Entropy **2015**, *17*, 5472-5502; doi:10.3390/e17085472

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

The Intrinsic Cause-Effect Power of Discrete Dynamical Systems—From Elementary Cellular Automata to Adapting Animats

Larissa Albantakis * and Giulio Tononi *

Department of Psychiatry, University of Wisconsin, Madison 53719, WI, USA

* Authors to whom correspondence should be addressed; E-Mails: albantakis@wisc.edu (L.A.); gtononi@wisc.edu (G.T.).

Academic Editors: Christoph Salge, Georg Martius, Keyan Ghazi-Zahedi and Daniel Polani

Received: 22 May 2015 / Accepted: 28 July 2015 / Published: 31 July 2015

Abstract: Current approaches to characterize the complexity of dynamical systems usually rely on state-space trajectories. In this article instead we focus on causal structure, treating discrete dynamical systems as directed causal graphs—systems of elements implementing local update functions. This allows us to characterize the system's intrinsic cause-effect structure by applying the mathematical and conceptual tools developed within the framework of integrated information theory (IIT). In particular, we assess the number of irreducible mechanisms (concepts) and the total amount of integrated conceptual information Φ specified by a system. We analyze: (i) elementary cellular automata (ECA); and (ii) small, adaptive logic-gate networks ("animats"), similar to ECA in structure but evolving by interacting with an environment. We show that, in general, an integrated cause-effect structure with many concepts and high Φ is likely to have high dynamical complexity. Importantly, while a dynamical analysis describes what is "happening" in a system from the extrinsic perspective of an observer, the analysis of its cause-effect structure reveals what a system "is" from its own intrinsic perspective, exposing its dynamical and evolutionary potential under many different scenarios.

Keywords: integration; information; causation; artificial evolution

---

1. Introduction

The term "dynamical system" encompasses a vast class of objects and phenomena—any system whose state evolves deterministically with time over a state space according to a fixed rule [1]. Since living systems must necessarily change their state over time, it is not surprising that many attempts have been made to model a wide range of living systems as dynamical systems [2–7], although identifying an accurate time-evolution rule and estimating the relevant state variables can be hard. On the other hand, even simple state-update rules can give rise to complex spatio-temporal patterns. This has been demonstrated extensively using a class of simple, discrete dynamical systems called "cellular automata" (CA) [8–12]. CA consist of a lattice of identical cells with a finite set of states. All cells evolve in parallel according to the same local update rule, which takes the states of neighboring cells into account (Figure 1A). In any dynamical system, the time-evolution rule determines the future trajectory of the system in its state space given a particular initial state. The aim of dynamical systems theory is to understand and characterize a system based on its long-term behavior, by classifying the geometry of its long-term trajectories. In this spirit, cellular automata have been classified according to whether their evolution for most initial states leads to fixed points, periodic cycles, or chaotic patterns associated with strange attractors [8,13] (Figure 1B). Despite the simplicity of the rules, this classification is undecidable for many CA with infinite or very large numbers of cells [14,15]. This is because determining the trajectory of future states is often computationally irreducible [16], meaning that it is impossible to predict the CA's long-term behavior in a more computationally efficient way than by actually running the system. In general, the relationship between the time-evolution rule, which describes the *local* behavior of each cell, and the *global* behavior of the entire CA remains indeterminate. Assuming an actual physical implementation of a finite size CA, the time-evolution rule is equivalent to a cell's mechanism, which determines its causal interactions with neighboring cells (Figure 1C, see below). Like a logic-gate, each cell computes its current state according to its rule, based on the inputs it receives from itself and its neighbors, and then outputs its state to itself and its neighbors in turn.

While the rich dynamical repertoire (dynamical complexity) of cellular automata has been studied extensively, their causal structure (causal complexity) has received little attention, presumably because it is assumed that all that matters causally reduces to the simple mechanism of the cells, and anything that may be interesting and complex is only to be found in the system's dynamic behavior. For the dynamical system itself, however, whether it produces interesting patterns or not might not make any causal difference.

Integrated information theory (IIT) [17,18] offers a mathematical framework to characterize the cause-effect structure specified by all the mechanisms of a system from its own intrinsic perspective, rather than from the perspective of an extrinsic observer. Table 1 provides an overview of all relevant IIT quantities. In IIT a mechanism is any system element or *combination* of elements with a finite number of states and an update rule, such as a CA cell, or a logic-gate, as long as it has irreducible cause-effect power within the system. This means that: (a) the mechanism must constrain the past and future states of the system by being in a particular state (information); and (b) the particular way in which it does so—the mechanism's cause-effect repertoire—must be irreducible to the cause-effect repertoire of its parts (integration), as measured by its integrated information φ.

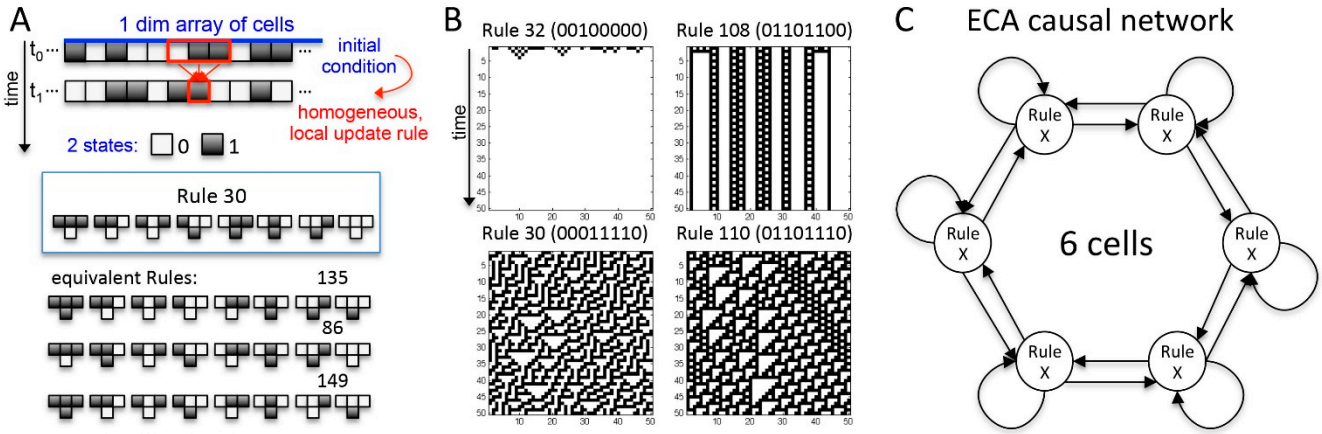

Figure 1. Elementary cellular automata (ECA). (**A**) ECA consist of a 1-dimensional chain of cells in either state 0 or 1, which are updated according to a time-evolution rule dependent on a cell's previous state and that of its two nearest neighbors. Given the $2^3 = 8$ possible past configurations, 256 different update rules exist, which are labeled by the decimal number of their rule in binary. "0" to "1" and/or left-right transformation of a rule (past and current state) lead to rules with equivalent behavior; (**B**) Example evolutions of four ECA (number of cells $N = 50$) with distinct long-term behavior over 50 time-steps for a random initial state and periodic boundary conditions; (**C**) ECA with $N = 6$ cells and periodic boundary conditions illustrated as a network of interacting elements. Edges denote connections between cells. Each cell in the ECA has a self-loop and connections to its nearest neighbors. Note, however, that depending on the cell's update rule some of the edges may not be causally effective, e.g., if the state of the left neighbor is irrelevant as in rule 136 (10001000).

Mathematically, the cause-effect repertoire (*CER*) of a mechanism M_t in its current state m_t is a set of two conditional probability distributions: the possible past states of a set of elements within the system (Z_{t-1}) and the possible futures states of a set of elements within the system (Z_{t+1}) conditioned on m_t:

$$CER(m_t, Z_{t\pm1}) = \{p_{cause}(z_{t-1}|m_t), p_{effect}(z_{t+1}|m_t)\} \quad (1)$$

By contrast to correlational measures, which use observed state distributions, in IIT cause-effect power is quantified taking all possible system perturbations into account with equal probability. In the Markovian systems discussed below, this corresponds to applying independent, maximum-entropy input states to all system elements (see [17,18] and Supplementary Methods for details).

To assess the integrated information φ of a mechanism $M_t = m_t$ for a given $Z_{t\pm1}$, its *CER* is partitioned into two independent parts by injecting independent noise into the connections between $\{M_{1,t}, Z_{1,t\pm1}\}$ and $\{M_{2,t}, Z_{2,t\pm1}\}$. φ is then measured as the distance between the intact *CER* and the product *CER* under the partition *P*:

$$\varphi(m_t, Z_{t\pm1}, P) = D\big(CER(m_t, Z_{t\pm1}), CER(m_{1,t}, Z_{1,t\pm1}) \times CER(m_{2,t}, Z_{2,t\pm1})\big) \quad (2)$$

To quantify the irreducibility of a mechanism over a particular set $Z_{t\pm1}$ it is important to determine φ for the "minimum information partition" (*MIP*), the partition that makes the least difference: $MIP = \underset{P}{\operatorname{argmin}} \big(\varphi(m_t, Z_{t\pm1}, P)\big)$. The maximally irreducible *CER* of a mechanism in a state $M_t = m_t$ over the particular set of system elements $Z^*_{t\pm1}$ that maximizes $\varphi(m_t, Z_{t\pm1}, MIP)$ is called its "concept"

(see Table 1). A concept determines the mechanism's causal role within the system from the intrinsic perspective of the system itself. Elementary mechanisms specify 1st-order concepts; mechanisms composed of several system elements specify higher-order concepts.

The set of all concepts is the "cause-effect structure" of a system S_t in state s_t. Integrated conceptual information Φ quantifies the irreducible cause-effect power of a system of mechanisms taken as a whole. It measures how irreducible a system's cause-effect structure $C(s_t)$ is compared to the cause-effect structure of the system when it is causally partitioned (unidirectionally) across its weakest link $C(s_t, MIP)$:

$$\Phi(s_t, MIP) = D(C(s_t) | C(s_t, MIP)) \quad (3)$$

where D is a distance measure between the two cause-effect structures (see Methods and [17]).

Over a particular set of elements, one can define any subset as a system from the extrinsic perspective of an observer, some of which may specify integrated cause-effect structures with $\Phi > 0$. From the intrinsic perspective taken by IIT, however, only non-overlapping sets of elements that specify local maxima of integrated conceptual information Φ^{Max} form "complexes" with self-defined causal boundaries. Only sets of elements with many specialized, but integrated concepts can achieve high values of Φ. Φ can thus be viewed as a measure of the intrinsic causal complexity of a system.

In this paper, we wish to investigate the relation between the dynamical properties of a system and its intrinsic cause-effect power, both in isolated systems and in agents that evolve in and interact with an environment of rich causal structure. To that end, we: (i) exhaustively characterize the cause-effect power of elementary cellular automata (ECA), one-dimensional CA with only nearest-neighbor interactions; and (ii) examine the causal and dynamical properties of small, adaptive logic-gate networks ("animats") evolving in task environments with different levels of complexity. While the state evolution of isolated systems, such as the ECA, must be a product of their intrinsic mechanisms, the dynamics of behaving agents, such as the animats, is at least partially driven by the inputs they receive from their environment. We predict that, to have a large dynamical repertoire, isolated systems must have an integrated cause-effect structure with many concepts. In particular, isolated systems that: (i) have few, unselective mechanisms with low φ^{Max}; (ii) lack composition, meaning their cause-effect structures lack higher-order concepts; or (iii) are reducible ($\Phi = 0$), should not be able to produce interesting global dynamics. In isolated systems, IIT measures characterizing intrinsic cause-effect power, such as Φ^{Max}, the number of concepts, and their $\Sigma\varphi^{Max}$, should thus correlate with general dynamical properties of the system, such as the maximal transient length to reach fixed points or periodic cycles. By contrast, non-isolated systems can exhibit complex reactive behavior driven by the environment. Analyzing the intrinsic cause-effect structure of a behaving agent can elucidate to what extent the agent itself has a complex structure, and to what extent it is merely reactive. Moreover, integrated systems with a rich cause-effect structure have adaptive advantages in environments that require context-sensitivity and memory [19]. Finally, the examples discussed in this article also reveal conceptual dissociations between the observable behaviors of discrete dynamical systems, which describe what a system "happens to be doing", and their intrinsic cause-effect structures, which describe what a system "is".

The material on evolving animats presented in this article includes and extends data from [19]. A condensed version of this article will appear as a book chapter in "From Matter to Life: Information and Causality".

Table 1. Overview of integrated information theory (IIT) quantities of mechanisms and systems of elements. For a more detailed mathematical formulation see the glossary of [17,18]. D: distances in IIT are measured using the earth mover's distance (see Methods).

