas were defined based on the atlas from Glasser and colleagues [201]. All ROIs were co-registered to individual structural Magnetic Resonance Imaging data. Source reconstruction was performed using LCMV beamformers based on leadfield matrices from 3-layer boundary element head-model (conductivity 0.3, 0.3, 0.006 S/m for scalp, brain, skull respectively) based on individual MRIs, using the covariance matrix (CM) of broadband data (275x275 sensors) and a regularization of 5% of CM. The source space was constrained to the cortical sheet with 4096 vertices per hemisphere, and source orientations chosen to maximize power at each vertex. To illustrate the spatial resolution of source reconstruction, in Fig. 3.S9A we plotted the correlation between LCMV spatial filters of neighboring sources vs distance, finding a very small correlation (< 0.02) at distances larger than 2.5cm (as expected from theoretical considerations [176]). To compute FIT and TE, the time-frequency representation of sensor data was projected into source space and averaged over vertices within the ROI (80,20,10 vertices for V1, V3A, LO3, respectively).

Parameters and details of the Information theoretic analyses For the analysis FIT and TE we used gamma-band instantaneous power obtained by computing the time-frequency representations of single-trial data via the multi-taper method and then averaging the obtained powers in the $[40-75]$ Hz band, exactly as described in the original publication [11]. To estimate the joint probability distributions of the neural activity and the stimulus (or the choice) used to compute the information theoretic quantities, we binned the MEG gamma power from each ROI into 2 equally populated bins and then computed empirically the frequency of occurrence of each response bin across all available trials. The stimulus features used for the information analyses was the average contrast of the 10 samples presented on each trial, discretized into two values. (We coded 0 the choice to report that the average contrast of the 10 sample gratings was below the reference contrast and 1 the choice to report that it was above the reference contrast). The choice feature used for the information analyses was the binary choice (average contrast of the sequence higher or lower than the reference contrast) reported by the participant in each trial. We computed the information quantities for both the feedforward and the feedback direction for the left and the right hemisphere separately and then averaged the two.

Unless otherwise stated, we computed information quantities using all available trials (correct and error trials). For the specific set of information analyses comparing correct and error trials (Fig. 3G,H), we randomly subsampled correct trials so that the number of correct and error trials used to compute the information quantities was the same for each session. In this way, the information values for correct and error trials can be compared fairly because their difference cannot reflect possible differences in limited-sampling biases due to different data numerosity [179].

Statistical analyses We established significance of the information measures in the time-delay space using a cluster-based nonparametric statistical test described in Section 3.8.1.7, see [177, 178].

To provide a quantification values of TE and FIT across participants, sessions and network links (Fig. 3D,F,H), we selected a rectangular region in the time-delay domain to select the TE and FIT values for the across sessions statistics, centered around the FIT significant cluster (FIT-specific region). We computed the average over delays and then picked the maximum over time within this region. This gave us one single TE and one single FIT value for each hemisphere in each session. The comparisons of values across participants, subjects and links was performed using two-tailed paired t-tests.

Additional results Here we list the results of a number of additional analyses that could not be inserted in the main text due to lack of space but that are helpful to better understand and support the conclusions presented in the main text.

The first set of results regards the encoding of information in individual regions of the visual network, rather than the transmission across regions of information about the stimulus. We reported temporal profiles of stimulus information in the three selected ROIs in Fig. 3.S9B. Instantaneous information profiles showed a clear lag in the onset of stimulus information that could not be explained by instantaneous source mixing (see Section 3.8.2.7). The amount of mutual information about the stimulus carried by the power of the gamma band in the visual cortical network is larger in the first half of the presentation of the stimulus ([0-500]ms peri-stimulus, shortened to 'early') than in second half of the presentation of the stimulus ([500-1000]ms peri-stimulus,denoted as 'late') within the trial (Fig. 3.S9C, left). This is why we concentrated the FIT analyses in the first part of the trial (the early window). In the early part of the trial, the gamma band activity in the visual cortical network carries more stimulus than choice information (Fig. 3.S9C, middle) and this information is higher in correct compared to error trials (Fig. 3.S9C, right) . These results are useful to confirm that stimulus information coding is of more prominent importance in the visual network and that the presence of this information is key to perform accurate perceptual discriminations.

The second set of results regards additional findings about the the transmission across regions of information about the stimulus. We produced network representation showing the relative strength of individual TE and stimulus FIT links

contributing to the observed differences in directionality (Fig. 3.S9D, compare with Fig.3D) and feedforward behavioral relevance (Fig. 3.S9E, compare with Fig.3E) of information transmission. We found no difference of FIT stimulus nor TE in the feedback direction between correct and error trials (Fig. 3.S9F). For the example pair V1-V3A we identified a significant cluster of stimulus FIT feedforward (V1 to V3A) but not feedback (V3A to V1) in the time-delay domain (Fig. 3.S9G; cluster statistics, $p<0.01$). These results suggest that feedforward propagation of stimulus information, but not feedback propagation, is specifically key for correct behavior.

In the main text we indicated that the time delay region in which FIT about the stimulus was significant was in the region 200 to 400ms after the stimulus onset, with an inter-area communication delay between 65 and 250ms. This statement is supported by the plot in Fig. 3.S9G of the time-delay map points that are significant according to the cluster permutation test.

Finally, we performed a control analysis to quantify TE in a time-delay region around which TE was maximal. This is of interested because in the main text analyses (Figure3E-H) we compared FIT and TE using a time-delay region around the peak of FIT. The TE panel (Fig. 3.S9H) shows that TE peaks in a different time-delay region with respect to stimulus FIT (Fig. 3C). Taking a DI-specific box centered around the TE peak in time-delay to select information values we could not assess the direction nor the behavioral relevance of information transmission (Fig. 3.S9I).

3.8.3.2 EEG data

Behavioral task and EEG recordings We next analyzed a publicly available EEG dataset [180]. Data are available at [https://datadryad.org/stash/dataset/doi:10.5061/dryad.8m2g3]. Full details (including details of approvals from Ethical Committees) are reported in the original publication. Here we summarize them briefly. The EEG data were recorded while participants (N=16) performed a face detection task. Participants were presented with an image hidden behind a bubble mask that was randomly generated in each trial. The presented image was a image of a face in half of the trials and a random texture in the other half of the trials. Participants were instructed to report whether a face was present or not. In our analyses, we only considered correct trials where the face was correctly detected by the participants (approximately 1000 trials per subject). Following the recommendations of the original publications analysing these data [180, 181], we excluded one participant from the analysis due to a poor EEG signal that did not contain significant eye visibility information in any of the electrodes. All analyses in our paper are based on the N=15 selected participants. EEGs were recorded by fitting participants with a Biosemi head cap comprising 128 EEG electrodes. EEG data were re-referenced offline to an average reference, band-pass filtered between 1 Hz and 30 Hz using a fourth order Butterworth filter, down-sampled to 500 Hz sampling rate and baseline corrected using the average activity between 300ms pre-stimulus and stimulus presentation. ICA was performed to

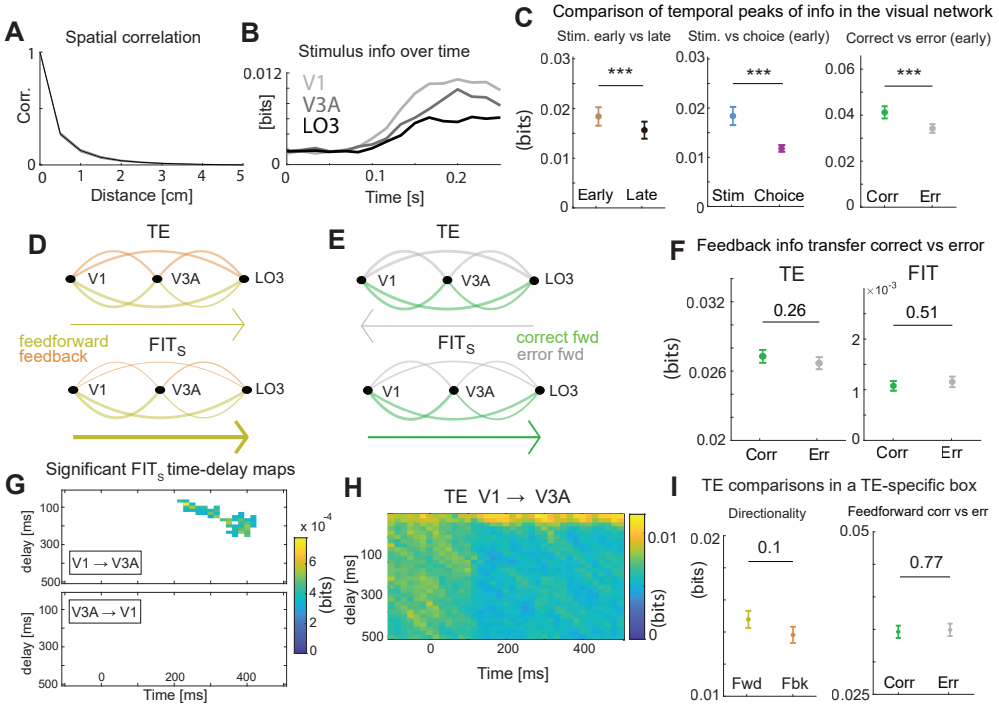

Figure 3.S9: Additional analyses of the MEG dataset. (A) Correlation between LCMV spatial filters of neighboring sources as a function of their distance. (B) Time course of stimulus information in each ROI. (C) Properties of mutual information between stimulus and MEG activity. Left: Peak values of stimulus information in the first (early time window, from 0 to 500ms post-stimulus onset) and second half (late time window, 500 to 1000ms after stimulus onset) of the stimulus window. Middle: Peak of stimulus and choice information in the early window. Right: Peak of stimulus information in the early window in correct and incorrect trials, respectively. D) Graphs representing the strength of feedforward (yellow) and feedback (orange) information transmission in the network for TE (top) and stimulus FIT (bottom). Links are weighted proportionally to the communication strength between each pair. The arrows on the bottom points toward the dominant direction of overall transmission, and are weighted proportionally to the difference between feedforward and feedback transmission. (E) Same as D but for feedforward transmission in correct (green) vs error (gray) trials. (F) Values of TE (left) and of FIT about the stimulus (right) computed in the feedback direction separately in correct and in error trials. (G) Plot of the points with significant values of the stimulus FIT between V1 and V3A (top) and V3A and V1 (bottom) according to a cluster permutation test. Only points that are significant are colored. Color scale is the same as the Fig. 3C. (H) TE time-delay map in the V1 to V3A direction. (I) Values of TE using a time-delay box centered around the TE peak in the time-delay map. In all panels, lines and image plots show averages and errorbars SEM across participants, experimental sessions and regions pairs (in case of FIT and TE) or regions (in case of mutual information). *: $p<0.05$, **: $p<0.01$, ***: $p<0.001$. All information-theoretic quantities were first computed separately for left and right hemisphere and then averaged.

reduce blink and eye- movement artifacts, as implemented in the infomax algorithm from EEGLAB [202]. Components representing blinks and eye movements were identified by visual inspection of their topographies, time courses, and amplitude spectra.

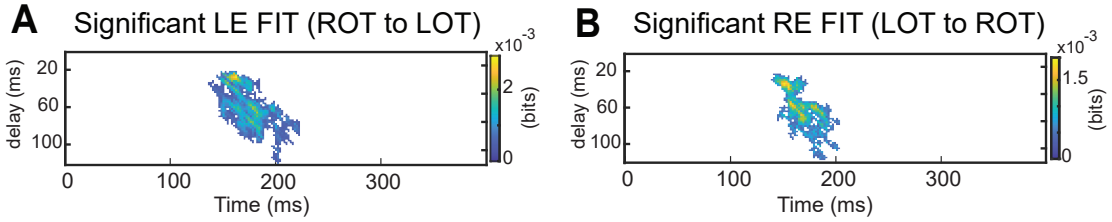

Figure 3.S10: A) Plot of the points with significant values of the left-eye (LE) visibility FIT between the EEG of Right Occipito-Temporal (ROT) and Left Occipito-Temporal (LOT) electrodes, according to a cluster permutation test. Only points that are significant are colored. B) Same as A but for the Left-Eye (LE) visibility FIT between the EEG of Left Occipito-Temporal (LOT) and Right Occipito-Temporal (ROT) electrodes. The color scale of panel A and B is the same as the corresponding plots in Figure 3B and C, respectively.

Details of the information theoretic analyses and additional results For the analyses of TE and FIT, we selected the EEG electrodes in the left and the right Occipito-Temporal regions that had the highest mutual information about the visibility of the contralateral eye, exactly as done in previous papers [181]. (Specifically, in Ref. [181], the authors used the following criteria to select one electrode in LOT and one in ROT for each participant, see Fig. 3.S13A. For LOT they selected the electrode with maximum right eye MI from electrodes on the radial axes of P07, P7, and TP7, excluding midline Oz and neighboring O1 radial axes. On the right hemisphere, for ROT the author selected the EEG electrode with maximum left eye information from sensors on the radial axes of PO8, P8, TP8, excluding midline Oz and neighboring O2 radial axes). We computed the first derivatives of the EEG signal for both Occipito-Temporal sensors and used both its absolute values and first derivatives to compute the information quantities, for consistency with the information-encoding analyses performed in a previous study [181].As stimulus feature for the computation of mutual information and FIT, we used the visibility of an eye (defined as the fraction of pixels within the eye region that were not hidden by the bubble mask). This feature was discretized using 2 equipopulated bins. We computed the information quantities for all combinations of directionality of flow across hemispheres (left to right, right to left) and eye identity (left or right eye). We computed significance of FIT in the time-delay using the cluster-based permutation test described in Section 3.8.1.7. This analysis revealed a significant cluster of FIT about the left eye in the right-to-left direction (Fig. 3.S10A) and a about the right eye in the left-to-right direction (Fig. 3.S10B). To provide a quantification values

of TE and FIT across participants (Fig. 4D), we selected a rectangular region in the time-delay domain to select the TE and FIT values for the across participants statistics, centered around the contralateral FIT significant cluster (same for both eyes, as they were significant in very similar time-delay regions). We computed the average over delays and then picked the maximum over time within this region. This gave us one single TE and one single FIT value for each subject. The comparisons of values across participants was performed using two-tailed paired t-tests.

3.8.3.3 Analysis of spiking activity in a thalamocortical network

Electrophsyiological experiments We analysed previously published [23] recordings of multi-unit spiking activity simultaneously obtained from electrodes placed in the primary visual cortex (V1), primary somatosensory cortex (S1), first-order visual thalamus (the lateral geniculate nucleus, LGN), and the first-order somatosensory thalamus (the ventral posteromedial nucleus, VPM) of anaesthetized rats (Fig. 3A). Data are made available with this NeurIPS submission as Supplemental Material. Full details (including details of approvals from Ethical Committees and Local Authorities) are reported in the original publication. Here we summarize them briefly.

These data were recorded from N=6 rats (using one-shank Silicon Michigan probes, Neuronexus Technologies; 100-μm intersite spacing) in three stimulation conditions: visual stimulation, whiskers tactile stimulation and bimodal stimulation (simultaneous visual and tactile). All experiments were conducted under urethane anesthesia. The visual stimuli consisted of a light flash (50-ms-long LED light flashes at 300 lux). The unimodal somatosensory stimulus consisted of a whisker deflection. For bimodal stimulation, whisker deflection and light flashes were applied in the same hemifield. Stimuli were randomly presented across trials. In our analysis, we considered only stimulation contralateral to the recorded brain areas. Each type of stimulus was presented 100 times. The non-stimulated eye was covered with an aluminum foil patch. Neural activity was recorded at a sampling rate of 32 kHz, bandpass filtered (0.1 Hz and 5 kHz) then down-sampled to 8 kHz. In the current work, we used the recordings from infragranular layers of S1 and V1 and from VPM and LGN. Multi-unit spike times were first detected from the band-passed (400–3,000 Hz, fourth-order IIR Butterworth Filter) extracellular potential in each electrode by threshold crossing (>3 SD). A spikes train was obtained for all channels using a temporal binning of 0.125 ms (1/8kHz). For each brain region, spiking activity was then pooled together using all recorded spikes form all electrodes related to that region.

Parameters and details of the information analyses To compute mutual information and FIT, we defined two different stimulus set of interest. To measure information related to tactile discrimination, we used a "tactile-discriminative set" made of the unimodal visual and the bimodal visual-tactile stimulus (the two stimuli in the set are discriminated by the presence or absence of a tactile stimulus). Similarly, to measure information related to visual discrimination, we used a

"visual-discriminative set" made of unimodal tactile and the bimodal visual-tactile stimulus (the two stimuli in the set are discriminated by the presence or absence of a visual stimulus).

As stimulus feature for the computation of mutual information and FIT for the tactile (visual) discriminative set, we used a binary value indicating either the delivery of a visual (tactile) unimodal or a bimodal stimulation. We computed the information quantities for all combinations of directionality of flow across the visual (LGN to V1 and V1 to LGN) and somatosensory (VPM to S1 and S1 to VPM) thalamo-cortical pathways and stimulus-discriminative sets (visual or tactile). To provide a quantification values of TE and FIT across animals (Fig. 3.S11E,F,H,I), we selected a rectangular region in the time-delay domain to select the TE and FIT values for the across animals statistics, centered around the FIT peaks about the tactile stimulus in the VPM to S1 direction (for the somatosensory pathway) and about the visual stimulus in the LGN to V1 direction (for the visual pathway). We computed the average over delays and then picked the maximum over time within this region. This gave us one single TE and one single FIT value for each animal. The comparisons of values across animals was performed using two-tailed paired t-tests.

Further details about the information-theoretic results We first focused on the somatosensory pathway (VPM and S1). We found that, on this pathway, the tactile FIT in the VPM to S1 direction was visibly higher than the visual FIT in the same direction and both tactile and visual FIT in the S1 to VPM direction (Fig. 3.S11B) We found that the timing of tactile FIT from VPM to S1 was consistent with the one of tactile information in neural activity, that was present in the 5-30ms and 15-30ms post-stimulus intervals in VPM and S1, respectively (Fig. 3.S11D left, top and bottom lines). We found that FIT revealed the directionality of tactile information flow, which was significantly larger in the feedforward (VPM to S1) compared to the feedback (S1 to VPM) direction (Fig. 3.S11E right, $p = 0.0065$). On the contrary, TE computed for the tactile set was not significantly different in the two directions (Fig. 3.S11E left, $p = 0.13$). FIT also revealed the content of communication from VPM to S1, being significantly larger for the tactile set than for the visual set (Fig. 3.S11F, right; $p = 0.0084$), while TE from VPM to S1 was not significantly different in the two directions (Fig. 3.S11F, left; $p = 0.12$).

Complementary results were found in the visual pathway (LGN and V1). On this pathway, the visual FIT in the LGN to V1 direction was visibly higher than the tactile FIT in the same direction and both tactile and visual FIT in the V1 to LGN direction (Fig. 3.S11C). FIT for the visual-discriminative set in the LGN to V1 direction peaked in a time interval of approximately 45 to 65ms after stimulus-onset and with a transfer delay of approximately 10-25ms Fig. 3.S11G). The visual FIT values were larger in the feedforward than in the feedback direction (Fig. 3.S11H, right; $p = 0.033$), while TE was not sensitive to the directionality of visual information (Fig. 3.S11H, left; $p = 0.15$). Moreover, visual FIT values were significantly

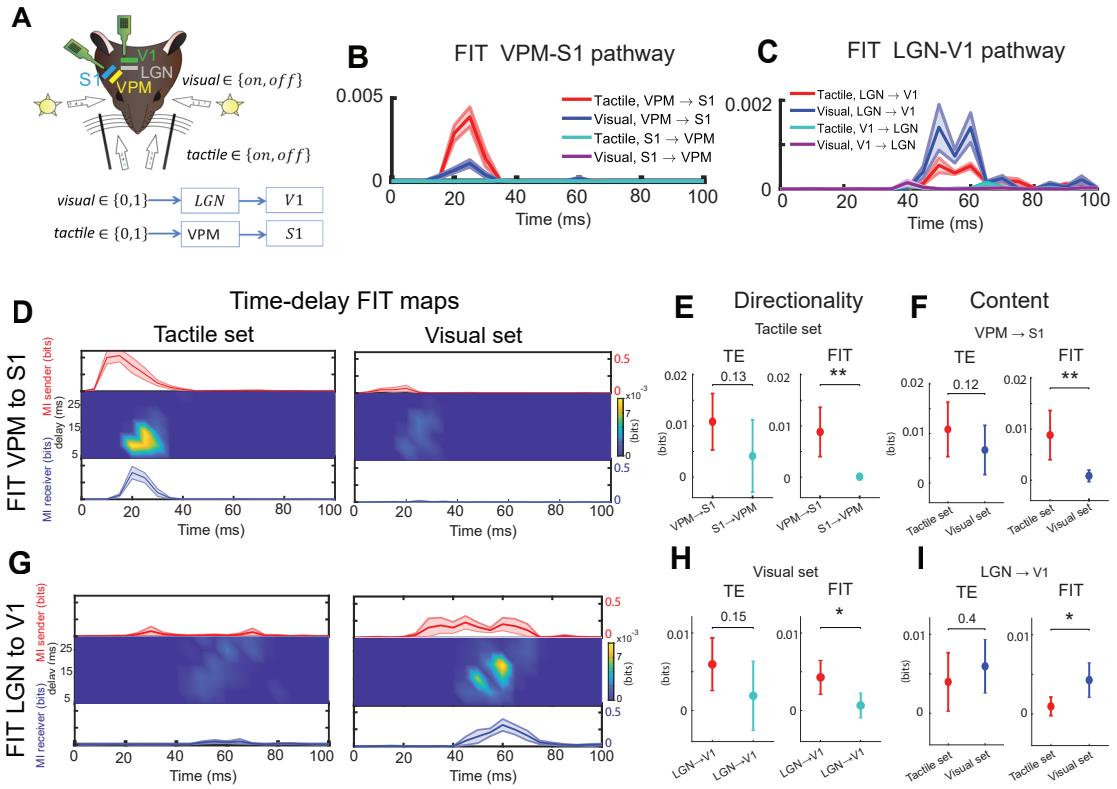

Figure 3.S11: Sensory related info transfer carried by multi unit actity (MUA). A) Schematic of the experimental setup. MUA was recorded in rats from the Ventral Posteromedial nucleus (VPM) of the thalamus, the Lateral Geniculate Nucleus (LGN) of the thalamus, the primary somatosensory (S1) and visual (V1) cortex simultaneously. During the recording either a unimodal tactile, a unimodal visual, or a bimodal (visual and tactile) stimulus was presented. B) FIT values averaged across delays for all combinations of transfer direction and stimulation type on the somatosensory pathway (VPM-S1). C) Same as B but for the visual pathway. D) FIT from VPM to S1 (mean across subjects) for each value of delay and post-stimulus time. Line plots above and below show Mutual Information (MI) between the presented stimulus and the recorded MUA in the VPM and S1, respectively. Shaded error bars show the SEM of the measure across subjects. The left panel reports values of information and FIT for the tactile-discriminative set, whereas the right panels report values of information and FIT about the visual stimulus set. (E) Directional sensitivity for TE (left) and FIT (right) between VPM and S1, for the Tactile stimulus-set. F) Comparisons between tactile- and visual-discriminative set, for the TE (left) and the FIT (right) from VPM to S1. G) Same as panel D but from LGN to V1. H) Same as panel E but between LGN and V1, for the Visual stimulus-set.

larger than the tactile ones from LGN to V1 (Fig. 3.S11I, right; p = 0.013), while TE did not capture these sensory modality-specific differences (Fig. 3.S11I, left; p = 0.4). Taken together, these results highlight the power of the FIT in revealing feature- and direction- specific transfers of information with high temporal preci-

sion, beyond what is achievable using methods that measure the total propagation of neural activity such as DI.

3.8.3.4 Applications of cFIT to real neural data

We tested the effectiveness of the conditioned FIT (cFIT) in the analysis of neurophysiological data by applying it to perform further analyses on the EEG and the MUA datasets (Fig. 3.S12). For simplicity, when computing cFIT from X to Y conditioned on the stimulus information of Z, we always considered the same communication delay (that is, the time difference between the present and past activity used to compute the information theoretic measures) for the past activity of the sender X and the past activity of the third region Z. However the definition of conditional FIT holds for an arbitrary representation of Z_{past}, potentially including multiple time points or a communication delay that is different from the one of X_{past}.

We first applied cFIT to the EEG dataset. We investigated whether the contralateral Occipito-Temporal electrodes used to compute TE and FIT in Figure 4 where the sole senders of eye-specific information across hemispheres. We selected two different sets of putative alternative senders of eye-specific information and used cFIT to remove the contribution of the putative alternative senders from the contralateral FIT that we measured. Namely, we selected the third location to be conditioned upon from either a set of weak or a set of strong alternative senders for both the left and for the right eye (Fig. 3.S13). For each participant, we defined the two weak alternative sender locations (one for the left an one for the right eye) as those electrodes carrying the lowest amount of stimulus information about the left or the right eye, respectively, in the frontal lobe of the brain. The expectation was that removing the contribution of these electrodes using cFIT would not change appreciably the results obtained with the contralateral unconditioned FIT reported in Figure 4.

For each participant, we defined the strong alternative senders locations (one for the left an one for the right eye) as those electrodes carrying the second-largest amount of information about the left eye in ROT or about the right eye in LOT (ROT and LOT defined as in Ref [181]). We found that FIT conditioned on the contralateral-eye information of one of the weak alternative senders (the orange lines in Fig. 3.S12A) did not reduce FIT, as the cFIT was virtually equal to the unconditioned FIT (the blue trends in Fig. 3.S12A). However, FIT conditioned on the contralateral-eye information of one of the strong alternative senders (the green trends in Fig. 3.S12A) was lower than unconditioned FIT. However, both cFIT given the weak and the strong alternative senders were significant (cluster statistics over time, $p<0.01$). The fact that cFIT was lower than FIT when conditioning on informative electrodes but not when conditioning on weakly informative electrodes suggests that cFIT is effective at removing influences related to similar feature-specific (but not un-specific) information present already in the past activity of other regions. The fact that the inter-hemisperic Occipito-Temporal controlateral-eye-specific cFIT is still highly significant and is only marginally smaller than the original uncondi-

tioned FIT suggests that most eye-specific information flows across hemispheres through the contralateral Occipito-Temporal electrodes selected in [181].

Next, we analysed the spiking activity dataset. We examined tactile-discriminative information flowing through the somatosensory pathway, and the visual-discriminative information flowing though the visual pathway.

We first used cFIT to test whether the tactile-discriminative FIT from the somatosensory thalamus VPM to the somatosensory cortex S1 could have actually been relayed through the visual thalamus LGN. The neurophysiological expectation is that all tactile-discriminative information flows within the somatosensory pathway, without contributions from visual stations. Consistent with this expectation, we found that the tactile-discriminative cFIT from VPM to S1 conditioned on the visaul thalamus LGN was equal to the unconditional tactile-discrminative FIT from VPM to S1.

We then used cFIT to test whether the visually-discriminative FIT from the visual thalamus LGN to the visual cortex V1 could have actually been relayed through the somatosensory thalamus. The neurophysiological expectation is that all visual-discriminative information flows within the visual pathway, without contributions from somatosensory stations. Consistent with this expectation, we found that the visually-discriminative cFIT from LGN to V1 conditioned on the somatosensory thalamus VPM was equal to unconditional visually-discriminative FIT from LGN to V1.

Together, these results suggest that cFIT is useful to remove contributions from alternative pathways specifically with regard to the transmission of feature-specific information.

3.8.4 Comparison with other possible or previously published measures

We examine how FIT differs with respect to other possible or previously published algorithms that were designed to identify the information flow across regions about behavioral or stimulus features of interest. We first consider two measures that implement the Wiener-Granger discounting of the information present in the past activity of the sender. We then consider two other methods, that did not implement this principle.

3.8.4.1 Comparison with variations in transfer entropy ΔTE

As mentioned in the main text, one simple-minded proxy for identifying feature-specific information flow could be quantifying how the total amount of transmitted information (TE) is modulated by the stimulus-feature [31]. For the case of two stimuli, this amounts to the difference of TE computed for each stimulus-feature value.

