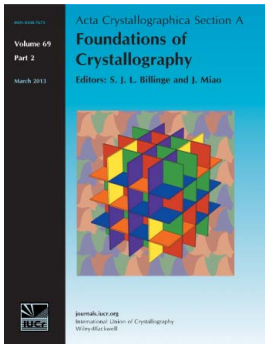
Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673
Editors: **S. J. L. Billinge and J. Miao**

&epsis;-Machine spectral reconstruction theory: a direct method for inferring planar disorder and structure from X-ray diffraction studies

D. P. Varn, G. S. Canright and J. P. Crutchfield

Acta Cryst. (2013). A**69**, 197–206



Acta Crystallographica Section A: Foundations of Crystallography covers theoretical and fundamental aspects of the structure of matter. The journal is the prime forum for research in diffraction physics and the theory of crystallographic structure determination by diffraction methods using X-rays, neutrons and electrons. The structures include periodic and aperiodic crystals, and non-periodic disordered materials, and the corresponding Bragg, satellite and diffuse scattering, thermal motion and symmetry aspects. Spatial resolutions range from the subatomic domain in charge-density studies to nanodimensional imperfections such as dislocations and twin walls. The chemistry encompasses metals, alloys, and inorganic, organic and biological materials. Structure prediction and properties such as the theory of phase transformations are also covered.

Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 19 April 2012
Accepted 11 November 2012

ϵ-Machine spectral reconstruction theory: a direct method for inferring planar disorder and structure from X-ray diffraction studies

D. P. Varn,[a,b,c]* G. S. Canright[c,d] and J. P. Crutchfield[a,b]*

[a]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA, [b]Complexity Sciences Center and Physics Department, University of California, Davis, One Shields Avenue, Davis, California 95616, USA, [c]Department of Physics and Astronomy, University of Tennessee, 1408 Circle Drive, Knoxville, Tennessee 37996, USA, and [d]Telenor Research and Development, 1331 Fornebu, Oslo, Norway. Correspondence e-mail: dpv@complexmatter.org, chaos@ucdavis.edu

In previous publications [Varn *et al.* (2002). *Phys. Rev. B*, **66**, 174110; Varn *et al.* (2007). *Acta Cryst.* B**63**, 169–182] we introduced and applied a new technique for discovering and describing planar disorder in close-packed structures directly from their diffraction patterns. Here, we provide the theoretical development behind those results, adapting computational mechanics to describe one-dimensional structure in materials. We show that the resulting statistical model of the stacking structure – called the ϵ-machine – allows the calculation of measures of memory, structural complexity and configurational entropy. The methods developed here can be adapted to a wide range of experimental systems in which power spectra data are available.

1. Introduction

Stacking faults (SFs) occur in crystal structures when one crystal plane, or more generally one modular layer (ML) (Varn & Canright, 2001), is displaced from another by a non-lattice vector. Many different kinds of SFs have been described in the literature, including growth faults, deformation faults and layer-displacement faults (Sebastian & Krishna, 1994). Provided the disorder is confined to the (dis)placement of otherwise undefected MLs, the specification of the crystal structure formally reduces to a one-dimensional list – called the *stacking sequence* – that gives the successive orientations of the MLs encountered as one moves along the stacking direction.

The problem of inferring planar disorder and structure in layered materials from their diffraction patterns (DPs) has a long history (Sebastian & Krishna, 1994; Estevez-Rams *et al.*, 2007). While early researchers were able to derive analytical expressions for the DP for specific types of SFs (Hendricks & Teller, 1942; Wilson, 1942), the inversion of a DP to find the disordered stacking sequence from which it arose has proven more challenging. This difficulty arises from the well known fact that the DP loses phase information. Indeed, there are many different stacking configurations for disordered crystals that give the same DP and, thus, it is not possible to invert the DP to find a unique configuration. Often, then, one chooses to find a statistical expression for an ensemble of disordered crystals, each of which could have given rise to the observed DP. If the disorder is not too pronounced, an underlying periodic structure is often retained, and it can be fruitful to examine the effects of various physically plausible SFs on the Bragg-like reflections.[1] By comparing the DPs from crystals with hypothetical SFs to experimental DPs, one can frequently deduce the kind and amount of SFs present. Techniques that use this approach are collectively called the *fault model* (FM) (Varn *et al.*, 2002). These approaches are *indirect*, because one begins with a set of postulated SFs and then sorts through them, searching for one or several that best fit the data. This is satisfactory, however, only if the disorder is sufficiently weak that it preserves the integrity of Bragg-like reflections.

Here we provide the theoretical foundations for a direct method of discovery and quantification of planar structure and disorder in close-packed structures (CPSs) that overcomes many of these difficulties. Although this technique has been previously applied to the detection and identification of disordered stacking sequences in ZnS (Varn *et al.*, 2002, 2007), the theoretical basis for that analysis has not been presented in detail, and we do that here. This method utilizes a type of stochastic finite-state automaton (Paz, 1971; Hopcroft &

[1] For crystals showing only mild disorder, we refer to *Bragg-like* regions of the DP that are centered on former Bragg reflections and that contain highly enhanced, yet diffuse scattering. They are distinguished by regions that would have received no refracted intensity were the crystal undefected. Regions that receive weak scattered intensity will be described as *broadband*.

Ullman, 1979) or hidden Markov model (Rabiner, 1989; Elliot *et al.*, 1995) to describe the crystal structure. It does not assume an underlying periodic stacking structure but instead finds the frequency of occurrence of all possible stacking sequences up to a given length and uses this to construct a model that captures the statistics of the stacking sequence. In this way, no *a priori* assumptions about the crystal structure or kind of disorder are required and, in this sense, it directly determines the stacking structure. This scheme for describing planar disorder unites both fault and crystal structure into a single framework. There is no need to treat each crystal structure or faulting scheme separately. This method treats any amount and kind of planar disorder present. Finally it quantitatively uses all of the information contained in the DP, both in the Bragg-like peaks and in the broadband scattering, to build a unique model of the stacking structure.

Computational mechanics (Crutchfield & Young, 1989; Shalizi & Crutchfield, 2001; Crutchfield, 2012) is an approach to discovering, describing and quantifying patterns. It provides for the construction of the minimal and unique model for a process that is optimally predictive; this model comes in the form of a directed graph called an ϵ-machine. A process's ϵ-machine is minimal in the sense of requiring the fewest model components to represent the process's structures and disorder; it is optimal in the sense that no alternative representation is more accurate; and it is unique in the sense that any alternative which is both minimal and optimally predictive is isomorphic to it. An ϵ-machine's algebraic structure captures a process's symmetries and approximate symmetries. From an ϵ-machine measures of a process's memory, entropy production and structural complexity can be found. We demonstrate elsewhere (Varn *et al.*, 2007) that knowledge of the ϵ-machine and the energy coupling between MLs allows one to calculate the average stacking energy for a disordered layered material.

Before being adapted to the present application of disorder in crystals, computational mechanics had been used to analyze structural complexity in a wide range of nonlinear processes, such as cellular automata (Hanson, 1993; Hanson & Crutchfield, 1997; Hordijk *et al.*, 2001) and the one-dimensional Ising model (Crutchfield & Feldman, 1997; Feldman & Crutchfield, 1998), as well as to experimental physical systems, such as the dripping faucet (Gonçalves *et al.*, 1998) and conformational dynamics of single molecules (Li *et al.*, 2008; Kelly *et al.*, 2012). Additionally, information-theoretic ideas and finite-state automata have been previously applied to the problem of polytypism and stacking disorder (Varn & Canright, 2001; Estevez-Rams, Aragon-Fernandez *et al.*, 2003; Estevez-Rams *et al.*, 2008).

Our development here is organized as follows. In §2 we give a detailed account of our procedure for discovering and quantifying pattern and disorder in CPSs; in §3 we discuss computational measures calculable from the reconstructed ϵ-machine and interpret architectural features of the ϵ-machine in terms of periodic stacking structures; and, lastly, in §4 we give our conclusions.