MECHANISM	cause-effect repertoire $CER(m_t, Z_{t\pm1})$	A set of two conditional probability distributions: $CER(m_t, Z_{t\pm1}) = \{p_{cause}(z_{t-1}\|m_t), p_{effect}(z_{t+1}\|m_t)\}$, describing how the mechanism M_t in its current state m_t constrains the past and future states of the sets of system elements Z_{t-1} and Z_{t+1}, respectively.
	partition *P*	$P = \{M_{1,t}, Z_{1,t\pm1}; M_{2,t}, Z_{2,t\pm1}\}$, where the connections between the parts $\{M_{1,t}, Z_{1,t\pm1}\}$ and $\{M_{2,t}, Z_{2,t\pm1}\}$ are injected with independent noise.
	integrated information φ ("small phi")	φ measures the irreducibility of a CER w.r.t. a partition *P*: $\varphi(m_t, Z_{t\pm1}, P) = D\big(CER(m_t, Z_{t\pm1}), CER(m_{1,t}, Z_{1,t\pm1}) \times CER(m_{2,t}, Z_{2,t\pm1})\big)$
	MIP	The partition that makes the least difference to a *CER*: $MIP = \underset{P}{\operatorname{argmin}}\big(\varphi(m_t, Z_{t\pm1}, P)\big)$.
	$Z^*_{t\pm1}$	The set of system elements $Z^*_{t\pm1} = \{Z^*_{t-1}, Z^*_{t+1}\}$, where $Z^*_{t-1} = \left\{\underset{Z_{t-1}}{\operatorname{argmax}}\big(\varphi_{cause}(m_t, Z_{t-1}, MIP_{cause})\big)\right\}$ and $Z^*_{t+1} = \left\{\underset{Z_{t+1}}{\operatorname{argmax}}\big(\varphi_{effect}(m_t, Z_{t+1}, MIP_{effect})\big)\right\}$.
	$\varphi^{Max}(m_t)$	The intrinsic cause-effect power of a mechanisms M_t: $\varphi^{Max}(m_t) = \varphi\big(m_t, Z^*_{t\pm1}, MIP\big) = \min\big(\varphi^{Max}_{cause}(m_t), \varphi^{Max}_{effect}(m_t)\big) = \min\big(\varphi_{cause}(m_t, Z^*_{t-1}, MIP_{cause}), \varphi_{effect}\big(m_t, Z^*_{t+1}, MIP_{effect}\big)\big)$
	concept	The maximally irreducible *CER(m_t)* with $\varphi^{Max}(m_t)$ over $Z^*_{t\pm1}$: $CER(m_t, Z^*_{t\pm1}) = \{p_{cause}(z^*_{t-1}\|m_t), p_{effect}(z^*_{t+1}\|m_t)\}$, describing the causal role of mechanism M_t within the system.
SYSTEM	cause-effect structure *C(s_t)*	The set of concepts specified by all mechanisms with $\varphi^{Max}(m_t) > 0$ within the system S_t in its current state s_t.
	$\Sigma\varphi^{Max}$	The sum of all $\varphi^{Max}(m_t)$ of *C(s_t)*.
	unidirectional partition $P_{\rightarrow}$	$P_{\rightarrow} = \{S_1, S_2\}$, where the connections from the set of elements S_1 to S_2 are injected with independent noise (for *t-1* → *t* and *t* → *t+1*).
	integrated conceptual information Φ ("big phi")	Φ measures the irreducibility of a cause-effect structure w.r.t. a partition $P_{\rightarrow}$: $\Phi(s_t, P_{\rightarrow}) = D(C(s_t)\|C(s_t, P_{\rightarrow}))$. Φ captures how much the *CER*s of the system's mechanisms are altered and how much φ^{Max} is lost by partitioning the system.
	MIP	The unidirectional system partition that makes the least difference to *C(s_t)*: $MIP = \underset{P_{\rightarrow}}{\operatorname{argmin}}\big(\Phi(s_t, P_{\rightarrow})\big)$.
	Φ^{Max}	The intrinsic cause-effect power of a system of elements S^*_t. $\Phi^{Max} = \Phi(s^*_t) > 0$ such that for any other S_t with $(S_t \cap S^*_t) \neq \emptyset$, $\Phi(s_t) \leq \Phi(s^*_t)$.
	complex	A set of elements S^*_t with $\Phi^{Max} = \Phi(s^*_t) > 0$. A complex thus specifies a maximally irreducible cause-effect structure.

2. Results and Discussion

Central to integrated information theory (IIT) is the postulate that, in order to exist, a system in a state must have cause-effect power, since there is no point in assuming that something exists if nothing can make a difference to it or it does not make a difference to anything. To exist from its own intrinsic perspective, the system moreover must have cause-effect power upon itself. To that end, the system must be comprised of mechanisms, elements that have cause-effect power on the system, alone or in combination.

Our objectives here are to assess whether and how much certain isolated and adaptive discrete dynamical systems exist (have irreducible cause-effect power) from their own intrinsic perspective and to determine how their cause-effect structures relate to their dynamic complexity. With our results we want to shed light on the distinction between the intrinsic perspective of the system itself and the extrinsic perspective of an observer, and highlight key aspects of the IIT formalism, such as the notions of causal selectivity, composition, and irreducibility, and their significance in the analysis of discrete dynamical systems.

2.1. Behavior and Cause-Effect Power of Elementary Cellular Automata

In order to analyze the cause-effect power of elementary cellular automata (ECA), we treat the ECA as directed causal graphs, meaning as systems of connected elements that each implement a particular function (Figure 1C). If not specified otherwise, all elements of a system implement the same ECA rule. It is assumed that the transition probabilities of all system states are known. In a discrete, deterministic system, such as the ECA, this corresponds to perturbing the system into all its possible states and recording all state-to-state transitions over one time-step (see Supplementary Methods). In the Methods section, we outline how the integrated conceptual information Φ and the cause-effect structure of an example ECA in a particular state are determined. Even for simple systems with a low number of cells N, evaluating the cause-effect structure and its Φ value for a given state is computationally costly. For this reason, the Φ values of the ECA rules presented below were calculated from only N+1 states out of the 2^N possible states with different numbers of cells in states "0" and "1", from which we obtained an average value ($<\Phi^{Max}>$). Since all elements in an ECA specify the same rule and symmetric states have redundant cause-effect structures, this sample of states is representative of a large number of the 2^N possible states (see Figure S1 for distributions of measured IIT quantities across all ECA with $N = 5$).

For further details on the mathematical and conceptual tools developed within the IIT framework see [17,18]. For the interested reader, the IIT website [20] allows calculating Φ and other IIT measures for small systems of logic gates.

2.1.1. Cause-Effect Structure of ECA *vs.* Wolfram Classes

There exist 256 possible ECA time-evolution rules, which can be grouped into 88 different equivalency classes. Each class contains maximally four rules, which show identical behavior under "0" to "1" transformations, left-right transformations, or both transformations applied at once (Figure 1A). Equivalent rules have identical Φ values and equivalent cause-effect structures for complementary states. In the remaining article we will thus label each equivalency class by its lowest-numbered member rule. In Figure 2 the average Φ^{Max} values of all 88 ECA rule classes are plotted against their respective number of concepts and $\Sigma\varphi^{Max}$ for systems with five and six cells. Φ^{Max} is the integrated conceptual information

of the main complex in the system, the set of elements S_t^* that is most irreducible. φ^{Max} measures the cause-effect power of individual concepts within the main complex (see Table 1 and Methods). Each equivalency class is color-coded according to its Wolfram classification I–IV for random initial states [9] according to whether almost all initial conditions lead to (I) the same uniform stable state (black); (II) a non-uniform stable state or periodic behavior (blue); (III) pseudo-random or chaotic behavior (green), or (IV) complex behavior with a mixture of randomness and order (red). Figure 1B shows the time evolution of four example rules from class I–IV, which are also highlighted in Figure 2. For ECA with five cells, simple rules from class I all have $\Phi = 0$ or low $<\Phi^{Max}>$, only a limited number of concepts, and a low value of $<\Sigma\varphi^{Max}>$. Rules from class II have low to intermediate values of $<\Phi^{Max}>$, but can have a high number of concepts, albeit typically with lower $<\Sigma\varphi^{Max}>$ than class III rules (Figure 2B). Class III rules show the highest values of $<\Phi^{Max}>$ and $<\Sigma\varphi^{Max}>$ with an intermediate number of concepts. For class III systems to have fewer concepts but higher $<\Sigma\varphi^{Max}>$ compared to class I and II systems, their concepts must typically be more selective and/or more irreducible (have more cause-effect power) than those of class I or II systems. Finally, class IV rules have high numbers of concepts with intermediate to high $<\Phi^{Max}>$ and $<\Sigma\varphi^{Max}>$ values.

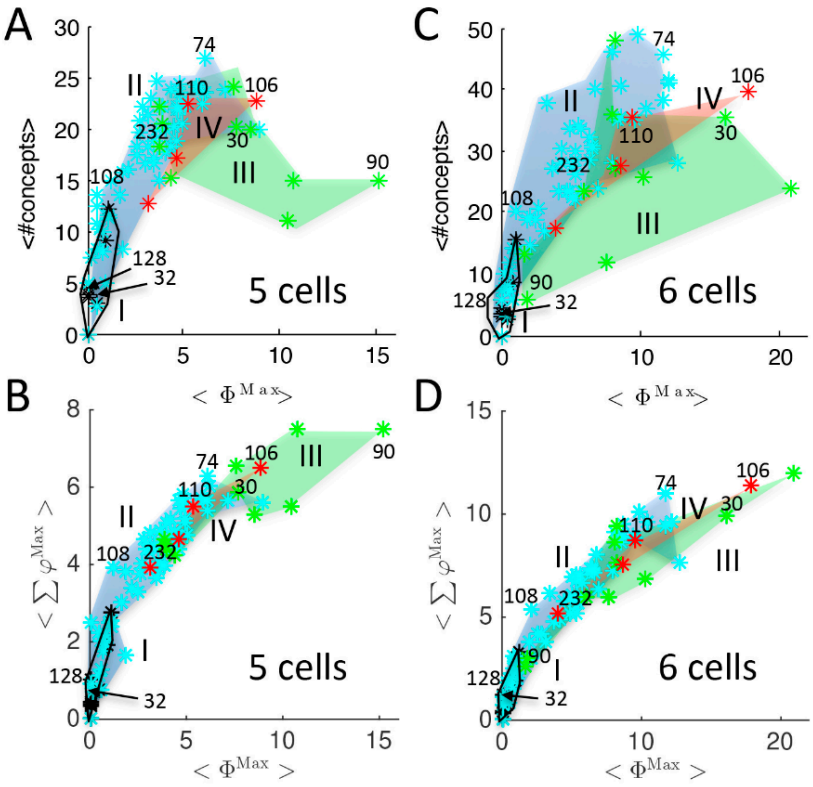

Figure 2. Relation of $<\Phi^{Max}>$, $<\#concepts>$, and $<\Sigma\varphi^{Max}>$ to Wolfram ECA classes I-IV. (**A**) Mean number of concepts $<\#concepts>$ plotted against $<\Phi^{Max}>$ for all 88 ECA equivalence classes for ECA implementations with $N = 5$ cells. Each rule is color coded by its Wolfram class: (I) uniform fixed points in black; (II) non-uniform fixed points and periodic behavior in blue, (III) random, chaotic behavior in green; and (IV) complex behavior in red; (**B**) $<\Sigma\varphi^{Max}>$ of all concepts in the system plotted against $<\Phi^{Max}>$ for $N = 5$ cells. (**C**,**D**) Same as (**A**,**B**), for $N = 6$ cells. Rules shown in Figure 1B and below in Figure 3 are labeled by their rule number.

The number of concepts, $<\Sigma\varphi^{Max}>$, and $<\Phi^{Max}>$ of ECA tends to grow with the number of cells N (see below). Nevertheless the distribution of rules on the $<\Phi^{Max}>$/<#concepts> and $<\Phi^{Max}>$/$<\Sigma\varphi^{Max}>$ planes for ECA with six cells stays the same for classes I, II, and IV, and most rules of class III. Exceptions are class III rules 18, 90, 105, and 150, which are strongly dependent on N being even or odd, with lower <#concepts> for even N.

An ECA's cause-effect structure can yield insights about the dynamic potential of the rule it is implementing by making its specific causal features explicit (Figure 3). The simplest rule, both dynamically and causally, is rule 0 (00000000), which maps every possible state to "all 0" and thus belongs to class I. Systems which implement rule 0 cannot specify information about the past state of the system (all system states are possible past states of state "all 0") and thus have no concepts (all $\varphi = 0$). Moreover, analyzed from the intrinsic perspective, these systems cannot form complexes and are always reducible ($\Phi = 0$) (Figure 3A). The latter also applies to class II rules 204 (11001100), the Identity rule, and its negation, rule 51 (00110011). This is because under these rules individual cells do not interact with each other and therefore can always be partitioned without loss of cause-effect power. Reducible systems cannot exist as a whole from the intrinsic perspective of the system itself, since $\Phi = 0$ means that one part of the system is not affected by the other. For the same reason, reducible systems cannot produce complex global dynamics.

Several class I and II rules, such as the AND rule 128, allow for complexes, but have only elementary mechanisms (only 1st-order concepts), which leads to low Φ values in all states (Figure 3B, Figure 4). The cause-effect structures of systems that implement these rules lack composition [17,18]: all sets of elements in these systems are reducible, meaning they do not have cause-effect power over and above the elementary mechanisms ($\varphi = 0$). In the example of Figure 3B, knowing that element A is in state "0" already specifies that the cells in its neighborhood A, B, and E must be "0" in the next state, since they are AND mechanisms. The state of element B or E does not have any additional effect with respect to specifying the system's future state. Another example of this type is the COPY rule 170 (10101010), which belong to class II. Note that all rules with such simple cause-effect structures can be expressed in terms of purely linear minimal Boolean functions over the cell neighborhood. In fact, all evaluated systems with <#concepts> $\leq N$ and $<\Phi^{Max}> \leq 1$ implement linearly separable rules (defined e.g., in [21]). Nevertheless, some linear rules still show a high number of concepts with low to medium φ^{Max} and moderate $<\Phi^{Max}>$ values, such as for example the Majority rule 232 (Figure 3C).

Nonlinearity is necessary, but not sufficient for class III and class IV behavior in ECA [21]. Rule 74 (01001010) for example belongs to Wolfram class II. In state "all 0" the system implementing rule 74 with five cells has only slightly higher Φ than the system implementing the linear rule 232 (Figure 3C,D). Averaged over N+1 states, however, $<\Phi^{Max}>$, <#concepts>, and $<\Sigma\varphi^{Max}>$ of rule 74 are substantially higher than rule 232 (Figures 2 and 4).

