From micro to macro units: a mathematical framework for identifying the causal grain of a system from its intrinsic perspective

William Marshall[1*], Graham Findlay[2,3], Larissa Albantakis[2], Giulio Tononi[2*]

1 Department of Mathematics and Statistics, Brock University, St. Catharines, Ontario, Canada
2 Department of Psychiatry, University of Wisconsin, Madison, Wisconsin, United States of America
3 Neuroscience Training Program, University of Wisconsin, Madison, Wisconsin, United States of America

* Corresponding authors: wmarshall@brocku.ca, gtononi@wisc.edu

Abstract

Integrated information theory (IIT) aims to account for the quality and quantity of consciousness in physical terms. It starts from the essential properties of experience, the theory's axioms, which it translates into postulates of cause-effect power—the ability of the system's units to "take and make a difference." Based on the theory's postulates, a substrate of consciousness must be a system of units that is a maximum of intrinsic, irreducible cause-effect power. Moreover, the grain of the substrate's units must be the one that ensures maximal intrinsic irreducibility. This work employs the mathematical framework of IIT 4.0 to assess cause-effect power at different unit grains according to the theory's postulates. Using simple, simulated systems, we show that the cause-effect power of a system of macro units can be higher than the cause-effect power of the corresponding micro units. Two examples highlight specific kinds of macro units, and how each kind can increase cause-effect power. The implications of the framework are discussed in the broader context of IIT, including how it provides a foundation for tests and inferences about consciousness.

1 Introduction

One goal of the scientific study of consciousness is to ascertain its neural substrate. Much attention has been given to the question of which regions of the brain support consciousness [9, 18]. No less important, but less often considered, is the question of the units constituting the substrate of consciousness and their "grain." Is it individual neurons, synapses, groups of neurons, or the smallest units that we can possibly manipulate and observe? Is it their state over a hundred milliseconds, one millisecond, or one second? These issues are not only empirical, but call for a theoretical understanding of why certain brain regions qualify as a substrate of consciousness, while others do not, and why the grain of the substrate's units is what it is.

Integrated information theory (IIT) aims to account for consciousness—its quality and quantity—by starting from phenomenology, identifying its essential properties, and then asking what physical substrate could support it [1]. Assuming the existence of consciousness as its 0^{th} "axiom", the theory characterizes a set of properties—the axioms of phenomenal existence—that are true of every conceivable experience: *intrinsicality*, *information*, *integration*, *exclusion*, and *composition*. These are translated into corresponding physical properties, called "postulates", that must be satisfied by the substrate of consciousness. Physical existence—IIT's 0^{th} postulate—is defined operationally in terms of cause-effect power—the ability to "take and make a difference." The postulates require that the substrate of consciousness has cause-effect power upon itself (intrinsicality), in a way that is specific (information), unitary (integration), definite (exclusion), and structured (composition). In principle, by evaluating whether and in what way a candidate substrate satisfies all of the postulates, one can evaluate whether and in what way it is conscious, with no additional ingredients.

Applying IIT's postulates to study the substrate of consciousness requires a mathematical framework for assessing cause-effect power. IIT's framework has been refined over time [22, 6, 19, 1], thanks to several developments [10, 7, 15] (for applications outside of consciousness science see [2, 16, 3]). The current framework—IIT 4.0 [1]—is the first complete account, with a tighter connection between the mathematical formalism and the postulates.

According to IIT, the substrate of consciousness is a set of units in a state, called a "complex", whose intrinsic cause-effect power is maximally irreducible, as measured by system integrated information (φ_s). In turn, the units' grain is the one at which the cause-effect power over the substrate is maximal. Initial work on determining the units' grain introduced the notion of a "macro unit"—a coarser unit derived from a set of finer units—and demonstrated that the cause-effect power of a system, as measured by effective information [23], could peak at a macro grain [12]. Further studies explored how and why macro grains could have higher cause-effect power than the micro grains on which they supervene [11, 14].

The goal of the current work is to provide a mathematical framework for measuring cause-effect power at macro grains that is based on IIT 4.0 [1] and that incorporates the theory's postulates explicitly. In Section 2, we review the mathematical framework for measuring the cause-effect power of systems of micro units, and then extend this framework to systems of macro units. In Section 3, the updated framework is applied to simple systems, demonstrating that macro-grain systems can have higher cause-effect power than the corresponding micro-grain systems. In Section 4, we provide a brief discussion of the importance of this framework for future work.

2 Theory

In this section, we first provide the details of IIT's mathematical framework that are necessary for expanding it to consider cause-effect power at macro grains. The IIT framework provides both the means to identify substrates of consciousness and to "unfold" a substrate's cause-effect power into a structure which corresponds to "what it is like" to be that substrate. Complete details on the definition and computation of integrated information and unfolding the cause-effect structure of a substrate are described elsewhere [15, 1]. Unfolding a substrate's cause-effect power, while a crucial aspect of IIT, is not necessary for expanding its framework to consider cause-effect power at macro grains. This is because the unfolding process is formally the same for all substrates, regardless of grain. For this reason, we will not discuss unfolding here, and refer interested readers to [1]. After briefly providing the relevant details of IIT's framework, we introduce a formal definition of macro units. Finally, we extend the mathematical framework to measure the cause-effect power of systems of macro units.

2.1 Background

According to IIT, something can be said to exist physically if it can "take and make a difference," i.e. bear a cause and produce an effect. Operationally, it must be possible to manipulate the system's units (change their state) and observe the result. To exist, then, a unit must have two available states ("this way" and "not this way").

Cause-effect structure at a particular grain must always be among units of that grain. Accordingly, it is assessed using partitions that "cut" causal connections among units. For this reason, from the intrinsic perspective of a complex, units must have exactly two states—"this way" and "not this way." One state is simply the complement of the other, whichever state is picked, with no further qualification. With more than two states, an internal structure distinguishing among "this way", "that way", and "the other way" is implied. This internal structure, rather than being among units, would be hidden within them. Therefore, it could not be assessed by observation and manipulation at the grain in question, only at a finer grain. In essence, with more than two states, causal structure from finer grains would be misattributed to coarser ones. Of course, from the extrinsic perspective of an experimenter unconcerned with strict isolation of grains, non-binary states are available for observation and manipulation, and can reveal important causal properties of a substrate.

The starting point of IIT's mathematical framework is thus a stochastic model for a physical universe $U = \{U_1, U_2, \ldots, U_n\}$ of n interacting binary units with state space $\Omega_U = \{0,1\}^n$. We define u as the set of units U in a particular state. More precisely, $u = \{(U_i, \text{state}(U_i)) : U_i \in U\}$ is a set of tuples, where each tuple contains a unit and the state of that unit. This formality allows us to define set operations over u that consider both the units and their states. We further denote Ω_U to be the set of all such tuple sets, corresponding to all the possible states of U. For IIT, physical existence is synonymous with having cause-effect power, the ability to take and make a difference. Consequently, a universe U with state space Ω_U is operationally defined by its potential interactions, assessed in terms of conditional probabilities. We denote the complete transition probability function of a universe U over a system update $\mathrm{u} \to \bar{\mathrm{u}}$ as

$$\mathcal{T}_U \equiv p(\bar{\mathrm{u}} \mid \mathrm{u}), \quad \mathrm{u}, \bar{\mathrm{u}} \in \Omega_U. \tag{1}$$

The individual random variables $U_i \in U$, given the preceding state of U, are conditionally independent from each other:

$$p(\bar{\mathrm{u}} \mid \mathrm{u}) = \prod_{i=1}^{n} p(\bar{u}_i \mid \mathrm{u}). \tag{2}$$

This amounts to a requirement that the model does not exhibit "instantaneous causation." In the discrete systems considered here, instantaneous interactions arise from incomplete knowledge, and therefore imply an incomplete causal model. For an extension of IIT to quantum systems, see [4].

Finally, $\mathcal{T}_U$ provides a complete description of the universe, which means that we can determine the conditional probabilities in (2) for every system state, with $p(\bar{\mathrm{u}} \mid \mathrm{u}) = p(\bar{\mathrm{u}} \mid \mathrm{do}(\mathrm{u}))$ [3, 13, 5, 20] (where the "do-operator" do(u) indicates that u is imposed by intervention). This implies that U corresponds to a causal network [3], and $\mathcal{T}_U$ is its transition probability matrix (TPM).

The TPM $\mathcal{T}_U$, which forms the starting point of IIT's analysis of cause-effect power, serves as an overall description of a universe's physical properties. It describes the probability that the universe will transition into each of its possible states upon being initialized into every possible state. Here, for simplicity, we assume that $\mathcal{T}_U$ is fixed and unchanging ("strict stationarity"). There is no additional role (or need) for intrinsic physical properties or laws of nature, but there is no issue if $\mathcal{T}_U$ evolves, and in fact it is expected[1].

For any candidate substrate (also called a "candidate system") $S \subseteq U$ in a state $s \subseteq u$, the IIT 4.0 framework defines its *system integrated information* $\varphi_s(s)$ [15, 1]. Based on the postulates of intrinsicality, information, and integration, $\varphi_s(s)$ quantifies how the system specifies a cause-effect state (its *intrinsic information* [8, 7]) as a whole, above and beyond how it specifies it as independent parts. Per the *principle of minimal existence*, which states that "nothing exists more than the least it exists", the comparison between the whole and its parts is performed by partitioning the system and evaluating the impact of the "minimum partition"—the partition over which the system is least irreducible ("a chain is only as strong as its weakest link") [1]. The system integrated information, φ_s, is defined as the intrinsic information of the whole, relative to the parts specified by its minimum partition. We do not present the full definition or algorithm for obtaining φ_s here, but only the parts that are relevant for extending the framework to macro units.

For any candidate system, $\varphi_s(s)$ is defined based on two system-specific transition probability functions/matrices, $\mathcal{T}_c$ and $\mathcal{T}_e$ (for describing causes and effects respectively). The system TPMs are computed by causally marginalizing the background units $W = U \setminus S$ conditional on the current state u, as described below. In this way, the TPMs capture intrinsic cause-effect power of the system within the context of a set of background conditions.