We now show, using simulations, that this measure can fail in capturing feature-related information flow. We performed simulations in a scenario having variable

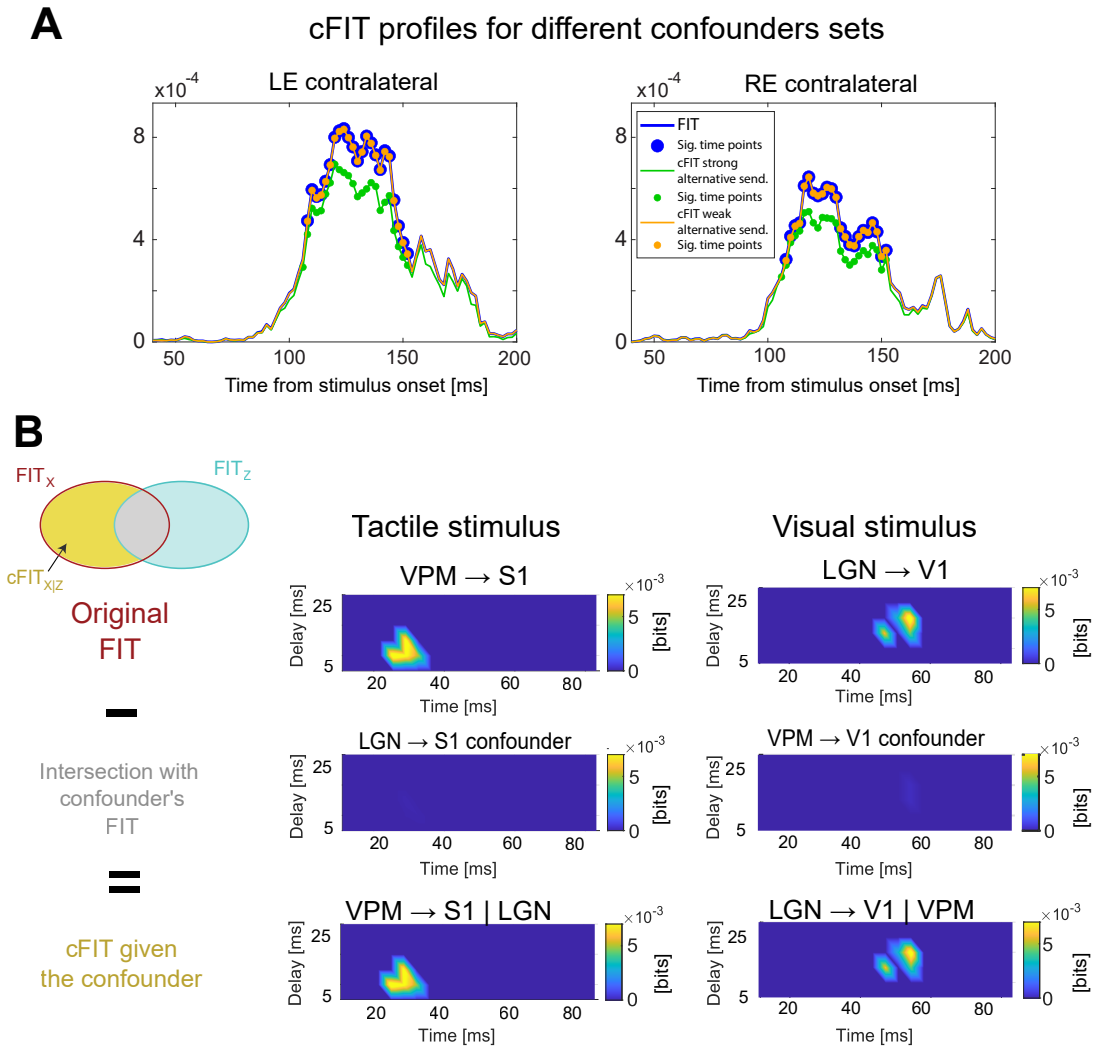

Figure 3.S12: Application of cFIT to experimental data. A) cFIT application to the EEG data. We conditioned the contralateral FIT about the left and the right eye (see Fig. 4) to the activity of two either weak or strong alternative eye-visibility information senders. Temporal profiles of unconditioned FIT (in blue) for about the left eye from ROT to LOT (left) and about the right eye from LOT to ROT (right). cFIT temporal profiles when conditioning on weak alternative senders (in orange) and on strong alternative senders (in green). The points where the measures were significant are indicated with a circular marker ($p<0.01$, cluster statistics). B) cFIT applied to MUA data. We conditioned tactile- (visual-) discriminative FIT through the somatosensory (visual) pathway (first row) to the activity of the visual (somatosensory) thalamus. The amount of unconditioned FIT that was shared with the FIT through the alternative sender (second row) was subtracted from the original FIT to obtain cFIT (third row). The left column shows results for the tactile-discriminative set, the right column for the visual-discriminative set.

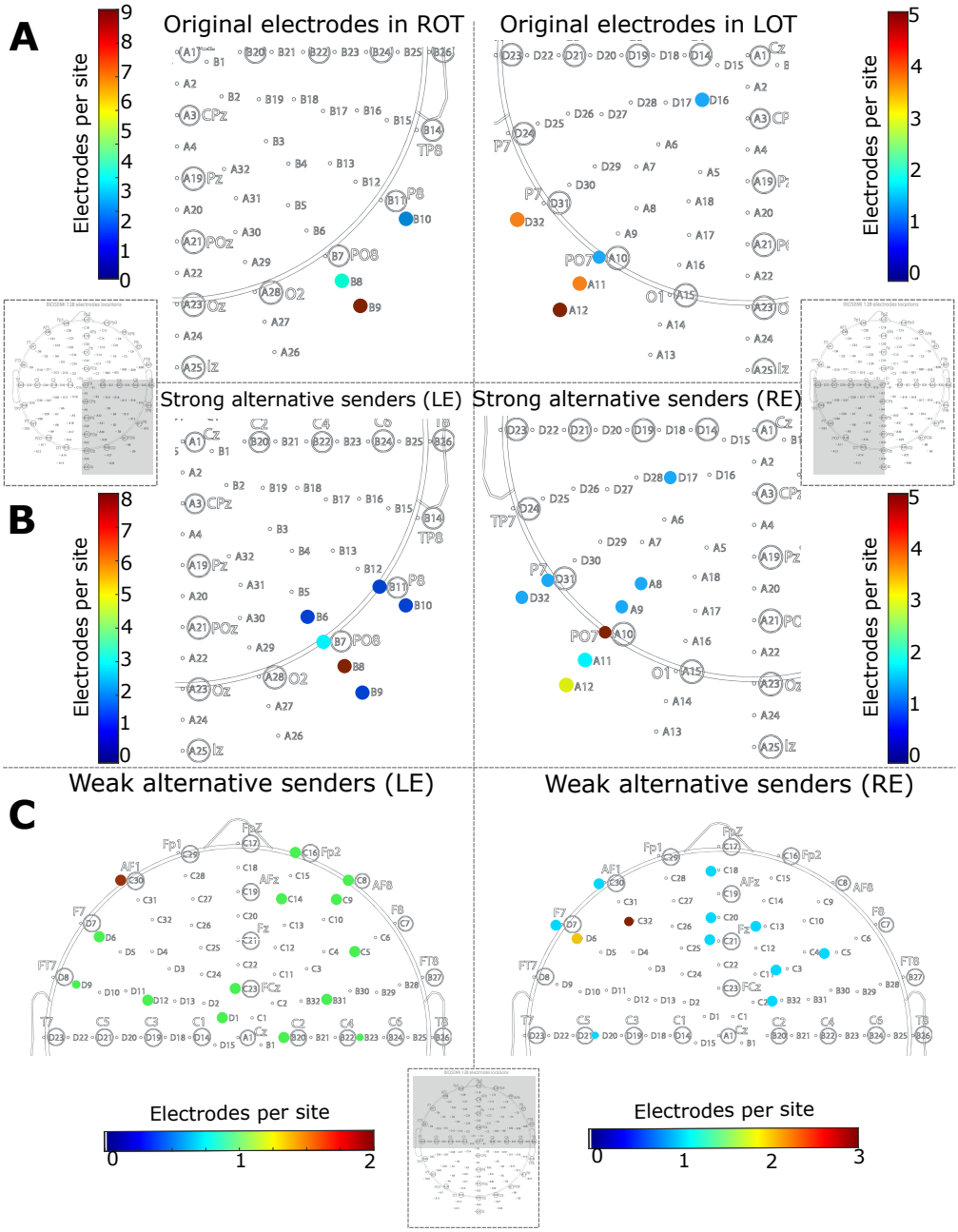

Figure 3.S13: Sets of electrodes used in the EEG data analysis. A) Electrodes used to measure FIT and TE and for all the analyses in Fig.4, same electrodes used in [181] B) Set of 'strong alternative senders', selected as the electrodes carrying the second maximum amount of information about the left eye in the ROT (left panel) and about the right eye in the LOT (right panel). C) Sets of 'weak alternative senders', selected as the electrodes in the frontal lobe carrying the minimal amount of information about the left eye (left panel) and about the right eye (right panel).

degrees of both feature-specific and feature-unrelated information transfer. The encoded and transmitted stimulus feature S was a stimulus-intensity integer value (1 or 2). The activity of the sender X was a two-dimensional variable with one stimulus-feature-informative X_{stim} and one stimulus-uninformative component X_{noise}. The feature-informative dimension had a temporally-localized stimulus-dependent bump in the activity (from 200 to 250ms) and additive Gaussian noise The stimulus-unrelated component was, at any time point, a zero-mean Gaussian noise. The activity of the receiver Y was the weighted sum of X_{stim} and X_{noise} with a delay δ, plus Gaussian noise. The delay δ was chosen randomly in each simulation repetition (N=50) in the range 40-60ms. We tested whether ΔTE across simulation repetitions was significantly different from zero using a two-tailed t-test.

As we did in Figs. 2 and 3.S3 for FIT and TE, we studied the behavior of ΔTE as a function of the simulation parameters W_{stim} (which increases the amount of information transferred about the stimulus feature) and W_{noise} (which increases the amount of feature-unspecific information that is transferred from X to Y). We found (Fig. 3.S14A) that ΔTE had almost no relationship with the values of W_{stim} and W_{noise}, unlike FIT which individuated stimulus-feature-specific transfer correctly because it increased with W_{stim} but not with W_{noise} (Fig. 3.S3A).

Note that we performed also simulations (that were exactly like those of Fig. 2, except that we had 2 rather than 4 stimulus intensity values) in which the noise in X was multiplicative rather than additive. In this case (results not shown) ΔTE increased with both W_{stim} and W_{noise}. Thus, ΔTE had limited capabilities of identifying some stimulus-feature-specific information transfer in some specific case, but it dot reflect it in general.

The reason that ΔTE cannot capture feature-specific information flow is, in our view, that ΔTE is a measure of variation of information strength across stimulus-feature conditions rather than a measure of stimulus-feature-specific information transfer.

Additionally, we tested ΔTE on MEG data. We first binarized the stimulus feature into two classes (average contrast either greater, S=1, or lower, S=0, than the reference contrast). We computed TE for all pairs of visual regions in the visual cortical network separately in trials with the same value of the binary stimulus we and computed the difference ΔTE between these values. Fig. 3.S14B shows that ΔTE in the visual cortical network had the same strength in the feedforward and feedback direction, unlike FIT that showed a clear directionality of communication of stimulus information (stronger in the feedforward than in the feedback direction). Finally, when computing TE on the spiking activity data of the rat thalamocortical network, we found that TE from thalamus to cortex did not vary between the tactile-discriminative and visually-discriminative stimuli set (see Fig. 3.S11E-I). In contrast, on the same data FIT could distinguish tactile-discriminative from visually-discriminative information flow from thalamus to cortex (see Fig. 3.S11E-I).

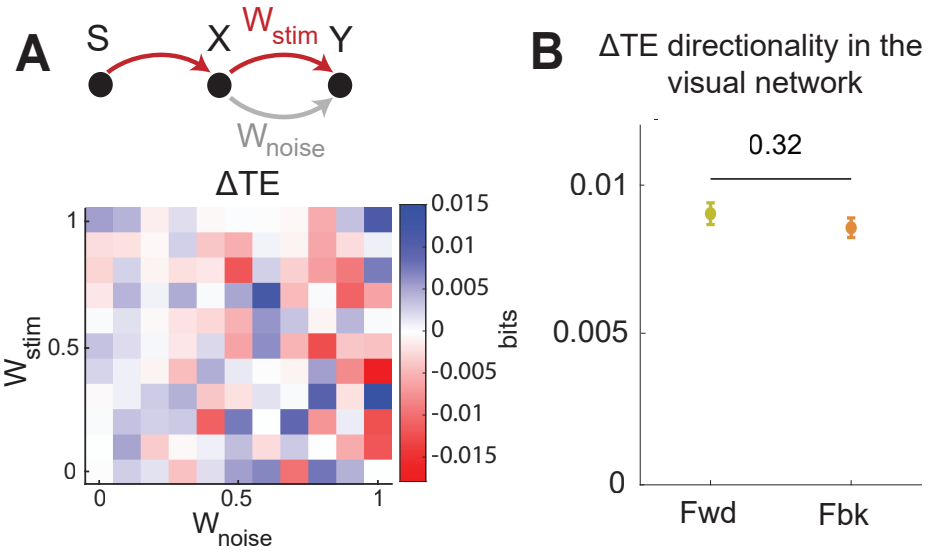

Figure 3.S14: Performance of the Transfer Entropy Difference across stimuli (ΔTE) on simulated and real MEG data. A) Simulated data. Values (mean $\pm$ SEM across 50 simulations) of ΔTE from X to Y computed on the same simulated data used in Fig 2X as a function of the simulation parameters W_{stim} (which increases the amount of information transferred about the stimulus feature) and W_{noise} (which increases the amount of feature-unspecific information that is transferred from X to Y). We found that ΔTE had almost no relationship with the values of W_{stim} and W_{noise}, unlike FIT which increases only with W_{stim}. B) Real MEG data. Average across participants, sessions and pairs of regions of the values of ΔTE (reported values were obtained taking the average over delays and then the maximum over time in the same time-delay region used for the results in Fig. 3D,F,H).

3.8.4.2 Comparisons with Directed Feature Information (DFI)

A previous study [183] defined a measure, Directed Feature Information (DFI), which computes feature-specific information redundant between the present activity of the receiver and the past activity of the sender, conditioned on the past activity of the receiver. However, DFI used a measure of redundancy that actually conflated the effects of redundancy and synergy (see Section 3.8.1.5 where we consider in detail its definition and its PID decomposition). Because of this, DFI can be negative and thus not interpretable as measure of information flow. Moreover, because DFI discounts only past activity of the sender rather than its feature-specific information, it is less precise and less conservative in localizing direction and timing of feature-specific information flow.

The above properties are expected from theoretical considerations but were also demonstrated by us in the following numerical simulations. We computed DFI in the two simulations described in Section 3.8.2.1 (Fig. 3.S15). We found that, in

general, DFI had a trend similar to FIT, increasing with the amount of stimulus-feature-related transfer from X to Y (W_{stim}) and decreasing with the amount of stimulus-unrelated transfer (W_{noise}). However, DFI had several false positives (cases when there was no transmission in the ground truth of the simulated data but it was detected as significant by the algorithm) and also had several false negatives (cases when there was transmission in the ground truth of the simulated data but it resulted as non significant by the algorithm). In comparison, FIT under the same conditions and same simulations had none, see Fig. 2A). More importantly, as a consequence of its inability to include only redundancy and discard synergy, DFI values were very often negative, and could be negative over time both at baseline and during stimulus-feature-related transmission (Fig. 3.S15B).

The limitations of DFI for individuating directed well time-resolved flow of information about specific stimulus features were further tested with the bidirectional information transfer simulations described in detail in Section 3.8.2.4. We remind briefly that in these simulations we simulated a scenario with bidirectional communication between X and Y with stimulus-feature-related transfer from X and stimulus-unrelated transfer from Y to X (Fig. 3.S5A). In brief, both X and Y received information directly from a feature-information-sending node S. X received feature information from S early on (between 50 and 90 ms) and Y received feature information from S at a later time (between 110 and 150 ms). X sent its entire activity to Y (therefore communicating its feature information when it became available). Y instead only sent to Y a part of its activity that did not carry feature information. We found that while DFI had a significant positive bump from X to Y in the $[60, 100]ms$ time window, it also had a significant negative bump from Y to X in the time window in which X encoded the feature $[50, 90]ms$. Crucially, the presence of significant DFI from Y to X preceding in time the DFI from X to Y would be interpreted that there is a bidirectional flow of stimulus-feature information, occurring first from Y to X and then from X to Y. Therefore, DFI could not capture correctly neither the directionality nor the timing of the stimulus-feature information flow that we put in the simulations.

For FIT, which is a non-negative measure, we always used one-tailed tests to determine whether the measured values were significantly larger than the 99th percentile of the null hypothesis distribution obtained as described in Section 3.8.1.7.

For DFI, which is an unsigned measure, we implemented a two-tailed test. Analogous to our method for FIT, we computed two null hypothesis distributions: one by shuffling S across all trials, and one by shuffling X for fixed values of S. We then tested whether DFI was either above the 99.5th percentile of the element-wise maximum or below the 0.5th percentile of the element-wise minimum of these two null hypothesis distributions. If one of these conditions was met, we assigned significance to DFI.

Lastly, we computed DFI on the three real datasets (MEG, EEG and spiking activity) presented in the main text. We found (Fig. 3.S5C-E) that the problems with DFI predicted by mathematics (see Section 3.8.1.5) and encountered simulations are

also found in the neural datasets. On real data, DFI was very often negative and it did not detect directionality or feature specificity in cases in which we would expect from previous literature that specificity or directionality should exist.

In the MEG dataset (Fig. 3.S5D), DFI was negative values and thus not interpretable as measure of information transfer. Unlike FIT, DFI could not detect that (as predicted by previous studies) stimulus information is stronger in the feedforward than in the feedback direction, and DFI could not detect that feedforward stimulus information is stronger in correct than error trials (an important result found by FIT).

In the EEG dataset (Fig. 3.S5C), DFI was negative and thus not interpretable as measure of information transfer. The comparison of the DFI results between eye visibility features and directionally of cross-hemispheric transfer could not support the conclusion (predicted by findings in previous literature and confirmed by the FIT analysis) that across-hemisphere information transfer is directional from contra- to -ipsilateral (DFI does not detected a leading direction of RE information transfer) and is feature specific (DFI does not detected a difference between LE and RE information in the R to L hemisphere communication).

In the thalamocortical spikes data (Fig. 3.S5E), DFI has mostly positive values which are thus interpretable in terms of information transmitted. DFI confirms (though with lower statistical power) the FIT results than in both the somatosensory and visual corticothalamic pathway more information is transmitted feedforward about the corresponding sensory modality (more visual than somatosensory information transmitted from visual thalamus to visual cortex, and more somatosensory than visual information transmitted from somatosensory thalamus to somatosensory cortex). However, DFI failed to demonstrate that, as expected from well-established neurophysiological findings, more information about such simple stimulus features is transmitted from thalamus to cortex than from cortex to thalamus.

In sum, our results lead us to conclude that the definition of redundancy used in DFI that, unlike the more refined one arising from PID, conflates synergistic and redundant effects, leads to major problems predicted by theory and confirmed by simulation and in real data. Our results suggest that DFI is not robust or refined enough to be applied generally and systematically to brain data, and that the advances provided by FIT with respect to DFI are important not only conceptually but also for the analysis of empirical datasets.

3.8.4.3 Comparison with measures not discounting past information in the receiver as in the Wiener-Granger Causality principle

We finally consider the suitability for identifying flow of information about specific features of possible alternative measures that, although have relevance to feature information coding across areas, *do not* implement the Wiener-Granger discounting of the information present in the past activity of the sender. In brief, methods that do not implement this (and thus just correlate past information of the sender with present information of the receiver), erroneously identify information already en-

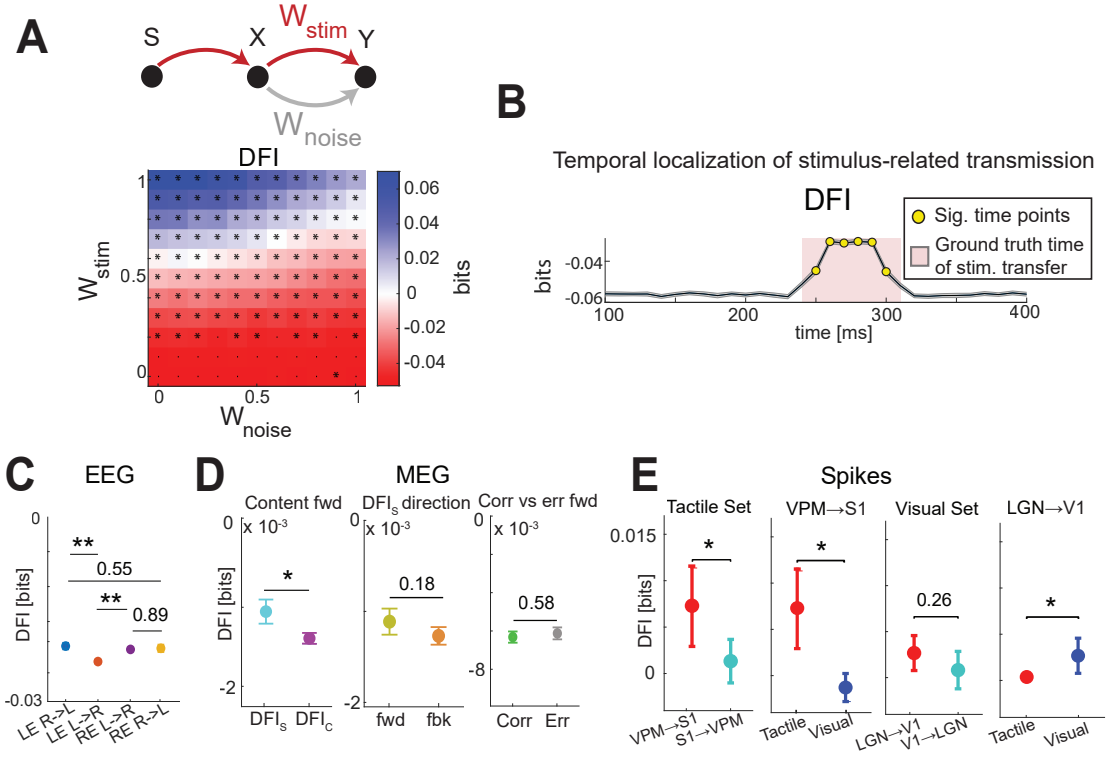

Figure 3.S15: DFI tested on simulated and real brain data. In Panels A,B DFI is tested on the same simulations used for Fig 2A-B. Panel C-E report DFI results on real brain data. A) DFI as function of stimulus-feature-related (W_{stim}) and -unrelated (w_{noise}) transmission strength. * indicate significant values ($p < 0.01$, permutation test) for the considered parameter set. B) Dynamics of DFI in a simulation with time-localized feature-information transmission. Red area shows the window of feature-related information transfer. Yellow dots show time points with significant information (p ¡ 0.01, permutation test). Results plot mean (lines) and SEM (shaded area) across 50 simulations (2000 trials each). C) EEG DFI between L, R hemisphere about L and R eye visibility (cf with FIT in Fig4D of main paper) D) MEG DFI about stimulus or choice feedback or feedforward and in correct vs error trials (cf with FIT in Fig. 3D-F) C) Spikes DFI for the thalamocortical pathway and tactile- or visual-discriminative stimulus set (cf with FIT in FigS8E-I). P-values: 2-tailed paired t-test. *: p<0.05, **: p<0.01.

coded in the past activity of the receiver as information transmitted from a sender. This general concern would apply in general to all measures that compute time-lagged cross-correlations of activity across areas [203]. In the following, we consider briefly some possible methods that have been used to infer feature-specific information transfer but that do not consider the Wiener-Granger Causality principle.

One possibility would be to measure the presence and timing of feature information (using mutual information between the feature feature and the activity of the individual area at each instant of time, as in Eq. 3.S5) and then inferring transfer of feature information from X to Y if information about S arises first in X and

Y. Inferring processing hierarchies on the basis of response selectivity latencies is a long-established practice in neuroscience [11, 197, 204]. However, the presence of time-lagged information in two areas does not mean that the information in the second area comes from the first area. Indeed, in the simulations to test FIT we created several such simulated scenarios of information present in each area with a different timing but without actual communication between the areas (Fig. 3.S7C, Section 3.8.2.6), and we created non-parametric tests to rule out this possibility based on FIT measures (Section 3.8.1.7). Thus, differential response latencies can be used to hypothesize the presence of processing hierarchies but not to prove transfer of information between specific nodes of the putative information processing network.

Another possibility would be to use PID to measure the presence of shared (or redundant) feature information encoded in both X and Y at different temporal lags. PID can specifically isolate only the information about a feature that is the same, e.g., redundantly encoded, in X and Y. Thus, measuring time-lagged shared information goes beyond computing a simple time-lagged correlation between the amount of reach-to-grasp information in X and Y, which would not consider whether the time-lagged information content is the same. However, this measure would not discount the presence of the same information in the past activity of Y, and it would thus be prone to detecting false communications in case the information was already present in the past of Y and thus could not have come from X. (In other words, it would erroneously identify information already encoded in the past activity of the receiver as information transmitted from a sender.) These types of problems of not discounting the past have been illustrated and discussed extensively in the Granger Causality literature. These considerations apply to a previous study [205] which used PID to attempt to define feature-specific information transmission using the so-called Intersection Information [104], computing information shared between the past activity of the sender and present activity of the receiver (not considering the information already present in the past activity of the receiver).

3.8.4.4 Computational resources

Each of the simulations in Figures 3.2, 3.S3, 3.S7, 3.S8, 3.S15, 3.S14, 3.S6, 3.S4 ran in approximately 30 minutes on a personal computer equipped with an Intel i7-10510U processor (4x 1.80GHz CPUs) and 16Gb of RAM, running Windows 10, using MATLAB R2021a. Simulations in Fig. 3.S5 took approximately 3 hours on the same machine.

Real neural dataset analyses ran on a server with an AMD Ryzen Threadripper 3970X processor (32x 3.7GHz CPUs) and 256Gb of RAM, running Ubuntu 18.04, using MATLAB R2019b. The EEG and MEG analyses ran in parallel (using the Parallel Computing Toolbox) over participants or links in the visual cortical network, respectively. Each analysis of the full real datasets (across all participants and experimental sessions) took 12-28 hours depending on the usage of the server.

Our MATLAB codes to compute Feature-specific Information Transfer are provided with this submission and are released under the MIT license.

Chapter 4

Information flow between motor cortex and striatum reverses during skill learning

The content of this Chapter was submitted for publication, and is currently under submission and being revised [51]. The analyses presented here correspond to those presented in the first submitted version of the paper prior to being revised.

4.1 Introduction

Skill learning is the process in which movements are selected and produced more consistently [206, 207] and automatically [208] with training. Skill learning is typically associated with the stabilization of neural activity patterns during execution of the learned skill [209–212], and the increased coordination of such activity patterns across the cerebral cortex, basal ganglia, and cerebellum [213–217]. In particular, the primary motor cortex (M1) and the dorsolateral striatum (DLS) – a region within the basal ganglia that is directly innervated by M1 – are thought to be central to skill learning [35, 128, 214, 218, 219].

Current evidence supports two opposing models of how M1 input to the DLS contributes to skill learning. The first model proposes that the importance of M1 input to the DLS increases during learning, playing a central role in shaping DLS activity patterns that control skilled movements. Supporting evidence includes the potentiation of synapses specifically from M1 neurons active during movement production onto DLS neurons during learning14, the emerging coordination of M1 and DLS movement-related activity during learning [214, 219], and demonstrations that learning requires plasticity at glutamatergic synapses in the DLS, which originate from either cortex or thalamus [35, 128, 219–221].