2. ϵ-Machine spectral reconstruction

Previous techniques of ϵ-machine reconstruction used a sequence of data produced by a process (Crutchfield & Young, 1989; Shalizi *et al.*, 2002). Here the experimental signal comes in the form of a power spectrum, and we need to develop a technique to infer the ϵ-machine from this type of data. To emphasize that we are using a power spectrum (in this case the DP) instead of a spatial or temporal sequence as has been done previously, we call this new class of inference algorithms *ϵ-machine spectral reconstruction* – abbreviated as ϵMSR and pronounced 'emissary'. We emphasize that our goal remains unchanged – to find the process's underlying description. It is only the inference procedure that is changed. In this section we give a detailed account of ϵMSR as applied to the problem of discovering pattern and disorder in CPSs.

We divide ϵMSR into five steps. First, we extract correlation information from a DP. Second, we use this to estimate stacking-sequence probabilities of a given length. Third, we reconstruct an ϵ-machine from this distribution. Fourth, we generate a DP from the ϵ-machine. Finally, we compare this ϵ-machine DP to the original. If there is insufficient agreement, we repeat the second through to fourth steps, estimating stacking-sequence probabilities at a longer length, building a new ϵ-machine, and again comparing with the original DP. In the final subsection, we give relations that can be used to determine the quality of experimental data.

2.1. Correlation functions from DPs

Let us make the following assumptions concerning DPs obtained from disordered, layered materials. We assume that:

(1) the MLs themselves are undefected and free of any distortions;

(2) the spacing between MLs does not depend on the local stacking arrangement;

(3) each ML has the same scattering power;

(4) the faults extend completely across the crystal; and

(5) the probability of finding a given stacking sequence in the crystal remains constant through the crystal.

The last assumption of stationarity guarantees spatial-translation invariance.

Let N be the number of hexagonal, close-packed MLs, with each ML occupying one of three orientations, denoted A, B or C (Sebastian & Krishna, 1994). We use three statistical quantities, $Q_{\rm c}(n)$, $Q_{\rm a}(n)$ and $Q_{\rm s}(n)$ (Yi & Canright, 1996): the two-layer *correlation functions* (CFs), where c, a and s stand for *cyclic*, *anti-cyclic* and *same*, respectively.[2] $Q_{\rm c}(n)$ is defined as the probability that any two MLs at a separation of n are cyclically related. By cyclic, we mean that if the ith ML is in orientation A (B, C), say, then the $(i+n)$th ML is in orientation B (C, A). $Q_{\rm a}(n)$ and $Q_{\rm s}(n)$ are defined in a similar fashion. Since these are probabilities, $0 \leq Q_{\alpha}(n) \leq 1$, where

[2] These are identical to the $Q(m)$, $R(m)$ and $P(m)$ often referred to in the literature (Kabra & Pandey, 1988; Shrestha & Pandey, 1996).

$\alpha \in \{c, a, s\}$. Additionally, at each n it is clear that $\sum_\alpha Q_\alpha(n) = 1$.

With these assumptions and definitions in place, the *total diffracted intensity* along the $10.\ell$ row can be written as[3] (Guinier, 1963; Berliner & Werner, 1986)

$$I(\ell) = \psi^2(\ell)\left(\frac{\sin^2(\pi N\ell)}{\sin^2(\pi\ell)} - 2\sqrt{3}\sum_{n=1}^{N}\left\{(N-n)\times \left[Q_c(n)\cos\left(2\pi n\ell + \frac{\pi}{6}\right) + Q_a(n)\cos\left(2\pi n\ell - \frac{\pi}{6}\right)\right]\right\}\right), \quad (1)$$

where ℓ is a continuous variable that indexes the magnitude of the perpendicular component of the diffracted wave, $k = 2\pi\ell/c$, and c is the spacing between adjacent MLs. The function $\psi^2(\ell)$ accounts for atomic scattering factors, the structure factor, dispersion factors or any other effects for which the experimentally obtained DPs may need to be corrected (Prince, 2006; Woolfson, 1997).

It is convenient to work with the intensity per ML, instead of the total intensity, so we define the corrected diffracted intensity per ML, $\mathsf{I}(\ell)$, as

$$\mathsf{I}(\ell) = \frac{I(\ell)}{\psi^2(\ell)N}. \quad (2)$$

We will always use $\mathsf{I}(\ell)$ unless otherwise noted and call this the *diffracted intensity* or simply the *diffraction pattern*. Observe that the diffracted intensity $\mathsf{I}(\ell)$ integrated over any unit ℓ interval is unity regardless of the particular values of the CFs (Varn, 2001). We may then use this fact to normalize experimental data.

The form of equations (1) and (2) suggests that the CFs can be found from the DP by Fourier analysis (Estevez-Rams, Martinez *et al.*, 2001; Estevez-Rams, Penton-Madrigal *et al.*, 2001; Varn, 2001).[4] Let us define $\mathsf{X}(n)$ and $\mathsf{Y}(n)$ as

$$\mathsf{X}(n) = \oint \mathsf{I}(\ell)\cos(2\pi n\ell)\,\mathrm{d}\ell \quad (3)$$

and

$$\mathsf{Y}(n) = \oint \mathsf{I}(\ell)\sin(2\pi n\ell)\,\mathrm{d}\ell, \quad (4)$$

where the small circle in the integral sign indicates that the integral is to be taken over a unit interval in ℓ. It is possible to show (Varn, 2001) that in the limit $N \to \infty$

$$Q_c(n) = \frac{1}{3} - \frac{1}{3}\left[\mathsf{X}(n) - \sqrt{3}\mathsf{Y}(n)\right] \quad (5)$$

and

$$Q_a(n) = \frac{1}{3} - \frac{1}{3}\left[\mathsf{X}(n) + \sqrt{3}\mathsf{Y}(n)\right]. \quad (6)$$

Thus, the CFs can be found by Fourier analysis of the DP.[5] Since the (corrected!) DP is periodic in ℓ with period one, there is freedom in the selection of the unit ℓ interval used for reconstruction. If the DP were not subject to any experimental error, then the choice of the integration interval would be driven solely by convenience. However, real DPs do have error and, thus, the selection of an appropriate unit ℓ interval for ϵMSR is of some importance. In §2.6 we offer guidance in this selection. We do note, however, that once the unit ℓ interval has been selected and the CFs determined, this represents the maximum information that can be extracted from the DP along the $10.\ell$ row. That is, this procedure uses both the Bragg-like reflections (should they exist), as well as the broadband diffracted intensity. To the extent that the DP is accurately represented by equation (1) and can be corrected for experimental effects [as incorporated in the $\psi^2(\ell)$ prefactor], we say that ϵMSR uses all of the information available from the DP.

2.2. Estimating the stacking-sequence distribution

In the second part of our approach, we estimate the distribution of stacking sequences from the two-layer CFs. First, though, we must consider what kind of information the CFs contain about stacking sequences. Therefore, let us define the *stacking process* as the effective stochastic process induced by scanning the stacking sequence along the stacking direction. We map the absolute orientations of the MLs $\{A, B, C\}$ onto a binary alphabet $\mathcal{A} = \{0, 1\}$ (Kabra & Pandey, 1988; Sebastian & Krishna, 1994). This is sometimes referred to as the Hägg notation. Transitions between MLs are labeled as '1' if the two MLs are cyclically related $[A \to B \to C \to A]$ and '0' if the two MLs are anti-cyclically related $[C \to B \to A \to C]$. Thus, the stacking of MLs in CPSs can be conveniently thought of as a binary 'spin chain', where each spin $s_i \in \mathcal{A}$ labels the interlayer transition from one ML to the next (Varn & Canright, 2001).

