Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Chemical Engineering

Chaotic crystallography: how the physics of information reveals structural order in materials☆

D P Varn and J P Crutchfield

We review recent progress in applying information-theoretic and computation-theoretic measures to describe material structure that transcends previous methods based on exact geometric symmetries. We discuss the necessary theoretical background for this new toolset and show how the new techniques detect and describe novel material properties. We discuss how the approach relates to well known crystallographic practice and examine how it provides novel interpretations of familiar structures. Throughout, we concentrate on disordered materials that, while important, have received less attention both theoretically and experimentally than those with either periodic or aperiodic order.

Address
Complexity Sciences Center, Physics Department, University of California, One Shields Avenue, Davis, CA 95616, USA

Corresponding authors: Varn, D.P. (dpv@complexmatter.org, http://www.wissenplatz.org) and Crutchfield, J.P. (chaos@ucdavis.edu, http://www.csc.ucdavis.edu/~chaos/)

Current Opinion in Chemical Engineering 2015, **7**:47–56

This review comes from a themed issue on **Material engineering**

Edited by **Jai A Sekhar**

For a complete overview see the Issue and the Editorial

Available online 4th December 2014

http://dx.doi.org/10.1016/j.coche.2014.11.002

Introduction

It is difficult to exaggerate the importance and influence of crystallography over the past century. Twenty-nine Nobel prizes have been awarded for discoveries either in or related to crystallography, with at least one prize per decade [1]. Crystallography strongly influences and is influenced by other fields, such as chemistry, biology, biochemistry, physics, materials science, mathematics, and geology, making it perhaps the quintessential interdisciplinary science.[1] So ingrained in other disciplines, it is now often thought of as a service science, in the sense that the techniques and theory developed in crystallography have become standard tools for researchers in these other fields. Often among the first questions in a research problem is 'What is the crystal structure of this material?'— or, more colloquially — '*Where are the atoms?*'

Unquestionably crystallography is a mature field. The *International Tables for Crystallography* consist of eight volumes (A-G, A1) and if printed out would, collectively, require nearly 6000 pages [2]. Together they coalesce and codify the combined knowledge of the worldwide crystallographic community. Additionally, there are at least a dozen major crystallographic databases, some cataloging hundreds of thousands of different solved crystal structures[2] with tens of thousands being added yearly.

As successful as this research program has been, there has been an inordinate interest in those material structures that possess *periodic order* and thus have discrete reflections in their diffraction patterns, called Bragg peaks.[3] Even in the early days of X-ray crystallography, though, some materials were known to have considerable diffuse scattering between the Bragg peaks [3] or even to lack Bragg peaks altogether [4]. While an observed broadband spectrum is sometimes a result of thermal agitations or limited experimental resolution, it can be and often is a signal of *disorder* within the material. And this disorder can be mild, preserving the integrity of the Bragg reflections, or it can be severe, where no identifiable long-range order is present. These cases have not, however, received nearly the same attention as those with "an essentially sharp diffraction pattern" [5,6] nor has the progress been nearly as impressive. Indeed, in some sense disordered structures have been defined to be outside the field of crystallography.[4]

☆ Santa Fe Institute Working Paper 14-09-036. arxiv.org:1409.5930 [cond-mat.stat-mech].

[1] Mackay [14••,45••] shows a 'concept-association network' of research areas as they relate to classical crystallography. There are considerable connections with other seemingly disparate fields.

[2] A few examples are the Crystallography Open Database (http://www.crystallography.net/), the Cambridge Structural Database (http://www.ccdc.cam.ac.uk/pages/Home.aspx), Pearson's Crystal Data (http://www.crystalimpact.com/pcd/Default.htm), and the Worldwide Protein Data Bank (http://www.wwpdb.org/). A list of databases is maintained by the International Union of Crystallographers at http://www.iucr.org/resources/data.

[3] In fact, until 1992 the defining feature of a crystal was the presence of periodic order, when this definition was changed due to the discovery of quasicrystals [54]. Now, any specimen that has "an essentially sharp diffraction pattern" [5,6] is officially classified as a crystal. Nonetheless, when we use the term 'crystal', we mean materials with periodic order.

[4] Mackay in particular has argued that the range of crystallography should extend outside its traditional boundaries [43]: "Crystallography is only incidentally concerned with crystals ... crystallography is rapidly becoming the science of structure at a particular level of organization, being concerned with structures bigger than those represented by simple atoms but smaller than those of, for example, the bacteriophage. It deals with form and function at those levels, particularly with the way in which large-scale form is the expression of local force."

Nonetheless crystallographers, defined broadly here as that community of researchers tasked with understanding and characterizing the atomic arrangement and composition of materials, have shown a persistent interest in them [4,7–9].

Researchers are increasingly discovering that disorder has profound effects on material properties and, perhaps surprisingly, disorder can improve their technological usefulness. For example, it was recently shown that significantly disordered graphene nanosheets are excellent candidates for use in high-capacity Li ion batteries due to their unusually high reversible capacities [10]. Theoretical investigations suggest that the band gap in $ZnSnP_2$, a promising candidate for high-efficiency solar cells, changes considerably (0.75–1.70 eV) as the material transitions from an ordered chalcopyrite structure to a disordered sphalerite structure [11].

The growing importance of disorder in materials, then, contrasts sharply with the lack of tools available to characterize disordered materials. And, just as researchers developed new conceptual models and theoretical techniques to understand the novel organizational structure in quasicrystals [12], new approaches are needed to characterize disordered materials. Here, we detail a recent initiative that exploits information- and computation-theoretic ideas to classify the structure of materials in a new way, one that can seamlessly bridge the gap between perfectly ordered materials, those materials with some disorder, and finally those that have no discernible underlying crystal structure.

Classical crystallography

Historically, crystals have been viewed as an unbounded repetition of atoms that fills three dimensional (3D) space.[5] Traditionally one divides this repetition into two parts: the *basis* and the *lattice*. The basis is a *fundamental structural unit* composed of one or more atoms. Although the basis can be simple in the extreme, as for example in Cu, Fe, and alkali metals where there is one atom in the basis, it can be also much more complicated, as for example in some inorganic crystals and proteins, where in the latter the basis can be composed of tens of thousands of atoms. Conversely, the lattice is a mathematical abstraction. It is defined as a regular periodic collection of points, such that if one translates from one lattice point to another, the entire arrangement of lattice points appears to be identical. There are only a finite number of ways that points can be so distributed in space.

