Integrated Information Theory in Complex Neural Systems

Pedro Antonio Martínez Mediano

Department of Computing
Imperial College London

A thesis submitted for the degree of
Doctor of Philosophy

December 2019

Declaration of Originality

I hereby declare that the contents of this thesis is my own work, except where specific reference is made to the work of others.

Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Pedro Antonio Martínez Mediano
December 2019

Abstract

This thesis concerns *Integrated Information Theory* (IIT), a branch of information theory aimed at providing a fundamental theory of consciousness. At its core, lie two powerful intuitions:

- That a system that is somehow *more than the sum of its parts* has non-zero integrated information, Φ; and
- That a system with non-zero integrated information is conscious.

The audacity of IIT's claims about consciousness has (understandably) sparked vigorous criticism, and experimental evidence for IIT as a theory of consciousness remains scarce and indirect. Nevertheless, I argue that IIT still has merits as a theory of informational complexity within complexity science, leaving aside all claims about consciousness. In my work I follow this broad line of reasoning: showcasing applications where IIT yields rich analyses of complex systems, while critically examining its merits and limitations as a theory of consciousness.

This thesis is divided in three parts. First, I describe three example applications of IIT to complex systems from the computational neuroscience literature (coupled oscillators, spiking neurons, and cellular automata), and develop novel Φ estimators to extend IIT's range of applicability. Second, I show two important limitations of current IIT: that its axiomatic foundation is not specific enough to determine a unique measure of integrated information; and that available measures do not behave as predicted by the theory when applied to neurophysiological data.

Finally, I present new theoretical developments aimed at alleviating some of IIT's flaws. These are based on the concepts of partial information decomposition and lead to a unification of both theories, *Integrated Information Decomposition*, or ΦID. The thesis concludes with two experimental studies on M/EEG data, showing that a much simpler informational theory of consciousness – the *entropic brain hypothesis* – can yield valuable insight without the mathematical challenges brought by IIT.

Acknowledgements

Damn, that was a blast!

These years have been a period of extreme personal and intellectual growth. I've honestly had a lot of fun, and I've met an unbelievable amount of fascinating people. Above all, there are four individuals that have played a critical role in this endeavour, without whom I would have either never started or never finished this programme. They are listed here by chronological order of appearance in this comedy of errors that was my PhD:

- Thanks to **Marc Deisenroth**, for his utterly unexplainable determination to help me get into this PhD programme, which I can only assume came from an abundance of compassion rather than from a surfeit of judgement.
- Thanks to **Murray Shanahan**, for his kindness and support, for his freedom and trust, and for shaping my thought and helping me find a love for philosophy I didn't know I had. I seriously cannot recall how many times I have recommended the first chapter of Embodiment and the Inner Life to people in pubs all over the planet.
- Thanks to **Adam Barrett**, for his extremely helpful guidance and his willingness to take this poor lost PhD student under his wing. Although Adam's name does not appear on this thesis as a supervisor, that is what he was. He has always been a role model for me, and being able to work with him has been an absolute priviledge, personally and professionally.
- Thanks to **Fernando Rosas**, for his unmatched passion, dedication, generosity, wisdom, and comradery. Working with Fernando has been the most enjoyable, most productive, and all-round best thing I've done in my PhD. You do not often get to meet someone who is so just like you, except better in every way.

I have many admirable scientists to thank for invaluable collaborations and friendships. Thanks to Henrik Jensen, for his enthusiasm and for welcoming me in the nourishing environment that is the Imperial Centre for Complexity Science. Thanks to Joe Lizier, for having such a tremendous impact on my research topic, and for sneaking me into the complex systems research community. In particular, special thanks to Neil Rabinowitz, for his honest advice, for his relentless optimism, for his irrational insistence that I do whatever it is I want to do, and for being so delightfully eccentric.

My friends from the Computational Neurodynamics Group and the Department of Computing have made Imperial an extraordinary environment to work in. Huge thanks to Kyriacos, Zafeirios, Marta, Christos, Nat, Matt, Claudia, Hugh, Matt, Kai, and Kostas for making lunches, dinners, coffee breaks, and Wednesday drinks (yes, Wednesday) something to look forward to every day.

I was going to keep this short but I'm already on two pages so who cares.

Thanks to my friends and colleagues at the Sackler Centre for Consciousness Science, especially Simon McGregor, Warrick Roseboom, Anil Seth, Michael Schartner, and Lionel Barnett. They were a major force behind my interest for consciousness science, and the Centre will always remain a reference point to look up to.

Special thanks to Brian Mitchell, for teaching me the ropes of academic life, to write like a proper scholar and speak like a stand-up comedian.

I've been incredibly lucky to travel and collaborate with people from all over the place, which has been an extremely gratifying growth experience. Thanks to my Emotech friends, especially Hongbin, Jan, Javi, Indy, and Pawel. Thanks to my Sydney friends, especially Conor and Leo. Thanks to my Valparaíso friends, especially Rodrigo, Rubén, and Marilyn. Thanks to my DeepMind friends, especially Avi, Andrea, and Francis. Thanks to the brilliant students ~~that have had to suffer me~~ with whom I've had the pleasure to collaborate, Juan Carlos, Tycho, Pietro, Lotte, and Dane.

Thanks to my flatmates Andy, Ibles, Chris, Dani, and Sergio for making daily life less miserable, and for showing a disproportionate amount of interest in my work. Damn, I can't believe they were fine with having one of my conference posters in the living room for more than a year. Thanks to Mada for never-ending psychedelic binaural stimuli that made working hours exceedingly groovy. Thanks to Julian for an astonishingly broad spectrum of interesting conversations, and for sharing the crushing existential angst of PhD life.

As if all of those weren't enough, I've also been incredibly lucky to have found a large group of friends with whom to share these years. I feel extremely grateful and they really are a lot, so I'll write their names in `\footnotesize`: Ignacio, Erica, Irene, Andrei, Alan, Álvaro, Burç, Lucy, Dianna, Azin, Andy, Hardik, Max, Gloria, Irina, Zoe, Klaus, Carlos, and Diana. Thanks to my friends in Cartagena, Héctor, Dani, Sandra, Víctor, Carmen, Alejandro, Pedro, David, and Nur, for not forgetting me despite my love-hate relationship with social media and other electronic forms of communication.

Thanks to Javier "Doctor" Valenzuela, for luring me into academic life, and for teaching me that research is, at the end of the day, about having fun.

Thanks to my parents and the rest of my family, for putting up with long periods away from home, with days of unexplainably bad mood, with my sheer stubbornness and poor life choices. Thanks for making my education their priority despite all of the above.

And again last, but not least, thanks to JJ for getting me into this business.

Contents

Abstract **4**

List of figures **15**

List of tables **17**

Analytical index **19**

1 Introduction **23**

1.1 Of bits and brains 24

1.2 Objectives and motivation 25

1.3 Thesis organisation 26

1.4 Publications 26

2 Information-theoretic foundations **29**

2.1 Integrated information theory: A historical account 30

2.2 Notation, convention and preliminaries 31

2.3 Measures of integrated information 33

2.3.1 Minimum information partition 34

2.3.2 Whole-minus-sum integrated information Φ 36

2.3.3 Integrated stochastic interaction $\tilde{\Phi}$ 38

2.3.4 Integrated synergy ψ 39

2.3.5 Decoder-based integrated information Φ^* 41

2.3.6 Geometric integrated information Φ_G 42

2.3.7 Causal density 44

2.3.8 Other measures 45

I Integrated information in complex systems **47**

3 Integrated information and metastability **49**
3.1 Introduction . . . 50
3.2 Methods . . . 51
3.2.1 Metastability . . . 52
3.3 Results . . . 52
3.3.1 Information-theoretic analysis . . . 53
3.3.2 Robustness of Φ against measurement noise . . . 56
3.4 Conclusion . . . 57

4 Integration and segregation in spiking neurons **59**
4.1 Introduction . . . 60
4.2 Methods . . . 61
4.2.1 Model specification . . . 61
4.2.2 Functional segregation and integration . . . 62
4.3 Results . . . 64
4.3.1 Model behaviour . . . 64
4.3.2 Avalanche statistics . . . 65
4.3.3 Information-theoretic analysis . . . 68
4.3.4 Criticality and linear interactions . . . 69
4.4 Conclusion . . . 71

5 Integrated information and distributed computation **73**
5.1 Introduction . . . 74
5.2 Methods . . . 75
5.2.1 Elementary automata and complexity classes . . . 75
5.2.2 Local information dynamics . . . 75
5.3 Results . . . 77
5.3.1 Integrated information and complexity class . . . 77
5.3.2 Integrated information at the edge of chaos . . . 78
5.3.3 Information is integrated by coherent structures . . . 79
5.4 Conclusion of Part I . . . 80

II Drawbacks and limitations of integrated information theory 83

6 Measuring integrated information 85

6.1 Introduction . . . 86
6.2 Methods . . . 87
6.2.1 Key quantities for computing integrated information measures . . . 88
6.3 Results . . . 89
6.3.1 Two-node network . . . 89
6.3.2 Eight-node networks . . . 90
6.3.3 Random networks . . . 94
6.4 Discussion . . . 96
6.4.1 Partition selection . . . 97
6.4.2 Continuous variables and the linear Gaussian assumption . . . 98
6.4.3 Empirical as opposed to maximum entropy distribution . . . 99
6.5 Final remarks . . . 99

7 Empirical evidence for and against IIT 101

7.1 Introduction . . . 102
7.2 Existing experimental evidence for IITC . . . 102
7.2.1 PCI and causal interventions . . . 104
7.3 Novel experimental evidence against IITC . . . 106
7.3.1 Datasets and methodology . . . 106
7.3.2 Broadband integrated information is lower in wakeful rest . . . 107
7.3.3 Band-specific integrated information is inconsistent . . . 108
7.4 Modelling the effect of unobserved activity on Φ . . . 110
7.5 Conclusion of Part II . . . 112

III Beyond integrated information: Developments and alternatives 113

8 Quantifying high-order interdependencies 115

8.1 Introduction . . . 116
8.2 Fundamentals of multivariate information . . . 116
8.2.1 Entropy and negentropy . . . 117

8.2.2 The two faces of interdependency . . . 118
8.3 O-information: Redundancy minus synergy . . . 120
8.3.1 Definition and basic properties . . . 120
8.3.2 Information decompositions . . . 122
8.3.3 Characterising extreme values of Ω . . . 126
8.3.4 Ω and high-order interactions in statistical mechanics . . . 127
8.4 Complexity and integrated information . . . 129
8.5 Conclusion . . . 131

9 Integrated information decomposition 133
9.1 Introduction . . . 134
9.2 Decomposing multivariate information . . . 134
9.2.1 Forward and backward information decomposition . . . 135
9.3 Integrated information decomposition: ΦID . . . 135
9.3.1 Double-redundancy lattice . . . 136
9.3.2 Redundancies and atoms . . . 137
9.3.3 Simple examples . . . 139
9.4 Results . . . 140
9.4.1 Limitations of conventional causal discovery methods . . . 140
9.4.2 Information processing in complex systems . . . 140
9.4.3 Different types of integration . . . 141
9.4.4 Measures of integrated information . . . 142
9.4.5 Why whole-minus-sum Φ can be negative . . . 143
9.4.6 Why unnormalised causal density can exceed TDMI . . . 145
9.5 Formalising causal emergence through ΦID . . . 146
9.6 Discussion . . . 148
9.6.1 Towards multi-dimensional measures of complexity . . . 148
9.6.2 Integration measures conflate transfer and synergy . . . 148
9.6.3 Causal analysis . . . 149
9.6.4 Limitations and future extensions . . . 149
9.6.5 Integrated information as a universal signature of emergence . . . 149

10 Consciousness and information content **151**
10.1 Introduction . . . 152
10.2 Entropy and Lempel-Ziv complexity . . . 152
10.3 Sensory stimuli and the psychedelic state . . . 154
10.3.1 Increased signal diversity under external stimulation . . . 155
10.3.2 Stronger stimulus weakens drug effect . . . 156
10.4 The improvisational state of mind . . . 157
10.4.1 Experiment and data acquisition . . . 158
10.4.2 Increased signal diversity in the improvisational state . . . 158
10.5 Discussion . . . 160
10.5.1 Lempel-Ziv complexity and Φ . . . 160
10.5.2 The 'brain' in 'entropic brain' . . . 161
10.6 Conclusion of Part III . . . 162

11 Conclusion **163**
11.1 Summary of thesis contributions . . . 164
11.2 Future work . . . 166
11.3 Parting thoughts . . . 167

Appendix A Information-theoretic foundations **169**
A.1 Derivation and concavity proof of I^* . . . 169
A.1.1 Derivation of I^* in Gaussian systems . . . 169
A.1.2 $\tilde{I}(\beta)$ is concave in β in Gaussian systems . . . 171
A.2 Bounds on causal density . . . 172
A.3 Properties of integrated information measures . . . 173
A.3.1 Whole-minus-sum integrated information Φ . . . 174
A.3.2 Integrated stochastic interaction $\tilde{\Phi}$ . . . 174
A.3.3 Integrated synergy ψ . . . 174
A.3.4 Decoder-based integrated information Φ^* . . . 175
A.3.5 Geometric integrated information Φ_G . . . 175
A.3.6 Causal density . . . 175

Appendix B Quantifying high-order interdependencies **177**

B.1 Statistical structures across scales 177

B.2 Ω as a superposition of tendencies 179

B.3 $R(\pi)$ decreases for finer partitions 181

B.4 Proof of Lemma 2 182

B.5 Proof of Proposition 1 182

B.6 Proof of Lemma 3 183

B.7 Proof of Proposition 2 183

B.8 Proof of Proposition 8 184

Appendix C Integrated information decomposition **187**

C.1 The product of two lattices is a lattice 187

C.2 Decomposing PID atoms 188

C.3 Computing the ΦID atoms 189

C.4 Results of section 'Different types of integration' 191

C.5 Results of section 'Measures of integrated information' 191

C.6 Results of section 'Why whole-minus-sum Φ can be negative' 194

C.7 Formalising causal emergence 194

Bibliography **199**

List of figures

3.1 Global synchrony and metastability in coupled oscillators 53
3.2 Integrated information Φ and coalition entropy H_c in the phase transition . . 54
3.3 Integrated information Φ and time-delayed mutual information 55
3.4 Robustness of Φ against measurement noise 56

4.1 Schematic diagram of modular spiking neural network 61
4.2 Sample runs of the network for different rewiring probabilities p 65
4.3 Avalanche size distributions for different rewiring probabilities p 66
4.4 Avalanche size distributions for three sample runs of the simulation 67
4.5 Information storage and transfer in a modular spiking neural network 68
4.6 Joint density distributions of pairs of modules 70
4.7 Linear and non-linear MI between pairs of modules 71

5.1 Integrated information grows monotonically with Wolfram class number . . 77
5.2 Integrated information peaks at the edge of chaos 79
5.3 Local information measures in cellular automata 80

6.1 Two-node AR process and integrated information measures 89
6.2 Integrated information measures for different coupling values 90
6.3 Integrated information measures for the Φ-optimal AR process 91
6.4 Networks used in the comparative analysis of integrated information measures 93
6.5 Integrated information measures for all networks in Figure 6.4 93
6.6 Integrated information measures for Erdős–Rényi random networks 95
6.7 Integrated information measures plotted against average correlation 96

7.1 Broadband integrated information in six datasets of interest 108

7.2 Integrated information measures for all datasets, resolved by frequency band 109
7.3 Integrated information in an AR thalamo-cortical drive model 111

8.1 Total information diagram 119
8.2 Double diamond diagram 123
8.3 O-information in Hamiltonian systems of increasing order 128
8.4 Relation between Ω, TC, DTC, and neural complexity 130

9.1 Lattice of nodes in $\mathcal{A}$ arranged according to the partial ordering $\preceq$. 136
9.2 Double-redundancy lattice for two predictors and two targets 137
9.3 Example systems of logic gates 141
9.4 Standard and corrected Φ in a two component noisy AND system 144

10.1 External stimuli increase LZ complexity 155
10.2 Stronger external stimulation reduces the entropic effect of LSD 156
10.3 Lempel-Ziv complexity in strict and improvised musical performances . . . 159

B.1 O-information and TC bounds at different scales 179
B.2 O-information bounds for a system with a binary symmetric channel 180

List of tables

2.1 Integrated information measures considered and original references. 34
2.2 Overview of properties of integrated information measures 34

6.1 Networks ranked by their value of each integrated information measure . . . 92
6.2 Integrated information measures considered and summary of results 97

7.1 Summary of existing experimental evidence for IITC 103
7.2 Datasets used in experimental validation of IITC 106

9.1 Sensitivity of integrated information measures to ΦID atoms 143

Analytical index

1. Introduction 24

The scientific study of consciousness is a supremely interesting endeavour. Among many proposals, ***Integrated Information Theory*** *(IIT) stands out as a candidate theory of consciousness, although its genuine contributions are often mixed with the speculations behind it. The main objective of this thesis is to push IIT as a useful measure of complexity, to fairly assess its support as a theory of consciousness, and to open developments and alternatives.*

2. Information-theoretic foundations 30

The main working tool of this thesis is Shannon's Information Theory. In this foundational chapter we lay out the basic building blocks of Shannon's theory we will use, and review the existing IIT literature with proposed measures of integrated information, Φ*. We outline the intuitions behind each of these measures, and provide detailed procedures describing how to compute them and how each element of the computation refers back to the ideas behind IIT.*

3. Integrated information and metastability 50

We begin our exploration of integrated information in complex systems with networks of coupled oscillators, frequently used as large-scale models of neural dynamics and capable of exhibiting a rich variety of so-called metastable chimera states. We bring metastability and integrated information together, by showing that these oscillators exhibit a critical peak of Φ *that coincides with peaks in other measures such as metastability and coalition entropy.*

4. Integration and segregation in spiking neurons 60

Next, we study a model network of spiking neurons with different coupling configurations. We find that information transfer and storage peak at two separate points for different values

of the coupling parameter and are balanced at an intermediate point, where avalanches follow a long-tailed, power law-like distribution and reproduce empirical findings in the biological brain. In this way, we link together the balance between functional integration and differentiation (à la IIT) with the appearance of power law-like avalanches.

5. Integrated information and distributed computation 74

As a final complexity case study, we consider distributed computation in cellular automata. We relate IIT to distributed computation in two ways: at a global scale, Φ is higher for complex, class IV *automata; and at a local scale Φ is higher for emergent coherent structures, like blinkers, gliders, and collisions. Together with the previous two examples, this suggests Φ is, empirically, a good candidate for a universal marker of complexity.*

6. Measuring integrated information 86

For it to be a fundamental theory of consciousness, IIT needs a robust axiomatic base linking phenomenology to one or more measurable quantities. We explore the properties of several proposed measures in simulation on simple, but non-trivial systems, and find a striking diversity in the behaviour of these measures – no two measures show consistent agreement across all analyses. We conclude that the axioms of IIT are underspecified, in the sense that multiple measures consistent with the axioms show qualitatively different behaviour in practice.

7. Empirical evidence for and against IIT 102

Regardless of its mathematical underpinnings, a minimum requirement for any theory of consciousness is to make successful predictions on adult human brains. We review existing experimental evidence for and against IIT as a theory of consciousness, and present new comprehensive analyses on several datasets. The evidence is mixed, and in some cases Φ, counterintuitively, is drastically increased *in the unconscious state and* reduced *in the psychedelic state. We discuss possible causes of this discrepancy and discuss the relevance of these results to IIT's current and future status.*

8. Quantifying high-order interdependencies 116

We revisit the mathematical basis of early IIT, with the aim of providing new, more suitable tools. Our investigation into the multivariate structure of information leads to a new measure, the O-information*, capable of characterising synergy- and redundancy-dominated systems. We compare the O-information against Φ's predecessor, the Tononi-*

Sporns-Edelman measure of neural complexity, and argue that the O-information is better able to capture the intuitions behind the origins of IIT.

9. Integrated information decomposition 134

We deepen the connection between information decomposition and IIT, by outlining a unified theory of Integrated Information Decomposition, ΦID. *Most importantly,* ΦID *reveals that what is typically referred to as 'integration' is actually an aggregate of several heterogeneous phenomena, and can help us understand and alleviate the limitations of existing* Φ *measures. Additionally, we link* ΦID *with fundamental principles of causal emergence, providing theoretical support to our claims relating IIT and complexity.*

10. Consciousness and information content 152

As an alternative to IIT, we consider another, much simpler informational theory of consciousness known as the Entropic Brain Hypothesis (EBH). We present two examples from the study of altered states of consciousness – musical improvisation and the psychedelic state – and interpret the results in the light of EBH. We argue that the simplicity and empirical success of EBH provide valuable lessons for IIT, and that a collective, open-minded engagement between these and other theories are key for a cohesive, mature, and productive science of consciousness.

Chapter 1

Introduction

Chapter summary

The scientific study of consciousness is a supremely interesting endeavour. Among many proposals, **Integrated Information Theory** (IIT) stands out as a candidate theory of consciousness, although its genuine contributions are often mixed with the speculations behind it. The main objective of this thesis is to push IIT as a useful measure of complexity, to fairly assess its support as a theory of consciousness, and to open developments and alternatives.

1.1 Of bits and brains

Consciousness. That's a *real* problem. The kind of problem you didn't realise you had, and once you do, keeps you up at night asking yourself "how is this a thing?"

After all, the brain is just a physical system, made of physical elements we call neurons. We know these neurons fairly well, thanks in part to the early work of Andrew Huxley – coincidentally, grandson to Thomas Huxley, who gave name to the building where this PhD research was carried out. And yet, these neurons go "click" in such precisely calculated ways that *it feels like something* to be this mush of two kilograms' worth of neurons between our ears.

In fact, not only it feels like something to be a brain, but this how-it-feels can be altered in outstanding ways. As a thought experiment, let us picture a computer, made of transistors, that just as our neurons, go "click" in very precisely calculated ways. Next, let us sprinkle a few micrograms of an alien substance, one the transistors have never encountered before. Now we turn on the computer – and, miraculously, instead of setting on fire, the computer produces a fast stream of sounds, images, and text it has never produced before.

As incredible as this seems, it is a powerful analogy for the brain effects of psychedelic drugs like LSD. The effect of literally a handful of molecules per neuron makes the owner of such a brain smell sounds, hear lights, and feel a profoundly different *kind* of consciousness.

So, how does this work? How is it that the interactions between neurons in the brain generate endless streams of experience and behaviour? During the last 40 years, a number of brave scientists have thrown their own brains at the problem, and have generated a constantly-evolving sequence of *theories of consciousness* attempting to give a satisfactory explanation to the problem.

Among these theories, one candidate stands out: the ***Integrated Information Theory* (IIT)**, developed by Giulio Tononi and a growing number of collaborators since 1994. At its core, lie two powerful intuitions:

- That a system that is somehow *more than the sum of its parts* has non-zero integrated information, Φ; and
- That a system with non-zero integrated information is conscious.

The core element of IIT is a measure of integrated information, Φ, that (broadly speaking) attempts to quantify the information that is contained in the interactions between the parts of a system and not within the parts themselves. At the moment there is no agreed-upon formula for Φ, although many proposals have been put forward. Over time, the conceptualisation of Φ has evolved: as a balance between integration and segregation, statistical interdependence between subsystems, or causal irreducibility – terms that will become clear in the following

chapters. Crucially, the formulation of Φ is linked to IIT's *axioms*, which provide the philosophical foundation that allows IIT to call itself a theory of consciousness.

For those that, like myself, come with a physicist's mindset, Φ opens many more doors than those of perception and consciousness. In particular, it brings strong reminiscences of ***complexity science***, the study of large systems made of locally interacting components giving rise to emergent behaviour. Complexity, as new of a science as that of consciousness, is also in search for its fundamental principles and unifying theories – a gap which, I argue, IIT is uniquely equipped to fill if we strip it of its unfettered claims about consciousness and keep its valuable information-theoretic contributions.

The development of IIT has been characterised by two parallel trends: a spark of genius intuitions, that have driven theoretical and experimental neuroscience research since their inception; coupled with a set of increasingly overambitious claims about the fundamental nature of consciousness. My hypothesis is that IIT's audacious claims about consciousness have kept other scientists away, blurring the line between fact and speculation, and thereby preventing some of its valuable ideas from reaching other areas of knowledge.

1.2 Objectives and motivation

This thesis is the result of a cocktail of one part fascination and one part frustration with IIT. The reader with sharp theory-of-mind skills will find throughout this text a constant tension between those two elements – hopefully, the kind of tension that leads to honest self-criticism and scientific progress.

With these considerations in mind, and with the aim of making an honest and positive contribution to neuroscience and complexity science, I set myself three research objectives:

1. To dissociate IIT's claims as a theory of consciousness from its information-theoretic contributions, and put the latter to use in complexity science;
2. To develop the mathematical basis of IIT; and
3. To provide an honest assessment of the evidence for and against IIT's claims about consciousness.

In doing so, I intend to bring IIT out of the secretive veil of the small theoretical neuroscience community where it was born, so it can learn from other sciences and contribute all the valuable insights it has to offer.

1.3 Thesis organisation

This thesis is organised in three parts, that form a cohesive story but can be read independently.

1. In **Part I**, I outline several example applications of IIT to the study of complex systems. In order to do so, I also describe a few developments regarding the estimation of Φ measures on non-linear systems. Through these applications I advocate for IIT as a valuable unifying theory of complexity, that encompasses other seemingly disconnected concepts like criticality, metastability, and distributed computation.

2. In **Part II**, I describe two limitations of IIT, both in theory and in practice. The theoretical limitation stems from the fact that the basic foundations of IIT do not uniquely determine a measure of integrated information; and the practical one is that most proposed Φ measures do not in fact show the expected difference when applied to conscious and unconscious brain states.

3. In **Part III**, I present developments and alternatives to standard IIT, on two fronts:

    - On the theoretical side, I re-examine the intuitions behind IIT and its predecessor, *neural complexity*, and recast them under the more mathematically solid framework of Partial Information Decomposition.
    - On the practical side, I describe two applications of a much simpler theory of consciousness, the *entropic brain hypothesis*, to musical improvisation and the psychedelic state. I argue that, as a much simpler theory, EBH can bring valuable insights without the mathematical challenges of IIT.

1.4 Publications

Content from the following publications is directly relevant to this thesis:

1. **P. Mediano**, J.C. Farah and M. Shanahan (2016). *Integrated Information and Metastability in Systems of Coupled Oscillators*. arXiv: 1606.08313.

2. **P. Mediano** and M. Shanahan (2017). *Balanced Information Storage and Transfer in Modular Spiking Neural Networks*. arXiv: 1708.04392.

3. **P. Mediano** and M. Shanahan (2015). *An Unexpected Discrepancy in a Well-known Problem: Kraskov Estimators Applied to Spiking Neural Networks*. Proceedings of the European Conference in Artificial Life (ECAL'15).

4. **P. Mediano**, A. Seth and A. Barrett (2018). *Measuring Integrated Information: Comparison of Candidate Measures in Theory and Simulation*. Entropy.

5. A. Barrett and **P. Mediano** (2019). *The Φ Measure of Integrated Information is not Well-defined for all Physical Systems*. Journal of Consciousness Studies.

6. F. Rosas, **P. Mediano**, M. Gastpar and H. Jensen (2019). *Quantifying High-order Interdependencies via Multivariate Extensions of the Mutual Information*. Physical Review E.

7. **P. Mediano***, F. Rosas*, R. Carhart-Harris, A. Seth and A. Barrett. *Beyond Integrated Information: A Taxonomy of Information Dynamics Phenomena*. arXiv: 1909.02297.

8. D. Dolan, H. Jensen, **P. Mediano**, M. Molina-Solana, H. Rajpal, F. Rosas and J. Sloboda (2018). *The Improvisational State of Mind: A Multi-disciplinary Study of an Improvisatory Approach to Classical Music Repertoire Performance*. Frontiers in Psychology.

In addition, these are other publications by the author that are relevant to the topic, but not explicitly covered in this thesis:

9. L. Novelli, P. Wollstadt, **P. Mediano**, M. Wibral and J. Lizier (2019). *Large-scale Directed Network Inference with Multivariate Transfer Entropy and Hierarchical Statistical Testing*. Network Neuroscience.

10. A. Tacchetti*, H. F. Song*, **P. Mediano*** *et al.* (2019). *Relational Forward Models for Multi-Agent Learning*. ICLR 2019.

11. F. Rosas, **P. Mediano**, M. Ugarte and H. Jensen (2018). *An Information-theoretic Approach to Self-organisation: Emergence of Complex Interdependencies in Coupled Dynamical Systems*. Entropy.

12. P. Wollstadt **et al.** (2018). *IDTxl: The Information Dynamics Toolkit xl: A Python Package for the Efficient Analysis of Multivariate Information Dynamics in Networks*. Journal of Open-Source Software.

13. S. McGregor and **P. Mediano** (2018). *Adaptation Is Not Just Improvement Over Time*. Artificial Life.

14. S. McGregor and **P. Mediano** (2018). *Measuring Fitness Effects of Agent-Environment Interactions*. Artificial Life.

15. T. Tax*, **P. Mediano*** and M. Shanahan (2017). *The Partial Information Decomposition of Generative Neural Network Models*. Entropy, 19(9), 474.

16. K. Nikiforou, **P. Mediano** and M. Shanahan (2017). *An Investigation of the Dynamical Transitions in Harmonically Driven Random Networks of Firing-Rate Neurons*. Cognitive Computation 3 (9).

17. X. Arsiwalla, **P. Mediano** and P. Verschure (2017). *Spectral Modes of Network Dynamics Reveal Increased Informational Complexity Near Criticality*. Procedia Computer Science 108.

18. N. Dilokthanakul, **P. Mediano**, M. Garnelo et al. (2016). *Deep Unsupervised Clustering with Gaussian Mixture Variational Autoencoders*. arXiv: 1611.02648.

Chapter 2

Information-theoretic foundations

Chapter summary

The main working tool of this thesis is Shannon's Information Theory. In this foundational chapter we lay out the basic building blocks of Shannon's theory we will use, and review the existing IIT literature with proposed measures of integrated information, Φ. We outline the intuitions behind each of these measures, and provide detailed procedures describing how to compute them and how each element of the computation refers back to the ideas behind IIT.

2.1 Integrated information theory: A historical account

The basic principles of IIT can be traced back to the fruitful collaboration between Tononi, Sporns, and Edelman (henceforth, TSE) in the early 90's. In alignment with the exciting complexity research emerging at the time [1], the trio set out to investigate what makes a system "more than the sum of its parts" – and how this related to cognition and the brain.

In their work, this notion of greater-than-the-sum was embodied in the concept of *dynamical complexity*, which was often portrayed as a balance between two competing tendencies:

- **integration**, to the extent that the system behaves as one; and
- **segregation**, to the extent that the parts of the system behave independently.

The concept of dynamical complexity evolved over time, and has been characterised using different mathematical tools. In the original work, which we will refer to as **IIT 0.1**,[1] TSE developed a measure of coexisting global integration and local segregation called *neural complexity* [2, 3]. Computing neural complexity involves weighting correlations within the system across multiple scales, to capture this interplay between local and global interactions.

The term "integrated information," and its characteristic symbol Φ, first appeared in 2003 with Ref. [4], which we refer to as **IIT 1.0**. This work introduced the idea of measuring integration by partitioning the system, measuring the correlations across the "cruelest cut."

The 2008 work of Balduzzi and Tononi [5], **IIT 2.0**, introduced a significant extension over IIT 1.0 by explicitly incorporating time into the theory. In IIT 2.0, two time series are integrated if they mutually affect each other's temporal evolution, and so the basic object of study becomes the mutual information of a system *across time*, instead of across space. This provides a stronger link with neural dynamics, as it views integrated information as an intrinsically *dynamical process*, as opposed to as a feature of static systems like IIT $\leq$ 1.0.

Finally, in its most recent iteration, Oizumi, Albantakis and Tononi [6] presented **IIT 3.0**, which represents a substantial departure from IIT 2.0 and standard information-theoretic tools. IIT 3.0 places a much stronger emphasis on causal interventions to elucidate the "intrinsic" informational properties of the system. However, this comes at a cost of a much more intricate definition, with more unjustified choices, and a measure that is computationally intractable except in the simplest or most carefully constructed cases.

Here, we focus on the earlier conceptions of integrated information because (i) they are more readily applicable to empirical time series; (ii) they remain conceptually powerful in theories of consciousness, and (iii) they promise general applicability to many other questions in neuroscience and beyond, in which part-whole relations are of interest.

[1] To better organise the different iterations of IIT, we will backpropagate the numerical labelling system introduced in 2014. It is unclear whether the 1994 work should be labelled as IIT 1.0 or 0.1 – clearly, the marketing value of a `major.minor` versioning system wasn't well appreciated back then. Since the term "integrated information" itself only appeared in 2003, we will refer to the 1994 work as IIT 0.1.

2.2 Notation, convention and preliminaries

In this section, we review the fundamental concepts needed to define and discuss the candidate measures of integrated information. For a more comprehensive introduction, see the standard textbook by Cover and Thomas [7]. In general, we will denote random variables with uppercase letters (e.g. X, Y) and particular instantiations with the corresponding lowercase letters (e.g. x, y). Variables can be either continuous or discrete, and we assume that continuous variables can take any value in $\mathbb{R}^n$ and that a discrete variable X can take any value in the finite set Ω_X. Whenever there is a sum involving a discrete variable X, we assume the sum runs for all possible values of X (i.e. the whole Ω_X). A partition $\mathcal{P} = \{M^1, M^2, ..., M^r\}$ divides the elements of system X into r non-overlapping, non-empty sub-systems (or parts), such that $X = M^1 \bigcup M^2 \bigcup ... \bigcup M^r$ and $M^i \bigcap M^j = \varnothing$, for any i, j. We denote each variable in X as X^i, and the total number of variables in X as n. When dealing with time series, time will be indexed with a subscript, e.g. X_t.

Entropy H quantifies the uncertainty associated with random variable X—i.e. the higher $H(X)$ the harder it is to make predictions about X—and is defined as

$$H(X) =: -\sum_x p(x) \log p(x). \tag{2.1}$$

In many scenarios, a discrete set of states is insufficient to represent a process or time series. This is the case, for example, with brain recordings, which come in real-valued time series and with no a priori discretisation scheme. In these cases, using a continuous variable $X \in \mathbb{R}$, we can similarly define the *differential entropy*,

$$H[p] =: -\int p(x) \log p(x) dx. \tag{2.2}$$

However, differential entropy is not as interpretable and well-behaved as its discrete-variable counterpart. For example, differential entropy is not invariant to rescaling or other transformations on X. Moreover, it is only defined if X has a density with respect to the Lebesgue measure dx; this assumption will be upheld throughout this thesis. We can also define the *conditional* and *joint* entropies as

$$\begin{aligned} H(X \,|\, Y) &=: \sum_y p(y) H(X \,|\, Y = y) \\ &= -\sum_y p(y) \sum_x p(x \,|\, y) \log p(x \,|\, y), \end{aligned} \tag{2.3}$$

$$H(X,Y) =: -\sum_{x,y} p(x,y) \log p(x,y), \tag{2.4}$$

respectively. Conditional and joint entropies can be analogously defined for continuous variables by appropriately replacing sums with integrals.

The Kullback–Leibler (KL) divergence quantifies the dissimilarity between two probability distributions p and q:

$$D_{KL}(p\|q) =: \sum_x p(x) \log \frac{p(x)}{q(x)}. \tag{2.5}$$

The KL divergence represents a notion of (non-symmetric) distance between two probability distributions. It plays an important role in information geometry, which deals with the geometric structure of manifolds of probability distributions.

Finally, mutual information I quantifies the interdependence between two random variables X and Y. It is the KL divergence between the full joint distribution and the product of marginals, but it can also be expressed as the average reduction in uncertainty about X when Y is given:

$$\begin{aligned} I(X;Y) &=: D_{KL}\left(p(X,Y) \parallel p(X)p(Y)\right) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y). \end{aligned} \tag{2.6}$$

Mutual information is symmetric in the two arguments X and Y. We make use of the following properties of mutual information:

1. $I(X;Y) = I(Y;X)$,
2. $I(X;Y) \geq 0$, and
3. $I(f(X);g(Y)) = I(X;Y)$ for any injective functions f, g.

We highlight one implication of property 3: I is upper-bounded by the entropy of both X and Y. This means that the entropy $H(X)$ of a random variable X is the maximum amount of information X can have about any other variable Y (or another variable Y can have about X).

Mutual information is defined analogously for continuous variables and, unlike differential entropy, it retains its interpretability in the continuous case.[2] Furthermore, one can track how much information a system preserves during its temporal evolution by computing the time-delayed mutual information (TDMI) $I(X_t;X_{t-\tau})$.

Next, we introduce notation and several useful identities to handle Gaussian variables. Given an n-dimensional real-valued system X, we denote its covariance matrix as $\Sigma(X)_{ij} =:$

[2]The formal derivation of the differential entropy proceeds by considering the entropy of a discrete variable with k states, and taking the $k \to \infty$ limit. The result is the differential entropy plus a divergent term that is usually dropped and is ultimately responsible for the undesirable properties of differential entropy. In the case of $I(X;Y)$ the divergent terms for the various entropies involved cancel out, restoring the useful properties of its discrete counterpart.

$\mathrm{cov}(X^i, X^j)$. Similarly, cross-covariance matrices are denoted as $\Sigma(X,Y)_{ij} =: \mathrm{cov}(X^i, Y^j)$. We will make use of the conditional (or partial) covariance formula,

$$\Sigma(X|Y) =: \Sigma(X) - \Sigma(X,Y)\Sigma(Y)^{-1}\Sigma(Y,X). \tag{2.7}$$

For Gaussian variables,

$$H(X) = \frac{1}{2}\log(\det\Sigma(X)) + \frac{1}{2}n\log(2\pi e)\,, \tag{2.8}$$

$$H(X|Y=y) = \frac{1}{2}\log(\det\Sigma(X|Y)) + \frac{1}{2}n\log(2\pi e)\,,\ \forall y\,, \tag{2.9}$$

$$I(X;Y) = \frac{1}{2}\log\left(\frac{\det\Sigma(X)}{\det\Sigma(X|Y)}\right). \tag{2.10}$$

All systems we deal with in this article are stationary and ergodic, so throughout the paper $\Sigma(X_t) = \Sigma(X_{t-\tau})$ for any τ.

2.3 Measures of integrated information

In this section, we review the theoretical underpinnings and practical considerations of several proposed measures of integrated information, and in particular how they relate to intuitions about segregation, integration and complexity. These measures are:

- Whole-minus-sum integrated information, Φ;
- Integrated stochastic interaction, $\tilde{\Phi}$;
- Integrated synergy, ψ;
- Decoder-based integrated information, Φ^*;
- Geometric integrated information, Φ_G; and
- Causal density, CD.

All of these measures (besides CD) have been inspired by the measure proposed by Balduzzi and Tononi in [5], which we call Φ_{2008}. Φ_{2008} was based on the information the current state contains about a hypothetical maximum entropy past state. In practice, this results in measures that are applicable only to discrete Markovian systems [8]. For broader applicability, it is more practical to build measures based on the ongoing spontaneous information dynamics—that is, based on $p(X_t, X_{t-\tau})$ without applying a perturbation to the system. Measures are then well-defined for any stochastic system (with a well-defined Lebesgue measure across the states), and can be estimated for real data using empirical

distributions if stationarity can be assumed. All of the measures considered in this thesis are based on a system's spontaneous information dynamics.

Table 2.1 contains a brief description of each measure and a reference to the original publication that introduced it.[3] We refer the reader to the original publications for more detailed descriptions of each measure. Table 2.2 contains a summary of properties of the measures considered, proven for the case in which the system is ergodic and stationary, and the spontaneous distribution is used.

Table 2.1: Integrated information measures considered and original references.

Measure	Description	Reference
Φ	Information lost after splitting the system	[5]
$\tilde{\Phi}$	Uncertainty gained after splitting the system	[8]
ψ	Synergistic predictive information between parts of the system	[9]
Φ^*	Past state decoding accuracy lost after splitting the system	[10]
Φ_G	Information-geometric distance to system with disconnected parts	[11]
CD	Average pairwise directed information flow	[12]

Table 2.2: Overview of properties of integrated information measures, proofs in Appendix A.3.

	Φ	$\tilde{\Phi}$	ψ	Φ^*	Φ_G	**CD**
Time-symmetric	✓	✓	×	?	✓	×
Non-negative	×	✓	✓	✓	✓	✓
Invariant to variable rescaling	✓	×	✓	✓	✓	✓
Upper-bounded by time-delayed mutual information	✓	×	✓	✓	✓	✓
Known estimators for arbitrary real-valued systems	✓	✓	×	×	×	✓
Closed-form expression in discrete and Gaussian systems	✓	✓	✓	×	×	✓

2.3.1 Minimum information partition

Key to all measures of integrated information is the notion of splitting or partitioning the system to quantify the effect of such split on the system as a whole. In that spirit, integrated information measures are defined through some measure of *effective information*, which

[3] Although the origins of causal density go as back as 1969, it hasn't been until the last decade that it has found its way into neuroscience. The paper referenced in the table acts as a modern review of the properties and behaviour of causal density. This measure is somewhat distinct from the others, but is still a measure of complexity based on information dynamics between the past and current state; therefore its inclusion here will be useful.

operationalises the concept of "information *beyond* a partition" $\mathcal{P}$. This typically involves splitting the system according to $\mathcal{P}$ and computing some form of information loss, via (for example) mutual information (Φ), conditional entropy ($\tilde{\Phi}$), or decoding accuracy (Φ^*) (see Table 2.1). Integrated information is then the effective information with respect to the partition that identifies the "weakest link" in the system, i.e. the partition for which the parts are least integrated. Formally, integrated information is the effective information beyond the *minimum information partition* (MIP), which, given an effective information measure $f[X;\tau,\mathcal{P}]$, is defined as

$$\mathcal{P}_{\text{MIP}} = \arg_{\mathcal{P}} \min \frac{f[X;\tau,\mathcal{P}]}{K(\mathcal{P})}, \tag{2.11}$$

where $K(\mathcal{P})$ is a normalisation coefficient. In other words, the MIP is the partition across which the (normalised) effective information is minimum, and integrated information is the (unnormalised) effective information beyond the MIP. The purpose of the normalisation coefficient is to avoid biasing the minimisation towards unbalanced bipartitions (recall that the extent of information sharing between parts is bounded by the entropy of the smaller part). Balduzzi and Tononi [5] suggest the form

$$K(\mathcal{P}) = (r-1)\min_k H(M_t^k). \tag{2.12}$$

However, not all contributions to IIT have followed Balduzzi and Tononi's treatment of the MIP. Of the measures listed above, Φ and $\tilde{\Phi}$ share this partition scheme, ψ defines the MIP through an *unnormalised* effective information, and Φ^*, Φ_G and CD are defined via the atomic partition without any reference to the MIP. These differences are a confounding factor when it comes to comparing measures—it becomes difficult to ascertain whether differences in behaviour of various measures are due to their definitions of effective information, to their normalisation factor (or lack thereof), or to their partition schemes. We return to this topic in the Discussion section below.

In the following, we present all measures as they were introduced in their original papers (see Table 2.1), although it is trivial to combine different effective information measures with different partition optimisation schemes. However, all results presented here are calculated by minimising each unnormalised effective information measure over even-sized bipartitions—i.e. bipartitions in which both parts have the same number of components. This is to avoid conflating the effect of the partition scan method with the effect of the integrated information measure itself.

2.3.2 Whole-minus-sum integrated information Φ

We next turn to the different measures of integrated information. As highlighted above, a primary difference among them is how they define the effective information beyond a given partition. Since most measures were inspired by Balduzzi and Tononi's Φ_{2008}, we start there.

For Φ_{2008}, the effective information φ_{2008} is written as (following notation from [13]) the KL divergence between $p_{\mathrm{c}}(X_0|X_1 = x)$ and $\Pi_k p_{\mathrm{c}}(M_0^k|M_1^k = m^k)$, where $p_c(X_0|X_1 = x)$ (and analogously $p_{\mathrm{c}}(M_0^k|M_1^k = m^k)$) is the conditional distribution for X_0 given $X_1 = x$ under the perturbation at time 0 into all states with equal probability—i.e. given that the joint distribution is given by $p_{\mathrm{ce}}(X_0, X_1) =: p(X_1|X_0)p_u(X_0)$, where p_u is the uniform (maximum entropy) distribution.[4]

Averaging φ_{2008} over all states x, the result can be expressed as either

$$I(X_0; X_1) - \sum_{k=1}^{r} I(M_0^k; M_1^k), \tag{2.13}$$

or

$$-H(X_0|X_1) + \sum_{k=1}^{r} H(M_0^k|M_1^k). \tag{2.14}$$

These two expressions are equivalent under the uniform perturbation, since they differ only by a factor that vanishes if $p(X_0)$ is the uniform distribution. However, they are *not* equivalent if the spontaneous distribution of the system is used instead—i.e. if $p(X_{t-\tau}, X_t)$ is used instead of $p_{\mathrm{ce}}(X_0, X_1)$. This means that, for application to spontaneous dynamics (i.e. without perturbation), we have two alternatives that give rise to two measures that are both equally valid analogs of Φ_{2008}.

We call the first alternative whole-minus-sum integrated information Φ (Φ_{E} in [8]). The effective information φ is defined as the difference in time-delayed mutual information between the whole system and the parts. The effective information of the system beyond a certain partition $\mathcal{P}$ is

$$\varphi[X; \tau, \mathcal{P}] =: I(X_{t-\tau}; X_t) - \sum_{k=1}^{r} I(M_{t-\tau}^k; M_t^k)\ . \tag{2.15}$$

[4] The c and e here stand respectively for cause and effect. Without an initial condition, here that the uniform distribution holds at time 0, there would be no well-defined probability distribution for these states. Further, Markovian dynamics are required for these probability distributions to be well-defined; for non-Markovian dynamics, a longer chain of initial states would have to be specified, going beyond just that at time 0.

We can interpret $I(X_{t-\tau};X_t)$ as how good the system is at predicting its own future or decoding its own past (which are equivalent because mutual information is symmetric). Then, φ here can be seen as the loss in predictive power incurred by splitting the system according to $\mathcal{P}$. The details of the calculation of Φ (and the MIP) are shown in Box 1.[5]

Φ is often regarded as a poor measure of integrated information because it can be negative [11, 10]. This is indeed conceptually awkward if Φ is seen as an absolute measure of integration between the parts of a system, though it is reasonable if Φ is interpreted as a "net synergy" measure [14] – quantifying to what extent the parts have shared or complementary information about the future (proven formally in Chapter 9). That is, if $\Phi > 0$, we infer that there is information in the whole which is not reducible to the parts (i.e. $\Phi > 0$ is a sufficient condition), but $\Phi \leq 0$ does not imply the opposite. Therefore, from an IIT perspective, a negative Φ can lead to the understandably confusing interpretation of a system having "negative integration," but, through a different lens (net synergy), it can be more easily interpreted as overall redundancy in the evolution of the system.

Box 1. Calculating whole-minus-sum integrated information Φ.

$$\Phi[X;\tau] = \varphi[X;\tau,\mathcal{B}^{\text{MIB}}] \tag{2.16a}$$

$$\mathcal{B}^{\text{MIB}} = \arg_{\mathcal{B}}\min \frac{\varphi[X;\tau,\mathcal{B}]}{K(\mathcal{B})} \tag{2.16b}$$

$$\varphi[X;\tau,\mathcal{B}] = I(X_{t-\tau};X_t) - \sum_{k=1}^{2} I(M^k_{t-\tau};M^k_t) \tag{2.16c}$$

$$K(\mathcal{B}) = \min\left\{H(M^1_t),H(M^2_t)\right\} \tag{2.16d}$$

1. For **discrete variables**:

$$I(X_{t-\tau};X_t) = \sum_{x,x'} p(X_{t-\tau}=x,X_t=x')\log\left(\frac{p(X_{t-\tau}=x,X_t=x')}{p(X_{t-\tau}=x)\;p(X_t=x')}\right)$$

2. For **continuous, linear-Gaussian variables**:

$$I(X_{t-\tau};X_t) = \frac{1}{2}\log\left(\frac{\det\Sigma(X_t)}{\det\Sigma(X_t\,|\,X_{t-\tau})}\right)$$

3. For **continuous variables** with an arbitrary distribution, we must resort to the nearest-neighbour methods introduced by [15]. See reference for details.

[5] Note that we use the symbol Φ to refer to both Eq. (2.16a) specifically, and to integrated information measures generically. The meaning should be clear from the context.

2.3.3 Integrated stochastic interaction $\tilde{\Phi}$

We next consider the second alternative for Φ_{2008} for spontaneous information dynamics: integrated stochastic interaction $\tilde{\Phi}$. Also introduced in Barrett and Seth [8], this measure embodies similar concepts as Φ, with the main difference being that $\tilde{\Phi}$ utilises a definition of effective information in terms of an *increase in uncertainty* instead of in terms of a *loss of information*.

$\tilde{\Phi}$ is based on *stochastic interaction* $\tilde{\varphi}$, introduced by Ay [16]. Akin to Equation (2.15), we define stochastic interaction beyond partition $\mathcal{P}$ as

$$\tilde{\varphi}[X;\tau,\mathcal{P}] =: \sum_{k=1}^{r} H(M_{t-\tau}^{k} \,|\, M_t^k) - H(X_{t-\tau} \,|\, X_t). \tag{2.17}$$

Stochastic interaction quantifies to what extent uncertainty about the past is increased when the system is split in parts, compared to considering the system as a whole. The details of the calculation of $\tilde{\Phi}$ are similar to those of Φ and are described in Box 2.

The most notable advantage of $\tilde{\Phi}$ over Φ as a measure of integrated information is that $\tilde{\Phi}$ is guaranteed to be non-negative. In fact, as mentioned above, φ and $\tilde{\varphi}$ are related through the equation

$$\tilde{\varphi}[X;\tau,\mathcal{P}] = \varphi[X;\tau,\mathcal{P}] + \mathrm{TC}(M_t^1;M_t^2;\ldots;M_t^r), \tag{2.18}$$

where TC is the *total correlation* [17] of the modules,

$$\mathrm{TC}(M_t^1;M_t^2;\ldots;M_t^r) = \sum_{k=1}^{r} H(M_t^k) - H(X_t). \tag{2.19}$$

This measure is also linked to *information destruction*, as presented by Wiesner et al. [18]. The quantity $H(X_{t-\tau}|X_t)$ measures the amount of irreversibly destroyed information, since $H(X_{t-\tau}|X_t) > 0$ indicates that more than one possible past trajectory of the system converged on the same present state, making the system irreversible and indicating a loss of information about the past states. From this perspective, $\tilde{\varphi}$ can be understood as the difference between the information that is considered destroyed when the system is observed as a whole, or split into parts. Note, however, that this measure is time-symmetric when applied to a stationary system; for stationary systems, total instantaneous entropy does not change with time. Furthermore, we know that $\tilde{\Phi}$ can exceed TDMI in some cases and that it quantifies a mixture of both causal and simultaneous influences [10].

Box 2. Calculating integrated stochastic interaction $\tilde{\Phi}$

$$\tilde{\Phi}[X;\tau] = \tilde{\varphi}[X;\tau,\mathcal{B}^{\text{MIB}}] \tag{2.20a}$$

$$\mathcal{B}^{\text{MIB}} = \arg_{\mathcal{B}} \min \frac{\tilde{\varphi}[X;\tau,\mathcal{B}]}{K(\mathcal{B})} \tag{2.20b}$$

$$\tilde{\varphi}[X;\tau,\mathcal{B}] = \sum_{k=1}^{2} H(M_{t-\tau}^{k} \,|\, M_{t}^{k}) - H(X_{t-\tau} \,|\, X_t) \tag{2.20c}$$

$$K(\mathcal{B}) = \min\left\{H(M_t^1), H(M_t^2)\right\} \tag{2.20d}$$

1. For **discrete variables**:

$$H(X_{t-\tau} \,|\, X_t) = -\sum_{x,x'} p(X_{t-\tau} = x, X_t = x') \log\left(\frac{p(X_{t-\tau} = x, X_t = x')}{p(X_t = x')}\right)$$

2. For **continuous, linear-Gaussian variables**:

$$H(X_{t-\tau} \,|\, X_t) = \frac{1}{2}\log\det\Sigma(X_{t-\tau} \,|\, X_t) + \frac{1}{2}n\log(2\pi e)$$

3. For **continuous variables** with an arbitrary distribution, we must resort to the nearest-neighbour methods introduced by [15]. See reference for details.

2.3.4 Integrated synergy ψ

Originally designed as a "more principled" integrated information measure [9], ψ shares some features with Φ and $\tilde{\Phi}$ but is grounded in a different branch of information theory, namely the Partial Information Decomposition (PID) framework [19]. In PID, the information that two (source) variables provide about a third (target) variable is decomposed into four non-negative terms as

$$I(X,Y;Z) = \texttt{Red}(X,Y;Z) + \texttt{Un}(X;Z|Y) + \texttt{Un}(Y;Z|X) + \texttt{Syn}(X,Y;Z)\ ,$$

where Un is the *unique information* one source but the other doesn't, Red is the *redundancy* between both sources and Syn is their *synergy*.

Integrated synergy ψ is the information that the parts provide about the future of the system that is exclusively synergistic—i.e. cannot be provided by any combination of parts

independently:

$$\psi[X;\tau,\mathcal{P}] =: I(X_{t-\tau};X_t) - \max_{\mathcal{P}} I_{\cup}(M^1_{t-\tau}, M^2_{t-\tau}, ..., M^r_{t-\tau}; X_t), \tag{2.21}$$

where

$$I_{\cup}(M^1_{t-\tau},\ldots,M^r_{t-\tau};X_t) =: \sum_{\mathcal{S}\subseteq\{M^1,\ldots,M^r\}} (-1)^{|\mathcal{S}|+1} I_{\cap}(\mathcal{S}^1_{t-\tau},\ldots,\mathcal{S}^{|\mathcal{S}|}_{t-\tau};X_t), \tag{2.22}$$

and $I_{\cap}(\mathcal{S}_1,\ldots,\mathcal{S}_{|\mathcal{S}|};Z)$ denotes the redundant information sources $\mathcal{S}_1,\ldots,\mathcal{S}_{|\mathcal{S}|}$ have about target Z. The challenge of PID is that it is, essentially, an underdetermined system of equations. For example, for the case of two sources, Shannon's information theory specifies three quantities ($I(X,Y;Z)$, $I(X;Z)$, $I(Y;Z)$), whereas PID specifies four (S, R, U_X, U_Y). Therefore, a complete operational definition of ψ requires a definition of redundancy from which to construct the partial information components [19]. In this sense, the main shortcoming of ψ, inherited from PID, is that there is no agreed consensus on a definition of redundancy [20, 21].

Here, we take Griffith's conceptual definition of ψ and we complement it with available definitions of redundancy (see Box 3). For the linear-Gaussian systems, we study here we use the minimum mutual information PID presented in Ref. [20].[6] Although we do not show any discrete examples here, for completeness, we provide complete formulae to calculate ψ for discrete variables using Griffith and Koch's redundancy measure [22]. Note that alternatives are available for both discrete and linear-Gaussian systems [19, 23–26].

[6] Barrett's derivation of the MMI-PID, which follows Williams and Beer and Griffith and Koch's procedure, gives this formula when the target is univariate. We generalise the formula here to the case of multivariate target in order to render ψ computable for Gaussians. This formula leads to synergy being the extra information contributed by the weaker source given the stronger source was previously known.

Box 3. Calculating integrated synergy ψ

$$\psi[X;\tau,\mathcal{P}] = I(X_{t-\tau};X_t) - \max_{\mathcal{P}} I_{\cup}(M^1_{t-\tau},\ldots,M^r_{t-\tau};X_t) \tag{2.23}$$

1. For **discrete variables**: (following Griffith and Koch's [22] PID scheme)

$$I_{\cup}(M^1_{t-\tau},\ldots,M^r_{t-\tau};X_t) = \min_{q} \sum_{x,x'} q(x,x') \log\left(\frac{q(x,x')}{q(x)\ q(x')}\right)$$
$$\text{s.t. } q(M^i_{t-\tau},X_t) = p(M^i_{t-\tau},X_t)$$

2. For **continuous, linear-Gaussian variables**:

$$I_{\cup}(M^1_{t-\tau},\ldots,M^r_{t-\tau};X_t) = \max_{k} I(M^k_{t-\tau};X_t)$$

3. For **continuous variables** with an arbitrary distribution: unknown.

2.3.5 Decoder-based integrated information Φ^*

Introduced by Oizumi et al. in Reference [10], decoder-based integrated information Φ^* takes a different approach from the previous measures. In general, Φ^* is given by

$$\Phi^*[X;\tau,\mathcal{P}] =: I(X_{t-\tau};X_t) - I^*[X;\tau,\mathcal{P}]\ , \tag{2.25}$$

where I^* is known as the *mismatched decoding information*, and quantifies how much information can be extracted from a variable if the receiver is using a suboptimal (or *mismatched*) decoding distribution [27, 28]. This mismatched information has been used in neuroscience to quantify the contribution of neural correlations in stimulus coding [29], and can similarly be used to measure the contribution of inter-partition correlations to predictive information.

