Local measures of information storage in complex distributed computation

Joseph T. Lizier,[1, 2, 3, *] Mikhail Prokopenko,[2] and Albert Y. Zomaya[3]
[1]*Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, 04103 Leipzig, Germany*
[2]*CSIRO Information and Communications Technology Centre, PO Box 76, Epping, NSW 1710, Australia*
[3]*School of Information Technologies, The University of Sydney, NSW 2006, Australia*

(Dated: March 12, 2012)

Information storage is a key component of intrinsic distributed computation. Despite the existence of appropriate measures for it (e.g. excess entropy), its role in interacting with information transfer and modification to give rise to distributed computation is not yet well-established. We explore how to quantify information storage on a local scale in space and time, so as to understand its role in the dynamics of distributed computation. To assist these explorations, we introduce the active information storage, which quantifies the information storage component that is directly in use in the computation of the next state of a process. We present the first profiles of local excess entropy and local active information storage in cellular automata, providing evidence that blinkers and background domains are dominant information storage processes in these systems. This application also demonstrates the manner in which these two measures of information storage are distinct but complementary. It also reveals other information storage phenomena, including the misinformative nature of local storage when information transfer dominates the computation, and demonstrates that the local entropy rate is a useful spatiotemporal filter for information transfer structure.

PACS numbers: 89.75.Fb, 89.75.Kd, 89.70.Cf, 05.65.+b
Keywords: information storage, intrinsic computation, complex systems, information theory, cellular automata, particles

I. INTRODUCTION

Information storage is considered an important aspect of the dynamics of many natural and man-made processes, for example: in human brain networks [22, 46] and artificial neural networks [4], synchronisation between coupled systems [5, 42], coordinated motion in autonomous [1] and modular robots [43], attractor dynamics of cellular automata [37], and in the dynamics of inter-event distribution times [13]. Despite the existence of suitable quantitative measures for information storage (e.g. the excess entropy [8]), the term is still often used rather loosely or in a qualitative sense. A major factor here is that there is not yet a well-established quantitative understanding of how information storage dynamically interacts with information transfer and modification to give rise to intrinsic distributed computation in multivariate systems.

In this paper we explore methods to quantify information storage in distributed computation. In particular, we focus on how information storage can be quantified on a *local scale* in space-time, which allows us to directly investigate the role of information storage in the *dynamics* of distributed computation.

We focus on cellular automata (CAs) which, as described in Section II, are a popular model of distributed computation in which the notion of information storage is qualitatively well-understood. Indeed, the emergent structures in CAs known as "blinkers" have been conjectured to implement information storage, and we hypothesise that appropriate quantification of the dynamics of information storage will align with this conjecture.

We define the concept of information storage as *the information in an agent, process or variable's past that can be used to predict its future.* We consider such storage in the *intrinsic computation* of the unfolding of a system's dynamics [11], which takes place regardless of whether it is explicitly interpretable as a computation to an observer. We describe in Section IV how *total* storage *relevant to the future* of a process is captured by the existing measure *statistical complexity* [9], while the *total* storage *actually used* in the future of a process is captured by the existing measure *excess entropy* [8]. We then introduce active information storage in Section V to capture the amount of storage that is *currently in use* by a process. Our perspective of distributed computation is important, providing the perspective that information can not only be stored internally by an agent or element implementing the computation, but also stored in its environment for later retrieval.

We present the first application of local profiles of the excess entropy and the active information storage to cellular automata in Section VI. As hypothesised above, these applications provide the first quantitative evidence that blinkers are the dominant information storage entities there. This result is significant in marrying these quantitative measures of information storage with the popularly-understood qualitative notion of its embodiment in distributed computation. These measures also reveal other important local information storage phenomena, including the misinformative nature of storage when information transfer dominates the computation. Our application also demonstrates the manner in which

[*] lizier@mis.mpg.de

these two measures of information storage are distinct but complementary, revealing different aspects of the information dynamics. Finally, we demonstrate that the entropy rate, a complementary measure to the active information storage, is a useful filter for viewing moving particle structures in CAs.

II. CELLULAR AUTOMATA

Information storage has been a topic of interest in the context of distributed computation in *cellular automata* (CAs). As we will describe, the notion of information storage is qualitatively well-understood with regard to CAs, and as such we choose these as our application domain.

CAs are discrete dynamical lattice systems involving an array of cells which synchronously update their states as a homogeneous deterministic function (or rule) of the states of their local neighbours [51]. Here we will use Elementary CAs (ECAs), which consist of a one-dimensional array of cells with binary states, with each updated as a function of the previous states of themselves and one neighbour either side (i.e. neighbourhood size 3 or range $r = 1$). CAs are widely used to study complex computation, since certain rules (e.g. ECA rules 110 and 54, defined using the Wolfram numbering scheme [51]) exhibit *emergent coherent structures* which are not discernible from their microscopic update functions but which provide the basis for understanding the macroscopic computations being carried out [39, 52].

These emergent structures are known as *particles*, *gliders* and *domains*. A domain may be understood as a set of background configurations in a CA, any of which will update to another such configuration in the absence of a disturbance. The most simple domain types involve periodic repetition of cell states in time and space. Domains are formally defined within the framework of computational mechanics [18] as spatial process languages in the CA. *Particles* are qualitatively considered to be moving elements of coherent spatiotemporal structure, in contrast to or against a background domain. *Gliders* are particles which repeat periodically in time while moving spatially, while stationary gliders are known as *blinkers*. Formally, particles are defined as a boundary between two domains [18]; as such, they can also be termed as *domain walls*, though this is typically used with reference to aperiodic particles (e.g. those in rule 18).

There are several long-held conjectures regarding the role of these emergent structures in the intrinsic distributed computation in CAs; i.e. how the cells process information in order to determine their collective future state [25, 39]. Blinkers are generally held to be the dominant information storage elements, since local pattern maintenance is an information storage process. In contrast, particles are held to be the dominant information transfer entities, since they communicate coherent information about the dynamics in one area of the system to another (indeed, we have provided the first direct quantitative evidence for this conjecture with a measure of local information transfer in [31]). Studies of the density classification task with rule ϕ_{par} [39–41] help our intuition here. They suggest a human-understandable computation, with stationary blinkers used to store information about the local density of "1"'s in one region of the CA, while moving gliders are used to transfer information about these local densities between regions.

The presence of such emergent structure in CAs is revealed by *filtering* techniques, which highlight particles against the background domain. Early methods were hand-crafted for specific CAs (relying on the user knowing the pattern of background domains) [14, 17], while later methods can be automatically applied to any given CA. These include: finite state transducers to recognise the regular spatial language of the CA using ϵ-machines [18, 19]; local information (i.e. local spatial entropy rate) [20]; the display of executing rules with the most frequently occurring rules filtered out [52]; and local statistical complexity (via a light-cone formulation) and local sensitivity [45].

Certainly, filtering is not a new concept, however the ability to *separately* filter different computational features is novel. For example, we have previously used local information *transfer* to filter *moving* particle structures in CAs [31], and a measure of local information *modification* to filter particle *collisions* in [32]. Here, we *hypothesise* that local measures of information *storage* should be useful filters for blinker structures: this would provide the first quantitative evidence for their computational role as dominant information storage structures in CAs. Together with our work regarding local information transfer [31] and information modification [32], the investigation of local information storage will show how these operations interrelate to give rise to complex behaviour, in comparison to other filters which give only a single view of where that complexity occurs. This would allow a more refined investigation than single measures, and should reveal interesting differences in the parts of the emergent structures that are highlighted. Furthermore, this set of measures will be generally applicable to any multivariate time-series, unlike some of the filtering measures above (e.g. spatial ϵ-machines [18, 19] and spatial entropy rate [20]) which are only applicable to lattice systems.

Similarly, we note the suggestion of alternative filters to ours for stored, transferred and modified information in CAs in [12], using redundant, unique and synergistic terms from the Partial Information Decomposition framework [48]. However while this framework certainly provides interesting insights into the sub-components of information storage and transfer (e.g. in [49]), and while these filters reveal unique insights into the dynamics of CAs, these filters are not readily interpretable as information storage and transfer measures. In particular regarding information storage, the use of spatial redundancy measures will only capture certain types of storage dynamics associated with *spatial* periodicities, e.g. it will not capture individualistic storage behaviour where each variable in a distributed system is exhibiting independent behaviour. This issue is underlined in that while these filters work well for the simple CA rule 184, the results for the more complex rule 54 do not cleanly separate for example gliders from the background; more importantly, the results for this rule are simply not consistent with the popular understanding of blinkers and domains as information storage and gliders as information transfer.

III. INFORMATION-THEORETIC PRELIMINARIES

Information theory [36] is the natural domain to describe information storage in distributed computation, and indeed information theory is proving to be a useful framework for the analysis and design of complex systems. The fundamental quantity is the (Shannon) *entropy*, which represents the uncertainty in a sample x of a random variable X: $H_X = -\sum_x p(x) \log_2 p(x)$ (all with units in bits). The *joint entropy* of two random variables X and Y is a generalization to quantify the uncertainty of their joint distribution: $H_{X,Y} = -\sum_{x,y} p(x,y) \log_2 p(x,y)$. The *conditional entropy* of X given Y is the average uncertainty that remains about x when y is known: $H_{X|Y} = -\sum_{x,y} p(x,y) \log_2 p(x|y)$. The *mutual information* between X and Y measures the average reduction in uncertainty about x that results from learning the value of y, or vice versa:

$$I_{X;Y} = H_X - H_{X|Y}. \tag{1}$$

The *conditional mutual information* between X and Y given Z is the mutual information between X and Y when Z is known: $I_{X;Y|Z} = H_{X|Z} - H_{X|Y,Z}$.