For evaluating effects, the state of the background units is fully determined by the current state of the universe ($w = u \setminus s$). The corresponding TPM, $\mathcal{T}_e$, is used to identify the effect of the current state:

$$\mathcal{T}_e \equiv p_e(\bar{\mathrm{s}} \mid \mathrm{s}) = p(\bar{\mathrm{s}} \mid \mathrm{s}, w), \quad \mathrm{s}, \bar{\mathrm{s}} \in \Omega_S, \tag{3}$$

For evaluating causes, knowledge of the current state is used to compute the probability distribution over potential prior states of the background units ($q(\bar{w})$), which is not necessarily uniform or deterministic. The corresponding TPM, $\mathcal{T}_c$, is used to evaluate the cause of the current state:

[1] An ontological principle of IIT not discussed here is the *principle of becoming*, which states that "powers become what powers do." An implication of this principle is that conditional probabilities in the TPM should update depending on what happens. There is an imperfect analogy between this principle and the Hebbian principle that "neurons which fire together wire together," though in the Hebbian case there are laws of nature that govern how probabilities in a neural network's TPM change depending on what happens, whereas no additional laws are needed to explain why $\mathcal{T}_U$ changes: conditional probabilities in $\mathcal{T}_U$ update because $\mathcal{T}_U$ itself is simply a record of which state transitions have transpired. No initial conditions or initial state need explaining, so long as there is some epsilon of fundamental indeterminism in $\mathcal{T}_U$ [21].

$$\mathcal{T}_c \equiv p_c(\mathrm{s} \mid \bar{\mathrm{s}}) = \prod_{i=1}^{|S|} \sum_{\bar{w}} p(s_i \mid \bar{s}, \bar{w}) q(\bar{w}) = \prod_{i=1}^{|S|} \sum_{\bar{w}} p(s_i \mid \bar{s}, \bar{w}) \left(\frac{\sum_{\hat{s}} p(u \mid \hat{s}, \bar{w})}{\sum_{\hat{u}} p(u \mid \hat{u})} \right), \quad \mathrm{s}, \bar{\mathrm{s}} \in \Omega_S. \tag{4}$$

Finally, according to the exclusion postulate, a substrate of consciousness must be definite: there must be a reason why it consists of these units, and not others. By the *principle of maximal existence*, which states that "what exists is what exists the most," we find the substrate that lays the greatest claim to existence as one entity, as measured by φ_s [1]. That is, s is a substrate of consciousness (also called a *"complex"*) if for any other $s' \subseteq u$,

$$s \cap s' \neq 0 \Rightarrow \varphi_s(s) > \varphi_s(s').$$

The above condition compares a candidate system $s \subseteq u$ to all other potential candidate systems $s' \subseteq u$, and ensures that its system integrated information is greater than any overlapping candidate system. Implicit in this definition is that all candidate systems are subsets of u, and thus share the same grain as u. However, according to the exclusion postulate, the units that constitute a complex should be definite, in the sense of having a definite grain. Practically, a complex should not only have greater φ_s than candidate systems at the same grain, but across all possible grains. Evaluating φ_s for candidate systems at all possible grains requires extending IIT's mathematical framework as follows.

2.2 Macro Units

A *micro unit* is conceived of as an "atom" of cause-effect power: it cannot be partitioned into finer constituents, its updates cannot be partitioned into finer updates, and it cannot have more than two states—the minimum necessary to bear a cause and produce an effect.

For the purpose of defining units at different grains, we assume that $U = \{U_1, U_2, \ldots, U_n\}$ is a set of micro units with update grain $\tau'_U = 1$. A macro unit J has three aspects:

$$J = (U^J, \tau'_J, g'_J\},$$

where $U^J \subseteq U$ are its micro constituents, $\tau'_J \in \mathbb{Z}^+$ is its update grain in terms of micro updates, and g'_J a mapping from the states of U^J over a sequence of τ'_J micro updates to the state of J:

$$g'_J : \Omega_{U^J}^{\tau'_J} \rightarrow \{0, 1\}.$$

From the intrinsic perspective of a complex, macro units (like micro units, and for the same reasons) have a repertoire of exactly two states—the minimum necessary to bear a cause and produce an effect, and to be an atom of macro cause-effect power.

It is often helpful to think of macroing as being "over units" (when a macro unit has more than one constituent unit), "over updates" (when a macro unit has $\tau'_J > 1$, even though it may not have more than one constituent unit), or both. Previous work referred to macroing as being "over space" and/or "over time" [12, 14], but we avoid these terms here, because of their metaphysical implications. The IIT framework does not require spacetime to be fundamental.

Constructing macro units directly from micro units is a special case of a more general framework. When integrated information is maximized at a macro grain, a macro unit at that ultimate grain may be built from constituents V^J that are themselves macro units at a finer grain—called *meso units*—and the same may be true for the meso units' constituents, and so forth. That is, a macro unit may be built from one or more levels of meso units sandwiched between it and its constituent micro units U^J, to each of which the postulates apply (Figure 1). Formally, there is no difference between macro units and meso units, but for clarity we will hereafter reserve the term "macro" for the grain of the complex (when it is not a micro grain), and "meso" for any intermediate grains.

To facilitate the distinction between a unit's micro constituents U^J and its direct constituents V^J—which may be meso units—we extend our definitions of J and g_J above to be completely general, covering cases where J is a micro, meso, or macro unit. As useful notation, we define a set function f that maps a set of units into the set of

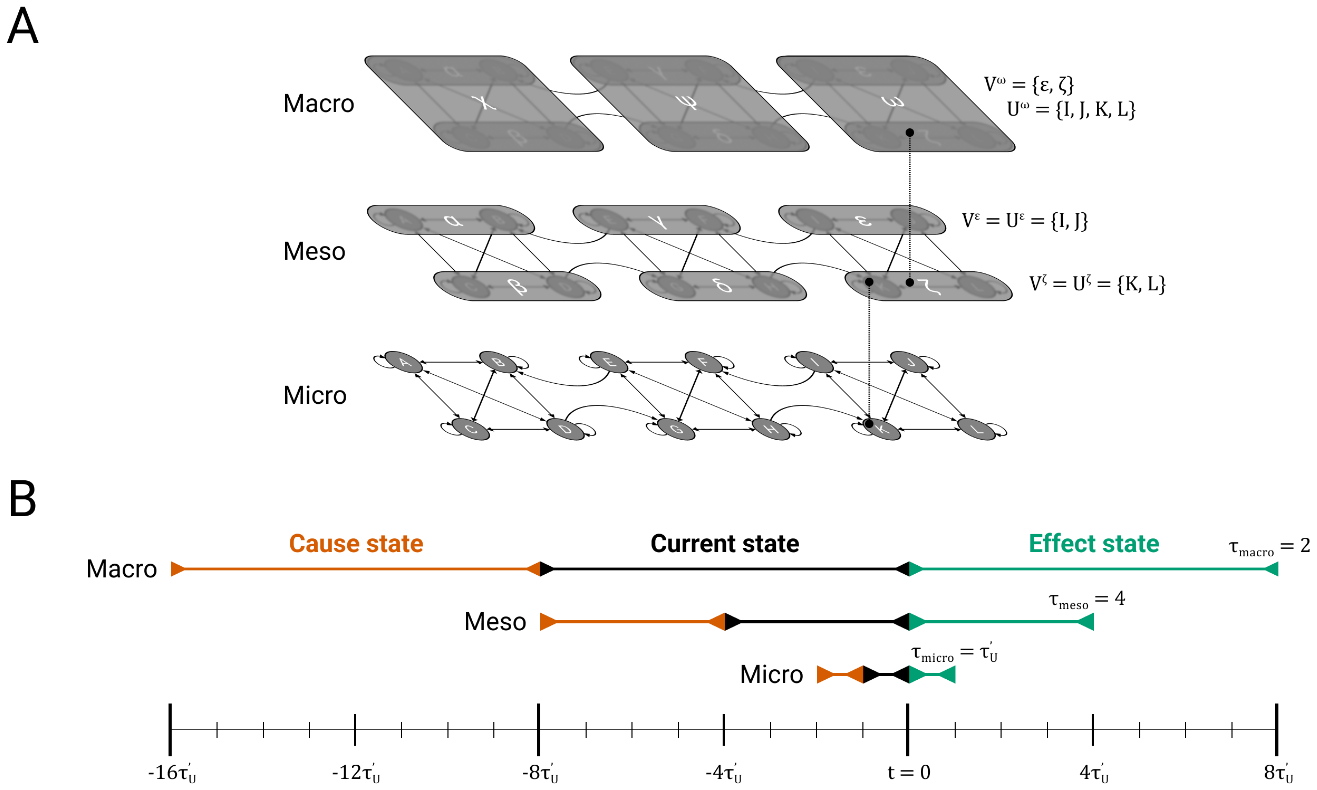


Figure 1: From micro to macro units. (**A**) A universe $U = \{A, B, C, D, E, F, G, H, I, J, K, L\}$, with unspecified transition probability function $\mathcal{T}_U$. Although some intrinsic cause-effect power may be associated with units at this micro grain, it is also possible that intrinsic cause-effect power is highest at a macro grain. For example, it might be maximal for the macro system $\{\chi, \psi, \omega\}$, which would mean that this system exists from its own perspective as a system of three macro units. The framework provided in this paper will allow us to assess if this is the case, including whether a macro unit, say ω, can be built using micro constituents $U^\omega = \{I, J, K, L\}$, possibly with intermediate meso constituents $V^\omega = \{\eta, \zeta\}$. (**B**) In addition to defining macro states over groups of units, it is also possible to define macro states over updates of U. We depict one hypothetical scenario in which macro units have an update grain equal to 2 meso updates ($\tau_{\text{macro}} = 2$), meso updates have an update grain equal to 4 micro updates ($\tau_{\text{meso}} = 4$), and the micro update ($\tau_{\text{micro}} = \tau_U = 1$) is inherited from U. The macro state of a complex is always defined looking back from the present micro instant. Thus, this macro state, while a function of several updates, can change every micro update, in a "sliding window" fashion.

all admissible units that can be defined from the original set, in its current state. Thus, $f(u)$ is the set of all units (micro, meso, or macro) that could be defined from u according to the requirements specified below. A unit $J \in f(u)$ has four aspects:

$$J = (U^J, V^J, \tau_J, g_J\},$$

where $U^J \subseteq U$ are its micro constituents, $V^J \subseteq f(u^J)$ are its constituents (which may be micro or meso units) with current state v^J and state space Ω_{V^J}, $\tau_J \in \mathbb{Z}^+$ is the update grain over which J's constituents are evaluated to define the state of J, and g_J a mapping from the states of V^J over a sequence of τ_J updates of V^J to the state of J:

$$g_J : \Omega_{V^J}^{\tau_J} \to \{0, 1\}.$$

In general, there are $2^{2^{\tau_J|V^J|}-1} - 1$ possible mappings from the state of constituents to the state of J. It is important to note that when J is constructed from a hierarchy of meso units of increasing grain, the update grain τ_J and the function g_J define a mapping across a *single* level of this hierarchy, from a sequence of states of V^J to the state of J. If V^J is a set of meso units, then τ_J is the number of *meso* updates that define J's state. There exist additional mappings between V^J's constituents and V^J, and so on, down to the micro constituents U^J. Thus, in addition to the update grain of J in terms of its direct constituents (τ_J), this hierarchical sequence of mappings can be used to define an update grain of J in terms of its *micro* constituents, which we label τ'_J. Similarly, we have a mapping g'_J from sequences of micro states to the state of J:

$$g'_J : \Omega_{U^J}^{\tau'_J} \to \{0, 1\}.$$

For example, in Figure 1B, $\tau_J = \tau_{\text{macro}} = 2$ and $\tau'_J = 8$, because J's state is defined over a sequence of 2 meso updates, each of which consists of 4 micro updates. Notice that unlike τ, τ' is non-decreasing as a function of the level in the hierarchy.

