

Between order and chaos

James P. Crutchfield

What is a pattern? How do we come to recognize patterns never seen before? Quantifying the notion of pattern and formalizing the process of pattern discovery go right to the heart of physical science. Over the past few decades physics' view of nature's lack of structure—its unpredictability—underwent a major renovation with the discovery of deterministic chaos, overthrowing two centuries of Laplace's strict determinism in classical physics. Behind the veil of apparent randomness, though, many processes are highly ordered, following simple rules. Tools adapted from the theories of information and computation have brought physical science to the brink of automatically discovering hidden patterns and quantifying their structural complexity.

One designs clocks to be as regular as physically possible. So much so that they are the very instruments of determinism. The coin flip plays a similar role; it expresses our ideal of the utterly unpredictable. Randomness is as necessary to physics as determinism—think of the essential role that 'molecular chaos' plays in establishing the existence of thermodynamic states. The clock and the coin flip, as such, are mathematical ideals to which reality is often unkind. The extreme difficulties of engineering the perfect clock¹ and implementing a source of randomness as pure as the fair coin testify to the fact that determinism and randomness are two inherent aspects of all physical processes.

In 1927, van der Pol, a Dutch engineer, listened to the tones produced by a neon glow lamp coupled to an oscillating electrical circuit. Lacking modern electronic test equipment, he monitored the circuit's behaviour by listening through a telephone ear piece. In what is probably one of the earlier experiments on electronic music, he discovered that, by tuning the circuit as if it were a musical instrument, fractions or subharmonics of a fundamental tone could be produced. This is markedly unlike common musical instruments—such as the flute, which is known for its purity of harmonics, or multiples of a fundamental tone. As van der Pol and a colleague reported in *Nature* that year², 'the turning of the condenser in the region of the third to the sixth subharmonic strongly reminds one of the tunes of a bag pipe'.

Presciently, the experimenters noted that when tuning the circuit 'often an irregular noise is heard in the telephone receivers before the frequency jumps to the next lower value'. We now know that van der Pol had listened to deterministic chaos: the noise was produced in an entirely lawful, ordered way by the circuit itself. The *Nature* report stands as one of its first experimental discoveries. Van der Pol and his colleague van der Mark apparently were unaware that the deterministic mechanisms underlying the noises they had heard had been rather keenly analysed three decades earlier by the French mathematician Poincaré in his efforts to establish the orderliness of planetary motion^{3–5}. Poincaré failed at this, but went on to establish that determinism and randomness are essential and unavoidable twins⁶. Indeed, this duality is succinctly expressed in the two familiar phrases 'statistical mechanics' and 'deterministic chaos'.

Complicated yes, but is it complex?

As for van der Pol and van der Mark, much of our appreciation of nature depends on whether our minds—or, more typically these days, our computers—are prepared to discern its intricacies. When confronted by a phenomenon for which we are ill-prepared, we often simply fail to see it, although we may be looking directly at it.

Perception is made all the more problematic when the phenomena of interest arise in systems that spontaneously organize.

Spontaneous organization, as a common phenomenon, reminds us of a more basic, nagging puzzle. If, as Poincaré found, chaos is endemic to dynamics, why is the world not a mass of randomness? The world is, in fact, quite structured, and we now know several of the mechanisms that shape microscopic fluctuations as they are amplified to macroscopic patterns. Critical phenomena in statistical mechanics⁷ and pattern formation in dynamics^{8,9} are two arenas that explain in predictive detail how spontaneous organization works. Moreover, everyday experience shows us that nature inherently organizes; it generates pattern. Pattern is as much the fabric of life as life's unpredictability.

In contrast to patterns, the outcome of an observation of a random system is unexpected. We are surprised at the next measurement. That surprise gives us information about the system. We must keep observing the system to see how it is evolving. This insight about the connection between randomness and surprise was made operational, and formed the basis of the modern theory of communication, by Shannon in the 1940s (ref. 10). Given a source of random events and their probabilities, Shannon defined a particular event's degree of surprise as the negative logarithm of its probability: the event's self-information is $I_i = -\log_2 p_i$. (The units when using the base-2 logarithm are bits.) In this way, an event, say i , that is certain ($p_i = 1$) is not surprising: $I_i = 0$ bits. Repeated measurements are not informative. Conversely, a flip of a fair coin ($p_{\text{Heads}} = 1/2$) is maximally informative: for example, $I_{\text{Heads}} = 1$ bit. With each observation we learn in which of two orientations the coin is, as it lays on the table.

The theory describes an information source: a random variable X consisting of a set $\{i = 0, 1, \dots, k\}$ of events and their probabilities $\{p_i\}$. Shannon showed that the averaged uncertainty $H[X] = \sum_i p_i I_i$ —the source entropy rate—is a fundamental property that determines how compressible an information source's outcomes are.

With information defined, Shannon laid out the basic principles of communication¹¹. He defined a communication channel that accepts messages from an information source X and transmits them, perhaps corrupting them, to a receiver who observes the channel output Y . To monitor the accuracy of the transmission, he introduced the mutual information $I[X; Y] = H[X] - H[X|Y]$ between the input and output variables. The first term is the information available at the channel's input. The second term, subtracted, is the uncertainty in the incoming message, if the receiver knows the output. If the channel completely corrupts, so

that none of the source messages accurately appears at the channel's output, then knowing the output Y tells you nothing about the input and $H[X|Y] = H[X]$. In other words, the variables are statistically independent and so the mutual information vanishes. If the channel has perfect fidelity, then the input and output variables are identical; what goes in, comes out. The mutual information is the largest possible: $I[X; Y] = H[X]$, because $H[X|Y] = 0$. The maximum input–output mutual information, over all possible input sources, characterizes the channel itself and is called the channel capacity:

$$C = \max_{P(X)} I[X; Y]$$

Shannon's most famous and enduring discovery though—one that launched much of the information revolution—is that as long as a (potentially noisy) channel's capacity C is larger than the information source's entropy rate $H[X]$, there is way to encode the incoming messages such that they can be transmitted error free¹¹. Thus, information and how it is communicated were given firm foundation.

How does information theory apply to physical systems? Let us set the stage. The system to which we refer is simply the entity we seek to understand by way of making observations. The collection of the system's temporal behaviours is the process it generates. We denote a particular realization by a time series of measurements: $\dots x_{-2}x_{-1}x_0x_1\dots$. The values x_t taken at each time can be continuous or discrete. The associated bi-infinite chain of random variables is similarly denoted, except using uppercase: $\dots X_{-2}X_{-1}X_0X_1\dots$. At each time t the chain has a past $X_t = \dots X_{t-2}X_{t-1}$ and a future $X_t = X_tX_{t+1}\dots$. We will also refer to blocks $X_{t'} = X_tX_{t+1}\dots X_{t'-1}$, $t < t'$. The upper index is exclusive.