We estimate the probability distribution $\Pr(\omega)$ of finding sequences ω averaged over the sample by considering a series of constraints on the sequence probabilities. Some of these constraints are simple consequences of the mathematics; some come from the CFs themselves. From conservation of probability, we have

$$\Pr(u) = \Pr(0u) + \Pr(1u) = \Pr(u0) + \Pr(u1), \quad (7)$$

[3] We use the standard notation conventions here. See Guinier (1963) and references therein for a complete discussion of typical geometries and notations. Note, however, that our definition of ℓ (see text) differs from that of many other authors (Sebastian & Krishna, 1994).

[4] While our development here is specifically fashioned for case of analyzing DPs from CPSs, the approach is much more general than it may seem. Under mild conditions, the Wiener–Khinchin theorem (Badii & Politi, 1997) guarantees that power spectra can be written in terms of autocorrelation functions, as is done in equation (1). Thus, this kind of decomposition is generic and relations connecting power spectra to autocorrelation functions and then to sequence probabilities (see the spectral equations in §2.2) typically exist.

[5] The method detailed here for finding CFs from DPs was used in previous analyses (Varn *et al.*, 2002, 2007). However, other techniques do exist (Estevez-Rams, Martinez *et al.*, 2001; Estevez-Rams, Penton-Madrigal *et al.*, 2001; Estevez-Rams, Aragon-Fernandez *et al.*, 2003; Estevez-Rams, Leoni *et al.*, 2003).

for all $u \in \mathcal{A}^r$, where $\mathcal{A}^r$ is the set of all sequences of length r in the Hägg notation.[6] Additionally, we require that the sum of all probabilities of sequences of length $r+1$ be normalized, *i.e.*

$$\sum_{\omega \in \mathcal{A}^{r+1}} \Pr(\omega) = 1. \tag{8}$$

Equations (7) and (8) together provide 2^r constraints among the 2^{r+1} possible stacking sequences of length $r+1$.

The remaining 2^r constraints come from relating CFs to sequence probabilities *via* the relations

$$Q_\alpha(n) = \sum_{\omega \in \mathcal{A}_\alpha^n} \Pr(\omega), \tag{9}$$

where $\mathcal{A}_\alpha^n$ is the subset of length-n sequences that generate a cyclic ($\alpha = \mathrm{c}$) or an anti-cyclic ($\alpha = \mathrm{a}$) rotation between MLs at separation n. A sequence generates a cyclic (anti-cyclic) rotation between MLs at separation n if $2m - n = 1 \pmod 3$, where m is the number of ones (zeros) in the sequence. We take as many of the relations in equation (9) as necessary to form a complete set of equations to solve for $\Pr(\omega)$. For $r = 1$ and $r = 2$ the sets of equations are linear and admit analytical solutions. At $r = 3$ the first nonlinearities appear due to the necessity of using CFs at $n = 5$ to obtain a complete set of equations. We rewrite the conditional probabilities at $n = 5$ in terms of those at $n = 4$ *via* relations of the form

$$\begin{aligned}
\Pr(s_0 s_1 s_2 s_3 s_4) &= \Pr(s_0 s_1 s_2 s_3)\Pr(s_4 | s_0 s_1 s_2 s_3) \\
&\simeq \Pr(s_0 s_1 s_2 s_3)\Pr(s_4 | s_1 s_2 s_3) \\
&= \frac{\Pr(s_0 s_1 s_2 s_3)\Pr(s_1 s_2 s_3 s_4)}{\Pr(s_1 s_2 s_3)} \\
&= \frac{\Pr(s_0 s_1 s_2 s_3)\Pr(s_1 s_2 s_3 s_4)}{\Pr(s_1 s_2 s_3 0) + \Pr(s_1 s_2 s_3 1)}.
\end{aligned} \tag{10}$$

In the second line we make the replacement $\Pr(s_4 | s_0 s_1 s_2 s_3) \simeq \Pr(s_4 | s_1 s_2 s_3)$. We refer to this approximation as *memory-length reduction*, as it effectively limits the memory (from four previous spins to three previous spins in this case) that we consider in order to obtain a complete set of equations. At fixed r the set of equations describes the stacking sequence as an rth-order Markov process.

We refer collectively to the set of equations (7), (8) and (9) as the *spectral equations* (*SEs*) *at a given r*. In Appendix *A* we give the analytical solutions for the $r = 1$ and $r = 2$ SEs, and we write out the SEs for $r = 3$. Although this latter set of equations is nonlinear, numerical techniques may be used to solve them for each particular set of CFs.

2.3. ϵ-Machine reconstruction from the stacking process

In the third part of our approach, we infer the stacking process's ϵ-machine from the estimated distribution of stacking sequences.

Suppose we know the probability distribution $\Pr(\omega)$ of stacking sequences $\omega = \ldots s_{-2} s_{-1} s_0 s_1 s_2 \ldots$, where $s_i \in \mathcal{A}$ and ω is a stacking sequence in the Hägg notation. Then at each ML we define the 'past' $\overleftarrow{\omega}$ as all the previous transitions s_i seen and the 'future' $\overrightarrow{\omega}$ as those transitions s_i yet to be seen: that is, $\omega = \overleftarrow{\omega}\overrightarrow{\omega}$.

Let $\overleftarrow{\omega}_1$ and $\overleftarrow{\omega}_2$ represent two distinct pasts. These two pasts are called equivalent if and only if each has the same probability distribution over possible futures, and we express this equivalence as $\overleftarrow{\omega}_1 \sim \overleftarrow{\omega}_2$. All pasts that are equivalent are grouped together, so that one need not keep track of each particular past, but rather just to the group to which each past belongs. More formally, the effective states or *causal states* (CSs) of the stacking process are defined as the *sets* of pasts $\overleftarrow{\omega}$ that lead to statistically equivalent futures,

$$\overleftarrow{\omega}_i \sim \overleftarrow{\omega}_j \text{ if and only if } \Pr(\overrightarrow{\omega}|\overleftarrow{\omega}_i) = \Pr(\overrightarrow{\omega}|\overleftarrow{\omega}_j), \tag{11}$$

for all futures $\overrightarrow{\omega}$, where $\Pr(\overrightarrow{\omega}|\overleftarrow{\omega}_i)$ is the conditional probability of seeing $\overrightarrow{\omega}$ having just seen $\overleftarrow{\omega}_i$ (Crutchfield & Young, 1989; Crutchfield, 1994).

As a default set of CSs, we initially assume that each history of length r forms a unique CS. So, for ϵMSR at r, we begin with 2^r CSs, each labeled by its unique length-r history. We refer to this set of CSs as *candidate causal states*, as they may not be the true CSs that describe the stacking process. If we label each past by the last r spins seen, then this implies that the only allowed state-to-state transitions are of the form $s_0 v \to v s$, where $v \in \mathcal{A}^{r-1}$ and $s \in \mathcal{A}$. All other transitions are taken to be zero.

As a specific example we treat the $r = 2$ case. The candidate CSs and their state-to-state transitions are shown in Fig. 1. Here all possible pasts of length $r = 2$ are distinguished ($2^2 = 4$ in this example) and all possible transitions between these candidate CSs are allowed (two transitions out of each state for $2 \times 4 = 8$ possible transitions). Each state is labeled by the last two observed spins. Consider the candidate CS K_{01} in Fig. 1. The next possible spin is either a 0 or a 1, leading to

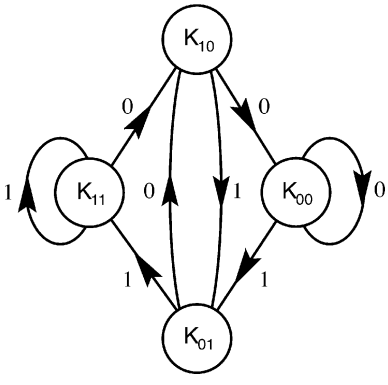


Figure 1
The topological state architecture of an $r = 2$ ϵ-machine showing all of the possible candidate causal states and allowed state-to-state transitions. Each state K_i is labeled by the last two observed spins. Notice that each state has two possible successor states, such that $s_0 v \to v s$, where $v \in \mathcal{A}^{r-1}$ and $s \in \{0, 1\}$. For example, we see that K_{01} has two possible successor states, namely K_{10} ($01 \to 10$) and K_{11} ($01 \to 11$).