The second model proposes that the importance of M1 input to the DLS diminishes during learning, resulting in DLS activity patterns that control skilled movements without reliance on M1 input. Supporting evidence includes demonstra-

tions that after learning a skill, M1 lesion or inactivation has little impact on skill production [20, 222]. Additionally, once a skill is learned, DLS activity can encode skilled movement kinematics after silencing DLS-projecting M1 neurons [223]. In this model, M1 input to the DLS during initial learning provides instructive signals that drive plasticity within the DLS, resulting in M1-independent control of skilled movements in the DLS [224].

To test these models, we measured how M1 and DLS encode and transfer information during learning of a reach-to-grasp skill in rats. The reach-to-grasp skill is an ethologically-relevant, evolutionarily-conserved behavior [225, 226] that requires both M1 and DLS activity to produce [214, 227–229] – making it particularly well-suited to examine the evolution of M1 and DLS interactions during learning. We developed and utilized an information theory framework to isolate the components of neural activity across M1 and DLS that encode and transmit information about the reach-to-grasp movement. We find that M1 input to the DLS does not strictly increase or decrease during skill learning, but rather the content of such input evolves. Our results support a hybrid model that reconciles previous studies by demonstrating that a bidirectional increase in overall information propagation between M1 and DLS during skill learning can cooccur with a reversal in the direction of behaviorally relevant information flow, from M1-to-DLS during naive movements to DLS-to-M1 during skilled movements.

4.2 Results

Neural signals, including local field potentials (LFP) and spiking activity, were monitored simultaneously in M1 and DLS in eight adult rats undergoing multi-day training of a reach-to-grasp skill (Figure 4.11a; data from six animals was included in previous work [128]). The reach-to-grasp skill has been used extensively in rodents to study how the brain controls movement [214, 215, 218, 227, 230–232]. On each day of training, rats performed 50-150 trials in a custom-built behavioral box [233]. The reach-to-grasp skill requires rats to reach through a small window in the behavioral box to grasp and retrieve a food pellet. With training, rats became more successful in retrieving the pellet and a range of kinematic features evolved (Figure 4.1b-d; Figure 4.S1; Table 4.S1). To find the kinematic features most relevant to success, we fit a generalized linear model to predict single-trial success during naive or skilled days from each of ten different reach features. We found that maximum reaching velocity and total reaching trajectory length were the best predictors of success (Figure 4.1e) and predicted success better for skilled movements compared to naive movements, indicating that these features captured learning- related changes in the reach-to-grasp movement related to improvements in success (Figure 4.1e). In this work, we computed information carried by M1 and DLS neural signals about these selected kinematic features. Importantly, maximum reaching velocity and reaching trajectory length both characterize the outward reaching component of the reaching and grasping action, which has been specifically associated with emerging

M1 and DLS coordinated activity. We will refer to information about these features as reach-to-grasp information.

4.2.1 Neural signal amplitude and encoding of reach-to-grasp information are dissociable.

We first sought to understand how M1 and DLS neural signals encoded reach-to-grasp information. To measure reach-to-grasp information we computed Shannon information encoded by single-trial neural signals about the selected kinematic features. Shannon information is a non-parametric measure that captures both the linear and non-linear encoding of information [75, 86] . Information was computed at individual time points throughout the reaching movement, aligned across trials to "pellet touch", i.e., the time in which the rat touches the reaching target, a food pellet, on each trial (Figure 4.1f). Trials were better aligned around the time of pellet touch, compared to other time points such as movement onset, with lower and more time-localized variance in hand position and velocity (Fig. 4.S2), indicating that trial alignment to pellet touch was better suited for studying the temporal profiles of information. Both LFP signals and spiking activity in M1 and DLS contained significant information about both naive and skilled movements with greater information for skilled, compared to naive, movements, suggesting a tighter relationship between neural activity and movement for skilled movements, consistent with previous work [230] (Figure 4.S3). The rationale for analyzing both spiking activity and LFP is that spiking activity represents a more direct measure of local neuron activity, while the advantage of analyzing LFP signals is that LFPs are more stable than spiking activity [234]. Reach-to-grasp information in LFP signals was localized to low (¡5Hz) frequencies, a frequency range previously associated with movement generation [57, 235] (Figure 4.S4).

Whether a single neuron or a neural population (e.g., measured by LFP) is involved in controlling movement is typically assessed by looking for a consistent change in the amplitude of the spiking rate or LFP signal during movement. We therefore first examined whether the amount of reach-to-grasp information encoded by a neural signal corresponded to the amplitude of that signal during movement – defined as the magnitude of deviation in a neural signal, either positive or negative, from baseline. Surprisingly, we found a dissociation between the encoding of reach-to-grasp information and neural signal amplitude, with peak reach-to-grasp information often encoded 50-500ms before maximum signal amplitude (Figure 1g&i). The intuitive interpretation of this timing difference is that the amplitude of a neural signal 50 to 500ms prior to the maximum signal amplitude covaried with trial-to-trial variations in movement kinematics while the maximum signal amplitude was invariant across trials (Figure 4.1h). This is consistent with previous work demonstrating that the largest amplitude component of M1 activity does not encode information about movement kinematics and instead simply reflects when a movement is occuring [236]. A similar dissociation has also been reported in sensory

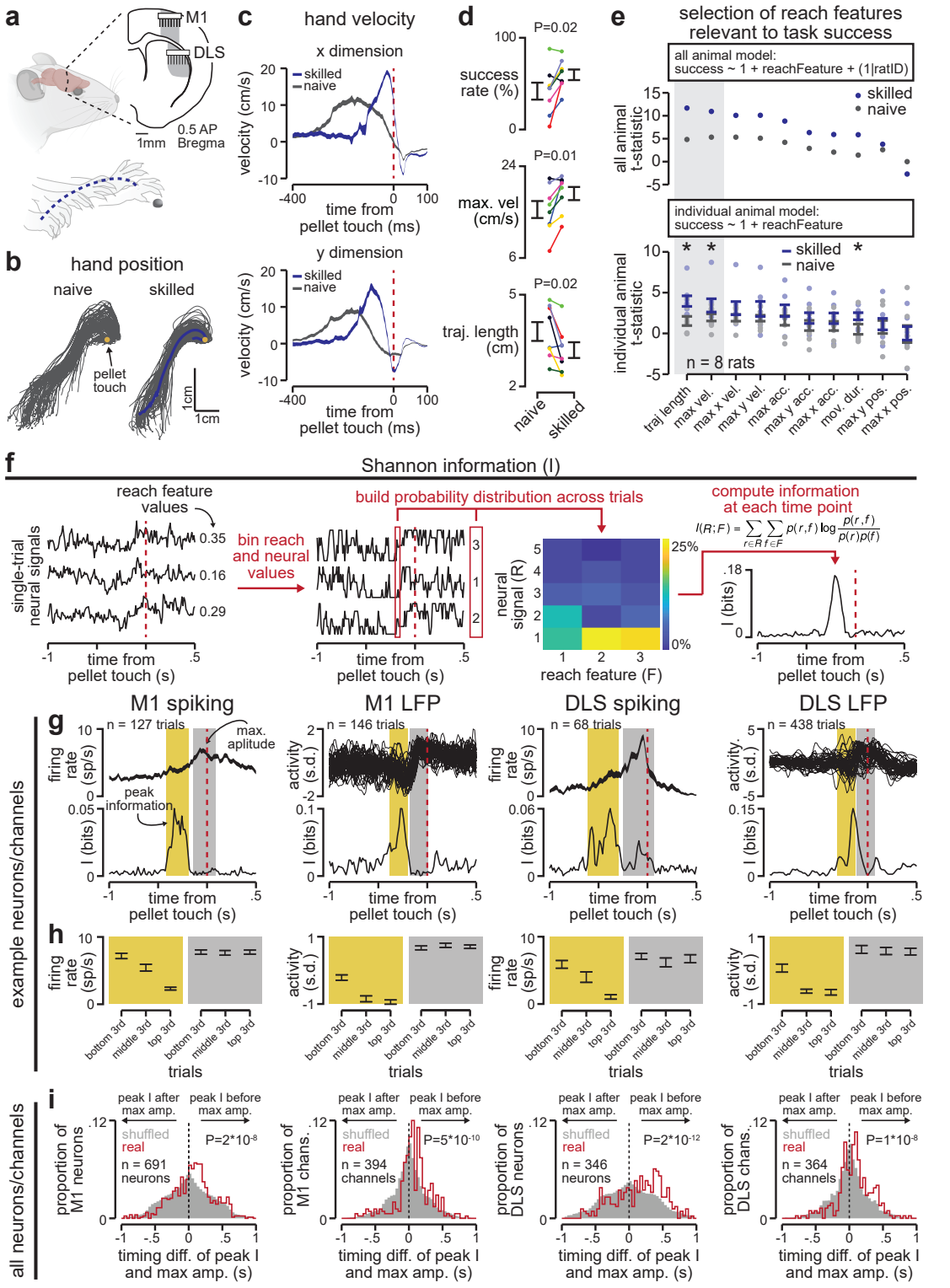

Figure 4.1: Neural signal amplitude and encoding of reach-to-grasp information are dissociable. (a) Depiction of reach-to-grasp task with neural recordings in M1 and DLS.(b) Example reach trajectories from naive and skilled days in example animal (individual trials in gray overlaid with mean in color).(c) Example velocity profiles from naive and skilled days in example animal (width of lines represents mean ± SEM).(d) Comparison of mean success rate and reach features from naive and skilled days across animals (success rate: naive: 41.4 ± 0.1%, skilled: 59.4 ± 0.1%, t(7)=-2.9, P=0.02; maximum velocity: naive: 15.6 ± 1.7cm/s, skilled: 18.7 ± 1.3cm/s, t(7)=-3.4, P=0.01; trajectory length: naive: 3.8 ± 0.3cm, skilled: 3.2 ± 0.3cm, t(7)=2.9, P=0.02; paired-sample t-test, n = 8 animals, individual animals in color and mean ± SEM in black).

(e) Comparison of t-statistics for models fit to predict single-trial success on naive or skilled days from different reach features. Top: model fit across animals. Bottom: model fit within individual animals (* denotes $p<0.05$). Gray shading denotes reach features selected for further analysis. (f) Schematic of Shannon information computation. (g) Comparison of trial-averaged spiking activity or single trial LFP signals (top) and corresponding encoded reach-to-grasp information about maximum reaching velocity (bottom) across example neurons and LFP channels in M1 and DLS. Yellow shaded area represents time bins around peak of reach-to-grasp information encoding and gray shaded area represents time bins around maximum neural signal amplitude. (h) Comparison of mean firing rate for example neurons and mean neural activity for example LFP channels during yellow shaded time bins, representing the time of peak of reach-to-grasp information encoding about maximum reaching velocity, and gray shaded time bins, representing the time of maximum neural signal amplitude. Each set of time bins is further separated according to the maximum reaching velocity of each trial (i.e., bottom 3 rd represents the third of trials with the lowest maximum velocity, middle 3rd represents the third of trials with the middle values of maximum velocity, and top 3rd represents the third of trials with the highest maximum velocity). (i) Distributions of timing differences between the time of peak reach-to-grasp information encoding and maximum trial-averaged neural signal amplitude, across all M1 or DLS neurons and all M1 or DLS LFP channels. Each distribution combines both naive and skilled days and both information about maximum reaching velocity and trajectory length. Real distribution is in red outline and shuffled distribution is in gray. P values denote comparison between real and shuffled distributions (M1 spiking: $P=2 \times 10^{-8}$, M1 LFP: $P=5 \times 10^{-10}$, DLS spiking: $P=2 \times 10^{-12}$, DLS LFP: $P=1 \times 10^{-8}$; two-sample Kolmogorov-Smirnov test).

systems, in which the peak information that a neuron encodes about a stimulus does not always correspond to values of that stimulus that result in maximum fing-rate responses [237].

4.2.2 Timing of reach-to-grasp information encoded in M1 and DLS with skill learning.

We next sought to identify learning-related changes in how M1 and DLS neural signals encode reach-to- grasp information. Given that during the production of a reach-to-grasp movement the encoding of reach- to-grasp information is dissociable from neural signal amplitude, we compared how both neural signal amplitude and reach-to-grasp information encoding evolved during learning. We first consider the simplest characterization of the relationship between the information encoding in M1 and DLS, that is the timing of information encoding. Although some earlier studies have considered the differences in information or response latencies across areas as an indication of the hierarchy of information processing [197], timing of information encoding cannot be taken per se as an indication of the direction of communication between areas (see Chapter 3 for examples of when timing does not reveal information processing hiearchies). However, studying how the timing

of information encoding differs between ares and across stages of learning is a first useful characterization of the dynamics of information processing between areas and how it changes with learning.

We found that the timing, relative to pellet touch, in which neural signal amplitude deviated from baseline did not consistently change from naive to skilled movements in either M1 or DLS (Figure 4.2a; Figure 4.S5a). In contrast, the timing in which M1 and DLS neural signals first encoded reach-to-grasp information shifted in opposite directions – with encoding in M1 shifting later and encoding in DLS shifting earlier, even if the shift in M1 was stronger (Figure 4.2b, Figure 4.S5b). These changes in the timing of reach-to-grasp information encoding were also captured by simpler linear regression models (Figure 4.S6). These results were consistent with a dissociation between the amplitude of a neural signal and the amount of behaviorally relevant information encoded by that signal and indicated that directly measuring how neural signals encode reach-to-grasp information was necessary to observe learning-related changes in M1 and DLS neural activity.

4.2.3 Cross-area timing relationship of shared reach-to-grasp information reverses during skill learning.

We next sought to characterize whether the same reach-to-grasp information was encoded at different times in M1 and DLS neural activity, or whether distinct reach-to-grasp information was encoded by each area. If the information was distinct, this would suggest that M1 and DLS may have separate, parallel learning-related mechanisms. Alternatively, if the same reach-to-grasp information was encoded in M1 and DLS neural activity at different times, this would suggest that reach-to-grasp information may flow between M1 and DLS. We used Partial Information Decomposition [100, 103, 110] (PID) to measure the presence of shared (or redundant) reach-to-grasp information encoded in both M1 and DLS neural signals at different temporal lags (Figure 4.3a). Unlike simpler measures of redundancy [87, 89], PID can specifically isolate the information about a reach feature that is the same, i.e., redundantly encoded, in distinct neural signals. Thus, measuring time-lagged shared information goes beyond computing a simple time-lagged correlation between the amount of reach-to-grasp information in M1 and DLS, which would not consider whether the information content is the same (Figure 4.S7). Given the low probability of recording from directly coupled pairs of neurons across areas, for the remaining cross-area analysis we focus on LFP signals. The relative stability of LFP signals, compared to spiking activity [234], allowed us to combine trials across naive or skilled days to perform more robust computations of cross- area information. To address challenges in interpreting LFP signals from non-laminar structures such as the striatum [28, 238] , we both utilized a previously established local referencing scheme to minimize the risk of volume conducted signals [128, 214] , as well as replicated our results with limited pairs of coupled M1 and DLS neurons recorded on individual naive and

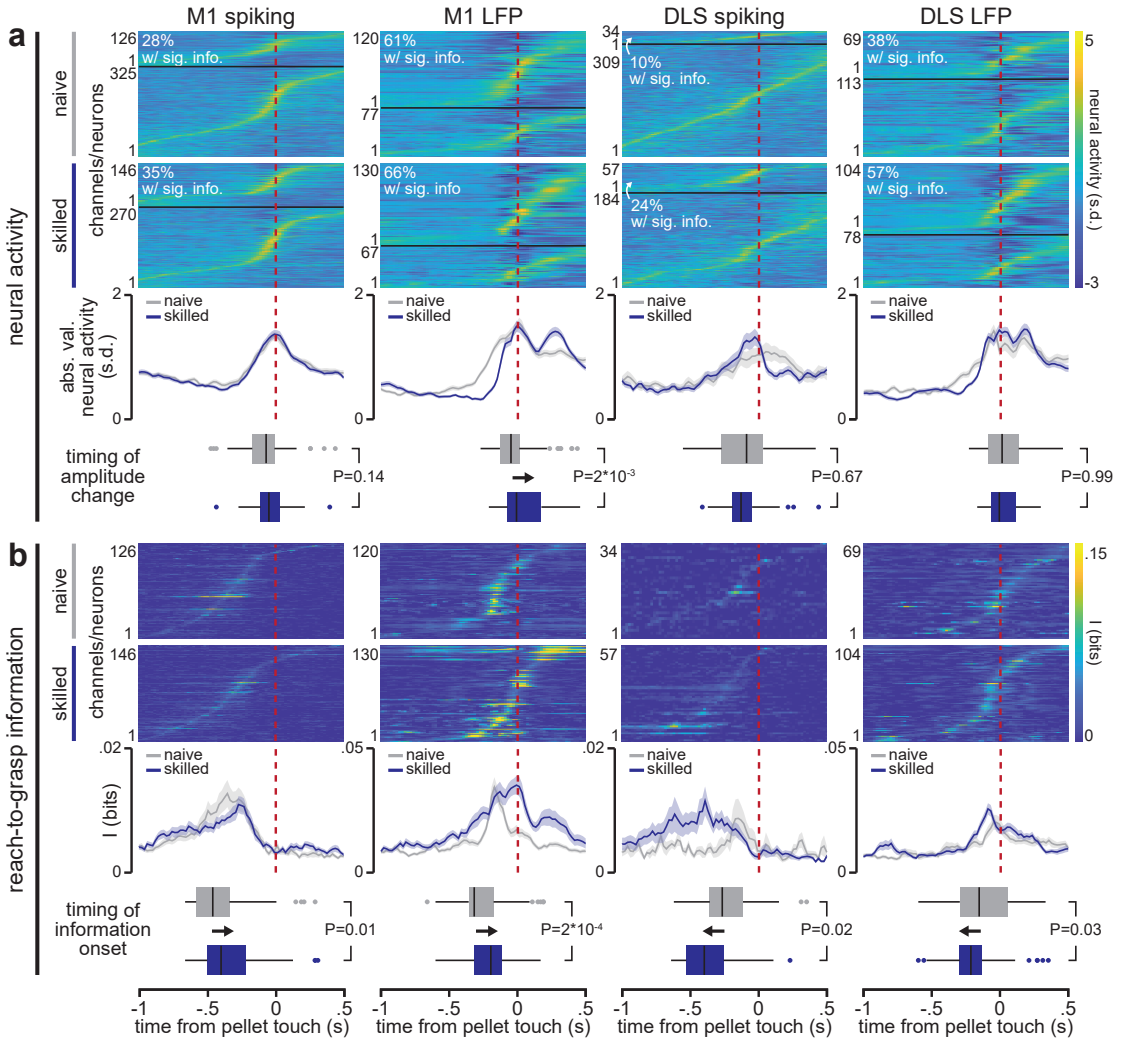

Figure 4.2: Timing of reach-to-grasp information encoding in M1 and DLS changes during skill learning. (a) Top: trial-averaged spiking activity and LFP signals, separated by neurons or channels encoding or not encoding reach-to-grasp information about maximum reaching velocity during naive or skilled movements, sorted by timing of maximum spiking activity or LFP signal. Middle: time courses of average absolute value of spiking activity or LFP signals during naive or skilled movements, across neurons or channels encoding information about maximum reaching velocity (mean ± SEM). Bottom: box plot representing the timing of when LFP signals or spiking activity initially deviate from baseline during naive or skilled movements, across channels or neurons encoding information about maximum reaching velocity (M1 spiking: P=0.14, M1 LFP: P=2×10^{-3} , DLS Spiking: P=0.67, DLS LFP: P=0.99; two-sided Wilcoxon rank sum test). (b) Top: reach-to-grasp information encoded by spiking activity or LFP signals about maximum reaching velocity during naive or skilled movements, neurons or channels sorted by timing of peak information. Middle: time course of information about maximum reaching velocity encoded by spiking activity or LFP signals during naive or skilled movements (mean ± SEM). Bottom: box plot representing the timing of when neural signals first encode information about maximum reaching velocity during naive or skilled movements (M1 spiking: P=0.01, M1 LFP: P=2×10^{-4} , DLS Spiking: P=0.02, DLS LFP: P=0.03; two-sided Wilcoxon rank sum test).

skilled days.

We found that, during naive movements, shared reach-to-grasp information was originally encoded in M1 neural signals and then emerged, on average, 50-100ms later in DLS – consistent with M1-to-DLS information flow. During skilled movements this temporal relationship reversed, with shared reach-to- grasp information originally encoded in DLS neural signals and then emerging 50-100ms later in M1 – consistent with DLS-to-M1 information flow (Figure 4.3b,e; Figure 4.S8b,e). Across pairs of M1 and DLS LFP channels, the temporal delays corresponding to maximum shared reach-to-grasp information encoded across areas shifted from a distribution centered on M1-to-DLS delays during naive movements to centered on DLS-to-M1 delays during skilled movements (4.3c; Figure 4.S8c). This reversal was robust across animals when comparing the change in percentage of channel pairs with maximum shared information encoded with either M1-to-DLS or DLS-to-M1 temporal delays from naive to skilled movements (Figure 4.3d; Supplemental 4.S8d). Consistent with the reversal observed in LFP signals, we found that shared reach-to-grasp information encoded between M1 and DLS neuron pairs peaked predominantly with M1-to-DLS temporal delays during naive movements and DLS-to-M1 delays during skilled movements (Figure 4.S9).

We next sought to determine whether it was necessary to isolate the behaviorally informative component of neural activity to observe a reversal in the flow of activity between M1 and DLS during skill learning. Alternatively, was there also a reversal in the ability to predict the overall neural activity of one brain area from the other brain area, without isolating the behaviorally relevant components of activity? To determine this, we computed time-lagged Shannon information between overall M1 and DLS neural activity (Figure 4.S10a). In contrast to the reversal in the cross-area timing relationship of shared reach-to-grasp information encoding, we observed an increase in time-lagged Shannon information between overall M1 and DLS neural activity in both the M1-to-DLS and DLS-to-M1 direction during skill learning, without evidence of a reversal (Figure 4.S10b-e). This suggested that breaking down overall neural activity into the behaviorally informative components was necessary to capture distinct learning-related changes in M1 and DLS neural activity.

4.2.4 The flow of reach-to-grasp information reverses during skill learning.

Given the mounting evidence that changes in reach-to-grasp information processing during learning were not captured by standard methods to measure learning-related changes in overall neural activity – including changes in the amplitude of neural signals or the overall coordination of neural signals across areas – we next sought to directly measure and compare the flow of reach-to-grasp informative neural activity and the general propagation of overall neural activity between areas. To equitably compare the cross-area flow of reach-to-grasp informative neural activity – which we will refer to as reach-to-grasp information flow – and the cross-area propagation of

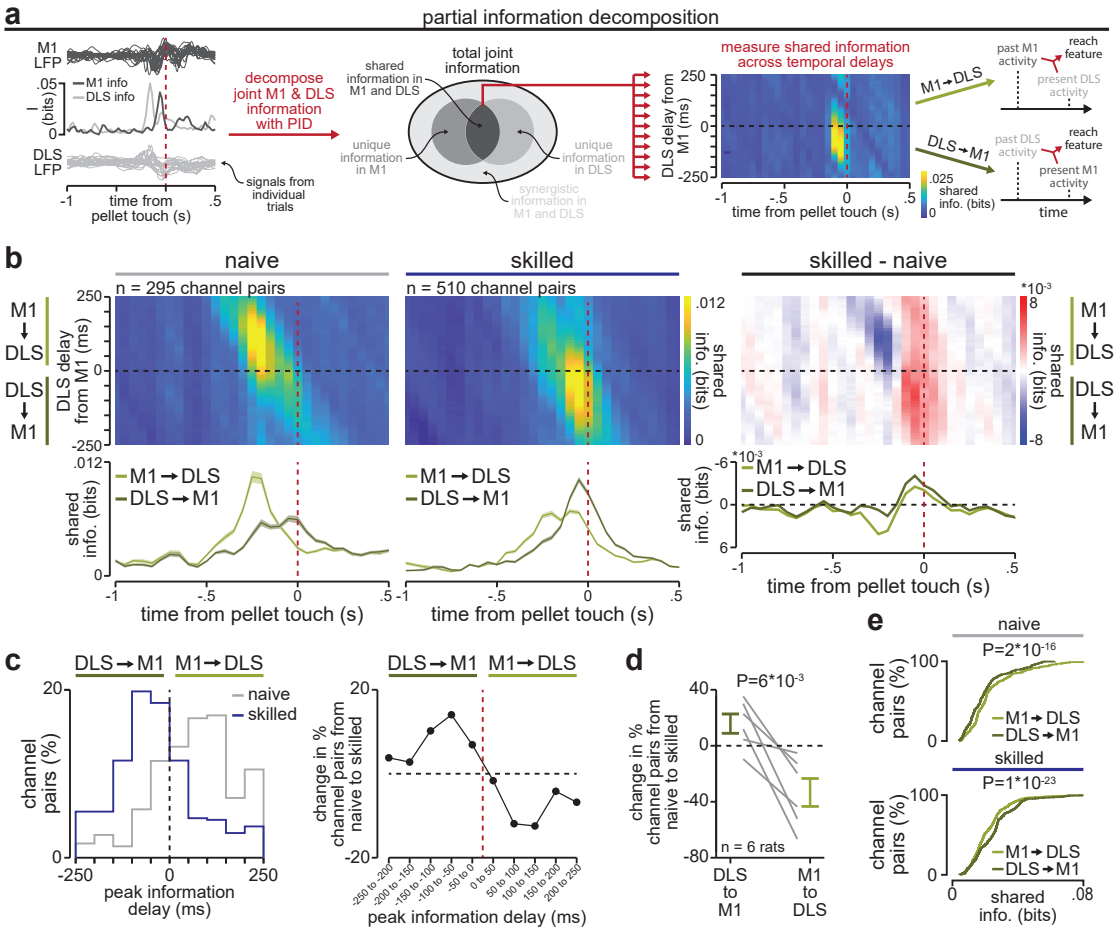

Figure 4.3: Cross-area timing relationship of shared reach-to-grasp information reverses during skill learning. (a) Schematic of partial information decomposition computation. (b) Top: mean time-lagged shared information in M1 and DLS LFP signals about maximum reaching velocity across temporal delays during naive movements, skilled movements, and the difference. Bottom: shared information averaged over positive (M1-to-DLS) and negative (DLS-to-M1) temporal delays during naive movements, skilled movements, and the difference. (c) Left: distribution of temporal delays corresponding to peak shared information about maximum reaching velocity across M1 and DLS LFP channel pairs for naive (gray) and skilled (blue) movements. Right: difference between naive and skilled distributions. (d) Change in percentage of M1 and DLS LFP channel pairs with peak shared information about maximum reaching velocity with a positive (M1-to-DLS) or negative (DLS-to-M1) temporal delay from naive to skilled movements, in each animal ($t(5)=-4.5$, $P=6\times10^{-3}$, two-sample t-test, n = 6 animals, 2 animals without significant information measured on either naive or skilled days). (e) Cumulative density functions comparing peak shared information about maximum reaching velocity across M1 and DLS LFP channel pairs, combined over positive (M1-to-DLS) or negative (DLS-to-M1) temporal delays. Top: naive movements ($P=2\times10^{-16}$, Wilcoxon signed rank test). Bottom: skilled movements ($P=1\times10^{-23}$, Wilcoxon signed rank test).

overall neural activity – which we will refer to as neural activity propagation – we utilized two analogous measures based on the Wiener-Granger causality principle [91, 239]. The Wiener-Granger causality principle states that the directed flow from signal X to signal Y can be measured as the ability to predict the current signal Y from the past of signal X, after discounting the self-prediction from the past of signal Y. Therefore, to compute neural activity propagation we measured the ability to predict neural activity values of a putative receiving area from the past neural activity of a putative sending area, discounted by the ability to self-predict the activity from the past of itself (Figure 4.4a). This method, termed transfer entropy (TE) [138], has been used extensively to study cross-area brain activity [31, 85, 166, 169] and extends the measure of time-lagged Shannon information (Figure 4.S10) by incorporating the Wiener-Granger causality principle. Correspondingly, to characterize reach-to-grasp information flow, we measured the reach-to-grasp information shared between the present activity of a putative receiving area and the past activity of a putative sending area, that is also unique with respect to the information encoded in the past activity of the receiving area (Figure 4b). We developed this method to extend the shared information results computed with PID (Figure 4.3; Figure 4.S8) by including the Wiener-Granger causality principle. We term this new measure Feature-specific Information Transfer (FIT; derivation in methods). In model simulations, we demonstrate that the established measure of neural activity propagation (TE) is insensitive to the content of neural activity transmitted between areas while our measure of reach-to-grasp information flow (FIT) is specifically sensitive to the transfer of neural activity encoding reach-to-grasp information (Figure 4.S11).

