A Case for Re-examining Synergy in the Brian

December 1, 2020

Abstract

The concept of synergy is not new to neuroscience; it has been theoretically and experimentally shown to arise in the brain. We argue in this paper that synergy is important for a renewed set of reasons. Firstly, recent work of ours has shown theoretically that it is not always possible to track information flow in the brain unless we account for synergy in the system. We provide a concrete demonstration of the importance of synergy in tracking information flow through a series of simulations. Secondly, we show that synergy may arise in ways that one may not expect, using a simulated case study on grid cells: the same neural activity may be encoding information about location either uniquely or synergistically, depending upon the resolution at which we interrogate location. Lastly, we believe that this is a ripe time at which to re-examine synergy as there have been significant recent advances in the information theory literature on developing well-motivated ways of *quantifying* synergy. We show how one such quantification may be applied in practice, using the aforementioned case study on grid cells.

1 Introduction

Synergy informally refers to the notion that a whole can be more than the sum of its parts. In information theory, this takes the form of two variables X and Y *jointly* conveying some information about a message M that cannot be obtained from any one of them *individually*. While this statement offers only a vague intuition for synergy, several recent works in the information theory literature have proposed concrete definitions for synergy (Williams and Beer, 2010; Harder et al., 2013; Bertschinger et al., 2014; see Lizier et al., 2018 for a recent review). Some of these definitions are rooted in strong operational interpretations, relying on ideas from statistical decision theory. There have also been significant efforts towards finding efficient and practical estimators for these definitions (Banerjee et al., 2018). These advances suggest that partial information measures—measures of unique, redundant and synergistic information—are ready to be used in neuroscience.

Synergy has been explored by many works in neuroscience over the last two and a half decades. For instance, Schneidman et al. (2003) identified three different kinds of "independence" in the neural code—activity independence, conditional independence and information independence—the last of these is related to the synergy between different cells about the stimulus. Gat and Tishby (1999) experimentally identified the presence of synergy in cats.. More recently, works by Timme and Lapish (2018) and Pica et al. (2017) have described how partial information measures such as unique, redundant and synergistic information may be used in neuroscience. We revisit the works of both Schneidman et al. (2003) and Timme and Lapish (2018) in a later section, providing additional context and contrasting some of the finer points of our results with theirs.

Despite the existence of past work addressing synergy in a multitude of ways, we believe that this concept deserves to be re-examined. Our argument rests on two points, both of which we illustrate in this paper using simulations:

1. A recent result of ours (Venkatesh et al., 2020b) shows that one *must* account for synergy in some form, in order to provably *track* how information about a stimulus flows through the brain. Understanding

such dynamical flows of information in the brain, in turn, is essential if we wish to intervene to modify such flows, particularly for the treatment of various brain diseases and disorders. Unless we account for synergy, there will always be instances where we cannot consistently identify the flow of information about a stimulus or response. In the present work, we provide concrete examples of such instances by simulating circuits at three different scales of neural processing.

2. We also show that synergy can arise in the brain in surprising ways, using a case study on grid cells. Borrowing from existing models of encoding and error correction in grid cells (Sreenivasan and Fiete, 2011), we build a model for the joint activity of three grid cell modules and examine how information about spatial location is decomposed between these modules. These simulations show that when interrogating information about spatially refined location, each module provides unique information with respect to the others; and when these grid modules possess the capacity for error correction, they provide redundant information with respect to the others. However, when interrogating information about location at a coarse spatial resolution, grid cells encode the same information synergistically.

These two results illustrate that synergy may be more common than previously assumed, and even unexpected in some instances, while at the same time being essential for understanding information flows. Our results also reveal interesting aspects of neural encoding and the nature of synergistic information. In conjunction with the aforementioned advances in developing well-motivated measures of synergistic information, these questions appear to be ripe for further investigation through experimental analyses.

2 Results

2.1 Background

Before we can explore the importance of synergy, we first provide an intuitive explanation of what synergy *means*. Then, we describe a few different ways in which prior literature in neuroscience has tried to understand synergy. Finally, we describe how synergy has been formalized through recent advances in the information theory literature. This section assumes that the reader is familiar with basic information-theoretic concepts such as Shannon entropy and mutual information (Cover and Thomas, 2012). For convenience, these are summarized in Table 1 at the end of this section.

Intuitively, synergy refers to the idea two variables X and Y can provide *more* information about some message M when taken *together*, than when considered *individually*. Schneidman et al. (2003) operationalized this intuition literally, by considering the difference of total and individual mutual informations. They defined a quantity that we denote $\mathrm{Syn}(M : X; Y)$, referring to the aforementioned difference:

$$\mathrm{Syn}(M : X; Y) \coloneqq I\big(M; (X, Y)\big) - I(M; X) - I(M; Y) \tag{1}$$

Schneidman et al. (2003) argued (correctly) that if this quantity was positive, then X and Y had synergistic information about M, and that if it was negative, then they had redundant information about M.

However, as we shall see, this definition suffered from an important issue, i.e., it did not allow for both synergy and redundancy to be present simultaneously. In fact, the aforementioned quantity was the difference of synergistic and redundant information: thus, while $\mathrm{Syn}(M : X; Y)$ was positive whenever synergy exceeded redundancy and negative whenever redundancy was greater, it would always *underestimate* each as long as the other was present. Moreover, synergy and redundancy could precisely balance each other, leading to a cancellation of the two quantities. Some of these issues were recognized by Schneidman et al. (2003), but there were no better ways of quantifying synergy and redundancy at that time.

Next, we present a more current understanding of synergy, arising out of the literature on Partial Information Decomposition (PID). The initial seeds of this understanding were first laid by Williams and Beer (2010), and was subsequently advanced through a series of works, including those of Harder et al. (2013), Griffith and Koch (2014) and Bertschinger et al. (2014) (Lizier et al., 2018 provide a recent review, and Timme and Lapish, 2018 explain how the PID may be used in the neuroscientific context).

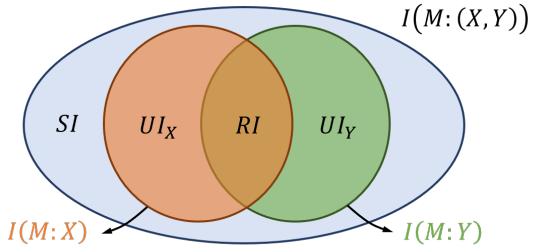


Figure 1: A Venn diagram summarizing the partial information quantities and their interactions, described in equations (2) and (3).

Williams and Beer (2010) suggested that synergy appears as part of a more general decomposition of the mutual information between M and (X, Y):

$$I\big(M;(X,Y)\big) = UI(M : X \setminus Y) + UI(M : Y \setminus X) + RI(M : X;Y) + SI(M : X;Y). \tag{2}$$

The four terms on the right hand side are respectively the information about M *uniquely* contained in X and not in Y, that uniquely contained in Y and not in X, that *redundantly* expressed in both X and Y, and that which only arises out of a *synergistic* combination of X and Y.

