research paper

Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 19 November 2012
Accepted 30 March 2013

Inferring planar disorder in close-packed structures *via* ε-machine spectral reconstruction theory: examples from simulated diffraction patterns

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

[a]Complexity Sciences Center and Physics Department, University of California, Davis, One Shields Avenue, Davis, California 95616, USA, [b]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, 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, geoffrey.canright@telenor.com, chaos@ucdavis.edu

A previous paper detailed a novel algorithm, ε-machine spectral reconstruction theory (εMSR), that infers pattern and disorder in planar-faulted, close-packed structures directly from X-ray diffraction patterns [Varn *et al.* (2013). *Acta Cryst.* A**69**, 197–206]. Here εMSR is applied to simulated diffraction patterns from four close-packed crystals. It is found that, for stacking structures with a memory length of three or less, εMSR reproduces the statistics of the stacking structure; the result being in the form of a directed graph called an ε-machine. For stacking structures with a memory length larger than three, εMSR returns a model that captures many important features of the original stacking structure. These include multiple stacking faults and multiple crystal structures. Further, it is found that εMSR is able to discover stacking structure in even highly disordered crystals. In order to address issues concerning the long-range order observed in many classes of layered materials, several length parameters are defined, calculable from the ε-machine, and their relevance is discussed.

1. Introduction

While crystallography has historically focused on the characterization of materials whose constituent parts are arranged in an orderly fashion, researchers have become increasingly interested in materials that display varying amounts of disorder, several examples being glasses, aerogels (Erenburg *et al.*, 2005) and amorphous metal oxides (Bataronov *et al.*, 2004). A broad range of layered materials called *polytypes* often show considerable disorder along the stacking direction and have been the subject of numerous theoretical and experimental investigations (Jagodzinski, 1949; Trigunayat, 1991; Sebastian & Krishna, 1994). Polytypism is the phenomenon in which a three-dimensional solid is built up by the stacking of identical (or nearly identical)[1] two-dimensional modular units (Price, 1983; Ferraris *et al.*, 2008), which we will refer to as *modular layers* (MLs). Each ML is itself crystalline and the only disorder comes from how adjacent MLs are stacked. Typically, energetic considerations restrict the number of ways two MLs can be stacked to a usually small set of orientations. Thus, the specification of a disordered polytype reduces to giving the one-dimensional list of the sequence of MLs called the *stacking sequence*.

Polytypes have attracted so much interest in part due to the multiple crystalline stacking sequences commonly observed – for two of the most polytypic materials, ZnS and SiC, there are 185 and 250 known periodic stacking structures, respectively (Mardix, 1986; Nasir *et al.*, 2012). Some of these crystalline structures have unit cells extending over 100 MLs (Sebastian & Krishna, 1994). This is in contrast to the calculated inter-ML interaction range of ∼1 ML in ZnS (Engel & Needs, 1990) and ∼3 MLs in SiC (Cheng *et al.*, 1987, 1988, 1990; Shaw & Heine, 1990). An important ancillary question is whether the *disordered* polytypes so commonly observed in annealed and as-grown crystals also possess coordination in the stacking of MLs over such a long range. Additionally, SiC has received considerable attention recently as a promising candidate for use as nanowires in advanced electronics devices (Wang *et al.*, 2011; Mélinon *et al.*, 2007). The electronic properties of SiC are dependent on both the polytype and the degree of disorder present.

Significant simplifications in the analysis of X-ray diffraction patterns (DPs) occur if the disorder in the crystal is restricted to one dimension and the constituent parts can assume only discrete positions. This is just the case that arises in the analysis of polytypes. While the general problem of

[1] See Trigunayat (1991) for a discussion of materials that have some variation in either the structure or stoichiometry of the composite layers. In the present study, we will assume that the composite layers are identical.

inverting DPs to obtain structure remains unsolved, this more restricted one-dimensional case has been much more amenable to theoretical analysis. We recently introduced a novel inference algorithm, ε-machine spectral reconstruction theory (εMSR or 'emissary') and applied it to the problem of inferring planar disorder from DPs for the special case of close-packed structures (CPSs) (Varn *et al.*, 2002, 2007, 2013).[2] Although we do not find the particular stacking sequence that generated the experimental DP, we do find a unique, statistical expression for an ensemble of stacking sequences each of which could have produced the observed DP. This statistical description comes in the compact form of an *ε-machine* (Crutchfield & Young, 1989; Shalizi & Crutchfield, 2001).