[5] Discussions of these well known concepts from crystallography are available in any standard text on condensed matter physics [13,55]. For a definitive exposition, see the *International Tables for Crystallography, Vols. A and A1*. For classical crystallography, we exclude the case of quasicrystals and, thus, define a crystal as a periodic arrangement of atoms (the pre-1992 definition) rather than by its diffraction pattern (the post-1992 definition).

In fact, there are fourteen lattice types in 3D, and these are gathered into seven systems: triclinic, monoclinic, orthorhombic, tetragonal, cubic, trigonal and hexagonal.

To form a *crystal structure* then, the basis is attached to each lattice point, with each basis having an identical orientation. This is conveniently summarized as [13]:

$$\textbf{crystal structure} = \textbf{basis} \times \textbf{lattice}. \quad (1)$$

Each crystal structure belongs to one of the 230 different *crystallographic space groups*, which are defined by the symmetries of the crystal, including translations, rotations, reflections, glides, and screw dislocations. Thus, the regular distribution of matter in space can be classified according to physical symmetry operations respected by the crystal structure. So important is this approach that is has been referred to as *classical crystallography* (ClC) [14••] and may be defined as *the categorization of material structures based on the geometric symmetries respected by the atoms and formally couched in the language of group theory*. Succinctly put then, given some material, a primary task of ClC is to identify the basis and to which of the 230 crystallographic space groups the crystal structure belongs. In doing so, ClC provides an answer to the question — *Where are the atoms?*

Towards a new crystallography

The exact symmetries captured by groups fail partially or utterly, however, depending on a material's degree of disorder. Thus, an alternative is required; one that naturally adapts to describe randomness and noisy, partial symmetries.

Processes defined: Consider an infinite sequence of random variables, as one might encounter from time series measurements or as one scans the positions of atoms along one direction in a material. Formally, we say that there is an ordered sequence of variables indexed by subscripts and written as $\{ \ldots X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots \}$. If we make an observation of this sequence, we observe a specific realization given in lower case: $\{ \ldots x_{-2}, x_{-1}, x_0, x_1, x_2, \ldots \}$. We define a *process* as *the collection of all the possible behaviors that the system may exhibit, , as the set of all possible realizations of the system.i.e.* The ensemble of all possible realizations implies a probability distribution over length-L sequences, at each finite L. We will find that *identifying the process that describes a material is analogous to determining the lattice in ClC*. We assume that all the processes considered here are *stationary*, in the weak sense that their sequence distributions are not functions of absolute position in space.

Information theory: Inherent in the notion of disorder is uncertainty, and the amount of uncertainty is quantified by *information theory* [17,15]. Imagine a random variable X that assumes discrete outcomes $x \in \mathcal{A}$, where the latter is the set of all possible outcomes. If before a measurement

the result is predicted, then there is no uncertainty in the outcome and one learns nothing by observing it. If all possible outcomes are equally likely (maximum ignorance) then, before the measurement, the result is maximally uncertain and much is learned by discovering the result. The genius of Claude Shannon was that this notion can be quantified and, subject to a few reasonable restrictions, one can define a unique function (up to an overall scaling factor) that measures the degree of uncertainty and hence the amount of information learned from a measurement. It is given by the *Shannon entropy* $H[X]$ as [17,15]:

$$H[X] = -\sum_{x \in \mathcal{A}} \Pr(x) \log_2 \Pr(x), \tag{2}$$

where $\Pr(x)$ is the probability of observing a particular realization x when the random variable X is measured. If the logarithm is taken to base 2, as done here, the units of the Shannon entropy are *bits*.

Shannon entropy has many multivariate extensions used to capture multivariate correlations. In particular, there are the oft-used *joint entropy* (the Shannon entropy of two or more variables), *conditional entropy* (the Shannon entropy of a variable conditioned on the outcome of one or more additional variables), and *mutual information* (the information shared between two or more variables). Other measures have been recently introduced in the literature that identify a new range of correlation types [18,19].

Computational mechanics: There is a well studied theory of correlated, discrete random variables called *computational mechanics* [20-22,23••]. Within computational mechanics many processes of interest are conveniently represented as a kind of hidden Markov model [24,25] known as an *ε-machine*. In turn, ε-machines can often be written as directed finite state automata (FSA) [26], where the nodes are called *causal states* (CSs) and are connected by directed arcs that represent *transitions* between the CSs. The arcs are labeled $s|p$, where s is the symbol emitted (observed) upon transition between CSs (which generally are not directly observable). The set of CSs, which we denote $\mathbb{S}$, together with the transition probabilities between them, the set of output symbols $\mathcal{A}$, and the initial state probability distributions define the ε-machine. Critically, instead of being described by group theory, such as one finds in the crystallographic space groups, the mathematical structure of the ε-machine is that of a *semi-group*. This relaxed mathematical construct allows the ε-machine to capture the approximate symmetries of the process in a natural and self-consistent manner. This becomes essential in disordered materials, where strict spatial symmetries may no longer exist.

Importantly, ε-machines have the *minimal* number of states, and all CSs have a unique successor CS upon transition with a particular symbol, a property called *unifilarity* [27]. It can be shown that the ε-machine for a process is *unique* — in the sense that any other minimal representation is isomorphic to it — and *optimal* — in the sense that no other representation captures more of the structure [22]. Figure 1 shows nine ε-machines that are important in crystallography. We call the arrangement of CSs and their transitions the *causal architecture* of the ε-machine, and the discovery, study, and interpretation of a process's causal architecture is one of the main goals of computational mechanics.

Measures of intrinsic computation: Glancing at Figure 1, one notices some obvious differences between the ε-machines: (i) some have more CSs than others and (ii) some have multiple outgoing transitions for some of their CSs. The first property relates to an intuitive notion of structure, which can be quantified in terms of the *statistical complexity* C_μ of the ε-machine, given by [20,22]:

$$C_\mu = -\sum_{\sigma \in \mathbb{S}} \Pr(\sigma) \log_2 \Pr(\sigma). \tag{3}$$

C_μ is simply the Shannon entropy of the state probability distribution Pr(\sigma) and represents the average amount of memory (in bits) that the process retains. As a general trend, the more CSs in an ε-machine, the larger C_μ and we say that the process is more structurally complex.