To understand intuitively what these four terms mean, consider the following example:

$$\begin{aligned} M &= [M_1, M_2, M_3, M_4] \\ X &= [M_1, M_3, M_4 \oplus Z] \\ Y &= [M_2, M_3, Z] \end{aligned}$$

where $M_1, M_2, M_3, M_4, Z \sim$ i.i.d. Ber(1/2), and $\oplus$ refers to the exclusive-OR (XOR) of two binary variables. Thus, M has four bits of entropy, spread evenly across M_1 through M_4. X has 1 bit of unique information about M, encapsulated in M_1, since this is information that cannot be extracted from Y. Similarly, Y has 1 bit of unique information about M not present in X, captured by M_2. M_3 constitutes 1 bit of redundant information which can be extracted from either X or Y. Finally, M_4 is present in neither X nor Y since $M \oplus Z$ and Z are both *individually* independent of M_4. However, when X and Y are taken *together*, we can reconstruct M_4 from the combination $[M_4 \oplus Z, Z]$. Thus, X and Y have 1 bit of synergistic information about M.

Intuition demands that the decomposition in (2) also imposes two other constraints:

$$\begin{aligned} I(M;X) &= UI(M : X \setminus Y) + RI(M : X;Y) \\ I(M;Y) &= UI(M : Y \setminus X) + RI(M : X;Y), \end{aligned} \tag{3}$$

since we expect that the total information about M present in X, $I(M;X)$, is the sum of the information about M uniquely present in X and the information redundantly encoded in both X and Y (which can be extracted from either). These constraints are summarized in the Venn diagram shown in Figure 1. Since we have four undefined partial information quantities and three constraints in equations (2) and (3), defining any *one* of the four partial information measures suffices to determine the rest.

Williams and Beer (2010) gave a formal definition for these partial information quantities, which is often called the Minimum Mutual Information (MMI) decomposition. Their decomposition was based on a definition for the redundant information:

$$RI_{MMI}(M : X;Y) \coloneqq \min\{I(M;X), I(M;Y)\}. \tag{4}$$

Unfortunately, this definition had some critical shortcomings, which is easily seen through a modification of the example given above. Consider the setting where $M = [M_1, M_2]$, $X = M_1$ and $Y = M_2$. Then, based on our intuition, we expect X and Y each to have 1 bit of unique information about M. However, the MMI

PID will find that redundant information is 1 bit, since $I(M;X) = I(M;Y) = 1$, the unique information in X and Y is therefore 0, by equation (3), and synergistic information is therefore 1 bit, from equation (2). Thus, the MMI PID finds 1 bit each of redundant and synergistic information in the simplest of examples where we intuitively expect 1 bit of unique information in X and Y each.

Unfortunately, being the easiest of the PID measures to compute, the MMI PID is the one that has been used most. Our work is the first, to our knowledge, which computes more complex PID measures in neuroscientific examples.

Bertschinger et al. (2014) proposed a better PID which satisfies our intuitive expectations in a larger number of cases and has better operational foundations, coming from statistical decision theory. Their decomposition defines the *unique* information about M present in X and not in Y as

$$UI(M : X \setminus Y) \coloneqq \min_{q \in \Delta_p} I_q(M; X \mid Y) \quad \text{where} \quad \Delta_p = \{q : q(m,x) = p(m,x), q(m,y) = p(m,y)\}. \tag{5}$$

The central intuition behind this definition arises from the following two points:

1. From equations (2) and (3), we have

$$I(M;X \mid Y) = I\big(M;(X,Y)\big) - I(M;Y) \tag{6}$$
$$= UI(M : X \setminus Y) + SI(M : X;Y) \tag{7}$$

   Thus, the conditional mutual information is the sum of the respective unique and synergistic components.

2. Bertschinger et al. (2014) argue that UI and RI should not depend on the full joint probability distribution $p(m,x,y)$, rather, they should depend only on the marginal $p(m)$ and the conditionals $p(x \mid m)$ and $p(y \mid m)$ (they justify this using a motivation from statistical decision theory). Since these marginals and conditionals are identical by definition for all $q \in \Delta_p$, UI and RI are constant over Δ_p. By taking the minimum conditional mutual information over this entire set, we are intuitively squeezing out the synergistic component (given that it is always non-negative), and defining what remains to be the unique information.

The PID framework described above helps us understand where the quantity used by Schneidman et al. (2003) falls short. Specifically, we have

$$\mathrm{Syn}(M : X;Y) = I\big(M;(X,Y)\big) - I(M;X) - I(M;Y) = SI(M : X;Y) - RI(M : X;Y). \tag{8}$$

Thus, while positive values of $\mathrm{Syn}(M : X;Y)$ may indicate the presence of synergy, and negative values the presence of redundancy, neither of these implies the absence of the other. It is also possible that both synergy and redundancy can be positive but exactly equal, cancelling each other out. An example of exactly this nature was seen above.

Finally, we note that Schneidman et al. (2003) also discuss a notion of *conditional* independence; however, their definition states that $I(X;Y \mid M) = 0$. This is very different from the notion of conditional mutual information that we will discuss in the context of information flow. We use $I(M;X \mid Y)$. This distinction will be highlighted later.

Notation	Meaning
$H(M)$	Shannon entropy of the random variable M
$I(M;X)$	Shannon mutual information between the random variables M and X
$UI(M : X \setminus Y)$	Unique information about M present in X and not in Y
$RI(M : X;Y)$	Redundant information about M present in X and in Y
$SI(M : X;Y)$	Synergistic information about M present between X and Y

Table 1
Information-theoretic notation used throughout the paper. All information quantities are measured in bits.

2.2 Synergy and Information Flow

Next, we look at how synergy is important for detecting information flow. Through examples, we show that unless we *account for synergy*, it is not always possible to track the paths along which information flows in neural circuits. Measures that account for synergy are those that use some form of *conditioning*, e.g., conditional mutual information, conditional correlation or partial correlation. Simpler measures based purely on Pearson correlation are unable to consistently track the paths along which information about a stimulus flows. We show this using simulated examples covering three different scales of neural information processing:

1. Neural circuits processing information encoded in single spikes
2. Circuits processing information encoded in spike trains
3. Information encoded at a population level, in the aggregate activity of multiple neurons

Our simulations are all based on networks of neurons; these rely on a reparametrized version of the Quadratic Integrate-and-Fire neuron model known as a "Theta" neuron model (Ermentrout and Kopell, 1986). Particulars of the simulation setup may be found in the Methods section.

2.2.1 Information Flow in a Simple XOR Circuit

We begin with a simple demonstration of how synergy may arise in a neural circuit, using the canonical example of exclusive-OR (XOR) operations. We implement an XOR operation using a network of three theta neurons. This is achieved as follows:

$$M \text{ XOR } Z = (M \text{ OR } Z) \text{ AND NOT } (M \text{ AND } Z) \tag{9}$$

The XOR operation is realized by dividing it into one OR and two AND operations, each of which is implemented using a theta neuron with synaptic weights set appropriately, in relation to the neuron's threshold (this is depicted in Figure 2a). In the above, M is a "message" (which can be thought of as a stimulus) that the network is trying to encode or convey, while Z is a noise variable representing an independent signal, or internal neural variability. M and Z are both encoded in the form of single spikes, i.e., if $M = 1$ in a given trial, a neuron receiving M as input would receive a single spike, and if $M = 0$, it would receive no spike.

In all that follows, the three-unit XOR network is condensed into a single "node" for the purpose of examining information flow. Using this XOR node, we design a circuit with the intent of demonstrating how accounting for synergy is essential when inferring information flow (shown in Figure 2b). This circuit consists of three nodes: the first node X_1 performs an XOR of M and Z; the second node X_2 acts as a delay element and preserves the noise variable Z along a separate path; and the third node X_3 performs another XOR operation, removing the noise Z and recovering the message M. We see that effectively, information about the message M is never lost in the circuit. However, at an intermediate time step, it is preserved synergisitcally in the form $[M \oplus Z, Z]$ along two separate edges. Thus, when inferring information flow at this time instant, it is necessary to account for synergy in the system. Otherwise, we lose track of where information about M is present in the system.