We claim in a companion paper (Varn *et al.*, 2013) that εMSR has many significant features, including the following: (i) εMSR does not assume any underlying crystal structure, nor does it require one to postulate *a priori* any particular candidate faulting structures. That is, there need not be any 'parent' crystal structure into which some preselected faulting is introduced. (ii) Consequently, εMSR can model specimens with multiple crystal or fault structures. (iii) Since εMSR does not require a parent crystal structure, it can detect and quantify stacking structure in samples with even highly disordered stacking sequences. (iv) εMSR uses all of the information available from the DP: Bragg, Bragg-like and broadband scattering. (v) εMSR results in a minimal and unique description of the stacking structure in the form of an ε-machine. From knowledge of the ε-machine, insight into the spatial organization of the stacking structure is possible. (vi) Parameters of physical interest, such as entropy density, hexagonality and memory length, are directly calculable from the ε-machine.

Our purpose here is fourfold: (i) We wish to validate the above assertions concerning the efficacy of εMSR by demonstrating its application to the discovery of pattern and disorder in layered materials from their DPs. (ii) As developed in Varn *et al.* (2013), εMSR can reconstruct processes up to third-order Markovian. We wish to test the robustness of εMSR by analyzing DPs from stacking sequences not describable as third-order Markovian. While we expect that εMSR will not recover the precise statistics of the original stacking sequence for these complicated stacking processes, we wish to understand how much it deviates in these cases. (iii) We wish to address the issue of long-range order in disordered polytypes. Thus, we also define length parameters calculable from the ε-machine and discuss their implication for finding long-range order in polytypes. (iv) Lastly, we wish to demonstrate how the architecture of the ε-machine provides an intuitive and quantitative understanding of the spatial organization of layered CPSs.

These goals are convincingly realized by analyzing DPs derived from simulated stacking sequences where there are no issues concerning experimental error. We are able to compare the ε-machine reconstructed from DPs with the ε-machine that describes the original stacking structure and, thus, we can explore how effectively εMSR captures the statistics of these complicated stacking structures. Additionally, this kind of analysis also allows us to identify possible difficulties that may arise when applying εMSR.

The present paper continues the discussion initiated in Varn *et al.* (2013) and readers desiring to fully grasp the details of the examples worked here are urged to consult that paper before continuing. Our development is organized as follows: in §2 we provide numerical details about the techniques we use to analyze the simulated DPs; in §3 we present our analysis of four simulated DPs using εMSR; in §4 we define several characteristic lengths calculable from a knowledge of the ε-machine and consider their implications for the long-range order so ubiquitous in polytypes; and in §5 we give our conclusions and directions for future work. In a companion paper we apply εMSR to DPs obtained from single-crystal X-ray diffraction experiments (Varn *et al.*, 2007).

2. Methods

We use the same notational conventions and definitions introduced previously (Varn *et al.*, 2013). For each example we begin with a stacking structure as described by an ε-machine. We generate a sample sequence from the ε-machine of length 400 000 in the Hägg notation. We map this spin sequence into a stacking orientation sequence in the *ABC* notation. We directly scan this latter sequence to find the two-layer *correlation functions* (CFs): $Q_c(n)$, $Q_a(n)$ and $Q_s(n)$ (Yi & Canright, 1996). For the disordered stacking sequences we treat here, the CFs typically decay to an asymptotic value of $1/3$ for large n. We set the CFs to $1/3$ when they reach $\simeq 1\%$ of this value, which usually occurs for $n \simeq 25$–100. We then calculate the *corrected diffracted intensity per ML*, $\mathsf{I}(\ell)$ (Varn *et al.*, 2013) in increments of $\Delta\ell = 0.001$ using equations (1) and (2) of Varn *et al.* (2013) with a stacking sequence of 10 000 MLs. Throughout, we refer to the corrected diffracted intensity per ML, $\mathsf{I}(\ell)$, as simply the DP. We now take this simulated DP as our 'experimental' DP.