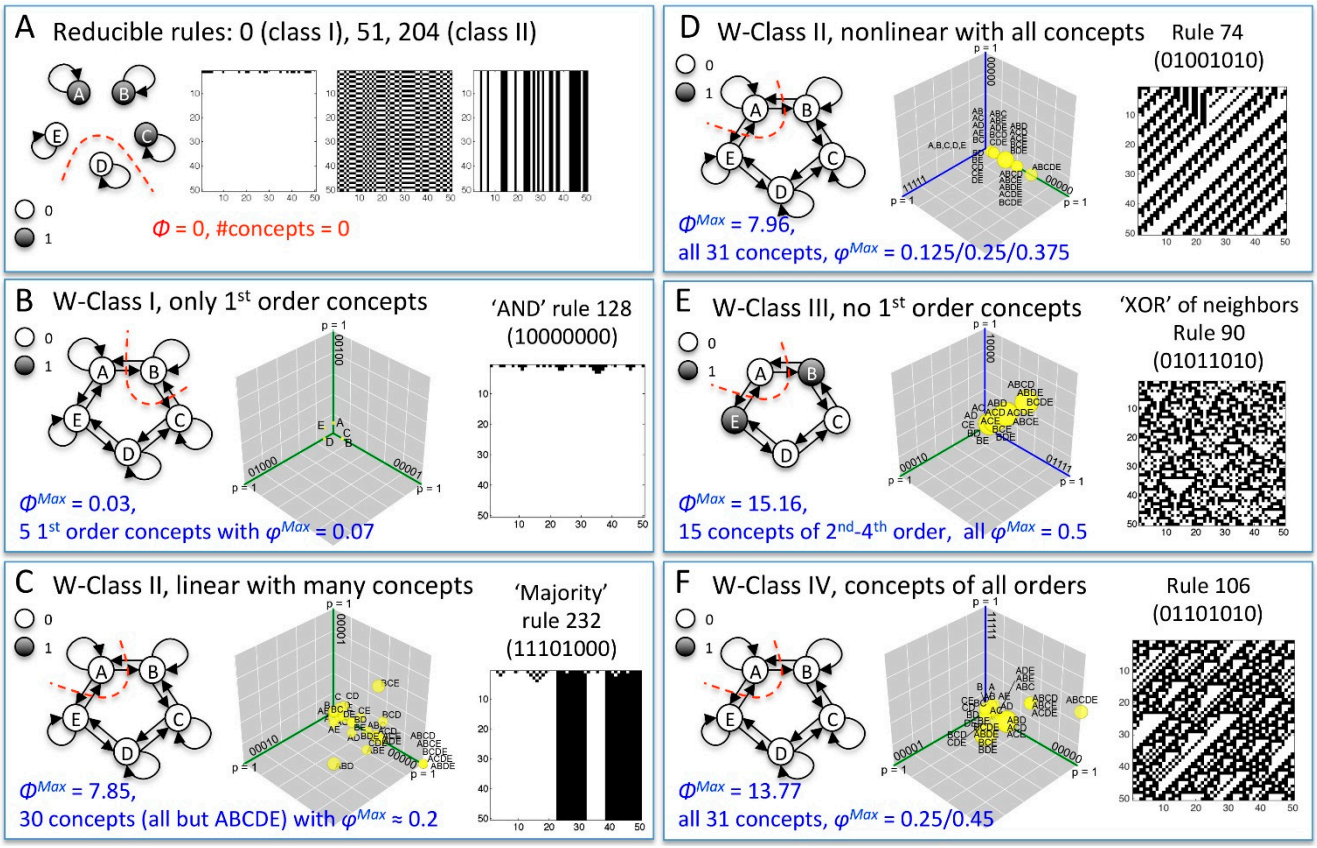

Figure 3. Examples of $N = 5$ cell ECA cause-effect structures. (**A**) Systems that implement reducible rules can be partitioned without loss of cause-effect power, as indicated by the system's network diagram (left). Note that the three reducible rules 0, 51, and 204 produce different patterns: class I, class II with periodic behavior, and class II with steady state behavior respectively; (**B**–**F**) For each example system, the effective network diagram in its current state is displayed on the left, together with the cause-effect structure in cause-effect space (middle), and a time evolution of the rule for 50 cells (right). The cause-effect structures are projected onto three dimensions of cause-effect space (see Methods). Blue axes indicate past system states, green axes futures system states. The φ^{Max} value of each concept is indicated by the size of its yellow circle. The coordinates for every concept in cause-effect space are the probabilities their cause-effect repertoire specifies for each system state; (**B**) The AND rule 128 system is a class I example for a system that does form a complex, but has only 1st order concepts. Its cause-effect structure lacks composition, leading to a low Φ^{Max} value; (**C**) Linear class II rules can have many concepts, albeit with rather low φ^{Max}; (**D**) Example for a non-linear class II rule with all concepts; (**E**) Example of the nonlinear class III rule 90, which does not have 1st order concepts, but highly irreducible concepts of higher orders and high Φ^{Max} for odd numbers of cells; (**F**) Example of a complex class IV rule with all possible concepts and high Φ^{Max}.

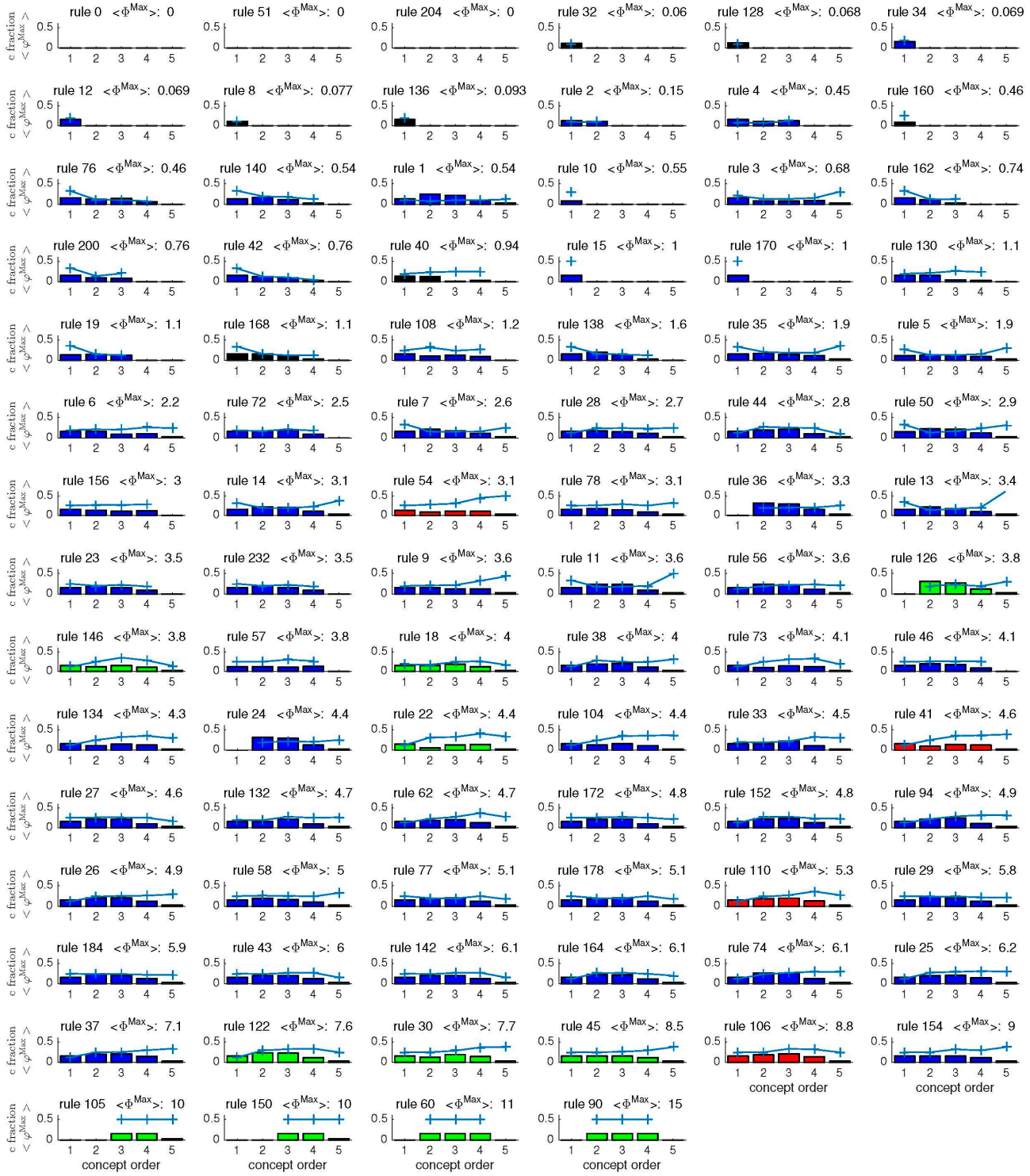

Figure 4. Distribution of <*#concepts*> and <φ^{Max}> across concept orders for $N = 5$ ECA. The 88 rule equivalency classes are ordered by their <$\boldsymbol{\Phi}^{Max}$>. For each rule, the bar plot shows the <*#concepts*> of each concept order (1st–5th) as a fraction of the maximum number of possible concepts (2^N-1 = 31 for $N = 5$). The blue line depicts <φ^{Max}> of all concepts at the respective concept order (same y axis as the fraction of concepts). Both <*#concepts*> and <φ^{Max}> are averages across N+1 system states. Bar colors indicate a rule's Wolfram class, as in Figure 2: (I) black; (II) blue; (III) green; and (IV) red. Note that there are maximally 5 possible 1st and 4th order concepts <φ^{Max}>, 10 possible 2nd and 3rd order concepts, and only one possible 5th order concept in an $N = 5$ system.

For nonlinear rules, it can be the case that elementary mechanisms by themselves have no cause-effect power ($\varphi = 0$). An example is the class III XOR rule 90 (01011010) (Figure 3E). In a rule 90 system, the fact that element A is for example in state "0" does not constrain the future state of the system at all. The states of A and C together, however, fully determine the future state of element B. Rule 90 is also interesting because $<\Phi^{Max}>$ and <#concepts> are strongly dependent on whether the number of cells N of the system is even or odd. In the case of even N, only N 2nd order concepts exist, which are composed of element pairs separated by one cell, such as AC, BD, *etc.* For odd N, as in Figure 3E, there can be additional concepts of higher order than 2.

Figure 3F shows the cause-effect structure of an $N = 5$ cell system in state "all 0" implementing rule 106, the class IV rule with the highest $<\Phi^{Max}>$ value. In this state the system specifies all possible 2^N-1 concepts. In general, for larger ECA systems certain concepts necessarily become reducible, due to the limited range of nearest neighbor interaction. Nevertheless, we found that all class IV rules have states with concepts of all possible orders, at least up to $N = 7$ cell systems. Based on this finding, we speculate that the capacity to have irreducible concepts at all orders might be a necessary condition for complex, universal class IV rules.

The examples presented in Figure 3 indicate that not only the overall number of concepts, but also the distribution of concepts across the different concept orders and their respective φ^{Max} values reflect general properties of a rule's cause-effect structure and its integrated conceptual information $<\Phi^{Max}>$. Figure 4 provides an overview of the conceptual profiles of all ECA rules for $N = 5$ systems, ordered by increasing $<\Phi^{Max}>$. Note that considering only causes and effects of individual cells (elementary, 1st order mechanisms), or of the ECA system as a whole (highest order) would not expose these differences in complexity across ECA. Our findings thus highlight the importance of assessing the causal composition of a system across all orders.

By definition, the reducible rules 0, 51, and 204 of Figure 3A with $<\Phi^{Max}> = 0$ do not have any concepts. As discussed above based on the example rule 128 (Figure 3B), several class I and II rules generally specify only 1st order concepts (e.g., rules 32, 34, 10, 15, *etc.*). Of these, rules 15 and 170 have the highest $<\Phi^{Max}> = 1$, since their 1st order concepts are very selective with $<\varphi^{Max}> = 0.5$. As shown for rule 90 (Figure 3D), some nonlinear class II and III rules (e.g., 36, 126, 105, 90, *etc.*) do not specify any 1st order concepts. In these systems only sets of elements can have cause-effect power. By contrast, all class IV rules have concepts of all orders. Overall, rules that specify cause-effect structures with higher $<\Phi^{Max}>$ tend to have more higher-order concepts and more selective concepts with high $<\varphi^{Max}>$ at all orders.

Taken together, certain general properties of the cause-effect structures of an ECA system are determined by its ECA rule and will hold for any number of cells. Moreover, the cause-effect structures of ECA systems with five and six cells suggest that a minimum number of concepts, $<\Sigma\varphi^{Max}>$, and $<\Phi^{Max}>$ may be necessary for rules to have the capacity for intricate class III and IV patterns. Nevertheless, certain rules that behaviorally lie in class II have only 1st order concepts and $<\Phi^{Max}> \leq 1$, and are thus intrinsically not more complex than class I rules. Other class II rules, however, have similar numbers of concepts and $<\Phi^{Max}>$ as rules with more complex or random behavior. This may indicate an evolutionary potential of these class II rules for complex behavior, meaning that only small changes are necessary to transform them into class III or IV rules. Class II rule 74, for example, has a relatively rich causal structure and differs in only one bit from class III rule 90 and class IV rule 106 (Figure 3D–F),

and rule 154, the class II rule with the highest $<\Phi^{Max}>$ value, has indeed been classified as (locally) chaotic by other authors [13,22].

2.1.2. Additional Causal Equivalencies

While Wolfram does not distinguish periodic from stationary long term behavior, others further subdivided Wolfram's class II into rules with stable states and rules with periodic behavior [13]. As shown above, in terms of the number of concepts, $<\Sigma\varphi^{Max}>$, and $<\Phi^{Max}>$, there is no inherent causal difference between simple periodic rules and rules with non-uniform stable states. The periodic rule 51, for example, is causally equivalent to the stationary rule 204 (Figures 3A and 4). In the same way, the Majority rule 232 (11101000), which evolves to a non-uniform stable state, is causally equivalent to rule 23 (00010111), which is periodic with cycle length 2. Rule 23 is the negation and reversion of rule 232; the same is true for rule pair 51 and 204. Analyzing the cause-effect structures of ECA here reveals additional equivalences between rules: all rules that under negation or reversion transform into the same rule are causally equivalent to their transformation (e.g., in class III 105 to 150, or in class II the equivalency classes of rules 12 and 34, *etc.*; see Figure 4, rules with identical $<\Phi^{Max}>$ and concept profiles). These additional symmetries between rules have recently been proposed in a numerical study by [23], which equates ECA rules if they show equivalent compressed state-to-state transition networks for finite ECA across different numbers of cells N. Since compressing the transition network is based on grouping states with equivalent causes and effects, the approach is related to IIT measures of cause-effect information, but lacks the notion of integration. Intrinsic causal equivalencies between rules that converge to fixed points and rules that show simple periodic behavior challenge the usefulness of a distinction based on these dynamical aspects. At least for the system itself it does not make a difference.