[6] The parameter r plays the same role in ϵMSR as the *Reichweite*, s, does in Jagodzinski's disorder model (Jagodzinski, 1949). They are both related to the number of previous MLs used to construct conditional probabilities, but differ somewhat because each is originally defined in a different nomenclature system. They also differ because Jagodzinski's disorder model assumes spin-flip invariance, while ϵMSR does not.

the sequences 010 or 011. These transitions should lead to states that are labeled by the last two spins seen, in this case 10 and 11, respectively. Examining the graph in Fig. 1, we see that the successor states to K_{01} are K_{10} (on a transition of a 0) and K_{11} (on a transition of a 1). These latter two candidate CSs correspond to the pasts 10 and 11, respectively, as they must.

We now estimate the state-to-state transition probabilities between candidate CSs as follows. Define the *transition matrices* $\mathsf{T}^{(s)}_{\mathcal{S}_i\to\mathcal{S}_j}$ as the probability of making a transition from a candidate CS $\mathcal{S}_i$ to a candidate CS $\mathcal{S}_j$ on seeing spin s. Then we can write the transition matrix as

$$\mathsf{T}^{(s)}_{\mathcal{S}_i\to\mathcal{S}_j} = \mathsf{T}^{(s)}_{s_0 v\to vs}. \tag{12}$$

We estimate these transition probabilities from the conditional probabilities,

$$\begin{aligned}\mathsf{T}^{(s)}_{s_0 v\to vs} &= \Pr(s|s_0 v)\\ &= \frac{\Pr(s_0 vs)}{\Pr(s_0 v)}.\end{aligned} \tag{13}$$

We now apply the equivalence relation (11) to merge histories with equivalent futures. The set of resulting CSs, along with the transitions between states, defines the process's ϵ-machine. This is the minimal, unique description of the stacking process that optimally produces the stacking distribution $\Pr(\omega)$. At this point we should refer to this as the *candidate* ϵ-machine, as it will reproduce the CFs used to find it, but it may fail to reproduce CFs at larger n satisfactorily. We address this issue of agreement between theory and experiment in §2.5.

2.4. CFs and DPs from the reconstructed ϵ-machine

In the fourth part, we use the reconstructed ϵ-machine to generate CFs and DPs *via* Monte Carlo methods (Berliner & Werner, 1986; Kabra & Pandey, 1988). Specifically we use the reconstructed ϵ-machine to generate a sample spin sequence M spins long in the Hägg representation and convert this to the ABC notation. CFs can be found directly by scanning this latter sequence. The DP is readily calculated from equations (1) and (2). It has been shown that for sufficiently large M, the DP for diffuse scattering scales as M (Varn, 2001), so that the number of MLs used to calculate the DP is not important, if M is sufficiently large (say, 10 000). To reduce the error due to fluctuations, it is desirable to use as long a sequence as possible to find the CFs.

2.5. Comparing experimental and theoretical CFs and DPs

In the fifth and final part we compare the CFs and DPs predicted by the ϵ-machine to those of the original DP. If there is not sufficient agreement, we increment r and repeat the reconstruction and comparison.

More precisely, in comparing the DP predicted by the reconstructed ϵ-machine (theory) with the original DP (experiment), we need a quantitative measure of the goodness-of-fit between them. We use the *profile* $\mathcal{R}$ *factor*,[7] which is defined as

$$\mathcal{R} = \frac{\oint |\mathsf{I}_{\epsilon M}(\ell) - \mathsf{I}_{\rm exp}(\ell)|\,{\rm d}\ell}{\oint \mathsf{I}_{\epsilon M}(\ell)\,{\rm d}\ell} \times 100\%, \tag{14}$$

where $\mathsf{I}_{\epsilon\rm M}(\ell)$ is the ϵ-machine DP and $\mathsf{I}_{\rm exp}(\ell)$ is the experimental DP. Notice that the denominator is unity due to normalization.

It is important, however, not to over-fit the original data, so we should not seek a fit that is closer than experimental error. Let us define $\delta\mathsf{I}_{\rm exp}(\ell)$ as the fluctuation-induced error in the DP. Then the *fluctuation error* $\mathcal{R}_{\rm err}$ can be defined as

$$\mathcal{R}_{\rm err} = \frac{\oint |\delta\mathsf{I}_{\rm exp}(\ell)|\,{\rm d}\ell}{\oint \mathsf{I}_{\rm exp}(\ell)\,{\rm d}\ell} \times 100\%. \tag{15}$$

Notice that the denominator once again reduces to unity due to normalization. $\mathcal{R}_{\rm err}$ gives a measure of how two DPs taken from the same sample over the same interval will differ from each other. Clearly we do not wish to estimate an ϵ-machine that gives better agreement than this. So, our criterion for stopping reconstruction is when $|\mathcal{R} - \mathcal{R}_{\rm err}| \leq \Gamma$, where the acceptable-error threshold Γ is set in advance.

2.6. Figures-of-merit from ϵMSR

An issue we have so far neglected is the CFs' independence. In order to solve the SEs, part 3 in ϵMSR (§2.2), we need 2^{r+1} independent constraints. It is therefore important to identify and avoid using any redundancies inherent in the CFs to solve the SEs. Rather than finding this a hindrance, any relations that CFs obey can be exploited to assess the quality of experimental data over a given ℓ interval. We find that as a result of stacking constraints and conservation of probability, there are two equalities that the CFs must satisfy. We develop and define these measures here.

We find the first by observing that, at $n = 1$, due to stacking constraints, $Q_{\rm c}(1) + Q_{\rm a}(1) = 1$. Adding equations (5) and (6) with $n = 1$ immediately gives $\mathsf{X}(1) = -1/2$. This suggests that we define a *figure-of-merit* γ as

$$\gamma = \oint \mathsf{I}(\ell)\cos(2\pi\ell)\,{\rm d}\ell. \tag{16}$$

γ can be used to evaluate the quality of experimental DPs. For an ideal, error-free DP, $\gamma = -1/2$. Since many DPs are known to contain systematic error (Pandey *et al.*, 1987; Sebastian & Krishna, 1994), the amount by which γ deviates from $-1/2$ can be used to assess how corrupt the data is over a given unit ℓ interval.

To find the second constraint, we observe that equation (7), with $r = 1$ and $u = 0$, gives $\Pr(01) = \Pr(10)$. We therefore find from equation (8) that $\Pr(00) + 2\Pr(01) + \Pr(11) = 1$. We can write $\Pr(01) = \Pr(1) - \Pr(11)$. This implies that

$$\Pr(00) + 2\Pr(1) - \Pr(11) = 1. \tag{17}$$

[7] Our definition of the profile $\mathcal{R}$ factor differs slightly from that used by other authors (Berliner & Werner, 1986). We perform an integral over a unit ℓ interval instead of summing the magnitude of the difference between theory and experiment. Also, we find it convenient to compare the *corrected diffraction intensities* $\mathsf{I}(\ell)$, rather than the *total diffracted intensity* $I(\ell)$ as is done elsewhere.

Table 1
The ϵMSR algorithm.

Here ω^r signifies the set of length-r sequences.

(1)	Set the acceptance threshold Γ.
(2)	Find the CFs from the DP.
	(*a*) Correct the DP for any experimental factors.
	(*b*) Calculate the figures-of-merit (§2.6) over possible ℓ intervals to find a suitable interval for ϵ-machine reconstruction.
	(*c*) Find the CFs over this interval.
	(*d*) Estimate the fluctuation error $\mathcal{R}_{\rm err}$ from the DP.
(3)	Estimate the stacking distribution $\Pr(\omega^r)$ from the CFs.
	(*a*) Set $r = 1$.
	(*b*) Solve the SEs (Appendix *A*) for $\Pr(\omega^r)$.
(4)	Reconstruct the ϵ-machine from the $\Pr(\omega^r)$.
	(*a*) Label candidate CSs by their length-r histories.
	(*b*) Estimate transition probabilities between states from sequence probabilities.
	(*c*) Merge histories with equivalent futures to form CSs.
(5)	Generate the CFs and the DP from the ϵ-machine.
(6)	Calculate the error $\Gamma(r) = \lvert\mathcal{R} - \mathcal{R}_{\rm err}\rvert$ between the experimental and ϵ-machine DPs:
	(*a*) If $\Gamma(r) \geq \Gamma$, replace r with $r+1$ and go to step (3*b*);
	(*b*) otherwise, stop.