Figure 2c shows the response of the network for all possible values of M and Z. If we analyze the information flow of M on different transmissions of this network, we find that around $t = 5$ms, spikes corresponding to M and Z are seen, each independent of the other and equally likely to be zero or one (corresponding to $H(M) = 1$ bit). Around $t = 24$ms, we find that X_3 shows perfect correlation with M, so that $I(M; X_3) = 1$ bit. Around $t = 14$ms, however, X_1 and X_2 both show no dependence on M, i.e., $I(M; X_1) = 0$ and $I(M; X_2) = 0$.

If we are aware of the underlying anatomy, this should strike us as perplexing, since M appears to have bypassed X_1 and X_2 to arrive at X_3, even though there are no other nodes in the system. The issue of course, lies with how we evaluate the "presence of information about M", which does not account for possible synergy between X_1 and X_2. Our previous theoretical work (Venkatesh et al., 2020b) proposed to resolve this issue by conditioning on X_2 when examining the information flow of M on X_1. Our definition of M-information flow (see Venkatesh et al., 2020b, Definition 4) states that an edge carries information flow about

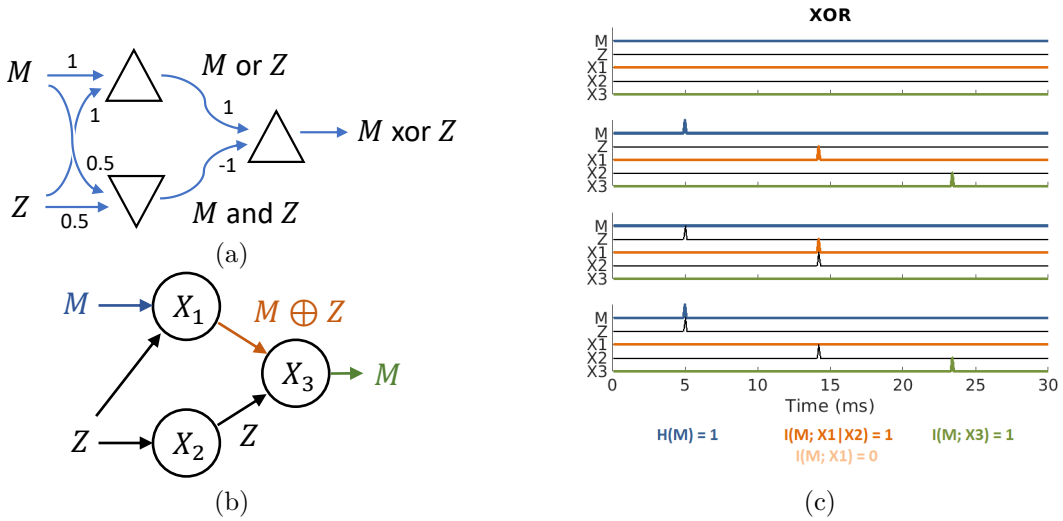


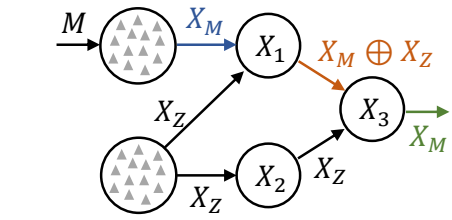
Figure 2: Simple XOR and OR networks demonstrating synergy in a neural circuit. (a) A depiction of the circuit designed to perform an XOR operation. Here, all neurons have a threshold of 1, and the numbers on edges are an indication of synaptic weight. The first uppermost neuron is excitatory and fires if either M or Z is active, thus it performs an OR operation. The lowermost neuron is inhibitory and fires only if both M and Z are active, and therefore performs an AND operation. Since the lower neuron is inhibitory, the neuron to the right fires only if the upper neuron fires and the lower one does not; effectively resulting in an exclusive-OR of M and Z.

M if its transmission depends on M, allowing for conditioning on other concurrent transmissions. This in fact reveals the flow of information about M on X_1, since $I(M; X_1 \mid X_2) = 1$ bit in this example.

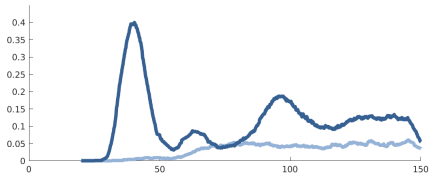
Interestingly, this also implies that X_2 has information flow of M, since $I(M; X_2 \mid X_1) = 1$ bit. Indeed, any time a transmission synergistically contributes to the information flow of M, our definition considers it to have information flow. The reason for this is that it is not easy to determine *which* of two edges that synergistically contribute to information about M is "actually" responsible for carrying that information. Indeed, this precise question was addressed in much greater theoretical depth in another work of ours (Venkatesh et al., 2020a). For the purposes of our discussion here, it suffices to note that this is not necessarily undesirable, and that there are pruning-based methods that can, in most cases, remove such edges if the need arises.

The idea that synergy can be expressed using an XOR operation, which can be operationalized using neuron models, is hardly new. Such examples, in the simplest setting, have also been pointed for instance, by Timme and Lapish (2018) and more broadly in the PID literature (Harder et al., 2013; Bertschinger et al., 2014). However, the idea that synergy plays an important role in inferring information flow has not been pointed out before, until our earlier work (Venkatesh et al., 2020b). What we have shown here is that the interplay of synergy and information flow may arise in real neural circuits.

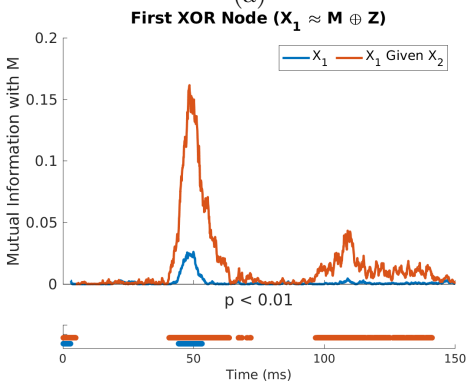
We should also note that examples where accounting for synergy is essential are not limited to the case of XOR's. Such situations may also arise in simpler settings, such as with excitatory addition followed by inhibitory cancellation. There are plenty of biological examples of such self-cancellation circuits, the most common being those of efference copies (von Holst and Mittelstaedt, 1971), or corollary discharge (Fukutomi and Carlson, 2020). There is also reason to believe that synergistic encoding is the product of a certain kind of information mixing, which is common in compression and error-control contexts (we will see one such example that uses grid cells in a later section). Such information mixing is likely to arise in the olfactory cortex, where there is evidence of a compressive-sensing–type circuit (Zhang and Sharpee, 2016). Lastly, coming back to XOR's, there has even been recent evidence to show that dendrites may compute XOR's (Gidon et al., 2020).