More than one outgoing arc at a CS suggests that there is some uncertainty about the next observed symbol. This notion of uncertainty can be quantified by the *Shannon entropy rate* h_μ and is directly calculable from the ε-machine as [22]:

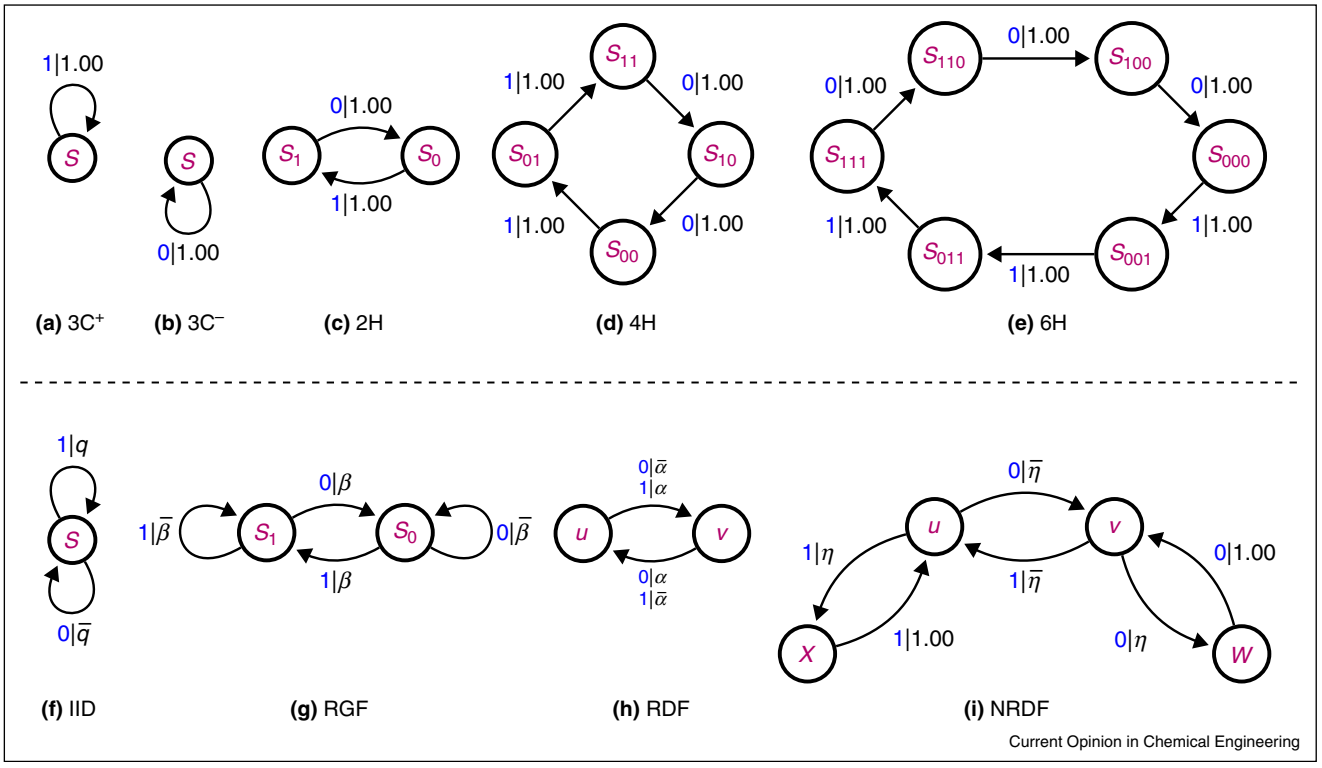
$$h_\mu = -\sum_{\sigma \in \mathbb{S}} \Pr(\sigma) \sum_{s \in \mathcal{A}} T^{(s)}_{\sigma \to \sigma'} \log_2 T^{(s)}_{\sigma \to \sigma'}. \tag{4}$$

The $T^{(s)}_{\sigma \to \sigma'}$ are the probabilities for a transition from CS σ to CS σ' on symbol s.[6] The Shannon entropy rate gives the average uncertainty per measurement when all correlations are accounted for. It has units of [bits/measurement].

While perhaps not obvious from casual examination, the ε-machines in Figure 1 imply different Markov orders — the range of interdependence. This is quantified by the *memory length* r_ℓ [28••], an integer parameter that measures the maximum range over which two symbols may carry nonredundant information about each other. That is, there may exist correlations between symbols that are not captured by the intervening symbols. It is possible

[6] When the $T^{(s)}_{\sigma \to \sigma'}$ are written as $m \times m$ matrices, with m being the number of CSs, these are the familiar *transition matrices* [16] from the study of Markov models of stochastic processes. Also, note that due to ε-machine's unifilarity the symbol s determines the unique destination state. And so, Eq. (4) does *not* need to sum over σ'.

Figure 1



Nine ε-machines that represent ordered (a-e; above the dashed line) and disordered (f-i; below the dashed line) material structures. For each the set of output symbols is chosen from $\mathcal{A} = \{0, 1\}$. The first seven ε-machines, (a-g), are *finite order Markov processes*, and the CSs are labeled by $\mathcal{S}$ with subscripts giving the minimum number of previous symbols necessary to uniquely place the process in that CS. In contrast, the last two ε-machines, (h) and (i), may require an indefinitely long history to place them in a particular CS. These ε-machines represent *strictly sofic* processes. The CSs are labeled with the symbols $\mathcal{U}, \mathcal{V}, \mathcal{W}, \mathcal{X}$. Arcs connecting CSs are labeled $s|p$, where s is the symbol emitted on transition and p is the probability of a transition. A bar over a transition probability is defined as $\bar{p} \equiv 1 - p$. (a) $3C^+$ crystal structure. (b) $3C^-$ crystal structure. (c) 2H crystal structure. (d) 4H crystal structure. (e) 6H crystal structure. (f) Independent and identically distributed (IID) process [15,16]. For $q = \bar{q} = 1/2$, the process is maximally random. (g) Random growth fault (RGF) process. For β small, we have a randomly twinned 3C structure and, for β large, there are random growth faults in the 2H structure. (h) Random deformation fault (RDF) process. For α small, we have random deformation faulting in 2H. (i) Nonrandom deformation fault (NRDF) process. For η small, this is nonrandom deformation faulting in 2H; for η large, this is a nonrandomly twinned 3C structure.

that, even if the set of states is finite, the memory length may be infinite; these are the *strictly sofic* processes [29].

Intrinsic computation is defined as how systems store, organize, and transform historical and spatial information [20,30]. Different processes may have quantitatively and qualitatively different kinds of intrinsic computation, and understanding these differences gives insight into how a system is structured [31]. In addition to the previous three measures of intrinsic computation, there are others such as *excess entropy* [32]; *transient information* and *synchronization time* [33]; *crypticity* [34]; *bound information* and *residual entropy*; and *elusive information* [18], each sensitive to different aspects of information processing and storage. Usefully, it has recently been shown that many of these information measures are directly calculable from the ε-machine [35,36].