Finally, the *entropy rate* is the limiting value[1] of the conditional entropy of the next state of X (i.e. measurements x_{n+1} of the random variable X') given knowledge of the previous k states of X (i.e. measurements $x_n^{(k)} = \{x_{n-k+1}, \dots, x_{n-1}, x_n\}$, up to and including time step n, of the random variable $X^{(k)}$):

$$H_{\mu X} = \lim_{k\to\infty} H_{X'|X^{(k)}} = \lim_{k\to\infty} H_{\mu X}(k), \tag{2}$$

$$H_{\mu X}(k) = H_{X^{(k+1)}} - H_{X^{(k)}}. \tag{3}$$

Note that $H_{\mu X}(k)$ denotes a finite-k estimate of $H_{\mu X}$.

We note that the above information-theoretic measures compute *expectation values* of each quantity over all possible configurations of the considered variables, or equivalently are *averages* over each observation. As such, we can consider the *local* or *pointwise* values of each quantity at each time step or observation. As an illustrative example of local measures, consider that the (average) entropy H_X defines the average uncertainty or information in X; the local entropy $h_X(n+1)$ quantifies the amount of information contained in a *specific observation* or realization x_{n+1} of the variable X at time step $n+1$ of the process:

$$h_X(n+1) = h(x_{n+1}) = -\log_2 p(x_{n+1}). \tag{4}$$

For a lattice system with spatially-ordered variables, we can write the local entropy for the value $x_{i,n+1}$ at time step $n+1$ of variable i as:

$$h(i, n+1) = h(x_{i,n+1}) = -\log_2 p(x_{i,n+1}). \tag{5}$$

As pointed out above, the average entropy is the expectation value of the local entropy; as such we have:

$$H_X = \langle h_X(n) \rangle_n, \tag{6}$$

$$H(i) = \langle h(i,n) \rangle_n, \tag{7}$$

and in homogeneous systems $H = \langle h(i,n) \rangle_{i,n}$.

Similarly, the *local temporal entropy rate* estimate is represented as:

$$h_{\mu X}(n+1) = \lim_{k\to\infty} h_{\mu X}(n+1,k), \tag{8}$$

$$h_{\mu X}(n+1,k) = h(x_{n+1} \mid x_n^{(k)}) = -\log_2 p(x_{n+1} \mid x_n^{(k)}). \tag{9}$$

In local lattice notation we can write the estimates as:

$$h_\mu(i, n+1, k) = h(x_{i,n+1} \mid x_{i,n}^{(k)}) = -\log_2 p(x_{i,n+1} \mid x_{i,n}^{(k)}). \tag{10}$$

Again, we have $H_{\mu X}(k) = \langle h_{\mu X}(n,k) \rangle_n$, $H_\mu(i,k) = \langle h_\mu(i,n,k) \rangle_n$ and in homogeneous systems $H_\mu(k) = \langle h_\mu(i,n,k) \rangle_{i,n}$. Importantly, since $p(x_{n+1} \mid x_n^{(k)}) \leq 1$ the local temporal entropy rate $h_{\mu X}(n+1,k) \geq 0$. As per Appendix A, the local entropy rate can only be shown to converge (i.e. reach a limiting value as $k \to \infty$) under more restrictive conditions[2] than the known conditions (e.g. see [7]) for convergence of the average entropy rate.

Importantly, local or pointwise values are directly attributable to the dynamics at each time step, in comparison to *sliding window averaging* techniques which provide a measure more localised than averages over long time periods but still blur out precise temporal dynamics. An additional problem occurs when, in an effort to try to keep the sliding window average well-localised, the observation set size is kept small (e.g. six samples for the large multivariate sample space in [12]) leading to undersampling.[3]

[1] We note that while these limiting values exist (i.e. the average entropy rate converges) as $k \to \infty$ for stationary processes (e.g. see [7]), there is no guarantee that such limits exist for non-stationary processes.

[2] I.e. convergence occurs for processes with finite order Markovian dependence on their past. Until less restrictive conditions can be established (if possible), one must be careful in using the local quantities since they may not converge as $k \to \infty$.

[3] Sliding window measurements could in principle use a larger

IV. EXCESS ENTROPY AS TOTAL INFORMATION STORAGE

Discussion of information storage or memory in CAs has often focused on periodic structures (particularly in construction of universal Turing machines), e.g. [25]. However, information storage does not necessarily entail periodicity. The **excess entropy** more broadly encompasses all types of structure and memory by capturing correlations across all lengths of time, including nonlinear effects. In Section IV A, we describe the excess entropy and the manner in which it measures the total information storage that is used in the future of a process. We then review in Section IV B how the excess entropy can be used to measure single-agent and collective information storage. We also discuss how information storage in an agent's environment in a distributed computation increases its information storage capacity beyond its internal capability. Subsequently in Section IV C we describe how the excess entropy can be localised in time and space.

A. Excess entropy

Grassberger [16] first noticed that a slow approach of the entropy rate to its limiting value was a sign of complexity. Formally, Crutchfield and Feldman [8] use the conditional entropy form of the entropy rate (Eq. (2))[4] to observe that at a finite block size k, the difference $H_{\mu X}(k) - H_{\mu X}$ represents the information carrying capacity in size k-blocks that is due to correlations. The sum over all k gives the total amount of structure in the system, quantified as **excess entropy**[5] (measured in bits):

$$E_X = \sum_{k=0}^{\infty} \left[H_{\mu X}(k) - H_{\mu X} \right]. \tag{11}$$

The excess entropy can also be formulated as the mutual information between the semi-infinite past and semi-infinite future of the system:

$$E_X = \lim_{k \to \infty} E_X(k), \tag{12}$$

$$E_X(k) = I_{X^{(k)};X^{(k^+)}} \tag{13}$$

where $X^{(k^+)}$ is the random variable (with measurements $x_{n+1}^{(k^+)} = \{x_{n+1}, x_{n+2}, \ldots, x_{n+k}\}$) referring to the k future states of X (from time step $n+1$ onwards). The two formulations of E_X in Eq. (11) and Eq. (12) are equal in the limit $k \to \infty$ (see Appendix 7 in [8]) though do not necessarily provide the same estimate at finite-k. Now, this second formulation in Eq. (13) is known as the **predictive information** [3], as it highlights that the excess entropy captures the information in a system's past which can be used to predict its future. From the perspective of distributed computation, the excess entropy can thus be seen to measure the information from the past of the system that is used at some point in the future. This is significant as it is explicitly consistent with the interpretation of the excess entropy as the amount of structure or memory in the system.

We note that the excess entropy is directly related to several known measures of complexity, including the Tononi-Sporns-Edelman complexity [47] (as described in [2]). Most important here is that it contrasts with the *statistical complexity* [9], which measures the amount of information in the past of a process that is *relevant* to the prediction of its future states. It is known that the statistical complexity $C_{\mu X}$ provides an upper bound to the excess entropy [44]; i.e. $E_X \leq C_{\mu X}$. This can be interpreted in that the statistical complexity measures *all* information stored by the system which *may be used* in the future, the excess entropy only measures that information which *is used* by the system at some point in the future. In this paper, we will focus on measuring only the excess entropy at local spatiotemporal points, since in our aim to understand the dynamics of information storage in distributed computation, we must focus here on how information storage is dynamically *used.* We plan to examine local profiles of the statistical complexity in future work. We note that the *light-cone formulation* of the statistical complexity has previously been examined at local spatiotemporal points [45], yet this formulation is not as readily interpretable as information storage for a single agent, and is not directly comparable to the single-agent perspective examined for excess entropy here.

B. Single-agent and collective excess entropy

We use the term *single-agent excess entropy* (or just excess entropy) to refer to measuring the quantity E_X for individual agents X or cells X_i using their one-dimensional time series of states. This is a measure of the *average* memory for *each agent.*

The predictive information form in Eq. (13), $E_X(k) = I_{X^{(k)};X^{(k^+)}}$, shows that the maximum excess entropy is the information capacity in sequences of k states $X^{(k)}$. In ECAs for example, this is 2^k bits. In the limit $k \to \infty$, this becomes infinite. Mathematically then, the information stored by a single agent can be larger than the information capacity of a single state. Where the agent

number of observations or samples to determine the relevant probability distribution functions (PDFs) than the number of evaluations averaged over for the measure (e.g. evaluate and average say five local entropy values, but with the PDFs estimated from 100 observations). However in most cases the observation and evaluation set is the same.

[4] $H_{\mu X}(k)$ here is equivalent to $h_\mu(L-1)$ in [8]. This means the sum in Eq. (11) starts from $k=0$ as equivalent to $L=1$.

[5] The excess entropy was originally termed the "effective measure complexity" by Grassberger in [16].

takes direct causal influence from only a single past state (as in CAs), the meaning of its information storage being larger than its information capacity is not immediately obvious. For instance, a cell in an ECA could not store more than 1 bit of information in isolation. However, the cells in a CA are participating in a *distributed computation*: cyclic causal paths (facilitated by *bidirectional* links) effectively allow cells to store extra information in neighbours (even beyond the immediate neighbours), and to subsequently retrieve that information from those neighbours at a later point in time. While measurement of the excess entropy does not explicitly look for such *self-influence* communicated through neighbours, it is indeed the channel through which a significant portion of information can be communicated. This self-influence between semi-infinite past and future blocks being conveyed via neighbours is indicated by the curved arrows in Fig. 1(a). There are parallels here to the use of stigmergy (indirect communication through the environment, e.g. see [23, 24]) to communicate with oneself, if one considers neighbours to be part of the environment of the agent. Indeed, because information may be stored and retrieved from one's neighbours, an agent can store information regardless of whether it is causally connected with itself.[6]

Information storage exceeding single-state information capacity is then a perfectly valid result. Indeed in an infinite CA, each cell has access to an infinite number of neighbours in which to store an infinite amount of information that can later be used to influence its own future. Since the storage medium is *shared* by all cells though, one should not think about the total memory as the total number of cells N multiplied by this amount (i.e. to give NE_X).