2.1.3. Comparison to Several Rule-Based Indicators of Dynamical Complexity

Integrated conceptual information Φ is related to several rule-based quantities that have been proposed as indicators of complex behavior during the past three decades and ultimately rely on basic notions of causal selectivity (see below). In Figure 5 we show the correlation between $<\Phi^{Max}>$, $<\#concepts>$, and $<\Sigma\varphi^{Max}>$ obtained from $N = 5$ ECA systems and three prominent measures [24], Langton's λ parameter [25], the sensitivity measure μ [26], and the Z-parameter [27]. For ECA, the λ parameter simply corresponds to the fraction of "0" or "1" outputs in the ECA rule. Rule 232 (11101000) for example has $\lambda = ½$. Class III and IV rules typically have high λ, and the rules with the highest $<\Phi^{Max}>$ values also have the maximum value of $\lambda = ½$ (Figure 5A, first panel). The parameter μ measures the sensitivity of a rule's output bit to a 1 bit change in the state of the neighborhood, counting across all possible neighborhood states [26]. Nonlinear rules that depend on the state of every cell in the neighborhood, such as the Parity rule 150 (10010110), have the highest values of μ. Finally, the Z parameter assesses the probability with which a partially known past neighborhood can be completed with certainty to the left or right side [27]. Sensitive rules with high μ also have high Z. However, Z can also be high for some simple rules, such as the Identity rule 204 (11001100), which besides has the highest $\lambda = ½$.

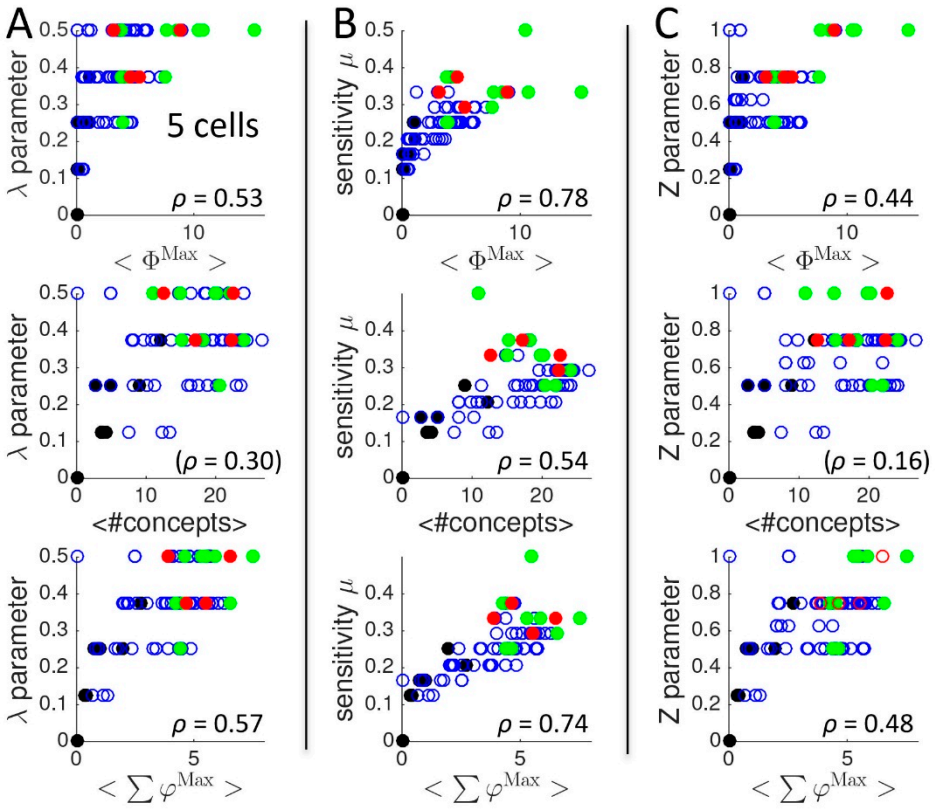

Figure 5. Relation of IIT measures to the rule based parameters λ, μ, and Z. Correlation of $<\Phi^{Max}>$, <#*concepts*>, and $<\Sigma\varphi^{Max}>$ with (**A**) Langton's λ parameter; (**B**) sensitivity μ, and (**C**) Wuensche's Z parameter (see text for details). ρ is Spearman's rank correlation coefficient, with $p < 0.001$ for all correlations, except for the λ and Z parameter *vs.* <#*concepts*>, which were not significantly correlated when corrected for multiple comparisons. Colors indicate Wolfram classes I-IV as in Figures 2 and 4: (I) uniform fixed points in black; (II) non-uniform fixed points and periodic behavior in blue; (III) random, chaotic behavior in green; and (IV) complex behavior in red.

All three rule-based quantities are related to each other and the IIT measures to some extent through the notion of causal selectivity. A mechanism is maximally selective if it is deterministic and non-degenerate, which means that its current state perfectly constrains the past and future state of the system [28]. Causal selectivity decreases with indeterminism (causal divergence) and/or degeneracy (causal convergence). While cellular automata are fully deterministic systems, many rules show degeneracy, which means they are deterministically convergent, mapping several past states into the same present state. Even in fully deterministic systems, individual mechanisms comprised of a subset of system elements typically cannot constrain the future state of the system completely, if there are degenerate mechanisms in the system: conditioning on $M_t = m_t$ in this case may still leave the remaining inputs to the elements in Z_{t+1} undetermined ("noised", *i.e.*, equally likely to be "0" or "1"). A single cell in an ECA that implements the Majority rule 232, for example, cannot completely determine its next state and that of its neighbors by itself (see Methods). Low values of λ, μ, and Z all indicate high degeneracy in the system. This means that, on average, the system's mechanisms and current states do not constrain the past and future states

of the system much. Unselective cause-effect repertoires lead to concepts with low φ^{Max}, less higher-order concepts in the system, and less integrated conceptual information Φ.

Of the three rule-based measures plotted in Figure 5, μ shows the strongest correlation with $<\Phi^{Max}>$, $<\#concepts>$, and $<\Sigma\varphi^{Max}>$. This is because, similar to the IIT measures, μ assesses the causal power of each cell in the rule neighborhood, by testing whether perturbing it makes a difference to the output. Unlike the λ and Z parameter, μ thus assigns low (but still not zero) values to rules with selective but trivially reducible causes and effects such as the Identity rule 204 or its negation rule 51, which only depend on a cell's own past value but not that of its neighbors (Figure 3A). However, compared to the IIT measures, the sensitivity parameter μ lacks the notion of causal composition, according to which higher order mechanisms can have irreducible cause-effect power. Generally, while λ, μ, and Z are largely based on empirical considerations, measures of information integration are derived from one underlying principle—intrinsic, irreducible cause-effect power.

2.1.4. Other Types of Classifications

Apart from rule-based measures and the classification of a rule's long term behavior, a CA's dynamical complexity can also be evaluated based on the morphological diversity [29] and Kolmogorov complexity [30] of its transient patterns. Morphological diversity measures the number of distinct 3 × 3 patterns in an ECA's evolution for a particular initial state. This is related to the ECA's cause-effect information for cell triplets, albeit inferred from the observed distribution rather than from the uniform distribution of all possible states. Again, $<\Phi^{Max}>$, $<\#concepts>$, and $<\Sigma\varphi^{Max}>$ correlate in a necessary, but not sufficient manner with morphological complexity, which can be low while the IIT measures are high (ρ = 0.60/0.35/0.53 Spearman's rank correlation coefficient, p <0.001/0.05/0.001 for $<\Phi^{Max}>/<\#concepts>/<\Sigma\varphi^{Max}>$ for N = 5 ECA systems) (see below Figure 6). Finally, there is also a significant correlation between the IIT measures and the Kolmogorov complexity of an ECA rule, approximated by its Block Shannon entropy (ρ = 0.65/0.40/0.59, p <0.001/0.005/0.001) or compressibility (ρ = 0.61/0.37/0.53, p < 0.001/0.05/0.001) averaged over several different initial conditions [30].

2.1.5. Being *vs.* Happening

To date, there is no universally agreed-upon classification of (elementary) cellular automata based on their dynamical behavior that uniquely assigns each rule to one class (but see [22]). Part of the problem is that depending on the initial conditions, a rule can show patterns of very different complexity. The left panel in Figure 6A, for example, displays the trivial evolution of three different rules for the typically applied initial condition of a single cell in state "1". Random initial conditions, however, reveal that only rules 0 and 128 typically tend to uniform stable states (Wolfram class I), while rule 232 belongs to Wolfram class II. Likewise, rules 2, 74, and 106 show the same simple time evolution starting from the initial condition with a single cell in state "1", but belong to different Wolfram classes: rules 2 and 74 to class II, while rule 106 belongs to class IV (Figure 6B).

Moreover, since cellular automata are deterministic, finite CA will eventually all arrive at a steady state or periodic behavior (for binary ECA after at most 2^N states). Small systems with few cells thus cannot unfold the full dynamical potential of their rule, although the local interactions of each cell are the same as in a larger system (Figure 6C). In order to predict the actual complexity of a CA's dynamical

evolution accurately, the initial state and the size of the system must be known in addition to the rule. This is why purely rule-based measures can, overall, only provide necessary but not sufficient conditions for complex behavior.

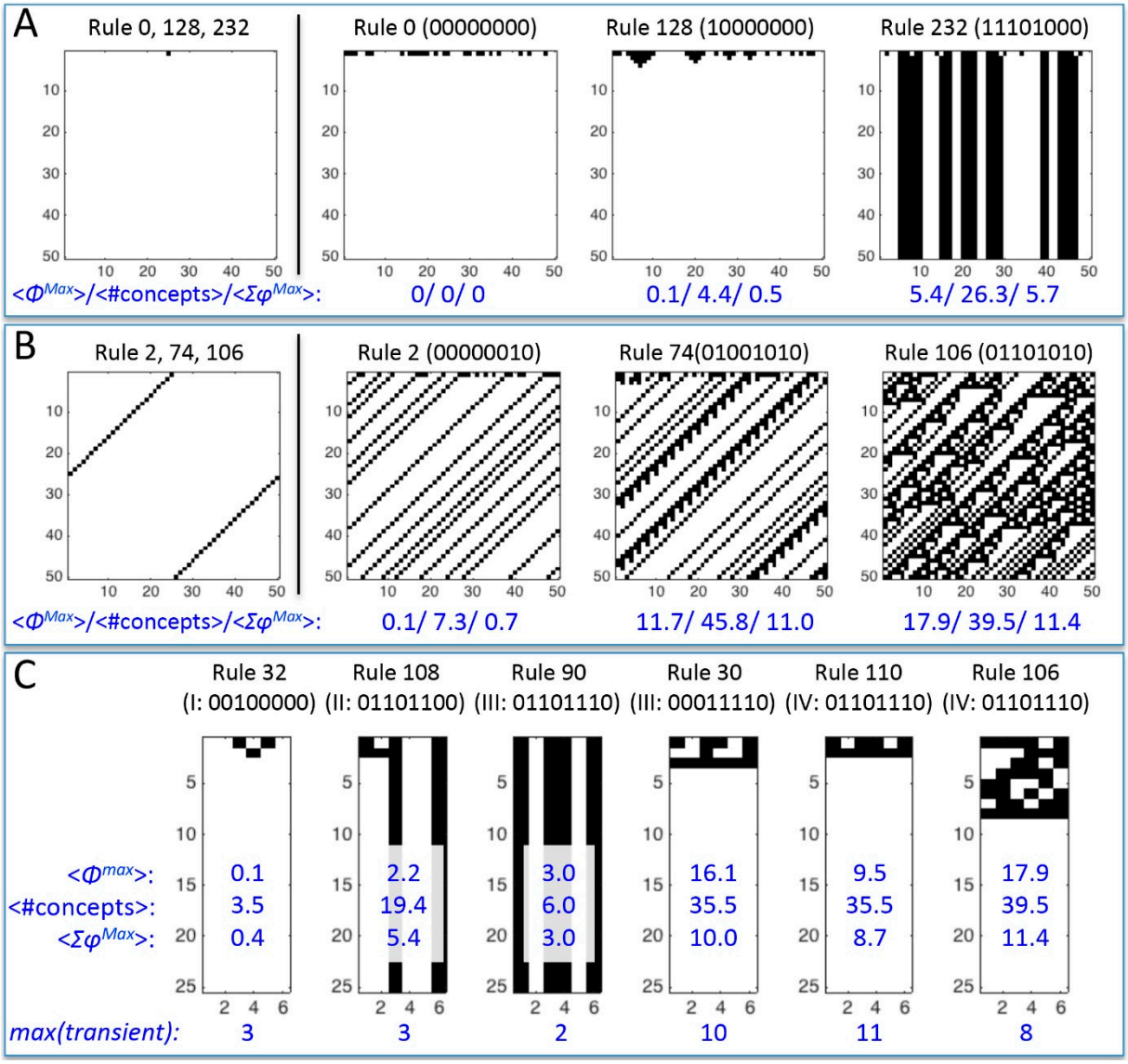

Figure 6. "Being" *vs.* "Happening". (**A**) Different rules show the same behavior under certain initial conditions. Rules 0, 128, and 232 all immediately converge to the state "all 0" for the standard initial condition of a single cell in state "1". Perturbing the system into random initial conditions, however, reveals that rules 0 and 128 belong to class I, while rule 232 belongs to class II; (**B**) Also rules 2, 74, and 106 show the same behavior for the standard initial condition, but belong to different classes under random initial conditions: rules 2 and 74 belong to class II, while rule 106 belongs to class IV; (**C**) Rules from the four different Wolfram classes all quickly converge to a steady state or periodic behavior for small ECA with $N = 6$ cells as indicated by their maximal transient length for $N = 6$ cells max*(transient)* (compare to Figures 1B and 3). The IIT measures reflect the classification ("being"), *i.e.*, the potential dynamical complexity, rather than the actual behavior ("happening"). The indicated IIT measures were obtained from the respective $N = 6$ ECA of each rule.

The problems encountered in classifying rules based on their dynamical complexity stems from a discrepancy between "being" and "happening" [18]. Patterns describe what is *happening* in the CA system following a particular initial state. A classification, however, is about what the system *is*. As the examples in Figure 6 show, classifying a rule for arbitrary initial conditions reveals its *potential* dynamical complexity, rather than the states actually observed in a particular dynamic evolution. Since it requires perturbing the system into many different initial states, this approach is somewhat related to the causal approach of IIT. More generally, the objective of IIT is to reveal and characterize how, and how much, a system of mechanisms in its current state exists ("is") from its own intrinsic perspective, rather than from the perspective of an external observer observing its temporal evolution. Intrinsic cause-effect power relies on causal composition and requires irreducibility [17,18]. According to IIT, the particular way the system exists, is given by its cause-effect structure, and to what extent it exists as a system, is given by its irreducibility, quantified as its integrated conceptual information Φ^{Max}.