Chaotic crystallography

Chaotic crystallography (ChC) [16,28••,37••,38•,39••,40•,41••] is the *application of information- and computation-theoretic methods to discover and characterize structure in materials*. The choice of the name is intended to be evocative: we retain the term 'crystallography' to emphasize continuity with past goals of understanding material structure; and we introduce the term 'chaotic' to associate this new approach with notions of disorder, complexity, and information processing.

The idea of appealing to information theory to describe material structure is not new, indeed Mackay has been a vocal and long-time proponent for such an approach [14••,42-44,45••]. Until recently, though, a comprehensive program to realize this vision was lacking. While ChC does realize this vision, it does not replace ClC, but rather augments it, providing a parallel, alternative view of structural organization in materials. In many

cases, especially for disordered materials, ChC gives a more consistent and comprehensive picture of material structure. We now show how these information-theoretic and computation-theoretic tools can be incorporated in a new view of material structure.

Quasi-one-dimensional materials: We specialize to the case where the periodic distribution of atoms is preserved in two dimensions (2D), but not necessarily in the third, as in the case of some *polytypes* such as ZnS and SiC (they are isostructural) [7]. A *modular layer* (ML) [46,47] is a sheet or plane of atoms organized in a regular 2D array. For closed-packed structures (CPSs), this is a hexagonal net. For ZnS in particular, at each lattice point in the net there is a Zn-S pair, separated by one-quarter of a body diagonal (as measured along the conventional unit cell) in the direction perpendicular to the plane of the net, called the *stacking direction*. Since spatial symmetries are absolutely respected within the MLs themselves, we can write the 2D version of Eq. (1) as:

$$\textbf{modular layer} = \textbf{basis} \times \textbf{2D lattice}. \tag{5}$$

For CPSs, each ML can assume only one of three possible positions, usually denoted A, B or C, and adjacent MLs stack according to the familiar closed-packed rule [13] that adjacent MLs may not have the same orientation. It is useful to take advantage of this *stacking constraint* and introduce the so-called Hägg notation, such that cyclic transitions ($A \to B \to C \to A$) between MLs are labeled '1' and anticyclical transitions ($A \to C \to B \to A$) are labeled '0' [48]. We define the *stacking sequence* [37••] as the sequence of MLs encountered as one scans the material along the stacking direction. The *stacking process* is defined as the effective stochastic process induced by sweeping the stacking sequence [37••], and we represent this in the Hägg notation over the binary symbols $\mathcal{A} = \{0, 1\}$.

Formally, for quasi-one-dimensional materials, ChC divides the task of describing material structure into two parts: (i) specify the structure of the fundamental unit, *i.e.*, the (crystalline) 2D MLs; and (ii) specify the mathematical construct that organizes the spatial distribution of the fundamental unit; *i.e.*, the kind and amount of intrinsic computation as captured by the ε-machine. The resulting material structure is referred to as a *chaotic crystal*. ChC's analogous relationship to ClC's Eq. (1) is:

$$\begin{pmatrix}\textbf{chaotic}\\ \textbf{crystal}\end{pmatrix} = \begin{pmatrix}\textbf{modular}\\ \textbf{layers}\end{pmatrix} \times (\varepsilon\textbf{–machine}). \tag{6}$$

Notice the tight parallels between ClC and ChC: the material structure (crystal versus chaotic crystal) is formed by taking a fundamental unit (basis or MLs) and distributing it through space according to some mathematical instruction (lattice or ε-machine). This close association between ClC and ChC is summarized in Table 1 (top).

Methods for detecting intrinsic computation: Determining a material's intrinsic computation, by calculating or estimating the ε-machine, is a primary goal of ChC, and several methods have been explored in the literature. Additionally, the causal architecture of the ε-machine provides invaluable information about the stacking process, and this is explored in the examples shortly. (i) One method to obtain the ε-machine is to postulate causal architectures based on theoretical grounds. Estevez *et al.* [49••] considered combined random growth and deformation faulting in closed-packed crystals and were able to generate a model that included both, called the *random growth and deformation faults* (RGDF) process [16,49••]. Although this model is not unifilar, and thus not an ε-machine, many of the techniques developed here can be adapted to analyze it [36,41••]. (ii) Another, statistical method is to simulate chaotic crystals, and use one of the reconstruction methods available in computational mechanics, such as the *subtree merging method* [20], *causal state splitting reconstruction* [50] or *Bayesian structural inference* [51], to find the appropriate model [38•]. (iii) Lastly, the approach that has received the most attention is *ε-machine spectral reconstruction theory* (εMSR) [28••,37••,39••,40•]. The importance of this technique is

Table 1

(top) A comparison of classical crystallography (ClC) and chaotic crystallography (ChC). Notice the close parallels between the two descriptions. (bottom) Measures of intrinsic computation for the ε-machines in Figure 1 and Figure 2(b). The units of C_μ are bits, h_μ are bits/ML, and r_ℓ are MLs. The abbreviations are: RT 3C = random twinned 3C; RD 2H = random deformation 2H; NRD 2H = nonrandom deformation 2H; NDT 3C = nonrandom (deformation and twinned) 3C.

	ClC	ChC
Material structure	Crystal	Chaotic crystal
Fundamental unit	Basis/unit cell	Modular layers
Organizational schema	Spatial symmetry	Intrinsic computation
Mathematical formalism	Group theory	Semi-Group theory
Symmetries	exact	approximate
Range of applicability	Crystalline	Crystalline or disordered

Example	Material structure	C_μ	h_μ	r_ℓ
1(a)	$3C^+$	0.00	0.00	0
1(b)	$3C^-$	0.00	0.00	0
1(c)	2H	1.00	0.00	1
1(d)	4H	2.00	0.00	2
1(e)	6H	2.58	0.00	3
1(f), $q = 0.50$	Random	0.00	1.00	0
1(g), $\beta = 0.10$	RT 3C	1.00	0.47	1
1(h), $\alpha = 0.10$	RD 2H	1.00	0.47	∞
1(i), $\eta = 0.10$	NRD 2H	1.44	0.43	∞
2(b), SK137	NDT 3C	2.7	0.65	3

that it uses experimentally obtained X-ray diffraction patterns to reconstruct the stacking process ε-machine. Using this, one directly calculates a stacking process's intrinsic computation. Table 1 (bottom) compares these for the nine machines in Figure 1 and that in Figure 2(b).