We apply εMSR (Table 1 of Varn *et al.*, 2013) to each experimental DP. Since these are simulated DPs, we find that the figures-of-merit, γ and β, are equal to their theoretical values within numerical error over all unit ℓ intervals. Therefore, we do not report γ and β, and instead perform εMSR over the interval $0 \leq \ell \leq 1$. Further, again since these are simulated DPs and hence have no error, we are not able to set an acceptable threshold error Γ in advance. Instead, each example, except for Example C, minimally requires the $r = 3$ solutions. Thus, we solve the *spectral equations* (SEs) at $r = 3$ (Appendix *A*3 of Varn *et al.*, 2013) *via* a Monte Carlo technique (Varn, 2001) to find sequence probabilities of length 4. We take the $r = 3$ ε-machine given in Fig. 2 of Varn *et al.* (2013) as our default or candidate ε-machine. All *causal states* (CSs) and allowed transitions between CSs are initially assumed present. From the sequence probabilities we estimate

[2] We note that there are no inherent obstacles to applying εMSR to materials with more complicated MLs or stacking rules (Brindley, 1980; Thompson, 1981; Varn & Canright, 2001). However, the case of CPSs is by no means merely academic, since several important polytypes, such as SiC and ZnS, are describable as CPSs.

the transition matrices, $\mathsf{T}^{(s)}_{\mathcal{S}_i \to \mathcal{S}_j}$, for making a transition from a candidate CS $\mathcal{S}_i$ to a candidate CS $\mathcal{S}_j$ on seeing a spin s. We apply the equivalence relation, equation (11) of Varn *et al.* (2013), to generate a final set of CSs. We refer to the resulting ε-machine as the reconstructed or 'theoretical' $r = 3$ ε-machine for the DP. In the event that the reconstructed ε-machine assigns to a CS an asymptotic state probability of less than 0.01, we take that CS to be nonexistent. We use the same method to find the CFs and the DP for the theoretical ε-machine as we did with the initial ε-machine.

We also calculate the information-theoretic quantities described in §3.1 of Varn *et al.* (2013) for each example and the reconstructed ε-machine.[3]

We find that the most calculationally intensive portion of εMSR is solving the SEs at $r = 3$. Even so, this is generally accomplished within a few minutes on a desktop computer with a ∼2 GHz Intel Core i7 processor and several GBs of RAM. The other steps in the reconstruction process require no more than a few seconds of computer time. It is likely that the $r = 4$ SEs (not treated here) can be solved in a comparable time frame. The SEs become quite cumbersome at $r = 5$ and it is unlikely that they are as easily solvable with currently available desktop technology. However, we strongly suspect that further investigation will reveal alternate algorithms that eliminate these calculational difficulties for larger r.

3. Analysis

3.1. Example A

We begin with the sample process given in Fig. 1. One possible way to interpret this ε-machine is to decompose it into *causal-state cycles* (CSCs) (Varn *et al.*, 2013), denoted here by the sequence of CSs enclosed in square brackets []. If one does this, then one might associate the following crystal and fault structures with CSCs:[4]

2H	$[\mathcal{S}_2\mathcal{S}_5]$
$3C^+$	$[\mathcal{S}_7]$
Deformation fault	$[\mathcal{S}_5\mathcal{S}_3\mathcal{S}_7\mathcal{S}_6]$, $[\mathcal{S}_2\mathcal{S}_4\mathcal{S}_0\mathcal{S}_1]$
Growth fault	$[\mathcal{S}_5\mathcal{S}_3\mathcal{S}_6]$, $[\mathcal{S}_2\mathcal{S}_4\mathcal{S}_1]$

where the '+' on 3C indicates that only the positive chirality (. . . 1111 . . .) structure is present. The faulting is given with reference to the 2H crystal.[5]

The DP from this process is shown in Fig. 2. The experienced crystallographer has little difficulty guessing the underlying crystal structure: the peaks at $\ell \simeq 0.50$ and at $\ell \simeq 1.00$ indicate 2H structure; while the peak at $\ell \simeq 0.33$ is characteristic of the 3C structure.