Making the identification from equation (9) that $Q_{\rm c}(1) = \Pr(1)$, $Q_{\rm a}(2) = \Pr(11)$ and $Q_{\rm c}(2) = \Pr(00)$ gives

$$2Q_{\rm c}(1) + Q_{\rm c}(2) - Q_{\rm a}(2) = 1. \tag{18}$$

This suggests that we define a second *figure-of-merit* β to be

$$\beta = 2Q_{\rm c}(1) + Q_{\rm c}(2) - Q_{\rm a}(2). \tag{19}$$

β should be unity for error-free data. This can also be used to evaluate the quality of the experimental data over a given unit ℓ interval.

Together, γ and β are the figures-of-merit over a unit ℓ interval for a DP. Therefore, in the first part of ϵMSR (§2.1) we evaluate each over candidate ℓ intervals and use this as a guide in selecting an appropriate interval for ϵ-machine reconstruction. The chosen interval should have figures-of-merit in good agreement with the theoretical values. The constraints γ and β imply that only two out of the first four CFs, $Q_{\rm c}(1)$, $Q_{\rm a}(1)$, $Q_{\rm c}(2)$ and $Q_{\rm a}(2)$, are independent. We choose to take the $n = 2$ terms as the independent parameters in the SEs.

This completes our presentation of ϵMSR. The overall procedure is summarized in Table 1.

3. Structure and intrinsic computation from ϵ-machines

In this section we briefly review several information- and computation-theoretic quantities of physical import that can be directly estimated from the reconstructed ϵ-machine. We also discuss physical interpretations of the reconstructed ϵ-machine, particularly how architectural features of the ϵ-machine correspond to known periodic stacking structures.

3.1. Measures of intrinsic computation

There are a number of different quantities in computational mechanics that describe the way information is processed and stored (Crutchfield & Feldman, 2003; Crutchfield *et al.*, 2009). We consider only the following.

Memory length r_ℓ: The value of r that results at the termination of ϵMSR is an estimate of the stacking process's *memory length*, denoted r_ℓ, since it is the number of MLs that one must use to optimally represent the process's sequence statistics, given the accuracy of the original DP.[8]

Statistical complexity C_μ: The minimum average amount of memory needed to statistically reproduce a process is known as the *statistical complexity* C_μ. Since this is a measure of memory, it has units of [bits]. It is the Shannon information stored in the set of CSs,

$$C_\mu = -\sum_{\sigma\in\mathcal{S}} \Pr(\sigma)\log_2\Pr(\sigma), \tag{20}$$

where $\mathcal{S}$ is the set of CSs for the process and $\Pr(\sigma)$ is the asymptotic probability of CS σ. The latter is the left eigenvector, normalized in probability, of the state-to-state transition matrix $\mathsf{T} = \sum_{s\in\mathcal{A}} \mathsf{T}^{(s)}$. Physically, the statistical complexity is related to the *average* number of previous spins needed to observe on scanning the spin sequence to make an optimal prediction of the next spin.

Entropy rate h_μ: The amount of irreducible randomness per ML after all correlations have been accounted for. It has units of [bits/ML]. It is also known as the *thermodynamic entropy density* in statistical mechanics and the *metric entropy* in dynamical systems theory. It is given by the average per-state spin uncertainty,

$$h_\mu = -\sum_{\sigma\in\mathcal{S}} \Pr(\sigma)\sum_{s\in\mathcal{A}} \mathsf{T}^{(s)}_{\sigma\to\sigma'}\log_2\mathsf{T}^{(s)}_{\sigma\to\sigma'}, \tag{21}$$

where σ' is the CS reached from σ upon seeing spin s. Physically, h_μ is a measure of the entropy density associated with the stacking process.

Excess entropy $\mathbf{E}$: The amount of *apparent* memory in a process. The units of $\mathbf{E}$ are [bits]. It is defined as the amount of Shannon information shared between the left and right halves of a stacking sequence,

$$\mathbf{E} = \sum_{\omega} \Pr(\omega)\log_2\left(\frac{\Pr(\omega)}{\Pr(\overleftarrow{\omega})\Pr(\overrightarrow{\omega})}\right). \tag{22}$$

Note that Feldman & Crutchfield (1998) and Crutchfield & Feldman (2003) have shown that for range-r Markov processes, these quantities are related by

$$C_\mu = \mathbf{E} + rh_\mu. \tag{23}$$

For general non-finite-range Markov processes, $\mathbf{E}$ can be calculated with the methods of Crutchfield *et al.* (2009).

3.2. Causal states cycles and crystal structures

Since the ϵ-machine reconstructed at r can distinguish at most only 2^r pasts, it can have no more than 2^r CSs. The most general reconstructed ϵ-machine of memory length r is topologically equivalent to a de Bruijn graph (Teubner, 1990) of order r. By 'most general' we mean that all length-r pasts are

[8] One should be careful not to confuse r_ℓ with a physical interaction length. Each represents a different quantity. The former originates from information-theoretic considerations of the stacking structure, while the latter represents the interaction length such as one would find in a Hamiltonian.

distinguished and all allowed transitions between CSs exist. Under these assumptions, the most general binary $r = 3$ ϵ-machine, which has $2^3 = 8$ CSs and $2^{3+1} = 16$ transitions, is shown in Fig. 2.

It is known that de Bruijn graphs can be broken into a finite number of closed, non-self-intersecting loops called *simple cycles* (SCs) (Canright & Watson, 1996). By analogy, we define a *causal-state cycle* (CSC) as a finite, closed, non-self-intersecting, symbol-specific path on an ϵ-machine. We denote a CSC by the sequence of CSs visited in square brackets [·]. The states themselves are labeled with a number that, when translated into binary notation, gives the sequence of the last r spins leading to that CS. For example, for an $r = 3$ reconstructed ϵ-machine, CS $\mathcal{S}_5$ means that 101 were the last three spins observed before reaching that CS. The *period* of the CSC is the number of CSs that comprise it.

We begin by noting that a purely crystalline structure is simply the repetition of a sequence of MLs. This is realized on an ϵ-machine as a CSC. That is, an ϵ-machine consisting of a single CSC repeats the same state sequence endlessly, giving a periodic stacking sequence, which physically is a crystal structure. It is therefore useful to catalog all of the possible CSCs on an $r = 3$ ϵ-machine, and this is done in Table 2. There are 19 CSCs on an $r = 3$ ϵ-machine, and each corresponds to a potential crystal structure. (These should be verified by tracing them out on Fig. 2.)

Table 2
The 19 CSCs on an $r = 3$ ϵ-machine.

In the first column, we give the CSC and, in the second, we show the stacking sequence in the Hägg notation implied by this CSC. If this CSC represents the sole causal architecture on the ϵ-machine, then we can interpret it as a crystal structure, as shown in the third column. Some CSCs come in pairs related by spin-inversion symmetry (Varn & Canright, 2001), *i.e.* $[\mathcal{S}_0]$ and $[\mathcal{S}_7]$ are both 3C structures, differing only in chirality. In cases where the Ramsdell notation is identical for different structures, we have attached a subscript to distinguish them. We list the period-8 hexagonal structures with a subscript to differentiate them from the more common 8H structure (00001111). One must perform ϵMSR at $r = 4$ to discover this latter 8H structure.

