PHILOSOPHICAL TRANSACTIONS A

royalsocietypublishing.org/journal/rsta

Research

Cite this article: Varley TF, Hoel E. 2022 Emergence as the conversion of information: a unifying theory. *Phil. Trans. R. Soc. A* **380**: 20210150.
https://doi.org/10.1098/rsta.2021.0150

Received: 31 May 2021
Accepted: 24 August 2021

One contribution of 17 to a theme issue 'Emergent phenomena in complex physical and socio-technical systems: from cells to societies'.

Subject Areas:
cybernetics

Keywords:
emergence, information, synergy, Boolean system, mutual information, partial information decomposition

Author for correspondence:
Erik Hoel
e-mail: hoelerik@gmail.com

THE ROYAL SOCIETY PUBLISHING

Emergence as the conversion of information: a unifying theory

Thomas F. Varley[1,2] and Erik Hoel[3]

[1]Department of Psychological and Brain Sciences, and [2]School of Informatics, Computing, and Engineering, Indiana University, Bloomington, IN, USA
[3]Allen Discovery Center, Tufts University, Medford, MA, USA

TFV, 0000-0002-3317-9882; EH, 0000-0002-2970-0057

Is reduction always a good scientific strategy? The existence of the special sciences above physics suggests not. Previous research has shown that dimensionality reduction (macroscales) can increase the dependency between elements of a system (a phenomenon called 'causal emergence'). Here, we provide an umbrella mathematical framework for emergence based on information conversion. We show evidence that coarse-graining can convert information from one 'type' to another. We demonstrate this using the well-understood mutual information measure applied to Boolean networks. Using partial information decomposition, the mutual information can be decomposed into redundant, unique and synergistic information atoms. Then by introducing a novel measure of the synergy bias of a given decomposition, we are able to show that the synergy component of a Boolean network's mutual information can increase at macroscales. This can occur even when there is no difference in the total mutual information between a macroscale and its underlying microscale, proving information conversion. We relate this broad framework to previous work, compare it to other theories, and argue it complexifies any notion of universal reduction in the sciences, since such reduction would likely lead to a loss of synergistic information in scientific models.

This article is part of the theme issue 'Emergent phenomena in complex physical and socio-technical systems: from cells to societies'.

1. Introduction

Reductionism is one of the classic principles of science. At the same time, science itself forms a diverse tree with elements at different spatio-temporal scales, such as quantum waves in physics, molecules in chemistry, cells in biology, all the way up to macroeconomics and sociology. Macroscale descriptions like biophysical models of cells, the machine code in computers, or organisms operating within a food web, are generally treated as if they reflect some intrinsic scale of function that cannot be neatly improved by reduction. The result is a contradiction between the theory and practice of science [1,2].

One resolution to this contradiction is the 'null hypothesis' of reductionism: that all macroscale descriptions, which broadly are some form of dimension reduction such as coarse-graining, are only useful due to computational constraints. This is because, according to this hypothesis, their underlying microscales contain all the information. That is, *information compression* is the only true benefit to analysing, modelling, or understanding a system at a macroscopic level. Compression of a given information source can be lossless or lossy, but can never lead to an overall gain of information [3]. Without any gain of information at a macroscale some have argued that macroscales cannot add anything above or beyond their underlying microscales and therefore should be considered epiphenomenal [4,5].

An alternative resolution to the contradiction between universal reductionism and science as practised is a formal theory of emergence. Such a formal theory should (a) directly and fairly compare microscales to macroscales, (b) offer a quantitative measurement of what a macroscale is providing in terms of information gain above and beyond compression and (c) enable the means to identify emergent scales in a given system or dataset. A non-trivial formal theory of emergence should allow for reduction or emergence on a case-by-case basis. Such a theory of emergence can solve longstanding problems in model choice for scientists, since it reveals the intrinsic scale of the function of systems.

Scientific work on emergence began to use information theory to examine things like parity checks in logic gates [6] or techniques like the Granger causality to look for emergence in time series [7]. The first explicit comparison between scales of a system was the theory of causal emergence, which claimed that causal relationships could be stronger at a macroscale (such as doing more work, being more predictive, or being more informative) [8]. To measure the informativeness of causal relationships, it made use of the effective information (*EI*), which is the mutual information between a set of interventions by an experimenter and their effects. More specifically, effective information is the mutual information following an experimenter intervening to set a system or part of a system to maximum entropy. It quantifies the number of YES/NO questions required to produce an output from an input, thus measuring the 'work' the system does in selecting that output [9].

The effective information has been shown mathematically to capture the causally relevant information in a system by being sensitive to the determinism (lack of uncertainty in state transitions) and degeneracy (similar or identical state transitions or dynamics, e.g. the necessity of a given state transition). Dimension reductions like coarse-graining (grouping elements of states in macro-elements or macro-states) or black-boxing (leaving elements or states exogenous) can increase the effective information [10]. Overall, the argument was that since macro-states were more deterministic and less degenerate, they did the most work to select the output of the system, and the effective information was able to identify the spatio-temporal scale at which this work peaked. According to the theory, this peak indicated that this scale, whether macro or micro, was the most causally relevant scale.

One possible criticism of the theory of causal emergence is that effective information is only one specific measure. The effective information was first proposed by Tononi & Sporns [11], effective information is mathematically well-described [8,9,12]. While it keeps being re-invented as a measure of causation [13,14], generally without acknowledgement of previous formulations or ongoing lines of research, the measure has not yet been proven

to be the unique measure of causation, and there are alternative proposals (many of which are mathematically related) for how to measure causation using information theory [15]. Indeed, the same general phenomena of causal emergence have been shown in integrated information [16,17] as the ϕ measure can increase at macroscales due to similar reasons of increasing determinism and decreasing degeneracy of state transitions. Measures that are not directly causal but capture aspects of causation, like assessing the entropy of random walkers on networks (indicating uncertainty of transitions), can also improve at macroscales in that random walkers are more deterministic in their dynamics [18]. The fact that causal emergence occurs across multiple measures indicates there is a broader phenomenon at work. Specifically, there is somehow more causally relevant information at macroscales (although it is currently unclear if this is captured by a unique measure of causation, or is better captured by a set of common measurements). Here, we explore broadly how such information gain is possible, and demonstrate it in the well-understood measure of the mutual information itself.

Of course, information cannot be created *ex nihilo*. Therefore, we propose that emergence is a form of *information conversion* at a higher scales. When measured in its totality in a given system, total information measures like the entropy of the distribution system states, the Kolmogorov information for describing the entirety of the system, or the total correlation in the form of the mutual information, all necessarily decrease at a macroscale (or at best, do not decrease, but can never increase). However, information can be converted from one type to another, with no change except for what scale the system is being modelled at, meaning there can be a gain of specific *types* of information at macroscales (causal emergence would then be a gain of causally relevant information in particular).

Herein, we seek to demonstrate general proof of information conversation across scales. In order to do so, we first eschew other measures and solely make use of the classic and well-understood measure of mutual information. Specifically, we consider the mutual information between the past and future of Boolean networks [19]. While total mutual information always decreases or remains constant at a macroscale (no information *ex nihilo*), this is not the full story. For the information itself can be decomposed into a set of partial information 'atoms' (PI atoms) that quantify how the total information is distributed over all of the elements of the system [20]. Herein, we show that, after coarse-graining, redundant information at the microscale can be converted into synergistic information at macroscales in an overall movement of information up the PI lattice, and that this effect exists even when no mutual information is lost. This indicates that the 'structure' of the mutual information is truly being converted from redundant to synergistic at macroscales.

