Evaluating the integrated information theory of consciousness in flies

Angus Chun-Kei Leung

BCompSc, BSc(Hons)

A thesis submitted for the degree of Doctor of Philosophy at
Monash University in 2021
School of Psychological Sciences and
Turner Institute for Brain and Mental Health

Copyright notice

Abstract

How physical interactions generate and support conscious experience is a key question in neuroscience. The integrated information theory of consciousness (IIT) attempts to address this question by taking a first-principles approach, introspecting our own conscious experience in order to identify its core aspects. Then, physical interactions must implement these aspects in order to be considered a physical substrate of consciousness. Following this approach, IIT identifies 5 core aspects of conscious experience. From these core aspects, the theory postulates the necessary physical interactions required to support them. Finally, the theory uses its postulates to derive measures which capture the extent to which a candidate system implements the core aspects of experience. Integrated information (Φ) captures the extent to which a system can be considered greater than the sum of its parts, while a corresponding cause-effect structure captures the information which is available intrinsically from the perspective of the system. Φ and the cause-effect structure are posited to correspond to level and contents of conscious experience, respectively, with Φ emerging at some spatiotemporal scale corresponding to the scale of conscious experience.

In this thesis, I work towards evaluating the validity of these measures as proposed by the IIT. As their computational complexity makes it infeasible to apply to complex systems such as the human brain, I apply them instead to recordings from relatively simple fly brains during wakefulness and anaesthesia. I test two predictions of IIT: that 1) its measures should be high during consciousness and low or zero during loss of consciousness, and 2) Φ should emerge to be maximal at some macro timescale, instead of at a micro timescale as would be expected from reductionism. Consistent with IIT's expectations, I find Φ and its associated cause-effect structure to be reduced during anaesthesia. I also find that the ratio of Φ during wakefulness to Φ during anaesthesia is maximal at a timescale of roughly 10 ms, which corresponds roughly to the timescale of neuronal interactions. Lastly, I work towards evaluating the advantage of IIT's approach in deriving measures of consciousness from first principles over discovering measures using a data-driven approach. To this end, I systematically apply a vast library of univariate time-series features to the fly recordings and evaluate their performances in discriminating wakefulness from anaesthesia. I find that relatively simple measures relating to signal variance and stationarity best discriminate levels of consciousness, with IIT's measures performing comparably or even slightly better. Together, these results demonstrate the utility of testing the expensive measures of IIT in simple biological systems, and provide early empirical support for IIT's first-principles, theoretical approach towards identifying the physical substrate of consciousness.

Thesis including published works declaration

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes 1 original paper published in a peer reviewed journal and 1 submitted publications. The core theme of the thesis is evaluating the empirical validity of the integrated information theory of consciousness. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the student, working within the School of Psychological Sciences and Turner Institute for Brain and Mental Health under the supervision of Naotsugu Tsuchiya.

(The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.)

In the case of Chapters 2 and 4, my contribution to the work involved the following:

Thesis Chapter	Publication Title	Status	Nature and % of student contribution	Co-author name(s) Nature and % of Co-author's contribution	Co-author(s), Monash student Y/N
2	Integrated information structure collapses with anesthetic loss of conscious arousal in Drosophila melanogaster	Published	80%. Conceptualisation, data analysis, writing	1) Dror Cohen, data provision, input into manuscript 5% 2) Bruno van Swinderen, data provision, input into manuscript 5% 3) Naotsugu Tsuchiya, conceptualisation, manuscript review and editing 10%	No
3	Emergence of integrated information at macro timescales in	Accepted for publication	90% Conceptualisation, data analysis, writing	1) Naotsugu Tsuchiya, conceptualisation, manuscript review and editing 10%	No

	real neural recordings				
4	Towards blinded classification of levels of consciousness: distinguishing wakefulness from general anesthesia in flies using a massive library of univariate time series analyses	Returned for revision	60%. Conceptualisation, data analysis, writing	1) Ahmed Mahmoud, initial exploratory analyses, writing initial draft 20% 2) Rhiannon Jeans, data provision, input to manuscript 5% 3) Ben Fulcher, input to manuscript 5% 4) Bruno van Swinderen, data provision, input to manuscript 3% 5) Naotsugu Tsuchiya, conceptualisation, input to manuscript 7%	No

I have not renumbered sections of submitted or published papers.

Student name: Angus Leung

Student signature: **Date**: 27/Apr/2022

I hereby certify that the above declaration correctly reflects the nature and extent of the student's and co-authors' contributions to this work. In instances where I am not the responsible author I have consulted with the responsible author to agree on the respective contributions of the authors.

Main Supervisor name: Naotsugu Tsuchiya

Main Supervisor signature: **Date**: 27/Apr/2022

Acknowledgements

I would like to acknowledge my supervisor, without whom this work would not have been possible. Thank you for your teaching and guidance, and for your constant inspiration that we can and should study consciousness in science! I would also like to give my thanks to my family, for their support, and to all the people – teachers, friends, colleagues, acquaintances, and so forth – who have shaped me into the person I am thus far. Maybe this thesis would not exist if I had interacted with different people… Lastly, I wish to acknowledge you, the reader/reviewer. At the end of the day, this thesis is only worth what you think of it!

This research was supported by an Australian Government Research Training Program (RTP) Scholarship.

Table of contents

Chapter 1 - Introduction

Consciousness is at the core of the universe. At the very least, it is the core of your universe, or your perspective of the universe - without conscious experience, the universe could not, and would not, exist to you. *You* wouldn't exist to you. It is only through your conscious experience that you can experience yourself, and a world around you. To lose your conscious experience would be to lose everything - nothing, including yourself, would exist. Despite the clear significance of conscious experience in keeping the universe extant, at least for ourselves, we struggle to understand just how it can be that systems like ourselves can even have conscious experiences.

Even though consciousness is a private, subjective experience, we generally extend the notion of consciousness to other beings - other people, and even other animals. In general day-to-day interaction, seemingly purposeful behaviours seem sufficient for most people to ascribe consciousness to some entity. When unsure, we might stimulate the being and use its reaction to judge how conscious it is.

Determining if a being is conscious is, however, often not straightforward. For example, conscious experience is possible without overt behavioural responsiveness, such as when dreaming and even under anaesthesia (Liu et al., 1991; Sebel et al., 2004). Meanwhile, physiologic responses to stress during consciousness under anaesthesia, such as increased heart rate, are often masked by accompanying drugs (Rani & Harsoor, 2012). From these cases, it becomes clear that neither the generally used behavioural (Guedel, 1937) or physiological (Rani & Harsoor, 2012) "signs" of consciousness are truly indicative of consciousness. Thus, a key goal in neuroscientific research is to identify the necessary conditions for consciousness - to find how consciousness arises from neural activity in the brain.

The scientific study of phenomenal consciousness may initially seem impossible. After all, consciousness deals in subjective experiences - the very thing which science tries to avoid. Indeed, early neuroscience, while acknowledging the question of consciousness (LeDoux et al., 2020) largely focussed on functional relationships between behaviour and neural activities. Thankfully, we have a starting point from which to begin some objective investigation into consciousness. That is, we all agree that individually, we are conscious beings, i.e. that we all have some kind of conscious experience. Specifically, you know that you are conscious, because you are having some conscious experience. This basic idea of consciousness as phenomenal experience is distinct from other phenomena which might also be referred to as “consciousness”, such as self-awareness (Morin, 2006), high-order thoughts (Edelman, 2003), or reportable access to one’s own experiences (Block, 1995). Though we may not always think about it, we all have some idea of what this basic idea of consciousness is - we all have experiences (usually of the world around us) which vanish when we go to sleep or are put under general anesthesia (and presumably when we die), and re-emerge when we wake up or dream. Indeed, this was recognised by (Crick & Koch, 1990), who put forward a framework to begin studying the neural basis of consciousness which invigorated modern consciousness research.

1.1 - Previous approach to understanding consciousness

The modern neuroscientific study of consciousness currently largely follows the overall framework proposed by (Crick & Koch, 1990), the search for “neural correlates of consciousness” (NCC). Their framework was to put aside the philosophical questions as to what consciousness is precisely, what function it serves, and so on. Instead, the framework first recognised that we all have conscious experience, regardless of what it is exactly. Then, having recognised the existence of consciousness, the framework proposed that we could work towards understanding what neural activities might generate consciousness by identifying the minimum neuronal mechanisms jointly sufficient for any one specific conscious percept (i.e. the NCC for that conscious percept; (Koch, 2004)). After identifying the NCC, we would then be in a better position to address the more difficult, traditionally philosophical questions surrounding consciousness.

Following this framework, researchers have worked towards identifying two main interpretations of the NCC, content-specific NCC and the full NCC (Koch et al., 2016). Content-specific NCC refers to the neuronal mechanisms underlying a particular phenomenal distinction within one’s conscious experience (for example, the experience of seeing a face on a computer screen). If identified, activating the content-specific NCC in a person’s brain should trigger the perception of that phenomenal distinction. Conversely, preventing the content-specific NCC from activating should prevent the person from experiencing the distinction when they otherwise would. The full NCC, on the other hand, refers to the neural activities which support consciousness in its entirety, regardless of what is being consciously experienced and so can be understood as the union of all content-specific NCC for all possible experiences.

The general strategy used for identifying neural activities which may be part of either type of NCC is to contrast neural activity when a certain percept is present or absent (i.e. perceived or not), for content-specific NCC (Breitmeyer & Ogmen, 2000; Blake & Logothetis, 2002; Tsuchiya & Koch, 2005; Imamoglu et al., 2012), or when consciousness is present or absent (wakefulness versus e.g. general anaesthesia, dreamless sleep, or coma), for the full NCC (Alkire et al., 2008; Maquet et al., 1997; Gosseries et al., 2014). This strategy has led to better understanding of the kinds of neural interactions which occur during specific conscious percepts, such as activation of the fusiform face area during face perception (Kanwisher & Yovel, 2006) and at different levels of consciousness. Accordingly, researchers have collated and assessed empirically reported neural interactions, and put forward theories which propose what interactions could constitute content-specific or the full NCC. Well-known theories include synchrony theory (Crick & Koch, 1990; Engel & Singer, 2001), recurrent processing theory (Lamme, 2006, 2010), predictive coding theory (Hohwy, 2013), and global neuronal workspace theory (Dehaene et al., 2011; Mashour et al., 2020). Undoubtedly, there has been great progress in discovering neural activities which are potentially related to consciousness. However, this approach towards understanding consciousness has significant drawbacks.

The first drawback is that many neural activities which at first seem promising as potential NCC are later found to not serve as reliable indicators of conscious experience. For example, synchronous activity and feedback interactions, similar to that put forward as potential NCCs by synchrony theory and recurrent processing theory, can be found in the cerebellum (Person & Raman, 2012; Witter et al., 2016), which is unlikely to contribute to consciousness (Yu et al., 2015), while the fusiform face area can be activated during perception of non-face stimuli (Gauthier et al., 1999, 2000).

Related is the potential lack of generalisability of proposed NCCs to systems other than those in which the associated neural activities were first observed. For example, the exact brain region activated during perception of human faces varies among humans, non-human primates, sheep, and dogs (Cuaya et al., 2016). The issue is more obvious when considering animals with very different brain architectures to humans, such as the octopus and fruit fly, both of which exhibit behaviours which might indicate the capacity for conscious experience (Medeiros et al., 2021; Tainton-Heap et al., 2021). Going even further, it is clear that biological NCCs cannot be used to assess whether an artificial system, such as a personal computer, is conscious.

A second drawback is the difficulty of disentangling whether some neural activity is varying with consciousness, or varying with other processes. For example, regarding content-specific NCC, process such as selective attention, expectation, self-monitoring, unconscious stimulus processing, planning, and reporting all occur with, or closely preceding or following conscious experience of some percept (Aru et al., 2012; de Graaf et al., 2012; Koivisto & Revonsuo, 2010; Miller, 2014), especially so in the context of being a participant in a study. Meanwhile, regarding the full NCC, neural activities might reflect changes in vigilance or attention rather than changes in conscious level itself (Hohwy, 2009). Consequently, it can be difficult to associate any neural activity with consciousness only, and not some other process.

Overall, though the framework of searching for NCCs can be useful to help identify the kinds of neural activities associated with consciousness, it is ultimately limited in that it does not result in any explanation as to why any particular proposed NCC should generate consciousness. This explanatory gap between physical, neural interactions and consciousness itself, known as the hard problem of consciousness (Chalmers, 1995), highlights the need for a principled theory of consciousness.

1.2 - Approaching consciousness from the intrinsic perspective

1.2.1 - Integrated information theory

The integrated information theory of consciousness (IIT; (Tononi, 2004, 2008; Oizumi et al., 2014)) differentiates itself from other theories of consciousness by taking a radically different approach towards finding the physical substrate of consciousness. Rather than constructing a theory based on observed neural activities, IIT introspects conscious experience in order to identify its fundamental properties. It then deduces the necessary physical mechanisms required to support these properties.

The fundamental properties of consciousness IIT identifies are as follows:

a) Intrinsic existence: an experience exists intrinsically to a conscious system, but not to external observers. Consequently, observing a conscious system having some experience does not give you the same experience.

b) Composition: an experience is composed of multiple aspects. For example, the experience of watching a movie is composed of visual and auditory aspects, and the experience of a face is composed of eyes, a nose, a mouth, etc.).

c) Information: an experience is specific. That is, it differs from every other possible experience that a conscious system could instead be having. For example, in reading this thesis you are consequently not experiencing all the other experiences you could possibly be having instead, such as watching a movie, cooking dinner, or reading another piece of writing).

d) Integration: an experience exists as a single whole which cannot be broken up into independent parts. For example, the experience of a red ball is integrated as one experience, rather than being two separate, independent experiences of redness, and a ball).

e) Exclusion: an experience is definite. That is, we do not have multiple separate experiences superposed among each other. Separate experiences, then, must be of separate conscious systems. This precludes any single system from having multiple conscious experiences, such as at different timescales.

From these fundamental properties of consciousness, IIT postulates a set of physical properties which are required to support consciousness (the full derivation is described in (Oizumi et al., 2014). Then, from these physical properties, IIT derives two key measures: integrated information Φ, whose magnitude is purported to reflect the level of consciousness in a system, and an associated conceptual structure, which is purported to reflect the system's experiential contents. This approach taken by IIT overcomes the previous limitations around the search for the NCC. Specifically, as its measures are derived from fundamental principles, instead of being discovered from observed neural activity, their applicability extends beyond humans, vertebrates, and even biological systems. In the next section, I aim to provide a concise overview of how IIT 3.0 (Oizumi et al., 2014) uses its axioms to arrive at its measures for consciousness in any given system.

1.2.2 - Formalisation of integrated information theory 3.0

From the fundamental properties of conscious experience posited by the theory, IIT 3.0 puts forth postulates - translations of the properties into the physical interactions required for a substrate to support them. Next, IIT 3.0 uses these postulates to derive measures which are purported to reflect the level and contents of consciousness in a given set of elements, (i.e. a system). As the full formalisation is given in (Oizumi et al., 2014), in this section I aim to provide a concise overview of the postulates and the key theoretical concepts behind how they come together to define a measure of consciousness. I give full details of the associated computations in Chapter 2 (specifically, I refer the reader to Figure 1, which illustrates all the concepts which I outline here).

Intrinsic existence

The intrinsic existence axiom states that an experience exists intrinsically to a conscious system. From this, IIT postulates that a system, and its constituent mechanisms, must have causal power on itself in order to be conscious. Physically, how can one observe the existence of anything? For example, why can this thesis exist to some reader? For this text to exist to a reader, the reader must be able to observe it. Physically, a reader can observe this thesis as photons (and lack thereof) travel from the text into their eyes. These photons cause visual receptors to fire, which then cause further activity through the visual stream and presumably across the brain. Conversely, if this text elicited no such change to the brain, how could a reader possibly observe it, and how could it exist to them? Hence, IIT reasons that as a conscious experience exists intrinsically to a conscious observer, the observer must be causing some change to itself. Hence, the constituent parts of the system, or *mechanisms*, must specify causes and effects (i.e. have "cause-effect power") within the system.

Composition

The composition axiom states that an experience is composed of multiple aspects. Accordingly, IIT postulates that mechanisms should combine to compose higher order mechanisms. Hence, a system of 3 elements {A, B, C} would have 7 candidate mechanisms {A, B, C, AB, AC, BC, ABC}, with AB being a "second order" mechanism composed of A and B, and ABC being a "third order" mechanism composed of A, B, and C, etc. Each candidate mechanism might or might not exist intrinsically to the system.

Postulates for the remaining axioms are then proposed at two levels: at the level of mechanisms, and at the level of a whole candidate system. At the level of mechanisms, the postulates aim to characterise how any part of a given system, i.e. mechanism, exists intrinsically to it. Meanwhile, the postulates applied at the level of a candidate system aim to determine whether it constitutes a conscious system, and to characterise its experience. For clarity, I introduce the terminology, "purview", here. A purview is any subset of the system. For example, a system of 3 elements {A, B, C} would have 7 possible purviews, {A, B, C, AB, AC, BC, ABC}.

Postulates at the level of mechanisms

Information

The information axiom states that an experience differs from every other possible experience that the conscious system could instead be having. In conjunction with the intrinsic existence postulate, which deals with the issue of to what or whom something exists, the information postulate deals with the nature of what it is that is existing. While the existence postulate deals in whether something exists intrinsically to a system, the information postulate deals in the nature of what that something is. Specifically, a candidate mechanism within a system exists in some way to the system if it specifies its causes and effects in the system. Meanwhile, how it specifies its causes and effects determines the information that it provides to the system.

This postulate is formalised in conjunction with the intrinsic existence postulate using mathematics similar to, yet also distinct from, that used in standard information theory (Shannon, 1948). Given a mechanism, its state specifies a cause and thus generates information about the past regarding the state of some subset of the system (i.e. "purview", which can be any subset of the system, including the mechanism itself) if it constrains the possible past states of that purview. How informative the state is depends on the extent of the constraints. If there are no constraints, and thus no information about the possible past states of the purview, then all possible states are equally probable. Conversely, the largest constraint is to specify that only one state could have possibly occurred in the past, leading to the mechanism's current state. Following this, the information that a mechanism gives about a purview's past, cause information (*cause_info*), is given by the distance between probability distributions:

(1) $cause\_info = D(\ P(purview_{past}), P(purview_{past} \mid mechanism)\ )$

Where $P(purview_{past})$ is the probability distribution of the purview's possible past states when no constraint is specified by the mechanism (i.e. when the purview's previous state is not constrained by the mechanism's current state; this is the distribution of all states being equally likely, i.e. the maximum entropy distribution), and $P(purview_{past} \mid mechanism)$ is the distribution of past purview states given the mechanism's current state. The earth mover's distance (EMD; (Rubner et al., 2000)) is used to quantify the distance between distributions. If the distance between the distributions is 0, (i.e. the mechanism's state does not constrain the possible pasts of the purview), it provides no information about the purview's past and does not specify a cause.

This same reasoning is applied also to the mechanism's effect. Whereas information about a purview's past is quantified by how its possible past states are constrained by the mechanism's current state, information about the purview's future, is quantified by how its future possible states are constrained by the mechanism's current state. So, the effect information (*effect_info*) specified by the mechanism is quantified as the distance between two other probability distributions:

(2) $effect\_info = D(\ P(purview_{future}), P(purview_{future} \mid mechanism)\ )$

Where $P(purview_{future})$ is the unconstrained distribution of possible future states of the purview, and $P(purview_{future} \mid mechanism)$ is the distribution of future purview states given the mechanism's current state. $P(purview_{future})$ is found by first perturbing the system into all possible states with equal likelihood, and for each of these starting states, obtaining the probability distributions of possible future states. Then, these distributions are marginalised across the starting states to obtain the distribution of possible future states given that nothing is known about the purview's present state. Similarly as for the mechanism's cause, if the distance between the distributions is 0, the mechanism specifies no information about the purview's future and does not specify an effect.

$P(purview_{past} \mid mechanism)$, $P(purview_{future} \mid mechanism)$, and $P(purview_{future})$ are all obtained by perturbing either the mechanism or the system into all possible states, and observing the resulting state for a given purview. In practice, however, they can be obtained by perturbing the system into all its possible states, and observing the probabilities of each system state transitioning into any other system state (“transition probabilities”). These system state transition probabilities can then be marginalised down to a specific mechanism and purview.

As stated earlier, a mechanism exists to the system if it specifies both its cause and effect. Consequently, if, across all purviews, it specifies either no cause information or no effect information (or both), it does not exist from the perspective of the system. If it does specify both, then $P(purview_{past} \mid mechanism)$ and $P(purview_{future} \mid mechanism)$ specify the nature of the information given by the mechanism - the “cause repertoire” and the “effect repertoire” respectively. In conjunction with each other, they specify the “cause-effect repertoire”.

Integration

The integration axiom states that an experience exists as a single whole which cannot be broken up into independent parts. While the composition postulate deals with how mechanisms can compose to form larger mechanisms, the integration postulate deals with whether these larger (i.e. higher order) mechanisms contribute anything above and beyond what their constituent mechanisms contribute to the system. Specifically, it postulates that mechanisms only exist if they specify information that is not given by (i.e. irreducible to) its independent components. In conjunction with the information postulate, a mechanism exists above and beyond its constituent parts if it specifies causes and effects which are irreducible to the causes and effects specified by its constituent parts.

Similar to how the information postulate is formalised comparing two distributions, the integration postulate is formalised again by comparing two distributions. This time, however, the distributions being compared are the mechanism’s cause or effect repertoire with the product of the cause or effect repertoires of the mechanism’s constituent parts. Integrated information φ then is the information which the mechanism has regarding some purview, above and beyond its constituent parts. This is quantified by partitioning the mechanism and purview such that each part of the mechanism can only influence one part of the purview:

$$(3)\ \varphi_{cause} = D(\ P(purview_{past} \mid mechanism),\\ P(purview^{A}_{past} \mid mechanism^{A}) \times P(purview^{B}_{past} \mid mechanism^{B})\ )$$

$$(4)\ \varphi_{effect} = D(\ P(purview_{future} \mid mechanism), \\ P(purview^{A}_{future} \mid mechanism^{A}) \times P(purview^{B}_{future} \mid mechanism^{B})\ )$$

Where φ_{cause} and φ_{effect} is the information which the mechanism has regarding some purview's possible past and future states respectively, above and beyond its independent parts. $P(purview^{A}_{past} \mid mechanism^{A})$ and $P(purview^{B}_{past} \mid mechanism^{B})$ are the probability distributions of the possible past states of each part of the purview, conditioned on the current states of each part of the mechanism (and likewise for $P(purview^{A}_{future} \mid mechanism^{A})$ and $P(purview^{B}_{future} \mid mechanism^{B})$, but for possible future states of each part of the purview). To assess if the mechanism has information above and beyond its constituents, all possible partitions need to be assessed. If any partition specifies the same causes or effects as the full mechanism, the full mechanism does not contribute anything further than its parts. Consequently, the partitions which result in repertoires which most closely approximate those of the full mechanism are used to quantify φ_{cause} and φ_{effect}. These partitions are referred to as minimum information partitions (MIPs). As specified in the information postulate, a mechanism must specify both causes and effects in order to exist to the system. Hence, if either φ_{cause} or φ_{effect} are 0 for all candidate cause purviews and all candidate effect purviews, it does not exist to the system. φ of the mechanism overall then is the minimum of φ_{cause} and φ_{effect}.

Exclusion

The exclusion axiom states that an experience cannot be superposed with other experiences. While the previous postulates have dealt with the information that a mechanism provides to the system, the exclusion postulate deals with the recipient of that information - what part of the system the mechanism specifies its cause and effect for. Specifically, it postulates that each mechanism only contributes at most one cause repertoire and one effect repertoire.

The exclusion postulate is formalised by, for a given mechanism, selecting the cause purview and its corresponding repertoire for which φ_{cause} is maximal (the "core cause"), and selecting the effect purview and its corresponding repertoire again for which φ_{effect} is maximal (the "core effect"). Taken together, they are the mechanism's maximally irreducible cause-effect repertoire (MICE). In conjunction with the previous postulates, a mechanism exists to the system if both its core cause and core effect have greater than 0 φ. If so, it constitutes a "concept".

Together, these postulates are applied to all mechanisms (i.e. subsets of elements) in the system. The postulates are then applied at the level of systems to quantify how the mechanisms come together to form a single consciousness.

Postulates at the level of systems

Information

At the mechanism level, the information postulate formalises what information, or causal constraints, a mechanism provides to the system. At the system level, the theory postulates that the set of mechanisms which exist (i.e., concepts) specify a "conceptual structure" (also referred to as a "cause-effect structure"; (Mayner et al., 2018)), which captures all the information that the system intrinsically has and distinguishes it from all other possible conceptual structures.

The postulate formally describes the conceptual structure as a "constellation" in concept space, a space with one axis for each possible future state and each past state of the system. In this space, the conceptual structure exists as a set of points, with each point corresponding to each concept in the system. For a particular concept, its location along each axis describes its core cause and effect (i.e. its coordinates describe its core cause and core effect repertoires; consequently each axis ranges from 0 to 1). A point in this space corresponds to a "null concept", that is, a mechanism which specifies no causes and effects and thus no information. This null concept has coordinates specifying the unconstrained distributions for possible past states and possible future states. The sum of distances of each concept from the null concept, each weighted by their φ, summarises the amount of information in the conceptual structure ("conceptual information"; *CI*).

$$(5)\quad CI = \sum \left(\varphi \times D(MICE,\ null) \right)$$

Where *D*(*MICE*, *null*) is the distance between the MICE specified by a concept and the unconstrained distributions specified by the null concept.

Integration

At the mechanism level, the integration postulate deals with whether higher order mechanisms exist above and beyond their constituent parts. At the system level, it deals with whether a system exists above and beyond its constituent parts. Specifically, the theory postulates that a set of elements can only be conscious if its mechanisms specify a conceptual structure which cannot be reduced to independent components.

This postulate is formalised by partitioning the system into smaller systems through noising connections such that the parts are connected at most in a unidirectional manner. Then, the conceptual structure of the whole system is compared to that of the partitioned system. The unidirectional partitioning ensures that, in the partitioned system, the mechanisms from one part can only specify either causes or effects to purviews in another part, but not both. The comparison is quantified as integrated information, Φ, in a similar way as for mechanisms - slightly simplified, as a sum of distances between the MICE of concepts in each conceptual structure:

$$(6)\ \Phi = \sum (\ \varphi \times D(MICE_{unpart}, MICE_{part})\ )$$

Where $D(MICE_{unpart}, MICE_{part})$ is the distance between the MICE specified by a given concept before and after partitioning the system.

Again, similar to the integration postulate for mechanisms, all possible partitions need to be assessed, including partitions which differ only by the direction by which parts are unidirectionally connected. The partition which gives the conceptual structure most similar to the unpartitioned system, i.e. gives the smallest Φ, is used to determine how integrated the system and its conceptual structure is. This partition is also referred to as the minimum information partition (MIP) as introduced at the mechanism level.

Exclusion

At the system level, the exclusion postulate deals with identifying the substrate of the conscious system. Specifically, it postulates that a conscious system is the set of elements which specifies a conceptual structure that is maximally irreducible - a maximally irreducible conceptual structure (MICS).

To find the MICS for a given candidate set of elements, all subsets, across all spatial (e.g. molecules vs neurons vs cortical regions) and temporal scales (e.g. microseconds versus milliseconds versus seconds), should be treated as candidate systems, and their conceptual structures assessed for irreducibility. The candidate system with the most irreducible conceptual structure, i.e. gives the greatest Φ (i.e. Φ^{max}) across elements, spatial scales, and temporal scales, is the system which forms the substrate of consciousness, and is referred to as a "complex".

1.2.3 - Integrated information as a measure of consciousness

In using all the postulates, IIT 3.0 arrives at two key measures of consciousness. Integrated information Φ is put forward as a measure of level of consciousness in a given system (which has some Φ^{max}), which can be used to search for the complex (i.e. the substrate of consciousness). Meanwhile, the complex's maximally irreducible conceptual structure, MICS, is purported to be identical to the conscious experience that a system has, or the contents of consciousness. However, the theory notes that its formulation is still incomplete with regards to how the MICS gives specific aspects of phenomenology, such as how different sensory modalities feel their own specific way.

In this section I have laid out the general ideas and framework of IIT, and how it is formalised in IIT 3.0. From its formalisation, it provides measures of consciousness, Φ and the associated MICS, to reflect the level and contents of consciousness in any given system. In the following sections, I will review core issues regarding the applicability of IIT to real data, and how it has actually been applied in the existing literature to test its ideas.

1.3 - Practical issues of integrated information theory

As laid out in the previous section, integrated information theory starts from first principles, identifying core aspects of consciousness and deriving postulates about the kinds of interactions needed to support it. These postulates arrive at clearly defined measures for consciousness which in principle can be applied to any system. Consequently, it serves as a potentially easy theory to falsify. Specifically, its key prediction that systems should have greater Φ when conscious than when not makes for an easy target to test. However, applying the theory to neural data can be difficult due to several issues.

Firstly, the theory's assessment of information requires the transition probabilities between every pair of system states. Obtaining this requires perturbing the system into all its possible states and observing the resulting state transitions. However, the number of system states grows exponentially with the number of elements. For example, while systems of 2 and 3 binary elements have 4 and 8 possible states respectively, a network of 5 binary elements has 32 possible states. Meanwhile, a system such as the human brain, with roughly 86 billion neurons (Azevedo et al., 2009) would have a staggering number of possible states (specifically, 2 to the power of 86 billion, when treating each neuron as a binary element). Consequently, empirically obtaining transition probabilities quickly becomes infeasible when considering larger and larger systems. Furthermore, the calculation of information uses assumptions which are only met for discrete elements (specifically, the use of the maximum entropy distribution when calculating the cause information of a mechanism; (Oizumi et al., 2016)), but not continuous elements. However, neural recordings often collate across populations of neurons, such as electroencephalographic recordings (EEG) and local field potentials (LFP), giving data of continuous-valued observables.

Secondly, the theory's assessment of integration across system elements in forming a single whole above and beyond its parts depends on identifying minimum information partitions (MIPs). However, when there is no prior knowledge of a system's causal connections (such as synaptic connections between neurons), identifying MIPs requires an exhaustive search across every possible partition of a given mechanism or system. Similar to the issue regarding the number of system states, the number of possible partitions for a group of elements grows, super-exponentially, with the number of elements. For example, at the system level, a system consisting of 4 elements can be partitioned in 15 ways, while a system consisting of 5 elements can be partitioned 52 ways (Aitken, 1933; Bell, 1934). Consequently, calculating Φ for a system of even 1000 elements is already estimated to take several magnitudes longer than the age of the universe (Toker & Sommer, 2019). Though toolboxes implementing IIT 3.0 reduce this problem somewhat by searching through bipartitions rather than all partitioning schemes, the search for the MIP (or, in this case, the minimum information bipartition, MIB; (Mayner et al., 2018)), still grows super-exponentially with system size.

Thirdly, in addressing the axiom of exclusion, the theory requires that, for Φ and a conceptual structure to be related to consciousness, it must be computed for a complex. The search for the complex, however, requires repeatedly computing conceptual structures and their associated Φ for all candidate systems (i.e. the power set) from a set of elements, at all possible spatial grainings and temporal resolutions. Each of these issues individually serve as significant hurdles for the application of IIT 3.0 proper, let alone all of them combined. Consequently, the application of IIT 3.0 has been limited to simulation studies involving small systems where analytical methods can be used to cheaply determine Φ. Meanwhile, for the application of IIT to real neural data, approximations and alternative derivations of Φ have been derived based on previous versions of the theory.

1.4 – Existing applications of integrated information theory

1.4.1 - Simulations

Applications of IIT 3.0 thus far have been mostly limited to simulation studies. Indeed, apart from applications of IIT 3.0 to characterise group interactions between humans (Shehata et al., 2021) and among cell-cycle states (Marshall et al., 2017), and leadership in fish (Niizato et al., 2020a, 2020b), all studies applying IIT 3.0 so far have been simulation studies. As simulation studies, they seek to characterise the behaviour of Φ and the MICS given certain system constraints and architectures, rather than directly dealing with the question of their empirical validity as measures of consciousness. For example, IIT 3.0 has been used in the context of answering why animals might evolve to have consciousness - given constraints on system resources, more complex behaviours can arise when connectivity among system elements allow for high Φ (Albantakis et al., 2014; Albantakis & Tononi, 2015). Other studies have utilised the exclusion postulate of IIT 3.0 to illustrate system architectures where causal power can emerge at macro spatial and temporal scales, rather than being causally complete at the most micro scale (Hoel et al., 2016; Marshall et al., 2018; Grasso et al., 2021). Though studies such as these can illustrate the utility of IIT 3.0 in the contexts of characterising efficiency or identifying ideal scales at which to describe physical interactions, applications which empirically test the measures of IIT 3.0 in real neural systems are ultimately lacking.

1.4.2 - Theoretical derivations

Given the difficulties in directly applying IIT to real data, alternative derivations of Φ have been proposed. For example, Φ_E (Barrett & Seth, 2011) was introduced to deal with the issues of using maximum entropy distributions in continuous elements, and its approach was extended in Φ* to better satisfy theoretical bounds of Φ (Oizumi et al., 2016). Reviews and comparisons of more well-known alternatives and comparisons of alternative versions of Φ which are "improved" to account for different assumptions underlying system elements can be found in (Tegmark, 2016; Mediano et al., 2019; Sevenius Nilsen et al., 2019).

Such measures have been applied to neural data, ignoring the issue of identifying the complex, and have given results largely consistent with IIT's main prediction of reduced system level integrated information with reduced level of consciousness. Findings include reduced integrated information in human EEG recordings with reduced behavioural responsiveness due to increased sevoflurane dosage (Kim & Lee, 2019), during general anaesthesia due to ketamine and propofol (Lee et al., 2008; Kim et al., 2018). Integrated information has also been reported to be reduced with deep sleep both in intracranial recordings from epilepsy patients (Chang et al., 2012) and EEG from infants (Isler et al., 2018).

These alternative integrated information measures have also been applied in the context of evaluating contents of consciousness. System level integrated information values from functional magnetic resonance imaging (fMRI) have been reported to increase with more meaningful stimuli (Boly et al., 2015), though IIT makes no prediction regarding how system level integrated information is related to the contents of consciousness. More closely applying the idea of the MICS, patterns of integrated information, obtained from calculating integrated information for all subsets of groups of electrocorticographic (ECoG) recordings have been reported to be able to distinguish whether a stimulus was perceived or not (A. M. Haun et al., 2017). The use of an alternative integrated information measure in fMRI has also resulted in a finding that the brain may functionally split into separate systems when concurrently performing unrelated tasks such as driving while listening to a radio show (Sasai et al., 2016).

Though alternative derivations of integrated information have allowed for applications and tests of IIT, they ultimately suffer from a key issue. Currently existing derivations are all variations of integrated information as proposed in previous versions of IIT (Tononi, 2004, 2008), which considered a system's ability to specify only its effects (instead of both causes and effects as postulated in IIT 3.0). Consequently, it is unclear whether findings reported using these measures would hold when using Φ as directly proposed in IIT 3.0. Further, they largely focus on providing an index of conscious level while ignoring and failing to provide a framework to account for experiential contents (relating to composition in IIT 3.0), and so it is unclear whether and how they can be extended to do so.

1.4.3 - IIT inspired practical measures

Another approach towards IIT has been to apply or construct completely new measures based on the core ideas of IIT. Along this approach, the perturbational complexity index (PCI; (Casali et al., 2013)) tries to capture IIT's key ideas of assessing information and integration through perturbing the system. PCI employs transcranial magnetic stimulation to perturb a part of the brain and analyses the resulting EEG spatiotemporal activation pattern. It characterises integration across brain regions by analysing the spatial extent of the resulting patterns (Massimini et al., 2005; Ferrarelli et al., 2010), and information as the extent to which the responses are differentiated among regions. It simultaneously characterises integration and information using Lempel-Ziv complexity (LZc), a relatively computationally cheap measure of how easily an image can be compressed. Activations patterns which are spatially restricted, reflecting lack of integration among regions are easily compressed by the Lempel-Ziv algorithm, as are patterns which spatially uniform which reflect lack of differentiation. Hence, activation patterns lacking either integration or differentiation give low LZc values.

The approach taken by PCI has seen success in applications of differentiating wakeful subjects and subjects in rapid eye movement (REM) sleep, a stage of sleep commonly associated with dreams, from subjects undergoing deep sleep as well as from subjects under general anaesthesia and patients with loss of consciousness due to brain injury (Massimini et al., 2010; Casali et al., 2013; Sarasso et al., 2014). It has also demonstrated utility in distinguishing between disorders of consciousness, such as distinguishing minimally conscious states and locked-in syndrome, where patients can be capable of purposeful behaviours and thoughts (Perrin et al., 2006; Schnakers et al., 2009), from vegetative states, where presumably patients are unconscious (Casarotto et al., 2016).

Outside the framework of PCI, LZc is used in neuroscience primarily as a measure of temporal complexity, rather than of spatiotemporal complexity, among other complexity measures as potential indices for depth of general anaesthesia (Zhang et al., 2001; Ferenets et al., 2006), or tools for revealing brain dynamics underlying altered states of consciousness (Mateos et al., 2018). In this vein, LZc somewhat captures IIT's notion of information, though the measure is not inspired by the theory. However, the apparent success of PCI has inspired the application of LZc to assess spatial aspects of activation, rather than just to the temporal aspects at individual regions. Applied in this manner, LZc seems to discriminate wakefulness from anesthesia when applied to functional magnetic resonance imaging (Hudetz et al., 2016), and gradually decrease from wakefulness to REM sleep, to deep sleep, when applied at the level of spike trains, local field potentials, and EEG (Abásolo et al., 2015; Andrillon et al., 2016), even without applying perturbation to the brain.

Though results from PCI and spatiotemporal LZc so far seem to support IIT and may hold clinical utility, they have similar drawbacks as alternative derivations of Φ. Specifically, their formulations avoid IIT's idea of intrinsic information, preferring to focus more on practical, clinical differentiation of levels of consciousness. Consequently, they do not directly show a link between intrinsic information and consciousness. Further, as measures which focus on distinguishing only levels of consciousness, it is unclear how they can be easily extended to take account also for contents of consciousness.

1.5 - Empirically evaluating IIT 3.0 in flies

Despite much progress towards translating IIT into practical methods which can be feasibly applied, as reviewed above, the question as to whether the constructs directly proposed by the theory itself can measure consciousness remains. As IIT is currently being updated so that its postulates more closely align with its identified axioms (Albantakis et al., 2019; A. Haun & Tononi, 2019; Barbosa et al., 2020, p. 20), having some benchmark performance of IIT 3.0 can help empirical assessment of the improvements posed in newer versions of the theory.

1.5.1 - Testing in flies

The most significant factor preventing faithful application of IIT 3.0 to real, neural data lies in how the cost of its measures grows with increasingly large systems. This cost makes it difficult to apply to large networks such as the human brain. Consequently, the question of where IIT can actually be applied, and its measures evaluated, arises. With this in mind, the fruit fly brain presents as a promising model in which to apply and evaluate the validity of IIT's measures.

The primary advantage which the fly brain provides, over a human or other mammalian brain, is its relatively simplicity. The smaller number of neurons in the fly brain, compared to more complex mammalian brains (10^5 compared to 10^8 for mice and 10^{11} for humans; (Alivisatos et al., 2012; Herculano-Houzel et al., 2006; Herculano-Houzel, 2009)) provides a system where computing IIT's measures across a large portion of neurons across the brain is more feasible to achieve. Specifically, the smaller number of neurons gives a smaller set of possible system states. In turn, this allows fewer observations are required to capture transition probabilities among all system states. The smaller brain size of the fly has already allowed for detailed imaging of neural circuits across large portions of the fly brain (Zheng et al., 2018; Scheffer et al., 2020). Detailed knowledge of connections among neurons can in the future be used to capture system transition probabilities, and inform computation of IIT constructs. For example, knowledge of connections can be used to reduce the set of disconnections to consider when computing φ or Φ.

Though the fly brain consists of a relatively small number of neurons, flies still exhibit complex behaviours. While simpler systems, such as the nematode or roundworm, have even fewer neurons than the fly, they exhibit simple behaviours which depend only on their immediate sensory environment (Barron & Klein, 2016). In contrast, the wakeful fly exhibits processes such as selective attention and spatial memory (Sareen et al., 2011; Swinderen, 2005). While these behaviours alone do not determine whether a system is conscious or not, they are useful for inferring consciousness in non-human animals (Mather, 2008). Flies additionally exhibit reductions in behavioural responsiveness, such as torpidness comparable to sleep states in mammals (Hendricks et al., 2000; Shaw et al., 2000). These periods of apparent sleep also have distinct sleeps similar to those observed in mammals (van Alphen et al., 2013; Tainton-Heap et al., 2021).

Given its relative simplicity in conjunction with complex behaviours, the fly brain is already extensively used as a model of anaesthetic loss of consciousness. Anaesthesia reduces behavioural responsiveness in flies at similar concentrations required for mammals (Allada & Nash, 1993; van Swinderen, 2006), with various observed molecular mechanisms of anaesthesia, such as decreased action potential amplitudes (Sandstrom, 2004; Wu et al., 2004), and effects on network dynamics such as reduced feedback connectivity from executive to sensory areas (Lee et al., 2009, 2013; Cohen et al., 2018), being conserved across animal species. Fly brains further appear to share graph-theoretical characteristics with mammalian brains (Shih et al., 2015) as well as cellular mechanisms (Littleton & Ganetzky, 2000), and fly LFPs share similarities with human electroencephalographic recordings (Nitz et al., 2002).

Finally, given the small size of the fly brain, multi-electrode methods can provide high quality population neural signals in both time and space which cannot be obtained using non-invasive measures available for humans. These high quality recordings can be obtained across the entirety of the brain, which is infeasible in mammalian brains using currently available recording techniques. While ultimately it is still impossible to evaluate Φ across the brain at the individual neuron level, such population recordings are more likely to be representing neural activity throughout the brain (when compared to e.g. electrodes on a linear probe inserted into the human brain), as electrical signals need to propagate through less physical media to reach and be recorded at the electrodes. Finally, the ratio of available local recordings used to compute Φ to the number of neurons which constitute the complex is likely larger for the fly brain, due to the human brain having magnitudes of order more neurons. Taken altogether, the fly serves as a useful model for applying and evaluating the measures of IIT.

1.5.2 - Thesis aims

In this thesis, I apply the measures of IIT 3.0 to real, neural data from flies in order to evaluate their validity. Specifically, each of the following chapters evaluates the following. In Chapter 2, I assess the construct validity of Φ and φ structures as measures of conscious level by estimating their values during wakefulness and anaesthesia. Next, in Chapter 3, I search for a temporal scale at which Φ is maximised, which should correspond to the timescale of conscious experience of the fly. Finally, in Chapter 4, I assess the divergent validity of Φ and the φ structures from univariate measures by applying a vast library of univariate time-series features to the same flies.

1.6 - References

Abásolo, D., Simons, S., da Silva, R. M., Tononi, G., & Vyazovskiy, V. V. (2015). Lempel-Ziv complexity of cortical activity during sleep and waking in rats. *Journal of Neurophysiology*, *113*(7), 2742–2752. https://doi.org/10.1152/jn.00575.2014

Aitken, A. (1933). A problem in combinations. *Edinburgh Mathematical Notes*, *28*, xviii–xxiii.

Albantakis, L., Hintze, A., Koch, C., Adami, C., & Tononi, G. (2014). Evolution of integrated causal structures in animats exposed to environments of increasing complexity. *PLoS Computational Biology*, *10*(12), e1003966. https://doi.org/10.1371/journal.pcbi.1003966

Albantakis, L., Marshall, W., Hoel, E., & Tononi, G. (2019). What caused what? A quantitative account of actual causation using dynamical causal networks. *Entropy*, *21*(5), 459.

Albantakis, L., & Tononi, G. (2015). The Intrinsic Cause-Effect Power of Discrete Dynamical Systems—From Elementary Cellular Automata to Adapting Animats. *Entropy*, *17*(8), 5472–5502. https://doi.org/10.3390/e17085472

Alivisatos, A. P., Chun, M., Church, G. M., Greenspan, R. J., Roukes, M. L., & Yuste, R. (2012). The brain activity map project and the challenge of functional connectomics. *Neuron*, *74*(6), 970–974. https://doi.org/10.1016/j.neuron.2012.06.006

Alkire, M. T., Hudetz, A. G., & Tononi, G. (2008). Consciousness and anesthesia. *Science (New York, N.Y.)*, *322*(5903), 876–880. PMC. https://doi.org/10.1126/science.1149213

Allada, R., & Nash, H. A. (1993). Drosophila melanogaster as a Model for Study of General Anesthesia: The Quantitative Response to Clinical Anesthetics and Alkanes. *Anesthesia & Analgesia*, *77*(1), 19–26.