While IIT measures do depend on the size of the system and its state, the average values obtained for ECA with a small number of cells and for a subset of states already reveal the general causal characteristics underlying different rules as demonstrated by the examples shown in Figure 3. In Figure 7, the average IIT measures of several rules from all Wolfram classes are plotted against N, the number of cells. Rules with the capacity for complex, or random behavior show relatively high values of $<\Phi^{Max}>$ already for small systems with 3–7 nodes. The way $<\Phi^{Max}>$, $<\#concepts>$, and $<\Sigma\varphi^{Max}>$ increase with increasing system size is also characteristic of a rule's cause-effect structure.

Class I rules 32 and 128 for example have only 1st-order concepts; no combination of system elements can have irreducible cause-effect power ($\varphi = 0$) (see also Figure 3B). Their maximum number of concepts thus increases linearly, while their $<\Phi^{Max}>$ values stay low, since no matter where the system is partitioned, only one concept is lost. By contrast, rules with higher order concepts at each subset size, such as rule 30, 106, or 110, show an almost exponential increase of $<\#concepts>$, $<\Sigma\varphi^{Max}>$ and $<\Phi^{Max}>$. The maximum number of potential concepts in any system is determined by the powerset of 2^N-1 candidate mechanisms (combinations of elements) in the system (indicated by the thin black dashed line in Figure 7B). While $<\#concepts>$ and $<\Phi^{Max}>$ will continue to grow for rules like rule 30, 106, and 110, the number of impossible concepts also increases, because ECA are limited to nearest neighbor interactions. In an $N = 6$ ECA system A-F, for example, the concepts AD, BE, and CF are impossible, since they are reducible by default with $\varphi = 0$. The reason is that, since their elements do not share any inputs or outputs, partitioning between them would thus never make any difference. Finally, as described above, for certain rules, the IIT measures depend on the parity of N, as for the XOR rule 90 (Figure 3E).

In many cases, the temporal evolution of CA is computationally irreducible [16], which makes it impossible to predict their dynamical behavior. Similarly, calculating the cause-effect structure of a system becomes computationally intractable already for a small number of elements. On the other hand, IIT measures can in principle be extended to higher dimensional CA with more than two states and larger neighborhoods. As Figure 7A–C show, general features of a rule's cause-effect structure can already be inferred from very small CA systems.

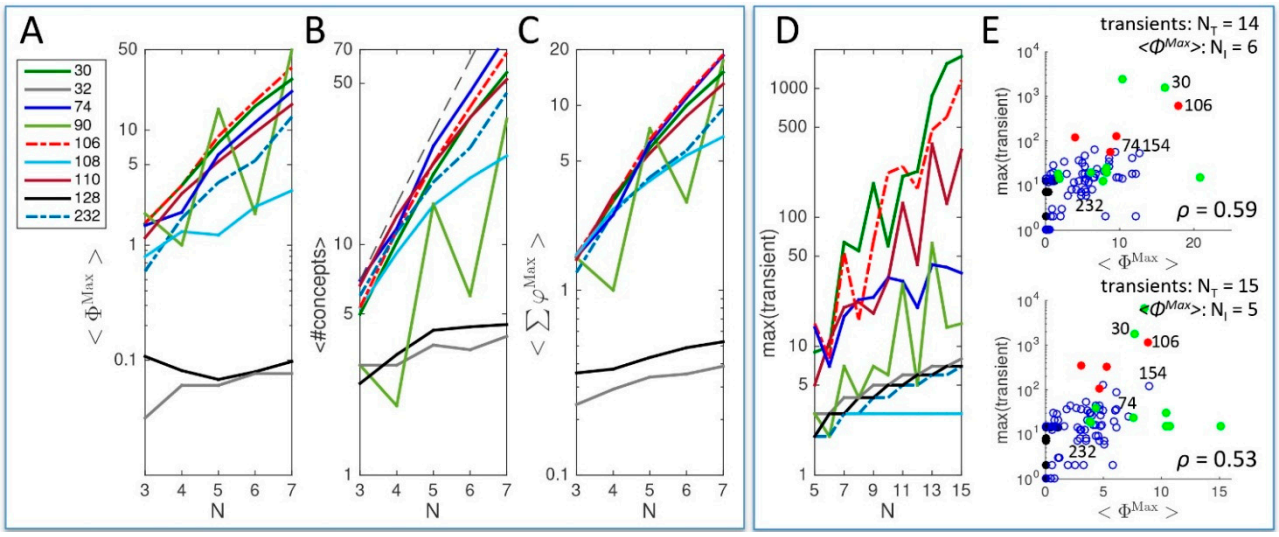

Figure 7. (**A**) $<\Phi^{Max}>$; (**B**) $<\#concepts>$; and (**C**) $<\Sigma\varphi^{Max}>$ across $N = 3$–7 cells. Already small systems reveal typical features of an ECA's cause-effect structure. How the IIT measures of an ECA rule depend on N is further indicative of the general properties of the rule's cause-effect structure (see main text); (**D**) Maximal transient length of the respective rules evaluated across all system states plotted against the number of cells N. Class I rules and class II rules with lower $<\Phi^{Max}>$ values (32, 108, 128, and 232) tend to have shorter transients as well; (**E**) Maximal transient length of all ECA equivalency classes plotted against $<\Phi^{Max}>$. The top panel shows the correlation of $N_T = 14$ ECA transients to the $<\Phi^{Max}>$ values obtained from the respective $N_I = 6$ ECA systems. The lower panel shows the correlation of $N_T = 15$ ECA transients to the $<\Phi^{Max}>$ values obtained from the respective $N_I = 5$ ECA systems. The labels indicate some of the example rules discussed above. (**A**–**E**) Note the logarithmic scale of the y-axis in all panels. Colors denote Wolfram classes as in Figures 2, 4, and 5.

Finite ECA necessarily converge to a steady state or periodic cycle after at most 2^N time steps. In Figure 7D,E we show the maximal transient length of ECA rules across all initial states of small $N = 5$–15 cell systems. While Wolfram's ECA classification is based on the long-term behavior of a rule for most initial states, the maximal transient length of some class I and II rules in these small systems can be of the same order as the maximal transient length of other class III or IV rules (compare e.g., rule 74 to rule 90 in Figure 7D), indicating a potential for complex dynamics under some initial conditions. In Figure 6E the maximum transient lengths of $N = 14$ and $N = 15$ cell systems are plotted against $<\Phi^{Max}>$ of their respective rules in $N = 6$ and $N = 5$ systems. Since the maximal transient length as well as $<\Phi^{Max}>$ of certain rules depends on the parity of N (see Figure 7A,D), we evaluated the correlation coefficients between the largest even/odd numbered systems for which computing maximal transient lengths ($N_T = 14/15$) and IIT measures ($N_I = 6/5$) was feasible for all rules. Similar results were obtained comparing all other pairs of system sizes. While the maximal transient length also correlated with $<\#concepts>$ and $<\Sigma\varphi^{Max}>$ ($\rho = 0.49/0.56$, $p < 0.001$ for $N_T = 14/N_I = 6$ and $\rho = 0.40/0.51$, $p < 0.01/0.001$ for $N_T = 15/N_I = 5$), it was most strongly correlated with the integrated conceptual information $<\Phi^{Max}>$ (Figure 7E, $\rho = 0.59/0.53$, $p < 0.001$ for both $N_T = 14/N_I = 6$ and $N_T = 15/N_I = 5$), indicating that irreducibility is a relevant factor for the dynamical complexity of a system.

In summary, the cause-effect structure of a system and its Φ describe what a system "is" from its intrinsic perspective, and thereby reveal a rule's potential for dynamical complexity. While we found strong relations between the IIT measures of ECAs and their Wolfram classes, as well as strong correlations with the ECAs' maximal transient lengths, having many concepts and high $<\Phi^{Max}>$ are not sufficient conditions for a system to actually show complex behavior under every initial condition. Nevertheless, employing IIT measures of causal complexity can significantly reduce the search space for complex rules, since they appear to be necessary for class III and IV behavior. Finally, class II rules with high $<\Phi^{Max}>$ and many concepts tend to exhibit long transients for some initial conditions and typically share certain rule-based features with more complex class III and IV rules, which can be interpreted as a high potential for complex behavior under small adaptive changes in an evolutionary context.

2.2. Behavior and Cause-Effect Power of Adapting Animats

Cellular automata are typically considered as isolated systems. In this section, we examine the cause-effect structures of small, adaptive logic-gate systems ("animats"), which are conceptually similar to discrete, deterministic cellular automata. By contrast to typical CA, however, the animats are equipped with two sensor and two motor elements, which allow them to interact with their environment (Figure 8A). Moreover, the connections between an animat's elements, as well as the update rules of its four hidden elements and two motors are evolved through mutation and selection within a particular task-environment over several 10,000 generations (Figure 8B). We demonstrated above, that isolated ECA require a sufficiently complex, integrated cause-effect structure for complex global dynamics. Given sufficiently complex inputs from the environment, the animats, however, can in principle exhibit complex dynamics even if their internal structure is causally trivial and/or reducible (e.g., unconnected COPY gates of the sensors). Consequently, the dynamic behavior and intrinsic cause-effect structures of these non-isolated systems may be dissociated. Nevertheless, in adaptive systems, evolution to an environment with a rich causal structure provides a link between the system's dynamics and its intrinsic cause-effect structures.

In the following, we review evidence from [19], which shows that, under constraints on the number of internal elements, environments that require context-sensitivity and memory favor the evolution of integrated systems with rich cause-effect structures. Building on these data, we show that, although the animats are very small systems, their average transient lengths in isolation from the environment tend to correlate with $<\Phi^{Max}>$, as observed in the isolated ECA systems. Finally, we discuss how the adaptive advantages of integrated animats, such as their higher economy in terms of mechanisms per element, and larger degeneracy in architecture and function, are related to the animats' behavioral and dynamical repertoire.

An animated example of animat evolution and task simulation can be found on the IIT website [31]. The task environments the animats were exposed to are variants of "Active Categorical Perception" (ACP) tasks, where moving blocks of different sizes have to be distinguished [19,32,33]. Solving the ACP tasks successfully requires combining different sensory inputs and past experience. Adaptation is measured as an increase in fitness: the percentage of correctly classified blocks ("catch" or "avoid").

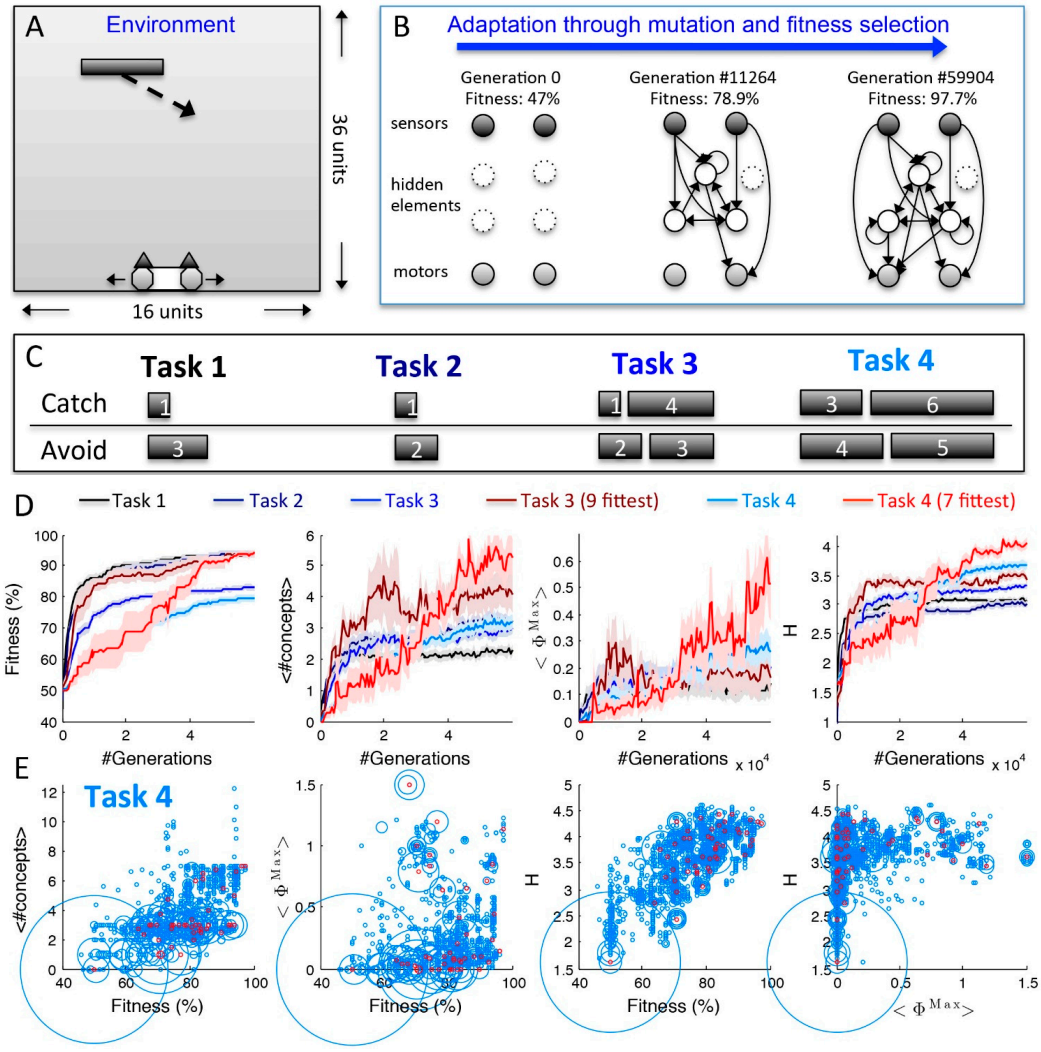

Figure 8. Fitness, <*#concepts*>, <Φ^{Max}>, and state entropy *H*, of animats adapting to four task environments with varying requirements for internal memory. (**A**) Schematic of animat in example environment. On each trial, the animat has to recognize the size of the downward moving block and either catch or avoid it. Blocks continuously move downward and either to the right or left, at a speed of one unit per time step (periodic boundary conditions). The animat has two sensors with a space of 1 unit between them and thus a total width of three units. Its two motors can move it one unit to the left or right, respectively; (**B**) Animat evolution. Each animat is initialized at generation 0 without connections between elements. Through mutation and fitness selection, the animats develop complex network structures with mechanisms that enable them to solve their task. Animats were let to evolve for 60,000 generations; (**C**) Illustration of the four Task environments with increasing difficulty and requirements for internal memory from left to right; (**D**) The final fitness achieved by the animats after 60,000 generations corresponds to the task difficulty. The two red traces show data from a subset of Task 3 and Task 4 trials with the same high average fitness as Task 1 and 2. Animats that evolved to the more difficult tasks, particularly Task 4, developed significantly more concepts, higher <Φ^{Max}> values, and more state entropy *H*, than those animats that evolved to Task 1. Shaded areas around curves denote SEM across 50 independent evolutions (LODs); (**E**) Scatter plots of all evaluated generations of animats

from all 50 LODs of Task 4 illustrating the relation of <*#concepts*>, <Φ^{Max}>, and H to fitness, and H to <Φ^{Max}>. The circle size is proportional to the number of animats with the same pair of values. Red dots denote the final generation of animats from all 50 independent evolutions. Panels A, C, and D were adapted with permission from [19].