Not every conceivable set of macro units $S \subseteq f(u)$ defines a valid macro system. More generally, although $f(u)$ yields the set of all admissible units that can be defined from U, not every subset of $f(u)$ is admissible as a candidate system. Valid candidate systems must satisfy IIT's requirement for physical existence, characterized operationally: it must be possible to manipulate the units (change their state) and observe the result, in order to assess their cause-effect power. Practically, this implies that the constituents of a candidate system must share a common update grain. Otherwise, the manipulation of one unit will interfere with the observation of another. Additionally, any macro units' constituents must not overlap in their micro constituents. Otherwise, one macro unit could not be manipulated or observed independently of another. Formally, the set of all admissible candidate systems (at any grain) defined from a set of micro constituents U has the form

$$\mathbb{P}(f(u)) = \{S \subseteq f(u) : U^{J_i} \cap U^{J_k} = \varnothing,\ \tau'_{J_i} = \tau'_{J_k}\ \forall\ J_i, J_k \in S\}. \tag{5}$$

Notice that these requirements are satisfied implicitly for sets of micro units.

We now introduce the main update to IIT's mathematical framework for measuring cause-effect power at macro grains, bringing it in line with IIT 4.0 and further connecting it with the postulates. Specifically, we require that a macro unit J_i in a candidate system

$$S = \{J_1, J_2, \ldots, J_{|S|}\}, \quad S \in \mathbb{P}(f(u))$$

with system micro constituents

$$U^S = \bigcup_{i=1}^{|S|} U^{J_i}$$

satisfies IIT's postulates, such that J_i needs to be a maximally irreducible constituent of a complex ("maximally irreducible within"), rather than a complex itself ("maximally irreducible within and without").

Specifically, a first consequence of the requirement that macro units satisfy the postulates is that both a macro unit and its constituents must satisfy IIT's requirement for physical existence, characterized operationally: it must be possible to manipulate the units (change their state) and observe the result. Practically, as is the case with candidate systems, and for the same reasons, this implies that the constituents of a macro unit must share a common update grain and must not overlap in their micro constituents. Put simply, we require that $V^J \in \mathbb{P}(f(u))$ (see Eqn. 5).

A second consequence of the requirement that macro units satisfy the postulates is that they must be irreducible. For a macro unit J_i with constituents $V^{J_i} \in \mathbb{P}(f(u))$ (in current state v^{J_i}) to satisfy the postulates of intrinsicality, information and integration, it must have cause-effect power that is intrinsic, specific, and irreducible. Put simply, we require that $\varphi_s(v^{J_i}) > 0$.

A third consequence of the requirement that macro units satisfy the postulates is that they must be definite. Recall that a complex must have higher integrated information than all other systems sharing even just one of its micro constituents. That is, if S were a complex:

$$\varphi_s(s) > \varphi_s(s'),\ \forall S' \in \mathbb{P}(f(u)) \text{ with } U^S \cap U^{S'} \neq \varnothing. \tag{6}$$

In contrast, a macro unit—when considered as a system of its constituents V^{J_i}—need only have higher integrated information than any other system that could be constructed from its micro constituents U^{J_i}:

$$\varphi_s(v^{J_i}) > \varphi_s(s'),\ \forall S' \neq V^{J_i} \in \mathbb{P}(f(u^{J_i})). \tag{7}$$

In other words, a macro unit must be a maximally irreducible constituent of a complex ("maximally irreducible within"), rather than a complex itself ("maximally irreducible within and without"). We thus require that the units from which a system with intrinsic, irreducible cause-effect power is built themselves have intrinsic, irreducible cause-effect power. To relax this requirement would allow reducible (or barely integrated) groups of units to be "hidden" inside macro units, creating the illusion of integration where there is none (Figure 2), effectively building something out of nothing (in terms of cause-effect power). An important consequence of the requirement that macro

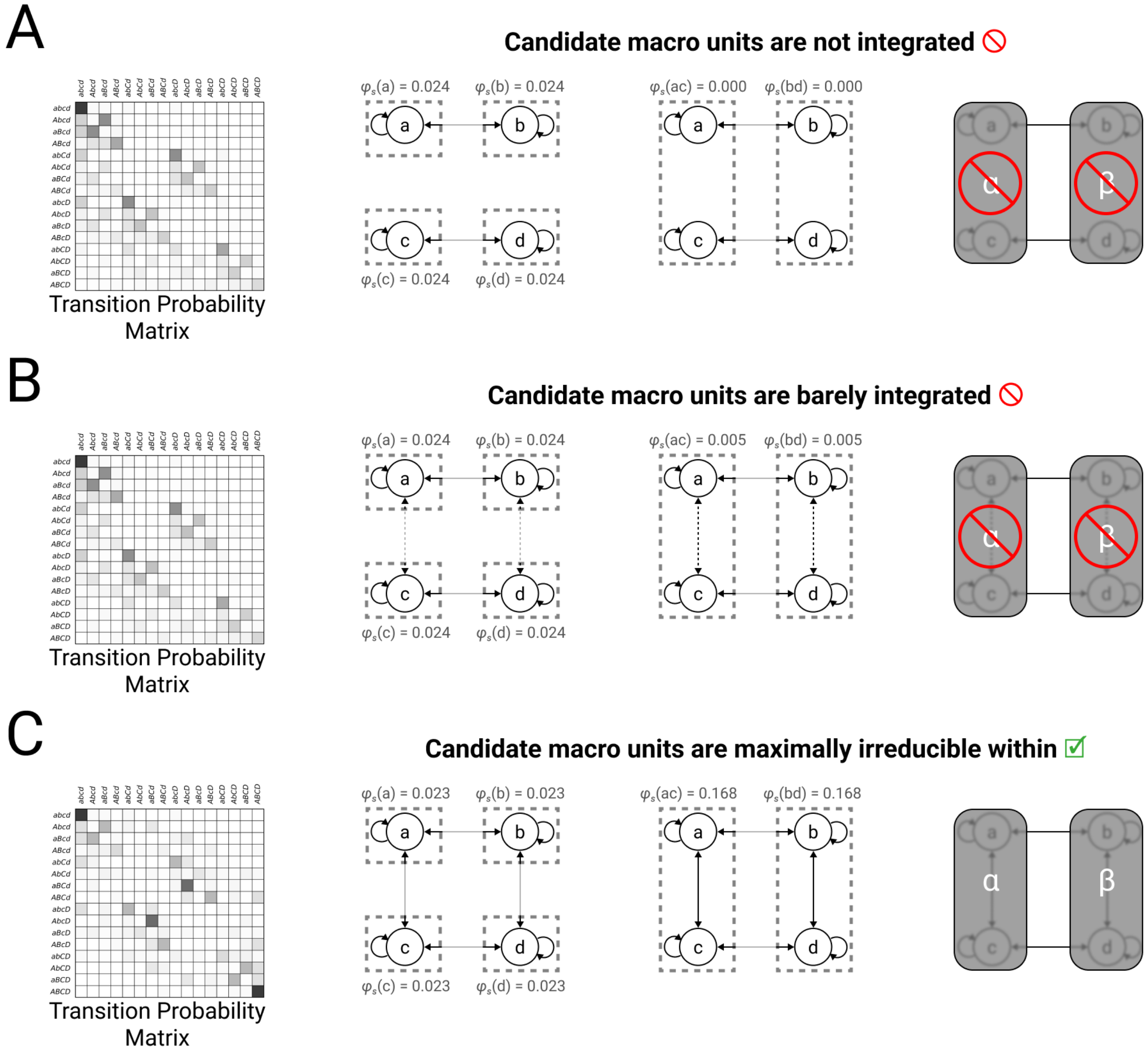

A
Candidate macro units are not integrated
abcd Abcd aBcd ABcd abCd AbCd aBCd ABCd abcD AbcD aBcD ABcD abCD AbCD aBCD ABCD
Transition Probability Matrix
φs(a) = 0.024
φs(b) = 0.024
φs(c) = 0.024
φs(d) = 0.024
φs(ac) = 0.000
φs(bd) = 0.000
a
b
c
d
α
β
B
Candidate macro units are barely integrated
abcd Abcd aBcd ABcd abCd AbCd aBCd ABCd abcD AbcD aBcD ABcD abCD AbCD aBCD ABCD
Transition Probability Matrix
φs(a) = 0.024
φs(b) = 0.024
φs(c) = 0.024
φs(d) = 0.024
φs(ac) = 0.005
φs(bd) = 0.005
a
b
c
d
α
β
C
Candidate macro units are maximally irreducible within
abcd Abcd aBcd ABcd abCd AbCd aBCd ABCd abcD AbcD aBcD ABcD abCD AbCD aBCD ABCD
Transition Probability Matrix
φs(a) = 0.023
φs(b) = 0.023
φs(c) = 0.023
φs(d) = 0.023
φs(ac) = 0.168
φs(bd) = 0.168
a
b
c
d
α
β