Andrillon, T., Poulsen, A. T., Hansen, L. K., Léger, D., & Kouider, S. (2016). Neural Markers of Responsiveness to the Environment in Human Sleep. *Journal of Neuroscience*, *36*(24), 6583–6596. https://doi.org/10.1523/JNEUROSCI.0902-16.2016

Aru, J., Bachmann, T., Singer, W., & Melloni, L. (2012). Distilling the neural correlates of consciousness. *Neuroscience & Biobehavioral Reviews*, *36*(2), 737–746. https://doi.org/10.1016/j.neubiorev.2011.12.003

Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., Farfel, J. M., Ferretti, R. E. L., Leite, R. E. P., Filho, W. J., Lent, R., & Herculano-Houzel, S. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology*, *513*(5), 532–541. https://doi.org/10.1002/cne.21974

Barbosa, L. S., Marshall, W., Streipert, S., Albantakis, L., & Tononi, G. (2020). A measure for intrinsic information. *Scientific Reports*, *10*(1), 18803. https://doi.org/10.1038/s41598-020-75943-4

Barrett, A. B., & Seth, A. K. (2011). Practical measures of integrated information for time-series data. *PLoS Computational Biology*, *7*(1), e1001052. https://doi.org/10.1371/journal.pcbi.1001052

Barron, A. B., & Klein, C. (2016). What insects can tell us about the origins of consciousness. *Proceedings of the National Academy of Sciences*, *113*(18), 4900–4908.

Bell, E. T. (1934). Exponential Polynomials. *Annals of Mathematics*, *35*(2), 258–277. https://doi.org/10.2307/1968431

Blake, R., & Logothetis, N. K. (2002). Visual competition. *Nature Reviews Neuroscience*, *3*(1), 13–21. https://doi.org/10.1038/nrn701

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*(2), 227–247. https://doi.org/10.1017/S0140525X00038188

Boly, M., Sasai, S., Gosseries, O., Oizumi, M., Casali, A., Massimini, M., & Tononi, G. (2015). Stimulus Set Meaningfulness and Neurophysiological Differentiation: A Functional Magnetic Resonance Imaging Study. *PLOS ONE*, *10*(5), e0125337. https://doi.org/10.1371/journal.pone.0125337

Breitmeyer, B. G., & Ogmen, H. (2000). Recent models and findings in visual backward masking: A comparison, review, and update. *Perception & Psychophysics*, *62*(8), 1572–1595. https://doi.org/10.3758/BF03212157

Casali, A. G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K. R., Casarotto, S., Bruno, M.-A., Laureys, S., Tononi, G., & Massimini, M. (2013). A theoretically based index of consciousness independent of sensory processing and behavior. *Science Translational Medicine*, *5*(198), 198ra105. https://doi.org/10.1126/scitranslmed.3006294

Casarotto, S., Comanducci, A., Rosanova, M., Sarasso, S., Fecchio, M., Napolitani, M., Pigorini, A., Casali, A. G., Trimarchi, P. D., Boly, M., Gosseries, O., Bodart, O., Curto, F., Landi, C., Mariotti, M., Devalle, G., Laureys, S., Tononi, G., & Massimini, M. (2016). Stratification of unresponsive patients by an independently validated index of brain complexity. *Annals of Neurology*, *80*(5), 718–729. https://doi.org/10.1002/ana.24779

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*(3), 200–219.

Chang, J.-Y., Pigorini, A., Massimini, M., Tononi, G., Nobili, L., & Van Veen, B. D. (2012). Multivariate autoregressive models with exogenous inputs for intracerebral responses to direct electrical stimulation of the human brain. *Frontiers in Human Neuroscience*, *6*, 317. PMC. https://doi.org/10.3389/fnhum.2012.00317

Cohen, D., van Swinderen, B., & Tsuchiya, N. (2018). Isoflurane impairs low frequency feedback but leaves high frequency feedforward connectivity intact in the fly brain. *ENeuro*, ENEURO.0329-17.2018.

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences*, *2*, 263–275.

Cuaya, L. V., Hernández-Pérez, R., & Concha, L. (2016). Our Faces in the Dog's Brain: Functional Imaging Reveals Temporal Cortex Activation during Perception of Human Faces. *PLoS ONE*, *11*(3), e0149431. PMC. https://doi.org/10.1371/journal.pone.0149431

de Graaf, T. A., Hsieh, P.-J., & Sack, A. T. (2012). The 'correlates' in neural correlates of consciousness. *Neuroscience & Biobehavioral Reviews*, *36*(1), 191–197. https://doi.org/10.1016/j.neubiorev.2011.05.012

Dehaene, S., Changeux, J.-P., & Naccache, L. (2011). The global neuronal workspace model of conscious access: From neuronal architectures to clinical applications. *Characterizing Consciousness: From Cognition to the Clinic?*, 55–84.

Edelman, G. M. (2003). Naturalizing consciousness: A theoretical framework. *Proceedings of the National Academy of Sciences*, *100*(9), 5520–5524. https://doi.org/10.1073/pnas.0931349100

Engel, A. K., & Singer, W. (2001). Temporal binding and the neural correlates of sensory awareness. *Trends in Cognitive Sciences*, *5*(1), 16–25. https://doi.org/10.1016/S1364-6613(00)01568-0

Ferenets, R., Lipping, T., Anier, A., Jantti, V., Melto, S., & Hovilehto, S. (2006). Comparison of entropy and complexity measures for the assessment of depth of sedation. *IEEE*

Transactions on Biomedical Engineering, *53*(6), 1067–1077. https://doi.org/10.1109/TBME.2006.873543

Ferrarelli, F., Massimini, M., Sarasso, S., Casali, A., Riedner, B. A., Angelini, G., Tononi, G., & Pearce, R. A. (2010). Breakdown in cortical effective connectivity during midazolam-induced loss of consciousness. *Proceedings of the National Academy of Sciences*, *107*(6), 2681–2686. https://doi.org/10.1073/pnas.0913008107

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nat Neurosci*, *3*(2), 191–197. https://doi.org/10.1038/72140

Gauthier, I., Tarr, M. J., Anderson, A. W., Skudlarski, P., & Gore, J. C. (1999). Activation of the middle fusiform 'face area' increases with expertise in recognizing novel objects. *Nat Neurosci*, *2*(6), 568–573. https://doi.org/10.1038/9224

Gosseries, O., Di, H., Laureys, S., & Boly, M. (2014). Measuring Consciousness in Severely Damaged Brains. *Annual Review of Neuroscience*, *37*(1), 457–478. https://doi.org/10.1146/annurev-neuro-062012-170339

Grasso, M., Albantakis, L., Lang, J. P., & Tononi, G. (2021). Causal reductionism and causal structures. *Nature Neuroscience*, *24*(10), 1348–1355. https://doi.org/10.1038/s41593-021-00911-8

Guedel, A. E. (1937). Inhalation Anesthesia: A Fundamental Guide. *Anesthesia & Analgesia*, *16*(2), 119–120.

Haun, A. M., Oizumi, M., Kovach, C. K., Kawasaki, H., Oya, H., Howard, M. A., Adolphs, R., & Tsuchiya, N. (2017). Conscious Perception as Integrated Information Patterns in Human Electrocorticography. *ENeuro*, *4*(5), ENEURO.0085-17.2017. https://doi.org/10.1523/ENEURO.0085-17.2017

Haun, A., & Tononi, G. (2019). Why does space feel the way it does? Towards a principled account of spatial experience. *Entropy*, *21*(12), 1160. https://doi.org/10.3390/e21121160

Hendricks, J. C., Finn, S. M., Panckeri, K. A., Chavkin, J., Williams, J. A., Sehgal, A., & Pack, A. I. (2000). Rest in Drosophila is a sleep-like state. *Neuron*, *25*(1), 129–138.

Herculano-Houzel, S. (2009). The human brain in numbers: A linearly scaled-up primate brain. *Frontiers in Human Neuroscience*, *3*, 31. https://doi.org/10.3389/neuro.09.031.2009

Herculano-Houzel, S., Mota, B., & Lent, R. (2006). Cellular scaling rules for rodent brains. *Proceedings of the National Academy of Sciences*, *103*(32), 12138–12143. https://doi.org/10.1073/pnas.0611396104

Hoel, E. P., Albantakis, L., Marshall, W., & Tononi, G. (2016). Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neuroscience of Consciousness*, *2016*(1), niw012. https://doi.org/10.1093/nc/niw012

Hohwy, J. (2009). The neural correlates of consciousness: New experimental approaches needed? *Consciousness and Cognition*, *18*(2), 428–438. https://doi.org/10.1016/j.concog.2009.02.006

Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press.

Hudetz, A. G., Liu, X., Pillay, S., Boly, M., & Tononi, G. (2016). Propofol anesthesia reduces Lempel-Ziv complexity of spontaneous brain activity in rats. *Neuroscience Letters*, *628*, 132–135. PMC. https://doi.org/10.1016/j.neulet.2016.06.017

Imamoglu, F., Kahnt, T., Koch, C., & Haynes, J.-D. (2012). Changes in functional connectivity support conscious object recognition. *NeuroImage*, *63*(4), 1909–1917. https://doi.org/10.1016/j.neuroimage.2012.07.056

Isler, J. R., Stark, R. I., Grieve, P. G., Welch, M. G., & Myers, M. M. (2018). Integrated information in the EEG of preterm infants increases with family nurture intervention, age, and conscious state. *PLOS ONE*, *13*(10), e0206237. https://doi.org/10.1371/journal.pone.0206237

Kanwisher, N., & Yovel, G. (2006). The fusiform face area: A cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*(1476), 2109–2128. https://doi.org/10.1098/rstb.2006.1934

Kim, H., Hudetz, A. G., Lee, J., Mashour, G. A., Lee, U., the ReCCognition Study Group, Avidan, M. S., Bel-Bahar, T., Blain-Moraes, S., Golmirzaie, G., Janke, E., Kelz, M. B., Picton, P., Tarnal, V., Vanini, G., & Vlisides, P. E. (2018). Estimating the Integrated Information Measure Phi from High-Density Electroencephalography during States of Consciousness in Humans. *Frontiers in Human Neuroscience*, *12*, 42. https://doi.org/10.3389/fnhum.2018.00042

Kim, H., & Lee, U. (2019). Criticality as a Determinant of Integrated Information Φ in Human Brain Networks. *Entropy*, *21*(10), 981. https://doi.org/10.3390/e21100981

Koch, C. (2004). *The Quest for Consciousness: A Neurobiological Approach*. Roberts and Company.

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: Progress and problems. *Nature Reviews Neuroscience*, *17*(5), 307–321. https://doi.org/10.1038/nrn.2016.22

Koivisto, M., & Revonsuo, A. (2010). Event-related brain potential correlates of visual awareness. *Neuroscience & Biobehavioral Reviews*, *34*(6), 922–934. https://doi.org/10.1016/j.neubiorev.2009.12.002

Lamme, V. A. F. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences*, *10*(11), 494–501. https://doi.org/10.1016/j.tics.2006.09.001

Lamme, V. A. F. (2010). How neuroscience will change our view on consciousness. *Cognitive Neuroscience*, *1*(3), 204–220. https://doi.org/10.1080/17588921003731586

LeDoux, J. E., Michel, M., & Lau, H. (2020). A little history goes a long way toward understanding why we study consciousness the way we do today. *Proceedings of the National Academy of Sciences*, *117*(13), 6976–6984. https://doi.org/10.1073/pnas.1921623117

Lee, U., Kim, S., Noh, G.-J., Choi, B.-M., Hwang, E., & Mashour, G. A. (2009). The directionality and functional organization of frontoparietal connectivity during consciousness and anesthesia in humans. *Consciousness and Cognition*, *18*(4), 1069–1078. https://doi.org/10.1016/j.concog.2009.04.004

Lee, U., Kim, S., Noh, G.-J., Choi, B.-M., & Mashour, G. A. (2008). Propofol Induction Reduces the Capacity for Neural Information Integration: Implications for the Mechanism of Consciousness and General Anesthesia. *Nature Precedings*, 1–1. https://doi.org/10.1038/npre.2008.1244.2

Lee, U., Ku, S., Noh, G., Baek, S., Choi, B., & Mashour, G. A. (2013). Disruption of frontal-parietal communication by ketamine, propofol, and sevoflurane. *Anesthesiology*, *118*(6), 1264–1275. https://doi.org/10.1097/ALN.0b013e31829103f5

Littleton, J. T., & Ganetzky, B. (2000). Ion channels and synaptic organization: Analysis of the Drosophila genome. *Neuron*, *26*(1), 35–43. https://doi.org/10.1016/s0896-6273(00)81135-6

Liu, W. H. D., Thorp, T. A. S., Graham, S. G., & Aitkenhead, A. R. (1991). Incidence of awareness with recall during general anaesthesia. *Anaesthesia*, *46*(6), 435–437. https://doi.org/10.1111/j.1365-2044.1991.tb11677.x

Maquet, P., Degueldre, C., Delfiore, G., Aerts, J., Péters, J.-M., Luxen, A., & Franck, G. (1997). Functional Neuroanatomy of Human Slow Wave Sleep. *Journal of*

Neuroscience, *17*(8), 2807–2812. https://doi.org/10.1523/JNEUROSCI.17-08-02807.1997

Marshall, W., Albantakis, L., & Tononi, G. (2018). Black-boxing and cause-effect power. *PLoS Computational Biology*, *14*(4), e1006114. https://doi.org/10.1371/journal.pcbi.1006114

Marshall, W., Kim, H., Walker, S. I., Tononi, G., & Albantakis, L. (2017). How causal analysis can reveal autonomy in models of biological systems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *375*(2109), 20160358. https://doi.org/10.1098/rsta.2016.0358

Mashour, G. A., Roelfsema, P., Changeux, J.-P., & Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*, *105*(5), 776–798. https://doi.org/10.1016/j.neuron.2020.01.026

Massimini, M., Ferrarelli, F., Huber, R., Esser, S. K., Singh, H., & Tononi, G. (2005). Breakdown of Cortical Effective Connectivity During Sleep. *Science*. https://doi.org/10.1126/science.1117256

Massimini, M., Ferrarelli, F., Murphy, M. J., Huber, R., Riedner, B. A., Casarotto, S., & Tononi, G. (2010). Cortical reactivity and effective connectivity during REM sleep in humans. *Cognitive Neuroscience*, *1*(3), 176–183. https://doi.org/10.1080/17588921003731578

Mateos, D. M., Guevara Erra, R., Wennberg, R., & Perez Velazquez, J. L. (2018). Measures of entropy and complexity in altered states of consciousness. *Cognitive Neurodynamics*, *12*(1), 73–84. https://doi.org/10.1007/s11571-017-9459-8

Mather, J. A. (2008). Cephalopod consciousness: Behavioural evidence. *Consciousness and Cognition*, *17*(1), 37–48. https://doi.org/10.1016/j.concog.2006.11.006

Mayner, W. G. P., Marshall, W., Albantakis, L., Findlay, G., Marchman, R., & Tononi, G. (2018). PyPhi: A toolbox for integrated information theory. *PLOS Computational Biology*, *14*(7), e1006343. https://doi.org/10.1371/journal.pcbi.1006343

Medeiros, S. L. de S., Paiva, M. M. M. de, Lopes, P. H., Blanco, W., Lima, F. D. de, Oliveira, J. B. C. de, Medeiros, I. G., Sequerra, E. B., de Souza, S., Leite, T. S., & Ribeiro, S. (2021). Cyclic alternation of quiet and active sleep states in the octopus. *IScience*, *24*(4), 102223. https://doi.org/10.1016/j.isci.2021.102223

Mediano, P. A. M., Seth, A. K., & Barrett, A. B. (2019). Measuring Integrated Information: Comparison of Candidate Measures in Theory and Simulation. *Entropy*, *21*(1), 17. https://doi.org/10.3390/e21010017

Miller, S. M. (2014). Closing in on the constitution of consciousness. *Frontiers in Psychology*, *5*, 1293. https://doi.org/10.3389/fpsyg.2014.01293

Morin, A. (2006). Levels of consciousness and self-awareness: A comparison and integration of various neurocognitive views. *Consciousness and Cognition*, *15*(2), 358–371. https://doi.org/10.1016/j.concog.2005.09.006

Niizato, T., Sakamoto, K., Mototake, Y., Murakami, H., Tomaru, T., Hoshika, T., & Fukushima, T. (2020a). Finding continuity and discontinuity in fish schools via integrated information theory. *PLOS ONE*, *15*(2), e0229573. https://doi.org/10.1371/journal.pone.0229573

Niizato, T., Sakamoto, K., Mototake, Y., Murakami, H., Tomaru, T., Hoshika, T., & Fukushima, T. (2020b). Four-Types of IIT-Induced Group Integrity of Plecoglossus altivelis. *Entropy*, *22*(7), 726. https://doi.org/10.3390/e22070726

Nitz, D. A., van Swinderen, B., Tononi, G., & Greenspan, R. J. (2002). Electrophysiological correlates of rest and activity in Drosophila melanogaster. *Current Biology*, *12*(22), 1934–1940. https://doi.org/10.1016/S0960-9822(02)01300-3

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Comput Biol*, *10*(5), e1003588. https://doi.org/10.1371/journal.pcbi.1003588

Oizumi, M., Amari, S., Yanagawa, T., Fujii, N., & Tsuchiya, N. (2016). Measuring integrated information from the decoding perspective. *PLoS Computational Biology*, *12*(1), e1004654. https://doi.org/10.1371/journal.pcbi.1004654

Perrin, F., Schnakers, C., Schabus, M., Degueldre, C., Goldman, S., Brédart, S., Faymonville, M.-E., Lamy, M., Moonen, G., Luxen, A., Maquet, P., & Laureys, S. (2006). Brain Response to One’s Own Name in Vegetative State, Minimally Conscious State, and Locked-in Syndrome. *Archives of Neurology*, *63*(4), 562–569. https://doi.org/10.1001/archneur.63.4.562

Person, A., & Raman, I. (2012). Synchrony and neural coding in cerebellar circuits. *Frontiers in Neural Circuits*, *6*, 97. https://doi.org/10.3389/fncir.2012.00097

Rani, D. D., & Harsoor, S. S. (2012). Depth of general anaesthesia monitors. *Indian Journal of Anaesthesia*, *56*(5), 437–441. PMC. https://doi.org/10.4103/0019-5049.103956

Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover’s distance as a metric for image retrieval. *International Journal of Computer Vision*, *40*(2), 99–121.

Sandstrom, D. J. (2004). Isoflurane depresses glutamate release by reducing neuronal excitability at the Drosophila neuromuscular junction. *The Journal of Physiology*, *558*(2), 489–502. https://doi.org/10.1113/jphysiol.2004.065748

Sarasso, S., Rosanova, M., Casali, A. G., Casarotto, S., Fecchio, M., Boly, M., Gosseries, O., Tononi, G., Laureys, S., & Massimini, M. (2014). Quantifying cortical EEG responses to TMS in (un)consciousness. *Clinical EEG and Neuroscience*, *45*(1), 40–49. https://doi.org/10.1177/1550059413513723

Sareen, P., Wolf, R., & Heisenberg, M. (2011). Attracting the attention of a fly. *Proceedings of the National Academy of Sciences*, *108*(17), 7230–7235. https://doi.org/10.1073/pnas.1102522108

Sasai, S., Boly, M., Mensen, A., & Tononi, G. (2016). Functional split brain in a driving/listening paradigm. *Proceedings of the National Academy of Sciences*, *113*(50), 14444–14449. https://doi.org/10.1073/pnas.1613200113

Scheffer, L. K., Xu, C. S., Januszewski, M., Lu, Z., Takemura, S., Hayworth, K. J., Huang, G. B., Shinomiya, K., Maitlin-Shepard, J., Berg, S., Clements, J., Hubbard, P. M., Katz, W. T., Umayam, L., Zhao, T., Ackerman, D., Blakely, T., Bogovic, J., Dolafi, T., … Plaza, S. M. (2020). A connectome and analysis of the adult Drosophila central brain. *ELife*, *9*, e57443. https://doi.org/10.7554/eLife.57443

Schnakers, C., Perrin, F., Schabus, M., Hustinx, R., Majerus, S., Moonen, G., Boly, M., Vanhaudenhuyse, A., Bruno, M.-A., & Laureys, S. (2009). Detecting consciousness in a total locked-in syndrome: An active event-related paradigm. *Neurocase*, *15*(4), 271–277. https://doi.org/10.1080/13554790902724904

Sebel, P. S., Bowdle, T. A., Ghoneim, M. M., Rampil, I. J., Padilla, R. E., Gan, T. J., & Domino, K. B. (2004). The Incidence of Awareness During Anesthesia: A Multicenter United States Study. *Anesthesia & Analgesia*, *99*(3), 833–839. https://doi.org/10.1213/01.ane.0000130261.90896.6c

Sevenius Nilsen, A., Juel, B. E., & Marshall, W. (2019). Evaluating approximations and heuristic measures of integrated information. *Entropy*, *21*(5), 525. https://doi.org/10.3390/e21050525

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(3), 379–423.

Shaw, P. J., Cirelli, C., Greenspan, R. J., & Tononi, G. (2000). Correlates of sleep and waking in Drosophila melanogaster. *Science*, *287*(5459), 1834–1837.

Shehata, M., Cheng, M., Leung, A., Tsuchiya, N., Wu, D.-A., Tseng, C., Nakauchi, S., & Shimojo, S. (2021). Team Flow Is a Unique Brain State Associated with Enhanced Information Integration and Interbrain Synchrony. *ENeuro*, *8*(5), ENEURO.0133-21.2021. https://doi.org/10.1523/ENEURO.0133-21.2021

Shih, C.-T., Sporns, O., Yuan, S.-L., Su, T.-S., Lin, Y.-J., Chuang, C.-C., Wang, T.-Y., Lo, C.-C., Greenspan, R. J., & Chiang, A.-S. (2015). Connectomics-based analysis of information flow in the Drosophila brain. *Current Biology*, *25*(10), 1249–1258. https://doi.org/10.1016/j.cub.2015.03.021

Swinderen, B. van. (2005). The remote roots of consciousness in fruit-fly selective attention? *BioEssays*, *27*(3), 321–330. https://doi.org/10.1002/bies.20195

Tainton-Heap, L. A. L., Kirszenblat, L. C., Notaras, E. T., Grabowska, M. J., Jeans, R., Feng, K., Shaw, P. J., & van Swinderen, B. (2021). A Paradoxical Kind of Sleep in Drosophila melanogaster. *Current Biology*, *31*(3), 578-590.e6. https://doi.org/10.1016/j.cub.2020.10.081

Tegmark, M. (2016). Improved measures of integrated information. *PLoS Computational Biology*, *12*(11), e1005123. https://doi.org/10.1371/journal.pcbi.1005123

Toker, D., & Sommer, F. T. (2019). Information integration in large brain networks. *PLoS Computational Biology*, *15*(2), e1006807. https://doi.org/10.1371/journal.pcbi.1006807

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, *5*(1), 42. https://doi.org/10.1186/1471-2202-5-42

Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *The Biological Bulletin*, *215*(3), 216–242. https://doi.org/10.2307/25470707

Tsuchiya, N., & Koch, C. (2005). Continuous flash suppression reduces negative afterimages. *Nature Neuroscience*, *8*(8), 1096–1101. https://doi.org/10.1038/nn1500

van Alphen, B., Yap, M. H. W., Kirszenblat, L., Kottler, B., & van Swinderen, B. (2013). A Dynamic Deep Sleep Stage in Drosophila. *The Journal of Neuroscience*, *33*(16), 6917. https://doi.org/10.1523/JNEUROSCI.0061-13.2013

van Swinderen, B. (2006). A succession of anesthetic endpoints in the Drosophila brain. *Journal of Neurobiology*, *66*(11), 1195–1211. https://doi.org/10.1002/neu.20300

Witter, L., Rudolph, S., Pressler, R. T., Lahlaf, S. I., & Regehr, W. G. (2016). Purkinje Cell Collaterals Enable Output Signals from the Cerebellar Cortex to Feed Back to Purkinje Cells and Interneurons. *Neuron*, *91*(2), 312–319. https://doi.org/10.1016/j.neuron.2016.05.037

Wu, X.-S., Sun, J.-Y., Evers, A. S., Crowder, M., & Wu, L.-G. (2004). Isoflurane inhibits transmitter release and the presynaptic action potential. *Anesthesiology*, *100*(3), 663–670. https://doi.org/10.1097/00000542-200403000-00029

Yu, F., Jiang, Q., Sun, X., & Zhang, R. (2015). A new case of complete primary cerebellar agenesis: Clinical and imaging findings in a living patient. *Brain*, *138*(6), e353. https://doi.org/10.1093/brain/awu239

Zhang, X.-S., Roy, R. J., & Jensen, E. W. (2001). EEG complexity as a measure of depth of anesthesia for patients. *IEEE Transactions on Biomedical Engineering*, *48*(12), 1424–1433. https://doi.org/10.1109/10.966601

Zheng, Z., Lauritzen, J. S., Perlman, E., Robinson, C. G., Nichols, M., Milkie, D., Torrens, O., Price, J., Fisher, C. B., Sharifi, N., Calle-Schuler, S. A., Kmecova, L., Ali, I. J., Karsh, B., Trautman, E. T., Bogovic, J. A., Hanslovsky, P., Jefferis, G. S. X. E., Kazhdan, M., … Bock, D. D. (2018). A complete electron microscopy volume of the brain of adult Drosophila melanogaster. *Cell*, *174*(3), 730-743.e22. https://doi.org/10.1016/j.cell.2018.06.019

Chapter 2 - Empirical validity of IIT's proposed measures

In this chapter, I aim to assess the construct validity of IIT's proposed measures of consciousness. To address this aim, I apply the theory's measures to recordings obtained from the fly brain during wakefulness and anaesthesia. For this chapter, I supply a manuscript published in PLOS Computational Biology.

2.1 - PLOS Computational Biology publication

This article was published in PLOS Computational Biology. It begins on the following page with its original page numbering.

Leung, A., Cohen, D., Swinderen, B. van, & Tsuchiya, N. (2021). Integrated information structure collapses with anesthetic loss of conscious arousal in Drosophila melanogaster. *PLOS Computational Biology*, *17*(2), e1008722.

RESEARCH ARTICLE

Integrated information structure collapses with anesthetic loss of conscious arousal in *Drosophila melanogaster*

Angus Leung[1]*, **Dror Cohen**[1,2], **Bruno van Swinderen**[3], **Naotsugu Tsuchiya**[1,2,4,5]*

1 School of Psychological Sciences, Monash University, Melbourne, Australia, **2** Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology (NICT), Osaka, Japan, **3** Queensland Brain Institute, The University of Queensland, Brisbane, Australia, **4** Monash Institute of Cognitive and Clinical Neuroscience (MICCN), Monash University, Melbourne, Australia, **5** Advanced Telecommunications Research Computational Neuroscience Laboratories, Kyoto, Japan

* angus.leung1@monash.edu (AL); naotsugu.tsuchiya@monash.edu (NT)

OPEN ACCESS

Citation: Leung A, Cohen D, van Swinderen B, Tsuchiya N (2021) Integrated information structure collapses with anesthetic loss of conscious arousal in *Drosophila melanogaster*. PLoS Comput Biol 17(2): e1008722. https://doi.org/10.1371/journal.pcbi.1008722

Editor: Daniele Marinazzo, Ghent University, BELGIUM

Received: May 15, 2020

Accepted: January 18, 2021

Published: February 26, 2021

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pcbi.1008722

Data Availability Statement: Preprocessed fly LFPs from which information structures were constructed are available on Figshare - doi: 10.26180/5ebe420ae8d89.

Abstract

The physical basis of consciousness remains one of the most elusive concepts in current science. One influential conjecture is that consciousness is to do with some form of causality, measurable through information. The integrated information theory of consciousness (IIT) proposes that conscious experience, filled with rich and specific content, corresponds directly to a hierarchically organised, irreducible pattern of causal interactions; i.e. an integrated informational structure among elements of a system. Here, we tested this conjecture in a simple biological system (fruit flies), estimating the information structure of the system during wakefulness and general anesthesia. Consistent with this conjecture, we found that integrated interactions among populations of neurons during wakefulness collapsed to isolated clusters of interactions during anesthesia. We used classification analysis to quantify the accuracy of discrimination between wakeful and anesthetised states, and found that informational structures inferred conscious states with greater accuracy than a scalar summary of the structure, a measure which is generally championed as the main measure of IIT. In stark contrast to a view which assumes feedforward architecture for insect brains, especially fly visual systems, we found rich information structures, which cannot arise from purely feedforward systems, occurred across the fly brain. Further, these information structures collapsed uniformly across the brain during anesthesia. Our results speak to the potential utility of the novel concept of an "informational structure" as a measure for level of consciousness, above and beyond simple scalar values.

Author summary

The physical basis of consciousness remains elusive. Efforts to measure consciousness have generally been restricted to simple, scalar quantities which summarise the complexity of a system, inspired by integrated information theory, which links a multi-dimensional, informational structure to the contents of experience in a system. Due to the complexity

Funding: AL was supported by an Australian Government Research Training Program (RTP) Scholarship. DC was supported by a JSPS international postdoctoral research fellowship and JSPS KAKENHI 18F18706. DC and NT were supported by TWCF0199 from Templeton World Charity Foundation, Inc. NT was funded by the Australian Research Council Future Fellowship FT120100619 and the Discovery Project DP180104128 and DP180100396 and the National Health and Medical Research Council APP1183280. BvS was funded by the National Health and Medical Research Council Project APP1103923. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

of the definition of the structure, assessment of its utility as a measure of conscious arousal in a system has largely been ignored. In this manuscript we evaluate the utility of such an information structure in measuring the level of arousal in the fruit fly. Our results indicate that this structure can be more informative about the level of arousal in a system than even the single-value summary proposed by the theory itself. These results may push consciousness research towards the notion of multi-dimensional informational structures, instead of traditional scalar summaries.

Introduction

The question of how subjective, conscious experience arises from physical interactions has been pondered by philosophers for centuries [1,2], and now has moved into the domain of cognitive neuroscience [3–5]. Because we are only able to experience our own individual consciousness, exact inference of others' conscious contents (i.e., what it is like to be a bat [1]) seems intractable. However, broader inference on levels of consciousness, ranging from low during coma and deep anesthesia to high in wakeful states seems possible across animals. Behaviors of animals, ranging from humans to insects, all seem to change in a similar manner from highly active wakefulness with marked high-level cognitive capability to loss of consciousness with negligible cognitive functions. Indeed, such inferences have been widely accepted across various losses of consciousness in brain damaged patients [6] and non-human mammals [7,8], and are now becoming applied to insects [9–12].

As stated, complex behavioural repertoires of animals, ranging from humans to insects, all seem to reduce in a similar manner from highly active wakefulness to loss of consciousness. During wakefulness, flies, for instance, have been shown to exhibit processes such as working memory [13–15], attention [16–18], and feature binding [19]. Flies also seem to experience varying states of arousal which are physiologically regulated in a similar manner to mammals, such as sleep [9,20,21] and anesthesia [11,12]. Despite these similarities, processing in the fly brain is largely thought to be feedforward, with potential exception of central structures such as the central complex and mushroom bodies [22–24]. This is in contrast to a more integrative view of the seemingly more complex brains of mammals, featuring both feedforward and feedback interactions during wakefulness from primary sensory areas to midbrain and executive areas [25–27].

The importance of feedback for conscious processing is emphasised in an influential view that consciousness arises with "integrated information" [28–30]. Integrated information, distinct from the standard notion of Shannon information [31], is defined as "differences that make a difference within a system" [32,29,30]. In other words, integrated information is concerned with how elements of a system causally influence each other such that information is accessible to the system itself (extrinsic information, conversely, concerns how states of a system causally influence states of another, separate system; see supporting information in [33]). Integrated information theory (IIT; [28,30,34]) provides a mathematical quantification of integrated information, and proposes that it is critical for consciousness to arise. Specifically, IIT describes how hierarchically organized elements uniquely and causally interact with other elements within a system in an integrated manner to produce information accessible to the system itself. According to IIT, the "maximally irreducible conceptual structure" [30] is hypothesised to directly correspond to the quantity and quality of consciousness. That is, the richer and more specific the informational structure of the system, the higher the level of consciousness in a system, and the richer the contents that the system consciously experiences.

Critically, the hierarchically organised elements must both exert effects on other elements and receive effects from others, all within the system, and thus these structures can only arise with the presence of both feedforward and feedback interactions.

While IIT offers a compelling theoretical account linking integrated information and consciousness, empirical applications of the theory remain rare [35]. Thus, whether *empirically* estimated integrated information structures relate to conscious arousal remains largely unknown. While we as yet cannot be certain of consciousness in flies, they pose an interesting system to apply the theory. In particular, regardless of consciousness per se, a purely feedforward brain should give zero integrated information and correspondingly a minimal informational structure. Thus, we address the following questions. First, how can we estimate informational structures from neural activity recorded from a biological system? Second, does the fly brain generate integrated information and non-minimal information structures? If it does, would the structures be collapsed during reduced arousal as manipulated using general anesthesia? And third, does integrated information and its associated information structures arise (and subsequently collapse during anesthesia) primarily in the central regions of the fly brain?

We address the above questions by analyzing neural recordings from the fruit fly, collected during wakefulness and isoflurane anesthesia [12,36]. We apply a novel construct, "integrated information structures" (IIS), to capture the level of arousal of the fly. We found that the structures which were present during wakefulness collapsed during anesthesia. Critically, they were better at classifying arousal states than a scalar summary (i.e. just a single number), a measure which is usually championed as "integrated information" in IIT, with their collapse occurring all throughout the fly brain. Our results indicate the presence of feedback interactions across the fly brain during wakefulness, and demonstrate the utility of information structure as a measure for level of arousal, above and beyond simple scalar values, opening the door for improved clinical measures of consciousness.

Results

Constructing integrated information structures from fly local field potentials

To construct the IIS, we used local field potentials (LFPs; hereafter referred to as "channels") recorded from the fruit fly brain (Fig 1A; see Methods; [12]). LFPs were recorded using a linear multi-electrode array, such that 15 channels covered both peripheral and central regions of the brain. We operationally defined the discrete state of each channel at each time by binarizing it with respect to the median voltage of that channel (Fig 1B; see also S1 Text for effect of binarizing using different thresholds).

Fig 1C–1I illustrates the steps to estimate the IIS of two channels, A and B. From the empirically observed time course of the two discretized channels, we first construct a transition probability matrix (TPM; Fig 1C). Each entry of the TPM gives the probability of a given channel taking some state in the future, given the current state of all channels in the system (see Methods). Thus, the TPM characterizes how the whole system (A and B) evolves over time, containing all necessary information for unfolding how subsets of the system (A, B, and AB) "causally" (in a statistical sense) interact to irreducibly specify the state of the whole system (AB). We refer to causality as statistically inferred from conditional probability distributions [37], which is not necessarily the same as perturbational causality [38]. We return to the issue of estimating the TPM from observed versus perturbed transitions in the Discussion. Importantly, we use the TPM to measure the information that each subset of the system specifies regarding some other subset of the system, as we describe below.

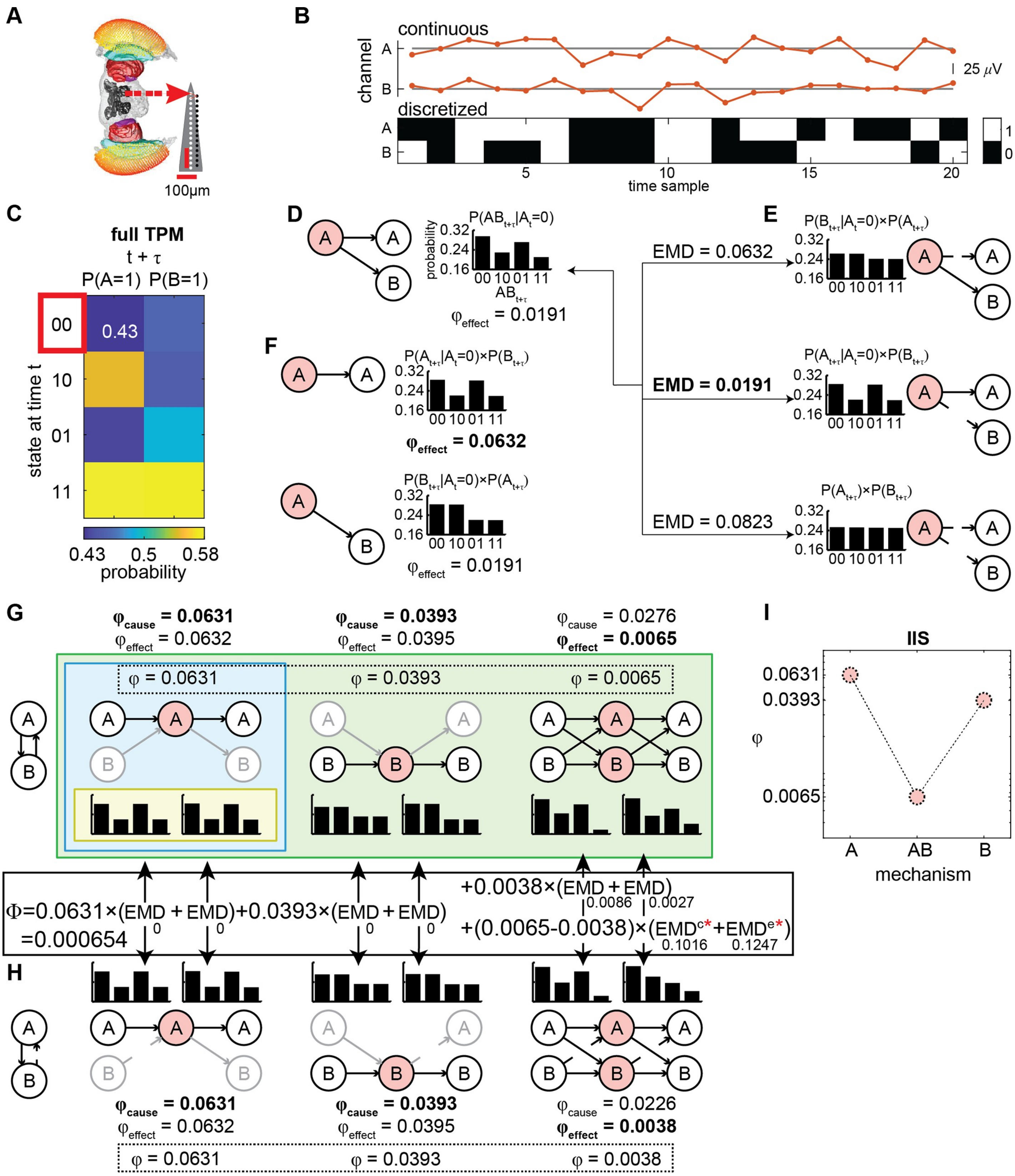
A
100μm
B
continuous
channel
A
B
25 μV
discretized
A
B
1
0
5
10
15
20
time sample
C
full TPM
t + τ
P(A=1) P(B=1)
00
10
01
11
0.43
state at time t
0.43 0.5 0.58
probability
D
P(AB_{t+τ}|A_t=0)
probability
0.32
0.24
0.16
00 10 01 11
AB_{t+τ}
φ_effect = 0.0191
E
P(B_{t+τ}|A_t=0)×P(A_{t+τ})
EMD = 0.0632
P(A_{t+τ}|A_t=0)×P(B_{t+τ})
EMD = 0.0191
P(A_{t+τ})×P(B_{t+τ})
EMD = 0.0823
F
P(A_{t+τ}|A_t=0)×P(B_{t+τ})
φ_effect = 0.0632
P(B_{t+τ}|A_t=0)×P(A_{t+τ})
φ_effect = 0.0191
G
φ_cause = 0.0631
φ_effect = 0.0632
φ_cause = 0.0393
φ_effect = 0.0395
φ_cause = 0.0276
φ_effect = 0.0065
φ = 0.0631
φ = 0.0393
φ = 0.0065
Φ=0.0631×(EMD + EMD)+0.0393×(EMD + EMD)
=0.000654
0 0 0 0
+0.0038×(EMD + EMD)
0.0086 0.0027
+(0.0065-0.0038)×(EMD^c* + EMD^e*)
0.1016 0.1247
H
φ_cause = 0.0631
φ_effect = 0.0632
φ_cause = 0.0393
φ_effect = 0.0395
φ_cause = 0.0226
φ_effect = 0.0038
φ = 0.0631
φ = 0.0393
φ = 0.0038
I
IIS
0.0631
0.0393
0.0065
φ
A
AB
B
mechanism

Fig 1. Summary of IIT and processing pipeline for computing the IIS from LFPs. (**A**) Multi-electrode probe recording of LFPs from the fly. (**B**) Continuous LFPs (red, top) are discretized (black/white, bottom) by comparing to the median voltage for each trial. Displayed is an example of 20 samples for a set of two channels A and B. (**C**) A state-by-channel transition probability matrix (TPM; see Methods) describes how the state of a system at time *t* specifies the possible future states of each channel at time t+τ (τ = 4 ms). For example, the top left entry of the full TPM is 0.43, which represents the probability of channel A being '1' at time t+τ given that channel A and B were both '0' at time *t*. (**D**) At a given state (e.g. A = '0' and B = '0' at time *t*, outlined in red in **C**, the effect information specified by a subset ("mechanism"; here A, in light red) over the future states of another subset ("purview"; here A and B, in white), is given by the probability distribution of the purview conditioned on the current state of the mechanism. (**E**) To compute integrated information (φ_{effect}) of mechanism A over purview AB, we find the disconnections (i.e. replacing connections with random-noise connections) between the mechanism and the purview (indicated by broken arrows) which best approximate the original probability distribution. We compare the disconnected probability distributions to the original distribution using the earth mover's distance (EMD; treating probabilities as "earth" to be moved). We interpret the minimum EMD (bolded) as irreducible information generated over the purview by the mechanism (i.e. φ_{effect}). (**F**) We compute φ_{effect} for every possible purview (A, B, and AB as in **D**, with values 0.0632, 0.0191, and 0.0191 respectively), and select the purview and its associated probability distribution which gives the maximally integrated effect (bold). As probability distributions (bar graphs), we display the distribution over both channels A and B, assuming the maximum entropy distribution and independence on the channels outside of the purview. (**G**) φ_{cause} is determined in the same manner as φ_{effect}, except looking at possible past states of the purview (at t-τ). Both φ_{cause} and φ_{effect}, and their associated probability distributions, are determined for every mechanism (A, B, and AB; left and right distributions are cause and effect probability distributions of the selected purviews; channels outside of the purview are greyed out). The overall φ generated by a mechanism is the minimum of φ_{cause} and φ_{effect} (bolded and in the dotted box). Yellow, blue, and green backgrounds (innermost, middle, and outermost rectangles) indicate correspondence with the IIT terminology of "cause-effect repertoire", "concept", and "cause-effect structure" (CES), respectively. (**H**) All φ values and associated probability distributions are re-computed for each possible uni-directional cut (again, replacing with random-noise connections) separating the channels into a feedforward interaction from one subset of channels to the remainder of the system. Broken lines here depict the cut removing channel B's input to A. System-level integrated information (Φ) is the sum of distances between cause and effect probability distributions specified by the full and (minimally) disconnected system, weighted by the φ value for each mechanism (hence Φ is the minimum across all possible system level cuts; solid box between **G** and **H**; see Methods for details on EMD^c and EMD^e which are marked by red asterisks). Note that distances between 1-channel mechanisms were 0, not contributing to Φ, which we found to be the case in general (see also S6 Text about the role of 1-channel mechanism in our results). (**I**) We take the φ values of each mechanism (within the dotted box in **G**) to form the integrated information structure (IIS).

https://doi.org/10.1371/journal.pcbi.1008722.g001

Fig 1D considers how subset A's current state (A = '0' at time *t*) specifies the future state of any subset of the system AB (at time t+τ; we use τ = 4 ms; we repeated analyses also at τ = 2 ms and 6 ms, see S2 Text). For brevity we will refer to the subset whose current status is analyzed (red circles in Fig 1D–1G) as a "mechanism", and the affected subset (white circles in Fig 1D–1G) as a "purview", following IIT terminology; [30]. Based on the TPM, we can compute a probability distribution over past and future purview states (the bar graphs in Fig 1D–1H), given the current state of the mechanism in consideration; for example, the bar graph in Fig 1D shows that if mechanism A is in state '0' at time *t*, AB is more likely to be '00' or '01' than '10' or '11' at time t+τ. Such a probability distribution specifies the information generated by a mechanism over a given purview.

Fig 1E illustrates the procedure to find "irreducibility" of the causal interaction from mechanism A to purview AB. To estimate how much the purview is irreducible, or uniquely determined by integrative interactions between A and the purview (according to IIT's integration axiom), we estimate probability distributions assuming that some causal interactions are "disconnected" (i.e. statistically noised; see S3 Text). We quantify the degree of causal interactions by computing the distance between the two probability distributions (distance is measured using the earth mover's distance, with probabilities being moved as "earth"; EMD; [39]). The distance between the full (Fig 1D) and disconnected distribution which best approximates (i.e. is closest to) the original full distribution (Fig 1E) quantifies integrated information φ. Here, φ of A on AB can be understood as the degree to which mechanism A generates information about purview AB, above and beyond independent parts. In Fig 1E, the disconnection from A to B minimally affects the distribution out of all the possible cuts, giving φ of 0.0191.