In §2, we overview how we are applying the mutual information in our model system, and also its decomposition in multi-element systems. We introduce a novel measure of redundancy/synergy bias in the mutual information, based on how PI atoms are distributed across the PI-lattice. In §3, we examine systems of Boolean networks across scales and accompanying changes in the partial information decomposition (PID). First, we look at common macroscales of logic gates and find a clear shift toward synergies in some systems (e.g. we show how an XOR is more synergistic than its underlying microscale logic gates, none of which are XORs). Next, we examine sets of Boolean networks wherein the mutual information is identical at both micro and macroscales, and show that even in these cases there can be an increase in synergy bias at the macroscale. This shows direct evidence of information conversion at macroscales in that the mutual information becomes more dominated by its synergistic component without changing its total bit value, just its decomposition. In §4, we connect our work to previous research by demonstrating how causal emergence can be thought of as a form of information conversion, as the total entropy of transitions is converted to the causally relevant form of effective information via dimension reduction. Overall, we conclude that information conversion offers an agnostic umbrella explanation for theories of emergence based in information theory.

2. Mutual information and its decomposition in discrete systems

Here, we detail our application of the mutual information to probabilistic Boolean networks. A Boolean network is a canonical model system in complex systems science: a directed network where every vertex can be in one of a finite number of 'states' and, as time flows, the state of each vertex changes according to a logical function of the states of all of the parent neurons: for example, a vertex with two parents may implement an AND function, where its state at time $t+1$ is the logical AND of both parents. Systems can be viewed as passing information from the past to the future over the channel of the present. The quantification of this information flow can be done by calculating the mutual information between the future and past joint states of the variables that make up a network. Specifically, X represents the past states of the Boolean network, while Y represents the future states of the network. The calculation of $I(X, Y)$ quantifies how much knowledge of the past state of the system reduces our uncertainty about the future state of the system. Specifically,

$$I(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P(x,y) \log_2 \left(\frac{P(x,y)}{P(x)P(y)} \right). \tag{2.1}$$

This calculation requires defining the distributions $P(X)$, $P(Y)$ and $P(X, Y)$. The joint distribution is given by the transition probability matrix (TPM) of the system (with each row weighted by the probability of that state), and $P(X)$ is an 'input' distribution. To not bias our measurements, $P(X)$ is the stationary distribution of the system (in cases of multiple stationary distributions, we use the one with the largest attractor, or design Boolean networks such that all states are included in a single attractor). The 'output' distribution ($P(Y)$) can then be calculated as the matrix product of $P(X)$ and $P(Y \mid X)$. Note that every 'state' in the support sets $\mathcal{X}$ and $\mathcal{Y}$ actually represents the joint state of multiple variables in the underlying Boolean network.

This application of the mutual information captures the total amount of information in the dynamics of the system (the calculation of which requires the system's stationary dynamics $P(X)$). For example, in networks where the stationary distribution contains only a single state in the form of a point attractor the mutual information is zero, since the system is like a source that sends only a single message over a channel: there is no uncertainty about the future to be reduced by knowledge of the past. In networks where each state is visited equally in the stationary distribution, and each state deterministically transitions to a unique state, the mutual information would be maximized as $\log_2(n)$, as every 'message' the system sends is as informative as possible.

(a) Partial information decomposition

A core limitation of mutual information when assessing systems with more than two variables is that it gives little direct insight into *how* information is distributed over sets of multiple interacting variables. Consider the classic case of two elements X_1 and X_2 that regulate a third variable Y: it is easy to determine the information shared between either X_i and Y as $I(X_i; Y)$, and it is possible to calculate the joint mutual information $I(X_1, X_2; Y)$, however, these measures leave it ambiguous as to what information is associated with which combination of variables. For example, if $X_1 \not\perp X_2$, then there is necessarily some information about Y that is redundantly shared between both X_1 and X_2. Similarly, it is possible that there is *synergistic* information about Y that is only disclosed by the joint states of X_1 and X_2 together and not retrievable from either variable independently (for example, if all elements are binary and $Y = X_1 \oplus X_2$, then $I(X_1; Y) = I(X_2; Y) = 0$ bit but $I(X_1, X_2; Y) = 1$ bit).

To address this issue, the PID framework was introduced [20]. It provides a method by which the mutual information between the joint state of multiple sources variables and a single target variable can itself be decomposed. For the example case detailed above, with two 'source'

variables (X_1 and X_2) and a single 'target' variable Y, the full PID breaks $I(X_1, X_2; Y)$ down into the following additive combination of 'partial information atoms':

$$I(X_1, X_2; Y) = \mathrm{Red}(X_1, X_2; Y) + \mathrm{Unq}(X_1; Y \mid X_2) + \mathrm{Unq}(X_2; Y \mid X_1) + \mathrm{Syn}(X_1, X_2; Y), \tag{2.2}$$

where $\mathrm{Red}(X_1, X_2; Y)$ is the information about Y that is *redundantly shared* between X_1 and X_2 (i.e. an observer could learn the same information about Y examining either X_1 or X_2), $\mathrm{Unq}(X_1; Y \mid X_2)$ refers to the information about Y that is uniquely present in X_1 and not in X_2, and $\mathrm{Syn}(X_1, X_2; Y)$ is the information about Y that is *only* disclosed by the joint states of X_1 and X_2 considered together. Furthermore, the bivariate mutual information can also be decomposed:

$$I(X_1; Y) = \mathrm{Red}(X_1, X_2; Y) + \mathrm{Unq}(X_1; Y \mid X_2) \tag{2.3}$$

and

$$I(X_2; Y) = \mathrm{Red}(X_1, X_2; Y) + \mathrm{Unq}(X_2; Y \mid X_1). \tag{2.4}$$

The result is that equations (2.2), (2.3) and (2.4) form an under-determined system of three equations with four unknowns (Red, Unq_1, Unq_2, Syn). Given an appropriate function with which to compute any of these three, the rest are trivial.

As the number of sources informing about a single target grows, the number of combinations of sources that must be considered grow super-exponentially. The seminal contribution of Williams and Beer was to realize that it is not necessary to brute-force search every combination in the power-set of sources, but rather, that meaningful combination of sources are naturally structured into a partially ordered lattice, known as the partial information (PI) lattice. Furthermore, for a particular set of sources α, the value of the associated partial information atom Π_α can be calculated recursively as the difference between the information redundantly shared across the sources of interest, and the sum of all PI atoms lower down on the lattice:

$$\Pi(\alpha, Y) = \mathrm{Red}(\alpha, Y) - \sum_{\beta \prec \alpha} \Pi(\beta, Y), \tag{2.5}$$

where $\mathrm{Red}(\alpha, Y)$ is the *redundancy function*, which quantifies the information about Y that is redundantly present in every element of α. For readers interested in the deeper mathematical details of the construction of the PI lattice, we refer to [20], and more recently [21] for a more in-depth discussion. For our purposes, it suffices to understand that there exists a partial ordering of PI atoms, with 'more redundant' atoms towards the bottom. For example, in the case of three sources, the bottom of the PI-lattice is given by {0}{1}{2}, which refers to the information about the target that is redundantly present in all three sources. At the top of the lattice is {012}, which gives the information about the target that is only accessible when considering the joint state of all three sources jointly, and not disclosed by any 'simpler' combination of sources. It is important to note that, for systems with more than two sources, the PI atoms no longer break down into neatly intuitive categories of 'redundant', 'unique' and 'synergistic': more exotic combinations of sources appear, for example, {0}{12}, which gives the information about the target that is redundantly present in X_0 and the joint state of X_1 and X_2 together. However, in general, the lower down on the PI-lattice a PI atom is, the more redundant the information is, while the higher on the lattice, the more synergistic.