Figure 2: Out of nothing, nothing comes. Consider four micro units $\{A, B, C, D\}$ in state $(0, 0, 0, 0)$, constituting universe U with transition probability matrix $\mathcal{T}_U$. For each micro unit U_i, when all its inputs are 0, the probability that its state will be 1 after the next update is 0.05. This probability is increased by 0.05 if u_i itself is currently 1, and further increased by 0.6 if the state of U_i's horizontal neighbor is 1. (**A**) Consider the case where there are no connections between vertical neighbors (left). Each micro unit has $\varphi_s = 0.024$ on its own (middle left), while each pair of vertical neighbors has $\varphi_s = 0$ (middle right). Because each pair of vertical neighbors is reducible, they are not valid macro elements (right). (**B**) Consider the case where vanishingly weak connections are introduced between vertical neighbors, such that the probability that u_i will be 1 after the next update is increased by 0.01 if U_i's vertical neighbor is 1 (left). Although each pair of vertical neighbors is now very weakly integrated with $\varphi_s = 0.005$ (middle right), they are not maximally irreducible within (e.g. $\varphi_s(a, b) < \varphi_s(a)$). The conclusion is the same as for (A): the vertical neighbors are not valid macro elements (right). (**C**) Finally, consider the case where strong connections are introduced between vertical neighbors, such that the probability that u_i will be 1 after the next update is increased by 0.25 if U_i's vertical neighbor is 1 (left). Integration between vertical neighbors is now sufficiently strong (middle right) that the "maximally irreducible within" criterion is satisfied (middle right vs middle left), so we can consider macro elements built from vertical neighbors (right). There is no guarantee that the macro system consisting of these elements $\{\alpha, \beta\}$ is a complex, but at least we may evaluate that possibility (not shown).

units satisfy the postulates is that if a complex of macro units was somehow deconstructed, we would find that the constituent macro units continue to exist in isolation — "out of nothing, nothing comes."

The requirement that macro units be maximally irreducible within applies whether macroing over units (Figure 3A-C), and/or over updates (Figure 3D), and applies to units at any grain. Altogether, we can now formally define $f(u)$ as

$$f(u) = \{(U^J, V^J, \tau_J, g_J) : V^J \in \mathbb{P}(f(u)),\ \varphi(v^J) > \varphi_s(s') \;\forall\; s' \in \mathbb{P}(f(u^J))\} \tag{8}$$

For a given set of micro constituents U^{J_i}, whether or not a macro unit J_i is built upon meso units (Figure 3E, bottom) rather than being built directly "in one shot" upon the micro units (Figure 3E, top) depends on which definition of V^{J_i} maximizes $\varphi_s(V^{J_i})$, making J_i maximally irreducible within. In general, having finer (e.g. micro) constituents means having a larger number of mappings available to J_i with which to maximize φ_s at the macro grain, but makes it harder for V^{J_i} to satisfy the requirement of being "maximally irreducible within." For finer grain systems with a large number of constituents, having high φ_s requires that these units are both highly selective (to support intrinsic information), and have a connectivity structure without fault lines (to support integration) [15]. By contrast, coaser systems of units (defined from the same micro constituents) have fewwe units, each of which can be flexibility defined through intermediate mappings to have the high selectivity and connectivity structure required to support high integrated information. Thus, although there is no strict requirement that a system of macro units be built up from meso units, there are good reasons to expect that many systems will have this property.

It is important to notice that whether a macro system is built up in levels, which precise macro and meso units it is built from, and which mappings define those units' state, are all ultimately determined by what maximizes φ_s at each level of the hierarchy. Thus, there is always a reason why a system and its constituents are precisely what they are: the principle of maximal existence.

Note that the construction of $f(u)$ (Eqn. 8) is non-trivial, due to its interdependence with $\mathbb{P}(f(u))$ (Eqn. 5): the set of candidate systems depends on the set of admissible macro units, and the set of admissible macro units depends on the set of candidate systems (for satisfying "maximally irreducible within"). Practically, the two sets need to be derived recursively. The starting point is that each micro unit U_i is a potential unit in $f(u)$, with $U^J = U_i$, $V^J = U_i$, $\tau'_J = 1$ and g_J the identity mapping. The set of micro units then defines a set of candidate systems to be included in $\mathbb{P}(f(u))$. Those candidate systems are then used as potential meso constituents for defining new potential units, which then leads to new candidate systems. The process can be repeated iteratively, until convergence, which is guaranteed: the requirement of non-overlapping macro units ensures finite possible candidate systems.

2.3 Macro TPMs

Having defined macro units, we next outline a general framework for assessing φ_s that applies regardless of whether a system's units are micro units or macro units. In essence, we define generalized TPMs $\mathcal{T}_c^S$ and $\mathcal{T}_e^S$ that take any macroing into account. These TPMs can then be used to compute $\varphi_s(s)$ as described in [1].

First, when dealing with macro update grains, we require some additional notation to define the current, cause, and effect states. We denote u_t to be a state of U,

$$u_t = (u_{1,t}, u_{2,t}, \ldots, u_{n,t}).$$

The current micro state of U is indicated by $t = 0$ ($u = u_0$), negative indices ($t < 0$) indicate the updates that led to the current micro state, and positive indices ($t > 0$) indicate the updates that follow the current micro state. Further, we define a sequence of micro updates, from $t = a$ to $t = b$ ($a < b$) as

$$u_{[a,b]} = (u_a, u_{a+1}, \ldots, u_{b-1}, u_b).$$

Similar notation applies for any set of units.

For a macro unit J with micro constituents U^J and macro update grain τ'_J, its current state is defined by looking at the current state of its micro constituents u^J, and the sequence of states that led to the current state. As depicted in Figure 1B, the current state of J is defined as

$$j = g'_J(u^J),$$

where

$$u^J = u^J_{[-\tau'_J+1,0]} = (u^J_{-\tau'_J+1}, u^J_{-\tau'_J+2}, \ldots, u^J_0).$$

As illustrated in Fig 1B, the effect state of a macro unit with update grain τ'_J is defined based on the τ'_J states that follow its current macro state,

$$\bar{u}^J_e = u^J_{[1,\tau'_J]} = (u^J_1, u^J_2, \ldots, u^J_{\tau'_J}),$$

and the cause state is defined by the τ'_J states that precede its current macro state,

$$\tilde{u}^J_c = u^J_{[-2\tau'_J+1,-\tau'_J]} = (u^J_{-2\tau'_J+1}, u^J_{-2\tau'_J+2}, \ldots, u^J_{-\tau'_J}).$$

Consider a system $S = \{J_1, \ldots, J_{|S|}\} \in \mathbb{P}(f(u))$, where each unit J_i (whether micro or macro) has micro constituents $U^{J_i} \subset U$, constituents $V^{J_i} \subset f(u^{J_i})$, a common update grain $\tau'_{J_i} = \tau' \in \mathbb{Z}^+$, and a mapping $g'_{J_i} : \Omega^{\tau'}_{U^{J_i}} \to \{0, 1\}$. The set of micro constituents for the system is

$$U^S = \bigcup_{i=1}^{|S|} U^{J_i},$$

and the background units (always treated at the micro grain) are

$$U^W = U \setminus U^S.$$

The current state of S is

$$s = (g'_{J_1}(u^{J_1}), \ldots, g'_{J_m}(u^{J_m})).$$

To extend the framework, the goal is to define generalized cause and effect TPMs $\mathcal{T}_c$ and $\mathcal{T}_e$ for S, from which the rest of the framework can be applied as usual. The process proceeds in three steps: (1) determine the probability distribution over the possible past and future states of the background units, which is needed for causal marginalization; (2) extrapolate the state-by-state micro TPM $\mathcal{T}_U$ into a pair of sequence-by-sequence TPMs $\mathcal{T}_c^{\circ\tau'}$ and $\mathcal{T}_e^{\circ\tau'}$ that describe the probability of the next sequence of τ' updates of U^S, given the previous sequence, while causally marginalizing the background (U^W) conditional on the current sequence of micro states (u^S); (3) use the mappings g'_{J_i} to compress these two sequence-by-sequence TPMs into a pair of generalized state-by-state TPMs $\mathcal{T}_c^S$ and $\mathcal{T}_e^S$. When no macroing is performed (i.e. $\tau' = 1$ and all g'_{J_i} are identity functions), (2) and (3) are trivial, and $\mathcal{T}_c^S$ and $\mathcal{T}_e^S$ work out to be (micro) TPMs as defined in [1].

2.3.1 Obtaining the conditional distributions of past and future background states

The first step is to determine the conditional distribution of background units given the current sequence of substrate states u^S. On the effect side, $\mathcal{T}_U$ is used to determine a probability distribution for the next $\tau' - 1$ updates of U, conditional on the current micro state sequence u_0 (because of the Markov property, only the final micro update of u^S is relevant). The micro constituents U^S are then marginalized to get a distribution for U^W. For $t = 0$ and $t = 1$, the marginal distributions of u^W_T are

$$q_0(w) = \begin{cases} 1 \text{ if } w = u^W_0 \\ 0 \text{ otherwise,} \end{cases}$$

and

$$q_1(w) = \sum_{s \in \Omega_{U_S}} \mathrm{P}(w, s \mid u_0),$$

respectively. Then for any $t > 1$, the marginal distribution of u^W_t is

$$q_t(w) = \sum_{s \in \Omega_{U_S}} \sum_{(u_1, \ldots, u_{t-1})} p(w, s \mid u_{t-1}) \prod_{T=1}^{t-1} p(u_T \mid u_{T-1}),$$

where the inner summation is over all possible sequences of substrate states between u_1 and u_{t-1}. On the cause side, $\mathcal{T}_U$ is used in combination with Bayes' rule to determine a probability distribution for the previous update of U ($u_{-\tau'}$), conditional on the current macro state. Again, because of the Markov property, only the distribution for $u^W_{-\tau'}$ is required to define the TPM. A uniform marginal distribution of the previous updates is assumed (i.e., maximum uncertainty about prior states). For $t = -\tau'$, the marginal distribution of $u^W_{-\tau'}$ is

$$q_{-\tau'}(w) = \frac{\sum_s p(u_{-\tau'+1} \mid w, s)}{\sum_u p(u_{-\tau'+1} \mid u)}$$