Next, Fig 1F illustrates the identification of the purview over which A generates the most integrated information, as dictated by IIT's exclusion axiom (the exclusion axiom in this context means that only the maximal information specified by A should be considered in order to avoid information being multiplied beyond necessity). The purview for which A generates the most integrated information is referred to as A's "core effect". Here, mechanism A has a set of candidate purviews: A, B and AB. Based on the current state of A ('0'), we repeat the process of measuring distances between full distributions and disconnected distributions among all purviews (i.e. all subsets which are potentially affected by A). In this particular case, purview A is the core effect (φ of A on A is 0.0632, compared to φ of A on B and φ of A on AB both being 0.0191). Next, we perform similar operations on the TPM, but now looking at information the mechanism generates about a purview's *past*, instead of future. This is done to estimate the core *cause* of A. According to the intrinsic existence axiom of IIT, we consider A's overall influence (i.e. the information it generates for the system), to be the minimum of A's cause and effect. Consequently, a mechanism which only provides outputs to its purview (i.e. only specifies its effects), or only takes inputs from its purview (i.e. only specifies causes), generates zero integrated information.

Repeating the procedure (Fig 1D–1F) for all candidate mechanisms (A, B, and AB), Fig 1G characterizes how all possible elements of the system specify the set of structured and integrated causal interactions, listing a full set of core causes and effects of all the mechanisms. The full set of distributions for all core causes and effects for all mechanisms, and their associated integrated information values is referred to in IIT as a cause-effect structure (CES).

Finally, Fig 1H explains how IIT arrives at a purported measure of level of consciousness, system-level integrated information Φ, through a system-level disconnection. The process of identifying core causes and effects for each mechanism is repeated after making unidirectional disconnections to the full system, in the same manner as disconnecting mechanisms from purviews. System-level integrated information is the sum of EMDs between the full CES and the CES of the statistically disconnected system, weighted by the integrated information φ of each mechanism in the full CES (as depicted in the calculation between Fig 1H and 1G; see Methods). Once again, as there are many possible ways of disconnecting the system, we select the disconnection which best approximates the CES of the full, whole system (i.e. which generates the smallest weighted EMD between the full CES and the disconnected CES). Consequently, a completely feedforward system generates zero system-level integrated information, as the unidirectional disconnection of feedback connections (which are actually non-existent) will yield identical probability distributions for all mechanisms and thus an identical CES as the fully connected system. In the case of the 2-channel system AB, the minimal disconnection is the disconnection from B to A. This disconnected CES is used to assess system-level integrated information (for details, see Methods).

One difficulty with Φ is the high computational cost due to the combinatorial explosion of all possible system cuts. To enable us to search through all possible cuts, we restricted analysis to 4 channels at a time, using every combination of 4 channels as a "system". This provided a good balance between spatial coverage for each set of channels and computation time.

We also considered a computationally cheaper alternative to Φ. Specifically, we assessed a set of φ values, which we term Integrated Information Structure (IIS; Fig 1I), as an alternative measure for discriminating level of consciousness. A set of mechanism-level φ values are faster to compute, as they are already obtained as part of the computation of Φ. The IIS is an approximation of the full cause-effect structure proposed by IIT [39]. While the cause-effect structure requires causal intervention for building the TPM, here we only observe interactions as they naturally occur over time. Further, the full cause-effect structure holds details beyond just integrated information values, specifically the purviews of each mechanism and their associated probability distributions, whereas for simplicity the IIS only considers the integrated information values themselves. As system-level integrated information and the IIS are obtained for each possible state of the system, we averaged across these states, weighting by the occurrences of each state [40].

Fig 2 shows an example IIS obtained from 1 fly, 1 channel set during both wakefulness and anesthesia, when extending this process to the 4-channel case. IIT provides two main hypotheses for this paper: 1) system-level integrated information (Φ) should be reduced by general

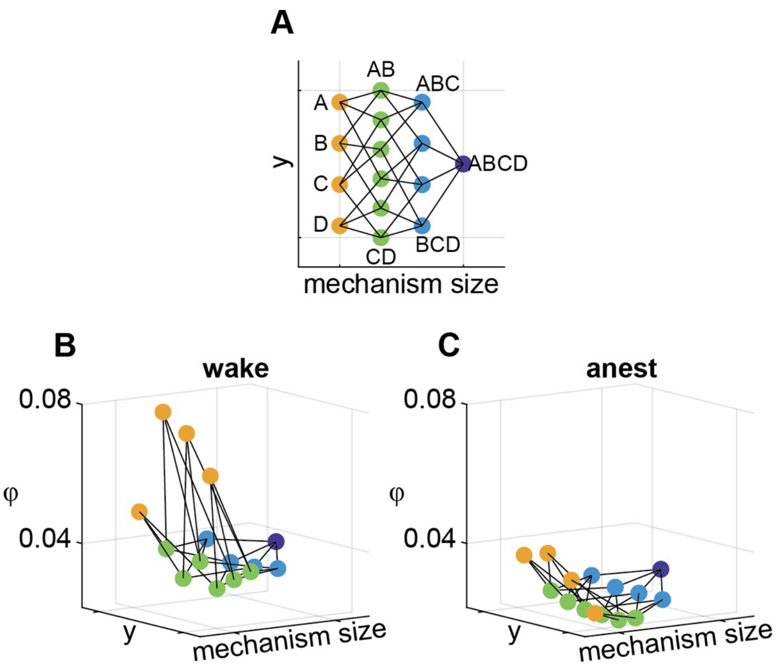

Fig 2. A 3D representation of an integrated information structure (IIS) for one channel set for one fly. (**A**) Top-down view of the IIS. Mechanism size refers to the number of channels that constitute each mechanism (yellow, green, light blue, and dark blue dots indicate mechanisms consisting of 1, 2, 3, and 4 channels respectively). The y-axis is arbitrarily set to give equal spacing between mechanisms. Lines indicate inclusion relations (e.g., mechanism AB consists of A and B). (**B**) An exemplar IIS from a single fly and channel set, during wakefulness. (**C**) An IIS from the same fly and channel set as in **B** during anesthesia. A 3D rotation video of the IIS is available at http://dx.doi.org/10.26180/5eb952457b48f.

https://doi.org/10.1371/journal.pcbi.1008722.g002

anesthesia, and 2) a set of mechanism-level integrated information (φ) values, the IIS, should also collapse during general anesthesia, reflected by reduced φ values for each mechanism (as opposed to increased φ for some mechanisms). While IIT does not explicitly predict the latter, we reasoned that level of consciousness should generally correlate with the richness of contents of consciousness. Note that these hypotheses here cannot directly confirm or invalidate IIT as a theory of consciousness, as the nature of insect consciousness is still unclear, and we do not apply every aspect of IIT (due to feasibility issues), which we expand on in the Discussion.

System-level integrated information reduces globally due to general anesthesia

We first checked the prediction that system-level integrated information (Φ), IIT's proposed measure of level of consciousness, was reduced during anesthesia. Using linear mixed effects analysis (to account for intra-fly channel set correlations; see Methods), we indeed found system-level integrated information to be significantly affected by anesthesia ($\chi^2(1) = 6.656 \times 10^3$, $p < .001$; likelihood ratio test, see Methods). Specifically, it was reduced during anesthesia (Figs 3 and 4A and 4B; β = -0.012, $t(12) = -2.525$, $p = .013$, one-tailed). This analysis also indicated that the fly LFPs did indeed generate non-minimal system-level integrated information during wakefulness. Across channel sets, system-level integrated information was significantly reduced during anesthesia for 12 of the 13 flies (S4 Text).

We next looked at whether anesthesia's effect depended on the spatial location of the channel sets. From the linear arrangement of channels from our recording setup, we characterised two features of each channel set, 1) the average location of channels in the set (relative to

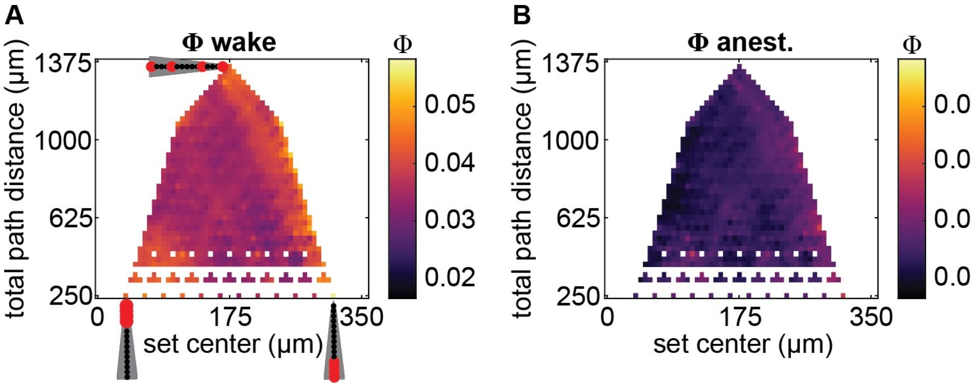

Fig 3. Spatial map of system-level integrated information Φ. (**A**) System-level integrated information Φ values during wakefulness, averaged across flies, as a function of average channel location relative to the position of the most central channel (x-axis; larger values indicate channel sets which on average are more peripherally located), and sum of pairwise distances between each pair of channels (total path distance; y-axis) within each channel set. Channel arrays, as in Fig 1A, indicate example locations of channels (in red) and their spacing along the two axes. Channel sets with identical centers and path distances were averaged. A subset of otherwise unfilled values in the map were linearly interpolated to reduce gaps in the map. (**B**) System-level integrated information Φ values during anesthesia.

https://doi.org/10.1371/journal.pcbi.1008722.g003

channel 1, the most central channel) and 2) the distance among channels in the set (the sum of pairwise distances between channel labels; total path distance).

Given the general view of the insect brain being largely feedforward [22,23], with potential exception of central brain structures which are responsible for integrating inputs from the periphery [10,24], we expected to find greater system-level integrated information for more centrally located channel sets. Opposite to our expectation, however, we found a trend indicating that peripheral channel sets tended to have slightly but significantly greater system-level integrated information ($\beta = 1.750 \times 10^{-2}$, $\chi^2(1) = 39.31$, $p < .001$). This trend was stronger during anesthesia, as indicated by a significant interaction between channel set location and wake/anesthesia condition ($\beta = 2.613 \times 10^{-2}$, $\chi^2(1) = 43.80$, $p < .001$). Thus, centrally located channel sets seemed to be more affected by anesthesia, despite having less system-level integrated information than peripheral channels. The latter finding is consistent with a view that central brains are more critical and sensitive to the level of arousal.

We also considered the effect of the spacing of channels within each channel set. If local recurrent connections drive the generation of integrated information, more "local" channel sets consisting of closely located channels would have greater system-level integrated information. Conversely, if long range recurrent connections are more important, more "global" channel sets consisting of widely spaced channels would have greater system-level integrated information, reflecting integration across the whole brain. We found system-level integrated information to increase slightly with greater distance among channels ($\beta = 1.364 \times 10^{-3}$, $\chi^2(1) = 5.351$, $p < .021$), with the direction of the trend being reversed during anesthesia (significant interaction between anesthesia and channel distance; $\beta = -2.714 \times 10^{-3}$, $\chi^2(1) = 10.59$, $p < .001$). Thus, during wakefulness, the more global sampling of channels tended to yield larger system-level integrated information, while anesthesia disrupted this effect to some extent.

Overall, system-level integrated information was reduced regardless of spatial location or distance among channels, suggesting the presence of both feedforward and feedback interactions all across the fly brain. So, for analysis on the multi-dimensional IIS, we analysed all channel sets together without dividing into groups based on location or distance among channels.

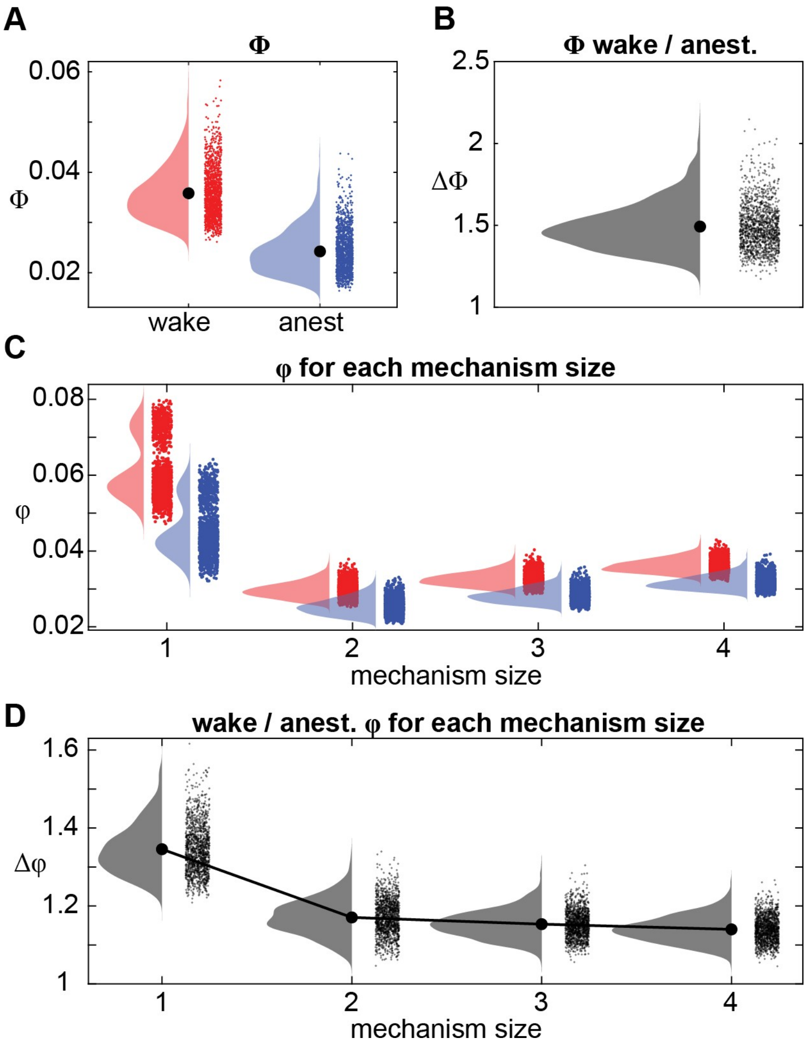

Fig 4. Effect of anesthesia on system-level integrated information (Φ) and the integrated information structure (IIS: a set of φ values). (**A**) Φ values during wakefulness (red) and anesthesia (blue) for each of 1365 channel sets, averaged across flies. (**B**) Ratio of Φ (wakeful / anesthetized), for all channel sets, averaged across flies. (**C**) φ values from the IIS for each mechanism size, for wake (red) and anesthesia (blue). We show the average value for each of 1365 channel sets averaged across flies for each mechanism size. (**D**) Ratio of wakeful φ to anesthetized φ (averaged across flies) for each mechanism size.

https://doi.org/10.1371/journal.pcbi.1008722.g004

Integrated information structure collapses due to general anesthesia

We next investigated integrated information (φ) for each mechanism during wakefulness, and compared them to those during anesthesia. First, we looked at the relationship between mechanism size and integrated information. Since larger mechanisms sample more sources of information, they have a greater capacity for integration, as compared to smaller mechanisms that sample fewer sources of information. Based on this, we reasoned that larger mechanisms will have greater integrated information.

We found integrated information values to significantly vary depending on the size of the mechanism ($\chi^2(3) = 1.512 \times 10^5$, $p < .001$; Fig 4C). Generally, we found that larger mechanisms generated greater integrated information (LME with two levels of mechanism size at a time, see Methods for details: 2-channel << 3-channel: $\beta = -2.941 \times 10^{-3}$, $t(12) = -18.73$, $p < .001$; 3-channel << 4-channel: $\beta = -3.544 \times 10^{-3}$, $t(12) = -28.27$, $p < .001$). However, 1-channel mechanisms by far had the greatest integrated information overall (compared to 4-channel mechanisms: $\beta = 0.025$, $t(12) = 6.49$, $p < .001$). A potential explanation for the large difference in integrated information between 1-channel mechanisms and the other mechanisms is that 1-channel mechanisms are inherently irreducible to smaller parts. We return to this in the Discussion, offering other possible explanations.

Next, looking at the effect of anesthesia, we found integrated information to reduce significantly across all mechanisms with loss of arousal ($\chi^2(1) = 3.092 \times 10^4$, $p < .001$; Fig 4C). We further found a significant interaction between anesthesia and mechanism size ($\chi^2(3) = 1.203 \times 10^4$, $p < .001$), indicating that the extent to which integrated information was reduced due to anesthesia depended on mechanism sizes. We break down this interaction further in the next section.

General anesthesia affects smaller mechanisms more than larger mechanisms

To understand the nature of the significant interaction between anesthesia and mechanism size, we next investigated how the different mechanism sizes were differentially affected by anesthesia. We expected that integrated information for larger mechanisms (consisting of more channels) would be affected more by anesthesia than smaller mechanisms. This is because anesthesia is known to preferentially disrupt global communication [12,36], and so its effect should be reflected more strongly in larger mechanisms involving many channels. To further illustrate, consider two pairs of strongly connected neurons, [AB] and [CD], where there is a very weak connection between the two pairs (i.e., [AB]- -[CD]). In such a case, integrated information for both the 2-channel pairs ([AB] and [CD]) and the 4-channel mechanism ([ABCD]) could be high. If during anesthesia the connections between the pairs are disrupted, then 2-channel integrated information of the individual pairs could remain high while the overall 4-channel integrated information would reduce to zero.

To test if larger mechanisms were more greatly affected by anesthesia, we first analyzed the degree of reduction in integrated information as a function of mechanism size. To account for the variation in integrated information among mechanism sizes, we compared the ratio of wakeful to anesthetized integrated information. A larger ratio corresponds to a larger decrease in integrated information due to anesthesia. We verified that the ratio of wakeful to anesthetized integrated information was also significantly different among mechanism sizes ($\chi^2(3) = 2.229 \times 10^4$, p < .001; Fig 4D). However, instead of finding larger mechanisms to have larger relative reductions in integrated information due to anesthesia, we found the opposite—larger mechanisms had smaller relative reductions ($\beta = 0.229$, t(12) = 3.816, p = .003, $\beta = 0.028$, t(12) = 2.248, p = .044, and $\beta = 0.017$, t(12) = 2.444, p = .031, for comparing 1-, to 2-, 2- to 3-, and 3- to 4-channel mechanisms respectively). Given that the IIS indeed collapsed during anesthesia, we next sought to determine whether larger mechanisms better discriminated conscious level than smaller mechanisms, possibly because of lower variability or noise.

Integrated information structure better distinguishes arousal level than system-level integrated information

Given that integrated information is reduced during anesthesia, we asked if this decrease is more reliable for larger mechanisms. We also sought to determine whether considering the entire IIS allows for better discrimination conscious level than just consideration of single mechanisms, i.e. is the pattern of integrated information useful above and beyond considering independent integrated information values in isolation? As IIT proposes the scalar system-level integrated information value as the measure of conscious level (whereas the multi-dimensional IIS should represent experiential contents), we further compared this to the reliability of the decrease in system-level integrated information. While IIT touts system-level integrated information as a measure of conscious level, we reasoned that, as level of consciousness should generally correlate with the richness of its contents, the IIS would either match or even exceed the classification accuracy of system-level integrated information.

To compare the reliability of decreased integrated information, the collapse of the IIS, and decreased system-level integrated information, we used classification analysis. This allowed us to compare the reliability of one-dimensional changes of integrated information and system-level integrated information with multidimensional changes of the IIS. We used support vector machines (SVMs) to classify the conscious arousal level of individual epochs within each fly (within-fly classification, repeated for each fly; leave-one-paired-epoch-out cross-validation for each channel set; see Methods). To compare integrated information of different sized mechanisms, we averaged accuracies obtained across mechanisms of the same size.

We were able to discriminate wakefulness from anesthesia in the majority of channel sets, using either integrated information values or system-level integrated information (Fig 5A). Further, classification accuracy varied significantly depending on what measure was used (LME testing for main effect of mechanism size (1- to 4-channels), IIS, and system-level integrated information; $\chi^2(5) = 1.300 \times 10^4$, $p < .001$). Overall, 1-channel mechanisms achieved the greatest classification performance, significantly greater than 2-channel mechanisms (β = 0.060, $t(12) = 5.473$, $p < .001$) and 3-channel mechanisms (β = 0.035, $t(12) = 3.945$, $p = .002$), but not 4-channel mechanisms (β = 0.018, $t(12) = 2.033$, $p = .065$). Unexpectedly, integrated information of 1-channel mechanisms also matched that achieved by system-level integrated information, exceeding it slightly but not significantly so (β = 0.008, $t(12) = 0.5843$, $p = .570$). 1-channel mechanisms outperforming other mechanisms is largely consistent with 1-channel mechanisms having the largest relative decrease in integrated information due to anesthesia (Fig 4D). However, 2- and 3-channel mechanisms performed worse than 4-channel mechanisms (β = -0.042, $t(12) = -3.237$, $p = .007$, and β = -0.017, $t(12) = -3.156$, $p = .008$) despite having larger relative decreases in integrated information due to anesthesia, indicating that the reduction in 4-channel mechanisms, while smaller than that for 2- and 3-channel mechanisms, is more reliable. Meanwhile, the full multi-dimensional IIS outperformed integrated information of individual mechanisms and system-level integrated information (β = 0.078, $t(12) = 11.36$ compared to individual 1-channel mechanisms, and β = 0.086, $t(12) = 6.644$ compared to system-level integrated information, $p < .001$), implying that the structure of integrated information may reflect quantity of consciousness better than the simple summary provided by system-level integrated information. To rule out the possibility that the IIS performed better simply because it uses more coefficients to fit the data, we also performed model selection analyses by using logistic regression and comparing Akaike Information Criterion values. Even after penalizing complexity of the model, we still found the IIS to outperform system-level integrated information (S5 Text).

We also tested whether the reductions in integrated information were reliable across flies. This is important because in certain clinical contexts, such as traumatic brain injury, there may be no baseline measurements available, ruling out within-subject assessment. We conducted the decoding analysis, this time repeating leave-one-fly-out cross-validation at each of 8 wake-anesthesia epoch pairs (see Methods). We found that the trend of results for discriminating wakefulness from anesthesia, among mechanism sizes, was similar to within-fly classification, though accuracies and performance differences were overall reduced (Fig 5B). As before, classification accuracy varied depending on what measure was used ($\chi^2(5) = 4451$, $p < .001$). In contrast to the within-fly analysis, we found that (1) the IIS performed similarly to system-level integrated information (β = 0.007, $t(7) = 1.396$, $p = .206$), (2) 1-channel integrated

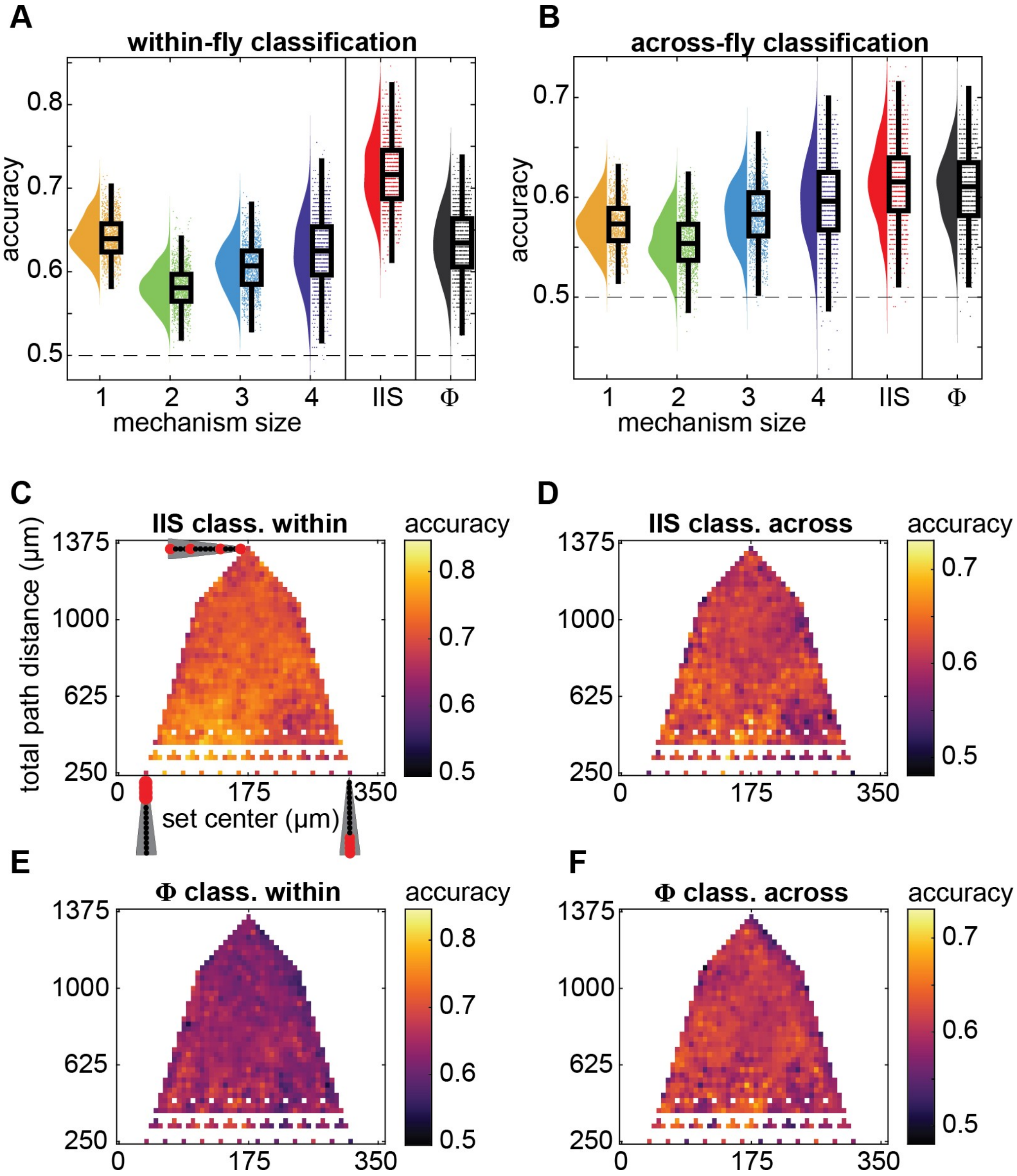

Fig 5. Classification of wakeful vs. anesthetized conditions using mechanism-level φ, system-level Φ, or the integrated information structure (IIS: a set of φ's). (**A**) Within-fly and (**B**) across-fly classification at each individual channel set using individual φ values for each mechanism size (orange, green, pale and dark blue are 1-, 2-, 3-, and 4-channel mechanisms, respectively; single-feature classification), when using the IIS (i.e. all mechanisms together, 15-feature classification, red), and when using Φ (single-feature classification, black). Individual points are classification accuracy of each channel set, after averaging accuracies across all mechanisms within the same mechanism size. Boxplots show median, 25th-75th percentiles, and whiskers are 1.5 interquartile below and above respectively. (**C**-**F**) Spatial map of classification accuracy (same format as in Fig 3). Within-fly (**C**) and across-fly (**D**) classification accuracies when using the IIS. Within-fly (**E**) and across-fly (**F**) classification accuracies when using Φ.

https://doi.org/10.1371/journal.pcbi.1008722.g005

information only outperformed 2-channel integrated information ($\beta = 0.018$, $t(7) = 3.161$, $p = .016$), with (3) 4-channel integrated information achieving greater accuracies than 1-, 2-, and 3-channel integrated information ($\beta = 0.023$, $t(12) = 2.515$, $p = 0.040$, $\beta = 0.042$, $t(7) = 7.976$, $p < .001$, and $\beta = 0.014$, $t(7) = 6.250$, $p < .001$, respectively). As for within-fly classification, system-level integrated information attained similar performance to the highest performing mechanism size, ($\beta = 0.011$, $t(7) = 1.711$, $p = .131$, compared to 4-channel integrated information). The pattern of reduced classification accuracy for smaller mechanisms suggests that the precise location of electrodes, or precise anatomical configuration, may not necessarily have been preserved across flies. Meanwhile, larger mechanisms may be less sensitive to the exact anatomical placement of channels.

Integrated information structure reliably collapses globally across the brain

Finally, we tested if the reliability of using the IIS to distinguish wakefulness from anesthesia depended on spatial features (Fig 5C and 5D). Similar to the trends for the raw system-level integrated information values previously, we found significant trends between classification accuracy and channel set location. Classification accuracy increased as channel sets moved closer to the central brain (with channel 1 being the most central in the brain), for both within- and across-fly classification (Table 1), and slightly decreased as channels became more spaced out, also for both within- and across-fly classification. Thus, while the IIS collapsed throughout the brain, it was most reliable for central regions. These same trends were present for classification when using the system-level integrated information values (Fig 5E and 5F), though the trend of decreasing accuracy with more spaced out channels was not significant. Overall, using the IIS to discriminate level of consciousness in the fly brain yielded better classification accuracies, while maintaining the same spatial pattern of results as system-level integrated information. These results suggest that multidimensional measures may hold greater promise in distinguishing arousal states than more traditional single scalar value summaries of conscious level.

Discussion

In this paper, we applied the measures derived from the Integrated Information Theory (IIT) of consciousness [30], one of the major quantitative theories of consciousness, to the neural recordings obtained from biological brains under two levels of arousal. We demonstrated the construction of integrated information structures (IIS), operationalised based on IIT 3.0, from real neural data to measure level of conscious arousal. We investigated how both system-level integrated information, the primary measure of conscious level put forward by IIT, and these information structures, consisting of a subset of the cause-effect structure (CES) proposed by

Table 1. Dependence of regressands on channel set location and distance among channels.

	Location			Distance		
Regressand	**β**[b]	**χ^2(1)**[c]	***p***	**β**[b]	**χ^2(1)**[c]	***p***
Φ Within	-9.16	44.30	< .001	-2.10	52.11	< .001
Φ Across	-5.76	16.77	< .001	-0.47	2.56	.110
IIS Within	-7.89	32.71	< .001	-2.09	51.05	< .001
IIS Across	-9.06	42.03	< .001	-0.90	9.31	.002

Regressands were the classification accuracies reported in Fig 5C–5F. Φ Within: within-fly classification accuracy using Φ, system-level integrated information. Φ Across: across-fly classification accuracy using Φ. IIS Within: within-fly classification accuracy using IIS, integrated information structure. IIS Across: across-fly classification accuracy using IIS. Location: average location of channels in a channel channel set. Distance: sum of pairwise distances between channels within a channel set.

b β from regressing *z*-scored classification accuracies; values are $\times 10^{-2}$

c The degree of freedom for all likelihood ratio tests was 1 (see Methods).

https://doi.org/10.1371/journal.pcbi.1008722.t001

IIT as corresponding to the structure of consciousness, varied with change in level of arousal in the fly.

To distinguish conscious arousal states of human subjects, previous studies have employed other measures, inspired by IIT [41,8,42,43], on neural data. However, rigorous assessment of IIT ultimately requires assessing its proposed measures, not proxies thereof. So far, empirical testing of IIT has been lacking in this regard. Instead, research has focussed on comparing varying operationalizations of system-level integrated information with regards to theoretical requirements [44,33,37,45,46] or specific network architectures [40,47,48]. Meanwhile, there are relatively few papers on testing system-level integrated information as a measure of consciousness in neural data [49]. Further, to our knowledge only one paper has empirically investigated the notion of information structures, but in the context of the correspondence between the structures and conscious content rather than level of arousal [35].

Consistent with IIT's predictions, we found system-level integrated information to be reduced during anesthesia, and this was accompanied with the collapse of the information structures as reflected by loss of integrated information across all mechanism sizes. Further, we found that the collapse in the information structure during anesthesia was more reliable than the reduction in system-level integrated information, allowing us to classify wake from anesthesia with greater accuracy than using the scalar summary measure. Finally, we found that both the reduction in system-level integrated information and the collapse of the information structures were fairly uniform across all the channel sets which we considered, as was the reliabilities of their respective reduction and collapse. Overall, these results suggest significant recurrent interactions across the whole fly brain, contrary to a general view that the fly brain is largely feedforward, and demonstrate the utility of using information structures to assess level of arousal, over a scalar measure such as system-level integrated information.

Global effect of anesthesia on system-level integrated information and the IIS

An influential view on the fly brain is that they are structured with largely feedforward and unidirectional synaptic connections, with possible exception in the central brain areas which have been identified as centers for integration [24,50,51,10]. From this view, we would expect minimal integrated information for peripheral regions and potentially greater integrated information for more centrally located channel sets, as system-level integrated information by design should be greater for those areas which have stronger recurrent connectivity as a whole (see S7 Text). There is however an emerging view that suggests that fly brains are densely connected in a hierarchical way much like mammalian brains [52]. This latter view implies that fly brains may be equipped with functionally recurrent and feedback computations like those of mammalian brains.

We found system-level integrated information to be slightly greater for more peripherally located channel sets (Fig 3A). Further, we found the decrease in system-level integrated information due to anesthesia, along with the collapse of the IIS to occur throughout the brain, regardless of location of the channel set or distance among channels in the set. Together, these results suggest that feedback interactions occur not only in the central areas of the fly brain, but also in more peripheral, sensory areas as well as across the whole brain. While this is in contrast to the general view of processing in the fly brain periphery being predominantly feedforward [53], feedback connections have indeed been reported in the fly brain (e.g.from the medulla to the lamina, and from the lamina to photoreceptors [54]). The finding of greater system-level integrated information in the peripheral, sensory processing, areas is also consistent with an indirect prediction of IIT that sensory areas are more important for consciousness [47], rather than higher order executive areas.

We acknowledge, however, potential limitations underlying our recordings and analyses. Firstly, it is conceivable that, due to the complexity of numerous brain structures in the centre the brain compared to the relative simplicity of fewer structures in the periphery [52], signals from a mix of many different structures may have cancelled each other out at the raw LFP level. Nonetheless, these central structures may have been more sensitive to the effects of anesthesia. Indeed, we found the effects of anesthesia on system-level integrated information and the IIS to be slightly more reliable for central channel sets (Fig 5C and 5E). Secondly, our method of discretising LFP voltages into binary states may not accurately represent the true space of real states of each of the channels, and also assumes equal probabilities of each state. Further, while IIT 3.0 focuses on moment-by-moment states, other methods, such as considering spectral power in time windows [55] may be more useful in describing the states of the channels, and so expanding IIT's framework to consider frequency domain data potentially is a promising avenue for future research [56]. Thirdly, we note that spurious high-order correlations can be found in partially observed multivariate systems and Markovian approximations of non-Markovian systems. These three limitations can be addressed through further investigation, especially with recordings at higher spatial resolutions than LFP, such as optical imaging or neuropixel probes [57], and expanding of IIT's theoretical framework.

Why are 1-channel mechanisms more affected by anesthesia?

We encountered unexpected results with regards to 1-channel mechanisms. They had larger decreases in integrated information due to anesthesia (Fig 4C and 4D) and higher classification accuracy (Fig 5A) than larger mechanisms composed of 2, 3, or 4 channels. We had suspected that, as integrated information is supposed to measure information which is generated above and beyond separate parts, it would reflect the strength of long-range connectivity, which has been shown to be disrupted by isoflurane anesthesia, in humans, rats, and flies [25,26,36,58]. Given this background, we had expected that larger mechanisms, which are more likely to reflect long range connectivity, would be much more reduced and more reliable in classifying conscious states than 1-channel mechanisms.

We see two ways of interpreting this. First, if we consider 1-channel mechanisms as providing information to the rest of the channel set, then disrupting communication among individual channels inevitably leads to disruption of larger mechanisms. A second interpretation is that the large decrease in integrated information for 1-channel mechanisms may primarily reflect disruption of strong self-connections present during wakefulness, rather than communication with other channels. Having said that, we note that 1-channel integrated information is not a well-developed theoretical construct. In fact 1-channel integrated information isn't clearly defined for earlier versions of IIT and its approximations [28,33,44,45,59]. Specifically, integrated information for a mechanism is assessed by comparing the information it generates before and after imposing some disconnection among its parts. 1-channel mechanisms however cannot be split and compared in this manner. While IIT 3.0 specifically considers disconnections between a mechanism and its purview, and so some disconnection can always be imposed for any mechanism-purview combination, disconnections must still separate the mechanism into independent parts (each affecting their own independent purviews) [60], and thus the problem remains. In Fig 1G, we illustrate that the purviews of mechanisms A and B were simply themselves. In this example, imposing a disconnection on these self-connections seemed to result in a relatively large loss of information (compared to mechanism AB). While further investigation is necessary to understand our finding regarding 1-channel integrated information (e.g. such as 1-channel integrated information being potentially related to autocorrelation; see S8 Text), our main results regarding the IIS are unaffected, as we verified that 1-channel mechanisms were not driving its classification performance (S6 Text).

Role of system-level integrated information

In light of better classification accuracy of wake and anesthesia achieved by the computationally cheaper IIS, one might question the relevance of system-level integrated information in measuring conscious level. However, system-level integrated information plays key roles in IIT other than measuring level of consciousness.

Specifically, system-level integrated information is critical for two key roles. Firstly, it is used for identifying the "complex", the set of parts which maximise system-level integrated information [61–63]. Identification of the complex is critical for determining the boundaries of a system. Once identified, the CES generated by the complex is the "maximally irreducible conceptual structure", which is proposed to directly correspond to contents of consciousness.

Secondly, system-level integrated information is critical in identifying the ideal description of the system across spatial and temporal scales (e.g. individual neurons versus populations of neurons versus LFPs, or ideal sampling rate or time delay τ), with the ideal description corresponding to the physical substrate of consciousness [48,64]. In the same vein, it can be used to identify e.g. the ideal function and/or threshold for binarizing states of the system (though we binarized voltages using the median to ensure equal entropy across channels and all epochs). The ideal description of the system is realised when system-level integrated information is maximal, and the IIS at that description is proposed to correspond to the experience of the system. These uses however require knowledge of all possible system elements, searching across many combinations of system elements, and searching across parameters for operationalising system states. Consequently, a proper, complete search remains infeasible for real neural data. We did however repeat our analyses at two different timescales, finding the same trend of results (see S2 Text).

Differences between perturbation and observation in building the TPM

In order to compute the IIS, we built transition probability matrices (TPMs) and measured the information generated by the system when it is in a particular state. While ideally the TPM should be built by perturbing the system into all states and observing the immediate transition at the next timestep, this is something which is not currently achievable in intact brains. Thus, we built the TPMs by observing the natural, spontaneous evolution of time courses. Natural observation and perturbation can provide the same TPM if a few assumptions are met. First, the correct descriptions of the system must be identified (e.g. at the ideal spatiotemporal scale, and operationalization of states). Second, all states need to be reached during natural observations. When these assumptions are not met, perturbation becomes necessary to obtain a complete TPM.

Perturbation should also be used for setting "background conditions" [30], which is required to distinguish common inputs into system parts from truly integrated parts. Consider an example of two flies, where both flies are stimulated identically. Without taking into account the common stimulation, neural activity in one fly may correspond to and predict neural activity in the other fly, and so system-level integrated information computed from a TPM built from natural observation may not be able to indicate the presence of two separate systems. To avoid this, explicit perturbation (e.g. forcing stimulation to only one fly at a time) can be conducted to separate out the common stimulation.

Applying IIT to loss of arousal in flies

We cannot be sure of the presence of consciousness in flies. Despite this uncertainty, we argue that the research program we are putting forward here is important and meaningful for several reasons. In particular, the fundamental approach of inferring consciousness in animals is to search for behavioural and physiological similarities between ourselves and the animal in question. With sufficiently strong behavioural and physiological similarities, we may at some point consider that the weight of evidence favours attributing consciousness to that particular animal [65,66]. In this context, there is accumulating evidence to suggest that flies indeed have varying levels of consciousness [20,21,10]. There is even some evidence to suggest similar psychological processes in flies as in humans, such as attention [16,21,67], memory [13–15] and feature binding [19]. Further similarities have been found for other insects, such as perception of illusory contours, metacognition, false memory, and long-term planning in bees [68–71].

Assessing the validity of IIT's constructs using recordings from the fly brain provides key advantages, compared to testing in humans. Firstly, using multi-electrode methods provides high quality population neural signals in both time and space, unaccessible with any non-invasive measures available in humans. Further, the small brain of the fly allows us to obtain recordings covering successive layers of visual processing simultaneously, from the retina to the central brain. Secondly, given how the computational cost of computing system-level integrated information and the associated information structures grows exponentially with the number of channels being considered simultaneously [39], the smaller number of neurons in the fly brain, compared to mammalian brains (10^5 compared to 10^8 for mice and 10^{11} for humans [72–74]), provides a system where computing these measures across a large majority of neurons is more feasible. The smaller brain size of the fly has already allowed for detailed imaging of neural circuits across large portions of the fly brain [75,76]. Detailed knowledge of connections among neurons can in the future help inform computation of IIT constructs, e.g. in reducing the set of disconnections to search through when computing integrated information of mechanisms or system-level integrated information. Thirdly, the fly brain is already extensively used as a model of anesthetic loss of consciousness, and various observed molecular mechanisms of anesthesia, such as decreased action potential amplitudes [77,78], and effects on network dynamics such as reduced feedback connectivity [12,26,79], seem to be conserved across species. Further, fly brains appear to share graph-theoretical characteristics with mammalian brains [52] as well as cellular mechanisms [80], and fly LFPs share similarities with human electroencephalographic recordings [81,82]. Taken together, the fly serves as a useful model for investigating the constructs of IIT.

Conclusion and future outlook

Our work opens up several future directions for empirically assessing mathematical approaches to consciousness, especially for IIT. It will be also important to test the generality of our finding across different modulations of consciousness, such as considering graded levels of anesthesia or sleep, as well as across datasets from different systems, such as in the more complex mammalian brain. Even without presuming consciousness for a given system, applying such approaches can inform biology, such as bringing focus to feedback interactions in a system which is largely considered feedforward. While we investigated the use of information structures in determining level of arousal, IIT links these structures more directly to contents of consciousness. As flies can demonstrate complex behaviors such as attentional selection [83] it would be interesting future research to see if the structures of consciousness in flies that can be reasonably inferred from behaviors would correlate with the structures of integrated information as in humans [35].

Methods

Experimental procedure

As the data have been published in [36], here we detail methods directly relevant to the current manuscript. Thirteen female laboratory-reared *Drosophila melanogaster* flies (Canton S wild type, 3–7 days post eclosion) were collected under cold anaesthesia and glued dorsally to a tungsten rod.

Linear silicon probes with 16 electrodes (Neuronexus Technologies) were inserted laterally into the fly's eye. Probes had an electrode site separation of 25 μm. Recordings were made using a Tucker-Davis Technologies multichannel data acquisition system with a 25 kHz sampling rate. Isoflurane was delivered from an evaporator onto the fly through a connected rubber hose. Actual concentration near the fly body was either 0 vol% (awake condition) or 0.6 vol% (isoflurane condition). Flies in the awake condition responded to air puffs by moving their legs and abdomen, but were rendered inert under the isoflurane condition. Importantly, they regained responsiveness when isoflurane was subsequently removed, ensuring that flies were alive during the anesthesia recording.

The experiment consisted of two blocks: one for the 0% isoflurane (air condition, followed by one for the isoflurane condition. Each block started with a series of air puffs, followed by 18 s of rest, 248 s of visual stimuli, another 18 s of rest, and finally a second series of air puffs. Isoflurane was administered immediately after completion of the first block (i.e. after the last air puff), and flies were left for 180 s to adjust to the new concentration before beginning the second block. We used data obtained in the 18 s period between the end of the first series of air puffs and the beginning of the visual stimuli.

Local field potential preprocessing

LFPs were downsampled to 1000 Hz from their original sampling rate of 25 kHz. Downsampled LFPs were bipolar re-referenced by subtracting neighbouring electrodes, resulting in 15 signals which we refer to as "channels". The 18 s of data for each condition was split into 2.25 s segments, giving 8 epochs of 2250 time-samples each. We removed line noise at 50 Hz using the function rmlinesmovingwinc.m function of the Chronux toolbox (http://chronux.org/; [84]) with three tapers, a window size of 0.75 s, and a step size of 0.375 s. Finally, we binarized voltages by taking the median voltage for each channel across all time-samples within a 2.25 s epoch, and then converting each time-sample in the epoch to 'on' if the voltage for that time-sample was greater than the median, and 'off' otherwise (for the effect of binarization threshold, see S1 Text).

IIS computation

Data processing for computing the IIS and system-level integrated information was conducted using Python 3.6.0 in MASSIVE (Multi-modal Australian ScienceS Imaging and Visualisation Environment), a high-performance computing facility. We calculated the measures using PyPhi (version 0.8.1; [39]), publicly available from https://github.com/wmayner/pyphi. Complete details of all the calculations can be found in [30,39].