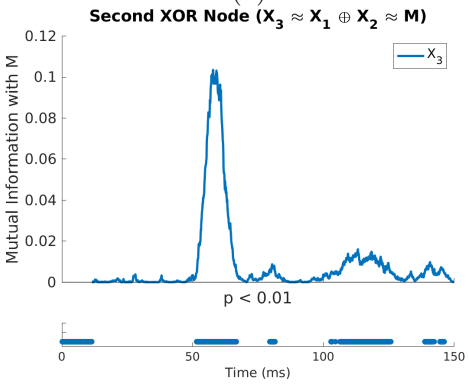
(a)



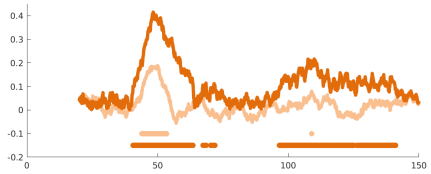
(b)



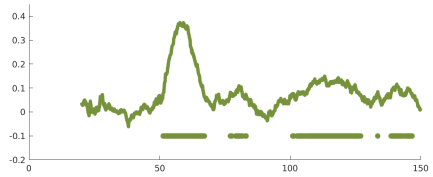
(c)



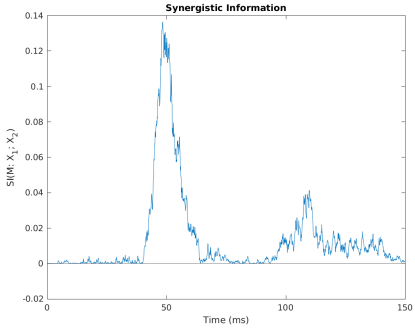
(d)



(e)



(f)



(g)

Figure 3

2.2.2 Information Flow in a Spike Train Encoding Model

The previous example was simplistic by design, to make the argument for synergy and information flow succinctly. It designed using single spikes in order to work cleanly with the XOR circuit model, which required somewhat precise timing of spike arrival to allow for cancellation of Z. Next, we consider a more complex scenario, to show that our conclusions about synergy and information flow are not affected by the aforementioned assumptions. We model a situation where the stimulus M is encoded by a sensory system in the form of a *spike train*. M is then passed on to a downstream region for further processing, and we are interested in how information about M flows in this downstream network.

Once again, to examine the importance of synergy, we take the downstream circuit to be the XOR network we examined before, where information about M is corrupted by noise Z, and is subsequently recovered through some form of cancellation. Therefore, the circuit being analyzed is the same as in Section 2.2.1, but M and Z are now encoded using spike trains rather than as single spikes. We use X_M to denote the spike train for M and X_Z to denote that for Z. The spike trains are generated by a randomly connected network of theta neurons with balanced excitation and inhibition, as shown in Figure 3a. The spike train X_M is the output of a single neuron from this network which encodes the value of M at most time instants. The spike train X_Z is taken to be very noisy, having an equal likelihood of firing and not firing at every time instant.

In this case, we first note that the message M is discernable from the spike train X_M only at certain distinct intervals of time. This is seen in the peristimulus time histogram (PSTH) of X_M shown in Figure 3b, where M is discernable from X_M only when the light and dark curves (corresponding to $M = 0$ and $M = 1$ respectively) are separated. During these time intervals when M is discernable, we examine whether or not the relevant nodes in the XOR network reveal information flow about M.

We measure information flow about the message M in a few different ways: first, we use the measure proposed in our earlier work (Venkatesh et al., 2020b). This is depicted in Figures 3c,d: observe that the transmissions of X_3 show statistically significant dependence with M in the 50–65ms and 100–125ms time periods in Figure 3d. This corresponds nicely with the time intervals where M is discernable in Figure 3b (approximately 30–50ms and 90-110ms respectively). Figure 3c shows that simple mutual information, $I(M; X_1)$, does not reveal statistically significant information flow about M in the transmissions of X_1, especially in the 95-115ms time interval. However, *conditioned* on X_2, we see strong conditional dependence in the transmissions of X_1, once again proving the importance of accounting for synergy when inferring information flow.

Since (conditional) mutual information is a difficult quantity to estimate in general, we also show how the same inferences can be obtained using a simpler adaptation of this measure. We use a correlation-based approximation of conditional mutual information that we call the mean absolute conditional correlation (MACC), defined as

$$\mathrm{MACC}(M : X; Y) \coloneqq \mathbb{E}_y \big| \rho(M, X \mid Y = y) \big|, \tag{10}$$

where $\rho(M, X \mid Y = y)$ refers to "conditional correlation", i.e., the correlation between M and X in the conditional distribution $p_{M,X \mid Y}(m, x \mid y)$, and the expectation in (10) is taken with respect to the marginal distribution of Y, i.e., $p_Y(y)$.

Figures 3e,f show analogous results to those we see for mutual information. Only upon conditioning are we able to track the paths along which information flows in this network. In particular, during the time interval 100–120ms, we see evidence of M at the output of X_3, but it is not clear how it got there when we use only mutual information to examine X_1.

We also show an estimate of the synergy in the system using the definition of Bertschinger et al. (2014), and the estimation methods of Banerjee et al. (2018); this can be seen in Figure 3g. This is the first concrete demonstration of a sophisticated partial information measure in a neuroscientific context. As we might expect, the synergy between X_1 and X_2 about M is large at precisely those time intervals when we see information flow.

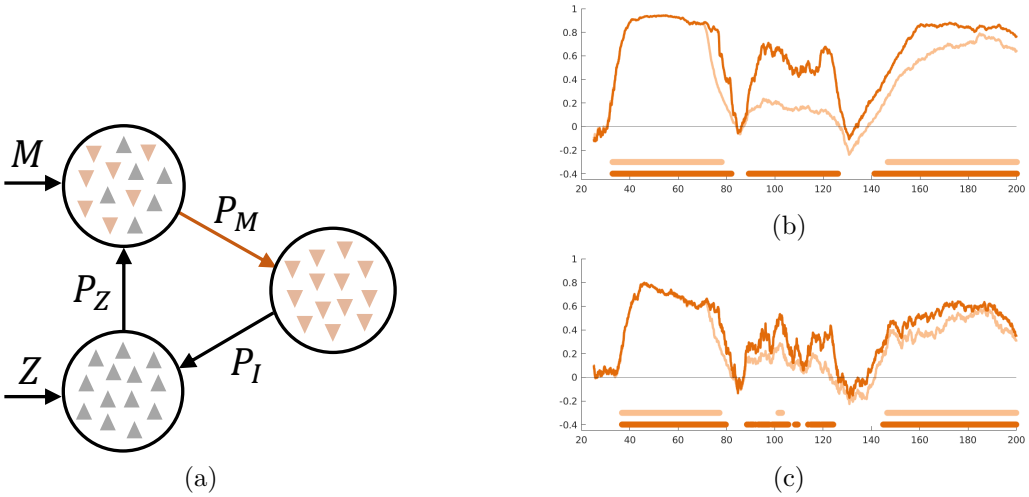


Figure 4

2.2.3 Information Flow in a Population Model

Lastly, we examine a scenario where the binary stimulus M is encoded by a *population* of neurons P_M in their average firing rate. This scenario is meant to emulate a setting where we use a multi-electrode array to record from a few different brain regions involved in a task. We also examine a setting where we subsample a fraction of the neurons in this region as might be the case with a multielectrode array.

Once again, to show how synergy might arise in such a system, we corrupt the message in P_M with noisy inputs arising from a second population P_Z, which encodes a continuous noise variable Z. Subsequently, if the average firing rate of P_M is large, a third population P_I, which is primarily inhibitory, suppresses P_Z after an axonal delay. This is depicted in Figure 4a. Our primary objective is to track which edges carry information about M at various time instants in this feedback network. In this setting, we measure information flow using a different approximation of conditional mutual information, namely, partial correlation.