We compared reach-to-grasp information flow to both neural activity propagation and a common measure of overall non-directed functional connectivity, LFP coherence. We labeled M1 and DLS as either the receiving or sending area based on the direction of maximum neural activity propagation or reach-to- grasp information flow. We found that reach-to-grasp information flow reversed, flowing primarily from M1-to-DLS during naive movements and from DLS-to-M1 during skilled movements (Figure 4.4c&d; Figure 4.S12c&d). This was consistent with the reversal in the cross-area timing relationship of shared reach-to-grasp information (Figure 4.3; Figure 4.S8). In contrast, neural activity propagation in both directions, as well as LFP coherence, increased from naive to skilled movements (Figure 4.4e-h; Figure 4.S12e-h). This was consistent with the bidirectional increase in time- lagged Shannon information between overall M1 and DLS neural activity (Figure 4.S10). These results demonstrated that isolating the component of M1 and DLS communication specifically relevant to the reach-to-grasp movement was required to identify a reversal in behaviorally relevant information flow. Interestingly, neural activity propagation also increased prior to movement during skill learning, suggesting that learning-related increases in coupling between M1 and DLS can be captured even during non-movement periods, consistent with prior work showing that increases in LFP coherence between M1 and DLS during sleep track skill learning [128] . In sum, by specifically isolating the components of neural activity across M1 and DLS that

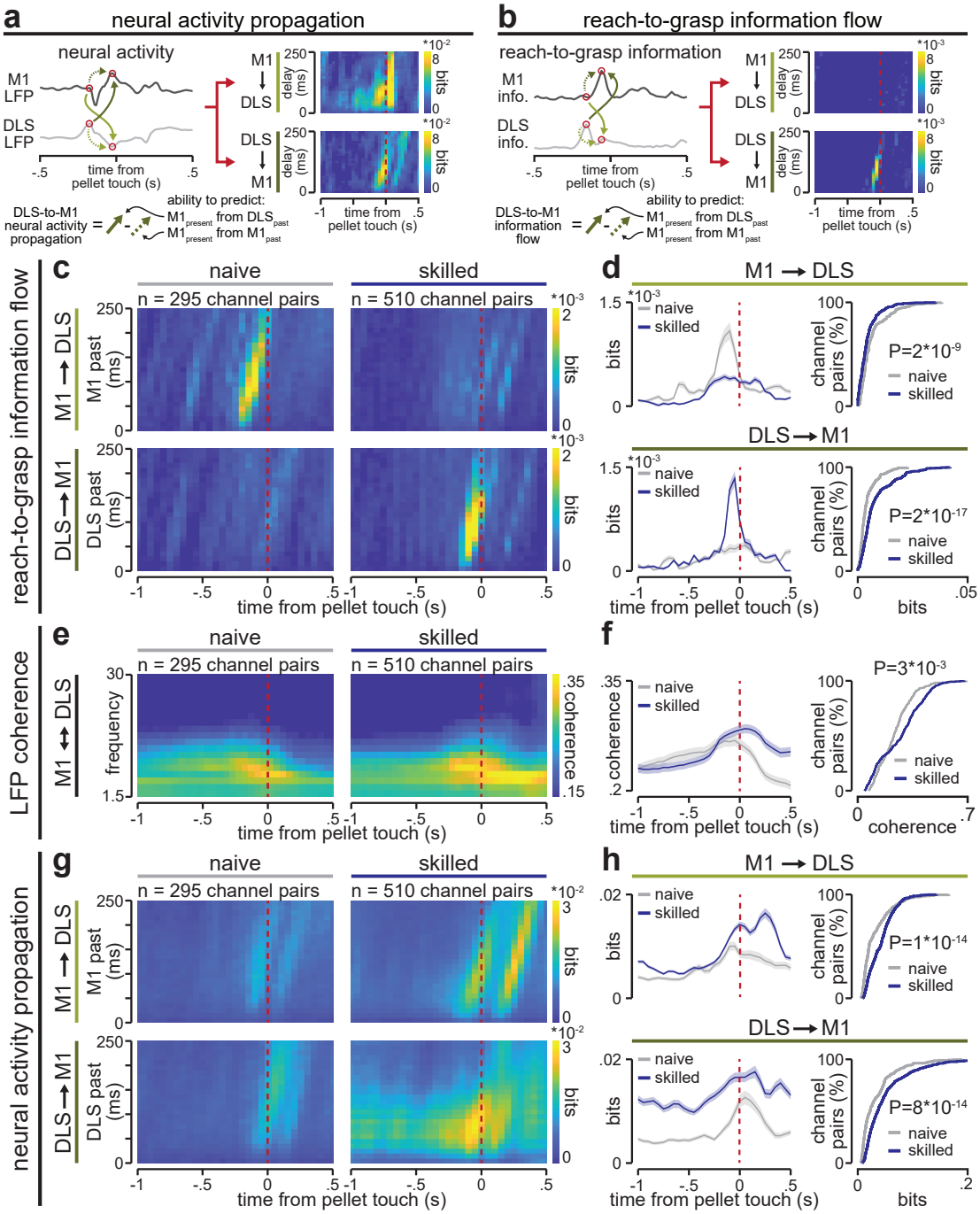

Figure 4.4: The flow of reach-to-grasp information reverses during skill learning. (a) Schematic of neural activity propagation computation. (b) Schematic of reach-to-grasp information flow computation. (c) Mean reach-to-grasp information flow about maximum reaching velocity between M1 and DLS LFP signals during naive and skilled movements. Top: M1-to-DLS temporal delays. Bottom: DLS- to-M1 temporal delays. (d) Left: time course of mean reach-to-grasp information flow about maximum reaching velocity between M1 and DLS LFP signals during naive and skilled movements, across M1-to-DLS temporal delays (top) and DLS-to-M1 temporal delays (bottom). Right: cumulative density functions comparing distributions of peak reach-to-grasp information flow about maximum reaching velocity between M1 and DLS LFP signals during naive and skilled movements, across M1-to-DLS temporal delays (top; $P=2 \times 10^{-9}$, two-sided Wilcoxon rank sum test) and DLS-to-M1 temporal delays (bottom; P=2 × 10−17 ; two-sided Wilcoxon rank sum test). (e) Mean LFP coherence between M1 and DLS LFP signals, across frequencies during naive and skilled movements, computed over same LFP channel pairs as (c).

(f) Left: time course of mean 3-15Hz LFP coherence between M1 and DLS LFP signals during naive and skilled movements. Right: cumulative density functions comparing distributions of peak 3- 15Hz LFP coherence between M1 and DLS LFP signals during naive and skilled movements ($P=3 \times 10^{-3}$, two-sided Wilcoxon rank sum test). (g) Mean neural activity propagation between M1 and DLS LFP signals during naive and skilled movements, computed over same LFP channel pairs as (c). Top: M1-to-DLS temporal delays. Bottom: DLS-to-M1 temporal delays. (h) Left: time course of mean neural activity propagation between M1 and DLS LFP signals during naive and skilled movements, across M1-to-DLS temporal delays (top) and DLS-to-M1 temporal delays (bottom). Right: cumulative density functions comparing distributions of peak neural activity propagation between M1 and DLS LFP signals during naive and skilled movements, across M1-to-DLS temporal delays (top; $P=1 \times 10^{-14}$, two-sided Wilcoxon rank sum test) and DLS- to-M1 temporal delays (bottom; $P=8 \times 10^{-14}$; two-sided Wilcoxon rank sum test).

encoded and transmitted information relevant to the reach-to- grasp movement, we uncovered an unexpected reversal in the flow of behaviorally relevant information from M1-to-DLS during naive movements to DLS-to-M1 during skilled movements.

4.3 Discussion

Our findings help reconcile paradoxical evidence suggesting that the importance of M1 inputs to the DLS can both increase and decrease during skill learning. We find that skill learning is not accompanied by either a simple increase or decrease in M1 input to the DLS. Instead, the nature of M1 input to the DLS evolves in how it contributes to information flow. We show that a bidirectional increase in overall information propagation between M1 and DLS during skill learning – consistent with both the potentiation of M1 inputs onto DLS neurons [218] and the emerging cross-area coordination of movement-related activity during learning [214, 219] – can cooccur with a reversal in the direction of behaviorally relevant information flow – consistent with a decreased reliance on M1 input to encode details of the skilled movements in the DLS during learning [223]. We propose a hybrid model in which, as learning progresses, the role of M1 input in conveying specific movement information to the DLS diminishes, while an overall excitatory M1 input to the DLS is maintained. In parallel, a new role emerges for DLS input that transfers information relevant to the skilled movement to M1.

4.3.1 The role of information flow during learning

What are the potential roles of reach-to-grasp information flow that could help explain a reversal between M1 and DLS during skill learning? One potential role of reach-to-grasp information flow is to instruct the adjustment of future behavior based on reward, both for naive movements during initial learning and for skilled movements if task parameters change. During naive movements, a role for M1-to-

DLS reach-to-grasp information flow could be to instruct plasticity at glutamatergic inputs to the DLS. This would be consistent with evidence that glutamatergic inputs to the DLS – including those originating from motor cortex – are potentiated during learning [222, 240–242], that disrupting NMDA receptor function in the DLS can disrupt learning [35, 214, 219], and that silencing DLS projecting M1 neurons disrupts learning [224]. During skilled movements, DLS-to-M1 information flow could play a role in guiding plasticity that is required to adjust skilled movements if task parameters change. This would be in line with evidence that adjustments to a skilled reaching and grasping action in rats produced in response to a change in food pellet location are reflected in coordinated changes across M1 and DLS neural signals [243]. An analogous role for basal ganglia input to cortex in guiding the adjustment of skilled movements exists in adult zebra finches, where input from area X in the basal ganglia to LMAN, a cortical analog region, promotes exploratory motor variability is required for reinforcement-driven adjustments to song production [244].

Another potential role of behaviorally relevant information flow between M1 and DLS is to combine sub-components of the reach-to-grasp action that may be controlled separately by M1 or DLS. M1 is strongly associated with the control of fine dexterous movements, such as grasping. Damage to M1 chronically disrupts such fine movements [214, 228]. In contrast, the DLS has been linked to the control of learned, non-dexterous, skilled movements, even after removal of M1 [223]. Therefore, one possibility is that DLS-to-M1 information flow emerges during learning to properly combine a DLS-controlled outward reach with an M1-controlled grasp. A prediction for this model would be that M1 inactivation specifically disrupts the grasping component of the skilled reaching and grasping action. However, acute M1 inactivation has been shown to completely block the production of a reach-to-grasp skill, even when applied after learning [227]. A potential explanation is that, although behaviorally relevant information flow from M1-to-DLS decreases during skill learning, an overall permissive excitatory input from M1 to DLS may still be required, for example to provide a generic excitatory drive that may initiate movement encoding in the DLS, rather than to carry specific reach to grasp information.

4.3.2 Beyond the Motor Cortex and Striatum

As M1 and DLS are part of a highly interconnected motor network in the brain, it is likely that inputs to M1 and DLS from other brain areas contribute to the observed information dynamics. A decrease in behaviorally relevant input to M1 during learning, coupled with an increase to the DLS, could explain the shift in origin of reach-to-grasp information flow from M1 to DLS. A potential explanation for this type of M1 to DLS shift in input is that, during reach-to-grasp skill learning, movements transition from goal-directed and sensory-driven to stimulus-response and automatic. Goal directed movements rely on the cortical integration of sensory evidence to guide movement [16, 245, 246], suggesting that inputs carrying

sensory information to M1 may be critical to initiate and plan naive movements. As learning proceeds, the requirement for rich sensory information may decrease, thus diminishing the role of inputs to M1. Instead, inputs to the basal ganglia and DLS, areas which are classically associated with control of stimulus-response and automatic movements [19, 247, 248] may increase. Potential inputs to the DLS include non-M1 cortical regions, consistent with evidence linking anterior cortical regions (e.g., M2) to the control of learned actions [241], or thalamic regions, consistent with work showing that silencing DLS-projecting thalamic neurons can disrupt learned actions [224]. This shift in the neural control of a behavior occurring within cortico-basal ganglia loops during learning is consistent with classic models of motor sequence learning [249]. Moreover, it is important to note that there are no direct connections from DLS to M1. Thus, it is possible that the observed increase in the flow of information from DLS to M1 is due to a strengthening of polysynaptic striatal projections to the cortex through the thalamus [128, 250]. Consequently, skill learning could be associated with the modulation of different pathways in the cortico-striatal-thalamo-cortical loop.

4.3.3 Isolating behaviorally relevant neural communication

Established analytical methods to study multi-area interactions – from simpler measures based on correlations or synchrony to more complex measures using dimensionality reduction [217, 251] or based on the Wiener-Granger causality principle [138, 160] – typically do not consider ongoing behavior (but see Ref. [252]). To overcome this limitation, we developed an analytical framework based on information theory that isolates the components of neural activity across brain areas that encode and transmit information relevant to a specific ongoing behavior. We demonstrate the utility of this approach by uncovering a reversal in behaviorally relevant information flow during skill learning that is unobservable using standard analytical methods. These results challenge the assumption that measuring the overall propagation of neural activity between brain areas is sufficient to understand the dynamics of neural communication relevant to ongoing behavior. Our results suggest that isolating behaviorally relevant components of neural activity across brain areas is a valuable approach for understanding the function of distributed brain networks [249].

4.4 Supplementary Material

4.4.1 Methods

4.4.1.1 Animal care and surgery

This study was performed in strict accordance with guidelines from the USDA Animal Welfare Act and United States Public Health Science Policy. Procedures were in accordance with protocols approved by the Institutional Animal Care and Use Committee at the San Francisco Veterans Affairs Medical Center. Experiments were

conducted with 8 male Long-Evans rats (approximately 12–16 weeks old) housed under controlled temperature and a 12 hr light/12 hr dark cycle with lights on at 6:00 a.m. All behavioral experiments were performed during the light period. Data from six of the animals in this study was included in a previous work [128]. Surgical procedures were performed using sterile techniques under 2-4% isoflurane. Animals were implanted with either microwire electrodes (n=7 animals; 32 or 64 channel 33 μm diameter Tungsten microwire arrays with ZIF-clip adapter; Tucker-Davis Technology) or high-density silicon probes (n=1 animal; custom-built silicon probe [253]) targeted to the forelimb area of M1 (centered at 3.5 mm lateral and 0.5 mm anterior to bregma and implanted in layer V at a depth of 1.5 mm) and the DLS (centered at 4 mm lateral and 0.5 mm anterior to bregma and implanted at a depth of 4 mm). Surgery involved exposure and cleaning of the skull, preparation of the skull surface (using cyanoacrylate), and implantation of skull screws for overall headstage stability. A reference screw was implanted posterior to lambda, contralateral to the neural recordings and a ground screw was implanted posterior to lambda, ipsilateral to the neural recordings. Craniotomy and durectomy were then performed, followed by implantation of neural probes and securing of the implant with Kwik-Sil (World Precision Instruments), C and B Metabond (Parkell, Product #S380), and Duralay dental acrylic (Darby, Product 8830630). Final location of electrodes was confirmed by electrolytic lesions. The postoperative recovery regimen included administration of buprenorphine at 0.02 mg/kg, meloxicam at 0.2 mg/kg, dexamethasone at 0.5 mg/kg, and trimethoprim/sulfadiazine at 15 mg/kg, administered postoperatively for 5 days. All animals recovered for at least 1 week before the start of behavioral training.

4.4.1.2 Reach-to-grasp task

Rats naive to any motor training were first tested for forelimb preference. This involved presenting approximately 10 food pellets to the animal and observing which forelimb was most often used to reach for the pellet. Rats then underwent neural probe implantation surgery in the hemisphere contralateral to the preferred hand. Following the recovery period, rats were trained on the reach-to-grasp task using an automated reach-box, controlled by custom MATLAB scripts and an Arduino microcontroller. This setup requires minimal user intervention, as described previously [233] . Each trial consisted of a pellet dispensed on the pellet tray followed by an alerting beep indicating that the trial was beginning, then the door would open. Animals had 15s to reach, grasp, and retrieve the pellet or the trial would automatically end, and the door would close. A real-time 'pellet detector' using an infrared sensor centered over the pellet would determine when the pellet was moved, indicating the trial was over and, after 2s, the door would close. Trials were separated by a 10s inter-trial interval. All trials were captured by a camera placed on the side of the behavioral box (n=3 animals monitored with a Microsoft LifeCam at 30 $frames/s$; n=5 animals monitored with a Basler ace acA640-750uc at 75 $frames/s$). Each animal underwent 5-14 days of training (100–150 trials per

day). Reach trajectories were captured from video using DeepLabCut [254] to track the center of the rat's hand as well as the food pellet. Trials were aligned to "pellet touch", which was classified as the frame in which the hand was closest to the pellet, before the pellet was displaced off the pellet holder. Only trials in which the pellet was displaced off the pellet holder were considered. Success was achieved if the rat retrieved the pellet from the pellet holder into the behavioral box.

4.4.1.3 In vivo electrophysiology

Throughout reach-to-grasp training, neural signals, including spiking activity and local field potential (LFP) signals were recorded using an RZ2 system (Tucker-Davis Technologies). For neural activity recorded with microwire electrode arrays, spiking data was sampled at 24,414 Hz and LFP/EMG data was sampled at 1017 Hz. LFP signals in M1 and DLS were locally referenced using common-mode referencing: at every time-point, the median signal across all electrodes in an area was calculated and subtracted from every electrode in that area to decrease common noise and minimize volume conduction. To detect spikes in microwire-implanted animals, an online threshold was set using a standard deviation of 4.5 (calculated over a 5 min baseline period). Waveforms and timestamps were stored for any event that crossed below that threshold. Spike sorting was performed separately on each day using Offline Sorter v.4.3.0 (Plexon) with a PCA-based clustering method followed by manual inspection. Single-neuron units were accepted based on waveform shape, clear cluster boundaries in PC space, and 99.5% of detected events with an ISI$> 2ms$. Neural activity recorded with silicon probes was recorded at 24,414 Hz. Spike times and waveforms were detected from the broadband signal using Offline Sorter v.4.3.0 (Plexon). Spike waveforms were then sorted using Kilosort2 (https://github.com/MouseLand/Kilosort2). We accepted units based on manual inspection using Phy (https://github.com/cortex-lab/phy) and 99.5% of detected events with an ISI¿$2ms$.

4.4.1.4 Selection of reach features

To calculate information-theoretic quantities that neural signals contain about the reaching action, we first selected features of the reaching action relevant for success on the reach-to-grasp task. We considered ten features that have a single output on each trial: maximum x-dimension position (i.e., maximum reach amplitude in x-dimension), maximum y-dimension position, maximum x-dimension velocity, maximum y- dimension velocity, mean maximum x- and y-dimension velocity, maximum x-dimension acceleration, maximum y-dimension acceleration, mean maximum x- and y-dimension acceleration, trajectory length (i.e., total distance that the hand travels from movement onset to pellet touch), and movement duration (i.e., time it takes from movement onset to pellet touch). Movement onset was defined as the first time bin in which velocity in the x-dimension was greater than 0 during the 1s preceding pellet touch. If velocity in the x-dimension was already greater than

0 before this time period, movement onset was not defined, leading to fewer trials considered for movement duration and trajectory length (Table 4.S1).

We compared the relevance of different reach features to success by comparing the fit of models used to predict success from each reach feature. We also compared how each reach feature predicted success for skilled and naive movements. Therefore, we fit separate models on the first 3-4 days of training (which we refer to as "naive" days) and the last 2-4 days of training (which we refer to as "skilled" days). The number of days pooled for naive and skilled days was limited by the total number of available recording days in each animal (Table 4.S1). Data was pooled across naive and skilled days to increase computational power of information-theoretic analyses. Two models were fit for both naive and skilled movements, separately, for each feature. The first was a generalized linear mixed-effects model using MATLAB function fitglme to predict success during naive or skilled trials across animals, with the reach feature as a fixed effect, rat ID as a random intercept, and a binomial distribution of the response variable. The second model was also fit with MATLAB function fitglme, but only included naive or skilled trials within each animal, with the reach feature as a fixed effect and a binomial distribution of the response variable.

4.4.1.5 Identification of learning-related changes in neural signal amplitude

To find the timing, relative to pellet touch, in which neural signal amplitude deviated from baseline during movement, we first aligned spiking activity and LFP signals across trials to the time of pellet touch and then binned signals in 10ms bins. Neural signals from 1s prior to pellet touch to .5s after pellet touch were averaged across trials and then z-scored. We then found the first time bin between .5s prior to pellet touch to .5s after pellet touch in which the normalized neural signals was greater than two standard deviations away from the mean.

4.4.2 Information theoretic tools

All analysis code used to compute information quantities (Shannon information, PID shared information, FIT, TE) from unprocessed neural signals is available at: https://gitlab.com/rmaffulli/lemkecelottoetal_codes_rep.

Computation of reach-to-grasp information with Shannon information We computed Shannon Information [75, 86] (I) to quantify the amount of information that neural signals (R) carry about specific reach-to-grasp features (F). Shannon information is a non-parametric measure that quantifies the full single-trial statistical relationship between two stochastic variables and captures the effect of all linear and nonlinear interactions. Shannon information is defined as follows:

$$I(R(t);F) = \sum_{r,f} p(r,f) \log \frac{p(r,f)}{p(r)p(f)} \tag{4.S1}$$

where $R(t)$ denotes the neural signal of a single LFP channel or a single neuron at time t, $p(r, f)$ is the joint probability of observing the neural response r and movement feature f, and $p(r)$ and $p(f)$ are the marginal probabilities of r and f, respectively.

For LFP signals, which are relatively more stable across days compared to spiking activity [234] , Shannon information was computed by combining trials (both successful and unsuccessful) across naive days and across skilled days, separately. In this way, we could enhance the statistical power of Shannon information and reduce sampling bias [179]. For spiking activity, combining trials across days was not possible as we did not track single neurons across days and Shannon information was computed separately on each day. Both LFP signals and spiking activity were aligned to the time of pellet touch and binned in 10ms bins. We then discretized the magnitude of the LFP signals on each trial into 5 discrete values using an equipopulated binning strategy [170]. We also discretized each movement feature across trials into 3 equipopulated bins. For spiking activity, any bin with more than 0 spikes was set to 1.

Computation of shared reach-to-grasp information with partial information decomposition We used partial information decomposition [100] (PID) to measure the shared (or redundant) information that M1 and DLS neural signals encode about a reach feature. In its general formulation, PID breaks down the joint mutual information that two or more source variables carry about one target variable into pieces (or atoms) of shared, unique and synergistic information. In the case of two source variables:

$$MI(F; X, Y) = SI(F : X, Y) + UI(F : X\ Y) + UI(F : Y\ X) + CI(F : X, Y) \quad (4.S2)$$

Where $MI(F; X, Y)$ is the joint mutual information that X and Y encode about F, $SI(F : X, Y)$ is the shared (or redundant) information about F that both X and Y individually encode, $UI(F : X\ Y)$ and $UI(F : Y\ X)$ are the unique pieces of information about F that are encoded only by X and Y, respectively, and $CI(F : X, Y)$ is the synergistic information about F that can be accessed only by considering both X and Y simultaneously. Simpler measures of redundancy only quantify the total effect of redundant versus synergistic information [87, 89]. Previous applications have demonstrated the benefit of using PID over these simpler measures to separate these information components [41, 83, 85]. To measure time-lagged shared reach-to-grasp information, we only made use of the shared information extracted from the PID, which quantifies the amount of redundant information that both source variables (in our case, the neural signals recorded from DLS and M1) separately carry about a target variable (in our case, the reaching feature). Both the unique and shared information terms of PID are used to compute the FIT measure outlined in the next subsection. We computed shared information using the "BROJA" definition [103]. This definition has theoretical advantages over alternative definitions

in the case of two source variables and one target, has been adopted in neuroscientific work [83], and has computationally efficient algorithms to compute it [255]. We computed shared information across a range of temporal delays between M1 and DLS neural signals (up to 250ms delay between each area) to determine whether the same reach-to-grasp information was present in M1 at time t and in DLS at time $t+\delta$. δ could be positive or negative, allowing either M1 or DLS to carry the shared information "first". To plot temporal profiles of M1-to-DLS or DLS-to-M1 shared information we averaged over 0-250ms M1 prior to DLS (M1-to-DLS) and 0-250ms DLS prior to M1 (DLS- to-M1). Shared information was computed specifically for pairs of M1 and DLS LFP signals which contained significant reach-to-grasp information that peaked during the 300ms preceding pellet touch (during the time bins which contained most of the reach-to-grasp information in LFP signals) and pairs of M1 and DLS neurons which contained significant reach-to-grasp information that peaked in the 1s preceding pellet touch (during the time bins that contained most of the reach-to-grasp information in spiking activity).

Computation of LFP coherence We computed LFP coherence between M1 and DLS LFP channels using the Chronux toolbox function cohgramc in MATLAB (http://chronux.org/). LFP was computed in 0.5s sliding windows, with shifts of 0.05s. LFP coherence was computed for the same LFP channels in which PID was computed, as described above.

Computation of neural activity propagation with transfer entropy To measure the propagation of neural activity between two simultaneously recorded neural signals X and Y we used transfer entropy [138] (TE). TE is an information theoretic measure of causal communication between signals based on the Wiener-Granger causality principle [91, 239]. The Weiner-Granger causality principle states that a putative sender X causally influences a receiver Y if the past state of X (X_{past}) predicts the present state of Y at time t (Y_t) beyond what can be predicted by the past state of Y (Y_{past}). TE is therefore an information-theoretic measure that is defined, according to the Wiener-Granger causality principle, as the conditional mutual information between X_{past} and Y_t conditioned to Y_{past}:

$$TE(X \rightarrow Y) = MI(X_{past}; Y_t | Y_{past}) \tag{4.S3}$$

In accordance with previous work, we used a single time-point implementation of Eq. 4.S3 where the past of the sender and the present of the receiver have the same temporal lag [161, 169, 256, 257]:

$$TE(X_{t-\delta} \rightarrow Y_t) = MI(X_{t-\delta}; Y_t | Y_{t-\delta}) \tag{4.S4}$$

TE was computed between M1 and DLS neural signals across a range of temporal delays, as outlined for PID. TE was also computed for the same LFP channels in which PID was computed, as described above.

Computation of reach-to-grasp information flow with feature-specific information transfer To measure the directed flow of reach-to-grasp information we developed a new measure called feature- specific information transfer [50] (FIT, see also Chapter three). This measure extends time-lagged shared information by incorporating into it the Wiener-Granger causality principle, which discounts the information present in the past of the putative sender. To define FIT, we extended time-lagged shared information (see subsection above) by adding to the PID an additional variable, the past activity of the putative receiver Y. We then used PID with three source variables and one target variable to compute $SUI(F : X_{past}, Y_t\ Y_{past})$, the information about F that is Shared between the past of X and the present of Y, that is Unique with respect to (i.e., not encoded in) the past of Y. This measure captures the directed flow of information about F from X to Y and has the desirable properties of being upper-bounded by the amount of reach-to-grasp information encoded in the past of the sender $MI(F; X_{past})$ and the reach-to-grasp information encoded in the present of the receiver $MI(F; Y_t)$ [50].

However, this measure is vulnerable to a mathematical problem called mechanistic redundancy [113, 193]. This term is used to indicate the problematic case in which shared information about a target variable may be measured from two independent source variables with no information transfer between them. In previous work on PID, we have developed an approach to remove the presence of mechanistic redundancies [113]. This approach is to take the minimum between PID information atoms with different targets of the decomposition and allows us to set the transfer entropy from X to Y as an upper bound on FIT. This ensures that if X and Y are independent, no FIT is measured [50]. Therefore, our definition of FIT is:

$$FIT(X \to Y \to F) = \min[SUI(F : X_{past}, Y_t\ Y_{past}); SUI(X_{past} : F, Y_t\ Y_{past})] \quad (4.S5)$$

where $SUI(F : X_{past}, Y_t\ Y_{past})$ is the information that is Shared between X_{past} and Y_t about F and is Unique with respect to Y_{past}, and $SUI(X_{past} : F, Y_t\ Y_{past})$ is the information shared by F and Y_t about X_{past} that is unique with respect to Y_{past}.