The faulting structure is less clear, however. It is known that various kinds of faults produce different effects on the Bragg peaks (Sebastian & Krishna, 1994). For instance, both growth and deformation faults broaden the peaks in the DP of the 2H structure. The difference is that growth faults broaden the integer-ℓ peaks three times more than the half-integer-ℓ peaks, while peaks broadened due to deformation faulting are about equal. The full width at half-maximum (FWHM) for the peaks are 0.028, 0.034 and 0.049 for $\ell \simeq 0.33$, 0.5 and 1, respectively. This gives, then, a ratio of about 1.4 for the integer-ℓ to half-integer-ℓ broadening, suggesting (perhaps) that deformation faulting is prominent. One expects there to be no shift in the position of the peaks for either growth or deformation faulting; which is not the case here. In fact, the two peaks associated with the 2H structure at $\ell \simeq 0.50$ and 1.00 are shifted by $\Delta\ell \simeq 0.006$ and 0.009, respectively. This analysis is, of course, only justified for one parent crystal in the overall structure; nonetheless if we neglect the peak shifts, the simple intuitive analysis appears to give good qualitative results here.

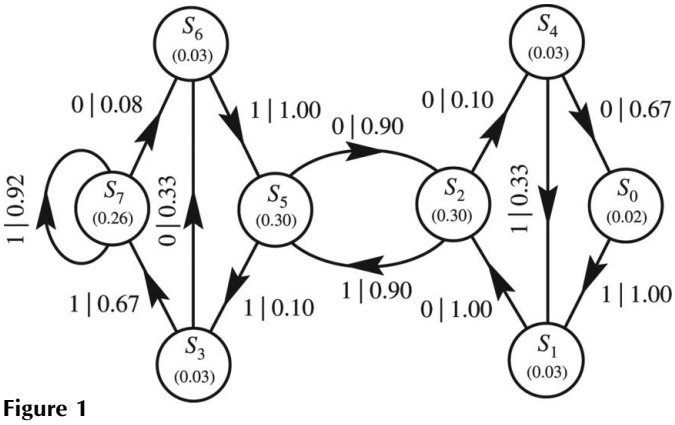

Figure 1
The $r = 3$ theoretical and experimental ε-machine for the Example A process. The nodes represent CSs and the directed arcs are transitions between them. The edge labels $s|p$ indicate that a transition occurs between the two CSs on symbol s with probability p. The asymptotic probabilities for each CS are given in parentheses. We label the states with the last three spins observed in base-10 notation. (A chart for converting base-10 into base-2 is given in Table 1.) The large probabilities to repeat the CSCs $[\mathcal{S}_7]$ and $[\mathcal{S}_2\mathcal{S}_5]$ suggest that one thinks of these cycles as crystal structure and everything else as faulting.

Table 1
A table for translating base-10 notation into binary notation of length 3.

This is useful for converting the base-10 subscripts of the CSs in Figs. 1, 5 and 13 into the corresponding binary spin sequences.

Base 10	Base 2
0	000
1	001
2	010
3	011
4	100
5	101
6	110
7	111

[3] Although we restrict our attention to those computational quantities defined in Varn *et al.* (2013), there have recently been other information-theoretic measures proposed to characterize complexity and natural information processing in materials (Cartwright & Mackay, 2012).

[4] A detailed survey of the structure of the ε-machine containing faulting sequences is a current topic of research. An approximate scheme relating faulting structures to causal-state paths on an ε-machine is given in Varn (2001).

[5] Here and elsewhere, we use the Ramsdell notation to specify crystalline stacking structures in CPSs. Recall that the ε-machine gives stacking sequences in terms of the Hägg notation. Also, we replace the usual '+' and '−' of the Hägg notation with '1' and '0', respectively.