An animat's behavior is deterministically guided by the sensory stimuli it receives from the environment. An animat sensor turns on if a block is located vertically above it; otherwise it is off. The hidden and motor elements are binary Markov variables, whose value is specified by a deterministic input-output logic. However, an animat's reaction to a specific sensor configuration is context-dependent, in the sense that it also depends on the current state of the animat's hidden elements, which can be considered as memories of previous sensor and hidden element configurations. In [19], we evaluated the cause-effect structure and integrated conceptual information Φ of animats evolved to four task environments that differed primarily in their requirements for internal memory (Figure 8C). For each task environment, we simulated 50 independent evolutions with 100 animats at each generation. The probability of an animat to be selected into the next generation was proportional to an exponential measure of the animat's fitness (roulette wheel selection) [19,33]. At the end of each evolution, the line of descent (LOD) of one animat from the final generation was traced back through all generations and the cause-effect structures of its ancestors were evaluated every 512 generations. For a particular animat generation in one LOD, the IIT measures were evaluated across all network states experienced by the animat during the 128 test trials, weighted by their probability of occurrence. As shown in Figure 8D, adapted from [19], in the more difficult task environments that required more internal memory to be solved (Task 3 and particularly Task 4), the animats developed overall more concepts and higher <Φ^{Max}> than in the simpler task environments (Task 1 and 2). This is even more evident when the tasks are compared at the same level of fitness (red lines in Figure 8D). Note that <*#concepts*> shown in Figure 8 was evaluated across all of the animat's elements including sensors and motors in order to capture all fitness relevant causal functions, while Φ^{Max} is the integrated conceptual information of the set of elements that forms the major complex (MC) in an animat's "brain" (values for the number of MC concepts behave similarly and can be found in [19]).

As an indicator for the dynamical repertoire (dynamical complexity) of the animats in their respective environments, we measured the state entropy $H = \sum_s p_s \log(p_s)$ of the animats' hidden and motor elements for the different task environments, displayed in the right panel of Figure 8D. The animats' state entropy increases with adaptation across generations, and also with task difficulty across the different task environments, similar to the IIT measures. The maximum possible entropy for six binary elements is $H = 6$, if all system states have equal probability to occur. Note that the animats are initialized without connections between elements and elements without inputs cannot change their state. During adaptation, the number of connected elements increases, particularly in the more difficult tasks that require more memory. More internal elements mean a greater capacity for memory, entropy, and also a higher number of concepts and integrated conceptual information. In this way, fitness, dynamical complexity, and causal complexity are tied together, particularly if the requirement for internal memory is high, even though, in an arbitrary, non-isolated system, the state entropy H could be dissociated from the system's cause-effect structure. This relation is illustrated in Figure 8E, where the <*#concepts*>, <Φ^{Max}>, and the state entropy H are plotted against fitness for every animat of all 50 LODs of Task 4.

All three measures are positively correlated with fitness ($\rho = 0.80/0.79/0.54$ Spearman's rank correlation coefficient for H/<#concepts>/$<\Phi^{Max}>$ with $p < 0.001$). Note that animats from the same LOD are related. The red dots in Figure 8E highlight the final generation of each LOD, which are independent of each other. Taking only the final generation into account, H and <#concepts> still correlate significantly with fitness. However, the correlation for $<\Phi^{Max}>$ is not significant after correcting for multiple comparisons ($\rho = 0.63/0.56$ for H/<#concepts> with $p < 0.001$), since having more $<\Phi^{Max}>$ even at lower fitness levels has no cost for the animats.

In contrast to the state entropy H, the entropy of the sensor states H_{Sen} is mostly task dependent: during adaptation H_{Sen} increases only slightly for Tasks 3 and 4 and decreases slightly for Tasks 1 and 2 (see Figure S4 of [19]). The entropy of the motor states H_{Mot} represents the behavioral repertoire (behavioral complexity) of the animats and is included in H. H_{Mot} increases during adaptation, but asymptotes at similar values (~1.6) for all tasks. This reflects the fact that the behavioral requirements ("catch" and "avoid") are similar in all task environments (see Figure S4 of [19]).

More elements allow for a higher capacity for state entropy H and also higher $<\Phi^{Max}>$. Nevertheless, H is also directly related to $<\Phi^{Max}>$, since the highest level of entropy for a fixed number of elements is achieved if, for each element, the probability to be in state "0" or "1" is balanced. As we saw above for elementary cellular automata, balanced rules that output "0" or "1" with equal probability are more likely to achieve high values of $<\Phi^{Max}>$ (Figure 5B, λ parameter). This is because mechanisms with balanced cause-effect repertoires have on average higher φ values and lead to more higher-order concepts, and thus cause-effect structures with higher $<\Phi^{Max}>$. Likewise, as shown for Task 4 in Figure 8E, right panel, animats with high $<\Phi^{Max}>$ also have high entropy H ($\rho = 0.66$, $p < 0.001$; taking only the final generation into account the correlation is still almost significant after correcting for multiple comparisons with $\rho = 0.44$, $p = 0.053$).

In the last section, we noted that for isolated ECA systems, having a certain level of $<\Phi^{Max}>$ and $<\#concepts>$ is necessary in order to have the potential for complex dynamics, and thus high state entropy. The animats, however, receive sensory inputs that can drive their internal dynamics. Consequently, also animats with modular, mainly feedforward structures ($\Phi = 0$) can have high state entropy H while they are behaving in their world. Keeping the sensory inputs constant, animats converge to steady states or periodic dynamics of small cycle length within at most seven time-steps. The average length of these transients, measured for the final generation of animats of all 50 LODs, tends to correlate with the average $<\Phi^{Max}>$ calculated from all states experienced during the 128 test trials especially in the simpler tasks 1 and 2 ($\rho = 0.45/0.46/0.43/0.39$ Spearman's rank correlation coefficient for Task 1–4 with $p = 0.04/0.03/0.067/0.19$ after correcting for multiple comparisons). Interestingly, there is no correlation between the transient length and the animats' fitness. This is because, in general, high fitness only requires a rich behavioral repertoire while interacting with the world, but not in isolation.

In addition to the state entropy, in [19] we also assessed how the sensory-motor mutual information (I_{SMMI}) [34] and predictive information (I_{Pred}) [35] of the animats as defined in [19,36] evolved during adaptation. I_{SMMI} measures the differentiation of the observed input-output behavior of the animats' sensors and motors. I_{Pred}, the mutual information between observed past and future system states, measures the differentiation of the observed internal states of the animats' hidden and motor elements. Both, high I_{SMMI} and high I_{Pred}, should be advantageous during adaptation to a complex environment, since they reflect the animats' behavioral and dynamical repertoire, in particular how deterministically

one state leads to another. I_{Pred} in the animats is indeed closely tied to the state entropy: it increases during adaptation with increasing fitness and a higher number of internal elements. I_{SMMI}, however, may actually decrease during adaptation in the animats, since an increase in internal memory may reduce the correlation between sensors and motors, which are restricted to two each (see Figure S4 of [19]). Both I_{SMMI} and I_{Pred}, are correlational measures, which depend on the observed distributions of system states. By contrast, analyzing the cause-effect structure of a system requires system perturbations that reveal the causal properties of the system's mechanisms under all possible initial states. The cause-effect structure thus takes the entire set of possible circumstances the animat might be exposed to into account and not just those observed in a given setting. As for cellular automata, an animat's cause-effect structures, evaluated by its $<\#concepts>$ and $<\Phi^{Max}>$ quantify its intrinsic causal complexity and its dynamical *potential*.

Under external constraints on the number of available internal elements, having many concepts and high integrated conceptual information Φ proved advantageous for animats in more complex environments (Figure 8D and [19]). While the simpler Tasks 1 and 2 could be solved (100% fitness) by animats with either integrated ($\Phi > 0$) or modular ($\Phi = 0$) network architectures, only animats with integrated networks reached high levels of fitness in the more difficult Tasks 3 and particularly Task 4, which required more internal computations and memory [19]. This is because integrated systems can implement more functions (concepts) for the same number of elements, since they can make use of higher-order concepts—irreducible mechanisms specified by combinations of elements.

When particular concepts are selected for during adaptation, higher-order concepts become available at no extra cost in terms of elements or wiring. This degeneracy in concepts may prove beneficial to respond to novel events and challenges in changing environments. Degeneracy here refers to different structures that perform the same function in a certain context [37,38]. Contrary to redundant structures, degenerate structures can diverge in function under different contexts. Animats with integrated networks with many degenerate concepts may already be equipped to master novel situations. In principle, this allows them to adapt faster to unpredicted changes in the environment than animats with modular structures, which first have to expand and rearrange their mechanisms and connectivity [39].

In the context of changing environments, large behavioral and dynamical repertoires are advantageous not only at the level of individual organisms, but also at the population level. In [19] we found that the variety of network connectomes, mechanisms, and distinct behaviors was much higher among animats that solved Task 1 and 2 perfectly with integrated network structures ($<\Phi^{Max}> > 0$, high degeneracy) than among animats with the same perfect fitness, but $<\Phi^{Max}> = 0$ (low degeneracy). In Task 1, for example, integrated solutions were encountered in six out of 50 lines of descent (LODs); modular solutions in seven out of 50 LODs. Nevertheless, analyzing all animats with perfect fitness across all generations and LODs, animats with $<\Phi^{Max}> > 0$ showed 332 different behavioral strategies, while animats with $<\Phi^{Max}> = 0$ only produced 44 different behavioral strategies. The reason is that integrated networks are more flexible and allow for neutral mutations that do not lead to a decrease in fitness. By contrast, modular networks showed very little variability once a solution was encountered. Having more potential solutions should give a probabilistic selective advantage to integrated networks, and should also lead to more heterogeneous populations, which provide an additional advantage in the face of environmental change.

Taken together, in causally rich environments that foster memory and sensitivity to context, integrated systems should have an adaptive advantage over modular systems. This is because under naturalistic constraints on time, energy, and substrates, integrated systems can pack more mechanisms for a given number of elements, exhibit higher degeneracy in function and architecture, and demonstrate greater sensitivity to context and adaptability. These prominent features of integrated systems also link intrinsic cause-effect power to behavioral and dynamical complexity at the level of individuals and populations.

3. Conclusions

One hallmark of living systems is that they typically show a wide range of interesting behaviors, far away from thermodynamic equilibrium (e.g., [40]). How living systems change their states in response to their environment can be seen as a form of natural computation [41]. Among the oldest model systems for the study of natural computation are small discrete dynamical systems, cellular automata [10,42]. CA have revealed that complex dynamical patterns can emerge from simple, local interactions of small homogeneous building blocks. It is thus not surprising that the extensive body of research dedicated to CA focused mainly on the systems' dynamical properties, investigating what is "happening" during their temporal evolution. While the scientists studying cellular automata may observe intriguing patterns computed by the system, it has been pointed out that these patterns have no relevance for the CA itself in the absence of some kind of "global self-referential mechanism" [43]. Integrated information theory (IIT) provides a framework for establishing precisely to what extent a system "makes a difference" to itself, from its own intrinsic perspective. The cause-effect structure of a system and its integrated conceptual information Φ characterize what a system "is"—how much and in which way it exists for itself, independent of an external observer—rather than what a system happens to be "doing". Consequently, even inactive systems, or systems in a steady state that do not appear to be "doing" anything from the extrinsic perspective, can nevertheless specify rich cause-effect structures [17] from their own intrinsic perspective. The cause-effect structure of a system can be taken as the causal foundation for the system's dynamic behavior, which may or may not manifest itself under the observed circumstances. For purposes of dynamical analysis, evaluating cause-effect structures may help to identify systems that are candidates for complex dynamic behavior.

4. Methods

In the following we illustrate the main principles and measures invoked by integrated information theory (IIT) by reference to a simple, elementary cellular automaton (ECA) with six cells and periodic boundary conditions implementing rule 232, the Majority rule (Figure 9A). Table 1 provides an overview over the general mathematical expressions of all relevant IIT measures. See [17] for more details. All IIT measures can be derived from the system's one time-step transition probability matrix (TPM). To obtain the TPM, we perturb the system into all possible states with equal probability and observe the resulting state distribution at the next time step (see Supplementary Methods).