Figure 4b is analogous to Figure 3c for the spike train model. It shows that, unless we are using partial correlation, we do not see statistically significant information flow about M during the 90–130ms time interval. Even upon significant subsampling of these populations, we find that the average firing rate robustly encodes the message; this can be seen in Figure 4c, where only 10% of all neurons are being sampled. We see that partial correlation picks up synergistic encoding even when recording from just 20 random excitatory neurons in P_M.

In this example, our computations assume that we know information about M is encoded in the average firing rate of the P_M population. In practice, one may need to determine the manifold along which information about M is encoded using dimensionality reduction approaches. For the case of the example provided here, canonical correlation analysis would reveal that the average activity of all neurons encodes the value of M. However, as long as information is encoded in a dense subspace of neural activity, and the subsampling mechanism is random with respect to this encoding mechanism, we would expect our results of subsampling to continue to hold.

2.2.4 Remarks on our Analysis and Assumptions

A caveat to partial correlation. In the spike train model, we used MACC as an approximation for conditional mutual information to measure information flow, while in the population model, we used the more well-known partial correlation. The reason we did this is that partial correlation does not yield statistically significant information flow in the spike train model: we found that $\rho(M, X_1 \mid X_2 = 0)$ and $\rho(M, X_1 \mid X_2 = 1)$ are nearly equal in magnitude and opposite in sign; therefore, upon taking an expectation with $p(X_2)$, the

two quantities cancel, leading to very small values of partial correlation, which are statistically indistinguishable from zero. This is why we use the mean *absolute* conditional correlation (MACC) instead, which takes absolute values to prevent cancellation. In general, one might want to start by trying to use partial correlation, which is a well-established method that often has easily available off-the-shelf implementations. If partial correlation reveals information flow, then one the analysis is complete, but if not, then one cannot conclude that there is *no* flow. Instead an alternative approach based on MACC or some other approximation for conditional mutual information should be pursued.

Discontinuous information flow. Throughout our examples, we stressed on the idea that we would like to be able to track the *paths* along which information about the message M flows. In particular, we wanted to be aware of which edges and which transmissions carried information about M at every instant of time. However, there were still many time instants when it was unclear where the message was: for example, in the single spike XOR model, the time intervals between spikes revealed no information flow of the message M. In fact, information about M was still present in the network, however, it was being communicated in the form of membrane voltages along axons or dendrites and could not be seen in spikes or firing rates at the cell body. This points to a crucial hidden variable in the system: namely, the voltages on the axonal and dendritic membranes. We will only be truly able to track information flow at the resolution of axonal delays if we also measure these variables (perhaps using voltage sensitive dyes). However, in practice, we find that we can get reasonably continuous and satisfactory estimates of flow (while accounting for synergy) due to random latencies and neural variability; as evidenced in both the spike train and population encoding models.

2.3 Synergy and Encoding in Grid Cells

In this section, we present a case study on entorhinal grid cells, showing how synergy may arise in interesting (and possibly surprising) ways in biological neural systems. We begin with a short introduction on how grid cells encode information about where an animal is spatially located.

2.3.1 A Brief Introduction to Grid Cells

Grid cells are neurons in the entorhinal cortex, which are thought to encode information about where an animal is spatially located (e.g., within a room). There are a few models of how grid cells might convey such information (Sreenivasan and Fiete, 2011; Wei et al., 2015); we refer to the work of Sreenivasan and Fiete (2011), which is briefly described in what follows.

Each grid cell has a distinct periodic firing pattern, in that its firing rate is modulated at periodic spatial intervals. Furthermore, grid cells are organized into groups, or "modules", that all have the same periodicity (or wavelength) in their firing patterns, though their patterns may be shifted with respect to the others'. Since the cells within a module all have the same *period* but different *phase offsets* in their firing fields, these cells can be thought of as constituting a *population code* encoding the *phase* of the animal's location within that module's wavelength. As a result, the joint activity of all cells within a module can only describe the animal's position *modulo* that module's wavelength.

In order to encode location beyond a single wavelength, the entorhinal cortex consists of multiple such grid cell modules, each with their own distinct wavelength. A simple yet effective way of understanding how grid modules jointly encode information about location is to visualize the *residual uncertainty* in an animal's location, given the activity of a module. A one-dimensional version of this is depicted in Figure 5a. If the prior distribution on the animal's location was uniform, then the posterior probability of location given the activity of a single module looks like a series of periodic peaks, separated by the wavelength of that module. Figure 5b shows how this posterior distribution shifts as the animal moves. The animal's location is uniquely determined only when we consider the joint activity of multiple modules: this is shown in Figure 5c. In particular, the posterior distributions of all grid modules *align* at the animal's true location, so that the *product* of these posterior distributions peaks at the true location.

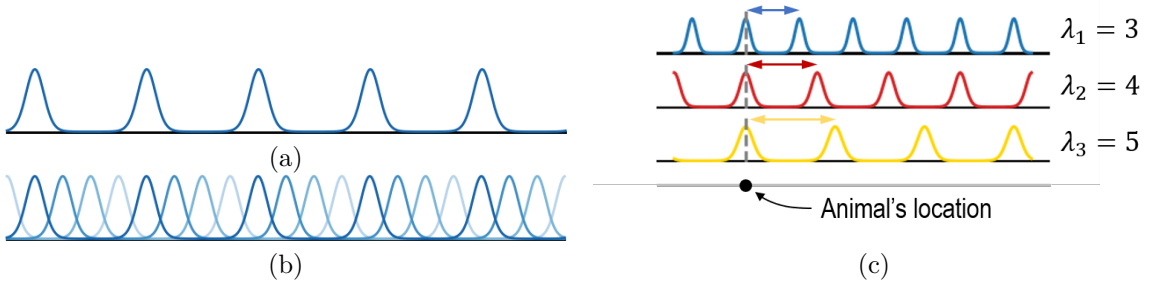


Figure 5

Sreenivasan and Fiete (2011) suggest that, in order to maximize the amount of space grid cells can encode, the wavelengths of different modules ought to be "co-prime" (or "incommensurate") with respect to each other. The expectation is that the total "range" that can be covered using an encoding scheme such as this is *exponential* in the number of modules, or more precisely, of the order of the *product of their wavelengths*. Sreenivasan and Fiete (2011) also claim that the maximum range that may be encoded using all modules is far greater than is likely to be necessary in an animal's lifetime; thus an animal would instead encode only a restricted range, so that any additional modules are effectively used as redundancy against neural variability.

A key takeaway from this depiction is that, given the activity of a single module, there is typically still a lot of residual uncertainty which is spread across the entire possible range of movement. It is only when we put information from several modules *together* that we get a refined understanding of the animal's location. Since information about location is encoded *jointly* by multiple modules, and no one module reveals this information on its own, this system provides an excellent opportunity for understanding how partial information measures may be useful in practice.

To understand the broader applicability of the PID framework, we examine situations encompassing all three types of partial information: unique, redundant and synergistic. However, in keeping with the central theme and motivation of the paper, we will focus on how synergy arises in grid cells, and what this teaches us about both synergy and information encoding. The introduction to grid cells above, as well as Figure 5 may suggest that information is primarily encoded synergistically, due to the fact that many modules come together to supply information about location: in what follows, we will see whether this is indeed the case.