To compute the IIS, transition probability matrices (TPMs) describing how the set of channels transition from one state to another across time are required. To estimate these, we first select a set of n channels of interest, for which there are 2^n possible states. For each channel in the set, we computed the empirical probability of being "on" at time $t+\tau$ given the state of the system at time t. This gives a $2^n \times n$ matrix (i.e. a "state-by-channel" matrix), which can then be directly fed to the PyPhi toolbox [39]. We use $\tau = 4$ ms as τ which is too small will not capture causal interactions which maximise integrated information, based on known physiology of synaptic interactions [85]. A comprehensive search across τ is infeasible due to computational cost (but see S2 Text for repeated analyses also at $\tau = 2$ ms and 6 ms).

The state-by-channel TPM is used in IIT 3.0, which assumes that there are no instantaneous interactions among the channels (i.e. the "conditional independence" assumption). In other words, the state of some channel being '1' or '0' at some time point is not affected by the state of other channels at the same time point. This assumption is reasonable for classical physical systems, but may not hold when not all units' interactions are considered (e.g. when there is common input to the system). As it is infeasible to obtain a full description of all parts and interactions of intact brains, this is a limitation of the current IIT 3.0 operationalisation of integrated information (note however that the issue is dealt with and resolved for a previous version of IIT by explicitly incorporating conditional dependence among system parts [33,37]).

We computed the state-by-channel TPMs for every possible, 4-channel subset out of the 15 channels (15choose4 = 1365 channel sets), repeating this procedure for each fly and epoch (obtaining one TPM per fly and 2.25s epoch). We selected 4 channels as this gave a reasonable balance between system-level integrated information and the IIS's strength of being a multivariate measure and their weakness of exponentially growing computation cost with system size [39].

To compute the IIS and system-level integrated information for a given set of 4 channels at a given epoch, we submitted its associated transition probabilities to PyPhi. Conceptually, PyPhi finds distances between the probability distribution of transitions specified by the full system with that of the disconnected system (Fig 1G and 1H). As there are 2^n possible states for a set of n-channels (16 states for 4-channels), we computed a set of 15 integrated information values (the IIS) and one system-level integrated information value for every state. Within each epoch, we first computed the within-epoch state-weighted average [40]. For the comparison of integrated information values between wakefulness and anesthesia, we further averaged these values across the 8 epochs.

In Fig 1G and 1H, we explained system-level integrated information (Φ) as the sum of distances between cause and effect probability distributions specified by the full and (minimally) disconnected system (i.e. the full CES and the disconnected CES). Distances for each mechanism are weighted by the mechanism's φ value, as φ is the "earth" which is being moved from the full to the disconnected system (consequently, the distance is weighted by the smaller φ out of the full CES and disconnected CES). Any differences in φ values between the full and disconnected system, such as for mechanism AB in Fig 1G and 1H, are "moved" to the maximally uninformative distributions (EMD^c and EMD^e; red asterisks between Fig 1G and 1H). This weighted summation is depicted in between Fig 1G and 1H, in the solid box.

Classification analysis

To assess the reliability of the effects of anesthesia on the IIS, we conducted classification analysis, which allows us to compare the multivariate IIS (15-features) with single mechanism integrated information (1-feature) and system-level integrated information (1-feature) values. We trained and tested SVMs for each channel set using LIBLINEAR (using default options, i.e. L2-regularized L2-loss support vector classification (dual) [86]) at two levels: a) classifying epochs within each fly (within-fly classification, repeated for each fly), and b) classifying flies at each trial (across-fly classification, repeated at each epoch).

For each measure (integrated information of individual mechanisms, IIS, or system-level integrated information), we conducted nested leave-one-pair-out cross-validation [87,88]. At each outer validation, we conducted an inner-cross-validation procedure on 7 epoch-pairs (within-fly classification; an epoch-pair consists of one wakeful and one anesthetized epoch) / 12 fly-pairs (across-fly classification; a fly-pair consists of one epoch of each wakefulness and anesthesia from the same fly), where we trained SVMs on 6 epoch-pairs / 11 fly-pairs at a time, and validated performance on the remaining epoch-pair / fly-pair. Training features (integrated information values or system-level integrated information values) were each z-scored before training, and testing features were z-scored using the mean and standard deviation of the training set. This was repeated at different cost hyperparameters (2^{-50} to 2^{50}, in steps of powers of 10).

We then trained a SVM on all 7 epoch-pairs / 12 fly-pairs used in the inner-cross-validation, repeating the z-scoring procedure, at the cost hyperparameter value which gave the greatest validation performance (in cases of tie conditions, we took the lower cost value), and tested the overall classifier on the remaining epoch-pair / fly-pair. For the majority of validations (~74% for within-fly classification), the lowest cost of 2^{-50} was selected as the cost value. This process was repeated for each fly / epoch (within-fly classification / across-fly classification), and we averaged across repeats to obtain a final classification accuracy for the channel set and measure. For accuracy of mechanisms with a given size, we report averaged accuracies across all mechanisms with the given size (e.g., we report 1-channel mechanism accuracy as the average accuracy across all 1-channel mechanisms).

Statistical analyses

We used linear mixed effects analysis (LME; [89,90]) to test for significant differences. LME allows us to account for within-fly correlations among channel sets and avoid averaging across either channel sets or flies. Thus we always included random intercepts for fly and the interaction between fly and channel set as random effects, unless otherwise specified. To test for statistical significance of an effect, we employed likelihood ratio tests, where we compared the log-likelihood of the full model with a model with the effect of interest removed. As the likelihood ratio statistic is χ^2 distributed when one model is nested in another, we report the likelihood ratio statistic with the associated degrees of freedom ($\chi^2(d.o.f.)$) corresponding to the difference in number of coefficients between the full model with the model with the effect of interest removed, as well as the corresponding p-value. To conduct pairwise comparisons (e.g. to compare 1-channel to 2-channel integrated information), we limited the effect of interest to two levels at a time and report the associated regression coefficient. As p-values associated with these regression coefficients were very small and potentially do not reflect the true degrees of freedom, we report the coefficients along with "classical" group-level t-tests (conducted after averaging across channel sets to obtain a single value per fly or, for across-fly classification, per epoch).

We first employed LME to compare system-level integrated information, Φ, between wakefulness and anesthesia, using the following model (in Wilkinson notation [90]):

$$\Phi \sim condition + (1|fly) + (1|fly : set) \tag{1}$$

Where condition is level of conscious arousal (wake or anesthesia; dummy coded to be treated as a categorical variable), fly is individual flies (treated as a nominal variable), and set is channel set (treated as a nominal variable). In Table 2, we summarize the amount of variance explained in each model as well as the intercepts for random effect.

To test for a relationship between system-level integrated information (Φ) values and channel set location or total path distance among channels, we regressed system-level integrated information values onto channel set location and distance among channels:

$$\Phi \sim condition + location + distance + location : condition + distance : condition + (1|fly) + (1|fly : set) \tag{2}$$

Table 2. Linear mixed effects model fit (adjusted R^2) and standard deviation (*SD*) of random effects.

	R^2	*SD*		
Random effect		**+ (1|*f*)#**	**+ (1|*f*:*n*)^**	**+ (1|*n*)&**
$\Phi \sim c$	.489	0.011	3.185×10^{-11}	
$\Phi \sim c + l + d + l{:}c + d{:}c$	.493	0.011	1.524×10^{-11}	
$\varphi \sim c + s + c{:}s$	.412	5.95×10^{-3}	4.07×10^{-3}	
$\Delta\varphi \sim s$	.372	0.235	0.121	
$a_W \sim F$	.476	0.019		
$a_A \sim F$	.309	0.020		
$a_{\Phi W} \sim \mathrm{l} + \mathrm{d}$	.562			0.683
$a_{\Phi A} \sim \mathrm{l} + \mathrm{d}$	.513			0.702
$a_{\varphi W} \sim \mathrm{l} + \mathrm{d}$	.555			0.686
$a_{\varphi A} \sim \mathrm{l} + \mathrm{d}$	.535			0.694

Model specifications are described in detail in Methods. Φ: system-level integrated information. *c*: level of arousal (wake/anesthesia). *l*: channel set location. *d*: sum of pairwise distances between channels within a channel set. φ: (mechanism-level) integrated information. *s*: mechanism size. Δφ: ratio of wakeful to anesthetized integrated information for mechanism-level integrated information. a_W: within-fly classification accuracy. *F*: feature used for classification (categorical variable; individual 1-, 2-, 3-, 4-channel mechanisms, 1 feature; IIS, 15 features; or system-level integrated information, 1 feature). a_A: across-fly classification accuracy. $a_{\Phi W}$: within-fly classification accuracy using system-level integrated information. $a_{\Phi A}$: across-fly classification using system-level integrated information. $a_{\varphi W}$: within-fly classification accuracy using the IIS. $a_{\varphi A}$: across-fly classification accuracy using the IIS.

# Random intercept for effect of fly.

^ Random intercept for interaction between fly and channel set.

& Random intercept for channel set.

https://doi.org/10.1371/journal.pcbi.1008722.t002

Where "location:condition" and "distance:condition" denote interaction terms between channel set location and condition, and distance among channels and condition, respectively. We describe the relationship between Φ and channel set location or distance among channels by reporting regression coefficients from *z*-scored Φ values in addition to the significance of the effect of location from the likelihood ratio test.

To compare integrated information (φ) values of the IIS between wakefulness and anesthesia and among mechanism orders, we used the model:

$$\varphi \sim condition + size + condition : size + (1|fly) + (1|fly, set) \tag{3}$$

Where size is mechanism size (1, 2, 3, or 4, dummy coded to be treated as a categorical variable). The number of observations among mechanism sizes differed due to each order having a different number of possible mechanisms (4, 6, 4, and 1, respectively for 1-, 2-, 3-, and 4-channel mechanisms). The term "condition:size" denotes an interaction between level of conscious arousal and mechanism size.

To compare the differential effects of anesthesia among mechanism sizes (breaking down the significant interaction between condition and size in the previous LME), we used the model:

$$\Delta\varphi \sim size + (1|fly) + (1|fly : set) \tag{4}$$

Where Δφ is the ratio of wakeful to anesthetized integrated information.

When comparing classification accuracy across flies across the different feature types (i.e. 1-, 2-, 3-, and 4-channel φ, the IIS, and Φ), classification accuracy was not nested within fly, thus we only included random intercepts for each channel set:

$$accuracy \sim feature + (1|set) \quad (5)$$

Where feature was dummy coded to be one of 1-, 2-, 3-, or 4-channel φ, the full IIS, or Φ.

To test for a relationship between classification performance and channel set location or distance among channels, we regressed accuracies onto the two spatial features:

$$accuracy \sim location + distance + (1|set) \quad (6)$$

Where accuracy is classification accuracies, averaged across flies or epochs (for within-fly and across-fly classification, respectively). As for the relationship between Φ and the spatial features, we describe the relationship between accuracies and the spatial features by reporting regression coefficients on *z*-scored accuracies in addition to the significance of the effect of location from the likelihood ratio test.

Supporting information

S1 Text. Effect of anesthesia is consistent at different binarization thresholds.
(PDF)

S2 Text. Effect of anesthesia is consistent using different timesteps.
(PDF)

S3 Text. "Disconnection" through statistical noising.
(PDF)

S4 Text. Effect of anesthesia on system-level integrated information for each fly.
(PDF)

S5 Text. IIS best predicts wakeful vs. anesthesia states.
(PDF)

S6 Text. 1-channel mechanisms do not drive classification performance of the IIS.
(PDF)

S7 Text. Recurrent activity is required for greater system-level integrated information.
(PDF)

S8 Text. Relation between 1-channel mechanisms and autocorrelation.
(PDF)

Acknowledgments

We would like to thank Andrew Haun, Leonardo Barbosa, and William Mayner for their comments on our manuscript. This work was supported by computational resources provided by the Australian Government through MASSIVE under the National Computational Merit Allocation Scheme.

Author Contributions

Conceptualization: Angus Leung, Naotsugu Tsuchiya.

Data curation: Dror Cohen.

Formal analysis: Angus Leung.

Funding acquisition: Bruno van Swinderen, Naotsugu Tsuchiya.

Investigation: Angus Leung, Dror Cohen.

Methodology: Dror Cohen, Bruno van Swinderen, Naotsugu Tsuchiya.

Project administration: Naotsugu Tsuchiya.

Resources: Dror Cohen, Bruno van Swinderen.

Software: Angus Leung.

Supervision: Naotsugu Tsuchiya.

Validation: Angus Leung.

Visualization: Angus Leung.

Writing – original draft: Angus Leung.

Writing – review & editing: Angus Leung, Dror Cohen, Bruno van Swinderen, Naotsugu Tsuchiya.

References

1. Nagel T. What is it like to be a bat? Philos Rev. 1974; 83: 435–450. https://doi.org/10.2307/2183914
2. Chalmers DJ. The conscious mind: In search of a fundamental theory. Oxford University Press; 1996.
3. Dehaene S. Consciousness and the brain: Deciphering how the brain codes our thoughts. Penguin; 2014.
4. Lamme V. The crack of dawn: perceptual functions and neural mechanisms that mark the transition from unconscious processing to conscious vision. Open MIND. Frankfurt am Main: MIND Group; 2014.
5. Koch C, Massimini M, Boly M, Tononi G. Neural correlates of consciousness: progress and problems. Nat Rev Neurosci. 2016; 17: 307–321. https://doi.org/10.1038/nrn.2016.22 PMID: 27094080
6. Laureys S, Gosseries O, Tononi G. The neurology of consciousness: cognitive neuroscience and neuropathology. Academic Press; 2015.
7. Yanagawa T, Chao ZC, Hasegawa N, Fujii N. Large-scale information flow in conscious and unconscious states: an ECoG study in monkeys. PloS One. 2013; 8: e80845. https://doi.org/10.1371/journal.pone.0080845 PMID: 24260491
8. Hudetz AG, Mashour GA. Disconnecting consciousness: is there a common anesthetic end-point? Anesth Analg. 2016; 123: 1228–1240. https://doi.org/10.1213/ANE.0000000000001353 PMID: 27331780
9. Shaw PJ, Cirelli C, Greenspan RJ, Tononi G. Correlates of sleep and waking in Drosophila melanogaster. Science. 2000; 287: 1834–1837. https://doi.org/10.1126/science.287.5459.1834 PMID: 10710313
10. Barron AB, Klein C. What insects can tell us about the origins of consciousness. Proc Natl Acad Sci. 2016; 113: 4900–4908. https://doi.org/10.1073/pnas.1520084113 PMID: 27091981
11. Zalucki O, Van Swinderen B. What is unconsciousness in a fly or a worm? A review of general anesthesia in different animal models. Conscious Cogn. 2016; 44: 72–88. https://doi.org/10.1016/j.concog.2016.06.017 PMID: 27366985
12. Cohen D, van Swinderen B, Tsuchiya N. Isoflurane impairs low frequency feedback but leaves high frequency feedforward connectivity intact in the fly brain. eNeuro. 2018; ENEURO.0329-17.2018. https://doi.org/10.1523/ENEURO.0329-17.2018 PMID: 29541686
13. Greenspan RJ, Van Swinderen B. Cognitive consonance: complex brain functions in the fruit fly and its relatives. Trends Neurosci. 2004; 27: 707–711. https://doi.org/10.1016/j.tins.2004.10.002 PMID: 15541510
14. Neuser K, Triphan T, Mronz M, Poeck B, Strauss R. Analysis of a spatial orientation memory in Drosophila. Nature. 2008; 453: 1244–1247. https://doi.org/10.1038/nature07003 PMID: 18509336
15. Lewis SA, Negelspach DC, Kaladchibachi S, Cowen SL, Fernandez F. Spontaneous alternation: a potential gateway to spatial working memory in Drosophila. Neurobiol Learn Mem. 2017; 142: 230–235. https://doi.org/10.1016/j.nlm.2017.05.013 PMID: 28559169

16. van Swinderen B. Attention in Drosophila. In: Atkinson N, editor. International Review of Neurobiology. Academic Press; 2011. pp. 51–85. https://doi.org/10.1016/B978-0-12-387003-2.00003-3 PMID: 21906536

17. Koenig S, Wolf R, Heisenberg M. Vision in flies: measuring the attention span. PLOS ONE. 2016; 11: e0148208. https://doi.org/10.1371/journal.pone.0148208 PMID: 26848852

18. de Bivort BL, van Swinderen B. Evidence for selective attention in the insect brain. Curr Opin Insect Sci. 2016; 15: 9–15. https://doi.org/10.1016/j.cois.2016.02.007 PMID: 27436727

19. Grabowska M, Jeans R, Steeves J, van Swinderen B. Oscillations in the central brain of Drosophila are phase locked to attended visual features. Proc Natl Acad Sci U S A. 2020;Forthcoming. https://doi.org/10.1073/pnas.2010749117 PMID: 33177231

20. Cirelli C, Bushey D. Sleep and wakefulness in Drosophila melanogaster. Ann N Y Acad Sci. 2008; 1129: 323–329. https://doi.org/10.1196/annals.1417.017 PMID: 18591491

21. Kirszenblat L, van Swinderen B. The yin and yang of sleep and attention. Trends Neurosci. 2015; 38: 776–786. https://doi.org/10.1016/j.tins.2015.10.001 PMID: 26602764

22. Reichardt W, Poggio T. Visual control of orientation behaviour in the fly: Part I. A quantitative analysis. Q Rev Biophys. 1976; 9: 311–375. https://doi.org/10.1017/s0033583500002523 PMID: 790441

23. Poggio T, Reichardt W. Visual control of orientation behaviour in the fly: Part II. Towards the underlying neural interactions. Q Rev Biophys. 1976; 9: 377–438. https://doi.org/10.1017/s0033583500002535 PMID: 790442

24. Farris SM. Evolution of insect mushroom bodies: old clues, new insights. Arthropod Struct Dev. 2005; 34: 211–234. https://doi.org/10.1016/j.asd.2005.01.008

25. Alkire MT, Hudetz AG, Tononi G. Consciousness and anesthesia. Science. 2008; 322: 876–880. https://doi.org/10.1126/science.1149213 PMID: 18988836

26. Lee U, Ku S, Noh G, Baek S, Choi B, Mashour GA. Disruption of frontal-parietal communication by ketamine, propofol, and sevoflurane. Anesthesiology. 2013; 118: 1264–1275. https://doi.org/10.1097/ALN.0b013e31829103f5 PMID: 23695090

27. Ranft A, Golkowski D, Kiel T, Riedl V, Kohl P, Rohrer G, et al. Neural correlates of sevoflurane-induced unconsciousness identified by simultaneous functional magnetic resonance imaging and electroencephalography. Anesthesiology. 2016; 125: 861–872. https://doi.org/10.1097/ALN.0000000000001322 PMID: 27617689

28. Tononi G. An information integration theory of consciousness. BMC Neurosci. 2004; 5: 42. https://doi.org/10.1186/1471-2202-5-42 PMID: 15522121

29. Tononi G. Consciousness as integrated information: a provisional manifesto. Biol Bull. 2008; 215: 216–242. https://doi.org/10.2307/25470707 PMID: 19098144

30. Oizumi M, Albantakis L, Tononi G. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. PLoS Comput Biol. 2014; 10: e1003588. https://doi.org/10.1371/journal.pcbi.1003588 PMID: 24811198

31. Shannon CE. A mathematical theory of communication. Bell Syst Tech J. 1948; 27: 379–423.

32. Bateson G. Steps to an ecology of mind: collected essays in anthropology, psychiatry, evolution, and epistemology. University of Chicago Press; 1972.

33. Oizumi M, Amari S, Yanagawa T, Fujii N, Tsuchiya N. Measuring integrated information from the decoding perspective. PLoS Comput Biol. 2016; 12: e1004654. https://doi.org/10.1371/journal.pcbi.1004654 PMID: 26796119

34. Tononi G, Boly M, Massimini M, Koch C. Integrated information theory: from consciousness to its physical substrate. Nat Rev Neurosci. 2016; 17: 450–461. https://doi.org/10.1038/nrn.2016.44 PMID: 27225071

35. Haun AM, Oizumi M, Kovach CK, Kawasaki H, Oya H, Howard MA, et al. Conscious perception as integrated information patterns in human electrocorticography. eNeuro. 2017; 4: ENEURO.0085-17.2017. https://doi.org/10.1523/ENEURO.0085-17.2017 PMID: 29085895

36. Cohen D, Zalucki OH, van Swinderen B, Tsuchiya N. Local versus global effects of isoflurane anesthesia on visual processing in the fly brain. eneuro. 2016; 3: ENEURO.0116-16.2016. https://doi.org/10.1523/ENEURO.0116-16.2016 PMID: 27517084

37. Oizumi M, Tsuchiya N, Amari S. Unified framework for information integration based on information geometry. Proc Natl Acad Sci. 2016; 113: 14817–14822. https://doi.org/10.1073/pnas.1603583113 PMID: 27930289

38. Pearl J. Causality. Cambridge: Cambridge University Press; 2009. https://doi.org/10.1017/CBO9780511803161

39. Mayner WGP, Marshall W, Albantakis L, Findlay G, Marchman R, Tononi G. PyPhi: a toolbox for integrated information theory. PLOS Comput Biol. 2018; 14: e1006343. https://doi.org/10.1371/journal.pcbi.1006343 PMID: 30048445

40. Albantakis L, Hintze A, Koch C, Adami C, Tononi G. Evolution of integrated causal structures in animats exposed to environments of increasing complexity. PLoS Comput Biol. 2014; 10: e1003966. https://doi.org/10.1371/journal.pcbi.1003966 PMID: 25521484

41. Abásolo D, Simons S, da Silva RM, Tononi G, Vyazovskiy VV. Lempel-Ziv complexity of cortical activity during sleep and waking in rats. J Neurophysiol. 2015; 113: 2742–2752. https://doi.org/10.1152/jn.00575.2014 PMID: 25717159

42. Casali AG, Gosseries O, Rosanova M, Boly M, Sarasso S, Casali KR, et al. A theoretically based index of consciousness independent of sensory processing and behavior. Sci Transl Med. 2013; 5: 198ra105. https://doi.org/10.1126/scitranslmed.3006294 PMID: 23946194

43. Sarasso S, Rosanova M, Casali AG, Casarotto S, Fecchio M, Boly M, et al. Quantifying cortical EEG responses to TMS in (un)consciousness. Clin EEG Neurosci. 2014; 45: 40–49. https://doi.org/10.1177/1550059413513723 PMID: 24403317

44. Barrett AB, Seth AK. Practical measures of integrated information for time-series data. PLoS Comput Biol. 2011; 7: e1001052. https://doi.org/10.1371/journal.pcbi.1001052 PMID: 21283779

45. Tegmark M. Improved measures of integrated information. PLoS Comput Biol. 2016; 12: e1005123. https://doi.org/10.1371/journal.pcbi.1005123 PMID: 27870846

46. Sevenius Nilsen A, Juel BE, Marshall W. Evaluating approximations and heuristic measures of integrated information. Entropy. 2019; 21: 525. https://doi.org/10.3390/e21050525 PMID: 33267239

47. Haun A, Tononi G. Why does space feel the way it does? Towards a principled account of spatial experience. Entropy. 2019; 21: 1160. https://doi.org/10.3390/e21121160

48. Marshall W, Albantakis L, Tononi G. Black-boxing and cause-effect power. PLoS Comput Biol. 2018; 14: e1006114. https://doi.org/10.1371/journal.pcbi.1006114 PMID: 29684020

49. Kim H, Hudetz AG, Lee J, Mashour GA, Lee U, the ReCCognition Study Group, et al. Estimating the integrated information measure Phi from high-density electroencephalography during states of consciousness in humans. Front Hum Neurosci. 2018; 12: 42. https://doi.org/10.3389/fnhum.2018.00042 PMID: 29503611

50. Anton S, Evengaard K, Barrozo RB, Anderson P, Skals N. Brief predator sound exposure elicits behavioral and neuronal long-term sensitization in the olfactory system of an insect. Proc Natl Acad Sci. 2011; 108: 3401–3405. https://doi.org/10.1073/pnas.1008840108 PMID: 21300865

51. Pfeiffer K, Homberg U. Organization and functional roles of the central complex in the insect brain. Annu Rev Entomol. 2014; 59: 165–184. https://doi.org/10.1146/annurev-ento-011613-162031 PMID: 24160424

52. Shih C-T, Sporns O, Yuan S-L, Su T-S, Lin Y-J, Chuang C-C, et al. Connectomics-based analysis of information flow in the Drosophila brain. Curr Biol. 2015; 25: 1249–1258. https://doi.org/10.1016/j.cub.2015.03.021 PMID: 25866397

53. Nériec N, Desplan C. Chapter Fourteen—From the eye to the brain: development of the Drosophila visual system. In: Wassarman PM, editor. Current Topics in Developmental Biology. Academic Press; 2016. pp. 247–271. https://doi.org/10.1016/bs.ctdb.2015.11.032 PMID: 26970623

54. Rivera-Alba M, Vitaladevuni SN, Mishchenko Y, Lu Z, Takemura S, Scheffer L, et al. Wiring economy and volume exclusion determine neuronal placement in the Drosophila brain. Curr Biol. 2011; 21: 2000–2005. https://doi.org/10.1016/j.cub.2011.10.022 PMID: 22119527

55. Hudson AE, Calderon DP, Pfaff DW, Proekt A. Recovery of consciousness is mediated by a network of discrete metastable activity states. Proc Natl Acad Sci. 2014; 111: 9283–9288. https://doi.org/10.1073/pnas.1408296111 PMID: 24927558

56. Cohen D, Sasai S, Tsuchiya N, Oizumi M. A general spectral decomposition of causal influences applied to integrated information. J Neurosci Methods. 2020; 330: 108443. https://doi.org/10.1016/j.jneumeth.2019.108443 PMID: 31732159

57. Jun JJ, Steinmetz NA, Siegle JH, Denman DJ, Bauza M, Barbarits B, et al. Fully integrated silicon probes for high-density recording of neural activity. Nature. 2017; 551: 232–236. https://doi.org/10.1038/nature24636 PMID: 29120427

58. Ku S-W, Lee U, Noh G-J, Jun I-G, Mashour GA. Preferential inhibition of frontal-to-parietal feedback connectivity is a neurophysiologic correlate of general anesthesia in surgical patients. PLoS ONE. 2011; 6: e25155. https://doi.org/10.1371/journal.pone.0025155 PMID: 21998638

59. Balduzzi D, Tononi G. Integrated information in discrete dynamical systems: motivation and theoretical framework. PLoS Comput Biol. 2008; 4: e1000091. https://doi.org/10.1371/journal.pcbi.1000091 PMID: 18551165

60. Albantakis L, Marshall W, Hoel E, Tononi G. What caused what? A quantitative account of actual causation using dynamical causal networks. Entropy. 2019; 21: 459. https://doi.org/10.3390/e21050459 PMID: 33267173

61. Hidaka S, Oizumi M. Fast and exact search for the partition with minimal information loss. PLoS ONE. 2018;13. https://doi.org/10.1371/journal.pone.0201126 PMID: 30204751

62. Kitazono J, Kanai R, Oizumi M. Efficient algorithms for searching the minimum information partition in integrated information theory. Entropy. 2018; 20: 173. https://doi.org/10.3390/e20030173 PMID: 33265264

63. Toker D, Sommer FT. Information integration in large brain networks. PLoS Comput Biol. 2019; 15: e1006807. https://doi.org/10.1371/journal.pcbi.1006807 PMID: 30730907

64. Hoel EP, Albantakis L, Marshall W, Tononi G. Can the macro beat the micro? Integrated information across spatiotemporal scales. Neurosci Conscious. 2016; 2016: niw012. https://doi.org/10.1093/nc/niw012 PMID: 30788150

65. Boly M, Seth AK, Wilke M, Ingmundson P, Baars B, Laureys S, et al. Consciousness in humans and non-human animals: recent advances and future directions. Front Psychol. 2013; 4. https://doi.org/10.3389/fpsyg.2013.00625 PMID: 24198791

66. Mashour GA, Alkire MT. Evolution of consciousness: Phylogeny, ontogeny, and emergence from general anesthesia. Proc Natl Acad Sci. 2013; 110: 10357–10364. https://doi.org/10.1073/pnas.1301188110 PMID: 23754370

67. Chittka L, Wilson C. Expanding consciousness. Amer Sci. 2019; 107: 364–369, Nov.

68. Horridge GA, Zhang S, O'Carroll D. Insect perception of illusory contours. Philos Trans R Soc Lond B Biol Sci. 1992; 337: 59–64. https://doi.org/10.1098/rstb.1992.0083

69. Perry CJ, Barron AB. Honey bees selectively avoid difficult choices. Proc Natl Acad Sci. 2013; 110: 19155–19159. https://doi.org/10.1073/pnas.1314571110 PMID: 24191024

70. Hunt KL, Chittka L. Merging of long-term memories in an insect. Curr Biol. 2015; 25: 741–745. https://doi.org/10.1016/j.cub.2015.01.023 PMID: 25728692

71. Gallo V, Chittka L. Cognitive aspects of comb-building in the honeybee? Front Psychol. 2018; 9. https://doi.org/10.3389/fpsyg.2018.00900 PMID: 29951014

72. Alivisatos AP, Chun M, Church GM, Greenspan RJ, Roukes ML, Yuste R. The brain activity map project and the challenge of functional connectomics. Neuron. 2012; 74: 970–974. https://doi.org/10.1016/j.neuron.2012.06.006 PMID: 22726828

73. Herculano-Houzel S, Mota B, Lent R. Cellular scaling rules for rodent brains. Proc Natl Acad Sci. 2006; 103: 12138–12143. https://doi.org/10.1073/pnas.0604911103 PMID: 16880386

74. Herculano-Houzel S. The human brain in numbers: a linearly scaled-up primate brain. Front Hum Neurosci. 2009; 3: 31. https://doi.org/10.3389/neuro.09.031.2009 PMID: 19915731

75. Zheng Z, Lauritzen JS, Perlman E, Robinson CG, Nichols M, Milkie D, et al. A complete electron microscopy volume of the brain of adult Drosophila melanogaster. Cell. 2018; 174: 730–743.e22. https://doi.org/10.1016/j.cell.2018.06.019 PMID: 30033368

76. Xu CS, Januszewski M, Lu Z, Takemura S, Hayworth KJ, Huang G, et al. A connectome of the adult Drosophila central brain. bioRxiv. 2020; 2020.01.21.911859. https://doi.org/10.7554/eLife.57443 PMID: 32880371

77. Wu X-S, Sun J-Y, Evers AS, Crowder M, Wu L-G. Isoflurane inhibits transmitter release and the presynaptic action potential. Anesthesiology. 2004; 100: 663–670. https://doi.org/10.1097/00000542-200403000-00029 PMID: 15108983

78. Sandstrom DJ. Isoflurane depresses glutamate release by reducing neuronal excitability at the Drosophila neuromuscular junction. J Physiol. 2004; 558: 489–502. https://doi.org/10.1113/jphysiol.2004.065748 PMID: 15169847

79. Lee U, Kim S, Noh G-J, Choi B-M, Hwang E, Mashour GA. The directionality and functional organization of frontoparietal connectivity during consciousness and anesthesia in humans. Conscious Cogn. 2009; 18: 1069–1078. https://doi.org/10.1016/j.concog.2009.04.004 PMID: 19443244

80. Littleton JT, Ganetzky B. Ion channels and synaptic organization: analysis of the Drosophila genome. Neuron. 2000; 26: 35–43. https://doi.org/10.1016/s0896-6273(00)81135-6 PMID: 10798390

81. Nitz DA, van Swinderen B, Tononi G, Greenspan RJ. Electrophysiological correlates of rest and activity in Drosophila melanogaster. Curr Biol. 2002; 12: 1934–1940. https://doi.org/10.1016/s0960-9822(02)01300-3 PMID: 12445387

82. Paulk AC, Zhou Y, Stratton P, Liu L, van Swinderen B. Multichannel brain recordings in behaving Drosophila reveal oscillatory activity and local coherence in response to sensory stimulation and circuit activation. J Neurophysiol. 2013; 110: 1703–1721. https://doi.org/10.1152/jn.00414.2013 PMID: 23864378

83. Van Swinderen B. Competing visual flicker reveals attention-like rivalry in the fly brain. Front Integr Neurosci. 2012; 6: 96. https://doi.org/10.3389/fnint.2012.00096 PMID: 23091453

84. Mitra PP, Bokil HS. Observed brain dynamics. Oxford University Press; 2007.

85. Koch C. Biophysics of Computation: Information processing in single neurons. Oxford, New York: Oxford University Press; 2004.

86. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J. LIBLINEAR: a library for large linear classification. J Mach Learn Res. 2008; 9: 1871–1874.

87. Tsamardinos I, Rakhshani A, Lagani V. Performance-estimation properties of cross-validation-based protocols with simultaneous hyper-parameter optimization. Int J Artif Intell Tools. 2015; 24: 1540023.

88. Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. NeuroImage. 2017; 145: 166–179. https://doi.org/10.1016/j.neuroimage.2016.10.038 PMID: 27989847

89. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Softw. 2015; 67: 1–48. https://doi.org/10.18637/jss.v067.i01

90. Harrison XA, Donaldson L, Correa-Cano ME, Evans J, Fisher DN, Goodwin CE, et al. A brief introduction to mixed effects modelling and multi-model inference in ecology. PeerJ. 2018; 6: e4794. https://doi.org/10.7717/peerj.4794 PMID: 29844961

2.2 - Supporting Information

I now provide the supporting information related to the article in Section 2.1, beginning on the next page.

S1 Text. Effect of anesthesia is consistent at different binarization thresholds

In the main text, we computed system-level integrated information and the integrated information structure (IIS) after first binarizing LFPs at each epoch based on the median voltage. While we can potentially operationalise the states of the brain signals in many different ways, binarization at the median is the simplest discretisation process, and also normalizes entropy across all epochs. This is important as it controls for potential changes in entropy levels between wakefulness and anesthesia [1].

To exclude the possibility that our results vary wildly depending on the specific threshold used for binarization, we used different binarization thresholds and computed system-level integrated information for sets of 2 channels at a time. We found the effect of anesthesia to be consistent across thresholds for all flies (Fig S1).

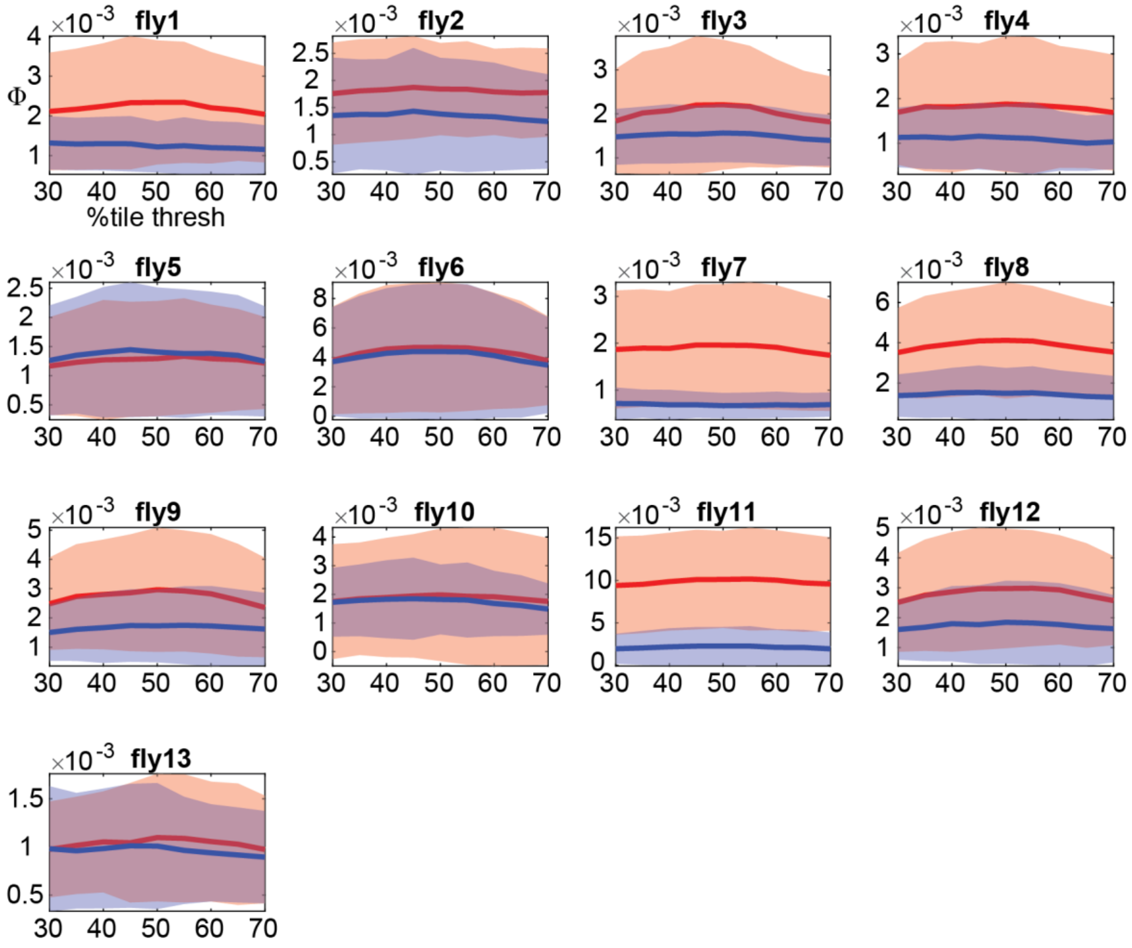

Fig S1. Effect of anesthesia on system-level integrated information is consistent across different binarization thresholds. We computed system-level integrated information (for sets of 2 channels at a time) after binarizing voltages of each channel at a given threshold (30th

up to 70th percentiles in steps of 5; voltages become ‘1’ if above the threshold, and ‘0’ otherwise). Plotted is mean and standard deviation (across 105 channel sets per fly) for wakeful (red) and anesthesia (blue) conditions.

References

1. Wollstadt P, Sellers KK, Rudelt L, Priesemann V, Hutt A, Fröhlich F, et al. Breakdown of local information processing may underlie isoflurane anesthesia effects. PLOS Comput Biol. 2017;13: e1005511. doi:10.1371/journal.pcbi.1005511

S2 Text. Effect of anesthesia is consistent using different timesteps

In the main text, we computed system-level integrated information and integrated information structures (IIS) from TPMs built at the timescale τ = 4 ms. We chose this timescale based on the known physiology of synaptic interactions between neurons. Specifically, if τ is too small, it will not capture causal interactions that maximise integrated information [1–3]. Thus we chose 4 ms as our timescale. A comprehensive search across τ values is infeasible due to the computational cost of system-level integrated information.

To exclude the possibility that our results vary wildly depending on the specific τ value selected, we selected a random sample of 200 channel sets (out of the total 1365 channel sets), and recomputed system-level integrated information and the IIS at two other τ values, 2 ms and 6 ms. As can be seen in Figs S2 and S3, the effect of anesthesia remains consistent with what we report in the main text (Figs 4 and 5).

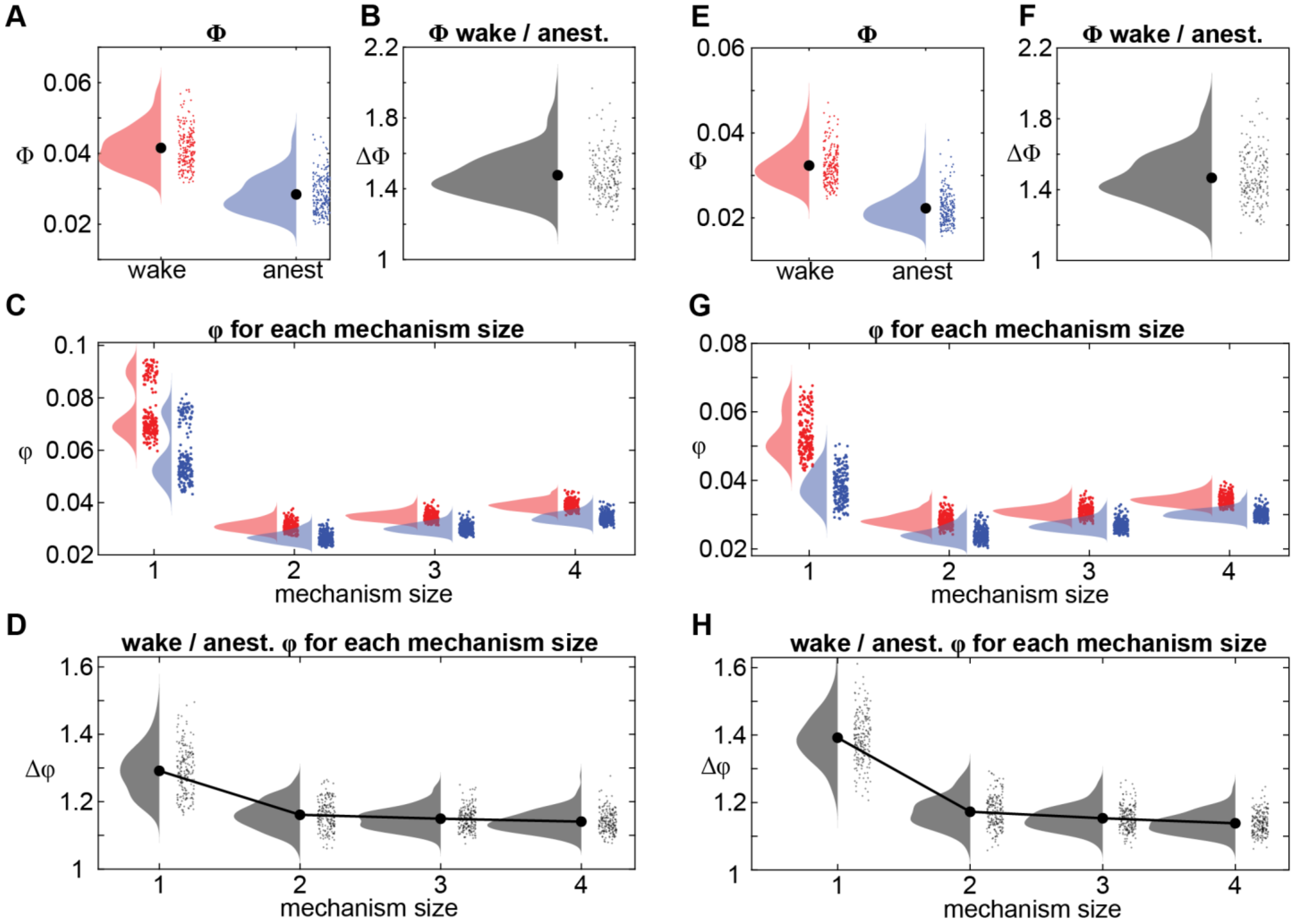

Fig S2. Effect of anesthesia on system-level integrated information (Φ) and the IIS is consistent for similar τ values. Format is the same as Fig 4 in the main article. (**A**-**D**) Results with τ = 2 ms. (**E**-**H**) Results with τ = 6 ms.

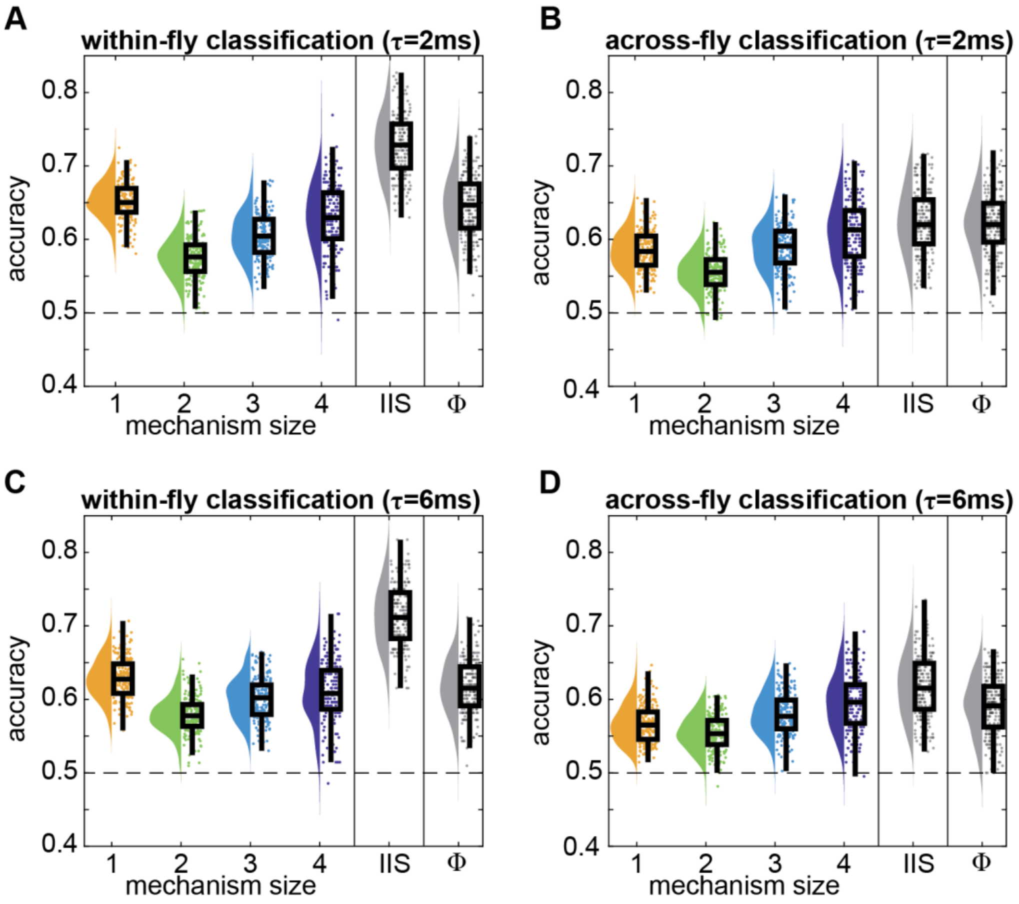

Fig S3. Classification accuracy of wake and anesthesia is consistent for similar τ values. Same format as Fig 5A and 5B in the main text. (**A**-**B**) Results with τ = 2 ms, (**C**-**D**) Results with τ = 6 ms.