To apply information theory to general stationary processes, one uses Kolmogorov's extension of the source entropy rate^{12,13}. This is the growth rate h_μ :

$$h_\mu = \lim_{\ell \rightarrow \infty} \frac{H(\ell)}{\ell}$$

where $H(\ell) = -\sum_{\{x_\ell\}} \Pr(x_\ell) \log_2 \Pr(x_\ell)$ is the block entropy—the Shannon entropy of the length- ℓ word distribution $\Pr(x_\ell)$. h_μ gives the source's intrinsic randomness, discounting correlations that occur over any length scale. Its units are bits per symbol, and it partly elucidates one aspect of complexity—the randomness generated by physical systems.

We now think of randomness as surprise and measure its degree using Shannon's entropy rate. By the same token, can we say what 'pattern' is? This is more challenging, although we know organization when we see it.

Perhaps one of the more compelling cases of organization is the hierarchy of distinctly structured matter that separates the sciences—quarks, nucleons, atoms, molecules, materials and so on. This puzzle interested Philip Anderson, who in his early essay 'More is different'¹⁴, notes that new levels of organization are built out of the elements at a lower level and that the new 'emergent' properties are distinct. They are not directly determined by the physics of the lower level. They have their own 'physics'.

This suggestion too raises questions, what is a 'level' and how different do two levels need to be? Anderson suggested that organization at a given level is related to the history or the amount of effort required to produce it from the lower level. As we will see, this can be made operational.

Complexities

To arrive at that destination we make two main assumptions. First, we borrow heavily from Shannon: every process is a communication channel. In particular, we posit that any system is a channel that

communicates its past to its future through its present. Second, we take into account the context of interpretation. We view building models as akin to decrypting nature's secrets. How do we come to understand a system's randomness and organization, given only the available, indirect measurements that an instrument provides? To answer this, we borrow again from Shannon, viewing model building also in terms of a channel: one experimentalist attempts to explain her results to another.

The following first reviews an approach to complexity that models system behaviours using exact deterministic representations. This leads to the deterministic complexity and we will see how it allows us to measure degrees of randomness. After describing its features and pointing out several limitations, these ideas are extended to measuring the complexity of ensembles of behaviours—to what we now call statistical complexity. As we will see, it measures degrees of structural organization. Despite their different goals, the deterministic and statistical complexities are related and we will see how they are essentially complementary in physical systems.

Solving Hilbert's famous *Entscheidungsproblem* challenge to automate testing the truth of mathematical statements, Turing introduced a mechanistic approach to an effective procedure that could decide their validity¹⁵. The model of computation he introduced, now called the Turing machine, consists of an infinite tape that stores symbols and a finite-state controller that sequentially reads symbols from the tape and writes symbols to it. Turing's machine is deterministic in the particular sense that the tape contents exactly determine the machine's behaviour. Given the present state of the controller and the next symbol read off the tape, the controller goes to a unique next state, writing at most one symbol to the tape. The input determines the next step of the machine and, in fact, the tape input determines the entire sequence of steps the Turing machine goes through.

Turing's surprising result was that there existed a Turing machine that could compute any input–output function—it was universal. The deterministic universal Turing machine (UTM) thus became a benchmark for computational processes.

Perhaps not surprisingly, this raised a new puzzle for the origins of randomness. Operating from a fixed input, could a UTM generate randomness, or would its deterministic nature always show through, leading to outputs that were probabilistically deficient? More ambitiously, could probability theory itself be framed in terms of this new constructive theory of computation? In the early 1960s these and related questions led a number of mathematicians—Solomonoff^{16,17} (an early presentation of his ideas appears in ref. 18), Chaitin¹⁹, Kolmogorov²⁰ and Martin-Löf²¹—to develop the algorithmic foundations of randomness.

The central question was how to define the probability of a single object. More formally, could a UTM generate a string of symbols that satisfied the statistical properties of randomness? The approach declares that models M should be expressed in the language of UTM programs. This led to the Kolmogorov–Chaitin complexity $KC(\mathbf{x})$ of a string \mathbf{x} . The Kolmogorov–Chaitin complexity is the size of the minimal program P that generates \mathbf{x} running on a UTM (refs 19,20):

$$KC(\mathbf{x}) = \operatorname{argmin}\{|P| : UTM \circ P = \mathbf{x}\}$$

One consequence of this should sound quite familiar by now. It means that a string is random when it cannot be compressed: a random string is its own minimal program. The Turing machine simply prints it out. A string that repeats a fixed block of letters, in contrast, has small Kolmogorov–Chaitin complexity. The Turing machine program consists of the block and the number of times it is to be printed. Its Kolmogorov–Chaitin complexity is logarithmic

in the desired string length, because there is only one variable part of P and it stores $\log \ell$ digits of the repetition count ℓ .

Unfortunately, there are a number of deep problems with deploying this theory in a way that is useful to describing the complexity of physical systems.

First, Kolmogorov–Chaitin complexity is not a measure of structure. It requires exact replication of the target string. Therefore, $KC(\mathbf{x})$ inherits the property of being dominated by the randomness in \mathbf{x} . Specifically, many of the UTM instructions that get executed in generating \mathbf{x} are devoted to producing the ‘random’ bits of \mathbf{x} . The conclusion is that Kolmogorov–Chaitin complexity is a measure of randomness, not a measure of structure. One solution, familiar in the physical sciences, is to discount for randomness by describing the complexity in ensembles of behaviours.

Furthermore, focusing on single objects was a feature, not a bug, of Kolmogorov–Chaitin complexity. In the physical sciences, however, this is a prescription for confusion. We often have access only to a system’s typical properties, and even if we had access to microscopic, detailed observations, listing the positions and momenta of molecules is simply too huge, and so useless, a description of a box of gas. In most cases, it is better to know the temperature, pressure and volume.

The issue is more fundamental than sheer system size, arising even with a few degrees of freedom. Concretely, the unpredictability of deterministic chaos forces the ensemble approach on us.

The solution to the Kolmogorov–Chaitin complexity’s focus on single objects is to define the complexity of a system’s process—the ensemble of its behaviours²². Consider an information source that produces collections of strings of arbitrary length. Given a realization $x_{:\ell}$ of length ℓ , we have its Kolmogorov–Chaitin complexity $KC(x_{:\ell})$, of course, but what can we say about the Kolmogorov–Chaitin complexity of the ensemble $\{x_{:\ell}\}$? First, define its average in terms of samples $\{x_{:\ell}^i : i = 1, \dots, M\}$:

$$KC(\ell) = \langle KC(x_{:\ell}) \rangle = \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{i=1}^M KC(x_{:\ell}^i)$$

How does the Kolmogorov–Chaitin complexity grow as a function of increasing string length? For almost all infinite sequences produced by a stationary process the growth rate of the Kolmogorov–Chaitin complexity is the Shannon entropy rate²³:

$$h_\mu = \lim_{\ell \rightarrow \infty} \frac{KC(\ell)}{\ell}$$

As a measure—that is, a number used to quantify a system property—Kolmogorov–Chaitin complexity is uncomputable^{24,25}. There is no algorithm that, taking in the string, computes its Kolmogorov–Chaitin complexity. Fortunately, this problem is easily diagnosed. The essential uncomputability of Kolmogorov–Chaitin complexity derives directly from the theory’s clever choice of a UTM as the model class, which is so powerful that it can express undecidable statements.