The average total memory stored for future use in a *collective* of agents (e.g. a set of neighbouring cells in a CA) is properly measured by the *collective excess entropy*. It is measured as *temporal* excess entropy of the agents using their two-dimensional time series of states. It is a joint temporal predictive information, i.e. the mutual information between the joint past $\mathbf{X}^{(k)}$ and future $\mathbf{X}^{(k^+)}$ of the agents:

$$E_{\mathbf{X}} = \lim_{k \to \infty} I_{\mathbf{X}^{(k)};\mathbf{X}^{(k^+)}}, \tag{14}$$

This collective measurement takes into account the inherent redundancy in the shared information storage medium (which NE_X does not). Collective excess entropy could be used for example to quantify the "undiscovered *collective memory* that may present in certain fish schools" [6].

Grassberger found divergent collective excess entropy for several CA rules, including rule 22 [15, 16].[7] This infinite amount of collective memory implies a highly complex process, since in using strong long-range correlations a semi-infinite sequence "could store an infinite amount of information about its continuation" [26]. On the other hand, infinite collective excess entropy can also be achieved for systems that only trivially utilise all of their available memory.[8] In attempting to quantify *local* information dynamics of distributed computation here, our focus is on information storage for *single agents or cells* rather than the joint information storage across the collective. Were the single-agent excess entropy found to be divergent (this has not been demonstrated), this may be more significant than for the collective case. This is because it would imply that all agents are *individually* strongly utilising the resources of the collective in a highly complex process. Crutchfield and Feldman measure the excess entropy in spatially-extended blocks in various CAs in an ensemble study in [11], however this is not an information storage since the measurement is not made temporally.

Our focus here is on *locally* quantifying information storage (temporally) in both *time* and space. We hypothesise this will provide much more detailed insights than single ensemble values into information storage structures and their involvement in distributed *computation.*

C. Local excess entropy

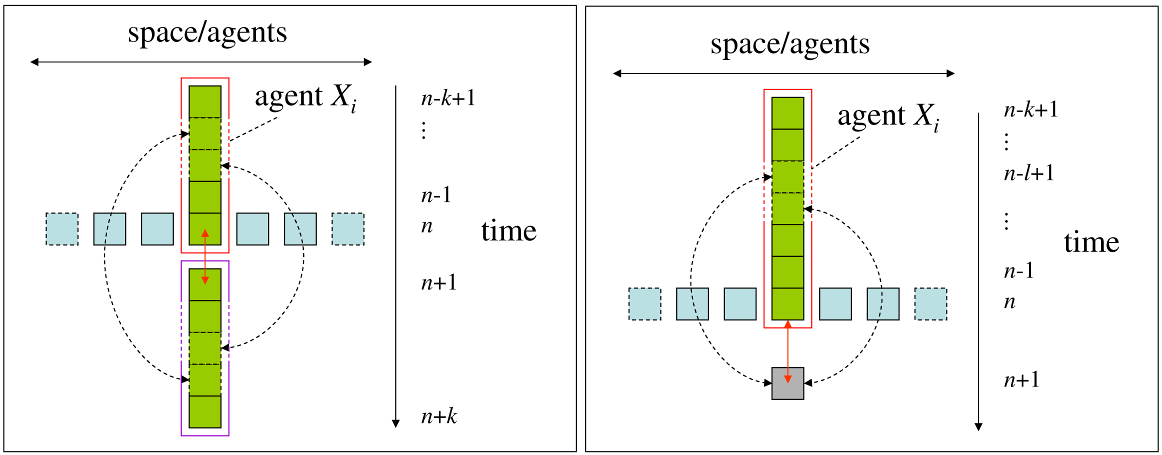
The **local excess entropy** is a measure of how much information a given agent is *currently* storing for future use at a particular point in time. To derive it, note that (as per Eq. (6)) the excess entropy of a process is actually the *expectation value* of the local excess entropy for the process at every time step [44]. This is as per Shalizi's original formulation of the local excess entropy in [44], however our presentation is for a single time-series rather than the light-cone formulation used there (such that we focus on the use of that time-series' past alone in the computation of its future).

Using the predictive information formulation from

[6] Information storage can also be measured here even when there is no causal path from an agent to itself via other agents (e.g. the agent may simply continually copy the value of another agent whose process involves information storage). Under our perspective of distributed computation, the measure simply captures *apparent* information storage which supports the process, whether that storage is in the agent itself, distributed via its neighbours, or is otherwise intrinsic to the process itself. The distinction between the perspectives of computation and causal effect is explored in [29].

[7] Lindgren and Nordahl [26] also measured excess entropy (referred to as effective measure complexity) for some ECAs. They measured spatial excess entropies however, and we note that it is only temporal excess entropies which are interpretable as information storage from our perspective of distributed computation.

[8] For example, a CA (with infinite width) that simply copied cell values to the right would have infinite collective excess entropy when started from random initial states, yet this is clearly a trivial use of this storage.



(a)Excess entropy

(b)Active information storage

FIG. 1: Measures of single-agent information storage in distributed systems. (a) Excess entropy: *total* information from the cell's past that is used at some point in its future. (b) Active information storage: the information storage that is *currently in use* in determining the next state of the cell. The stored information can be conveyed directly through the cell itself or via neighbouring cells.

Eq. (12), the local excess entropy $e_X(n+1)$ of a process is simply the local mutual information between the semi-infinite past $x_n^{(k)}$ and future $x_{n+1}^{(k^+)}$ at the given time step $n+1$:

$$E_X = \langle e_X(n+1) \rangle_n \, , \tag{15}$$

$$e_X(n+1) = \lim_{k \to \infty} i(x_n^{(k)}; x_{n+1}^{(k^+)}), \tag{16}$$

$$e_X(n+1) = \lim_{k \to \infty} \log_2 \frac{p(x_n^{(k)}, x_{n+1}^{(k^+)})}{p(x_n^{(k)}) p(x_{n+1}^{(k^+)})}. \tag{17}$$

Certainly the formulation of entropy rate overestimates in Eq. (11) could be used to directly form alternative localisations, e.g. $e'_X(n+1) = \sum_{k=0}^{\infty} [h_{\mu X}(n+1,k) - h_{\mu X}(n+1)]$. Indeed, in the limit $k \to \infty$ the averages of these localisations E_X will be the same. However, these two localisations will produce different results for each time step n in general. We measure only the local formulation from the predictive information in Eq. (17), because this form uniquely, explicitly captures the total information stored for future use at a particular temporal point, which is our quantity of interest here due to our focus on local information storage.[9]

The limit $k \to \infty$ is an important part of the above definition (carried over from Eq. (12)), since correlations at all time scales should be included in the computation of information storage. Since this is not computationally feasible in general, we retain the following notation for finite-k estimates:

$$E_X(k) = \langle e_X(n+1,k) \rangle_n \, , \tag{18}$$

$$e_X(n+1,k) = i(x_n^{(k)}; x_{n+1}^{(k^+)}), \tag{19}$$

$$= \log_2 \frac{p(x_n^{(k)}, x_{n+1}^{(k^+)})}{p(x_n^{(k)}) p(x_{n+1}^{(k^+)})}, \tag{20}$$

$$e_X(n+1) = \lim_{k \to \infty} e_X(n+1,k). \tag{21}$$

We discuss the requirements for $e_X(n,k)$ to converge to a limiting value with $k \to \infty$ in Appendix A.

The notation is generalised for lattice systems (such as CAs) with *spatially-ordered* agents to represent the local excess entropy for cell i at time $n+1$ as:

$$e(i,n+1) = \lim_{k \to \infty} \log_2 \frac{p(x_{i,n}^{(k)}, x_{i,n+1}^{(k^+)})}{p(x_{i,n}^{(k)}) p(x_{i,n+1}^{(k^+)})}, \tag{22}$$

$$= \lim_{k \to \infty} e(i,n+1,k). \tag{23}$$

Local excess entropy is defined for every spatiotemporal point (i,n) in the system (where i is a spatial index and n is a time index). Note that the collective excess entropy $E_{\mathbf{X}}$ can also be localised, but only in time, to have $e_{\mathbf{X}}(n+1,k)$.

While the average excess entropy is always positive, the local excess entropy may in fact be positive or *negative*, meaning the past history of the cell can either positively inform us or actually *misinform* us about its future. An observer is misinformed where the semi-infinite past and future are relatively unlikely to be observed together as compared to their independent likelihoods. In other words, an observer is misinformed by the past when the

[9] We note that this is in spite of the fact that the finite-k estimates $E_X(k)$ from Eq. (11) are better estimators than those from Eq. (12) [8]: interpretability of information storage properties is more important for our purposes.

observed future is conditionally less likely given the observed past than without considering the past. In this situation we have $p(x_{n+1}^{(k^+)} \mid x_n^{(k)}) < p(x_{n+1}^{(k^+)})$ making the denominator of Eq. (17) greater than the numerator, and giving a negative value for $e_X(n+1)$.

V. ACTIVE INFORMATION STORAGE AS STORAGE DIRECTLY IN USE

The excess entropy measures the total stored information which will be used *at some point* in the future of the state process of an agent. This information will possibly but not necessarily be used at the next time step $n+1$. Since the dynamics of computation unfold one step at a time, we are quite interested in how much of the stored information is actually *in use* at the next time step when the new process value is computed. As can be seen in extensions of this work [31, 32], this is particularly important in understanding how stored information interacts with information transfer in information processing. As such, we derive **active information storage** A_X as the average mutual information between the semi-infinite past of the process $X^{(k)}$ and its *next state* X', as opposed to its whole (semi-infinite) future:

$$A_X = \lim_{k\to\infty} A_X(k), \tag{24}$$

$$A_X(k) = I_{X^{(k)};X'}. \tag{25}$$

We use $A_X(k)$ to represent finite-k estimates. The active information storage is represented in Fig. 1(b). Of course, one could also define a collective active information storage $A_{\mathbf{X}} = \lim_{k\to\infty} I_{\mathbf{X}^{(k)};\mathbf{X}'}$.