2.3.2 Obtaining micro sequence TPMs from micro state TPMs

The second step is to use $\mathcal{T}_U$ to create micro-level TPMs $\mathcal{T}_c^{\circ\tau'}$ and $\mathcal{T}_e^{\circ\tau'}$ that describe the probability of the next sequence of τ' updates of U^S, given the previous sequence, while causally marginalizing U^W conditional on $u_{-\tau'+1,0]}$. As part of the causal marginalization, the probability that each micro unit goes through each possible state sequence is computed independently of the other units, and then these probabilities are combined as a product to determine the joint distribution over sequences. This renders the macro TPM conditionally independent (i.e. no "instantaneous causation" between units; see also [14]). This also eliminates potential influences between the micro units of distinct macro units that are not captured by the macro states. While such interactions may be relevant to determine the dynamical evolution of a system from the extrinsic perspective, they should not be taken into account when evaluating the intrinsic cause-effect power of the macro system. For evaluating effects

$$\mathcal{T}_e^{\circ\tau'} \equiv p_e^{\tau'}(\bar{u}_e^S \mid u^S) = \prod_{i=1}^{|U^S|} \prod_{t=0}^{\tau'-1} \sum_w p(u^S_{i,t+1} \mid w, u^S_t) q_t(w),$$

where $\bar{u}_e^S$ is a sequence of τ' effect states. Then for evaluating causes

$$\mathcal{T}_c^{\circ\tau'} \equiv p_c^{\tau'}(u^S \mid \tilde{u}_c^S) = \prod_{i=1}^{|U_S|} \left(\sum_w p(u^S_{i,-\tau'+1} \mid w, u^S_{-\tau'}) q_{-\tau'+1}(w) \right) \prod_{t=-\tau'+1}^{-1} p(u^S_{i,t+1} \mid u^W_t, u^S_t),$$

where $\tilde{u}_c^S$ is the sequence of τ' cause states.

2.3.3 Using a mapping to compress the micro sequence TPMs into generalized TPMs

The final step is to use the mappings g'_{J_i} to compress the pair of τ'-step micro TPMs (one for evaluating causes, one for evaluating effects) into a pair of corresponding generalized TPMs for S. Effectively, we will combine rows of a sequence TPM that lead to the same effect state. For each unit $J_i \in S$, let $D_{J_i}(j)$ be the domain of the state $J_i = j$ ($j \in \{0, 1\}$), e.g., the set of sequences of micro states u^S that map to $J_i = j$

$$D_{J_i}(j) = \{u^S \in \Omega^{\tau'}_{U^S} : g'_{J_i}(u^S) = j\}.$$

Then for any current state $S = s$, and any effect state $\bar{s} = (\bar{j}_1, \bar{j}_2, \ldots, \bar{j}_{|S|})$, we can define the effect transition probabilities,

$$\mathrm{P}_e(J_i = \bar{j}_i \mid S = s) = \sum_{u^S \in D_S(s)} \sum_{\bar{u}_e^S \in D_{J_i}(j)} \frac{1}{|D_S(s)|} p_e^{\tau'}(\bar{u}_e^S \mid u^S),$$

which can then be combined to define the effect TPM for S

$$\mathcal{T}_e^S \equiv p_e^S(\bar{s} \mid s) = \prod_{i=1}^{|S|} \mathrm{P}_e(J_i = \bar{j}_i \mid S = s).$$

Similarly, for any current state $s = (j_1, j_2, \ldots, j_{|S|})$ and cause state $\tilde{s}$, we can define transition probabilities,

$$\mathrm{P}_c(J_i = j \mid S = \bar{s}) = \sum_{\tilde{u}_c^S \in D_S(s)} \sum_{u^S \in D_{J_i}(j)} \frac{1}{|D_S(s)|} p_c^{\tau'}(u^S \mid \tilde{u}_c^S),$$

and then a cause TPM for S

$$\mathcal{T}_c^S \equiv p_c^S(s \mid \tilde{s}) = \prod_{i=1}^{|S|} \mathrm{P}_c(J_i = j_i \mid S = \tilde{s}).$$

Having defined $\mathcal{T}_c^S$ and $\mathcal{T}_e^S$, we can use the two TPMs to compute $\varphi_s(s)$ as described in [1].

3 Results

In this section, the framework is applied to two example systems. The examples demonstrate that intrinsic cause-effect power can be higher for a system of macro units than for any system of the corresponding micro units, extending results from earlier work [11, 14] to the updated framework [1]. Computations of integrated information were performed using PyPhi [17]. In what follows, we omit the state as input to φ_s (e.g., $\varphi_s(\{A, B\})$) when the state of the units can be inferred from the context of the example.

3.1 Example 1

Consider four micro units $\{A, B, C, D\}$ in state $(0, 0, 0, 0)$, constituting universe U with transition probability matrix $\mathcal{T}_U$ (Figure 4A). Each micro unit U_i has the same function: When all its inputs are 0, the probability that its state will be 1 after the next update is 0.05. This probability is increased by 0.01 if u_i itself is currently 1. Thus, there is a very weak tendency for a unit that is 1 to remain 1. The probability that u_i will be 1 after the next update is increased by 0.1 if the state of U_i's horizontal neighbor is 1. For example, A is more likely to be 1 after the next update if B is currently 1. Finally, the probability that u_i will be 1 after the next update is increased by 0.8 if *both* its vertical neighbor and its diagonal neighbor are currently 1. For example, A is very likely to be 1 after the next update if both C and D are currently 1. Thus, each micro unit approximates a noisy logical AND function over its vertical and diagonal neighbors, with a weak independent influence from its horizontal neighbor, and very weak self-influence. Because each unit's future state is mostly dictated by its vertical and diagonal neighbors (e.g. A's future state depends most heavily on the current states of C and D), and because horizontal neighbors share the same vertical and diagonal neighbors, (e.g. both A and B are dominated by C and D), we expect that macroing horizontal neighbors into macro units will reduce the indeterminism (e.g. neither A nor B are perfect AND functions of C and D) and degeneracy (both A and B are likely to be in the same state) associated with the micro units, and thereby increase cause-effect power [12, 11, 14].

To confirm this intuition, we first assess the system integrated information φ_s of all possible candidate systems of micro units (Figure 4B). At this micro level, the two candidate systems with the most irreducible cause-effect power are $\{A, B\}$ and $\{C, D\}$, both with $\varphi_s = 0.044$. Because these candidate systems are maximally irreducible within (e.g. $\varphi_s(\{A, B\}) > \varphi_s(s) \quad \forall S \subseteq \{A, B\}$), they satisfy Eqn. (7) and can be considered as macro units. Notice that although $\{A, B, C, D\}$ as a whole has irreducible cause-effect power ($\varphi_s = 0.020$), it is not maximally irreducible within (e.g. $\varphi_s(\{A, B\}) > \varphi_s(\{A, B, C, D\})$) and *cannot* be considered as a macro unit.

Let macro unit α be defined from micro constituents $\{A, B\}$, and β from $\{C, D\}$. There are 14 possible mappings for each macro unit (Figure 3C). In particular, the mapping shown in Figure 4C seems promising, because it ought to decrease both the indeterminism and the degeneracy that are present in the micro system. This class of mapping, in which the state of the macro unit is a simple function of the number constituents in state 1, has also been referred to as "coarse-graining" [12, 11]. Coarse-graining corresponds to the typical notion of a macro state in statistical physics [14]. Under the mapping shown in Figure 4C, each macro unit's state is 1 if-and-only-if both its micro constituents are 1. When one macro unit is 1, odds are that the other macro unit will be 1 after the next update. When one macro unit is 0, odds are that the other macro unit will be 0 after the next update. Thus, the macro system behaves something like two reciprocally connected COPY gates, with some additional complexity provided by the connections between horizontal neighbors at the micro level. This is reflected in the macro system's TPM (Figure 4C, middle). Indeed, when we measure the system integrated information of $\{\alpha, \beta\}$, we find $\varphi_s(\{\alpha, \beta\}) = 1.004$, demonstrating that this system of macro units has more irreducible, intrinsic cause-effect power than any candidate system built without macro units (Figure 4C, right).

3.2 Example 2

Consider eight micro units $\{A, B, C, D, E, F, G, H\}$ in state $(1, 1, 1, 1, 1, 1, 1, 1)$, constituting universe U with transition probability matrix $\mathcal{T}_U$ (Figure 5A). The left half of the system and the right half of the system ($\{A, B, C, D\}$ and $\{E, F, G, H\}$, respectively) are mirror images of each other, so for simplicity consider the left half. For every unit U_i, the probability that its state will be 1 after the next update is marginally higher if its current state is 1. C approximates a noisy logical OR function of A and B, which in turn approximate a noisy logical COPY function of C's image G. When A, B, or C are 1, the probability that D is 1 after the next update increases linearly. D's current state also has weak influence on the future state of A and B. Roughly speaking, we can think of the two halves of the system as copying each other's state, but whereas a disruption to any of the connections *within* either half will moderately disrupt this function, a disruption to any of the connections *between* halves will severely disrupt it. It is reasonable to expect that macroing the left half of the system and the right half of the system into separate macro units, and treating the macro units' states as a simple function of C and G's states, will increase intrinsic cause-effect power. This class of mapping, in which the state of the macro unit is determined only by the state of specific constituents, ignoring others, has also been referred to as "black-boxing." Black boxes correspond to the typical notion of macro units in the special sciences, because they are constituted of heterogeneous micro units that are often compartmentalized and have highly specific functions, which would be muddled by averaging [14].

To compare the cause-effect power of the micro and macro systems, we first assess the system integrated information φ_s of all possible candidate systems of micro units (Figure 5B). Note that, in addition to the candidate systems shown in Figure 5B, all candidate systems of five units (e.g. $\{A, B, C, D, E\}$), six units (e.g. $\{A, B, C, D, E, F\}$), and seven units (e.g. $\{A, B, C, D, E, F, G\}$) were evaluated (not shown). At the micro grain, the maximum value of φ_s is 0.135 ($S = \{A, C, E, G\}$ and symmetric systems). At the micro grain, the two candidate systems that we hypothesized would make good macro units ($\{A, B, C, D\}$ and $\{E, F, G, H\}$) are maximally irreducible within, with $\varphi_s = 0.030$. Because these candidate systems are maximally irreducible within (e.g. $\varphi_s(\{A, B, C, D\}) > \varphi_s(s) \quad \forall s \subseteq \{A, B, C, D\}$), they satisfy condition (7) and can be considered as macro units.