To calculate Φ^*, we formulate a restricted model q in which the correlations between partitions are ignored,

$$q(X_t|X_{t-\tau}) = \prod_i p(M^i_t|M^i_{t-\tau}), \tag{2.26}$$

and we calculate I^* for the case where the sender is using the full model p as an encoder and the receiver is using the restricted model q as a decoder. The details of the calculation of Φ^* and I^* are shown in Box 4. Unlike the previous measures shown in this section, Φ^* does not

have an interpretable formulation in terms of simpler information-theoretic functionals like entropy and mutual information.

Calculating I^* involves a one-dimensional optimisation problem, which is straightforwardly solvable if the optimised quantity, $\tilde{I}(\beta)$, has a closed form expression [27]. For systems with continuous variables, it is in general very hard to estimate $\tilde{I}(\beta)$. However, for continuous linear-Gaussian systems and for discrete systems, $\tilde{I}(\beta)$ has an analytic closed form as a function of β if the covariance or joint probability table of the system are known, respectively. In Appendix A.1, we derive the formulae.[7] Conveniently, in both the discrete and the linear-Gaussian case, $\tilde{I}(\beta)$ is concave in β (proofs in Reference [27] and in Appendix A.1, respectively), which makes the optimisation significantly easier.

Box 4. Calculating decoder-based integrated information Φ^*

$$\Phi^*[X;\tau,\mathcal{P}] = I(X_{t-\tau};X_t) - I^*[X;\tau,\mathcal{P}] \tag{2.27a}$$
$$I^*[X;\tau,\mathcal{P}] = \max_{\beta} \tilde{I}(\beta;X,\tau,\mathcal{P}) \tag{2.27b}$$

1. For **discrete variables**:

$$\tilde{I}(\beta;X,\tau,\mathcal{P}) = -\sum_{x'} p(X_t = x') \log \sum_{x} p(X_{t-\tau} = x) q(X_t = x' | X_{t-\tau} = x)^{\beta} + \sum_{x,x'} p(X_{t-\tau} = x, X_t = x') \log q(X_t = x' | X_{t-\tau} = x)^{\beta}$$

2. For **continuous, linear-Gaussian variables**: (see appendix for details)

$$\tilde{I}(\beta;X,\tau,\mathcal{P}) = \frac{1}{2}\log\left(|Q||\Sigma_x|\right) + \frac{1}{2}\operatorname{tr}\left(\Sigma_x R\right) + \beta \operatorname{tr}\left(\Pi_{x|\tilde{x}}^{-1}\Pi_{x\tilde{x}}\Pi_x^{-1}\Sigma_{\tilde{x}x}\right)$$

3. For **continuous variables** with an arbitrary distribution: unknown.

2.3.6 Geometric integrated information Φ_G

In Reference [11], Oizumi et al. approach the notion of dynamical complexity via yet another formalism. Their approach is based on *information geometry* [30, 31], which studies families of probability distributions as differentiable manifolds. The natural metric in information geometry is the Fisher information metric, and the KL divergence provides a

[7] Note that the version written down in Reference [10] is incorrect, although their simulations match our results; we checked results from our derived version of the formulae versus results obtained from numerical integration, and confirmed that our derived formulae are the correct ones.

natural measure of (asymmetric) distance between probability distributions. Information geometry is the application of differential geometry to the relationships and structure of probability distributions.

To quantify integrated information, Oizumi et al. [11] consider the divergence between the complete model $p(X_{t-\tau}, X_t)$ and a *restricted model* $q(X_{t-\tau}, X_t)$ in which links between the parts of the system have been severed. This is known as the *M-projection* of the system onto the manifold of restricted models $Q = \{q : q(M_t^i|X_{t-\tau}) = q(M_t^i|M_{t-\tau}^i)\}$, and

$$\Phi_G[X;\tau,\mathcal{P}] =: \min_{q\in Q} D_{KL}\left(p(X_{t-\tau},X_t)\|q(X_{t-\tau},X_t)\right). \tag{2.28}$$

Key to this measure is that, in the partitioned system, it is only the connections that are cut; correlations between the parts are still allowed on the partitioned system. Although conceptually simple, Φ_G is very hard to calculate (see Box 5): there is no known closed form solution for any system, and we can only find approximate numerical estimates for some systems. In particular, for discrete and linear-Gaussian variables, we can formulate Φ_G as the solution of a constrained multivariate convex optimisation problem [32].

Box 5. Calculating geometric integration Φ_G

$$\Phi_G[X;\tau,\mathcal{P}] = \min_q D_{KL}(p\|q) \tag{2.29a}$$

$$\text{s.t. } q(M_{t+\tau}^i\,|\,X_t) = q(M_{t+\tau}^i\,|\,M_t^i). \tag{2.29b}$$

1. For **discrete variables**: numerically optimise the objective $D_{KL}(p\|q)$ subject to the constraints

$$\sum_{x,x'} q(X_{t-\tau}=x', X_t=x) = 1 \qquad \text{and} \qquad q(M_t^i\,|\,X_{t-\tau}) = q(M_t^i\,|\,M_{t-\tau}^i)\ \forall i.$$

2. For **continuous, linear-Gaussian variables**: numerically optimise the objective

$$\Phi_G[X;\tau,\mathcal{P}] = \min_{\Sigma(E)'} \frac{1}{2}\log\frac{|\Sigma(E)'|}{|\Sigma(E)|},$$

where $\Sigma(E) = \Sigma(X_t|X_{t-1})$, and subject to the constraints

$$\Sigma(E)' = \Sigma(E) + (A-A')\Sigma(X)(A-A')^{\mathrm{T}} \qquad \text{and}$$
$$(\Sigma(X)(A-A')\Sigma(E)'^{-1})_{ii} = 0.$$

3. For **continuous variables** with an arbitrary distribution: unknown.

2.3.7 Causal density

Causal density (CD) is somewhat distinct from the other measures considered so far, in the sense that it is a sum of information transfers rather than a direct measure of the extent to which the whole is greater than the parts. Nevertheless, we include it here because of its relevance and use in the dynamical complexity literature.

CD was originally defined in terms of Granger causality [33, 34], but here we write it in terms of Transfer Entropy (TE), which provides a more general information-theoretic definition [35]. The conditional transfer entropy from X to Y conditioned on Z is defined as

$$\mathrm{TE}_\tau(X \to Y \,|\, Z) =: I(X_t; Y_{t+\tau} \,|\, Z_t, Y_t). \tag{2.30}$$

With this definition of TE, we define CD as the average pairwise conditioned TE between all variables in X,

$$\mathrm{CD}[X; \tau, \mathcal{P}] =: \frac{1}{r(r-1)} \sum_{i \neq j} \mathrm{TE}_\tau(M^i \to M^j \,|\, M^{[ij]}), \tag{2.31}$$

where $M^{[ij]}$ is the subsystem formed by all variables in X except for those in parts M^i and M^j.

In a practical sense, CD has many advantages. It has been thoroughly studied in theory [36] and applied in practice, with application domains ranging from complex systems to neuroscience [37–39]. Furthermore, there are off-the-shelf algorithms that calculate TE in discrete and continuous systems [40]. For details of the calculation of CD, see Box 6.

Causal density is a principled measure of dynamical complexity, as it vanishes for purely segregated or purely integrated systems. In a highly segregated system, there is no information transfer at all, and, in a highly integrated system, there is no transfer from one variable to another beyond the rest of the system [12]. Furthermore, CD is non-negative and upper-bounded by the total time-delayed mutual information (proof in Appendix A.2), therefore satisfying what other authors consider an essential requirement for a measure of integrated information [11].

Box 6. Calculating causal density CD

$$\mathrm{CD}[X;\tau,\mathcal{P}] = \frac{1}{r(r-1)} \sum_{i \neq j} \mathrm{TE}_\tau(M^i \to M^j \,|\, M^{[ij]}) \tag{2.32}$$

1. For **discrete variables**:

$$TE_\tau(X^i \to X^j \,|\, X^{[ij]}) =$$
$$\sum_{x,x'} p\left(X^j_{t+\tau} = x'^j, X_t = x\right) \log \left(\frac{p\left(X^j_{t+\tau} = x'^j \,|\, X_t = x\right)}{p\left(X^j_{t+\tau} = x'^j \,|\, X^j_t = x^j, X^{[ij]}_t = x^{[ij]}\right)} \right)$$

2. For **continuous, linear-Gaussian variables**:

$$TE_\tau(X^i \to X^j \,|\, X^{[ij]}) = \frac{1}{2} \log \left(\frac{\det \Sigma\left(X^j_{t+\tau} \,|\, X^j_t \oplus X^{[ij]}_t\right)}{\det \Sigma\left(X^j_{t+\tau} \,|\, X_t\right)} \right)$$

3. For **continuous variables** with an arbitrary distribution, we must resort to the nearest-neighbour methods introduced by [15]. See reference for details.

2.3.8 Other measures

As already mentioned, all the measures reviewed here (besides CD) were inspired by the Φ_{2008} measure, which arose from the version of IIT laid out in Ref. [5]. The most recent version of IIT [6] is conceptually distinct, and the associated Φ-3.0 is consequently different to the measures we consider here. The consideration of perturbation of the system, as well as all of its subsets, in both the past and the future renders Φ-3.0 considerably more computationally expensive than other Φ measures. We do not here attempt to consider the construction of an analogue of Φ-3.0 for spontaneous information dynamics.

Recently, Tegmark [41] developed a comprehensive taxonomy of all integrated information measures that can be written as a distance between a probability distribution pertaining to the whole and one obtained as a product of probability distributions pertaining to the parts. Tegmark further identified a shortlist of candidate measures, based on a set of explicit desiderata. This shortlist overlaps with the measures we consider here, and also contains other measures which are minor variants. Of Tegmark's shortlisted measures, ϕ^{M} is equivalent to $\tilde{\Phi}$ under the system's spontaneous distribution, $\phi^{\mathrm{M}}_{kk'}$ is its state-resolved version, ϕ^{oak} is transfer entropy (which we cover here through CD), and ϕ^{npk} is not defined for continuous variables. The measures Φ_G and ψ are outside of Tegmark's classification scheme.

Part I

Integrated information in complex systems

Chapter 3

Integrated information and metastability in systems of coupled oscillators

Chapter summary

We begin our exploration of integrated information in complex systems with networks of coupled oscillators, frequently used as large-scale models of neural dynamics and capable of exhibiting a rich variety of so-called metastable chimera states. We bring metastability and integrated information together, by showing that these oscillators exhibit a critical peak of Φ that coincides with peaks in other measures such as metastability and coalition entropy.

3.1 Introduction

We begin our exploration of integrated information in complex systems by considering systems of coupled oscillators. Systems of coupled oscillators are ubiquitous both in nature and in the human-engineered environment, making them of considerable scientific interest [42]. A variety of mathematical models of such systems have been devised and their synchronisation properties have been the subject of much study. Typical studies of this sort, such as the classic work of Kuramoto [43], have examined the conditions under which systems of identical oscillators converge on a stable state of either full synchronisation or desynchronisation, perhaps identifying an order parameter that determines a critical phase transition from one state to the other. The collection of known attractors of such systems was enlarged with the discovery of so-called chimera states, in which the system of oscillators partitions into two stable subsets, one of which is fully synchronised while the other remains permanently desynchronised [44]. In addition to being mathematically interesting, chimera states have turned out to be empirically relevant in a wide variety of contexts, including neural dynamics [45–47], superconducting materials [48], earthquakes [49], animal flocking behaviour [50], and power grid networks [51].

Although systems of coupled oscillators that converge on stable states are both mathematically interesting and scientifically relevant, they are by no means representative of all real-world synchronisation phenomena. For example, the brain exhibits synchronous rhythmic activity on multiple spatial and temporal scales, but never settles into a stable state. Although it enters chimera-like states of high partial synchronisation, these are only temporary. Likewise, the chimera-like states of partial synchrony entered into by real flocks, herds, and swarms are always transient, and never truly stable. A system of coupled oscillators that continually moves from one highly synchronised state to another under its own intrinsic dynamics is said to be metastable.[1] In Ref. [53], it was shown that a modular network of phase-lagged Kuramoto oscillators will exhibit metastable chimera states under certain conditions. Variants of this model have since been used to replicate the statistics of the brain under a variety of conditions, including the resting state [45], cognitive control [46], and anaesthesia [47], as well as the dynamics of superconducting materials [48].

The dynamical properties of metastable oscillators, however, have rarely been discussed in conjunction with their informational properties. In particular, for the purposes of this thesis we are interested in the capacity of these coupled oscillators to integrate information, as measured by the tools introduced in Chapter 2. In this context, we can think of Φ measures as a proxy for the concept ot *dynamical complexity* [3]: a system is said to have high dynamical complexity if it exhibits a balance of integrated and segregated activity, where a

[1] Note that a system that traverses multiple highly synchronised states does not necessarily show chimera states, see e.g. [52].

system's activity is integrated to the extent that its parts influence each other and segregated to the extent that its parts act independently [54].

In this chapter, we connect these two lines of enquiry by demonstrating that modular networks of coupled oscillators of the sort described in Ref. [53] not only exhibit metastable chimera states, but also have high dynamical complexity according to IIT. Specifically, we show that measures of both phenomena peak in the narrow critical region of the parameter space wherein the system is poised between order and disorder. To our knowledge, this is the first description of a dynamical system in which the three major complexity indicators of criticality [55], metastability [56], and integrated information coincide. Moreover, as our work shows, IIT offers a rich picture of dynamical complexity in this critical regime, and constitute the first demonstration in this thesis that practical measures of integrated information have application outside theoretical neuroscience and can be applied to the analysis of complex systems.

3.2 Methods

We examine a system of coupled Kuramoto oscillators, extensively used to study non-linear dynamics and synchronisation processes [57]. We build upon the work of Shanahan [53] with a community-structured network of oscillators. The network is composed of 8 communities of 32 oscillators each, with every oscillator being coupled to all other oscillators in its community with probability 1 and to each oscillator in the rest of the network with probability $1/32$. The state of each oscillator i is captured by its phase θ_i, the evolution of which is governed by the equation

$$\frac{\mathrm{d}\theta_i}{\mathrm{d}t} = \omega + \frac{1}{\kappa+1}\sum_j K_{ij}\sin\left(\theta_j - \theta_i - \alpha\right) \ , \tag{3.1}$$

where ω is the *natural frequency* of each oscillator, κ is the average degree of the network, K is the *connectivity matrix* and α is a global *phase lag*. We set $\omega = 1$ and $\kappa = 63$. To reflect the community structure, the coupling between two oscillators i, j is $K_{ij} = 0.6$ if they are in the same community or $K_{ij} = 0.4$ otherwise. We tune the system by modifying the value of the phase lag, parametrised by $\beta = \pi/2 - \alpha$. We note that the system is fully deterministic, i.e. there is no noise injected in the dynamical equations.

We will analyse the trajectories of this system from two perspectives – through the concept of metastability from dynamical systems theory, and through the tools of information theory described in Chapter 2. In particular, we will use the measure of whole-minus-sum Φ given in Eq. (2.16).

3.2.1 Metastability

In this section we review the basic concepts behind metastability and how it is quantified, following a similar description to that of Ref. [53].

The building block of the dynamical quantities we study in this chapter is the *instantaneous synchronisation* R, that quantifies the dispersion in θ-space of a given set of oscillators. In general, we denote as $R_c(t)$ the instantaneous synchronisation of a community c of oscillators at time t, given by

$$R_c(t) = \left|\langle e^{i\theta_j(t)}\rangle_{j\in c}\right| . \tag{3.2}$$

To quantify metastability, we use the *metastability index* λ, which is defined as the average temporal variance of the synchrony of each community c, i.e.

$$\lambda_c = \mathrm{var}_t\, R_c(t) \tag{3.3a}$$

$$\lambda = \langle \lambda_c \rangle_c . \tag{3.3b}$$

Last, we also define *global synchrony* ξ as the average across time and space of instantaneous synchrony,

$$\xi = \left|\left\langle R_c(t)\right\rangle_{t,c}\right| . \tag{3.4}$$

According to Eq. (3.2), R_c (and therefore ξ) is bounded in the $[0,1]$ interval. $R_c(t)$ will be 1 if all oscillators in c have the same phase at time t, and will be 0 if they are maximally spread across the unit circle. This $[0,1]$ bound on R allows us to place an upper bound on λ, namely $\lambda_{max} = 1/9$. See Ref. [58] for details.

As defined in Eq. (3.3a), λ_c represents the size of the fluctuations in the internal synchrony of a community. A system that is either hypersynchronised or completely desynchronised will have a very small λ_c, whereas one whose elements fluctuate in and out of synchrony will have a high λ_c. In other words, a system of oscillators exhibits metastability if its elements remain in the vicinity of a synchronised state without falling into such a state permanently.

3.3 Results

We ran 1500 simulations with values of β distributed uniformly at random in the range $[0, 2\pi)$ using RK4 with a stepsize of 0.05 for numerical integration. Each simulation was run for 5×10^6 timesteps, of which the first 10^4 are discarded to avoid effects from transient states. All information-theoretic measures are reported in bits.

We first study the system from a purely dynamical perspective, following the analysis in Ref. [53]. Global synchrony and metastability are shown in Fig. 3.1. The first characteristic we observe is that there are two well differentiated dynamical regimes – one of hypersyn-

chronisation and one of complete desynchronisation, with strong metastability appearing in the narrow transition bands between one and the other.

It is in this transition region where the oscillators operate in a critical regime poised between order and disorder and complex phenomena appear. As the system moves from desynchronisation to full synchronisation there is a sharp increase in metastability, followed by a smoother decrease as the system becomes hypersynchronised. In the region $0 < \beta < \pi/8$, the system remains in a complex equilibrium between an ordered and a disordered phase.

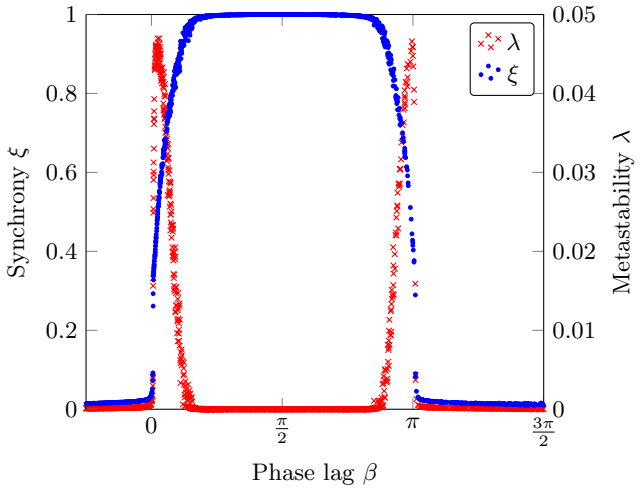


Figure 3.1: Global synchrony and metastability for different phase lags β for the whole $[0, 2\pi)$ range. Rapid increase of metastability marks the onset of the phase transition.

3.3.1 Information-theoretic analysis

One feature of Φ (and of any other information-theoretic measure), compared to λ, is that it need not be calculated directly from the state of the system. In other words, Φ is *substrate-agnostic*, meaning that the relevant quantity for the calculation of Φ is not the physical state of the system, but some *informational state* – the configuration of the system that we consider to contain information. Therefore, we must define an *informational state mapping*, that extracts the information-bearing symbols from the physical state of the system.

Although calculating Φ on the real-valued phases is possible, for simplicity we choose the *coalition configuration* of the system as the informational state, defined as the set of communities that are highly internally synchronised. To calculate the coalition configuration at time t we calculate $R_c(t)$ of each community and threshold it, such that

$$X_t^c = \begin{cases} 1 & \text{if } R_c(t) > \gamma \\ 0 & \text{otherwise.} \end{cases}$$

We refer to γ as the *coalition threshold*. After calculating the coalitions, the history of the system is reduced to a time series with 8 binary variables. Having a multivariate discrete time series, it is now tractable to compute Φ. We use $\gamma = 0.8$ in all our analyses shown here.

As depicted in Fig. 3.2, Φ shows a similar behaviour to λ – it peaks in the transition regions and shrinks in the fully ordered and the fully disordered regimes. We also compare Φ with arguably the simplest information-theoretic measure – entropy H. The entropy of the state of the network calculated on the coalitions X_t forms the *coalition entropy* H_c.

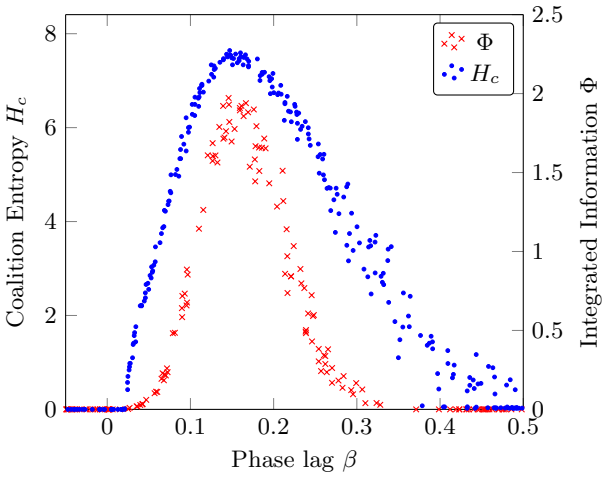


Figure 3.2: Integrated information Φ and coalition entropy H_c in the phase transition. Within the broad region between order and disorder in which H_c rises there is a narrower band in which complex spatiotemporal patterns generate high Φ.

Both Φ and H_c peak precisely at the same point. Although both measures depend on the chosen coalition threshold γ, the results are qualitatively the same for a wide range of thresholds. Although it peaks in the same region as λ and H_c, we note that Φ reveals new properties of the system by virtue of incorporating temporal information in its definition. That is, to have a high Φ a system must exhibit complex spatial *and* temporal patterns. We can verify this by performing a random time-shuffle on the time series. This shuffling leaves λ and H_c unaltered, as they don't explicitly depend on time correlations, but has a high impact on Φ, which shrinks to zero. This indicates that Φ is sensitive to properties of the system that are not reflected by other measures.

A continuous-time system, like the one considered here, can be integrated to an arbitrarily fine temporal resolution. This gives us an opportunity to use Φ to investigate the behaviour of the system at multiple timescales. Figure 3.3 shows the behaviour of Φ for several values of τ, and compares it with standard time-delayed mutual information (TDMI) $I(X_{t-\tau};X_t)$.

The first thing we note is that Φ and TDMI have opposite trends with τ. TDMI decreases for longer timescales, while Φ increases. At short timescales the system is highly predictable – thus the high TDMI – but this short-term evolution does not involve any system-wide

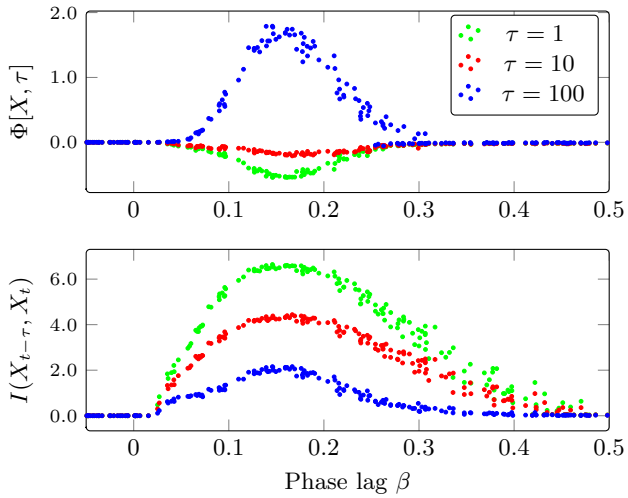


Figure 3.3: Integrated information Φ and time-delayed mutual information $I(X_{t-\tau};X_t)$ for several timescales τ. See text for details.

interaction – thus the low Φ. Furthermore, at these timescales Φ is negative, which can be interpreted as an indication of *redundancy* [20] in the evolution of the system: the parts share some information, such that they separately contain more information about their past than the whole system about its past. For larger τ TDMI decreases, as the evolution of the system becomes harder to track. Simultaneously, Φ becomes higher, indicating that the remaining TDMI has a stronger integrated component that is not accounted for by the TDMI of the partitions of the system. Overall, we see a clear trend of TDMI diminishing at longer timescales but becoming progressively more integrated in nature.

Finally, it is interesting to combine the insights from the dynamical and information-theoretic analyses. Inspecting Figs. 3.1 and 3.2 we see that the peak in Φ is much narrower than the peaks in λ and H_c. While some values of β do give rise to non-trivial dynamics, it is only at the centre of the critical region that these dynamics give rise to integration. A certain degree of internal variability is necessary to establish integrated information, but not all configurations with high internal variability lead to a high Φ. This means that Φ is sensitive to more complicated dynamic patterns than the other measures considered, and is in that sense more discriminating.

We note that λ is a community-local quantity – that is, the calculation of λ_c for each community is independent of the rest. Conversely, Φ relies exclusively on the irreducible interaction between communities. These two quantities are nevertheless intrinsically related, insofar as internal variability enables the system to visit a larger repertoire of states in which system-wide interaction can take place.

3.3.2 Robustness of Φ against measurement noise

We will now consider the impact of measurement noise on Φ, wherein the system runs unchanged but our recording of it is imperfect. For this experiment we run the (deterministic) simulation as presented in the previous section and take the binary time series of coalition configurations. We then emulate the effect of uncorrelated measurement noise by flipping each bit in the time series with probability p, yielding the corrupted time series $\hat{X}$. Finally we recalculate Φ on the corrupted time series, and show the results in Fig. 3.4. To quantify how fast Φ changes we calculate the ratio between the corrupted and the original time series,

$$\eta = \frac{\Phi[\hat{X}, \tau]}{\Phi[X, \tau]} \,. \tag{3.5}$$

In order to avoid instabilities as $\Phi[X, \tau]$ gets close to zero, we calculate η only in the region within 0.5 rad of the centre of the peak, where $\Phi[X, \tau]$ is large. The inset of Fig. 3.4 shows the mean and standard deviation of η at different noise levels p.

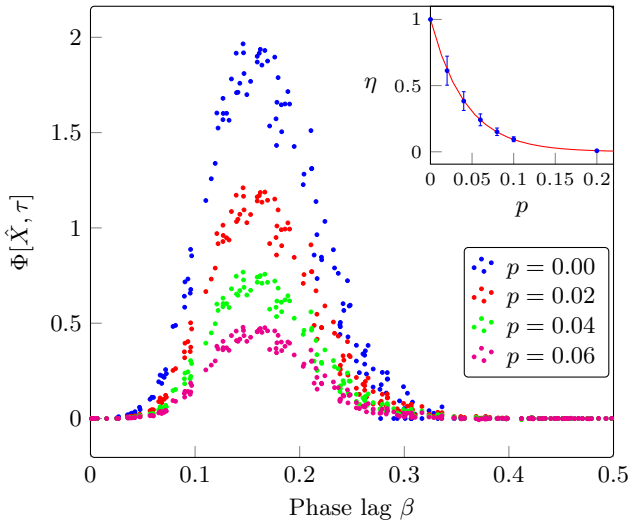


Figure 3.4: Integrated information Φ for different levels of measurement noise p. Inset: (blue) Mean and variance of the ratio η between Φ of the corrupted and the original time series. (red) Exponential fit $\eta = \exp(-p/\ell)$, with $\ell \approx 0.04$.

We find that Φ monotonically decays with p, reflecting the gradual loss of the precise spatiotemporal patterns characteristic of the system. The distortion has a greater effect on time series with greater Φ, but preserves the dominant peak in $\beta \approx 0.15$. The inset shows that both the mean and variance of η decay as a clean exponential with p. Φ is highly sensitive to noise and undergoes a rapid decline, as a measurement noise of 5% wipes out 70% of the perceived integrated information of the system.

3.4 Conclusion

We have presented a community-structured network of Kuramoto oscillators and discussed their collective behaviour in terms of metastability [53] and integrated information [5]. We showed that the system undergoes a phase transition whose critical region presents a sharp, clear peak of integrated information Φ that coincides with a strong increase in metastability. To our knowledge, this is the first description of a dynamical system in which the three major complexity indicators of criticality, high metastability, and high integrated information all appear. The resulting confluence of two major research directions in complexity science suggests that this is a system that merits further study.

In the context of the present model, the high internal variability of the system's components enables system-wide interaction, which in turn leads to high Φ. As we have seen, the system presents a region of high metastability, but notably it is only within an even narrower band that we find strong integrated information. The temporal profile of Φ gives us insight about the effective interaction timescale between parts of the system, and may be used in more heterogeneous systems to assess interdependencies more effectively. In this way we provide evidence that complex dynamics – as quantified by the metastability index λ – are a necessary but not sufficient condition for complex information processing – as quantified by integrated information Φ.

Dynamical and information-theoretic measures provide different lenses through which we can understand a system, and offer complementary views of its behaviour. Our findings support the claim that Φ is a valuable tool for understanding complex spatial and temporal behaviour in dynamical systems, particularly when combined with other analysis techniques.

Chapter 4

Balanced integration and segregation in modular spiking neural networks

Chapter summary

Next, we study a model network of spiking neurons with different coupling configurations. We find that information transfer and storage peak at two separate points for different values of the coupling parameter and are balanced at an intermediate point, where avalanches follow a long-tailed, power law-like distribution and reproduce empirical findings in the biological brain. In this way, we link together the balance between functional integration and differentiation (à la IIT) with the appearance of power law-like avalanches.

4.1 Introduction

Information theory has been an invaluable tool for neuroscience, and in the past few decades it has been making great contributions to our understanding of neural computation and coding [59]. This has inaugurated a whole research field bringing together both disciplines [60, 61]. The broad goal of this research is to describe neural computation in abstract terms, to dissociate the cognitive process from its neural implementation. This has proven to be a fruitful and interesting endeavour, since an abstract account would allow us to compare the brain with other cognitive systems, both biological and artificial.

However, in deliberately ignoring the physical substrate of computations (much like we did in Chapter 3), a purely information-theoretic view of the brain misses some interesting research questions: what kinds of dynamical states lead to what kinds of computations? Does a particular process make use of all resources available to the neurons? Could a given computation be instantiated by a different dynamical process? To address these and other questions we must consider the specific mechanisms by which the neural physical substrate gives rise to emergent computations.

For these reasons we advocate a hybrid view of neural computation, in which information and dynamics are two sides of the same coin [62]. Along these lines, several authors have established connections between information-theoretic and dynamical properties of neural networks at several scales: specific single-neuron-level mechanisms have been found to be informationally optimal in some sense [63, 64], and on a larger scale criticality has been linked to increased information transfer [65].

With these considerations in mind, in this chapter we proceed with our study of integrated information in complex systems, now with an explicit focus on the link between information processing and its neural physical substrate. Specifically, we study the effect of local and global coupling in a modular network of spiking neurons [54], and match the findings in our model with observed experimental data. As a further difference with respect to the previous chapter, instead of using a single Φ-like measure to quantify the balance between functional integration and segregation, we wish to *dissociate* these two, by using two separate measures and comparing them across model configurations.

We find that information transfer and storage, acting as proxy measures for functional integration and segregation respectively, peak at two separate points for different values of the coupling parameter, and are balanced at an intermediate point. In this configuration, avalanches in the network follow a long-tailed, power law-like distribution. Furthermore, the avalanche statistics at this point reproduce empirical findings in the biological brain [66], suggesting that the brain (or parts of it) may be operating in this balanced regime.

4.2 Methods

4.2.1 Model specification

We consider a system similar to the one shown in Ref. [54]. The network consists of a total of 1000 neurons, comprising one population of 200 inhibitory neurons and $n = 8$ populations (or *modules*) of 100 excitatory neurons each.

A total of 1000 internal one-directional connections (or *synapses*) are added to each excitatory module, such that any given pair of neurons are connected with probability 0.1 – resulting in modules of 10% edge density. Synapses from excitatory to inhibitory neurons are focal, with every 4 excitatory neurons in the same module projecting to the same inhibitory neuron. Every inhibitory neuron is connected to all other neurons in the network. The delay of each excitatory-excitatory synapse is sampled uniformly at random from the $[1,20]$ms interval, and the delay of all other synapses is fixed at 1 ms.

Once initialised, the network is subject to a *rewiring process*, akin to the one proposed by Watts and Strogatz [67]. Watts and Strogatz's key result is that the network undergoes a transition regime in which strong clustering coexists with short path lengths, making the network simultaneously segregated and integrated – termed a 'small-world' network. Here we seek to investigate how such small-world topological properties affect the dynamical and informational behaviour of the network.

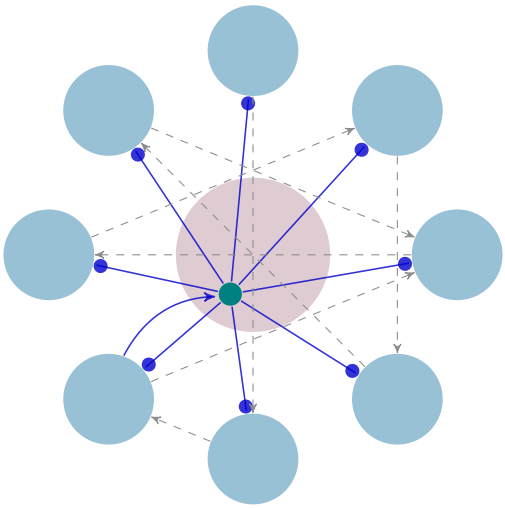


Figure 4.1: Schematic diagram of the model network. There are 8 excitatory modules (light blue) connected to one another and to a larger inhibitory pool (light purple). Inhibitory neurons have diffuse connections to all the network (blue round arrows), and each excitatory module has focal connections to a different set of inhibitory neurons (blue pointed arrow, green dot) and long-range connections to other excitatory modules (dashed gray arrows).

The rewiring process is only applied to the 800 excitatory neurons, and is implemented as follows. With probability p, each synapse is detached from its target neuron, and assigned

a new target picked uniformly at random from any excitatory module, thus introducing inter-module synapses. This rewiring probability p effectively regulates the balance between local intra-module coupling and long-range inter-module coupling, and is the main object of analysis in this chapter.

Once the topology of the network is set, we add a dynamical model to simulate the spiking behaviour of biological neurons. The dynamics of each neuron are simulated using the Izhikevich model [68],

$$\frac{dv}{dt} = 0.04v^2 + 5v + 140 - u + I \tag{4.1a}$$

$$\frac{du}{dt} = a(bv - u)\ , \tag{4.1b}$$

where v is the membrane potential (or voltage) of the neuron, u is an auxiliary recovery variable and I is the incoming current from ingoing synapses or external sources. All quantities are in arbitrary units. When the voltage of any given neuron goes above a certain threshold we record a discrete spike event, such that

$$\text{if } v \geq 30, \text{then} \begin{cases} v \leftarrow c \\ u \leftarrow u + d\ . \end{cases} \tag{4.2}$$

When a pre-synaptic neuron i spikes, all of its post-synaptic neurons j receive an instantaneous pulse with current equal to the weight W_{ij} of the synapse between them after a delay D_{ij} associated with that synapse. The values for the a, b, c, d parameters for both excitatory and inhibitory neurons are taken from Ref. [68]. All populations are slightly heterogeneous, as neuron parameters are randomised.

Once the network topology and the neuron parameters are set, the network can be simulated by numerically integrating Eqs. (4.1) and (4.2). We store all the spiking events from excitatory neurons for future analysis and ignore the spikes in the inhibitory population.

4.2.2 Functional segregation and integration

In this section we introduce some further information-theoretic methods we will use to analyse the system, complementing those in Chapter 2. The overall idea is that, instead of using a Φ-like measure to quantify the balance between functional integration and segregation, we wish to *dissociate* these two by using two separate measures inspired by recent research on information processing in complex systems [69]. In particular, we use *active information storage* [70] as a proxy for functional segregation; and a modified version of Causal Density (c.f. Chapter 2) as a proxy for functional integration.

Let us first describe active information storage, following Ref. [70]. The aim of AIS is to quantify the information in a time series X_t that can be retrieved by measuring its entire previous history. Mathematically,

$$\mathrm{AIS}(X_t) = \lim_{k\to\infty} I(X_t; X_{t-1}, X_{t-2}, \ldots, X_{t-k}) \ . \tag{4.3}$$

In other words, AIS quantifies how much information about the history of the system is useful in predicting the system's next state. As the $k \to \infty$ limit in Eq. (4.3) is (for obvious reasons) intractable, we write the finite-k approximation of the AIS of module i as

$$\mathrm{AIS}_k(M_{i,t}) = I(M_{i,t}^{(k)}; M_{i,t+1}) \ , \tag{4.4}$$

where $X_t^{(k)}$ is the k-dimensional embedding vector of X at time t, that contains the past k values of X_t up to and including time t. The aim of this embedding vector is to perform a state-space reconstruction of the underlying dynamical process [71], and is particularly useful when dealing with non-Markovian systems like the one in this chapter.[1] Further discussion of the AIS can be found in Refs. [39, 70, 72].

In the context of IIT, we interpret AIS loosely as a measure of segregated functional activity – information that is stored within each module separately.

In a similar fashion, to quantify functional integration we use Causal Density (CD), earlier introduced in Chapter 2. However, following the same rationale as for AIS_k, we formulate an embedding version of CD applicable to non-Markovian systems. In particular, the standard conditional transfer entropy is replaced by the conditional mutual information between the embedded time series,

$$\mathrm{CTE}_k(X \to Y|Z) = I(X_t^{(k)}, Y_{t+1}|Y_t^{(k)}, Z_t^{(k)}) \ , \tag{4.5}$$

which is then used to define an embedded version of causal density, akin to Eq. (2.32) but formulated in terms of vector embedings,

$$\mathrm{CD}_k(S) = \frac{1}{n(n-1)} \sum_{ij} \mathrm{CTE}_k(M_i \to M_j|S_{[ij]}) \ , \tag{4.6}$$

where M_i represents the activity of module i and $S_{[ij]}$ represents the whole system with variables M_i and M_j removed.

[1] Note that an Izhikevich spiking neuron is Markovian at the level of dynamical variables v, u, but not at the level of spikes.

4.3 Results

We generate 400 networks with different values of p sampled uniformly at random in the $[0,1)$ interval, and another 200 with p sampled exponentially at random in the $(10^{-2},1)$ interval. This is to have dense coverage of the parameter space at the low end of the range. The activity of each network is simulated using the NeMo library [73] for 200 s using the RK4 method with a timestep of 0.2 ms, and subsampled to a resolution of 1 ms. The first 1 s of simulation is discarded to avoid transient effects. Information-theoretic quantities are calculated using the implementation in Ref. [74] and are reported in bits.

4.3.1 Model behaviour

In this section we give a qualitative summary of the behaviour of the model that will help interpretation of quantitative findings described in the rest of the chapter. Spike raster plots of representative runs of the network with different values of p are shown in Fig. 4.2.

We begin with the fully modular network, the $p = 0$ case (bottom panel in Fig. 4.2). In this setting there are no direct connections between the excitatory modules. When any neurons in an excitatory module become active, the high density of intra-module synapses ensures that all neurons in the module quickly become activated.

Through the focal excitatory-inhibitory synapses, the active module feeds charge to the subset of inhibitory neurons assigned to it. These start spiking rapidly, and because of the diffuse connections they shut down the activity in all other excitatory modules. This results in competitive multistable dynamics, as one module gaining control of the network prevents all others from doing so. In computational neuroscience this kind of competition mechanism is known as Winner-Take-All (WTA). Subsequently, the active module saturates and the refractory period of the neurons makes it cease firing, so that other module can take over.

At the other end of the parameter range, at $p = 0.9$ (top panel in Fig. 4.2), the dynamics are very different. Topologically, this setting is closer to a fully random, Erdős-Rényi network (which is the case exactly for $p = 1$). There is no notion of modules anymore, and all excitatory neurons are statistically equivalent. The result is an interaction between a uniform population of excitatory neurons with a smaller group of inhibitory neurons. This is reminiscent of a known mechanism of oscillation generation – a PING architecture [75]. The interplay between excitation and inhibition and the synaptic delays between them make the whole system oscillate. In this regime the modules are strongly correlated and cooperate in maintaining the global oscillation.

Finally, at intermediate values of p (middle panel in Fig. 4.2), these two opposite trends coexist. The dynamics of the system are more chaotic and there is no clear pattern. Local and long-range coupling are balanced and both affect the emergent dynamics (we recall that

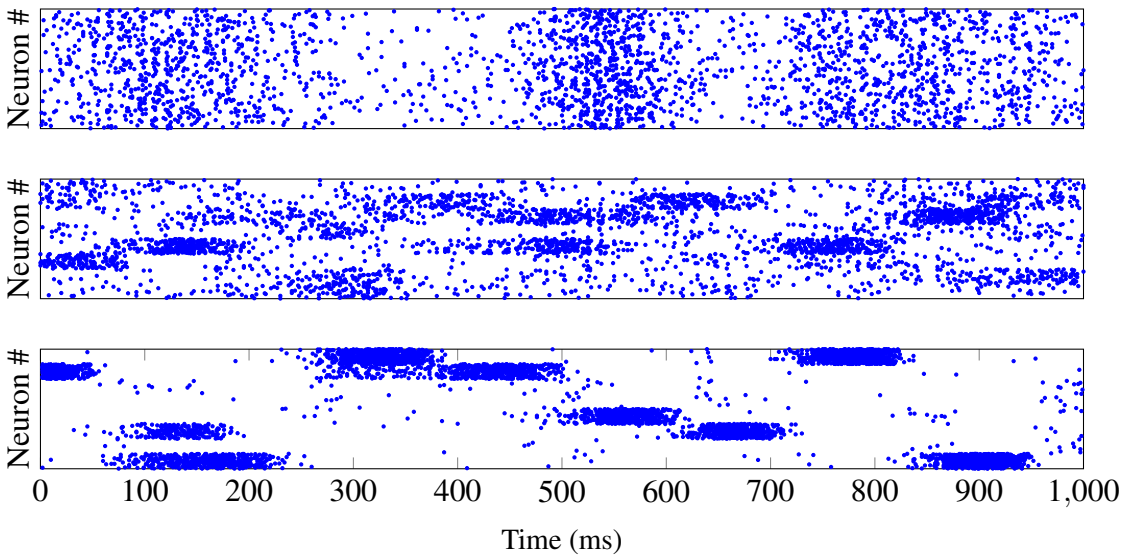


Figure 4.2: Sample runs of the network for different values of the rewiring probability p. The values are $p = 0.9$ (top), 0.2 (middle) and 0 (bottom). As p increases the system transitions from multistable competitive dynamics to oscillatory cooperative dynamics.

the total number of connections is fixed, so an increase in long-range coupling is always at the expense of a weaker intra-module coupling).

This transition is also interpretable as an emergent synchronisation phenomenon. For low p the WTA mechanism pushes the modules out of phase, and the network is maximally desynchronised. Conversely, for high p the modules blend together and the synchrony between them increases.

In summary, the model we described features a transition from a competitive to cooperative regime, controlled by a continuous parameter. This model naturally interpolates between two neural circuits ubiquitously present in the cortex: PING oscillators and multistable WTA circuits. As we describe below, it is between these two extremes where critical dynamics and complex information processing take place.

4.3.2 Avalanche statistics

The seminal work of Beggs and Plenz [66] set out the search for criticality in neural systems, in particular through the analysis of avalanche statistics. A *neural avalanche* is defined as a period of continued spiking activity – i.e. a period in which the activity of a neural population is continuously above a certain *avalanche threshold*. The *avalanche size* is the total number of spikes fired by all neurons in the population between any two points of below-threshold activity.

By counting occurrences of avalanches in the network and recording their size we obtain the *avalanche size distribution*, a very relevant mathematical construct subject of much study in statistical physics and complex systems research. A common signature of critical dynamics and phase transitions is that avalanche sizes follow a *power law distribution* [76], defined as

$$P(s) \propto s^{-\alpha} \,, \tag{4.7}$$

where α is called the *critical exponent*. Beggs and Plenz's key result is that measured activity in the biological brain consistently follows a power law avalanche size distribution – leading to the hypothesis that the brain operates in a critical regime. Although their claims on criticality have been contested [77], their empirical finding of power law distributions in neural recordings is widely accepted.

In this section we present the avalanche analysis of the resulting activity of our model. As a general methodological note, we mention that estimating and evaluating power laws when working with empirical data is remarkably complicated. In this analysis we use the methods and implementation provided in Refs. [78, 79].

We generate and run networks for many values of p as described above and measure avalanches in each module. To do this we calculate the mean firing rate of each module over 1 ms bins and run the analysis with an avalanche threshold of 3 spikes/ms. The distributions of the 8 modules in the same run of the experiment are aggregated together to improve statistics. Log-log avalanche size histograms are shown for evenly spaced values of p in Fig. 4.3.

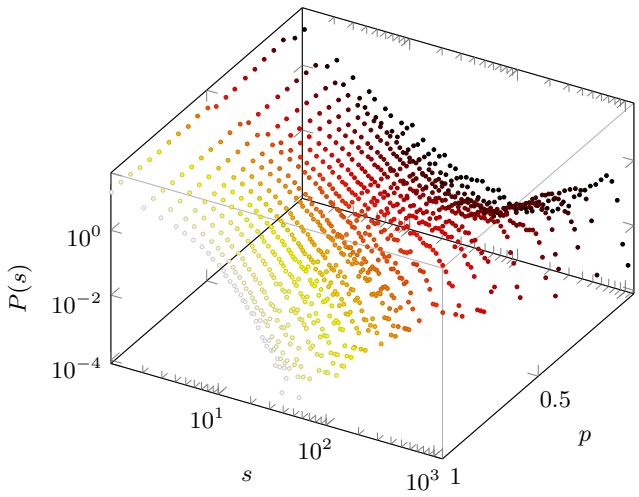


Figure 4.3: Avalanche size distributions for different rewiring probabilities p. As connections become delocalised (increasing p), the system shifts from supercritical (high probability of large avalanches) to subcritical (low probability of large avalanches). The p-axis is reversed for visualisation purposes (p decreases going into the page).

For low values of p, when connections are highly localised, the system is supercritical – the avalanche size distribution is characterised by a prominent peak at the far tail, which indicates that a disproportionately large fraction of the avalanches are strongly energetic and saturate the modules.

Conversely, at high values of p the system is subcritical. Avalanches are weak and the avalanche size distribution has a short exponential tail. This is probably caused by the diffuseness of the connectivity pattern – rewiring keeps the global synaptic strength fixed, but the influence of each burst of activity is spread across the whole network instead of focalised in one single module.

It is at middle that the activity of the modules resembles the activity of a critical system. Avalanche size distributions show power law-like statistics, with a characteristic straight line in the log-log histogram and a small protuberance at the end, result of finite-size effects. To test the claim that the behaviour of the system is closest to a power law at an intermediate value of p, we perform a maximum-likelihood power law fit to each trial and calculate the 1-sample Kolmogorov-Smirnov (KS) statistic between the measured data and the fitted power law. The results, together with three representative histograms are shown in Fig. 4.4.

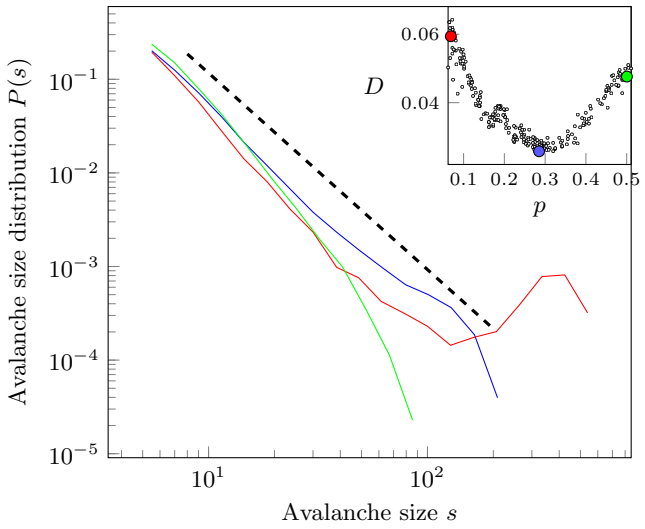


Figure 4.4: Avalanche size distributions for three runs of the simulation, supercritical (red), subcritical (green) and critical (blue). Black line: reference $\alpha = 2$ power law as reported by [66, Fig. 3A] for 1 ms-binned LFP data. Inset: Kolmogorov-Smirnov statistic D comparing the data against a theoretical power law with the estimated parameters. Filled colour circles in the inset correspond to the runs shown in the main plot.

This figure more clearly shows the difference between critical, subcritical and supercritical behaviour; and the KS statistic determines that at $p = 0.3$ the system's avalanche size distribution is closest to a power law. Furthermore, at that point the critical exponent of the maximum-likelihood fit is consistent with the $\alpha \approx 2$ value found by Beggs and Plenz for

1 ms-binned local field potential (LFP) data,[2] by de Arcangelis and Herrmann in simulated scale-free networks, and by Levina and Priesemann in recent studies of developing neurons *in vivo* [66, 65, 80].

4.3.3 Information-theoretic analysis

In practice, the challenge behind computing information-theoretic measures amounts to estimating probability densities for the involved quantities (e.g. $p(M_{i,t+1}|M_{i,t}^{(k)})$ and $p(M_{i,t+1})$ in the case of AIS). For our analyses we use the nearest-neighbour estimators devised by Kraskov, Stögbauer and Grassberger [15]. The KSG estimators are non-parametric and make only weak assumptions on the local neighbourhoods of the estimated probability density, which makes them a robust, flexible tool. Reported results are corrected with surrogate data methods [81].

More importantly, we measure information storage and transfer with AIS and CD and show the results in Fig. 4.5. AIS is calculated separately for each module and then the 8 modules are averaged for each run. Nonparametric CD is calculated as described in Eq. (4.6). The embedding dimension k is fixed at 5 for all calculations.

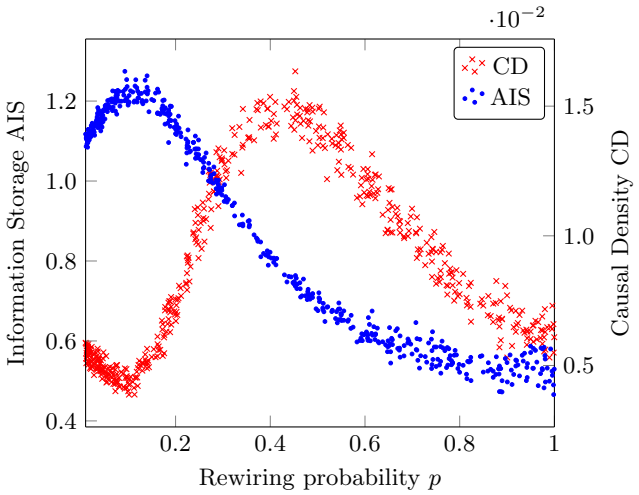


Figure 4.5: Active information storage (blue, left axis) and nonparametric causal density (red, right axis) for different rewiring probabilities p. Each measure peaks at one side of the critical region around $p = 0.3$ where the system shows power law-like statistics (see Fig. 4.4).

First we note that information storage dominates the low-p regime. Because of the WTA competition mechanism, if a module is inactive it tends to remain inactive, whereas if it is

[2]Note that one of Beggs and Plenz's results is that the exponent depends on the bin size, and that the well-known 3/2 exponent later publicised corresponds to a bin size equal to the average inter-spike interval, which is larger than 1 ms.

active it will most likely saturate and cease activity shortly after. This means that the recent history of the module's activity is highly informative of their future.

Regarding transfer, CD has a prominent peak in the mid-p region. As expected, there is little transfer in the $p = 0$ or $p = 1$ extremes, away of the neighbourhood around the critical transition. This is because in the low-p regime the modules are completely disconnected; and in the high-p regime the modules are so correlated that module i no longer provides information about j after conditioning on the rest of the modules.

More interesting is the neighbourhood around $p = 0.3$, where storage and transfer are maximally balanced. This coincides with the point where the avalanche dynamics are closest to a power law, as measured by the KS statistic and shown in Fig. 4.4. This suggests that there is a configuration of the system in which the balance between local and global coupling results in a balance between local information storage and long-range information transfer, which moreover is accompanied by a near-critical avalanche distribution. This finding links together three complementary views on neural computation: topological, informational and dynamical complexity.

4.3.4 Criticality and linear interactions

As we have argued above, the system exhibits a transition from competitive to cooperative dynamics as coupling shifts from short- to long-range. This transition is accompanied, in the large scale, by power law-like avalanche statistics and a balance of information storage and transfer. In this section we explore the signatures of the transition in the pairwise relationships between the activations of different modules in the network.

To study the spectral aspects of the model's dynamics, we analyse the time series of module firing rates under three filtering conditions:

- Raw (unfiltered) data.
- After first-order differencing ($X'_t = X_t - X_{t-1}$).
- After second-order differencing ($X''_t = X'_t - X'_{t-1}$).

Since time-differencing is essentially a highpass filter, by taking successive differences we are effectively exploring higher regions of the network's frequency spectrum. Figure 4.6 shows aggregated histograms of the activity of all pairs of modules for growing values of p and the three filtering conditions.

For high p these joint distributions visually appear Gaussian, suggesting that the relationships between module activations are mostly linear in this regime. For lower p, however, the WTA dynamics are clearly visible and distributions are heavily nonlinear. Interestingly, the transition between the linear and nonlinear regimes lies in the $p \in (0.2, 0.4)$ range, where information processing is most diverse and avalanches exhibit power law-like statistics.

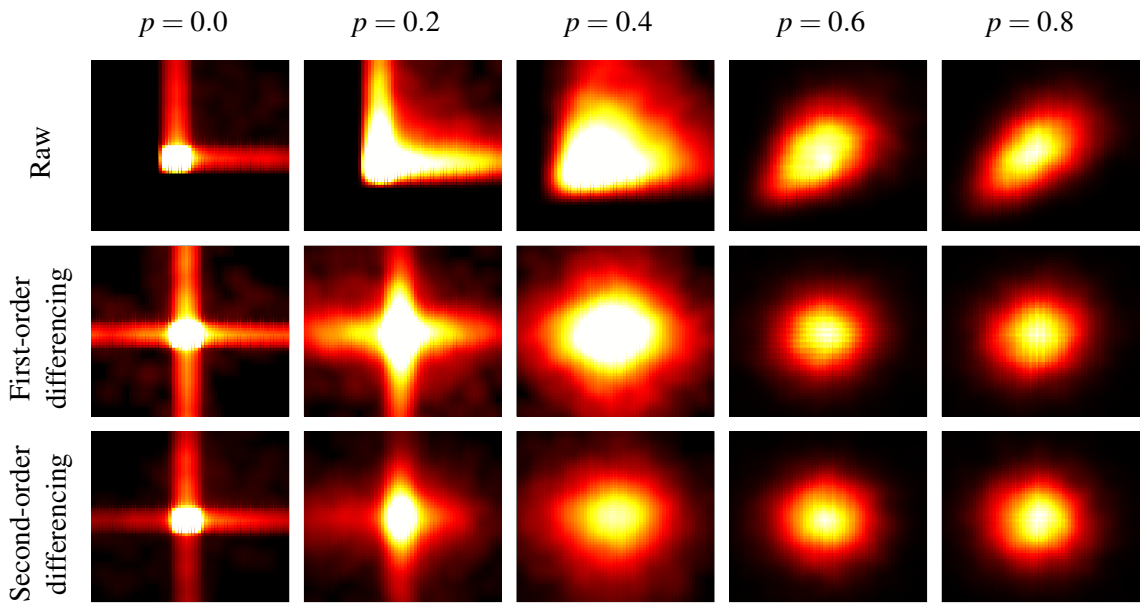


Figure 4.6: Heatmaps depicting the joint density distribution of the activations of two modules M_t^i, M_t^j (i.e. for each timestep, activation of module i in the x axis and module j in the y axis, aggregated for all pairs $i \neq j$) for the raw time series (top), after first-order time differencing (middle) and after second-order time differencing (bottom). As the rewiring probability p increases these distributions become increasingly Gaussian, indicating that the relationship between the activations becomes more linear.

As a rough quantitative measure for nonlinearity, we calculate how much information is accounted for by linear relationships. To do so we compare MI between modules using two methods: the nonparametric KSG estimator, and a parametric estimator under the assumption that all interactions are linear with Gaussian noise.

The latter is referred to as the *linear-Gaussian* estimator, and it assumes that all variables in the system are jointly distributed as a multivariate Gaussian distribution (i.e. assumes that all histograms in Fig. 4.6 are Gaussian). In this case all relevant information-theoretic quantities can be calculated analytically from the joint covariance matrix of the system [7, Chapter 9]. Under this assumption, the nonlinear component of the distribution is ignored. Therefore, the difference between the KSG and linear-Gaussian estimators is a good proxy for how much weight the nonlinear component of a distribution carries.[3]

To illustrate the effect of this assumption-breaking on informational measures, in Fig. 4.7 we show the average MI between all pairs of modules (i.e. the MI between the two variables shown in the histograms of Fig. 4.6) calculated with the linear-Gaussian estimator and with the nonparametric KSG estimator.