A. Local active information storage

Following our local approach in Eq. (6), the **local active information storage** $a_X(n+1)$ is then a measure of the amount of information storage in use by the process at a particular time-step $n+1$. It is the local mutual information (or pointwise mutual information) between the semi-infinite past of the process and its next state:

$$A_X = \langle a_X(n+1)\rangle_n, \tag{26}$$

$$a_X(n+1) = \lim_{k\to\infty} a_X(n+1,k), \tag{27}$$

$$A_X(k) = \langle a_X(n+1,k)\rangle_n, \tag{28}$$

$$a_X(n+1,k) = \lim_{k\to\infty} \log_2 \frac{p(x_n^{(k)}, x_{n+1})}{p(x_n^{(k)})p(x_{n+1})}, \tag{29}$$

$$= \lim_{k\to\infty} i(x_n^{(k)}; x_{n+1}). \tag{30}$$

As for the excess entropy, note that we have retained notation for finite-k estimates here.

Again, we generalise the measure for agent X_i in a lattice system as:

$$a(i,n+1) = \lim_{k\to\infty} \log_2 \frac{p(x_{i,n}^{(k)}, x_{i,n+1})}{p(x_{i,n}^{(k)})p(x_{i,n+1})}, \tag{31}$$

$$= \lim_{k\to\infty} a(i,n+1,k). \tag{32}$$

We note that the local active information storage is defined for every spatiotemporal point (i,n) in the lattice system. We have $A(i,k) = \langle a(i,n,k)\rangle_n$. For systems of homogeneous agents where the probability distribution functions are estimated over all agents, it is appropriate to average over all agents also, giving:

$$A(k) = \langle a(i,n,k)\rangle_{i,n}. \tag{33}$$

The average active information storage will always be positive (as for the excess entropy), but is limited by the amount of information that can be used in the next state. This is, it is bounded above by the average information capacity of a single state (e.g. $\log_2 b$ bits where the agent only takes b discrete states). The *local* active information storage is not bound in this manner however, with larger values indicating that the particular past of an agent provides strong positive information about its next state.

Furthermore, the local active information storage can be *negative*, where the past history of the agent is actually *misinformative* about its next state. Similar to the local excess entropy, an observer is misinformed where the probability of observing the given next state in the context of the past history, $p(x_{n+1} \mid x_n^{(k)})$, is lower than the probability $p(x_{n+1})$ of observing that next state without considering the past.

B. Relation to entropy rate and excess entropy

The *average* information required to predict the next state X' of an element is simply the single cell entropy $H_{X'}$. We use the mutual information expansion of Eq. (1) to express the entropy in terms of the active information storage and entropy rate estimates[10]:

$$H_{X'} = I_{X';X^{(k)}} + H_{X'|X^{(k)}}, \tag{34}$$

$$H_{X'} = A_X(k) + H_{\mu X}(k), \tag{35}$$

$$H_{X'} = A_X + H_{\mu X}. \tag{36}$$

Logically, we can restate this as: the information to compute or predict a given state is the amount predictable

[10] We emphasise that these equations are correct not only in the limit $k\to\infty$ (i.e. for Eq. (36)) but for estimates with any value of $k \geq 1$ (i.e. for Eq. (35)). We note also that Eq. (36) demonstrates that A_X converges to a limiting value when such limits exist for $H_{\mu X}$ (e.g. for stationary processes [7]).

from its past (the active memory) plus the remaining uncertainty after examining this memory.

This equation makes explicit our interpretation of information storage in distributed computation. It is the information in the past that is *observed* to contribute to the computation of the next state. Whether this information is actually causal for that next state, either directly or indirectly through neighbouring agents, is irrelevant for this perspective.[11]

This relationship between the entropy, active information storage and entropy rate can be expressed in local notation also:

$$h_X(n+1) = a_X(n+1,k) + h_{\mu X}(n+1,k), \qquad (37)$$
$$h(i,n+1) = a(i,n+1,k) + h_\mu(i,n+1,k). \qquad (38)$$

Appendix A describes the manner in which Eq. (37) can be used to show that the local active information storage converges to a limiting value with $k \to \infty$ when the local entropy rate exhibits such convergence (which can only be demonstrated under more strict conditions than for convergence of the average entropy rate).

It is also possible to express the excess entropy in terms of finite-k estimates of the active information storage. We rearrange Eq. (35) to get $H_{\mu X}(k) = H_{X'} - A_X(k)$, and then substitute this into Eq. (11), getting:

$$E_X = \sum_{k=0}^{\infty} \left[A_X - A_X(k) \right]. \qquad (39)$$

This expression shows that the excess entropy is the sum of *underestimates* of the active information storage at each finite history length $k \geq 0$ (with $A_X(0) = 0$).

This relationship is displayed graphically in Fig. 2 in a similar fashion to the plot for the excess entropy in terms of entropy rate estimates in Fig. 3 of [8].[12] Note that $A(k)$ is non-decreasing with k; this is because increasing the time examined in history widens the scope of temporal correlations that the measure can capture.

VI. LOCAL INFORMATION STORAGE IN CELLULAR AUTOMATA

In this section, we evaluate the local measures within sample runs for the CA rules described in Section II. To do so, we estimate the required probability distribution functions from CA runs of 10 000 cells, initialised from random states, with 600 time steps retained (after the first 30 time steps were eliminated to allow the CA to settle). Periodic boundary conditions were used. Observations taken at every spatiotemporal point in the CA were used in estimating the required probability distribution functions, since the cells in the CA are homogeneous agents. All conclusions were confirmed by multiple runs from different initial states, and all CA plots were generated using modifications to [50].

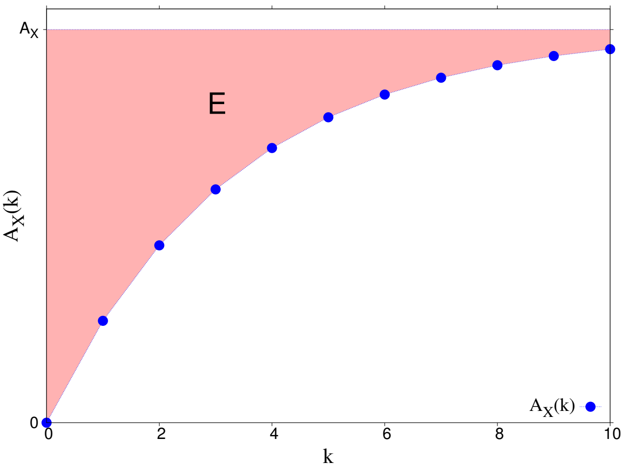


FIG. 2: Active information storage convergence: a plot of estimates $A_X(k)$ versus history length k as they converge to the limiting value A_X. The shaded area is the excess entropy E.

In the following we discuss the key results here:

- that one should use as large a history length k as possible to adequately measure information storage (Sections VI A and VI B);
- the dominant storage entities in CAs are blinkers and domains (Section VI B);
- negative information storage at particles represents the misinformation conveyed by the storage there (Section VI C);
- local entropy rate highlights the location of moving particles (Section VI G); and
- observations of other minor information storage phenomena (Sections VI D, VI E and VI F).

A. Appropriate history lengths

As previously stated, these measures are only completely correct in the limit $k \to \infty$, however this limit is not computationally achievable. A logical question is what history length k is reasonable to use, noting that setting $k = 1$ is something of a default approach (e.g. for the excess entropy in [1]).

Fig. 3 presents the average active information storage $A(k)$ in several ECA rules as a function of history length k. In particular, we observe that using too small a value

[11] The distinction between causal effect and the perspective of computation is explored in [29].

[12] Note that our sum and the plot start from $k = 0$, unlike the expressions and plots in [8] which start from $L = 1$. The difference is that we have adopted $k = L - 1$ to keep a focus on the number of steps k in the past history, which is important for our computational view.

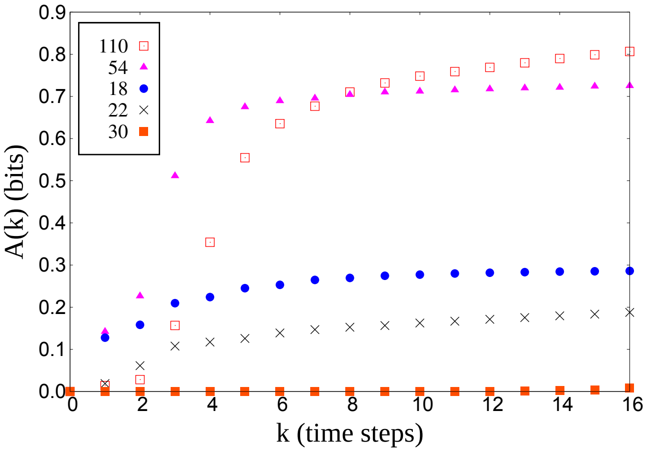


FIG. 3: Active information storage $A(k)$ versus history length k for several ECA rules.

for k (e.g. $k < 5$ for rule 110) can lead one to substantially underestimate the information storage. Even in a system as apparently simple as a CA, the default $k = 1$ is clearly inadequate. Obviously in measuring information storage one wants to capture all of the temporal correlations and so use $k \to \infty$. However, the selection of k is limited not only by the amount of history available to examine, but also by the number of observations available for probability distribution function estimation. If k is made too large, the mutual information will be artificially inflated due to under-sampling.