$[\mathcal{S}_0]$	$(0)^*$	3C
$[\mathcal{S}_7]$	$(1)^*$	3C
$[\mathcal{S}_2\mathcal{S}_5]$	$(01)^*$	2H
$[\mathcal{S}_1\mathcal{S}_3\mathcal{S}_6\mathcal{S}_4]$	$(0011)^*$	4H
$[\mathcal{S}_1\mathcal{S}_3\mathcal{S}_7\mathcal{S}_6\mathcal{S}_4\mathcal{S}_0]$	$(000111)^*$	6H
$[\mathcal{S}_5\mathcal{S}_2\mathcal{S}_4\mathcal{S}_1\mathcal{S}_3\mathcal{S}_7]$	$(001101)^*$	$6H_a$
$[\mathcal{S}_2\mathcal{S}_5\mathcal{S}_3\mathcal{S}_7\mathcal{S}_4\mathcal{S}_1]$	$(110010)^*$	$6H_a$
$[\mathcal{S}_5\mathcal{S}_2\mathcal{S}_4\mathcal{S}_0\mathcal{S}_1\mathcal{S}_3\mathcal{S}_7\mathcal{S}_6]$	$(00011101)^*$	$8H_a$
$[\mathcal{S}_2\mathcal{S}_5\mathcal{S}_3\mathcal{S}_7\mathcal{S}_6\mathcal{S}_4\mathcal{S}_0\mathcal{S}_1]$	$(11100010)^*$	$8H_a$
$[\mathcal{S}_3\mathcal{S}_6\mathcal{S}_5]$	$(011)^*$	9R
$[\mathcal{S}_4\mathcal{S}_1\mathcal{S}_2]$	$(100)^*$	9R
$[\mathcal{S}_7\mathcal{S}_6\mathcal{S}_5\mathcal{S}_3]$	$(0111)^*$	12R
$[\mathcal{S}_0\mathcal{S}_1\mathcal{S}_2\mathcal{S}_4]$	$(1000)^*$	12R
$[\mathcal{S}_3\mathcal{S}_6\mathcal{S}_4\mathcal{S}_0\mathcal{S}_1]$	$(00011)^*$	15R
$[\mathcal{S}_4\mathcal{S}_1\mathcal{S}_3\mathcal{S}_7\mathcal{S}_6]$	$(11100)^*$	15R
$[\mathcal{S}_5\mathcal{S}_2\mathcal{S}_4\mathcal{S}_0\mathcal{S}_1\mathcal{S}_3\mathcal{S}_6]$	$(0001101)^*$	$21R_a$
$[\mathcal{S}_2\mathcal{S}_5\mathcal{S}_3\mathcal{S}_7\mathcal{S}_6\mathcal{S}_4\mathcal{S}_1]$	$(1110010)^*$	$21R_a$
$[\mathcal{S}_3\mathcal{S}_6\mathcal{S}_4\mathcal{S}_0\mathcal{S}_1\mathcal{S}_2\mathcal{S}_5]$	$(0001011)^*$	$21R_b$
$[\mathcal{S}_4\mathcal{S}_1\mathcal{S}_3\mathcal{S}_7\mathcal{S}_6\mathcal{S}_5\mathcal{S}_2]$	$(1110100)^*$	$21R_b$

3.3. SFs on ϵ-machines

Suppose an ϵ-machine has a single CSC that occasionally allows for deviations from strictly periodic stacking. That is, instead of each CS having a unique successor state as before, there is some small probability that a transition will occur from one of the CSs that interrupts the periodic repetition of CSs. Further suppose that this interruption is brief and that the CS path followed along this deviation quickly returns to the dominant CSC. Physically, this break in the periodic repetition of MLs corresponds to an SF. By studying the placement, frequency and length of these interruptions, one can associate well known SFs with these alternate paths.

Note, however, that nothing in the development of ϵMSR requires that there be a single dominant CSC, or that the interruptions be small. In fact, while this aids in the interpretation of the subsequent ϵ-machine, the number of CSs and the transition probabilities between them can be such that no CSC is dominant, resulting in a highly disordered crystal. Additionally, there could exist more than one kind of deviation from a dominant CSC. In this way, an ϵ-machine describes a greater range of crystal structures than simply a particular crystal structure interspersed with a particular kind of SF.

A systematic interpretation of the CS architecture of ϵ-machines reconstructed from diffuse DPs, and hence describing disordered stacking sequences, is a current topic of research.

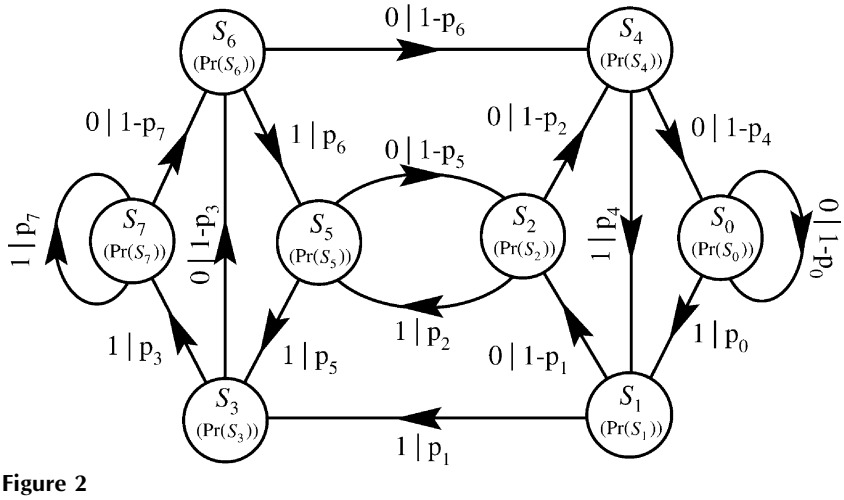


Figure 2
The most general $r = 3$ ϵ-machine. We show only the recurrent portion of the ϵ-machine, as the transient part is not physically relevant (at this stage). Here, the CSs are labeled by the last three spins seen, *i.e.* $\mathcal{S}_5$ means that 101 were the last three spins seen. The numbers in parentheses are the asymptotic CS probabilities. The edge label $s|p$ indicates a transition on spin s with probability p.

4. Summary

ϵMSR offers a number of significant features in the discovery and description of planar disorder. (i) There is no need to assume an underlying periodic (crystalline) stacking structure. Indeed, ϵMSR makes no assumption at all about what periodic or fault structure may be present. In its current

formulation ϵMSR is, however, limited to Markov models. (ii) ϵMSR is *not* limited to stacking structures that contain only weak faulting. That is, ϵMSR can be used to describe the stacking for any amount and kind of ordered and disordered sequence that a material may contain. (iii) ϵMSR uses all of the information in the diffraction pattern (DP) – Bragg-like and broadband – instead of considering only the effect disorder has on the Bragg-like peaks alone. (iv) ϵMSR defines two figures-of-merit – β and γ – that can be used to evaluate the error in experimental DPs. (v) ϵMSR results in the minimal and unique statistical expression of the stacking sequence—the ϵ-machine. (vi) Finally, from the reconstructed ϵ-machine, parameters of physical interest such as the entropy per ML, the statistical complexity, various length parameters and the average stacking-fault energy for disordered stacking sequences are directly calculable (Varn *et al.*, 2007).

In addition to its application in discovering and describing disordered stacking in ZnS single crystals (Varn *et al.*, 2007), ϵMSR is well suited to treat other materials. For example, high quality DPs (Boulle *et al.*, 2009, 2010; Dompoint *et al.*, 2011, 2012) for single-crystal SiC over large intervals in reciprocal space have recently been obtained. SiC is isostructural to ZnS and of considerable current interest. As a large-band-gap semiconductor, its electrical properties are highly influenced by its stacking structure, both ordered and disordered. ϵMSR is also applicable to the study of DPs from rare-earth compounds of the composition R_2Co_{17} (R = rare earth). Studied for their possible use as permanent magnets, at least one of these materials (Gd_2Co_{17}) shows significant disorder along the stacking direction that is not readily understood in terms of a simple random faulting model (Estevez-Rams, Martinez *et al.*, 2001). Since Estevez-Rams and co-workers (Estevez-Rams, Martinez *et al.*, 2001; Estevez-Rams, Penton-Madrigal *et al.*, 2001; Estevez-Rams, Aragon-Fernandez *et al.*, 2003; Estevez-Rams, Leoni *et al.*, 2003) have given a procedure to find the CFs between MLs directly from their powder DP, steps 3–6 of ϵMSR (see Table 1) could be used to reconstruct the ϵ-machine, possibly shedding light on the nature of these SFs. We anticipate that there are many other layered materials of theoretical and practical interest that show planar disorder and that would be amenable to an ϵMSR analysis as presented here.