While the PID framework provides the scaffold on which information can be decomposed, it fails to provide the specific keystone necessary to actually calculate it: the redundancy function that forms the base of the PI-lattice. Williams and Beer proposed the *specific information* as a plausible redundancy function typically denoted as I_{WB}:

$$I_{\mathrm{WB}}(\mathbf{X}; Y) = \sum_{y \in \mathcal{Y}} P(y) \min_{X_i \in \mathbf{X}} I(X_i; y). \tag{2.6}$$

The specific information quantifies the average minimum amount of every element of $\mathbf{X}$ discloses about Y. The term $\min_{X_i \in \mathbf{X}} I(X_i; y)$ calculates the minimal amount of information any $X_i \in \mathbf{X}$ provides about the specific state $Y = y$. Across all $y \in \mathcal{Y}$, I_{WB} quantifies the expected minimum amount of information that $\mathbf{X}$ will disclose about Y. As a redundancy function, I_{WB} has

a number of appealing quantities: in contrast to other redundancy functions, it will only return values greater than or equal to zero bit. Given the perennial difficulties of interpreting negatively valued information quantities, this is a key property, one not shared by most other redundancy functions. Furthermore, I_{WB} is both conceptually and computationally simple: being based on the specific information, it is a 'pure' information-theoretic measure and does not require leveraging theory or algorithms from fields like information geometry [22], game theory [23] or decision theory [24,25] and is one of the fastest-running functions in the *Discrete Information Theory* toolbox [26]. I_{WB} is also arguably the most well-used redundancy function in the scientific literature, having been the function of choice in [27–31].

Williams and Beer's redundancy function has been critiqued for behaving in an 'unintuitive' manner in some cases [32], and does not readily localize the way that the mutual information function does [33]. Since PID was initially proposed, considerable work has gone into developing an 'optimal' redundancy function, resulting in a plethora of measures [22,32–39]. So far, no single measure has emerged as the accepted 'gold standard', although they share many commonalities. Each function comes with it is own limitations (for example, only being defined for two variables, or occasionally returning negative quantities of partial information, or violating some intuitions about how such a measure 'should' behave), so care is necessary when deciding which one to use.

In this work, we used the original measure put forth by Williams and Beer, as it was necessary that whatever measure we chose never return negative quantities of information, be applicable to systems with more than two sources, and run in reasonable times. I_{WB} remains widely used and the most studied. We used the *Discrete Information Theory* package [26] for the implementation of I_{WB} and all PID calculations. In future work, we aim to replicate these findings using alternative redundancy functions and related frameworks.

(b) PID of temporal mutual information

PID is usually applied to situations like those given in the example above, where a set of sources (neurons, perceptrons, etc.) synapse onto a single target and is often applied in such cases [27–31]. Here, we detail our application of the PID of the mutual information between the past and the future in Boolean networks.

Consider a Markovian system with two interacting elements that is evolving in time. Following the convention given above, we will say that $\mathbf{X} = \{X_1, X_2\}$ indicates the past states of both elements of our system, and $\mathbf{Y} = \{Y_1, Y_2\}$ indicates the future states of both elements of our system. We can then adapt the classic PID framework by defining our 'sources' as every $X_i \in \mathbf{X}$, and our single target as the joint future state $\mathbf{Y}$. The PID of this two-element dynamical system is then given by

$$I(\mathbf{X}, \mathbf{Y}) = \mathrm{Red}(X_1, X_2; \mathbf{Y}) + \mathrm{Unq}(X_1; \mathbf{Y} \mid X_2) + \mathrm{Unq}(X_2; \mathbf{Y} \mid X_1) + \mathrm{Syn}(X_1, X_2; \mathbf{Y}). \tag{2.7}$$

This decomposition details how information about the next joint-state of the system is distributed [40].

(c) Introducing synergy and redundancy biases

In our two toy examples above, we relied heavily on the categorical distinction between redundant, unique and synergistic information. These classifications are useful for building intuition, but do not readily generalize to systems with more than two elements. To address this, we introduce the construct of a *partial information spectrum*, from which one can calculate the relative top- or bottom-heaviness of a PI lattice without directly having to define well-delineated 'pools' of redundant, unique, or synergistic information (figure 1).

Recall that the value of a given PI atom is calculated recursively from the sums of every PI atom lower than it down on the lattice; PI atoms higher on the lattice contain information that is increasingly synergistic and cannot be extracted from any simpler combination of sources. Because the PI lattice is partially ordered, there are collections of PI atoms that are at the same

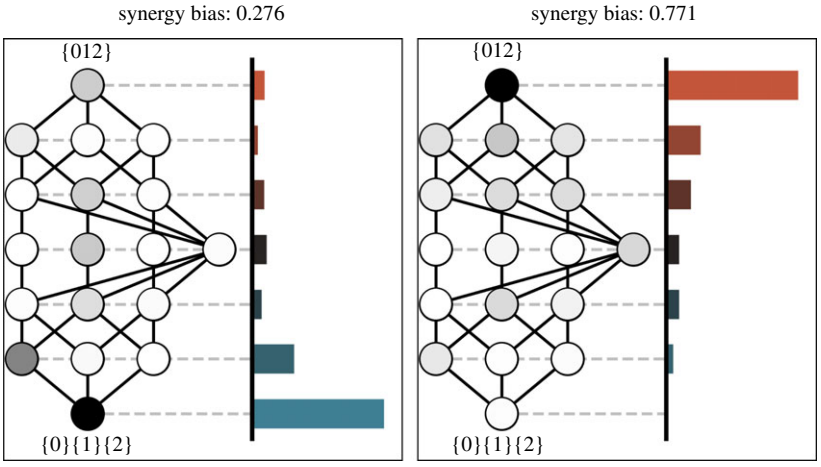

Figure 1. The partial information spectrum. The PI lattices for two, three-element systems. For each system, from the PI lattice we can create a PI spectrum, which gives the proportion of the total mutual information present in all PI atoms at a given 'height' on the lattice. Left: this system has a low synergy bias (high redundancy bias): the majority of the mutual information about the joint future state is redundantly shared across all the elements ({0}{1}{2}), or other highly redundant PI atoms (e.g. {0}{1}). Right: this three-element system has a high degree of synergy bias, with the majority of the information about the joint future present only in the joint state of all three elements ({012}). (Online version in colour.)

'height' on the lattice relative to the bottom (the maximum redundancy atom) or the top (the maximum synergy atom). We claim that these atoms comprise a 'layer' of the lattice defined by some ratio of redundancy to synergy. The PI spectrum S is then defined as an ordered list where the ith bin in the spectrum is given by the proportion of total mutual information present in all PI atoms in the ith layer.

Once the PI spectrum (S) has been calculated, it is easy to determine how top-heavy it is using a measure similar to the Earth Mover's Distance. We define the synergy bias ($B_{\mathrm{syn}}(S)$) as the amount of normalized partial information in each layer (S_i being the sum of all PI atoms in the ith layer, divided by the joint mutual information), weighted by the number of steps that layer is from the bottom.

$$B_{\mathrm{syn}}(S) = \sum_{i=0}^{|S|} \frac{i}{|S|} S_i, \tag{2.8}$$

where i indexes the layer (starting from the bottom, maximally redundant layer) and $|S|$ gives the total number of layers in the lattice. By normalizing by the total number of layers, we can compare the synergy bias between two different sized lattices, since we are looking at the proportion of the total lattice height moved, rather than counting the actual number of layers.

The redundancy bias is defined equivalently, although the reference point is at the top of the lattice, rather than the bottom:

$$B_{\mathrm{red}}(S) = 1 - \sum_{i=0}^{|S|} \frac{i}{|S|} S_i. \tag{2.9}$$

Conveniently, since both measures are symmetric and normalized by the total joint mutual information, $B_{\mathrm{syn}} + B_{\mathrm{red}} = 1$, so we only ever have to calculate one and both are greater than zero.

The synergy and redundancy biases allow us to compare how top- and bottom-heavy two different PI lattices are: a high synergy bias indicates that most of the partial information is present in synergistic relationships between elements, while a high redundancy bias indicates that most of the partial information is redundantly present across multiple individual elements. Since it is a normalized measure, we can compare the top- and bottom-heaviness of systems with different numbers of elements (and consequently, different sized lattices) and thus can measure synergy/redundancy bias across scales.

3. Evidence of information conversion across scales

(a) Macroscales can increase the synergy bias of information

We begin with a well-known type of macroscale as our model system: that of a single logic gate, which itself is some dimension reduction of a larger collection of networked gates. By breaking down three basic logic gates with distinct mechanisms (AND, OR and XOR) into collections of microscale gates with simpler mechanisms, we can directly and fairly compare the microscales and macroscales in terms of their respective distribution of partial information. This provides the first showcase of information conversion across scales.