One approach to making a complexity measure constructive is to select a less capable (specifically, non-universal) class of computational models. We can declare the representations to be, for example, the class of stochastic finite-state automata^{26,27}. The result is a measure of randomness that is calibrated relative to this choice. Thus, what one gains in constructiveness, one loses in generality.

Beyond uncomputability, there is the more vexing issue of how well that choice matches a physical system of interest. Even if, as just described, one removes uncomputability by choosing a less capable representational class, one still must validate that these, now rather specific, choices are appropriate to the physical system one is analysing.

At the most basic level, the Turing machine uses discrete symbols and advances in discrete time steps. Are these representational choices appropriate to the complexity of physical systems? What about systems that are inherently noisy, those whose variables are continuous or are quantum mechanical? Appropriate theories of computation have been developed for each of these cases^{28,29}, although the original model goes back to Shannon³⁰. More to the point, though, do the elementary components of the chosen representational scheme match those out of which the system itself is built? If not, then the resulting measure of complexity will be misleading.

Is there a way to extract the appropriate representation from the system’s behaviour, rather than having to impose it? The answer comes, not from computation and information theories, as above, but from dynamical systems theory.

Dynamical systems theory—Poincaré’s qualitative dynamics—emerged from the patent uselessness of offering up an explicit list of an ensemble of trajectories, as a description of a chaotic system. It led to the invention of methods to extract the system’s ‘geometry from a time series’. One goal was to test the strange-attractor hypothesis put forward by Ruelle and Takens to explain the complex motions of turbulent fluids³¹.

How does one find the chaotic attractor given a measurement time series from only a single observable? Packard and others proposed developing the reconstructed state space from successive time derivatives of the signal³². Given a scalar time series $x(t)$, the reconstructed state space uses coordinates $y_1(t) = x(t)$, $y_2(t) = dx(t)/dt, \dots, y_m(t) = d^m x(t)/dt^m$. Here, $m + 1$ is the embedding dimension, chosen large enough that the dynamic in the reconstructed state space is deterministic. An alternative is to take successive time delays in $x(t)$ (ref. 33). Using these methods the strange attractor hypothesis was eventually verified³⁴.

It is a short step, once one has reconstructed the state space underlying a chaotic signal, to determine whether you can also extract the equations of motion themselves. That is, does the signal tell you which differential equations it obeys? The answer is yes³⁵. This sound works quite well if, and this will be familiar, one has made the right choice of representation for the ‘right-hand side’ of the differential equations. Should one use polynomial, Fourier or wavelet basis functions; or an artificial neural net? Guess the right representation and estimating the equations of motion reduces to statistical quadrature: parameter estimation and a search to find the lowest embedding dimension. Guess wrong, though, and there is little or no clue about how to update your choice.

The answer to this conundrum became the starting point for an alternative approach to complexity—one more suitable for physical systems. The answer is articulated in computational mechanics³⁶, an extension of statistical mechanics that describes not only a system’s statistical properties but also how it stores and processes information—how it computes.

The theory begins simply by focusing on predicting a time series $\dots X_{-2} X_{-1} X_0 X_1 \dots$. In the most general setting, a prediction is a distribution $\Pr(X_t | x_{:t})$ of futures $X_t = X_t X_{t+1} X_{t+2} \dots$ conditioned on a particular past $x_{:t} = \dots x_{t-3} x_{t-2} x_{t-1}$. Given these conditional distributions one can predict everything that is predictable about the system.

At root, extracting a process’s representation is a very straightforward notion: do not distinguish histories that make the same predictions. Once we group histories in this way, the groups themselves capture the relevant information for predicting the future. This leads directly to the central definition of a process’s effective states. They are determined by the equivalence relation:

$$x_{:t} \sim x_{:t'} \Leftrightarrow \Pr(X_t | x_{:t}) = \Pr(X_t | x_{:t'})$$

The equivalence classes of the relation \sim are the process's causal states \mathcal{S} —literally, its reconstructed state space, and the induced state-to-state transitions are the process's dynamic \mathcal{T} —its equations of motion. Together, the states \mathcal{S} and dynamic \mathcal{T} give the process's so-called ϵ -machine.

Why should one use the ϵ -machine representation of a process? First, there are three optimality theorems that say it captures all of the process's properties^{36–38}: prediction: a process's ϵ -machine is its optimal predictor; minimality: compared with all other optimal predictors, a process's ϵ -machine is its minimal representation; uniqueness: any minimal optimal predictor is equivalent to the ϵ -machine.

Second, we can immediately (and accurately) calculate the system's degree of randomness. That is, the Shannon entropy rate is given directly in terms of the ϵ -machine:

$$h_\mu = - \sum_{\sigma \in \mathcal{S}} \Pr(\sigma) \sum_{\{x\}} \Pr(x|\sigma) \log_2 \Pr(x|\sigma)$$

where $\Pr(\sigma)$ is the distribution over causal states and $\Pr(x|\sigma)$ is the probability of transitioning from state σ on measurement x .

Third, the ϵ -machine gives us a new property—the statistical complexity—and it, too, is directly calculated from the ϵ -machine:

$$C_\mu = - \sum_{\sigma \in \mathcal{S}} \Pr(\sigma) \log_2 \Pr(\sigma)$$

The units are bits. This is the amount of information the process stores in its causal states.

Fourth, perhaps the most important property is that the ϵ -machine gives all of a process's patterns. The ϵ -machine itself—states plus dynamic—gives the symmetries and regularities of the system. Mathematically, it forms a semi-group³⁹. Just as groups characterize the exact symmetries in a system, the ϵ -machine captures those and also 'partial' or noisy symmetries.

Finally, there is one more unique improvement the statistical complexity makes over Kolmogorov–Chaitin complexity theory. The statistical complexity has an essential kind of representational independence. The causal equivalence relation, in effect, extracts the representation from a process's behaviour. Causal equivalence can be applied to any class of system—continuous, quantum, stochastic or discrete.

Independence from selecting a representation achieves the intuitive goal of using UTMs in algorithmic information theory—the choice that, in the end, was the latter's undoing. The ϵ -machine does not suffer from the latter's problems. In this sense, computational mechanics is less subjective than any 'complexity' theory that per force chooses a particular representational scheme.

To summarize, the statistical complexity defined in terms of the ϵ -machine solves the main problems of the Kolmogorov–Chaitin complexity by being representation independent, constructive, the complexity of an ensemble, and a measure of structure.

In these ways, the ϵ -machine gives a baseline against which any measures of complexity or modelling, in general, should be compared. It is a minimal sufficient statistic³⁸.