One simple recommended heuristic is to select k to have at least three times as many samples as possible state configurations $(x_{n+1}, x_n^{(k)})$ [35], which would suggest keeping $k \leq 19$ here. More formally however, we select k to have at least M samples on average for each observation in the *typical set* of state configurations [7, 38]. The typical set refers to the set of state configurations where the "sample entropy is close to the true entropy" of that joint state [7], and can be thought of as the set of state configurations likely to be encountered frequently enough to contribute to that entropy. For k length blocks of binary variables, the size of the typical set can be approximated as $2^{h_\mu k}$. For our purposes with $A(k)$, we are considering k length blocks plus the next state, so calculate the typical set for our state configurations as $2^{h_\mu (k+1)}$. With $h_\mu = 0.18$ estimated using $k = 16$ for rule 110 [33], we find the size of the typical set grows much more slowly with k than the set of possible state configurations. This means that fulfilling a desire for $M > 10$ for reasonable accuracy is easily fulfilled for rule 110 and many other rules using $k \leq 18$. Indeed, it is only for rules with $h_\mu \to 1$ (e.g. rule 30, see [33]) that M even approaches 10 with $k \leq 18$; for most rules $k \leq 18$ results in a much larger average number of samples $M >> 10$ and therefore larger accuracy. We elect to continue our CA investigations with the more strict condition $k \leq 16$ however, for comparison to our related work in [31–33].

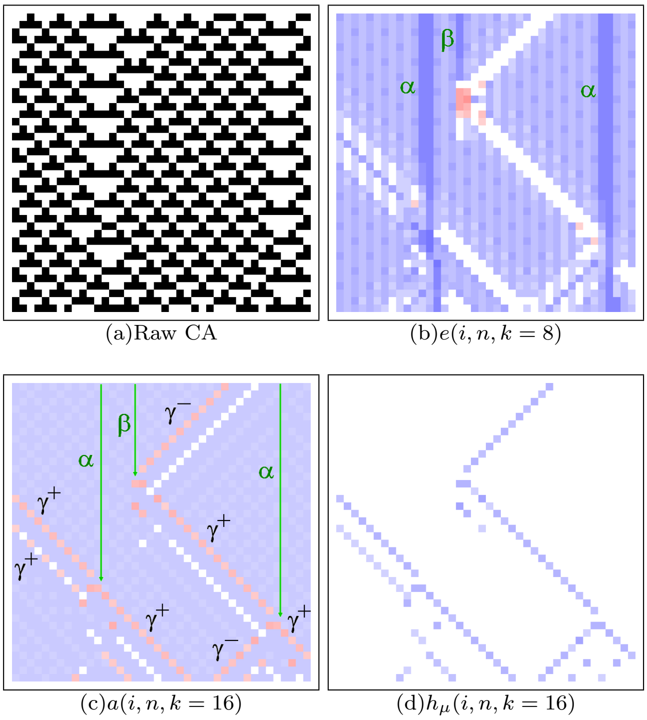
Fig. 3 suggests for example that the majority of the information storage for rule 110 is captured with $k \geq 7$ or so, however this examination of the average values of $A(k)$ does not show explicitly why this is the case. Furthermore, these average values tell us nothing about whether blinkers are dominant information storage structures, and if so whether the information storage in them has been captured at these history lengths. To understand these issues, we begin to examine the information storage on a local scale in the next section.

B. Periodic blinker and domain processes as dominant storage

We begin by examining the local profiles for rules 54 and 110, which are known to contain regular gliders against periodic background domains. For the CA evolutions in Fig. 4(a) and Fig. 5(a), the local profiles of $e(i, n, k = 8)$ are displayed in Fig. 4(b) and Fig. 5(b), and the local profiles of $a(i, n, k = 16)$ in Fig. 4(c) and Fig. 5(c).

It is quite clear that positive information storage is concentrated in the vertical gliders or blinkers, and the domain regions. As expected in our hypothesis in Section II, these results provide quantitative evidence that the **blinkers are the dominant information storage entities**. This is because the cell states in the blinkers are strongly predictable from their past history, since they are temporally periodic. It is only the local profiles that demonstrate the strong information storage at these entities though. That **the domain regions for these rules also contain significant information storage** should not be surprising, since these too are periodic and so their past does indeed store information about their future. In fact, the local values for each measure form spatially and temporally periodic patterns in these regions, due to the underlying periodicities exhibited there.

As expected, these two measures provide useful filters for information storage structure here. Yet **the local active information storage and local excess entropy yield subtly different results here**. While $a(i, n, k = 16)$ indicates a similar amount of stored information in use to compute each space-time point in both the domain and blinker areas, $e(i, n, k = 8)$ reveals a larger *total* amount of information is stored in the blinkers. For the blinkers known as α and β in rule 54 [21] this is because the temporal sequences of the centre columns of the blinkers (0-0-0-1, with $e(i, n, k = 8)$ in the range 5.01 to 5.32 bits) are more complex than those in the domain (0-0-1-1 and 0-1, with $e(i, n, k = 8)$ in the range 1.94 to 3.22 bits), even where they are of the same period. In principle, we could define a threshold i_B to differentiate between the blinkers and domain using the local excess entropy. Note that we have the total stored information $e(i, n, k = 8) > 1$ bit in these regions due to the distributed information storage supported by bidirectional communication (as discussed in Section IV B).



(a)Raw CA

(b)$e(i, n, k = 8)$

(c)$a(i, n, k = 16)$

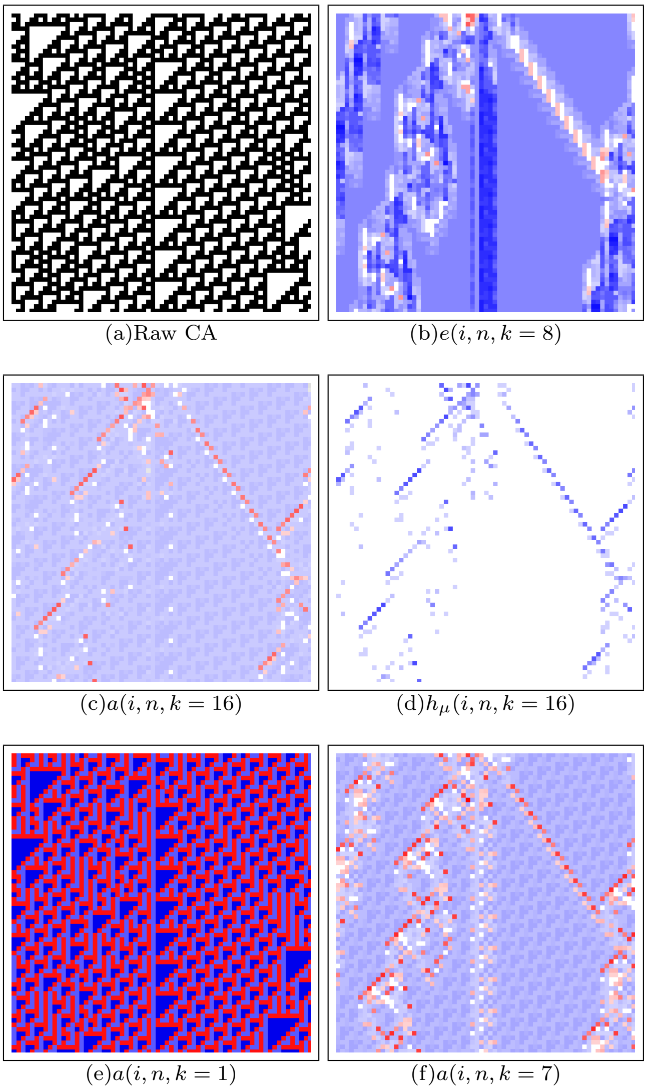
(d)$h_\mu(i, n, k = 16)$

FIG. 4: (colour online) Local information storage in **rule 54** (40 time steps displayed for 40 cells, time increases down the page for all CA plots). **Profiles are discretised into 16 levels, with blue for positive values and red for negative**. (b) Local excess entropy, max. 11.79 bits, min. -12.35 bits; (c) Local active information storage, max. 1.07 bits, min. -12.27 bits; (d) Local temporal entropy rate, max. 13.20 bits, min. 0.00 bits. Note that we mark the positions (in (b) and (d)) of the blinker types α and β and glider types γ^+ and γ^- (named following [21]).

This mechanism supports these periodic sequences being longer than two time steps (the maximum period a binary cell could sustain in isolation).

Importantly, we also confirm the information storage capability of the blinkers and domains in the human understandable computation of the ϕ_{par} density classification rule [40, 41] (not shown, see additional results in [34]).

To further investigate the appropriate history length k for use with the information storage measures, we examine the profiles of $a(i, n, k = 1)$ and $a(i, n, k = 7)$ for rule 110 in Fig. 5(e) and Fig. 5(f). As per the low average value for $A(k = 1)$ for rule 110 in Fig. 3, Fig. 5(e) demonstrates that the use of $k = 1$ is inadequate here since it does not capture the strong information storage in the gliders and domain regions that we see for the profiles with $k = 16$. On the other hand, Fig. 5(f) shows that the use of $k = 7$ does capture most of this strong information storage (compare to $k = 16$ in Fig. 5(c)), in alignment with the average value for $A(k = 7)$ for rule 110 approaching the limiting value in Fig. 3. This is because the blinker and domain regions for rule 110 are both of period 7. To understand why setting k at this period is effective, consider first an *infinite* temporally periodic process with period p. The next state of that process is completely predictable from its (infinite) past. In fact, the number of past states an observer must examine to correctly determine the next state is limited by p (as per the synchronisation time in [10]). Using $k > p - 1$ does not add any extra information about the next state than is already contained in the $p-1$ previous states. However, using $k < p - 1$ may not provide sufficient information



(a)Raw CA

(b)$e(i, n, k = 8)$

(c)$a(i, n, k = 16)$

(d)$h_\mu(i, n, k = 16)$

(e)$a(i, n, k = 1)$

(f)$a(i, n, k = 7)$

FIG. 5: (colour online) Local information dynamics in **rule 110** (67 time steps displayed for 67 cells). (b) Local excess entropy, max. 10.01 bits, min. -10.35 bits; (c) Local active information storage, max. 1.22 bits, min. -9.21 bits; (d) Local temporal entropy rate, max. 10.43 bits, min. 0.00 bits. Local active information storage with short history lengths: (e) with $k = 1$, max. 0.24 bits, min. -0.21 bits; (f) with $k = 7$, max. 1.22 bits, min. -5.04 bits.

for the prediction. Using $k = p-1$ is a sufficient (Markovian) condition for infinitely periodic processes. **Where a process contains *punctuated* periodic sequences** (e.g. the periodic blinkers and domains in a single cell's time series here), **setting $k = p-1$ will capture the information storage related to the period of these sequences and is a useful minimum value**. However, it will still ignore important longer-range correlations (e.g. encountering one type of glider in the near past may be strongly predictive of encountering a different type of glider in the near future). There is no general limit on the range of such self-influence, so in theory the limit $k \to \infty$ should be taken in measuring these quantities. This is why $k = 7$ captures much, but not all of the active information storage for rule 110.