Note that here we are technically leaving elements exogenous in time in our macroscale (since the microscale networks require multiple timesteps to run), and all mechanisms have been coarse-grained into the single mechanism (again, broadly, these are all referred to as forms of dimension reduction). Such example systems of logic gates have no stationary dynamics, since they are not closed, but are open systems. Therefore, to calculate the mutual information, in all cases (both micro and macro) the same input distribution to either the macroscale mechanism's inputs or the inputs to the network of microscale logic gates that underlies them. For solely explanatory purposes, we make use of the maximum entropy as our input distribution in all cases, and this means that $P(X)$ is identical for both macro and microscales in our comparisons.

By calculating the mutual information of the macro and microscales with the same inputs, and decomposing the result using PID, we can see that the synergy bias of the system is not constant between scales. Consider the XOR gate, which can be decomposed into a network of one NAND gate, one AND gate and one OR gate (as well as two inputs A and B). The resulting system has five elements compared to the macroscale, which has three elements (XOR, A and B). We found that, at the microscale, the system had mutual information of 2.5 bit, while it had the expected 1 bit of mutual information at the macro scale. However, while the total mutual information decreased for the XOR gate macroscale, the synergy of that information increased, from 0.52 at the microscale to 0.83 at the macroscale. The same was true for both the OR and the AND gates, although to a lesser degree (final values can be found in table 1). Note that while the AND and OR gates have the same macroscale mutual information and synergy bias (since they are isomorphic), they have different microscale values, which reflects the different number and structure of NAND gates required to implement them.

This suggests that, while dimension reduction reduces the overall amount of information in a system, the 'leftover' information can move higher on the macroscale PID lattice. This can be seen directly in figure 2, which shows the PI spectra based off on the PI lattices for three macroscale mechanisms and their underlying microscales of networked logic gates.

When considering our logic gate result, it is clear that dimension reductions like coarse-graining can alter the distribution of information in the PI lattice of a system, even when both scales are simply a different description of the same system. Note that this result fits intuitively with the idea that something is being gained by modelling an XOR gate as an actual XOR gate, even if it is made of a set of underlying logic gates that, like NAND gates, are not themselves XORs. What is gained is a distinctive bias toward synergy in the information flowing through the system. Additionally, it is intuitive that XORs should greatly demonstrate this effect, like ANDs and ORs demonstrate it to only a slight degree. While not shown, it should be obvious that there can be reverse cases; for example, some macroscales may be more redundant than their underlying microscales, since there is no limit to how complex a microscale can be.

(b) Redundant to synergistic information conversion

In the cases of classic logic gate composition above in §a, it could be argued by a skeptic that while the synergy biases are indeed increasing at the macroscale, this is because all the information at the bottom of the lattice is being removed by dimension reduction. This may be true in some instances. Luckily, we can provide direct evidence that the effect we have observed is not just

Table 1. The macro and microscale temporal mutual information and their respective synergy biases.

gate	microscale MI	macroscale MI	microscale syn. bias	macroscale syn. bias
AND	1.623 bit	0.811 bit	0.533	0.578
OR	2.811 bit	0.811 bit	0.518	0.578
XOR	2.5 bit	1 bit	0.52	0.833

For three logic gates (AND, OR and XOR), this table shows the effects that going up a level of abstraction has on the temporal mutual information and the synergy biases. It is important to understand that while the micro and macroscale implement the same function, the temporal mutual information can be very different.

that some types of information (like redundant information at the bottom of the lattice) are lost at macroscales. Rather, there is evidence that information is being converted from one type to another (or more precisely, information is moving up the PI spectrum from redundant to more synergistic at the macroscale).

To showcases of information conversion, we developed a method by which the mutual information can be kept constant across scales. Since the mutual information in terms of total bits is not decreasing at the macroscale, any change in synergy bias must be from the conversion of information, not its loss.

First, it is important to note that a neglected aspect of fairly comparing micro and macroscales is making sure that the macroscales are viable models of the system. It is therefore critical that the dynamics between a macroscale model and a microscale model are either identical (as in our cases), or highly similar. That is, dimension reduction shouldn't lead to significant differences in dynamics, nor to responses to interventions, or else the macroscale is a poor model of the system. This has been called 'consistency' and has been explored in structural equation modelling specifically of equivalence classes [41]. Given that precise consistency is not always possible, it is possible to measure the amount of inconsistency as the difference between the dynamics of the microscale and that of the macroscale, and an informational measure of inconsistency was therefore introduced that can analyse how consistent a wide variety of systems are [18]. It is worth noting that in this work, because of the use of equivalence classes in our method of expansion, in addition to the mutual information being constant, all scales used here also have zero inconsistency according to this measure. Such perfectly consistent macroscales do not necessarily need to be constructed via equivalence classes in order to ensure consistency, as we do here, since there are various types of macroscales that can give complete consistency [18].

What follows is the description of our method to hold the mutual information constant and ensure consistency between scales. This 'expansion method' introduced here is based on generating a Boolean network (represented by some TPM). Assuming this Boolean network is a macroscale, we can then bifurcate nodes in the network in such a way as to create an equivalence class. When we split a single node in an N node network, we go from a system with 2^N joint states to a microscale system with 2^{N+1} joint states. By re-allocating the probabilities of transitions across the new, expanded state-space, we can 'fix' the total joint, temporal mutual information, despite increasing the dimensionality of the overall system. Effectively, this allows for the generation of microscales from a given macroscale, an inversion of the normal process of finding macroscales from a given microscale [8]. Relevant Python code can be found at https://github.com/EI-research-group/synergistic_information_emergence. This process allows us to create different arbitrary microscales for a given system, while the mutual information remains fixed between the two scales. Therefore, any change in bias on the PI spectrum comes from the conversion of information from one type to another.

To illustrate this, we constructed 200 TPMs that were positive-Gaussian by initially generating 8×8 random Gaussian matrices from a distribution with a mean of 0 and a standard deviation of 1, taking the absolute value of every entry, and finally normalizing the rows to define discrete probability distributions. The resulting TPMs describe the stochastic dynamics of 200 distinct,