As for TE (see Eq. 4.S4), we used a single time-point implementation of FIT where the past of the sender and the present of the receiver have the same temporal lag. Additionally, we computed FIT using the I_{min} definition of shared information [100] as it provides nonnegative results for PID with three source variables and one target.

We validated the ability of FIT in disentangling the flow of neural activity encoding reach-to-grasp information and non-informative neural activity in a simple simulated scenario (Supplemental Figure 11). We simulated the transmission of neural activity from a sender of information X to a receiver of information Y and parametrically controlled the degree to which such flow contained information about a reach variable F. On each trial, the value of X was drawn from a Gaussian distribution with mean equal to 0 and variance equal to 1. Activity of Y was the sum of a motor dimension Y_{reach} and a non-motor dimension $Y_{no-reach}$, both dimensions

were the sum of a Gaussian variable with mean 0 and variance 0.5 plus an input from X. We denote w_{reach} and $w_{no-reach}$ the weights of the links between X and Y_{reach} or $Y_{no-reach}$, respectively. F was computed on each trial by taking the sign of Y_{reach}. The past value of Y_{reach} and of $Y_{no-reach}$ was drawn, independently, from a Gaussian distribution with mean equal to 0. The activity of X was discretized into 4 equipopulated bins while Y_{reach} and $Y_{no-reach}$ were discretized in 2 bins, so that the joint variable $Y = (Y_{reach}, Y_{no-reach})$ had 4 possible outcomes. Similarly, the past of Y_{reach} and $Y_{no-reach}$ were separately discretized in 2 bins and then pooled in a joint variable. We ran the simulation varying the w_{reach} and w_{reach} parameters independently between 0 and 1 in steps of 0.1, with 10000 trials per combination of parameters. We show the values of $FIT(X \to Y \to F)$ and $TE(X \to Y)$ on the grid made of all combinations of w_{reach} and w_{reach} in Figure 4.S11b&c. TE is equally sensitive to the increment of w_{reach} and w_{reach}, while FIT is only sensitive to the increment of w_{reach}, but not to $w_{no-reach}$, as expected from a measure of feature-specific information transmission.

We computed FIT for the same M1 and DLS LFP channels in which we calculated PID and TE, as described above. FIT was also computed between M1 and DLS neural signals across a range of temporal delays, as outlined for PID. We used the direct method to sample the four-dimensional joint probability distribution $p(X_{past}, Y_{past}, Y_t, F)$.

Statistical significance of information theoretic measures To determine the statistical significance of information theoretic measures (Shannon information, shared information, FIT, and TE) we used an approach described in Ref. [178]. In brief, we used a non-parametric cluster permutation technique [177]. We permuted the reach feature values across trials 100 times and recomputed information values to generate a shuffled distribution of each measure. We assigned significance to the largest information "clusters" within the non-permuted data based on a threshold determined by the information values in the shuffled distributions. We determined the 95th percentile value from all shuffled dataset information values. We created information clusters in the original and shuffled datasets by summing together all adjacent information values above the 95th percentile threshold. We then determined a null distribution for information clusters using the maximum cluster value from each shuffled dataset. Finally, we assigned significance to clusters in the original dataset if their value was larger than the 95th percentile of the clusters null distribution.

Significance for Shannon information was determined separately for each LFP channel and neuron. Significance for shared information, FIT, and TE was determined separately for each M1 and DLS LFP channel pair or each pair of M1 and DLS neurons. We subtracted the average shuffled information from all information-theoretic measures in the original dataset. This step conservatively removed information in the original dataset that was due to limited sampling bias [67, 198]. To further ensure that limited sampling bias did not impact our results, we matched

the number of trials between naive and skilled days for all the information-theoretic analyses.

Rat ID	**Electrodes Implanted**	**Training days**	**# of trials naive/skilled**	**# of trials naive/skilled for mov. duration and traj. length**
Rat_1	Microwire array: 16 M1 electrodes & 16 DLS electrodes	8 training days	146 naive trials from days 1-4 146 skilled trials from days 6-8	117 naive trials from days 1-4 117 skilled trials from days 6-8
Rat_2	Microwire array: 16 M1 electrodes & 16 DLS electrodes	5 training days	89 naive trials from days 1-3 89 skilled trials from days 4-5	82 naive trials from days 1-3 82 skilled trials from days 4-5
Rat_3	Microwire array: 16 M1 electrodes & 16 DLS electrodes	5 training days	87 naive trials from days 1-3 87 skilled trials from days 4-5	69 naive trials from days 1-3 69 skilled trials from days 4-5
Rat_4	Microwire array: 16 M1 electrodes & 16 DLS electrodes	9 training days	232 naive trials from days 1-3 232 skilled trials from days 8-9	222 naive trials from days 1-3 222 skilled trials from days 8-9
Rat_5	Microwire array: 32 M1 electrodes & 16 DLS electrodes	14 training days	500 naive trials from days 1-4 500 skilled trials from days 11-14	432 naive trials from days 1-4 432 skilled trials from days 11-14
Rat_6	Microwire array: 32 M1 electrodes & 32 DLS electrode	12 training days	282 naive trials from days 1-3 282 skilled trials from days 10-12	269 naive trials from days 1-3 269 skilled trials from days 10-12
Rat_7	Microwire array: 32 M1 electrodes & 32 DLS electrode	10 training days	336 naive trials from days 1-4 336 skilled trials from days 7-10	332 naive trials from days 1-4 332 skilled trials from days 7-10
Rat_8	Custom-built silicon probe: 37 M1 electrodes & 38 DLS electrodes	6 training days	438 naive trials from days 1-3 438 skilled trials from days 4-6	404 naive trials from days 1-3 404 skilled trials from days 4-6

Table 4.S1: Supplemental Table 1: Experimental animal information

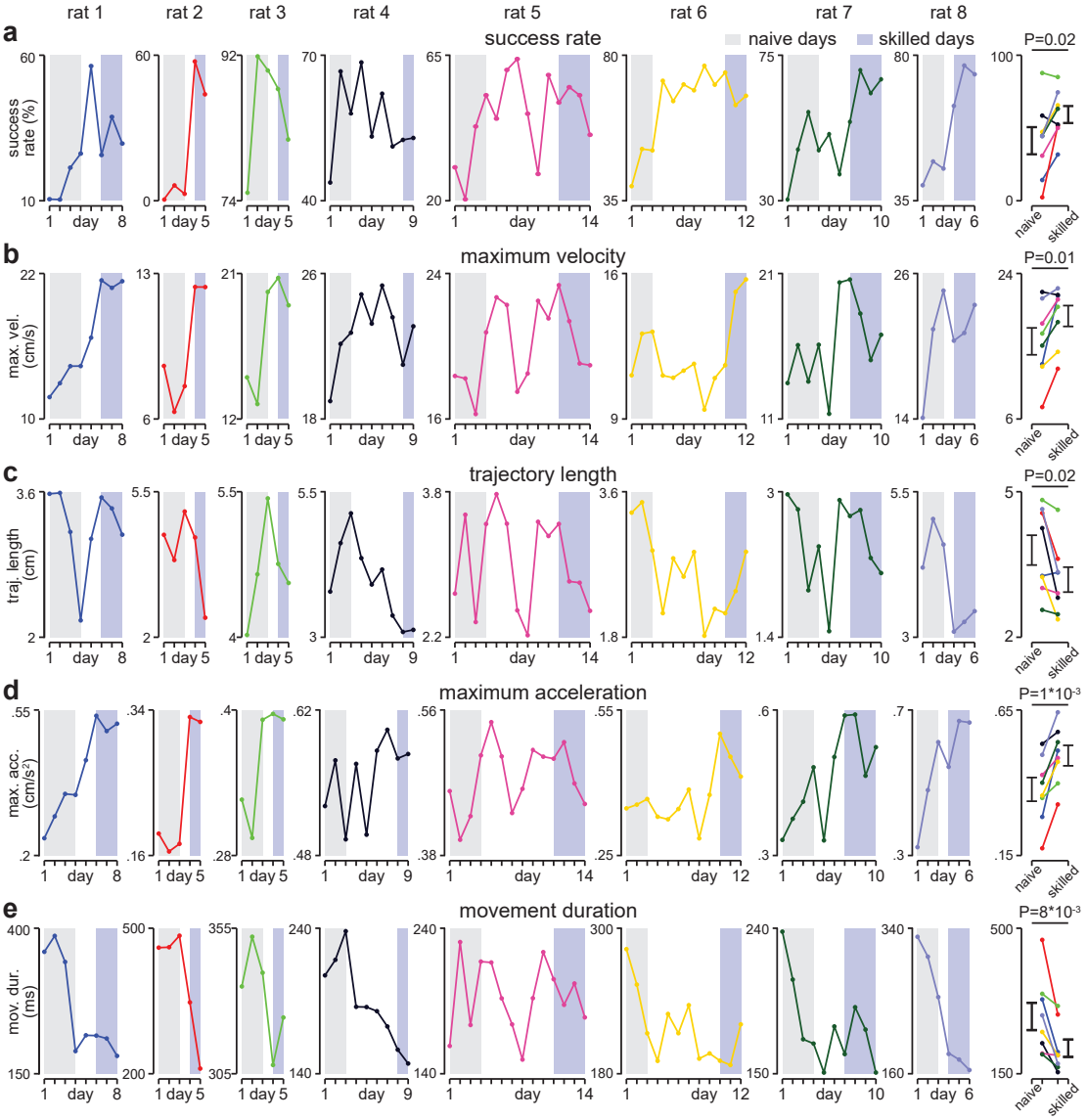

Figure 4.S1: Time course of reach features during skill learning in individual animals. (a) Success rate: naive: 41.4 ± 0.1%, skilled: 59.4 ± 0.1%, t(7)=-2.9, P=0.02, paired-sample t-test. (b) Maximum reaching velocity: naive: 15.6 ± 1.7cm/s, skilled: 18.7 ± 1.3cm/s, t(7)=-3.4, P=0.01, paired-sample t-test. (c) Reaching trajectory length: naive: 3.8 ± 0.3cm, skilled: 3.2 ± 0.3cm, t(7)=2.9, P=0.02, paired- sample t-test. (d) Maximum acceleration: naive: 0.38 ± 0.04cm/s^2, skilled: 0.49 ± 0.04cm/s^2 , t(7)=-5.2, P=1×10^{-3} , paired-sample t-test. (e) Movement duration: naive: 287.9 ± 33ms, skilled: 211.3 ± 21ms, t(7)=3.6, P=8×10^{-3} , paired- sample t-test.

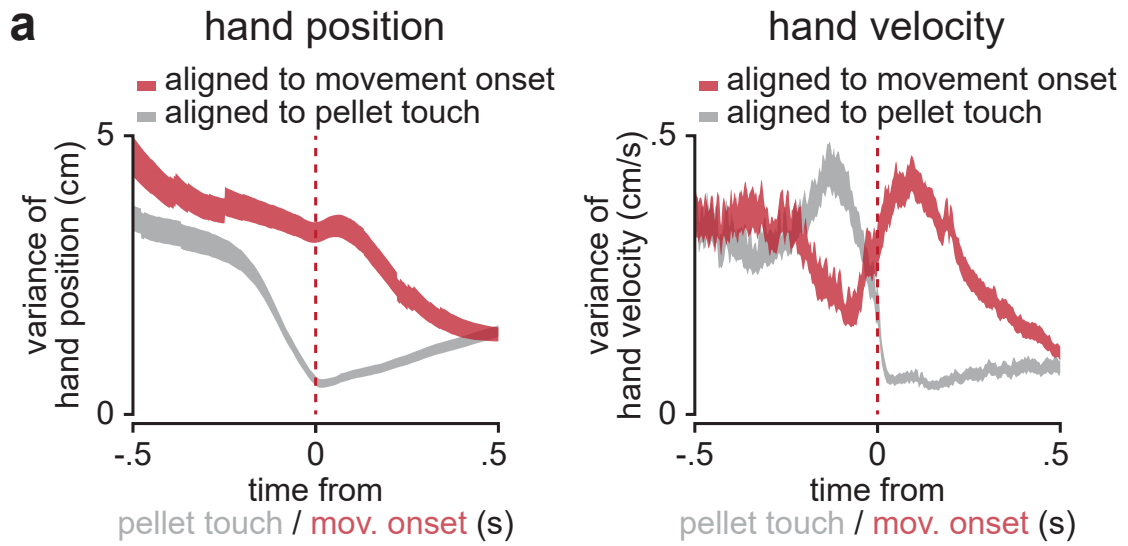

Figure 4.S2: Comparison of variance in hand position and velocity across trials aligned to pellet touch or movement onset. Figure 4.S2: Comparison of variance in hand position and velocity across trials aligned to pellet touch or movement onset. (a) Comparison of across-trial variance of hand position and velocity between trials aligned to pellet touch (gray) and movement onset (red). The width of each line represents mean ± SEM of variance values computed for each animal, in the x and y dimension separately, on naive and skilled days separately (n = 32 variance values from 8 animals with 2 dimensions each for naive and skilled days).

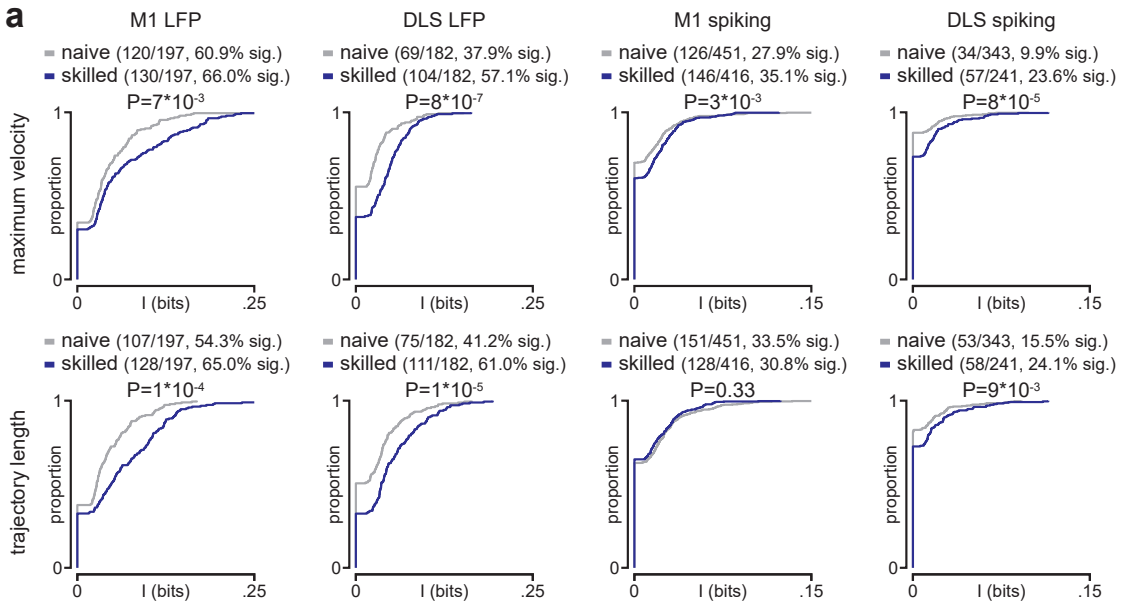

Figure 4.S3: Reach-to-grasp information encoded by M1 and DLS neural signals increases from naive to skilled movements. (a) Cumulative density functions of peak reach-to-grasp information encoded by M1 or DLS LFP signals or spiking activity about maximum reaching velocity (top; M1 LFP: P=7×10^{-3} ; DLS LFP: P=8×10^{-7} ; M1 Spiking: P=3×10^{-3} ; DLS Spiking: P=8×10^{-5} ; Wilcoxon signed rank test) or reaching trajectory length (M1 LFP: P=1×10^{-4} ; DLS LFP: P=1×10^{-5} ; M1 Spiking: P=0.33; DLS Spiking: P=9×10^{-3} ; Wilcoxon signed rank test).

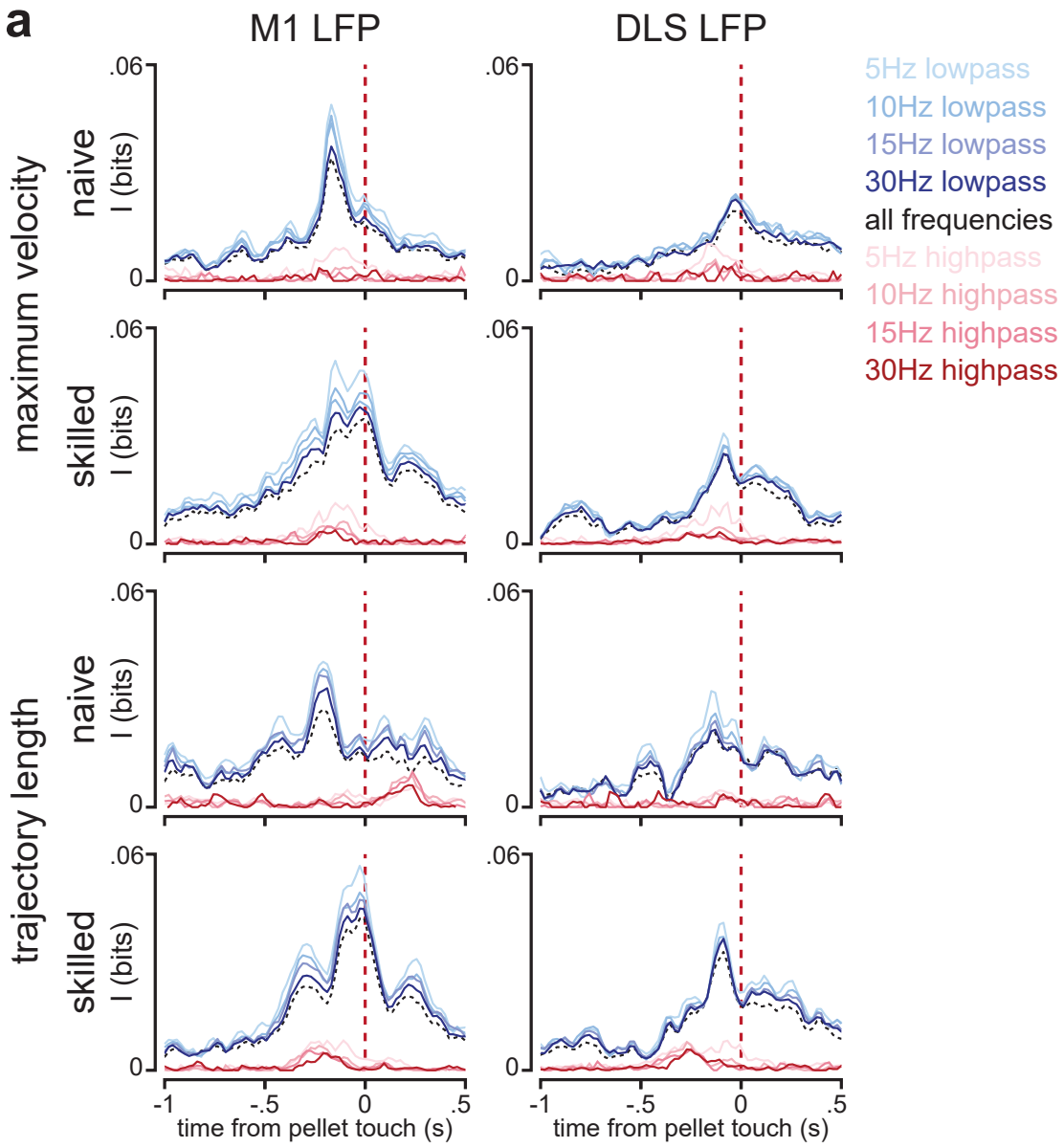

Figure 4.S4: (a) Comparison of mean Shannon information encoded by LFP signals about maximum reaching velocity and reaching trajectory length for M1 and DLS LFP signals high-pass or low-pass filtered at different frequencies.

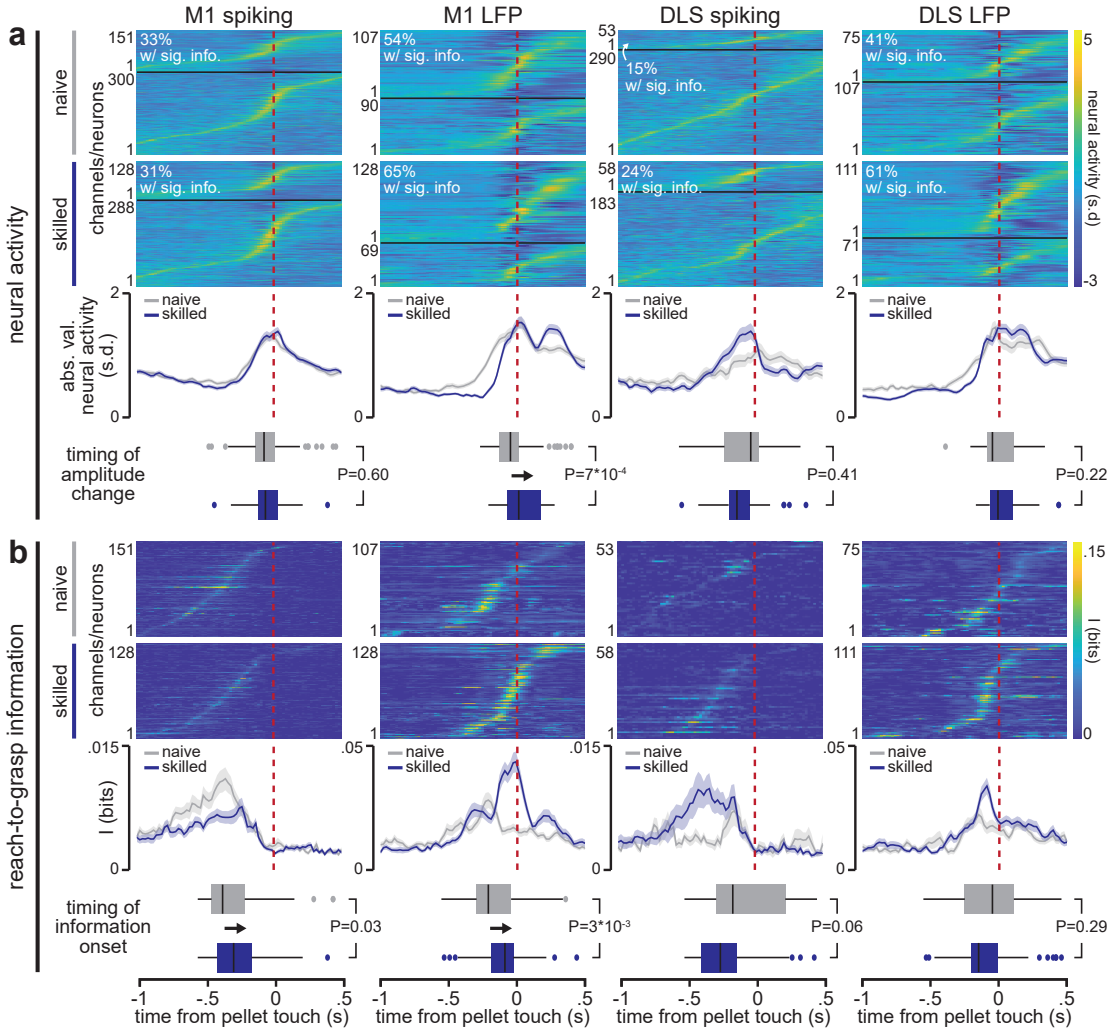

Figure 4.S5: Timing of reaching trajectory length information encoding in M1 and DLS neural signals changes during skill learning. (a) Top: trial-averaged spiking activity and LFP signals, separated by neurons or channels encoding or not encoding reach-to-grasp information about reaching trajectory length during naive or skilled movements, sorted by timing of maximum spiking activity or LFP signal. Middle: time courses of average absolute value of spiking activity or LFP signals during naive or skilled movements, across neurons or channels encoding information about reaching trajectory length (mean ± SEM). Bottom: box plot representing the timing of when LFP signals or spiking activity initially deviate from baseline during naive or skilled movements, across channels or neurons encoding information about reaching trajectory length (M1 spiking: P=0.60, M1 LFP: P=7×10^{-4} , DLS Spiking: P=0.41, DLS LFP: P=0.22; two-sided Wilcoxon rank sum test). (b) Top: reach-to-grasp information encoded by spiking activity or LFP signals about reaching trajectory length during naive or skilled movements, neurons or channels sorted by timing of peak information. Middle: time course of information about reaching trajectory length encoded by spiking activity or LFP signals during naive or skilled movements (mean ± SEM). Bottom: box plot representing the timing of when neural signals first encode information about reaching trajectory length during naive or skilled movements (M1 spiking: P=0.03, M1 LFP: P=3×10^{-3} , DLS Spiking: P=0.06, DLS LFP: P=0.29; two-sided Wilcoxon rank sum test)

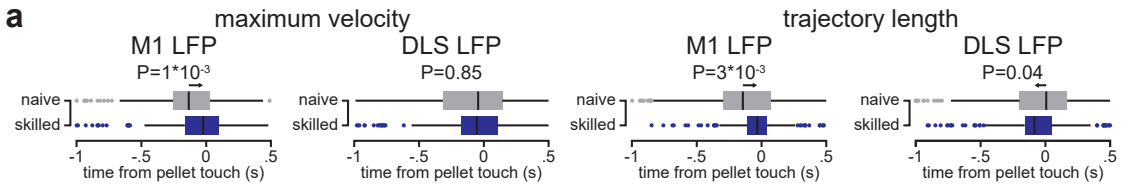

Figure 4.S6: M1 and DLS timing of best linear regression fit changes from naive to skilled movements. (a) Left: Box plots representing distributions of timings of best linear regression model fits between individual M1 or DLS LFP signals and maximum reaching velocity, comparison between naive and skilled days (M1: P=1×10^{-3} and DLS: P=0.85, two-sided Wilcoxon rank sum test). Right: same as Left for reaching trajectory length (M1: P=3×10^{-3} and DLS: P=0.04, two-sided Wilcoxon rank sum test).