Let macro unit α be defined from micro constituents $\{A, B, C, D\}$, and β from $\{E, F, G, H\}$. Our mapping of interest, where the state of α is dictated by the state of its output unit C over two micro updates ($\tau = 2$), is shown in Figure 5C. Under this mapping, the macro system behaves something like two reciprocally connected COPY gates, with some additional complexity provided by the connections within each macro unit. This is reflected in the macro system's TPM (Figure 5C, middle), which is very similar to the macro TPM obtained in the previous example (Figure 4C, middle). Indeed, when we measure the system integrated information of $\{\alpha, \beta\}$, we find $\varphi_s(\{\alpha, \beta\}) = 1.118$, demonstrating that this system of macro units has more irreducible, intrinsic cause-effect power than any candidate system built without macro units (Figure 5C, right).

4 Discussion

In this work, we extend IIT's mathematical framework for assessing cause-effect power (IIT 4.0 [1]) to systems of macro units. We provide a single framework, explicitly grounded on IIT's postulates (existence, intrinsicality, information, integration, exclusion, and composition), that handles both macroing over units and over updates. We further demonstrate that macro-grain systems can have higher cause-effect power than the corresponding micro-grain systems, as measured by system integrated information (φ_s).

IIT's existence postulate requires that macro units have cause-effect power (that they "take and make a difference"), as established operationally by manipulating and observing their state. Practically, this implies that the constituents of a macro unit must share a common update grain and must not overlap in their micro constituents. The next four postulates require that the cause-effect power of macro units be intrinsic, specific, irreducible ($\varphi_s > 0$), and definite. Based on IIT's principle of maximal existence (among competing existents, the one that actually exists is the one that exists the most), definiteness implies that the macro units and their state are defined such that (i) each unit is maximally irreducible "within" (it has greater φ_s than any combination of its constituents) and (ii) taken together, they constitute a complex (they maximize φ_s over their substrate). From the perspective of the complex whose structure they compose, macro units have no internal structure of their own, and exist in one of two alternative macro states.

In the case of micro units, which are subject to the same requirements as macro and meso units, the postulates of existence and intrinsicality highlight the need for indeterminism. To satisfy existence and intrinsicality, each micro unit must be able to take and make a difference from and to itself (reflexively, through a "self-loop" in the language of causal models) regardless of background conditions. This implies that each unit must have a degree of intrinsic indeterminism, because both states must always be available with non-zero probability [1].

Searching across grains for maxima (of φ_s) implicitly assumes that cause-effect power can be highest at macro grains [12, 11, 14]. The examples presented in this work employing the IIT 4.0 framework demonstrate that this is indeed possible. Specifically, we show that a macro system can have greater cause-effect power than the corresponding micro system if it is associated with reduced indeterminism and degeneracy of state transitions, such that the selectivity of causes and effects is correspondingly increased (see also [12, 11]). In IIT 4.0, this increased selectivity is captured naturally by φ_s because of its formulation in terms of intrinsic information [8, 7, 15]. Increased selectivity can arise at the macro grain because the analysis of cause-effect power treats each macro state of the macro units as equally likely, corresponding to a non-uniform distribution of micro states. Moreover, a system of macro units can have greater cause-effect power than the corresponding micro units if integration is higher at the macro level (see also [14]).

To achieve a high value of φ_s, systems of any grain must balance integration with differentiation. Whether φ_s will increase with a larger number of units depends on a balance between how much additional cause and effect information the system can specify (because its state repertoire has expanded), how much the selectivity of causes and effects within the system is reduced (because cause and effect information is spread over additional states, even more so if the additional units bring increased noise), and how well integrated the additional units are with the rest of the system [8, 7, 15]. Thus, a system of many units can only "hang together well" as an intrinsic entity if its units are themselves highly integrated and are appropriately interconnected, say as a dense, directed lattice [1]. We conjecture that macro units built upon a hierarchy of meso units may play a crucial role in allowing large systems to exist as maxima of intrinsic, irreducible cause-effect power. Hierarchies of this sort appear to be a common feature of biological systems.

In general, macro grains with φ_s values higher than *most* finer or coarser grains—that is, local or "extrinsic" maxima of integration and causal efficacy [12]—are likely to capture relevant levels of substrate organization by "carving nature at its joints." In the brain, for example, these might correspond to proteins, ion channels, organelles, synaptic vesicles, synapses, neurons, groups of tightly interconnected neurons, and so on. Such "extrinsic units," well-suited to manipulations and observations by neuroscientists, are critical for understanding how the system works. However, according to IIT, there is a critical difference between these locally maximal grains and the absolute maximal grain whose "intrinsic units" maximize φ_s within and without: only the latter constitutes the substrate of consciousness and contributes to the way the experience feels—all other levels of organization do not exist from the intrinsic perspective.

Another consequence of IIT's intrinsic framework has to do with update grains. Assuming the grain of intrinsic units is that of neurons, the update grain might be, for instance, on the order of 30 milliseconds (in line with estimates of the duration below which non-simultaneous sensory stimuli are perceived as being simultaneous, or changing stimuli are perceived as static, [24]). From the extrinsic perspective of an experimenter, several update grains may be critical to understand different kinds of causal interactions—finer grains for events such as ion channel opening, quantal release of transmitters, and the like—and longer grains for low-frequency synchronization, the induction of plastic changes, and so on. But again, while these faster and slower time scales are critical for understanding how the system works, only one time scale matters intrinsically—from the perspective of the conscious subject. Accordingly, IIT predicts that experience should only change if there is a change in the state of intrinsic units at their intrinsic update grain. Any other changes will affect the brain, but not experience. Even more stringently, the requirement that intrinsic units have binary macro states implies that any change in their micro state that does not translate into a switch of their macro state will not affect experience. For example, changes in the timing of neuronal firing, or in the rate of firing, may have clear-cut effects on the rest of the brain, but if they map onto the same intrinsic macro state, they will not have effects on the experience.

Data Availability

All code used to obtain figures and examples can be found at https://github.com/CSC-UW/Marshall_et_al_2024.

Funding

This project was made possible through the support of a grant from Templeton World Charity Foundation (TWCF0216, G.T.). In addition, this research was supported by the David P White Chair in Sleep Medicine at the University of Wisconsin-Madison, by the Tiny Blue Dot Foundation (UW 133AAG3451; G.T.), and by the Natural Science and Engineering Research Council of Canada (NSERC; RGPIN-2019-05418; W.M.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

G.T. holds an executive position and has a financial interest in Intrinsic Powers, Inc., a company whose purpose is to develop a device that can be used in the clinic to assess the presence and absence of consciousness in patients. This does not pose any conflict of interest with regard to the work undertaken for this publication.

Author contributions

	William Marshall	Graham Findlay	Larissa Albantakis	Giulio Tononi
Conceptualization	■	■	■	■
Data curation	■	■		
Formal analysis	■	■		
Funding acquisition	■			■
Investigation	■	■		
Methodology	■		■	■
Project administration				
Resources				
Software		■		
Supervision				■
Validation				
Visualization		■		
Writing – original draft	■	■		
Writing – review & editing	■	■	■	■

References

[1] Larissa Albantakis, Leonardo Barbosa, Graham Findlay, Matteo Grasso, Andrew M. Haun, William Marshall, William GP. Mayner, Alireza Zaeemzadeh, Melanie Boly, Bjorn E. Juel, Shuntaro Sasai, Keiko Fujii, Isaac David,

Jeremiah Hendren, Jonathan P. Lang, and Giulio Tononi. Integrated information theory (iit) 4.0: formulating the properties of phenomenal existence in physical terms. *PLoS Comp. Biol.*, 2023.

[2] Larissa Albantakis, Arend Hintze, Christof Koch, Christoph Adami, and Giulio Tononi. Evolution of Integrated Causal Structures in Animats Exposed to Environments of Increasing Complexity. *PLoS computational biology*, 10(12):e1003966, dec 2014.

[3] Larissa Albantakis, William Marshall, Erik Hoel, and Giulio Tononi. What caused what? A quantitative account of actual causation using dynamical causal networks. *Entropy*, 21(5):459, may 2019.

[4] Larissa Albantakis, Robert Prentner, and Ian Durham. Measuring the integrated information of a quantum mechanism. *Entropy*, 25, 2023.

[5] Nihat Ay and Daniel Polani. Information Flows in Causal Networks. *Advances in Complex Systems*, 11(01):17–41, feb 2008.

[6] David Balduzzi and Giulio Tononi. Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Comput Biol*, 4(6):e1000091, jun 2008.

[7] Leonardo S Barbosa, William Marshall, Larissa Albantakis, and Giulio Tononi. Mechanism Integrated Information. *Entropy*, 23(3):362, March 2021.

[8] Leonardo S Barbosa, William Marshall, Sabrina Streipert, Larissa Albantakis, and Giulio Tononi. A measure for intrinsic information. *Scientific Reports*, 10(1):18803, 2020.

[9] Melanie Boly, Marcello Massimini, Naotsugu Tsuchiya, Bradley R Postle, Christof Koch, and Giulio Tononi. Are the neural correlates of consciousness in the front or in the back of the cerebral cortex? clinical and neuroimaging evidence. *Journal of Neuroscience*, 37(40):9603–9613, 2017.

[10] Andrew M Haun and Giulio Tononi. Why Does Space Feel the Way it Does? Towards a Principled Account of Spatial Experience. *Entropy*, 21(12):1160, nov 2019.

[11] Erik P. Hoel, Larissa Albantakis, William Marshall, and Giulio Tononi. Can the macro beat the micro? Integrated information across spatiotemporal scales. *Neuroscience of Consciousness*, 2016(1), 2016.

[12] Erik P. Hoel, Larissa Albantakis, and Giulio Tononi. Quantifying causal emergence shows that macro can beat micro. *PNAS*, 110(49):19790–19795, nov 2013.

[13] Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf. Quantifying causal influences. *The Annals of Statistics*, 41(5):2324–2358, oct 2013.

[14] William Marshall, Larissa Albantakis, and Giulio Tononi. Black-boxing and cause-effect power. *PLOS Computational Biology*, 14(4):e1006114, apr 2018.