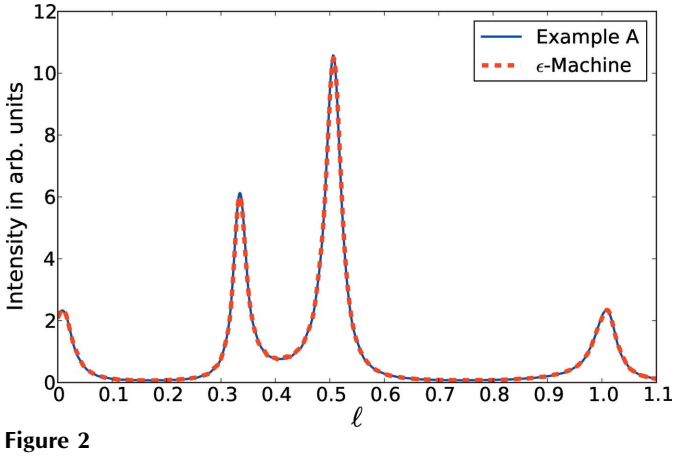

Figure 2
A comparison between the DPs $\mathrm{I}(\ell)$ generated by Example A (blue solid line) and by the $r = 3$ spectrally reconstructed ε-machine (red dashed line). The differences between the DP for Example A and the $r = 3$ reconstructed ε-machine are too small to be seen. We calculate $\mathcal{R} = 2\%$, but this is largely due to numerical error (see text.) The peak at $\ell \simeq 1/3$ corresponds to the 3C structure and the two peaks at $\ell \simeq 1/2$ and $\ell \simeq 1$ to the 2H structure.

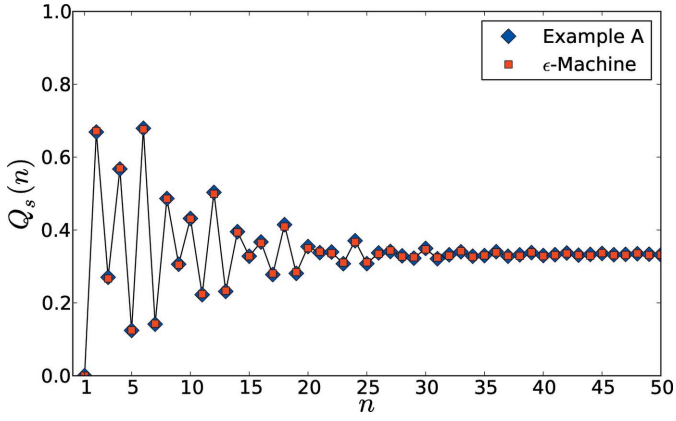

Figure 3
A comparison of the CFs $Q_s(n)$ between the Example A process (blue diamonds) and the $r = 3$ reconstructed ε-machine (red squares). As with the DPs the differences are too small to be seen on the graph. As an aid to the eye, here and in other graphs showing CFs, we connect the values of adjacent CFs with straight lines. The CFs, of course, are defined only for integer values of n.

With the 3C peak, both deformation and growth faults produce a broadening, the difference being that the broadening is asymmetrical for the growth faults. One also expects there to be some peak shifting for the deformation faulting. There is a slight shift ($\Delta\ell \simeq 0.002$) for the $\ell \simeq 0.33$ peak and the broadening seems (arguably) symmetric, so one is tempted to guess that deformation faulting is important here. Indeed, $[\mathcal{S}_7\mathcal{S}_6\mathcal{S}_5\mathcal{S}_3]$ is consistent with deformation faulting in the 3C crystal. Heuristic arguments, while not justified here, seem to give qualitative agreement with the known structure.

The ε-machine description does better. The reconstructed ε-machine is equivalent to the original one, with CS probabilities and transition probabilities typically within 0.1% of their original values, except for the transition probability from from $\mathcal{S}_4$ to $\mathcal{S}_1$, $\mathsf{T}^{(1)}_{\mathcal{S}_4\to\mathcal{S}_1} = 0.33$, which was 1% too small. Not surprisingly, the process shown in Fig. 1 *is* the reconstructed ε-machine and so we do not repeat it.