Examples

Periodic stacking sequences: ClC is well suited to describe periodic stacking sequences. Being periodic, spatial symmetries are strictly obeyed, and crystal structures are often specified using the Ramsdell notation *nX*, where *n* refers to the period of the repeated stacking sequence and *X* to the crystal system [48]. Commonly encountered crystal systems for CPSs include the cubic (C), hexagonal (H) and rhombohedral (R). Examples are $3C^+$ (...*ABCABC*...), 2H (...*ABABAB*...) and 6H (...*ABCACB*...) or in the Hägg notation these are (...111111...), (...101010...) and (...111000...), respectively.

ChC describes these familiar crystalline stacking structures in the form of an ε-machine. For example, the $3C^+$ stacking structure is compactly given in Figure 1(a): an ε-machine with but a single CS and a single transition. The 2H stacking structure, Figure 1(c), is slightly more involved: there are a pair of CSs connected by a pair of transitions. More involved still is the 6H stacking structure, Figure 1(e), requiring six CSs and six transitions. Indeed, for each of the first five ε-machines in Figure 1(a-e), each CS allows only one outgoing transition, and the ε-machine describes periodicity. It should be apparent that any such periodic repetition of CSs generates some crystal structure and that crystal structures can only come from this kind of causal architecture. Closed, finite, nonself-intersecting, symbol-specific paths on an ε-machine such as these are referred to as *causal state cycles*, and they are often specified by putting in square brackets [·] the sequence of causal states visited.

The measures of intrinsic computation defined in ChC quantify crystal structure and organization. Intuitively, we expect that the 6H is more complex than say $3C^+$ and indeed, by direct application of Eq. (3), we find the statistical complexities to be $C_\mu^{(6H)} = 2.58$ bits and $C_\mu^{(3C^+)} = 0$ bits. Thus, as we might expect on purely physical grounds, the 6H stacking structure requires more computational memory than $3C^+$. Additionally, we observe that for each of these three examples, direct calculation of the Shannon entropy rate using Eq. (4) finds that $h_\mu^{(3C^+)} = h_\mu^{(2H)} = h_\mu^{(6H)} = 0$ bits/ML, as we would expect for perfect crystal structures. Lastly, we might imagine that somehow the 6H stacking structure requires coordination between MLs at a greater length than that of either the $3C^+$ or 2H stacking structures. This notion is captured by the memory length, and we find that for these three structures, $r_\ell^{(3C^+)} = 0$ ML, $r_\ell^{(2H)} = 1$ ML, and $r_\ell^{(6H)} = 3$ ML, confirming our intuition.

Nonperiodic stacking sequences: When one moves beyond periodic stacking sequences, strict symmetries are no longer maintained, but instead are approximate. Mathematics based in the language of semi-groups — specifically ε-machines — is therefore more suitable than that of groups, which describe strict symmetries.

We begin with a pedagogical example. Suppose that the stacking of MLs is random, in the sense that other than respecting the CPS stacking constraints, there is no correlation between MLs. If we allow for a bias in the stacking process — *i.e.*, $\Pr(0) \neq \Pr(1)$ — then the process is described as being *independent and identically distributed* (IID) [15]. This process has been studied, for example by Guinier [4], as a simple model of disorder. The ε-machine for the IID process is shown in Figure 1(f). One notes a striking similarity with two of the periodic processes, namely the $3C^+$ and $3C^-$ in Figure 1(a) and (b). The one free parameter in the IID process is $q \in [0, 1]$, and adjusting it lets one scan from $q = 1$, giving a $3C^+$ stacking structure, to $q = 1/2$, giving an entirely disordered structure, to lastly $q = 0$, giving the crystal structure $3C^-$. From a ChC point of view then, *the crystal structures $3C^+$ and $3C^-$ are nothing more than special cases of a general IID model* and *this same IID model can also generate completely disordered stacking structures*. This is perhaps the clearest illustration of how perfectly crystalline and disordered materials may be computationally similar. However, although they share nearly identical causal architectures, measures of intrinsic computation do distinguish them. While we find $C_\mu^{(\text{random})} = C_\mu^{(3C^+)} = 0$, echoing their identical computational requirements; we also find $h_\mu^{(\text{random})}(q = 1/2) = 1.0$ bit/ML $\neq h_\mu^{(3C^+)} = 0$. This also illustrates the ease with which ChC seamlessly encompasses both crystalline and disordered structures.

Random versus nonrandom stacking faults: Many technologically useful materials, such as SiC and GaP, are subject to stacking faults (SFs). And, considerable effort is expended to characterize and understand them, often with the intention of avoiding them during manufacturing. Let's see how the ε-machines in the last three panels of Figure 1(g), (h), and (i) characterize various SFs in CPSs.

The ε-machine in Figure 1(g) represents the *random growth fault* (RGF) process. For β large, the RGF usually oscillates between the two CSs, $\mathcal{S}_0$ and $\mathcal{S}_1$, giving 2H crystal structure. With some small probability, an additional 1 or 0 is inserted into the stacking sequence and, physically, this corresponds to a *growth fault* of the 2H structure [49••]. At the other extreme when β is small, the RGF usually transits the state self-loops on each of the CSs and, physically, repetition of each of these loops gives one of the 3C stacking structures. (Compare with Figure 1(a) and (b)). We recognize this as the 3C stacking structure with randomly distributed twin faults. And, as

we saw before, ChC connects these two chaotic crystal structures (2H with random growth faults and twinned 3C) into a common causal architecture, the only difference being in the transition probabilities. Transformations between 2H and 3C are observed in ZnS [7] and, while a more complex causal architecture is needed to describe the transformation, we see that in principle ChC provides very simple models to transform from one crystal structure to another.

The ε-machine in Figure 1(h) represents the *random deformation faulting* (RDF) process as it models random deformation faults in the 2H crystal structure [49••]. The introduction of deformation faults in 2H crystals is often modeled by *Glauber dynamics* [38•,52] that corresponds to changing 1 to 0 or 0 to 1. The ε-machine for the RGF does this randomly, with some small probability α.