2.3.2 Model setup

Next, we briefly describe how we setup a model for a few different grid modules encoding a one-dimensional location. To keep our simulation simple and to focus on parameters of importance, we forego a spiking neuron model and instead directly model the activity of an entire grid module. We do this by assuming a conditional distribution for the residual uncertainty in location, given each module's activity. In order to account for neural variability, we let these conditional distributions have different degrees of "variance" (see methods for details).

We consider a total of three modules: in our simulations and analyses, we use wavelengths of 9, 10 and 11 units; for the purpose of illustration, we use wavelengths of 3, 4 and 5 units (this is explicitly mentioned where needed). The conditional distributions are discretized to simplify the implementation of computing information measures. Once again, we use the PID of Bertschinger et al. (2014), and compute partial information measures using the implementation by Banerjee et al. (2018).

2.3.3 Unique, Synergistic and Redundant Information in Grid Cells

First, we examine the extent of unique, redundant and synergistic information between each module and the other two, in a setting where they encode the maximum encoding range of $9 \times 10 \times 11 = 990$. Figure 6 shows the unique, redundant and synergistic information in each module with respect to the other two, as

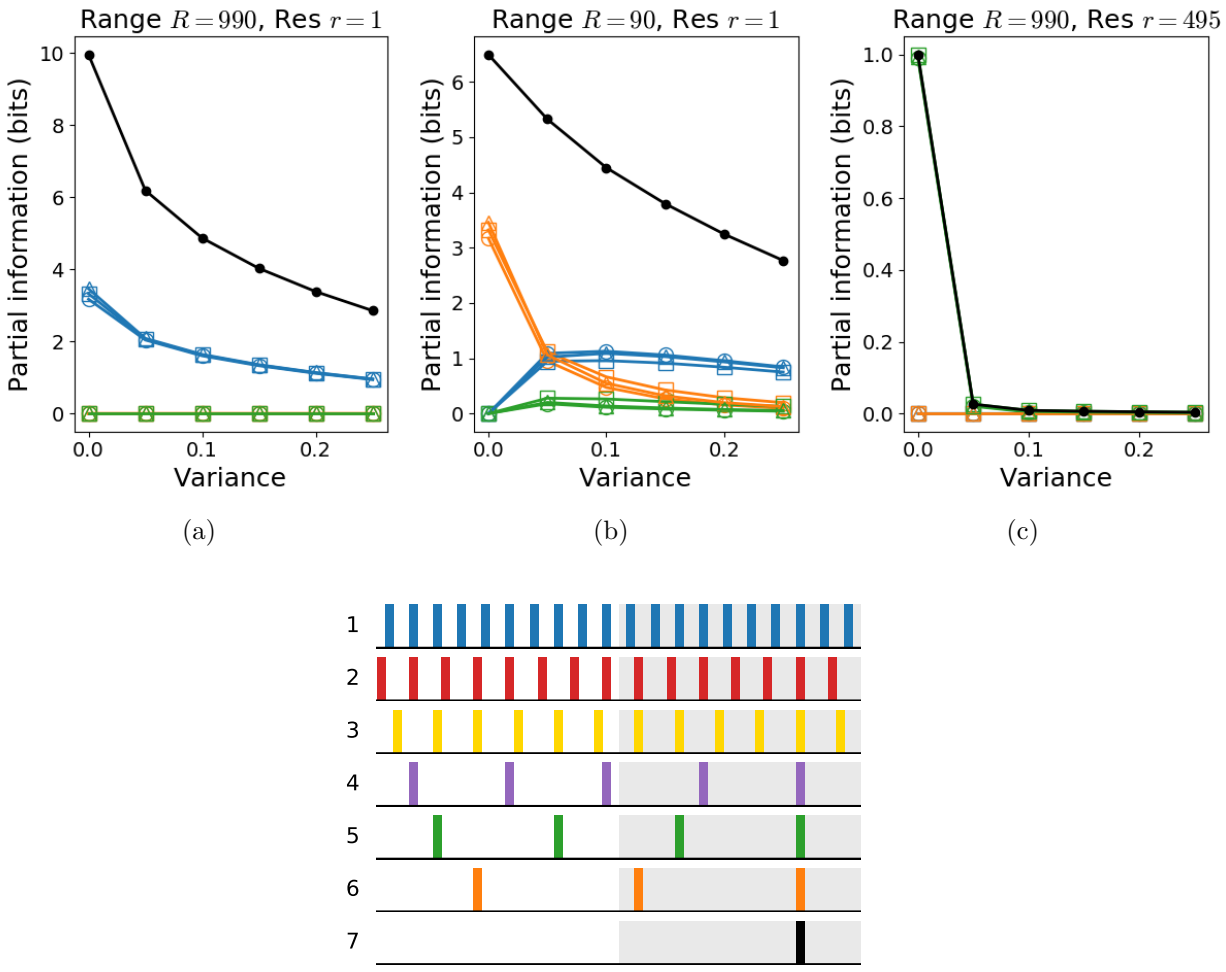


Figure 6

a function of increasing neural variability. The main takeaway from this figure is that all of the information content is actually *unique* to each module. In other words, while each module conveys little about location by itself, there are no *synergistic* effects. Thus, the joint information from two different modules is *not* greater than their sum. What this shows, therefore, is that each module reduces uncertainty about location in a way that is "orthogonal" in some sense to the others. This may even be expected, considering that we are operating at maximum "capacity", when encoding the maximum possible range.

Next, we examine a setting where we allow for a reduced encoding range. We consider a reduced range of $9 \times 10 = 90$, and compute unique, redundant and synergistic information about location between each module and the other two, as before (see Figure 6). Here, as expected, we find that in the absence of neural variability, each module contains purely redundant information with respect to the other two (since any two other modules suffice to encode this range). On the other hand, as variability rises, the redundancy drops shaprly, while uniqueness rises, and total mutual information drops more slowly. This suggests that error correction is in effect; indeed the presence of a combination of redundant and unique information could indicate some form of error correction in other settings as well. However, there is no synergy even in this setting, at any noise level.

Finally, to understand how synergy can arise in such a system, we change the "question" we are asking:

instead of looking at information about *precise* location, we consider information about *coarse* location, for example, is the animal in the left or right half of the room? When we change the question, or in effect, the *message* under consideration, we find that synergy arises in this system. The intuition for this is explained in Figure 6. We find that the residual uncertainty in location, given the activity of one module, spans both left and right halves of the room equally. Thus the uncertainty in left-vs-right remains as it did before we knew the activity of a module. The same applies when we know the activities of any *pair* of modules. Indeed, it is only when the activities of all three modules are known that the residual distribution collapses into one of the two halves of the room. This is indeed close to the canonical example for synergy: each module individually gives little to no information about a message, but jointly they explain everything about the message.

The main takeaway from this analysis is that synergy can arise in a circuit in unexpected ways: in this instance, changing the message changed how it was represented between different modules.

We believe that measuring partial information quantities may help distinguish between hypotheses such as that of Sreenivasan and Fiete (2011) and Wei et al. (2015).