The two-layer CFs $Q_s(n)$ *versus* n from the process and from the reconstructed ε-machine are shown in Fig. 3. The differences are too small to be seen on the graph. We calculate the profile $\mathcal{R}$ factor (Varn *et al.*, 2013) to compare the experimental DP (Example A) to the theoretical DP (reconstructed ε-machine) and find a value of $\mathcal{R} = 2\%$. If we generate several DPs from the same process, we find profile $\mathcal{R}$ factors of similar magnitude. This error then must be due to sampling. It stems from the finite spin sequence length we use to calculate the CFs and our method for setting them equal to their asymptotic value. This can be improved by taking longer sample sequence lengths and refining the procedure for setting the CFs to their asymptotic value. Since profile $\mathcal{R}$ factors comparing theory and experiment are typically much larger than this, at present, this does not seem problematic. A comparison of the two DPs is shown in Fig. 2. This kind of agreement is typical of εMSR from any process that can be represented as an $r = 3$ ε-machine (Varn, 2001).

We find by direct calculation from the ε-machine that both Example A and the reconstructed process have a configurational entropy of $h_\mu = 0.44$ bits/ML, a statistical complexity of $C_\mu = 2.27$ bits and an excess entropy of $\mathbf{E} = 0.95$ bits.

Example A illustrates several points. (i) For processes that are representable as an $r = 3$ ε-machine, the reconstruction procedure typically reproduces the ε-machine within numerical error. (ii) The ε-machine can accommodate two distinct crystal structures (3C and 2H) and the faulting between them. (iii) Although the faulting was weak enough so that the DP retained Bragg-like peaks, we did not need to incorporate this information into the reconstruction procedure. εMSR detected the underlying crystal structures without explicit intervention. (iv) εMSR has no difficulty modeling crystals that are *not* spin inversion symmetric (*i.e.* $0 \Leftrightarrow 1$) (Varn & Canright, 2001). That is, εMSR correctly found that only the positive chirality 3C structure was present.

3.2. Example B

Upon annealing, solid-state transformations can occur in many polytypes (Sebastian & Krishna, 1994). Here, we attempt to model twinned 3C structures in the presence of 6H structure. However, since two crystal structures represented on an ε-machine cannot share a CS, an $r = 3$ ε-machine has an insufficient memory to capture simultaneously both 3C and 6H structures. For example, on an $r = 3$ ε-machine, $3\mathrm{C}^+$ and 6H share the stacking sequence 111 and, hence, each must visit $\mathcal{S}_7$. The transition probabilities from this latter CS cannot specify both that the next symbol have a high probability of being 1 (and thus create $3\mathrm{C}^+$) and a high probability of being 0 (and thus generate 6H). In fact, it is necessary to use an $r = 4$ ε-machine to encompass both structures.

The $r = 4$ ε-machine in Fig. 4 does just this. $[\mathcal{R}_1\mathcal{R}_3\mathcal{R}_7\mathcal{R}_{14}\mathcal{R}_{12}\mathcal{R}_8]$ is the CSC associated with 6H, although the probability of repeating this CSC more than once is low