[15] William Marshall, Matteo Grasso, William GP Mayner, Alireza Zaeemzadeh, Leonardo S Barbosa, Erick Chastain, Graham Findlay, Shuntaro Sasai, Larissa Albantakis, and Giulio Tononi. System Integrated Information. *Entropy*, 25, 2023.

[16] William Marshall, Hyunju Kim, Sara I Walker, Giulio Tononi, and Larissa Albantakis. How causal analysis can reveal autonomy in models of biological systems. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 375(2109):20160358, dec 2017.

[17] William G.P. Mayner, William Marshall, Larissa Albantakis, Graham Findlay, Robert Marchman, and Giulio Tononi. PyPhi: A toolbox for integrated information theory. *PLoS Computational Biology*, 14(7):e1006343, jul 2018.

[18] Brian Odegaard, Robert Knight, and Hakwan Lau. Should a few null findings falsify prefrontal theories of conscious perception? *Journal of Neuroscience*, 37, 2017.

[19] Masafumi Oizumi, Larissa Albantakis, and Giulio Tononi. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Computational Biology*, 10(5):e1003588, may 2014.

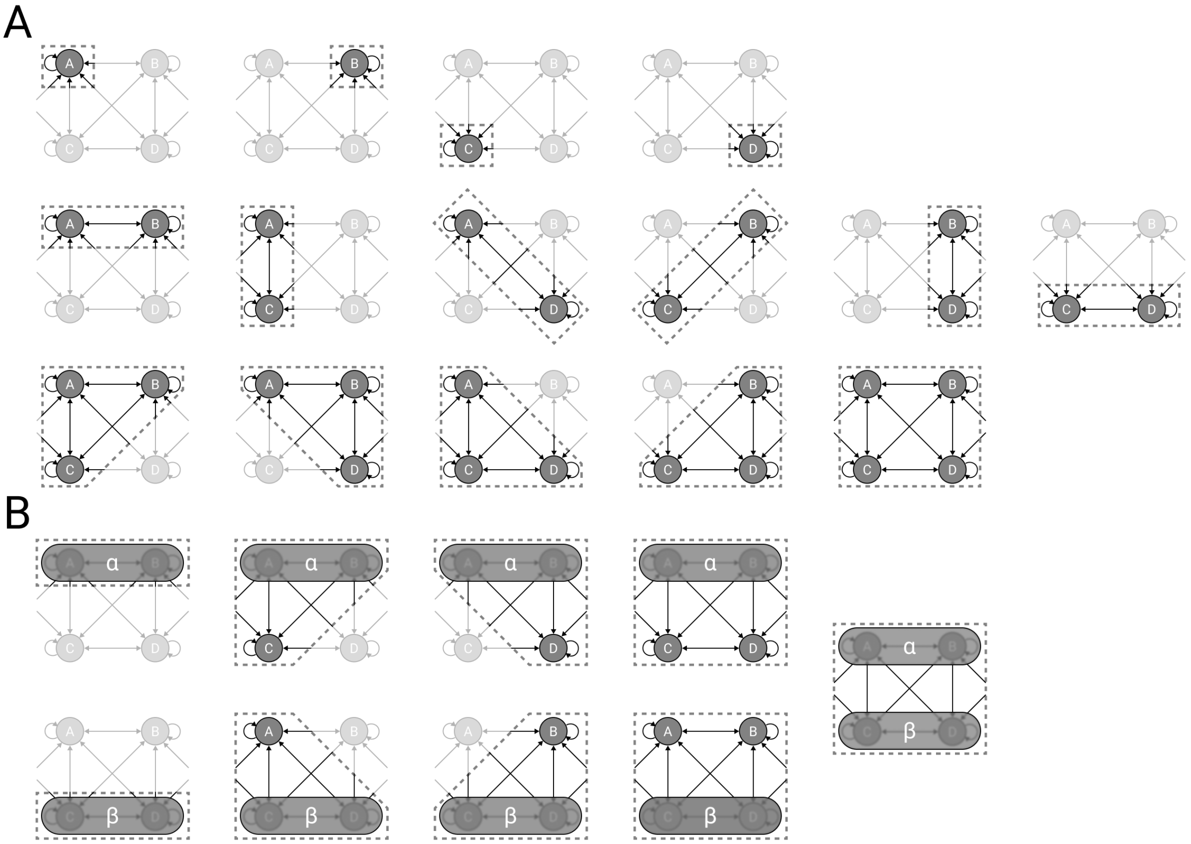
[20] J Pearl. *Causality: models, reasoning and inference*, volume 29. Cambridge Univ Press, 2000.

[21] G Tononi. *On being*. forthcoming.

[22] Giulio Tononi. An information integration theory of consciousness. *BMC neuroscience*, 5:42, nov 2004.

[23] Giulio Tononi and Olaf Sporns. Measuring information integration. *BMC neuroscience*, 4(31):1–20, 2003.

[24] Peter A. White. Is conscious perception a series of discrete temporal frames? *Consciousness and Cognition*, 60:98–126, 2018.



C

α = {(A, B), (A, B), τ=1, g}

g: $\Omega_{(A,\,B)}{}^{\tau=1} \rightarrow \{0, 1\}$

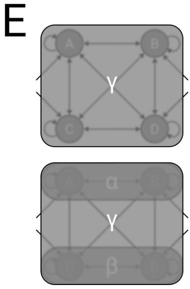
$AB_{t=0}$ →	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}	g_{12}	g_{13}	g_{14}
00	1	0	1	0	1	0	1	0	1	0	1	0	1	0
10	0	1	1	0	0	1	1	0	0	1	1	0	0	1
01	0	0	0	1	1	1	1	0	0	0	0	1	1	1
11	0	0	0	0	0	0	0	1	1	1	1	1	1	1

D

α = {(A, B), (A, B), τ=2, g}

g: $\Omega_{(A,\,B)}{}^{\tau=2} \rightarrow \{0, 1\}$

$AB_{t=-1}$	$AB_{t=0}$ →	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}	g_{12}	…	g_{65534}
00	00	1	0	1	0	1	0	1	0	1	0	1	0		0
00	10	0	1	1	0	0	1	1	0	0	1	1	0		1
00	01	0	0	0	1	1	1	1	0	0	0	0	1		1
00	11	0	0	0	0	0	0	0	1	1	1	1	1		1
10	00	0	0	0	0	0	0	0	0	0	0	0	0		1
10	10	0	0	0	0	0	0	0	0	0	0	0	0		1
10	01	0	0	0	0	0	0	0	0	0	0	0	0		1
10	11	0	0	0	0	0	0	0	0	0	0	0	0	…	1
01	00	0	0	0	0	0	0	0	0	0	0	0	0		1
01	10	0	0	0	0	0	0	0	0	0	0	0	0		1
01	01	0	0	0	0	0	0	0	0	0	0	0	0		1
01	11	0	0	0	0	0	0	0	0	0	0	0	0		1
11	00	0	0	0	0	0	0	0	0	0	0	0	0		1
11	10	0	0	0	0	0	0	0	0	0	0	0	0		1
11	01	0	0	0	0	0	0	0	0	0	0	0	0		1
11	11	0	0	0	0	0	0	0	0	0	0	0	0		1



γ = {(A, B, C, D), (A, B, C, D), τ, g_γ}
τ=1: 32,727 possible g_γ
τ=2 → ~5.78x10^{76} possible g_γ

α = {(A, B), (A, B), τ, g_α}
β = {(C, D), (C, D), τ, gβ}
γ = {(A, B, C, D), (α, β), τ, g_γ}
τ=1 → 7 possible g_γ
τ=2 → 32,727 possible g_γ

Figure 3: Defining macro units. Four micro units $\{A, B, C, D\}$ are embedded within a larger universe U, with unspecified transition probability function $\mathcal{T}_U$. We wish to know if a macro unit γ can be built using micro constituents $U^\gamma = \{A, B, C, D\}$, possibly with intermediate meso constituents $V^\gamma \neq U^\gamma$. (**A**) To ask if $\gamma = \{(A, B, C, D), ...\}$ is admissible as a macro unit, we must first check integrated information φ_s for every subset of micro units. (**B**) Suppose we find that $\{A, B\}$ and $\{C, D\}$ are maximally irreducible within (i.e. they satisfy Eqn. 7). This means that they are potential meso units, labeled α and β respectively. To continue verifying that $\gamma = \{(A, B, C, D), ...\}$ is admissible as a macro unit, we must now check integrated information for every subset of units that include α and β as well. (**C**) Let S be the macro system containing α. For a given candidate unit, say $\alpha = \{(A, B), (A, B), \tau_\alpha = 1, g_\alpha)\}$, there are many potential mappings g_α from the states of $V^\alpha = \{A, B\}$ over a sequence of $\tau_\alpha = 1$ updates to the state of α, but only one (here unspecified) will maximize $\varphi_s(s)$. (**D**) Same as (C), but over a sequence of $\tau_\alpha = 2$. Note that it is not only the ultimate state of the micro constituents that determine the macro unit's state, but the precise sequence of micro states. (**E**) Depending on which of $\{A, B, C, D\}$ or $\{\alpha, \beta\}$ (or some mixture) is maximally irreducible, γ's constituents V^γ might be $\{A, B, C, D\}$ or $\{\alpha, \beta\}$ (or some mixture), which in turn will dictate the set of potential mappings from which g_γ can be defined, for any given τ. There are far fewer mappings that need to be considered for a macro unit whose constituents are meso units, because the mapping of the macro unit (γ) is constrained by the mappings of its meso constituents (α, β).