To address one remaining question, let us make explicit the connection between the deterministic complexity framework and that of computational mechanics and its statistical complexity. Consider realizations $\{x_\ell\}$ from a given information source. Break the minimal UTM program P for each into two components: one that does not change, call it the 'model' M ; and one that does change from input to input, E , the 'random' bits not generated by M . Then, an object's 'sophistication' is the length of M (refs 40,41):

$$\text{SOPH}(x_\ell) = \text{argmin}\{|M| : P = M + E, x_\ell = \text{UTM} \circ P\}$$

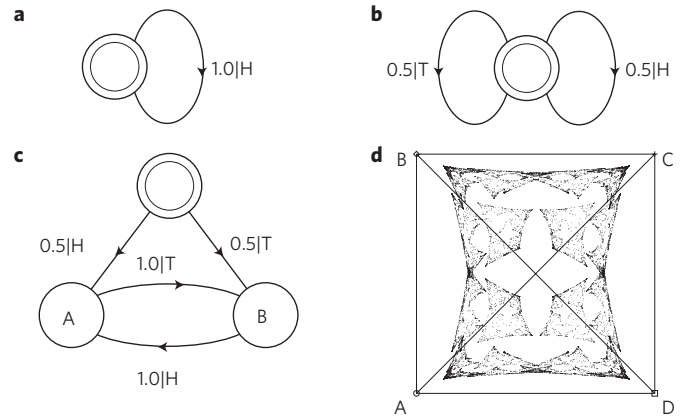


Figure 1 | ϵ -machines for four information sources. **a**, The all-heads process is modelled with a single state and a single transition. The transition is labelled $p|x$, where $p \in [0; 1]$ is the probability of the transition and x is the symbol emitted. **b**, The fair-coin process is also modelled by a single state, but with two transitions each chosen with equal probability. **c**, The period-2 process is perhaps surprisingly more involved. It has three states and several transitions. **d**, The uncountable set of causal states for a generic four-state HMM. The causal states here are distributions $\Pr(A; B; C; D)$ over the HMM's internal states and so are plotted as points in a 4-simplex spanned by the vectors that give each state unit probability. Panel **d** reproduced with permission from ref. 44, © 1994 Elsevier.

As done with the Kolmogorov–Chaitin complexity, we can define the ensemble-averaged sophistication (SOPH) of 'typical' realizations generated by the source. The result is that the average sophistication of an information source is proportional to its process's statistical complexity⁴²:

$$KC(\ell) \propto C_\mu + h_\mu \ell$$

That is, $\langle \text{SOPH} \rangle \propto C_\mu$.

Notice how far we come in computational mechanics by positing only the causal equivalence relation. From it alone, we derive many of the desired, sometimes assumed, features of other complexity frameworks. We have a canonical representational scheme. It is minimal and so Ockham's razor⁴³ is a consequence, not an assumption. We capture a system's pattern in the algebraic structure of the ϵ -machine. We define randomness as a process's ϵ -machine Shannon-entropy rate. We define the amount of organization in a process with its ϵ -machine's statistical complexity. In addition, we also see how the framework of deterministic complexity relates to computational mechanics.

Applications

Let us address the question of usefulness of the foregoing by way of examples.

Let's start with the Prediction Game, an interactive pedagogical tool that intuitively introduces the basic ideas of statistical complexity and how it differs from randomness. The first step presents a data sample, usually a binary times series. The second asks someone to predict the future, on the basis of that data. The final step asks someone to posit a state-based model of the mechanism that generated the data.

The first data set to consider is $x_{:0} = \dots \text{HHHHHHHH}$ —the all-heads process. The answer to the prediction question comes to mind immediately: the future will be all Hs, $x_{:1} = \text{HHHHH}\dots$. Similarly, a guess at a state-based model of the generating mechanism is also easy. It is a single state with a transition labelled with the output symbol H (Fig. 1a). A simple model for a simple process. The process is exactly predictable: $h_\mu = 0$

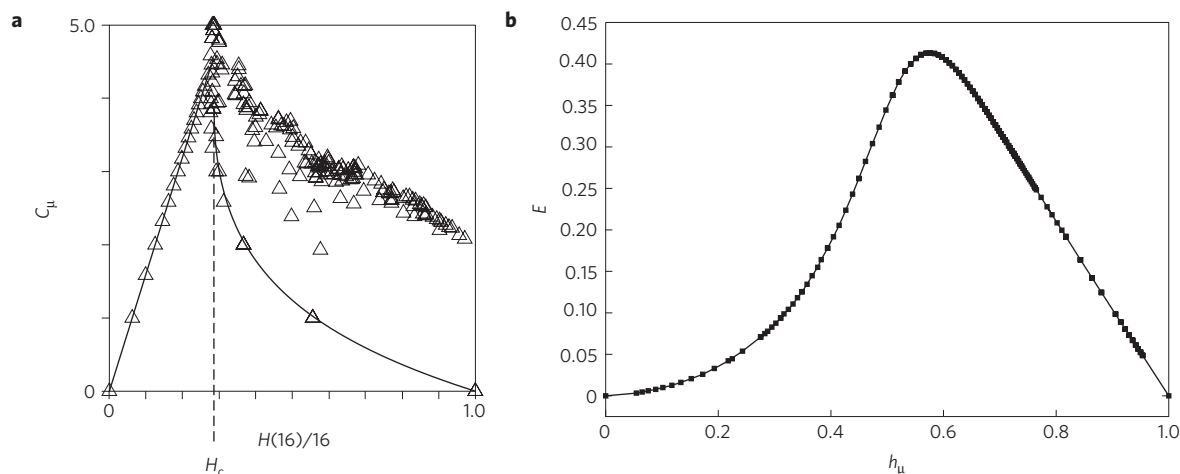


Figure 2 | Structure versus randomness. **a**, In the period-doubling route to chaos. **b**, In the two-dimensional Ising-spinsystem. Reproduced with permission from: **a**, ref. 36, © 1989 APS; **b**, ref. 61, © 2008 AIP.

bits per symbol. Furthermore, it is not complex; it has vanishing complexity: $C_\mu = 0$ bits.

The second data set is, for example, $x_0 = \dots\text{THTHTTHTHH}$. What I have done here is simply flip a coin several times and report the results. Shifting from being confident and perhaps slightly bored with the previous example, people take notice and spend a good deal more time pondering the data than in the first case.

The prediction question now brings up a number of issues. One cannot exactly predict the future. At best, one will be right only half of the time. Therefore, a legitimate prediction is simply to give another series of flips from a fair coin. In terms of monitoring only errors in prediction, one could also respond with a series of all Hs. Trivially right half the time, too. However, this answer gets other properties wrong, such as the simple facts that Ts occur and occur in equal number.

The answer to the modelling question helps articulate these issues with predicting (Fig. 1b). The model has a single state, now with two transitions: one labelled with a T and one with an H. They are taken with equal probability. There are several points to emphasize. Unlike the all-heads process, this one is maximally unpredictable: $h_\mu = 1$ bit/symbol. Like the all-heads process, though, it is simple: $C_\mu = 0$ bits again. Note that the model is minimal. One cannot remove a single ‘component’, state or transition, and still do prediction. The fair coin is an example of an independent, identically distributed process. For all independent, identically distributed processes, $C_\mu = 0$ bits.

In the third example, the past data are $x_0 = \dots\text{HTHTHTHTH}$. This is the period-2 process. Prediction is relatively easy, once one has discerned the repeated template word $w = \text{TH}$. The prediction is $x_1 = \text{THTHTHTH} \dots$. The subtlety now comes in answering the modelling question (Fig. 1c).