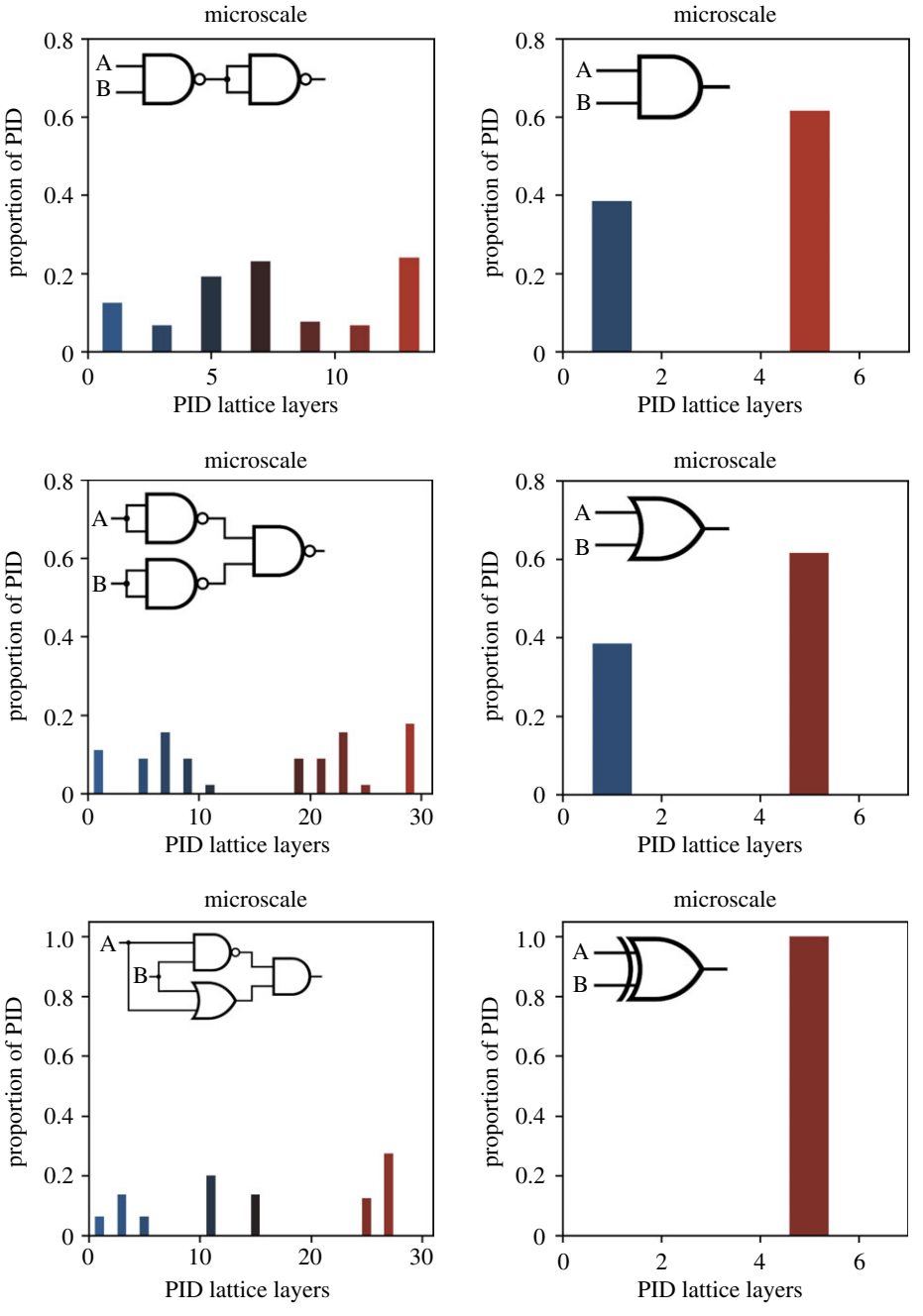

Figure 2. Partial information spectra for logic gates. Partial information spectra for three logic gates: AND (top), OR (middle), XOR (bottom) at the macroscale (right) and the microscale (left). The synergy biases and temporal mutual information values can be viewed in table 1. (Online version in colour.)

fully connected three-element systems with binary states. These are our starting macroscales. For each of these 200 systems, we split one element to create a four-element system, two of which are from our initial macroscale, and two of which are 'children' of the initial split macroscale element. We then re-calculate the PI spectrum and synergy bias for our new microscale. We can carry out this process of creating children in an equivalence class more than once: if so, we call the

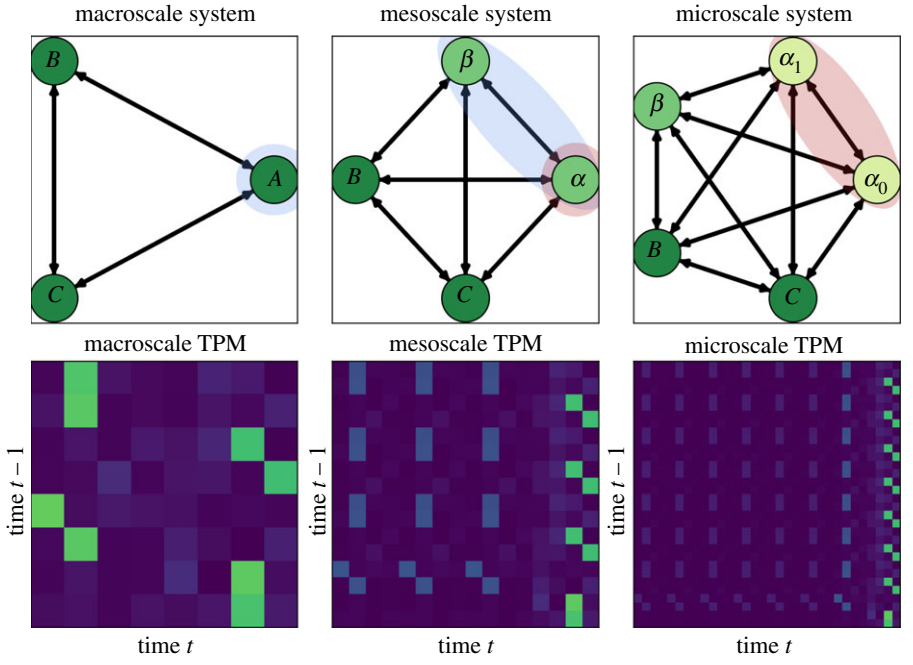

Figure 3. Multi-scale analysis. Here, we can see how to construct equivalence class microscales from a given macroscale such that mutual information is fixed. Left: a three-element system (top), and its associated TPM. We select a single node (A) to expand. Middle: expanding node A into nodes α and β results in a four-element system, which crucially preserves the mutual information from the joint-past to the joint future. We can select another node (α) to expand again, resulting in Right: the final microscale expansion of our system. Note that the original node A has been expanded twice, while the overall mutual information dynamics are preserved in all cases. (Online version in colour.)

four-node network a 'mesoscale' and the five-node network a 'microscale'. An example system is shown in figure 3, as well as details of the entire process as it is expanded into a mesoscale and then microscale (non-macroscale nodes in figure 3 are referred to as α and β in this process).

First, all of the 200 Gaussian systems showed an increased bias toward synergy at the macroscale, despite the mutual information being unchanged across scales. In general, our hypothesis was that the higher the synergy bias of the macroscale, the more that synergy would be lost at the microscale. This appears to be true in these model systems; in figure 4, we correlated the gain in synergy bias following the conversion of the microscale to the macroscale, against the macroscale synergy bias. Pearson's product-moment correlation found a highly significant positive correlation between macroscale synergy bias and the gain in synergy bias under repeated coarse-graining of the microscales (see figure 4, left, $\rho = 0.819$, $p < 10^{-10}$). This suggests that, for random stochastic systems, even when total mutual information is constant across scales, the systems have more redundant information at the microscale than at the macroscale. This is proof that dimensionality reductions exist that increase the overall synergy of the system by converting information to be more synergistic.

In addition to the Gaussian matrices, we also constructed 185 'deterministic' systems, where a single joint state led predictably to another single joint state with probability 0.99 (the remaining probability was evenly distributed to ensure the system was ergodic, did not have fixed point attractor, and that the stationary distribution involved all system states), which we expanded in the same manner as described above into both 'meso' and microscales. By exploring both highly stochastic and deterministic systems, we can generate a richer sample of the space of all three-element Boolean systems and identify more universal patterns.

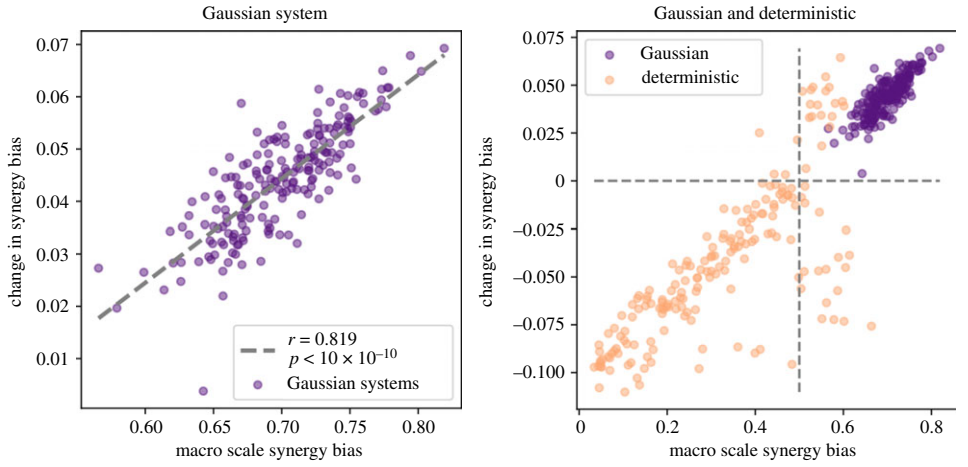

Figure 4. Change in synergy bias across scales. Right: the change in synergy from microscale to macroscale plotted against the starting macroscale synergy for Gaussian systems. There is a clear positive correlation. Indeed, all systems show an increase in synergy at the macroscale and an increase in redundancy at the microscale. The equivalence class structure used to construct the systems holds the mutual information steady across scales, so an increase in synergistic information at the macroscale can only come from a decrease in the redundant information of the microscale. Left: the same plot, although here we display both the Gaussian and deterministic systems. Note that, in contrast to the Gaussian systems, the deterministic systems start at a much lower synergy bias. Typically, these lower synergy systems actually *lose* synergy bias after coarse-graining, although the linear relationship between how synergistic the macroscale is and how much synergy is lost at the microscale remains. Interestingly, visual examination suggests that for both classes of system, the relationship between the macroscale synergy bias and the change in synergy bias is generally linear and, for both systems, lies along a common line of best fit. This suggests that, while different systems behave differently under coarse-graining, the broader relationship is preserved. (Online version in colour.)