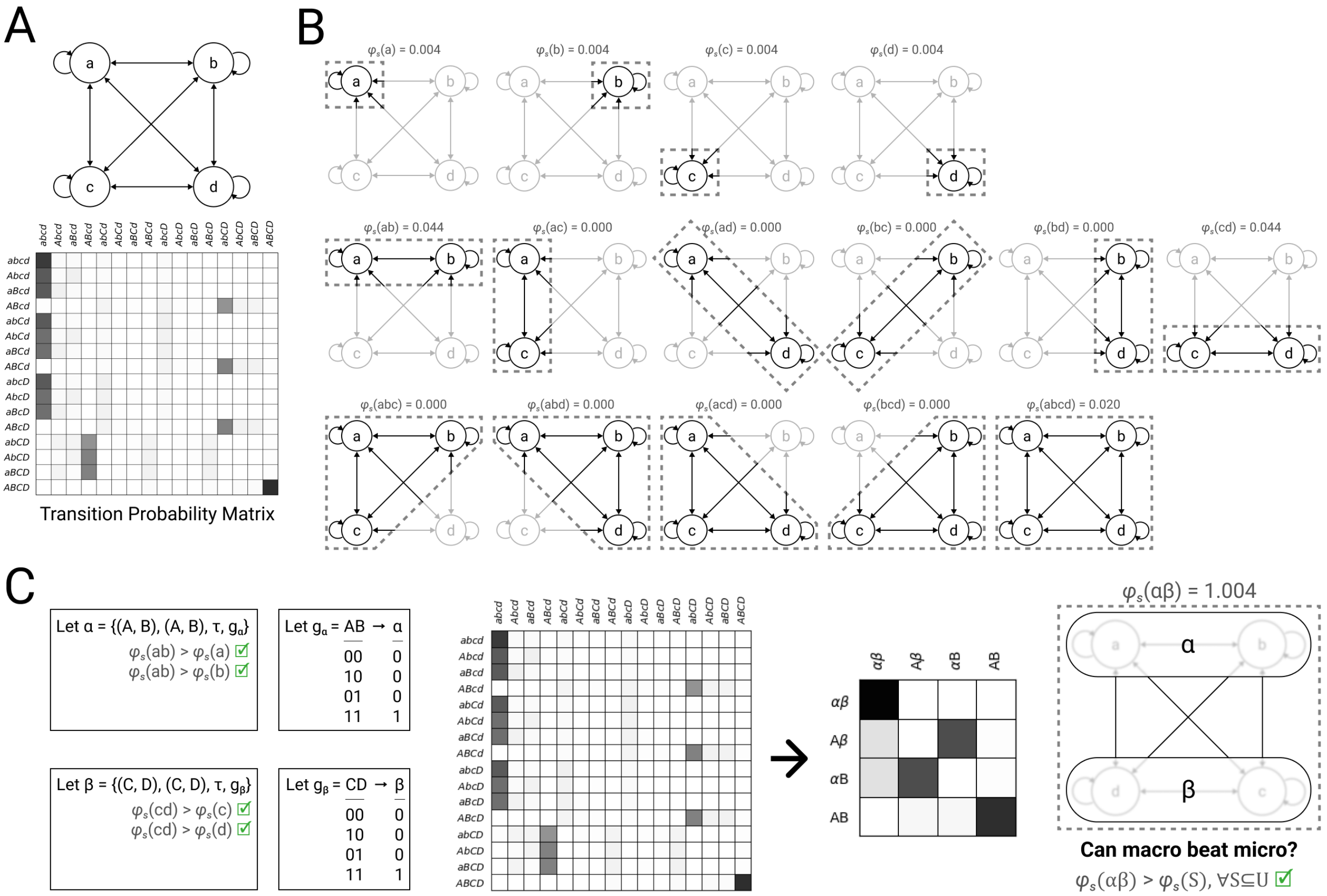


Figure 4: Example 1. (**A**) Consider four micro units $\{A, B, C, D\}$ in state $(0, 0, 0, 0)$, with transition probability matrix $\mathcal{T}_U$. For illustrative purposes, capitalization denotes the state of each unit, both in causal network diagrams and transition probability matrix state labels (e.g. state $(0, 1, 0, 0)$ is written $aBcd$). (**B**) System integrated information $\varphi_s(s)$ must be checked for each subset of micro units $S \in \mathbb{P}(\{A, B, C, D\})$. Greyed-out units are background. Notice that $\{A, B\}$ and $\{C, D\}$ are maximally irreducible within. In the case of $\{A, B\}$: $\varphi_s(\{A, B\}) = 0.044$, greater than either $\varphi_s(A) = 0.004$ or $\varphi_s(\{B\}) = 0.004$. (**C**) Since $\{A, B\}$ and $\{C, D\}$ are maximally irreducible within, we may consider their potential macro units, labeled α and β respectively. One possible pair of mappings for these macro units are g_α and g_β, resulting in macro TPM $\mathcal{T}^S$. This candidate system $\{\alpha, \beta\}$ in state $(0, 0)$ (given by g_α, g_β) has system integrated information $\varphi_s = 1.004$, greater than any of the micro level candidate systems in (B). Thus, although we would have to check all other valid macro unit definitions and mappings in order to determine whether this macro system is *maximally* irreducible relative to all others, we can conclude that intrinsic cause-effect power will be higher at a macro level than at the micro level—we know that we can do at least as well as $\varphi_s = 1.004$.

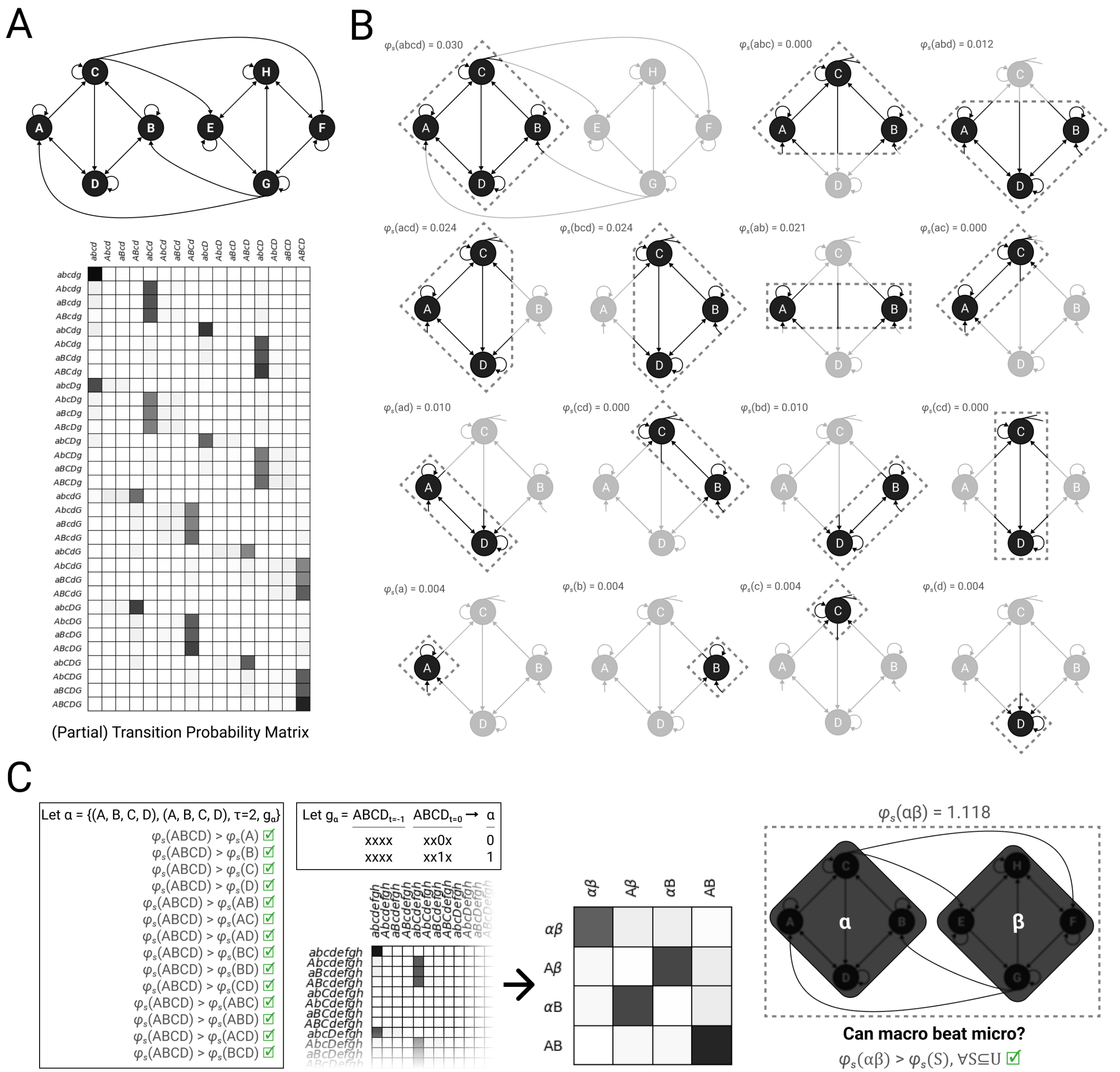


Figure 5: Example 2. (**A**) Consider eight micro units $\{A, B, C, D, E, F, G, H\}$ in state $(1, 1, 1, 1, 1, 1, 1, 1)$, with transition probability matrix (TPM) $\mathcal{T}_U$. Because of space limitations in this and subsequent panels, we illustrate some analysis steps for the left half of the system only (i.e. $\{A, B, C, D\}$), but all calculations were done using the full eight-unit system. For example, although the full TPM is used for all calculations, a partial TPM illustrating the behavior of $\{A, B, C, D\}$ is shown here. Rows are past system states and columns are future states. (**B**) System integrated information $\varphi_s(s)$ must be checked for each subset of micro units $S \in \mathbb{P}(\{A, B, C, D, E, F, G, H\})$. Here, because of space limitations, we illustrate checks for $S \in \mathbb{P}(\{A, B, C, D\})$. Notice that $\{A, B, C, D\}$ is maximally irreducible within. (**C**) Since $\{A, B, C, D\}$ is maximally irreducible within, we may consider its potential macro unit, labeled α. One possible mapping for α with $\tau = 2$ is shown. Since the full system is symmetric, $\{E, F, G, H\}$ can be considered as a potential macro unit β with analogous g_β, resulting in macro TPM $\mathcal{T}^S$. This candidate system $\{\alpha, \beta\}$ in in state $(1, 1)$ (given by g_α, g_β) has system integrated information $\varphi_s = 1.118$, greater than all of the micro level candidate systems (max $\varphi_s = 0.135$, not shown, but see (B) for a subset). Thus, although we would have to check all other valid macro unit definitions and mappings in order to determine whether this macro system is *maximally* irreducible relative to all others, we can conclude that intrinsic cause-effect power will be higher at a macro level than at the micro level—we know that we can do at least as well as $\varphi_s = 1.118$.