Since ϵMSR relies on short-range correlations to construct the ϵ-machine, materials which exhibit effectively infinite-range interactions between MLs, such as are found in some metals, may fall outside the range of applicability of ϵMSR. Although it may be possible for infinite-range interactions to generate stacking structures which are well described by a small memory length, care should be taken. We discuss practical issues with using ϵMSR, and especially the relationship between the memory length, physical interaction lengths and other characteristic length parameters, in a forthcoming paper (Varn *et al.*, 2013).

Additionally, ϵMSR also contributes to the machine-learning side of computational mechanics. ϵMSR is novel in that we use a power spectrum to reconstruct the ϵ-machine instead of a temporal or spatial data sequence, as prior algorithms have.

There are, however, some limitations to ϵMSR, as developed here. We only attempted ϵ-machine reconstruction up to $r = 3$. It has recently been shown that a model for a simple solid-state transformation from the 2H to the 3C structure in CPSs results in stacking sequences that imply an infinite memory length (Varn & Crutchfield, 2004). While in principle one can attempt ϵMSR for any r, there are computational complexity difficulties. We feel that the general case of $r = 4$ is tractable. We also suspect that there are alternative algorithms that will greatly reduce the computational complexity of finding solutions.

APPENDIX *A*
The spectral equations

A1. r = 1

The SEs at $r = 1$ are linear and admit an analytical solution. Specifically, we write out equations (7), (8) and (9) for $r = 1$ and solve them. We find

$$\begin{aligned}\Pr(11) &= Q_a(2),\\ \Pr(01) &= \frac{1}{2}[1 - Q_c(2) - Q_a(2)],\\ \Pr(00) &= Q_c(2).\end{aligned}$$

A2. r = 2

Similarly, the SEs at $r = 2$ are linear and can be solved analytically. Again, we write out equations (7), (8) and (9) for $r = 2$ and solve them. We find

$$\begin{aligned}\Pr(000) &= [3Q_c(2) - 2Q_c(3) - 3Q_a(2) - 4Q_a(3) + 3]/6,\\ \Pr(001) &= [3Q_c(2) + 2Q_c(3) + 3Q_a(2) + 4Q_a(3) - 3]/6,\\ \Pr(010) &= [-3Q_c(2) - 2Q_c(3) - 3Q_a(2) - Q_a(3) + 3]/3,\\ \Pr(011) &= [3Q_c(2) + 4Q_c(3) + 3Q_a(2) + 2Q_a(3) - 3]/6,\\ \Pr(100) &= [3Q_c(2) + 2Q_c(3) + 3Q_a(2) + 4Q_a(3) - 3]/6,\\ \Pr(101) &= [-3Q_c(2) - Q_c(3) - 3Q_a(2) - 2Q_a(3) + 3]/3,\\ \Pr(110) &= [3Q_c(2) + 4Q_c(3) + 3Q_a(2) + 2Q_a(3) - 3]/6,\\ \Pr(111) &= [-3Q_c(2) - 4Q_c(3) + 3Q_a(2) - 2Q_a(3) + 3]/6.\end{aligned}$$

A3. r = 3

At $r = 3$, we require 16 relations to constrain the length-4 binary-sequence probabilities. Although we now encounter nonlinearities, the SEs may solved numerically.

At $r = 3$, equation (7) implies the following seven equations:

$$\begin{aligned}\Pr(0111) &= \Pr(1110),\\ \Pr(0001) &= \Pr(1000),\\ \Pr(0011) + \Pr(1011) &= \Pr(0111) + \Pr(0110),\end{aligned}$$

$$\Pr(0101) + \Pr(1101) = \Pr(1011) + \Pr(1010),$$
$$\Pr(0010) + \Pr(1010) = \Pr(0101) + \Pr(0100),$$
$$\Pr(0001) + \Pr(1001) = \Pr(0011) + \Pr(0010),$$
$$\Pr(0100) + \Pr(1100) = \Pr(1001) + \Pr(1000).$$

Equation (8) provides for normalization, providing one additional constraint. Finally, the remaining eight SEs are found by relating sequence probabilities to CFs as described by equation (9). We further reduce the last two relations which involve sequence probabilities of length-5 to those of length-4 *via* memory-length reduction. We find

$$\begin{aligned}
Q_c(2) &= \Pr(0000) + \Pr(0001) + \Pr(0010) + \Pr(0011),\\
Q_a(2) &= \Pr(1100) + \Pr(1101) + \Pr(1110) + \Pr(1111),\\
Q_c(3) &= \Pr(0110) + \Pr(0111) + \Pr(1010) + \Pr(1011)\\
&\quad + \Pr(1100) + \Pr(1101),\\
Q_a(3) &= \Pr(0010) + \Pr(0011) + \Pr(0100) + \Pr(0101)\\
&\quad + \Pr(1000) + \Pr(1001),\\
Q_c(4) &= \Pr(1111) + \Pr(1000) + \Pr(0100) + \Pr(0010)\\
&\quad + \Pr(0001),\\
Q_a(4) &= \Pr(0000) + \Pr(0111) + \Pr(1011) + \Pr(1101)\\
&\quad + \Pr(1110),\\
Q_c(5) &= \frac{\Pr^2(0000)}{\Pr(0000) + \Pr(0001)} + \frac{\Pr(0011)\Pr(0111)}{\Pr(0111) + \Pr(0110)}\\
&\quad + \frac{\Pr(0101)\Pr(1011)}{\Pr(1011) + \Pr(1010)} + \frac{\Pr(0110)\Pr(1101)}{\Pr(1101) + \Pr(1100)}\\
&\quad + \frac{\Pr(0111)\Pr(1110)}{\Pr(1110) + \Pr(1111)} + \frac{\Pr(1001)\Pr(0011)}{\Pr(0011) + \Pr(0010)}\\
&\quad + \frac{\Pr(1010)\Pr(0101)}{\Pr(0101) + \Pr(0100)} + \frac{\Pr(1011)\Pr(0110)}{\Pr(0110) + \Pr(0111)}\\
&\quad + \frac{\Pr(1100)\Pr(1001)}{\Pr(1001) + \Pr(1000)} + \frac{\Pr(1101)\Pr(1010)}{\Pr(1010) + \Pr(1011)}\\
&\quad + \frac{\Pr(1110)\Pr(1100)}{\Pr(1100) + \Pr(1101)},\\
Q_a(5) &= \frac{\Pr^2(1111)}{\Pr(1111) + \Pr(1110)} + \frac{\Pr(1100)\Pr(1000)}{\Pr(1000) + \Pr(1001)}\\
&\quad + \frac{\Pr(1010)\Pr(0100)}{\Pr(0100) + \Pr(0101)} + \frac{\Pr(1001)\Pr(0010)}{\Pr(0010) + \Pr(0011)}\\
&\quad + \frac{\Pr(1000)\Pr(0001)}{\Pr(0001) + \Pr(0000)} + \frac{\Pr(0110)\Pr(1100)}{\Pr(1100) + \Pr(1101)}\\
&\quad + \frac{\Pr(0101)\Pr(1010)}{\Pr(1010) + \Pr(1011)} + \frac{\Pr(0100)\Pr(1001)}{\Pr(1001) + \Pr(1000)}\\
&\quad + \frac{\Pr(0011)\Pr(0110)}{\Pr(0110) + \Pr(0111)} + \frac{\Pr(0010)\Pr(0101)}{\Pr(0101) + \Pr(0100)}\\
&\quad + \frac{\Pr(0001)\Pr(0011)}{\Pr(0011) + \Pr(0010)}.
\end{aligned}$$

We thank D. P. Feldman and E. Smith for useful conversations; and A. Mills, P. M. Riechers and three anonymous referees for helpful comments on the manuscript. This work was supported at the Santa Fe Institute under the Networks Dynamics Program funded by the Intel Corporation and under the Computation, Dynamics and Inference Program *via* SFI's core grants from the National Science and MacArthur Foundations. Direct support was provided by NSF grants DMR-9820816 and PHY-9910217 and DARPA Agreement F30602-00-2-0583. DPV's visit to SFI was partially supported by the NSF.