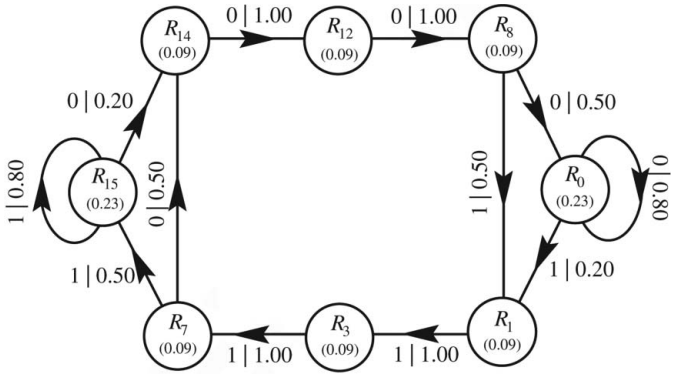

Figure 4
The experimental ε-machine for Example B. Since it has a memory of $r_\ell = 4$, we label the states with the last four spins observed: *i.e.*, $\mathcal{R}_{12}$ means that 1100 were the last four spins. (A chart for converting base 10 into base 2 is given in Table 2.) The CSCs $[\mathcal{R}_{15}]$ and $[\mathcal{R}_0]$ give rise to 3C structure and the CSC $[\mathcal{R}_1\mathcal{R}_3\mathcal{R}_7\mathcal{R}_{14}\mathcal{R}_{12}\mathcal{R}_8]$ generates 6H structure.

Table 2
A table for translating base-10 notation into binary notation of length 4.

This is useful for converting the base-10 subscripts of the CSs in Fig. 4 into the corresponding binary spin sequences.

Base 10	Base 2
0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111
8	1000
9	1001
10	1010
11	1011
12	1100
13	1101
14	1110
15	1111

owing to the transition probabilities out of CSs $\mathcal{R}_7$ and $\mathcal{R}_8$. $[\mathcal{R}_0]$ and $[\mathcal{R}_{15}]$ give the twinned 3C structures.

Employing spectral reconstruction, we find the $r = 3$ ε-machine shown in Fig. 5. All CSs are present and all transitions, save those that connect $\mathcal{S}_2$ and $\mathcal{S}_5$, are present. A comparison of the CFs for the original process and the reconstructed ε-machine is given in Fig. 6. The agreement is remarkably good. It seems that the $r = 3$ ε-machine picks up most of the structure in the original process.

There is similar, though not as good, agreement in the DPs, as Fig. 7 shows. The most notable discrepancies are in the small rises at $\ell \simeq 0.17$ and $\ell \simeq 0.83$. We calculate a profile $\mathcal{R}$ factor of $\mathcal{R}$ = 12% between the DPs for Example B and the reconstructed ε-machine. The $r = 3$ ε-machine has difficulty reproducing the 6H structure in the presence of 3C structure, as expected.

Given the good agreement between the CFs and the DPs generated by Example B and the $r = 3$ ε-machine, we are led to ask what the differences between the two are. In Table 3 we

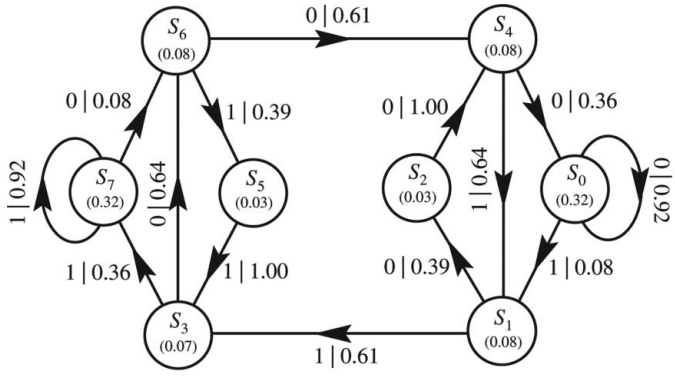

Figure 5
The reconstructed (theoretical) ε-machine at $r = 3$ for Example B. The absence of arcs connecting the $\mathcal{S}_2$ and $\mathcal{S}_5$ CSs indicate that εMSR has correctly identified that there is no 2H structure in this crystal.