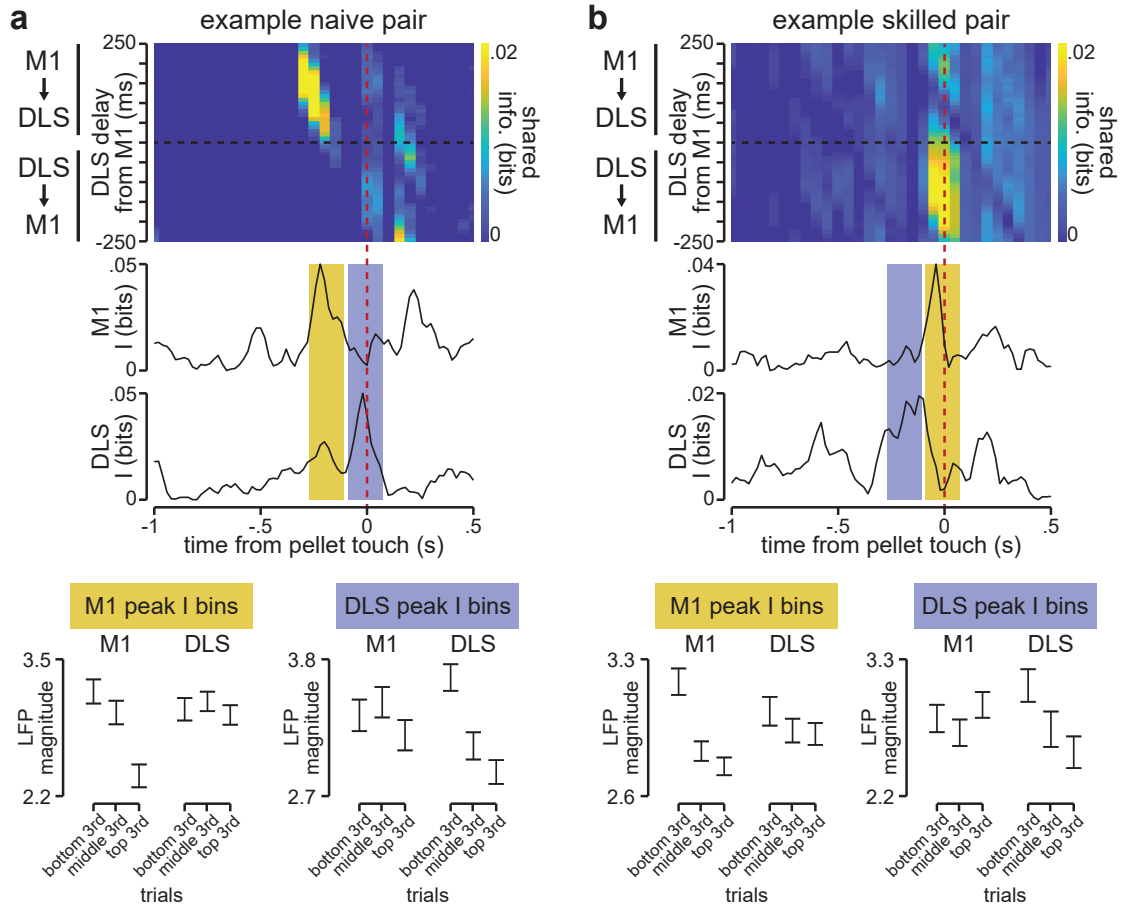

Figure 4.S7: Examples of M1 and DLS LFP pairs encoding shared reach-to-grasp information during naive or skilled movement. (a) Top: shared reach-to-grasp information about maximum reaching velocity between an example pair of M1 and DLS LFP signals during naive movement. Middle: Shannon information about maximum reaching velocity in corresponding individual M1 and DLS LFP signals. Bottom: LFP signal magnitude during time bins in which M1 encoded high reach-to-grasp information (yellow) or during time bins in which DLS encoded high reach-to-grasp information (blue). Each set of time bins is further separated according to the maximum reaching velocity of each trial (i.e., bottom 3 rd represents the third of trials with the lowest maximum velocity, middle 3 rd represents the third of trials with the middle values of maximum velocity, and top 3 rd represents the third of trials with the highest maximum velocity). (b) Same as (a) for example M1 and DLS LFP pair during skilled movement

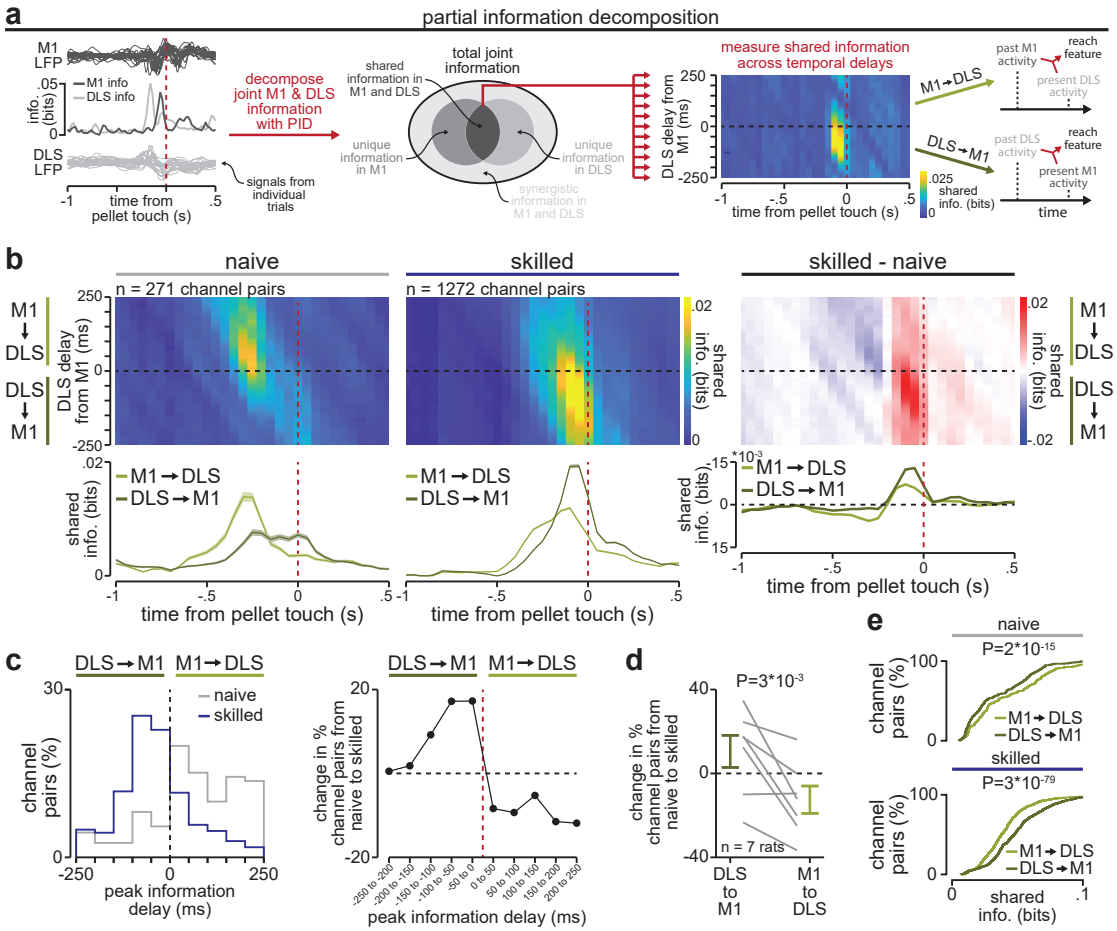

Figure 4.S8: Cross-area timing relationship of shared reach-to-grasp information about reaching trajectory length reverses during skill learning. (a) Schematic of partial information decomposition computation. (b) Top: mean time-lagged shared information in M1 and DLS LFP signals about trajectory reaching length across temporal delays during naive movements, skilled movements, and the difference. Bottom: shared information averaged over positive (M1-to-DLS) and negative (DLS-to-M1) temporal delays during naive movements, skilled movements, and the difference. (c) Left: distribution of temporal delays corresponding to peak shared information about trajectory reaching length across M1 and DLS LFP channel pairs for naive (gray) and skilled (blue) movements. Right: difference between naive and skilled distributions. (d) Change in percentage of M1 and DLS LFP channel pairs with peak shared information about trajectory reaching length with a positive (M1-to-DLS) or negative (DLS-to-M1) temporal delay from naive to skilled movements, in each animal ($t(6)=-4.7$, $P=3\times10^{-3}$, two-sample t-test, n = 7 animals, 1 animal without significant information measured on either naive or skilled days). (e) Cumulative density functions comparing peak shared information about trajectory reaching length across M1 and DLS LFP channel pairs, combined over positive (M1-to-DLS) or negative (DLS-to-M1) temporal delays. Top: naive movements ($P=2\times10^{-15}$, Wilcoxon signed rank test). Bottom: skilled movements ($P=3\times10^{-79}$, Wilcoxon signed rank test).

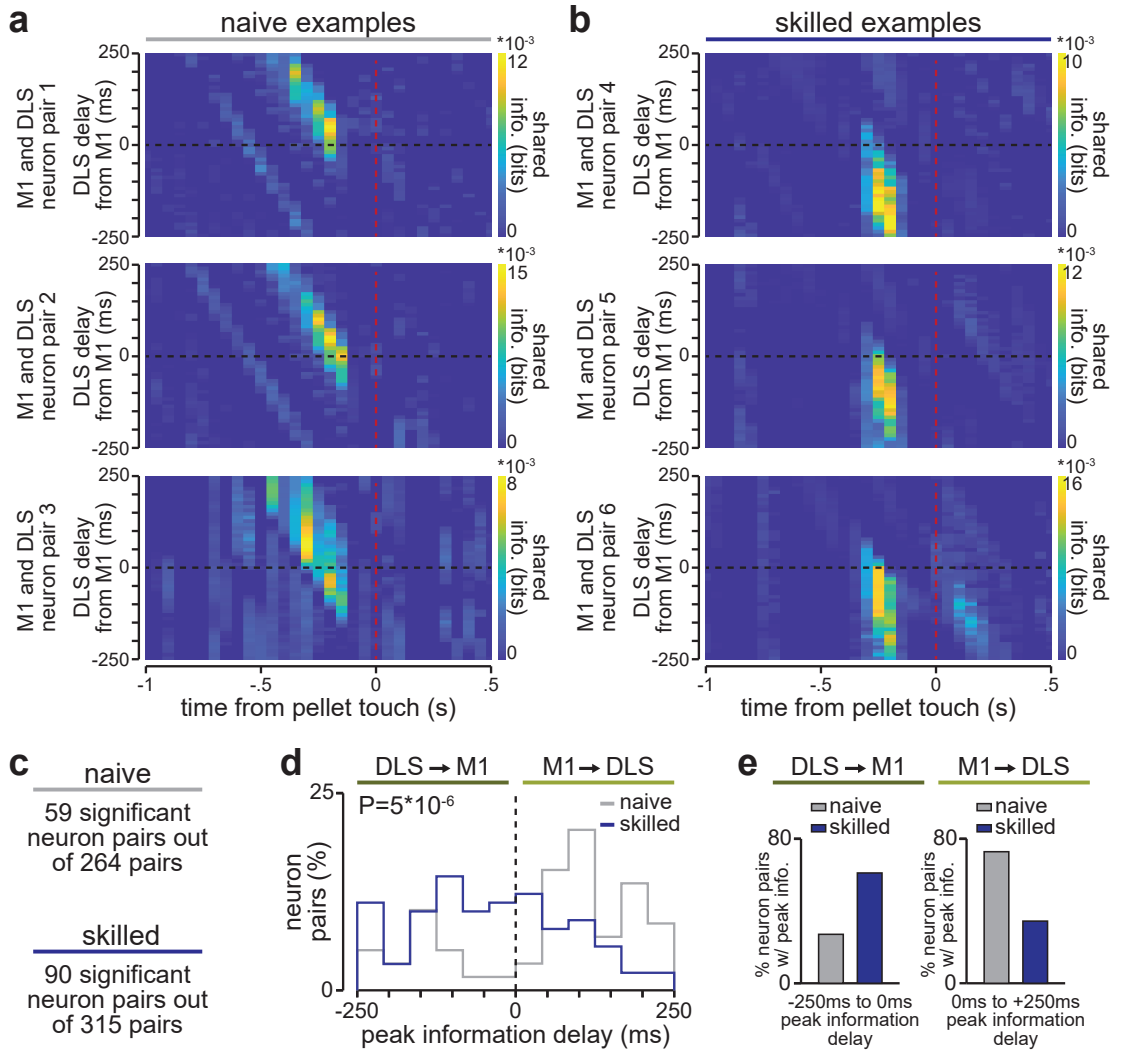

Figure 4.S9: Temporal delays corresponding to peak shared reach-to-grasp information between M1 and DLS neurons reverses from primarily M1-to-DLS delays during naive movements to DLS-to-M1 delays during skilled movements. (a) Time-lagged shared reach-to-grasp information about maximum reaching velocity between example M1 and DLS neuron pairs across temporal delays, during naive movements. (b) Same as (a) for example M1 and DLS neuron pairs during skilled movements (c) Ratio of M1 and DLS neuron pairs encoding significant shared reach-to-grasp information about maximum reaching velocity, out of total possible M1 and DLS neuron pairs that both individually encoded significant Shannon Information about maximum reaching velocity. (d) Distribution of temporal delays corresponding to peak shared reach-to-grasp information about maximum reaching velocity across M1 and DLS neuron pairs during naive (gray) and skilled (blue) movements ($P=5\times10^{-6}$, two-sample Kolmogorov-Smirnov test). (e) Percentage of M1 and DLS neuron pairs with positive (M1-to-DLS) or negative (DLS-to-M1) temporal delays corresponding to peak shared reach-to-grasp information about maximum reaching velocity during naive (gray) and skilled (blue) movements.

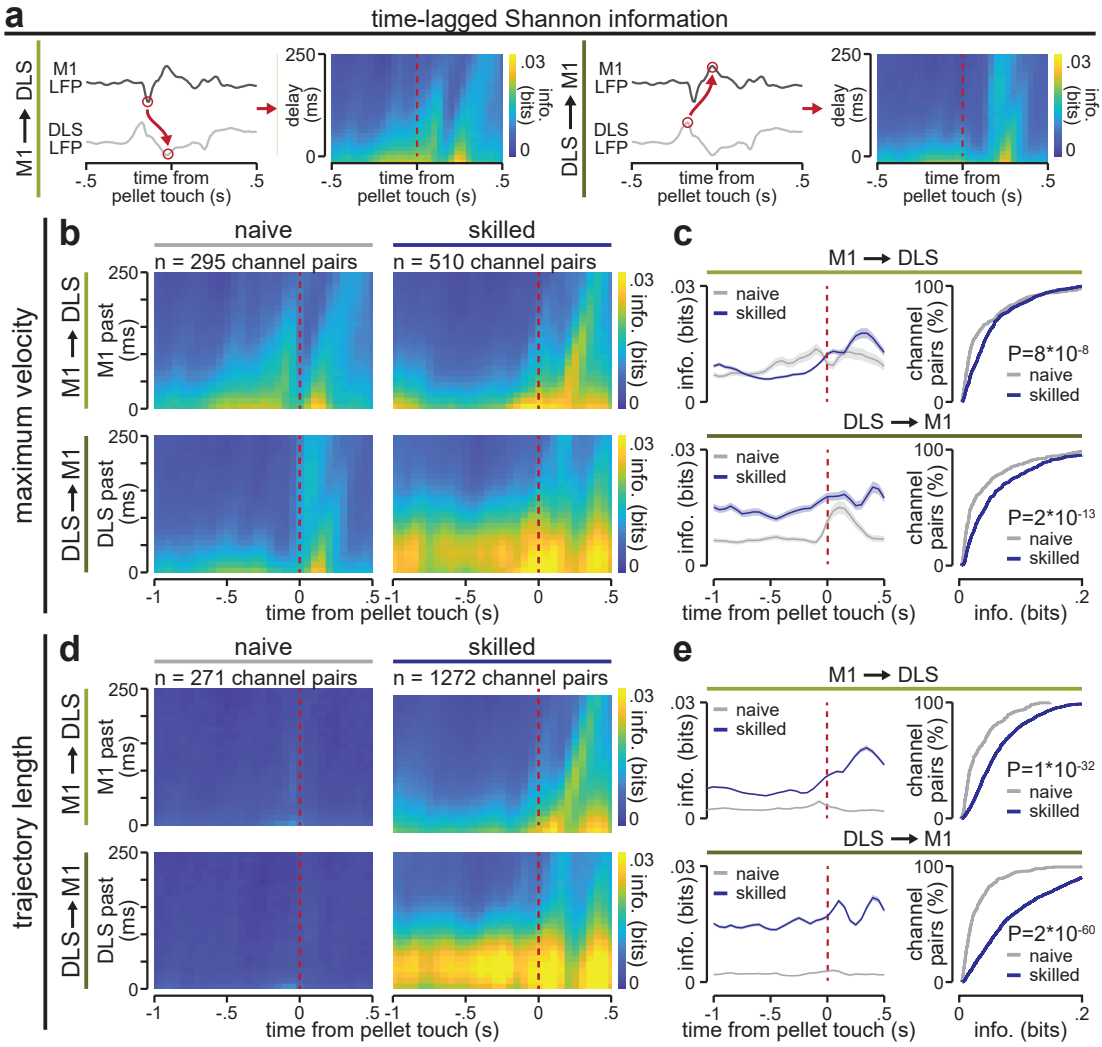

Figure 4.S10: Time-lagged Shannon information between M1 and DLS LFP signals increases bidirectionally during skill learning. (a) Schematic of time-lagged Shannon information computation. (b) Mean time-lagged Shannon information between M1 and DLS LFP signals, averaging across LFP channel pairs that both individually encode information about maximum reaching velocity, during naive and skilled movements. Top: M1-to-DLS temporal delays. Bottom: DLS-to-M1 temporal delays. (c) Left: time courses of mean time-lagged Shannon for LFP channels pairs in (b) during naive and skilled movements, for M1-to-DLS temporal delays (top) and DLS-to-M1 temporal delays (bottom). Right: cumulative density functions comparing distributions of peak time-lagged Shannon information across M1 and DLS LFP signals during naive and skilled movements. Top: M1-to-DLS temporal delays: $P=8 \times 10^{-8}$; Bottom: DLS-to-M1 temporal delays: $P=2 \times 10^{-13}$; two-sided Wilcoxon rank sum test) (d) Same as (b) for M1 and DLS LFP channels encoding information about reaching trajectory length. (e) Same as (c) for M1 and DLS LFP channels encoding information about reaching trajectory length. Top: M1-to-DLS temporal delays: $P=1 \times 10^{-32}$; Bottom: DLS-to-M1 temporal delays: $P=2 \times 10^{-60}$; two- sided Wilcoxon rank sum test).

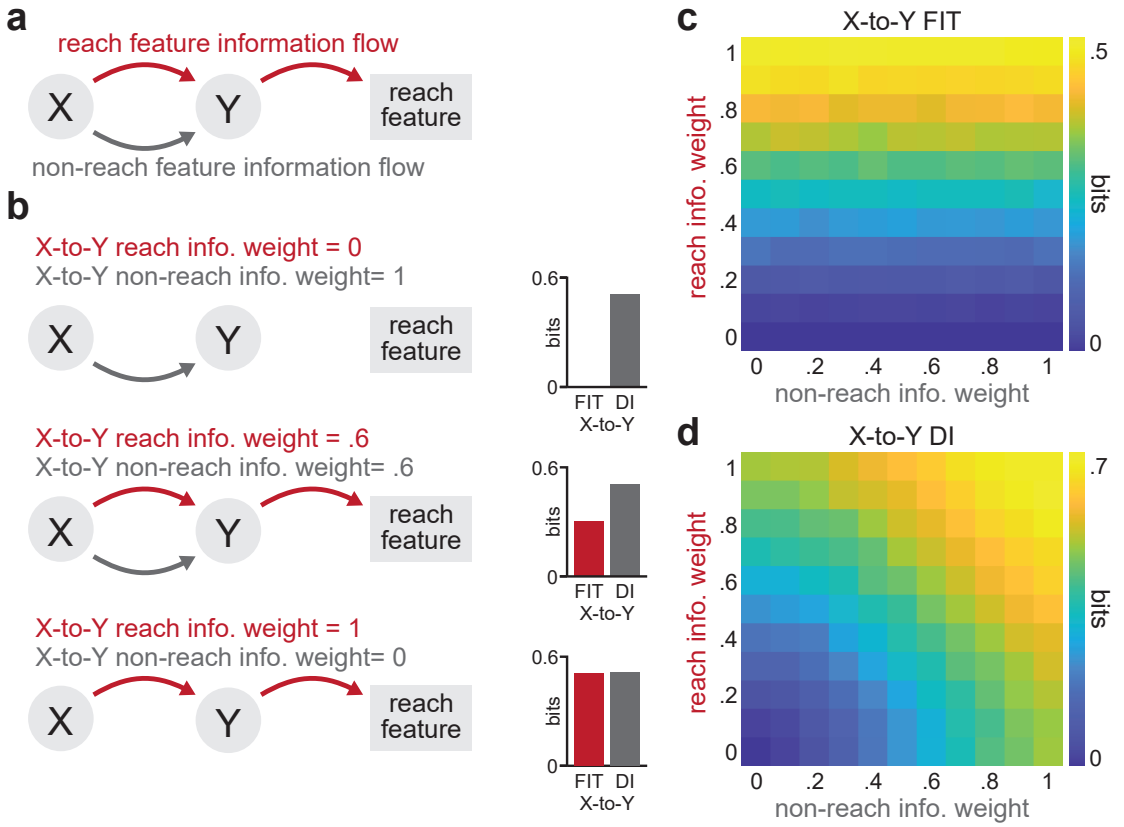

Figure 4.S11: FIT can uniquely capture the flow of neural activity that is informative about a specific behavioral feature, while TE is insensitive to information content. (a) Schematic of simulation model. X simultaneously transmits reach-feature specific and non-reach- feature specific information to Y. We adjusted the amount of reach-feature specific and non-reach- feature specific information flow by changing the values of the X-to-Y reach info. weight (red arrow from X to Y) and of the X-to-Y non-reach info. weight (grey arrow from X to Y). (b) Three example cases with different ratios of reach info. and non-reach info. weights showing the resulting amount of flow measured with FIT (i.e., our measure of reach-to-grasp information flow) and TE (i.e., our measure of overall neural activity propagation). (c) X-to-Y FIT (i.e., our measure of reach-to-grasp information flow) over a range of reach info. and non-reach info. weights, demonstrating that FIT is specifically sensitive to reach feature information flow. (d) X-to-Y TE (i.e., our measure of overall neural activity propagation) over a range of reach info. and non-reach info. weights, demonstrating that TE is insensitive information content and captures both reach feature information flow and non-reach feature information flow.

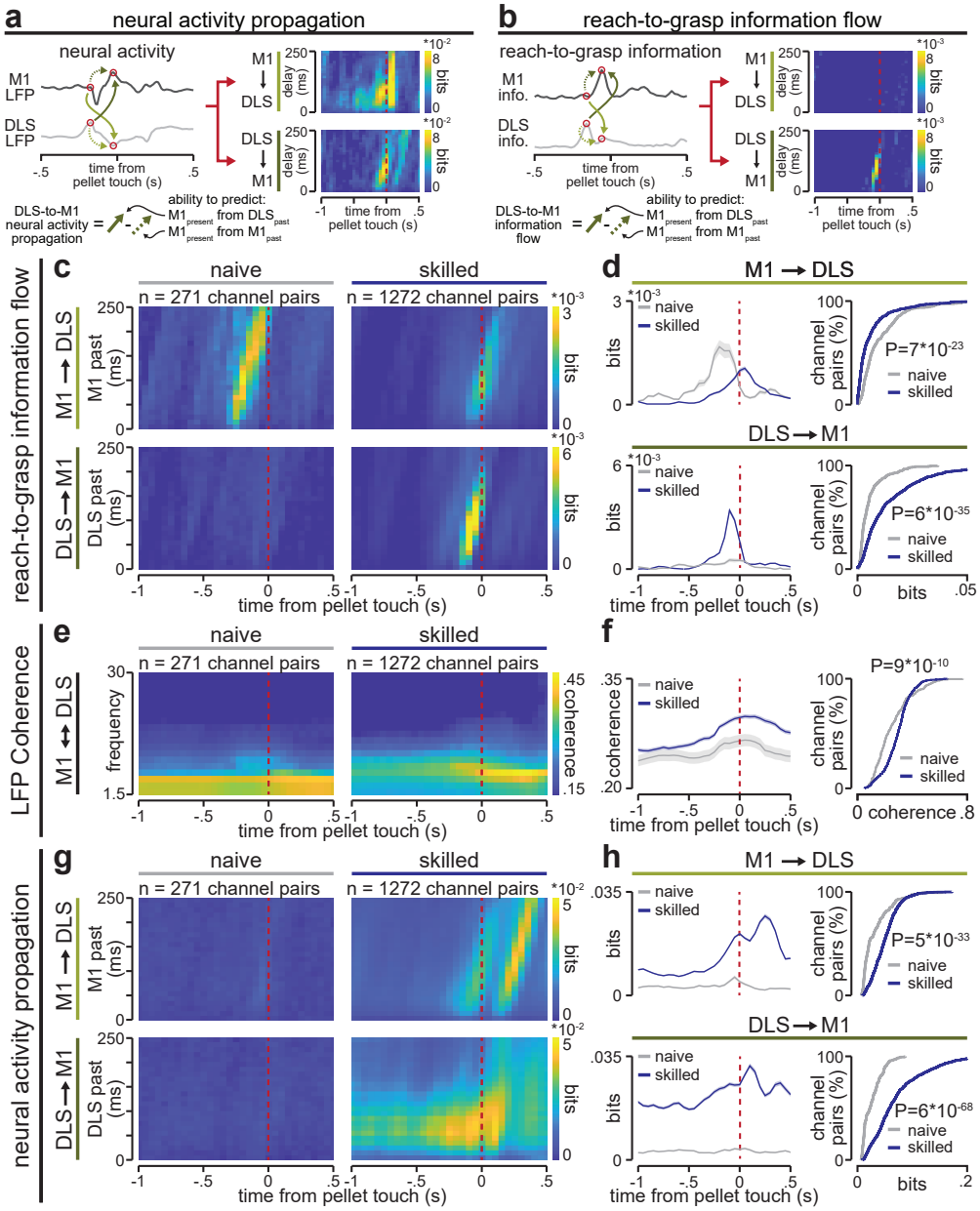

Figure 4.S12: The flow of reach-to-grasp information about reaching trajectory length reverses during skill learning. (a) Schematic of neural activity propagation computation. (b) Schematic of reach-to-grasp information flow computation. (c) Mean reach-to-grasp information flow about reaching trajectory length between M1 and DLS LFP signals during naive and skilled movements. Top: M1-to-DLS temporal delays. Bottom: DLS-to- M1 temporal delays. (d) Left: time course of mean reach-to-grasp information flow about reaching trajectory length between M1 and DLS LFP signals during naive and skilled movements, across M1-to-DLS temporal delays (top) and DLS-to-M1 temporal delays (bottom). Right: cumulative density functions comparing distributions of peak reach-to-grasp information flow about reaching trajectory length between M1 and DLS LFP signals during naive and skilled movements, across M1-to-DLS temporal delays (top; $P=7\times10^{-23}$, two-sided Wilcoxon rank sum test) and DLS-to-M1 temporal delays (bottom; $P=6\times10^{-35}$; two-sided Wilcoxon rank sum test), during naive and skilled movements. (e) Mean LFP coherence between M1 and DLS LFP signals, across frequencies during naive and skilled movements, computed over same LFP channel pairs as (c).

(f) Left: time course of mean 3-15Hz LFP coherence between M1 and DLS LFP signals during naive and skilled movements. Right: cumulative density functions comparing distributions of peak 3- 15Hz LFP coherence between M1 and DLS LFP signals during naive and skilled movements (P=$9 \times 10 - 10$, two-sided Wilcoxon rank sum test). (g) Mean neural activity propagation between M1 and DLS LFP signals during naive and skilled movements, computed over same LFP channel pairs as (c). Top: M1-to-DLS temporal delays. Bottom: DLS-to-M1 temporal delays. (h) Left: time course of mean neural activity propagation between M1 and DLS LFP signals during naive and skilled movements, across M1-to-DLS temporal delays (top) and DLS-to-M1 temporal delays (bottom). Right: cumulative density functions comparing distributions of peak neural activity propagation between M1 and DLS LFP signals during naive and skilled movements, across M1-to-DLS temporal delays (top; P=5×10^{-33} , two-sided Wilcoxon rank sum test) and DLS- to-M1 temporal delays (bottom; P=6×10^{-68} ; two-sided Wilcoxon rank sum test).