[3]Note, however, that the linear-Gaussian MI is not a lower bound to the true MI [82].

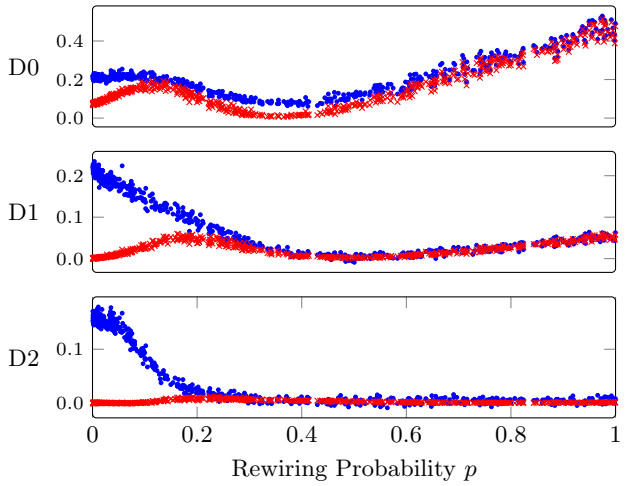


Figure 4.7: Mutual information between pairs of modules using linear-Gaussian (red) and nonparametric (blue) estimators. The system has a stronger non-linear component in the higher frequencies of the low-p regime, as indicated by the discrepancy between the two estimators.

As expected, the linear estimator always lies (up to random fluctuations) below the KSG. By considering the linear effects only, linear methods effectively provide a lower bound of the true MI. The linear-Gaussian estimator is very close to the KSG for high p but consistently below it in the low-p range, which validates our claim that interactions shift from nonlinear to linear with increasing p. Furthermore, the gap between both estimators is more prominent in the differenced time series, indicating that lower frequencies, which are more strongly suppressed by time-differencing, are mostly responsible for the linear component of the mutual information between modules.

4.4 Conclusion

In this chapter we studied a simple modular spiking neural network and used it to explore the relation between dynamics, information processing and underlying network topology. The fully modular setting implements a WTA mechanism, whereas the fully random setting is comparable to a PING oscillator – both of which are ubiquitous neural circuits in biological brains. This model gives us a way of interpolating between the two in a continuous fashion by varying a long-range connectivity parameter, p.

We find that for intermediate values of p the network passes through a near-critical regime in which avalanches display power law-like statistics, with the same critical exponent as found in biological brains [66]. Measures of information storage and transfer peak at either side of the critical point, and the point where they are maximally balanced coincides with the point where avalanches are closest to a power law.

This transition can also be understood as a breakdown of linearity, with cooperative linear interaction being prevalent when connectivity is global and delocalised, and competitive winner-take-all interaction more prominent when connectivity is local.

Taken together, these findings link together three complementary views on neural computation: topological, informational, and dynamical complexity.

Chapter 5

Integrated information and distributed computation in cellular automata

Chapter summary

As a final complexity case study, we consider distributed computation in cellular automata. We relate IIT to distributed computation in two ways: at a global scale, Φ is higher for complex, class IV automata; and at a local scale Φ is higher for emergent coherent structures, like blinkers, gliders, and collisions. Together with the previous two examples, this suggests Φ is, empirically, a good candidate for a universal marker of complexity.

5.1 Introduction

Cellular Automata (CA) are an important class of discrete dynamical models widely used in the study of complex systems and distributed computation. They have been used in the study of neural dynamics [83], ecological phenomena [84], and biological evolution [85], among many other disciplines; and are a quintaessential example of complex behaviour arising from simple local interactions [86]. Some particularly simple automata, most famously elementary automaton 110, are known to be capable of universal computation [87].

Historically, these automata have been at the core of early information-theoretic studies of complexity [88]. Their simplicity and flexibility make them excellent subjects of study – even simple rules can display a broad diversity of behaviour, and their discrete state-space makes it (relatively) easy to estimate probability distributions. It is through studying these automata and similar systems that information theory has been successfully used to formalise previously fuzzy concepts around complexity and self-organisation [89, 90].

Part of the early work on CA was due to their appeal as theoretical models of *distributed computation* – they are Turing-complete, in the sense that they can emulate any Turing machine; but they are radically unlike Turing machines in the sense that *there is no centralised element (like the machine's "head") that performs the computation*. Instead, it is groups of cells that, by each following local rules, are able to *collectively* perform the computation. The idea that the brain may be studied as such a distributed computing device has been tremendously influential throughout the cognitive sciences, embodied in the work on connectionism and parallel distributed processing [91]. They have also influenced the study of consciousness, nowhere more explicitly than in Baars' global workspace theory [92, 93].

One peculiar feature of CA is that this distributed computation is instantiated by distinct *emergent structures*, typically thought of as *particles* – coherent spatio-temporal structures that may move and collide with one another, thereby instantiating certain kinds of computation [94]. In his brilliant PhD thesis, Lizier [69] provided a comprehensive study of information processing through emergent structures in ECA, which he achieved through the formulation of local information-theoretic measures.

In this chapter we build on Lizier's work by applying measures of integrated information to the analysis of cellular automata. We show that, on a global scale, automata capable of more sophisticated computation tend to have higher Φ; and, using a novel local measure of integrated information, we show that this information is integrated in the aforementioned emergent structures. These results support our claim that integrated information is a landmark feature of the distributed computation and emergent dynamics observed in cellular automata.

5.2 Methods

5.2.1 Elementary automata and complexity classes

Our analysis begins with Elementary Cellular Automata (ECA), which constitute a particular subclass of CA. In ECA, agents (or *cells*) are arranged in a one-dimensional cyclic array (or *tape*). The state of each cell at a given time step has two possible states, 0 or 1, and is a boolean function of the state of itself and its immediate neighbours at the previous time step. The same boolean function dictates the time evolution of all cells, inducing a spatial translation symmetry. Hence, in the case of ECA, each of the 256 different boolean functions of three binary inputs induces a different *evolution rule*. Rules are then enumerated from 0 to 255 and each ECA, irrespective of its number of agents, can be classified by its rule. For a more detailed description of ECA and their numbering system, see Ref. [86].

Early influential work by Wolfram [95] introduced the notion of *complexity class* to classify ECA in four groups. According to Wolfram, an ECA belongs to each class I-IV if when started from random initial conditions it generates:

I A simple attractor with a very small number of states.

II Simple periodic attractors.

III Chaotic, seemingly random patterns.

IV Complex, highly structured patterns with persistent structures.[1]

In this way, rules of increasing class number display increasingly complex behaviour. In fact, Wolfram hypothesised that all class IV automata may be Turing-complete – and, although some of them have been proven to be so [87, 97, 98, 94], the problem of determining to which class a particular CA belongs to is in general undecidable [99, 100], casting doubt on the classification scheme itself. Nonetheless, for our purposes here we will consider an automaton's class number as an approximate indicator of the complexity of its behaviour, and study its relation with Φ and other information-theoretic quantities.

5.2.2 Local information dynamics

In his PhD thesis and accompanying articles, Lizier [69, 101, 70, 102] used information theory to study the local information dynamics of cellular automata, and convincingly showed that computation is implemented in coherent, emergent structures known as *particles*. Key to this finding was the focus on *local* (or *pointwise*) information measures, evaluated throughout the temporal evolution of an ECA.

[1] This classification, originally made loosely by Wolfram based on visual inspection, was later formalised in terms of fixed points, limit cycles, and strange attractors [96].

As an example, the local mutual information is given by

$$i(x;y) = \log \frac{p(x,y)}{p(x)p(y)} \,, \tag{5.1}$$

which is the integrand of Eq. (2.6), such that the expected value of i is the standard mutual information, $I(X;Y) = \mathbb{E}[i(x;y)]$. By evaluating i on every x,y pair, we can elucidate which particular combinations of symbols are responsible for the measured correlation between X and Y. This reasoning has been successfully used recently in the context of information decomposition [103], which will be explored further in Chapters 8 and 9.

Complementing the formulation of local measures, Lizier proposed a taxonomy of distributed information processing as composed of *storage*, *transfer* and *modification*.[2] In this framework, storage is identified with *excess entropy*, $E = I(X_t^{(k)}; X_{t+1}^{(k^+)})$, and transfer with *transfer entropy* (c.f. Eq. (2.30)) – with their corresponding local versions

$$e_k(x_t) = \log \frac{p(x_t^{(k)}, x_{t+1}^{(k^+)})}{p(x_t^{(k)})p(x_{t+1}^{(k^+)})} \,, \tag{5.2}$$

$$t_k(y_t \to x_t) = \log \frac{p(x_{t+1}|x_t^{(k)}, y_t^{(k)})}{p(x_{t+1}|x_t^{(k)})} \,, \tag{5.3}$$

where a subscript (k) indicates the vector embedding as used in Chapter 4, and $x_t^{(k^+)}$ is a future embedding vector $x_t^{(k^+)} = \{x_t, x_{t+1}, ..., x_{t+k-1}\}$.

In this chapter we extend the standard formulation of integrated information measures in two ways: first, we add embedding vectors, so they are applicable to non-Markovian systems; and second, we formulate pointwise measures to be applied on a local spatio-temporal scale.

Mathematically, these modifications are straighforward: we reformulate the effective information in Eq. (2.16) to add embedding vectors,

$$\varphi_k[X;\tau,\mathcal{B}] = I(X_{t-\tau}^{(k)}; X_t) - \sum_{j=1}^{2} I(M_{t-\tau}^{j,(k)}; M_t^j) \,, \tag{5.4}$$

and apply the same partition schemes described in Sec. 2.3.1 to obtain a 'big-phi' integrated information, Φ_k. Additionally, the equation above can be readily made into a local measure by replacing mutual information with its local counterpart,

$$\phi_k[x_t;\tau,\mathcal{B}] = i(x_{t-\tau}^{(k)}; x_t) - \sum_{j=1}^{2} i(m_{t-\tau}^{j,(k)}; m_t^j) \,. \tag{5.5}$$

[2] Further discussion on this taxonomy, specifically about information modification, will be presented in Chapter 9.

5.3 Results

For all the results presented below, we follow the same simulation parameters used by Lizier in his study of local information dynamics in ECA [69]. Specifically, we initialise a tape of length 10^4 with random i.i.d. binary variables, discard the first 100 steps of simulation, and run 600 more steps that are used to estimate the probability distributions used in the information-theoretic measures described below.

5.3.1 Integrated information and complexity class

As a first experiment, we calculate the average integrated information of each ECA, Φ_k, separating each automaton by its complexity class. We used the rules given in Wolfram's original article [95], as well as other clear-cut rules, and excluded border cases which did not neatly fit into one single category. The results are shown in Figure 5.1.

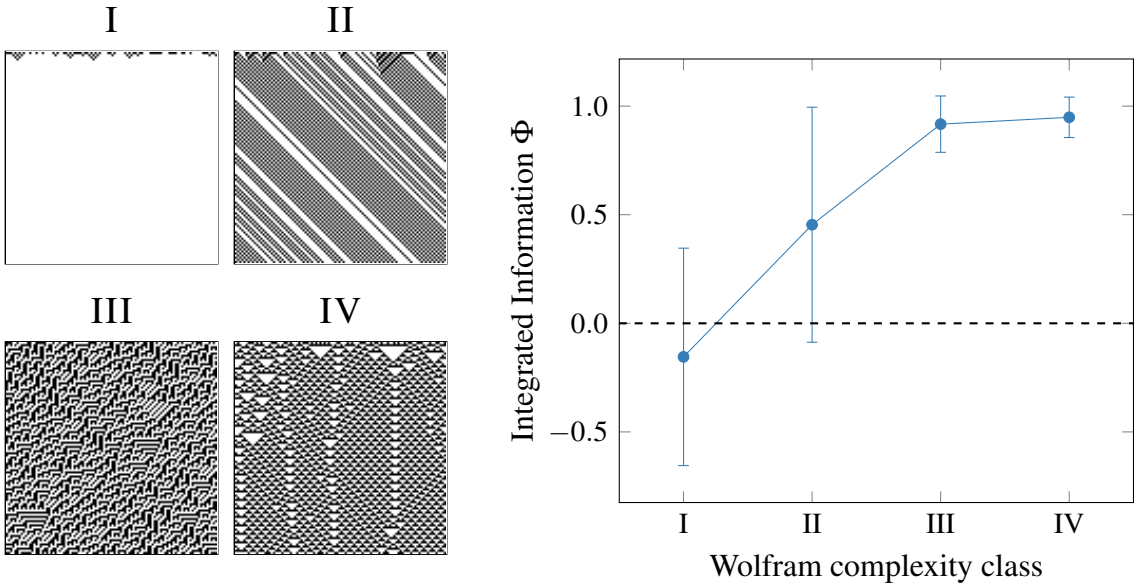


Figure 5.1: Integrated information grows monotonically with Wolfram class number. (Left) Examples of each complexity class (ECA rules 32, 56, 75, and 54, respectively), showing noticeable differences in behaviour. Notice the presence of localised particles in the class IV rule. (Right) Correspondingly, Φ is highest for the more complex classes IV and III, and lower (and often negative) for the simpler behaviours in classes I and II.

The clear result is that complexity, as measured by Φ, correlates strongly with complexity as discussed by Wolfram – automata of higher classes have consistently higher Φ than automata of lower classes, and the difference between classes I,II and III,IV is stark. In fact, some low-complexity rules (especially in class I, but also in class II) have *negative* Φ, indicating a prevalence of redundancy in their dynamics (c.f. Chapter 9).

It is worth noting the small difference between classes III and IV. This is likely related to the blurriness of the line separating both classes – visually, it is hard to judge whether structures are "coherent enough" and, formally, the problem of determining whether a particular rule belongs to class III or IV is considered undecidable [99, 104]. Based on this, we may temptatively suggest that the capacity to integrate information is a neccesary, but not sufficient, condition for universal computation.

5.3.2 Integrated information at the edge of chaos

In his seminal 1990 article, Langton [96] took a step beyond Wolfram's classification, and argued that the complexity and universality observed in ECA may reflect a broader phenomenon he called *computation at the edge of chaos*. In this view, computation is made possible by indefinitely long transient states, a manifestation of *critical slowing-down* [105], that form the particle-like structures seen in class IV rules.

Langton's argument starts by defining a parameter λ, which represents the fraction of neighbourhoods in a CA's rule table that map to a non-quiescent state (i.e. a non-white colour). Then, by initialising one automaton with an empty rule table and progressively filling it with non-quiescent states, one can observe a transition point with exponentially long, particle-like transients (Fig. 5.2a). Here we repeat Langton's experiments using a 6-colour, range-2 CA, compute its average Φ, and show the results in Figure 5.2.

Interestingly, and in agreement with Langton's argument, we see a peak of integrated information for intermediate values of λ, coinciding with the automata's transition to a chaotic regime (Fig. 5.2b). It is rules in this critical region where computation is possible that have the highest Φ, showing *peak integrated information at the edge of chaos*.

Another unusual feature of Fig. 5.2b is that there is a region where complex, high-Φ automata coexist with simpler ones. This phenomenon was reported in Ref. [96] already: different automata will make a "transition" at different values of λ. This motivated Langton to analyse measures of complexity as a function of $\Delta\lambda$, the distance from the transition event for that particular automaton. As expected, when aligned by their critical λ value and plotted against $\Delta\lambda$ (Fig. 5.2c), all curves align onto a consistent picture of integrated information across the λ range.

For completeness, it is worth mentioning why at the right side of Figs. 5.2b and 5.2c Φ does not vanish for high λ (as one may expect, given that the single-cell autocorrelation does [96]). This is essentially due to the determinism and locality of the automaton's rule: given a spatially extended set of cells, it is always possible to predict the middle ones with perfect certainty. At the same time, cutting the system with a bipartition will reduce the spatial extent of this predictable region, so that the predictability of the whole is greater than the predictability of the parts, and thus $\Phi > 0$.

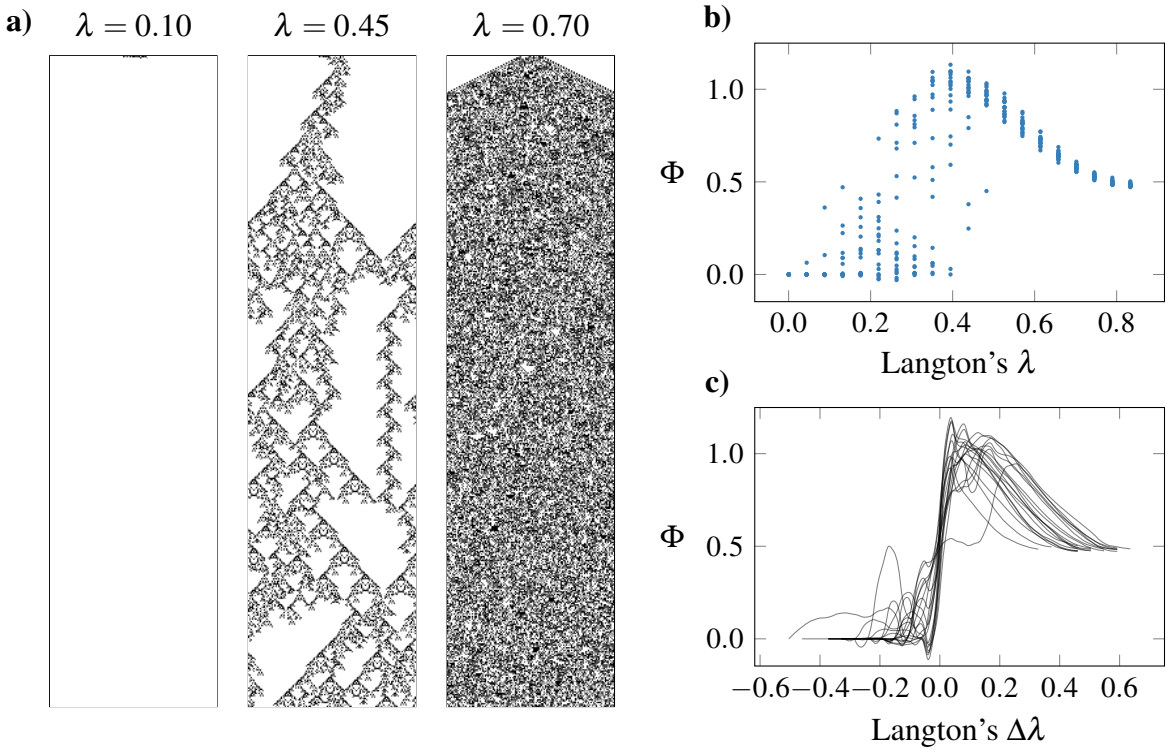


Figure 5.2: Integrated information peaks at the edge of chaos. **(a)** Sample runs from a random cellular automaton with different λ values, starting from a blank tape with 20 randomised cells in the middle. **(b)** Integrated information Φ peaks at an intermediate level of λ. **(c)** When plotted against $\Delta\lambda$, the distance from a transition event, all runs align on a similar Φ profile.

5.3.3 Information is integrated by coherent structures

In the experiments above we have shown that more complex automata integrate more information. However, this is not enough to make a case for Φ as a marker of distributed computation – it may just be the case that medium-λ CA have higher Φ due to general properties of their rule tables, or for some other reasons. In this section we address this possible counter-argument by showing that the increase of Φ is due to the emerging particles, and therefore can be directly associated with distributed computation.

To show this, we run large simulations of ECA rules 54 and 110, and evaluate several local information measures in small fragments of both (Fig. 5.3). Specifically, we compute Lizier's measures of storage (excess entropy, e_k) and transfer (transfer entropy, t_k), as well as our new local integrated information ϕ_k. Note that to measure transfer in either direction, we compute the local TE from a cell to its left neighbour and to its right neighbour and show the maximum of the two.

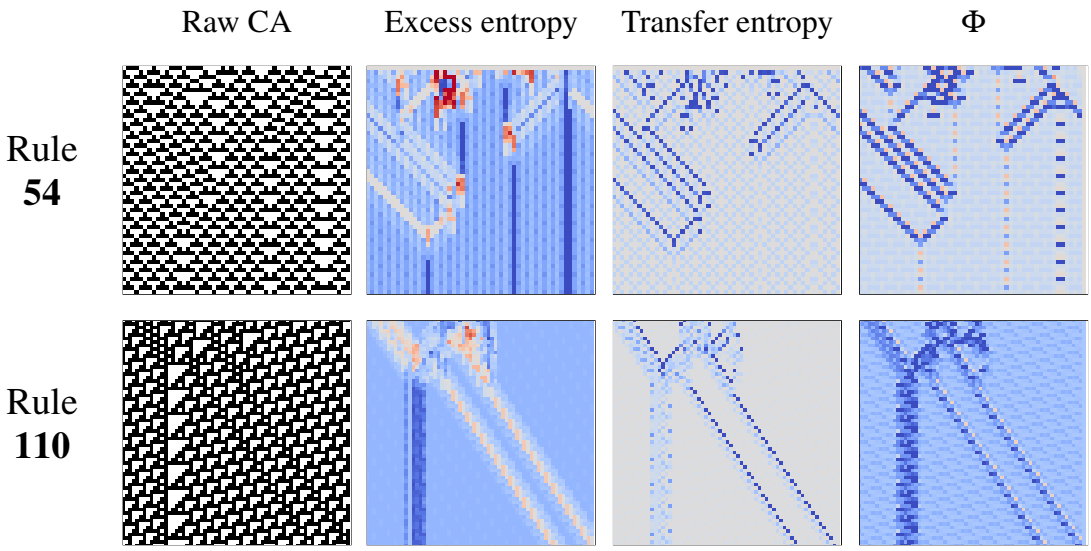


Figure 5.3: Local information measures in cellular automata, applied to two simulations of rule 54 (top) and 110 (bottom). Excess entropy is high for static particles, transfer entropy for moving particles, and integrated information Φ for both.

As expected, TE is high in gliders (moving particles), while excess entropy is high in blinkers (static particles). More interestingly for our purposes, is that Φ is high in *all of them* – gliders, blinkers, and the collisions between them.

When studied at a local scale in space and time, we see that information integration encompasses the three categories – storage, transfer, and modification – and that Φ can detect all of them *without having been explicitly designed to do so*. This reinforces our claim that Φ is a generic marker of emergent dynamics, and is connected with known measures of information processing. This connection will be made mathematically rigorous in Chapter 9 when we provide a decomposition of Φ and link it explicitly to TE and excess entropy.

5.4 Conclusion of Part I

In the last three chapters we have presented three case studies of integrated information in complex systems of different sorts: coupled oscillators, spiking neurons, and cellular automata. In all cases, Φ agreed with intuitive notions of complexity. This shows that Φ represents the broad notion of complexity, *both* in the Tononi-Sporns-Edelman sense of balanced integration and differentiation; *and* in the Wolfram sense of distributed computation.

It is worth noting that Φ is by no means the only quantity that peaks with the system's complexity – in cellular automata one could use the autocorrelation of a single cell, and in coupled oscillators the variance of Kuramoto's order parameter. The feature that makes Φ

unique is that *it is universally applicable across the board*, and yields the desired results in different kinds of systems without formulating idiosyncratic, ad-hoc measures.

This concludes our *pragmatic* argument that Φ can act as a unifying measure of complexity. It is important to emphasise that we have presented an purely empirical case, showing examples where Φ behaves as expected, but have given no theoretical explanation for why this should be the case. Such a theoretical explanation will come in Chapter 9, when we decompose Φ and identify some of its components as *causally emergent* dynamics.

In the next Part of this thesis, we switch gears back and return to considering IIT as a theory of consciousness. We will argue that, although IIT may be a suitable theory of complexity, there are limitations that (in its current form) stand in its way as a theory of consciousness.

Part II

Drawbacks and limitations of integrated information theory

Chapter 6

Measuring integrated information: Comparison of candidate measures in theory and simulation

Chapter summary

For it to be a fundamental theory of consciousness, IIT needs a robust axiomatic base linking phenomenology to one or more measurable quantities. We explore the properties of several proposed measures in simulation on simple, but non-trivial systems, and find a striking diversity in the behaviour of these measures – no two measures show consistent agreement across all analyses. We conclude that the axioms of IIT are underspecified, in the sense that multiple measures consistent with the axioms show qualitatively different behaviour in practice.

6.1 Introduction

Measures of integrated information of the sort we used in Chapters 3–5 seek to quantify the extent to which a whole system generates more information than the sum of its parts as it transitions between states. Since the concept of integrated information can be operationalised in many different ways, a whole range of distinct integrated information measures have come into being in the literature [8, 9, 11, 10], as we reviewed in Chapter 2. Several of them are beginning to see application to empirical data [106], or to large-scale simulations [107, 108], yet a systematic comparison of the behaviour of the various measures on non-trivial network models has not previously been performed.

All of these measures, however, have the potential to behave in ways which are not obvious a priori, and in a manner difficult to express analytically. While some simulations of some of the measures (Φ, $\tilde{\Phi}$, CD) on networks have been performed [8, 12], and some analytical understanding has been achieved for Φ and $\tilde{\Phi}$ [10, 41], other measures (ψ, Φ^*, Φ_G) have not previously been computed on any model consisting of more than two components. Since the ultimate goal of these measures is to provide a quantitative index of consciousness in biological brains and other physical systems, we need to understand what features of the target system drive the measures' behaviour. Furthermore, given the plethora of measures available, we need to understand their similarities and differences, and under what conditions they agree or disagree.

In this chapter we provide such a comparison of the full suite of measures on non-trivial network models, in order to shed light on their comparative practical utility. We consider eight-node networks with a range of different architectures, animated with basic noisy vector autoregressive dynamics. We examine how each measure is affected by network topology, coupling strength and noise correlation, as well as its relation with simpler dynamical controls like overall correlation. Based on these comparisons, we discuss the extent to which each measure captures the co-existence of integration and segregation central to the concept of dynamical complexity.

The main result of this exploration is a striking diversity in the behaviour of the measures – no two of them show consistent agreement across all analyses. This is particularly worrying, since all of them were proposed as representing the same notion of balanced integration and segregation. Therefore, the main takeaway of this exercise is that *all of these measures behave very differently, despite all being consistent with the IIT principles*. In other words, the axiomatic basis of IIT is not specific enough to meaningfully narrow down a measure of integrated information.

6.2 Methods

In this study, we consider the following six measures, previously introduced in Sec. 2.3:

- Whole-minus-sum integrated information, Φ,
- Integrated stochastic interaction, $\tilde{\Phi}$,
- Decoder-based integrated information, Φ^*,
- Geometric integrated information, Φ_G,
- Integrated synergy, ψ,
- Causal density, CD.

We use models based on stochastic linear auto-regressive (AR) processes with Gaussian variables. These constitute appropriate models for testing the measures of integrated information. They are straightforward to parameterise and simulate, and are amenable to the formulae presented in the previous section. Mathematically, we define an AR process (of order 1) by the update equation

$$X_{t+1} = AX_t + \varepsilon_t, \tag{6.1}$$

where ε_t is a serially independent random sample from a zero-mean Gaussian distribution with given covariance $\Sigma(\varepsilon)$, usually referred to as the *noise* or *error term*. A particular AR process is completely specified by the coupling matrix or *network A* and the noise covariance matrix $\Sigma(\varepsilon)$. An AR process is stable, and stationary, if the spectral radius of the coupling matrix is less than 1 [109]. (The spectral radius is the largest of the absolute values of its eigenvalues.) All the example systems we consider are calibrated to be stable, so the Φ measures can be computed from their stationary statistics.

We shall consider how the measures vary with respect to: (i) the strength of connections, i.e. the magnitude of non-zero terms in the coupling matrix; (ii) the topology of the network, i.e. the arrangement of the non-zero terms in the coupling matrix; (iii) the density of connections, i.e. the density of non-zero terms in the coupling matrix; and (iv) the correlation between noise inputs to different system components, i.e. the off diagonal terms in $\Sigma(\varepsilon)$. The strength and density of connections can be thought of as reflecting, in different ways, the level of integration in the network. The correlation between noise inputs reflects (inversely) the level of segregation, in some sense. We also, in each case, compute the control measures

- Time-delayed mutual information (TDMI), $I(X_{t-\tau};X_t)$; and
- Average absolute correlation $\bar{\Sigma}$, defined as the average absolute value of the non-diagonal entries in the system's correlation matrix.

These simple measures quantify straightforwardly the level of interdependence between elements of the system, across time and space, respectively. TDMI captures the total information generated as the system transitions from one time-step to the next, and $\bar{\Sigma}$ is another basic measure of the level of integration.

We report the unnormalised measures minimised over even-sized bipartitions—i.e. bipartitions in which both parts have the same number of components. In doing this, we avoid conflating the effects of the choice of definition of effective information with those of the choice of partition search (see discussion on MIP in Sec. 2.3.1). See the Discussion for more details on this topic.

6.2.1 Key quantities for computing integrated information measures

To compute the integrated information measures, the stationary covariance and lagged partial covariance matrices are required. By taking the expected value of $X_t^T X_t$ with Equation (6.1) and given that ε_t is white noise, uncorrelated in time, one obtains that the stationary covariance matrix $\Sigma(X)$ is given by the solution to the discrete-time Lyapunov equation,

$$\Sigma(X_t) = A\,\Sigma(X_t)\,A^{\mathrm{T}} + \Sigma(\varepsilon_t). \tag{6.2}$$

This can be easily solved numerically, for example in Matlab via use of the `dlyap` command. The lagged covariance can also be calculated from the parameters of the AR process as

$$\Sigma(X_{t-1}, X_t) = \langle X_t (A X_t + \varepsilon_t)^{\mathrm{T}} \rangle = \Sigma(X_t) A^{\mathrm{T}}, \tag{6.3}$$

and partial covariances can be obtained by applying Eq. (2.7). Finally, we obtain the analogous quantities for the partitions by the marginalisation properties of the Gaussian distribution. Given a bipartition $X_t = \{M_t, N_t\}$, we write the covariance and lagged covariance matrices as

$$\begin{aligned} \Sigma(X_t) &= \begin{pmatrix} \Sigma(X_t)_{mm} & \Sigma(X_t)_{mn} \\ \Sigma(X_t)_{nm} & \Sigma(X_t)_{nn} \end{pmatrix}, \\ \Sigma(X_{t-1}, X_t) &= \begin{pmatrix} \Sigma(X_{t-1}, X_t)_{mm} & \Sigma(X_{t-1}, X_t)_{mn} \\ \Sigma(X_{t-1}, X_t)_{nm} & \Sigma(X_{t-1}, X_t)_{nn} \end{pmatrix}, \end{aligned} \tag{6.4}$$

and we simply read the partition covariance matrices as

$$\begin{aligned} \Sigma(M_t) &= \Sigma(X_t)_{mm}\ , \\ \Sigma(M_{t-1}, M_t) &= \Sigma(X_{t-1}, X_t)_{mm}\ . \end{aligned} \tag{6.5}$$

6.3 Results

6.3.1 Two-node network

We begin with the simplest non-trivial AR process,

$$A = \begin{pmatrix} a & a \\ a & a \end{pmatrix}, \tag{6.6a}$$

$$\Sigma(\varepsilon) = \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix}. \tag{6.6b}$$

Setting $a = 0.4$, we obtain the same model as depicted in Figure 3 in Reference [10]. We simulate the AR process with different levels of noise correlation c and show results for all the measures in Figure 6.1. Note that, as c approaches 1, the system becomes degenerate, so some matrix determinants in the formulae become zero causing some measures to diverge.

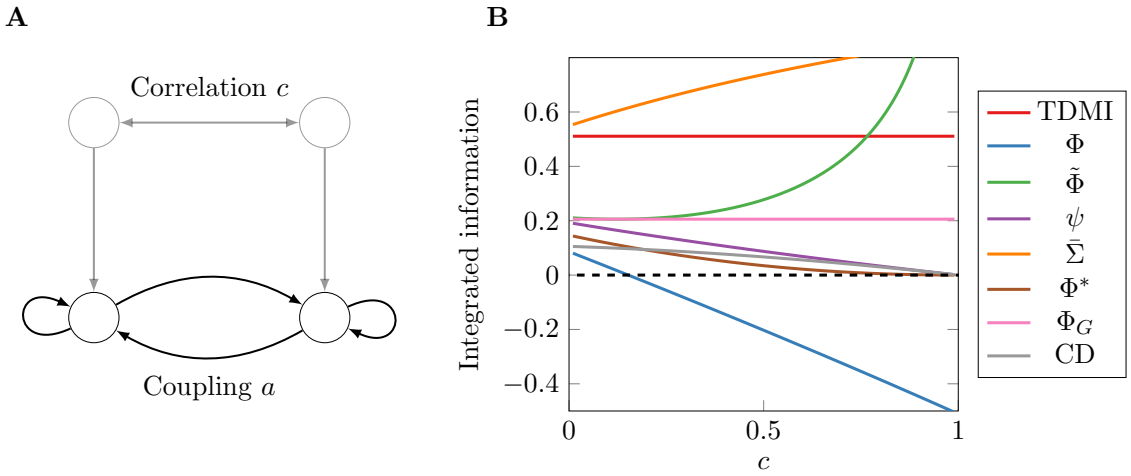


Figure 6.1: (**A**) graphical representation of the two-node AR process described in Eq. (6.6). Two connected nodes with coupling strength a receive noise with correlation c, which can be thought of as coming from a common source; (**B**) all integrated information measures for different noise correlation levels c.

Inspection of Figure 6.1 immediately reveals a wide variability of behaviour among the measures, in both value and trend, even for this minimally simple model. A good candidate measure of (dynamical) integrated information should tend to 0 as the noise tends to becoming perfectly correlated ($c \to 1$) because, in that instance, the whole just becomes a collection of copies of the parts (we don't consider $c = 1$ because the Gaussian model becomes singular in this limit). Only the measures ψ, Φ^*, and CD achieve this. Φ_G is

here unaffected by noise correlation,[1] and $\tilde{\Phi}$ grows monotonically with c. Furthermore, $\tilde{\Phi}$ diverges to infinity as $c \to 1$. On the other hand, Φ also decreases monotonically but becomes negative for large enough c.

In Figure 6.2, we analyse the same system, but now varying both noise correlation c and coupling strength a. As per the stability condition presented above, any value of $a \geq 0.5$ makes the system's spectral radius greater than or equal to 1, so the system becomes non-stationary and variances diverge. Hence, in these plots, we evaluate all measures for values of a below the limit $a = 0.5$.

Again, the measures behave very differently. In this case, TDMI and Φ_G remain unaffected by noise correlation, and grow with increasing coupling strength as expected. In contrast, $\tilde{\Phi}$ and $\bar{\Sigma}$ increase with both a and c. Φ decreases with c but shows non-monotonic behaviour with a. Three of the measures, ψ, Φ^*, and CD, show properties consistent with capturing conjoined segregation and integration—they monotonically decrease with noise correlation and increase with coupling strength.

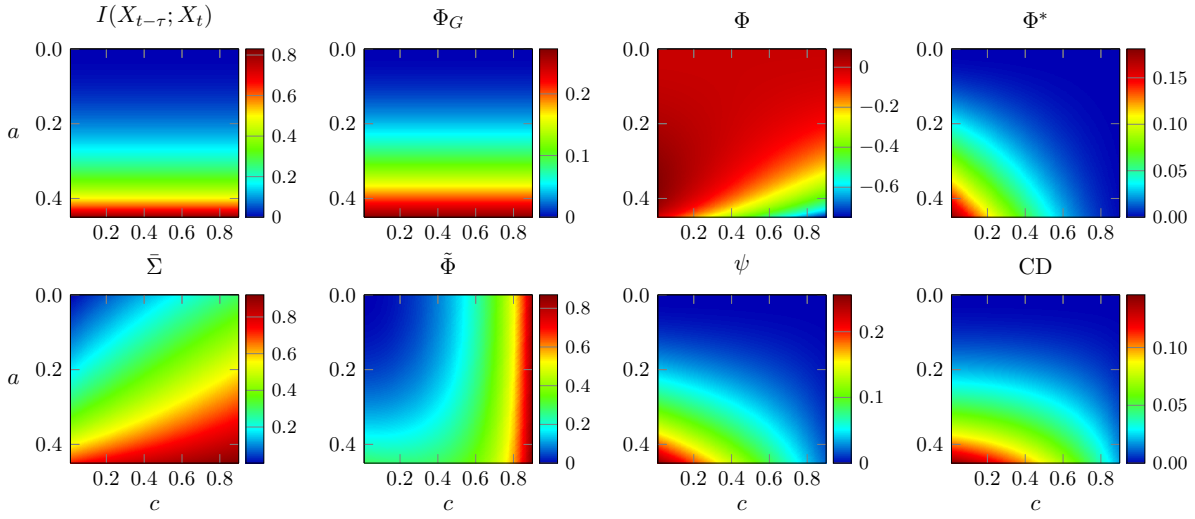


Figure 6.2: All integrated information measures for the two-node AR process described in Equation (6.6), for different coupling strengths a and noise correlation levels c. The vertical axis is inverted for visualisation purposes.

6.3.2 Eight-node networks

We now turn to networks with eight nodes, enabling examination of a richer space of dynamics and topologies.

[1] According to an anonymous reviewer, Φ_G does decrease with noise correlation in discrete systems, although in this article we focus exclusively in Gaussian systems.

We first analyse a weighted network optimised using a genetic algorithm to yield high Φ (Figure 2b in [8]). The noise covariance matrix has ones in the diagonal and c everywhere else, and now a is a global factor applied to all edges of the network. The (weighted) adjacency matrix is scaled such that its spectral radius is 1 when $a = 1$. Similar to the previous section, we evaluate all measures for multiple values of a and c and show the results in Figure 6.3.

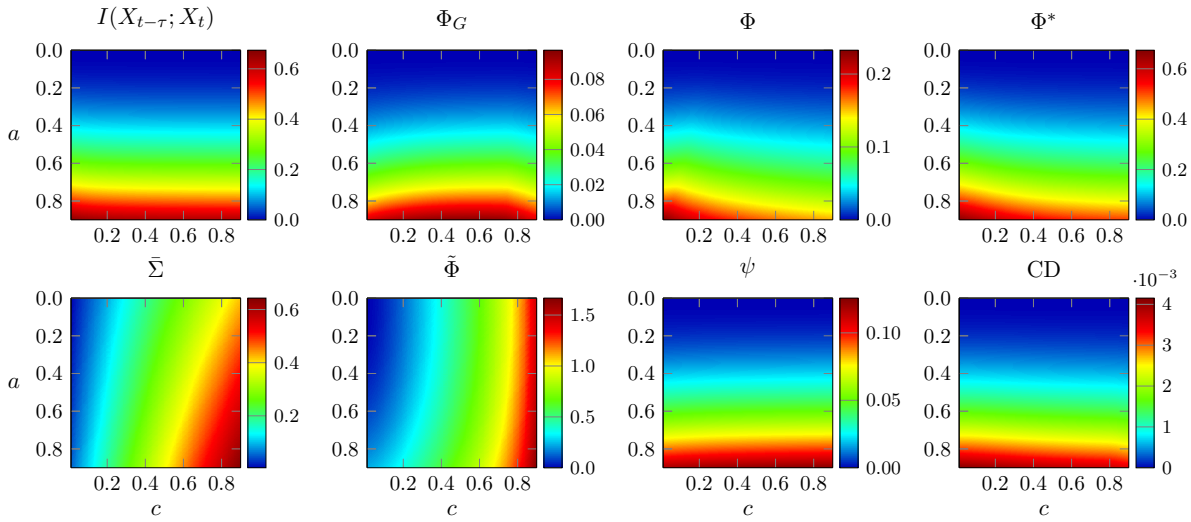


Figure 6.3: All integrated information measures for the Φ-optimal AR process proposed by [8], for different coupling strengths a and noise correlation levels c. Vertical axis is inverted for visualisation purposes.

Moving to a larger network mostly preserves the features highlighted above. TDMI is unaffected by c; $\tilde{\Phi}$ behaves like $\bar{\Sigma}$ and diverges for large c; and Φ^* and CD have the same trend as before, although now the decrease with c is less pronounced. Interestingly, ψ and Φ_G increase slightly with c, and Φ does not show the instability and negative values seen in Figure 6.2. Overall, in this more complex network, the effect of increasing noise correlation on Φ, ψ, Φ^*, and CD is not as pronounced as in simpler networks, where these measures decrease rapidly towards zero with increasing c.

Thus far, we have studied the effect of AR dynamics on integrated information measures, keeping the topology of the network fixed and changing only global parameters. We next examine the effect of network topology, on a set of six networks:

A A fully connected network without self-loops.

B The Φ-optimal binary network presented in [8].

C The Φ-optimal weighted network presented in [8].

D A bidirectional ring network.

E A "small-world" network, formed by introducing two long-range connections to a bidirectional ring network.

F A unidirectional ring network.

In each network, the adjacency matrix has been normalised to a spectral radius of 0.9. As before, we simulate the system following Equation (6.1), and here set noise input correlations to zero ($c = 0$) so the noise input covariance matrix is just the identity matrix. Figure 6.4 shows connectivity diagrams of the networks for visual comparison, and Figure 6.5 shows the values of all integrated information measures evaluated on all networks.

As before, there is substantial variability in the behaviour of all measures, but some general patterns are apparent. Intriguingly, the unidirectional ring network consistently scores highest for all measures (except for $\tilde{\Phi}$ and CD), followed in most cases by the weighted Φ-optimal network.[2] On the other end of the spectrum, the fully connected network **A** consistently scores lowest, which is explained by the large correlation between its nodes as shown by $\bar{\Sigma}$.

The results here can be summarised by comparing the rank assigned to the networks by each measure (see Table 6.1). Inspecting this table reveals a remarkable alignment between TDMI, Φ_G, Φ^*, and ψ, especially given how much their behaviour diverges when varying a and c. Although the particular values are different, the measures largely agree on the ranking of the networks based on their integrated information. This consistency of ranking is initially encouraging with regard to empirical application.

Table 6.1: Networks ranked according to their value of each integrated information measure (highest value to the left). We add small-world index as a dynamics-agnostic measure of network complexity.

Measure	**Ranking**					
$I(X_t, X_{t+\tau})$	F	C	D	E	B	A
Φ_G	F	C	D	E	B	A
Φ	F	C	B	E	D	A
Φ^*	F	C	B	E	D	A
$\bar{\Sigma}$	C	B	A	E	D	F
$\tilde{\Phi}$	C	F	B	D	E	A
ψ	F	C	D	E	B	A
CD	C	F	B	D	E	A
SWI	C	E	B	A	D	F

[2] Note that in Figure 6.5 the Φ-optimal networks **B** and **C** score much less than simpler network **F**. This is because all networks have been scaled to a spectral radius of 0.9 – when the networks are normalised to a spectral radius of 0.5, as in the original paper, then **B** and **C** are, as expected, the networks with highest Φ.

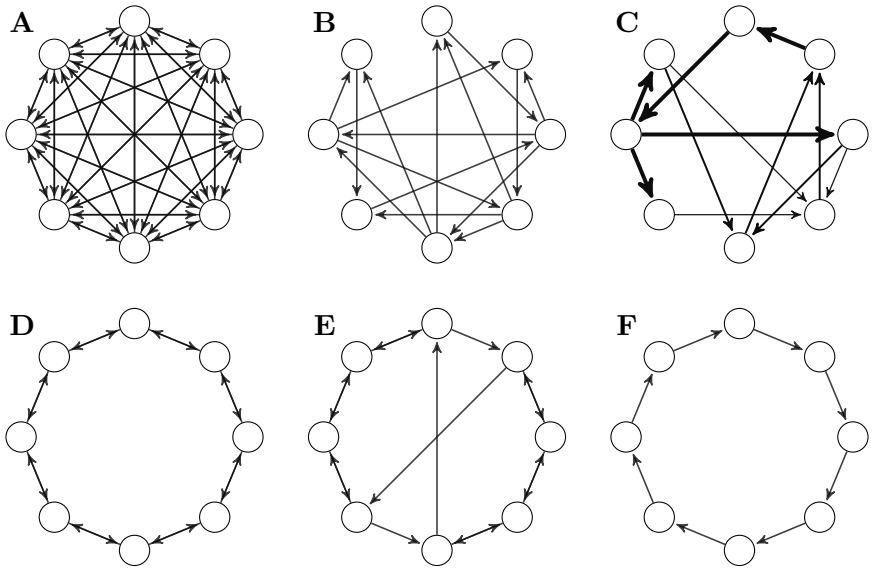


Figure 6.4: Networks used in the comparative analysis of integrated information measures. (**A**) fully connected network; (**B**) Φ-optimal binary network from [8]; (**C**) Φ-optimal weighted network from [8]; (**D**) bidirectional ring network; (**E**) small world network; and (**F**) is a unidirectional ring network.

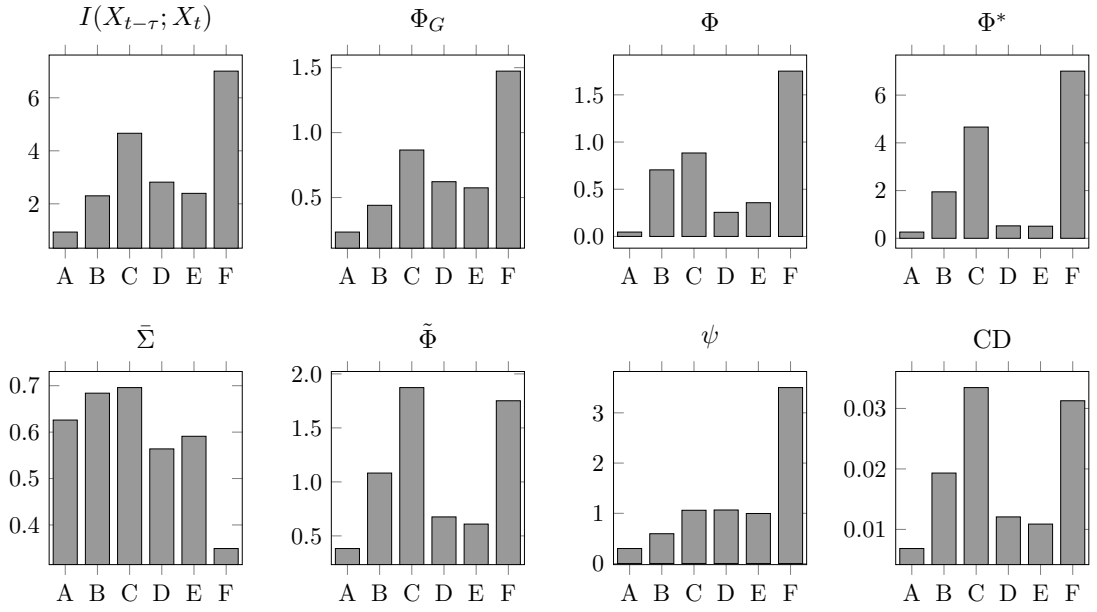


Figure 6.5: Integrated information measures for all networks in the suite shown in Figure 6.4, normalised to spectral radius 0.9 and under the influence of uncorrelated noise. The ring and weighted Φ-optimal networks score consistently at the top, while denser networks like the fully connected and the binary Φ-optimal networks are usually at the bottom. Most measures disagree on specific values but agree on the relative ranking of the networks.

However, the ranking is not what might be expected from topological complexity measures from network theory. If we ranked these networks by e.g. small-world index (SWI) [110, 111],[3] we expect networks **B**, **C**, and **E** to be at the top and networks **A**, **D**, and **F** to be at the bottom—very different from any of the rankings in Table 6.1. In fact, the Spearman correlation between the ranking by small-world index and those by TDMI, Φ_G, Φ^*, and ψ is around -0.4, leading to the somewhat paradoxical conclusion that more structurally complex networks integrate *less* information. We note that these rankings are very robust to noise correlation (results not shown) for all measures except Φ. Across all simulations in this study, the behaviour of Φ is erratic, undermining prospects for empirical application. (This behaviour is even more prevalent if Φ is optimised over all bipartitions, as opposed to over even bipartitions.)

6.3.3 Random networks

We next perform a more general analysis of the performance of measures of integrated information, using Erdős–Rényi random networks. We consider Erdős–Rényi random networks parametrised by two numbers: the edge density of the network ρ and the noise correlation c (defined as above), both in the $[0,1)$ interval. To sample a network with a given ρ, we generate a matrix in which each possible edge is present with probability ρ and then remove self-loops. The stochasticity in the construction of the Erdős–Rényi network induces fluctuations on the integrated information measures, such that, for each (ρ, c), we calculate the mean and variance of each measure.

First, we generate 50 networks for each point in the (ρ, c) plane and take the mean of each integrated information measure evaluated on those 50 networks. As before, the adjacency matrices are normalised to a spectral radius of 0.9. Results are shown in Figure 6.6.

Φ_G increases markedly with ρ and moderately with c, $\bar{\Sigma}$ increases sharply with both and the rest of the measures can be divided in two groups, with Φ, ψ and CD that decrease with c and TDMI, $\tilde{\Phi}$ and Φ^* that increase. Notably, all integrated information measures except Φ_G show a band of high value at an intermediate value of ρ. This demonstrates their sensitivity to the level of integration. The decrease when ρ is increased beyond a certain point is due to the weakening of the individual connections in that case (due to the fixed overall coupling strength, as quantified by spectral radius).

Secondly, in Figure 6.7, we plot each measure against the average correlation of each network, following the rationale that dynamical complexity should (as a necessary, but not sufficient condition) peak at an intermediate value of $\bar{\Sigma}$—i.e. it should reach its maximum

[3] The small-world index of a network is defined as the ratio between its clustering coefficient and its mean minimum path length, normalised by the expected value of these measures on a random network of the same density. Since the networks we consider are small and sparse, we use the 4th order cliques (instead of triangles, which are 3rd order cliques) to calculate the clustering coefficient.

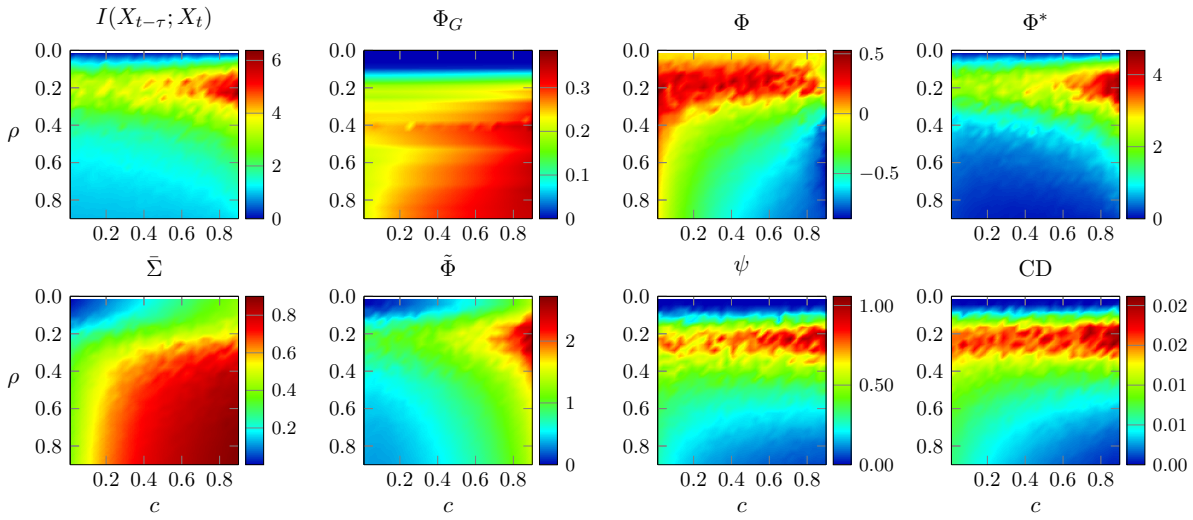


Figure 6.6: Average integrated information measures for Erdős–Rényi random networks with given density ρ and noise correlation c. The vertical axis is inverted for consistency with Figures 6.2 and 6.3.

value in the middle range of $\bar{\Sigma}$. To obtain this figure, we sampled a large number of Erdős–Rényi networks with random (ρ, c), and evaluated all integrated information measures, as well as their average correlation $\bar{\Sigma}$.

Figure 6.7 shows that some of the measures have this intermediate peak, in particular: Φ^*, ψ, Φ_G, and CD. Although also showing a modest intermediate peak, $\tilde{\Phi}$ has a stronger overall positive trend with $\bar{\Sigma}$, and Φ an overall negative trend. These analyses further support the notion that Φ^*, ψ, Φ_G, and CD reflect some form of dynamical complexity, although the relation between them remains unclear and not always consistent in other scenarios.

One might worry that these peaks could be due to a biased sampling of the $\bar{\Sigma}$ axis—if our sampling scheme were obtaining many more samples in, say, the $0.2 < \bar{\Sigma} < 0.4$ range, then the points with high Φ we see in that range could be explained by the fact that the high-Φ tails of the distribution are sampled better in that range than in the rest of the $\bar{\Sigma}$ axis. However, the histogram at the bottom of Figure 6.7 shows this is not the case—on the contrary, the samples are relatively uniformly spread along the axis. Therefore, the peaks shown by Φ^*, ψ, Φ_G, and CD are not sampling artefacts.

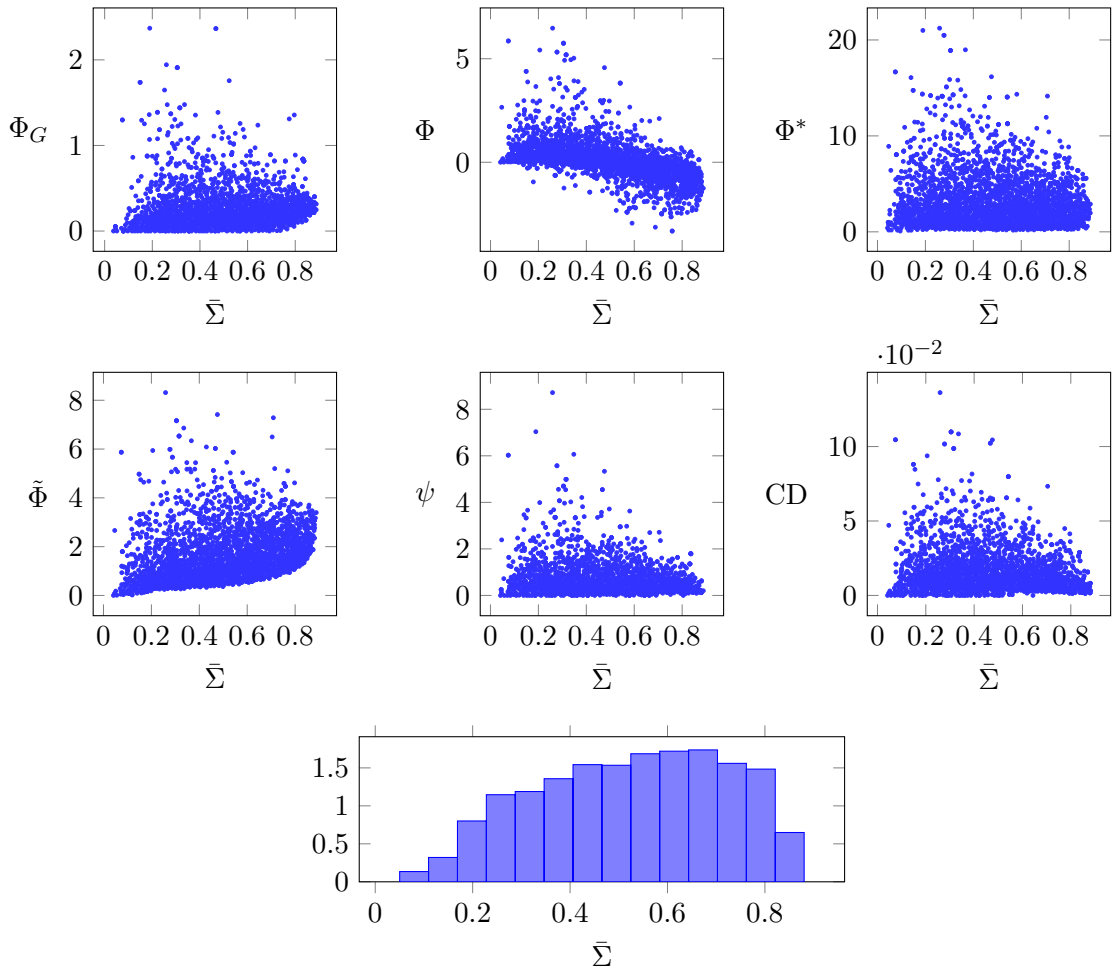


Figure 6.7: Integrated information measures of random Erdős–Rényi networks, plotted against the average correlation $\bar{\Sigma}$ of the same network; (bottom) normalised histogram of $\bar{\Sigma}$ for all sampled networks.

6.4 Discussion

In this study, we compared several candidate measures of integrated information in terms of their theoretical construction, and their behaviour when applied to the dynamics generated by a range of non-trivial network architectures. We found that no two measures had precisely the same basic mathematical properties (see Table 2.2). Empirically, we found a striking variability in the behaviour among the measures even for simple systems; see Table 6.2 for a summary. Three of the measures, ψ, Φ^* and CD, capture conjoined segregation and integration on small networks, when animated with Gaussian linear AR dynamics (Figure 6.1). These measures decrease with increasing noise input correlation and increase with increasing coupling strength (Figure 6.3). Furthermore, on random networks with fixed overall coupling strength (as quantified by spectral radius), they achieve their highest scores

Table 6.2: Integrated information measures considered and brief summary of our results.