We found similar results with the deterministic systems, although several clear differences are immediately apparent upon inspection (see figure 4, right). First, it's clear that deterministic systems of the kind we are creating have a lower synergy bias overall than using the Gaussian method of creating systems. In cases where the macroscales are indeed synergistic, like they are in Gaussian systems (above 0.5 in synergy bias), it is also the case that the underlying microscales are more redundant. However, in deterministic macroscales that are biased toward redundancy (below 0.5 in synergy bias), they can actually be expanded into microscales that are themselves more synergy biased (and the correlation between the change in synergy and the macroscale synergy remained positive).

4. Causal emergence as information conversion

Given the evidence that information conversion can occur at macroscales, even in a well-understood baseline measure like the mutual information, it is next necessary to establish what it means and how it relates to other information-theoretic approaches to emergence. How does it play out in measures like the effective information, which was specifically designed to capture causal influence and can peak at a macroscale?

A hint comes from the already well-established fact that the effective information (*EI*) changes across scales due to a change in determinism and/or degeneracy [8,10,18]. Indeed, it is already proven that:

$$EI(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y)\log_2\left(\frac{P(x,y)}{P(x)P(y)}\right)\Bigg| P(X) = H^{\max}, \tag{4.1}$$

which differs from the normal mutual information calculation in that $P(X)$ is set to $H^{\max}$, and can be rewritten as

$$\text{Effective information} = \text{determinism} - \text{degeneracy}. \tag{4.2}$$

In this interpretation of the effective information, the determinism is based on the information lost via uncertainty in state transitions:

$$\text{determinism} = \log_2(N) - \langle H(p(y) \mid P(X) = H^{\max})\rangle. \tag{4.3}$$

The term $\log_2(N)$ can be understood as the uncertainty about the future state of a maximally entropic system with N unique states. The second term $\langle H(p(y) \mid P(X) = H^{\max})\rangle$ gives the average uncertainty about the future state of our real system X (note that this is applied over a single timestep, e.g. t to t_{+1}). The average uncertainty is a function of the noise, wherein $\langle H(p(y) \mid P(X) = H^{\max})\rangle$ is zero if there is no noise in any transition (and the system is therefore deterministic).

The difference between the two terms (the hypothetical maximum entropy and the empirical entropy) gives us a measure of how much better we are at predicting the future of X than we would be in the 'worst case scenario'. If effective information is increasing at a macroscale due to an increase in the determinism term, then the entropy term in the determinism must itself be necessarily decreasing. This is because $\log_2(N)$ also necessarily decreases at the macroscale, so therefore any increases in the determinism term must come from a greater decrease in $\langle H(p(y) \mid P(X) = H^{\max})\rangle$ than $\log_2(N)$. Figure 4(left) shows this decrease in the information (the uncertainty of transitions), which can lead to an increase in effective information at a macroscale.

The degeneracy contains a similar entropy term and a size term:

$$\text{degeneracy} = \log_2(N) - H(\langle p(y) \mid P(X) = H^{\max}\rangle). \tag{4.4}$$

The degeneracy term is very similar structurally to the determinism term: once again the $\log_2(N)$ represents the maximally entropic 'reference' system. The term $H\langle(p(y) \mid P(X) = H^{\max})\rangle$ quantifies the uncertainty in retrodiction of a previous state, given a current state. That is, degeneracy is the amount of information lost (in a single timestep) if the system is 'reversed' in time.

The degeneracy can be thought of as the amount of information about the past that is lost when multiple causal paths 'run together'. For instance, we can imagine a system where two unique states both lead to the same subsequent state. In this case, information about the past is lost because it is impossible to determine which of the two prior states preceded the current one from just the current state alone. It can also be thought of as quantification of how different each state's transitions are. If in a system every state has a unique transition, then the degeneracy $H\langle(p(y) \mid P(X) = H^{\max}\rangle)$ term is zero, and degeneracy maximal.

The case of degeneracy is more complicated, since here the entropy term can increase at macroscales. However, as in determinism, the $\log_2(N)$ term always decreases. And it is actually the decreasing of $\log_2(N)$ that can lead to a decrease in degeneracy at the macroscale, e.g. if two states both deterministically transition to a single state, then a grouping over those (or an equivalent drop out of one) leads to an decrease in degeneracy since $\log_2(N)$ decreases. This leads to an increase in *EI* because the degeneracy term itself is subtracted in the *EI* equation.

Notably, it is well-known that the effective information cannot increase at a macroscale if determinism is maximal and degeneracy is minimal. Why? Because there is no information to convert, neither in terms of the size of the system $\log_2(N)$ nor in terms of the uncertainty of transitions $\langle H(p(y) \mid P(X) = H^{\max})\rangle$. Therefore, causal emergence can be conceived as the conversion of causally irrelevant information (like the uncertainty of state transitions) to causally relevant information (the effective information), fitting into the umbrella theory of emergence as information conversion.

We have offered forth an umbrella theory of emergence based on how changes in scale lead to the conversion of one type of information to another. That is, dimension reduction does not necessarily leave all types of information invariant. We claim that the best way to consider

these questions is to examine how changes in modelling, such as dimension reduction, change information type. While some information measures, like the total correlation between past and future measured by the mutual information, can only stay constant or decrease at macroscales, such measures can still demonstrate information conversion (such as here from redundant to synergistic information). We have shown this effect in Boolean networks, such as comparing a macroscale XOR to its underlying logic gates, none of which are XORs at the microscale. We have also shown it in cases of equivalence classes where the mutual information is held constant across scales and yet still information can become more biased toward synergy at the macroscale, proving information conversion.

Further future work may be examining things like at what scale synergistic information peaks, or how to find scales that maximally convert information while minimally losing information. Though we have shown evidence that some redundant information must become synergistic (or vice versa), it remains to be understood exactly *which* information changes form. Recent work on decomposing the local mutual information into directed local entropies may provide an interesting path forward [42]. Another promising future direction of research would be to introduce local information analysis to this pipeline [33,43].

More broadly, this umbrella theory of emergence reveals that there are measurable benefits to macroscale models. If so, this is likely to have been selected for in science, i.e.members of the special sciences can be viewed as converting redundant information into a more useful form. Indeed, it has even been shown that synergistic information processing can be key to certain games like to a successful poker strategy [44]. Our hypothesis is that the special sciences, and macroscale models in general, involve the conversion of redundant information into synergistic and unique information, making such macroscales useful for experimenters above and beyond their degree of compression.

Tying this to previous research, macroscale modelling can also convert causally irrelevant information to causally relevant information by making causal relationships between variables in a model more dependent (by increasing determinism or decreasing degeneracy), i.e. causal emergence. Note that there are clear advantages to identifying scales at which variables are more dependent. For instance, it has been shown that biological networks show more causal emergence than comparable technological or social networks [18]. This is probably because there are multiple advantages for macroscales, ranging from a lower entropy of random walkers to greater global efficiency at macroscales [45]. Some preliminary research examining the protein interactomes of over 1000 species shows that macroscales have become more likely to demonstrate causal emergence over evolutionary time [46]. This may even be one reason that controlling biological systems is so difficult: they are cryptic by having an intrinsic functional scale be a difficult-to-discover macroscale, making biological networks more robust to failure and less likely to be controlled from the outside [47].