The ε-machine in Figure 1(i) is similar to the previous one, since for small η it too represents deformation faulting in the 2H structure, but now the SFs are distributed *nonrandomly* through the stacking sequence. We call this the *Nonrandom deformation faulting* (NRDF) process. It is a simplified version of a previous model obtained from simulation experiments of the 2H $\rightarrow$ 3C transformation in ZnS [38•]. The critical difference between the RDF and the NRDF processes is the addition of two CSs 'on the wings' of the 2H CSs — $\mathcal{U}$ and $\mathcal{V}$. These extra CSs have the effect of preventing sequences that have an *even* number of 1s or 0s. Physically this implies that the occurrence of one deformation fault *suppresses* the occurrence of an adjacent deformation fault, and this is observed in experiment [53]. Also, as the fault parameter η grows, the chaotic crystal becomes increasingly dominated by odd-length sequence domains of 1s and 0s. Thus, this ε-machine reflects that the chaotic crystal transforms into a nonrandomly twinned 3C crystal.

Here then, we see two important points: (i) the causal architecture of the ε-machines for chaotic crystals can sensitively reflect the structural organization of the stacking process; and (ii) the ε-machine seamlessly connects apparently different kinds of stacking processes into a single causal architecture, facilitating the study of solid-state phase transitions. A major task in ChC is the interpretation of the ε-machine in terms of physical mechanisms that result in observed stacking processes.

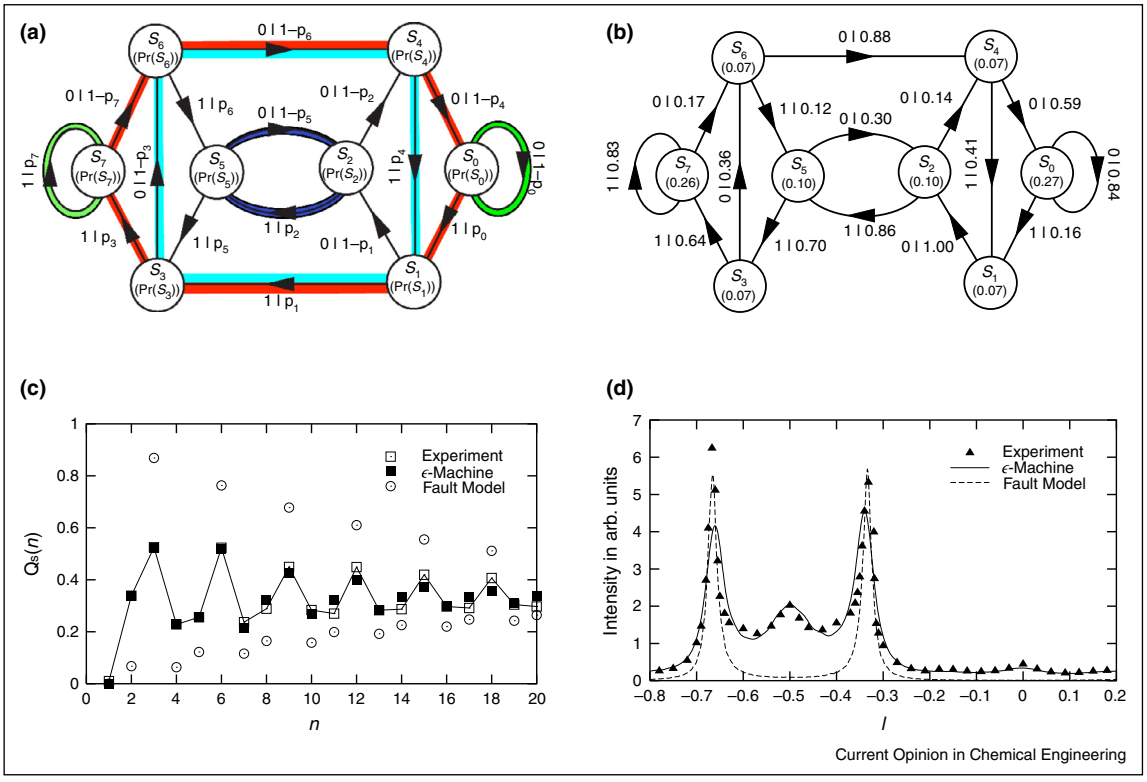
ε-Machine spectral reconstruction theory: A significant source of information about crystals is X-ray diffraction, and ChC has a method of discovering intrinsic computation from this source, much as ClC uses X-ray diffraction studies to determine crystal structure. εMSR [28••,37••,39••,40•] employs Fourier analysis over a unit interval in frequency space to extract information about the pairwise correlations between the MLs and then solves a set of equations for sequence probabilities. The algorithm initially considers low-order Markov processes and compares the diffraction pattern calculated from the model with the experimental one. If the agreement is unsatisfactory, the order of the Markov model is increased, and the comparison is repeated. This incremental process has been accomplished up to third-order Markov models. The most general order-3 Markov model is shown in Figure 2(a).

The triangles (▲) in Figure 2(d) show the diffraction pattern (corrected for experimental effects) along the $10.\ell$ row of an as-grown disordered specimen of ZnS [39••]. The degraded Bragg reflections at $\ell \approx -0.67$ and $\ell \approx -0.33$ are highly suggestive of twinned 3C structure, but there is also considerable diffuse scattering, especially in the region near $\ell \approx -0.5$. This is where one would expect to observe a Bragg reflection if 2H structure were present, suggesting that in the disorder there may be some stacking sequences reminiscent of 2H character. εMSR was performed on this diffraction pattern over the interval $\ell \in [-0.80,\ 0.20]$, and the resulting reconstructed ε-machine is shown in Figure 2(b). The diffraction pattern calculated from the reconstructed ε-machine is shown in a solid line (—) in Figure 2(d) as well as the diffraction pattern calculated from a competing model of disorder, the fault model,[7] in a dashed line (- - -). Clearly, the ε-machine is successful in capturing the broadband scattering near $\ell \approx -0.5$ and it also reproduces the Bragg-like reflections near $\ell \approx -0.67$ and $\ell \approx -0.33$, though the peak intensities are somewhat less than that observed in experiment. From other processes reconstructed from diffuse diffraction patterns, it is known εMSR can sometimes have difficulty faithfully reproducing the line profiles [39••].

The correlation function for the probability that two MLs at separation n have the same absolute orientation (either A, B, or C) extracted from the experimental diffraction pattern (□) are shown in Figure 2(c), along with those from the reconstructed ε-machine (■) and an alternative description of the disorder, the fault model (⊙). For small n, the agreement between the correlation function calculated from the ε-machine and experiment is rather good, but becomes less so at larger n. One explanation for this is that there are correlations between MLs that a third-order Markov model has difficulty reproducing. Indeed, simulation studies on solid-state phase transitions in ZnS [38•] suggest that no finite-order Markov model is capable of exactly capturing all the structure.