C. Negative informative storage as misinformation at particles

Negative values of $a(i, n, k = 16)$ for rules 54 and 110 are also displayed in Fig. 4(c) and Fig. 5(c). Interestingly, **negative local components of active information storage are concentrated in the travelling glider areas** (e.g. γ^+ and γ^- for rule 54 [21]), **providing a good spatiotemporal filter of the glider structure**. **This is because when a travelling glider is encountered at a given cell, the past history of that cell** (being part of the background domain) **is *misinformative* about the next state**, since the domain sequence was more likely to continue than be interrupted.

For example, see the marked positions of the γ gliders in Fig. 6. These positions are part of a domain wall (or particle) by definition, since the temporal pattern of the background domain breaks down at these points. At the points in the glider marked ×, we have $p(x_{n+1} \mid x_n^{(k=16)}) = 0.25$ and $p(x_{n+1}) = 0.52$: since the next state occurs relatively infrequently after the given history, that history provides a misinformative $a(n, k = 16) = -1.09$ bits about the next state. This is juxtaposed with the points four time steps before those marked ×, which have the same history $x_n^{(k=16)}$ but remain part of the domain. There we have $p(x_{n+1} \mid x_n^{(k=16)}) = 0.75$ and $p(x_{n+1}) = 0.48$ giving $a(n, k = 16) = 0.66$ bits, quantifying the positive information storage there.

Note that the points with misinformative storage are not necessarily those selected by other filtering techniques as part of the gliders. For example, the finite state transducers technique from computational mechanics (using left to right *spatial* scanning by convention) [19] would identify points 3 cells to the right of those marked × as part of the γ^+ glider. While that technique has the perspective of spatial pattern recognition, we take the temporal perspective of unfolding computation.

The local excess entropy also produced some negative values around travelling gliders (see Fig. 4(b) and Fig. 5(b)), though these were far less localised on the gliders themselves and less consistent in occurrence than for the local active information storage. This is because the local excess entropy, as measure of *total* information storage used in the future, is more loosely tied to the dynamics at the given spatiotemporal point. The effect of a glider encounter on $e(i, n, k)$ is smeared out in time, and in fact the dynamics may store more positive information in total than the misinformation encountered at the specific location of the glider. For example, parallel glider pairs in Fig. 4(b) have positive total information storage, since a glider encounter becomes much more likely in the wake of a previous glider. We also note negative values of $e(i, n, k)$ near the start and end of blinkers (see Fig. 4(b)), since the encountering of new periodic behaviour here can be misinformative until enough of that behaviour is established in the past of the time series to facilitate prediction of the future.

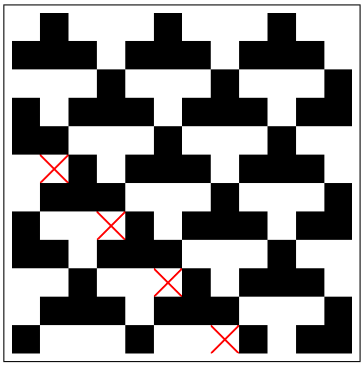


FIG. 6: Close up of raw states of rule 54. × marks some positions in a γ^+ glider. As discussed in Section VI C these points are part of the glider because the temporal domain pattern 0-0-1-1 breaks down at these points. Section VI C explains that we have *negative* local information storage at these points, because the past history of the cell misinforms an observer regarding an expected continuation of the background domain.

D. Particles create new information storage

There is also strong positive information storage in the "wake" of the more complex gliders in rule 110 (e.g. see the gliders at the left of Fig. 5(b) and Fig. 5(c)). This indicates that while the leading edge of the gliders cause the cell states to become unpredictable from their past, the subsequent activity (before a domain pattern is established) is predictable given the glider encounter. **The leading edge of the gliders can thus be seen to store information in the cell about its new behaviour**. The presence of this information storage is shown by both measures, although the relative strength of the total information storage is again revealed only by the local excess entropy. We will observe a similar creation of new information storage by domain walls in rule 18 in Section VI F.

E. Structured information storage in domain of rule 18

There is also interesting information storage structure in ECA rule 18, which contains domain walls against a seemingly irregular background domain. The local profiles for $e(i,n,k=8)$ and $a(i,n,k=16)$ are plotted in Fig. 7(b) and Fig. 7(c) for the raw states of rule 18 displayed in Fig. 7(a). In contrast to rules 54 and 110, the background domain for rule 18 contains points with both positive and negative local active information storage. Considering these components together, we observe a pattern to the background domain of spatial and temporal period 2 corresponding to the period-2 ϵ-machine generated to recognise the background domain for ECA rule 18 by Hanson and Crutchfield [18]. Every second site in the domain is a "0", and contains a small positive $a(i,n,k=16)$ ($\approx$ 0.43 to 0.47 bits); information storage of this primary temporal phase of the period is sufficient to predict the next state here. The alternate site is either a "0" or a "1", and contains either a small negative $a(i,n,k=16)$ at the "0" sites ($\approx$ -0.45 to -0.61 bits) or a larger positive $a(i,n,k=16)$ at the "1" sites ($\approx$ 0.98 to 1.09 bits). Information storage of the cell being in the alternate temporal phase is strongly in use or active in computing the "1" sites, since the "1" sites only occur in the alternate phase. However, the information storage indicating the alternate temporal phase is misleading in computing the "0" sites since they occur more frequently with the primary phase. Indeed, encountering a "0" at the alternate sites creates ambiguity in the future (since it makes determination of the phase more difficult) so in this sense it can be seen as detracting from the overall storage. The background domain should contain a consistent level of excess entropy at 1 bit to store the temporal phase information, and this occurs for most points.[13] Again, this resembles a smearing out of the local periodicity of the storage in use, and **highlights the subtle differences between the excess entropy and active information storage**.

F. Misinformation and new storage creation by domain walls

The domain walls in rule 18 are points where the spatiotemporal domain pattern is violated. **Strong negative components of the local active information storage reveal the temporal violations, which occur when the domain wall moves** or travels into a new cell and the past of that cell cannot then predict the next state successfully. This misinformation is analogous to our observations for regular gliders in rules 54 and 110 in Section VI C. Importantly, these negative values of $a(i,n,k=16)$ (which are less than -2.5 bits) are much stronger than those in the background domain, and are strongly localised on the domain walls. Again, the negative components of $a(i,n,k=16)$ appear to be a useful filter for moving coherent spatiotemporal structure.

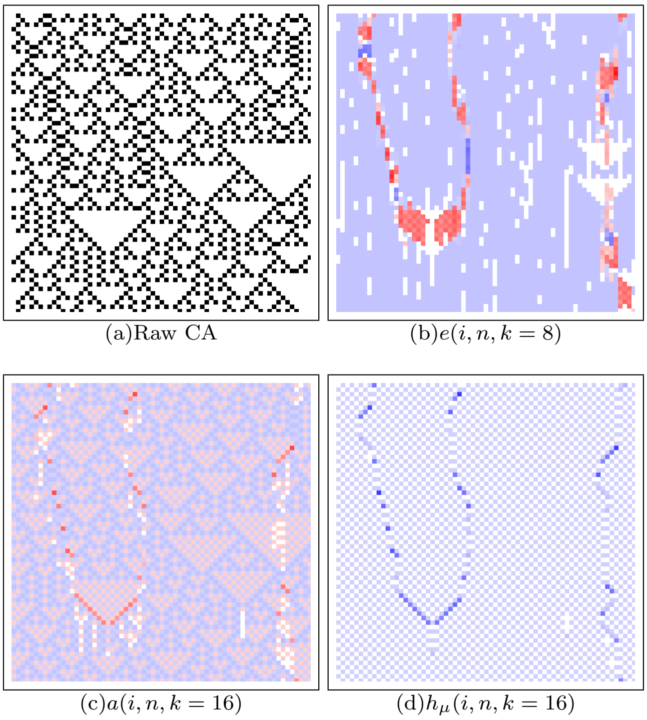


FIG. 7: Local information storage in **rule 18** (67 time steps displayed for 67 cells). (b) Local excess entropy, max. 4.62 bits, min. -8.65 bits; (c) Local active information storage, max. 1.98 bits, min. -9.92 bits; (d) Local temporal entropy rate, max. 11.90 bits, min. 0.00 bits.

The local excess entropy profile on the other hand contains both positive and negative values (Fig. 7(b)) for the domain walls. As per the results for gliders, these negative values are less specifically localised on the domain walls than observed for $a(i,n,k)$. The strong positive values of $e(i,n,k=8)$ are observed to occur where the domain wall makes several changes of direction during the k steps but is somewhat motionless on average. This is because a domain wall encounter is much more likely in the wake of previous domain wall movement than elsewhere in the CA. This has analogies to both the new information storage creation by gliders in Section VI C, and the storage in stationary blinkers in Section VI B.

G. Local temporal entropy rate highlights moving particles

The local temporal entropy rate profiles $h_\mu(i,n,k)$ are displayed in Fig. 4(d) for rule 54, Fig. 5(d) for rule 110

[13] The exceptions are where long temporal chains of 0's occur, disturbing the memory of the phase due to finite-k effects.

and Fig. 7(d) for rule 18. These are the first known *temporal* entropy rate profiles published for CAs, though *spatial* entropy rate profiles can be seen in [20] (referred to as *local information* there). Clearly **these local temporal entropy rate profiles are useful spatiotemporal filters for moving emergent structure, highlighting all of the moving particles in each system**.