3 Methods

3.1 Details of Simulations for Information Flow

3.1.1 Neuron model

Simulations used the theta model for neurons (Ermentrout and Kopell, 1986). The theta model is a change of variables from the standard Quadratic Integrate-and-Fire (QIF) model that expresses the voltage in terms of an angle on the unit circle, $V(t) = \tan(\theta/2)$. A neuron spikes when $\theta = \pi$ ($V \to \infty$) and is reset by subtracting 2π ($\theta \to -\pi$, $V \to -\infty$), thereby removing the discontinuity of the QIF model. Each neuron is governed by the set of differential equations

$$\frac{d\theta^k}{dt} = 1 - \cos(\theta^k) + (1 + \cos(\theta^k))(I_0^k + w_e^k s_e - w_i^k s_i) + \sigma\epsilon \tag{11}$$

$$\frac{ds_k}{dt} = \frac{1}{\tau_k}(-s_k + f_k) \tag{12}$$

where k indicates the type of neuron (excitatory or inhibitory), $\epsilon \sim \mathcal{N}(0, 1)$ is independently drawn for every neuron at every time step, and f_k is the number of presynaptic neurons of type k that fired at the last time step. The constant parameters are the input current I_0^k, strength of excitatory (inhibitory) synapses w_e^k (w_i^k), strength of noise σ, and synaptic decay time τ_k. The equations were numerically integrated using Euler's method ($dt = 0.1$ms).

3.1.2 Connectivity models

We considered three main connectivity models. As an intermediate step, three theta neurons were arranged to perform the XOR operation, as shown in Figure 2a. Our approach differs from the XOR gate in Timme and Lapish (2018) in that we rely on different connection weights to produce the desired effect rather than a constant background inhibition. Recent results from Gidon et al. (2020) suggest that cortical dendrites possess an activation function capable of computing XOR with individual neurons, thus we consider each XOR gate as a single node regardless of its exact implementation. Two of these gates and an excitatory neuron were used to produce the first network with synergistic information. Figure 2b shows this network unrolled in time, where X_1 and X_3 are the XOR gates and X_2 is the excitatory neuron. Table 2 gives the parameter values for these neurons.

The binary message variable M and the noise variable Z were represented by spike trains produced by large, sparsely connected networks of theta neurons (Table 3). To encode $M = 1$, a constant stimulus

$(I_0 \to I_0 + 0.05)$ was applied to a single excitatory neuron, raising its firing rate and propagating the message through the network. On the other hand, when $M = 0$ there was no such added stimulus. In both cases, the output of a different excitatory neuron in the network was used as the spike train input for M. Figure 3b shows spike histograms produced for M over 1000 trials. To encode the noise variable Z, the level of noise within the network (σ) was raised so that there was an almost 50% chance of a neuron spiking within each time bin. The resulting spike trains were uncorrelated from trial to trial.

Simulations were run for 150 ms and divided into 10 ms time bins. Correlations with the message, $\rho(M, X_i)$, were calculated for every time bin over 1000 trials, where X_i is the number of spikes produced by that node in the given time bin. To reveal synergistic information, we also calculated the conditional correlations $\rho(M, X_i \mid X_j = 0)$ and $\rho(M, X_i \mid X_j > 0)$. Calculation of p-values was done using the built-in Matlab function, and a significance level of $p = 0.01$ is shown on all figures for reference.

With population encoding, the nodes of the network became populations of theta neurons, and M and Z were encoded in the average firing rate of a population. Figure 4a shows the arrangement of these populations. P_M is the encoding population and receives input from M throughout the simulation/starting at 30 ms. M is again a binary message variable and determines whether neurons in P_M receive a constant input current, thereby determining the average firing rate (Table 4). P_Z is the noise population and receives input from Z after 40 ms/70 ms. Z is independently drawn from a uniform distribution between 0 and 1, and samples are drawn until the magnitude of the correlation between M and Z across trials is less than 0.0001. Z then determines the input current to X_2 (Table 5). P_I is the inhibition population, and it is designed so that it easily sustains input from P_M and provides strong inhibition to P_Z, with appropriate delays to better see the flow of information (Tables 6 and 7).

Simulations were run for 200 ms with a 30 ms transient period and divided into 10 ms time bins using a moving window. As with single neuron encoding, correlations with the message, $\rho(M, X_i)$, were calculated for every time bin over 100 and/or 1000 trials, where X_i was now the average firing rate of the population in a given time bin. Since the firing rate is nearly a continuous variable, partial correlations (instead of conditional correlations) were calculated to reveal synergistic information. Calculation of p-values was again done using the built-in Matlab function, and a significance level of $p = 0.01$ is shown on all figures.

3.2 Details of Simulations of Grid Cells

The conditional distribution for location given a grid module's activity was assumed to be a von Mises distribution (this is what is shown in Figure 5). For the purpose of computing partial information measures, however, we require discrete distributions; therefore we use a discretized version of the von Mises distribution. Neural variability affects the width of the resulting conditional distribution, and is parameterized using the circular variance of the von Mises distribution.

We compute partial information measures where the message is taken to be the discretized location (one of 990 possible locations when considering the full encoding range), and the two constituent variables are the activity of one module and the joint activity of the two others.

4 Discussion and Conclusion

Our simulations show that synergy may be prevalant in neural circuits: the XOR examples (both based on individual spikes as well as using spike trains) and the population coding example show that synergy is essential for inferring information flow; on the other hand the grid cell simulation shows that synergy may arise in a system when we change the message.

In other words, the grid cell example shows that even if we have previously examined and understood a system, novel stimuli may engender unexpected synergistic responses. Furthermore, unless we are able to identify and account for possible synergy (through conditioning of some form), we will be unable to track the paths along which information flows.

We also showed that synergy and its associated partial information measures of uniqueness and redundancy can be estimated in fairly complex settings using novel definitions and algorithms. The same applies to information flow: although our original definitions were based on conditional mutual information, one can often arrive at the same inferences using simpler measures such as partial and conditional correlation.

Our paper therefore makes a case for consciously examining the possibility of synergistic encoding in neural circuits and systems: because synergy may arise in ways we do not expect, because it affects our determinations of information flow, and because we now have the tools to measure it.

5 Supplementary Material

Parameter	Value
Constant input current to all neurons (I_0)	-0.03
Synaptic weight, large (w_L)	0.80
Synaptic weight, small (w_S)	0.40
Synaptic decay time for all connections (τ)	2
Strength of noise, XOR network (σ)	0.03
Strength of noise, excitatory/inhibitory network (σ)	0.04

Table 2
Parameter values for the single neuron encoding networks

Parameter	Value
Number of excitatory neurons	30
Number of inhibitory neurons	30
Probability of connection among all neurons	0.10
Constant input current to excitatory neurons (I_0^e)	0
Constant input current to inhibitory neurons (I_0^i)	0
Synaptic weight from excitatory to excitatory neurons (w_e^e)	0.30
Synaptic weight from excitatory to inhibitory neurons (w_e^i)	0.15
Synaptic weight from inhibitory to excitatory neurons (w_i^e)	0.50
Synaptic weight from inhibitory to inhibitory neurons (w_i^i)	0.20
Synaptic decay time for excitatory connections (τ_e)	2
Synaptic decay time for inhibitory connections (τ_i)	8
Strength of noise for M, XOR network (σ_M)	0.03
Strength of noise for M, excitatory/inhibitory network (σ_M)	0.04
Strength of noise for Z (σ_Z)	0.25

Table 3
Parameter values for the networks that generate M and Z spike trains in single neuron encoding