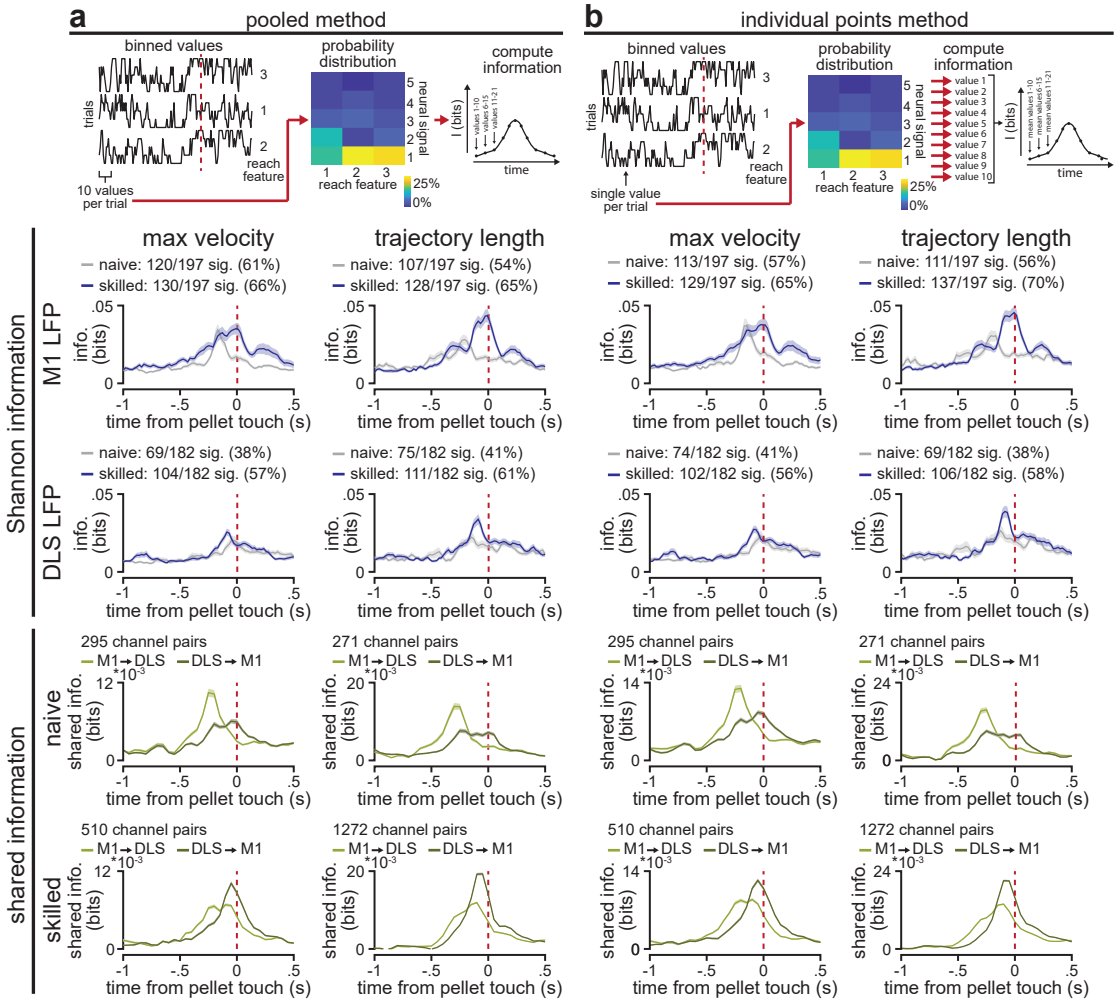

Figure 4.S13: Comparison of "pooled" and "individual points" method for computing Shannon information and shared information. (a) Top: schematic of "pooled" method. Middle: mean Shannon information about maximum reaching velocity and reaching trajectory length encoded in M1 and DLS LFP signals during naive and skilled movements. Bottom: shared reach-to-grasp information about maximum reaching velocity and reaching trajectory length encoded in M1 and DLS LFP signals averaged over positive (M1-to-DLS) and negative (DLS-to-M1) temporal delays during naive and skilled movements. (b) Same as (a) for "individual points" method.

Chapter 5

Conclusions

5.1 Activity and information

Previous literature already studied in detail the relationship between activity and information levels, showing how high activity amplitude in response to a sensory stimulus or concurrent to motor execution [236, 237, 258] does not imply carrying high information about such stimuli or specific kinematic variables. In this work, we delved deeper into the dissociation between neural activity and information, particularly in the context of communication between brain areas.

In Chapters three and four, we proved, on simulated and real data, that looking at the overall information propagated by neural activity can hide some important characteristics of feature-specific neural information processing. The advancement provided by FIT over previous methodologies are due to several, concurrent effects. One first factor limiting traditional measures of causal communication (such as TE) is the inability, by construction, to disambiguate which among several features are transmitted over a neural pathway, and rank them by amount of feature information flow. This first limitation is exemplified in Fig. 3.4, in which the same amount of overall information is propagated across hemispheres in the two directions. FIT can be used to identify the specificity of information transmitted in the two directions, with information about specific eyes being transmitted along contra-lateral pathways, while TE cannot, by construction, discern the specificity of transmitted information.

A second limiting factor is that traditional measures have lower effects size in measuring directed communication. This is possibly because they are less conservative in discarding information already present in the past of the receiver (see e.g. Fig. 3.S8) by capturing components of synergistic information transfer [167]. Another factor adding noise to overall information propagation measures is the larger susceptibility, compared to FIT, to general confounding effects occurring in the total activity space, such as co-fluctuations in feature-unrelated components of the activity (e.g. due to global network co-variations, transmission related to ongoing regulatory operations or to other task-related variables independent from the feature

of interest). This second limitation is exemplified in Fig. 3.3 where TE has lower effect size compared to stimulus FIT in determining that information transmitted feedforward in the visual network is stronger than information transmitted feedback.

Since TE captures the superposition of all task-related and unrelated components of communication, the dynamics of TE could be completely dissociated from the one of FIT. This scenario is best exemplified in the dramatic differences we observed in the evolution of TE dynamics compared to reach-to-grasp FIT during motor learning (see Fig 4.4). Indeed, TE increased bidirectionally, while FIT showed a clear reversal in the area originating and transmitting reaching information in the cortico-striatal network with learning.

Importantly, we are not claiming that TE has any conceptual flaw. We are rather saying that, by design, this measure answers to different questions with respect to FIT. Overall, the development and application of FIT contributed clarifying which are the conceptual advancements of measuring content-specific information flow compared to just the overall propagated information. To get an exhaustive understanding of communication between brain regions, it will be important to combine the complementary perspective provided by TE and FIT.

5.2 What do we mean by shared information?

Similarly to how Shannon information theory gave a quantitative definition of the intuitive yet elusive concept of information [75], PID tries to quantify what we mean by carrying the "same" information, i.e. quantifying redundancy in the encoding. In practice, this problem has proven to be more complicated than what was originally thought. Several PID redundancy measures have been proposed in the literature [100, 103, 112, 259], each satisfying different mathematical properties, such as non-negativity or additivity for independent sources [260]. However, still no consensus exists about which measures have the best theoretical properties and which work best when applied to real neuroscientific data. For example, the redundancy measure we used to compute FIT, I_{min}, is the original one defined when introducing PID [100] and quantifies redundancy as the similarity between the source variables in discriminating individual values of the target [50]. The measure we used to quantify time-lagged shared information in Chapter four [103], which is also commonly used in neuroscience [85, 101], quantifies redundancy by maximizing a Shannon Information quantity measuring the difference in S-related shared and synergistic dependencies between X_1 and X_2. This is done in the probability space of all distributions $Q(X_1, X_2, S)$ that preserves shared dependencies about S by keeping the marginals $P(X_1, S)$ and $P(X_2, S)$ fixed to the original ones.

Despite the conceptual and mathematical differences between these measures, some studies have started establishing analogies between these metrics, e.g. showing their equivalence when applied to multivariate Gaussian systems [261].

Since many aspects of PID are still not fully understood, its application to real data poses significant challenges to the community of theoretical and computational

neuroscientists, which will need to collaborate closely to ensure that theoretical progress aligns with the requirements of real-world neuroscientific questions that PID could address [76, 106, 108]. In this work, we contributed proving that even if further refinements of the theory are required, PID is already a valid tool that can be used to drive advancements in the study of neural information processing [41, 85, 101]. For the PID metrics to be widely applied in neuroscience, it will be fundamental to refine their conceptual interpretation in terms of the properties of neural activity. This includes further investigating the relationship between PID shared information and similarity of tuning curves (see e.g. Fig. 3.2C-E) or correlations between pairs of neurons [89].

In future studies, it will also be interesting to build upon the work presented in Chapter two and use PID-based metrics of functional connectivity to estimate pairwise structural connectivity. For instance, one could analyze all possible triplets of cells that include a specific pair of interest. Within each triplet, one computes the unique information between the activity of the two main cells, in the context of the third neuron's activity. The average unique information between pairs of genes, across triplets, has indeed proven to be a beneficial metric in estimating pairwise interactions in gene networks [107].

5.3 The role of noise in inferring content-specific communication

One key conceptual point that emerged during the development of FIT is that, to infer significant content information transfer, it is necessary to have time-lagged single-trial correlations between the sender and the receiver that are not induced by the feature itself (also known as noise correlations [89]). Although it might sound paradoxical, time-lagged correlations in the feature-unrelated noise between feature-encoding dimensions of neural activity is the only way to identify real feature-specific communication. Indeed, if real feature-specific communication occurs, when the sender fails in encoding the feature, the receiver should similarly fail [9]. This concept is at the basis of the permutation null hypothesis we developed for FIT, in which we destroy single-trial correlations between the sender and the receiver while preserving the overall amount of feature information individually encoded by the two regions (which could, in principle, be explained as time-lagged independent encoding of the feature, see e.g. Fig. 3.S7D). Future refinements in methodologies designed to capture feature-specific information flow will benefit from capturing these types of time-lagged noise correlations.

5.4 Parametric versions of FIT to study population-level information flow

One of the greatest advantages of using information theory to study neural information processing is that it is nearly assumption-free (it does not make any assumption on noise distributions, nor on the type of interaction between variables under analysis). While on one hand this model-free approach has clear benefits, on the other it comes with intrinsic limitations that parametric, model-based approaches can help overcome.

A first limitation of information theory is that it is data hungry [86], since it relies on estimating the full multivariate probability distributions from real data. Directly sampling the frequency of each combination of events (i.e., neural responses and features of interest) across trials works well when dealing with low-dimensional neural signals [179]. However, applying this approach to the high-dimensional neural population data collected in modern experiments [45, 46] becomes challenging due to the curse of dimensionality [86]. For this reason, methods based on the joint application of model-based decoders and information theory have been developed to quantify feature encoding at a neural population level [67, 86]. Additionally, estimation of multivariate probability distributions is an active field of research [187].

A second limitation of information theory is that, given its model-free nature, it does not provide models. While it can be used to rule out candidate neural codes — because if a neural response carries zero information about a feature, the accuracy of any model trained to decode that feature would be at chance [86, 104] — and to put constraints on models [256], it cannot be used to directly decode features or estimate information transfer in individual trials.

To ease the application of FIT to population-recordings it will be important to understand how to best combine FIT and data-dimensionality reduction techniques, or develop new model-based versions of FIT which could also provide insights about the single-trial efficacy of feature information transfer. Such techniques would allow capturing elements of feature-information transfer emerging at the population-level [55] when estimating communication between neural populations [203, 251, 262] or populations of different cell types, such as neurons and astrocytes [67]. However, we predict that, outside the information-theoretic framework, it will be challenging to establish precise relationships between measures of feature information flow and either single-region encoding or overall propagated information.

Lastly, it is important to stress that brain areas are highly interconnected, with activity being simultaneous propagated in different directions over large brain networks [30, 85]. It is therefore necessary to develop tools that can deal with the presence of confounding variables, such as ruling out the role of other simultaneously recorded brain areas in sending information to the putative sender and receiver with a temporal lag [94]. In causality analyses, this is typically done by conditioning out the effect of other areas when inferring causal links, but the practical implemen-

tation of these methods is challenging due to limited amount of data [86] and it is an active topic of research [263]. While we provided a definition of the conditional FIT [50], further computational and theoretical developments, including the definition of parametric measures of feature-information flow, will ease the computation of multivariate cFIT, allowing to condition the FIT between two brain regions over the past activity of multiple other recorded regions.

Bibliography

1. Bullmore, E. & Sporns, O. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat Rev Neurosci* **10,** 186–198 (2009).
2. Bassett, D. & Gazzaniga, M. Understanding complexity in the human brain. *Trends Cogn Sci* **15,** 200–209 (2011).
3. Lennie, P. The cost of cortical computation. *Current Biology* **13,** 493–497 (2003).
4. Howarth, C., Gleeson, P. & Attwell, D. Updated energy budgets for neural computation in the neocortex and cerebellum. *Journal of Cerebral Blood Flow Metabolism* **32,** 1222–1232 (2012).
5. Miller, E. & Cohen, J. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience* **24,** 167–202 (2001).
6. Phelps, E. Emotion and cognition: insights from studies of the human amygdala. *Annual Review of Psychology* **57,** 27–53 (2006).
7. Wolpert, D., Doya, K & Kawato, M. A unifying computational framework for motor control and social interaction. *Philosophical Transactions of the Royal Society B* **358,** 593–602 (2003).
8. Grillner, S. Biological pattern generation: the cellular and computational logic of networks in motion. *Neuron* **52,** 751–766 (2006).
9. Panzeri, S., Harvey, C. D., Piasini, E., Latham, P. E. & Fellin, T. Cracking the Neural Code for Sensory Perception by Combining Statistics, Intervention, and Behavior. *Neuron* **93,** 491–507 (2017).
10. Duncan, J. The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. *Trends in cognitive sciences* **14,** 172–179 (2010).
11. Wilming, N., Murphy, P. R., Meyniel, F. & Donner, T. H. Large-scale dynamics of perceptual decision information across human cortex. *Nature Communications* **11,** 5109 (2020).
12. Kira, S., Safaai, H., Morcos, A., Panzeri, S. & Harvey, C. A distributed and efficient population code of mixed selectivity neurons for flexible navigation decisions. *Nature Communications* **14,** 2121 (2023).

13. Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S. A. & Hudspeth, A. J. *Principles of neural science* 5th ed. (McGraw-Hill Medical, New York, 2013).

14. Kayser, C., Körding, K. P. & König, P. Processing of complex stimuli and natural scenes in the visual cortex. *Current Opinion in Neurobiology* **14,** 468–473 (2004).

15. Ball, T. *et al.* The role of higher-order motor areas in voluntary movement as revealed by high-resolution EEG and fMRI. *NeuroImage* **10,** 682–694 (1999).

16. Allen, W. *et al.* Global Representations of Goal-Directed Behavior in Distinct Cell Types of Mouse Neocortex. *Neuron* **94,** 891–907.e6 (2017).

17. Runyan, C. A., Piasini, E., Panzeri, S. & Harvey, C. D. Distinct timescales of population coding across cortex. *Nature* **548,** 92–96 (2017).

18. Benarroch, E. E. The central autonomic network: functional organization, dysfunction, and perspective. *Mayo Clinic Proceedings* **68,** 988–1001 (1993).

19. Graybiel, A. The basal ganglia and chunking of action repertoires. *Neurobiology of Learning and Memory* **70,** 119–136 (1998).

20. Kawai, R. *et al.* Motor cortex is required for learning but not for executing a motor skill. *Neuron* **86,** 800–812 (2015).

21. Sur, M, Garraghty, P. & Roe, A. Experimentally induced visual projections into auditory thalamus and cortex. *Science* **242,** 1437–1441 (1988).

22. Diamond, M. E., Armstrong-James, M. & Ebner, F. F. Somatic sensory responses in the rostral sector of the posterior group (POm) and in the ventral posterior medial nucleus (VPM) of the rat thalamus. *The Journal of Comparative Neurology* **318,** 462–476 (1992).

23. Bieler, M., Xu, X., Marquardt, A. & Hanganu-Opatz, I. L. Multisensory integration in rodent tactile but not visual thalamus. *Scientific Reports* **8,** 15684 (2018).

24. Wise, R. A. Dopamine, learning and motivation. *Nature Reviews Neuroscience* **5,** 483–494 (2004).

25. Aston-Jones, G & Cohen, J. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annual Review of Neuroscience* **28,** 403–450 (2005).

26. Breton-Provencher, V., Drummond, G., Feng, J., Li, Y. & Sur, M. Spatiotemporal dynamics of noradrenaline during learned behaviour. *Nature* **606,** 732–738 (2022).

27. Dayan, P. & Abbott, L. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems* (The MIT Press, Cambridge, 2001).

28. Buzsáki, G., Anastassiou, C. & Koch, C. The origin of extracellular fields and currents - EEG, ECoG, LFP and spikes. *Nature Reviews Neuroscience* **13,** 407–420 (2012).

29. Buonomano, D. & Maass, W. State-dependent computations: spatiotemporal processing in cortical networks. *Nature Reviews Neuroscience* **10,** 113–125 (2009).

30. Bosman, C. A. *et al.* Attentional Stimulus Selection through Selective Synchronization between Monkey Visual Areas. *Neuron* **75,** 875–888 (2012).

31. Besserve, M., Lowe, S. C., Logothetis, N. K., Schölkopf, B. & Panzeri, S. Shifts of Gamma Phase across Primary Visual Cortical Sites Reflect Dynamic Stimulus-Modulated Information Transfer. *PLOS Biology* **13,** e1002257 (2015).

32. Shadlen, M. N. & Newsome, W. T. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *Journal of Neuroscience* **18,** 3870–3896 (1998).

33. Faisal, A. A., Selen, L. P. & Wolpert, D. M. Noise in the nervous system. *Nature Reviews Neuroscience* **9,** 292–303 (2008).

34. Citri, A. & Malenka, R. C. Synaptic Plasticity: Multiple Forms, Functions, and Mechanisms. *Neuropsychopharmacol.* **33,** 18–41 (2007).

35. Koralek, A., Jin, X., Long, J. n., Costa, R. & Carmena, J. Corticostriatal plasticity is necessary for learning intentional neuroprosthetic skills. *Nature* **483,** 331–335 (2012).

36. Hansen, E. C. A., Battaglia, D., Spiegler, A., Deco, G. & Jirsa, V. K. Functional connectivity dynamics: modeling the switching behavior of the resting state. *NeuroImage* **105,** 525–535 (2015).

37. Preti, M., Bolton, T. & Van De Ville, D. The dynamic functional connectome: State-of-the-art and perspectives. *NeuroImage* **160,** 41–54 (2017).

38. Brovelli, A. *et al.* Dynamic reconfiguration of visuomotor-related functional connectivity networks. *Journal of Neuroscience* **37,** 839–853 (2017).

39. Honey, C. J., 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* **104,** 10240–10245 (2007).

40. Ostojic, S., Brunel, N. & Hakim, V. How Connectivity, Background Activity, and Synaptic Properties Shape the Cross-Correlation between Spike Trains. *Journal of Neuroscience* **29,** 10234–10253 (2009).

41. Luppi, A. *et al.* A synergistic core for human brain evolution and cognition. *Nature Neuroscience* **25,** 771–782 (2022).

42. Sporns, O. The complex brain: connectivity, dynamics, information. *Trends in Cognitive Sciences* **26,** 1066–1067 (2022).

43. Feldman, D. E. The spike-timing dependence of plasticity. *Neuron* **75,** 556–571 (2012).

44. Celotto, M., Lemke, S. & Panzeri, S. Inferring the temporal evolution of synaptic weights from dynamic functional connectivity. *Brain Informatics* **9,** 28 (2022).

45. Jun, J. J. *et al.* Fully integrated silicon probes for high-density recording of neural activity. *Nature* **551,** 232–236 (2017).

46. Sofroniew, N. J., Flickinger, D., King, J. & Svoboda, K. A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. *Elife* **5,** e14472 (2016).

47. Boto, E, Holmes, N, Leggett, J, *et al.* Moving magnetoencephalography towards real-world applications with a wearable system. *Nature* **555,** 657–661 (2018).

48. Hilbert, M. & López, P. The world's technological capacity to store, communicate, and compute information. *Science* **332,** 60–65 (2011).

49. Varghese, B. & Buyya, R. Next generation cloud computing: New trends and research directions. *Future Generation Computer Systems* **79,** 849–861 (2018).

50. Celotto, M. *et al.* An Information-theoretic quantification of the content of communication between brain regions. *Advances in Neural Information Processing Systems 37 (in press)* (2023).

51. Lemke, S. M., Celotto, M, Maffulli, R, Ganguly, K & Panzeri, S. Information flow between motor cortex and striatum reverses during skill learning. *Submitted* (2023).

52. Petersen, R. S., Panzeri, S. & Diamond, M. E. Population coding of stimulus location in rat somatosensory cortex. *Neuron* **32,** 503–514 (2001).

53. Mainen, Z. F. & Sejnowski, T. J. Reliability of spike timing in neocortical neurons. *Science* **268,** 1503–1506 (1995).

54. Schneidman, E., Berry, M. J., Segev, R. & Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **440,** 1007–1012 (2006).

55. Panzeri, S., Moroni, M., Safaai, H. & Harvey, C. D. The structures and functions of correlations in neural population codes. *Nature Reviews Neuroscience* **23,** 551–567 (2022).

56. Kayser, C., Montemurro, M. A., Logothetis, N. K. & Panzeri, S. Spike-phase coding boosts and stabilizes information carried by spatial and temporal spike patterns. *Neuron* **61,** 597–608 (2009).

57. Hall, T., de Carvalho, F. & Jackson, A. A common structure underlies low-frequency cortical dynamics in movement, sleep, and sedation. *Neuron* **83,** 1185–1199 (2014).

58. Belitski, A. *et al.* Low-frequency local field potentials and spikes in primary visual cortex convey independent visual information. *Journal of Neuroscience* **28,** 5696–5709 (2008).

59. Bastos, A. M. *et al.* Visual Areas Exert Feedforward and Feedback Influences through Distinct Frequency Channels. *Neuron* **85,** 390–401 (2015).

60. Pachitariu, M., Steinmetz, N. A., Kadir, S. N., Carandini, M. & Harris, K. D. *Fast and accurate spike sorting of high-channel count probes with KiloSort* in *Advances in Neural Information Processing Systems* **29** (Curran Associates, Inc., 2016).

61. Einevoll, G. T., Kayser, C., Logothetis, N. K. & Panzeri, S. Modelling and analysis of local field potentials for studying the function of cortical circuits. *Nat Rev Neurosci* **14,** 770–785 (2013).

62. Wei, Z *et al.* A comparison of neuronal population dynamics measured with calcium imaging and electrophysiology. *PLoS Computational Biology* **16,** e1008198 (2020).

63. Chen, T. W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499,** 295–300 (2013).

64. Schummers, J., Yu, H. & Sur, M. Tuned responses of astrocytes and their influence on hemodynamic signals in the visual cortex. *Science* **320,** 1638–1643 (2008).

65. Araque, A. *et al.* Gliotransmitters travel in time and space. *Neuron* **81,** 728–739 (2014).

66. Perea, G., Yang, A., Boyden, E. S. & Sur, M. Optogenetic astrocyte activation modulates response selectivity of visual cortex neurons in vivo. *Nature Communications* **5,** 3262 (2014).

67. Curreli, S., Bonato, J., Romanzi, S., Panzeri, S. & Fellin, T. Complementary encoding of spatial information in hippocampal astrocytes. *PLoS Biology* **20,** e3001530 (2022).

68. Desimone, R & Duncan, J. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience* **18,** 193–222 (1995).

69. Betley, J. *et al.* Neurons for hunger and thirst transmit a negative-valence teaching signal. *Nature* **521,** 180–185 (2015).

70. Pillow, J. *et al.* Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* **454,** 995–999 (2008).

71. Cox, D. & Savoy, R. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage* **19,** 261–270 (2003).

72. Livezey, J. A. & Glaser, J. I. Deep learning approaches for neural decoding across architectures and recording modalities. *Briefings in Bioinformatics* **22,** 1577–1591 (2021).

73. Cover, T. M. & Thomas, J. A. *Elements of Information Theory* (Wiley & Sons, Hoboken, New Jersey, 2006).

74. McCulloch, W. & Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* **5,** 115–133 (1943).

75. Shannon, C. A mathematical theory of communication. *The Bell System Technical Journal* **27,** 379–423 (1948).

76. Wibral, M., Lizier, J. T. & Priesemann, V. Bits from Brains for Biologically Inspired Computing. *Frontiers in Robotics and AI* **2** (2015).

77. McCulloch, W. S. & Pfeiffer, J. Of Digital Computers Called Brains. *The Scientific Monthly* **69,** 368–376 (1949).

78. MacKay, D. & McCulloch, W. The limiting information capacity of a neuronal link. *Bulletin of Mathematical Biophysics* **14,** 127–135 (1952).

79. Werner, G. & Mountcastle, V. Neural activity in mechanoreceptive cutaneous afferents: Stimulus-response relations, weber functions, and information transmission. *Journal of Neurophysiology* **28,** 359–397 (1965).

80. Perkel, D. & Bullock, T. *Neural Coding: By Donald H. Perkel and Theodore Holmes Bullock* (1968).

81. Strong, S. P., Koberle, R, De Ruyter Van Steveninck, R & Bialek, W. Entropy and information in neural spike trains. *Physical Review Letters* **80,** 197 (1998).

82. Nigam, S., Pojoga, S. & Dragoi, V. Synergistic Coding of Visual Information in Columnar Networks. *Neuron* **104,** 402–411.e4 (2019).

83. Francis, N. A. *et al.* Sequential transmission of task-relevant information in cortical neuronal networks. *Cell Reports* **39,** 110878 (2022).

84. Vicente, R., Wibral, M., Lindner, M. & Pipa, G. Transfer entropy-a model-free measure of effective connectivity for the neurosciences. *Journal of computational neuroscience* **30,** 45–67 (2011).

85. Varley, T., Sporns, O., Schaffelhofer, S., Scherberger, H. & Dann, B. Information-processing dynamics in neural networks of macaque cerebral cortex reflect cognitive state and behavior. *Proceedings of the National Academy of Sciences* **120,** e2207677120 (2023).

86. Quiroga, R. Q. & Panzeri, S. Extracting information from neuronal populations: information theory and decoding approaches. *Nature Reviews Neuroscience* **10,** 173–185 (2009).

87. Schneidman, E., Bialek, W. & Berry Jr, M. J. Synergy, Redundancy, and Independence in Population Codes. *Journal of Neuroscience* **23,** 11539–11553 (2003).

88. Ince, R. *et al.* A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. *Human Brain Mapping* **38,** 1541–1573 (2017).

89. Pola, G., Thiele, A., Hoffmann, K.-P. & Panzeri, S. An exact method to quantify the information transmitted by different mechanisms of correlational coding. *Network: Computation in Neural Systems* **14,** 35–60 (2003).

90. Petersen, C. Sensorimotor processing in the rodent barrel cortex. *Nature Reviews Neuroscience* **20,** 533–546 (2019).

91. Granger, C. W. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37,** 424–438 (1969).

92. Wiener, N. in *Modern Mathematics for Engineers* (E. F. Beckenbach, New York: McGraw-Hill, 1956).

93. Bressler, S. L. & Seth, A. K. Wiener-Granger causality: A well-established methodology. *NeuroImage* **58,** 323–329 (2011).

94. Barnett, L. & Seth, A. K. The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference. *Journal of Neuroscience Methods* **223,** 50–68 (2014).

95. Rohenkohl, G., Bosman, C. A. & Fries, P. Gamma Synchronization between V1 and V4 Improves Behavioral Performance. *Neuron* **100,** 953–963 (2018).

96. Sheikhattar, A. *et al.* Extracting neuronal functional network dynamics via adaptive Granger causality analysis. *Proceedings of the National Academy of Sciences* **115,** E3869–E3878 (2018).

97. Barnett, L., Barrett, A. B. & Seth, A. K. Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables. *Physical Review Letters* **103,** 238701 (2009).

98. Wibral, M. *et al.* Transfer entropy in magnetoencephalographic data: Quantifying information flow in cortical and cerebellar networks. *Progress in Biophysics and Molecular Biology* **105,** 80–97 (2011).

99. Nirenberg, S. & Latham, P. E. Decoding neuronal spike trains: How important are correlations? *Proceedings of the National Academy of Sciences* **100,** 7348–7353 (2003).

100. Williams, P. L. & Beer, R. D. Nonnegative decomposition of multivariate information. *arXiv:1004.2515* (2010).

101. Koçillari, L. *et al. Measuring Stimulus-Related Redundant and Synergistic Functional Connectivity with Single Cell Resolution in Auditory Cortex* in *BI 2023. Lecture Notes in Computer Science* **13974** (Springer Nature Switzerland, Cham, 2023), 45–56.

102. Bell, A. *The co-information lattice* in *Proceedings of the 4th International Symposium Independent Component Analysis and Blind Source Separation* (Nara, Japan, 2003), 921–926.