In fact, these profiles are quite similar to those of the negative values of local active information storage.[14] This is not surprising since $h_\mu(i,n,k)$ and $a(i,n,k)$ are seen to be complementary in Eq. (37). Where $a(i,n,k)$ is negative, $h_\mu(i,n,k)$ must be strongly positive since the local single cell entropy $h(i,n)$ averages close to 1 bit for these examples. That is, where the information storage $a(i,n,k)$ is misinformative about the next state of a cell, there is a high uncertainty $h_\mu(i,n,k)$ in this next state given its past history.

Similarly, $h_\mu(i,n,k)$ only highlights *travelling* particles: as we have already seen, stationary coherent elements such as blinkers are information storage entities for which there is little to no uncertainty in the next state given the past.

In deterministic systems such as CAs, any extra information $h_\mu(i,n,k)$ about the next state *must* come from the neighbouring cells. **The temporal entropy rate therefore represents a *collective information transfer* from the neighbours in deterministic systems** (see further discussion of this point taken up in [32]). Importantly though, it cannot identify *individual sources* of information transfer: we directly investigate information transfer based filtering (identifying specific sources of information) in [31].

H. Absence of coherent information storage structure

Finally, we note that profiles of $e(i,n,k=8)$ and $a(i,n,k=16)$ for ECA rule 22 and chaotic rule 30, not shown here, are included in additional results in [33, 34]. While storage is certainly observed to occur for these rules, these plots **provide further evidence that there is no coherent structure to the information storage for rule 22** (or indeed 30). This is particularly important for rule 22, since we recall from Section IV B that it was suggested to have infinite collective excess entropy [15, 16], however it had no known coherent structural elements [45]. Thus this is another clear example of the utility of examining local information dynamics over ensemble values.

VII. CONCLUSION

In this paper we have introduced and contrasted the local excess entropy and local active information storage in Sections IV and V. We have demonstrated that they provide complementary insights into information storage dynamics, because the excess entropy measures total storage used in the future while the active information storage measures storage in use in computing the next state. As such, their results are often similar in general, but do reveal subtly different aspects of the dynamics.

Importantly, in Section VI we showed that both are useful filters for information storage structure, and provide the first quantitative evidence that blinkers and domains are dominant information storage entities in cellular automata. In particular, the excess entropy revealed that blinkers in the CAs investigated here stored more information in total than domains. (See a summary of the application to CAs, and how our local information-theoretic measures could be used to classify emergent structure, in Table I). While both measures provide useful insights, the local active information storage is the most useful in a *real-time* sense, since calculation of the local excess entropy requires knowledge of the dynamics an arbitrary distance into the future.[15] Also, it also provides the most specifically *localised* insights, including highlighting moving elements of coherent spatiotemporal structure.

Furthermore as hinted by its relationship with the temporal entropy rate, the focus of the active information storage on computation of the *next state* of a process is particularly important in understanding how stored information interacts with information transfer in information processing. As such, we demonstrated that its complementary quantity, the entropy rate, reveals gliders as coherent information transfer structures here in CAs and is a useful filter for such moving structure. This being said, none of these measures are capable of identifying the *source* of information transfer in moving coherent structures, nor where information is modified in distributed computation; investigations of these operations on information are considered in our related work [31–33]. Importantly, this paper along with our related work [31–33] can thus be seen to form a complete framework for analysis of the local information dynamics of distributed computation in complex systems: i.e. the space-time dynamics of how information is stored, transferred and modified in distributed computation. This framework thus provides complete evidence for the com-

[14] Recall though from Section V B that $h_\mu(i,n,k)$ itself is never negative.

[15] As described in Section IV C, while there are alternative formulations of the local excess entropy which can be computed from past observations alone, they cannot be interpreted as the total information storage at the given time point. A similar concept would be the partial localisation (see [29]) $I(x_n^{(k)};X^{(k^+)})$, which quantifies how much information from the past is *likely* to be used in the future.

	$e_X(n) > i_B$	$e_X(n) < i_B$
$a_X(n) > 0$	blinkers, stationary domain walls (Section VI B and VI F)	periodic domain (Section VI B)
$a_X(n) < 0$	gliders, moving domain walls (Section VI C and VI F)	

TABLE I: Classification of emergent structures in cellular automata via their specific information storage properties. As discussed in Section VI B, i_B is a hypothetical threshold which could be used to differentiate between the blinkers and background domain using the local excess entropy.

putational roles of emergent structure in CAs: blinkers and domains as information storage (in this paper), particles as information transfer [31], and particle collisions as information modification [32].

We note finally that our local measures used here are able to properly reveal and explain all of the information storage dynamics (e.g. regarding the dominant structures) in ECA rule 54, unlike the alternative (spatial) redundancy measure suggested for storage in [12] (which for example cannot distinguish "storage" dynamics between gliders and the background domain). This is because the measures used here specifically capture information storage rather than spatially redundant information, which does not necessarily directly relate to all types of information storage dynamics. Similarly, the alternative measure for transfer used in [12] does not capture the dominant transfer components in gliders in ECA rule 54, because it measures only unique information from transfer sources (whereas gliders are only distinguished by a *synergistic* interaction of the source with destination history as history length k is made large enough [31]). Furthermore, short sliding window averaging was used for localisation in [12], which both blurs out local dynamics and also significantly undersamples the relevant probability distribution functions. The partial information decomposition approach certainly reveals very interesting properties of the dynamics, however properly capturing information storage and transfer dynamics requires specific measures of these operations, as provided by our framework.

In future work, we plan to compare our findings to local profiles of the statistical complexity which, while not measuring the information storage dynamically *in use* in the future, does measure the total information storage that is *relevant* to the future of the process [9]. We are also examining the relationship between network structure and the dynamic information storage capabilities of nodes on that network, e.g. finding that storage capability is directly related to locally clustered structure [28], specifically feedback and feedforward loop motifs [27].

Acknowledgments

A brief exposition of the local active information storage was previously presented in [30]. The CA diagrams were generated using enhancements to the software of Wójtowicz [50]. We thank Nils Bertschinger and Nihat Ay for discussions regarding convergence of local values.

APPENDIX A: CONVERGENCE OF LOCAL QUANTITIES

Here we consider conditions for convergence of the local entropy rate $h_{\mu X}(n, k)$, and consequently the local active information storage $a_X(n, k)$, as $k \to \infty$.

First, we note that when the average entropy rate converges (which occurs for example for stationary processes [7]), the difference between the estimates $H_{\mu X}(k+1) - H_{\mu X}(k)$ approaches 0 as $k \to \infty$. This difference can be expressed as an average conditional mutual information (MI) $\langle i(x_{n+1}; x_{n-k} \mid x_n^{(k)}) \rangle$ between the next state x_{n+1} and the state $k+1$ time steps beforehand, x_{n-k}, conditioned on the previous k states, $x_n^{(k)}$. We note that the difference between the *local* entropy rate estimates at each time step n, $h_{\mu X}(n, k+1) - h_{\mu X}(n, k)$, is equal to the local values of this conditional MI at that time step, i.e. $i(x_{n+1}; x_{n-k} \mid x_n^{(k)})$.

Now, under convergence of the average entropy rate, this average conditional MI $\langle i(x_{n+1}; x_{n-k} \mid x_n^{(k)}) \rangle$ must *converge* to zero as $k \to \infty$. Suppose this average conditional MI was *equal* to zero. Since we know that the Kullback-Leibler divergence between two conditional distributions $p(a \mid b)$ and $q(a \mid b)$ is equal to zero if and only if $p(a \mid b) = q(a \mid b)$ for all a and b with $p(b) > 0$ [7, p.27], then under such equality of the average, each local term $i(x_{n+1}; x_{n-k} \mid x_n^{(k)})$ must also be *equal* to zero. Now, being *equal* to zero is a stronger condition than *convergence* to zero; the equality $\langle i(x_{n+1}; x_{n-k} \mid x_n^{(k)}) \rangle = 0$ holds for processes of finite Markovian order k, but cannot be said to do so in general.[16]

[16] In general, one can construct local conditional MI values of arbitrarily large value as the average conditional MI converges to zero. Without the relevant Markovian condition, this is possible for $i(x_{n+1}; x_{n-k} \mid x_n^{(k)})$. We *conjecture* however that the *variance* of local conditional MI terms converge to zero as the average does so. Were this to be the case (which we will attempt to prove in future work), then one could show that the probability of encountering any non-zero local conditional MI value converges to zero when the average does so.

As such, we conclude that for processes with convergent average entropy rates, and of a finite Markovian order k, the incremental differences between the local entropy rate estimates vanish as $k \to \infty$, i.e. the local entropy rate converges. We will investigate whether less strict conditions can be established in future work. Until this can be done, one must be careful in using the local quantities since they may not converge as $k \to \infty$.

We note that the local active information storage $a_X(n, k)$ converges under the same conditions as above, since it is a simple function of the local entropy rate and the local entropy (see Eq. (37)).

Finally, a similar argument can be used for the convergence of the (predictive information form of the) local excess entropy in Eq. (16). We note that the incremental difference between the averaged estimates $E_X(k+1) - E_X(k)$ is the sum of two average conditional MI terms (which account for the extra future state x_{n+k+1} and past state x_{n-k}): $\langle i(x_{n+k+1}; x_n^{(k)} \mid x_n^{(k^+)}) \rangle + \langle i(x_{n-k}; x_n^{(k+1^+)} \mid x_n^{(k)}) \rangle$. As above, when the average excess entropy converges to a limiting value (i.e. the incremental difference vanishes), then both of these average conditional MI terms must converge to zero (since their sum must be zero, and neither can be negative). As above, under the stronger condition of equality to zero (met by processes of finite Markovian order), all of the local terms of each conditional MI are equal to zero, meaning the incremental differences $e_X(n, k+1) - e_X(n, k)$ between the *local* excess entropy estimates also converges to zero. As such, our local excess entropy terms converge to a limiting value for processes of a finite Markovian order (whose averages do so).