Measure	Summary of Results
Φ	Erratic behaviour, negative when nodes are strongly correlated.
$\tilde{\Phi}$	Mostly reflects noise input correlation, not sensitive to changes in coupling.
ψ	Reflects both segregation and integration.
Φ^*	Reflects both segregation and integration.
Φ_G	Mostly reflects changes in coupling, not sensitive to noise input correlation.
CD	Reflects both segregation and integration.

when an intermediate number of connections are present (Figure 6.6). They also obtain their highest scores when the average correlation across components takes an intermediate value (Figure 6.7).

In terms of network topology, none of the measures strongly reflect complexity of the network structure in a graph theoretic sense. At fixed overall coupling strength, a simple ring structure (Figure 6.4) leads in most cases to the highest scores. Among the other measures: $\tilde{\Phi}$ is largely determined by the level of correlation amongst the noise inputs, and is not very sensitive to changes in coupling strength; Φ_G depends mainly on the overall coupling strength, and is not very sensitive to changes in noise input correlation; and Φ generally behaves erratically.

Considered together, our results motivate the continued development of ψ, Φ^* and CD as theoretically sound and empirically adequate measures of dynamical complexity.

6.4.1 Partition selection

Integrated information is typically defined as the effective information beyond the minimum information partition [5, 4]. However, when a particular measure of integrated information has been first introduced, it is often with a new operationalisation of both effective information and the minimum information partition. In this paper, we have restricted attention to comparing different choices of measure of effective information, while keeping the same partition selection scheme across all measures. Specifically, we restricted the partition search to even-sized bipartitions, which has the advantage of obviating the need for introducing a normalisation factor when comparing bipartitions with different sizes. For uneven partitions, normalisation factors are required to compensate for the fact that there is less capacity for information sharing as compared to even partitions. However, such factors are known to introduce instabilities, both under continuous parameter changes, and in terms of numerical errors [8]. Further research is needed to compare different approaches to defining the minimum information partition, or finding an approximation to it in reasonable computation time [112].

In terms of computation time, performing the most thorough search, through all partitions, as in the early formulation of Φ by Balduzzi and Tononi [5] requires time $\mathcal{O}(n^n)$. Restricting attention to bipartitions reduces this to $\mathcal{O}(2^n)$, whilst restricting to even bipartitions reduces this further to $\mathcal{O}(n^2)$. These observations highlight a trade-off between computation time and comprehensive consideration of possible partitions. Future comparisons of integrated information measures may benefit from more advanced methods for searching among a restricted set of partitions to obtain a good approximation to the minimum information partition. For example, Toker and Sommer use graph modularity, stochastic block models or spectral clustering as informed heuristics to suggest a small number of partitions likely to be close to the MIP, and then take the minimum over those. With these approximations, they are able to calculate the MIP of networks with hundreds of nodes [112, 106]. Alternatively, Hidaka and Oizumi make use of the submodularity of mutual information to perform efficient optimisation and find the bipartition across which there is the least instantaneous mutual information of the system [113]. Presently, however, their method is valid only for instantaneous mutual information and is therefore not applicable to finding the bipartition that minimises any form of normalised effective information as described above in the section dedicated to the MIP.

Furthermore, each measure carries special considerations regarding partition search. For example, for ψ, taking the minimum across all partitions is equivalent to taking it across bipartitions only, thanks to the properties of $I_\cap$ [19, 20, 23]. Arsiwalla and Verschure [114] used $\tilde{\Phi}$ and suggested always using the atomic partition on the basis that it is fast, well-defined, and, for $\tilde{\Phi}$ specifically, it can be proven to be the partition of *maximum* information; and thus it provides a quickly computable upper bound for the measure.

6.4.2 Continuous variables and the linear Gaussian assumption

We have compared the various integrated information measures only on systems whose states are given by continuous variables with a Gaussian distribution. This is motivated by measurement variables being best characterised as continuous in many domains of potential application. Future research should continue the comparison of these measures on a test-bed of systems with discrete variables. Moreover, non-Gaussian continuous systems should also be considered because the Gaussian approximation is not always a good fit to real data. For example, the spiking activity of populations of neurons typically exhibit exponentially distributed dynamics [59]. Systems with discrete variables are in principle straightforward to deal with, since calculating probabilities (following the most brute-force approach) amounts simply to counting occurrences of states. General continuous systems, however, are less straightforward. Estimating generic probability densities in a continuous domain is challenging, and calculating information-theoretic quantities on these is difficult [15, 115]. The AR systems we have studied here are a rare exception, in the sense that their

probability density can be calculated and all relevant information-theoretic quantities have an analytical expression. Nevertheless, the Gaussian assumption is common in biology, and knowing now how these measures behave on these Gaussian systems will inform further development of these measures, and motivate their application more broadly.

6.4.3 Empirical as opposed to maximum entropy distribution

We have considered versions of each measure that quantify information with respect to the empirical, or spontaneous, stationary distribution for the state of the system. This constitutes a significant divergence from the supposedly fundamental measures of intrinsic integrated information of IIT versions 2 and 3 [5, 6]. Those measures are based on information gained about a hypothetical past moment in which the system was equally likely to be in any one of its possible states (the "maximum entropy" distribution). However, as pointed out previously [8], it is not possible to extend those measures, developed for discrete Markovian systems, to continuous systems. This is because there is no uniquely defined maximum entropy distribution for a continuous random variable (unless it has hard-bounds, i.e. a closed and bounded set of states). Hence, quantification of information with respect to the empirical distribution is the pragmatic choice for construction of an integrated information measure applicable to continuous time-series data.

The consideration of information with respect to the empirical, as opposed to maximum entropy, distribution does, however, have an effect on the concept underlying the measure of integrated information—it results in a measure not of mechanism, but of dynamics [116]. That is, what is measured is not information about what the possible mechanistic causes of the current state *could be*, but rather what the likely preceding states *actually are*, on average, statistically; see [8] for further discussion. Given the diversity of behaviour of the various integrated information measures considered here even on small networks with linear dynamics, one must remain cautious about considering them as generalisations or approximations of the proposed "fundamental" Φ measures of IIT versions 2 or 3 [5, 6].

A remaining important challenge, in many practical scenarios, is the identification of stationary epochs. For a relatively long data segment, it can be unrealistic to assume that all the statistics are constant throughout. For shorter data segments, one can not be confident that the system has explored all the states that it potentially would have, given enough time.

6.5 Final remarks

The further development, and empirical application of Integrated Information Theory requires a good understanding of the various potential operational measures of information integration. During the last few years, several measures have been proposed, but their be-

haviour in any but the simplest cases has not been extensively characterised or compared. In this study, we have reviewed several candidate measures of (dynamical/empirical) integrated information, and provided a comparative analysis on simulated data, generated by simple Gaussian dynamics applied to a range of network topologies.

Assessing the degree of dynamical complexity, integrated information, or co-existing integration and segregation exhibited by a system remains an important outstanding challenge. Progress in meeting this challenge will have implications not only for theories of consciousness, such as Integrated Information Theory, but more generally in situations where relations between local and global dynamics are of interest. The review presented here identifies promising theoretical approaches for designing adequate measures of integrated information. Furthermore, our simulations demonstrate the need for empirical investigation of such measures, since measures that share similar theoretical properties can behave in substantially different ways, even on simple systems.

Chapter 7

Empirical evidence for and against IIT

Chapter summary

Regardless of its mathematical underpinnings, a minimum requirement for any theory of consciousness is to make successful predictions on adult human brains. We review existing experimental evidence for and against IIT as a theory of consciousness, and present new comprehensive analyses on several datasets. The evidence is mixed, and in some cases Φ, counterintuitively, is drastically *increased* in the unconscious state and *reduced* in the psychedelic state. We discuss possible causes of this discrepancy and discuss the relevance of these results to IIT's current and future status.

7.1 Introduction

So far, our discussion of IIT as a theory of consciousness (and of its limitations as such) has remained in abstract terms: the level of specification of its axioms, the robustness of its definition, and other mathematical properties. And yet, we haven't thus far addressed the obvious question: is Φ in unconscious brains *actually lower* than in conscious ones? The goal of this chapter is to attempt to provide an answer to that question.

In the spirit of Part I and our stated objectives (Sec. 1.2), throughout this chapter we will be careful to dissociate the information-theoretic tools employed from any statements about consciousness. In particular, we will distinguish between **IIT**, the collection of information-theoretic tools and insights discussed earlier in this thesis; and **IITC**, the set of claims concerning IIT *as a theory of consciousness*.

The chapter is structured as follows. We first discuss existing experimental evidence for IITC, and argue that results have been overinterpreted and important details ignored. Next, we present a comprehensive direct evaluation of Φ measures on altered states of consciousness, and reach the unexpected result that Φ measures often show the *opposite* behaviour to that predicted by IITC. Finally, we propose a small toy model that may help us interpret these unexpected results, and suggests new avenues for methodological research in consciousness neuroscience. Taken together, these pieces of evidence draw a complex picture with no clear-cut conclusion at this stage. Empirical evidence does not seem to match the theory's predictions, although our statistical tools are poorly suited to fairly test those predictions. We end this Part with a discussion of IIT's current and future status, including past hurdles and upcoming opportunities.

7.2 Existing experimental evidence for IITC

In this section we review the published empirical evidence for IITC, with the goal of summarising the experimental findings, as well as putting them in context with IITC's predictions and with each other. The key results of the papers discussed here are presented in Table 7.1. Note that this is by no means an exhaustive list of experimental applications of IIT. We focus only on studies with direct relevance to the study of consciousness, and exclude others – like the application of Φ measures to simulations of the *C. elegans* brain [106], to neurological disease [117], or to other simulated systems [118] – that are not clearly linked to IITC's key claims about consciousness in healthy human brains.

First, let us comment on the early predictions of IITC put forward in some of the original theoretical papers (e.g. Ref. [5]). These predictions include statements like the cerebellum not playing a role in consciousness due to its feed-forward architecture; or Φ being low in neurally inactive and hyperactive states like coma or seizures.

Table 7.1: Summary of existing experimental evidence for IITC. Note that these are very succinct summaries that leave aside many important details of all the listed papers. See the original references for a more nuanced picture of the available evidence.

Contribution / experiment set-up	Key result	Caveats	First author & reference
Theoretical formulation of Φ	Consciousness is not in the cerebellum, fades in inactive or hyperactive states	Claims are non-specific, most theories of consciousness make similar predictions	Balduzzi [5]
Φ-like measure on EEG data of patients under propofol anaesthesia	Strong reductions in Φ in the γ band	Broadband Φ still higher under anaesthesia, version of Φ used was modified for unjustified reasons	Lee [120]
$\tilde{\Phi}$ evaluated on intracranial electrode data	$\tilde{\Phi}$ is lower in NREM slow-wave sleep	Only three subjects, experiment used direct current stimulation	Chang [121]
Φ^* computed on fMRI of subjects watching a movie, scrambled movie, and noise	Φ^* is lower in scrambled movie and pixel noise conditions	Unclear connection between "stimulus set meaningfulness" and level of consciousness	Boly [122]
Φ^* computed on human ECoG data watching supra- and subliminal stimuli	Classifier trained on Φ^* performs better than if trained on H, MI	Does not report Φ^* being actually higher for conscious percepts	Haun [123]
Φ^* evaluated on monkey resting-state ECoG data	Φ^* peaks at intermediate timescale	No mention of higher Φ^* in same dataset with anaesthetised monkey	Oizumi [10]
Φ^* computed on human EEG	Φ^* and EEG connectivity discriminate between stages of anaesthesia	Φ^* lower in α but higher in all other bands, inconsistent with Ref [120]	Kim [124]

These predictions, while correct and reasonably justified, are not enough to garner support for IITC, because virtually every theory of consciousness has something to say about those basic phenomena.[1] In that sense, while the predictions are accurate, they are not *specific* – they do not set IITC apart from many other proposals in a meaningful way.

With respect to the actual empirical results, two overarching patterns emerge:

1. Results are overall inconsistent; and
2. IIT methodology is only loosely applied.

Let us elaborate first on two examples showing inconsistent pieces of evidence. Consider the results in Ref. [123] showing that Φ^* is better than alternatives at classifying whether a visual stimulus was consciously or unconsciously perceived. This is an important result, that sets IITC apart from other theories by making statements about the structure of conscious percepts. The limitation of Ref. [123], however, is they do not show Φ^* to be actually *higher* for consciously perceived stimuli, as opposed to just better as a classifier feature.

In principle, one could argue this is still valid – subjects are in wakeful rest, so their "overall" Φ^* may remain approximately unchanged. *But*, when considered in conjunction with the earlier result that Φ^* is higher for more "meaningful" stimuli [122], then we would expect Ref. [123] to report higher Φ^* when a face (instead of a random pixel mask) is consciously perceived – which is not reported.

A similar, more obvious case of inconsistency is in the results of Refs. [120, 124]. Comparing subjects with and without propofol anaesthesia, Ref. [120] reports decreased Φ under anaesthesia in the gamma band, while Ref. [124] reports lower Φ^* in the alpha band, but *higher* Φ^* in all other bands. It should be noted that these two studies use different measures of integrated information, which might explain the discrepancy (but it highlights again the problem of underspecification discussed in Chapter 6).

The second, and perhaps more insidious problem, is that IIT methodology is loosely applied in the experimental work mentioned, in such a way that the hypotheses being tested are only surrogates for the central IITC claims. One example of this loose interpretation is Casali *et al.*'s PCI experiment [125]. This case plays such a prominent role in the defence of IITC that it deserves extended discussion in the subsection below.

7.2.1 PCI and causal interventions

In their landmark 2013 paper, Casali and colleagues made a remarkable contribution to applied consciousness science by introducing the Perturbational Complexity Index (PCI) [125].

[1] For example, a supporter of higher-order thought [119] or global workspace [93] theories will arrive at the same conclusions through equally simple arguments.

The procedure to estimate PCI consists of applying a TMS pulse to a subject's brain, measuring its response through EEG, and analysing the response via Lempel-Ziv complexity (c.f. Chapter 10). Casali's central result is that PCI correlates very well with clinically-defined levels of consciousness, being higher for healthy awake subject, and lower for individuals sleeping or suffering from disorders of consciousness. Empirically, PCI is extremely successful, and it remains the undisputed state-of-the-art of neurophysiological markers of consciousness in clinical settings.

Proponents of IITC often mention PCI as a successful prediction of IITC [6, 126, 127], typically presented as simply stating the obvious. Common points in the argument include that "PCI can gauge the intrinsic cause-effect power of the cortex," and that it "is high only if brain responses are both integrated and differentiated, corresponding to a distributed spatio-temporal pattern of causal interactions that is complex and hence not very compressible" [126, p. 459].

Despite the excitement, and despite its unquestionable empirical success, I argue that *PCI does not constitute positive evidence to support the case for IITC*, for multiple reasons:

1. In PCI, the perturbation used is of a very different kind compared to that in IIT 3.0. In fact, one of the remarkable features of PCI is that the locus and intensity of the TMS pulse do not matter – which is radically unlike the maximum-entropy perturbation employed in IIT 3.0, a long procedure that involves carefully setting the system in each one of its possible states and analysing the cause and effect information of each one of its subsets.

2. Perturbational analyses tend to arrive at qualitatively similar result as their resting-state analogues, albeit with a greater signal-to-noise ratio.[2] Therefore, insofar as PCI uses LZ, it is strongly correlated with LZ, and LZ is purely a measure of entropy (see Chapter 10), the case for PCI as a measure of integrated information becomes much weaker.

3. While having a high PCI does require the brain to be in a state with balanced integration and differentiation, as we have extensively argued in Chapter 6 this restriction is not specific enough. All the measures reviewed in Chapter 6 require this integration-differentiation balance, but they do not enjoy the empirical success of PCI.

On the basis of these results, it is inappropriate to hold the PCI results as evidence for IITC; and *equating PCI with Φ only perpetuates the trend of underspecificity and ambiguity in IIT research*.

[2]Empirically, this really seems to be the case: Casali's PCI paper gets to essentially the same result as Schartner *et al.*'s studies using spontaneous LZ [47, 128, 129]; and Deco and Tagliazucchi's results on intrinsic ignition [130] also point towards essentially the same outcome as their earlier work [131] suggesting that widespread correlation is stronger in wake than in N3 sleep.

Far from being a setback, we see this as an opportunity for IIT research. Investigating the source of the increase in signal-to-noise ratio due to the TMS is a promising avenue for research in computational neuroscience of consciousness, which could inform both future theory and experiment.

7.3 Novel experimental evidence against IITC

7.3.1 Datasets and methodology

For our own analysis, we pool together six different datasets related to markedly different states of consciousness. These datasets span across multiple recording modalities (M/EEG, ECoG), states of consciousness (sleep, anaesthesia, psychedelic state), and research centres. Relevant details of each dataset can be found in Table 7.2.[3]

Table 7.2: Datasets used in experimental validation of IITC

Dataset name	Subjects	*N*	Modality	Positive condition	Negative condition	Reference
KTMD anaesthesia	Macaque monkeys	4	ECoG	Awake	Sedated	[132]
Propofol anaesthesia	Healthy volunteers	7	EEG	Awake	Sedated	[47, 133]
Psilocybin (PSIL)	Healthy volunteers	14	MEG	Drug	Placebo	[134]
Ketamine (KET)	Healthy volunteers	19	MEG	Drug	Placebo	[135]
LSD	Healthy volunteers	15	MEG	Drug	Placebo	[136]
Sleep	Epilepsy patients	10	EEG	Awake	Slow-wave sleep	[137]

With the aim of facilitating comparisons, we select two conditions (i.e. states of consciousness) from each dataset, and re-label them as *positive* or *negative*. This re-labelling is

[3]These analyses were done as part of a project in collaboration with Michael Schartner and Adam Barrett.

shown in Table 7.2, and is done in such a way that, according to the predictions of IITC, integrated information should be higher in the positive condition across all datasets.[4]

All datasets were cleaned and pre-processed as described in the original references, in all cases including pre-selection, artefact rejection, and notch filtering. For the band-resolved measures in Sec. 7.3.3, we use second-order Butterworth filters with the following cutoff frequencies: delta, 1–4 Hz; theta, 4–7 Hz; alpha, 8–13 Hz; beta, 14–25 Hz; and gamma, 25–50 Hz. Finally, data was split into pseudo-stationary 2 s epochs.

The procedure followed to compute all measures was similar to that in Ref. [124], and is as follows: first, a large sample of M sets of K different channels is selected at random. Typically, $M = 200$ and $K = \{4, 8, 12\}$.[5] For each of these subsets of K channels, measures are computed according to the descriptions in Sec. 2.3 (i.e. including the bipartition search), and bias-corrected using surrogate data methods [81]. Finally, each measure is averaged across the M subsets to obtain a single scalar number for each subject and condition. Since all the experiments employed a within-subjects design, we can make use of paired t-tests to report p-values and effect sizes comparing the positive and the negative condition.

7.3.2 Broadband integrated information is lower in wakeful rest

We begin with a direct test of the central claim of IITC that integrated information is higher during conscious states than during unconscious states. We evaluate the measures of integrated information reviewed in Chapter 2 on the datasets described above, for several values of integration timescale τ, and show the results in Fig. 7.1. Reported t-scores correspond to the diference between the positive and the negative condition of each dataset, such that scores greater than zero support IITC's predictions.

It seems painfully ironic that all the measures that in Chapter 6 did not agree on a single analysis now do agree, and they go *against* our predictions. In fact, these are most of the times very strong effects, noticeably beyond the standard significance threshold of $|t| = 1.96$.

When inspected for each measure separately, results do not look much better: it is the more recent measures (Φ^*, Φ_G), that behave more strongly against the theoretical predictions. The only measure that does to some extent follow the expected trend (for low τ) is whole-minus-sum Φ – which, also ironically, has been the most criticised in the recent IIT literature due to it being negative in some circumstances [9, 11, 10].

[4] This is only for the purpose of facilitating a comparison of results concerning the key claims of IITC. The very concept of "level of consciousness" has been called into question, and it is particularly contentious to speak of the psychedelic state as a "higher" state of consciousness [138].

[5] The limitation on low values of K comes from the computational demands imposed by the computation of Φ measures, and from the difficulty of estimating joint probability distributions over many variables from short segments of data.

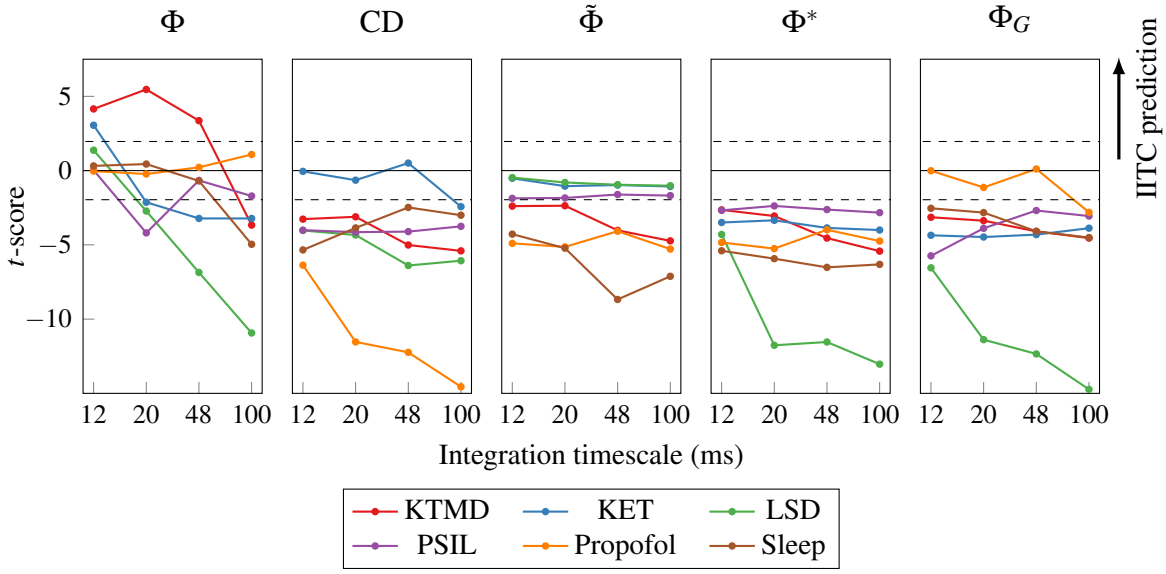


Figure 7.1: Broadband integrated information in six datasets of interest, for several integration timescales τ. Reported t-scores correspond to the diference between positive and negative conditions of each dataset, such that scores greater than zero support IITC's predictions. All measures, with the exception of whole-minus-sum Φ, go *against* IITC's prediction, being higher for the unconscious than the conscious state. (Dashed lines represent $t = \pm 1.96$.)

7.3.3 Band-specific integrated information is inconsistent

In addition to the broadband analysis above, we present the results of all measures evaluated on data pre-filtered for each frequency band (Fig. 7.2). Motivated by the result in Fig. 7.1 that measures tend to be closer to IITC's prediction for lower timescales, we report t-scores for $\tau = 12$ ms (although for all other timescales the results are qualitatively similar).

Unfortunately, this breakdown of integrated information by frequency band does not reach any strong conclusions. Although some measures do align with IITC's predictions in some frequency bands, the results are inconsistent and no clear pattern emerges.

Take, for example, alpha-band Φ_G under KTMD anaesthesia. This is a strong effect ($t > 5$), which would, looking at the KTMD dataset alone, survive multiple comparisons and provide a compelling result. However, this does not seem to generalise: alpha-band Φ_G is again in disagreement with IITC in the Propofol and Sleep datasets, calling into question the KTMD result. Similar inconsistencies appear in gamma-band Φ or beta-band Φ^*, among others. One might argue that these comparisons are unwarranted, and that the mechanism behind the loss of integrated information in natural sleep is different from that under KTMD anaesthesia. While that argument is valid, it does not get through the basic result that, in their current state, integrated information measures do not provide a successful index of consciousness in cases as clear as slow-wave sleep or general anaesthesia.

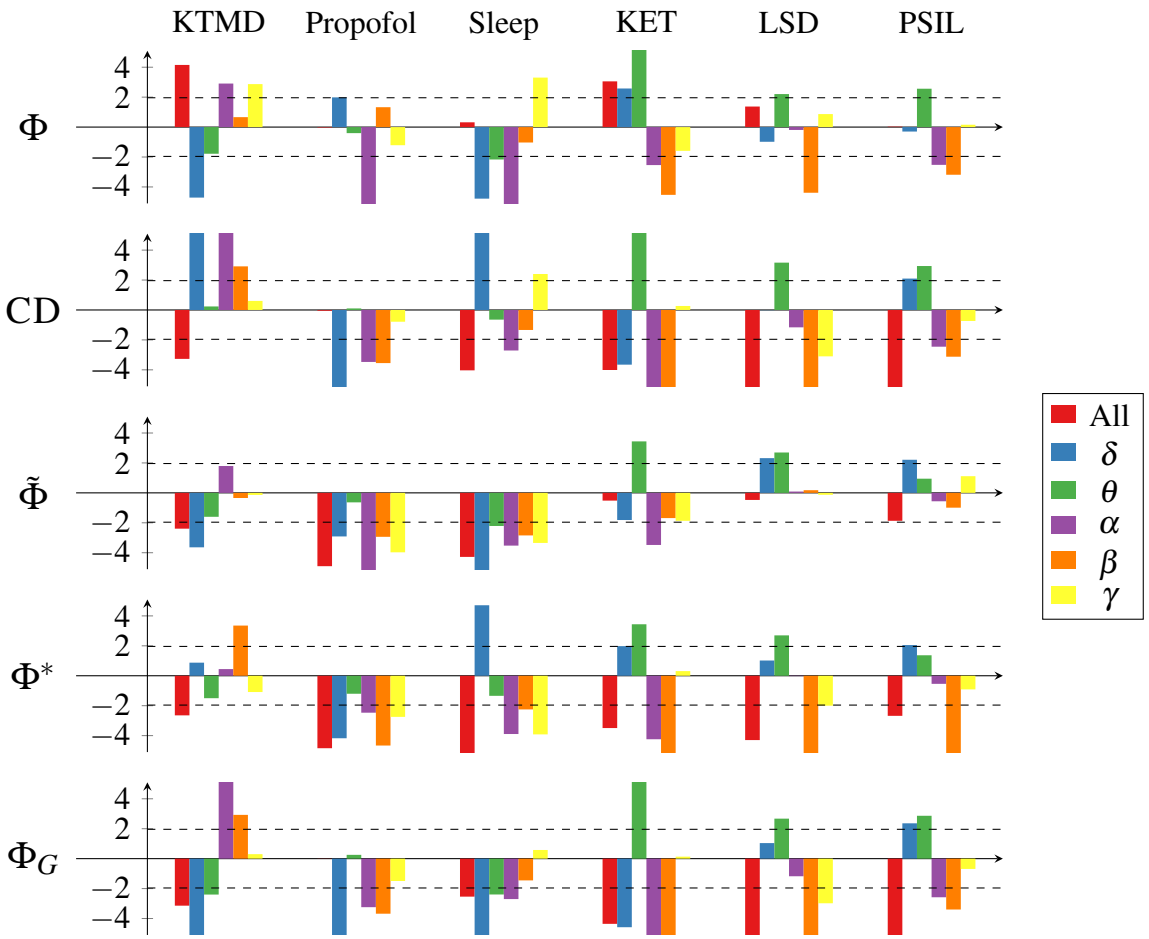


Figure 7.2: Integrated information measures for all datasets, resolved by frequency band. No consistent pattern emerges, and measures behave inconsistently across frequencies. (Dashed lines represent $t = \pm 1.96$.)

Interestingly, one consistent pattern does appear with psychedelics: most measures, in all three psychedelic substances (with the exception of $\tilde{\Phi}$ under PSIL) show a significant increase with respect to a placebo in the θ band. This is a puzzling result, since the literature on psychedelic neuroscience has not highlighted any special role of theta activity in the psychedelic state, and instead tends to focus on the disruptions to alpha activity [136, 139].

Finally, note that our results do not reproduce Lee's results of lower gamma-band Φ under propofol [120], Chang's results of reduced broadband $\tilde{\Phi}$ in sleep [121], or Kim's results of reduced alpha-band Φ^* under propofol or ketamine anaesthesia [124].

7.4 Modelling the effect of unobserved activity on Φ

One particular requirement of IIT (especially in its later iteration [6]) is that one must study systems at their smallest scale, examine all of their possible coarse-grainings, and pick the one where its intrinsic cause-effect power is maximised – i.e. "where the action happens" [140].

This proposal is theoretically appealing, but it is disappointingly far from our current access to neuronal activity. One single M/EEG sensor aggregates the activity of thousands or millions of neurons, and to avoid estimation problems we discard even more data and consider only a handful of M/EEG channels at a time. Throughout this process we are heavily *subsampling* the system of interest, which may have unintended consequences.

In other fields of neuroscience it is well known that subsampling can introduce undesired measurement artifacts.[6] However, in multivariate information theory subsampling has not been studied in detail until recently [142], and, to the best of our knowledge, in IIT subsampling has not been studied at all. In this section we put forward a temptative, exonerating hypothesis: that the values of integrated information reported above may still be compatible with the central IITC claims, but are heavily distorted by subsampling artefacts.

In support of our hypothesis, we study a bespoke toy model of "loss of consciousness."[7] Our model is a simple AR network (c.f. Chapter 6), made to vaguely resemble a subcortical thalamic region coupled to two cortical hemispheres. Cortical nodes are strongly connected within hemispheres and sparsely connected across them, and thalamic nodes have strong connections with each other and weak diffuse connections to the whole cortex (Fig. 7.3a).

This model is inspired by the known role of thalamo-cortical drive in slow-wave sleep: slow oscillations in deep sleep arise from the strengthening of thalamo-cortical feedback loops, that entrain distant parts of the brain into coherent delta oscillations [143, 144]. Accordingly, in our model we fix the intra-cortical and intra-thalamic coupling matrices, and simulate loss of consciousness by increasing thalamo-cortical coupling (Fig. 7.3b).

When we compute Φ on the whole cortex we observe that, as expected, Φ decreases as the thalamo-cortical coupling strengthens. However, if we run the same simulation, but now measure only one node from each hemisphere, a radically different picture appears: in the subsampled system, Φ now *increases* as thalamo-cortical coupling is strengthened. This spells a bleak future for practical applications of IIT – if we need access to all neurons in the brain to be confident in our Φ estimates, the prospect of having a meaningful empirical test of IITC in the near future becomes much less likely.

[6] For example, debate continues to rage in the statistical physics community as to whether neuronal avalanches are critical or driven subcritical based on subsampling arguments [80, 141].

[7] The model is so simplistic that every neuroscientific reference around it deserves bold-font scare quotes. We will omit these scare quotes in the interest of readability.

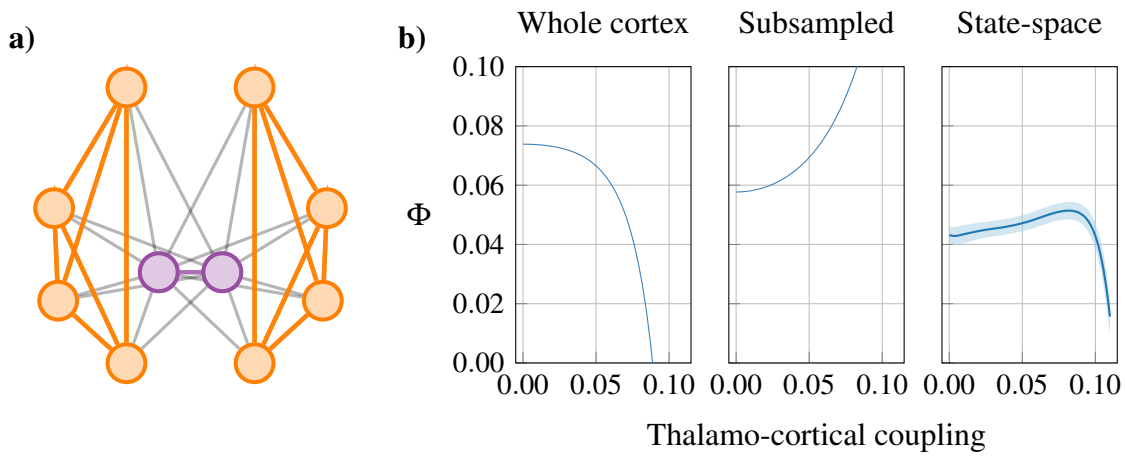


Figure 7.3: Integrated information in an AR thalamo-cortical drive model. (**a**) Diagram of the model, that includes a thalamus (purple) and two hemispheres (orange), with weak coupling between them. (**b**) As we raise the thalamo-cortical coupling, Φ of the cortex is reduced (left), although if we only measure one node in each hemisphere we will wrongly conclude it has increased (middle). The problem can be alleviated by using appropriate state-space reconstruction methods, which can recover the original downward trend (right).

Taken's acclaimed theorem [71] shows that we can reconstruct the topology of a system's attractor through the delay embeddings of one variable, through a procedure known as *state-space reconstruction*.[8] In an attempt to alleviate the subsampling problem, we use Barnett & Seth's method of state-space estimates for Granger causality [146] and build a new state-space estimate of Φ that we can evaluate on the same two nodes.

As shown in the right pane of Fig. 7.3b, with state-space reconstruction we are able to (at least partially) recover the downward trend from the whole-cortex scenario by measuring only two nodes. This suggests a possible way forward for practical IIT research, both in terms of further experimental and simulation analyses, as well as formal theoretical work linking statistical information theory with unobserved hidden variables [147].

It should be noted that this is in no way a serious model of sleep, thalamo-cortical interactions or loss of consciousness. It is merely a proof of concept, loosely based on established neuroscentific findings, that even in very simple systems subsampling may have strong effects on observed experimental Φ results. Of course, this model is not a full explanation of the results in Fig. 7.1, nor does it provide strong support for IITC, but it does show those results may still be *compatible* with the IITC predictions.

[8] This methodology has already been employed successfully in the study of consciousness, although not within IIT [145].

7.5 Conclusion of Part II

After the applications to complex systems studied in Part I, in this Part we have focused on properties of IIT relevant for a theory of consciousness. As a result, we have revealed two limitations of IITC in its current form.

First, Chapter 6 presented a theoretical limitation: we reviewed and benchmarked multiple Φ measures consistent with the principles of IIT and found a striking diversity in their behaviour. This shows that the mathematical basis of IIT is not restrictive enough to determine a unique measure of integrated information, and hints at the fact that there may not be such a thing as a single universal measure of integration.

Second, Chapter 7 evaluated the experimental merit of IIT as a theory of consciousness, both reviewing published literature and presenting new evidence. The result is that the question is far from settled – published articles show little agreement, and Φ often has the *opposite* trend from that predicted by IITC. Through toy models we show that these paradoxical results may at least in part be attributed to the large number of unobserved variables in contemporary neuroimaging techniques, and point to state-space reconstruction as a potential way forward.

These two limitations, both theoretical and experimental, call for a more pragmatic and evidence-driven research programme in IIT. In other words, it is best that we focus our efforts on solving the *actual* problems of IIT, like making a theory that is applicable and practical, and elucidating the conditions in which it can be meaningfully tested. In particular, we argue that efforts to formulate (i) more specific measures of integrated information with explicit operational meaning, and (ii) robust estimators from neuroimaging data, are likely to benefit the development of IIT. The IITC goal is a formidable one, and as such it is to be expected that various sorts of auxiliary tools will need to be developed before a meaningful test can take place.

Part III

Beyond integrated information: Developments and alternatives

Chapter 8

Quantifying high-order interdependencies via multivariate extensions of the mutual information

Chapter summary

We revisit the mathematical basis of early IIT, with the aim of providing new, more suitable tools. Our investigation into the multivariate structure of information leads to a new measure, the *O-information*, capable of characterising synergy- and redundancy-dominated systems. We compare the O-information against Φ's predecessor, the Tononi-Sporns-Edelman measure of neural complexity, and argue that the O-information is better able to capture the intuitions behind the origins of IIT.

8.1 Introduction

As argued in Chapter 6, some of the limitations encountered by IIT stem from the ambiguity of the underlying concepts of integration and differentiation. Ultimately, these concepts come from our intuitions from standard information theory, which (typically) deals with bivariate systems. Indeed, in bivariate systems these concepts can be unambiguosly defined: two variables are differentiated if they are statistically independent, or integrated if they are perfectly correlated. However, these intuitions do not easily generalise to multivariate systems with more than two elements.

We argue that, in essence, the mathematical challenges of IIT lie in the formalisation of a *multivariate theory of information*. In an attempt to formalise the intuitive concepts of integration and differentiation behind early IIT work [2, 3], in this chapter we take a step back and re-examine what it means for a multivariate stochastic system to be "more than the sum of its parts." In particular, we contextualise our work from the perspective of the ***Partial Information Decomposition*** (PID), which distinguishes different "types" of information that multiple predictors convey about a target variable [19]. Two types of particular interest for this work are *redundancy*, information that is conveyed by either predictor, and *synergy*, information conveyed by both predictors jointly but not by either of them separately.

The main contribution of this study is the formulation of a new quantity, the *O-information*, a measure able to distinguish between redundancy-dominated scenarios where three or more variables have copies of the same information, and synergy-dominated systems characterised by high-order patterns untraceable from low-order marginals. In contrast with existing quantities that require a division between predictors and target variables [148], the O-information is – to the best of our knowledge – the first symmetric quantity that can measure intrinsic synergy in systems of more than three variables. Furthermore, we argue that the O-information represents a more principled alternative to Tononi, Sporns and Edelman's *neural complexity* [2], and that the apparent trade-off between integration and differentiation can be resolved through lessons from PID.

8.2 Fundamentals of multivariate information

The crux of multivariate interdependencies is that their information-theoretic descriptions are not straighforward, as extensions of Shannon's classical results to multivariate settings have proven elusive [149]. The most established multivariate extensions of Shannon's mutual information are the *total correlation* [17] and the *dual total correlation* [150], which provide suitable metrics of overall correlation strength. Their values, however, differ in ways that are hard to understand, even gaining the adjective of "enigmatic" among scholars [151]. Another popular extension of the mutual information is the *interaction information* [152],

which is a signed measure obtained by applying the inclusion-exclusion principle to the Shannon entropy [153, 154]. Although this metric provides insighful results when applied to three variables, its is not easily interpretable when applied to larger groups [19].

As a prelude for the definition of the O-information, below we discuss multivariate interdependency via two dual persectives: as *shared randomness* and as *collective constraints*, and describe their relation to these known extensions of the mutual information.

8.2.1 Entropy and negentropy

> *For every outside there is an inside and for every inside there is an outside. And although they are different, they always go together.*
>
> Alan Watts, *Myth of myself*

Following the Bayesian interpretation of information theory, we define the *information contained in a system* as the average amount of data that an observer would gain after determining its configuration – i.e. after measuring it [155]. If each possible configuration is to be represented by a distinct sequence of bits, source coding theory [7, Ch. 5] shows that the optimal (i.e. shortest) labelling depends on prior information available before the measurement. Information, hence, refers to how the observer's state of knowledge changes after measurement, quantifying the amount of bits that are revealed through this process.

For concreteness, consider an observer measuring a system of n discrete variables, $\boldsymbol{X}^n = (X_1, ..., X_n)$. If the observer only knows that each variable X_j can take values over a finite alphabet $\mathcal{X}_j$, the amount of bits needed to specify the state of X_j is $\log|\mathcal{X}_j|$ (logarithms are in base 2 unless specified otherwise). In contrast, if the observer knows that the system's behaviour follows a probability distribution $p_{\boldsymbol{X}^n}$, the average amount of information in the system reduces to the *entropy* $H(\boldsymbol{X}^n) := -\sum_{\boldsymbol{x}^n} p_{\boldsymbol{X}^n}(\boldsymbol{x}^n) \log p_{\boldsymbol{X}^n}(\boldsymbol{x}^n)$ [155]. The difference,

$$\mathcal{N}(\boldsymbol{X}^n) := \sum_{j=1}^{n} \log|\mathcal{X}_j| - H(\boldsymbol{X}^n) \,, \tag{8.1}$$

is known as *negentropy* [156], and corresponds to the information about the system that is disclosed by the knowledge of the statistics, before any measurement takes place [23].

Probability distributions are, from this perspective, a compendium of soft and hard constraints that reduce the effective phase space that the system can explore (hard constraints completely forbid some configurations; soft constraints make them improbable). Consequently, a given distribution divides the phase space in an admisible region quantified by the entropy, and an inadmissible region quantified by the negentropy. Each part describes the system's structure from a different point of view: the entropy refers to what the system can do, and the negentropy to what it can't.

8.2.2 The two faces of interdependency

Collective constraints

In the same way as $\mathcal{N}(\boldsymbol{X}^n)$ quantifies the strength of the overall constraints that rule the system, the constraints that affect individual variables are captured by the *marginal negentropies* $\mathcal{N}(X_j) \coloneqq \log|\mathcal{X}_j| - H(X_j)$. Intuitively, the constraints that affect the whole system are richer than individual constraints, as the latter do not take into account collective effects. Their difference,

$$\begin{aligned} \mathrm{TC}(\boldsymbol{X}^n) &\coloneqq \mathcal{N}(\boldsymbol{X}^n) - \sum_{j=1}^{n} \mathcal{N}(X_j) \\ &= \sum_{j=1}^{n} H(X_j) - H(\boldsymbol{X}^n) \ , \end{aligned} \tag{8.2}$$

quantifies the strength of the "collective constraints." This quantity is known as *total correlation* [17] (or *multi-information* [157]), and was briefly introduced earlier in Sec. 2.3.3. By re-writing this relationship as $\mathcal{N}(\boldsymbol{X}^n) = \sum_j \mathcal{N}(X_j) + \mathrm{TC}(\boldsymbol{X}^n)$ one finds that the constraints prescribed by the distribution are of two types: constraints confined to individual variables, and collective constraints that restrict groups of two or more variables.

Example 1. *Consider X_1 and X_2 to be binary random variables with $p_{X_1,X_2}(0,1) = p_{X_1,X_2}(1,0) = 1/2$. This distribution divides the total information (two bits) into $H(X_1,X_2) = 1$ and $\mathcal{N}(X_1,X_2) = 1$. Moreover, $\mathcal{N}(X_1) = \mathcal{N}(X_2) = 0$ and therefore $\mathrm{TC}(X_1,X_2) = \mathcal{N}(X_1,X_2) = 1$, confirming that the constraints act on both X_1 and X_2. As a contrast, consider Y_1 and Y_2 binary random variables with distribution $p_{Y_1,Y_2}(0,0) = p_{Y_1,Y_2}(1,0) = 1/2$. In this case $\mathcal{N}(Y_1) = 0$ while $\mathcal{N}(Y_2) = \mathcal{N}(Y_1,Y_2) = 1$, showing that the only constraint in this system acts solely over Y_2. Accordingly, for this case $\mathrm{TC}(Y_1,Y_2) = 0$.*

Shared randomness

As we did for $\mathcal{N}(\boldsymbol{X}^n)$, let us decompose $H(\boldsymbol{X}^n)$ in individual and collective components. To do this, we introduce the quantity $R_j = H(X_j|\boldsymbol{X}^n_{-j})$ as a metric of how independent X_j is from the rest of the system $\boldsymbol{X}^n_{-j} = (X_1,...,X_{j-1},X_{j+1},...,X_n)$. According to distributed source coding theory [149, Ch. 10.5], R_j corresponds to the data contained in X_j that cannot be extracted from measurements of other variables.[1] The quantity $\sum_{j=1}^n R_j$ is known as the *residual entropy* [158] (originally introduced under the name of *erasure entropy* [159]), and quantifies the total information that can only be accessed by measuring a specific variable,

[1] In fact, a direct calculation shows that the variables $\boldsymbol{X}^n$ are independent if and only if $\sum_j R_j = H(\boldsymbol{X}^n)$.

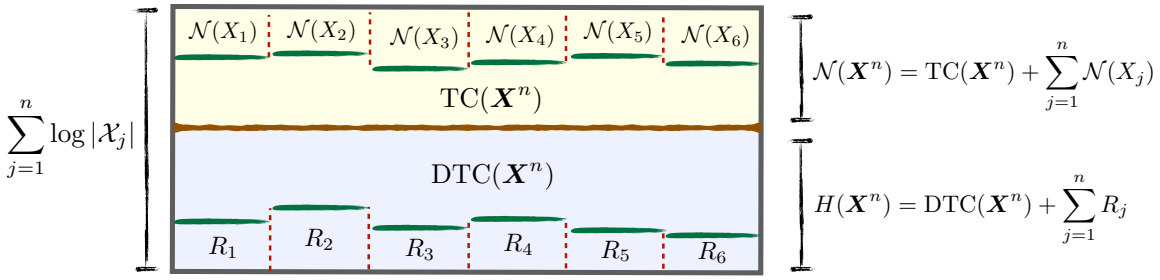


Figure 8.1: The total information that can be stored in the system $\boldsymbol{X}^n$ ($\sum_{j=1}^{n} \log|\mathcal{X}_j|$) is decomposed by a given state of knowledge (i.e. a probability distribution) into two parts: what is determined by the constraints (the negentropy, $\mathcal{N}(\boldsymbol{X}^n)$), and what is not instantiated until an actual measurement takes place (the entropy, $H(\boldsymbol{X}^n)$). Both terms can be further decomposed into their individual and collective components, yielding different perspectives on interdependency seen as either collective constraints (measured by the total correlation $\mathrm{TC}(\boldsymbol{X}^n)$) or shared randomness (corresponding to the dual total correlation $\mathrm{DTC}(\boldsymbol{X}^n)$).

i.e. the amount of "non-shared randomness." Accordingly, the difference

$$\mathrm{DTC}(\boldsymbol{X}^n) := H(\boldsymbol{X}^n) - \sum_{j=1}^{n} R_j \tag{8.3}$$

is known as *dual total correlation* [150] (being also known as *binding information* [158] and *excess entropy* [160]), and refers to the part of the joint entropy that is shared by two or more variables – equivalently, information that can be obtained by measuring more than one specific variable. As the entropy corresponds to the randomness within the system, the dual total correlation quatifies the "shared randomness" that exists among the variables.

Example 2. *Let us consider X_1, X_2 and Y_1, Y_2 from Example 1. For the former system one finds that $R_1 = R_2 = 0$ and hence $\mathrm{DTC}(X_1, X_2) = H(X_1, X_2) = 1$, which means that the randomness within the system can be retrieved from measuring either X_1 or X_2. In contrast, when considering Y_1, Y_2 one finds that $R_2 = 0$ and $R_1 = H(Y_1, Y_2) = 1$, and hence $\mathrm{DTC}(Y_1, Y_2) = 0$. This implies that the randomness of the system can be retrieved by measuring only Y_1.*

Wrapping up, one can re-write Eq. (8.1) using Eqs. (8.2) and (8.3) and express the total information encoded in the system $\boldsymbol{X}^n$ in terms of constraints and randomness (Fig. 8.1):

$$\begin{aligned} \sum_{j=1}^{n} \log|\mathcal{X}_j| &= \mathcal{N}(\boldsymbol{X}^n) + H(\boldsymbol{X}^n) \\ &= \underbrace{\left[\mathrm{TC}(\boldsymbol{X}^n) + \sum_{j=1}^{n} \mathcal{N}(X_j)\right]}_{\substack{\text{Collective and individual} \\ \text{constraints}}} + \underbrace{\left[\mathrm{DTC}(\boldsymbol{X}^n) + \sum_{j=1}^{n} R_j\right]}_{\substack{\text{Shared and private} \\ \text{randomness}}}. \end{aligned}$$

8.3 O-information: Redundancy minus synergy

8.3.1 Definition and basic properties

The TC and DTC provide complementary metrics of interdependence strength. Following Occam's Razor, one might ask which of these perspectives allows for a shorter (i.e. more parsimonious) description. This is answered by the following definition:

Definition 1. *The O-information (shorthand for "information about Organisational structure") of the system described by the random vector $\boldsymbol{X}^n$ is defined as*

$$\begin{aligned}\Omega(\boldsymbol{X}^n) &:= \mathrm{TC}(\boldsymbol{X}^n) - \mathrm{DTC}(\boldsymbol{X}^n) \\ &= (n-2)H(\boldsymbol{X}^n) + \sum_{j=1}^{n} \left[H(X_j) - H(\boldsymbol{X}^n_{-j})\right].\end{aligned} \tag{8.4}$$

Intuitively, $\Omega(\boldsymbol{X}^n) > 0$ indicates that the interdependencies can be more efficiently explained as shared randomness, while $\Omega(\boldsymbol{X}^n) < 0$ implies that viewing them as collective constraints can be more convenient. Note that $\Omega(\boldsymbol{X}^n)$ was first introduced as "enigmatic information" in Ref. [151], although now that its properties have been revealed we choose to give it a more appropriate name.

To develop some insight about the O-information, let us compare it with the *interaction information*,[2] which is a signed metric defined according to the inclusion-exclusion principle by

$$I(X_1;X_2;\ldots;X_n) := \sum_{\boldsymbol{\gamma}\subseteq\{1,\ldots,n\}} (-1)^{|\boldsymbol{\gamma}|+1} H(\boldsymbol{X}^{\boldsymbol{\gamma}}) \,, \tag{8.5}$$

where the sum is over all the subsets of indices $\boldsymbol{\gamma} \subseteq \{1,\ldots,n\}$, with $|\boldsymbol{\gamma}|$ being the cardinality of $\boldsymbol{\gamma}$ and $\boldsymbol{X}^{\boldsymbol{\gamma}}$ the vector of all variables with indices in $\boldsymbol{\gamma}$. For $n = 2$, Eq. (8.5) reduces to the well-known mutual information,

$$I(X_1;X_2) = H(X_1) + H(X_2) - H(X_1,X_2) \,.$$

For $n = 3$, Eq. (8.5) gives

$$\begin{aligned}I(X_1;X_2;X_3) &= I(X_i;X_j) - I(X_i;X_j|X_k) \\ &= I(X_i;X_j) + I(X_i;X_k) - I(X_i;X_j,X_k)\end{aligned} \tag{8.6}$$

for $\{i,j,k\} = \{1,2,3\}$, which, seen from a PID angle, is known to measure the difference between synergy and redundancy in $\boldsymbol{X}^3$ [19]. Interestingly, although PID is an asymmetric construction (predictors and targets cannot be swapped), Eq. (8.6) is fully symmet-

[2] The interaction information is closely related to the *I-measures* [154], the *co-information* [161], and the *multi-scale complexity* [162].

ric under permutations of all three variables. Specifically, redundancy dominates when $I(X_1;X_2;X_3) \geq 0$; e.g. if X_1 is a Bernoulli random variable with $p = 1/2$ and $X_1 = X_2 = X_3$, then $I(X_1;X_2;X_3) = 1$. In contrast, synergy dominates when $I(X_1;X_2;X_3) \leq 0$, corresponding to statistical structures that are present in the full distribution but not in the pairwise marginals. For example, if Y_1 and Y_2 are independent Bernoulli variables with $p = 1/2$ and $Y_3 = Y_1 + Y_2 \pmod 2$ (i.e. an `xor` logic gate) then $I(Y_1;Y_2;Y_3) = -1$, since these variables are pairwise independent while globally correlated [23]. Unfortunately, for $n \geq 4$ the co-information no longer reflects the balance between redundancy and synergy [19, Section V].

In contrast with the interaction information, the next Lemma presents some basic properties of Ω.

Lemma 1. *The O-information satisfies the following properties:*

(i) Ω *does not depend on the order of* $X_1, \ldots, X_n$.

(ii) $\Omega(X_1, X_2) = 0$ *for any* $p_{X_1X_2}$.

(iii) $\Omega(X_1, X_2, X_3) = I(X_1;X_2;X_3)$ *for any* $p_{\boldsymbol{X}^3}$.

Property (i) shows that Ω reflects an intrinsic property of the system, without the need of dividing the variables in groups with differentiated roles (e.g. targets vs predictors, or input vs output). Property (ii) confirms that Ω captures only interactions that go beyond pairwise relationships. Finally, Property (iii) shows that when $n = 3$ the O-information is equal to $I(X_1;X_2;X_3)$. Interestingly, a direct calculation shows that if $n > 3$ then in general $\Omega(\boldsymbol{X}^n) \neq I(X_1;X_2;\ldots;X_n)$.

At this stage, one might wonder if the O-information could provide a metric for quantifying the balance of redundancy and synergy, as the interaction information does for $n = 3$. Intutively, one could expect redundant systems to have small $\mathrm{DTC}(\boldsymbol{X}^n)$ due to the multiple copies of the same information that exist in the system, while having large values of $\mathrm{TC}(\boldsymbol{X}^n)$ because of the constraints that are needed to ensure that the variables remain correlated. On the other hand, synergistic systems are expected to have small values of $\mathrm{TC}(\boldsymbol{X}^n)$ due to the few high-order constraints that rule the system, while having larger values of $\mathrm{DTC}(\boldsymbol{X}^n)$ due to the weak low-order structure. These insights are captured in the following definition, which is supported by multiple findings presented in the following sections.

Definition 2. *If* $\Omega(\boldsymbol{X}^n) > 0$ *we say that the system is redudancy-dominated, while if* $\Omega(\boldsymbol{X}^n) < 0$ *we say it is synergy-dominated.*

8.3.2 Information decompositions

This section presents information decompositions that deepen our understanding of the O-information, and will help us prove some of its useful properties. In the following, we first introduce the partition lattice, which is then used to build decompositions of the TC, DTC, and Ω.

The lattice of partitions

Let us characterise the possible ways in which one can sequentially decompose the system described by $\boldsymbol{X}^n$. For this, consider partitions $\pi = (\boldsymbol{\alpha}_1|\boldsymbol{\alpha}_2|...|\boldsymbol{\alpha}_m)$ of the set of indices $\{1,...,n\}$, which are collections of *cells* $\boldsymbol{\alpha}_j = \{\alpha_j^1,...,\alpha_j^{l(j)}\} \subset \{1,...,n\}$ that are disjoint and satisfy $\bigcup_{j=1}^m \boldsymbol{\alpha}_j = \{1,...,n\}$. The collection of all possible partitions of $\{1,...,n\}$, denoted by $\mathcal{P}_n$, has a lattice structure[3] enabled by the partial ordering introduced by the refinement relationship, in which $\pi_2 \succeq \pi_1$ if π_2 is *finer*[4] than π_1 (or, equivalently, if π_1 is *coarser* than π_2). A partition π_2 is said to *cover* π_1 if $\pi_2 \succeq \pi_1$ and it is not possible to find another partition π_3 such that $\pi_2 \succeq \pi_3 \succeq \pi_1$.[5] For this partial order relationship, $\pi_{\text{source}} = (12...n)$ is the unique infimum of $\mathcal{P}_n$, and $\pi_{\text{sink}} = (1|2|...|n)$ is the unique supremum of $\mathcal{P}_n$.

A directed acyclic graph (DAG) $\mathcal{G}_n$ can be built, where the nodes are the partitions in $\mathcal{P}_n$, and a directed edge exists from π_1 to π_2 if and only if π_2 covers π_1.[6] A path $\mathtt{p}$ in $\mathcal{G}_n$ joining two partitions π_a and π_b is a sequence of nodes $\mathtt{p} = (\pi_1,...,\pi_L)$, where $\pi_1 = \pi_\mathrm{a}$, $\pi_L = \pi_\mathrm{b}$, and π_{i+1} covers π_i for all $i \in \{1,...,L-1\}$. The collection of all paths from π_a to π_b is denoted by $\mathtt{P}(\pi_\mathrm{a},\pi_\mathrm{b})$.[7] If the edge joining π_1 and π_2 has a weight $v(\pi_1,\pi_2)$ associated, then the corresponding *path weight* of $\mathtt{p} = (\pi_1,...,\pi_L)$ is merely the summation of all edge weights along $\mathtt{p}$:

$$W(\mathtt{p};v) := \sum_{k=1}^{L-1} v(\pi_k,\pi_{k+1}) \,. \tag{8.7}$$

[3] A lattice is a partially ordered set with a unique infimum and supremum. For more details on this construction, see Ref. [163]

[4] If $\pi_1,\pi_2 \in \mathcal{P}_n$ with $\pi_1 = (\boldsymbol{\alpha}_1|...|\boldsymbol{\alpha}_r)$ and $\pi_2 = (\boldsymbol{\beta}_1|...|\boldsymbol{\beta}_s)$, π_1 is *finer* than π_2 if for each $\boldsymbol{\alpha}_i$ exists $\boldsymbol{\beta}_k$ such that $\boldsymbol{\alpha}_i \subset \boldsymbol{\beta}_k$.

[5] It is direct to see that π_2 covers π_1 if and only if it is an "elementary refinement," i.e. π_2 can be obtained from π_1 by dividing one cell of π_1 in two. Hence, if π_2 covers π_1 then $|\pi_2| = |\pi_1| + 1$, where $|\pi|$ is the number of (non-empty) cells of π.