There are three causal states. This requires some explanation. The state at the top has a double circle. This indicates that it is a start state—the state in which the process starts or, from an observer’s point of view, the state in which the observer is before it begins measuring. We see that its outgoing transitions are chosen with equal probability and so, on the first step, a T or an H is produced with equal likelihood. An observer has no ability to predict which. That is, initially it looks like the fair-coin process. The observer receives 1 bit of information. In this case, once this start state is left, it is never visited again. It is a transient causal state.

Beyond the first measurement, though, the ‘phase’ of the period-2 oscillation is determined, and the process has moved into its two recurrent causal states. If an H occurred, then it

is in state A and a T will be produced next with probability 1. Conversely, if a T was generated, it is in state B and then an H will be generated. From this point forward, the process is exactly predictable: $h_\mu = 0$ bits per symbol. In contrast to the first two cases, it is a structurally complex process: $C_\mu = 1$ bit. Conditioning on histories of increasing length gives the distinct future conditional distributions corresponding to these three states. Generally, for p -periodic processes $h_\mu = 0$ bits symbol⁻¹ and $C_\mu = \log_2 p$ bits.

Finally, Fig. 1d gives the ϵ -machine for a process generated by a generic hidden-Markov model (HMM). This example helps dispel the impression given by the Prediction Game examples that ϵ -machines are merely stochastic finite-state machines. This example shows that there can be a fractional dimension set of causal states. It also illustrates the general case for HMMs. The statistical complexity diverges and so we measure its rate of divergence—the causal states’ information dimension⁴⁴.

As a second example, let us consider a concrete experimental application of computational mechanics to one of the venerable fields of twentieth-century physics—crystallography: how to find structure in disordered materials. The possibility of turbulent crystals had been proposed a number of years ago by Ruelle⁵³. Using the ϵ -machine we recently reduced this idea to practice by developing a crystallography for complex materials^{54–57}.

Describing the structure of solids—simply meaning the placement of atoms in (say) a crystal—is essential to a detailed understanding of material properties. Crystallography has long used the sharp Bragg peaks in X-ray diffraction spectra to infer crystal structure. For those cases where there is diffuse scattering, however, finding—let alone describing—the structure of a solid has been more difficult⁵⁸. Indeed, it is known that without the assumption of crystallinity, the inference problem has no unique solution⁵⁹. Moreover, diffuse scattering implies that a solid’s structure deviates from strict crystallinity. Such deviations can come in many forms—Schottky defects, substitution impurities, line dislocations and planar disorder, to name a few.

The application of computational mechanics solved the longstanding problem—determining structural information for disordered materials from their diffraction spectra—for the special case of planar disorder in close-packed structures in polytypes⁶⁰. The solution provides the most complete statistical description of the disorder and, from it, one could estimate the minimum effective memory length for stacking sequences in close-packed structures. This approach was contrasted with the so-called fault

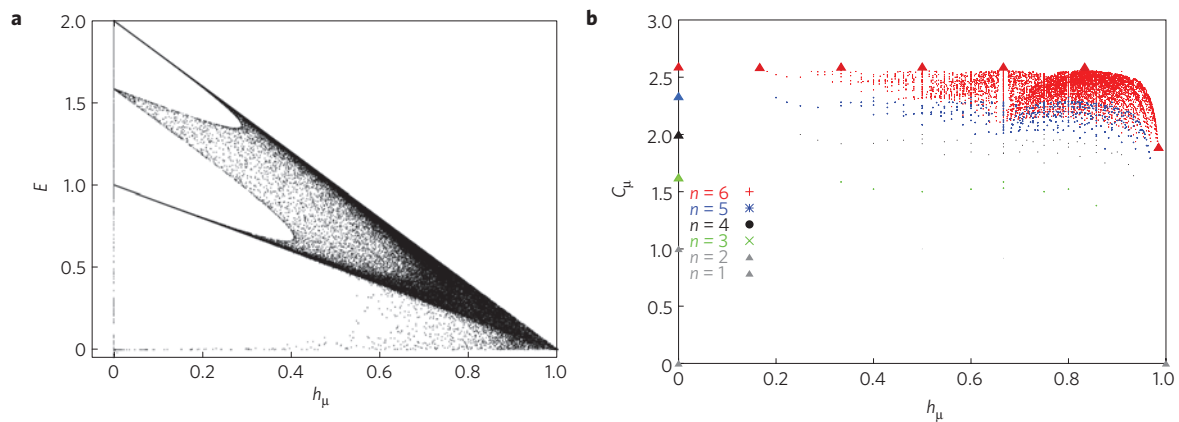


Figure 3 | Complexity–entropy diagrams. **a**, The one-dimensional, spin-1/2 antiferromagnetic Ising model with nearest- and next-nearest-neighbour interactions. Reproduced with permission from ref. 61, © 2008 AIP. **b**, Complexity–entropy pairs (h_μ, C_μ) for all topological binary-alphabet ϵ -machines with $n = 1, \dots, 6$ states. For details, see refs 61 and 63.

model by comparing the structures inferred using both approaches on two previously published zinc sulphide diffraction spectra. The net result was that having an operational concept of pattern led to a predictive theory of structure in disordered materials.

As a further example, let us explore the nature of the interplay between randomness and structure across a range of processes. As a direct way to address this, let us examine two families of controlled system—systems that exhibit phase transitions. Consider the randomness and structure in two now-familiar systems: one from nonlinear dynamics—the period-doubling route to chaos; and the other from statistical mechanics—the two-dimensional Ising-spin model. The results are shown in the complexity–entropy diagrams of Fig. 2. They plot a measure of complexity (C_μ and E) versus the randomness ($H(16)/16$ and h_μ , respectively).

One conclusion is that, in these two families at least, the intrinsic computational capacity is maximized at a phase transition: the onset of chaos and the critical temperature. The occurrence of this behaviour in such prototype systems led a number of researchers to conjecture that this was a universal interdependence between randomness and structure. For quite some time, in fact, there was hope that there was a single, universal complexity–entropy function—coined the ‘edge of chaos’ (but consider the issues raised in ref. 62). We now know that although this may occur in particular classes of system, it is not universal.

It turned out, though, that the general situation is much more interesting⁶¹. Complexity–entropy diagrams for two other process families are given in Fig. 3. These are rather less universal looking. The diversity of complexity–entropy behaviours might seem to indicate an unhelpful level of complication. However, we now see that this is quite useful. The conclusion is that there is a wide range of intrinsic computation available to nature to exploit and available to us to engineer.

Finally, let us return to address Anderson’s proposal for nature’s organizational hierarchy. The idea was that a new, ‘higher’ level is built out of properties that emerge from a relatively ‘lower’ level’s behaviour. He was particularly interested to emphasize that the new level had a new ‘physics’ not present at lower levels. However, what is a ‘level’ and how different should a higher level be from a lower one to be seen as new?

We can address these questions now having a concrete notion of structure, captured by the ϵ -machine, and a way to measure it, the statistical complexity C_μ . In line with the theme so far, let us answer these seemingly abstract questions by example. It turns out that we already saw an example of hierarchy, when discussing intrinsic computation at phase transitions.