[1] Ay, N., Bertschinger, N., Der, R., Güttler, F., and Olbrich, E. (2008) Predictive information and explorative behavior of autonomous robots. *European Physical Journal B*, **63**, 329–339.

[2] Ay, N., Olbrich, E., Bertschinger, N., and Jost, J. (2011) A geometric approach to complexity. *Chaos*, **21**, 037103+.

[3] Bialek, W., Nemenman, I., and Tishby, N. (2001) Complexity through nonextensivity. *Physica A*, **302**, 89–99.

[4] Boedecker, J., Obst, O., Mayer, N. M., and Asada, M. (2009) Initialization and self-organized optimization of recurrent neural network connectivity. *HFSP Journal*, **3**, 340–349.

[5] Ceguerra, R. V., Lizier, J. T., and Zomaya, A. Y. (2011) Information storage and transfer in the synchronization process in locally-connected networks. *Artificial Life (ALIFE), 2011 IEEE Symposium on*, pp. 54–61, IEEE.

[6] Couzin, I., James, R., Croft, D., and Krause, J. (2006) Social organization and information transfer in schooling fishes. C., B., Laland, K., and Krause, J. (eds.), *Fish Cognition and Behavior*, pp. 166–185, Fish and Aquatic Resources, Blackwell Publishing.

[7] Cover, T. M. and Thomas, J. A. (1991) *Elements of Information Theory*. John Wiley & Sons.

[8] Crutchfield, J. P. and Feldman, D. P. (2003) Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos*, **13**, 25–54.

[9] Crutchfield, J. P. and Young, K. (1989) Inferring statistical complexity. *Physical Review Letters*, **63**, 105.

[10] Feldman, D. P. and Crutchfield, J. P. (2004) Synchronizing to periodicity: The transient information and synchronization time of periodic sequences. *Advances in Complex Systems*, **7**, 329–355.

[11] Feldman, D. P., McTague, C. S., and Crutchfield, J. P. (2008) The organization of intrinsic computation: Complexity-entropy diagrams and the diversity of natural information processing. *Chaos*, **18**, 043106.

[12] Flecker, B., Alford, W., Beggs, J. M., Williams, P. L., and Beer, R. D. (2011) Partial information decomposition as a spatiotemporal filter. *Chaos*, **21**, 037104+.

[13] Goh, K. I. and Barabási, A. L. (2008) Burstiness and memory in complex systems. *Europhysics Letters*, **81**, 48002.

[14] Grassberger, P. (1983) New mechanism for deterministic diffusion. *Physical Review A*, **28**, 3666.

[15] Grassberger, P. (1986) Long-range effects in an elementary cellular automaton. *Journal of Statistical Physics*, **45**, 27–39.

[16] Grassberger, P. (1986) Toward a quantitative theory of self-generated complexity. *International Journal of Theoretical Physics*, **25**, 907–938.

[17] Grassberger, P. (1989) Information content and predictability of lumped and distributed dynamical systems. *Physica Scripta*, **40**, 346.

[18] Hanson, J. E. and Crutchfield, J. P. (1992) The attractor-basin portait of a cellular automaton. *Journal of Statistical Physics*, **66**, 1415–1462.

[19] Hanson, J. E. and Crutchfield, J. P. (1997) Computational mechanics of cellular automata: An example. *Physica D*, **103**, 169–189.

[20] Helvik, T., Lindgren, K., and Nordahl, M. G. (2004) Local information in one-dimensional cellular automata. Sloot, P. M., Chopard, B., and Hoekstra, A. G. (eds.), *Proceedings of the International Conference on Cellular Automata for Research and Industry, Amsterdam*, Berlin/Heidelberg, vol. 3305 of *Lecture Notes in Computer Science*, pp. 121–130, Springer.

[21] Hordijk, W., Shalizi, C. R., and Crutchfield, J. P. (2001) Upper bound on the products of particle interactions in cellular automata. *Physica D*, **154**, 240–258.

[22] Kitzbichler, M. G., Smith, M. L., Christensen, S. R., and Bullmore, E. (2009) Broadband criticality of human brain network synchronization. *PLoS Computational Biology*, **5**, e1000314.

[23] Klyubin, A. S., Polani, D., and Nehaniv, C. L. (2004) Tracking information flow through the environment: Simple cases of stigmergy. Pollack, J., Bedau, M., Husbands, P., Ikegami, T., and Watson, R. A. (eds.), *Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems (ALife IX), Boston, USA*, Cambridge, MA, USA, pp. 563–568, MIT Press.

[24] Korošec, P., Šilc, J., and Filipič, B. (2012) The differential ant-stigmergy algorithm. *Information Sciences*, **192**, 82–97.

[25] Langton, C. G. (1990) Computation at the edge of chaos: phase transitions and emergent computation. *Physica D*, **42**, 12–37.

[26] Lindgren, K. and Nordahl, M. G. (1988) Complexity measures and cellular automata. *Complex Systems*, **2**, 409–440.

[27] Lizier, J. T., Atay, F. M., and Jost, J. (2011) Information storage, loop motifs and clustered structure in complex networks, under submission.

[28] Lizier, J. T., Pritam, S., and Prokopenko, M. (2011) Information dynamics in small-world Boolean networks. *Artificial Life*, **17**, 293–314.

[29] Lizier, J. T. and Prokopenko, M. (2010) Differentiating information transfer and causal effect. *European Physical Journal B*, **73**, 605–615.

[30] Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2007) Detecting non-trivial computation in complex dynamics. Almeida e Costa, F., Rocha, L. M., Costa, E., Harvey, I., and Coutinho, A. (eds.), *Proceedings of the 9th European Conference on Artificial Life (ECAL 2007), Lisbon, Portugal*, Berlin / Heidelberg, vol. 4648 of *Lecture Notes in Artificial Intelligence*, pp. 895–904, Springer.

[31] Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2008) Local information transfer as a spatiotemporal filter for complex systems. *Physical Review E*, **77**, 026110.

[32] Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2010) Information modification and particle collisions in distributed computation. *Chaos*, **20**, 037109.

[33] Lizier, J. T., Prokopenko, M., and Zomaya, A. Y. (2012) Coherent information structure in complex computation. *Theory in Biosciences*, in press.

[34] Lizier, J. T. (2010) *The local information dynamics of distributed computation in complex systems*. Ph.D. thesis, The University of Sydney.

[35] Lungarella, M., Pegors, T., Bulwinkle, D., and Sporns, O. (2005) Methods for quantifying the informational structure of sensory and motor data. *Neuroinformatics*, **3**, 243–262.

[36] MacKay, D. J. (2003) *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

[37] Maji, P. and Pal Chaudhuri, P. (2008) Non-uniform cellular automata based associative memory: Evolutionary design and basins of attraction. *Information Sciences*, **178**, 2315–2336.

[38] Marton, K. and Shields, P. C. (1994) Entropy and the consistent estimation of joint distributions. *The Annals of Probability*, **22**, 960–977.

[39] Mitchell, M. (1998) Computation in cellular automata: A selected review. Gramss, T., Bornholdt, S., Gross, M., Mitchell, M., and Pellizzari, T. (eds.), *Non-Standard Computation*, pp. 95–140, VCH Verlagsgesellschaft.

[40] Mitchell, M., Crutchfield, J. P., and Das, R. (1996) Evolving cellular automata with genetic algorithms: A review of recent work. Goodman, E. D., Punch, W., and Uskov, V. (eds.), *Proceedings of the First International Conference on Evolutionary Computation and Its Applications, Moscow*, Russia, Russian Academy of Sciences.

[41] Mitchell, M., Crutchfield, J. P., and Hraber, P. T. (1994) Evolving cellular automata to perform computations: Mechanisms and impediments. *Physica D*, **75**, 361–391.

[42] Morgado, R., Cieśla, M., Longa, L., and Oliveira, F. A. (2007) Synchronization in the presence of memory. *Europhysics Letters*, **79**, 10002.

[43] Prokopenko, M., Gerasimov, V., and Tanev, I. (2006) Evolving spatiotemporal coordination in a modular robotic system. Nolfi, S., Baldassarre, G., Calabretta, R., Hallam, J., Marocco, D., Meyer, J.-A., and Parisi, D. (eds.), *Proceedings of the Ninth International Conference on the Simulation of Adaptive Behavior (SAB'06), Rome*, vol. 4095 of *Lecture Notes in Artificial Intelligence*, pp. 548–559, Springer Verlag.

[44] Shalizi, C. R. (2001) *Causal Architecture, Complexity and Self-Organization in Time Series and Cellular Automata*. Ph.D. thesis, University of Wisconsin-Madison.

[45] Shalizi, C. R., Haslinger, R., Rouquier, J.-B., Klinkner, K. L., and Moore, C. (2006) Automatic filters for the detection of coherent structure in spatiotemporal systems. *Physical Review E*, **73**, 036104.

[46] Sporns, O. (2011) *Networks of the brain*. MIT Press.

[47] Tononi, G., Sporns, O., and Edelman, G. (1994) A measure for brain complexity: Relating functional segregation and integration in the nervous system. *Proceedings of the National Academy of Sciences*, **91**, 5033–5037.

[48] Williams, P. L. and Beer, R. D. (2010) Nonnegative decomposition of multivariate information. arXiv:1004.2515.

[49] Williams, P. L. and Beer, R. D. (2011) Generalized measures of information transfer. arXiv:1102.1507.

[50] Wójtowicz, M. (2002), Java cellebration v.1.50. Online software.

[51] Wolfram, S. (2002) *A New Kind of Science*. Wolfram Media.

[52] Wuensche, A. (1999) Classifying cellular automata automatically: Finding gliders, filtering, and relating spacetime patterns, attractor basins, and the Z parameter. *Complexity*, **4**, 47–66.