[6] Put simply, there is an edge from π_1 to π_2 if π_2 results from taking π_1 and splitting one of its cells in two.

[7] It is direct to check that $\pi_\mathrm{b} \succ \pi_\mathrm{a}$ if and only if $\mathtt{P}(\pi_\mathrm{a},\pi_\mathrm{b}) \neq \varnothing$. Moreover, all $\mathtt{p} \in \mathtt{P}(\pi_\mathrm{a},\pi_\mathrm{b})$ have the same length, given by $|\mathtt{p}| = ||\pi_\mathrm{b}| - |\pi_\mathrm{a}||$, where $|\mathtt{p}|$ is the number of edges in the path.

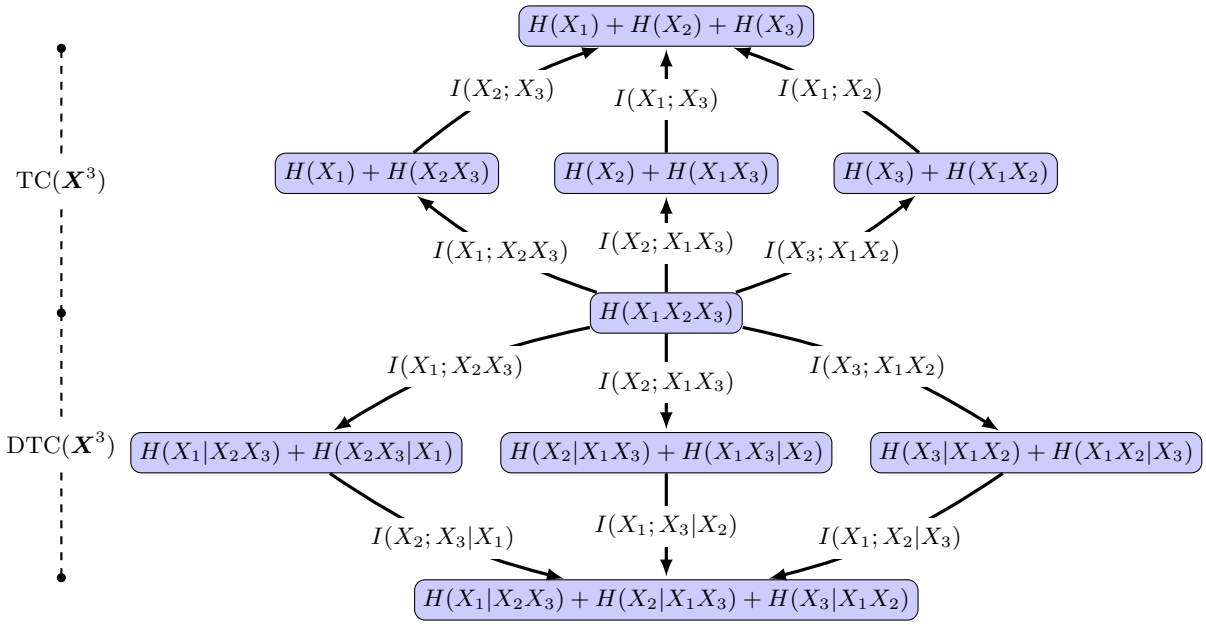


Figure 8.2: *Double diamond diagram* with the possible sequences of binary partitions of three variables. Every path from the source node ($H(\boldsymbol{X}^3)$ to the two sink nodes ($H(X_1)+H(X_2)+H(X_3)$ and $H(X_1|X_2X_3)+H(X_2|X_1X_3)+H(X_3|X_1X_2)$) corresponds to a decomposition of either $\mathrm{TC}(\boldsymbol{X}^3)$ or $\mathrm{DTC}(\boldsymbol{X}^3)$.

Lattice decompositions of $\mathrm{TC}(\boldsymbol{X}^n)$ and $\mathrm{DTC}(\boldsymbol{X}^n)$

Let us build some useful weight functions over $\mathcal{G}_n$. We first assign to each node $\pi = (\boldsymbol{\alpha}_1|...|\boldsymbol{\alpha}_L) \in \mathcal{P}_n$ the value

$$H(\pi) := H\Big(\prod_{j=1}^{L} p_{\boldsymbol{X}^{\boldsymbol{\alpha}_j}}\Big) = \sum_{j=1}^{L} H\big(\boldsymbol{X}^{\boldsymbol{\alpha}_j}\big)$$

with $\boldsymbol{X}^{\boldsymbol{\alpha}_j} = (X_{\alpha_j^1},...,X_{\alpha_j^{l(j)}})$, which corresponds to the entropy of the probability distribution $\prod_{j=1}^{L} p_{\boldsymbol{X}^{\boldsymbol{\alpha}_j}}$ that includes interdependencies within cells, but not across cells. To each edge of $\mathcal{G}_n$ we assign a weight

$$v_{\mathrm{h}}(\pi_1,\pi_2) := H(\pi_2) - H(\pi_1) \ . \tag{8.8}$$

Since $H(\pi_{\mathrm{a}}) \geq H(\pi_{\mathrm{b}})$ if $\pi_{\mathrm{a}} \succeq \pi_{\mathrm{b}}$, one can represent $\mathcal{G}_n$ under v_{h} by placing nodes with more cells in higher layers (see the upper half of Figure 8.2).

Alternatively, let us now consider the residual entropy of $\pi = (\boldsymbol{\alpha}_1|...|\boldsymbol{\alpha}_m) \in \mathcal{P}_n$, which is given by $R(\pi) := \sum_{k=1}^{m} R_{\boldsymbol{\alpha}_k}$, with

$$R_{\boldsymbol{\alpha}_k} := H(\boldsymbol{X}^{\boldsymbol{\alpha}_k}|\boldsymbol{X}^{\boldsymbol{\alpha}_1},...,\boldsymbol{X}^{\boldsymbol{\alpha}_{k-1}},\boldsymbol{X}^{\boldsymbol{\alpha}_{k+1}},...,\boldsymbol{X}^{\boldsymbol{\alpha}_m}).$$

The above quantity generalises the notion of residual entropy per individual variable given in Section 8.2.2.[8] With this, we introduce weights to each edge of $\mathcal{G}_n$ based on residuals, given by

$$v_{\mathrm{r}}(\pi_1, \pi_2) := R(\pi_1) - R(\pi_2) \ . \tag{8.9}$$

As residual entropy decreases when the partition is refined (see Appendix B.3), in this case one can illustrate the corresponding DAG by placing nodes with more cells in lower positions (see lower half of Figure 8.2).

Conveniently, for every edge v_h and v_r correspond to a mutual information or a conditional mutual information term, respectively. This is illustrated in the edges of Figure 8.2 and formalised in the Appendix.

The next result shows that the weights v_{h} and v_{r} provide decompositions for $\mathrm{TC}(\boldsymbol{X}^n)$ and $\mathrm{DTC}(\boldsymbol{X}^n)$, respectively.

Lemma 2. *Every path* $\mathrm{p} \in P(\pi_{source}, \pi_{sink})$ *provides the following decompositions:*

$$\begin{aligned} \mathrm{TC}(\boldsymbol{X}^n) &= W(\mathrm{p}; v_h) \\ \mathrm{DTC}(\boldsymbol{X}^n) &= W(\mathrm{p}; v_r) \ . \end{aligned}$$

Proof. See Appendix B.4. □

Let us now leverage these results to develop a decomposition for the O-information. For this, let us first introduce a new assignment of weights for the edges of $\mathcal{G}_n$, given by

$$v_{\mathrm{s}}(\pi_1, \pi_2) := v_{\mathrm{h}}(\pi_1, \pi_2) - v_{\mathrm{r}}(\pi_1, \pi_2) \ . \tag{8.10}$$

In contrast with Eqs. (8.8) and (8.9), these weights can attain negative values. The following key result shows that the weights v_s provide a decomposition of $\Omega(\boldsymbol{X}^n)$.

Proposition 1. *Every path* $\mathrm{p} \in P(\pi_{source}, \pi_{sink})$ *provides the following decomposition:*

$$\Omega(\boldsymbol{X}^n) = W(\mathrm{p}; v_s) \ . \tag{8.11}$$

Moreover, Eq. (8.11) *is a sum of interaction information terms of the form in Eq.* (8.6).

Proof. See Appendix B.5. □

This finding extends property (iii) of Lemma 1 by showing that the O-information can always be expressed as a sum of interaction information terms of three sets of variables (see Corollary 1 below for an explicit example of this). As a consequence, the O-information inherits the capabilities of the triple interaction information for reflecting the balance between

[8] In effect, $R_{\boldsymbol{\alpha}_k}$ represents the portion of the entropy of the k-th cell that is not shared with other cells

synergies and redundancies, and is applicable to systems of any size. This decomposition of the O-information is analogous to the one introduced in Ref. [164] for the redundancy-synergy index.

An inconventient feature of partition lattices is that they grow super-exponentially with system size,[9] and hence heuristic methods for exploring them are necessary. A particularly interesting sub-family of $\mathtt{P}(\pi_{\text{source}}, \pi_{\text{sink}})$ are the "assembly paths," which have the form (up to re-labelling)

$$\mathtt{p}_{\text{a}} = \{(12...n), (12...(n-1)|n), ..., (1|2|...|n)\}. \tag{8.12}$$

These paths can be thought of as the process of first connecting X_1 and X_2, then connecting X_3 to $\boldsymbol{X}^2$, and so on – i.e. as assembling the system by sequentially placing its pieces together. The following corollary of Proposition 1 presents useful decompositions of $\text{TC}(\boldsymbol{X}^n)$, $\text{DTC}(\boldsymbol{X}^n)$, and $\Omega(\boldsymbol{X}^n)$ in terms of assembly paths.

Corollary 1. *For an assembly path as given in Eq. (8.12), the corresponding decompositions of the TC, DTC and O-information are*

$$\text{TC}(\boldsymbol{X}^n) = \sum_{i=2}^{n} I(X_i; \boldsymbol{X}^{i-1}) \ , \tag{8.13}$$

$$\text{DTC}(\boldsymbol{X}^n) = I(X_n; \boldsymbol{X}^{n-1}) + \sum_{j=2}^{n-1} I(X_j; \boldsymbol{X}^{j-1} | \boldsymbol{X}_{j+1}^n), \tag{8.14}$$

$$\Omega(\boldsymbol{X}^n) = \sum_{k=2}^{n-1} I(X_k; \boldsymbol{X}^{k-1}; \boldsymbol{X}_{k+1}^n) \ , \tag{8.15}$$

with $\boldsymbol{X}_k^n = (X_k, X_{k+1}, ..., X_n)$ *and* $\boldsymbol{X}^k = (X_1, ..., X_k)$.

As a concluding remark, let us note that the decompositions presented by Corollary 1 are valid for any relabeling of the indices (i.e. any ordering of the system's variables). This property is a direct consequence of the lattice construction developed in this subsection, which plays an important role in the following sections.

[9] The number of the nodes of $\mathcal{G}_n$ grows with the *Bell numbers*, known for their super-exponential growth rate [165]. To find the number of paths in $\mathtt{P}(\pi_{\text{source}}, \pi_{\text{sink}})$, note that if one starts from the sink and moves towards the source, every step corresponds to merging two cells into one. Therefore, as selecting two out of m cells gives $\binom{m}{2}$ choices, the total number of paths is given by

$$|\mathtt{P}(\pi_{\text{source}}, \pi_{\text{sink}})| = \sum_{m=2}^{n} \binom{m}{2} = \frac{n!(n-1)!}{2^{n-1}} \ ,$$

which grows faster than the Bell numbers.

8.3.3 Characterising extreme values of Ω

Let us explore the range of values that the O-information can attain. As a first step, Lemma 3 provides bounds for $\mathrm{TC}(\boldsymbol{X}^n)$, $\mathrm{DTC}(\boldsymbol{X}^n)$, and $\Omega(\boldsymbol{X}^n)$.

Lemma 3. *The following bounds hold:*

- $(n-1)\log|\mathcal{X}| \geq \mathrm{TC}(\boldsymbol{X}^n) \geq 0$,
- $(n-1)\log|\mathcal{X}| \geq \mathrm{DTC}(\boldsymbol{X}^n) \geq 0$,
- $n\log|\mathcal{X}| \geq \mathrm{TC}(\boldsymbol{X}^n) + \mathrm{DTC}(\boldsymbol{X}^n) \geq 0$,
- $(n-2)\log|\mathcal{X}| \geq \Omega(\boldsymbol{X}^n) \geq (2-n)\log|\mathcal{X}|$.

Where we use the shorthand notation $|\mathcal{X}| := \max_{j=1,\ldots,n}|\mathcal{X}_j|$ for the cardinality of the largest alphabet in $\boldsymbol{X}^n$. Moreover, these bounds are tight.

Proof. See Appendix B.6. □

Let us introduce some nomenclature. A random binary vector $\boldsymbol{X}^n$ is said to be a "n-bit copy" if X_1 is a Bernoulli random variable with parameter $p = 1/2$ (i.e. a *fair coin*) and $X_1 = X_2 = \ldots = X_n$. Also, a random binary vector $\boldsymbol{X}^n$ is said to be a "n-bit `xor`" if $\boldsymbol{X}^{n-1}$ are i.i.d. fair coins and $X_n = \sum_{j=1}^{n-1} X_j \pmod 2$. Our next result shows that these two distributions attain the upper and lower bounds of the O-information.

Proposition 2. *Let $\boldsymbol{X}^n$ be a binary vector with $n \geq 3$. Then,*

1. *$\Omega(\boldsymbol{X}^n) = n-2$, if and only if $\boldsymbol{X}^n$ is a n-bit copy.*
2. *$\Omega(\boldsymbol{X}^n) = 2-n$, if and only if $\boldsymbol{X}^n$ is a n-bit `xor`.*

Proof. See Appendix B.7. □

Corollary 2. *The same proof can be used to confirm that for variables with $|\mathcal{X}_1| = \ldots = |\mathcal{X}_n| = m$, the maximum $\Omega(\boldsymbol{X}^n) = (n-2)\log m$ is attained by variables which are a copy of each other, while the minimum $\Omega(\boldsymbol{X}^n) = (2-n)\log m$ corresponds to when $\boldsymbol{X}^{n-1}$ are independent and uniformly distributed and $X_n = \sum_{j=1}^{n-1} X_j \pmod m$.*

Proposition 2 points out an important difference betwen the O-information and the interaction information: if $\boldsymbol{X}^n$ is an n-bit `xor` then $\Omega(\boldsymbol{X}^n) = 2-n$ is consistently negative and decreasing with n, while $I(X_1;\ldots;X_n) = (-1)^{n+1}$ oddly oscillates between -1 and $+1$. This result also points out the convenience of merging $\mathrm{TC}(\boldsymbol{X}^n)$ and $\mathrm{DTC}(\boldsymbol{X}^n)$ into $\Omega(\boldsymbol{X}^n)$, as only the latter has the n-bit copy and the n-bit `xor` as unique extremes.

Finally, note that Ω is continuous over small changes in $p_{\boldsymbol{X}^n}$, as it can be expressed as a linear combination of Shannon entropies (see Definition 1), which are themselves continuous. Therefore, Proposition 2 guarantees that distributions that are similar to a n-bit copy have a positive O-information, while distributions close to a n-bit `xor` have negative O-information.

In summary, we have proposed a new multivariate extension of the mutual information, $\Omega(\boldsymbol{X}^n)$, and argued that it *quantifies the difference between overall redundancy and synergy in* $p(\boldsymbol{X}^n)$. The main arguments are three:

1. For $n = 3$, $\Omega(X_1, X_2, X_3) = I(X_1; X_2; X_3)$.
2. For $n > 3$, $\Omega(\boldsymbol{X}^n)$ is a sum of redundancies minus synergies between subsystems of $\boldsymbol{X}^n$.
3. $\Omega(\boldsymbol{X}^n)$ is maximised by a system consisting of 1 random bit and $n-1$ copies of it, and minimised by a system composed by successive `xor` gates.

In the rest of this chapter we investigate the relation between Ω and other notions of high-order effects coming from statistical mechanics and IIT 0.1. Further results on the O-information, including bounds and further examples, are presented in Appendices B.1 and B.2.

8.3.4 Ω and high-order interactions in statistical mechanics

In a completely orthogonal line of research from PID, a popular approach to represent high-order interactions in the statistical physics literature is via Hamiltonians that include interaction terms with three or more variables [166]. For example, systems of n spins (i.e. $\mathcal{X}_i = \{-1, 1\}$ for $i = 1, \ldots, n$) that exhibit k-th order interactions are usually represented by probability distributions of the form

$$p_{\boldsymbol{X}^n}(\boldsymbol{x}^n) = \frac{e^{-\beta \mathcal{H}_k(\boldsymbol{x}^n)}}{Z} \,, \tag{8.16}$$

where β is the inverse temperature, Z is a normalization constant (also known as the partition function), and $\mathcal{H}(\boldsymbol{x}^n)$ is a Hamiltonian given by

$$\begin{aligned} \mathcal{H}_k(\boldsymbol{x}^n) = &- \sum_{i=1}^{n} J_i x_i - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} J_{i,j} x_i x_j \\ &\ldots - \sum_{|\boldsymbol{\gamma}|=k} J_{\boldsymbol{\gamma}} \prod_{i \in \boldsymbol{\gamma}} x_i \,, \end{aligned}$$

where the last sum runs over all subsets $\boldsymbol{\gamma} \subseteq \{1,\ldots,n\}$ of size $|\boldsymbol{\gamma}| = k$. For example, a Hamiltonian with $k = 2$ corresponds to the standard Ising model.

According to Eq. (8.16), configurations with lower $\mathcal{H}_k(\boldsymbol{x}^n)$ are more likely to be visited. Note that J_i quantify external influences acting over individual spins, while $J_{\boldsymbol{\gamma}}$ for $|\boldsymbol{\gamma}| \geq 2$ represent the strength of the interactions; in particular, if $J_{i,k} > 0$ then the pair X_i, X_k tend to be aligned, while if $J_{i,k} < 0$ they tend to be anti-aligned. As a matter of fact, $\boldsymbol{X}^n$ are independent if and only if $J_{\boldsymbol{\gamma}} = 0$ for all $\boldsymbol{\gamma}$ with $|\boldsymbol{\gamma}| \geq 2$. Models with k-th order interactions have been studied via the maximum entropy principle [166], information geometry [167] and PID [168].

Considering the results presented in previous sections, one could expect that systems with high-order interactions (i.e. large k) should attain lower values of Ω than systems with low-order interactions (i.e. small k). To confirm this hypothesis, we studied ensembles of systems with k-th order interactions, and analised how the value of Ω is influenced by k. For this, we considered random Hamiltonians with $J_{\boldsymbol{\gamma}}$ drawn i.i.d. from a standard normal distribution and $\beta = 0.1$.

In agreement with intuition, results show that Ω is usually very close to zero for $k = 2$, and becomes negative as k grows (Figure 8.3). These results suggest that the notion of synergy measured by Ω is consistent with the traditional ideas of high-order interactions from statistical physics.

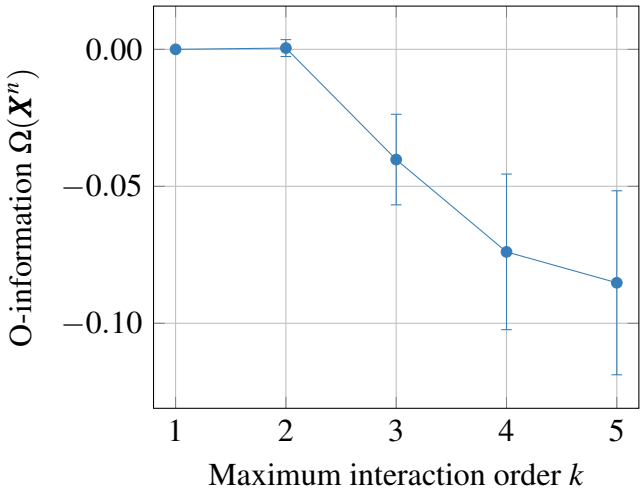


Figure 8.3: Mean value and 95% confidence intervals of Ω for ensembles of systems of $n = 5$ spins with randomly generated Hamiltonians. By including high-order interaction terms, net synergy increases and Ω decreases.

8.4 Complexity and integrated information

In their seminal 1994 article, Tononi, Edelman, and Sporns devised a measure of neural complexity (also known as *TSE complexity*) to describe the interplay between local segregation and global integration [2, 3]. The TSE complexity is defined as

$$C_{\mathrm{TSE}}(\boldsymbol{X}^n) := \sum_{k=1}^{n} \left[\frac{k}{n} \mathrm{TC}(\boldsymbol{X}^n) - C_n(k) \right] , \tag{8.17}$$

where $C_n(k) = \binom{n}{k}^{-1} \sum_{|\boldsymbol{\gamma}|=k} \mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\gamma}})$ is the average total correlation of the subsets $\boldsymbol{\gamma} \subseteq \{1,\dots,n\}$ of size $|\boldsymbol{\gamma}| = k$. By measuring the convexity of $C_n(k)$ as function of k, the TSE complexity attempts to distinguish scenarios that exhibit "relative statistical independence of small subsets of the system [...] and significant deviations from independence of large subsets" [2, Abstract], in the same spirit as our motivation behind Ω above.

To study the relationship between the TSE complexity and the O-information, it is useful to consider an alternative expression of the former:

$$C_{\mathrm{TSE}}(\boldsymbol{X}^n) = \sum_{k=1}^{\lfloor n/2 \rfloor} \binom{n}{k}^{-1} \sum_{|\boldsymbol{\gamma}|=k} I(\boldsymbol{X}^{\boldsymbol{\gamma}}; \boldsymbol{X}^n_{-\boldsymbol{\gamma}}) , \tag{8.18}$$

where $X^n_{-\boldsymbol{\gamma}}$ represents all the variables that are not in $\boldsymbol{\gamma}$, and $\lfloor \cdot \rfloor$ is the floor function. Motivated by this expression, let us introduce the quantity[10]

$$\begin{aligned} \Sigma(\boldsymbol{X}^n) &:= \mathrm{TC}(\boldsymbol{X}^n) + \mathrm{DTC}(\boldsymbol{X}^n) \\ &= \sum_{i=1} I(X_i; \boldsymbol{X}^n_{-i}) . \end{aligned} \tag{8.19}$$

By noting the similarities between Eqs. (8.18) and (8.19), together with the fact that $C_{\mathrm{TSE}}(\boldsymbol{X}^3) = \frac{1}{3}\left[C(\boldsymbol{X}^3) + \mathrm{DTC}(\boldsymbol{X}^3)\right]$, we can hypothesise that, qualitatively,

$$C_{\mathrm{TSE}}(\boldsymbol{X}^n) \propto \Sigma(\boldsymbol{X}^n) . \tag{8.20}$$

Monte Carlo simulations show that this approximation is justified: when evaluated on distributions $p_{\boldsymbol{X}^n}$ sampled uniformly at random from the probability simplex, the correlation between Σ and C_{TSE} is consistently above 0.97 (Fig. 8.4a). Moreover, Σ outperforms other proposed approximations of the TSE complexity.[11]

[10] The quantity TC + DTC has been introduced in the context of time series analysis as *local exogenous information*, and given the suitable nickname of "very mutual information" [151].

[11] In [3, Fig. 2] the DTC (under the name "interaction complexity") is proposed as a metric "related but not identical to neural complexity." Numerical evaluations show that the sum of TC and DTC, as proposed in (8.20), is a more accurate approximation for the TSE complexity (results not shown).

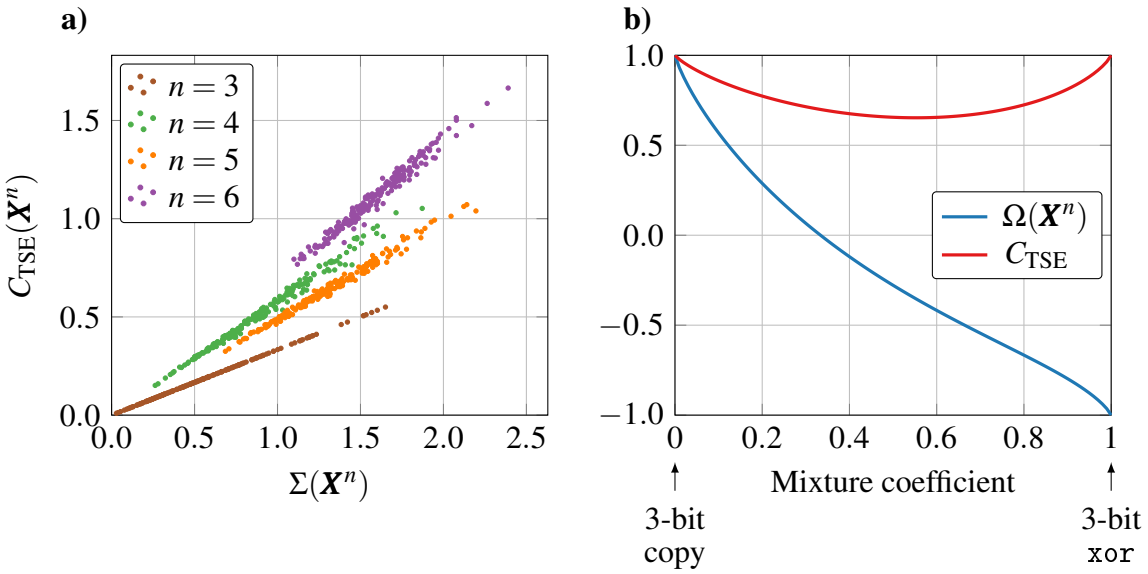


Figure 8.4: **(a)** The sum of the TC and DTC (denoted by Σ) is an accurate approximation of the TSE complexity. Each dot corresponds to probability distribution over n binary variables, which are sampled uniformly at random from the corresponding probability simplex. **(b)** C_{TSE} (upper line) and Ω (lower line) evaluated on a distribution resulting from a linear mixture between a copy (left) and an `xor` (right). Figure shows the case $n = 3$, but results are qualitatively similar for larger systems. Both of these results show that the TSE complexity conflates synergy and redundancy, and is better thought of as a measure of overall correlation strength.

Figure 8.4a and Eq. (8.20) suggest that the TSE complexity is large when either the shared randomness or the collective constraints are large. As a more direct example, we evaluate C_{TSE} in a distribution given by a linear mixture of the distributions of a 3-bit copy and a 3-bit `xor`, showing that C_{TSE} has exactly the same value in both extremes, and hence that it conflates redundancy with synergy (Fig. 8.4b).

Taken together, our results show that the TSE complexity is a good metric of overall integration between parts of the system, but it generally fails to discriminate high- from low-order phenomena. Overall, the fact that

$$\begin{aligned} \Omega &= \mathrm{TC} - \mathrm{DTC}\,, \\ C_{\mathrm{TSE}} &\propto \mathrm{TC} + \mathrm{DTC}\,, \end{aligned} \tag{8.21}$$

suggests that the TSE complexity and the O-information are complementary, corresponding to an insightful "change of basis" from an elementary constraints vs randomness representation. Effectively, while both TC and DTC provide two measures of roughly the same phenomenon (interdependency strength), Ω and C_{TSE} refer to different aspects: C_{TSE} gives an overarching account of the strength of the interdependencies within $\boldsymbol{X}^n$, and Ω indicates whether these correlations are predominantly redundant or synergistic.

8.5 Conclusion

We introduced the *O-information*, Ω, as the difference between strength of the collective constraints and the shared randomness in a multivariate system. We argued that Ω captures the net balance between synergy and redundancy, since (i) it is a sum of triple interaction informations, (ii) it is maximised (minimised) by an n-bit copy (`xor`), and (iii) it correlates with other notions of high-order effects from the statistical mechanics literature.

In the context of IIT, we propose the O-information as a modern, more principled analogue to neural complexity [2] – our results suggest that the O-information is a better characterisation of Tononi, Sporns and Edelman's insightful intuitions of coexisting local independence and global information sharing. In fact, we found that neural complexity does not measure statistical synergy as such, but total correlation strength. This suggests that Ω and TSE are complementary metrics: neural complexity gives an overarching account of the strength of the interdependencies within the system, and the O-information reveals whether these interdependencies are predominantly redundant or synergistic. We take this as a step towards a multi-dimensional framework that allows for a finer and more subtle taxonomy of complex systems.

In the next chapter we fast-forward from 1994 to 2008, and extend the PID principles to formulate a full-fledged decomposition of the integrated information measures in IIT 2.0.

Chapter 9

Integrated information decomposition

Chapter summary

We deepen the connection between information decomposition and IIT, by outlining a unified theory of Integrated Information Decomposition, ΦID. Most importantly, ΦID reveals that what is typically referred to as 'integration' is actually an aggregate of several heterogeneous phenomena, and can help us understand and alleviate the limitations of existing Φ measures. Additionally, we link ΦID with fundamental principles of causal emergence, providing theoretical support to our claims relating IIT and complexity.

9.1 Introduction

As a final theoretical contribution, in this chapter we repeat the procedure we presented in Chapter 8 with IIT 1.0 and C_{TSE}, this time with IIT 2.0 and Φ. Again, this will require us to reconsider the concepts of integration and differentiation, now in a dynamical setting.

According to the IIT line of reasoning, two systems are said to be integrated if they mutually affect each other's temporal evolution. From this perspective, it would seem that integration can be stronger or weaker, and hence could be represented by a single scalar number. This chapter shows that such a characterisation is incomplete, as integration can differ not only in quantity *but also in quality*. We present a framework to decompose different 'modes' of information dynamics, showing that even in small systems there are many different effects at play, including storage, copy, transfer, erasure, and upward and downward causation.

The key to this development is a substantial extension to PID, that allows us to formulate a *multi-target* information decomposition. Applying PID to a stochastic dynamical system setting, we consider the decomposition of the whole set of 'cause'- and 'effect'-type informational relationships, and obtain what we call the ***Integrated Information Decomposition***, ΦID.

This new framework enables a great deal of advancements in the study of IIT and information processing in complex systems: it sheds light on modes of information dynamics that have not been previously reported; it explains some of the alleged 'flaws' of early Φ measures; it allows us to extend Lizier's taxonomy with specific proposals for information modification; and it provides a suggestive link with basic principles of causal emergence.

9.2 Decomposing multivariate information

Consider two interdependent processes that are measured at regular time intervals. The *excess entropy* [169] of these processes, $\mathbf{E}$, is the total amount of (Shannon) information that is transferred through these processes from past to future, which is a well-known metric to assess dynamical complexity [88]. While $\mathbf{E}$ is in general hard to compute [170], for Markovian systems it simplifies to

$$\mathbf{E} = I(X_1, X_2;\, Y_1, Y_2)\ , \tag{9.1}$$

where $\boldsymbol{X} = \{X_1, X_2\}$ and $\boldsymbol{Y} = \{Y_1, Y_2\}$ denote the past and future state of the system respectively, and the subscript denotes variable index. We consider the decomposition of $\mathbf{E}$ into modes of information dynamics, focusing on systems with Markovian dynamics, leaving extensions to processes with memory for future work.

9.2.1 Forward and backward information decomposition

Our approach is to decompose $\mathbf{E}$ using principles of the PID framework [19], that also guided our formulation of Ω in Chapter 8. By focusing on how information flows from past to future, one can consider a *forward PID* that decomposes the information provided by the two past variables, X_1 and X_2, about the joint future (Y_1, Y_2) as

$$\begin{aligned}\mathbf{E} = &\,\mathtt{Red}(X_1, X_2; Y_1Y_2) + \mathtt{Un}(X_1; Y_1Y_2|X_2) \\ &+ \mathtt{Un}(X_2; Y_1Y_2|X_1) + \mathtt{Syn}(X_1, X_2; Y_1Y_2).\end{aligned}$$

Intuitively, $\mathtt{Red}(X_1, X_2; Y_1Y_2)$ corresponds to *redundant information* provided by both X_1 and X_2 about Y_1Y_2; $\mathtt{Un}(X_1; Y_1Y_2|X_2)$ (resp. $\mathtt{Un}(X_2; Y_1Y_2|X_1)$) refers to the *unique information* that only X_1 (resp. X_2) provides about Y_1Y_2; and finally, $\mathtt{Syn}(X_1, X_2; Y_1Y_2)$ accounts for the information that X_1 and X_2 provide about Y_1, Y_2 only when they are observed together, henceforth called *synergistic information*. Note that this synergistic information $\mathtt{Syn}(X_1, X_2; Y_1Y_2)$ is the ψ integrated information measure in Sec. 2.3.4 proposed by Griffith [9].

Similarly, an equivalent decomposition can be built by considering the information that Y_1 and Y_2 contain about the past state (X_1, X_2). Correspondingly, a *backward PID* is given by

$$\begin{aligned}\mathbf{E} = &\,\mathtt{Red}(Y_1, Y_2; X_1X_2) + \mathtt{Un}(Y_1; X_1X_2|Y_2) \\ &+ \mathtt{Un}(Y_2; X_1X_2|Y_1) + \mathtt{Syn}(Y_1, Y_2; X_1X_2).\end{aligned}$$

The forward and backward PID are related to the notions of *cause* (forward) and *effect* (backward) information in IIT. These two information decompositions provide complementary, but overlapping descriptions of the system's dynamics. The next section explains how they can be unified in a single and encompassing description.

9.3 Integrated information decomposition: ΦID

This section develops the mathematical framework of our contribution. The goal is to provide a decomposition of $\mathbf{E}$ similar to the two above, but that applies to both cause and effect information simultaneously. To do this, we solve PID's limitation of having only one single target variable, in order to allow for multi-target information decompositions. This decomposition will be then applied to the measures in Sec. 2.3 and discussed in the broader context of IIT.

9.3.1 Double-redundancy lattice

Let us begin by first considering the redundancy lattice [19], which is used in PID to formalise our intuitive understanding of redundancy. Let $\mathcal{A}$ be the collection given by

$$\mathcal{A} := \{\{1\},\{2\},\{1,2\},\{\{1\},\{2\}\}\}, \tag{9.2}$$

which are all the sets of subsets of $\{1,2\}$ where no element is contained in another.[1]

The elements of $\mathcal{A}$ have a natural (partial) order relationship: for $\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathcal{A}$, one says that $\boldsymbol{\alpha} \preceq \boldsymbol{\beta}$ if for all $b \in \boldsymbol{\beta}$ there exists $a \in \boldsymbol{\alpha}$ such that $a \subset b$ [19]. The lattice that encodes the relationship $\preceq$ is known as the redundancy lattice (Fig. 9.1), and guides the construction of the four terms in the PID.

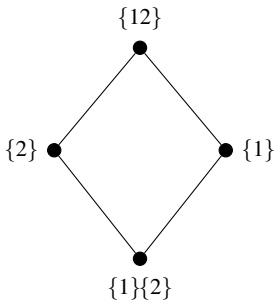


Figure 9.1: Lattice of nodes in $\mathcal{A}$ arranged according to the partial ordering $\preceq$.

Our first step is to build a *product lattice* over $\mathcal{A} \times \mathcal{A}$, in order to extend the notion of redundancy from PID to the case of multiple source and target variables (here X_1, X_2 and Y_1, Y_2 respectively). Extending Williams and Beer's [19] notation, we denote sets of sources and targets using their indices only, with an arrow going from past to future.[2] Hence, the nodes of the product lattice are denoted as $\boldsymbol{\alpha} \to \boldsymbol{\beta}$ for $\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathcal{A}$, and a partial ordering relationship among them is defined by

$$\boldsymbol{\alpha} \to \boldsymbol{\beta} \preceq \boldsymbol{\alpha}' \to \boldsymbol{\beta}' \quad \text{iff} \quad \boldsymbol{\alpha} \preceq \boldsymbol{\alpha}' \text{ and } \boldsymbol{\beta} \preceq \boldsymbol{\beta}'. \tag{9.3}$$

This relationship guarantees a lattice structure with 16 nodes, which is shown in Figure 9.2 and proven in Appendix C.1. An intuitive understanding of the product lattice is developed in the sections below.

[1] In a general N-variable case, $\mathcal{A}$ is the set of antichains on the lattice $(\mathcal{P}(\{1,\ldots,N\}),\subseteq)$, discussed in Ref. [19]. We focus on the bivariate case for clarity, although our results hold for any N.

[2] Note that although we use an arrow and we talk about past and future, this formalism is symmetric in sources and targets.

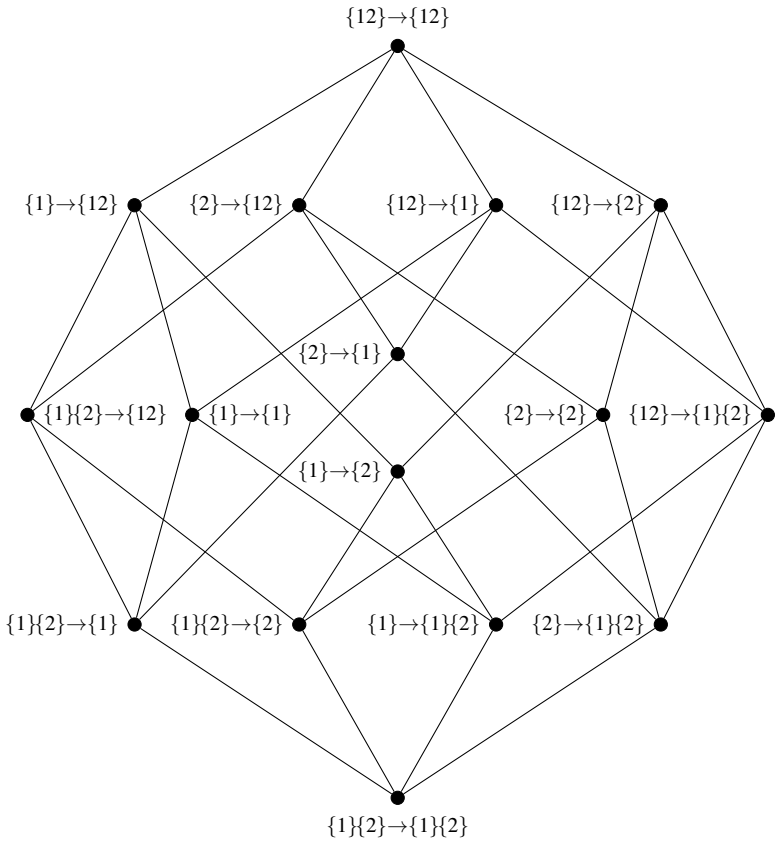


Figure 9.2: The double-redundancy lattice for two predictors and two targets, which is the product of two lattices as shown in Figure 9.1.

9.3.2 Redundancies and atoms

The next ingredient in the PID recipe is a *redundancy function*, $I_\cap$, that quantifies the 'overlapping' information about the target that is common to a set of sources $\boldsymbol{\alpha} \in \mathcal{A}$ [19]. Intuitively, $I_\cap^{\{1\}\{2\}}$ is the information about the target that is in either source, $I_\cap^{\{i\}}$ the information in source i, and $I_\cap^{\{12\}}$ the information that is in both sources together. The partial ordering relation between these quantities is the basis for the lattice in Fig. 9.1.

In this subsection, we extend the notion of overlapping information to the multi-target setting. The key to do so is to define certain axioms that a multi-target information decomposition must satisfy, and that inherit from the corresponding axioms in standard PID.

For a given $\boldsymbol{\alpha} \to \boldsymbol{\beta} \in \mathcal{A} \times \mathcal{A}$, the overlapping information that is common to sources $\boldsymbol{\alpha}$ and can be seen in targets $\boldsymbol{\beta}$ is denoted as $I_\cap^{\boldsymbol{\alpha}\to\boldsymbol{\beta}}$ and referred to as the *double-redundancy function*. In the following, we assume that the double-redundancy function satisfies two axioms:

- **Axiom 1 (compatibility)**: if $\boldsymbol{\alpha} = \{\alpha_1, ..., \alpha_J\}$ and $\boldsymbol{\beta} = \{\beta_1, ..., \beta_K\}$ with $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{A}$ and α_j, β_k non-empty subsets of $\{1, ..., N\}$, then the following cases can be reduced to the redundancy of PID or the mutual information:[3]

$$I_\cap^{\boldsymbol{\alpha}\to\boldsymbol{\beta}} = \begin{cases} \mathtt{Red}(\boldsymbol{X}^{\alpha_1}, ..., \boldsymbol{X}^{\alpha_J}; \boldsymbol{Y}^{\beta_1}) & \text{if} \quad K = 1, \\ \mathtt{Red}(\boldsymbol{Y}^{\beta_1}, ..., \boldsymbol{Y}^{\beta_K}; \boldsymbol{X}^{\alpha_1}) & \text{if} \quad J = 1, \\ I(\boldsymbol{X}^{\alpha_1}; \boldsymbol{Y}^{\beta_1}) & \text{if} \quad J = K = 1. \end{cases}$$

- **Axiom 2 (partial ordering)**: if $\boldsymbol{\alpha} \to \boldsymbol{\beta} \preceq \boldsymbol{\alpha}' \to \boldsymbol{\beta}'$ then $I_\cap^{\boldsymbol{\alpha}\to\boldsymbol{\beta}} \leq I_\cap^{\boldsymbol{\alpha}'\to\boldsymbol{\beta}'}$.

By exploiting these axioms, one can define 'atoms' that belong to each of the nodes via the Moebius inversion formula. Concretely, the *integrated information atoms* $I_\partial^{\boldsymbol{\alpha}\to\boldsymbol{\beta}}$ are defined as the quantities that guarantee the following condition for all $\boldsymbol{\alpha} \to \boldsymbol{\beta} \in \mathcal{A} \times \mathcal{A}$:

$$I_\cap^{\boldsymbol{\alpha}\to\boldsymbol{\beta}} = \sum_{\boldsymbol{\alpha}'\to\boldsymbol{\beta}' \preceq \boldsymbol{\alpha}\to\boldsymbol{\beta}} I_\partial^{\boldsymbol{\alpha}'\to\boldsymbol{\beta}'} \,. \tag{9.4}$$

In other words, $I_\partial^{\boldsymbol{\alpha}\to\boldsymbol{\beta}}$ corresponds to the information contained in node $\boldsymbol{\alpha} \to \boldsymbol{\beta}$ and not in any node below it in the lattice. These are analogues to the redundant, unique, and synergistic atoms in the forward and backward PID above, but using the product lattice as a scaffold. By inverting this relationship, one can find a recursive expression for calculating I_∂ as

$$I_\partial^{\boldsymbol{\alpha}\to\boldsymbol{\beta}} = I_\cap^{\boldsymbol{\alpha}\to\boldsymbol{\beta}} - \sum_{\boldsymbol{\alpha}'\to\boldsymbol{\beta}' \prec \boldsymbol{\alpha}\to\boldsymbol{\beta}} I_\partial^{\boldsymbol{\alpha}'\to\boldsymbol{\beta}'} \,. \tag{9.5}$$

With all the tools at hand, we can deliver the promised decomposition of $\mathbf{E}$ in terms of atoms of integrated information.

Definition 3. *The Integrated Information Decomposition (ΦID) of a system with Markovian dynamics is the collection of atoms I_∂ defined from the redundancies $I_\cap$ via Eq. (9.5), which satisfy*

$$\mathbf{E} = I(\boldsymbol{X}; \boldsymbol{Y}) = \sum_{\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathcal{A}} I_\partial^{\boldsymbol{\alpha}\to\boldsymbol{\beta}} \,. \tag{9.6}$$

It is direct to see that ΦID of two time series gives 16 atoms that correspond to the lattice shown in Figure 9.2, which are computed from a linear transformation over the 16 redundancies. Interestingly, Axioms 1 and 2 allow us to compute all the $I_\cap$ terms once a

[3] We use the shorthand notation $\boldsymbol{X}^\alpha = (X_{i_1}, ..., X_{i_K})$ for $\alpha = \{i_1, ..., i_K\}$.

single-target PID redundancy function $\mathtt{Red}(\cdot)$ has been chosen, with the sole exception of $I_\cap^{\{1\}\{2\}\to\{1\}\{2\}}$. All this is summarised in the following result.

Proposition 3. *Axioms 1 and 2 provide unique values for the 16 atoms of the product lattice (see Figure 9.2) after one defines (i) a single-target redundancy function $\mathtt{Red}(\cdot)$, and (ii) an expression for $I_\partial^{\{1\}\{2\}\to\{1\}\{2\}}$.*

In the same way as in PID the definition of $\mathtt{Red}(\cdot)$ gives 3 other terms (unique and synergy) as side-product, Proposition 3 shows that in ΦID the addition of the double-redundancy function $I_\partial^{\{1\}\{2\}\to\{1\}\{2\}}$ gives 15 other terms for free.[4]

Throughout the rest of the article we outline how ΦID can be used to revise theories of information dynamics and integrated information, and how it can provide more detailed analyses of systems of interest.

9.3.3 Simple examples

To start developing our intuition about the ΦID atoms, let us decompose the mutual information between the present of one variable, X_i, and its own future, Y_i, i.e. the information storage in variable i [69]:

$$\begin{aligned} I(X_i;Y_i) = I_\partial^{\{1\}\{2\}\to\{1\}\{2\}} + I_\partial^{\{1\}\{2\}\to\{i\}} \\ + I_\partial^{\{i\}\to\{1\}\{2\}} + I_\partial^{\{i\}\to\{i\}} \,. \end{aligned} \tag{9.7}$$

Here, $I_\partial^{\{1\}\{2\}\to\{1\}\{2\}}$ corresponds to redundant information in the sources that is present in both targets; $I_\partial^{\{1\}\{2\}\to\{i\}}$ is the redundant information in the sources that is eliminated from the j-th source ($j \neq i$) and hence is only conserved in Y_i; and similarly for the remaining atoms.

As another example, consider the transfer entropy from i to j (with $i \neq j$):

$$\begin{aligned} I(X_i;Y_j|X_j) = I_\partial^{\{12\}\to\{1\}\{2\}} + I_\partial^{\{12\}\to\{j\}} \\ + I_\partial^{\{i\}\to\{1\}\{2\}} + I_\partial^{\{i\}\to\{j\}} \,. \end{aligned} \tag{9.8}$$

As before, $I_\partial^{\{12\}\to\{1\}\{2\}}$ is the synergistic information present in the joint past (X_1, X_2) that can be read through either Y_1 or Y_2, and similarly for the rest of the terms.

In the following section we explore the possibilities offered by this decomposition, and its implications for causal analysis, IIT, and complex systems in general.

[4] Note that our framework does not prescribe a particular formula for $I_\partial^{\{1\}\{2\}\to\{1\}\{2\}}$. A discussion on this issue can be found in the supplementary material.

9.4 Results

9.4.1 Limitations of conventional causal discovery methods

Mutual information and transfer entropy (or linear variants of them, to which our conclusions also apply) are the building blocks of most popular methods of statistical causal discovery. We now show that these metrics have two kinds of limitations: they conflate multiple effects in counterintuitive ways, and they fail to capture some effects altogether.

First, let us focus on the decomposition of information storage in Eq. (9.7). Note that, although X_2, Y_2 are not in this mutual information, $I(X_1;Y_1)$ shares the term $I_\partial^{\{1\}\{2\}\to\{1\}\{2\}}$ with $I(X_2;Y_2)$ by virtue of them being considered part of the same multivariate stochastic process. Therefore, if one uses simple mutual information as a measure of storage one may include information that is not stored exclusively in a given variable, which may lead to paradoxical conclusions such as the sum of individual storages being greater than $\mathbf{E}$.

Next, consider the terms in the decomposition of transfer entropy in Eq. (9.8). Note that, of these, $I_\partial^{\{i\}\to\{j\}}$ is the only 'genuine' transfer term – all others correspond to redundant or synergistic effects involving both variables in past or future. Furthermore, one of the 'extra' terms ($I_\partial^{\{i\}\to\{1\}\{2\}}$) is shared with $I(X_i;Y_i)$, in a somewhat counterintuitive overlap between storage and transfer. Similar concerns have been discussed in the literature [171], showing that transfer entropy *per se* cannot be taken as a pure measure of information transfer.

Finally, from the decompositions of the mutual information and conditional mutual information as shown above, it is clear that none of these quantities are able to capture the ΦID terms of the form $I_\partial^{\alpha\to\{12\}}$. These terms correspond to 'synergistic effects' (i.e. causes whose effects only manifest on groups, rather than individual variables) and are neglected by standard causal discovery methods.

9.4.2 Information processing in complex systems

Based on ΦID, and building on Lizier's work [69], we propose an extended taxonomy of information dynamics, with 6 disjoint and qualitatively distinct phenomena:

Storage: Information that remains in the same source set, even if it includes collective effects. Includes $I_\partial^{\{1\}\{2\}\to\{1\}\{2\}}$, $I_\partial^{\{1\}\to\{1\}}$, $I_\partial^{\{2\}\to\{2\}}$, and $I_\partial^{\{12\}\to\{12\}}$.

Copy: Information that becomes duplicated. Includes $I_\partial^{\{1\}\to\{1\}\{2\}}$, and $I_\partial^{\{2\}\to\{1\}\{2\}}$.

Transfer: Information that moves between variables. Includes $I_\partial^{\{1\}\to\{2\}}$, and $I_\partial^{\{2\}\to\{1\}}$.

Erasure: Duplicated information that is pruned. Includes $I_\partial^{\{1\}\{2\}\to\{1\}}$, and $I_\partial^{\{1\}\{2\}\to\{2\}}$.

Downward causation: Collective properties that define individual futures. Includes $I_\partial^{\{12\}\to\{1\}}$, $I_\partial^{\{12\}\to\{2\}}$, and $I_\partial^{\{12\}\to\{1\}\{2\}}$.

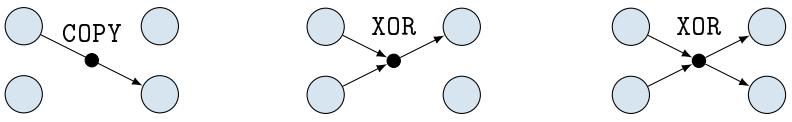
Upward causation: Collective properties that are defined by individuals. Includes $I_\partial^{\{1\}\to\{12\}}$, $I_\partial^{\{2\}\to\{12\}}$, and $I_\partial^{\{1\}\{2\}\to\{12\}}$.

While downward causation has been discussed in the past [171], upward causation and synergistic storage ($I_\partial^{\{12\}\to\{12\}}$) have not been reported in the literature. This revised taxonomy leads to less ambiguous, quantifiable descriptions of information dynamics in complex systems, in addition to grounding abstract concepts such as upward and downward causation, as well as notions like integrated information.

9.4.3 Different types of integration

One important conceptual result of our framework is that there are multiple qualitatively different ways in which a multivariate dynamical process can integrate information through combinations of redundant, unique, or synergistic effects. As elementary examples, consider the following systems of 2 binary variables:

- A **copy transfer** system, in which x_1, x_2, y_1 are i.i.d. fair coin flips, and $y_2 = x_1$ (i.e. one bit is shifted).
- The **downward XOR**, in which x_1, x_2, y_2 are independent identically distributed fair coin flips, and $y_1 \equiv x_1 + x_2 \pmod 2$.
- The **parity-preserving random** (PPR), in which x_1, x_2 are i.i.d. fair coin flips, and $x_1 + x_2 \equiv y_1 + y_2 \pmod 2$ (i.e. $\boldsymbol{y}$ is a random string of the same parity as $\boldsymbol{x}$).



$\Phi = 1$ $\quad$ $\Phi = 1$ $\quad$ $\Phi = 1$

$I_\partial^{\{1\}\to\{2\}} = 1$ $\quad$ $I_\partial^{\{12\}\to\{1\}} = 1$ $\quad$ $I_\partial^{\{12\}\to\{12\}} = 1$

Figure 9.3: Example systems of logic gates. All of them have the same integrated information, but their information dynamics are different. This difference is captured by a full ΦID decomposition, that shows the only non-zero atoms are transfer (left), downward causation (centre), and synergistic storage (right).

These three systems (Fig. 9.3) are 'equally integrated,' in the sense that the dynamics of the whole cannot be perfectly predicted from the parts alone and the integrated information measure $\Phi = 1$ for all of them [172, 8]. However, they integrate information in qualitatively different ways: in effect, the integration in the copy system is entirely due to transfer dynamics ($I_\partial^{\{1\}\to\{2\}}$); the downward XOR integrates information due to downward causation ($I_\partial^{\{12\}\to\{1\}}$); and PPR due to synergistic storage ($I_\partial^{\{12\}\to\{12\}}$). All the other ΦID atoms in each of these systems are zero (proofs in Appendix C.5).

9.4.4 Measures of integrated information

In Chapter 6 we showed that many of the different measures of integrated information that have been proposed in the literature behave inconsistently, even in very simple systems, but we did not explain *why* this is the case. In this subsection we use ΦID to dissect and compare four of the measures of integrated information (Φ, ψ, Φ_G) and dynamical complexity (CD) introduced earlier in this thesis (c.f. Chapter 2).

Importantly, not all of these measures are able to detect the presence of non-trivial statistical structure in the previously mentioned systems. In fact, although all measures detect the integration in the copy transfer system, CD and even more recent measures like Φ^* and Φ_G [11, 10] are zero for the PPR system above – i.e. they do not detect all synergistic effects. Of the currently available measures, only Φ and Griffith's ψ [9] are able to capture the integration in the PPR system.

As a systematic exploration, one can determine which measures are sensitive to which modes of information dynamics by calculating whether each measure is zero, positive, or negative for a system consisting of only one particular ΦID atom (Table 9.1; proofs in Appendix C.5). The main result is that each measure captures a different combination of ΦID atoms: although generally most of them capture synergistic effects and avoid (or penalise) redundant effects, they differ substantially. In particular, they differ most in their treatment of atoms that include synergistic effects, $I_\partial^{\alpha\to\{12\}}$, which had not been previously reported in the literature.

The key conclusion is that these measures are not simply different approximations of a unique concept of integration, but that they are capturing intrinsically different aspects of the system's information dynamics. While aggregate measures like these can be empirically useful, it is important to remember that they are measuring different combinations of different effects within the system's information dynamics. Echoing the conclusions of Chapter 6: these measures behave differently not only in practice, but also *in principle*.

Table 9.1: Sensitivity of integrated information measures to ΦID atoms. For each measure, entries indicate whether the value is positive (+), negative (-) or 0 in a system in which the given ΦID atom is the only non-zero atom.

ΦID atoms	Measures			
	Φ	CD	ψ	Φ_G
$I_\partial^{\{1\}\{2\}\to\{1\}\{2\}}$	-	0	0	0
$I_\partial^{\{1\}\{2\}\to\{i\}}$	0	0	0	0
$I_\partial^{\{1\}\{2\}\to\{12\}}$	+	0	0	0
$I_\partial^{\{i\}\to\{1\}\{2\}}$	0	+	0	+
$I_\partial^{\{i\}\to\{i\}}$	0	0	0	0
$I_\partial^{\{i\}\to\{j\}}$	+	+	0	+
$I_\partial^{\{i\}\to\{12\}}$	+	0	0	0
$I_\partial^{\{12\}\to\{1\}\{2\}}$	+	+	+	+
$I_\partial^{\{12\}\to\{i\}}$	+	+	+	+
$I_\partial^{\{12\}\to\{12\}}$	+	0	+	0

9.4.5 Why whole-minus-sum Φ can be negative

The ΦID framework can be further leveraged to provide elegant explanations of certain behaviours of integrated information and dynamical complexity measures. For example, whole-minus-sum Φ (Sec. 2.3.2), which is calculated as

$$\Phi = I(X_1, X_2; Y_1, Y_2) - I(X_1; Y_1) - I(X_2; Y_2) \tag{9.9}$$

for a bivariate process, can sometimes take negative values. This feature, which has been used as an argument to discard Φ as a suitable measure of integrated information [9, 11], can be explained as follows. By applying the decomposition in Eq. (9.4), one finds that

$$\begin{aligned}
\Phi = \quad & -I_\partial^{\{1\}\{2\}\to\{1\}\{2\}} && \Big\} \text{Red} \\
& +\mathtt{Syn}(X_1, X_2; Y_1 Y_2) + I_\partial^{\{1\}\{2\}\to\{12\}} && \Big\} \text{Syn} \\
& +I_\partial^{\{1\}\to\{12\}} + I_\partial^{\{2\}\to\{12\}} && \\
& +I_\partial^{\{1\}\to\{2\}} + I_\partial^{\{2\}\to\{1\}} . && \Big\} \text{Un}
\end{aligned}$$

Hence, Φ accounts for all the synergies in the system (the seven terms in Fig. 9.2 with $\{12\}$ in either side), the unique information transferred between parts of the system, and, importantly, the negative of the bottom node of the ΦID lattice. The presence of this negative

double-redundancy term shows that in highly redundant systems Φ can be negative. This is akin to Williams and Beer's [19] explanation of the negativity of the interaction information, applied to multivariate processes. Based on this insight, one can formulate a 'corrected' Φ by adding back the double-redundancy:

$$\Phi^{c} := \Phi + I_{\partial}^{\{1\}\{2\}\to\{1\}\{2\}} ,$$

which now includes only synergistic and unique transfer terms.