[7] We do not discuss the fault model in detail here, but we note that this model is based on ClC, where one assumes a perfect crystal 'corrupted' by some fault structure [37••]. While often useful for weakly faulted specimens, it is not tenable when the disorder is large, such that the crystallographic symmetries are appreciably broken.

Figure 2



(a) The most general $r = 3$ ε-machine, with several of the more common causal state cycles shown in color: in green, $[\mathcal{S}_7]$ and $[\mathcal{S}_0]$ give the $3C^+$ and $3C^-$ crystal structures, respectively; in blue, $[\mathcal{S}_5\mathcal{S}_2]$ gives 2H; in cyan, $[\mathcal{S}_3\mathcal{S}_6\mathcal{S}_4\mathcal{S}_1]$ gives 4H; and in red, $[\mathcal{S}_7\mathcal{S}_6\mathcal{S}_4\mathcal{S}_0\mathcal{S}_1\mathcal{S}_3]$ gives 6H. (Adapted from Varn *et. al.* [28••], used with permission.) **(b)** The ε-machine that results from εMSR when the experimental diffraction pattern (▲) in panel (d) is analyzed. (From Varn *et al.* [39••], used with permission.) **(c)** A comparison of the pairwise correlation function between MLs as obtained from experiment (□), the reconstructed ε-machine (■) and an alternative description of the disorder, the fault model (⊙). The $Q_s(n)$ are the probabilities that two MLs at separation n have the same absolute orientation (either *A*, *B* or *C*.) The correlation functions are only defined for discrete values of n, and the line connecting adjacent points serves as an aid for the eye. (From Varn *et al.* [39••], used with permission.) **(d)** Comparison of the diffraction pattern calculated from the reconstructed ε-machine (—) and the fault model (- - -) to the experimental diffraction pattern (▲). (From Varn *et al.* [39••], used with permission.)

Examining the reconstructed ε-machine in Figure 2(b) we observe the high state probabilities for the CSs $\mathcal{S}_0$ and $\mathcal{S}_7$ as well as their large self-loop transition probabilities, confirming that this is a twinned 3C crystal, albeit with considerable disorder. Notably, the next most visited CSs are $\mathcal{S}_2$ and $\mathcal{S}_5$, and they do have a relatively small but nonetheless nonnegligible inter-state transition probability between them. This causal state cycle would give 2H crystal structure, if it were more strongly represented. So, there does seem to be some 2H character in the stacking process, although it is weak. The remaining states represent transitions between these two structures; *i.e.*, they are faulting structures. For highly disordered specimens, such as this one, it is often difficult to unambiguously assign a particular fault or crystal structure to specific architectural features [37••,39••] and a more nuanced investigation, coupled with simulation studies is required. It is clear that for many real crystals, however, that the disorder can be profound and not as simply represented as the processes of Figure 1 might imply.

This is an example of the kind of analysis that is possible with ChC. Close coupling between experimental investigations, simulation studies, and theoretical reconstruction procedures is promising as a highly effective tool for discovering, characterizing, and explaining disordered stacking structures.

Future directions

While ChC is still in its infancy, it has potential to significantly impact the way disordered structures are understood, discovered, and described. Since the modeling procedure is based in the mathematics of (probabilistic) semi-groups, it can naturally accommodate inexact or approximate symmetries such as those found in disordered materials, where ClC loses applicability.

Future directions include expanding on recent developments in understanding spectral properties of ε-machines [16,35,36,41••], where they can be a powerful quantitative tool. In particular, calculating material properties,

such as thermal and electronic transport through disordered media via their ε-machine representation, offers a way to systematically search the space of disordered processes for interesting and useful phenomena. Additionally, measures of intrinsic computation, so closely linked to structure, are likely to strongly correlate with material properties.

Another research direction is applying ChC to materials in higher dimensions; *i.e.*, treating 2D materials. Although the formalism as reviewed here concentrated on quasi-one-dimensional materials, the basic notions transfer to higher dimensions, and this is an area of current research.

Lastly, we return to one of the initial motivations of crystallography, as encapsulated in the question we began with — *Where are the atoms?* ClC gives an unambiguous answer in the form the material's crystal structure. In its use of probabilities, it seems perhaps that ChC has failed to reach this goal. The answer offered by ChC, however, is at once both new and informative in a different way. ChC finds and examines the process that describes the material, and this may not only be a more convenient, but a more insightful answer. From the process, computational and physical parameters are calculable; and the space of possible configurations is given a kind of order, permitting systematic investigation. This is because ChC does not necessary tell where each and every atom is (although it does in the case of periodic processes), but rather it defines an ensemble of configurations, as specified by the ε-machine, that statistically represents the material. And often, this is enough.

Acknowledgements

The authors thank Julyan Cartwright, Chris Ellison, Alan Mackay, John Mahoney, Tara Michels-Clark, Paul Riechers and Richard Welberry for comments on the manuscript and the Santa Fe Institute for its hospitality during visits. JPC is an SFI External Faculty member. This material is based upon work supported by, or in part by, the U.S. Army Research Laboratory and the U. S. Army Research Office under contract W911NF-13-1-0390.