103. Bertschinger, N., Rauh, J., Olbrich, E., Jost, J. & Ay, N. Quantifying Unique Information. *Entropy* **16,** 2161–2183 (2014).

104. Pica, G. *et al. Quantifying how much sensory information in a neural code is relevant for behavior* in *Advances in Neural Information Processing Systems 30* (2017), 3689–3699.

105. Mediano, P. A. *et al.* Towards an extended taxonomy of information dynamics via Integrated Information Decomposition. *arXiv:2109.13186* (2021).

106. Wibral, M., Priesemann, V., Kay, J. W., Lizier, J. T. & Phillips, W. A. Partial information decomposition as a unified approach to the specification of neural goal functions. *Brain and cognition* **112,** 25–38 (2017).

107. Chan, T., Stumpf, M. & Babtie, A. Gene Regulatory Network Inference from Single-Cell Data Using Multivariate Information Measures. *Cell Systems* **5,** 251–267.e3 (2017).

108. Beer, R. D. & Williams, P. L. Information Processing and Dynamics in Minimally Cognitive Agents. *Cognitive Science* **39,** 1–38 (2015).

109. Tax, T., Mediano, P. & Shanahan, M. The Partial Information Decomposition of Generative Neural Network Models. *Entropy* **19,** 474 (2017).

110. Lizier, J., Bertschinger, N., Jost, J. & Wibral, M. Information Decomposition of Target Effects from Multi-Source Interactions: Perspectives on Previous, Current and Future Work. *Entropy* **20,** 307 (2018).

111. Finn, C. & Lizier, J. Pointwise Partial Information Decomposition Using the Specificity and Ambiguity Lattices. *Entropy* **20,** 297 (2018).

112. Kolchinsky, A. A Novel Approach to the Partial Information Decomposition. *Entropy* **24,** 403 (2022).

113. Pica, G., Piasini, E., Chicharro, D. & Panzeri, S. Invariant Components of Synergy, Redundancy, and Unique Information among Three Variables. *Entropy* **19,** 451 (2017).

114. Sherrill, S., Timme, N., Beggs, J. & Newman, E. Correlated activity favors synergistic processing in local cortical networks in vitro at synaptically relevant timescales. *Network Neuroscience* **4,** 678–697 (2020).

115. Park, H., Ince, R., Schyns, P., Thut, G. & Gross, J. Representational interactions during audiovisual speech entrainment: Redundancy in left posterior superior temporal gyrus and synergy in left motor cortex. *PLoS Biology* **16,** e2006558 (2018).

116. Delis, I., Ince, R., Sajda, P. & Wang, Q. Neural Encoding of Active Multi-Sensing Enhances Perceptual Decision-Making via a Synergistic Cross-Modal Interaction. *Journal of Neuroscience* **42,** 2344–2355 (2022).

117. Celotto, M., Lemke, S. & Panzeri, S. *Estimating the Temporal Evolution of Synaptic Weights from Dynamic Functional Connectivity* in *Brain Informatics* (eds Mahmud, M., He, J., Vassanelli, S., van Zundert, A. & Zhong, N.) (Springer International Publishing, Cham, 2022), 3–14.

118. Izhikevich, E. Polychronization: Computation with Spikes. *Neural Comput.* **18,** 245–282 (2006).

119. Swadlow, H. A. Physiological properties of individual cerebral axons studied in vivo for as long as one year. *Journal of Neurophysiology* **54,** 1346–62 (1985).

120. Peron, S. *et al.* Recurrent interactions in local cortical circuits. *Nature* **579,** 256–259 (2020).

121. Kuan, A. T. *et al.* Synaptic wiring motifs in posterior parietal cortex support decision-making. *bioRxiv: 2022.04.13.488176* (2022).

122. Mastrogiuseppe, F. & Ostojic, S. Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks. *Neuron* **99,** 609–623 (2018).

123. Goñi, J. *et al.* Resting-brain functional connectivity predicted by analytic measures of network communication. *Proceedings of the National Academy of Sciences* **111,** 833–838 (2014).

124. Tononi, G. & Cirelli, C. Sleep and synaptic homeostasis: a hypothesis. *Brain Res Bull* **62,** 143–150 (2003).

125. Tononi, G. & Cirelli, C. Sleep and the Price of Plasticity: From Synaptic and Cellular Homeostasis to Memory Consolidation and Integration. *Neuron* **81,** 12–34 (2014).

126. De Vivo, L. *et al.* Ultrastructural evidence for synaptic scaling across the wake/sleep cycle. *Science* **355,** 507–510 (2017).

127. Yang, G. *et al.* Sleep promotes branch-specific formation of dendritic spines after learning. *Science* **344,** 1173–1178 (2014).

128. Lemke, S. M. *et al.* Coupling between motor cortex and striatum increases during sleep over long-term skill learning. *eLife* **10,** e64303 (2021).

129. Vahdat, S., Fogel, S., Benali, H. & Doyon, J. Network-wide reorganization of procedural memory during NREM sleep revealed by fMRI. *Elife* **6,** e24987 (2017).

130. Genzel, L., Kroes, M., Dresler, M. & Battaglia, F. Light sleep versus slow wave sleep in memory consolidation: a question of global versus local processes? *Trends in Neurosciences* **37,** 10–19 (2014).

131. Kim, J., Gulati, T. & Ganguly, K. Competing Roles of Slow Oscillations and Delta Waves in Memory Consolidation versus Forgetting. *Cell* **179,** 514–526.e13 (2019).

132. Fasoli, D., Faugeras, O. & Panzeri, S. A formalism for evaluating analytically the cross-correlation structure of a firing-rate network model. *The Journal of Mathematical Neuroscience* **5,** 6 (2015).

133. Ito, S. *et al.* Extending Transfer Entropy Improves Identification of Effective Connectivity in a Spiking Cortical Network Model. *PLoS ONE* **6,** e27431 (2011).

134. Pastore, V. P., Massobrio, P., Godjoski, A. & Martinoia, S. Identification of excitatory-inhibitory links and network topology in large-scale neuronal assemblies from multi-electrode recordings. *PLOS Computational Biology* **14,** e1006381 (2018).

135. Izhikevich, E. Simple model of spiking neurons. *IEEE T. Neural Networ.* **14,** 1569–1572 (2003).

136. Tauste Campo, A. *et al.* Feed-forward information and zero-lag synchronization in the sensory thalamocortical circuit are modulated during stimulus perception. *Proceedings of the National Academy of Sciences* **116,** 7513–7522 (2019).

137. Cutts, C. S. & Eglen, S. J. Detecting Pairwise Correlations in Spike Trains: An Objective Comparison of Methods and Application to the Study of Retinal Waves. *Journal of Neuroscience* **34,** 14288–14303 (2014).

138. Schreiber, T. Measuring information transfer. *Physical Review Letters* **85,** 461–464 (2000).

139. Hlavackovaschindler, K, Palus, M, Vejmelka, M & Bhattacharya, J. Causality detection based on information-theoretic approaches in time series analysis. *Phys. Rep.* **441,** 1–46 (2007).

140. Wibral, M. *et al.* Measuring Information-Transfer Delays. *PLoS ONE* **8,** e55809 (2013).

141. Davis, J. & Goadrich, M. The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd international conference on Machine learning - ICML '06,* 233–240 (2006).

142. Hindriks, R. *et al.* Can sliding-window correlations reveal dynamic functional connectivity in resting-state fMRI? *NeuroImage* **127,** 242–256 (2016).

143. Fino, E, Deniau, J. & Venance, L. Cell-specific spike-timing-dependent plasticity in GABAergic and cholinergic interneurons in corticostriatal rat brain slices. *The Journal of Physiology* **586,** 265–282 (2008).

144. Perez, S *et al.* Striatum expresses region-specific plasticity consistent with distinct memory abilities. *Cell Reports* **38,** 110521 (2022).

145. Mongillo, G., Barak, O. & Tsodyks, M. Synaptic theory of working memory. *Science* **319,** 1543–1546 (2008).

146. Kozachkov, L. *et al.* Robust and brain-like working memory through short-term synaptic plasticity. *PLoS Computational Biology* **18,** e1010776 (2022).

147. Goris, R., Movshon, J. & Simoncelli, E. Partitioning neuronal variability. *Nature Neuroscience* **17,** 858–865 (2014).

148. Vidaurre, D. *et al.* Discovering dynamic brain networks from big data in rest and task. *NeuroImage* **180,** 646–656 (2018).

149. Baker, A. P. *et al.* Fast transient networks in spontaneous human brain activity. *eLife* **3,** e01867 (2014).

150. Pan, S., Mayoral, S. R., Choi, H. S., Chan, J. R. & Kheirbek, M. A. Preservation of a remote fear memory requires new myelin formation. *Nature Neuroscience* **23,** 487–499 (2020).

151. Kobayashi, R. *et al.* Reconstructing neuronal circuitry from parallel spike trains. *Nature Communications* **10,** 4468 (2019).

152. Jaynes, E. T. Information Theory and Statistical Mechanics. *Phys. Rev.* **106,** 620–630 (1957).

153. Landau, I. D., Egger, R., Dercksen, V. J., Oberlaender, M. & Sompolinsky, H. The Impact of Structural Heterogeneity on Excitation-Inhibition Balance in Cortical Networks. *Neuron* **92,** 1106–1121 (2016).

154. Bock, D. *et al.* Network anatomy and in vivo physiology of visual cortical neurons. *Nature* **471,** 177–182 (2011).

155. Bressler, S. L. & Menon, V. Large-scale brain networks in cognition: emerging methods and principles. *Trends in cognitive sciences* **14,** 277–290 (2010).

156. Van Vugt, B. *et al.* The threshold for conscious report: Signal loss and response bias in visual and frontal cortex. *Science* **360,** 537–542 (2018).

157. Varela, F., Lachaux, J.-P., Rodriguez, E. & Martinerie, J. The brainweb: Phase synchronization and large-scale integration. *Nature Reviews Neuroscience* **2,** 229–239 (2001).

158. Brovelli, A. *et al.* Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by Granger causality. *Proceedings of the National Academy of Sciences* **101,** 9849–9854 (2004).

159. Seth, A. K., Barrett, A. B. & Barnett, L. Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience* **35,** 3293–3297 (2015).

160. Massey. Causality, feedback and directed information (1990).

161. Besserve, M, Schölkopf, B, Logothetis, N. & Panzeri, S. Causal relationships between frequency bands of extracellular signals in visual cortex revealed by an information theoretic analysis. *Journal of Computational Neuroscience* **29,** 547–566 (2010).

162. Li, M. *et al.* Transitions in information processing dynamics at the whole-brain network level are driven by alterations in neural gain. *PLoS computational biology* **15,** e1006957 (2019).

163. Stramaglia, S. *et al.* Synergetic and Redundant Information Flow Detected by Unnormalized Granger Causality: Application to Resting State fMRI. *IEEE Transactions on Biomedical Engineering* **63,** 2518–2524 (2016).

164. Van Kerkoerle, T. *et al.* Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proceedings of the National Academy of Sciences* **111,** 14332–14341 (2014).

165. Vinck, M. *et al.* How to detect the Granger-causal flow direction in the presence of additive noise? *NeuroImage* **108,** 301–318 (2015).

166. Ramirez-Villegas, J. F. *et al.* Coupling of hippocampal theta and ripples with pontogeniculooccipital waves. *Nature* **589,** 96–102 (Nov. 2021).

167. James, R. G., Barnett, N. & Crutchfield, J. P. Information Flows? A Critique of Transfer Entropies. *Physical Review Letters* **116,** 238701 (2016).

168. Palmigiano, A., Geisel, T., Wolf, F. & Battaglia, D. Flexible information routing by transient synchrony. *Nature Neuroscience* **20,** 1014–1022 (2017).

169. Oever, S. T., Sack, A. T., Oehrn, C. R. & Axmacher, N. An engram of intentionally forgotten information. *Nature Communications* **12,** 6443 (2021).

170. Magri, C., Whittingstall, K., Singh, V., Logothetis, N. K. & Panzeri, S. A toolbox for the fast information analysis of multiple-site LFP, EEG and spike train recordings. *BMC Neuroscience* **10** (2009).

171. Montalto, A., Faes, L. & Marinazzo, D. MuTE: A MATLAB Toolbox to Compare Established and Novel Estimators of the Multivariate Transfer Entropy. *PLoS ONE* **9,** e109462 (2014).

172. Hadjipapas, A., Lowet, E., Roberts, M., Peter, A. & Weerd, P. D. Parametric variation of gamma frequency and power with luminance contrast: A comparative study of human MEG and monkey LFP and spike responses. *NeuroImage* **112,** 327–340 (2015).

173. Ray, S. & Maunsell, J. H. R. Differences in Gamma Frequencies across Visual Cortex Restrict Their Possible Use in Computation. *Neuron* **67,** 885–896 (2010).

174. Donner, T. H. & Siegel, M. A framework for local cortical oscillation patterns. *Trends in Cognitive Sciences* **15,** 191–199 (2011).

175. Michalareas, G. *et al.* Alpha-Beta and Gamma Rhythms Subserve Feedback and Feedforward Influences among Human Visual Cortical Areas. *Neuron* **89,** 384–397 (2016).

176. Gross, J., Timmermann, L., Kujala, J., Salmelin, R. & Schnitzler, A. Properties of MEG tomographic maps obtained with spatial filtering. *NeuroImage* **19,** 1329–1336 (2003).

177. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG-and MEG-data. *Journal of neuroscience methods* **164,** 177–190 (2007).

178. Combrisson, E. *et al.* Group-level inference of information-based measures for the analyses of cognitive brain networks from neurophysiological data. *NeuroImage* **258,** 119347 (2022).

179. Panzeri, S., Senatore, R., Montemurro, M. A. & Petersen, R. S. Correcting for the sampling bias problem in spike train information measures. *Journal of Neurophysiology* **98,** 1064–1072 (2007).

180. Rousselet, G. A., Ince, R. A. A., van Rijsbergen, N. J. W. & Schyns, P. G. Eye coding mechanisms in early human face event-specific potentials. *Journal of vision* **14,** 7 (2014).

181. Ince, R. A. A. *et al.* The Deceptively Simple N170 Reflects Network Information Processing Mechanisms Involving Visual Feature Coding and Transfer Across Hemispheres. *Cerebral Cortex* **26,** 4123–4135 (2016).

182. Sherman, S. M. & Guillery, R. W. The role of the thalamus in the flow of information to the cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **357,** 1695–1708 (2002).

183. Ince, R. A. A. *et al.* Tracing the Flow of Perceptual Features in an Algorithmic Brain Network. *Scientific Reports* **5,** 17681 (2015).

184. De Ruyter van Steveninck, R. R., Lewen, G. D., Strong, S. P., Koberle, R. & Bialek, W. Reproducibility and Variability in Neural Spike Trains. *Science* **275,** 1805–1808 (1997).

185. Chelaru, M. I. *et al.* High-order interactions explain the collective behavior of cortical populations in executive but not sensory areas. *Neuron* **109,** 3954–3961 (2021).

186. Ganmor, E., Segev, R. & Schneidman, E. Sparse low-order interaction network underlies a highly correlated and learnable neural population code. *Proceedings of the National Academy of Sciences* **108,** 9679–9684 (2011).

187. Safaai, H, Onken, A, Harvey, C. & Panzeri, S. Information estimation using nonparametric copulas. *Physical Review E* **98,** 053302 (2018).

188. Griffith, V. & Koch, C. in *Guided Self-Organization: Inception* 159–190 (Springer, Berlin, Heidelberg, 2014).

189. Barrett, A. B. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. *Physical Review E* **91,** 052802 (2015).

190. Shahbazi, F., Ewald, A., Ziehe, A. & Nolte, G. *Constructing Surrogate Data to Control for Artifacts of Volume Conduction for Functional Connectivity Measures* in *17th International Conference on Biomagnetism Advances in Biomagnetism – Biomag2010* **28** (2010), 207–210.

191. Freud, E., Culham, J. C., Plaut, D. C. & Behrmann, M. The large-scale organization of shape processing in the ventral and dorsal pathways. *eLife* **6,** e27576 (2017).

192. Saur, D. *et al.* Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences* **105,** 18035–18040 (2008).

193. Harder, M., Salge, C. & Polani, D. Bivariate measure of redundant information. *Physical Review E* **87,** 012130 (2013).

194. Chicharro, D. Quantifying multivariate redundancy with maximum entropy decompositions of mutual information. *arXiv:1708.03845* (2017).

195. Makkeh, A., Chicharro, D., Theis, D. O. & Vicente, R. MAXENT3D_PID: An Estimator for the Maximum-Entropy Trivariate Partial Information Decomposition. *Entropy* **21,** 862 (2019).

196. Chicharro, D. & Panzeri, S. Synergy and Redundancy in Dual Decompositions of Mutual Information Gain and Information Loss. *Entropy* **19,** 71 (2017).

197. Schmolesky, M. T. *et al.* Signal Timing Across the Macaque Visual System. *Journal of Neurophysiology* **79,** 3272–3278 (1998).

198. Panzeri, S. & Treves, A. Analytical estimates of limited sampling biases in different information measures. *Network: Computation in Neural Systems* **7,** 87–107 (1996).

199. Montemurro, M. A., Senatore, R. & Panzeri, S. Tight data-robust bounds to mutual information combining shuffling and model selection techniques. *Neural Comput* **19,** 2913–2957 (2007).

200. AM, A. M. B. & Schoffelen, J.-M. A Tutorial Review of Functional Connectivity Analysis Methods and Their Interpretational Pitfalls. *Front Syst Neurosci* **9,** 175 (2016).

201. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536,** 171–178 (2016).

202. Delorme, A. & Makeig, S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods* **134,** 9–21 (2004).

203. Gokcen, E. *et al.* Disentangling the flow of signals between populations of neurons. *Nature Computational Science* **2,** 512–525 (2022).

204. Siegle, J. H., Jia, X. & et al., S. D. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature* **592,** 86–92 (2021).

205. Pica, G., Soltanipour, M. & Panzeri, S. Using intersection information to map stimulus information transfer within neural networks. *BioSystems* **185,** 104028 (2019).

206. Diedrichsen, J. & Kornysheva, K. Motor skill learning between selection and execution. *Trends in Cognitive Sciences* **19,** 227–233 (2015).

207. Shmuelof, L. & Krakauer, J. Are we ready for a natural history of motor learning? *Neuron* **72,** 469–476 (2011).

208. Haith, A. & Krakauer, J. The multiple effects of practice: skill, habit and reduced cognitive load. *Current Opinion in Behavioral Sciences* **20,** 196–201 (2018).

209. Kargo, W. & Nitz, D. Improvements in the signal-to-noise ratio of motor cortex cells distinguish early versus late phases of motor skill learning. *Journal of Neuroscience* **24,** 5560–5569 (2004).

210. Peters, A., Chen, S. & Komiyama, T. Emergence of reproducible spatiotemporal activity during motor learning. *Nature* **510,** 263–267 (2014).

211. Li, Q. *et al.* Refinement of learned skilled movement representation in motor cortex deep output layer. *Nature Communications* **8,** 15834 (2017).

212. Makino, H. *et al.* Transformation of Cortex-wide Emergent Properties during Motor Learning. *Neuron* **94,** 880–890.e8 (2017).

213. Koralek, A., Costa, R. & Carmena, J. Temporally precise cell-specific coherence develops in corticostriatal networks during learning. *Neuron* **79,** 865–872 (2013).

214. Lemke, S., Ramanathan, D., Guo, L., Won, S. & Ganguly, K. Emergent modular neural control drives coordinated motor actions. *Nature Neuroscience* **22,** 1122–1131 (2019).

215. Fleischer, P. *et al.* Emergent Low-Frequency Activity in Cortico-Cerebellar Networks with Motor Skill Learning. *eNeuro* **10** (2023).

216. Wagner, M. *et al.* Shared Cortex-Cerebellum Dynamics in the Execution and Learning of a Motor Task. *Cell* **177,** 669–682.e24 (2019).

217. Veuthey, T., Derosier, K., Kondapavulur, S. & Ganguly, K. Single-trial cross-area neural population dynamics during long-term skill learning. *Nature Communications* **11,** 4057 (2020).

218. Hwang, F.-J. *et al.* Motor learning selectively strengthens cortical and striatal synapses of motor engram neurons. *Neuron* **110,** 2790–2801.e5 (2022).

219. Santos, F., Oliveira, R., Jin, X. & Costa, R. Corticostriatal dynamics encode the refinement of specific behavioral variability during skill learning. *Elife* **4,** e09423 (2015).

220. Dang, M. *et al.* Disrupted motor learning and long-term synaptic plasticity in mice lacking NMDAR1 in the striatum. *Proceedings of the National Academy of Sciences* **103,** 15254–15259 (2006).

221. Jin, X. & Costa, R. Start/stop signals emerge in nigrostriatal circuits during sequence learning. *Nature* **466,** 457–462 (2010).

222. Hwang, E. *et al.* Disengagement of motor cortex from movement control during long-term learning. *Science Advances* **5,** eaay0001 (2019).

223. Dhawale, A., Wolff, S., Ko, R. & Ölveczky, B. The basal ganglia control the detailed kinematics of learned motor skills. *Nature Neuroscience* **24,** 1256–1269 (2021).

224. Wolff, S., Ko, R. & Ölveczky, B. Distinct roles for motor cortical and thalamic inputs to striatum during motor skill learning and execution. *Science Advances* **8,** eabk0231 (2022).

225. Sacrey, L.-A., Alaverdashvili, M. & Whishaw, I. Similar hand shaping in reaching-for-food (skilled reaching) in rats and humans provides evidence of homology in release, collection, and manipulation movements. *Behavioural Brain Research* **204,** 153–161 (2009).

226. Iwaniuk, A. & Whishaw, I. On the origin of skilled forelimb movements. *Trends in Neurosciences* **23,** 372–376 (2000).

227. Guo, J.-Z. *et al.* Cortex commands the performance of skilled movement. *Elife* **4,** e10774 (2015).

228. Alaverdashvili, M. & Whishaw, I. Motor cortex stroke impairs individual digit movement in skilled reaching by the rat. *European Journal of Neuroscience* **28,** 311–322 (2008).

229. Whishaw, I., O'Connor, W. & Dunnett, S. The contributions of motor cortex, nigrostriatal dopamine and caudate-putamen to skilled forelimb use in the rat. *Brain* **109,** 805–843 (1986).

230. Sauerbrei, B. *et al.* Cortical pattern generation during dexterous movement is input-driven. *Nature* **577,** 386–391 (2020).

231. Bova, A. *et al.* Precisely timed dopamine signals establish distinct kinematic representations of skilled movements. *Elife* **9,** e61591 (2020).

232. Albarran, E., Raissi, A., Jáidar, O., Shatz, C. & Ding, J. Enhancing motor learning by increasing the stability of newly formed dendritic spines in the motor cortex. *Neuron* **109,** 3298–3311.e4 (2021).

233. Wong, C., Ramanathan, D., Gulati, T., Won, S. & Ganguly, K. An automated behavioral box to assess forelimb function in rats. *Journal of Neuroscience Methods* **246,** 30–37 (2015).

234. Flint, R., Scheid, M., Wright, Z., Solla, S. & Slutzky, M. Long-Term Stability of Motor Cortical Activity: Implications for Brain Machine Interfaces and Optimal Feedback Control. *Journal of Neuroscience* **36,** 3623–3632 (2016).

235. Ramanathan, D. *et al.* Low-frequency cortical activity is a neuromodulatory target that tracks recovery after stroke. *Nature Medicine* **24,** 1257–1267 (2018).

236. Kaufman, M. *et al.* The Largest Response Component in the Motor Cortex Reflects Movement Timing but Not Movement Type. *eNeuro* **3** (2016).

237. Butts, D. & Goldman, M. Tuning curves, neuronal variability, and sensory coding. *PLoS Biology* **4,** e92 (2006).

238. Tanaka, T. & Nakamura, K. Focal inputs are a potential origin of local field potential (LFP) in the brain regions without laminar structure. *PLoS One* **14,** e0226028 (2019).

239. Wiener, N. Nonlinear prediction and dynamics, 249–254 (1956).

240. O'Hare, J. *et al.* Pathway-Specific Striatal Substrates for Habitual Behavior. *Neuron* **89,** 472–479 (2016).

241. Rothwell, P. *et al.* Input- and Output-Specific Regulation of Serial Order Performance by Corticostriatal Circuits. *Neuron* **88,** 345–356 (2015).

242. Yin, H. *et al.* Dynamic reorganization of striatal circuits during the acquisition and consolidation of a skill. *Nature Neuroscience* **12,** 333–341 (2009).

243. Kondapavulur, S. *et al.* Transition from predictable to variable motor cortex and striatal ensemble patterning during behavioral exploration. *Nat. Commun.* **13,** 2450 (2022).

244. Kojima, S., Kao, M., Doupe, A. & Brainard, M. The Avian Basal Ganglia Are a Source of Rapid Behavioral Variation That Enables Vocal Motor Exploration. *J. Neurosci.* **38,** 9635–9647 (2018).

245. Morandell, K. & Huber, D. The role of forelimb motor cortex areas in goal-directed action in mice. *Scientific Reports* **7,** 15759 (2017).

246. Omlor, W. *et al.* Context-dependent limb movement encoding in neuronal populations of motor cortex. *Nature Communications* **10,** 4812 (2019).

247. Ito, M. & Doya, K. Distinct neural representation in the dorsolateral, dorsomedial, and ventral parts of the striatum during fixed- and free-choice tasks. *Journal of Neuroscience* **35,** 3499–3514 (2015).

248. Balleine, B. & O'Doherty, J. Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology* **35,** 48–69 (2010).

249. Hikosaka, O. *et al.* Parallel neural networks for learning sequential procedures. *Trends Neurosci.* **22,** 464–471 (1999).

250. Aoki, S. *et al.* An open cortico-basal ganglia loop allows limbic control over motor output via the nigrothalamic pathway. *eLife* **8,** e49995 (2019).

251. Semedo, J., Zandvakili, A., Machens, C., Yu, B. & Kohn, A. Cortical Areas Interact through a Communication Subspace. *Neuron* **102,** 249–259.e4 (2019).

252. Sani, O., Abbaspourazad, H., Wong, Y., Pesaran, B. & Shanechi, M. Modeling behaviorally relevant neural dynamics enabled by preferential subspace identification. *Nature Neuroscience* **24,** 140–149 (2021).

253. Egert, D. *et al.* Cellular-scale silicon probes for high-density, precisely localized neurophysiology. *Journal of Neurophysiology* **124,** 1578–1587 (2020).

254. Mathis, A. *et al.* DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience* **21,** 1281–1289 (2018).

255. Makkeh, A., Theis, D. & Vicente, R. BROJA-2PID: A Robust Estimator for Bivariate Partial Information Decomposition. *Entropy* **20,** 271 (2018).

256. Safaai, H, Neves, R., Eschenko, O, Logothetis, N. & Panzeri, S. Modeling the effect of locus coeruleus firing on cortical state dynamics and single-trial sensory processing. *Proceedings of the National Academy of Sciences* **112,** 12834–12839 (2015).

257. Giordano, B. *et al.* Contributions of local speech encoding and functional connectivity to audio-visual speech perception. *Elife* **6** (2017).

258. Kaufman, M. T., Churchland, M. M., Ryu, S. I. & Shenoy, K. V. Cortical activity in the null space: permitting preparation without movement. *Nature Neuroscience* **17,** 440–8 (2014).

259. Ince, R. Measuring Multivariate Redundant Information with Pointwise Common Change in Surprisal. *Entropy* **19,** 318 (2017).

260. Rauh, J., Banerjee, P. K., Olbrich, E., Montúfar, G. & Jost, J. Continuity and additivity properties of information decompositions. *International Journal of Approximate Reasoning* **161,** 108979 (C 2023).

261. Barrett, A. B. Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems. *Phys. Rev. E* **91,** 052802 (5 2015).

262. Barrett, A. B., Barnett, L. & Seth, A. K. Multivariate Granger causality and generalized variance. *Physical Review E* **81,** 041907 (2010).

263. Novelli, L, Wollstadt, P, Mediano, P, Wibral, M & Lizier, J. Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing. *Network Neuroscience* **3,** 827–847 (2019).