We computed Φ^{c} numerically for a simple example, using an extension of the PID presented by James et al. [173]. Mimicking the setting in Fig. 6.1 but with discrete variables, let us consider a system in which y_1, y_2 are noisy AND gates of x_1, x_2 and the correlation between the noise components of y_1 and y_2 is a free parameter. We calculated Φ and Φ^{c} with respect to the system's stationary distribution. Plots of the standard and corrected Φ for this system are shown in Fig, 9.4, and details of the computation can be found in Appendix C.6.

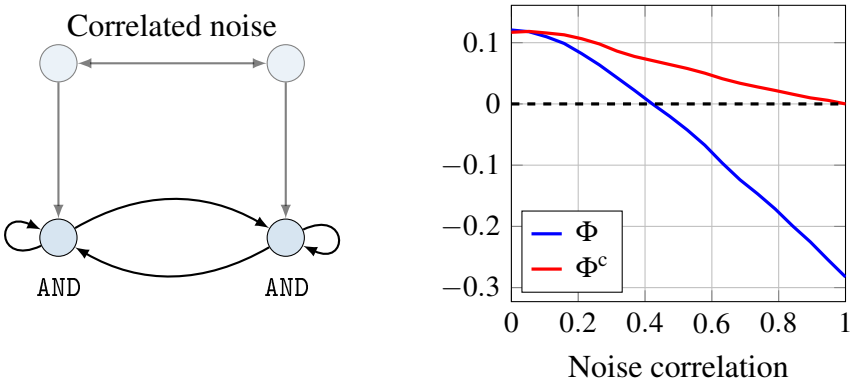


Figure 9.4: Standard and corrected Φ in a two component noisy AND system with varying correlation in the noise to each component. As the correlation increases, Φ drops below zero reflecting an increase in redundancy. Adding back the double-redundancy term results in a non-negative measure that, as expected, vanishes to zero in the fully correlated system.

As expected, Φ drops below zero as synergy decreases and redundancy increases with noise correlation, replicating the results in Fig. 6.1. Interestingly, however, after adding the double-redundancy term, the corrected version Φ^{c} tends to zero for high noise correlation, which is more similar to some of the other measures highlighted in Chapter 6, e.g. CD and Φ^{*}. This example shows that using ΦID we can formulate new measures of integrated information, that capture specific sets of information effects and that may have different properties to suit the target application.

9.4.6 Why unnormalised causal density can exceed TDMI

In Oizumi et al. [11], the authors correctly point out that the sum of conditional pairwise transfer entropies (or unnormalised Causal Density; uCD) in a system can exceed the total mutual information $I(\boldsymbol{X};\boldsymbol{Y})$, which is problematic for considering this as a measure of integrated information [11].[5] This quantity, given by

$$\begin{aligned} \text{uCD} &= \text{TE}_{Z_1 \to Z_2} + \text{TE}_{Z_2 \to Z_1} \\ &= I(X_1; Y_2 | X_2) + I(X_2; Y_1 | X_1) \ , \end{aligned} \tag{9.10}$$

can also be readily decomposed using ΦID. By applying Eq. (9.4) to the expression of uCD, we find that

$$\begin{aligned} \text{uCD} = &\texttt{Un}(X_1; Y_2 | X_2) + I_\partial^{\{12\}\to\{2\}} \\ &+ \texttt{Un}(X_2; Y_1 | X_1) + I_\partial^{\{12\}\to\{1\}} \\ &+ 2I_\partial^{\{12\}\to\{1\}\{2\}} \ . \end{aligned}$$

Besides the unique and synergistic terms that one would expect in a measure of information transfer [174], there is in addition a *double-counting* of a downward causation ΦID atom, $I_\partial^{\{12\}\to\{1\}\{2\}}$. Specifically, uCD double-counts synergistic information in the past that is transferred redundantly to the future, and this can cause uCD to be greater than $I(X_1, X_2; Y_1, Y_2)$. This is because in the calculation of both transfer entropies the targets are considered independent of each other.

This finding makes it straightforward to design systems for which uCD is maximal, highlighting this effect (i.e. a system that has only $I_\partial^{\{12\}\to\{1\}\{2\}} > 0$): for example, a system where x_1, x_2 are independent fair coin flips and $y_1 = y_2 = x_1 \oplus x_2$. Indeed, for this system uCD $= 2$ bit $> I(X_1, X_2; Y_1, Y_2) = 1$ bit. Hypothetically, if one wanted to devise a uCD-based measure of integrated information that is upper-bounded by $I(\boldsymbol{X};\boldsymbol{Y})$, we could simply subtract $I_\partial^{\{12\}\to\{1\}\{2\}}$ from Eq. (9.10), as we did earlier with Φ and $I_\partial^{\{1\}\{2\}\to\{1\}\{2\}}$ in Figure 9.4.

Furthermore, this decomposition also shows that there are many common atoms in the ΦID expansions of CD and Φ^{WMS}, which might explain why CD has sometimes been considered together with measures of integrated information [12, 172].

[5] Note that the original definition of causal density is normalised by $L(L-1)$, and has been proven to be bounded by mutual information (Appendix A.3).

9.5 Formalising causal emergence through ΦID

Perhaps the most fundamental theoretical problem in complexity science is that of providing a rigorous theory of *causal emergence*. The philosophy literature, for example, is rife with proposals for how to taxonomise and reason about emergence. One standard approach is to distinguish between two types of phenomena:

- *strong emergence*, that corresponds to the somewhat paradoxical case of supervenient properties with irreducible causal power; and
- *weak emergence*, that occurs when a collective property cannot be derived from the interactions between elements at lower levels via explanatory shortcuts, but only via exhaustive simulation [175].[6]

While the idea of weak emergence is usually accepted [176], strong emergence is widely considered to be either impossible, or "uncomfortably like magic" [175]. This judgement is not fully unfounded – a property that is simultaneously supervenient (i.e. that can be computed from the microstate of the system), and that has irreducible causal power (i.e. that "tells us something" that the microstate doesn't) can indeed seem paradoxical. Nonetheless, as we show here this paradox is no more than a lack of imagination, and a lack of familiarity with the (admittedly counterintuitive) laws of multivariate information dynamics.

Inspired by Seth [177], we present a practically useful and ontologically innocent framework to study causal emergence based on ΦID. We take the perspective of an experimentalist who has no prior knowledge of the phenomenon of interest, but has sufficient data to generate an accurate statistical description of it. We put forward a definition of causal emergence which is logically and statistically consistent, and provides a practical tool to test (and possibly reject) hypotheses about emergence in scenarios of interest, based purely on data. Here we provide a brief outline of the theory, and leave the details for a future publication.

Let us begin directly by stating our proposed definition of of causal emergence:

Definition 4. *For a system with past state $\boldsymbol{X}$ and future state $\boldsymbol{Y}$, a supervenient feature $V = f(\boldsymbol{X})$ is said to exhibit statistical causal emergence iff*

$$\mathtt{Un}(V;\boldsymbol{Y}|X_1,...,X_n) > 0. \tag{9.11}$$

This definition directly represents the intuitions behind previous formulations of emergence – Eq. (9.11) implies that V predicts something about the future state of the system that cannot be predicted from any single subsystem. Note that this requires us to provide a definition the unique information given a *set* of variables – which is presented, along with all other proofs for this section, in Appendix C.7.

[6] This is strongly related to Wolfram's concept of *computational irreducibility* [86].

Based on Definition 4, we can use PID to derive a criterion for emergence that does not depend on our knowledge of the emergent feature V:

Proposition 4. *A system has a casually emergent feature iff* $\texttt{Syn}(X_1,...,X_n;\boldsymbol{Y}) > 0$.

The above result has an interesting consequence: by inspecting the system's statistics we can determine whether or not it has any emergent properties, *without knowing what those properties are*. With this feature-agnostic criterion of emergence, a direct calculation with ΦID allows us to decompose this synergy into two terms,

$$\texttt{Syn}(X_1,...,X_n;\boldsymbol{Y}) := \mathcal{D}(\boldsymbol{X}_t) + \mathcal{G}(\boldsymbol{X}_t) \ , \tag{9.12}$$

where $\mathcal{D}$ and $\mathcal{G}$ denote *downward causation* and *causal decoupling* respectively, and are essentially generalisations of the concepts of downward causation and synergistic storage introduced for $n = 2$ in Sec. 9.4.2. These two quantities satisfy the following propositions:

Proposition 5. *A system has downward causation* $\mathcal{D}(\boldsymbol{X}_t) > 0$ *iff there exists an emergent feature* V *such that* $\texttt{Un}(V;Y_k|X_1,...,X_n) > 0$ *for some* $k \in \{1,...,n\}$.

Proposition 6. *A system has causal decoupling* $\mathcal{G}(\boldsymbol{X}_t) > 0$ *if there exists an emergent feature* V *such that* $\texttt{Un}(V;V'|X_1,...,X_n) > 0$, *where* $V = f(\boldsymbol{X})$ *and* $V' = f(\boldsymbol{Y})$.

Matching our intuitions, a feature with downward causation is a collective property with a sort of "enslaving" effect on single agents. Perhaps most interestingly, a causally decoupled feature is one that has "a life of its own" – a sort of *statistical ghost*, that perpetuates itself over time without any individual agent influencing or being influenced by it.

This proposal, while theoretically appealing, suffers from all the challenges of PID, including the estimation of joint probabilities over many variables and the computation of the ΦID atoms themselves. As an alternative, let us introduce the quantity

$$\Psi(V) := I(V;V') - \sum_j I(X_j;V') \ , \tag{9.13}$$

as a general measure of information about the dynamics of V that cannot be accounted for by individual agents. Using this quantity we can formulate a practical criterion to detect statistical causal emergence:

Proposition 7. $\Psi(V) > 0$ *is a sufficient condition for* V *to be causally emergent.*

Therefore, although calculating whether a system has any emergent feature (with Proposition 4) may be computationally difficult, if we have a candidate feature V we believe may be emergent, we can compute the simple quantity in Eq. (9.13), that depends only on standard mutual information and pairwise marginals.

9.6 Discussion

We propose ΦID as a novel information-theoretic framework to study high-order interactions in time-series data. By unifying integrated information theory (IIT) and partial information decomposition (PID), the ΦID framework allows us to decompose information flow in a multivariate stochastic process into interpretable, disjoint parts. This allows systematic studies of unexplored modes of information dynamics – including modes of synergistic storage, and upward and downward causation – in a purely data-driven fashion.

9.6.1 Towards multi-dimensional measures of complexity

Besides the importance of having an encompassing taxonomy of information dynamics phenomena, this frameworks suggests, following Feldman and Crutchfield [178], that there is no theoretical basis to a purported all-encompassing scalar measure of dynamical complexity. The richness of complex dynamics is vast, and the prospect of subsuming all into a single number is unreasonable. Scalar measures might still have great practical value in certain experimental or theoretical contexts;[7] nevertheless, a general theory of complex systems (biological or otherwise) cannot be reduced to a single, one-size-fits-all measure.

Instead, we argue that the real strength of a complexity measure is in its specificity – what is it measuring, and how? A well-determined measure captures one specific aspect of a system's dynamics, and it does so clearly and explicitly. In this sense, our framework provides a solid basis upon which to design principled measures of complexity, and can allow us to disentangle and understand other measures.

9.6.2 Integration measures conflate transfer and synergy

Using ΦID, one is able to inspect previous measures of integrated information, explaining similarities and differences between them, and fixing some of their shortcomings. Most importantly, we have shown that what is usually refer to as 'integration' is in fact an aggregate of several different information effects, typically including transfer and synergy phenomena. Moreover, different measures capture different effects in various proportions, which explains the heterogeneity among existing measures reported in Chapter 6. By employing ΦID one can tailor measures for targeting specific mixtures of effects, according to the information processes one wishes to analyse.

[7] For example, measures that accurately discriminate between neural configurations corresponding to conscious and unconscious states in a particular experimental paradigm [125].

9.6.3 Causal analysis

As presented, ΦID is a generic tool to decompose multivariate mutual information, which can be directly used to perform causal analysis. Most integrated information measures can be roughly divided between those that describe integration in a system based on its causal properties [6], and those that use the system's attractor statistics, known as *dynamical* integration measures [172].[8] Given a system's conditional probability distribution $p(\boldsymbol{Y}|\boldsymbol{X})$, one can use ΦID to perform either a causal or a dynamical analysis by using the stationary attractor distribution $p(\boldsymbol{X})$, or a maximum entropy distribution on $\boldsymbol{X}$. However, note that a few additional assumptions need to hold to interpret the results in a strict causal sense; in particular, the conditional distribution $p(\boldsymbol{Y}|\boldsymbol{X})$ needs to be equivalent to a `do()` distribution in Pearl's sense [179], and the system must satisfy the faithfulness and causal Markov conditions.[9]

9.6.4 Limitations and future extensions

Naturally, our method inherits some of the limitations of PID. In particular, there is still no consensus on which of the various proposed decompositions is preferable for performing numerical evaluations of the PID atoms [173].

Given the long-standing challenges and the formidable mathematical difficulty of PID, it is natural to think that a consensus on ΦID is unlikely to be achieved in the near future. Yet, just as the theory of power series becomes simpler when considered over the (more complicated) complex plane, it is our hope that the dynamical insights and challenges introduced by the ΦID framework might serve to provide new intuitions to tackle the current limitations of PID.

9.6.5 Integrated information as a universal signature of emergence

In Chapters 3-5 we presented examples of integrated information in complex systems, and made a *pragmatic* case that, empirically, it matches our intuitions on what constitutes complex, emergent behaviour.

In this chapter, we have outlined a rigorous theory of causal emergence, both in the aggregate (Proposition 4), and in terms of decomposed downward and decoupled causation (Propositions 5 and 6). Importantly, we were able to identify these concepts with specific synergistic atoms in ΦID, which are the ones captured by (at least some of) the integrated

[8] Intuitively, a causal analysis reveals what the system *could do*, while a dynamical analysis based on attractor statistics reveals what the system *actually does*.

[9] Also, note that the maximum entropy distributions employed by some causal integration frameworks are well-defined for discrete Markovian systems, but in general may not always exist [180].

information measures introduced in Chapter 2. Therefore, those measures constitute viable tools to detect causally emergent dynamics in general complex systems.

In particular, we highlight the original whole-minus-sum Φ is especially attractive for this purpose, since includes all relevant ΦID atoms (Table 9.1), and, unlike PID-based measures like ψ, it does not depend on the specifics of a PID measure, allowing us to establish sufficiency conditions using standard information-theoretic tools.

In short, the decomposition of Φ, and the formulation of a rigorous theory of causal emergence, provide theoretical support to our argument for *integrated information as a universal signature of causal emergence*.

Chapter 10

Consciousness and information content: The entropic brain hypothesis

Chapter summary

As an alternative to IIT, we consider another, much simpler informational theory of consciousness known as the Entropic Brain Hypothesis (EBH). We present two examples from the study of altered states of consciousness – musical improvisation and the psychedelic state – and interpret the results in the light of EBH. We argue that the simplicity and empirical success of EBH provide valuable lessons for IIT, and that a collective, open-minded engagement between these and other theories are key for a cohesive, mature, and productive science of consciousness.

10.1 Introduction

In Chapter 7 we argued that, although theoretically appealing, the predictions of IITC have proven difficult to evaluate using Φ measures on neuroimaging data. These difficulties arise from multiple fronts: computational challenges associated with estimating Φ measures in high-dimensional spaces, unpredictable behaviour of Φ measures in partially observed systems, and possible confounding factors in the dynamics of conscious and unconscious brains.

As an alternative to IITC, in this chapter we consider a much simpler theory of consciousness, the *Entropic Brain Hypothesis* (EBH) [181]. The EBH puts forward a bold proposition: that richness of psychological content is correlated with richness of signal diversity, or information content, in the neural activity. This "signal diversity" can be operationalised using standard entropy estimators, making the EBH predictions straightforward to test on neurophysiological data.

In the study of altered states of consciousness, the predictions of EBH clearly align with our intuitions and with experimental results: brain entropy is lower in NREM sleep [137] and in disorders of consciousness like vegetative and minimally conscious states [125]; and it is an accurate marker of epileptic seizures [182] and depth of anaesthesia [183]. Furthermore, entropy is dramatically increased under the effects of psychedelic drugs [129], and, empirically, is an unreasonably effective tool – results are robust to different data preprocessing and filtering methods, as well as to different types of surrogate data controls [137]. Overall, EBH is a simple, interpretable theory with clear predictions, and very successful experimental results in agreed-upon paradigmatic cases.

In this chapter we present two brief examples of analyses conducted and interpreted following the EBH. In the first study, we show that brain entropy is increased with stronger external stimulation, and that, in agreement with subjective reports, the entropy-enhancing effects of LSD are larger in inward-focused states of lower external stimulation. In the second study, this one focused on musical performance, we argue for musical improvisation as a distinct state of mind in its full right, and describe it in terms of EBH's *primary states*.[1] Finally, we conclude with a discussion of the relation between EBH and IIT, and argue that the simplicity and empirical success of EBH provide valuable lessons for future developments of IIT.

10.2 Entropy and Lempel-Ziv complexity

The key requirement to empirically test EBH's predictions is a practical measure of signal diversity. To this end, the *Lempel-Ziv complexity* (LZ), originally introduced in 1976 [185],

[1] A full description of the study on musical improvisation can be found in the original reference [184].

is a popular tool to quantify the uncertainty contained in time series data. Intuitively, LZ measures the diversity of the patterns present in a particular signal. In this section we lay out the basics of the LZ algorithm and discuss its interpretation in relation to the EBH.

LZ is calculated in two steps. First, the value of a given signal X_t of length T is binarised, typically using its mean value as a threshold.[2] The resulting binary sequence is scanned sequentially and divided in distinct patterns (more on this below), and the resulting number of patterns is the LZ complexity itself, $C(X_t)$. In complexity science jargon, we say that LZ identifies signal complexity with *richness of content* [186] – a signal is regarded as complex if it is not possible to provide a brief (i.e. compressed) representation of it.

A popular way of interpreting LZ is as an approximation to the Kolmogorov complexity – the length of the shortest program able to reproduce a given sequence of symbols in a Turing machine [187]. However, we argue that this view is brittle in theory and of limited use in practice: first, the Kolmogorov complexity of a finite string of any length can differ by an *arbitrarily large* constant depending on the Turing machine used [7], making comparisons between Kolmogorov complexities estimated from finite data essentially meaningless. Second, we do not *actually* think the discretised neural signal is generated by a Turing machine, which makes the experimental explanatory power of this interpretation debatable.

Instead, a simpler and more parsimonious interpretation of LZ is as an efficient estimator of the entropy rate of X_t [188]. Mathematically,

$$\lim_{T\to\infty} \frac{C(X_t)}{\Delta_T} = H(X_t|\boldsymbol{X}_{-\infty}^{t-1}) \qquad \text{with} \quad \Delta_T = \frac{T}{\log_2 T}\,. \tag{10.1}$$

The entropy rate is an established measure of unpredictability in a time series: one half of the entropy rate is the probability of making an error with the best informed guess about the next sample [189]. Throughout the rest of the chapter we use this normalised quantity as an estimator of brain entropy, and refer to it generically as LZ.

One element we have not yet addressed, and that is common cause of confusion in the neuroscience literature, is how exactly to divide the signal into patterns to compute $C(X_t)$. This comes from the fact that there are multiple LZ compression algorithms, most prominently LZ76 [185], LZ77 [190], and LZ78 [191]. The common implementation, using a dictionary to explicitly store all the patterns, corresponds to the LZ78 algorithm. In contrast, LZ76 and LZ77 use a sliding window construction to compute the number of patterns without the need to store them. In this work we use the original LZ76 algorithm, as described by Kaspar and Schuster [192]. While it is possible to obtain entropy rate estimates from LZ77 or LZ78, these are substantially less straightforward than simply using Eq. (10.1).

[2] Common alternatives are thresholding over the median, after detrending the signal, or using the Hilbert transform. To the best of our knowledge, no solid theoretical reasons or comprehensive experimental studies exist to prefer one over the other.

10.3 Sensory stimuli and the psychedelic state

Psychedelic substances, such as LSD and psilocybin, are able to induce profound changes in subjects' perception, cognition, and conscious experience. In addition to their long-known uses for self-exploration and introspection, there is now evidence that psychedelics may enable effective treatments for various mental health conditions [193].

The therapeutic mechanisms of psychedelics, however, are not yet fully understood, although they are thought to be related to their acute positive effect on brain entropy [129]. A recent conceptual framework to understand these phenomena is the "RElaxed Beliefs Under pSychedelics" (REBUS) model [194], that posits that the entropy-enhancing effects of psychedelics help reorganise tight and possibly self-damaging beliefs (or *priors*) the subject may have about themselves or the world around them. This model, while still speculative, is consistent with other mainstream theories of brain function, such as the free-energy principle [195], and fits experimental data, including Lebedev's findings that acute entropy increase under psychedelics predicts later personality changes [196]. Given the importance of these entropic effects, understanding what internal and external factors mediate them may prove useful towards developing safe and effective therapies.

To this extent, in this first study we investigate whether changes in external stimuli have a measurable effect on the entropy of brain dynamics in human subjects. We use the data originally presented by Carhart-Harris *et al.* [136]. Twenty subjects participated in the study, and attended two experimental sessions – one in which they received a placebo, and one in which they received an i.v. administration of 75 µg of LSD. The order of the sessions was randomised separated by two weeks and blind to the participants.

MEG recordings of several minutes of length were collected under four conditions:

- Resting state, eyes closed.
- Listening to music, eyes closed.
- Resting state, eyes open.
- Watching a video, eyes open.

The music tracks were taken from the album "Yearning" by Robert Rich and Lisa Moskow; and the video was composed of segments of the "Planet Earth" documentary series produced by the BBC. Throughout this paper we will refer to these conditions as 'Closed', 'Music', 'Open', and 'Video'.

Segments with excessive artefacts were identified by visual inspection and removed, and remaining artefacts were cleaned with ICA. Data was notch-filtered at 50 Hz, and source-reconstructed in a dense mesh throughout the cortical sheet matching the standard template in the Fieldtrip toolbox [197]. Source time series (or *virtual sensors*) were reconstructed, segmented in 2 s epochs, and compressed with the LZ76 algorithm outlined above.

10.3.1 Increased signal diversity under external stimulation

Our first result is that richer external stimuli induce a strong increase in MEG signal diversity, as measured by LZ complexity. Hotelling's T^2 tests reject the null hypothesis that all stimuli have the same effect ($p < 0.001$) and trends are clearly visible, both at the group and subject level (Fig. 10.1). Similar outcomes are obtained using other statistical procedures, such as ANOVA or the linear mixed-effects model discussed below.

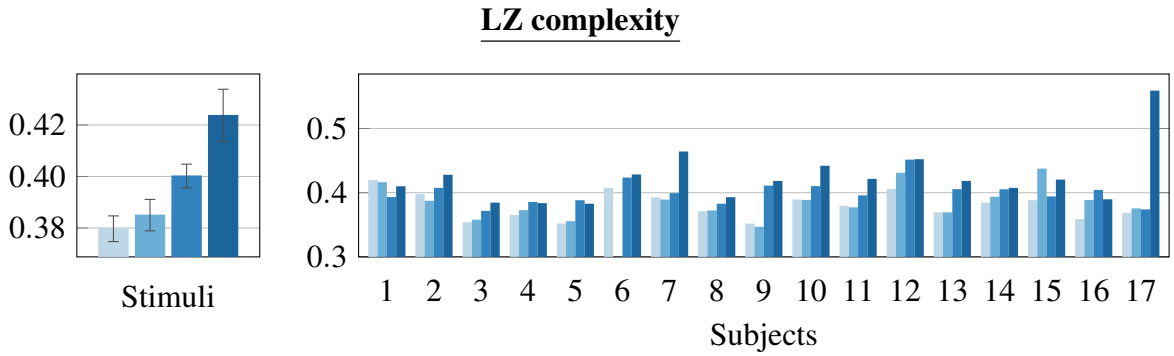


Figure 10.1: Data from the non-drug baseline condition, showing external stimulation increases LZ complexity. For each measure, data are provided per subject (right) and averaged across subjects (left; error bar is standard error). LZ complexity is reported in bits/sample.

As expected, there is a large LZ increase between the closed- and open-eyes conditions, observed throughout the whole brain. However, we also find measurable differences between the conditions with and without music, eyes closed (paired t-test, $p < 0.01$) and with and without video, eyes open (paired t-test, $p < 0.001$).

Exploring the effect of stimuli more rigorously, we now test whether stimulus richness has an effect *beyond* the mere effect of opening one's eyes. To do that we formulate two binary variables, eyes open and stimulus presence, based on the experimental conditions and according to the following table:

Experimental conditions	Model variables	
	Eyes open	Stimulus
Eyes closed, no music	0	0
Eyes closed, music	0	1
Eyes open, no video	1	0
Eyes open, video	1	1

We then fit a linear mixed-effects (LME) model using the presence of stimulus as predictor variable, and subject identity and eyes open as random effects. This model reveals a significant positive effect of stimulus on LZ (log-likelihood ratio test against null model without stimulus, $p < 0.01$). Therefore, the trends in Fig. 10.1 cannot be explained merely

by the presence or absence of visual stimulation, and must be related to the structure of such stimulation – in this case, the music and the video being played.

This result seems to differ from a recent analysis by Bola *et al.* [198], which reported finding no correlation between EEG signal diversity and 'informativeness' of auditory stimuli (i.e. the speed at which a speech recording was played). We hypothesise that Bola *et al.* did not find significant results because the difference between the experimental conditions used were too subtle. Nonetheless, this suggests that exploring which features of perceptual stimuli elicit an increase in signal diversity is an open research problem.

10.3.2 Stronger stimulus weakens drug effect

After finding that richer stimuli drive an increase in MEG signal diversity, we set out to investigate how this relationship is modulated under the effect of LSD. As before, we compute LZ complexity and fit a LME model, this time using drug and stimulus as predictor variables, and subject identity as a random effect (Fig. 10.2a). For the model shown in the figure we assumed the four experimental conditions were ordered (i.e. had integer values), but all our results remain if this assumption is relaxed (i.e. conditions are one categorical variable).

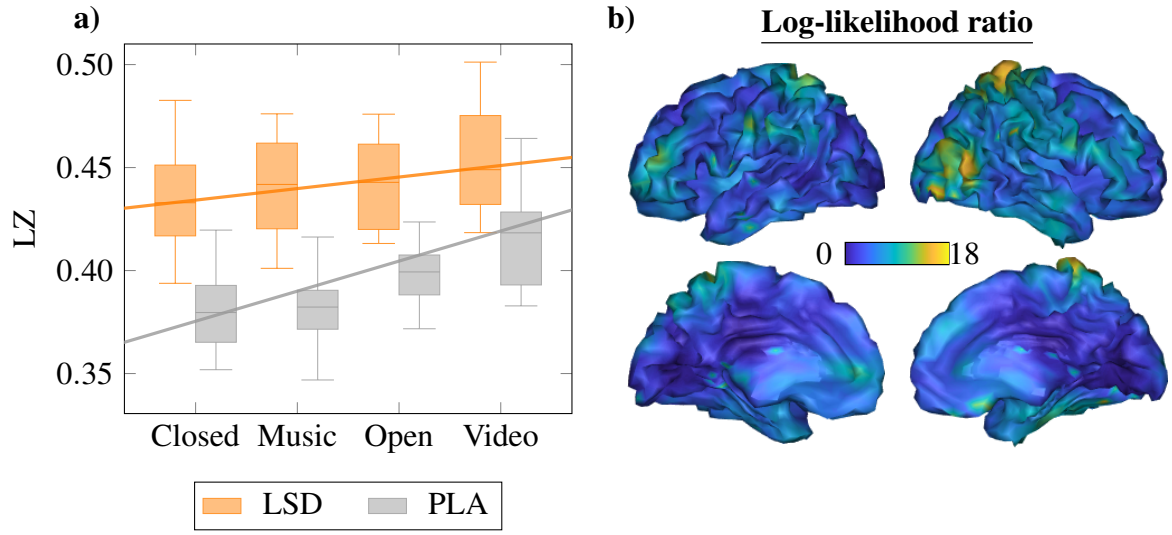


Figure 10.2: Stronger external stimulation reduces the enhancing effect of LSD on brain signal diversity. **(a)** Linear mixed-effects models fit to LZ complexity, illustrating the significant drug×stimulus interaction term ($p < 0.001$). **(b)** Drug×stimulus log-likelihood ratio (LLR) for the LME model. For reference, the $p = 0.05$ threshold in this comparison corresponds to LLR ≈ 9, and higher LLR values imply lower p-values. In regions with a with high LLR (bright yellow hue) external stimuli such as music or video strongly reduced the effect of LSD.

The first result is that LZ increases dramatically under the effects of LSD (drug term of the LME is significantly above 0, with $p < 0.001$), replicating previously reported results [129]. This effect is strong and widespread, observable in all stimulus conditions. Additionally, and in agreement with the results above, this LME model also reveals a strong positive effect of stimuli on LZ irrespective of drug ($p < 0.001$).

Finally, and in what constitutes the main result of this study, we assess the effect of the drug×stimulus interaction. We do this by formulating two models: a full model where all terms are present, and a pruned model where the interaction term has been removed. We fit both models to the data, and the full model is strongly preferred over the null model by log-likelihood ratio tests ($p < 0.01$), which indicates a significant interaction effect.[3] Furthermore, this interaction effect is negative – i.e. increased external stimulation *reduces* the effect of the drug. Spatially, the effect is located predominantly in the posterior temporal lobe, although in the case of LZ the effect is also seen throughout the parietal lobe (Fig. 10.2b).

These results agree with the multitude of studies and reports that have highlighted the importance of the environment in which the drug is taken and the subject's mood and expectations prior to the substance use – factors commonly known as *set and setting*. More broadly, these results have implications for our understanding of the psychedelic state and the application of psychedelics in psychotherapy. If the therapeutic mechanisms of psychedelics do depend on their acute entropy-enhancing effects, these findings could provide therapists with guidelines to modulate the patient's entropy by means of music, video, or a balance between inward- and outward-focused states.

10.4 The improvisational state of mind

In this study, we focus on another conjecture of the EBH: that brain entropy not only is higher in states of richer conscious content, but that this difference is due to the interplay between two kinds of conscious states. EBH makes the distinction between *secondary states*, characteristic of the experience of adult humans; and *primary states* to which the mind regresses under specific conditions, such as in response to severe stress, psychedelic drugs or in REM sleep. Physiologically, brain entropy is increased in primary states, which correlates with greater diversity and richness of experiential content; and is suppressed in secondary states, which give rise to more regular and stable cognitive processes enabling metacognitive functions like reality-testing and self-awareness.

Although primary consciousness may be a suboptimal mode of cognition, they seem to be more than a mere psychological atavism. Plenty of reports show how events involving

[3] We additionally performed 2-way ANOVA tests, which also detected a significant interaction effect ($p < 0.05$).

primary states can bring deep experiences and have profound therapeutic effects [199, 200]. In effect, the high entropy of primary states seems to allow one to overcome the inability to think and behave in a flexible manner, narrow-mindedness and aggressive self-critical attitudes.

In this study we take the EBH out of its home base of psychedelic state, and ask if it is applicable to the domain of musical experience, and in particular musical improvisation. Is the improvisational state of mind a primary state? Could one find traces of primary states in musicians and audience during improvisational performance?

10.4.1 Experiment and data acquisition

The experiment consisted of a live chamber music concert by a professional trio (piano, flute, singer), in the presence of an invited audience of 22 adults with varying levels of musical expertise. During the experiment each piece was performed twice: once in strict mode (corresponding to a prepared interpretation, as close as possible to the written score), and once in what they described as an improvised, or "let-go," mode (corresponding to the improvisatory approach, in which performers are free to deviate from the score in extemporised gestures).

Raw EEG signals of the three performers and four audience members were measured using CE-certified devices with electrodes positioned according to the 10–20 electrode positioning system [201]. The reference electrode was placed behind Cz and the ground electrode on the forehead. All locations were cleaned with abrasive gel and conductive gel was used to ensure low skin impedance. EEG data were collected at 250 Hz, and bandpass filtered between 2 and 40 Hz. Bad channels and bad epochs were visually identified and removed from the analysis.

The neural signal was split in segments of 2 s, which provides enough data points to have an accurate estimation of LZ while being short enough to not compromise the stationarity of the data. The values of each segment were then binarised using the corresponding median value as a threshold. The LZ was finally calculated for each temporal segment of each electrode, and then averaged across time and electrodes to obtain one LZ value per subject per condition. Statistical significance is determined with t-tests (paired when possible, and unpaired elsewhere) and effect sizes are measured with Cohen's d.

10.4.2 Increased signal diversity in the improvisational state

Based on the properties of LZ outlined above, we investigated the complexity of the measured EEG signals of all 7 subjects in both conditions, under a working hypothesis that LZ is higher during the improvised than during the prepared condition. Our main result is that

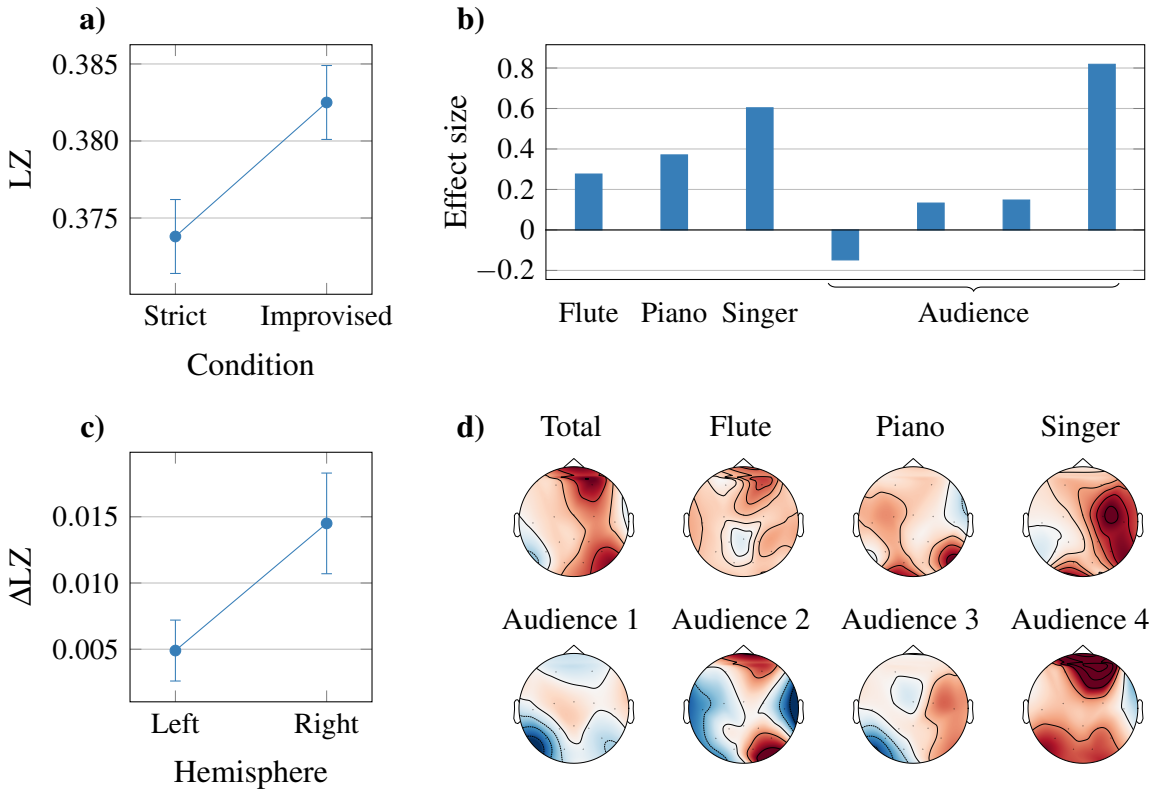


Figure 10.3: Lempel-Ziv complexity in strict and improvised musical performances. **(a)** LZ increase in the improvised performance. **(b)** 6 of the 7 subjects show effect sizes consistent with the overall trend. **(c)** The increase comes from the right hemisphere. **(d)** Topographical maps for the LZ increase, with red for positive values and blue for negative ones. Colour map range: $-1 \leq d \leq 1$.

LZ does in fact increase in the improvised condition, by a difference of 0.010 bit (95% CI: 0.002–0.016, $N = 7$, two-sample t-test $p = 0.024$), shown in Figure 10.3a.

The small p-value for the group-level test, despite the very small number of subjects in the study, is caused by the fact that the observed LZ increase is very consistent across subjects, with 6 of the 7 participants showing changes in the same (positive) direction (Fig. 10.3b). While results among the audience are mixed, all three musicians show substantial increases in LZ during the improvised performance, and this effect is most significant in the singer and the pianist.

Following up on our main result, and in agreement with common neuroscientific theories, we find that the LZ increase is mainly localised in the right hemisphere (average difference in LZ increase between right and left hemisphere: 0.010 bit, 95% CI: 0.004–0.016, $p = 0.003$). The right hemisphere is conventionally associated with cognitive processes like creativity and divergent thinking, which may indicate that musicians were more engaged in a creative process during the improvised performance, and were less likely to enter the logic-driven

and rule-following states usually associated with the left hemisphere. Figure 10.3c shows the average difference in LZ increase and Figure 10.3d its spatial distribution.

In the context of the EBH, we interpret these results as a quantitative signature of a shift from a secondary toward a primary state of consciousness during musical improvisation, characterised by a more flexible creative process.

This opens an interesting connection between primary states and states of *flow* [202]. In a more speculative tone, we raise the tentative idea that all states of flow are primary states (but not vice-versa). In other words, all the descriptions associated with feelings of flow are consistent with the characteristics of primary states of cognition, while it is clear that not all primary states involve flow. If true, this could add LZ to the small set of known biomarkers of flow states [203].

10.5 Discussion

These studies, together with the myriad other reports in the literature, strongly suggest that the bold intuitions behind EBH are not entirely unfounded. While it does not make any predictions about its structure, EBH's claim that the richness of *conscious content* is correlated with richness in *information content* is strongly supported by the evidence. In light of this success, it is worth examining the relation between EBH and IIT so we can discover, synergistically, what we can learn from both together that we wouldn't learn from either of them on their own.

10.5.1 Lempel-Ziv complexity and Φ

Methodologically, this chapter diverges substantially from all others in this thesis: instead of the dynamic, high-order correlations discussed everywhere else, in this chapter we used simple, single-channel entropy, and found it a clear and robust neural correlate of consciousness. Given the huge disparity between the LZ presented and reviewed here, and the Φ results in Chapter 7, it is natural to ask whether there is any relation between the two, and if LZ can help us formulate principled and more empirically powerful Φ measures. Interestingly, a recent simulation study showed that LZ and Φ are in fact strongly correlated [204] (although this has only been shown in the small logic-gate networks on which Φ-3.0 can be computed).

As a start, we can try to speculate why it is that Φ and LZ are so correlated in these simple systems. One possible hint comes from a standard result in symbolic dynamics [205]: when a Markovian system is coarse-grained, the resulting dynamics are in general non-Markovian. In IIT language, this could mean that when we partition the system during the calculation of

Φ, we are not "being fair" to the parts, because we are unjustly maiming their predictive power. It could be that the system is perfectly predictable, but with longer memory.

Maybe this hints at the role of *non-Markovianity* of the neural system. Perhaps then it is not so surprising that embedding methods able to deal with non-Markovian dynamics have been so successful in consciousness neuroscience [145]. If this line of reasoning holds any ground, it would be bad news for a fundamental, causal version of IIT (since, as we have argued before [180], these kinds of Φ measures are not defined for non-Markovian processes), but it would open an exciting line of research at the interface between these two informational theories of consciousness.

10.5.2 The 'brain' in 'entropic brain'

A common criticism to the entropic brain hypothesis (e.g. Ref. [206]) is that a naive interpretation of it might suggest that conscious level is maximised in a maximally entropic state, in which the neurons are disconnected and therefore statistically independent – akin to the disordered phase of an Ising-style model. In this dynamical regime no substantial information processing is possible [207], and is evidently not how the brain operates. This reasoning might also suggest that an ideal gas is somehow "maximally conscious" amongst other thermodynamical systems, a claim that would make any philosopher of mind shiver.

However, this is an example of the danger of taking analogies with simple models too seriously. In a critical Ising model, if the spins are disconnected (e.g. by setting the coupling to $J = 0$) the system shifts to the disordered, maximally entropic state. In contrast, in a slice of actual brain tissue, if the neurons are disconnected (e.g. through pharmacological agents), the dynamical range of neurons is drastically reduced [208] – i.e. *they just shut down*. In plain words: predictions based on an Ising model simply do not hold.

A more charitable way of interpreting the claims of the entropic brain hypothesis is that *subject to the physiological constraints of the human brain*, states of higher entropy are states of richer contents of consciousness. This weaker form of the EBH does seem supported by the evidence, as argued in the two studies above and in the broader literature.

This move, however, comes at a cost: we are confining EBH's domain of applicability, in a somewhat arbitrary way. In contrast, modern IIT remains insistent on tackling the "hard" "problem," and being a *fundamental* theory of consciousness. Learning from EBH, it may be beneficial to whip up a *weak* IITC, as opposed to the dominant *strong* IITC, that aims at explaining consciousness "only" as created by human brains. To some, conceptual elegance, in exchange for actual predictive power, might seem like a worthwile price to pay.

10.6 Conclusion of Part III

In this Part, we have presented developments and alternatives to IIT, both as a collection of information-theoretic tools, and as a theory of consciousness.

In Chapters 8 and 9 we proposed Partial Information Decomposition [19] as a framework to understand high-order interactions, which are of crucial importance for IIT. Using PID, we made two direct contributions to IIT:

- First, we showed that neural complexity, the flagship measure of IIT 0.1 does not faithfully represent a balance between global integration and local segregation. We proposed a new measure, the O-information, that represents these intuitions more directly.
- Second, we presented a unification of IIT and PID, ΦID, and used it to explain the seemingly paradoxical results of Chapter 6. In addition, ΦID allowed us to alleviate the shortcomings of previously proposed measures in IIT 2.0.

Taken together, these contributions show the benefits of a principled, structured approach to the problem of studying integrated information – instead of imposing a large number of arbitrary choices on a measure of integrated information, we started from a more basic theory of multivariate information and used it to refine our intuitions of integration and segregation.

Finally, we presented two studies analysing M/EEG data of subjects in altered states of consciousness, and interpreted the results using the Entropic Brain Hypothesis [181]. We argued that a marker as simple as brain entropy is extremely effective in practice, in stark contrast with the elaborate Φ measures of Chapters 6 and 7, for which experimental evidence is mixed and inconclusive.

Chapter 11

Conclusion

11.1 Summary of thesis contributions

The results presented in this thesis have achieved the objectives laid out in Section 1.2: we have applied Integrated Information Theory to the study of complex neural systems, we have reviewed existing and novel evidence regarding IIT as a theory of consciousness, and have presented new developments to its mathematical basis. In the following, we list our contributions chapter by chapter.

Literature review of integrated information measures. We have provided a comprehensive review of proposed measures of integrated information, including a comparison of their basic properties (and the corresponding proofs). For each measure, we gave a general information-theoretic description, avoiding, to the extent possible, the specialised jargon of recent IIT.

Link between Φ and criticality in metastable coupled oscillators. We showed that a network of coupled Kuramoto oscillators, when tuned close to its critical point, exhibits both high metastability and high integrated information. This is the first description of a dynamical system in which the three major complexity indicators of criticality, metastability, and integrated information coincide.

Link between Φ and power law-like avalanches in spiking neurons. We presented a simple model network of spiking neurons which, by tuning its topology through a single parameter, undergoes a transition between two dynamical regimes, characterised by dominant information storage or transfer, respectively. The intermediate point, with balanced integration and segregation, is the point where neural avalanches most closely approximate power law statistics, and estimated exponents match empirical findings in *in vitro* recordings.

Link between Φ and distributed computation in cellular automata. We first showed that Φ in cellular automata correlates with intuitive notions of complexity, as given by Wolfram's complexity classes, and later strengthened this result through extensive simulations in the context of Langton's *edge of chaos* hypothesis. We then analysed a novel pointwise version of Φ at a local scale in space and time, and confirmed that information is integrated by coherent structures like gliders, blinkers, and collisions.

Taken together, these results support our proposal of Φ *as a universal marker of dynamical complexity in complex systems*. This idea was first presented and argued qualitatively through the examples in Chapters 3-5, and later formalised with the full decomposition of Φ and the identification of a theory of causal emergence within it in Chapter 9.

Comprehensive comparison study of Φ measures in simulation. We provided the most extensive simulation study of integrated information measures to date, and observed a striking variety of behaviour among proposed measures even in simple systems. From these results, we pushed forward the conclusion that the axiomatic basis of IIT is *underspecified*, in the sense that measures that roughly represent the same intuitions in theory behave very differently in practice.

Literature review and novel evidence regarding IIT as a theory of consciousness. We performed, for the first time, a literature review of Φ measures applied to neuroimaging data, and concluded that at this stage experimental evidence does not warrant a direct association between Φ and levels of consciousness. However, a simple model suggests this discrepancy may be attributed to the large number of unobserved variables in contemporary neuroimaging techniques, and points to state-space reconstruction as a potential way forward.

These two limitations, both theoretical and experimental, call for a more pragmatic and evidence-driven research programme in IIT. In particular, we argue that efforts to formulate (i) more specific measures of integrated information with explicit operational meaning, and (ii) robust estimators from neuroimaging data, are likely to benefit the development of IIT.

New measure of high-order interdependecies in large systems. We proposed a novel measure of high-order interdependecies, the *O-information*, that is more scalable and interpretable than the alternatives. We showed that neural complexity, the flagship measure of IIT 0.1 does not faithfully represent a balance between global integration and local segregation, while the O-information represents these intuitions more directly.

Decomposition of integrated information and explanation of previous results. We unified IIT with another branch of information theory, Partial Information Decomposition, to provide a full decomposition of integrated information, or ΦID. Most importantly, ΦID reveals that what is typically referred to as 'integration' is actually an aggregate of several heterogeneous phenomena, and can help us understand and alleviate the limitations of existing Φ measures.

Case studies of information theory in altered states of consciousness. We presented two experimental studies of altered states of consciousness – the psychedelic state and musical improvisation – analysed and interpreted using the (comparatively much simpler) framework of the Entropic Brain Hypothesis. We argued that the considerable success of the simple measures in this framework, such as Lempel-Ziv complexity, should be learnt from and incorporated in future IIT research.

11.2 Future work

Several directions for future work have been proposed throughout this thesis, coming both from the limitations highlighted in Part II and from the new developments presented in Part III. Here we summarise these ideas and link them to relevant ongoing work in the research community.

First, in order to eliminate the ambiguity in our intuitions of integration and differentiation, as highlighted in Chapter 6, it might help to *formulate measures of integration with operational meaning*. Measures with operational meaning are more grounded, and easier to handle and reason about. The recent work on operational definitions of redundancy may be relevant [209, 210], and the axiomatic basis of IITC will need to be re-evaluated [211].

Once a suitable, unambiguous measure (or set of measures) of integrated information has been identified, we can relate it to the decomposition in Chapter 9 and *investigate not only the amount of integrated information, but also its structure*. For example, the double-redundancy lattice (Fig. 9.2) gives a structure of dynamical information that is different from the one suggested in other IIT literature [6, 11]. An exploration of their differences, mathematical properties and empirical success is a very interesting follow-up to this work.

Overall, this is part of a broader trend to move beyond IIT as a theory of conscious level (focused on 'Big Phi' as a scalar index), and towards a more comprehensive theory of conscious content. This could be related to more elaborate mathematical structures, for example coming from category theory [212].

On a separate front, we have much to learn from the unreasonable effectiveness of LZ as a marker of conscious level. As suggested by Ref. [204], in certain small systems there is an empirical correlation between LZ and Φ-3.0, although the origin of this correlation and its relation to neural dynamics is not established. Therefore, it is worthwile to *investigate the role of non-Markovianity in neural dynamics*, as a way to bridge the gap between Φ and LZ. For example, this could be done through Takens vector embeddings [213] or explicit state-space models [146].

On this same line of research, there is plenty of space for theoretical work to *extend, in a principled manner, the functionality of LZ as an analysis tool*. For example, one important limitation of the LZ algorithm is that it only returns the estimated value of the entropy rate, and nothing else – it does not provide an *explicit* model that can be evaluated at will. Such a model could be extremely useful, since it would allow us to do things like obtain a time-resolved (i.e. pointwise) estimate of surprise, allowing us to inspect LZ-like entropy with temporal granularity; to fit the model in one time series and evaluate it in another, giving us a notion of relative entropy or divergence; and to inspect the model itself to analyse not only how many, but what kinds of patterns appear. To this end, *variable-order Markov models* [214] are promising candidates as general data-efficient models of time series data.

11.3 Parting thoughts

It is hard not to feel overwhelmed by the unsurmountableness of the problems tackled here. After all, when it comes to consciousness science, this thesis has brought more questions than answers: If, upon closer inspection, integration is so ambiguous and ill-defined, did the original IIT 0.1 intuitions have any value to begin with? If we follow the leads in Part III and try to fix IIT with PID, given the lack of consensus in the PID literature doesn't ΦID just perpetuate the ambiguity of Φ, but sixteen times over? More fundamentally, if experimental evidence for IITC is so thin, and the maths are so hard, why keep working on it?

To answer these difficult questions, I shall steal the words of the accomplished epistemic anarchist Paul Feyerabend. In his outstanding polemic *Against Method* [215], Feyerabend convincingly argues that *there is no such thing as a "scientific method,"* and that the advancement of knowledge is a *fundamentally irrational endeavour*.

This is necessarily the case, for any new cosmology or world-view is bound to clash with the status quo, and be deemed irrational by the standards of its time. In this sense, Copernicus and Galileo's ideas were irrational at their times, and so were Einstein's and many others' – or, closer to home, those of Shannon himself. As beautifully portrayed by Hamming [216]:

> *Courage is one of the things that Shannon had supremely. You have only to think of his major theorem. He wants to create a method of coding, but he doesn't know what to do so he makes a random code. Then he is stuck. And then he asks the impossible question, "What would the average random code do?" He then proves that the average code is arbitrarily good, and that therefore there must be at least one good code. Who but a man of infinite courage could have dared to think those thoughts? That is the characteristic of great scientists; they have courage. They will go forward under incredible circumstances; they think and continue to think.*

While we advocate for an open and transparent engagement with experimental evidence, we cannot reasonably demand the same level of adherence to "facts" from a theory of consciousness as we demand from theories in other branches of science. In the words of Feyerabend, "this is like arranging a fight between an infant and a grown man, and announcing triumphantly, what is obvious anyway, that the man is going to win" [215, p. 108]. Therefore, a new science of consciousness must find its own path, and be allowed to make mistakes, *as a child playing air-guitar plays no wrong notes*.

In one of his most inspiring passages, Feyerabend calls [215, p. 113]:

> *This need to* wait*, and to* ignore *large masses of critical observations and measurements, is hardly ever discussed in our methodologies. Disregarding the possibility that a new physics or a new astronomy might have to be judged by a new theory of knowledge and might require entirely new tests, empirically inclined scientists at once confront it with the* status quo *and announce triumphantly that 'it is not in agreement with facts and received principles'. They are of course right, and even trivially so, but not in the sense intended by them. For at an early stage of development the contradiction only indicates that the old and the new are* different *and* out of phase*. It does not show which view is the* better *one. [...] How shall we proceed in order to bring about such a fair comparison? The first step is clear: we must* retain *the new cosmology until it has been supplemented by the necessary auxiliary sciences. We must retain it in the face of plain and unambiguous refuting facts.*

If we could listen to Feyerabend, he would no doubt prompt us to disregard the overwhelming odds against, and *follow our gut*. Radical ideas that transform science do not come from capital-R Reason. Radical ideas fight, and survive – and they survive because of "*prejudice, passion, conceit, errors, sheer pigheadedness*" [215, p. 116].

If History is of any guide, it is *naive curiosity* and *fascination* that move us forward. And, after all, the fact remains that ***the problem of consciousness is so devilishly interesting***, that despite the ups and downs, the highs and lows, one can't but continue staying up at night asking oneself "how is this a thing?"

Appendix A

Information-theoretic foundations

A.1 Derivation and concavity proof of I^*

A.1.1 Derivation of I^* in Gaussian systems

Here, we provide a closed-form expression for the mismatched decoding information in a Gaussian dynamical system. For clarity, we omit the $X, \tau, \mathcal{P}$ arguments of $\tilde{I}$ and write it as a function of β only. The formula for $\tilde{I}(\beta)$ for a stationary continuous random process is

$$\tilde{I}(\beta) = -\int \mathrm{d}x\, p(x) \log \int \mathrm{d}\tilde{x}\, p(\tilde{x}) q(x|\tilde{x})^{\beta} + \int \mathrm{d}\tilde{x} \int \mathrm{d}x\, p(x,\tilde{x}) \log q(x|\tilde{x})^{\beta}, \tag{A.1}$$

where $p(x)$ is the distribution for X_t, $p(x,\tilde{x})$ is the joint distribution for $(X_t, X_{t-\tau})$, and $q(x|\tilde{x})$ is the conditional distribution for X_t given $X_{t-\tau}$ under the partitioning in question. The function $\tilde{I}(\beta)$ also depends on X_t, τ and $\mathcal{P}$, but for the sake of clarity we omit all arguments except for β, which is the parameter of interest here. When X_t is Gaussian with covariance matrix Σ_X (and mean 0 without loss of generality), we have

$$p(x) = (2\pi)^{-n/2} |\Sigma_X|^{-1/2} \exp\left[-\frac{1}{2} \psi\left(x, \Sigma_X^{-1}\right) \right], \tag{A.2}$$

where we define

$$\psi(x, M) =: x^{\mathrm{T}} M x \tag{A.3}$$

for a vector x and a matrix M. Furthermore,

$$q(x|\tilde{x}) \;=\; (2\pi)^{-n/2} |\Pi_{X|\tilde{X}}|^{-1/2} \exp\left[-\frac{1}{2} \psi\left(x - \Pi_{X\tilde{X}} \Pi_{\tilde{X}}^{-1} \tilde{x}, \Pi_{X|\tilde{X}}^{-1}\right) \right], \tag{A.4}$$

where Π_X is the block diagonal covariance matrix for X_t under the partition, $\Pi_{X\tilde{X}} =: \Sigma_q(X_t, X_{t-\tau}) = \Pi_{\tilde{X}X}^{\mathrm{T}}$ is the block diagonal auto-covariance matrix associated with the parti-

tion, and $\Pi_{X|\tilde{X}}$ is the partial covariance

$$\Pi_{X|\tilde{X}} = \Pi_X - \Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}\Pi_{\tilde{X}X}\,. \tag{A.5}$$

We start with the integral

$$\int \mathrm{d}\tilde{x}\, p(\tilde{x})q(x|\tilde{x})^\beta = (2\pi)^{-n\beta/2}|\Pi_{X|\tilde{X}}|^{-\beta/2}(2\pi)^{-n/2}\,|\Sigma_X|^{-1/2}\int \mathrm{d}\tilde{x}\exp(\mathcal{E})\,, \tag{A.6}$$

where

$$\mathcal{E} = \frac{1}{2}\tilde{x}^{\mathrm{T}}\Sigma_X^{-1}\tilde{x} - \frac{\beta}{2}\tilde{x}^{\mathrm{T}}\Pi_X^{-1}\Pi_{\tilde{X}X}\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_X^{-1}\tilde{x} + \beta x^{\mathrm{T}}\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_X^{-1}\tilde{x} - \frac{\beta}{2}x^{\mathrm{T}}\Pi_{X|\tilde{X}}^{-1}x\,. \tag{A.7}$$

If we write

$$\mathcal{E} = -\frac{1}{2}(\tilde{x}-Bx)^{\mathrm{T}}Q(\tilde{x}-Bx) - \frac{1}{2}x^{\mathrm{T}}R_1 x\,, \tag{A.8}$$

then

$$Q = \Sigma_X^{-1} + \beta\Pi_X^{-1}\Pi_{\tilde{X}X}\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_X^{-1}\,, \tag{A.9a}$$

$$B^{\mathrm{T}} = \beta\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_X^{-1}Q^{-1}\,, \tag{A.9b}$$

$$R_1 = \beta\Pi_{X|\tilde{X}}^{-1} - \beta^2\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_X^{-1}Q^{-1}\Pi_X^{-1}\Pi_{\tilde{X}X}\Pi_{X|\tilde{X}}^{-1}\,, \tag{A.9c}$$

so

$$\begin{aligned}\int \mathrm{d}\tilde{x}\exp(\mathcal{E}) &= \exp\left(-\frac{1}{2}x^{\mathrm{T}}R_1x\right)\int \mathrm{d}y\exp\left(-\frac{1}{2}y^{\mathrm{T}}Qy\right)\\ &= \exp\left(-\frac{1}{2}x^{\mathrm{T}}R_1x\right)(2\pi)^{n/2}|Q|^{-1/2}\,.\end{aligned} \tag{A.10}$$

Hence, using Equations (A.2) and (A.6), we obtain the first term in Equation (A.1):

$$-\int \mathrm{d}x\, p(x)\log\int \mathrm{d}\tilde{x}\, p(\tilde{x})q(x|\tilde{x})^\beta = \frac{n\beta}{2}\log 2\pi + \frac{1}{2}\log\left(|Q|\cdot|\Sigma_X|\cdot|\Pi_{X|\tilde{X}}|^\beta\right) + \frac{1}{2}\mathrm{tr}(\Sigma_X R_1)\,. \tag{A.11}$$

Now, moving on to the second term in Equation (A.1),

$$\int \mathrm{d}\tilde{x}\int \mathrm{d}x\, p(x,\tilde{x})\log q(x|\tilde{x})^\beta = -\frac{\beta n}{2}\log 2\pi - \frac{\beta}{2}\log|\Pi_{X|\tilde{X}}| - \frac{\beta}{2}I_1\,, \tag{A.12}$$

where

$$\begin{aligned} I_1 &= \int \mathrm{d}\tilde{x} \int \mathrm{d}x\, p(x,\tilde{x})\, \psi\left(x - \Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}\tilde{x}, \Pi_{X|\tilde{X}}^{-1}\right) \\ &= \int \mathrm{d}x\, p(x)\, \psi\left(x, \Pi_{X|\tilde{X}}^{-1}\right) + \int \mathrm{d}\tilde{x}\, p(\tilde{x})\, \psi\left(\tilde{x}, \Pi_{\tilde{X}}^{-1}\Pi_{\tilde{X}X}\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}\right) \\ &\quad - 2\int \mathrm{d}\tilde{x} \int \mathrm{d}x\, p(x,\tilde{x})\, x^{\mathrm{T}}\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}\tilde{x} \\ &= \mathrm{tr}\left(\Pi_{X|\tilde{X}}^{-1}\Sigma_X\right) + \mathrm{tr}\left(\Pi_{\tilde{X}}^{-1}\Pi_{\tilde{X}X}\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}\Sigma_X\right) - 2\,\mathrm{tr}\left(\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}\Sigma_{\tilde{X}X}\right), \end{aligned} \tag{A.13}$$

where $\Sigma_{\tilde{X}X} =: \Sigma(X_{t-\tau}, X_t)$. Thus, the second term in Equation (A.1) is given by

$$\begin{aligned} \int \mathrm{d}\tilde{x} \int \mathrm{d}x\, p(x,\tilde{x}) \log q(x|\tilde{x})^{\beta} = &-\frac{\beta n}{2}\log 2\pi - \frac{\beta}{2}\log|\Pi_{X|\tilde{X}}| \\ &+ \frac{1}{2}\mathrm{tr}(\Sigma_X R_2) + \beta\,\mathrm{tr}\left(\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}\Sigma_{\tilde{X}X}\right), \end{aligned} \tag{A.14}$$

where

$$R_2 = -\beta\Pi_{X|\tilde{X}}^{-1} - \beta\Pi_{\tilde{X}}^{-1}\Pi_{\tilde{X}X}\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}. \tag{A.15}$$

Finally, putting all the terms in Equations (A.11) and (A.14) together, we obtain

$$\tilde{I}(\beta) = \frac{1}{2}\log\left(|Q|\cdot|\Sigma_X|\right) + \frac{1}{2}\mathrm{tr}(\Sigma_X R) + \beta\,\mathrm{tr}\left(\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}\Sigma_{\tilde{X}X}\right), \tag{A.16}$$

where

$$Q = \Sigma_X^{-1} + \beta\Pi_{\tilde{X}}^{-1}\Pi_{\tilde{X}X}\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}, \tag{A.17}$$

$$R = -\beta\Pi_{\tilde{X}}^{-1}\Pi_{\tilde{X}X}\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1} - \beta^2\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}Q^{-1}\Pi_{\tilde{X}}^{-1}\Pi_{\tilde{X}X}\Pi_{X|\tilde{X}}^{-1}. \tag{A.18}$$

We note that this formula for $\tilde{I}(\beta)$ has been verified with numerical methods, and it is not the same as the formula reported by Oizumi et al. [10].