References

1. Hoel EP, Albantakis L, Tononi G. Quantifying causal emergence shows that macro can beat micro. Proc Natl Acad Sci. 2013;110: 19790–19795. doi:10.1073/pnas.1314922110

2. Hoel EP, Albantakis L, Marshall W, Tononi G. Can the macro beat the micro? Integrated information across spatiotemporal scales. Neurosci Conscious. 2016;2016: niw012. doi:10.1093/nc/niw012

3. Marshall W, Albantakis L, Tononi G. Black-boxing and cause-effect power. PLoS Comput Biol. 2018;14: e1006114. doi:10.1371/journal.pcbi.1006114

S3 Text. “Disconnection” through statistical noising

In the main text, we assessed the irreducibility of mechanisms (and of the system), by “disconnecting” connections between the mechanism and its purview such that some part of the mechanism affects only some part of the purview (and same for their complements). We carry out this “disconnection” by statistically noising the connections between the mechanism and the purview. Here we provide an example of this procedure in detail.

Consider a mechanism consisting of two channels, A and B, and a purview, consisting of channel C. For simplicity, we consider a case where both A and B have to be ‘1’ simultaneously to make C take the state ‘1’ at the next time step. This system’s state-by-channel (i.e., AB-by-C) TPM is shown in Fig S4A.

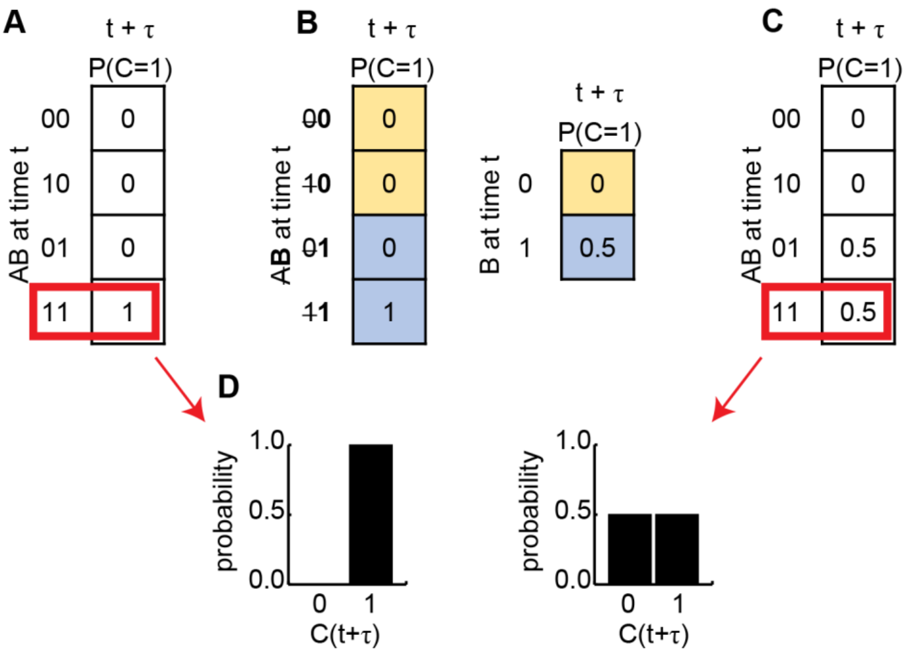

Fig S4. We estimate the effects of “disconnections” through statistical noising of the connection (not through physical disconnection). (**A**) An example state-by-channel transition probability matrix (TPM) for a mechanism AB and purview C. C becomes ‘1’ at time *t*+*τ* if both A and B are ‘1’ at time *t*. (**B**) To “disconnect” A from C, we replace the connection from A with noise by marginalising across the states of A. Colors indicate marginalising within each state of B. (**C**) Expanding the TPM marginalised over states of A returns the TPM to the original space of states of both A and B. (**D**) For the state of AB = 11 in red boxes in panel A and C, we obtain the probability distributions of C, before and after the imposed

disconnection. We compare these distributions (using earth mover's distance) to obtain integrated information.

To assess the irreducible effect of AB on C, we carry out a "disconnection" (i.e. noising). Following the process as illustrated in Fig 1D and 1E, we want to compare the probability distribution of the purview, P(C=1) at $t+\tau$, when both A(t) and B(t) are known, to when only A(t) is known or only B(t) is known. To consider the case when only B(t) is known, we replace A(t) with noise, by marginalising over the possible states of A(t). This gives us a disconnected TPM, as in Fig S4B. If we expand the disconnected TPM to again consider the possible states of B(t) (which now give no information at all about C($t+\tau$), due to the prior marginalisation), we obtain Fig S4C.

Then, for a given state (e.g. AB='11'), IIT 3.0 uses earth mover's distance (EMD) to quantify the distance between the original probability distribution of C($t+\tau$), from the original TPM, and the probability distribution of C($t+\tau$) from the "disconnected" TPM where knowledge of some part of the mechanism has been factored out (in this example, knowledge of B(t); Fig S4D).

S4 Text. Effect of anesthesia on system-level integrated information for each fly

In the main text, we compared the IIS to system-level integrated information, across all flies. Here, we show the effect of anesthesia on system-level integrated information per fly (Fig S5).

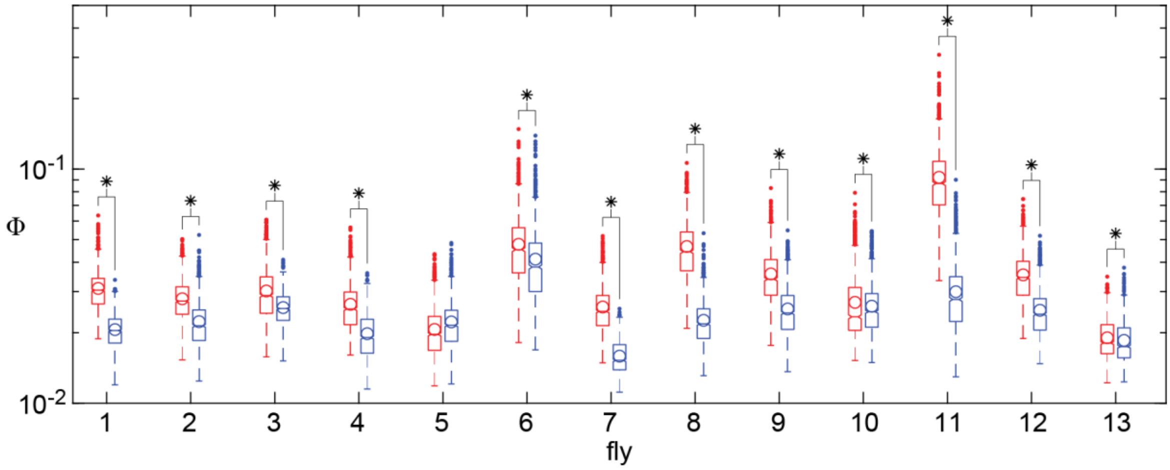

Fig S5. System-level integrated information, Φ (in log scale), during wakefulness (red) and anesthesia (blue) per fly. Boxes indicate 25th, 50th, and 75th percentiles across 1365 channel sets per fly. Circles indicate the mean across channel sets. Asterisks indicate significant one-tailed t-tests (system-level integrated information greater during wakefulness) across channel sets, $p < .001$.

S5 Text. IIS best predicts wakeful vs. anesthesia states

In the main text, to assess the utility of the integrated information structure (IIS) in classifying wakeful vs. anesthetized states, we trained and tested support vector machines (SVMs) on either system-level integrated information (1 feature), the IIS (15 features), or integrated information of individual mechanisms (1 feature). While we found that SVMs trained on the IIS generally outperformed others, this might have been simply due to the IIS having more features to train on.

To exclude such a trivial interpretation, we conducted a complementary analysis with logistic regression, where we systematically compared the goodness of fits among models using an information theoretic model selection procedure (Akaike Information Criterion; AIC; [1]). AIC is defined as:

(S1) $\text{AIC} = -2\mathit{ln}(L) + 2k$

where L is the maximum value of the likelihood function for the model and k is the number of fitted parameters in the model. As the likelihood increases, the first term on the right-hand side becomes smaller, thus a smaller AIC is favoured. However, as a model includes more parameters, it gets penalised by the second term on the right-hand side, as the $2k$ term becomes larger. Thus, given two models with equal likelihoods, AIC selects the model with fewer parameters. Using AIC, we took into account the number of regressors (specifically, models with more regressors are penalised) and compared different model architectures (SII vs. other models with different numbers of mechanisms associated with integrated information).

We used the MATLAB implementation of logistic regression (fitglm.m) to regress a binary level of arousal (either wakeful or anesthetized) onto integrated information values. Specifically, we regressed the arousal level onto either 1) system-level integrated information (giving a single regressor, excluding the intercept), 2) the full IIS, where we used all 15 integrated information values associated with all mechanisms (giving 15 regressors) or 3) integrated information values of one of 1-, 2-, 3-, or 4-channel mechanisms (respectively giving 4, 6, 4, or 1 regressors). As a null model, we also regressed the arousal level onto only an intercept.

To interpret the results in the context of our SVM classification (Fig 5A in main text), we built a model per fly using all 1365 channel sets as observations (2 × 1365 observations per model; Fig S6). The results were consistent with our conclusion with the SVM classification: the IIS performed better than the system-level integrated information even after accounting for the number of available variables fitted.

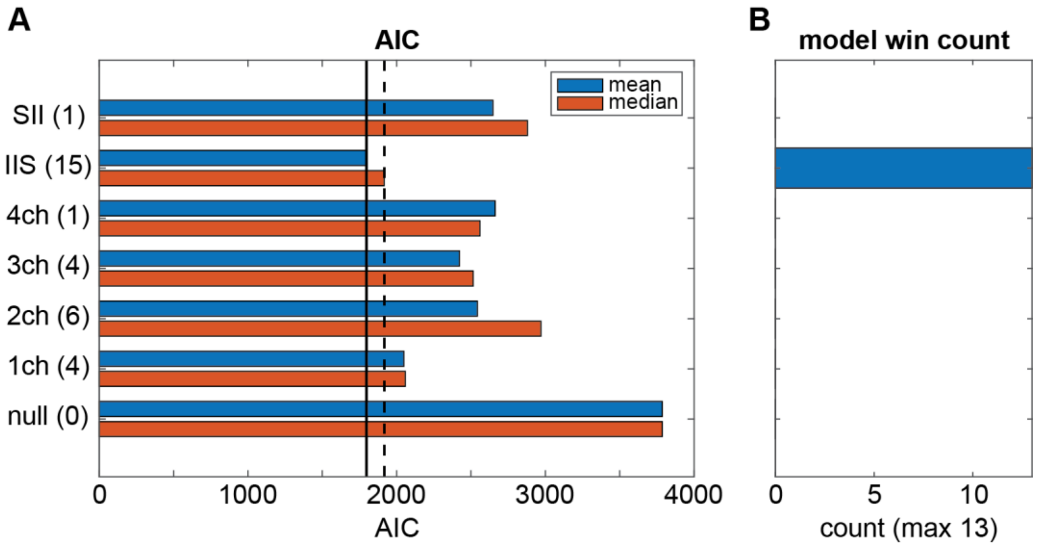

Fig S6. AIC values for logistic regression of the binary level of arousal onto different mechanism sizes, and system-level integrated information. Null models regressed the level of arousal onto only an intercept. (**A**) One AIC value was obtained from fitting models from 1365 channel sets per fly (observations per model = 1365 channel sets × 2 conditions). Shown are mean (blue) and median (red) of the 13 AIC values obtained from each of 13 flies. Solid and dashed lines indicate that the full IIS model performed the best (i.e. gave the smallest AIC) in terms of the mean of the median. Y-labels give the features used for fitting the model and the associated number of coefficients fitted (in parentheses), excluding the intercept. (**B**) The full IIS model was chosen as the best model, which gave the minimal AIC in all 13 flies, while the SII model was never chosen as the best model.

References

1. Burnham KP, Anderson DR. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. 2nd ed. New York: Springer-Verlag; 2002. doi:10.1007/b97636

S6 Text. 1-channel mechanisms do not drive classification performance of the IIS

As shown in Fig 4C, 1-channel mechanisms were associated with a higher magnitude of integrated information compared to other higher-order mechanisms. To quantify the degree of contribution of these 1-channel integrated information for the IIS classification, we repeated the classification analysis without integrated information associated with 1-channel mechanisms (Fig S7), which demonstrated no substantial contribution of 1-channel integrated information.

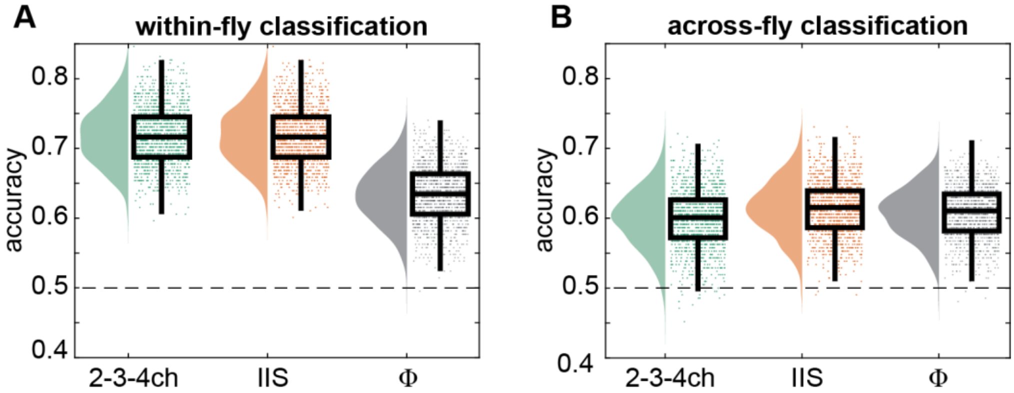

Fig S7. Classification accuracy between wakeful vs. anesthetized conditions using the IIS without 1-channel integrated information (green). (**A**) Within-fly and (**B**) across-fly classification. We replot the same results for the IIS and system-level integrated information for comparison (the same data as in Fig 5A and 5B).

For within-fly classification, the restricted IIS consisting of only 2-, 3-, and 4-channel mechanisms did not achieve significantly different performance to the full IIS ($\chi^2(1) = 0.5691$, $p = 0.451$ using LME model (5), see Methods, where feature had two levels, restricted IIS, i.e. lacking 1-channel mechanisms, or full IIS including all mechanisms; pairwise comparison of restricted IIS to full IIS, $\beta = 9.862 \times 10^{-5}$, $t(7) = -0.658$, $p = .523$; Fig S7A). For across-fly classification, the restricted IIS achieved worse performance than the full IIS ($\chi^2(1) = 306.5$, $p < .001$; $\beta = -0.0155$, $t(7) = -4.401$, $p = .003$). Taken together with the AIC results (S5 Text), we conclude that while 1-channel mechanisms contributed to the IIS, they were not driving its classification performance.

S7 Text. Recurrent connectivity is required for greater system-level integrated information

In the main text, we infer that recurrent connections throughout the fly brain is reduced by general anesthesia based on our observation that integrated information is reduced during anesthesia. A potential concern is that integrated information may be high in a nonlinear system even in the absence of recurrent connections. Here we provide a simulation to demonstrate that recurrent connectivity is required for greater system-level integrated information.

Here, we compare 2-channel integrated information among 10 simulation runs of 3 auto-regressive models with a nonlinear component: 1) a model with 2 channels that are not physically connected, 2) a model with 2 channels where one channel sends output to the other unidirectionally through a physical connection, and 3) a bidirectionally connected model (the model specifications are given below). Given these models, we would expect system-level integrated information to be greater than zero for model 3 and zero for models 1 and 2, as system-level integrated information requires bidirectional connectivity as explained extensively in [1].

The general form of these models is specified as:

- $X_{t+1} = -0.1X_t + AY_t + e_X$
- $Y_{t+1} = -0.1Y_t + BX_t + e_Y$
- Innovations covariance: diagonal 0.5, off-diagonals 0

(1) In the completely disconnected model:

- $A = 0$
- $B = 0$

(2) In the unidirectionally connected model:

- $A = 0$
- $B = 0.9$ if $X_t >$ threshold; 0 otherwise
    - (i.e., X only influences Y if X is above a certain threshold)
- threshold = 0.9

(3) In the bidirectionally connected model:

- $A = 0.9$ if $Y_t >$ threshold; 0 otherwise
- $B = 0.9$ if $X_t >$ threshold; 0 otherwise
- threshold = 0.9

We compute system-level integrated information on the simulated time series in the same way as in the main text: we 1) binarise the simulated time series based on the median, 2) obtain a TPM, then 3) use PyPhi to compute integrated information, which involves several steps as described in the main text (Fig 1 and [2]).

We find that system-level integrated information is, as expected, much greater for the bidirectionally connected model than the other two models (which are much closer to 0; Fig S8). As integrated information is always above or equal to 0, it is positively biased. While here we included a simple nonlinearity in our model (thresholds), further work should be conducted to assess the behaviour of integrated information also in partially observed systems and non-markovian systems approximated through a Markovian assumption, where spurious high-order correlations might affect the measure.

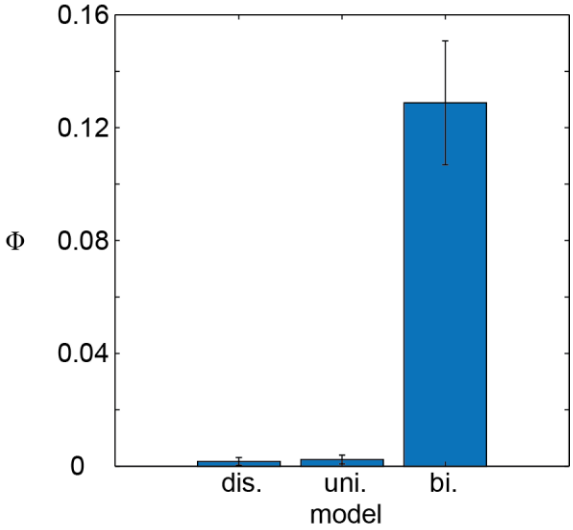

Fig S8. System-level integrated information for three simple nonlinear autoregressive models. System-level integrated information is close to 0 when the system is disconnected or unidirectionally connected. Meanwhile, system-level integrated information is much greater than 0 for the bidirectionally connected system. Shown are mean and standard deviation across 10 simulation runs of each model. For each run, a TPM was built such that each row of the TPM was obtained from observing 200 state transitions. We used these TPMs to compute system-level integrated information in the same way as we describe in the main text.

References

1. Oizumi M, Albantakis L, Tononi G. From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. PLoS Comput Biol. 2014;10: e1003588. doi:10.1371/journal.pcbi.1003588

2. Mayner WGP, Marshall W, Albantakis L, Findlay G, Marchman R, Tononi G. PyPhi: a toolbox for integrated information theory. PLOS Comput Biol. 2018;14: e1006343. doi:10.1371/journal.pcbi.1006343

S8 Text. Relation between 1-channel mechanisms and autocorrelation

In the main text, we consider 1-channel mechanisms to be unclear theoretical constructs of IIT. One possible interpretation is that 1-channel integrated information reflects self-connectivity. Here we quantified the contribution of autocorrelation to 1-channel integrated information.

We directly compared differences (wake minus anesthesia) in 1-channel φ and difference in single-channel autocorrelation (Fig S9). To compute autocorrelation for a given channel, we correlated each LFP time series (of 2.25 s) with itself, shifted by τ = 4 ms (corresponding to τ = 4 ms for our integrated information results). Fig S9A plots autocorrelation values against 1-channel φ values for one fly during wakefulness. Note that each channel only has one autocorrelation value but multiple 1-channel φ values (each from a different set of 4 channels; 14 choose 3 = 364 channel sets containing the channel; error bars in Fig S9A are standard deviations across 364 1-channel φ values). Thus, some fixed autocorrelation value (x-axis of Fig S9A) of a given channel corresponds to multiple, highly varied 1-channel φ values (y-axis). This is expected theoretically, because 1-channel φ has to reflect on how the channel is embedded in and interacts with the other three channels.

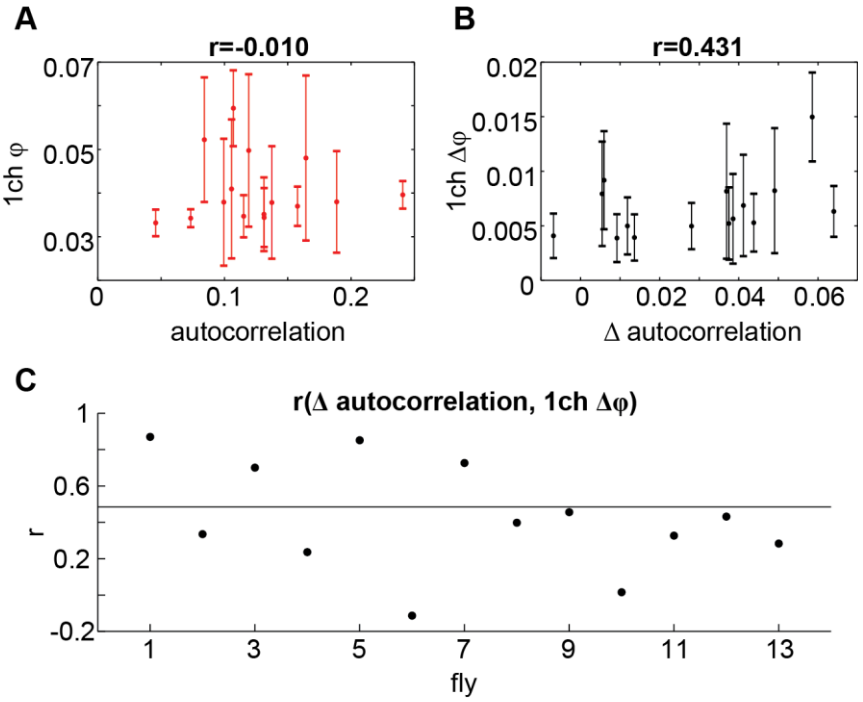

Fig S9. Relationship between 1-channel integrated information and autocorrelation, at τ = 4 ms. (**A**) Single channel autocorrelation plotted against 1-channel integrated information, for a

representative fly during wakefulness. Each point corresponds to 1-channel. Error bars are standard deviations of 1-channel φ for a given channel (each channel is contained in 364 out of all 1365 sets of 4 channels). Title gives the correlation coefficient between autocorrelation and 1-channel φ for the fly. (**B**) Difference (wake - anesthesia) in Fisher *z* transformed single-channel autocorrelation (Δ autocorrelation) plotted against difference in 1-channel integrated information (Δφ), for the same fly. Title gives the correlation coefficient between Δ autocorrelation and Δφ for the fly. (**C**) Correlation coefficients between Δ autocorrelation and Δφ for each individual fly. Solid line indicates the average correlation coefficient across flies (coefficients were averaged after Fisher *z* transform, plotted is inverse transform of the mean).

We next subtracted Fisher *z* transformed autocorrelation values during anesthesia from those during wakefulness (Δ autocorrelation). Fig S9B shows Δ autocorrelation plotted against Δφ (wake φ minus anesthetized φ), for the same fly as Fig S9A. Correlations at each fly, between Δ autocorrelation and average Δφ values of each channel, indicated that there is some positive correlation between the two measures at the group level (Fig S9C). We confirmed this using a one-sample t-test comparing Fisher *z* transformed correlation coefficients to 0 ($M = 0.424$, $SD = 0.443$, $t(12) = 4.308$, $p = .001$).

In sum, while there seems to be some relationship between the two measures, we conclude that 1-channel integrated information reflects something above and beyond its autocorrelation, namely, the informational and (statistical) causal interactions between that channel with the rest of the channels in the considered system. Whether this is an ideal property for integrated information may need further theoretical exploration in the future.

Chapter 3 - Searching for intrinsic timescale

In this chapter, I aim to search for a temporal scale at which Φ is maximised. To address this aim, I apply Φ first to simulated model data, and then to recordings from the fly brain previously described in Chapter 2. For this chapter, I supply a manuscript currently accepted for publication in Entropy. The manuscript begins on the following page.

3.1 - Entropy submission

This manuscript was accepted for publication in Entropy. It begins on the following page.

Title: Emergence of integrated information at macro timescales in real neural recordings

+Leung A[1], +Tsuchiya N[1,2,3]

+ Corresponding authors

1. School of Psychological Sciences and Turner Institute for Brain and Mental Health, Monash University, Melbourne, Victoria, Australia
2. Center for Information and Neural Networks (CiNet), Japan
3. Advanced Telecommunications Research Computational Neuroscience Laboratories, Japan.

Author Contributions: Conceptualization, A.L. and N.T.; methodology, A.L. and N.T.; formal analysis, A.L.; writing-original draft preparation, A.L.; writing-review and editing, A.L. and N.T.; supervision, N.T.; funding acquisition, N.T. All authors have read and agreed to the published version of the manuscript.

Funding: A.L. was supported by an Australian Government Research Training Program (RTP) Scholarship. NT was funded by the National Health and Medical Research Council Ideas Grant GNT1183280, TWCF0199 from Templeton World Charity Foundation, Inc., Australian Research Council Future Fellowship FT120100619 and Discovery Projects DP180104128 and DP180100396, and the National Health and Medical Research Council APP1183280. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Data Availability Statement: Pre-processed fly LFPs are available on Figshare - doi: 10.26180/5ebe420ae8d89. Simulation and data analysis codes are available at https://github.com/Prototype003/phi3_timescale_sim.

Conflicts of Interest: The authors declare no conflict of interest.

Abstract

How a system generates conscious experience remains an elusive question. One approach towards answering this is to consider the information available in the system from the perspective of the system itself. Integrated information theory (IIT) proposes a measure to capture this, integrated information (Φ). While Φ can be computed at any spatiotemporal scale, IIT posits that it be applied at the scale at which the measure is maximised. Importantly, Φ in conscious systems should emerge to be maximal not at the smallest spatiotemporal scale, but at some macro scale where system elements or timesteps are grouped into larger elements or timesteps. Emergence in this sense has been demonstrated in simple example systems composed of logic gates, but it remains unclear whether it occurs in real neural recordings which are generally continuous and noisy. Here we first utilise a computational model to confirm that Φ becomes maximal at the temporal scales underlying its generative mechanisms. Second, we search for emergence in local field potentials from the fly brain recorded during wakefulness and anaesthesia, finding that normalised Φ (wake/anaesthesia), but not raw Φ values, peaks at 5 ms. Lastly, we extend our model to investigate why raw Φ values themselves did not peak. This work extends the application of Φ to simple artificial systems consisting of logic gates towards searching for emergence of a macro spatiotemporal scale in real neural systems.

Introduction

Integrated information theory tackles the question of how physical interactions can support consciousness by introspecting conscious experience [1,2]. It then deduces postulates, the necessary physical interactions to support conscious experience, and from these derives a numerical measure of consciousness which should be high in a conscious system, and low otherwise. We previously applied the measures proposed by integrated information theory (IIT) 3.0 to local field potentials (LFPs) from the fly brain, testing the hypotheses that system-level integrated information Φ and its associated conceptual structure should be reduced during reduced level of consciousness as induced by anaesthesia [3]. As expected from the theory, both Φ and associated conceptual structures computed from the LFPs were indeed reduced during anaesthesia. However, we were unable to apply all of IIT's postulates exactly as they are put forward by the theory. Specifically, we did not fully apply IIT's exclusion postulate, which states that only one set of overlapping sets of elements, the complex, can be conscious.

To identify the complex, IIT's exclusion postulate requires searching across all subsets of system elements, recomputing Φ for each subset. However, this search quickly becomes computationally infeasible for larger numbers of elements, due to the rapidly increasing cost of repeatedly identifying the minimum information partition (MIP; [2]) for all subsets. IIT's exclusion postulate also requires searching for the complex across spatial and temporal scales. As LFPs are an aggregate measure which summate electrical activity arising from neurons' cell bodies, axons, and dendrites at a scale much coarser than that of individual neurons [4], searching for the potential spatial scale of the complex did not seem to be a promising avenue for investigation. However, searching for the temporal scale is a feasible and likely more fruitful endeavour, given the high temporal resolution of LFPs.

IIT provides a clear expectation as to the temporal scale of the conscious complex. Specifically, IIT's exclusion postulate ties the complex to the scale at which our experiences occur. Through introspection, it is apparent that, for humans, an instance of experience occurs most likely at the scale of milliseconds - we are unable to perceive events which occur at too short a timescale, such as events which occur at the scale of microseconds. This intuitive scale is backed empirically by psychophysics studies, with humans being able to discern events at the scale of tens of milliseconds, but not shorter [5]. It is also unlikely to be at longer timescales such as seconds or longer, where we can differentiate multiple instances of experience. Consequently, through its exclusion postulate, IIT predicts that Φ should be maximal at some particular scale, in the order of milliseconds. Conversely, it should be lower, both at micro timescales which are too short and at macro timescales which are too long to correspond to the timescale of conscious experience. While the emergence of maximal Φ at some intermediate scale has been previously illustrated in example binary systems consisting of logic gate elements [6–8], it is unclear whether it occurs in real neural data which is typically continuous in nature.

In this paper, we will test the above prediction using both real neural recordings and related computational generative models. First, by using a toy auto-regressive model, for which temporal interactions among system elements are known a priori [9], we verify that Φ identifies the timescale of system interactions from continuous data generated by the model. Next, we apply Φ to the fly recordings previously analysed in [3], to search for a potential temporal scale of interactions in the complex. However, we find Φ to either increase or decrease in a monotonic fashion with changes in temporal scale both when flies were wakeful and anaesthetised, depending on how the recordings were pre-processed to characterise timescale. Meanwhile, the ratio of wakeful to anaesthetised Φ identifies a potential temporal scale of interactions, again depending on how timescale is characterised. Given these findings, in the last section of this paper, we expand the auto-regressive simulation to explore limitations of our application of Φ to the fly recordings, namely non-Markovianity and partial observation.

Results

Integrated information identifies the timescale of interactions in a nonlinear autoregressive process

Example interactions between system elements leading to maximal Φ occurring not at the most fine-grained scale, but rather at a macro scale, have previously been illustrated in toy systems with binary elements. However, this illustration has not been extended to systems with continuous elements. So, to check the in-principle feasibility of searching for emergence of Φ at a macro scale in continuous data, we first utilised a toy autoregressive model, where the value of each time-sample is determined by values at previous times [10]. We modelled a bivariate, bidirectionally connected system as follows:

$$X(t) = aX(t-l) + bY(t-l) + \varepsilon_X(t)$$
$$Y(t) = cY(t-l) + dX(t-l) + \varepsilon_Y(t)$$

Where *X*(*t*) and *Y*(*t*) are voltages for two system elements (which we refer to as channels) at a given time *t*. *a* and *c* are autoregressive coefficients representing self-connections, while *b* and *d* are autoregressive coefficients representing cross-connections between the two elements *X* and *Y*. We set both *a* and *c* as -0.1, simulating self-inhibition. *l* is the time delay between self- and cross-connections among system elements, which we set to be 10. $\varepsilon_X(t)$ and $\varepsilon_Y(t)$ represent uncorrelated Gaussian noise, with mean 0 and variance both set to 0.5.

As the neural mechanisms underlying LFPs are known to have nonlinear dynamics, we included a nonlinearity in the model. We set the cross-connectivity to be dependent on the voltages of *X* and *Y*:

$$b = \begin{cases} 0.9, & Y(t-l) > threshold \\ 0, & otherwise \end{cases}$$
$$d = \begin{cases} 0.9, & X(t-l) > threshold \\ 0, & otherwise \end{cases}$$

Where *threshold* was set as 0.9. This adds a nonlinear dynamic which simulates reliable neural communication through bursting [11,12]. Note that, overall, the system elements are bidirectionally connected, and only interact with a delay of 10 timesteps. Consequently, we expected Φ to be non-zero for this system (which we previously reported for *l* = 1 in [3], S7 Text), and critically, maximal at the timescale corresponding to 10 timesteps.

To check that Φ does indeed identify this timescale of 10 timesteps, we simulated the model for 10 runs (see Methods). For each run we operationalised the state of each channel at a given time point by binarising it with respect to the median voltage of that channel. Then we constructed a transition probability matrix (TPM) by finding, for each state of the system at time *t*, the empirical probabilities of each channel being in an “on” state at *t* + τ. This is the same method we previously used to compute TPMs for real neural recordings in [3] (see also Discussion in [3] for issues regarding observation versus perturbation in constructing TPMs). From this method, which we hereafter refer to as the “skipping” method, timescale is characterised as the delay τ. We repeatedly computed TPMs for exponentially increasing values of τ (Figure 1B and 1C) and, from these TPMs, computed Φ values at each τ value.

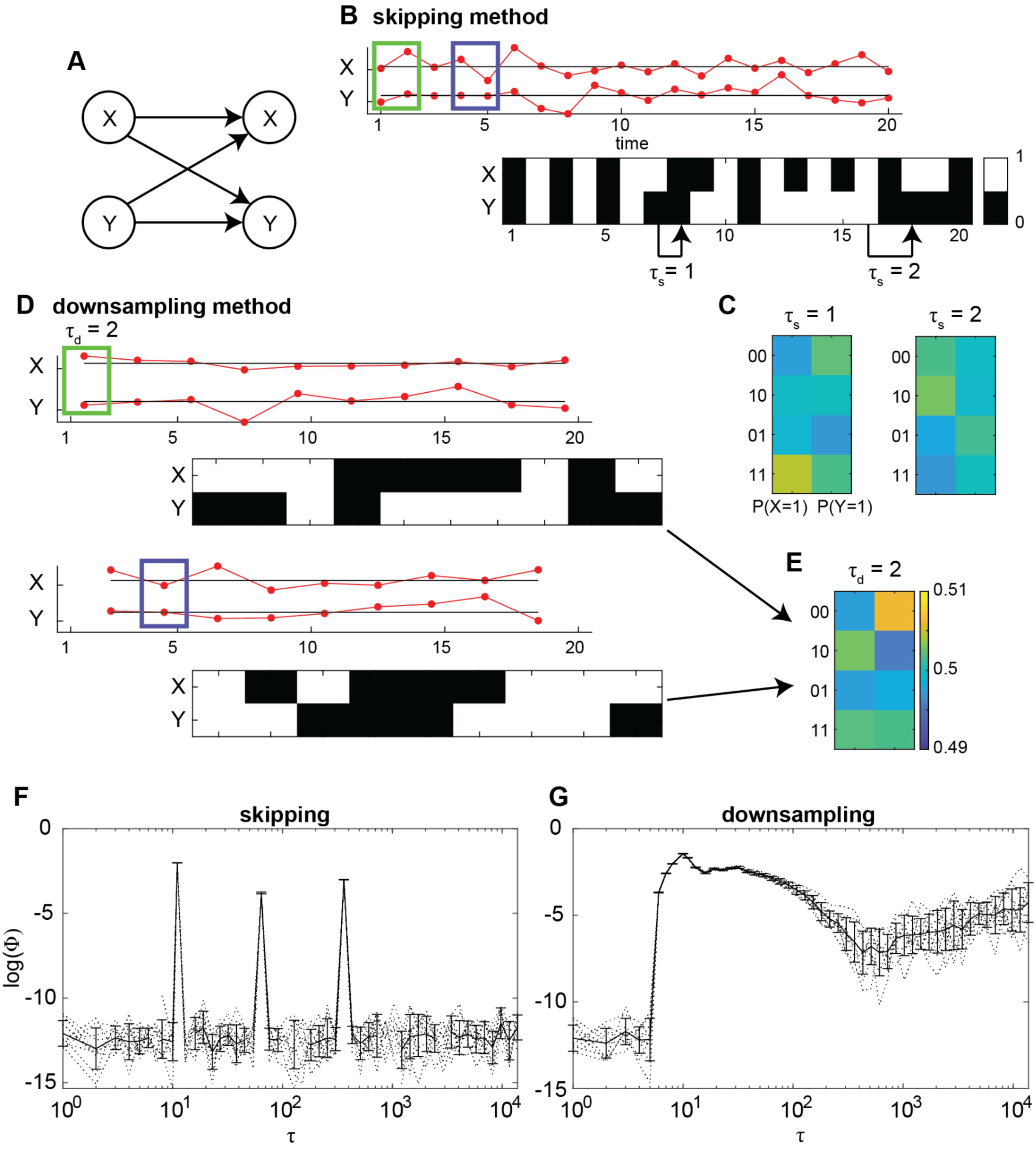

Figure 1. Relationship between integrated information Φ and timescale τ in a system with nonlinearity. (**A**) We generate continuous time-series by modelling a nonlinear, bidirectionally connected system. (**B**) For the skipping method, continuous time-series values (red, top) are discretised into binary states (black/white, bottom) by comparing to the median value for each run. Displayed is an example of 20 samples from 1 run. (**C**) State-by-node transition probability matrices (TPM) are constructed using the skipping method for increasing τ. For each possible system state at time *t* (each row in the TPM), each entry describes the probability a node will take state '1' at time *t* + τ. (**D**) For the downsampling method, τ contiguous time-series values are averaged together to form coarse-grained time-series. Multiple downsampled time-series are obtained by offsetting the time sample from which to begin coarse-graining, from 0 up to τ - 1 samples. Green rectangles indicate the first bin of contiguous time samples, for τ = 2, from the original time-series in **A** which are averaged together, for the first offset (of 0 samples). Blue rectangles indicate the second bin for the second offset. Coarse-grained time samples are then discretised into binary states by comparing the median value for each offset, at each run. (**E**) TPMs are constructed for the downsampling method using all transitions across all offsets. Each entry describes the probability a node will have a coarse-grained state '1' at a coarse-grained time *t* + 1, given the system state at *t*. (**F**) Φ values in relation to τ when using the skipping method. Dotted and solid lines indicate individual simulation runs and the average across runs, respectively. Error bars indicate standard deviation across runs. (**G**) Same as **F**, but for Φ values computed using the downsampling method.

Figure 1F shows the trend of Φ with respect to τ when using the skipping method. While there existed multiple local maxima of Φ (peaks at roughly τ = 10, 60, and 360 timesteps), Φ was, as expected, maximal at τ = 10 timesteps, corresponding to the time delay *l* = 10 in the model.

While the skipping method is consistent with how empirical estimates of Φ from previous versions of IIT have been applied [13–15], simulation papers illustrating maximal Φ at macro temporal scales have utilised different methods [6–8]. Specifically, they utilise coarse-graining or black-boxing, whereby micro timesteps are collated together to form macro timesteps. Following this approach, we characterised timescale in a second way, by averaging voltages in bins of size τ (green and blue rectangles in Figure 1B and 1D). Then, in the same manner as to the skipping method, we operationalised states of each channel by binarising the resulting downsampled voltages based on their medians to construct TPMs for increasing τ (Figure 1E). We refer to this method as the "downsampling" method.

The downsampling method has a notable drawback compared to the skipping method. Specifically, for a given time-series, when τ is increased by some factor, the number of time-samples available for building a TPM is decreased by that factor. For example, doubling τ would result in half the original number of samples in the time-series. Due to the fewer number of samples and thus fewer overall state transitions, empirical transition probabilities in the TPM can rapidly become unreliable as τ increases. To address this, we constructed TPMs from multiple rounds of downsampling, by offsetting the starting time sample of each bin (Figure 1D) before downsampling and then binarising voltages. In this manner a TPM for a given τ was constructed using all transitions from all offsets. Using this method, the number of transitions used to construct a TPM was equal to the number of transitions used in the skipping method.

Figure 1G shows the trend of Φ with respect to τ when using the downsampling method. While Φ seemed to be non-minimal for a larger range of τ when compared to the skipping method, it was, again as expected, maximal at $\tau = 10$, corresponding to the time delay $l = 10$ in the model. These results indicate that, Φ identifies the timescale of interactions among continuous processes both when using the skipping and downsampling methods.

Normalised empirical integrated information identifies a timescale of interactions

We next sought to find some timescale in neural recordings at which Φ is maximised. We utilised 15 local field potentials (LFPs, hereafter referred to also as "channels") recorded from across the brains of 13 fruit flies using a linear multi-electrode array as previously described in [16,3] (see Methods).

As we did previously for the simulation, we operationalised the state of each channel by binarising voltages based on the median voltage for the channel before then constructing a TPM at increasing values of τ (skipping method), as well as by repeatedly binarising voltages based on median voltages after downsampling at increasing values of τ (downsampling method). Given the computational cost of computing Φ, and needing to repeatedly compute

Φ at each τ value, we restricted analysis to 2 channels at a time, treating every pair of channels as a system.

Figure 2 shows the trend of Φ (log transformed) with increasing τ across the flies, for both methods of characterising timescale. On average across all the channel pairs, there was no visual indication of Φ being maximal at a timescale other than the smallest or largest timescales. Rather, Φ trended such that it tended to be larger for smaller timescales when using the skipping method (Figure 2A), and larger for larger timescales when using the downsampling method (Figure 2B).

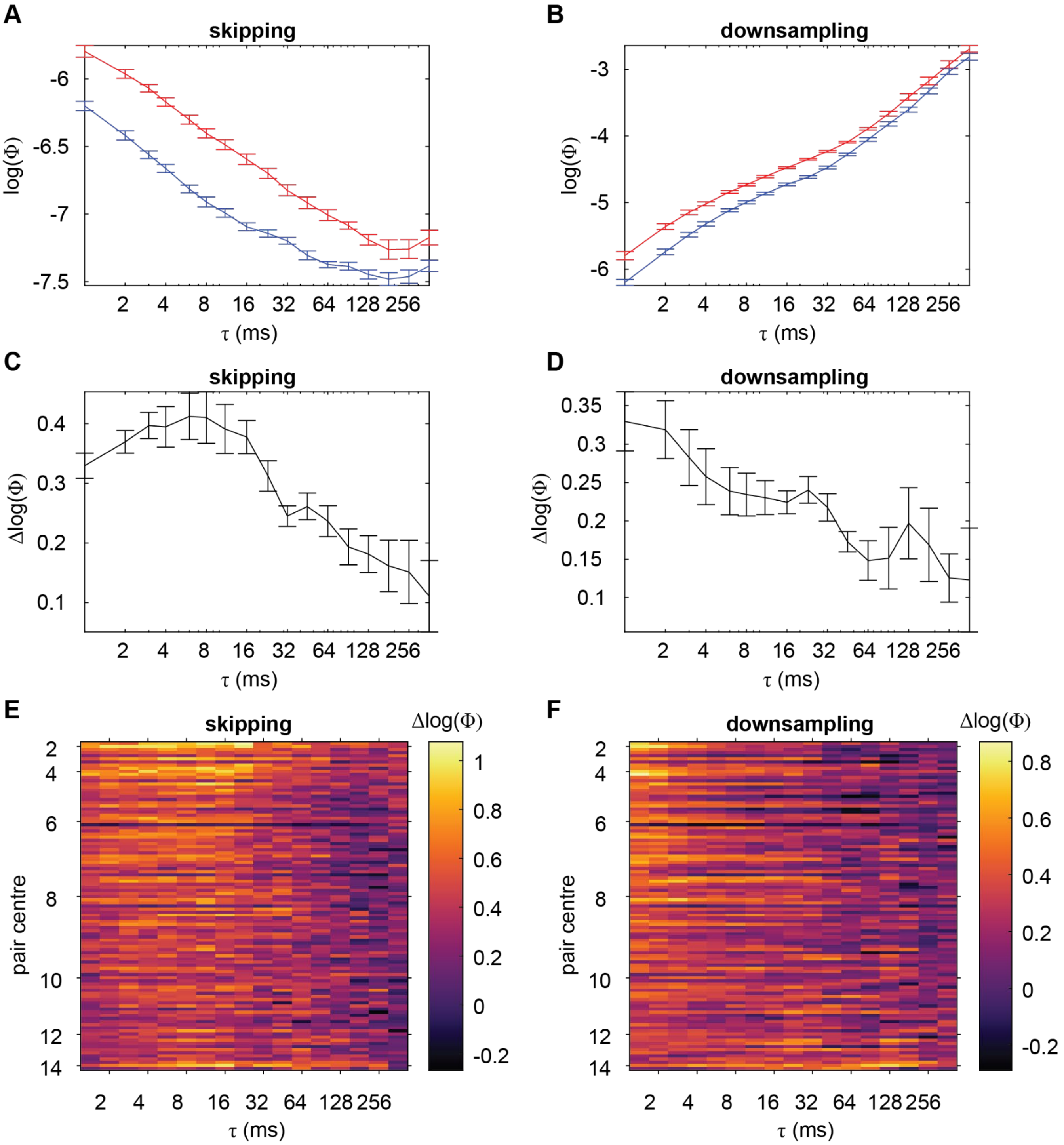
A
skipping
log(Φ)
-6
-6.5
-7
-7.5
2 4 8 16 32 64 128 256
τ (ms)
B
downsampling
log(Φ)
-3
-4
-5
-6
2 4 8 16 32 64 128 256
τ (ms)
C
skipping
Δlog(Φ)
0.4
0.3
0.2
0.1
2 4 8 16 32 64 128 256
τ (ms)
D
downsampling
Δlog(Φ)
0.35
0.3
0.25
0.2
0.15
0.1
2 4 8 16 32 64 128 256
τ (ms)
E
skipping
Δlog(Φ)
pair centre
2
4
6
8
10
12
14
1
0.8
0.6
0.4
0.2
0
-0.2
2 4 8 16 32 64 128 256
τ (ms)
F
downsampling
Δlog(Φ)
pair centre
2
4
6
8
10
12
14
0.8
0.6
0.4
0.2
0
-0.2
2 4 8 16 32 64 128 256
τ (ms)

Figure 2. Relationship between integrated information Φ and timescale τ in fly recordings. (**A**) Log transformed Φ values, averaged across channel pairs and flies, as a function of timescale when using the skipping method. Red and blue are values during wakefulness and anaesthesia, respectively. Error bars indicate within-subject standard error [17,18]. (**B**) Log transformed Φ values, as in **A**, but for when using the downsampling method. (**C-D**) Difference between wakeful and anaesthetised log transformed Φ, Δlog(Φ), as a function of timescale when using the skipping and downsampling methods respectively. (**E-F**) Δlog(Φ) as a function of timescale for each channel pairing when using the skipping and downsampling methods respectively. Channel pairs are sorted by the average position of the channels in the pair (y-axis, larger values indicate pairs which on average are located more in the periphery). Pairs with the same average position are sorted by the distance between the channels, with larger distances being lower in the y-axis. τ (x-axis) increases in an exponential manner.