References

1. International Union of Crystallography, http://www.iucr.org/people/nobel-prize (2014a), [online; accessed 04 August 2014].
2. International Union of Crystallography, http://it.iucr.org/services/guidedtour/ (2014b), [online; accessed 05 August 2014].
3. Welberry TR: **Metall Mater Trans A** 2014, **45**:75.
4. Guinier A: *X-Ray Diffraction in Crystals, Imperfect Crystals, and Amorphous Bodies*. New York: W.H. Freeman and Company; 1963.
5. **Report of the executive committee for 1991**, *Acta Crystallogr. A* 48 (1992) 922.
6. International Union of Crystallography, http://reference.iucr.org/dictionary/Crystal (2014c), [online; accessed 06 August 2014].
7. Sebastian MT, Krishna P: *Random, Non-Random and Periodic Faulting in Crystals*. The Netherlands: Gordon and Breach; 1994.
8. Welberry TR: *Diffuse X-Ray Scattering and Models of Disorder, International Union of Crystallography*. Vol. 16, Oxford: Oxford University Press; 2004.
9. Egami T, Billinge SJL: *Underneath the Bragg Peaks: Structural Analysis of Complex Materials*. 2nd ed. **Pergamon Materials Series**. Vol. 16, New York: Pergamon; 2013.
10. Pan D, Wang S, Zhao B, Wu M, Zhang H, Wang Y, Jiao Z: **Chem Mater** 2009, **21**:3136.
11. Scanlon DO, Walsh A: **Appl Phys Lett** 2012, **100**:251911.
12. Janssen T, Janner A: **Acta Crystallogr B** 2014, **70**:617.
13. Kittel C: *Introduction to Solid State Physics*. 8th ed. New York: John Wiley & Sons; 2005.
14. Mackay AL: **Comput Math Appl B** 1986, **12**:21.
•• Mackay summarizes ClC in the statement: GROUP (MOTIF) = PATTERN. He then suggests that this should be replaced with PROGRAM (MOTIF) = STRUCTURE. The introduction of computation forces crystallographers to consider information as a critical component of material organization.
15. Cover TM, Thomas JA: *Elements of Information Theory*. 2nd ed. Hoboken: John Wiley & Sons; 2006.
16. Riechers PM, Varn DP, Crutchfield JP: Santa Fe Institute Working Paper 2014-08-026 (2014), arXiv:1407.7159 [cond-mat.mtrl-sci].
17. Shannon CE: **Bell Syst Technical J** 1948, **27**:379.
18. James RG, Ellison CJ, Crutchfield JP: **Chaos** 2011, **21**:037109.
19. James RG, Burke K, Crutchfield JP: **Phys Lett A** 2014, **378**:2124.
20. Crutchfield JP, Young K: **Phys Rev Lett** 1989, **63**:105.
21. Crutchfield JP, Feldman DP: **Phys. Rev. E.** 1997, **55**:R1239.
22. Shalizi CR, Crutchfield JP: **J. Stat. Phys.** 2001, **104**:817.
23. Crutchfield JP: **Nat Phys** 2012, **8**:17.
•• A recent review of a computational mechanics that surveys the major developments and applications is presented.
24. Rabiner LR: **IEEE Proc** 1989, **77**:257.
25. Elliot RJ, Aggoun L, Moore JB: *Hidden Markov Models: Estimation and Control, Applications of Mathematics*. Vol. 29, New York: Springer; 1995.
26. Hopcroft JE, Ullman JD: *Introduction to Automata Theory, Languages, and Computation*. Reading: Addison-Wesley; 1979.
27. Ellison CJ, Mahoney JR, Crutchfield JP: **J Stat Phys** 2009, **136**:1005.
28. Varn DP, Canright GS, Crutchfield JP: **Acta Crystallogr A** 2013, **69**:197.
•• The authors give a complete exposition of εMSR.
29. Weiss B: **Monatsh Math** 1973, **77**:462.
30. Feldman DP, McTague CS, Crutchfield JP: **Chaos** 2008, **18**:043106.
31. Crutchfield JP: **Physica D** 1994, **75**:11.
32. Crutchfield JP, Packard NH: **Physica D** 1983, **7**:201.
33. Feldman DP, Crutchfield JP: **Adv Complex Syst** 2004, **07**:329.
34. Crutchfield JP, Ellison CJ, Mahoney JR: **Phys Rev Lett** 2009, **103**:094101.
35. Crutchfield JP, Ellison CJ, Riechers PM: Santa Fe Institute Working Paper 2013-09-028 (2013), arXiv:1309.3792 [cond-mat.stat-mech].
36. Riechers PM, Crutchfield JP: **Spectral decomposition of structural complexity: Meromorphic functional calculus of nondiagonalizable dynamics**, (2014), manuscript in preparation.
37. Varn DP, Canright GS, Crutchfield JP: **Phys Rev B** 2002, **66**:174110.
••

The authors give the first demonstration of εMSR from experimental diffraction patterns.

38. Varn DP, Crutchfield JP: **Phys Lett A** 2004, **324**:299.
•
A model of a solid state transformation in ZnS is presented, demonstrating how ε-machines can be found from simulation studies.

39. Varn DP, Canright GS, Crutchfield JP: **Acta Crystallogr. B** 2007,
•• **63**:169.
εMSR is applied to experimental ZnS diffraction patterns.

40. Varn DP, Canright GS, Crutchfield JP: **Acta Crystallogr A** 2013,
• **69**:413.
A demonstration of the efficacy of εMSR as applied to simulated diffraction patterns is presented.

41. Riechers PM, Varn DP, Crutchfield JP: **Diffraction patterns of**
•• **layered close-packed structures from hidden Markov models**. *Santa Fe Institute Working Paper 2014-10-038*. 2014, arxiv.org:1410.5028.
Although not discussed in this review, the authors introduce closed-form expressions for diffraction patterns of layered materials in terms of the hidden Markov models that describe the stacking sequence. They show how the eigenvalues of the transition matrices dictate the placement of Bragg-like reflections. These techniques are likely extendable to other quantities of physical interest.

42. Mackay AL: **Chimia** 1969, **23**:433.

43. Mackay AL: **Izvj Jugosl Centr Kryst** 1975, **10**:15.

44. Mackay AL: **Struct Chem** 2002, **13**:215.

45. Cartwright JHE, Mackay AL: **Phil Trans R Soc A** 2012, **370**:2807.
••
The authors argue for 'the convergence of crystallography, materials science and biology,' and present the case that information theory is an important component to a successful synthesis.

46. Ferraris G, Makovicky E, Merlino S: *Crystallography of Modular Materials*. Oxford University Press; 2008.

47. Varn DP, Canright GS: **Acta Crystallogr A** 2001, **57**:4.

48. Ortiz AL, Sánchez-Bajo F, Cumbrera FL, Guiberteau F: **J Appl Crystallogr** 2013, **46**:242.

49. Estevez-Rams E, Welzel U, Madrigal AP, Mittemeijer EJ: **Acta**
•• **Crystallogr A** 2008, **64**:537.
A careful derivation of a HMM that describes random deformation and growth faults in CPSs is given, demonstrating how analysis of a transformation mechanism might be developed into a quantitative model.

50. Shalizi CR, Shalizi KL, Crutchfield JP: Santa Fe Institute Working Paper 02-10-060; arXiv.org/abs/cs.LG/0210025. (2002).

51. Strelioff CC, Crutchfield JP: **Phys Rev E** 2014, **89**:042119.

52. Kabra VK, Pandey D: **Phys Rev Lett** 1988, **61**:1493.

53. Sebastian MT, Krishna P: **Cryst Res Technol** 1987, **22**:929.

54. Shechtman D, Blech I, Gratias D, Cahn JW: **Phys Rev Lett** 1984, **53**:1951.

55. Ashcroft NW, Mermin ND: *Solid State Physics*. New York: Saunders College Publishing; 1976.