Specifically, higher-level computation emerges at the onset of chaos through period-doubling—a countably infinite state ϵ -machine⁴²—at the peak of C_μ in Fig. 2a.

How is this hierarchical? We answer this using a generalization of the causal equivalence relation. The lowest level of description is the raw behaviour of the system at the onset of chaos. Appealing to symbolic dynamics⁶⁴, this is completely described by an infinitely long binary string. We move to a new level when we attempt to determine its ϵ -machine. We find, at this ‘state’ level, a countably infinite number of causal states. Although faithful representations, models with an infinite number of components are not only cumbersome, but not insightful. The solution is to apply causal equivalence yet again—to the ϵ -machine’s causal states themselves. This produces a new model, consisting of ‘meta-causal states’, that predicts the behaviour of the causal states themselves. This procedure is called hierarchical ϵ -machine reconstruction⁴⁵, and it leads to a finite representation—a nested-stack automaton⁴². From this representation we can directly calculate many properties that appear at the onset of chaos.

Notice, though, that in this prescription the statistical complexity at the ‘state’ level diverges. Careful reflection shows that this also occurred in going from the raw symbol data, which were an infinite non-repeating string (of binary ‘measurement states’), to the causal states. Conversely, in the case of an infinitely repeated block, there is no need to move up to the level of causal states. At the period-doubling onset of chaos the behaviour is aperiodic, although not chaotic. The descriptive complexity (the ϵ -machine) diverged in size and that forced us to move up to the meta- ϵ -machine level.

This supports a general principle that makes Anderson’s notion of hierarchy operational: the different scales in the natural world are delineated by a succession of divergences in statistical complexity of lower levels. On the mathematical side, this is reflected in the fact that hierarchical ϵ -machine reconstruction induces its own hierarchy of intrinsic computation⁴⁵, the direct analogue of the Chomsky hierarchy in discrete computation theory⁶⁵.

Closing remarks

Stepping back, one sees that many domains face the confounding problems of detecting randomness and pattern. I argued that these tasks translate into measuring intrinsic computation in processes and that the answers give us insights into how nature computes.

Causal equivalence can be adapted to process classes from many domains. These include discrete and continuous-output HMMs (refs 45,66,67), symbolic dynamics of chaotic systems⁴⁵,

molecular dynamics⁶⁸, single-molecule spectroscopy^{67,69}, quantum dynamics⁷⁰, dripping taps⁷¹, geomagnetic dynamics⁷² and spatiotemporal complexity found in cellular automata^{73–75} and in one- and two-dimensional spin systems^{76,77}. Even then, there are many remaining areas of application.

Specialists in the areas of complex systems and measures of complexity will miss a number of topics above: more advanced analyses of stored information, intrinsic semantics, irreversibility and emergence^{46–52}; the role of complexity in a wide range of application fields, including biological evolution^{78–83} and neural information-processing systems^{84–86}, to mention only two of the very interesting, active application areas; the emergence of information flow in spatially extended and network systems^{74,87–89}; the close relationship to the theory of statistical inference^{85,90–95}; and the role of algorithms from modern machine learning for nonlinear modelling and estimating complexity measures. Each topic is worthy of its own review. Indeed, the ideas discussed here have engaged many minds for centuries. A short and necessarily focused review such as this cannot comprehensively cite the literature that has arisen even recently; not so much for its size, as for its diversity.

I argued that the contemporary fascination with complexity continues a long-lived research programme that goes back to the origins of dynamical systems and the foundations of mathematics over a century ago. It also finds its roots in the first days of cybernetics, a half century ago. I also showed that, at its core, the questions its study entails bear on some of the most basic issues in the sciences and in engineering: spontaneous organization, origins of randomness, and emergence.

The lessons are clear. We now know that complexity arises in a middle ground—often at the order–disorder border. Natural systems that evolve with and learn from interaction with their immediate environment exhibit both structural order and dynamical chaos. Order is the foundation of communication between elements at any level of organization, whether that refers to a population of neurons, bees or humans. For an organism order is the distillation of regularities abstracted from observations. An organism's very form is a functional manifestation of its ancestor's evolutionary and its own developmental memories.

A completely ordered universe, however, would be dead. Chaos is necessary for life. Behavioural diversity, to take an example, is fundamental to an organism's survival. No organism can model the environment in its entirety. Approximation becomes essential to any system with finite resources. Chaos, as we now understand it, is the dynamical mechanism by which nature develops constrained and useful randomness. From it follow diversity and the ability to anticipate the uncertain future.

There is a tendency, whose laws we are beginning to comprehend, for natural systems to balance order and chaos, to move to the interface between predictability and uncertainty. The result is increased structural complexity. This often appears as a change in a system's intrinsic computational capability. The present state of evolutionary progress indicates that one needs to go even further and postulate a force that drives in time towards successively more sophisticated and qualitatively different intrinsic computation. We can look back to times in which there were no systems that attempted to model themselves, as we do now. This is certainly one of the outstanding puzzles⁹⁶: how can lifeless and disorganized matter exhibit such a drive? The question goes to the heart of many disciplines, ranging from philosophy and cognitive science to evolutionary and developmental biology and particle astrophysics⁹⁶. The dynamics of chaos, the appearance of pattern and organization, and the complexity quantified by computation will be inseparable components in its resolution.

Received 28 October 2011; accepted 30 November 2011;
published online 22 December 2011