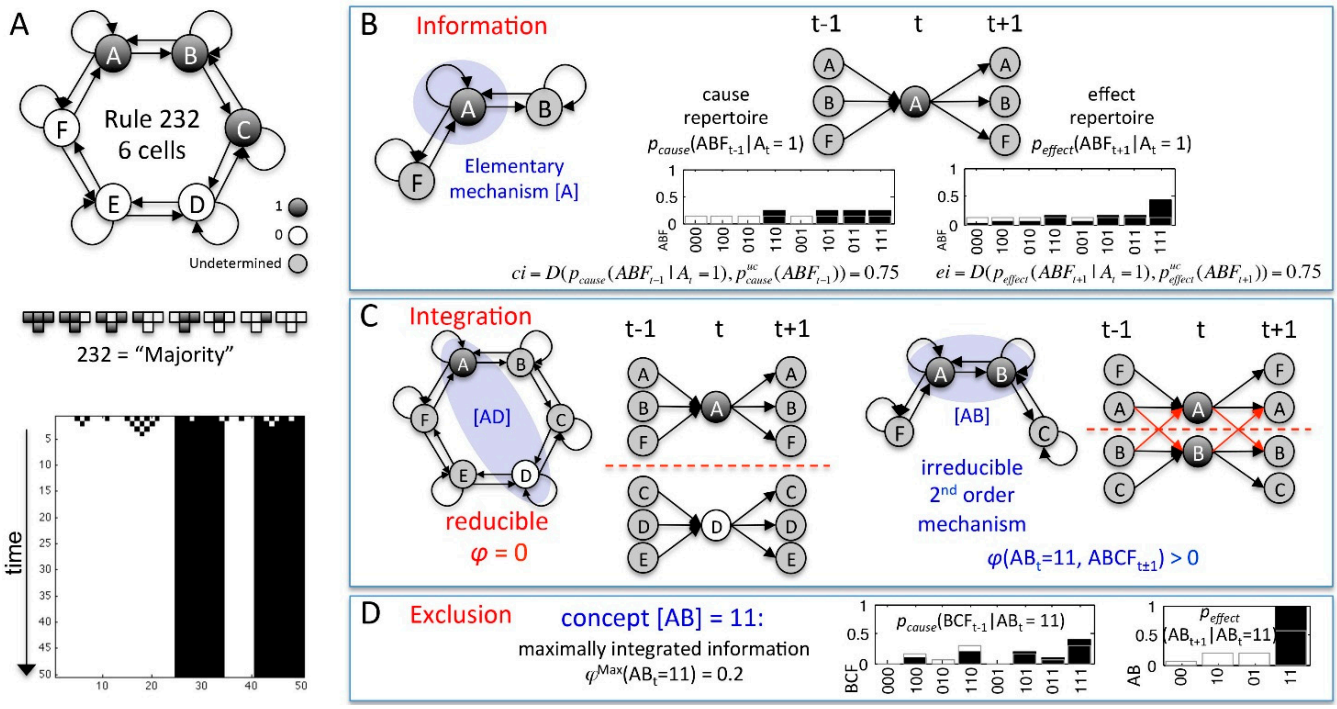

Figure 9. Information, integration, and exclusion postulate at the level of mechanisms. (**A**) ECA 232 with $N = 6$ cells A-F illustrated as a network of interacting elements in the current state A-F_t = 111000. All cells implement rule 232, the Majority rule; (**B**) Information: Element A_t in state 1 constrains the past and future states of $ABF_{t\pm1}$ compared to their unconstrained (flat) distribution (overlaid in gray); (**C**) Integration: Elements A and D do not form a higher order mechanism, since AD_t is reducible to its component mechanisms A_t and D_t. AB_t in state 11, however, does form a higher order mechanism, since AB_t specifies both irreducible causes and irreducible effects, as evaluated by integrated information φ (see text); (**D**) Exclusion: Cause and effect repertoires are evaluated over all sets of system elements. The cause and effect repertoire that are maximally irreducible and their φ^{Max} value form the concept of a mechanism. In case of AB, the maximally irreducible cause-repertoire is over elements BCF_{t-1}; the maximally irreducible effect-repertoire is over elements AB_{t+1}.

4.1. Mechanisms and Concepts

To determine whether a set of elements M_t in state m_t forms a mechanism with cause-effect power on the system, we first need to evaluate whether (i) the past state of the system makes a difference to the set M_t and (ii) the state of the set $M_t = m_t$ makes a difference to the future state of the system. Condition (i) is satisfied if M_t, by being in its current state m_t, constrains the probability distribution of possible past states of a set of system elements Z_{t-1}, specified by the cause repertoire $p_{cause}(z_{t-1}|m_t)$. The cause-repertoire of element A_t in state "1" of the ECA shown in Figure 9A over its inputs ABF_{t-1}, for example, reflects A's update rule: only states in which two or more of A's inputs are "1" are possible past states of $A_t = 1$ (see Supplementary Methods for details on the calculation of the cause repertoire). The *cause information (ci)* of $A_t = 1$ quantifies the distance between its cause repertoire and the unconstrained (uc) cause repertoire in which each past state of ABF_{t-1} is assumed equally likely (see Figure 9B).

$$ci(A_t = 1, ABF_{t-1}) = D\big(p_{cause}(ABF_{t-1}|A_t = 1), p^{uc}_{cause}(ABF_{t-1})\big) = 0.75 \quad (4)$$

Distances in IIT are measured by the so-called earth mover's distance (emd) [44,45]. The emd quantifies the cost of transforming one probability distribution into another, using the Hamming distance between states as the underlying metric. (For the cause repertoire of Figure 9B, for example, $p = 1/8$ has to be moved from each of the four states 110, 101, 011, and 111 to the remaining states, twice over a Hamming distance of 1 (e.g., 101→100) and twice over a Hamming distance of 2 (e.g., 110→ 000), resulting in a value of $2 \times 2 \times 1/8+2 \times 1 \times 1/8=0.75$).

Condition (ii) is satisfied if M_t, by being in its current state m_t, constrains the probability distribution of possible future states of a set of system elements Z_{t+1}, specified by the effect repertoire $p_{effect}(z_{t+1}|m_t)$. Figure 9B shows the effect repertoire of $A_t = 1$ over its outputs ABF_{t+1} (see Supplementary Methods for details on the calculation of the effect repertoire). The *effect information (ei)* of $A_t = 1$ quantifies the distance between its effect repertoire and the unconstrained effect repertoire, which considers all input states to all system elements to be equally likely.

$$ei(A_t = 1, ABF_{t+1}) = D\big(p_{effect}(ABF_{t+1}|A_t = 1), p^{uc}_{effect}(ABF_{t+1})\big) = 0.75 \quad (5)$$

Note that, although the ECA is deterministic, $A_t = 1$ by itself cannot perfectly constrain its outputs ABF_{t+1}. Since the state of the other inputs to ABF_{t+1} is unconstrained (perturbed into "0" and "1" with equal probability), the effect of $A_t = 1$ by itself is merely to increase the probability of its output elements to be in state "1" at t + 1 compared to the unconstrained distribution, as specified by the effect repertoire shown in Figure 9B.

In order to have cause-effect power, the *cause-effect information (cei)* of M_t, the minimum of its *ci* and *ei*, must be positive. Here, for $A_t = 1$:

$$cei(A_t = 1, ABF_{t\pm1}) = \min\big(ci(A_t = 1, ABF_{t-1}), ei(A_t = 1, ABF_{t+1})\big) = 0.75 \quad (6)$$

Even a set of elements with positive *cei* does not have cause-effect power of its own, and is thus not a mechanism, if it is reducible to its sub-mechanisms. Consider, for example, elements A and D in the ECA system of Figure 9. A_t and D_t together do not constrain the past or future states of the system more than A_t and D_t taken separately. Irreducibility is tested by partitioning connections, which means making connections causally ineffective, and can be thought of as "injecting independent noise" into the connections. Partitioning between A_t and D_t and their respective inputs and outputs does not make a difference to AD_t's cause or effect repertoire ($\varphi = 0$). The elements A and B together, however, do form a higher-order mechanism AB_t, since no matter how the mechanism's connections are partitioned, AB_t's cause-effect repertoire is changed. To assess the irreducibility of AB_t, over $ABCF_{t-1}$ we first calculate the *integrated cause information* φ_{cause} for all possible partitions of AB_t and $ABCF_{t-1}$, as the distance between the intact cause repertoire and the product cause repertoire of the partition. In this way, we find the partition that makes the least difference to the cause repertoire, the minimum information partition $MIP_{cause} = \underset{P}{\operatorname{argmin}}\big(\varphi_{cause}(AB_t = 11, ABCF_{t-1}, P)\big)$. Here:

$$\begin{aligned} &\varphi_{cause}(AB_t = 11, ABCF_{t-1}, MIP_{cause}) \\ &= D\big(p_{cause}(ABCF_{t-1}|AB_t = 11), (p_{cause}(C_{t-1}|\emptyset) \times (p_{cause}(ABF_{t-1}|AB_t = 11)\big) = 0.2 \end{aligned} \quad (7)$$

where ∅ denotes the empty set (a partition with $p(\emptyset|\emptyset)$, however, is not a valid partition). Likewise, we calculate the *integrated effect information* φ_{effect} for all possible partitions to find $MIP_{effect} = \underset{P}{\operatorname{argmin}}\left(\varphi_{effect}(AB_t = 11, ABCF_{t+1}, P)\right)$. Here:

$$\begin{aligned}\varphi_{effect}(AB_t = 11, ABCF_{t+1}, MIP_{effect}) \\ = D\big(p_{effect}(ABCF_{t+1}|AB_t = 11), (p_{effect}(C_{t+1}|\emptyset) \\ \times (p_{effect}(ABF_{t+1}|AB_t = 11)\big) = 0.25\end{aligned} \quad (8)$$

An irreducible set of elements $M_t = m_t$ with positive *integrated information* $\varphi(m_t, Z_{t\pm1}, MIP)$, the minimum of φ_{cause} and φ_{effect}, is a mechanism that has cause-effect power on the system. Note that, depending on whether $\varphi = \varphi_{cause}$ or $\varphi = \varphi_{effect}$, M_t's overall $MIP = \underset{P}{\operatorname{argmin}}\left(\varphi(m_t, Z_{t\pm1}, P)\right)$ is either MIP_{cause} or MIP_{effect}. The set of M_t's cause and effect repertoire describe how M_t constrains the past and future of the system elements $Z_{t\pm1}$ and is termed "cause-effect repertoire" $CER(m_t, Z_{t\pm1})$.

Finally, φ_{cause} and φ_{effect} can be measured for cause-effect repertoires over all possible sets of system elements $Z_{t\pm1}$. The cause-effect repertoire of $M_t = m_t$ that is maximally irreducible over the sets of elements $Z^*_{t\pm1}$, with $max(\varphi_{cause})$ and $max(\varphi_{effect})$ (see Table 1), and its integrated information φ^{Max} form M_t's "concept". It can be thought of as the causal role of the mechanism in its current state from the intrinsic perspective of the system itself. The concept of AB in state "11" in the ECA system, for example, is specified by the cause repertoire over the set of elements BCF and the effect repertoire over the set of elements AB, with $\varphi^{Max} = \min(\varphi^{Max}_{cause}, \varphi^{Max}_{effect}) = 0.2$.

$$\begin{aligned}\varphi^{Max}_{cause}(AB_t = 11)) \\ = D\big(p_{cause}(BCF_{t-1}|AB_t = 11), (p_{cause}(C_{t-1}|\emptyset) \times (p_{cause}(BF_{t-1}|B_t = 1)\big) = 0.2\end{aligned} \quad (9)$$

$$\begin{aligned}\varphi^{Max}_{effect}(AB_t = 11)) \\ = D\big(p_{effect}(AB_{t+1}|AB_t = 11), (p_{effect}(\emptyset|A_t = 1) \times (p_{effect}(AB_{t+1}|B_t = 1)\big) = 0.5\end{aligned} \quad (10)$$

For comparison, the cause-effect repertoire over all inputs and outputs of AB is less irreducible with $\varphi = 0.17$ (see Equations (7) and (8)) and thus not further considered.

4.2. Cause-Effect Structures

As stated above, to have intrinsic cause-effect power, the system must have elements with cause-effect power upon the system. In other words, the system must have concepts. The set of concepts of all mechanisms within a system shapes its "cause-effect structure". The 6-node, rule 232 ECA A-F_t in state "111000" for example has 14 concepts, six elementary and eight higher-order concepts (Figure 10A).

As with a mechanism, a system only exists as a whole from the intrinsic perspective if it is irreducible to its parts. Specifically, every subset of the system must have cause-effect power upon the rest of the system, otherwise there may be system elements that never affect the system or are never affected by the system. This is tested by unidirectionally partitioning the connections between every subset of the system and its complement (see Table 1). *Integrated conceptual information* Φ quantifies the distance between the cause-effect structure $C(s_t)$ of the intact system $S_t = s_t$ and the cause-effect structure of the partitioned system $C(s_t, P_{\rightarrow})$. Again, we test all possible partitions $P_{\rightarrow}$ in order to find the minimum information partition (*MIP*), the partition that makes the least difference to the cause-effect structure of S_t. The

distance between two cause-effect structures is evaluated using an extended version of the earth mover's distance (emd). It quantifies the cost of transforming one cause-effect structure into another, taking into account how much the cause-effect repertoires and φ^{Max} values of all concepts change through the partition, see [17].