The sort of above analyses are just a fraction of the applications of a formal theory of emergence, which is ultimately a toolkit for identifying the intrinsic scale of the function of complex systems. This issue of identifying intrinsic scale crops up all across the sciences, such as modelling gene regulatory networks in biology [47], understanding whether the brain functions at the scale of neurons or minicolumns [48,49], answering what level of abstraction is appropriate for modelling and comparing deep neural networks [50], or even examining the best scale to model cardiac systems at [51]. Previous research has shown how intrinsic scale comes about via growth rules, e.g. networks that develop causally emergent macroscales only do so once the network is no longer growing in a 'scale-free' manner [18]. Ultimately, by tracking information conversion, experimenters and modellers can close in on the intrinsic scale of function for a given system.

(a) Comparison to other theories of emergence

To help the nascent field of formal theories of emergence, it is important to discuss exactly what type of emergence we mean here, and compare and contrast to other definitions. For

example, there is no doubt that simple laws and interactions in systems can lead to the emergence of complex, interesting, or unexpected dynamics, such as in cases of symmetry breaking [52] or simple rule iteration [53]. Sometimes this is referred to as 'emergence'. However, this phenomenon of complexity emerging from simplicity is not conceptually mysterious, and is already well-understood mathematically.

Another use of the term 'emergence' comes from thinking about joint effects, which is a 'whole versus parts' emergence. Ultimately, this is motivated by the fact that elements in a system can exhibit behaviour, dynamics or functions that would not take place if they were partitioned or isolated from the rest of the system. One such measure that captures how much information is lost by partitioning individual elements off from a given system is Integrated Information Theory [54,55]. However, the mere fact that joint sets of elements behave differently compared to isolated elements in terms of effects or information flow does not by itself capture what is lost by reduction to some lower scale of explanation. IIT contains an explicit built-in distinction between 'whole vs. parts' emergence and 'macro vs. micro' emergence (the latter is when some supervenient model of the system is compared to its underlying model, such as a higher scale to a lower scale) [16].

There are also extensions of IIT, such as the 'integrated information decomposition' (ΦID) [40,56]. As with the work described here, the ΦID framework actually takes as its starting point the decomposition of the mutual information between past and future. However, the ΦID framework constructs a double PI lattice that describes how information moves from one PI atom to another through time. In this framework, the authors define 'emergent' information as when a supervenient feature (a dimension reduction in our language) contains unique information across time that is not present in the unique information of its independent underlying microscale parts. However, this means leaving out the synergistic and redundant information in the comparison between macro and micro, and a fair comparison would involve these, since the underlying microscale will almost certainly have joint effects. Additionally, purely unique information generally vanishes as a system grows in size.

By focusing on an explicit comparison between the fully modelled micro and macroscales of systems, this allows us to examine the kind of emergence that is traditionally discussed in analytic philosophy, which involves issues of supervenience, multiple-realizablity and causality [57]. There it is sometimes called 'synchronic' emergence [58], although we eschew this term as confusing for scientific usage, and continue to use 'macro vs. micro' to refer to this sort of emergence. As we have shown, such a mathematical theory of emergence is part of modeller and experimenter toolkits when it comes to identifying intrinsic scales of function, as well as modelling practices. This process involves explicit modelling of different scales of model or physical systems, followed by their comparison.

While related to philosophical discussions of emergence, note that our proposal of emergence as information conversion does not fit cleanly into the traditional strong/weak emergence dichotomies in philosophy [59]. Supervenience is not violated when information is converted from one type to another (such as redundant mutual information becoming synergistic, uncertainty of transitions being transformed into effective information, etc). In this view, the reduction is always possible when supervenience holds, and therefore there is always an identifiable procedure to map one scale to another. However, such reduction can lead to a real and measurable loss of a given type of information. This offers a subtle but powerful explanation as to what advantages macroscale models provide above and beyond compression, and may explain the necessary existence of the special sciences.

Data accessibility. All scripts necessary to replicate the results and figures reported here are available in supplementary material.

Authors' contributions. E.H. conceptualized and supervised the study, T.F.V. wrote the scripts, performed the data analysis, and made visualizations. Both authors drafted and edited the manuscript. Both authors read and approved the manuscript.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. The authors declare that they have no competing interests.

Funding. This work was supported and made possible by grant #W911NF2010243 from the USA Army Research Office and the NSF-NRT (grant no. 1735095), Interdisciplinary Training in Complex Networks and Systems at Indiana University Bloomington.

Acknowledgements. We would like to acknowledge Dr Olaf Sporns for his support, and Dr Fernando Rosas for his insightful comments about our manuscript.

References

1. Fodor JA. 1974 Special sciences (or: The disunity of science as a working hypothesis). *Synthese* **28**, 97–115. (doi:10.1007/BF00485230)
2. Hoel EP. 2018 Agent above, atom below: how agents causally emerge from their underlying microphysics. In *Wandering towards a goal* (eds A Aguirre, B Foster, Z Merali), pp. 63–76. New York, NY: Springer.
3. Cover TM, Thomas JA. 2012 *Elements of information theory*. Hoboken, NJ: John Wiley & Sons.
4. Kim J. 1998 *Mind in a physical world: an essay on the mind-body problem and mental causation*. Cambridge, MA: MIT Press.
5. Bontly TD. 2002 The supervenience argument generalizes. *Phil. Stud.* **109**, 75–96. (doi:10.1023/A:1015786809364)
6. Bar-Yam Y. 2004 A mathematical theory of strong emergence using multiscale variety. *Complexity* **9**, 15–24. (doi:10.1002/cplx.20029)
7. Seth A. 2008 Measuring emergence via nonlinear Granger causality. In *Alife* (eds S Bullock, J Nobel, R Watson, M Bedau), pp. 545–552. Cambridge, MA: MIT Press.
8. Hoel EP, Albantakis L, Tononi G. 2013 Quantifying causal emergence shows that macro can beat micro. *Proc. Natl Acad. Sci. USA* **110**, 19 790–19 795. (doi:10.1073/pnas.1314922110)
9. Balduzzi D. 2011 Information, learning and falsification. (http://arxiv.org/abs/1110.3592)
10. Hoel EP. 2017 When the map is better than the territory. *Entropy* **19**, 188. (doi:10.3390/e19050188)
11. Tononi G, Sporns O. 2003 Measuring information integration. *BMC Neurosci.* **4**, 1–20. (doi:10.1186/1471-2202-4-31)
12. Balduzzi D, Tononi G. 2008 Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Comput. Biol.* **4**, e1000091. (doi:10.1371/journal.pcbi.1000091)
13. Korb KB, Hope LR, Nyberg EP. 2009 Information-theoretic causal power. In *Information theory and statistical learning* (eds F Emmert-Streib, M Dehmer), pp. 231–265. New York, NY: Springer.
14. Griffiths PE, Pocheville A, Calcott B, Stotz K, Kim H, Knight R. 2015 Measuring causal specificity. *Phil. Sci.* **82**, 529–555. (doi:10.1086/682914)
15. Albantakis L, Marshall W, Hoel E, Tononi G. 2019 What caused what? A quantitative account of actual causation using dynamical causal networks. *Entropy* **21**, 459. (doi:10.3390/e21050459)
16. Hoel EP, Albantakis L, Marshall W, Tononi G. 2016 Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neurosci. Consciousness* **2016**, niw012. (doi:10.1093/nc/niw012)
17. Marshall W, Albantakis L, Tononi G. 2018 Black-boxing and cause-effect power. *PLoS Comput. Biol.* **14**, e1006114. (doi:10.1371/journal.pcbi.1006114)
18. Klein B, Hoel E. 2020 The emergence of informative higher scales in complex networks. *Complexity* **2020**, 8932526 . (doi:10.1155/2020/8932526)
19. Kauffman S. 1969 Homeostasis and differentiation in random genetic control networks. *Nature* **224**, 177–178. (doi:10.1038/224177a0)
20. Williams PL, Beer RD. 2010 Nonnegative decomposition of multivariate information. (http://arxiv.org/abs/1004.2515 [math-ph, physics:physics, q-bio])
21. Gutknecht AJ, Wibral M, Makkeh A. 2020 Bits and pieces: understanding information decomposition from part-whole relationships and formal logic. (http://arxiv.org/abs/2008.09535)
22. Harder M, Salge C, Polani D. 2013 Bivariate measure of redundant information. *Phys. Rev. E* **87**, 012130. (doi:10.1103/PhysRevE.87.012130)
23. Ay N, Polani D, Virgo N. 2019 Information decomposition based on cooperative game theory. (http://arxiv.org/abs/1910.05979)