References

1. Press, W. H. Flicker noises in astronomy and elsewhere. *Comment. Astrophys.* **7**, 103–119 (1978).
2. van der Pol, B. & van der Mark, J. Frequency demultiplication. *Nature* **120**, 363–364 (1927).
3. Goroff, D. (ed.) in *H. Poincaré New Methods of Celestial Mechanics, 1: Periodic And Asymptotic Solutions* (American Institute of Physics, 1991).
4. Goroff, D. (ed.) *H. Poincaré New Methods Of Celestial Mechanics, 2: Approximations by Series* (American Institute of Physics, 1993).
5. Goroff, D. (ed.) in *H. Poincaré New Methods Of Celestial Mechanics, 3: Integral Invariants and Asymptotic Properties of Certain Solutions* (American Institute of Physics, 1993).
6. Crutchfield, J. P., Packard, N. H., Farmer, J. D. & Shaw, R. S. Chaos. *Sci. Am.* **255**, 46–57 (1986).
7. Binney, J. J., Dowrick, N. J., Fisher, A. J. & Newman, M. E. J. *The Theory of Critical Phenomena* (Oxford Univ. Press, 1992).
8. Cross, M. C. & Hohenberg, P. C. Pattern formation outside of equilibrium. *Rev. Mod. Phys.* **65**, 851–1112 (1993).
9. Manneville, P. *Dissipative Structures and Weak Turbulence* (Academic, 1990).
10. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27** 379–423; 623–656 (1948).
11. Cover, T. M. & Thomas, J. A. *Elements of Information Theory* 2nd edn (Wiley–Interscience, 2006).
12. Kolmogorov, A. N. Entropy per unit time as a metric invariant of automorphisms. *Dokl. Akad. Nauk. SSSR* **124**, 754–755 (1959).
13. Sinai, Ja G. On the notion of entropy of a dynamical system. *Dokl. Akad. Nauk. SSSR* **124**, 768–771 (1959).
14. Anderson, P. W. More is different. *Science* **177**, 393–396 (1972).
15. Turing, A. M. On computable numbers, with an application to the Entscheidungsproblem. *Proc. Lond. Math. Soc.* **2** **42**, 230–265 (1936).
16. Solomonoff, R. J. A formal theory of inductive inference: Part I. *Inform. Control* **7**, 1–24 (1964).
17. Solomonoff, R. J. A formal theory of inductive inference: Part II. *Inform. Control* **7**, 224–254 (1964).
18. Minsky, M. L. in *Problems in the Biological Sciences* Vol. XIV (ed. Bellman, R. E.) (Proceedings of Symposia in Applied Mathematics, American Mathematical Society, 1962).
19. Chaitin, G. On the length of programs for computing finite binary sequences. *J. ACM* **13**, 145–159 (1966).
20. Kolmogorov, A. N. Three approaches to the concept of the amount of information. *Probab. Inform. Trans.* **1**, 1–7 (1965).
21. Martin-Löf, P. The definition of random sequences. *Inform. Control* **9**, 602–619 (1966).
22. Brudno, A. A. Entropy and the complexity of the trajectories of a dynamical system. *Trans. Moscow Math. Soc.* **44**, 127–151 (1983).
23. Zvonkin, A. K. & Levin, L. A. The complexity of finite objects and the development of the concepts of information and randomness by means of the theory of algorithms. *Russ. Math. Survey* **25**, 83–124 (1970).
24. Chaitin, G. *Algorithmic Information Theory* (Cambridge Univ. Press, 1987).
25. Li, M. & Vitanyi, P. M. B. *An Introduction to Kolmogorov Complexity and its Applications* (Springer, 1993).
26. Rissanen, J. Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory* **IT-30**, 629–636 (1984).
27. Rissanen, J. Complexity of strings in the class of Markov sources. *IEEE Trans. Inform. Theory* **IT-32**, 526–532 (1986).
28. Blum, L., Shub, M. & Smale, S. On a theory of computation over the real numbers: NP-completeness, Recursive Functions and Universal Machines. *Bull. Am. Math. Soc.* **21**, 1–46 (1989).
29. Moore, C. Recursion theory on the reals and continuous-time computation. *Theor. Comput. Sci.* **162**, 23–44 (1996).
30. Shannon, C. E. Communication theory of secrecy systems. *Bell Syst. Tech. J.* **28**, 656–715 (1949).
31. Ruelle, D. & Takens, F. On the nature of turbulence. *Comm. Math. Phys.* **20**, 167–192 (1974).
32. Packard, N. H., Crutchfield, J. P., Farmer, J. D. & Shaw, R. S. Geometry from a time series. *Phys. Rev. Lett.* **45**, 712–716 (1980).
33. Takens, F. in *Symposium on Dynamical Systems and Turbulence*, Vol. 898 (eds Rand, D. A. & Young, L. S.) 366–381 (Springer, 1981).
34. Brandstater, A. et al. Low-dimensional chaos in a hydrodynamic system. *Phys. Rev. Lett.* **51**, 1442–1445 (1983).
35. Crutchfield, J. P. & McNamara, B. S. Equations of motion from a data series. *Complex Syst.* **1**, 417–452 (1987).
36. Crutchfield, J. P. & Young, K. Inferring statistical complexity. *Phys. Rev. Lett.* **63**, 105–108 (1989).