As a control, we also computed Φ for the channel pairs when the flies were anaesthetised. We reasoned that, during loss of consciousness, Φ should not have a clear maximum at some timescale. Rather, assuming that there is no consciousness under anaesthesia, it should be minimal at all timescales. Any variations in Φ across timescale should correspond not to a potential complex of consciousness, but instead to other things such as background neural activity which does not support consciousness (or supports some minimal consciousness) or issues regarding empirical observations of state transitions which are used to build the TPM (which we expand on in the Discussion). Blue lines in Figure 2A and 2B show the trend of Φ with increasing τ during anaesthesia respectively when using the skipping and downsampling methods. While the magnitude of Φ tended to be overall reduced across all τ when compared to wakefulness, consistent with our previous results [3], the trends of Φ with respect to τ, for both the skipping and downsampling methods, appeared to be the same as for wakefulness.

Given that the trends of Φ with respect to τ during anaesthesia was similar to during wakefulness, we considered that the trends during wakefulness could also be reflecting issues of empirical observation of TPMs. Meanwhile, any trend of Φ related to the timescale of interactions underlying the complex could be masked by these trends. To address this, we considered using Φ values during anaesthesia as a baseline. Specifically, we investigated how the difference (wake minus anaesthesia) in log transformed Φ values (Δlog(Φ); corresponding to taking the ratio of wakeful to anaesthetised values in the natural scale) varied with τ.

Figure 2C and 2E show the trend of Δlog(Φ) across τ when using the skipping method. Unlike raw Φ values, visual inspection indicated a peak of Δlog(Φ) in the range of τ = 8 to τ = 16 ms. While this peak was most prominent for the most centrally located channel pairs, it appeared to extend across the fly brain. To confirm that there was indeed a peak within this range of τ, we utilised mixed effects analysis (to account for intra-fly channel pair correlations, see Methods), regressing Δlog(Φ) onto a quadratic term τ^2. The turning point of the fitted quadratic would indicate a peak of Δlog(Φ) at some timescale other than the smallest or largest ones if: 1) it is a local maximum (corresponding to the fitted coefficient for τ^2, β_2, being negative) and 2) it occurs at some intermediate timescale. We first statistically confirmed previous visual inspection that no such peak occurred in the raw Φ values during wakefulness or anaesthesia (Table S1). Meanwhile, the observed peak in Δlog(Φ) when using the skipping method was indeed statistically significant, with fitted coefficients indicating a local maximum at roughly 5 ms ($\chi^2(1) = 663.99$, $\beta_2 = -9.18 \times 10^{-3}$, $\beta_1 = 4.441 \times 10^{-2}$, $\beta_0 = 0.433$).

We next checked if this result could also be found using the downsampling method (Figure 2D and 2F). Given the previous simulation results, we expected to find a similar peak to when using the skipping method. However, visual inspection indicated that the greatest Δlog(Φ) occurred at the smallest τ. The lack of a peak at some intermediate timescale was statistically confirmed by a positive regression coefficient for regressing Δlog(Φ) onto τ^2 (Table S1).

Integrated information identifies the timescale of interactions under non-Markovianity

Though we found some indication of a temporal peak, for Δlog(Φ) when using the skipping method, we were unable to identify such a peak in the raw Φ values themselves, or for Δlog(Φ) when using the downsampling method. So, we next considered whether particular limitations regarding the application of IIT to neural data could have directly prevented any such finding. Specifically, we considered the limitations which we previously identified in [3] regarding the validity of Φ when there are potential of spurious correlations among system elements, which can occur in non-Markovian systems and when multivariate systems are only partially observed.

We first investigated the issue of non-Markovianity. Specifically, non-Markovianity may be problematic for Φ as IIT 3.0 is entirely constructed for Markovian systems where the state of a system depends only on its immediately previous state. To test if non-Markovianity immediately invalidates the application of Φ with regards to identifying the timescale of system interactions, we extended the previous nonlinear autoregressive model by modifying the lag term *l*. Specifically, we set the lag term *l* to be jittered among 9, 10, and 11 in a probabilistic manner. This way, the system cannot be described as a purely Markovian system where its state at time *t* is completely determined by its state at time *t-l* for some fixed *l*. For simulation, we initialised processes *X* and *Y* to uncorrelated Gaussian noise with mean 0 and variance both set to 0.5:

$$X(t) = \varepsilon_X(t)$$
$$Y(t) = \varepsilon_Y(t)$$

Then, for each timepoint *t*:

$$aX(t) \rightarrow X(t + l_a), \qquad bY(t) \rightarrow X(t + l_b)$$
$$cY(t) \rightarrow Y(t + l_c), \qquad dX(t) \rightarrow Y(t + l_d)$$

Where "->" denotes updating the right-hand value by adding the value on the left. We added non-Markovianity here by probabilistically choosing l_a, l_b, l_c, and l_d to be 9, 10, or 11 timesteps, all independently of one another, with probability 0.25, 0.5, and 0.25 respectively. Consequently, each time sample could have been determined by either 1, 2, or 3 individual timepoints from the past. This simulates variability in neural spike or burst timings [19,20]. Note that, while the model is now non-Markovian, the system elements still clearly interact at a timescale of roughly 10 timesteps. The cross-connection strengths were, as for the first simulation, dependent on a threshold voltage:

$$b = \begin{cases} 0.9, & Y(t) > threshold \\ 0, & otherwise \end{cases}$$

$$d = \begin{cases} 0.9, & X(t) > threshold \\ 0, & otherwise \end{cases}$$

With *threshold* again being 0.9.

Figure 3 shows the trend of Φ when computed from the time-series generated by this model, for both the skipping and downsampling methods. For both methods, Φ was maximal at the scale of 10 timesteps, corresponding to the timescale of system interactions. Hence, non-Markovianity per se does not prevent Φ from identifying the timescale of system interactions. However, non-Markovianity did appear to affect the magnitude of Φ values when using the skipping method. Specifically, Φ was an order of magnitude lower than in the first simulation (maximum Φ being ~0.025 in Figure 3A, compared to ~0.13 in Figure 1F). This drastic reduction in Φ did not occur when using the downsampling method (Figure 3B).

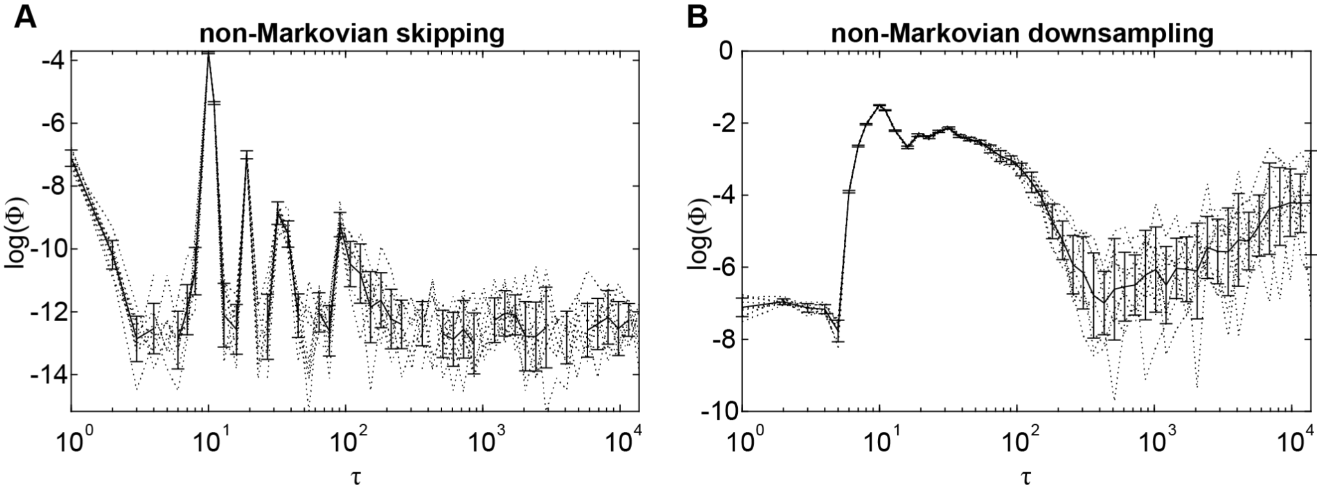

Figure 3. Relationship between integrated information Φ and timescale τ in a nonlinear system (Figure 1) extended with non-Markovianity. (**A**) Log transformed Φ values in relation to τ when using the skipping method. Dotted and solid lines indicate individual simulation runs and the average across runs, respectively. Error bars indicate standard deviation across runs. (**B**) Same as **A**, but for Φ values computed using the downsampling method.

Integrated information identifies the timescale of interactions under partial observation

Given that we were able to identify the timescale of interactions even with a violation of Markovianity, we next turned towards the issue of partial observations. As Φ computed from partial observations (i.e. not across the full system, or complex) is not postulated to correspond to consciousness per se, it could be the case that the timescale is smoothed out through delayed effects from interactions from non-observed system elements. To test this, we extended our non-Markovian system by introducing a third element, giving a total of three system elements. Specifically:

$$X(t) = \varepsilon_X(t)$$
$$Y(t) = \varepsilon_Y(t)$$
$$Z(t) = \varepsilon_Z(t)$$

Z(*t*) is the third system element, and $\varepsilon_Z(t)$ represents Gaussian noise, with mean 0 and variance both set to 0.5. Then, at each timepoint *t*, time samples affected future time points through self- and cross-connections with some lag *l*:

$$\begin{array}{lll} aX(t) \rightarrow X(t+l_a), & bY(t) \rightarrow X(t+l_b), & gZ(t) \rightarrow X(t+l_g) \\ cY(t) \rightarrow Y(t+l_c), & dX(t) \rightarrow Y(t+l_d), & hZ(t) \rightarrow Y(t+l_h) \\ eZ(t) \rightarrow Z(t+l_e), & fX(t) \rightarrow Z(t+l_f), & iY(t) \rightarrow Z(t+l_i) \end{array}$$

Where *e* is the self-connection of *Z*, and *f*, *g*, *h*, and *i* are the new cross-connections connecting all three system elements *X*, *Y* and *Z* bidirectionally. In this model, we set all self-connections *a*, *c*, and *e* to -0.1, and all cross-connections to 0.4. All the lag terms $l_{\text{a-i}}$ were independent and probabilistic, taking values again of 9, 10, or 11 with probabilities 0.25, 0.5, and 0.25. Again, cross-connection strengths were dependent on a threshold voltage

$$\pi = \begin{cases} 0.4, & \Pi(t) > threshold \\ 0, & otherwise \end{cases}$$

Where π is a cross-connection coefficient (b, d, f, g, h, or i), and $\Pi(t)$ is the voltage for the associated channel ($X(t)$, $Y(t)$, or $Z(t)$; e.g. for cross-connection $\pi = b$, the associated channel is $\Pi(t) = Y(t)$). *threshold* was again 0.9, for all connections.

We first confirmed that our previous findings regarding Φ identifying the timescale of system interactions in the two channel case extends to the three channel case (Figure 4A-B), by computing Φ for two channels at a time. Ideally, background conditions (i.e. the states of channels outside those being used to compute Φ) should be fixed. However, in real neural data, doing so drastically limits the number of observations available to build a TPM. Further, the number of possible background conditions to consider grows exponentially with the number of channels. Consequently, fixing background conditions to compute Φ is infeasible for real neural data, and so we also did not fix background conditions in this simulation. As expected, Φ was maximal at the timescale corresponding to 10 timesteps. Though the magnitude of Φ at this peak was lower than in the previous 2-channel simulations, this was expected from a fully connected system. Specifically, system states in a fully connected system have low specificity about their causes and effects, and this should result in low Φ [21,22]. Though our 2-channel simulations were also fully connected, the only other way of connecting 2 channels is using a unidirectional connection, which would result in minimal Φ (see [3] S7 Text).

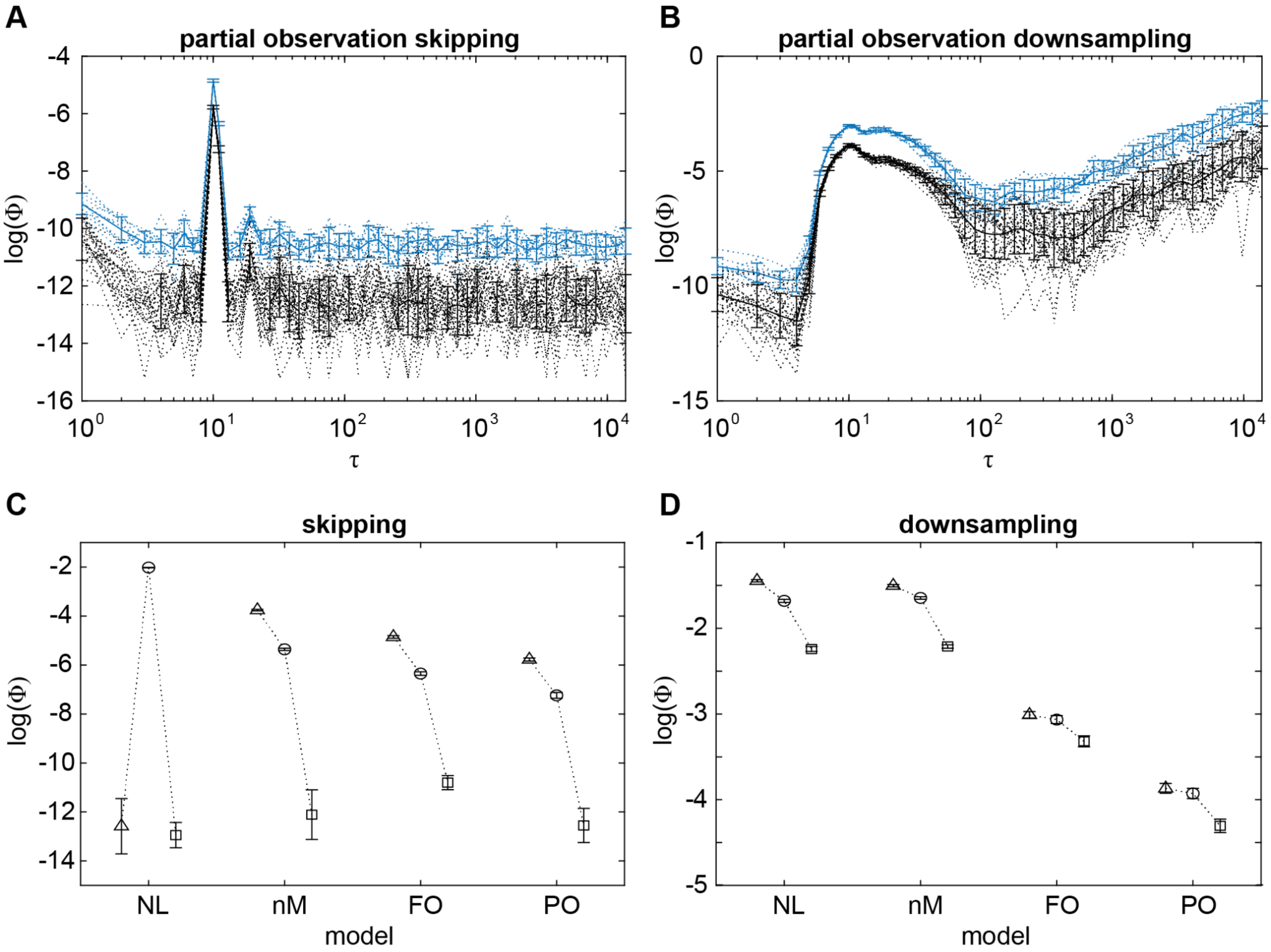

Figure 4. Relationship between integrated information Φ and timescale τ in a system with nonlinearity and non-Markovianity, under partial observation. (**A**) Log transformed Φ values computed from all 3 channels (full observation; blue), and for values computed from 2 channels at a time (partial observation; black), in relation to τ when using the skipping method. Dotted lines indicate individual pairs (partial observation) and runs, while solid lines indicate the mean across pairs and runs, respectively. Error bars indicate standard deviation across pairs (partial observations) and runs. (**B**) Same as **A**, but for Φ values computed using the downsampling method. (C) Summary of maximal Φ values computed using the skipping method (log transformed) for each simulation (NL nonlinear; nM non-Markovian; FO full observation; PO partial observation). Triangles, circles, and squares indicate log(Φ) at τ = 10, 11, and 13 ms respectively. Error bars represent standard deviation across pairs (partial observation) and runs.

To test whether partial observation prevents Φ from identifying the timescale of system interactions, we then computed Φ on 2 channels, out of the 3, at a time. This simulates the case of not being able to observe the states of all neurons in the brain. Or, as previously in the fly LFPs, the case of not being able to compute Φ using all available observations. Figure 4A and 4B also show the trend of Φ when computed from 2 channels at a time, in relation to timescale. Similarly to non-Markovianity, the magnitude of Φ was again reduced by an order of magnitude, this time for both the skipping and downsampling methods. However, Φ was still maximal at the timescale of 10 timesteps, suggesting that partial observation per se also does not in principle prevent Φ from identifying the timescale of system interactions.

Discussion

Here we applied the measure Φ to simple autoregressive models and real neural data, both with continuous system elements. Φ has been proposed by integrated information theory 3.0 (IIT) to be maximal at a temporal scale corresponding to that of conscious experience. Here, we demonstrated that for a nonlinear system, Φ can be maximal to the timescale corresponding to that at which system elements interact. We also applied Φ to neural data, finding that the measure, when normalised, peaks at a timescale of roughly 5 ms. Finally, in follow-up simulations we demonstrated that Φ still peaks at the timescale at which system elements interact, even when certain assumptions of IIT, namely Markovianity and full observation of the system, are not met.

The emergence of a temporal peak of Φ has previously been illustrated in simulation studies utilising systems consisting of binary elements [6,23]. These studies focused on utilising the framework provided by IIT to question the common view posed by reductionism – that the causal structure of a system is fully captured at the most fine-grained level, with there being no room for causal contribution from macro spatiotemporal scales. Rather, they posit that Φ can capture and describe causal emergence, whereby interactions at a macro scale contribute to the causal structure of a system beyond those at the most fine-grained level. The simulation results presented here extend their illustration of causal emergence across temporal scales, as captured by Φ, to systems with continuous elements.

Why is there a peak in normalised II but not directly in II?

Though we found Φ to clearly peak at the timescale of interactions among system elements in the autoregressive models, we observed no such peak in fly LFPs during wakefulness or anaesthesia. Instead, we found a temporal peak to manifest for normalised Φ, the ratio of Φ during wakefulness to anaesthesia. Why this was the case is not immediately clear, but there are some considerations which may have prevented Φ from clearly peaking at some intermediate timescale, as was the case for $\Delta\log(\Phi)$ when using the skipping method to characterise timescale.

One potential explanation regards the effects of non-Markovianity and partial observation. While the peaks in Φ for the simulated systems reliably matched the timescale at which their elements interacted with one another, the systems were designed to have clear temporal dynamics. Specifically, elements interacted with a consistent delay of around 10 timesteps. However, the temporal dynamics of the brain are much less clear, where the effects of non-Markovianity and partial observation are likely to be much greater than in the models used here. For example, autoregressive models fit to LFPs from monkeys have been fit to the 10th or 20th order with timesteps of 5 ms [24,25], with many historical timesteps potentially influencing any one given time sample. Consequently, any one ideal temporal scale may be greatly blurred. Indeed, in the simulations here, peak Φ values reduced as non-Markovianity and partial observation were incrementally added to the models. As these factors are present both during wakefulness and anaesthesia, it is conceivable that normalising wakeful Φ by anaesthetised Φ cancels them out to some extent.

A second potential explanation regards the TPMs used for computing Φ, which were constructed at each timescale. For a given TPM, the number of transitions used to construct it depended on its associated timescale. Specifically, for n time samples, the number of transitions that can be used to construct the TPM is n-τ when using the skipping method, and n-τ-1 when using the downsampling method. Consequently, each entry of the TPM is determined using fewer samples as τ increases, with probabilities becoming less reliable and more likely to take more deterministic values (i.e. probabilities closer to 0 or 1). This in turn may cause Φ values to increase systematically with τ, as more deterministic probabilities allow for greater information in each system state. While we observed this trend for the downsampling method, the skipping method however revealed an opposite trend. At this point, it is unclear how less reliable but more deterministic seeming TPMs would result in both increasing and decreasing Φ values, depending on the method used to characterise timescale. However, the systematic effect may further hide the temporal scale of a system, while meanwhile being cancelled out by normalising wakeful Φ values by anaesthetised Φ values.

Why do skipping and downsampling methods give different peaks?

The autoregressive simulation results presented here indicated that Φ would be maximal at the timescale corresponding to that at which system elements interact, regardless of whether the skipping or downsampling methods were used. Specifically, Φ computed from both methods should identify the same timescale. However, this was not the case in fly data for Δlog(Φ), where Δlog(Φ) peaked at roughly 5 ms when using the skipping method but not the downsampling method. While it is not immediately clear as to why only one method would identify a peak, here we provide a potential interpretation of this result.

While the simulations we used here had very clear dynamics at a particular, specific timescale, it is conceivable that interactions in the brain take place at multiple timescales. Multiple timescales may exist by virtue of the skipping and downsampling methods capturing different types of timescales. Specifically, the skipping method captures the delay between system elements being in some particular state affecting others. An example of different timescales of this type might be short and long range connections having shorter and longer delays respectively. Meanwhile, the downsampling method instead tries to capture the temporal size of the states the system elements can take. Different timescales of this type could manifest as, for example, both neuronal bursting and individual neuronal spikes being states which influence other neurons. Further simulations incorporating the above considerations may be required to understand how Φ or Δlog(Φ) behaves when system elements interact across multiple such timescales.

Taking into account the above considerations, the peak in Δlog(Φ) computed using the skipping method at 5 ms may reflect just one timescale at which neuronal interactions occur. This timescale sits between two neurophysiologically reported timescales. The first is that of axon conduction delays, the delay in firing between connected neurons, which is known to be on the order of single-digit milliseconds [26]. The second is that of critical flicker fusion frequency, the frequency at which a flickering visual stimulus is indistinguishable from a constant stimulus. For flies, the critical flicker fusion frequency has been reported, using electroretinograms, to be at 57 Hz [27], with each individual flicker lasting ~18 ms. We note however that critical flicker fusion frequencies are known to vary, at least in humans, depending on a variety of factors such as stimulus size and intensity and perceptual load [28,29], and that flicker fusion frequencies have not to our knowledge been validated in flies using a behavioural paradigm.

Meanwhile, $\Delta\log(\Phi)$ computed using the downsampling method peaking in the shorter timescales (1-2 ms; Figure 3D) may correspond more directly to the shorter timescale of axon conduction delays. While regressing $\Delta\log(\Phi)$ computed using this method onto timescale did not reveal a negative parabolic trend with a global maximum, this may have been due to not having higher sampling rate data. Thus, it is unclear whether this peak is a potential global maximum or just a general trend of $\Delta\log(\Phi)$ increasing with shorter timescales.

Conclusion and future directions

This work is to our knowledge the first direct application of IIT to search for a potential timescale of consciousness using neural data. While a previous study characterised a proxy measure of Φ, Φ_{AR}, across timescales in electroencephalographic recordings from infants [30], the Φ values were negative for most of the timescales investigated, making their interpretation unclear within the framework of the theory [14]. Meanwhile, here we identified a timescale which aligns with neural physiology and potentially flies' behaviourally and phenomenologically (if any) relevant flicker fusions. However, this comes with the caveat that raw Φ values from the fly recordings either increased or decreased monotonically across timescales, depending on the pre-processing method used. Consequently, more work, utilising both simulation and neural recordings with higher temporal and spatial resolution, is required to confirm whether Φ peaks uniquely at this identified timescale or at varying timescales depending on the method used for characterising timescale. Within this line of work, other methods of characterising timescale should be explored in neural data, such as grouping micro states with logical operations or through black-boxing [6,7]. There is also the further question of whether the peak identified here persists across differing spatial scales, such as at the single neuron level. Finally, behavioural paradigms which capture the temporal scale of conscious experience in a system would be required to more strongly link this potential peak in Φ to consciousness.

Methods

As the fly LFPs analysed here are the same data as described and analysed in [3], we refer the reader there instead of repeating the details here. Details regarding the algorithm for computing Φ from TPMs are also identical to those provided in [3] (albeit for 2 channels at a time, instead of 4 channels, due to the extra computational cost of repeatedly computing Φ at different timescales). So, here we provide only the details regarding generating data from the autoregressive models described in the Results section, and statistical analyses of the LFPs.

Autoregressive simulation

Model simulation and data analyses were conducted using MATLAB 2019b. For each of the three autoregressive models (each additionally including nonlinearity, non-Markovianity, and partial observations), we simulated 20,000 timepoints, for each of 10 runs. The initial conditions for each run were determined by the uncorrelated noise terms ε_X, ε_Y and ε_Y, as described in the Results.

Φ computation

Data processing for computing Φ was conducted using Python 3.6.0 in MASSIVE (Multi-modal Australian ScienceS Imaging and Visualisation Environment), a high-performance computing facility. We calculated the measures using PyPhi (version 0.8.1; [31]), publicly available at https://github.com/wmayner/pyphi. Detailed description regarding the computation of Φ from the TPM, are provided in [3,2,31].

Statistical analyses

We used linear mixed effects analysis (LME; [32,33]) to statistically test for a peak in Φ at some intermediate timescale (i.e. not corresponding to the shortest or longest timescales). This allowed us to account for within-fly correlations among channel pairs without averaging across channel pairs or flies, by including random intercepts for fly and the interaction between fly and pairs as random effects. As Φ was positively skewed, we analysed log transformed values. To test for a potential peak in Φ at some intermediate timescale, we first assessed the significance of a quadratic fit by comparing the quadratic models (Table 2) to models with only a linear term (in Wilkinson notation [33]):

$$\Phi \sim \tau + (1|fly) + (1|fly{:}pair)$$

Where Φ is one of Φ_{SW}, Φ_{SA}, $\Phi_{S\Delta}$, Φ_{DW}, Φ_{DA}, $\Phi_{D\Delta}$, and τ is timescale (see Supplementary Table 2). Subscripts *S* and *D* indicate Φ computed using the skipping and downsampling methods, respectively, while *A* and *W* indicate Φ computed during wakefulness or anaesthesia. Subscript Δ indicates $\Delta\log(\Phi)$. As we searched through exponentially increasing τ, we fitted to $\log_2(\tau)$ values. To assess significance, we used likelihood ratio tests, comparing the log-likelihood of the quadratic model to the log-likelihood of the linear model. As the likelihood ratio is χ^2 distributed when one model is nested within another with degrees of freedom equal to the difference in number of coefficients between the models, we report $\chi^2(d.o.f)$ and its corresponding *p*-value. We summarise the amount of variance explained by each fitted quadratic model, and by each random effect, in Table S2.

Given that the quadratic term significantly increased the variance explained by the model, we then calculated the turning points of each fitted model. We considered there to be an intermediate peak in Φ if the turning point was a local maximum and occurred at some τ other than the most extreme timescales.

References

1. Tononi, G. Consciousness as Integrated Information: A Provisional Manifesto. *The Biological Bulletin* **2008**, *215*, 216–242, doi:10.2307/25470707.

2. Oizumi, M.; Albantakis, L.; Tononi, G. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Comput Biol* **2014**, *10*, e1003588, doi:10.1371/journal.pcbi.1003588.

3. Leung, A.; Cohen, D.; Swinderen, B. van; Tsuchiya, N. Integrated Information Structure Collapses with Anesthetic Loss of Conscious Arousal in Drosophila Melanogaster. *PLOS Computational Biology* **2021**, *17*, e1008722, doi:10.1371/journal.pcbi.1008722.

4. Buzsáki, G.; Anastassiou, C.A.; Koch, C. The Origin of Extracellular Fields and Currents — EEG, ECoG, LFP and Spikes. *Nat Rev Neurosci* **2012**, *13*, 407–420, doi:10.1038/nrn3241.

5. Holcombe, A.O. Seeing Slow and Seeing Fast: Two Limits on Perception. *Trends in Cognitive Sciences* **2009**, *13*, 216–221, doi:10.1016/j.tics.2009.02.005.

6. Hoel, E.P.; Albantakis, L.; Marshall, W.; Tononi, G. Can the Macro Beat the Micro? Integrated Information across Spatiotemporal Scales. *Neurosci Conscious* **2016**, *2016*, niw012, doi:10.1093/nc/niw012.

7. Marshall, W.; Albantakis, L.; Tononi, G. Black-Boxing and Cause-Effect Power. *PLoS Computational Biology* **2018**, *14*, e1006114, doi:10.1371/journal.pcbi.1006114.

8. Gomez, J.D.; Mayner, W.G.P.; Beheler-Amass, M.; Tononi, G.; Albantakis, L. Computing Integrated Information (Φ) in Discrete Dynamical Systems with Multi-Valued Elements. *Entropy* **2021**, *23*, 6, doi:10.3390/e23010006.

9. Cohen, D.; Sasai, S.; Tsuchiya, N.; Oizumi, M. A General Spectral Decomposition of Causal Influences Applied to Integrated Information. *J Neurosci Methods* **2020**, *330*, 108443, doi:10.1016/j.jneumeth.2019.108443.

10. Lütkepohl, H. *New Introduction to Multiple Time Series Analysis*; Springer Science & Business Media, 2005; ISBN 978-3-540-27752-1.

11. Zeldenrust, F.; Wadman, W.J.; Englitz, B. Neural Coding With Bursts—Current State and Future Perspectives. *Front. Comput. Neurosci.* **2018**, *12*, 48, doi:10.3389/fncom.2018.00048.

12. Constantinou, M.; Elijah, D.H.; Squirrell, D.; Gigg, J.; Montemurro, M.A. Phase-Locking of Bursting Neuronal Firing to Dominant LFP Frequency Components. *Biosystems* **2015**, *136*, 73–79, doi:10.1016/j.biosystems.2015.08.004.

13. Barrett, A.B.; Seth, A.K. Practical Measures of Integrated Information for Time-Series Data. *PLoS Computational Biology* **2011**, *7*, e1001052, doi:10.1371/journal.pcbi.1001052.

14. Oizumi, M.; Amari, S.; Yanagawa, T.; Fujii, N.; Tsuchiya, N. Measuring Integrated Information from the Decoding Perspective. *PLoS Computational Biology* **2016**, *12*, e1004654, doi:10.1371/journal.pcbi.1004654.

15. Kim, H.; Hudetz, A.G.; Lee, J.; Mashour, G.A.; Lee, U.; the ReCCognition Study Group; Avidan, M.S.; Bel-Bahar, T.; Blain-Moraes, S.; Golmirzaie, G.; et al. Estimating the Integrated Information Measure Phi from High-Density Electroencephalography during States of Consciousness in Humans. *Frontiers in Human Neuroscience* **2018**, *12*, 42, doi:10.3389/fnhum.2018.00042.

16. Cohen, D.; van Swinderen, B.; Tsuchiya, N. Isoflurane Impairs Low Frequency Feedback but Leaves High Frequency Feedforward Connectivity Intact in the Fly Brain. *eNeuro* **2018**, ENEURO.0329-17.2018.

17. Cousineau, D. Confidence Intervals in Within-Subject Designs: A Simpler Solution to Loftus and Masson's Method. *Tutorials in Quantitative Methods for Psychology* **2005**, *1*, 42–45.

18. O'Brien, F.; Cousineau, D. Representing Error Bars in Within-Subject Designs in Typical Software Packages. *Tutorials in Quantitative Methods for Psychology* **2014**, *10*, 56–67.

19. Kepecs, A.; Lisman, J. Information Encoding and Computation with Spikes and Bursts. *Network* **2003**, *14*, 103–118, doi:10.1088/0954-898X/14/1/306.

20. Rauske, P.L.; Chi, Z.; Dave, A.S.; Margoliash, D. Neuronal Stability and Drift across Periods of Sleep: Premotor Activity Patterns in a Vocal Control Nucleus of Adult Zebra Finches. *J. Neurosci.* **2010**, *30*, 2783–2794, doi:10.1523/JNEUROSCI.3112-09.2010.

21. Mediano, P.A.M.; Seth, A.K.; Barrett, A.B. Measuring Integrated Information: Comparison of Candidate Measures in Theory and Simulation. *Entropy* **2019**, *21*, 17, doi:10.3390/e21010017.

22. Sarasso, S.; Rosanova, M.; Casali, A.G.; Casarotto, S.; Fecchio, M.; Boly, M.; Gosseries, O.; Tononi, G.; Laureys, S.; Massimini, M. Quantifying Cortical EEG Responses to TMS in (Un)Consciousness. *Clinical EEG and Neuroscience* **2014**, *45*, 40–49, doi:10.1177/1550059413513723.

23. Grasso, M.; Albantakis, L.; Lang, J.P.; Tononi, G. Causal Reductionism and Causal Structures. *Nat Neurosci* **2021**, *24*, 1348–1355, doi:10.1038/s41593-021-00911-8.

24. Brovelli, A.; Ding, M.; Ledberg, A.; Chen, Y.; Nakamura, R.; Bressler, S.L. Beta Oscillations in a Large-Scale Sensorimotor Cortical Network: Directional Influences Revealed by Granger Causality. *PNAS* **2004**, *101*, 9849–9854, doi:10.1073/pnas.0308538101.

25. Hoerzer, G.; Liebe, S.; Schloegl, A.; Logothetis, N.; Rainer, G. Directed Coupling in Local Field Potentials of Macaque V4 during Visual Short-Term Memory Revealed by Multivariate Autoregressive Models. *Frontiers in Computational Neuroscience* **2010**, *4*, 14, doi:10.3389/fncom.2010.00014.

26. Gaudry, Q.; Hong, E.J.; Kain, J.; de Bivort, B.L.; Wilson, R.I. Asymmetric Neurotransmitter Release Enables Rapid Odour Lateralization in Drosophila. *Nature* **2013**, *493*, 424–428, doi:10.1038/nature11747.

27. Buschbeck, E.K.; Ehmer, B.; Hoy, R.R. The Unusual Visual System of the Strepsiptera: External Eye and Neuropils. *J Comp Physiol A* **2003**, *189*, 617–630, doi:10.1007/s00359-003-0443-x.

28. Hecht, S.; Shlaer, S. Intermittent Stimulation by Light: V. The Relation between Intensity and Critical Frequency for Different Parts of the Spectrum. *Journal of General Physiology* **1936**, *19*, 965–977, doi:10.1085/jgp.19.6.965.

29. Carmel, D.; Saker, P.; Rees, G.; Lavie, N. Perceptual Load Modulates Conscious Flicker Perception. *Journal of Vision* **2007**, *7*, 14, doi:10.1167/7.14.14.

30. Isler, J.R.; Stark, R.I.; Grieve, P.G.; Welch, M.G.; Myers, M.M. Integrated Information in the EEG of Preterm Infants Increases with Family Nurture Intervention, Age, and Conscious State. *PLOS ONE* **2018**, *13*, e0206237, doi:10.1371/journal.pone.0206237.

31. Mayner, W.G.P.; Marshall, W.; Albantakis, L.; Findlay, G.; Marchman, R.; Tononi, G. PyPhi: A Toolbox for Integrated Information Theory. *PLOS Computational Biology* **2018**, *14*, e1006343, doi:10.1371/journal.pcbi.1006343.

32. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting Linear Mixed-Effects Models Using Lme4. *Journal of Statistical Software* **2015**, *67*, 1–48, doi:10.18637/jss.v067.i01.

33. Harrison, X.A.; Donaldson, L.; Correa-Cano, M.E.; Evans, J.; Fisher, D.N.; Goodwin, C.E.; Robinson, B.S.; Hodgson, D.J.; Inger, R. A Brief Introduction to Mixed Effects Modelling and Multi-Model Inference in Ecology. *PeerJ* **2018**, *6*, e4794.

3.2 - Supporting Information

I now provide the supporting information related to the manuscript in Section 3.1, beginning on the next page.

Table S1

Table S1. Dependence of regressands on timescale τ.

	β_2[b]	**β_1[c]**	**β_0[d]**	**$\chi^2(1)$[e]**	**τ_{TP}[f]**
Φ_{SW}	8.56×10^{-3}	-0.260	-5.717	394.51	37983
Φ_{SA}	1.77×10^{-2}	-0.305	-6.151	1795.09	384
$\Phi_{S\Delta}$	-9.18×10^{-3}	4.441×10^{-2}	0.433	308.79	5
$\Phi_{D\mathrm{W}}$	1.16×10^{-2}	0.237	-5.63	853.28	0
Φ_{DA}	1.05×10^{-2}	0.279	-6.026	663.99	0
$\Phi_{D\Delta}$	1.10×10^{-3}	-4.26×10^{-2}	0.399	5.20 (p=.022)	655125

Φ_{SW}: integrated information calculated using the skipping method during wakefulness. Φ_{SA}: integrated information calculated using the skipping method during anaesthesia. $\Phi_{S\Delta}$: integrated information ratio (wakeful to anaesthetised) calculated using the skipping method. Φ_{DW}, Φ_{DA}, and $\Phi_{D\Delta}$: same as Φ_{SW}, Φ_{SA}, and $\Phi_{S\Delta}$, but for integrated information computed using the downsampling method.

[b] β from regressing onto $log_2(\tau)^2$ (see Methods).

[c] β from regressing onto $log_2(\tau)$.

[d] Intercept term from regression.

[e] The degree of freedom for all likelihood ratio tests was 1 (likelihood ratio tests comparing linear models to quadratic models; see Methods). p << .001 for all comparisons, except for $\Phi_{D\Delta}$ where p = .022

[f] τ value (ms) at which turning points for fitted models occur, as determined by each of the regressands β_0, β_1, and β_2**.**

Table S2

Table S2. Linear mixed effect model fit (adjusted R^2) and standard deviation (SD) of random effects.

	***R*^{2}**	***SD***	
Random effect		**+ (1|*f*)$^\#$**	**+ (1|*f*:*n*)$^\wedge$**
$\Phi_{SW} \sim \tau + \tau^2$	.827	0.502	0.445
$\Phi_{SA} \sim \tau + \tau^2$	.756	0.278	0.415
$\Phi_{DW} \sim \tau + \tau^2$	.880	0.233	0.295
$\Phi_{DA} \sim \tau + \tau^2$	.891	0.209	0.329
$\Phi_{S\Delta} \sim \tau + \tau^2$	.601	0.415	0.367
$\Phi_{D\Delta} \sim \tau + \tau^2$	.327	0.164	0.247

Model specifications are described in detail in Methods. Φ_{SW}: integrated information calculated using the skipping method during wakefulness. Φ_{SA}: integrated information calculated using the skipping method during anaesthesia. $\Phi_{S\Delta}$: integrated information ratio (wakeful to anaesthetised) calculated using the skipping method during wakefulness. Φ_{DW}, Φ_{DA}, and $\Phi_{D\Delta}$: same as Φ_{SW}, Φ_{SA}, and $\Phi_{S\Delta}$, but for integrated information computed using the downsampling method. τ: delay between timesteps (log(ms)).

$^\#$ Random intercept for effect of fly.

$^\wedge$ Random intercept for interaction between fly and channel pair.

Chapter 4 - Discovering measures of consciousness using a data-driven approach

In Chapter 2 and 3, I evaluated integrated information theory (IIT) 3.0 by applying its measures to local field potentials (LFPs) from the fly brain and testing its predictions with regards to how its measures should change with respect to level of consciousness and timescale. In this chapter, I seek to begin comparing the performance of integrated information (Φ) and the associated integrated information structure (IIS), measures which are linked to consciousness through its derivation from first principles, to other, potentially simpler measures of conscious level. Specifically, I use a data-driven strategy, systematically evaluating the performance of many candidate measures in discriminating wakefulness from anaesthesia. I employ a toolbox, highly comparative time-series analysis (*hctsa*; Fulcher & Jones, 2017), to provide these candidate measures. *hctsa* provides a library of 7702 univariate time-series features originating from various fields of research which are not linked to consciousness through first principles.

Given the highly exploratory nature of the data-driven approach to finding potential measures for discriminating conscious level, I take a registered report approach for this Chapter. With this approach, I first conduct initial analyses on the dataset presented previously in Chapters 2 and 3 (which I refer to in the main text of this chapter as the discovery flies), by training and cross-validating a classifier for each time-series feature in *hctsa*. Next I test these trained classifiers on a small pilot subset of flies in a new dataset (which I refer to as pilot evaluation flies). After in-principle acceptance by a journal, I then complete the analyses (testing the trained classifiers for each time-series feature) on the remaining set of flies in the new dataset (final evaluation flies).

4.1 - Registered report

For this chapter I supply a Stage 1 manuscript currently submitted to and under revision at PLOS Biology. The manuscript begins on the following page.

Registered report

Title: Towards blinded classification of levels of consciousness: distinguishing wakefulness from general anesthesia in flies using a massive library of univariate time series analyses

*+Leung A[1], *Mahmoud A[1], Jeans R[2], Fulcher BD[3], van Swinderen B[2], +Tsuchiya N[1,4,5]

* Equal contribution
+ Corresponding authors
1. School of Psychological Sciences and Turner Institute for Brain and Mental Health, Monash University, Melbourne, Victoria, Australia
2. Queensland Brain Institute, The University of Queensland
3. School of Physics, The University of Sydney, Camperdown, NSW 2006, Australia.
4. Center for Information and Neural Networks (CiNet), Japan
5. Advanced Telecommunications Research Computational Neuroscience Laboratories, Japan.

Author Contributions: The plans for the original experiment were conceived by BvS. The plans for the original analysis for this paper was conceived by NT. The original data was collected by Dror Cohen and RJ. The data blinding was performed by RJ. The analysis was performed by MA and AL. The first draft was written by MA, AL and NT. BF provided critical consultation during the manuscript development. All authors contributed to the final draft.

Abstract

The neural mechanisms of consciousness remain elusive. Previous studies on both human and non-human animals, through manipulation of level of conscious arousal, have reported that specific time-series features correlate with level of consciousness, such as spectral power in certain frequency bands. However, such features often lack principled, theoretical justifications as to why they should be related with level of consciousness. This raises two significant issues: firstly, many other types of times-series features which could also reflect conscious level have been ignored due to researcher biases towards specific analyses; and secondly, it is unclear how to interpret identified features to understand the neural activity underlying consciousness, especially when they are identified from recordings which summate activity across large areas such as electroencephalographic recordings. To address the first concern, here we propose a new approach: in the absence of any theoretical priors, we should be maximally agnostic and treat as many known features as feasible as equally promising candidates. To apply this approach we use *h*ighly *c*omparative *t*ime-*s*eries *a*nalysis (*hctsa*), a toolbox which provides over 7,700 different univariate time-series features originating from different research fields. To address the second issue, we employ *hctsa* to high-quality neural recordings from a relatively simple brain, the fly brain (*Drosophila melanogaster*), extracting features from local field potentials during wakefulness and general anesthesia. For each feature, we constructed a classifier for discriminating the two conditions in a discovery group of flies (*N*=13). In this registered report, we will assess their performance on a blinded evaluation group of flies (*N*=12). While the full details of the experimental methods are unknown to the data analysis team at the time of submission of this Stage 1 manuscript, they will be reported upon in-principle acceptance. Pilot results indicate that the performance of only a small subset of features (up to 590, depending on recording location) successfully generalises to an independent dataset. Features which successfully generalise can be fruitful avenues to explore towards robust discoveries of the neural correlates of consciousness.