24. Bertschinger N, Rauh J, Olbrich E, Jost J, Ay N. 2014 Quantifying unique information. *Entropy* **16**, 2161–2183. (doi:10.3390/e16042161)
25. Kolchinsky A. 2019 A novel approach to multivariate redundancy and synergy. (http://arxiv.org/abs/1908.08642)
26. James R, Ellison C, Crutchfield J. 2018 dit: a Python package for discrete information theory. *J. Open Sour. Softw.* **3**, 738. (doi:10.21105/joss.00738)
27. Timme N, Alford W, Flecker B, Beggs JM. 2014 Synergy, redundancy, and multivariate information measures: an experimentalist's perspective. *J. Comput. Neurosci.* **36**, 119–140. (doi:10.1007/s10827-013-0458-4)
28. Timme NM, Ito S, Myroshnychenko M, Nigam S, Shimono M, Yeh F-C, Hottowy P, Litke AM, Beggs JM. 2016 High-degree neurons feed cortical computations. *PLoS Comput. Biol.* **12**, e1004858. (doi:10.1371/journal.pcbi.1004858)
29. Faber SP, Timme NM, Beggs JM, Newman EL. 2018 Computation is concentrated in rich clubs of local cortical networks. *Netw. Neurosci.* **3**, 1–21. (doi:10.1101/290981)
30. Sherrill SP, Timme NM, Beggs JM, Newman EL. 2020 Correlated activity favors synergistic processing in local cortical networks in vitro at synaptically relevant timescales. *Netw. Neurosci. (Cambridge, Mass.)* **4**, 678–697. (doi:10.1162/netn_a_00141)
31. Tax TMS, Mediano PAM, Shanahan M. 2017 The partial information decomposition of generative neural network models. *Entropy* **19**, 474. (doi:10.3390/e19090474)
32. Griffith V, Koch C. 2014 Quantifying synergistic mutual information. (http://arxiv.org/abs/1205.4265)
33. Finn C, Lizier JT. 2018 Pointwise partial information decomposition using the specificity and ambiguity lattices. *Entropy* **20**, 297. (doi:10.3390/e20040297)
34. Bertschinger N, Rauh J, Olbrich E, Jost J. 2013 Shared information—new insights and problems in decomposing information in complex systems, pp. 251–269. (http://arxiv.org/abs/1210.5902)
35. Griffith V, Chong EKP, James RG, Ellison CJ, Crutchfield JP. 2014 Intersection information based on common randomness. *Entropy* **16**, 1985–2000. (doi:10.3390/e16041985)
36. Olbrich E, Bertschinger N, Rauh J. 2015 Information decomposition and synergy. *Entropy* **17**, 3501–3517. (doi:10.3390/e17053501)
37. Goodwell AE, Kumar P. 2017 Temporal information partitioning: characterizing synergy, uniqueness, and redundancy in interacting environmental variables. *Water Resour. Res.* **53**, 5920–5942. (doi:10.1002/2016WR020216)
38. Ince RAA. 2017 Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy* **19**, 318. (doi:10.3390/e19070318)
39. James RG, Emenheiser J, Crutchfield JP. 2018 Unique information via dependency constraints. *J. Phys. A: Math. Theor.* **52**, 014002. (doi:10.1088/1751-8121/aaed53)
40. Mediano PAM, Rosas F, Carhart-Harris RL, Seth AK, Barrett AB. 2019 Beyond integrated information: a taxonomy of information dynamics phenomena. (http://arxiv.org/abs/1909.02297)
41. Rubenstein PK, Weichwald S, Bongers S, Mooij JM, Janzing D, Grosse-Wentrup M, Schölkopf B. 2017 Causal consistency of structural equation models. (http://arxiv.org/abs/1707.00819)
42. Finn C, Lizier JT. 2018 Probability mass exclusions and the directed components of mutual information. *Entropy* **20**, 826. (doi:10.3390/e20110826)
43. Lizier JT. 2013 *The local information dynamics of distributed computation in complex systems*. Berlin, Heidelberg: Springer Berlin Heidelberg. (Springer theses).
44. Frey S, Albino DK, Williams PL. 2018 Synergistic information processing encrypts strategic reasoning in poker. *Cogn. Sci.* **42**, 1457–1476. (doi:10.1111/cogs.12632)
45. Griebenow R, Klein B, Hoel E. 2019 Finding the right scale of a network: efficient identification of causal emergence through spectral clustering. (http://arxiv.org/abs/1908.07565)
46. Hoel E, Klein B, Swain A, Griebenow R, Levin M. 2020 Evolution leads to emergence: an analysis of protein interactomes across the tree of life. *bioRxiv*. (doi:10.1101/2020.05.03.074419)
47. Hoel E, Levin M. 2020 Emergence of informative higher scales in biological systems: a computational toolkit for optimal prediction and control. *Commun. Integr. Biol.* **13**, 108–118. (doi:10.1080/19420889.2020.1802914)
48. Buxhoeveden DP, Casanova MF. 2002 The minicolumn hypothesis in neuroscience. *Brain* **125**, 935–951. (doi:10.1093/brain/awf110)

49. Yuste R. 2015 From the neuron doctrine to neural networks. *Nat. Rev. Neurosci.* **16**, 487–497. (doi:10.1038/nrn3962)
50. Cao R, Yamins D. 2021 Explanatory models in neuroscience: part 1–taking mechanistic abstraction seriously. (http://arxiv.org/abs/2104.01490)
51. Ashikaga H, Prieto-Castrillo F, Kawakatsu M, Dehghani N. 2018 Causal scale of rotors in a cardiac system. *Front. Phys.* **6**, 30. (doi:10.3389/fphy.2018.00030)
52. Anderson PW. 1972 More is different. *Science* **177**, 393–396. (doi:10.1126/science.177.4047.393)
53. Wolfram S. 2002 *A new kind of science*, vol. 5. Champaign, IL: Wolfram Media.
54. Tononi G. 2008 Consciousness as integrated information: a provisional manifesto. *Biol. Bull.* **215**, 216–242. (doi:10.2307/25470707)
55. Oizumi M, Albantakis L, Tononi G. 2014 From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Comput. Biol.* **10**, e1003588. (doi:10.1371/journal.pcbi.1003588)
56. Rosas FE, Mediano PAM, Jensen HJ, Seth AK, Barrett AB, Carhart-Harris RL, Bor D. 2020 Reconciling emergences: an information-theoretic approach to identify causal emergence in multivariate data. *PLoS Comput. Biol.* **16**, e1008289. (doi:10.1371/journal.pcbi.1008289)
57. Humphreys P. 2016 *Emergence: a philosophical account*. Oxford, UK: Oxford University Press.
58. Humphreys P. 2008 Synchronic and diachronic emergence. *Minds and Machines* **18**, 431–442. (doi:10.1007/s11023-008-9125-3)
59. Chalmers DJ. 2006 Strong and weak emergence. *The re-emergence of emergence* (eds P Clayton, P Davies), pp. 244–256. New York, NY: Oxford University Press.