References

Badii, R. & Politi, A. (1997). *Complexity: Hierarchical Structures and Scaling and Physics*, Cambridge Nonlinear Science Series, Vol. 6. Cambridge University Press.
Berliner, R. & Werner, S. A. (1986). *Phys. Rev. B*, **34**, 3586–3603.
Boulle, A., Aube, J., Galben-Sandulache, I. G. & Chaussende, D. (2009). *Appl. Phys. Lett.* **94**, 201904.
Boulle, A., Dompoint, D., Galben-Sandulache, I. & Chaussende, D. (2010). *J. Appl. Cryst.* **43**, 867–875.
Canright, G. S. & Watson, G. (1996). *J. Stat. Phys.* **84**, 1095–1131.
Crutchfield, J. P. (1994). *Physica D*, **75**, 11–54.
Crutchfield, J. P. (2012). *Nature Phys.* **8**, 17–24.
Crutchfield, J. P., Ellison, C. J. & Mahoney, J. R. (2009). *Phys. Rev. Lett.* **103**, 094101.
Crutchfield, J. P. & Feldman, D. P. (1997). *Phys. Rev. E*, **55**, R1239–R1242.
Crutchfield, J. P. & Feldman, D. P. (2003). *Chaos*, **13**, 25–54.
Crutchfield, J. P. & Young, K. (1989). *Phys. Rev. Lett.* **63**, 105–108.
Dompoint, D., Boulle, A., Galben-Sandulache, I. G. & Chaussende, D. (2012). *Nucl. Instrum. Methods Phys. Res. B*, **284**, 19–22.
Dompoint, D., Boulle, A., Galben-Sandulache, I. G., Chaussende, D., Hoa, L. T. M., Ouisse, T., Eyidi, D., Demenet, J. L., Beaufort, M. F. & Rabier, J. (2011). *J. Appl. Phys.* **110**, 053508.
Elliot, R. J., Aggoun, L. & Moore, J. B. (1995). *Hidden Markov Models: Estimation and Control*. *Applications of Mathematics*, Vol. 29. New York: Springer.
Estevez-Rams, E., Aragon-Fernandez, B., Fuess, H. & Penton-Madrigal, A. (2003). *Phys. Rev. B*, **68**, 064111.
Estevez-Rams, E., Leoni, M., Scardi, P., Aragon-Fernandez, B. & Fuess, H. (2003). *Philos. Mag.* **83**, 4045–4057.
Estevez-Rams, E., Martinez, J., Penton-Madrigal, A. & Lora-Serrano, R. (2001). *Phys. Rev. B*, **63**, 054109.
Estevez-Rams, E., Penton-Madrigal, A., Lora-Serrano, R. & Martinez-Garcia, J. (2001). *J. Appl. Cryst.* **34**, 730–736.
Estevez-Rams, E., Penton-Madrigal, A., Scardi, P. & Leoni, M. (2007). *Z. Kristallogr.* (Suppl.). **26**, 99–104.
Estevez-Rams, E., Welzel, U., Pentón Madrigal, A. & Mittemeijer, E. J. (2008). *Acta Cryst.* A**64**, 537–548.
Feldman, D. P. & Crutchfield, J. P. (1998). *Discovering Non-critical Organization: Statistical Mechanical, Information Theoretic, and Computational Views of Patterns in One-Dimensional Spin Systems*, Santa Fe Institute Working Paper 98-04-026.
Gonçalves, W. M., Pinto, R. D., Sartorelli, J. C. & de Oliveira, M. J. (1998). *Physica A*, **257**, 385–389.
Guinier, A. (1963). *X-ray Diffraction in Crystals, Imperfect Crystals, and Amorphous Bodies*. New York: W. H. Freeman and Company.
Hanson, J. E. (1993). *Computational Mechanics of Cellular Automata*. PhD thesis, University of California, Berkeley.
Hanson, J. E. & Crutchfield, J. P. (1997). *Physica D*, **103**, 169–189.
Hendricks, S. & Teller, E. (1942). *J. Chem. Phys.* **10**, 147–167.
Hopcroft, J. E. & Ullman, J. D. (1979). *Introduction to Automata Theory, Languages and Computation*. Reading: Addison-Wesley.
Hordijk, W., Crutchfield, J. P. & Shalizi, C. R. (2001). *Physica D*, **154**, 240–258.
Jagodzinski, H. (1949). *Acta Cryst.* **2**, 208–214.
Kabra, V. K. & Pandey, D. (1988). *Phys. Rev. Lett.* **61**, 1493–1496.

Kelly, D., Dillingham, M., Hudson, A. & Wiesner, K. (2012). *PLoS ONE*, **7**, e29703.
Li, C.-B., Yang, H. & Komatsuzaki, T. (2008). *Proc. Natl Acad. Sci. USA*, **105**, 536–541.
Pandey, D., Prasad, L., Lele, S. & Gauthier, J. P. (1987). *J. Appl. Cryst.* **20**, 84–89.
Paz, A. (1971). *Introduction to Probilistic Automata*. New York: Academic Press.
Prince, E. (2006). Editor. *International Tables for Crystallography*, Vol. C, *Mathematical, Physical and Chemical Tables*. 1st online ed. Chester: International Union of Crystallography. doi: 10.1107/97809553602060000103.
Rabiner, L. R. (1989). *IEEE Proc.* **77**, 257.
Sebastian, M. T. & Krishna, P. (1994). *Random, Non-Random and Periodic Faulting in Crystals*. The Netherlands: Gordon and Breach.
Shalizi, C. R. & Crutchfield, J. P. (2001). *J. Stat. Phys.* **104**, 817–881.
Shalizi, C. R., Shalizi, K. L. & Crutchfield, J. P. (2002). *Pattern Discovery in Time Series, Part I: Theory, Algorithm, Analysis, and Convergence*, Santa Fe Institute Working Paper 02-10-060. http://arXiv.org/abs/cs.LG/0210025.
Shrestha, S. P. & Pandey, D. (1996). *Acta Mater.* **44**, 4949–4960.
Teubner, M. (1990). *Physica A*, **169**, 407–420.
Varn, D. P. (2001). *Language Extraction from ZnS*. PhD thesis, University of Tennessee, Knoxville, USA.
Varn, D. P. & Canright, G. S. (2001). *Acta Cryst.* A**57**, 4–19.
Varn, D. P., Canright, G. S. & Crutchfield, J. P. (2002). *Phys. Rev. B*, **66**, 174110.
Varn, D. P., Canright, G. S. & Crutchfield, J. P. (2007). *Acta Cryst.* B**63**, 169–182.
Varn, D. P., Canright, G. S. & Crutchfield, J. P. (2013). *Acta Cryst.* A. Submitted.
Varn, D. P. & Crutchfield, J. P. (2004). *Phys. Lett. A*, **324**, 299–307.
Wilson, A. J. C. (1942). *Proc. R. Soc. Ser. A*, **180**, 277–285.
Woolfson, M. M. (1997). *An Introduction to X-ray Crystallography*. Cambridge University Press.
Yi, J. & Canright, G. S. (1996). *Phys. Rev. B*, **53**, 5198–5210.