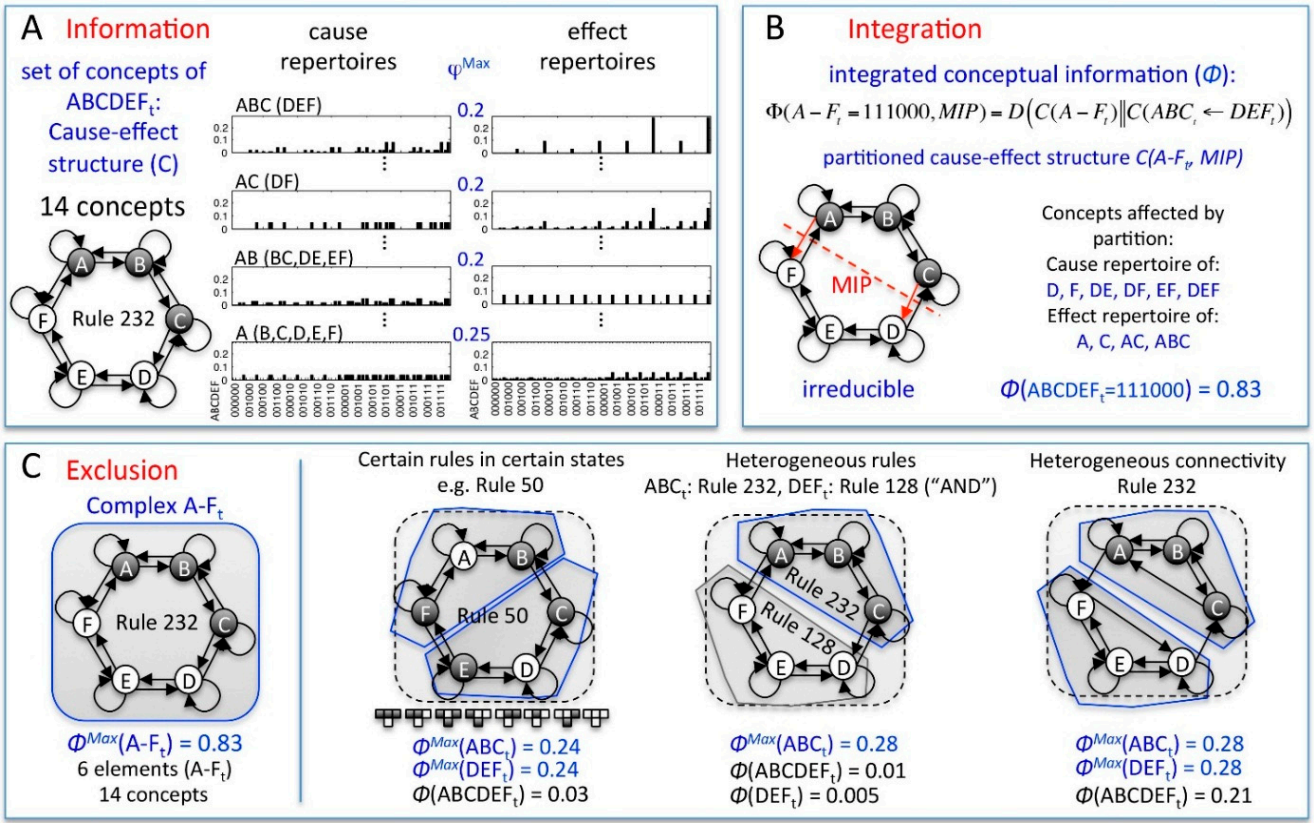

Figure 10. Information, integration, and exclusion postulate at the level of systems of mechanisms. (**A**) Information: The system A-F$_t$ in state "111000" has 14 concepts: six elementary concepts A$_t$,B$_t$,C$_t$,D$_t$,E$_t$ and F$_t$, four 2nd order concepts of adjacent elements, as well as AC$_t$ and DF$_t$, and the 3rd order concepts ABC$_t$ and DEF$_t$; (**B**) Integration: the cause-effect structures of A-F$_t$ is irreducible to the cause-effect structure of its minimum information partition (*MIP*), which eliminates the effects of subset ABC$_t$ onto subset DEF$_t$, as measured by the integrated conceptual information Φ, which quantifies the distance between the whole and partitioned cause-effect structure; (**C**) Exclusion: Cause-effect structures and their Φ values are evaluated over all sets of system elements. The set of elements S_t^* with the maximally irreducible cause-effect structure forms a complex; here it is the whole system A-F$_t$. This is not always the case. For different types of rules, or more heterogeneous connections or rules, system subsets may be maximally irreducible complexes. Complexes cannot overlap. In the three examples on the right, DEF$_t$ thus may form another complex beneath the main complex ABC$_t$; the whole system, however, is excluded from being a complex, as are subsets of elements within a complex.

In the example system A-F$_t$ the *MIP* renders the connections from subset ABC to DEF causally ineffective. Although this partition is the one that least affects the cause-effect structure of A-F$_t$, it still

alters the cause-effect repertoires of many concepts (Figure 9B). Only the concepts $B_t = 1$, $E_t = 0$, $AB_t = 11$, and $BC_t = 11$ remain intact. A-F_t's cause-effect structure is thus irreducible with:

$$\Phi(A - F_t = 111000, MIP) = D\big(C(A - F_t)\big|C(ABC_t \leftarrow DEF_t)\big) = 0.83. \quad (1)$$

Even if a system is irreducible, there may be subsets (or supersets, if the system were embedded in a larger set of elements) that also specify irreducible cause-effect structures. When subsets of elements of a larger system are considered, Φ is calculated for the set, treating the other elements as fixed background conditions. To avoid causal over-determination (counting the cause-effect power of the same mechanism multiple times) and thereby the multiplication of "entities" beyond necessity (Occam's razor), only one of all overlapping irreducible cause-effect structures is postulated to exist—the one that is maximally irreducible with the highest Φ value (Φ^{Max}). A set of elements that has Φ^{Max} and is thus maximally irreducible is called a "complex". In our example, the whole set A-F_t is a complex, since it has the highest Φ value compared to all its subsets. This is not always the case: given different ECA rules, systems with heterogeneous rules, or systems with slightly different connectivity, smaller sets may have the most irreducible cause-effect structure (Figure 10C). Once the major complex with Φ^{Max} is identified, non-overlapping sets of elements can form additional, minor complexes, as shown for the system with heterogeneous rules (Figure 10C, middle). The whole system A-F_t, however, is excluded from being a complex in this example, since it overlaps with the major complex ABC_t.

The cause-effect structure of a system can be illustrated as a constellation in cause-effect space. In cause-effect space, every axis is a past or future state of the system resulting in a total of 2×2^N axes in a system with N cells and binary elements. Each system concept corresponds to a point in this high-dimensional space. The coordinates are given by the probabilities specified in the concept's cause-effect repertoire for each past and future system state. The size of each point corresponds to the concept's φ^{Max} value. To plot cause-effect structures in cause-effect space, we project the high dimensional structure onto the three past or future states (blue or green, respectively) for which the concepts' cause-effect repertories have the highest variance (see Figure 3C for a $N = 5$ cell Rule 232 system in state "all 0"). In this way, the selectivity and distribution of concepts in the state space of their system can be visualized in a straightforward manner. On the IIT webpage [20] all example cause-effect structures of this article can be viewed in 3D.

In sum, IIT postulates that a set of elements exists for itself to the extent that it has maximally irreducible cause-effect power, evaluated by Φ^{Max}. In order to have high cause-effect power, it is necessary to have many internal mechanisms, which form a rich cause-effect structure with many elementary and higher order concepts.

4.3. Transient Lengths

Maximal transient lengths of ECA reported in Figures 6C and 7D,E were obtained from ECA systems of N cells with periodic boundary conditions, taking all possible system states into account. To that end, the ECA were started in a particular system state and evolved in time until a previous state was repeated. In finite ECA systems, which are discrete, deterministic dynamical systems, this must happen after at most 2^N time steps, corresponding to the number of possible states of the system. Transient lengths for

the animats were obtained in the same way, for all possible sensor states, keeping the sensor inputs to the animat constant while the internal states and motors evolved in time.

4.4. Statistics and IIT Code

All correlation coefficients ρ presented in this study are Spearman rank correlation coefficients. The *p*-values were corrected for multiple comparisons by multiplying the *p*-value from every single comparison with the number of evaluated coefficients, $N_{\text{coeff}} = 40$.

Custom-made MATLAB software was used for data analysis. The PyPhi package used to calculate all IIT measures is available at [46]. All examples presented in this paper can also be viewed and recalculated online at the IIT website [20]. EMD calculations within the IIT program were performed using the open source fast code of Pele and Werman [45].

Acknowledgments

We thank Will Mayner for help with the Python IIT software (PyPhi) and William Marshall for helpful discussions. This work has been supported by the Templeton World Charities Foundation (Grant #TWCF 0067/AB41).

Author Contributions

L.A. and G.T. conceived and designed the experiments; L.A. performed the experiments and analyzed the data; L.A. wrote the paper and G.T. edited it.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Nykamp, D. Dynamical System Definition. Available online: http://mathinsight.org/dynamical_system_definition (accessed on 15 May 2015).
2. Ermentrout, G.B.; Edelstein-Keshet, L. Cellular automata approaches to biological modeling. *J. Theor. Biol.* **1993**, *160*, 97–133.
3. De Jong, H. Modeling and simulation of genetic regulatory systems: A literature review. *J. Comput. Biol.* **2002**, *9*, 67–103.
4. Nowak, M.A. *Evolutionary Dynamics*; Harvard University Press: Cambridge, MA, USA, 2006.
5. Sumpter, D.J.T. The principles of collective animal behaviour. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **2006**, *361*, 5–22.
6. Kaneko, K. *Life: An Introduction to Complex Systems Biology*; Springer: Berlin/Heidelberg, Germany, 2006.
7. Izhikevich, E.M. *Dynamical Systems in Neuroscience*; MIT Press: Cambridge, MA, USA, 2007.
8. Wolfram, S. Universality and complexity in cellular automata. *Physica D* **1984**, *10*, 1–35.
9. Wolfram, S. *A New Kind of Science*; Wolfram: Champaign, IL, USA, 2002; Volume 5.

10. Von Neumann, J. Theory of self-reproducing automata. In *Essays on Cellular Automata*; University of Illinois Press: Champaign, IL, USA, 1966.
11. Gardner, M. Mathematical games: The fantastic combinations of John Conway's new solitaire game "life." *Sci. Am.* **1970**, *223*, 120–123.
12. Knoester, D.B.; Goldsby, H.J.; Adami, C. Leveraging Evolutionary Search to Discover Self-Adaptive and Self-Organizing Cellular Automata. **2014**, arXiv:1405.4322.
13. Li, W.; Packard, N. The structure of the elementary cellular automata rule space. *Complex Syst.* **1990**, *4*, 281–297.
14. Culik, K., II; Yu, S. Undecidability of CA classification schemes. *Complex Syst.* **1988**, *2*, 177–190.
15. Sutner, K. On the Computational Complexity of Finite Cellular Automata. *J. Comput. Syst. Sci.* **1995**, *50*, 87–97.
16. Wolfram, S. Undecidability and intractability in theoretical physics. *Phys. Rev. Lett.* **1985**, *54*, 735–738.
17. Oizumi, M.; Albantakis, L.; Tononi, G. From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0. *PLoS Comput. Biol.* **2014**, *10*, e1003588.
18. Tononi, G. Integrated information theory. *Scholarpedia* **2015**, *10*, 4164.
19. Albantakis, L.; Hintze, A.; Koch, C.; Adami, C.; Tononi, G. Evolution of Integrated Causal Structures in Animats Exposed to Environments of Increasing Complexity. *PLoS Comput. Biol.* **2014**, *10*, e1003966.
20. Online IIT Calculation. Available online: http://integratedinformationtheory.org/calculate.html (accessed on 9 April 2015).
21. Chua, L.O.; Yoon, S.; Dogaru, R. A Nonlinear Dynamics Perspective of Wolfram's New Kind of Science Part I: Threshold of Complexity. *Int. J. Bifurc. Chaos* **2002**, *12*, 2655–2766.
22. Schüle, M.; Stoop, R. A full computation-relevant topological dynamics classification of elementary cellular automata. *Chaos* **2012**, *22*, 043143.
23. Krawczyk, M.J. New aspects of symmetry of elementary cellular automata. **2013**, arXiv:1304.5771.
24. De Oliveira, G.M.B.; de Oliveira, P.P.B.; Omar, N. Guidelines for dynamics-based parameterization of one-dimensional cellular automata rule spaces. *Complexity* **2000**, *6*, 63–71.
25. Langton, C.G. Computation at the edge of chaos: Phase transitions and emergent computation. *Physica D* **1990**, *42*, 12–37.
26. Binder, P. A phase diagram for elementary cellular automata. *Complex Syst.* **1993**, *7*, 241–247.
27. Wuensche, A.; Lesser, M. *The Global Dynamics of Cellular Automata: An Atlas of Basin of Attraction Fields of One-dimensional Cellular Automata*; Santa Fe Institute Studies in the Sciences of Complexity: Reference Volume I; Addison-Wesley: Boston, MA, USA, 1992.
28. Hoel, E.P.; Albantakis, L.; Tononi, G. Quantifying causal emergence shows that macro can beat micro. *PNAS* **2013**, *110*, 19790–19795.
29. Adamatzky, A.; Martinez, G.J. On generative morphological diversity of elementary cellular automata. *Kybernetes* **2010**, *39*, 72–82.
30. Zenil, H.; Villarreal-Zapata, E. Asymptotic Behaviour and Ratios of Complexity in Cellular Automata. *Int. J. Bifurc. Chaos* **2013**, *23*, 1350159.
31. Online Animat Animation. Available online: http://integratedinformationtheory.org/animats.html (accessed on 5 May 2015).

32. Beer, R.D. The Dynamics of Active Categorical Perception in an Evolved Model Agent. *Adapt. Behav.* **2003**, *11*, 209–243.
33. Marstaller, L.; Hintze, A.; Adami, C. The evolution of representation in simple cognitive networks. *Neural Comput.* **2013**, *25*, 2079–2107.
34. Ay, N.; Bertschinger, N.; Der, R.; Güttler, F.; Olbrich, E. Predictive information and explorative behavior of autonomous robots. *Eur. Phys. J. B* **2008**, *63*, 329–339.
35. Bialek, W.; Nemenman, I.; Tishby, N. Predictability, complexity, and learning. *Neural Comput.* **2001**, *13*, 2409–2463.
36. Edlund, J.A.; Chaumont, N.; Hintze, A.; Koch, C.; Tononi, G.; Adami, C. Integrated information increases with fitness in the evolution of animats. *PLoS Comput. Biol.* **2011**, *7*, e1002236.
37. Tononi, G.; Sporns, O.; Edelman, G.M. Measures of degeneracy and redundancy in biological networks. *Proc. Natl. Acad. Sci. USA* **1999**, *96*, 3257–3262.
38. Edelman, G. *Neural Darwinism: The Theory of Neuronal Group Selection*; Basic Books: New York, NY, USA, 1987.
39. Albantakis, L.; Tononi, G. Advantages of integrated cause-effect structures in changing environments; An *in silico* study on evolving animats. **2015**, in preparation.
40. Schrödinger, E. *What is Life?: With Mind and Matter and Autobiographical Sketches*; Cambridge University Press: Cambridge, UK, 1992.
41. Still, S.; Sivak, D.A.; Bell, A.J.; Crooks, G.E. Thermodynamics of Prediction. *Phys. Rev. Lett.* **2012**, *109*, 120604.
42. Kari, J. Theory of cellular automata: A survey. *Theor. Comput. Sci.* **2005**, *334*, 3–33.
43. Pavlic, T.; Adams, A.; Davies, P.; Walker, S. Self-Referencing Cellular Automata: A Model of the Evolution of Information Control in Biological Systems. In *Artificial Life 14: Proceedings of the Fourteenth International Conference on the Synthesis and Simulation of Living Systems*; The MIT Press: Cambridge, MA, USA, 2014; Volume 14, pp. 522–529.
44. Rubner, Y.; Tomasi, C.; Guibas, L. The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Vis.* **2000**, *40*, 99–121.
45. Pele, O.; Werman, M. Fast and Robust Earth Mover's Distances. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009.
46. PyPhi Package. Available online: https://github.com/wmayner/pyphi (accessed on 5 May 2015).