A.1.2 $\tilde{I}(\beta)$ is concave in β in Gaussian systems

Throughout this proof, we will rely multiple times on the the book *Convex Optimization* by Boyd and Vandenberghe [32]. Our aim is to show that $\tilde{I}(\beta)$ is concave in β, which means it has a unique maximum and can be treated with standard convex optimisation tools. Throughout this proof, we follow Boyd and Vandenberghe's notation: a function f is said to be convex, convex downwards or concave upwards if $f(ax+by) \leq af(x)+bf(y)$, for all real non-negative a, b with $a+b=1$.

We start with the second term in Equation (A.1),

$$\int \mathrm{d}\tilde{x} \int \mathrm{d}x\, p(x,\tilde{x}) \log q(x|\tilde{x})^{\beta} = \beta \int \mathrm{d}\tilde{x} \int \mathrm{d}x\, p(x,\tilde{x}) \log q(x|\tilde{x}), \tag{A.19}$$

which is linear in β. Moving to the first term, using Equation (A.4) it can be rewritten as

$$\begin{aligned} -\int \mathrm{d}x\, p(x) \log \left[\int \mathrm{d}\tilde{x}\, p(\tilde{x}) q(x|\tilde{x})^{\beta} \right] = -\int \mathrm{d}x\, p(x) \left[-\tfrac{n\beta}{2} \log 2\pi - \tfrac{\beta}{2} \log |\Pi_{X|\tilde{X}}| \right] \\ -\int \mathrm{d}x\, p(x) \log \left[p(\tilde{x}) \exp\left(-\beta f(x,\tilde{x})\right) \mathrm{d}\tilde{x} \right] . \end{aligned}$$

We see that the only nonlinear term in $\tilde{I}(\beta)$ is

$$-\int \mathrm{d}x\, p(x) \log \left[\int \mathrm{d}\tilde{x}\, p(\tilde{x}) \exp(-\beta f(x,\tilde{x}) \right], \tag{A.20}$$

where

$$f(x,\tilde{x}) = \frac{1}{2} \psi \left(x - \Pi_{X\tilde{X}} \Pi_{\tilde{X}}^{-1} \tilde{x}, \Pi_{X|\tilde{X}}^{-1} \right). \tag{A.21}$$

Now, we draw from two lemmas presented in [32]:

- An affine function preserves concavity, in the sense that a linear combination of convex (concave) functions is also convex (concave).
- A non-negative weighted sum preserves concavity. Since $p(x) > 0$, the outer integral in Equation (A.20) preserves concavity,

With these two remarks, we know that, to prove the concavity of $\tilde{I}(\beta)$, we just need to prove the concavity of

$$-\log \left[\int \mathrm{d}\tilde{x}\, p(\tilde{x}) \exp\left(-\beta f(x,\tilde{x})\right) \right]. \tag{A.22}$$

This is known as a log-sum-exp function, which, as per Section 3.1.5 of [32], is convex in β. Finally, the minus sign in the last equation flips the convexity and we conclude that $\tilde{I}(\beta)$ is concave in β.

A.2 Bounds on causal density

We now prove that causal density is upper-bounded by time-delayed mutual information, satisfying what other authors have considered a fundamental requirement for a measure of integrated information [11]. As before, we omit the arguments to CD for clarity. We begin

by writing down CD in terms of mutual information:

$$\begin{aligned} \mathrm{CD} &= \frac{1}{n(n-1)} \sum_{i \neq j} \mathrm{TE}_\tau(X^i \to X^j | X^{[ij]}) \\ &= \frac{1}{n(n-1)} \sum_{i \neq j} I(X_t^i; X_{t+\tau}^j | X_t^{[i]}), \end{aligned} \tag{A.23}$$

where as before $X_t^{[i]}$ represents the set of all variables in X_t except X_t^i. We will use the chain rule of mutual information [7],

$$I(X;Y,Z) = I(X;Z) + I(X;Y|Z). \tag{A.24}$$

Using this chain rule and the non-negativity of mutual information, we can state that $I(X_t^i; X_{t+\tau}^j | X_t^{[i]}) \leq I(X_t; X_{t+\tau}^j)$, and therefore

$$\mathrm{CD} \leq \frac{1}{n(n-1)} \sum_{i \neq j} I(X_t; X_{t+\tau}^i). \tag{A.25}$$

Also by the same chain rule, it is easy to see that $I(X_t; X_{t+\tau}^i) \leq I(X_t; X_{t+\tau})$. Then,

$$\mathrm{CD} \leq \frac{1}{n(n-1)} \sum_{i \neq j} I(X_t; X_{t+\tau}) \ . \tag{A.26}$$

Given that the sum runs across all $n(n-1)$ pairs, we arrive at our result

$$\mathrm{CD} \leq I(X_t; X_{t+\tau}). \tag{A.27}$$

A.3 Properties of integrated information measures

We prove the properties of in Table 2.2 of the main text. We will make use of the properties of mutual information introduced in Section 2.2, repeated here for convenience:

MI-1 $I(X;Y) = I(Y;X)$,

MI-2 $I(X;Y) \geq 0$,

MI-3 $I(f(X); g(Y)) = I(X;Y)$ for any injective functions f, g,

A.3.1 Whole-minus-sum integrated information Φ

Time-symmetric Follows from (MI-1).

Non-negative Proof by example. If $X_t^i = X_t^j$, we have $\Phi = (1-N)I(X_t^i; X_{t-\tau}^i) \leq 0$.

Rescaling-invariant Follows from (MI-3) when Balduzzi and Tononi's [5] normalisation factor is not used.

Bounded by TDMI Follows from (MI-2).

A.3.2 Integrated stochastic interaction $\tilde{\Phi}$

Time-symmetric Follows from $H(X_t|H_{t-\tau}) = H(X_{t-\tau}|H_t)$, which can be proved starting from the system temporal joint entropy

$$\begin{aligned} H(X_t, X_{t-\tau}) &= H(X_t|X_{t-\tau}) + H(X_{t-\tau}) \\ &= H(X_{t-\tau}, X_t) = H(X_{t-\tau}|X_t) + H(X_t), \end{aligned}$$

and using the fact that by the ergodic property $H(X_t) = H(X_{t-\tau})$. The same logic applies to all parts of the system:

Non-negative Follows from the fact that $\tilde{\Phi}$ is an M-projection (see Reference [11]).

Rescaling-invariant Follows from the non-invariance of differential entropy [7] (regardless of whether a normalisation factor is used).

Bounded by TDMI Proof by counterexample. In the two-node AR process of the main text $\tilde{\Phi} \to \infty$ as $c \to 1$, although TDMI remains finite.

A.3.3 Integrated synergy ψ

Time-symmetric Proof by counterexample—for the AR system with

$$A = \begin{pmatrix} a & a \\ 0 & 0 \end{pmatrix} \quad , \quad \Sigma(\varepsilon) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We have $\psi = \frac{1}{2}\log\left(1+a^2\right)$, while, for the time-reversed process, $\psi = \frac{1}{2}\log\left(1+a^4\right)$. Note that this proof applies only to the MMI-PID used in this paper and presented in [20].

Non-negative Follows from $I_\cup(X, Y; Z) < I(\{X, Y\}; Z)$ [19].

Rescaling-invariant Follows from (MI-3) and the fact that $I_\cap$ is also invariant (see property (**Eq**) in Section 5 of [9]).

Bounded by TDMI Follows from the non-negativity of $I_\cup$ [19].

A.3.4 Decoder-based integrated information Φ^*

Non-negative Follows from $I^*[X;\tau,\mathcal{P}] \leq I(X_t;X_{t-\tau})$, proven in Reference [28].

Rescaling-invariant Assume that the measure is computed on a time series of rescaled data $X_t^r = X_t A$, where A is a diagonal matrix with positive real numbers. Then, its covariance is related to the covariance of the original time series as $\Sigma_X^r = \mathbb{E}\left[X_t^{r\mathrm{T}} X_t^r\right] = \mathbb{E}\left[A^{\mathrm{T}} X_t^{\mathrm{T}} X_t A\right] = A^2 \Sigma_X$. We can analogously calculate $\Pi_X, \Pi_{X\tilde{X}}, \Pi_{X|\tilde{X}}$ and easily verify that all A's cancel out, proving the invariance.

Bounded by TDMI Follows from $I^*[X;\tau,\mathcal{P}] \geq 0$, proven in Reference [28].

A.3.5 Geometric integrated information Φ_G

Time-symmetric Follows from the symmetry in the constraints that define the manifold of restricted models Q [11].

Non-negative Follows from the fact that Φ_G is an M-projection [11].

Rescaling-invariant Given a Gaussian distribution p with covariance Σ_p, its M-projection in Q is another Gaussian with covariance Σ_q. Given a new distribution p' formed by rescaling some of the variables in p, the M-projection of p' is a Gaussian with covariance $A^2\Sigma_q$ with A a diagonal positive matrix (see above), which satisfies $D_{KL}(p\|q) = D_{KL}(p'\|q')$ and therefore Φ_G is invariant to rescaling.

Bounded by TDMI TDMI can be defined as the M-projection of the full model p to a manifold of restricted models $Q^{MI} = \{q : q(X_t, X_{t-\tau}) = q(X_t)q(X_{t-\tau})\}$ [11]. The bound $\Phi_G \leq I(X_t;X_{t-\tau})$ follows from the fact that $Q^{MI} \subset Q$.

A.3.6 Causal density

Time-symmetric Follows from the non-symmetry of transfer entropy [217].

Non-negative Re-writing CD as a sum of conditional MI terms, follows from (MI-2).

Rescaling-invariant Follows from (MI-3).

Bounded by TDMI Proven in Section A.2.

Appendix B

Quantifying high-order interdependencies via multivariate extensions of the mutual information

B.1 Statistical structures across scales

In this section we study how the O-information is related to statistical structures of subsets of $\boldsymbol{X}^n$ – i.e. structures at different scales of the system. For simplicity, we assume in this section that $|\mathcal{X}|$ is finite.

In the next proposition we present some fundamental restrictions between the total correlation of subsystems and the value of $\Omega(\boldsymbol{X}^n)$.

Proposition 8. *If* $\Omega(\boldsymbol{X}^n) \geq 0$*, then for all* $m \in [n-1]$

$$\min_{|\boldsymbol{\gamma}|=m} \mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\gamma}}) \geq \Omega(\boldsymbol{X}^n) - (n-m-1)\log|\mathcal{X}| \,. \tag{B.1}$$

If $\Omega(\boldsymbol{X}^n) \leq 0$*, then for all* $m \in [n-1]$

$$\max_{|\boldsymbol{\gamma}|=m} \mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\gamma}}) \leq \Omega(\boldsymbol{X}^n) + (n-2)\log|\mathcal{X}| \,. \tag{B.2}$$

Both bounds are tight if $|\Omega| \geq (n-m+1)\log|\mathcal{X}|$.

Proof. See Appendix B.8. □

Corollary 3. *The following bounds hold for all $\boldsymbol{\gamma} \subseteq \{1,\dots,n\}$ with $|\boldsymbol{\gamma}| = m$:*

$$\begin{aligned}\min\left\{m-1, \frac{\Omega(\boldsymbol{X}^n)}{\log|\mathcal{X}|} + (n-2)\right\} &\geq \frac{\mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\gamma}})}{\log|\mathcal{X}|} \\ &\geq \max\left\{0, \frac{\Omega(\boldsymbol{X}^n)}{\log|\mathcal{X}|} - (n-m-1)\right\}.\end{aligned}$$

Corollary 3 shows that positive values of Ω constrain subgroups to be correlated: if $\Omega(\boldsymbol{X}^n) \geq (n-m-1)\log|\mathcal{X}|$ then all groups of m or more variables must have some statistical dependency. Negative values of Ω, on the other hand, impose limits on the allowed correlation strength: if $\Omega(\boldsymbol{X}^n) \leq -(n-m-1)\log|\mathcal{X}|$ then the correlation of all groups of m or more variables is upper-bounded. As an example, for $|\mathcal{X}| = 2$ and $m = 2$ the bounds given in Corollary 3 are

$$\begin{aligned}\max\{1, \Omega(\boldsymbol{X}^n) + n - 2\} &\geq I(X_i;X_j) \\ &\geq \min\{0, \Omega(\boldsymbol{X}^n) - (n-3)\}\ ,\end{aligned}$$

for all $i, j \in \{1,\dots,n\}$, which shows that the bounds related to Ω are only active when $n-3 \leq |\Omega| \leq n-2$.

In conclusion, the sign of Ω determines whether the constraint is a lower or upper bound, and $|\Omega|$ determines which scales of the system are affected, with smaller groups being harder to constrain – i.e. requiring higher absolute values of Ω. The relationship between the system's scales and the values of Ω is illustrated in Figure B.1.

The next result corresponds to the converse of Corollary 3, and shows how interactions at different scales limit the achievable values of Ω.

Corollary 4. *For a given $\boldsymbol{\gamma} \subset \{1,\dots,n\}$ with $|\boldsymbol{\gamma}| = m$, the following bounds on Ω hold:*

$$n-m-1+\frac{\mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\gamma}})}{\log|\mathcal{X}|} \geq \frac{\Omega(\boldsymbol{X}^n)}{\log|\mathcal{X}|} \geq -(n-2)+\frac{\mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\gamma}})}{\log|\mathcal{X}|}.$$

By comparing it with Lemma 3, this result shows that a large $\mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\gamma}})$ does not allow Ω to reach its lower bound. On the other hand, small values of $\mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\gamma}})$ decrease the upper bound, forbidding high values of Ω. Additionally, note that fixing the value of only one subset of m variables reduces the range of values of Ω from $2(n-2)$ to $2(n-2)-(m-1)$. The following example illustates these findings.

Example 3. *Let us consider a system $\boldsymbol{X}^n$ of binary variables, two of which are related by the marginal distribution*

$$p_{X_1X_2}(x_1,x_2) = \frac{(1-\eta)^{1-|x_1-x_2|}\eta^{|x_1-x_2|}}{2}\ .$$

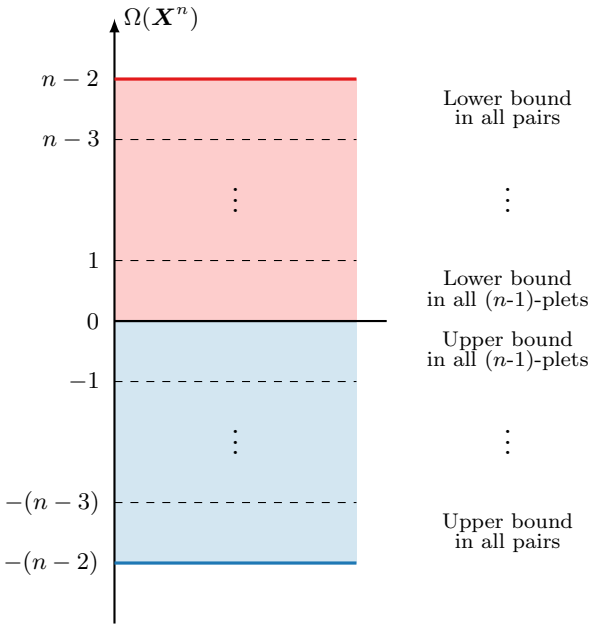


Figure B.1: Diagram of how values of the O-information impose limits on the strength of interactions – as measured by $\mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\gamma}})$ – at different scales. Positive (negative) values of Ω put lower (upper) bounds on subsets of $\boldsymbol{X}^n$, and higher absolute values of Ω put bounds on subsystems of smaller sizes.

That is, X_1 and X_2 are fair coins linked by a binary symmetric channel with crossover probability η [7, Sec. 7]. Hence, $\mathrm{TC}(\boldsymbol{X}^2) = I(X_1;X_2) = 1-H(\eta)$, with $H(\eta) = -\eta\log\eta - (1-\eta)\log(1-\eta)$ being the binary entropy function. By considering $m=2$, Corollary 4 states that

$$n-2-H(\eta) \geq \Omega(\boldsymbol{X}^n) \geq -\big(n-3+H(\eta)\big)\ ,$$

which is illustrated in Figure B.2. Moreover, using Eq. (8.15) one can verify that the upper bound (solid red line) is attained when $X_2 = X_3 = \ldots = X_n$, while the lower bound (solid blue line) is attained when $X_3, \ldots, X_{n-1}$ are independent fair coins and $X_n = \sum_{j=1}^{n-1} X_j \pmod 2$.[1]

B.2 Ω as a superposition of tendencies

This section explores sufficient conditions that make a system have a small O-information. As a preliminary step, the next result shows that Ω is additive for systems with independent subsystems.

[1] Interestingly, despite the correlation between X_1 and X_2, an n-bit `xor` still enables the most synergistic configuration attainable.

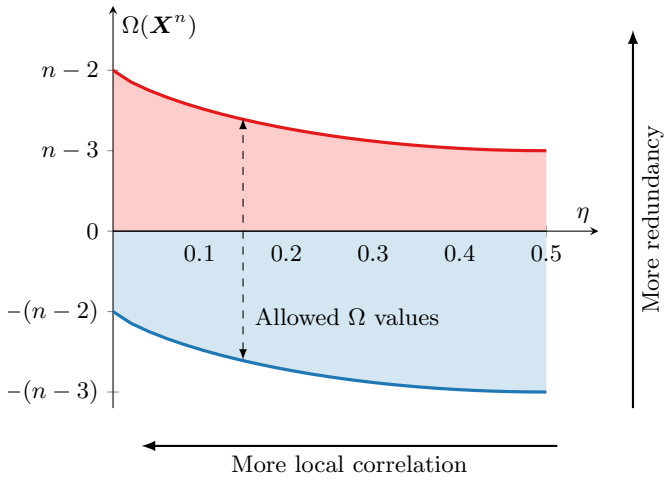


Figure B.2: Bounds of the O-information when two variables are connected via a binary symmetric channel with crossover probability η (see Example 3).

Lemma 4. *If* $p_{\boldsymbol{X}^n}(\boldsymbol{x}^n) = \prod_{k=1}^{m} p_{\boldsymbol{X}^{\boldsymbol{\alpha}_k}}(\boldsymbol{x}^{\boldsymbol{\alpha}_k})$ *for some partition* $\pi = (\boldsymbol{\alpha}_1|\ldots|\boldsymbol{\alpha}_m)$*, then*

$$\Omega(\boldsymbol{X}^n) = \sum_{k=1}^{m} \Omega(\boldsymbol{X}^{\boldsymbol{\alpha}_k}) \ .$$

Proof. Let us consider the case $\pi = (\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$, as the general case is then guaranteed by induction. Using Eqs. (8.13) and (8.14) it is direct to check that, due to the independence, $\mathrm{TC}(\boldsymbol{X}^n) = \mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\alpha}_1}) + \mathrm{TC}(\boldsymbol{X}^{\boldsymbol{\alpha}_2})$ and $\mathrm{DTC}(\boldsymbol{X}^n) = \mathrm{DTC}(\boldsymbol{X}^{\boldsymbol{\alpha}_1}) + \mathrm{DTC}(\boldsymbol{X}^{\boldsymbol{\alpha}_2})$. Then, the desired result follows from induction on the number of cells and the definition of Ω. □

Corollary 5. $\Omega(\boldsymbol{X}^n) = 0$ *for all systems whose joint distribution can be factorised as*

$$p_{\boldsymbol{X}^n}(\boldsymbol{x}^n) = \prod_{k=1}^{n/2} p_{X_{2k-1}X_{2k}}(x_{2k-1}, x_{2k}) \ . \tag{B.3}$$

Proof. Using Eq. (B.3) and Lemma 4 we find that

$$\Omega(\boldsymbol{X}^n) = \sum_{k=1}^{n/2} \Omega(X_{2k-1}, X_{2k}) = 0 \ \ ,$$

where the last equality is a consequence of the O-information being zero for sets of two variables, as shown in Proposition 1. □

Corollary 5 states that having disjoint pairwise interactions is a sufficient condition for $\Omega = 0$ to hold. However, this condition is not necessary: from Lemma 4 we can see that a system composed by redundant ($\Omega > 0$) and synergistic ($\Omega < 0$) subsystems can attain zero net O-information due to "destructive interference."

As a consequence, the O-information can be understood as the result of a superposition of behaviours of subsystems. Therefore, $\Omega = 0$ can take place in two qualitatively different scenarios: systems in which redundancies and synergies are balanced, or systems with only disjoint pairwise effects. Some of these cases can be resolved by considering the information diagram of $\mathrm{TC}(\boldsymbol{X}^n)$ and $\mathrm{DTC}(\boldsymbol{X}^n)$ (c.f. Figure 8.2), or by studying the O-information of parts of the system. However, it is important to remark that redudancy and synergy can coexist either in disjoint subsystems or within the same variables. An insightful example of the latter case can be found in Ref. [218, Section 2].

As a final remark, note that systems where pairwise interdependencies are overlapping (e.g. pairwise maximum entropy models [219]) cannot be factorised as required by Corollary 5, and hence can have either positive or negative O-information.[2]

B.3 $R(\pi)$ decreases for finer partitions

Lemma 5. *Let us consider two partitions* $\pi_a = (\boldsymbol{\alpha}_1|...|\boldsymbol{\alpha}_K)$ *and* $\pi_b = (\boldsymbol{\beta}_1|...|\boldsymbol{\beta}_J)$ *such that* $\pi_b \succeq \pi_a$. *Then,* $R(\pi_b) \leq R(\pi_a)$.

Proof. Let us assume that $\pi_\mathrm{a} = (\boldsymbol{\alpha}_1|...|\boldsymbol{\alpha}_K\}$, $\pi_\mathrm{b} = (\boldsymbol{\beta}_1|...|\boldsymbol{\beta}_J)$ such that $\pi_\mathrm{b} \succeq \pi_\mathrm{a}$, and consider a path $\mathtt{p} = (\pi_1, ..., \pi_L)$ in $\mathtt{P}(\pi_\mathrm{a}, \pi_\mathrm{b})$ so that $\pi_1 = \pi_\mathrm{a}$ and $\pi_L = \pi_\mathrm{b}$. To prove the Lemma suffices to show that $R(\pi_{j+1}) \leq R(\pi_j)$ for $j = 1, ..., L-1$. As $\pi_1, ..., \pi_n$ are related by covering relationships, one just needs to prove the inequality for two partitions such that one covers the other.

Consider $\pi_1, \pi_2 \in \mathcal{P}_n$ such that π_2 covers π_1. As both partitions differ only in one elementary refinement, let us without loss of generality assume that the refinement is done on the last cell of π_1; i.e. $\pi_1 = (\boldsymbol{\alpha}_1|...|\boldsymbol{\alpha}_m)$ and $\pi_2 = (\boldsymbol{\alpha}_1|...|\boldsymbol{\alpha}_{m-1}|\tilde{\boldsymbol{\alpha}}_m|\tilde{\boldsymbol{\alpha}}_{m+1})$ so that $\tilde{\boldsymbol{\alpha}}_m \cup \tilde{\boldsymbol{\alpha}}_{m+1} = \boldsymbol{\alpha}_m$ and $\tilde{\boldsymbol{\alpha}}_m \cap \tilde{\boldsymbol{\alpha}}_{m+1} = \varnothing$. Then

$$\begin{aligned} R(\pi_1) - R(\pi_2) &= R_{\boldsymbol{\alpha}_m} - (R_{\tilde{\boldsymbol{\alpha}}_m} + R_{\tilde{\boldsymbol{\alpha}}_{m+1}}) \\ &= I(\boldsymbol{X}^{\tilde{\boldsymbol{\alpha}}_m}; \boldsymbol{X}^{\tilde{\boldsymbol{\alpha}}_{m+1}} | \boldsymbol{X}^{\boldsymbol{\alpha}_1} ... \boldsymbol{X}^{\boldsymbol{\alpha}_{m-1}}) \\ &\geq 0 \ , \end{aligned}$$

proving the desired result. □

[2] For a detailed discussion of this issue for the case of three variables see [23, Sec. 5].

B.4 Proof of Lemma 2

Proof. Consider a path $\mathtt{p} \in \mathtt{P}(\pi_{\text{source}}, \pi_{\text{sink}})$, so that $\mathtt{p} = (\pi_1, ..., \pi_L)$ with $\pi_1 = \pi_{\text{source}}$ and $\pi_L = \pi_{\text{sink}}$. Then, by using Eqs. (8.7) and (8.8), a direct calculation shows that

$$\begin{aligned} W(\mathtt{p}; v_h) &= \sum_{j=1}^{L-1} \left[H(\pi_{j+1}) - H(\pi_j) \right] \\ &= H(\pi_{\text{sink}}) - H(\pi_{\text{source}}) \\ &= \sum_{i=1}^{n} H(X_i) - H(\boldsymbol{X}^n) \ . \end{aligned}$$

Similarly, using Eqs. (8.7) and (8.9) gives

$$\begin{aligned} W(\mathtt{p}; v_{\mathrm{r}}) = &\sum_{j=1}^{L-1} \left[R(\pi_j) - R(\pi_{j+1}) \right] \\ =&R(\pi_{\text{source}}) - R(\pi_{\text{sink}}) \\ =&H(\boldsymbol{X}^n) - \sum_{i=1}^{n} H(X_i | \boldsymbol{X}^n_{-i}) \ . \end{aligned}$$

Both results make use of the fact that $W(\mathtt{p}; v_{\mathrm{h}})$ and $W(\mathtt{p}; v_{\mathrm{r}})$ are telescopic sums and all but the first and last terms cancel out. □

B.5 Proof of Proposition 1

Proof. Let us consider a path $\mathtt{p} \in \mathtt{P}(\pi_{\text{source}}, \pi_{\text{sink}})$. Then,

$$\begin{aligned} W(\mathtt{p}; v_{\mathrm{s}}) &= \sum_{j=1}^{L} v_{\mathrm{s}}(\pi_j, \pi_{j+1}) \\ &= \sum_{j=1}^{L} v_{\mathrm{h}}(\pi_j, \pi_{j+1}) - \sum_{k=1}^{L} v_{\mathrm{r}}(\pi_k, \pi_{k+1}) \\ &= \mathrm{TC}(\boldsymbol{X}^n) - \mathrm{DTC}(\boldsymbol{X}^n) = \Omega(\boldsymbol{X}^n), \end{aligned} \tag{B.4}$$

which proves the first part of the theorem.

Thanks to Eq. (B.4), one can prove the second part of the Theorem by showing that if $\pi_{\mathrm{a}}, \pi_{\mathrm{b}} \in \mathcal{P}_n$ such that $\pi_{\mathrm{b}} \succeq \pi_{\mathrm{a}}$, then $v_{\mathrm{s}}(\pi_1, \pi_2)$ is equal to an interaction information. To show this, first note that if $\pi_{\mathrm{b}} \succeq \pi_{\mathrm{a}}$ then both partitions differ only in one elementary refinement. Without no loss of generality, we assume that the refinement is done on the last cell, such

that $\pi_{\mathrm{a}} = (\boldsymbol{\alpha}_1|...|\boldsymbol{\alpha}_m)$ and $\pi_{\mathrm{b}} = (\boldsymbol{\alpha}_1|...|\boldsymbol{\alpha}_{m-1}|\tilde{\boldsymbol{\alpha}}_m|\tilde{\boldsymbol{\alpha}}_{m+1})$ such that $\tilde{\boldsymbol{\alpha}}_m \cap \tilde{\boldsymbol{\alpha}}_{m+1} = \varnothing$ and $\tilde{\boldsymbol{\alpha}}_m \cup \tilde{\boldsymbol{\alpha}}_{m+1} = \boldsymbol{\alpha}_m$. Then,

$$\begin{aligned} v_{\mathrm{s}}(\pi_{\mathrm{a}},\pi_{\mathrm{b}}) &= v_{\mathrm{h}}(\pi_{\mathrm{a}},\pi_{\mathrm{b}}) - v_{\mathrm{r}}(\pi_{\mathrm{a}},\pi_{\mathrm{b}}) \\ &= \big[H(\pi_{\mathrm{b}}) - H(\pi_{\mathrm{a}})\big] - \big[R(\pi_{\mathrm{a}}) - R(\pi_{\mathrm{b}})\big] \\ &= I(\boldsymbol{X}^{\tilde{\boldsymbol{\alpha}}_m};\boldsymbol{X}^{\tilde{\boldsymbol{\alpha}}_{m+1}}) \\ &\quad - I(\boldsymbol{X}^{\tilde{\boldsymbol{\alpha}}_m};\boldsymbol{X}^{\tilde{\boldsymbol{\alpha}}_{m+1}}|\boldsymbol{X}^{\boldsymbol{\alpha}_1}...\boldsymbol{X}^{\boldsymbol{\alpha}_{m-1}}) \\ &= I(\boldsymbol{X}^{\tilde{\boldsymbol{\alpha}}_m};\boldsymbol{X}^{\tilde{\boldsymbol{\alpha}}_{m+1}};\boldsymbol{X}^{\boldsymbol{\alpha}_1}...\boldsymbol{X}^{\boldsymbol{\alpha}_{m-1}}) \ , \end{aligned}$$

which proves the desired result. □

B.6 Proof of Lemma 3

Proof. Let us first note that

$$\log|\mathcal{X}| \geq I(X_i;X_j|X_k) \geq 0 \ , \tag{B.5}$$

$$\log|\mathcal{X}| \geq I(X_i;X_j;X_k) \geq -\log|\mathcal{X}| \ , \tag{B.6}$$

for all $i,j,k \in \{1,...,n\}$. Above, Eq. (B.6) follows from noting that $I(X_i;X_j;X_k) = I(X_i;X_j) - I(X_i;X_j|X_k)$, and applying the bounds in Eq. (B.5). The lemma is proved by applying these inequalities on Eqs. (8.13), (8.14), (8.15), and (8.19). Finally, the tightness of the bounds is a direct consequence of the tightness of Eqs. (B.5) and (B.6). □

B.7 Proof of Proposition 2

Proof. Let us first prove the first statement. By considering $\boldsymbol{X}^n$ to be a n-bit copy, a direct calculation using Eqs. (8.13) and (8.14) shows that $\mathrm{TC}(\boldsymbol{X}^n) = n-1$ and $\mathrm{DTC}(\boldsymbol{X}^n) = 1$, and therefore the upper bound is attained. To prove the converse, let us start by assuming that $\Omega(\boldsymbol{X}^n) = n-2$. By applying (B.6) to each term in (8.15), is clear that $I(X_j;\boldsymbol{X}^{j-1};\boldsymbol{X}^n_{j+1}) = 1$ holds for all $j \in \{1,...,n\}$. In particular $I(X_2;X_1;\boldsymbol{X}^n_3) = 1$ holds, which due to Eq. (8.15) implies that $I(X_2;X_1|\boldsymbol{X}^n_3) = 0$ and hence $I(X_2;X_1) = 1$, which in turns implies that X_1 and X_2 are Bernoulli distributed with parameter $p = 1/2$, and also that $X_1 = X_2$. By relabelling the variables and following the same rationale one can prove that every pair of variables are equal to each other, which proves that $\boldsymbol{X}^n$ is a n-bit copy.

Let us prove the second statement. By considering now $\boldsymbol{X}^n$ to be a n-bit `xor`, using Eqs. (8.13) and (8.14) it is direct to check that $\mathrm{TC}(\boldsymbol{X}^n) = 1$ and $\mathrm{DTC}(\boldsymbol{X}^n) = n-1$, and hence the lower bound is attained. To prove the converse, let us assume that $\boldsymbol{X}^n$ is such that

$\Omega(\boldsymbol{X}^n) = 2-n$. By considering the bounds given by Eq. (B.6) in Eq. (8.15), this implies that $I(X_j;\boldsymbol{X}^{j-1};\boldsymbol{X}_{j+1}^n) = -1$ for all $j \in \{2,\ldots,n-1\}$, and in particular $I(\boldsymbol{X}^{n-2};X_{n-1};X_n) = -1$. Due to Eq. (B.6), this implies in turn that $I(\boldsymbol{X}^{n-2};X_{n-1}) = 0$, and via relabeling one can prove that $\boldsymbol{X}^{n-1}$ are jointly independent. Moreover, $I(\boldsymbol{X}^{n-2};X_{n-1};X_n) = -1$ also implies that $I(X_{n-1};X_n|\boldsymbol{X}^{n-2}) = 1$, which implies that

$$I(\boldsymbol{X}^{n-1};X_n) = I(X_{n-1};X_n|\boldsymbol{X}^{n-2}) + I(\boldsymbol{X}^{n-2};X_n) = 1.$$

This equality implies that X_n is Bernoulli distributed with $p = 1/2$, and that X_n is a deterministic function of $\boldsymbol{X}^{n-1}$. Moreover, the fact that $I(X_1;X_n|\boldsymbol{X}_2^{n-1}) = 1$ implies that for given $\boldsymbol{X}_2^{n-1}$ then X_n is a function of X_1, while via relabelling one finds that $I(X_1;X_n) = 0$. Since the only functions with these properties are functions isomorphic to an n-variate `xor`, this proves the desired result. □

B.8 Proof of Proposition 8

The following proof uses Lemma 6, which is stated and proved afterwards in this Appendix.

Proof. To prove Eq. (B.1), first note that

$$\Omega(\boldsymbol{X}^n) = \mathrm{TC}(\boldsymbol{X}^{n-1}) - \mathrm{DTC}(\boldsymbol{X}^{n-1}|X_n) \leq \mathrm{TC}(\boldsymbol{X}^{n-1}) \ .$$

Then, the inequality follows form a direct application of Lemma 6. As $\mathrm{TC}(\boldsymbol{X}^m) \geq 0$, the equality becomes non-trivial when

$$\Omega(\boldsymbol{X}^n) - (n-m-1)\log|\mathcal{X}| \geq 0 \ .$$

To prove Eq. (B.2), note that by using Eqs. (8.13), (8.14), and (8.15) one can find that

$$\begin{aligned}\Omega(\boldsymbol{X}^n) =& \mathrm{TC}(\boldsymbol{X}^m) - \mathrm{DTC}(\boldsymbol{X}^m|\boldsymbol{X}_{m+1}^n) \\ &+ \sum_{j=m+1}^{n-1} I(X_j;\boldsymbol{X}^{j-1};\boldsymbol{X}_{j+1}^n) \\ \geq& \mathrm{TC}(\boldsymbol{X}^m) - (n-2)\log|\mathcal{X}|.\end{aligned}$$

Above, the inequality is due to $I(X_j;\boldsymbol{X}^{j-1};\boldsymbol{X}_{j+1}^n) \leq \log|\mathcal{X}|$ and $\mathrm{DTC}(\boldsymbol{X}^m|\boldsymbol{X}_{m+1}^n) \leq (m-1)\log|\mathcal{X}|$. As the above relationship does not depend on the labelling of the X's, this proves Eq. (B.2). As $\mathrm{TC}(\boldsymbol{X}^m) \leq (m-1)\log|\mathcal{X}|$, the equality becomes non-trivial when

$$\Omega(\boldsymbol{X}^n) + (n-2)\log|\mathcal{X}| \leq (m-1)\log|\mathcal{X}| \ .$$

□

Lemma 6. *If* $|\mathcal{X}| = \min_{i=1,\dots,n} |\mathcal{X}_i|$, *then*

$$\min_{|\boldsymbol{\gamma}|=m} \mathrm{TC}(\boldsymbol{X}^{\gamma}) \geq \mathrm{TC}(\boldsymbol{X}^n) - (n-m)\log|\mathcal{X}| \ .$$

Proof. A direct calculation using Eq. (8.13) shows that

$$\begin{aligned}\mathrm{TC}(\boldsymbol{X}^n) &= \mathrm{TC}(\boldsymbol{X}^m) + \sum_{j=m+1}^{n} I(X_j; \boldsymbol{X}^{j-1}) \\ &\leq \mathrm{TC}(\boldsymbol{X}^m) + (n-m)\log|\mathcal{X}|.\end{aligned}$$

As the labelling of the indices can be modified without changing this result, this suffices to prove the desired result. □

Appendix C

Integrated information decomposition

C.1 The product of two lattices is a lattice

A lattice is a partially ordered set $(\mathcal{A}, \preceq)$ for which every pair of elements a, b has a well-defined *meet* $a \wedge b$ and *join* $a \vee b$, which correspond to their common greatest lower bound (infimum) and common least upper bound (supremum), respectively [163]. Here we prove that, if $(\mathcal{A}, \preceq)$ is a lattice, then the product lattice $(\mathcal{A} \times \mathcal{A}, \preceq^*)$ equipped with the order relationship

$$\alpha \to \beta \preceq^* \alpha' \to \beta' \quad \text{if and only if} \quad \alpha \preceq \alpha' \text{ and } \beta \preceq \beta', \tag{C.1}$$

is also a lattice, where $\alpha, \beta, \alpha', \beta' \in \mathcal{A}$. As a corollary of this, given that the set and partial ordering relationship used in PID are a lattice [19, 220], then the set and partial ordering relationship used in ΦID are also a lattice.

For compactness, let us use the notation $\gamma = \alpha \to \beta$ and $\gamma' = \alpha' \to \beta'$ for $\gamma, \gamma' \in \mathcal{A} \times \mathcal{A}$. To prove the lattice structure of $(\mathcal{A} \times \mathcal{A}, \preceq^*)$ it suffices to show that

1. $\gamma \wedge^* \gamma' := \alpha \wedge \alpha' \to \beta \wedge \beta'$ is a valid meet; and
2. $\gamma \vee^* \gamma' := \alpha \vee \alpha' \to \beta \vee \beta'$ is a valid join.

Note that the fact that $(\mathcal{A}, \preceq)$ is a lattice implies that $\alpha \wedge \beta$ and $\alpha \vee \beta$ are well-defined for all $\alpha, \beta \in \mathcal{A}$.

Let us begin with the meet, for which we use $m = \gamma \wedge^* \gamma'$ as a shorthand notation. First, one can directly check that $m \preceq^* \gamma$ and $m \preceq^* \gamma'$, given the definition of $\preceq^*$ above and the fact that $\alpha \wedge \alpha' \preceq \alpha$ (and similarly for α', β, and β'). Next, we need to prove that for any $\gamma'' = \alpha'' \to \beta'' \in \mathcal{A} \times \mathcal{A}$ such that $\gamma'' \preceq^* \gamma$ and $\gamma'' \preceq^* \gamma'$, we have $\gamma'' \preceq^* m$ (i.e. that m is the greatest lower bound of γ and γ'). To see this, note that the conditions $\gamma'' \preceq^* \gamma$ and $\gamma'' \preceq^* \gamma'$

imply the following four statements:

$$\begin{aligned}\alpha'' &\preceq \alpha\ ,\\ \alpha'' &\preceq \alpha'\ ,\\ \beta'' &\preceq \beta\ ,\\ \beta'' &\preceq \beta'\ .\end{aligned}$$

Using these relationships and the $\wedge$ operator from $\mathcal{A}$, one can show that $\alpha'' \preceq \alpha \wedge \alpha'$ and $\beta'' \preceq \beta \wedge \beta'$, which in turn implies that $\gamma'' \preceq^* m$. Finally, the proof for the join is analogous, replacing $\wedge$ with $\vee$ and $\preceq$ with $\succeq$.

C.2 Decomposing PID atoms

Equation (4) in the main text shows how to decompose redundancies in the product lattice in terms of ΦID atoms. Here we provide a more general statement, that allows us to decompose not only redundancies, but also other PID atoms. The goal of this appendix is to build stronger connections between PID and ΦID, and to extend Proposition 1 to allow greater flexibility for specifying a ΦID function.

For the forward PID, and borrowing the notation from Williams and Beer [19], given a non-empty set of 'future' variables $F \in \mathcal{P}(\{Y_1,...,Y_N\})$ and an an element of the redundancy lattice $\alpha \in \mathcal{A}$, let us denote by $\Pi_F(\alpha;F)$ the α atom of the PID decomposition for $I(\boldsymbol{X};F)$, such that

$$I(\boldsymbol{X};F) = \sum_{\alpha\in\mathcal{A}} \Pi_F(\alpha;F)\ . \tag{C.2}$$

We use an analogous notation for the backward PID, with a corresponding non-empty set of 'past' variables $P \in \mathcal{P}(\{X_1,...,X_N\})$ and $\beta \in \mathcal{A}$, such that

$$I(P;\boldsymbol{Y}) = \sum_{\beta\in\mathcal{A}} \Pi_B(P;\beta)\ . \tag{C.3}$$

Then, these quantities can be further decomposed in ΦID atoms as

$$\Pi_F(\alpha;F) = \sum_{\gamma\preceq F} I_\partial^{\alpha\to\gamma}\ , \tag{C.4a}$$

$$\Pi_B(P;\beta) = \sum_{\gamma\preceq P} I_\partial^{\gamma\to\beta}\ . \tag{C.4b}$$

Note that the sum runs only across one of the sets (instead of both as it does in Eq. (4) of the main text), and that every element in $\mathcal{P}(\{1,...,N\})$ is also in $\mathcal{A}$, and hence the partial order relationship in the sums above is well-defined. As a few examples, in a bivariate system the following forward PID atoms decompose as:

$$\texttt{Red}(X_1,X_2;Y_i) = \Pi_F(\{1\}\{2\};Y_i) = I_\partial^{\{1\}\{2\}\to\{1\}\{2\}} + I_\partial^{\{1\}\{2\}\to\{i\}} \ ,$$

$$\texttt{Syn}(X_1,X_2;Y_i) = \Pi_F(\{12\};Y_i) = I_\partial^{\{12\}\to\{1\}\{2\}} + I_\partial^{\{12\}\to\{i\}} \ ,$$

$$\texttt{Un}(X_1;Y_1Y_2|X_2) = \Pi_F(\{1\};Y_1Y_2) = I_\partial^{\{1\}\to\{1\}\{2\}} + I_\partial^{\{1\}\to\{1\}} + I_\partial^{\{1\}\to\{2\}} + I_\partial^{\{1\}\to\{12\}} \ .$$

These decompositions can be used to prove Proposition 1 of the main text. Adopting a view of ΦID as a linear system of equations, one needs 16 independent equations to solve for the 16 unknowns that are the ΦID atoms. Of those, 9 are given by standard Shannon mutual information (specifically, $I(X_i;Y_j)$, $I(X_1X_2;Y_i)$, $I(Y_1Y_2;X_i)$, and $I(X_1X_2;Y_1Y_2)$, for $i,j=\{1,2\}$) decomposed with Eq. (4) of the main text, and 6 are given by the single-target PIDs ($\texttt{Red}(X_1,X_2;Y_1)$, $\texttt{Red}(X_1,X_2;Y_2)$, and $\texttt{Red}(X_1,X_2;Y_1Y_2)$, as well as the 3 corresponding backward PIDs) decomposed by the expression above. Finally, one only need to add one individual ΦID atom to make the 16 equations needed, and the system can be solved for all other atoms.

Taking these results together, Proposition 1 in the main text can be generalised as follows: a valid ΦID can be defined not only in terms of redundancy, but also in terms of unique information or synergy. This is equivalent to the case of PID, for which decompositions based on unique information [173] or synergy [221, 210] have been proposed. In fact, for the numerical results in Fig. 5 of the main text we use a ΦID based on unique information defined below.

C.3 Computing the ΦID atoms

In Ref. [173], James, Emenheiser and Crutchfield introduce a PID based on a new measure of unique information, I_{dep}, which we succinctly describe here. To define I_{dep}, they first define a *constraint lattice* $\mathcal{L}$ on a set of variables (formally defined as the set of antichain covers with the natural partial ordering). Specifically, given a constraint σ and a probability distribution p, consider the set $\Delta_p(\sigma)$ of distributions that match marginals in σ with p:

$$\Delta_p(\sigma) = \{q : p(\gamma) = q(\gamma), \gamma \in \sigma\} \ .$$

For example, the constraint $\sigma = \{(X,Y),(X,Z)\}$ determines the set of distributions q such that $q(x,y) = p(x,y)$ and $q(x,z) = p(x,z)$. In addition, the elements of $\mathcal{L}$ (i.e. the nodes in

the lattice) have an associated value of an information-theoretic measure $f[p_\sigma]$ evaluated on $p_\sigma = \arg\max\{H[q] : q \in \Delta_p(\sigma)\}$.

Let us focus on the bivariate PID: denote by L the collection of edges of the constraint lattice for the variables X, Y, Z, and let f be the joint mutual information $I(XY;Z)$. For a link $(\sigma_1, \sigma_2) \in L$, one can evaluate the change in f along the link via the operator $\Delta_{\sigma_2}^{\sigma_1}$; e.g. $\Delta_{\sigma_2}^{\sigma_1} I(XY;Z) = I_{\sigma_1}(XY;Z) - I_{\sigma_2}(XY;Z)$. Additionally, for any $\gamma \in \mathcal{P}(\{X,Y,Z\})$ let us define $E(\gamma)$ to be the set of all links that contain γ only at one side, i.e.

$$E(\gamma) = \{(\sigma_1, \sigma_2) \in L : \gamma \in \sigma_1, \gamma \notin \sigma_2\}. \tag{C.5}$$

Then, the unique information is defined by

$$I_{\text{dep}}(X \to Z|Y) = \min_{(\sigma_1,\sigma_2)\in E(X,Z)} \Delta_{\sigma_2}^{\sigma_1} I(XY;Z)\ . \tag{C.6}$$

That is, the unique information is the smallest perturbation that is seen when adding the dependency between X and Z. For further details, and a more pedagogical introduction, we refer the reader to the original paper [173].

This measure can be naturally generalized to the ΦID setting by replacing $I(XY;Z)$ above with the full joint mutual information $I(\boldsymbol{X};\boldsymbol{Y})$ and formulating the appropriate constraint lattice for $(\boldsymbol{X},\boldsymbol{Y})$. More precisely:

Definition 5. ***Double-unique information based on dependencies***. *For a given set of variables $(\boldsymbol{X},\boldsymbol{Y})$, and two indices i and j, the double-unique information based on dependencies is defined as*

$$I_{\partial,\text{dep}}^{\{i\}\to\{j\}} := \min_{(\sigma_1,\sigma_2)\in E(X_i,Y_j)} \Delta_{\sigma_2}^{\sigma_1} I(\boldsymbol{X};\boldsymbol{Y}). \tag{C.7}$$

This definition is applicable to any probability distribution, on either both discrete and continuous random variables. In practice, the difficulty of calculating I_{dep} amounts to the difficulty of calculating maximum-entropy projections, which for Gaussian and discrete distributions is easily done with off-the-shelf software – in the case of discrete variables, for example using the `dit` package [171]. Once the double-unique information has been calculated, the same lattice can be reused to compute the unique information atoms for all 6 single-target PIDs, and together with the 9 MIs, these 16 numbers fully determine the numerical values of every ΦID atom.

It is important to recall that, as mentioned in the main body of the paper, the two axioms of ΦID do not uniquely determine $I_\partial^{\{i\}\to\{j\}}$. An exploration of alternative decompositions and their theoretical and practical implications for ΦID will be covered in a separate publication.

C.4 Results of section 'Different types of integration'

Here we present calculations for the example systems in Fig. 4 of the main text. These proofs hold for all ΦID that satisfy the partial ordering axiom of $I_\cap^{\alpha\to\beta}$ (Axiom 2 in the main text), have a non-negative double-redundancy function $I^{\{1\}\{2\}\to\{1\}\{2\}} \geq 0$, and satisfy the following bound that follows from the basic properties of PID (c.f. [23]):

$$\texttt{Red}(X,Y;Z) \leq \min\{I(X;Z), I(Y;Z)\} \,. \tag{C.8}$$

Let us examine the three systems in turn:

- For the copy transfer system, $Y_2 = X_1$, while X_2 and Y_2 are independent i.i.d. fair coin flips. Since Y_2 is independent from the rest of the system, $\texttt{Red}(X_1,X_2;Y_2) = \texttt{Red}(X_1,X_2;Y_2) = 0$, and due to partial ordering $I_\cap^{\{1\}\{2\}\to\{1\}\{2\}} = 0$. Finally, using the Moebius inversion formula it follows that $I_\partial^{\{1\}\to\{2\}} = I(X_1;Y_2) = 1$ and all other atoms are zero.

- In the downward XOR system, X_1 and X_2 are i.i.d. fair coin flips, $Y_1 = X_1 \oplus X_2$, and Y_2 is independent of the rest. Then, it is clear that $I(X_1,X_2;Y_1,Y_2) = I(X_1,X_2;Y_1) = 1$, while $I(X_1;Y_1) = I(X_2;Y_1) = 0$. Additionally, note that $I_\cap^{\{12\}\to\{1\}\{2\}} = 0$, since $\texttt{Red}(Y_1,Y_2;X_1X_2) \leq I(Y_2;X_1X_2) = 0$. All this implies that all the redundancies (and hence all the atoms) below $\{12\}\to\{1\}$ are zero, and hence $I_\partial^{\{12\}\to\{1\}} = 1$ due to the Moebius inversion formula.

- Finally, consider the PPR system where X_1, X_2, Y_1 are i.i.d. fair coin flips and Y_2 is such that $X_1 \oplus X_2 = Y_1 \oplus Y_2$. Then $I(X_1,X_2;Y_1) = I(X_1,X_2;Y_2) = I(X_1;Y_1,Y_2) = I(X_2;Y_1,Y_2) = 0$. This implies that all redundancies (and hence atoms) except $I_\cap^{\{12\}\to\{12\}}$ are zero, and hence using again the Moebius inversion formula $I_\partial^{\{12\}\to\{12\}} = I(X_1,X_2;Y_1,Y_2) = 1$.

C.5 Results of section 'Measures of integrated information'

In this appendix we prove the results in Table 1 of the main text, that shows whether each of four measures of integrated information (Φ, CD, ψ, Φ_G) are positive, negative, or zero in a system containing only one

ΦID atom. A succinct definition of each measure is given below, and a comprehensive review and comparison of these and other measures can be found in Ref. [172].

Throughout this section we focus on bivariate systems, and use i, j as variable indices, with $i \neq j$. To complete the proof we will first show that it is possible to build systems with exactly one bit of information in one ΦID atom, and we will then compute the four measures on those systems.

Let us begin with the design of systems with one specific ΦID atom. Intuitively, this can be accomplished with a suitable combination of COPY and XOR gates for redundant and synergistic sets of variables, respectively. More formally, the procedure to build a system with $I_\partial^{\alpha\to\beta} = 1$ and all other atoms equal to zero is as follows:

1. Sample w from a Bernoulli distribution with $p = 0.5$.
2. Sample $\boldsymbol{x}$ based on α:
   - If $\alpha = \{1\}\{2\}$, then $x_1 = x_2 = w$.
   - If $\alpha = \{i\}$, then $x_i = w$ and x_j is sampled from a Bernoulli distribution with $p = 0.5$.
   - If $\alpha = \{12\}$, then $\boldsymbol{x}$ is a random string with parity w.
3. Sample $\boldsymbol{y}$ based on β analogously.

In all cases there will be one bit of information (w) shared between $\boldsymbol{X}$ and $\boldsymbol{Y}$, hence $I(\boldsymbol{X};\boldsymbol{Y}) = 1$ for any choice of α, β. This can be proven using the fact that for any α, β, one has $H(W) = 1$, $H(W|\boldsymbol{X}) = H(W|\boldsymbol{Y}) = 0$, and $p(\boldsymbol{x}, \boldsymbol{y}, w) = p(\boldsymbol{x}|w)p(\boldsymbol{y}|w)p(w)$. To do so, let us start from the mutual information chain rule:

$$\begin{aligned} I(\boldsymbol{X};\boldsymbol{Y}W) &= I(\boldsymbol{X};W) + I(\boldsymbol{X};\boldsymbol{Y}|W) \\ &= I(\boldsymbol{X};\boldsymbol{Y}) + I(\boldsymbol{X};W|\boldsymbol{Y}) \ . \end{aligned}$$

Rearranging the above terms, one can find that

$$I(\boldsymbol{X};\boldsymbol{Y}) = I(\boldsymbol{X};W) + I(\boldsymbol{X};\boldsymbol{Y}|W) - I(\boldsymbol{X};W|\boldsymbol{Y}) \ ,$$

where $I(\boldsymbol{X};W) = H(W) - H(W|\boldsymbol{X}) = 1$ and $I(\boldsymbol{X};\boldsymbol{Y}|W) = 0$. Finally, one finds that

$$\begin{aligned} I(\boldsymbol{X};W|\boldsymbol{Y}) &= H(\boldsymbol{X}|\boldsymbol{Y}) + H(W|\boldsymbol{Y}) - H(\boldsymbol{X}W|\boldsymbol{Y}) \\ &= H(\boldsymbol{X}|\boldsymbol{Y}) + H(W|\boldsymbol{Y}) - (H(\boldsymbol{X}|\boldsymbol{Y}) + H(W|\boldsymbol{X}\boldsymbol{Y})) = 0 \ , \end{aligned}$$

which concludes the proof that $I(\boldsymbol{X};\boldsymbol{Y}) = 1$. Furthermore, following a procedure similar to those in the previous section, it can be shown that any ΦID that satisfies the axioms described above (partial ordering, non-negative double-redundancy, and upper-bounded redundancy) correctly assigns 1 bit of information to $I_\partial^{\alpha\to\beta}$, and 0 to all other atoms.