Introduction

The question of how physical mechanisms generate conscious arousal is a longstanding question in neuroscience. Understanding the mechanisms that support consciousness will have significant impacts in clinical assessment of loss of consciousness [1]. Historically, researchers have approached this question through identifying electrophysiological differences in brain recordings between differing levels of consciousness, such as wakefulness and anesthesia. This approach has resulted in the discovery of multiple time-series features as markers of level of consciousness, including spectral power in different frequency bands [2–8] and measures of signal complexity in spontaneous recordings [9–12]. Despite these developments, performance in distinguishing levels of consciousness using such markers remains limited [13,14].

The limited performance of previously identified markers in distinguishing levels of consciousness, and failure to extend to new conditions, may be due to a lack of theoretical expectations as to what they should be. Historically, candidate markers were often found through visually contrasting electrophysiological recordings, such as electroencephalograms, obtained at varying levels of consciousness [15,16]. Though this approach has led to well-known markers of depth of anesthesia, it is limited by biases towards groups of time-series features for which differences in level of consciousness are visually clear (for a related issue in sleep research, see [17]). While newer markers have moved away from features which are visually clear, a similar problem applies, wherein researchers investigate features selected based on their own expertise of particular facets of time-series structure. Consequently, a vast range of other time-series features will have been ignored as potential markers for conscious level.

One way of removing the bias inherent to selecting individual features to investigate as markers of conscious level is to be maximally agnostic about the types of time-series properties which can map to consciousness. Then, we can systematically test and compare as many time-series features as feasible. This approach consists of two main components.

Firstly, "all" potential features of some given neurophysiological time series should be investigated. While comparison of multiple features has been done previously on well-established features [13,14], "all" features should be compared, not only those determined by or related to visual inspection or individual expertise in particular time-series features. While at first this may seem like a daunting task, this is feasible using *h*ighly *c*omparative *t*ime *s*eries *a*nalysis (*hctsa*; [18]). *hctsa* is a computational framework which extracts from a given time series a massive number (>7000) of univariate time-series features. These features are taken from a multitude of research fields, and include measures such as basic statistics of the distribution of time samples, linear correlations among timepoints, stationarity, entropy measures, among others. This library has been applied previously to find meaningful time-series features for such applications as detecting falls [19] to identifying physiological dynamics underlying neurological disorders [20].

Secondly, to avoid overfitting to a particular dataset, features should be validated on datasets independent from the original dataset from which the features were originally identified [21]. While "cross-validation", a method which splits data into training and testing sets is common, it is rarer to use completely independent and unseen data to test models [14,22]. The use of independent, unseen data is particularly rare in consciousness research (but see [23]), likely due to the cost and clinical problems of obtaining independent datasets. This is especially true for data from human participants which involve manipulations of level of consciousness through general anesthesia [24]. Ethics further limits recruitment of healthy participants for which there is no medical reason for inducing anesthesia or obtaining recordings.

The issue of data availability in human anesthesia recordings can be circumvented by first applying our approach to simpler brains, such as fly brains. Recordings from flies can be obtained relatively cheaply with no clinical concerns, and, due to the relatively small brain, (~10^5 neurons compared to 10^{11} for human brains; [25,26]), neural activity can be obtained simultaneously throughout the whole brain. Consequently, we can obtain high-quality recordings from many healthy flies. Using high-quality recordings from a relatively simple system also offers an advantage. That is, the identified time-series features can be more directly interpreted to understand underlying neural phenomena (compared to features identified from e.g. recordings from the human scalp). Despite seemingly different neural architecture compared to mammals, flies seem to experience varying states of arousal, regulated in a similar way to mammals, such as sleep [27–30] and anesthesia [31,32]. Given these similarities and advantages described above, the fly serves as a useful model to begin to apply new data-driven approaches to discriminating consciousness levels from univariate neural time series (see also [33,12,34]).

In this registered report, we aim to evaluate a massive, comprehensive set of individual time-series features, coming from multiple research fields, as potential markers of level of consciousness. Which univariate time-series features accurately and reliably distinguish between conscious levels? And do they correspond to previously proposed univariate measures of conscious levels? Or are there some conceptually unexplored time-series features which perform better? If no features reliably distinguish conscious levels, this would highlight the need for bivariate or multivariate features. These would include features such as coherence, Granger causality, [32,35], transfer entropy [36], Lempel-Ziv complexity [37], perturbational complexity index [38], etc.. Alternatively, new measures derived from theories, such as integrated information, may be necessary [34,39]. Indeed, many theories of consciousness rely on interactions among parts, and would predict univariate features to be uninformative of conscious level.

Here, we compare the most comprehensive available set of scientific features, made available in the *hctsa* toolkit [18], searching for features that may warrant further exploration in the future as potential markers of consciousness. First, we search for features which reliably distinguish wakefulness from anesthesia, and generalise to a blinded, independent dataset. Second, we search for features for which the direction of the effect of anesthesia (i.e., yield consistently higher or lower values in anesthesia versus wakefulness) is consistent across datasets. These directionally consistent measures could be useful in assessing level of consciousness when a subject's baseline is known. For these purposes, we apply and compare the *hctsa* features systematically. Critically, we validate them on recordings obtained from an independent set of flies which are completely blinded to the analysis team. At the time of submitting this registered analysis, our early results indicate that the performances of many features which have statistically significant performance in classifying wakefulness and anesthesia in one dataset (N = 13 flies) do not generalise to the second independent dataset, highlighting the importance of evaluating measures on independent datasets. Despite this, across the datasets, many features maintain their direction of the effect of anesthesia across the flies.

Summary Table

Research question	What univariate time-series features (from *hctsa*) can serve as markers of level of consciousness ACROSS individuals?	What univariate time-series features (from *hctsa*) can serve as markers of level of consciousness WITHIN individuals?
Hypotheses	1 hypothesis for each hctsa feature at each channel: ● Feature X classifies wake/anesthesia above chance	1 hypothesis for each hctsa feature at each channel: ● Direction of effect of anesthesia for feature X is more consistent than chance
Sampling plan	Use existing data: ● 13 discovery flies x 8 2.25s epochs each of wake/anesthesia (published previously); ● 2 pilot evaluation flies x 112 2.25s epochs each of wake/anesthesia (unpublished); ● 10 final validation flies (unpublished, details undisclosed to data analysis team, but expecting same/similar to pilot evaluation flies)	
Statistical analyses	Classification analysis, using a nearest-median classifier **trained on the discovery flies**. ● Obtain classifier accuracy on discovery flies (leave-one-fly-out validation) and evaluation flies ● Obtain significance by comparing classifier performance to random classification distribution (α = 0.05) ● FDR correction at each channel (*q* = 0.05)	Consistency of wakeful epochs being greater/less than anesthesia epochs at each fly, based on direction of anesthesia effect **in the discovery flies** (see Methods Section “Within-fly effect direction consistency”) ● Obtain significance by comparing consistency to random consistency distribution (α = 0.05) ● FDR correction at each channel (*q* = 0.05)
Pre-specified outcomes	The performance of feature X in discriminating wakefulness/anesthesia shows significant generalisation across individuals and the feature is worth future investigation as a marker of conscious level if: ● It performs significantly in the discovery flies AND ● It performs significantly in the evaluation flies	The within-individual effect of anesthesia for feature X shows significant generalisation across individuals, and the feature is worth future investigation as a marker of conscious level if: ● Consistency of the direction of the effect of anesthesia is significantly above chance in the discovery flies AND ● Consistency is significantly above chance in the evaluation flies, for the same direction as the discovery flies

Methods

Data and preprocessing

We use already-collected local field potentials (LFPs) from fruit fly brains during wakefulness and during isoflurane anesthesia. We use two independent datasets: (i) a *discovery dataset* for initially identifying features which perform well at discriminating wakefulness from anesthesia; and (ii) a blinded *evaluation dataset* for assessing the generalisability of these features to a separate dataset. Figure 1 illustrates our data analysis pipeline for the two sets of flies. As our discovery dataset, we use previously published data from 13 flies [32,12,34]. As our blinded evaluation dataset, we use data from an additional 12 flies collected by RJ and BvS which is delabelled before being provided to AM, AL, and NT for analysis (initials refer to authors of this registered report). At time of submission, 2 of the evaluation flies were provided and used for pilot analysis, with the remaining 10 flies being withheld for final evaluation.

Discovery flies

For this dataset, we provide details relevant to this registered report (for full details see [32]). 13 laboratory-reared female *Drosophila melanogaster* (Canton S wild type 3-7 days post eclosion) were collected under cold anesthesia and glued dorsally to a tungsten rod. Linear silicon probes (Neuronexus Technologies) were inserted laterally into the fly's eye. Each linear probe consisted of 16 electrodes separated with a site separation of 25 μm, and covered approximately half of the fly brain. Recordings were made with a sampling rate of 25 kHz using a Tucker-Davis Technologies multichannel data acquisition system and downsampled to 1000 Hz.

Recordings for each fly were obtained from two blocks: one block with 0 vol% isoflurane at the fly body (wake condition), followed by a block with 0.6 vol% isoflurane (anesthesia condition). Isoflurane was delivered from an evaporator to the fly through a rubber hose. Each block followed a series of air puffs, and consisted of 18 s of rest, 248 s of visual stimuli, another 18 s of rest, and a second series of air puffs. Isoflurane was administered following the last air puff of the first block, and flies were left to adjust to the new concentration for 180 s before beginning the second block. Flies in the wake condition responded to air puffs by moving their legs and abdomen, but were inert during the anesthesia condition. Flies regained responsiveness after isoflurane was removed, ensuring that flies were alive during the anesthesia recordings [33]. We use the data obtained in the 18 s period of each block corresponding to the rest period preceding the visual stimuli.

We bipolar re-referenced the LFPs by subtracting adjacent electrodes to acquire 15 signals which we refer to as "channels". Channel 1 refers to the channel positioned furthest into the fly brain. Finally, we segmented the 18 s period into 2.25 s segments, giving 8 epochs per fly and condition.

Pilot evaluation flies

On 14/06/2019, the data-analysis team (AM, AL, and NT) was provided with 56 segments of 20 second spontaneous activity recordings from the data collection team (RJ and BvS). The 56 segments were known to the data analysis team as coming from 2 flies and from varied levels of anesthesia. The analysis team was initially blinded to the labelling of the segments, such that the source condition and fly of each segment was unknown. Further, the analysis team was blinded as to the distribution of segments coming from each fly or anesthesia condition (e.g., whether the 56 segments had an equal number of wake and anesthetized segments, or an equal number of segments from each fly), and to the specific variant of fly and the context in which the data had originally been collected.

However, these labels and information were made available (in June 2019) after early analyses using 18 s segments (corresponding to the original length of the segments from the discovery flies, instead of 2.25 s segments). We later deemed the 18 s approach inappropriate as we would be generalising across-fly classification performance to within-fly classification performance (applying classifiers trained on a single epoch each of wakefulness and anesthesia from each discovery fly to multiple epochs from an individual pilot evaluation fly; see Section "Classification of conscious level"), before finalising the full methods and parameters. With the labels, it was revealed that the 56 segments were equally divided into 14 segments of wakefulness or anesthesia for each of the two flies. It was also revealed that the flies were of a w2202 background (also called isoCJ1), which has a similar isoflurane sensitivity profile to the Canton-S wild-type fly (CS; [40]).

The following technical details of the recordings were available to the data analysis team, to enable equal pre-processing of signals. Electrophysiological data were recorded at 25 kHZ, down sampled to 1000 Hz. Next, LFPs were bipolar re-referencing by subtracting adjacent unipolar channels (*n*=16) to acquire 15 channels.
The exact details originally provided to the analysis team are available at https://osf.io/bq5ry/?view_only=3789097395c1419db2a9eb615bc1effe.

Final evaluation flies

Delabelled final evaluation data will be provided to the data analysis team after in-principle acceptance of the manuscript. The following information was disclosed by the data collection team (RJ and BvS) to the data analysis team at the time of writing of this registered analysis. At the point of submission, the analysis strategy was fixed, and final evaluation data had not been provided to the analysis team. The teams agreed that disclosing the following information would not affect the outcome of the results.

The final evaluation dataset will consist of 10 additional flies (same type as the pilot evaluation flies). The data analysis team will again be provided with 14 x 20 s segments of wakefulness or anesthesia for each fly, shuffled and delabelled, with the same pre-processing applied as for the pilot evaluation flies. At the time of submission of this registered report, the data analysis team was not made aware of any other potential details regarding this final evaluation dataset. We intend to update this section when further recording details are revealed to the data analysis team upon in-principle acceptance. We will repeat the methods pertaining to the pilot evaluation flies, described in the following sections, on the final evaluation flies upon in-principle acceptance.

Local field potential pre-processing

The data analysis team subtracted the mean voltage from each epoch of the discovery and pilot evaluation flies, and then removed line noise from each epoch using the **rmlinesc.m** function of the Chronux toolbox (http://chronux.org/; [41]) with 9 tapers, a time-bandwidth product of 5, and zero-padding factor 2. As a sanity check, we performed visual inspection of power spectrum plots after pre-processing to confirm the removal of line noise. These same pre-processing steps will be applied to the final evaluation flies.

Feature-based time-series analysis using *hctsa*

We extracted 7702 time-series features from each epoch and bipolar re-referenced channel of the discovery and pilot evaluation flies using *hctsa* (v1.03; [18]) on MATLAB 2017b. For a given time series, *hctsa* extracts a vast set of 7702 univariate time-series features from analysis methods developed in a wide range of scientific disciplines, including nonlinear physics, biomedicine, economics, and neuroscience.

Not all of the available time-series features could be extracted successfully from our datasets. For example, the class of features derived from the *hctsa* function `DN_CompareKSFit` includes fits of the data to a beta distribution, which assumes values between 0 and 1, an assumption that is not fulfilled by our data and consequently returns missing (`NaN`) values. To filter out these cases, we excluded any feature which returned `NaN` across all time series for a given channel in the discovery flies. This reduced the set of features down to an average of 6860 features across the 15 channels (ranging from 6657 to 7004). We further excluded features which returned a constant value across all time series for a given channel in the discovery flies because they are uninformative for classification, reducing the set of features again to on average 6764 features across the 15 channels (ranging from 6560 to 6908).

While we analyse raw *hctsa* features, the range of values varies greatly across features, and some features include infinity values (which we keep as they can be used in our classification analysis, see Section: “Classification of conscious level”). Where specified, we visualise scaled feature values using an outlier-robust sigmoidal transformation, which maps values of all epochs for a given feature to the unit interval [42].

Classification of conscious level

We use single-feature classification analysis at each channel to compare the performance of each individual time-series feature in distinguishing wakefulness from anesthesia. If a feature distinguishes conscious level, it should have high classification performance in the discovery flies generalises to the evaluation flies. To account for features which can return infinity values, we employ nearest-centroid classifiers, with class medians as centres.

We first trained and cross-validated features on the discovery flies. For a given channel and feature, we employed a leave-one-fly-out cross-validation procedure on the evaluation flies. Specifically, at each cross-validation iteration, we trained a classifier on all 8 epochs of wake and anesthesia from 12 flies, and tested on 8 epochs of wake and anesthesia on the remaining fly. Each classifier consists of: 1) a threshold, the middle point between the median feature value for wakefulness and anesthesia as obtained from the training set, and 2) a direction indicating whether points above the threshold should be classified as wakeful or anesthetized (and vice versa).

After obtaining cross-validation accuracies on the discovery flies, we finally obtained classifiers for each feature and channel by training on all epochs from all flies in the discovery dataset. We validate and report the performance of these classifiers on the pilot evaluation (*N*=2, in the submitted manuscript) and final evaluation (*N*=10, after in-principle acceptance) evaluation datasets. For the final evaluation dataset, the data analysis team will be provided with shuffled, delabelled epochs upon which the trained classifiers are applied. Only after all epochs in the dataset are classified are the true labels revealed to the data analysis team.

We determined if a feature classifies wake and anesthesia significantly better than chance by comparing each feature to a random-classification distribution at the $\alpha = 0.05$ level. We corrected for multiple comparisons at each channel using the false discovery rate (FDR) correction [43]. We obtained random-classification distributions for the discovery and pilot evaluation flies by repeatedly classifying discovery or evaluation epochs randomly, with equal probability (as there are 7702 potentially available features in *hctsa*, we repeated this random classification $N = 7702$ times to build the distribution). We expect that features which reflect some process underlying change in conscious level will have significant classification performance which persists through cross-validation on the discovery flies to the final validation flies, ideally across channels.

Within-fly effect direction consistency

In assessing generalisability, it is possible that the effect of anesthesia (relative to wake) is highly consistent within individuals, even when features do not classify well across subjects. For example, feature values in the evaluation flies might be entirely outside the range of values in the discovery flies (e.g., all wake and anesthesia values in the evaluation flies being above both the wake and anesthesia values in the discovery flies). In such a case, all feature values would be classified to a single class, even if wake and anesthesia values are separated in the evaluation flies. This is relevant in scenarios where, such as in this registered report, there may be variability in the exact placement of electrodes among individuals. Values may further vary among individuals due to factors such as exact experimental setups and baseline arousal states. To address this, we assessed a weaker form of generalisation – whether a feature is predictive of the relative difference between conscious levels within an individual fly – and report the consistency of the direction of the effect of anesthesia (after receiving the correct wake/anesthesia labels in the case of the final evaluation flies).

Specifically, at a given feature, fly, and channel, we obtained for each wakeful datapoint the proportion of anesthetized datapoints which lie below it. Because the direction of the effect of anesthesia is not necessarily the same across features and channels, we first assigned directionality labels based on the median wakeful and anesthesia values in the 13 discovery flies. For a given feature and channel, we gave a label of 1 if the median wakeful value was greater than for anesthesia, and -1 otherwise. We then multiplied feature values by these labels, flipping the direction of the effect of anesthesia when the median wakeful value is lesser than the median anesthesia value and making the analysis uniform across features and channels. Finally, we report the average proportion across all wakeful epochs and flies.

In a similar way as for testing for significance of classification performance, we used permutation testing to determine if the within-fly effect direction consistency of a feature was significantly better than chance. We obtained reference chance distributions for the discovery flies and pilot evaluation flies by repeatedly ($N = 7702$) randomly assigning the portion of anesthesia epochs which are below each wakeful epoch, with equal probability, and averaging across wakeful epochs and flies. We compare each feature to the distribution at the $\alpha = 0.05$ level, correcting for multiple comparisons at each channel using FDR correction.

Pilot Results

We investigate if any of the time-series features in *hctsa* individually serve as a potential measure of level of conscious arousal in independently obtained recordings from fly brains. We first assessed the performance of hctsa features which we applied to a discovery dataset of previously published fly brain recordings ($N = 13$) [33,32,12,34]. Then, to assess generalisability, we apply classifiers trained on the discovery flies to recordings obtained from an independent set of pilot evaluation flies ($N = 2$). Upon in-principle acceptance of this registered analysis, we will repeat the analyses conducted on the pilot evaluation flies on a final set of evaluation flies ($N = 10$), reporting the features which consistently perform well in distinguishing wakefulness from anesthesia at all recording locations across all the flies.

(Note that upon in-principle acceptance, we will repeat the analysis as described in Methods and Pilot Results, extending and updating the analyses done on the pilot evaluation flies to the remaining 10 evaluation flies. Figures 2b-e, 3b-e, 4, and the corresponding text in results concerning the pilot evaluation flies, will be updated to convey the full analyses on all 12 evaluation flies. We will also provide supplementary materials giving classification performance and consistencies for significant features in the pilot evaluation flies to reflect the full set of 12 evaluation flies.)

Classification of conscious level

We first extracted 7702 time-series features from the initial discovery flies using *hctsa*, yielding 6560 to 6908 valid features across the 15 channels (*M* = 6764). Figure 2a shows a matrix of feature values extracted from Channel 6 in the discovery dataset. We first visually inspected this feature matrix to inspect trends across features and flies. To facilitate interpretation, we first sorted the order of the features according to hierarchical clustering using correlation distances between features, across time-series. This revealed two clear clusters of features, one with values which are generally greater during wakefulness (columns roughly 500 to 1500), and one with values which are generally greater during anesthesia (columns roughly 4500 to 5500). Features in each of these clusters would likely achieve similar classification accuracies.

Having reordered the features, groupings across rows corresponding to epochs from individual flies became apparent. This indicated strong within-fly correlations of feature values but weak correlations across flies, suggesting that few features, if any, would generalise across all the flies. Overall, our visual inspection of the similarities across features and similarities across flies suggested that many features could individually achieve better-than-chance classification accuracy. However, there appeared to be no clear cluster of features which would perfectly discriminate wakefulness from anesthesia in all of the flies.

While Figure 2a set up our global expectations visually, there may have been features outside the visually clear clusters which also distinguish wakefulness from anesthesia extremely well. To reveal such features, we next quantified the across-fly classification performance (within the discovery flies). For each feature, we classified wakeful from anesthetized epochs using a nearest-median classification rule. We assessed the statistical significance of the cross-validation accuracy of each feature by comparing it to a distribution of accuracies resulting from random classification (see Methods). For Channel 6, this yielded 3089 features which performed significantly better than chance ($p < .018$). The best-performing feature, an index of mean stationarity (*hctsa* feature: `StatAvl250`; [44]), achieved a mean classification accuracy of 76% (*SD* = 12% across 13 cross-validations; Figure 2b). Upon performing the classification analysis for each of the remaining 14 channels, we found features to perform heterogeneously across the channels. Overall, the average classification accuracy achieved across channels tended to be much lower than that achieved by individual channels. For example, the across-channel average of the mean cross-validated accuracy of `StatAvl250` was 63% (*SD* = 6% across 15 channels).

Indeed, the number of significant features varied greatly across the channels, ranging from 14 to 3089, with channels closer to the peripheral tending to have fewer significantly performing features (Figure 2b). We found the greatest number of significant features, along with the most accurately classifying features, to occur at Channels 5 and 6, corresponding roughly to the protocerebrum. This is consistent with our previous analyses on this dataset, which reported better discrimination between wakefulness and anesthesia in some but not all channels [12,32].

We next sought to determine how well the performance of features would generalise to an independent evaluation set of flies. While the overall recording procedure was known by the data analysis team to be similar to that of the discovery flies, the exact experimental methods were not revealed at the time of submitting this registered analysis (see Methods). We finalised the training of classifiers by obtaining thresholds based on all 13 discovery flies. As a pilot for this registered analysis, we applied these classifiers to recordings from 2 flies (out of a total of 12 evaluation flies). Across the 15 channels, we found an additional 48 to 416 (*M* = 180) features to either output a `NaN` or have a constant value across epochs in the pilot evaluation flies.

Figure 2c shows how classification performance at Channel 6 in the discovery flies generalizes into the pilot evaluation flies. The best performing feature at Channel 6 in the discovery flies, `StatAvl250`, attained a much reduced accuracy of 63% (green circle and arrow). Meanwhile, several features attained higher performance than in the discovery flies. The feature with the best performance at Channel 6 in the pilot evaluation flies quantifies the relative low-frequency power via a Fourier power spectrum (*hctsa* feature: `SP_Summaries_welch_rect_logarea_2_1`) attained 76% accuracy, despite attaining 62% (*SD* = 14% across cross-validations) in the discovery flies (red circle and arrow). The best performing features across all the channels in the pilot evaluation flies were related to signal variance at Channel 5 (including root-mean-square, *hctsa* feature `rms`, and standard deviation, `standard_deviation`), and also had greater performance than in the discovery flies, attaining 91% accuracy, compared to 71% (*SD* = 21% across 13 cross-validations, for both features). While we leave the interpretation of high-performing features until after the final analysis, it is notable that all these features are related to the variance of the voltage fluctuations, which is consistent with previous literature on the effects of anesthesia on the fly LFPs [32].

Generally, however, we found a drastic drop in the number of features with statistically significant classification accuracy across the channels (Figure 2b). This suggested that features overall performed worse in the evaluation flies, and that their performance was again heterogeneous across channels. Across the channels, the number of significantly performing features was substantially less in pilot evaluation flies, ranging now from 9 to 1885. Further, after restricting to the set of features which yielded significant cross-validation accuracies in the discovery flies, the number of significant features dropped even further, ranging across the channels from 0 to 590. This result alerts us to the danger of interpreting cross-validation accuracies of the discovery flies as an estimate of the true generalisation accuracies, which can only be evaluated using an independent dataset. We will discuss the implication of this finding in Discussion in the Stage 2 manuscript. Upon final data analysis, we will provide the classification performance of each significant feature, at each channel in Supplementary material S1.

Within-fly effect direction consistency

Given that the across-flies classification performance of many features in the discovery flies did not generalise to the pilot evaluation flies, we next assessed a weaker form of generalisation. Even though features may not classify well across subjects, features for which the effect of anesthesia is highly consistent within individuals may still be useful for clinical assessment of conscious level. This is especially true for individual subjects whose baseline neural activity is available (e.g., before anesthetic induction). So, for each feature, we assessed the within-fly effect direction consistency of anesthesia. For an individual fly, consistency would be 1 if the effect of anesthesia is totally consistent for every pairing of wake and anesthesia epochs. (see Methods). In other words, for a given feature in a given channel, if the value for wake minus anesthesia is always above 0 for any pair of one wake and one anesthesia epoch (or vice versa), then we consider such a feature as a perfect measure of consciousness for that particular fly.

Figure 3a illustrates the within-fly effect direction consistency for each feature for the discovery flies, again at Channel 6. Overall, consistencies appeared to be reliable within flies, as we expected. However, strikingly, consistency seemed to be reliable even across flies, indicating that, for many features, the direction of the effect of anesthesia would be consistent across individual flies despite mediocre classification performance (e.g. due to differing baseline values at each fly). Visually, there appeared to be two clusters of features (from column 500 to 1500 and from 4500 to 5000) with high consistency.

We assessed the statistical significance of each feature, this time by comparing its consistency to a distribution of consistencies for randomly labelled epochs (see Methods). For Channel 6, this gave 3923 features which were more consistent than chance ($p < .027$). The feature with the highest consistency at Channel 6, as well as on average across all the channels (a measure of variation in the differences between consecutive time samples, *hctsa* feature: `MD_rawHRVmeas_SD2`; [45]) had a consistency of 0.94 (i.e., on average, each wakeful epoch from an individual fly had a greater value than 94% of anesthesia epochs from the same fly), which previously attained a across-flies cross-validation accuracy of 68% (*SD* = 20%). Across the 15 channels, in general we found many more features to have significant within-fly consistency (2486 to 3923; Figure 3b), compared to across-flies classification.

We next assessed how the within-fly effect direction consistencies generalised to the pilot evaluation flies. We computed consistencies in the pilot evaluation flies, taking into account the direction of the effect of anesthesia observed in the discovery flies. Hence, if wakeful and anesthesia epochs were perfectly separable in the same direction as the discovery flies, consistency would be 1. However, if they were perfectly separable in the opposite direction to the discovery flies, consistency in pilot evaluation flies would be 0.

Figure 3c shows how within-fly effect direction consistency at Channel 6 in the discovery flies generalises to the pilot evaluation flies. Unlike for across-flies classification, within-fly consistencies between the discovery and pilot evaluation flies seemed to be strongly positively correlated. This indicates that the within-fly consistency of many more features generalised to the pilot evaluation flies. The feature with the highest consistency in the evaluation flies, root-mean-square (*hctsa* feature: `rms`), achieved high consistency in the pilot evaluation flies (0.91, red circle and arrow). Across the 15 channels, the number of significantly consistent features seemed to vary more in the pilot evaluation flies, ranging from 478 to 4030 (Figure 3b). This range reduced to 177 to 3125 after restricting to the set of features which also had significant consistency in the discovery flies.

Notably, the decrease in the number of significant features for within-fly consistency in the pilot evaluation flies was less pronounced than for across-fly classification performance. Like across-fly classification, there were more significant features for within-fly consistency at the central than peripheral channels. Overall, these results indicate that many features could be informative and consistent in terms of changes within a single fly due to anesthesia without being strong, absolute measures of conscious level across flies. We will revisit the implication of this finding in Discussion after the final data analysis. Upon final data analysis, we will provide the consistencies of each significant feature, at each channel in Supplementary material S2.

Timeline

We can immediately begin the analyses on the full set of evaluation flies upon in-principle acceptance. All classification models for candidate time-series features have already been trained on the discovery flies. We anticipate the computation of *hctsa* features and final analyses on the evaluation flies as outlined in our manuscript, and writing up of results and discussion to take 2-3 months from in-principle acceptance.

Data Availability Plan

Pre-processed data from the discovery flies are available on Figshare:
https://doi.org/10.26180/5ebe420ae8d89
Pre-processed data from the pilot evaluation flies and blinding procedure are available on OSF: https://osf.io/bq5ry/?view_only=3789097395c1419db2a9eb615bc1effe
Pre-processed data for the full set of evaluation flies will be made available in the OSF project linked above.

Figures

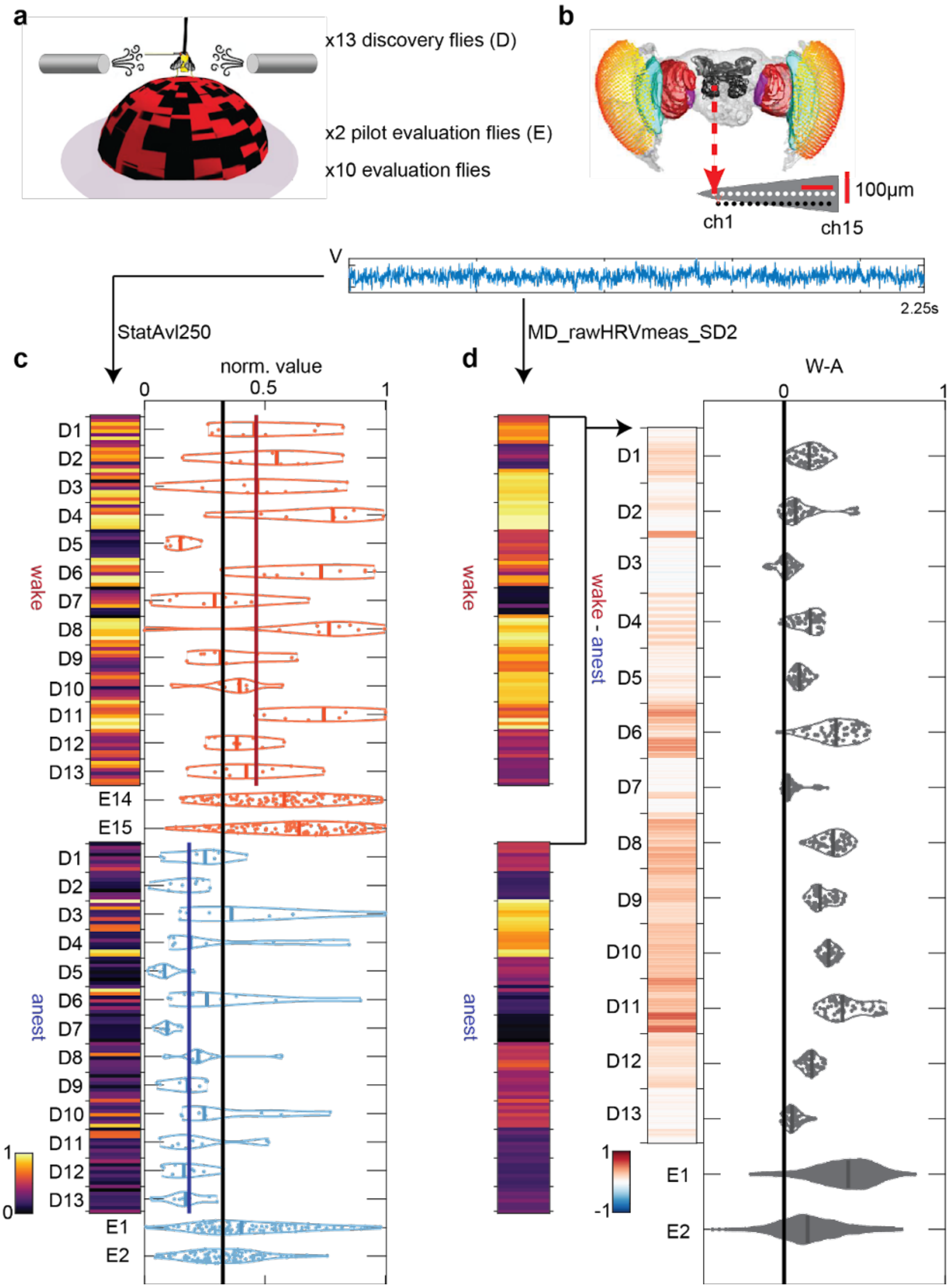

Figure 1. Analysis pipeline for individual features in *hctsa*.

a) Flies were dorsally fixed to a rod and placed on an air-supported ball. Isoflurane was administered through a rubber hose. We use a discovery dataset of 13 flies to identify time-series features which discriminate wakefulness from anesthesia. We assess how the performances of these features generalise to an independent evaluation dataset consisting of

12 flies. We use 2 of these flies to obtain pilot generalisation performance for registering this analysis. **b)** Local field potentials (LFPs) are obtained during wakefulness and anesthesia using linear multi-electrode arrays inserted laterally into the fly brain. **c)** At a given channel and time-series feature (here we show the feature StatAvl250), we compute feature values for every epoch from each fly (each entry in the image plot corresponds to a scaled feature value from one epoch). We train a nearest-median classifier using the discovery (D) flies, where the threshold for classifying wakefulness (red) versus anesthesia (blue) is the middle point (black vertical line) between the median values of the two conditions (red and blue vertical lines). We assess the feature's across-fly classification performance on the discovery flies using a leave-one-fly-out cross-validation procedure. We assess the generalisation of the feature's performance by classifying epochs from the evaluation (E) flies using its threshold as obtained from all the discovery flies. **d)** As a weaker form of generalisation, we also assess within-fly effect direction consistency by finding, for each wake epoch, the proportion of anesthesia epochs which have greater or lesser (depending on the direction of the effect of anesthesia for the feature as illustrated in **c**) feature values. We visualise this here for a second feature by showing the within-fly differences between scaled feature values. Each entry in the rightmost image plot gives the difference between every combination of one wake epoch and one anesthesia epoch from the same fly.

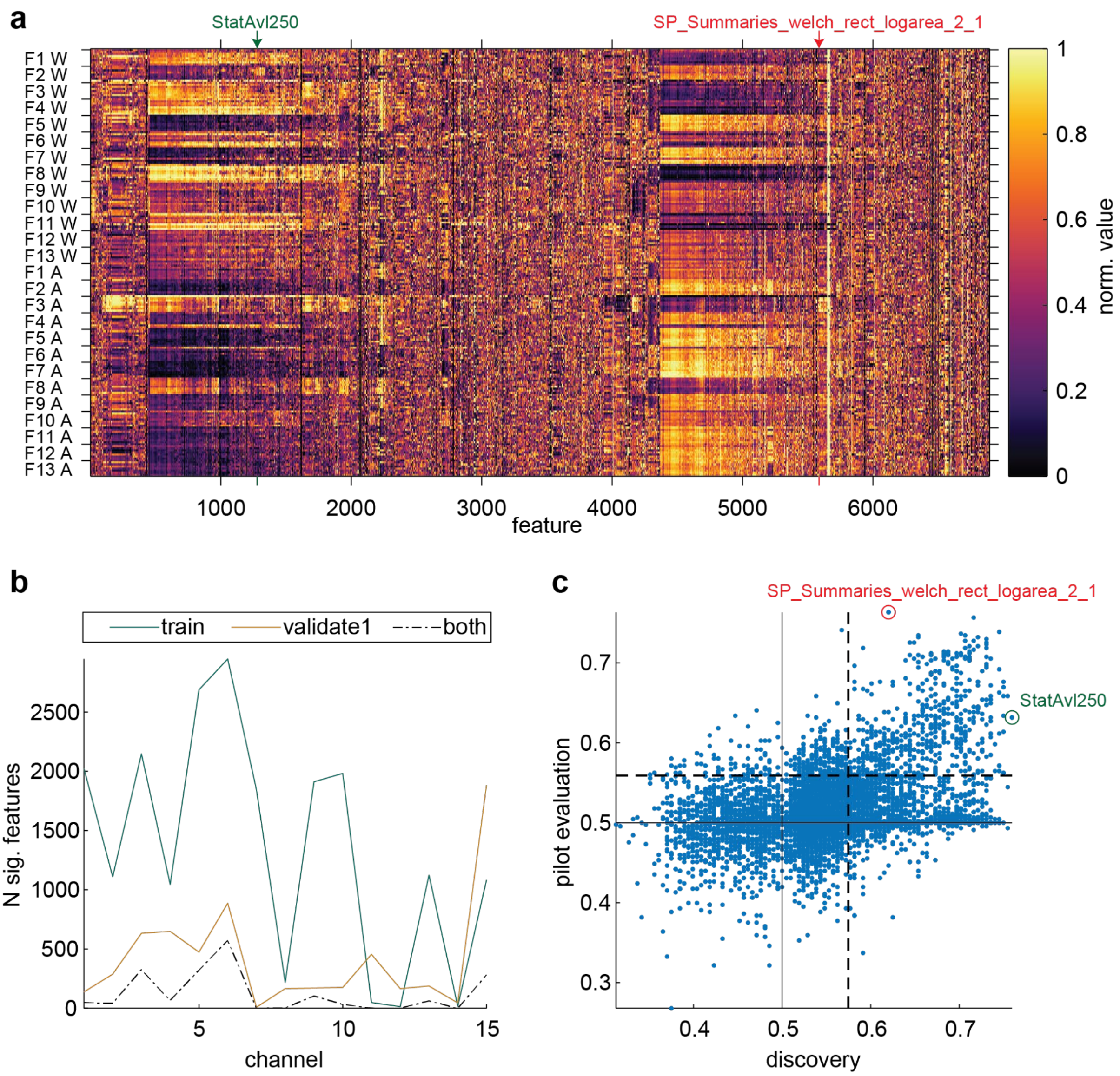

Figure 2. Classification performance of *hctsa* features.

a) Values of *hctsa* features in the discovery flies, at Channel 6. Each row corresponds to an individual 2.25s epoch, from 13 flies (F) during wakefulness (W) and anesthesia (A). Each row displays scaled values for all valid features for the channel. Features (columns) are ordered based on hierarchical clustering using correlation (across time series) distance between normalised features. This ordering places features with highly correlated values across the dataset close to each other. Arrows indicate the features which attained the highest classification performance in the discovery (green) and pilot evaluation (red) flies. **b)** Number of features which achieved statistically significant classification performance at each channel, in the discovery flies blue line, pilot evaluation flies (orange line), and in both the discovery flies and pilot evaluation flies (broken black line). **c)** Correlation of classification performances between the discovery (x-axis) and pilot evaluation flies (y-axis). Each dot represents the classification performance of one of the 6800 features shown in **a**). Solid horizontal and vertical lines indicate chance classification performance (= 0.5). Dashed horizontal and vertical lines indicate the thresholds for statistically significant across-fly classification performance in each set of flies (see Methods). Dots located in the top right quadrant are the features which successfully classified wake from anesthesia across both the discovery and pilot evaluation flies. Circled are the features which attained the highest performance in the discovery (green) and pilot evaluation (red) flies.

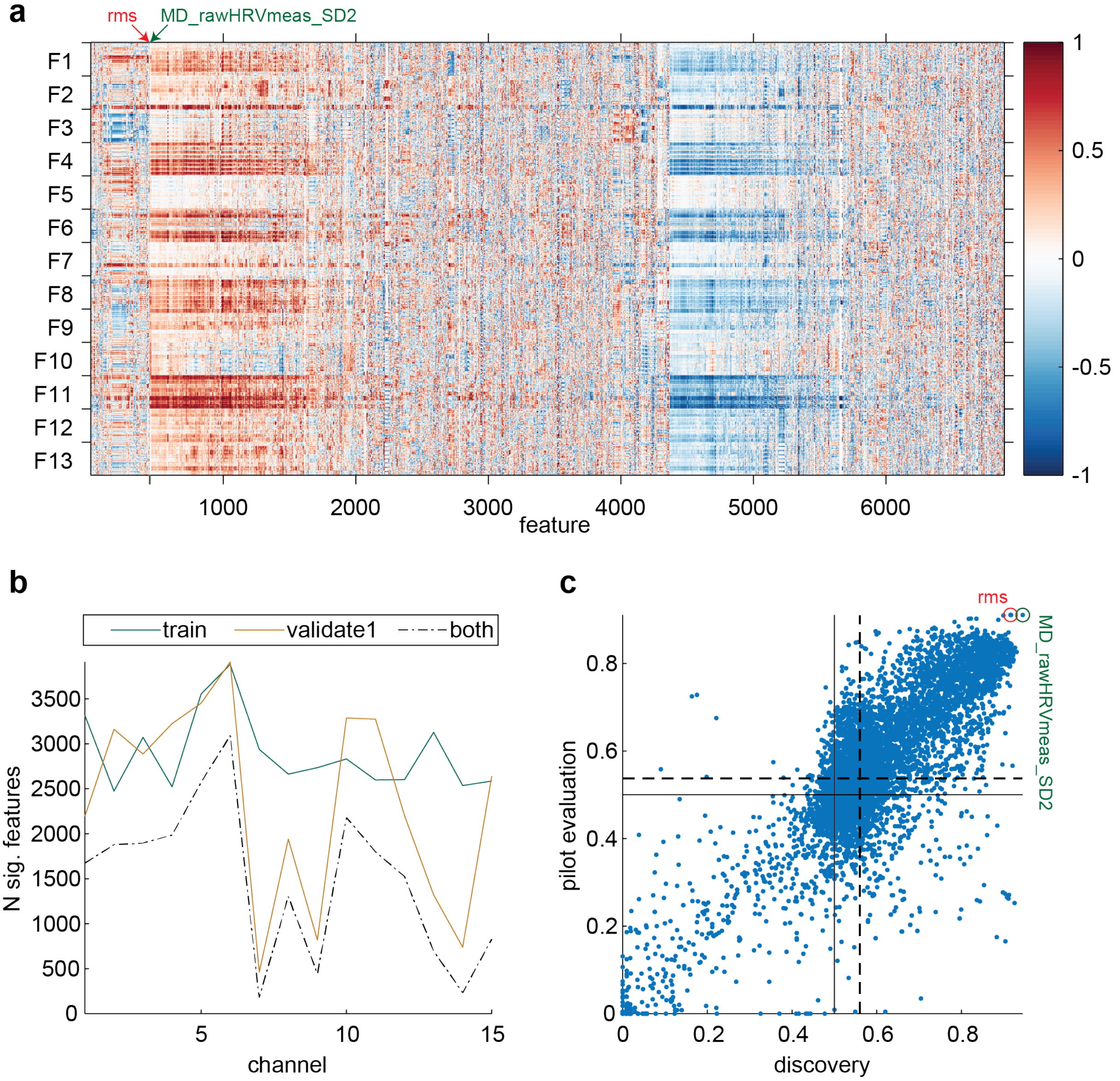

Figure 3. Within-fly effect direction (wake - anesthesia) consistency of *hctsa* features. **a)** Differences in scaled *hctsa* values between wakefulness and anesthesia in the 13 discovery flies (F), at Channel 6. Each row displays the difference between a wakeful and anesthesia epoch from the same fly, for all valid features for the channel. Features (columns) have the same ordering as in Figure 2a. Arrows indicate the features which attained the highest classification performance in the discovery (green) and pilot evaluation (red) flies. **b)** Number of features which achieved statistically significant consistency at each channel, in the discovery flies, pilot evaluation flies, and in both the discovery flies and pilot evaluation flies. **c)** Correlation of within-fly effect direction consistencies between the discovery (x-axis) and pilot evaluation flies (y-axis). Each dot represents the consistency of one of the 6800 features shown in **a**). Solid horizontal and vertical lines indicate chance consistency (= 0.5).

Dashed horizontal and vertical lines indicate the thresholds for statistically significant consistency in each set of flies (see Methods). Dots located in the top right quadrant are the features which were significantly consistent across both the discovery and pilot evaluation flies. Circled are the features which attained the highest consistency in the discovery (green) and pilot evaluation (red) flies.