Parameter	Value
Number of excitatory neurons	200
Number of inhibitory neurons	200
Probability of connection among all neurons	0.10
Constant input current to excitatory neurons, $M = 0$ (I_0^e)	0
Constant input current to inhibitory neurons, $M = 0$ (I_0^i)	0
Constant input current to excitatory neurons, $M = 1$ (I_0^e)	0.01
Constant input current to inhibitory neurons, $M = 1$ (I_0^i)	0.005
Constant input current start time	0 ms/30 ms
Synaptic weight from excitatory to excitatory neurons (w_e^e)	0.30
Synaptic weight from excitatory to inhibitory neurons (w_e^i)	0.15
Synaptic weight from inhibitory to excitatory neurons (w_i^e)	0.50
Synaptic weight from inhibitory to inhibitory neurons (w_i^i)	0.20
Synaptic decay time for excitatory connections (τ_e)	2
Synaptic decay time for inhibitory connections (τ_i)	8
Strength of noise (σ)	0.10

Table 4
Parameter values for the encoding population (X_1)

Parameter	Value
Number of excitatory neurons	100
Number of inhibitory neurons	100
Probability of connection among all neurons	0.20
Constant input current to excitatory neurons (I_0^e)	$Z \sim \text{Uniform}(0, 1)$
Constant input current to inhibitory neurons (I_0^i)	$\frac{1}{2} I_0^e$
Constant input current start time	40 ms/70 ms
Synaptic weight from excitatory to excitatory neurons (w_e^e)	0.30
Synaptic weight from excitatory to inhibitory neurons (w_e^i)	0.15
Synaptic weight from inhibitory to excitatory neurons (w_i^e)	0.50
Synaptic weight from inhibitory to inhibitory neurons (w_i^i)	0.20
Synaptic decay time for excitatory connections (τ_e)	2
Synaptic decay time for inhibitory connections (τ_i)	8
Strength of noise (σ)	0.10

Table 5
Parameter values for the noise population (X_2)

Parameter	Value
Number of excitatory neurons	100
Number of inhibitory neurons	300
Probability of connection from excitatory to excitatory neurons	0.50
Probability of connection from excitatory to inhibitory neurons	0.75
Probability of connection from inhibitory to excitatory neurons	0.01
Probability of connection from inhibitory to inhibitory neurons	0.01
Constant input current to excitatory neurons (I_0^e)	-0.050
Constant input current to inhibitory neurons (I_0^i)	-0.025
Constant input current start time	0 ms
Synaptic weight from excitatory to excitatory neurons (w_e^e)	0.30
Synaptic weight from excitatory to inhibitory neurons (w_e^i)	0.15
Synaptic weight from inhibitory to excitatory neurons (w_i^e)	0.50
Synaptic weight from inhibitory to inhibitory neurons (w_i^i)	0.20
Synaptic decay time for excitatory connections (τ_e)	2
Synaptic decay time for inhibitory connections (τ_i)	8
Strength of noise (σ)	0.10

Table 6
Parameter values for the inhibition population (X_3)

Parameter	Value
X_1 to X_3 connection probability (excitatory to excitatory neurons)	0.10
X_3 to X_2 connection probability (inhibitory to excitatory neurons)	0.75
X_2 to X_1 connection probability (excitatory to excitatory neurons)	0.10
Delay between X_1 and X_3	10 ms
Delay between X_3 and X_2	15 ms
Delay between X_2 and X_1	0 ms

Table 7
Inter-population parameter values

References

Pradeep Kr Banerjee, Johannes Rauh, and Guido Montúfar. Computing the unique information. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 141–145. IEEE, 2018.

Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014. ISSN 1099-4300. doi: 10.3390/e16042161. URL http://www.mdpi.com/1099-4300/16/4/2161.

Thomas M Cover and Joy A Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

Bard Ermentrout and Nancy Kopell. Parabolic bursting in an excitable system coupled with a slow oscillation. *SIAM*, 46(2):233–253, 1986.

Matasaburo Fukutomi and Bruce A. Carlson. A history of corollary discharge: Contributions of mormyrid weakly electric fish. *Frontiers in Integrative Neuroscience*, 14:42, 2020. ISSN 1662-5145. doi: 10.3389/fnint.2020.00042. URL https://www.frontiersin.org/article/10.3389/fnint.2020.00042.

Itay Gat and Naftali Tishby. Synergy and redundancy among brain cells of behaving monkeys. In *Advances in Neural Information Processing Systems*, pages 111–117, 1999.

Albert Gidon, Timothy Adam Zolnik, Pawel Fidzinski, Felix Bolduan, Athanasia Papoutsi, Panayiota Poirazi, Martin Holtkamp, Imre Vida, and Matthew Evan Larkum. Dendritic action potentials and computation in human layer 2/3 cortical neurons. *Science*, 367(6473):83–87, 2020. ISSN 0036-8075. doi: 10.1126/science.aax6239. URL https://science.sciencemag.org/content/367/6473/83.

Virgil Griffith and Christof Koch. *Quantifying Synergistic Mutual Information*, pages 159–190. Springer Berlin Heidelberg, 2014. doi: 10.1007/978-3-642-53734-9_6.

Malte Harder, Christoph Salge, and Daniel Polani. Bivariate measure of redundant information. *Phys. Rev. E*, 87:012130, Jan 2013. doi: 10.1103/PhysRevE.87.012130. URL https://link.aps.org/doi/10.1103/PhysRevE.87.012130.

Joseph T Lizier, Nils Bertschinger, Jürgen Jost, and Michael Wibral. Information decomposition of target effects from multi-source interactions: Perspectives on previous, current and future work. *Entropy*, 20(4): 307, 2018.

Giuseppe Pica, Eugenio Piasini, Houman Safaai, Caroline Runyan, Christopher Harvey, Mathew Diamond, Christoph Kayser, Tommaso Fellin, and Stefano Panzeri. Quantifying how much sensory information in a neural code is relevant for behavior. In *Advances in Neural Information Processing Systems*, pages 3686–3696, 2017.

Elad Schneidman, William Bialek, and Michael J Berry. Synergy, redundancy, and independence in population codes. *Journal of Neuroscience*, 23(37):11539–11553, 2003.

Sameet Sreenivasan and Ila Fiete. Grid cells generate an analog error-correcting code for singularly precise neural computation. *Nature neuroscience*, 14(10):1330, 2011.

Nicholas M Timme and Christopher Lapish. A tutorial for information theory in neuroscience. *eNeuro*, 5 (3), 2018.

Praveen Venkatesh, Sanghamitra Dutta, and Pulkit Grover. How else can we define information flow in neural circuits? In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2879–2884. IEEE, 2020a.

Praveen Venkatesh, Sanghamitra Dutta, and Pulkit Grover. Information flow in computational systems. *IEEE Transactions on Information Theory*, 66(9):5456–5491, September 2020b. URL https://doi.org/10.1109/TIT.2020.2987806.

Erich von Holst and Horst Mittelstaedt. The principle of reafference: Interactions between the central nervous system and the peripheral organs. *Perceptual processing: Stimulus equivalence and pattern recognition*, pages 41–72, 1971.

Xue-Xin Wei, Jason Prentice, and Vijay Balasubramanian. A principle of economy predicts the functional architecture of grid cells. *Elife*, 4:e08362, 2015.

Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv:1004.2515 [cs.IT]*, 2010.

Yilun Zhang and Tatyana O Sharpee. A robust feedforward model of the olfactory system. *PLoS computational biology*, 12(4):e1004850, 2016.