37. Crutchfield, J. P. & Shalizi, C. R. Thermodynamic depth of causal states: Objective complexity via minimal representations. *Phys. Rev. E* **59**, 275–283 (1999).
38. Shalizi, C. R. & Crutchfield, J. P. Computational mechanics: Pattern and prediction, structure and simplicity. *J. Stat. Phys.* **104**, 817–879 (2001).
39. Young, K. *The Grammar and Statistical Mechanics of Complex Physical Systems*. PhD thesis, Univ. California (1991).
40. Koppel, M. Complexity, depth, and sophistication. *Complexity* **1**, 1087–1091 (1987).
41. Koppel, M. & Atlan, H. An almost machine-independent theory of program-length complexity, sophistication, and induction. *Information Sciences* **56**, 23–33 (1991).
42. Crutchfield, J. P. & Young, K. in *Entropy, Complexity, and the Physics of Information* Vol. VIII (ed. Zurek, W.) 223–269 (SFI Studies in the Sciences of Complexity, Addison-Wesley, 1990).
43. William of Ockham *Philosophical Writings: A Selection, Translated, with an Introduction* (ed. Philotheus Boehner, O. F. M.) (Bobbs-Merrill, 1964).
44. Farmer, J. D. Information dimension and the probabilistic structure of chaos. *Z. Naturf.* **37a**, 1304–1325 (1982).
45. Crutchfield, J. P. The calculi of emergence: Computation, dynamics, and induction. *Physica D* **75**, 11–54 (1994).
46. Crutchfield, J. P. in *Complexity: Metaphors, Models, and Reality* Vol. XIX (eds Cowan, G., Pines, D. & Melzner, D.) 479–497 (Santa Fe Institute Studies in the Sciences of Complexity, Addison-Wesley, 1994).
47. Crutchfield, J. P. & Feldman, D. P. Regularities unseen, randomness observed: Levels of entropy convergence. *Chaos* **13**, 25–54 (2003).
48. Mahoney, J. R., Ellison, C. J., James, R. G. & Crutchfield, J. P. How hidden are hidden processes? A primer on crypticity and entropy convergence. *Chaos* **21**, 037112 (2011).
49. Ellison, C. J., Mahoney, J. R., James, R. G., Crutchfield, J. P. & Reichardt, J. Information symmetries in irreversible processes. *Chaos* **21**, 037107 (2011).
50. Crutchfield, J. P. in *Nonlinear Modeling and Forecasting* Vol. XII (eds Casdagli, M. & Eubank, S.) 317–359 (Santa Fe Institute Studies in the Sciences of Complexity, Addison-Wesley, 1992).
51. Crutchfield, J. P., Ellison, C. J. & Mahoney, J. R. Time's barbed arrow: Irreversibility, crypticity, and stored information. *Phys. Rev. Lett.* **103**, 094101 (2009).
52. Ellison, C. J., Mahoney, J. R. & Crutchfield, J. P. Prediction, retrodiction, and the amount of information stored in the present. *J. Stat. Phys.* **136**, 1005–1034 (2009).
53. Ruelle, D. Do turbulent crystals exist? *Physica A* **113**, 619–623 (1982).
54. Varn, D. P., Canright, G. S. & Crutchfield, J. P. Discovering planar disorder in close-packed structures from X-ray diffraction: Beyond the fault model. *Phys. Rev. B* **66**, 174110 (2002).
55. Varn, D. P. & Crutchfield, J. P. From finite to infinite range order via annealing: The causal architecture of deformation faulting in annealed close-packed crystals. *Phys. Lett. A* **234**, 299–307 (2004).
56. Varn, D. P., Canright, G. S. & Crutchfield, J. P. *Inferring Pattern and Disorder in Close-Packed Structures from X-ray Diffraction Studies, Part I: ϵ -machine Spectral Reconstruction Theory* Santa Fe Institute Working Paper 03-03-021 (2002).
57. Varn, D. P., Canright, G. S. & Crutchfield, J. P. Inferring pattern and disorder in close-packed structures via ϵ -machine reconstruction theory: Structure and intrinsic computation in Zinc Sulphide. *Acta Cryst. B* **63**, 169–182 (2002).
58. Welberry, T. R. Diffuse x-ray scattering and models of disorder. *Rep. Prog. Phys.* **48**, 1543–1593 (1985).
59. Guinier, A. *X-Ray Diffraction in Crystals, Imperfect Crystals and Amorphous Bodies* (W. H. Freeman, 1963).
60. Sebastian, M. T. & Krishna, P. *Random, Non-Random and Periodic Faulting in Crystals* (Gordon and Breach Science Publishers, 1994).
61. Feldman, D. P., McTague, C. S. & Crutchfield, J. P. The organization of intrinsic computation: Complexity-entropy diagrams and the diversity of natural information processing. *Chaos* **18**, 043106 (2008).
62. Mitchell, M., Hraber, P. & Crutchfield, J. P. Revisiting the edge of chaos: Evolving cellular automata to perform computations. *Complex Syst.* **7**, 89–130 (1993).
63. Johnson, B. D., Crutchfield, J. P., Ellison, C. J. & McTague, C. S. *Enumerating Finitary Processes* Santa Fe Institute Working Paper 10-11-027 (2010).
64. Lind, D. & Marcus, B. *An Introduction to Symbolic Dynamics and Coding* (Cambridge Univ. Press, 1995).
65. Hopcroft, J. E. & Ullman, J. D. *Introduction to Automata Theory, Languages, and Computation* (Addison-Wesley, 1979).
66. Upper, D. R. *Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models*. Ph.D. thesis, Univ. California (1997).
67. Kelly, D., Dillingham, M., Hudson, A. & Wiesner, K. Inferring hidden Markov models from noisy time sequences: A method to alleviate degeneracy in molecular dynamics. Preprint at <http://arxiv.org/abs/1011.2969> (2010).
68. Ryabov, V. & Nerukh, D. Computational mechanics of molecular systems: Quantifying high-dimensional dynamics by distribution of Poincaré recurrence times. *Chaos* **21**, 037113 (2011).
69. Li, C.-B., Yang, H. & Komatsuzaki, T. Multiscale complex network of protein conformational fluctuations in single-molecule time series. *Proc. Natl Acad. Sci. USA* **105**, 536–541 (2008).
70. Crutchfield, J. P. & Wiesner, K. Intrinsic quantum computation. *Phys. Lett. A* **372**, 375–380 (2006).
71. Goncalves, W. M., Pinto, R. D., Sartorelli, J. C. & de Oliveira, M. J. Inferring statistical complexity in the dripping faucet experiment. *Physica A* **257**, 385–389 (1998).
72. Clarke, R. W., Freeman, M. P. & Watkins, N. W. The application of computational mechanics to the analysis of geomagnetic data. *Phys. Rev. E* **67**, 160–203 (2003).
73. Crutchfield, J. P. & Hanson, J. E. Turbulent pattern bases for cellular automata. *Physica D* **69**, 279–301 (1993).
74. Hanson, J. E. & Crutchfield, J. P. Computational mechanics of cellular automata: An example. *Physica D* **103**, 169–189 (1997).
75. Shalizi, C. R., Shalizi, K. L. & Haslinger, R. Quantifying self-organization with optimal predictors. *Phys. Rev. Lett.* **93**, 118701 (2004).
76. Crutchfield, J. P. & Feldman, D. P. Statistical complexity of simple one-dimensional spin systems. *Phys. Rev. E* **55**, 239R–1243R (1997).
77. Feldman, D. P. & Crutchfield, J. P. Structural information in two-dimensional patterns: Entropy convergence and excess entropy. *Phys. Rev. E* **67**, 051103 (2003).
78. Bonner, J. T. *The Evolution of Complexity by Means of Natural Selection* (Princeton Univ. Press, 1988).
79. Eigen, M. Natural selection: A phase transition? *Biophys. Chem.* **85**, 101–123 (2000).
80. Adami, C. What is complexity? *BioEssays* **24**, 1085–1094 (2002).
81. Frenken, K. *Innovation, Evolution and Complexity Theory* (Edward Elgar Publishing, 2005).
82. McShea, D. W. The evolution of complexity without natural selection—A possible large-scale trend of the fourth kind. *Paleobiology* **31**, 146–156 (2005).
83. Krakauer, D. Darwinian demons, evolutionary complexity, and information maximization. *Chaos* **21**, 037111 (2011).
84. Tononi, G., Edelman, G. M. & Sporns, O. Complexity and coherency: Integrating information in the brain. *Trends Cogn. Sci.* **2**, 474–484 (1998).
85. Bialek, W., Nemenman, I. & Tishby, N. Predictability, complexity, and learning. *Neural Comput.* **13**, 2409–2463 (2001).
86. Sporns, O., Chialvo, D. R., Kaiser, M. & Hilgetag, C. C. Organization, development, and function of complex brain networks. *Trends Cogn. Sci.* **8**, 418–425 (2004).
87. Crutchfield, J. P. & Mitchell, M. The evolution of emergent computation. *Proc. Natl Acad. Sci. USA* **92**, 10742–10746 (1995).
88. Lizier, J., Prokopenko, M. & Zomaya, A. Information modification and particle collisions in distributed computation. *Chaos* **20**, 037109 (2010).
89. Flecker, B., Alford, W., Beggs, J. M., Williams, P. L. & Beer, R. D. Partial information decomposition as a spatiotemporal filter. *Chaos* **21**, 037104 (2011).
90. Rissanen, J. *Stochastic Complexity in Statistical Inquiry* (World Scientific, 1989).
91. Balasubramanian, V. Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Comput.* **9**, 349–368 (1997).
92. Glymour, C. & Cooper, G. F. (eds) in *Computation, Causation, and Discovery* (AAAI Press, 1999).
93. Shalizi, C. R., Shalizi, K. L. & Crutchfield, J. P. *Pattern Discovery in Time Series, Part I: Theory, Algorithm, Analysis, and Convergence* Santa Fe Institute Working Paper 02-10-060 (2002).
94. MacKay, D. J. C. *Information Theory, Inference and Learning Algorithms* (Cambridge Univ. Press, 2003).
95. Still, S., Crutchfield, J. P. & Ellison, C. J. Optimal causal inference. *Chaos* **20**, 037111 (2007).
96. Wheeler, J. A. in *Entropy, Complexity, and the Physics of Information* volume VIII (ed. Zurek, W.) (SFI Studies in the Sciences of Complexity, Addison-Wesley, 1990).

Acknowledgements

I thank the Santa Fe Institute and the Redwood Center for Theoretical Neuroscience, University of California Berkeley, for their hospitality during a sabbatical visit.

Additional information

The author declares no competing financial interests. Reprints and permissions information is available online at <http://www.nature.com/reprints>.