References

1. Laureys S, Gosseries O, Tononi G. The neurology of consciousness: cognitive neuroscience and neuropathology. Academic Press; 2015.
2. Panayiotopoulos CP, Obeid T, Waheed G. Differentiation of typical absence seizures in epileptic syndromes: A video EEG study of 224 seizures in 20 patients. Brain. 1989;112: 1039–1056. doi:10.1093/brain/112.4.1039
3. Thomsen CE, Rosenfalck A, Christensen KN. Assessment of anaesthetic depth by clustering analysis and autoregressive modelling of electroencephalograms. Comput Methods Programs Biomed. 1991;34: 125–138. doi:10.1016/0169-2607(91)90038-U
4. Panayiotopoulos CP, Chroni E, Daskalopoulos C, Baker A, Rowlinson S, Walsh P. Typical absence seizures in adults: clinical, EEG, video-EEG findings and diagnostic/syndromic considerations. J Neurol Neurosurg Amp Psychiatry. 1992;55: 1002. doi:10.1136/jnnp.55.11.1002
5. Vuilleumier P, Assal F, Blanke O, Jallon P. Distinct Behavioral and EEG Topographic Correlates of Loss of Consciousness in Absences. Epilepsia. 2000;41: 687–693. doi:10.1111/j.1528-1157.2000.tb00229.x
6. Goldfine AM, Victor JD, Conte MM, Bardin JC, Schiff ND. Determination of awareness in patients with severe brain injury using EEG power spectral analysis. Clin Neurophysiol. 2011;122: 2157–2168. doi:10.1016/j.clinph.2011.03.022
7. Murphy M, Bruno M-A, Riedner BA, Boveroux P, Noirhomme Q, Landsness EC, et al. Propofol Anesthesia and Sleep: A High-Density EEG Study. Sleep. 2011;34: 283–291. doi:10.1093/sleep/34.3.283
8. Colombo MA, Napolitani M, Boly M, Gosseries O, Casarotto S, Rosanova M, et al. The spectral exponent of the resting EEG indexes the presence of consciousness during unresponsiveness induced by propofol, xenon, and ketamine. NeuroImage. 2019;189: 631–644. doi:10.1016/j.neuroimage.2019.01.024
9. Bruhn J, Röpcke H, Rehberg B, Bouillon T, Hoeft A. Electroencephalogram Approximate Entropy Correctly Classifies the Occurrence of Burst Suppression Pattern as Increasing Anesthetic Drug Effect. Anesthesiology. 2000;93: 981–985. doi:10.1097/00000542-200010000-00018
10. Liang Z, Wang Y, Ouyang G, Voss LJ, Sleigh JW, Li X. Permutation auto-mutual information of electroencephalogram in anesthesia. J Neural Eng. 2013;10: 026004. doi:10.1088/1741-2560/10/2/026004
11. Liang Z, Wang Y, Sun X, Li D, Voss LJ, Sleigh JW, et al. EEG entropy measures in anesthesia. Front Comput Neurosci. 2015;9: 16. doi:10.3389/fncom.2015.00016
12. Muñoz RN, Leung A, Zecevik A, Pollock FA, Cohen D, van Swinderen B, et al. General anesthesia reduces complexity and temporal asymmetry of the informational structures derived from neural recordings in Drosophila. Phys Rev Res. 2020;2: 023219. doi:10.1103/PhysRevResearch.2.023219
13. Sitt JD, King J-R, El Karoui I, Rohaut B, Faugeras F, Gramfort A, et al. Large scale screening of neural signatures of consciousness in patients in a vegetative or minimally conscious state. Brain. 2014;137: 2258–2270. doi:10.1093/brain/awu141
14. Engemann DA, Raimondo F, King J-R, Rohaut B, Louppe G, Faugeras F, et al. Robust EEG-based cross-site and cross-protocol classification of states of consciousness. Brain. 2018;141: 3179–3192. doi:10.1093/brain/awy251
15. Forrest FC, Tooley MA, Saunders PR, Prys-Roberts C. Propofol infusion and the suppression of consciousness: the EEG and dose requirements. Br J Anaesth. 1994;72: 35–41. doi:10.1093/bja/72.1.35
16. Schwilden H, Stoeckel H, Schüttler J. Closed-loop feedback control of propofol

anaesthesia by quantitative EEG analysis in humans. Br J Anaesth. 1989;62: 290–296. doi:10.1093/bja/62.3.290

17. Decat N, Walter J, Koh ZH, Sribanditmongkol P, Fulcher BD, Windt JM, et al. Beyond traditional visual sleep scoring: massive feature extraction and unsupervised clustering of sleep time series. bioRxiv. 2021; 2021.09.08.458981. doi:10.1101/2021.09.08.458981
18. Fulcher BD, Jones NS. hctsa: A Computational Framework for Automated Time-Series Phenotyping Using Massive Feature Extraction. Cell Syst. 2017;5: 527-531.e3. doi:10.1016/j.cels.2017.10.001
19. Nahian MdJA, Ghosh T, Banna MdHA, Aseeri MA, Uddin MN, Ahmed MR, et al. Towards an Accelerometer-Based Elderly Fall Detection System Using Cross-Disciplinary Time Series Features. IEEE Access. 2021;9: 39413–39431. doi:10.1109/ACCESS.2021.3056441
20. Schreglmann SR, Wang D, Peach RL, Li J, Zhang X, Latorre A, et al. Non-invasive suppression of essential tremor via phase-locked disruption of its temporal coherence. Nat Commun. 2021;12: 363. doi:10.1038/s41467-020-20581-7
21. Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI. Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci. 2009;12: 535–540. doi:10.1038/nn.2303
22. Ferdinandy B, Gerencsér L, Corrieri L, Perez P, Újváry D, Csizmadia G, et al. Challenges of machine learning model validation using correlated behaviour data: Evaluation of cross-validation strategies and accuracy measures. PLOS ONE. 2020;15: e0236092. doi:10.1371/journal.pone.0236092
23. Wong W, Noreika V, Móró L, Revonsuo A, Windt J, Valli K, et al. The Dream Catcher experiment: blinded analyses failed to detect markers of dreaming consciousness in EEG spectral power. Neurosci Conscious. 2020;2020. doi:10.1093/nc/niaa006
24. Mashour GA, Lydic R. Neuroscientific Foundations of Anesthesiology. Oxford University Press, USA; 2011.
25. Alivisatos AP, Chun M, Church GM, Greenspan RJ, Roukes ML, Yuste R. The brain activity map project and the challenge of functional connectomics. Neuron. 2012;74: 970–974. doi:10.1016/j.neuron.2012.06.006
26. Herculano-Houzel S. The human brain in numbers: a linearly scaled-up primate brain. Front Hum Neurosci. 2009;3: 31. doi:10.3389/neuro.09.031.2009
27. Shaw PJ, Cirelli C, Greenspan RJ, Tononi G. Correlates of sleep and waking in Drosophila melanogaster. Science. 2000;287: 1834–1837.
28. Cirelli C, Bushey D. Sleep and wakefulness in Drosophila melanogaster. Ann N Y Acad Sci. 2008;1129: 323–329. doi:10.1196/annals.1417.017
29. Kirszenblat L, van Swinderen B. The yin and yang of sleep and attention. Trends Neurosci. 2015;38: 776–786. doi:10.1016/j.tins.2015.10.001
30. Geissmann Q, Beckwith EJ, Gilestro GF. Most sleep does not serve a vital function: Evidence from Drosophila melanogaster. Sci Adv. 2019;5: eaau9253. doi:10.1126/sciadv.aau9253
31. Zalucki O, Van Swinderen B. What is unconsciousness in a fly or a worm? A review of general anesthesia in different animal models. Conscious Cogn. 2016;44: 72–88.
32. Cohen D, van Swinderen B, Tsuchiya N. Isoflurane impairs low frequency feedback but leaves high frequency feedforward connectivity intact in the fly brain. eNeuro. 2018; ENEURO.0329-17.2018.
33. Cohen D, Zalucki OH, van Swinderen B, Tsuchiya N. Local versus global effects of isoflurane anesthesia on visual processing in the fly brain. eneuro. 2016;3: ENEURO.0116-16.2016. doi:10.1523/ENEURO.0116-16.2016

34. Leung A, Cohen D, Swinderen B van, Tsuchiya N. Integrated information structure collapses with anesthetic loss of conscious arousal in Drosophila melanogaster. PLOS Comput Biol. 2021;17: e1008722. doi:10.1371/journal.pcbi.1008722

35. Boly M, Moran R, Murphy M, Boveroux P, Bruno M-A, Noirhomme Q, et al. Connectivity changes underlying spectral EEG changes during propofol-induced loss of consciousness. J Neurosci. 2012;32: 7082–7090. doi:10.1523/JNEUROSCI.3769-11.2012

36. Lee U, Blain-Moraes S, Mashour GA. Assessing levels of consciousness with symbolic analysis. Philos Trans R Soc Math Phys Eng Sci. 2015;373: 20140117. doi:10.1098/rsta.2014.0117

37. Schartner M, Seth A, Noirhomme Q, Boly M, Bruno M-A, Laureys S, et al. Complexity of Multi-Dimensional Spontaneous EEG Decreases during Propofol Induced General Anaesthesia. PLOS ONE. 2015;10: e0133532. doi:10.1371/journal.pone.0133532

38. Casarotto S, Comanducci A, Rosanova M, Sarasso S, Fecchio M, Napolitani M, et al. Stratification of unresponsive patients by an independently validated index of brain complexity. Ann Neurol. 2016;80: 718–729. doi:10.1002/ana.24779

39. Lee U, Kim S, Noh G-J, Choi B-M, Hwang E, Mashour GA. The directionality and functional organization of frontoparietal connectivity during consciousness and anesthesia in humans. Conscious Cogn. 2009;18: 1069–1078. doi:10.1016/j.concog.2009.04.004

40. Zalucki O, Day R, Kottler B, Karunanithi S, van Swinderen B. Behavioral and electrophysiological analysis of general anesthesia in 3 background strains of Drosophila melanogaster. Fly (Austin). 2015;9: 7–15. doi:10.1080/19336934.2015.1072663

41. Mitra PP, Bokil HS. Observed brain dynamics. Oxford University Press; 2007.

42. Fulcher BD, Little MA, Jones NS. Highly comparative time-series analysis: the empirical structure of time series and their methods. J R Soc Interface. 2013;10: 20130048. doi:10.1098/rsif.2013.0048

43. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B Methodol. 1995;57: 289–300. doi:10.1111/j.2517-6161.1995.tb02031.x

44. Pincus SM, Cummins TR, Haddad GG. Heart rate control in normal and aborted-SIDS infants. Am J Physiol-Regul Integr Comp Physiol. 1993;264: R638–R646. doi:10.1152/ajpregu.1993.264.3.R638

45. Brennan M, Palaniswami M, Kamen P. Do existing measures of Poincare plot geometry reflect nonlinear features of heart rate variability? IEEE Trans Biomed Eng. 2001;48: 1342–1347. doi:10.1109/10.959330

4.2 - Discussion

In this stage 1 registered report, I applied a vast library of univariate time-series features as candidate measures of conscious level. In the pilot results, I find that many features are able to distinguish wakefulness from anaesthesia in the discovery set of flies (the same set of flies as previously analysed in the previous chapters). However, many of these features fail to generalise to a new set of flies, where details regarding recording and experimental methods are not known a priori. Meanwhile, the direction of the effect of anaesthesia, within each individual fly, is consistent across both sets of flies for many more features. While a full discussion on promising time-series features as measures of consciousness will depend on completing the analyses in the final evaluation flies after in-principle acceptance of the registered report, here I provide some early interpretations regarding how the current results compare to IIT's measures. I focus on the results from discovery flies, as these were the only flies analysed in the previous chapters.

The results presented here provide a benchmark performance for assessing the utility of IIT's measures in distinguishing wakefulness from anaesthesia. The cross-validated classification accuracy of the top performing features in the discovery flies (76% at channel 6) was higher than the greatest accuracy achieved by integrated information (Φ; < 75% in Figure 5A of Chapter 2), but lower than the greatest accuracies achieved by the associated integrated information structure (IIS; > 80% in Figure 5A of Chapter 2). This comparison between the classification performances of the top performing features and IIT's measures demonstrate the advantage of the first-principles approach taken by IIT. Specifically, the first-principles approach provides a meaningful measure which already performs comparable to or outperforms an arbitrary but vast selection of alternative candidate measures. I note however that the classification accuracies from the two chapters are not conceptually comparable due to two key differences in analysis methods.

The first difference is in the exact analysis framework used. In Chapter 2, I trained classifiers either using only multiple epochs from an individual fly at a time (within-fly classification; 8 epochs), or using single epochs from multiple flies at a time (across-fly classification; 13 epochs). However, in the present chapter, for the purpose of registering a more streamlined (and standalone) analysis, I trained a single classifier for each time-series feature using multiple epochs from multiple flies (8 × 13 epochs). Consequently, each individual classifier in Chapter 2 was trained on many fewer epochs. This may have reduced their overall performance (Kotsiantis et al., 2006; Zhu et al., 2016). This issue can be addressed by conducting within-fly and across-fly classification on the time-series features as additional exploratory analyses, after in-principle acceptance of the registered report.

The second difference is the comparison of the univariate time-series features to IIT's multivariate measures. IIT's measures are computed from multiple channels, with each Φ value or IIS being computed from the time-series of the channels being considered. Consequently, their performance in distinguishing between wakefulness and anaesthesia is linked to a particular set of channels. Meanwhile, the accuracies that I report for each time-series feature is linked to a particular channel. Thus, comparing IIT's measures to each univariate time-series feature entails comparing the distributions of accuracies achieved by Φ and the associated IIS across channel sets to the accuracies achieved by the time-series feature at each individual channel. A more direct comparison might be possible if the time-series features are collated across multiple channels. However, there are many conceivable ways to collate across channels, such as by averaging (such as in Sitt et al., 2014, and Engemann et al., 2018), or by directly training multivariate classifiers. Another direction to address this would be to evaluate multivariate time-series features, such as cross-correlation or mutual information (Sitt et al., 2014; Cliff et al., 2021), though to my knowledge there is currently no multivariate analogue to *hctsa* which provides a library of multivariate measures.

To conclude, IIT's measures, specifically its cause-effect structures, discriminates wakefulness from anaesthesia with performance comparable to or even exceeding the best performing univariate time-series features. While differences in analysis methods muddy the comparison of classification accuracies presented here and in Chapter 2, there are clear paths to take in the immediate future to address them. The registered report format used here also provides a future avenue to further evaluate IIT's measures. Specifically, the registered report format mitigates issues relating to publication bias or p-hacking (Botvinik-Nezer et al., 2020; Soderberg et al., 2021). Thus, testing generalisability by applying predetermined analyses to independent, blinded datasets can be applied as a more rigorous framework for evaluating the validity and utility of Φ and associated cause-effect structures.

4.3 – References

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88.

Cliff, O. M., Novelli, L., Fulcher, B. D., Shine, J. M., & Lizier, J. T. (2021). Assessing the significance of directed and multivariate measures of linear dependence between time series. *Physical Review Research*, *3*(1), 013145.

Engemann, D. A., Raimondo, F., King, J.-R., Rohaut, B., Louppe, G., Faugeras, F., Annen, J., Cassol, H., Gosseries, O., Fernandez-Slezak, D., Laureys, S., Naccache, L., Dehaene, S., & Sitt, J. D. (2018). Robust EEG-based cross-site and cross-protocol classification of states of consciousness. *Brain*, *141*(11), 3179–3192.

Fulcher, B. D., & Jones, N. S. (2017). hctsa: A Computational Framework for Automated Time-Series Phenotyping Using Massive Feature Extraction. *Cell Systems*, *5*(5), 527-531.e3.

Kotsiantis, S. B., Zaharakis, I. D., & Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, *26*(3), 159–190.

Sitt, J. D., King, J.-R., El Karoui, I., Rohaut, B., Faugeras, F., Gramfort, A., Cohen, L., Sigman, M., Dehaene, S., & Naccache, L. (2014). Large scale screening of neural signatures of consciousness in patients in a vegetative or minimally conscious state. *Brain*, *137*(8), 2258–2270.

Soderberg, C. K., Errington, T. M., Schiavone, S. R., Bottesini, J., Thorn, F. S., Vazire, S., Esterling, K. M., & Nosek, B. A. (2021). Initial evidence of research quality of registered reports compared with the standard publishing model. *Nature Human Behaviour*, *5*(8), 990–997.

Zhu, X., Vondrick, C., Fowlkes, C. C., & Ramanan, D. (2016). Do We Need More Training Data? *International Journal of Computer Vision*, *119*(1), 76–92.

Chapter 5 - General Discussion

5.1 - Summary of findings

In the preceding chapters, I applied the measures put forward by the integrated information theory of consciousness (IIT) 3.0, to neural recordings obtained from the fly brain during wakefulness and anaesthesia. In Chapter 2, I tested IIT's principle prediction that integrated information (Φ) should be high during wakefulness and low during loss of consciousness. Consistent with this expectation, I did indeed find Φ to be greater in the fly brain during wakefulness, when compared to during anaesthesia. I further found associated integrated information structures (IIS), a proxy measure of the associated cause-effect structures proposed by IIT as being linked to contents of conscious experience, to also be reduced in the fly brain during anaesthesia. In Chapter 3, I tested IIT's prediction that Φ should be maximal at the temporal scale corresponding to conscious experience, rather than at some minimal, micro timescale as might be expected from reductionist views. Again, meeting with IIT's expectation, I found Φ to be maximal at a scale consistent with the timescale of neurophysiological interactions. Finally, in Chapter 4, I applied a vast library of univariate time-series features as alternate candidate measures, to evaluate any advantage of IIT's measures which result from IIT's introspective and theoretical approach. While the best performing univariate time-series features distinguished wakefulness from anaesthesia with similar performance as IIT's measures, their simplicity makes them difficult to link back to how consciousness is generated from physical interactions. While these results provide empirical support for IIT, there is still much more to investigate.

5.2 - Integrated information measures in the fly brain behave as expected without perturbation or identifying the complex

While I have attempted to apply IIT 3.0's measures as they are directly operationalised by the theory, clearly it is currently infeasible to apply them exactly as they are defined. Two issues prevent a completely faithful application of the measures. The first is the requirement of perturbation of a system into all its states. The second is the search for the set of elements which constitute the complex across spatial and temporal scales. Both of them are impractical to apply to neural data from biological brains. However, from the results I have presented in this thesis, the measures proposed by IIT 3.0 seem to perform largely as expected by the theory even without directly perturbing the system or searching for the complex.

An optimistic interpretation of this might be that Φ and the IIS can have utility in determining conscious level even when applied only to subsets of a candidate system, and at spatiotemporal scales which do not necessarily maximise Φ. However, it is unclear whether the measures would still behave as expected when applying the same analyses to larger systems. Rather, despite not being applied completely faithfully to the theory, Φ and the IIS behaving as expected could be attributed to the use of the fly brain. Specifically, the spatial scale and temporal dynamics of the small number of local field potentials (LFPs) analysed in this thesis may already be relatively close to the spatial and temporal scales of the complex in the fly brain. Also, the extent of the recordings, which spanned throughout the fly brain, could be close to capturing all the interactions of the complex. However, for a larger system, such as even the mouse brain, small numbers of LFPs might not be as representative of dynamics throughout the whole brain. Consequently, observable LFPs may only capture a much smaller subset of dynamics within the complex. Meanwhile, it could be the case that recordings at a larger spatial scale, such as electrocorticographic or electroencephalographic recordings, may be too coarse grained. Consequently, whether the results here will generalise to larger brains, when applying similar methods to a similar number of recordings, remains unclear.

Even with the hypothetical capability to capture more recordings spanning a larger brain, the application of IIT's measures remain limited. This is due to the computational cost of searching for minimum information partitions (MIPs). Specifically, when computing Φ, the search for the MIP at the system level requires repeatedly computing the cause-effect structures for every partition of the system. However, in Chapter 2, I found that the IIS was able to discriminate between wakefulness and anaesthesia more reliably than Φ. This suggests that the repeated computation of cause-effect structures may be unnecessary for the practical purpose of measuring the level of consciousness in a subject, when ignoring the precise location of the complex. While this already substantially reduces computational costs of applying IIT, a similar, expensive, search for MIPs, at the level of mechanisms, is still required for every mechanism (in combination with every purview) in order to obtain the IIS. To identify these MIPs, one must search through all partitions of a mechanism and its purview (Oizumi et al., 2014). However, this is only necessary when the connectivity among elements is unknown a priori. Knowledge of how elements are connected can allow for the deduction or more restricted search of the MIP (Kitazono et al., 2018; Hidaka & Oizumi, 2018), allowing one to avoid a costly search. The fly brain provides an opportunity to take advantage of this, specifically through consistent progress towards a comprehensive mapping of the fly connectome (Zheng et al., 2018; Scheffer et al., 2020). Knowledge of connections also allows for an easier search for the spatiotemporal scale at which Φ is maximised. Particular connectivity patterns have been used to illustrate emergence of Φ at macro scales (Hoel et al., 2013, 2016; Marshall et al., 2018), and these may serve as candidate network motifs to search for through the fly brain connectome.

Thus, given current experimental and computational capabilities for computing Φ and the IIS, the fly brain serves as a key system where the measures of IIT can be most faithfully applied. Its relative simplicity with regards to the number of neurons it consists of, and comprehensive knowledge of how they are connected gives a promising system in which to continue evaluating the ideas of IIT in the future.

5.3 - Assessing varied levels of consciousness

In this thesis, I have focussed on applying and evaluating IIT in the fly brain during two levels of consciousness - wakefulness and anaesthesia. Specifically, anaesthesia was induced at a specific concentration of isoflurane anaesthesia, and for this particular administration of anaesthesia, I found Φ and the associated IIS to be reduced. However, as measures of consciousness, this reduction should generalise to reduced levels of consciousness induced by other means.

Firstly, the reduction should generalise to loss of consciousness as induced by other anaesthetic drugs. Whether it does, though, is not necessarily immediately clear. Different anaesthetic drugs induce anaesthesia through differing molecular mechanisms (Hines & van Swinderen, 2021). Consequently, neural signals recorded during anaesthesia can have different temporal properties depending on the type of drug used (Purdon et al., 2015). IIT's measures may be sensitive to these differences, as the transition probability matrices (TPMs) used for computing the measures are constructed directly from such recordings. On the other hand, a common endpoint for general anaesthesia is the disruption of communication across the brain. If IIT's measures ultimately reflect global communication, small differences in time-series properties may be abstracted out by IIT's measures. Meanwhile, univariate time-series features which are not linked to consciousness through first principles would be sensitive to such differences in temporal properties. So, while different features might successfully distinguish wakefulness from anaesthesia for one particular type of anaesthesia, they may fail to generalise across anaesthetic drugs.

Secondly, the reduction in IIT's measures should also be observed not only with complete loss of consciousness but also in graded reductions in consciousness. General anaesthesia is not binary, with a subject being only either clearly awake or clearly unconscious. Depending both on the type of anaesthetic drug and its dosage, subjects can enter a variety of states, such as being sedated but still behaviourally responsive (Guedel, 1937), being paradoxically excited (Jeong et al., 2011; Zuleta-Alarcon et al., 2014), being intraoperatively aware (Mashour et al., 2011), or dreaming (Brandner et al., 1997; Leslie et al., 2009; Sarasso et al., 2015). These states involve conscious experience, albeit perhaps a reduced, less rich experience compared to that during full wakefulness. Correspondingly, IIT expects its measures to vary with regard to the exact depth and state of anaesthesia. For example, as a subject gradually is sedated with increasing dosage until they lose consciousness, Φ should gradually decrease. Similar as for anaesthesia induced through different drugs, these states are associated with electrophysiological recordings with different spectral properties. Again, these differences might be abstracted out in IIT's measures, but not in non-theory driven measures. For example, in Chapter 4 I found univariate measures dealing in variance of a time-series to perform well in distinguishing wakefulness from consciousness. However, the amplitudes of recordings during anaesthesia can be greater or lower than that during wakefulness depending on the depth of anaesthesia (Purdon et al., 2015). Consequently, variance related measures might fail in correctly identifying both states of isoelectricity (where amplitude fluctuations are much lower than during wakefulness) and surgical levels of anaesthesia (where amplitude fluctuations are greater than during wakefulness) as states of anaesthesia rather than wakefulness.

Extending beyond anaesthesia, IIT's measures should also be reduced in the case of sleep. Stages of sleep are associated with graded levels of consciousness, similar to how different anaesthetic drugs and dosages can induce both reduced and complete loss of consciousness. In humans, rapid eye movement (REM) sleep for example is associated with dreaming, while deep, slow wave sleep (SWS) is associated with complete loss of consciousness (Windt et al., 2016). Flies are known to exhibit sleep-like behaviours similar to mammals (Hendricks et al., 2000; Shaw et al., 2000), with distinct stages of sleep (van Alphen et al., 2013; Tainton-Heap et al., 2021). These stages include a stage of active sleep, where the fly brain exhibits activity similar to during wakefulness while being disconnected from the external environment, and deep sleep, where the brain exhibits activity similar to SWS in mammals. If active and deep sleep in the fly induce reduced and loss of consciousness similar to REM and SWS sleep in mammals, then Φ ought to gradually reduce from wakefulness to active sleep and finally deep sleep. Such a graded decrease has been observed in humans and rats using measures inspired from IIT (Massimini et al., 2010; Abásolo et al., 2015; Andrillon et al., 2016), but not for the measures as defined directly by IIT itself.

5.4 - Assessing contents of consciousness

While in this thesis I have focussed only on distinguishing levels of consciousness, IIT deals also with the contents of consciousness - any conscious system must after all be experiencing and conscious of something. IIT posits that the maximally irreducible cause-effect structure (MICS) characterises what a given system is experiencing. As seen in Chapter 2, the proxy of the MICS which I use, the IIS, does indeed reduce during anaesthesia. However, does it also correspond to the experiences of the fly?

This is currently a difficult question to address. While the fly brain provides many advantages over human brains for applying the ideas of IIT, it has the severe drawback in that we may be unable to characterise the phenomenology of the fly. Indeed, this problem extends already to essentially any non-human system (Nagel, 1974), where we lack the ability to communicate in order to understand a subject's experience. One might try to make progress towards characterising experience through behaviours in response to different stimuli. Such basic psychophysics has been employed in flies, characterising perception such as brightness, colour, size, and figure-ground discrimination (Menne & Spatz, 1977; Reichert & Bicker, 1979; Fresquet & Médioni, 1993; Grabowska et al., 2018; Aptekar et al., 2015).

However, current methods of understanding a subject's experience require more than first-order behavioural responses to stimuli. For example, in humans, blindsight patients can discriminate facial emotions without consciously perceiving them (Morris et al., 2001), and even successfully navigate around obstacles despite lacking visual experience of them (de

Gelder et al., 2008; Striemer et al., 2009). For such cases, second-order report, metacognition, where subjects report how confident they are in their task choices, is used in addition to first-order responses to help inform whether stimuli are consciously perceived or not (Fleming & Lau, 2014). While metacognition has not been investigated in flies, insects such as honey bees have been reported to demonstrate it (Perry & Barron, 2013). As the neural substrate proposed for supporting metacognition in honey bees, specifically neural interactions in the mushroom bodies (Perry & Barron, 2013), also exists in flies, investigating metacognition in flies might yield progress towards understanding their experiential contents.

Meanwhile, an additional problem is that of directly relating the cause-effect structures put forward by IIT with phenomenology. Specifically, for cause-effect structures to be a compelling measure, they ought to go beyond simply correlating with specific conscious experiences. In a similar manner as for Φ in the context of varied levels of consciousness, cause-effect structures need to generalise beyond simple correlations such as those which might also be obtained by non-theory driven measures. One proposed approach to address this is to use category theory (Tsuchiya & Saigo, 2021), whereby a strong notion of equivalence, categorical equivalence, can be assessed between experiences and cause-effect structures. Critical to this approach is the characterisation of relationships between different experiences, together with relationships between cause-effect structures, above simply matching cause-effect structures to experiences. Similarity ratings, subjective reports reflecting how similar an experience is to another (Hiramatsu et al., 2011; Kriegeskorte & Kievit, 2013), have been proposed as a potential way of characterising these relationships (Fink et al., 2021; Tsuchiya & Saigo, 2021). While, to my knowledge, similarity ratings have not been obtained from flies, it might be possible to infer similarities of perceived stimuli by finding just noticeable differences (Stevens, 1957) between pairs of stimuli, in combination with metacognition.

Overall, assessing whether the MICS as proposed by IIT corresponds to conscious experiences is potentially more difficult than assessing whether Φ corresponds to levels of consciousness. Current and prospective methods for understanding the experiences of a system rely on introspective, behavioural reports which are most understandable for more complex systems similar to ourselves. However, whether IIT's ideas can be applied faithfully for these systems is unclear. Meanwhile for the fly brain, one currently needs to rely solely on behaviours which are unrelatable to ourselves to infer the fly's experiences. Developing paradigms to capture metacognition and similarity ratings in flies may allow us to better understand the conscious experience of a fly and then evaluate the construct validity of the MICS. Failing this, a middle ground for evaluating the construct validity of the MICS might be slightly larger, but still relatively small and simple systems, such as the honey bee brain.

5.5 - Updates to integrated information theory

It should be noted that IIT 3.0 is not the end-all of integrated information theory. Theoretical concepts which are identified as problematic have been and are being addressed in order to build a more sound and comprehensive theory. Consequently, IIT is actively being revised and updated, and updated versions of IIT also require empirical evaluation.

One such issue which has since been revised is that regarding the partitioning of 1-channel mechanisms. As brought up in Chapter 2, IIT 3.0 does not require partitioning a mechanism in order to assess integrated information φ for that mechanism. Rather, partitions are only required to separate a mechanism-purview combination such that some part of the purview has its connection from the mechanism noised. This seems to be logically inconsistent with the aim of IIT 3.0's integration postulate, which is to assess whether a mechanism contributes more than its constituent parts. Accordingly, the requirements for partitions used to assess φ have since been updated (Albantakis et al., 2019). Specifically, as a single channel has no constituent parts, the only partition which should be used to assess φ of 1-channel mechanisms is that where all connections from the mechanism to the purview are noised.

Another issue deals with the uniqueness of IIT's postulates. Though the core aspects of experience, as identified by IIT have gone fairly unchallenged (beyond relatively small issues of semantics regarding how the axioms should be interpreted; Bayne, 2018), the postulates and derivation of measures put forward by IIT have faced more criticism. Specifically, the postulates of IIT 3.0 do not all uniquely follow from axioms. One example is the use of the earth mover's distance (EMD) to characterise "information" (distance between a probability distributions). While EMD is explicitly chosen over standard information theoretical measures, such as Kullback-Leibler divergence as used in earlier versions of IIT (Tononi, 2004, 2008), due to its metric properties (Oizumi et al., 2014), there is no principled reason from IIT's axioms as to why it should be used over some other measure (Cerullo, 2015; Tegmark, 2016; Barrett & Mediano, 2019). To address this, a new measure of information has been proposed, intrinsic difference, which is derived taking into account IIT's intrinsic axiom (Barbosa et al., 2020), and which will likely be used in future versions of IIT in place of EMD.

A third type of revision is the addition of ideas which are lacking in IIT 3.0. While IIT 3.0 derives a measure of conscious contents, a cause-effect structure consisting of separate mechanisms which exist intrinsically to the system, it is not explicit as to *how* they exist to the system. Specifically, how do mechanisms, being in some particular states, come together to form a particular, unitary experience? To try and address this question, the notion of "relations", which were originally already described to a small extent in previous versions of IIT (Tononi, 2008) but left out in IIT 3.0, have been formalised in a more updated framework (Haun & Tononi, 2019). Relations aim to describe how mechanisms overlap with each other to give a unified experience with complex components such as, for example, some arbitrarily shaped boundary in the visual field. Updates such as this are likely to be very important in the context of assessing the link between cause-effect structures and experiences, as they provide tools to better describe how IIT's ideas come together to support some specific kind of experience.

All the above updates seemingly only add to the complexities, and consequently computational costs, of IIT. So, a last "improvement" for the theory would be the provision of alternate, cheaper measures which still capture the core ideas of those directly put forward by it. While previous versions of IIT have had multiple "improved" derivations of Φ proposed, allowing for greater feasibility in applying IIT to real data, to my knowledge no such alternate derivation of Φ based on IIT 3.0 has been put forward. Existing alternate measures (Barrett & Seth, 2011; Oizumi, Tsuchiya, et al., 2016; Oizumi, Amari, et al., 2016; Tegmark, 2016; Mediano et al., 2019) lack the improvements of IIT 3.0, such as the consideration of both causes and effects of a system instead of only its causes (Tononi, 2004, 2008), and completely ignore the contents of consciousness, focussing only on approximating Φ. As the infeasible-to-apply-fully IIT becomes even more infeasible due to revisions and additions such as those described above, "improved" derivations of Φ and the MICS for IIT 3.0 and subsequent versions of the theory would be of great benefit towards applying the more up-to-date notions of IIT to real data.

5.6 - Concluding remarks

The integrated information theory of consciousness approaches the question of how physical interactions can result in consciousness by taking an introspective approach. By identifying core aspects of consciousness and from these postulating the necessary physical interactions to support them, it gives clear predictions regarding the behaviour of its proposed measures as consciousness varies in a system.

In this thesis, I have worked towards assessing the construct validity of the theory's measures, by applying the measures proposed in its current iteration, IIT 3.0, to recordings from the fly brain. Consistent with expectations from the theory, its measures, both Φ and its associated cause-effect structure, distinguish consciousness from anaesthesia, and Φ seems to be maximal at some non-minimal temporal scale. However, in the context of practicality, much simpler measures seem to also distinguish wakefulness from anaesthesia with similar performance, despite theoretical links between them and consciousness being unclear.

Overall, there remains much work to be done. As a theory of consciousness, IIT deals not only with levels of consciousness, but also contents of consciousness. Whether its proposed measure for contents of consciousness, the cause-effect structure of a system, accurately captures highly varied conscious experience needs to be assessed. Next, the theory itself is currently undergoing updates and revisions so that its postulates more accurately deal with and reflect the core aspects of conscious experience. Measures incorporating these updates would need empirical assessment. Finally, IIT faces issues with practicality. Even for the fly brain, it is currently infeasible to apply the theory's measures to the whole brain at the level of individual neurons. Whether this issue can be overcome through breakthroughs in theory or technology and computing remains to be seen.

5.7 - References

Abásolo, D., Simons, S., da Silva, R. M., Tononi, G., & Vyazovskiy, V. V. (2015). Lempel-Ziv complexity of cortical activity during sleep and waking in rats. *Journal of Neurophysiology*, *113*(7), 2742–2752.

Albantakis, L., Marshall, W., Hoel, E., & Tononi, G. (2019). What caused what? A quantitative account of actual causation using dynamical causal networks. *Entropy*, *21*(5), 459.

Andrillon, T., Poulsen, A. T., Hansen, L. K., Léger, D., & Kouider, S. (2016). Neural Markers of Responsiveness to the Environment in Human Sleep. *Journal of Neuroscience*, *36*(24), 6583–6596.

Aptekar, J. W., Keleş, M. F., Lu, P. M., Zolotova, N. M., & Frye, M. A. (2015). Neurons Forming Optic Glomeruli Compute Figure–Ground Discriminations in Drosophila. *Journal of Neuroscience*, *35*(19), 7587–7599.

Barbosa, L. S., Marshall, W., Streipert, S., Albantakis, L., & Tononi, G. (2020). A measure for intrinsic information. *Scientific Reports*, *10*(1), 18803.

Barrett, A. B., & Mediano, P. A. M. (2019). The Phi Measure of Integrated Information is not Well-Defined for General Physical Systems. *Journal of Consciousness Studies*, *26*(1–2), 11–20.

Barrett, A. B., & Seth, A. K. (2011). Practical measures of integrated information for time-series data. *PLoS Computational Biology*, *7*(1), e1001052.

Bayne, T. (2018). On the axiomatic foundations of the integrated information theory of consciousness. *Neuroscience of Consciousness*, *2018*(1), niy007.

Brandner, B., Blagrove, M., McCallum, G., & Bromley, L. M. (1997). Dreams, images and emotions associated with propofol anaesthesia. *Anaesthesia*, *52*(8), 750–755.

Cerullo, M. A. (2015). The Problem with Phi: A Critique of Integrated Information Theory. *PLOS Computational Biology*, *11*(9), e1004286.

de Gelder, B., Tamietto, M., van Boxtel, G., Goebel, R., Sahraie, A., van den Stock, J., Stienen, B. M. C., Weiskrantz, L., & Pegna, A. (2008). Intact navigation skills after bilateral loss of striate cortex. *Current Biology*, *18*(24), R1128–R1129.

Fink, S. B., Kob, L., & Lyre, H. (2021). A structural constraint on neural correlates of consciousness. *Philosophy and the Mind Sciences*, *2*.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*, 443.

Fresquet, N., & Médioni, J. (1993). Effects of ageing on visual discrimination learning in Drosophila melanogaster. *The Quarterly Journal of Experimental Psychology Section B*, *46*(4), 399–412.

Grabowska, M. J., Steeves, J., Alpay, J., Van De Poll, M., Ertekin, D., & van Swinderen, B. (2018). Innate visual preferences and behavioral flexibility in Drosophila. *Journal of Experimental Biology*, *221*(23), jeb185918.

Guedel, A. E. (1937). Inhalation Anesthesia: A Fundamental Guide. *Anesthesia & Analgesia*, *16*(2), 119–120.

Haun, A., & Tononi, G. (2019). Why does space feel the way it does? Towards a principled account of spatial experience. *Entropy*, *21*(12), 1160.

Hendricks, J. C., Finn, S. M., Panckeri, K. A., Chavkin, J., Williams, J. A., Sehgal, A., & Pack, A. I. (2000). Rest in Drosophila is a sleep-like state. *Neuron*, *25*(1), 129–138.

Hidaka, S., & Oizumi, M. (2018). Fast and exact search for the partition with minimal information loss. *PLoS ONE*, *13*(9), e0201126.

Hines, A. D., & van Swinderen, B. (2021). Tracking Single Molecule Dynamics in the Adult Drosophila Brain. *ENeuro*, *8*(3), ENEURO.0057-21.2021.

Hiramatsu, C., Goda, N., & Komatsu, H. (2011). Transformation from image-based to perceptual representation of materials along the human ventral visual pathway. *NeuroImage*, *57*(2), 482–494.

Hoel, E. P., Albantakis, L., Marshall, W., & Tononi, G. (2016). Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neuroscience of Consciousness*, *2016*(1), niw012.

Hoel, E. P., Albantakis, L., & Tononi, G. (2013). Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*, *110*(49), 19790–19795.

Jeong, S., Lee, H. G., Kim, W. M., Jeong, C. W., Lee, S. H., Yoon, M. H., & Choi, J. I. (2011). Increase of paradoxical excitement response during propofol-induced sedation in hazardous and harmful alcohol drinkers. *BJA: British Journal of Anaesthesia*, *107*(6), 930–933.

Kitazono, J., Kanai, R., & Oizumi, M. (2018). Efficient algorithms for searching the minimum information partition in integrated information theory. *Entropy*, *20*(3), 173.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412.

Leslie, K., Sleigh, J., Paech, M. J., Voss, L., Lim, C. W., & Sleigh, C. (2009). Dreaming and Electroencephalographic Changes during Anesthesia Maintained with Propofol or Desflurane. *Anesthesiology*, *111*(3), 547–555.

Marshall, W., Albantakis, L., & Tononi, G. (2018). Black-boxing and cause-effect power. *PLoS Computational Biology*, *14*(4), e1006114.

Mashour, G. A., Orser, B. A., Avidan, M. S., & Warner, D. S. (2011). Intraoperative Awareness: From Neurobiology to Clinical Practice. *Anesthesiology*, *114*(5), 1218–1233.

Massimini, M., Ferrarelli, F., Murphy, M. J., Huber, R., Riedner, B. A., Casarotto, S., & Tononi, G. (2010). Cortical reactivity and effective connectivity during REM sleep in humans. *Cognitive Neuroscience*, *1*(3), 176–183.

Mediano, P. A. M., Seth, A. K., & Barrett, A. B. (2019). Measuring Integrated Information: Comparison of Candidate Measures in Theory and Simulation. *Entropy*, *21*(1), 17.

Menne, D., & Spatz, H.-C. (1977). Colour vision inDrosophila melanogaster. *Journal of Comparative Physiology*, *114*(3), 301–312.

Morris, J. S., DeGelder, B., Weiskrantz, L., & Dolan, R. J. (2001). Differential extrageniculostriate and amygdala responses to presentation of emotional faces in a cortically blind field. *Brain*, *124*(6), 1241–1252.

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*(4), 435–450.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Comput Biol*, *10*(5), e1003588.

Oizumi, M., Amari, S., Yanagawa, T., Fujii, N., & Tsuchiya, N. (2016). Measuring integrated information from the decoding perspective. *PLoS Computational Biology*, *12*(1), e1004654.

Oizumi, M., Tsuchiya, N., & Amari, S. (2016). Unified framework for information integration based on information geometry. *Proceedings of the National Academy of Sciences*, *113*(51), 14817–14822.

Perry, C. J., & Barron, A. B. (2013). Honey bees selectively avoid difficult choices. *Proceedings of the National Academy of Sciences*, *110*(47), 19155–19159.

Purdon, P. L., Sampson, A., Pavone, K. J., & Brown, E. N. (2015). Clinical Electroencephalography for Anesthesiologists: Part I: Background and Basic Signatures. *Anesthesiology*, *123*(4), 937–960.

Reichert, H., & Bicker, G. (1979). A visual learning study of brightness perception in two mutants of Drosophila melanogaster. *Journal of Comparative Physiology*, *133*(4), 283–290.

Sarasso, S., Boly, M., Napolitani, M., Gosseries, O., Charland-Verville, V., Casarotto, S., Rosanova, M., Casali, A. G., Brichant, J.-F., Boveroux, P., Rex, S., Tononi, G., Laureys, S., & Massimini, M. (2015). Consciousness and Complexity during Unresponsiveness Induced by Propofol, Xenon, and Ketamine. *Current Biology*, *25*(23), 3099–3105.

Scheffer, L. K., Xu, C. S., Januszewski, M., Lu, Z., Takemura, S., Hayworth, K. J., Huang, G. B., Shinomiya, K., Maitlin-Shepard, J., Berg, S., Clements, J., Hubbard, P. M., Katz, W. T., Umayam, L., Zhao, T., Ackerman, D., Blakely, T., Bogovic, J., Dolafi, T., … Plaza, S. M. (2020). A connectome and analysis of the adult Drosophila central brain. *ELife*, *9*, e57443.

Shaw, P. J., Cirelli, C., Greenspan, R. J., & Tononi, G. (2000). Correlates of sleep and waking in Drosophila melanogaster. *Science*, *287*(5459), 1834–1837.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review*, *64*(3), 153–181.

Striemer, C. L., Chapman, C. S., & Goodale, M. A. (2009). “Real-time” obstacle avoidance in the absence of primary visual cortex. *Proceedings of the National Academy of Sciences*, *106*(37), 15996–16001.

Tainton-Heap, L. A. L., Kirszenblat, L. C., Notaras, E. T., Grabowska, M. J., Jeans, R., Feng, K., Shaw, P. J., & van Swinderen, B. (2021). A Paradoxical Kind of Sleep in Drosophila melanogaster. *Current Biology*, *31*(3), 578-590.e6.

Tegmark, M. (2016). Improved measures of integrated information. *PLoS Computational Biology*, *12*(11), e1005123.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, *5*(1), 42.

Tononi, G. (2008). Consciousness as integrated information: A provisional manifesto. *The Biological Bulletin*, *215*(3), 216–242.

Tsuchiya, N., & Saigo, H. (2021). A relational approach to consciousness: Categories of level and contents of consciousness. *Neuroscience of Consciousness*, *2021*(2), niab034.

van Alphen, B., Yap, M. H. W., Kirszenblat, L., Kottler, B., & van Swinderen, B. (2013). A Dynamic Deep Sleep Stage in Drosophila. *The Journal of Neuroscience*, *33*(16), 6917.

Windt, J. M., Nielsen, T., & Thompson, E. (2016). Does Consciousness Disappear in Dreamless Sleep? *Trends in Cognitive Sciences*, *20*(12), 871–882.

Zheng, Z., Lauritzen, J. S., Perlman, E., Robinson, C. G., Nichols, M., Milkie, D., Torrens, O., Price, J., Fisher, C. B., Sharifi, N., Calle-Schuler, S. A., Kmecova, L., Ali, I. J., Karsh, B., Trautman, E. T., Bogovic, J. A., Hanslovsky, P., Jefferis, G. S. X. E., Kazhdan, M., … Bock, D. D. (2018). A complete electron microscopy volume of the brain of adult Drosophila melanogaster. *Cell*, *174*(3), 730-743.e22.

Zuleta-Alarcon, A., Castellon-Larios, K., Moran, K. R., Soghomonyan, S., Kurnutala, L. N., & Bergese, S. D. (2014). Anesthesia-related perioperative seizures: Pathophysiology, predisposing factors and practical recommendations. *Austin Journal of Anesthesia and Analgesia*, *2*(4), 1026.