Now that we have built these 16 single-atom systems, let us move to the integration measures of interest. For CD, ψ, and Φ, we will proceed by decomposing them in terms of ΦID atoms and checking whether each atom is positive (+), negative (–), or absent (0) from the decomposition to obtain the results in Table 1 of the article. Let us begin with CD, defined as the sum of transfer entropies from one variable to the other:

$$\begin{aligned} \mathrm{CD} &= \frac{1}{2}\sum_{i=1}^{2} I(X_i;Y_j|X_j) \\ &= \frac{1}{2}\sum_{i=1}^{2}\left(I_\partial^{\{i\}\to\{1\}\{2\}} + I_\partial^{\{i\}\to\{j\}} + I_\partial^{\{12\}\to\{1\}\{2\}} + I_\partial^{\{12\}\to\{j\}}\right) . \end{aligned} \tag{C.9}$$

Similarly, for ψ the atoms can be extracted from the decomposition of $\mathtt{Syn}(X_1,X_2;Y_1Y_2)$ in Eq. (C.4a):

$$\begin{aligned} \psi &= \mathtt{Syn}(X_1,X_2;Y_1Y_2) \\ &= I_\partial^{\{12\}\to\{1\}\{2\}} + I_\partial^{\{12\}\to\{1\}} + I_\partial^{\{12\}\to\{2\}} + I_\partial^{\{12\}\to\{12\}} . \end{aligned} \tag{C.10}$$

For Φ, the atoms can be extracted from the decomposition of Eq. (9) in the main text:

$$\begin{aligned} \Phi &= I(X_1X_2;Y_1Y_2) - I(X_1;Y_1) - I(X_2;Y_2) \\ &= -I_\partial^{\{1\}\{2\}\to\{1\}\{2\}} + I_\partial^{\{1\}\{2\}\to\{12\}} + \psi + \sum_{i=1}^{2}\left(I_\partial^{\{i\}\to\{j\}} + I_\partial^{\{i\}\to\{12\}}\right) . \end{aligned} \tag{C.11}$$

The Φ_G case is slightly more involved, since it is not easily decomposable into a sum of ΦID atoms. According to the definition of Φ_G [11], for a system given by the joint probability distribution $p(\boldsymbol{X},\boldsymbol{Y})$ one has

$$\Phi_G = \min_{q\in\mathcal{M}_G} D_{\mathrm{KL}}(p\|q) ,$$

where $\mathcal{M}_G$ is the manifold of probability distributions that satisfy the constraints

$$q(Y_i|\boldsymbol{X}) = q(Y_i|X_i) . \tag{C.12}$$

Therefore, it suffices to check whether the probability distribution of the system satisfies the constraints in Eq. (C.12) — if it does, then $\Phi_G = 0$, and otherwise $\Phi_G > 0$ —, which can be easily verified for each system separately to obtain the Φ_G column in Table 1, concluding the proof.

C.6 Results of section 'Why whole-minus-sum Φ can be negative'

In this appendix we describe the details of the noisy AND system and how to compute its ΦID to yield the results shown in Figure 4 of the main text.

Given the past state of the system x_1x_2, the next state is given by

$$y_1 = (x_1 \cdot x_2) \oplus n_1$$
$$y_2 = (x_1 \cdot x_2) \oplus n_2 \ ,$$

where n_1, n_2 are two auxiliary noise variables sampled from Bernoulli distributions with parameter $p = 0.2$, and they are sampled independently with probability $1 - c$ and set to be identical to each other with probability c. This results in a system that, for $c = 0$, consists of two separate AND gates with some noise, and for $c = 1$ a system of two perfectly correlated components that at each time step change state with probability 0.2. All information-theoretic functionals are computed with respect to the system's stationary distribution.

To compute the ΦID atoms we follow the procedure described above based on James et al.'s I_{dep} measure. To minimise numerical problems with the maximum-entropy projections involved, instead of computing all relevant quantities separately we compute one single constraint lattice for the whole system $X_1X_2Y_1Y_2$ and read off all relevant quantities:

- 9 values of mutual information, which can be directly read from the corresponding nodes in the lattice;
- 6 values of single-target PID unique information, which can be obtained as the minimum of suitable subsets of the lattice according to Eq. (C.6); and
- One ΦID double-unique information according to Eq. (C.7).

Together, these 16 numbers fully determine all 16 ΦID atoms, and the resulting linear system of equations can be easily solved.

C.7 Formalising causal emergence

We begin by generalising the notion of synergy for the case of n variables as follows:

$$\texttt{Syn}(X_1, ..., X_n; Y) := \sum_{\boldsymbol{\alpha} \in \mathcal{S}} I_\partial^{\boldsymbol{\alpha}} \ , \tag{C.13}$$

where $\mathcal{S} = \{\boldsymbol{\alpha} \in \mathcal{A} : \{j\} \notin \boldsymbol{\alpha}, j = 1, ..., n\}$. In other words, $\mathcal{S}$ is contains all sets of sources $\boldsymbol{\alpha}$, such that in any $\boldsymbol{\alpha}$ there is no singleton source. i.e. $|a| > 1 \ \forall a \in \boldsymbol{\alpha}$. Intuitively, $\mathcal{S}$ is

the collection of PID atoms that correspond to the information about the target that is not contained in any individual agent.[1]

Next, let us define the set $\mathcal{R}(k) = \{\boldsymbol{\alpha} \in \mathcal{A} : \exists j \neq k, \{j\} \in \boldsymbol{\alpha}, \{k\} \in \boldsymbol{\alpha}\}$, that denotes the collection of atoms that correspond to the information that X_k holds redundantly with at least one other agent X_j.[2] With this set, we can define an analogue of the unique information, as

$$\mathtt{Un}(X_n; Y | X_1, ..., X_{n-1}) := I(X_n; Y) - \sum_{\boldsymbol{\alpha} \in \mathcal{R}(n)} I_\partial^{\boldsymbol{\alpha}} \,. \tag{C.14}$$

Let us denote the set of atoms in this quantity as $\mathcal{U}(k) = \{\boldsymbol{\alpha} \in \mathcal{A} : \boldsymbol{\alpha} \preceq \{k\}, \boldsymbol{\alpha} \notin \mathcal{R}(k)\}$. This corresponds to all the atoms where $\{k\}$ is the only singleton – which, importantly, is in general not just $I_\partial^{\{k\}}$. Intuitively, this is the information that X_n has access to, that no other agent has access to *on their own* (although groups of other agents may).[3] In the following, we denote the set complement by a superscript c, for example $\mathcal{S}^c = \{\boldsymbol{\alpha} \in \mathcal{A} : \boldsymbol{\alpha} \notin \mathcal{S}\}$.

In addition, we will need the following lemmas:

Lemma 7. *Data processing inequality for unique information: if $Z - X_1 - X_2, ..., X_t^n, Y$ is a Markov chain, then*

$$\mathtt{Un}(X_1; Y | X_2, ..., X_n) \geq \mathtt{Un}(Z; Y | X_2, ..., X_n) \,. \tag{C.15}$$

Proof. Consider the set of sources $\boldsymbol{\beta}$ to be equal to $\boldsymbol{\alpha}$ but with the index 1 replaced by the index correpsonding to Z. Using this notation, let us assume that the redundancy function satisfies *predictor monotonicity*: if $Z - X_1 - X_2, ..., X_n$, then (by assumption)

$$I_\partial^{\boldsymbol{\alpha}} \geq I_\partial^{\boldsymbol{\beta}} \,. \tag{C.16}$$

Furthermore, let us denote by $\tilde{\mathcal{U}}(1)$ the set of all the $\boldsymbol{\beta}$'s that corresponds to the $\boldsymbol{\alpha}$'s that belong to $\mathcal{U}(1)$. Then, a direct calculation shows that

$$\mathtt{Un}(X_1; Y | X_2, ..., X_n) = \sum_{\boldsymbol{\alpha} \in \mathcal{U}(n)} I_\partial^{\boldsymbol{\alpha}} \tag{C.17}$$

$$\geq \sum_{\boldsymbol{\alpha} \in \tilde{\mathcal{U}}(n)} I_\partial^{\boldsymbol{\beta}} = \mathtt{Un}(Z; Y | X_2, ..., X_n) \,. \tag{C.18}$$

□

[1] For example, for $n = 2$ we obtain the standard synergy $\mathcal{S} = \{\{12\}\}$, and for $n = 3$ we have $\mathcal{S} = \{\{12\}, \{13\}, \{23\}, \{12\}\{13\}, \{12\}\{23\}, \{13\}\{23\}, \{12\}\{13\}\{23\}, \{123\}\}$.

[2] Again, in the $n = 2$ case we recover the standard redundancy $\mathcal{R}(1) = \mathcal{R}(2) = \{\{1\}\{2\}\}$; and as an example for $n = 3$ we have $\mathcal{R}(1) = \{\{1\}\{2\}, \{1\}\{3\}, \{1\}\{2\}\{3\}\}$.

[3] And again, for $n = 2$ we recover $\mathcal{U}(i) = \{\{i\}\}$, but for $n = 3$ we have e.g. $\mathcal{U}(1) = \{\{1\}, \{1\}\{23\}\}$.

Lemma 8. *The following equality holds:*

$$\texttt{Un}(\boldsymbol{X}^n; Y | X_1, ..., X_n) = \texttt{Syn}(X_1, ..., X_n; Y) \ . \tag{C.19}$$

Proof. Begin by considering the PID of n sources $X_1, ..., X_n$ on the lattice $\mathcal{A}^n$, and define its set $\mathcal{S}$ as above. Now we add an additional $n+1^{\text{st}}$ variable that is simply all of them concatenated, $X_{n+1} = \boldsymbol{X}^n$, and build a PID on the lattice $\mathcal{A}^{n+1}$.

Note that the nodes in $\mathcal{A}^{n+1}$ that precede $\{n+1\}$ are those in $\mathcal{A}$, but with the singleton $\{n+1\}$ appended to them; and by the *Deterministic Equality* axiom of PID [209], we have that $\mathcal{U}(n+1) = \mathcal{S}$. Then, it is direct to see that

$$\texttt{Un}(\boldsymbol{X}^n; Y | X_1, ..., X_n) = \sum_{\boldsymbol{\alpha} \in \mathcal{U}(n+1)} I_\partial^{\boldsymbol{\alpha}} = \sum_{\boldsymbol{\alpha} \in \mathcal{S}} I_\partial^{\boldsymbol{\alpha}} = \texttt{Syn}(X_1, ..., X_n; Y) \ . \tag{C.20}$$

□

Lemma 9. *Consider a feature V that exhibits causal emergence over $\boldsymbol{X}$. Then, there exists no deterministic function $g(\cdot)$ such that $V = g(X_k)$ for any $k = 1, ..., n$.*

Proof. Proof by contrapositive. Let us assume that $g(\cdot)$ does exist, and $H(V|X_k) = 0$ for some k. Then, this implies that

$$\begin{aligned} \texttt{Un}(V; \boldsymbol{Y} | X_1, ..., X_n) &\leq \texttt{Un}(V; \boldsymbol{Y} | X_k) && \text{(C.21)} \\ &\leq I(V; \boldsymbol{Y} | X_k) && \text{(C.22)} \\ &= 0, && \text{(C.23)} \end{aligned}$$

which shows that V_t cannot exhibit emergent behaviour. □

Now we begin with the proofs of our statements about causal emergence in Sec. 9.5. With these definitions and lemmas in hand, we can link the original Definition 4 of causal emergence in terms of unique information, with the equivalent criterion in terms of synergy in Proposition 4.

Proof of Proposition 4. If $\texttt{Syn}(X_1, ..., X_n; \boldsymbol{Y}) > 0$, then by Lemma 8 it is clear that the feature $V = \boldsymbol{X}$ exhibits causal emergence.

To prove the converse, note that if $\texttt{Syn}(X_1, ..., X_n; \boldsymbol{Y}) = 0$ then by Lemma 8 we have $\texttt{Un}(\boldsymbol{X}^n; Y | X_1, ..., X_n) = 0$, and by Lemma 7 we have $\texttt{Un}(V; Y | X_1, ..., X_n) \leq \texttt{Un}(\boldsymbol{X}^n; Y | X_1, ..., X_n) = 0$ for any $V = f(\boldsymbol{X})$. □

Since we can decompose PID atoms in terms of ΦID atoms (Appendix C.2), now we can split this aggregate causal emergence into downward and decoupled components:

$$\mathtt{Syn}(X_1,...,X_n;\boldsymbol{Y}) = \sum_{\boldsymbol{\alpha}\in\mathcal{S},\boldsymbol{\beta}\in\mathcal{A}} I_\partial^{\boldsymbol{\alpha}\to\boldsymbol{\beta}} \tag{C.24}$$

$$= \sum_{\boldsymbol{\alpha}\in\mathcal{S},\boldsymbol{\beta}\in\mathcal{S}^c} I_\partial^{\boldsymbol{\alpha}\to\boldsymbol{\beta}} + \sum_{\boldsymbol{\alpha}\in\mathcal{S},\boldsymbol{\beta}\in\mathcal{S}} I_\partial^{\boldsymbol{\alpha}\to\boldsymbol{\beta}} \,. \tag{C.25}$$

The result follows from identifying $\mathcal{D}$ and $\mathcal{G}$ with the two terms in the RHS:

$$\mathcal{D}(\boldsymbol{X}_t) \coloneqq \sum_{\boldsymbol{\alpha}\in\mathcal{S},\boldsymbol{\beta}\in\mathcal{S}^c} I_\partial^{\boldsymbol{\alpha}\to\boldsymbol{\beta}} \tag{C.26}$$

$$\mathcal{G}(\boldsymbol{X}_t) \coloneqq \sum_{\boldsymbol{\alpha}\in\mathcal{S},\boldsymbol{\beta}\in\mathcal{S}} I_\partial^{\boldsymbol{\alpha}\to\boldsymbol{\beta}} \,. \tag{C.27}$$

Where $\mathcal{S}^c$ is the set complement of $\mathcal{S}$, and therefore contains all the sources in which there is at least one singleton.

Using these definitions, we can now complete the remaining proofs regarding necessary and sufficient conditions for downward causation and decoupled causality.

Proof of Proposition 5. Note that, by the properties of ΦID and Lemma 8, the following bounds hold

$$\mathtt{Un}(\boldsymbol{X};Y_k|X_1,...,X_n) = \mathtt{Syn}(\boldsymbol{X};Y_k) \leq \mathcal{D}(\boldsymbol{X}_t) \leq \sum_{k=1}^n \mathtt{Syn}(\boldsymbol{X};Y_k) = \sum_{k=1}^n \mathtt{Un}(\boldsymbol{X};Y_k|X_1,...,X_n)$$

Then, from the left side of this expression and Lemma 7, it is direct to see that just one non-zero $\mathtt{Un}(\boldsymbol{X};Y_k|X_1,...,X_n)$ suffices to make $\mathcal{D}(\boldsymbol{X}_t)$ non-zero, and therefore there exists an emergent feature $V=\boldsymbol{X}$. Similarly, from the right side of this expression and Lemma 7, it is direct to see that if all $\mathtt{Un}(V;Y_k|X_1,...,X_n)$ are zero, then $\mathcal{D}(\boldsymbol{X}_t)=0$. □

Proof of Proposition 6. We prove that if $\mathcal{G}(\boldsymbol{X}_t)>0$, then there exists a V,V' such that $\mathtt{Un}(V;V'|X_1,...,X_n)>0$. If $\mathcal{G}(\boldsymbol{X}_t)>0$, then there must be at least one positive ΦID atom $I_\partial^{a\to b}>0$, for some $a,b\in\mathcal{S}$. Define the set $\boldsymbol{\gamma}$ as the least upper bound of a and b that is a subset of $\{1,...,n\}$, and the feature $V=f(\boldsymbol{X})=\boldsymbol{X}^{\boldsymbol{\gamma}}$, with the corresponding $V'=f(\boldsymbol{Y})=\boldsymbol{Y}^{\boldsymbol{\gamma}}$. Then, $\mathtt{Un}(V;V'|X_1,...,X_n)\geq I_\partial^{a\to b}>0$. □

Proof of Proposition 7. Consider the quantity $\Xi = I(\boldsymbol{X};V') - \sum_k I(X_k;V')$, which is Ψ but using $\boldsymbol{X}$ instead of V. By the data processing inequality, and since $V-\boldsymbol{X}-\boldsymbol{Y}-V'$ is a Markov chain, we have $\Psi\leq\Xi$. At the same time, by performing a normal PID decomposition on Ξ

with target V' we have

$$\Xi = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} I_\partial^{\boldsymbol{\alpha}} - \sum_{k=1}^{n} \sum_{\boldsymbol{\alpha} \preceq \{k\}} I_\partial^{\boldsymbol{\alpha}} \leq \sum_{\boldsymbol{\alpha} \in \mathcal{S}} I_\partial^{\boldsymbol{\alpha}} \,, \tag{C.28}$$

where the inequality comes from the fact that some terms (in fact, all except the $\{k\}$) are double-counted in the negative component of Ξ, i.e. in $\sum_k I(X_k; V')$. Then, by the PID data processing inequality, each one of the atoms of the decomposition of with target V' is less than or equal to the same atom of the decomposition with target $\boldsymbol{Y}$, and therefore by the definition of $\mathcal{S}$ we have $\Xi \leq \mathtt{Syn}(X_1, ..., X_n; \boldsymbol{Y})$. Hence, if $\Psi > 0$ then $\mathtt{Syn}(X_1, ..., X_n; \boldsymbol{Y}) > 0$ and by Proposition 4 the system is causally emergent. □

Bibliography

[1] M. M. Waldrop, *Complexity: The Emerging Science at the Edge of Order and Chaos*. Simon and Schuster, 1993.

[2] G. Tononi, O. Sporns, and G. M. Edelman, "A measure for brain complexity: Relating functional segregation and integration in the nervous system," *Proceedings of the National Academy of Sciences* **91** no. 11, (May, 1994) 5033–7.

[3] G. Tononi, G. M. Edelman, and O. Sporns, "Complexity and coherency: Integrating information in the brain," *Trends in Cognitive Sciences* **2** no. 12, (Dec, 1998) 474–484.

[4] G. Tononi and O. Sporns, "Measuring information integration," *BMC Neuroscience* **4** no. 3, (2003) .

[5] D. Balduzzi and G. Tononi, "Integrated information in discrete dynamical dystems: Motivation and theoretical framework," *PLoS Computational Biology* **4** no. 6, (Jun, 2008) e1000091.

[6] M. Oizumi, L. Albantakis, and G. Tononi, "From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0," *PLoS Computational Biology* **10** no. 5, (May, 2014) e1003588.

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley, Hoboken, 2006.

[8] A. B. Barrett and A. K. Seth, "Practical measures of integrated information for time-series data," *PLoS Computational Biology* **7** no. 1, (Jan, 2011) e1001052.

[9] V. Griffith, "A principled infotheoretic ϕ-like measure," `arXiv:1401.0978`.

[10] M. Oizumi, S.-i. Amari, T. Yanagawa, N. Fujii, and N. Tsuchiya, "Measuring integrated information from the decoding perspective," *PLoS Computational Biology* **12** no. 1, (2016) e1004654.

[11] M. Oizumi, N. Tsuchiya, and S.-i. Amari, "Unified framework for information integration based on information geometry," *Proceedings of the National Academy of Sciences* **113** no. 51, (2016) 14817–14822.

[12] A. K. Seth, A. B. Barrett, and L. Barnett, "Causal density and integrated information as measures of conscious level," *Philosophical Transactions A* **369** no. 1952, (Jan, 2011) 3748–67.

[13] S. Krohn and D. Ostwald, "Computing integrated information," arXiv:1610.03627.

[14] P. Mediano, F. Rosas, R. L. Carhart-Harris, A. K. Seth, and A. B. Barrett, "Beyond integrated information: A taxonomy of information dynamics phenomena," arXiv:1909.02297.

[15] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Physical Review E* **69** no. 6, (Jun, 2004) 066138.

[16] N. Ay, "Information geometry on complexity and stochastic interaction," *Entropy* **17** no. 4, (Apr, 2015) 2432–2458.

[17] S. Watanabe, "Information theoretical analysis of multivariate correlation," *IBM Journal of Research and Development* **4** no. 1, (1960) 66–82.

[18] K. Wiesner, M. Gu, E. Rieper, and V. Vedral, "Information-theoretic bound on the energy cost of stochastic simulation," arXiv:1110.4217.

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

[20] A. B. Barrett, "Exploration of synergistic and redundant information sharing in static and dynamical Gaussian systems," *Physical Review E* **91** no. 5, (2015) 052802.

[21] N. Bertschinger, J. Rauh, E. Olbrich, and J. Jost, "Shared Information – New Insights and Problems in Decomposing Information in Complex Systems," in *Proceedings of the European Conference on Complex Systems 2012*, T. Gilbert, M. Kirkilionis, and G. Nicolis, eds., Springer Proceedings in Complexity. Springer International Publishing, 2012. arXiv:1210.5902.

[22] V. Griffith and C. Koch, "Quantifying synergistic mutual information," arXiv:1205.4265.

[23] F. Rosas, V. Ntranos, C. Ellison, S. Pollin, and M. Verhelst, "Understanding interdependency through complex information sharing," *Entropy* **18** no. 2, (Jan, 2016) 38.

[24] R. A. A. Ince, "Measuring multivariate redundant information with pointwise common change in surprisal," *Entropy* **19** no. 7, (Feb, 2017) 318, arXiv:1602.05063.

[25] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay, "Quantifying unique information," *Entropy* **16** no. 4, (Apr, 2014) 2161–2183, arXiv:1311.2852.

[26] M. Harder, C. Salge, and D. Polani, "Bivariate measure of redundant information," *Physical Review E* **87** no. 1, (2013) 012130.

[27] P. E. Latham and S. Nirenberg, "Synergy, redundancy, and independence in population codes, revisited," *The Journal of Neuroscience* **25** no. 21, (May, 2005) 5195–206.

[28] N. Merhav, G. Kaplan, A. Lapidoth, and S. Shamai Shitz, "On information rates for mismatched decoders," *IEEE Transactions on Information Theory* **40** no. 6, (1994) 1953–1967.

[29] M. Oizumi, T. Ishii, K. Ishibashi, T. Hosoya, and M. Okada, "Mismatched decoding in the brain," *The Journal of Neuroscience* **30** no. 13, (Mar, 2010) 4815–26.

[30] S.-i. Amari and H. Nagaoka, *Methods of Information Geometry*. American Mathematical Society, Jan, 2000.

[31] S.-i. Amari, "Information geometry in optimization, machine learning and statistical inference," *Frontiers of Electrical and Electronic Engineering in China* **5** no. 3, (Jul, 2010) 241–260.

[32] S. S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004. arXiv:1111.6189v1.

[33] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica* **37** no. 3, (Aug, 1969) 424.

[34] A. K. Seth, "Causal connectivity of evolved neural networks during behavior," *Network: Computation in Neural Systems* **16** no. 1, (2005) 35–54. PMID: 16350433.

[35] L. Barnett, A. B. Barrett, and A. K. Seth, "Granger causality and transfer entropy are equivalent for Gaussian variables," *Physical Review Letters* **103** no. 23, (Dec, 2009) 238701.

[36] L. Barnett and A. K. Seth, "Behaviour of Granger causality under filtering: Theoretical invariance and practical application," *Journal of Neuroscience Methods* **201** no. 2, (2011) 404–419.

[37] M. Lindner, R. Vicente, V. Priesemann, and M. Wibral, "TRENTOOL: A Matlab open source toolbox to analyse information flow in time series data with transfer entropy," *BMC Neuroscience* **12** no. 1, (Jan, 2011) 119.

[38] J. T. Lizier, J. Heinzle, A. Horstmann, J.-D. Haynes, and M. Prokopenko, "Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity," *Journal of Computational Neuroscience* **30** no. 1, (Aug, 2010) 85–107.

[39] P. A. M. Mediano and M. P. Shanahan, "Balanced information storage and transfer in modular spiking neural networks," arXiv:1708.04392.

[40] L. Barnett and A. K. Seth, "The MVGC Multivariate Granger Causality toolbox: A new approach to Granger-causal inference," *Journal of Neuroscience Methods* **223** (2014) 50–68.

[41] M. Tegmark, "Improved measures of integrated information," *PLoS Computational Biology* **12** no. 11, (2016) e1005123.

[42] A. Pikovsky, M. Rosenblum, and J. Kurths, *Synchronization: A Universal Concept in Nonlinear Sciences*. Cambridge University Press, Cambridge, 2001.

[43] Y. Kuramoto, *Chemical Oscillations, Waves and Turbulence*. Dover Publications, 1984.

[44] M. J. Panaggio and D. M. Abrams, "Chimera states: Coexistence of coherence and incoherence in networks of coupled oscillators," *Nonlinearity* **28** no. 3, (Mar, 2015) R67–R87, arXiv:1403.6204.

[45] J. Cabral, E. Hugues, O. Sporns, and G. Deco, "Role of local network oscillations in resting-state functional connectivity," *NeuroImage* **57** no. 1, (Jul, 2011) 130–9.

[46] P. J. Hellyer, G. Scott, M. Shanahan, D. J. Sharp, and R. Leech, "Cognitive flexibility through metastable neural dynamics is disrupted by damage to the structural connectome," *The Journal of Neuroscience* **35** no. 24, (Jun, 2015) 9050–63.

[47] M. Schartner, A. K. Seth, Q. Noirhomme, M. Boly, M.-A. Bruno, S. Laureys, and A. B. Barrett, "Complexity of multi-dimensional spontaneous EEG decreases during propofol induced general anaesthesia," *PloS ONE* **10** no. 8, (Jan, 2015) e0133532.

[48] N. Lazarides, G. Neofotistos, and G. P. Tsironis, "Chimeras in SQUID metamaterials," *Physical Review B* **91** no. 5, (Feb, 2015) 054303.

[49] K. Vasudevan, M. Cavers, and A. Ware, "Earthquake sequencing: Chimera states with Kuramoto model dynamics on directed graphs," *Nonlinear Processes in Geophysics* **22** no. 5, (2015) 499–512.

[50] N. Kruk, Y. Maistrenko, and H. Koeppl, "Self-propelled chimeras," *Physical Review E* **98** no. 3, (2018) 032219.

[51] E. A. Martens, S. Thutupalli, A. Fourrière, and O. Hallatschek, "Chimera states in mechanical oscillator networks," *Proceedings of the National Academy of Sciences* **110** no. 26, (2013) 10563–10567.

[52] S. Petkoski, A. Spiegler, T. Proix, P. Aram, J.-J. Temprado, and V. K. Jirsa, "Heterogeneity of time delays determines synchronization of coupled oscillators," *Physical Review E* **94** no. 1, (Jul, 2016) 012209.

[53] M. Shanahan, "Metastable chimera states in community-structured oscillator networks," *Chaos* **20** no. 1, (Mar, 2010) 013108, arXiv:0908.3881.

[54] M. Shanahan, "Dynamical complexity in small-world networks of spiking neurons," *Physical Review E* **78** no. 4, (2008) 041924.

[55] D. R. Chialvo, "Emergent complex neural dynamics," *Nature Physics* **6** no. 10, (Oct, 2010) 744–750, arXiv:1010.2530.

[56] E. Tognoli and J. A. S. Kelso, "The metastable brain," *Neuron* **81** no. 1, (Jan, 2014) 35–48.

[57] D. M. Abrams, R. Mirollo, S. H. Strogatz, and D. A. Wiley, "Solvable model for chimera states of coupled oscillators," *Physical Review Letters* **101** no. 8, (Aug, 2008) 084103.

[58] R. Sharma, M. Gupta, and G. Kapoor, "Some better bounds on the variance with applications," *Journal of Mathematical Inequalities* **4** no. 3, (2010) 355–363.

[59] P. Dayan and L. F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press, Cambridge, MA, 2001.

[60] A. Borst and F. E. Theunissen, "Information theory and neural coding," *Nature Neuroscience* **2** no. 11, (Nov, 1999) 947–957.

[61] D. H. Johnson, C. M. Gruner, K. Baggerly, and C. Seshagiri, "Information-theoretic analysis of neural coding," *Journal of Computational Neuroscience* **10** no. 1, (2001) 47–69.

[62] R. D. Beer and P. L. Williams, "Information processing and dynamics in minimally cognitive agents," *Cognitive Science* **39** no. 1, (Jul, 2014) 1–38.

[63] T. Lochmann and S. Denève, "Information transmission with spiking Bayesian neurons," *New Journal of Physics* **10** no. 5, (May, 2008) 055019.

[64] G. Hennequin, W. Gerstner, and J.-P. Pfister, "STDP in adaptive neurons gives close-to-optimal information transmission," *Frontiers in Computational Neuroscience* **4** (Jan, 2010) 143.

[65] L. de Arcangelis and H. J. Herrmann, "Learning as a phenomenon occurring in a critical state," *Proceedings of the National Academy of Sciences* **107** no. 9, (Mar, 2010) 3977–81.

[66] J. M. Beggs and D. Plenz, "Neuronal avalanches in neocortical circuits," *The Journal of Neuroscience* **23** no. 35, (2003) 11167–11177.

[67] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature* **393** no. 6684, (Jun, 1998) 440–442.

[68] E. M. Izhikevich, "Simple model of spiking neurons," *IEEE Transactions on Neural Networksetworks* **14** no. 6, (Jan, 2003) 1569–72.

[69] J. T. Lizier, *The Local Information Dynamics of Distributed Computation in Complex Systems*. Springer Theses. Springer, Berlin, Heidelberg, 2010.

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

[71] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence*, D. Rand and L.-S. Young, eds., pp. 366–381. Springer, Berlin, Heidelberg, 1981.

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

[73] A. K. Fidjeland, E. B. Roesch, M. P. Shanahan, and W. Luk, "NeMo: A platform for neural modelling of spiking neurons using GPUs," in *2009 20th IEEE International Conference on Application-specific Systems, Architectures and Processors*, pp. 137–144. IEEE, Jul, 2009.

[74] J. T. Lizier, "JIDT: An information-theoretic toolkit for studying the dynamics of complex systems," *Frontiers in Robotics and AI* **1** (Dec, 2014) 37, arXiv:1408.3270.

[75] G. Buzsáki and X.-J. Wang, "Mechanisms of gamma oscillations," *Annual Review of Neuroscience* **35** no. 1, (2012) 203–225.

[76] G. Pruessner, *Self-Organised Criticality: Theory, Models and Characterisation*. Cambridge University Press, Cambridge, UK, 2012.

[77] V. Priesemann, M. Wibral, M. Valderrama, R. Pröpper, M. Le Van Quyen, T. Geisel, J. Triesch, D. Nikolić, and M. H. J. Munk, "Spike avalanches in vivo suggest a driven, slightly subcritical brain state," *Frontiers in Systems Neuroscience* **8** no. 108, (Jan, 2014) 108.

[78] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM Review* **51** no. 4, (2009) 661–703, arXiv:0706.1062.

[79] J. Alstott, E. T. Bullmore, and D. Plenz, "Powerlaw: A Python package for analysis of heavy-tailed distributions," *PloS ONE* **9** no. 1, (May, 2014) e85777.

[80] A. Levina and V. Priesemann, "Subsampling scaling," *Nature Communications* **8** (2017) 15140.

[81] J. H. Lucio, R. Valdés, and L. R. Rodríguez, "Improvements to surrogate data methods for nonstationary time series," *Physical Review E* **85** no. 5, (May, 2012) 056202.

[82] D. V. Foster and P. Grassberger, "Lower bounds on mutual information," *Physical Review E* **83** no. 1, (2011) 010101.

[83] A. Haimovici, E. Tagliazucchi, P. Balenzuela, and D. R. Chialvo, "Brain organization into resting state networks emerges at criticality on a model of the human connectome," *Physical Review Letters* **110** no. 17, (2013) 178101.

[84] B. Drossel and F. Schwabl, "Self-organized critical forest-fire model," *Physical Review Letters* **69** no. 11, (1992) 1629.

[85] J. Werfel, D. E. Ingber, and Y. Bar-Yam, "Programed death is favored by natural selection in spatial systems," *Physical Review Letters* **114** no. 23, (2015) 238103.

[86] S. Wolfram, *A New Kind of Science*. Wolfram Media, 2002.

[87] M. Cook, "Universality in elementary cellular automata," *Complex Systems* **15** no. 1, (2004) 1–40.

[88] P. Grassberger, "Toward a quantitative theory of self-generated complexity," *International Journal of Theoretical Physics* **25** no. 9, (1986) 907–938.

[89] M. Prokopenko, F. Boschetti, and A. J. Ryan, "An information-theoretic primer on complexity, self-organization, and emergence," *Complexity* **15** no. 1, (2009) 11–28.

[90] F. Rosas, P. A. Mediano, M. Ugarte, and H. J. Jensen, "An information-theoretic approach to self-organisation: Emergence of complex interdependencies in coupled dynamical systems," *Entropy* **20** no. 10, (2018) .

[91] J. L. McClelland and D. E. Rumelhart, *Parallel Distributed Processing*, vol. 2. MIT Press, 1987.

[92] B. J. Baars, "In the theatre of consciousness: Global workspace theory, a rigorous scientific theory of consciousness," *Journal of Consciousness Studies* **4** no. 4, (1997) 292–309.

[93] S. Dehaene, L. Charles, J.-R. King, and S. Marti, "Toward a computational theory of conscious processing," *Current Opinion in Neurobiology* **25** (2014) 76–84.

[94] A. Adamatzky and J. Durand-Lose, *Collision-based Computing*. Springer, 2012.

[95] S. Wolfram, "Universality and complexity in cellular automata," *Physica D* **10** (1984) 1–35.

[96] C. G. Langton, "Computation at the edge of chaos: Phase transitions and emergent computation," *Physica D: Nonlinear Phenomena* **42** no. 1-3, (1990) 12–37.

[97] A. R. Smith III, "Simple computation-universal cellular spaces," *Journal of the ACM* **18** no. 3, (1971) 339–353.

[98] K. Lindgren and M. G. Nordahl, "Universal computation in simple one-dimensional cellular automata," *Complex Systems* **4** no. 3, (1990) 299–318.

[99] K. Culik II and S. Yu, "Undecidability of CA classification schemes," *Complex Systems* **2** no. 2, (1988) 177–190.

[100] J. Kari, "Decidability and undecidability in cellular automata," *International Journal of General Systems* **41** no. 6, (2012) 539–554.

[101] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, "Information transfer by particles in cellular automata," in *Australian Conference on Artificial Life*, pp. 49–60, Springer. 2007.

[102] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya, "Information modification and particle collisions in distributed computation," *Chaos: An Interdisciplinary Journal of Nonlinear Science* **20** no. 3, (2010) 037109.

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

[104] G. J. Martinez, J. C. Seck-Tuoh-Mora, and H. Zenil, "Computation and universality: Class IV versus class III cellular automata," *Journal of Cellular Automata* **7** (2013) 393–430.

[105] M. Scheffer, J. Bascompte, W. A. Brock, V. Brovkin, S. R. Carpenter, V. Dakos, H. Held, E. H. Van Nes, M. Rietkerk, and G. Sugihara, "Early-warning signals for critical transitions," *Nature* **461** no. 7260, (2009) 53.

[106] D. Toker and F. T. Sommer, "Information integration in large brain networks," *PLoS Computational Biology* **15** no. 2, (2019) e1006807.

[107] P. A. M. Mediano, J. C. Farah, and M. P. Shanahan, "Integrated information and metastability in systems of coupled oscillators," arXiv:1606.08313.

[108] E. Tagliazucchi, "The signatures of conscious access and its phenomenology are consistent with large-scale brain communication at criticality," *Consciousness and Cognition* **55** (2017) 136–147.

[109] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. New York, 2005.

[110] M. D. Humphries and K. Gurney, "Network 'small-world-ness:' A quantitative method for determining canonical network equivalence," *PLoS ONE* **3** no. 4, (Jan, 2008) e0002051.

[111] H. Yin, A. R. Benson, and J. Leskovec, "Higher-order clustering in networks," arXiv:1704.03913.

[112] D. Toker and F. Sommer, "Moving past the minimum information partition: How to quickly and accurately calculate integrated information," arXiv:1605.01096.

[113] S. Hidaka and M. Oizumi, "Fast and exact search for the partition with minimal information loss," arXiv:1708.01444.

[114] X. D. Arsiwalla and P. F. M. J. Verschure, "Integrated information for large complex networks," in *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, Aug, 2013.

[115] Q. Wang, S. R. Kulkarni, and S. Verdu, "Divergence estimation for multidimensional densities via k-nearest-neighbor distances," *IEEE Transactions on Information Theory* **55** no. 5, (May, 2009) 2392–2405.

[116] A. B. Barrett and L. Barnett, "Granger causality is designed to measure effect, not mechanism," *Frontiers in Neuroinformatics* **7** (2013) 6.

[117] A.-M. v. C. van Walsum, Y. Pijnenburg, H. Berendse, B. van Dijk, D. Knol, P. Scheltens, and C. Stam, "A neural complexity measure applied to MEG data in Alzheimer's disease," *Clinical Neurophysiology* **114** no. 6, (2003) 1034–1040.

[118] L. Albantakis and G. Tononi, "The intrinsic cause-effect power of discrete dynamical systems – from elementary cellular automata to adapting animats," *Entropy* **17** no. 8, (2015) 5472–5502.

[119] H. Lau and D. Rosenthal, "Empirical support for higher-order theories of conscious awareness," *Trends in Cognitive Sciences* **15** no. 8, (2011) 365–373.

[120] U. Lee, G. A. Mashour, S. Kim, G.-J. Noh, and B.-M. Choi, "Propofol induction reduces the capacity for neural information integration: Implications for the mechanism of consciousness and general anesthesia," *Consciousness and Cognition* **18** no. 1, (2009) 56–64.

[121] J.-Y. Chang, A. Pigorini, M. Massimini, G. Tononi, L. Nobili, and B. D. Van Veen, "Multivariate autoregressive models with exogenous inputs for intracerebral responses to direct electrical stimulation of the human brain," *Frontiers in Human Neuroscience* **6** (2012) 317.

[122] M. Boly, S. Sasai, O. Gosseries, M. Oizumi, A. Casali, M. Massimini, and G. Tononi, "Stimulus set meaningfulness and neurophysiological differentiation: a functional magnetic resonance imaging study," *PloS ONE* **10** no. 5, (2015) e0125337.

[123] A. M. Haun, M. Oizumi, C. K. Kovach, H. Kawasaki, H. Oya, M. A. Howard, R. Adolphs, and N. Tsuchiya, "Conscious perception as integrated information patterns in human electrocorticography," *eNeuro* **4** no. 5, (2017) .

[124] H. Kim, A. G. Hudetz, J. Lee, G. A. Mashour, U. Lee, M. S. Avidan, T. Bel-Bahar, S. Blain-Moraes, G. Golmirzaie, E. Janke, *et al.*, "Estimating the integrated information measure Phi from high-density electroencephalography during states of consciousness in humans," *Frontiers in Human Neuroscience* **12** (2018) 42.

[125] A. G. Casali, O. Gosseries, M. Rosanova, M. Boly, S. Sarasso, K. R. Casali, S. Casarotto, M.-A. Bruno, S. Laureys, G. Tononi, and M. Massimini, "A theoretically based index of consciousness independent of sensory processing and behavior," *Science Translational Medicine* **5** no. 198, (Aug, 2013) 198ra105.

[126] G. Tononi, M. Boly, M. Massimini, and C. Koch, "Integrated information theory: From consciousness to its physical substrate," *Nature Reviews Neuroscience* **17** no. 7, (2016) 450.

[127] G. Tononi and C. Koch, "Consciousness: Here, there and everywhere?," *Philosophical Transactions of the Royal Society B: Biological Sciences* **370** no. 1668, (2015) 20140167.

[128] M. M. Schartner, *On the relation between complex brain activity and consciousness*. PhD thesis, University of Sussex, 2016.

[129] M. M. Schartner, R. L. Carhart-Harris, A. B. Barrett, A. K. Seth, and S. D. Muthukumaraswamy, "Increased spontaneous MEG signal diversity for psychoactive doses of ketamine, LSD and psilocybin," *Scientific Reports* **7** (Apr, 2017) 46421.

[130] G. Deco, E. Tagliazucchi, H. Laufs, A. Sanjuán, and M. L. Kringelbach, "Novel intrinsic ignition method measuring local-global integration characterizes wakefulness and deep sleep," *eNeuro* **4** no. 5, (2017) .

[131] E. Tagliazucchi and H. Laufs, "Decoding wakefulness levels from typical fMRI resting-state data reveals reliable drifts between wakefulness and sleep," *Neuron* **82** no. 3, (2014) 695–708.

[132] T. Yanagawa, Z. C. Chao, N. Hasegawa, and N. Fujii, "Large-scale information flow in conscious and unconscious states: An ECoG study in monkeys," *PloS ONE* **8** no. 11, (2013) e80845.

[133] M. Murphy, M.-A. Bruno, B. A. Riedner, P. Boveroux, Q. Noirhomme, E. C. Landsness, J.-F. Brichant, C. Phillips, M. Massimini, S. Laureys, *et al.*, "Propofol anesthesia and sleep: A high-density EEG study," *Sleep* **34** no. 3, (2011) 283–291.

[134] S. D. Muthukumaraswamy, R. L. Carhart-Harris, R. J. Moran, M. J. Brookes, T. M. Williams, D. Errtizoe, B. Sessa, A. Papadopoulos, M. Bolstridge, K. D. Singh, *et al.*, "Broadband cortical desynchronization underlies the human psychedelic state," *Journal of Neuroscience* **33** no. 38, (2013) 15171–15183.

[135] S. D. Muthukumaraswamy, A. D. Shaw, L. E. Jackson, J. Hall, R. Moran, and N. Saxena, "Evidence that subanesthetic doses of ketamine cause sustained disruptions of NMDA and AMPA-mediated frontoparietal connectivity in humans," *Journal of Neuroscience* **35** no. 33, (2015) 11694–11706.

[136] R. L. Carhart-Harris, S. Muthukumaraswamy, L. Roseman, M. Kaelen, W. Droog, K. Murphy, E. Tagliazucchi, E. E. Schenberg, T. Nest, C. Orban, *et al.*, "Neural correlates of the LSD experience revealed by multimodal neuroimaging," *Proceedings of the National Academy of Sciences* **113** no. 17, (2016) 4853–4858.

[137] M. M. Schartner, A. Pigorini, S. A. Gibbs, G. Arnulfo, S. Sarasso, L. Barnett, L. Nobili, M. Massimini, A. K. Seth, and A. B. Barrett, "Global and local complexity of intracranial EEG decreases during NREM sleep," *Neuroscience of Consciousness* **2017** no. 1, (01, 2017) . niw022.

[138] T. Bayne and O. Carter, "Dimensions of consciousness and the psychedelic state," *Neuroscience of Consciousness* **2018** no. 1, (2018) niy008.

[139] C. Timmermann, L. Roseman, M. Schartner, R. Milliere, L. Williams, D. Erritzoe, S. Muthukumaraswamy, M. Ashton, A. Bendrioua, O. Kaur, *et al.*, "Neural correlates of the DMT experience as assessed via multivariate EEG," *bioRxiv* (2019) 706283.

[140] E. P. Hoel, L. Albantakis, and G. Tononi, "Quantifying causal emergence shows that macro can beat micro," *Proceedings of the National Academy of Sciences* **110** no. 49, (2013) 19790–19795.

[141] T. L. Ribeiro, S. Ribeiro, H. Belchior, F. Caixeta, and M. Copelli, "Undersampled critical branching processes on small-world and random networks fail to reproduce the statistics of spike avalanches," *PloS ONE* **9** no. 4, (2014) e94992.

[142] C. Koutlis, V. K. Kimiskidis, and D. Kugiumtzis, "Identification of hidden sources by estimating instantaneous causality in high-dimensional biomedical time series," *International Journal of Neural Systems* **29** no. 4, (2019) 1850051–1850051.

[143] M. Steriade, D. A. McCormick, and T. J. Sejnowski, "Thalamocortical oscillations in the sleeping and aroused brain," *Science* **262** no. 5134, (1993) 679–685.

[144] M. Bazhenov, I. Timofeev, M. Steriade, and T. J. Sejnowski, "Model of thalamocortical slow-wave sleep oscillations and transitions to activated states," *Journal of Neuroscience* **22** no. 19, (2002) 8691–8704.

[145] S. Tajima, T. Yanagawa, N. Fujii, and T. Toyoizumi, "Untangling brain-wide dynamics in consciousness by cross-embedding," *PLoS Computational Biology* **11** no. 11, (2015) e1004537.

[146] L. Barnett and A. K. Seth, "Granger causality for state-space models," *Physical Review E* **91** no. 4, (2015) 040101.

[147] O. M. Cliff, M. Prokopenko, and R. Fitch, "An information criterion for inferring coupling of distributed dynamical systems," *Frontiers in Robotics and AI* **3** (2016) 71.

[148] I. Gat and N. Tishby, "Synergy and redundancy among brain cells of behaving monkeys," in *Advances in Neural Information Processing Systems*, pp. 111–117. 1999.

[149] A. El Gamal and Y.-H. Kim, *Network Information Theory*. Cambridge University Press, 2011.

[150] T. S. Han, "Linear dependence structure of the entropy space," *Information and Control* **29** no. 4, (1975) 337–368.

[151] R. G. James, C. J. Ellison, and J. P. Crutchfield, "Anatomy of a bit: Information in a time series observation," *Chaos: An Interdisciplinary Journal of Nonlinear Science* **21** no. 3, (2011) 037109.

[152] W. J. McGill, "Multivariate information transmission," *Psychometrika* **19** no. 2, (1954) 97–116.

[153] H. K. Ting, "On the amount of information," *Theory of Probability and its Applications* (1962) 439–447.

[154] R. W. Yeung, "A new outlook on Shannon's information measures," *IEEE Transactions on Information Theory* **37** no. 3, (1991) 466–474.

[155] E. T. Jaynes, *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[156] L. Brillouin, "The negentropy principle of information," *Journal of Applied Physics* **24** no. 9, (1953) 1152–1163.

[157] M. Studený and J. Vejnarová, "The multiinformation function as a tool for measuring stochastic dependence," in *Learning in Graphical Models*, M. Jordan, ed., pp. 261–297. MIT Press, Cambride, US, Feb, 1999.

[158] S. A. Abdallah and M. D. Plumbley, "A measure of statistical complexity based on predictive information with application to finite spin systems," *Physics Letters A* **376** no. 4, (2012) 275–281.

[159] S. Verdú and T. Weissman, "Erasure entropy," *IEEE International Symposium on Information Theory* (2006) 98–102.

[160] E. Olbrich, N. Bertschinger, N. Ay, and J. Jost, "How should complexity scale with system size?" *The European Physical Journal B* **63** no. 3, (2008) 407–415.

[161] A. J. Bell, "The co-information lattice," in *Proceedings of the Fifth International Workshop on Independent Component Analysis and Blind Signal Separation: ICA*, vol. 2003. 2003.

[162] Y. Bar-Yam, "Multiscale complexity/entropy," *Advances in Complex Systems* **7** no. 01, (2004) 47–63.

[163] R. P. Stanley, *Enumerative Combinatorics*, vol. Vol. 1 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press;, 2012.

[164] L. M. Bettencourt, V. Gintautas, and M. I. Ham, "Identification of functional information subgraphs in complex networks," *Physical Review Letters* **100** no. 23, (2008) 238701.

[165] L. Comtet, *Advanced Combinatorics: The Art of Finite and Infinite Expansions*. Springer Science & Business Media, 2012.

[166] E. Schneidman, S. Still, M. J. Berry, and W. Bialek, "Network information and connected correlations," *Physical Review Letters* **91** no. 23, (2003) 238701.

[167] S.-I. Amari, "Information geometry on hierarchy of probability distributions," *IEEE Transactions on Information Theory* **47** no. 5, (2001) 1701–1711.

[168] E. Olbrich, N. Bertschinger, and J. Rauh, "Information decomposition and synergy," *Entropy* **17** no. 5, (2015) 3501–3517.

[169] J. P. Crutchfield and D. P. Feldman, "Regularities unseen, randomness observed: Levels of entropy convergence," *Chaos: An Interdisciplinary Journal of Nonlinear Science* **13** no. 1, (2003) 25–54.

[170] J. P. Crutchfield, C. J. Ellison, and J. R. Mahoney, "Time's barbed arrow: Irreversibility, crypticity, and stored information," *Physical Review Letters* **103** no. 9, (2009) 094101.

[171] R. G. James, N. Barnett, and J. P. Crutchfield, "Information flows? A critique of transfer entropies," *Physical Review Letters* **116** no. 23, (2016) 238701.

[172] P. Mediano, A. Seth, and A. Barrett, "Measuring integrated information: Comparison of candidate measures in theory and simulation," *Entropy* **21** no. 1, (2019) 17.

[173] R. James, J. Emenheiser, and J. Crutchfield, "Unique information via dependency constraints," *Journal of Physics A: Mathematical and Theoretical* (2018) .

[174] P. L. Williams and R. D. Beer, "Generalized measures of information transfer," `arXiv:1102.1507`.

[175] M. A. Bedau, "Weak emergence," *Noûs* **31** (1997) 375–399.

[176] F. E. Turkheimer, P. Hellyer, A. A. Kehagia, P. Expert, L.-D. Lord, J. Vohryzek, J. D. F. Dafflon, M. Brammer, and R. Leech, "Conflicting emergences. Weak vs. strong emergence for the modelling of brain function," *Neuroscience & Biobehavioral Reviews* **99** (2019) 3–10.

[177] A. K. Seth, "Measuring autonomy and emergence via Granger causality," *Artificial Life* **16** no. 2, (2010) 179–196.

[178] D. P. Feldman and J. P. Crutchfield, "Measures of statistical complexity: Why?" *Physics Letters A* **238** no. 4-5, (1998) 244–252.

[179] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

[180] A. Barrett and P. Mediano, "The Phi measure of integrated information is not well-defined for general physical systems," *Journal of Consciousness Studies* **26** no. 1-2, (2019) 11–20.

[181] R. L. Carhart-Harris, R. Leech, P. J. Hellyer, M. Shanahan, A. Feilding, E. Tagliazucchi, D. R. Chialvo, and D. Nutt, "The entropic brain: A theory of conscious states informed by neuroimaging research with psychedelic drugs," *Frontiers in Human Neuroscience* **8** (2014) 20.

[182] N. Radhakrishnan and B. Gangadhar, "Estimating regularity in epileptic seizure time-series data," *IEEE Engineering in Medicine and Biology Magazine* **17** no. 3, (1998) 89–94.

[183] X.-S. Zhang, R. J. Roy, and E. W. Jensen, "EEG complexity as a measure of depth of anesthesia for patients," *IEEE transactions on Biomedical Engineering* **48** no. 12, (2001) 1424–1433.

[184] D. Dolan, H. J. Jensen, P. Mediano, M. Molina-Solana, H. Rajpal, F. Rosas, and J. A. Sloboda, "The improvisational state of mind: A multidisciplinary study of an improvisatory approach to classical music repertoire performance," *Frontiers in Psychology* **9** (2018) 1341.

[185] A. Lempel and J. Ziv, "On the complexity of finite sequences," *IEEE Transactions on Information Theory* **22** no. 1, (1976) 75–81.

[186] M. Mitchell, *Complexity: A Guided Tour*. Oxford University Press, 2009.

[187] P. M. Vitanyi and M. Li, *An Introduction to Kolmogorov Complexity and its Applications*, vol. 34. Springer Heidelberg, 1997.

[188] J. Ziv, "Coding theorems for individual sequences," *IEEE Transactions on Information Theory* (1978) .

[189] M. Feder and N. Merhav, "Relations between entropy and error probability," *IEEE Transactions on Information Theory* **40** no. 1, (1994) 259–266.

[190] J. Ziv and A. Lempel, "A universal algorithm for sequential data compression," *IEEE Transactions on Information Theory* **23** no. 3, (1977) 337–343.

[191] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE transactions on Information Theory* **24** no. 5, (1978) 530–536.

[192] F. Kaspar and H. Schuster, "Easily calculable measure for the complexity of spatiotemporal patterns," *Physical Review A* **36** no. 2, (1987) 842.

[193] G. Scott and R. L. Carhart-Harris, "Psychedelics as a treatment for disorders of consciousness," *Neuroscience of Consciousness* **2019** no. 1, (04, 2019) . niz003.

[194] R. L. Carhart-Harris, "How do psychedelics work?" *Current Opinion in Psychiatry* **32** no. 1, (2019) 16–21.

[195] K. Friston, "The free-energy principle: A unified brain theory?," *Nature Reviews Neuroscience* **11** no. 2, (2010) 127.

[196] A. Lebedev, M. Kaelen, M. Lövdén, J. Nilsson, A. Feilding, D. Nutt, and R. Carhart-Harris, "LSD-induced entropic brain activity predicts subsequent personality change," *Human Brain Mapping* **37** no. 9, (2016) 3203–3213.

[197] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen, "Fieldtrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data," *Computational Intelligence and Neuroscience* **2011** (2011) 1.

[198] M. Bola, P. Orłowski, K. Baranowska, M. Schartner, and A. Marchewka, "Informativeness of auditory stimuli does not affect EEG signal diversity," *Frontiers in Psychology* **9** (2018) 1820.

[199] R. R. Griffiths, W. A. Richards, M. W. Johnson, U. D. McCann, and R. Jesse, "Mystical-type experiences occasioned by psilocybin mediate the attribution of personal meaning and spiritual significance 14 months later," *Journal of Psychopharmacology* **22** no. 6, (2008) 621–632.

[200] K. A. MacLean, M. W. Johnson, and R. R. Griffiths, "Mystical experiences occasioned by the hallucinogen psilocybin lead to increases in the personality domain of openness," *Journal of Psychopharmacology* **25** no. 11, (2011) 1453–1461.

[201] G. H. Klem, H. O. Lüders, H. Jasper, C. Elger, *et al.*, "The ten-twenty electrode system of the International Federation," *Electroencephalography and Clinical Neurophysiology* **52** no. 3, (1999) 3–6.

[202] M. Csikszentmihalyi, *Finding Flow: The Psychology of Engagement with Everyday Life*. Basic Books, 1997.

[203] L. Noy, N. Levit-Binun, and Y. Golland, "Being in the zone: Physiological markers of togetherness in joint improvisation," *Frontiers in Human Neuroscience* **9** (2015) 187.

[204] A. S. Nilsen, B. E. Juel, and W. Marshall, "Evaluating approximations and heuristic measures of integrated information," *Entropy* **21** no. 5, (2019) .

[205] B.-L. Hao, *Elementary Symbolic Dynamics and Chaos in Dissipative Systems*. World Scientific, 1989.

[206] D. R. Chialvo, "Life at the edge: Complexity and criticality in biological function," arXiv:1810.11737.

[207] L. Barnett, J. T. Lizier, M. Harré, A. K. Seth, and T. Bossomaier, "Information flow in a kinetic Ising model peaks in the disordered phase," *Physical Review Letters* **111** no. 17, (2013) 177203.

[208] W. L. Shew, H. Yang, T. Petermann, R. Roy, and D. Plenz, "Neuronal avalanches imply maximum dynamic range in cortical networks at criticality," *Journal of Neuroscience* **29** no. 49, (2009) 15595–15600.

[209] A. Kolchinsky, "A novel approach to multivariate redundancy and synergy," arXiv:1908.08642.

[210] B. Rassouli, F. Rosas, and D. Gündüz, "Latent feature disclosure under perfect sample privacy," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7, IEEE. 2018.

[211] H. Morch, "Is consciousness intrinsic?: A problem for the integrated information theory," *Journal of Consciousness Studies* **26** no. 1-2, (2019) 133–162.

[212] G. Northoff, N. Tsuchiya, and H. Saigo, "Mathematics and the brain – A category theoretic approach to go beyond the neural correlates of consciousness," *bioRxiv* (2019) 674242.

[213] S. Tajima and R. Kanai, "Integrated information and dimensionality in continuous attractor dynamics," *Neuroscience of Consciousness* **2017** no. 1, (2017) nix011.

[214] R. Begleiter, R. El-Yaniv, and G. Yona, "On prediction using variable order Markov models," *Journal of Artificial Intelligence Research* **22** (2004) 385–421.

[215] P. Feyerabend, *Against Method*. Verso, 1993.

[216] R. Hamming, "You and your research." Lecture at the Bell Communications Research Colloquium Seminar, 1986.

[217] T. Schreiber, "Measuring information transfer," *Physical Review Letters* **85** no. 2, (Jul, 2000) 461–464.

[218] R. G. James and J. P. Crutchfield, "Multivariate dependence beyond Shannon information," *Entropy* **19** no. 10, (2017) 531.

[219] R. Cofré, C. Maldonado, and F. Rosas, "Large deviations properties of maximum entropy markov chains from spike trains," *Entropy* **20** no. 8, (2018) .

[220] J. Crampton and G. Loizou, "The completion of a poset in a lattice of antichains," *International Mathematical Journal* **1** no. 3, (2001) 223–238.

[221] R. Quax, O. Har-Shemesh, and P. Sloot, "Quantifying synergistic information using intermediate stochastic variables," *Entropy* **19** no. 2, (2017) 85.