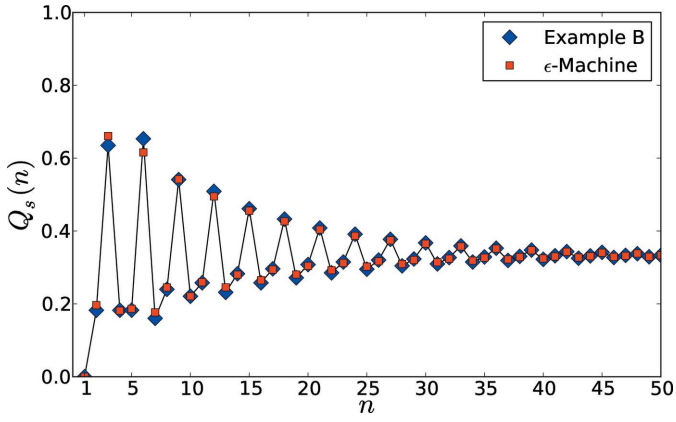

Figure 6
A comparison of the CFs $Q_s(n)$ generated by the $r = 3$ reconstructed ε-machine (red squares) and generated by Example B (blue diamonds). The agreement is generally quite good, except for perhaps $n = 3, 6, 7$.

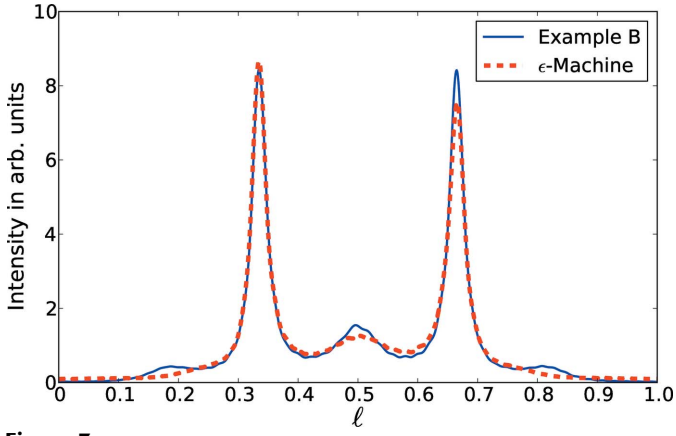

Figure 7
A comparison of the DPs $I(\ell)$ between the $r = 3$ reconstructed ε-machine (red dashed line) and the process of Example B (solid blue line). The agreement is surprisingly good; we calculate a profile $\mathcal{R}$ factor of $\mathcal{R}$ = 12%. The small enhanced scattering at $\ell \simeq 1/6$ and $\ell \simeq 5/6$ corresponds to the 6H structure. The $r = 3$ ε-machine has difficulty in reproducing these because the 6H and the 3C structure both share the $\mathcal{S}_7$ and $\mathcal{S}_0$ CSs and so require an ε-machine reconstructed at $r = 4$ to properly disambiguate them.

Table 3
The frequencies of length-3 sequences obtained from Example B and the ε-machine reconstructed at $r = 3$.

Sequence	Example B	εMSR
111	0.318	0.324
110	0.091	0.081
101	0.000	0.027
100	0.091	0.076
011	0.091	0.070
010	0.000	0.026
001	0.091	0.076
000	0.318	0.322

give the frequencies of the eight length-3 sequences generated by each process. The agreement is excellent. They both give nearly the same probabilities ($\sim$0.32) for the most common length-3 sequences, 111 and 000. Example B does forbid two length-3 sequences, 101 and 010, which the reconstructed $r = 3$ ε-machine allows with a small probability ($\sim$0.03). At the level of length-3 sequences, the ε-machine is capturing most of the structure in the stacking sequence.

A similar analysis allows us to compare the probabilities of the 16 length-4 sequences generated by each; the results are given in Table 4. There are more striking differences here. The frequencies of the two most common length-4 sequences in Example B, $\Pr^{(\mathrm{Ex})}(1111) = \Pr^{(\mathrm{Ex})}(0000) = 0.227$, are overestimated by the $r = 3$ ε-machine, which assigns them a probability of $\Pr^{(\mathrm{Th})}(1111) \simeq \Pr^{(\mathrm{Th})}(0000) \simeq 0.30$ each. Similarly, sequences forbidden by Example B – 1101, 1011, 1010, 1001, 0110, 0101, 0100, 0010 – are not necessarily forbidden by the $r = 3$ ε-machine. In fact, the $r = 3$ ε-machine forbids only two of them, 0101 and 1010. This implies that the $r = 3$ ε-machine can find spurious sequences that are not in the original stacking sequence. This is to be expected. But the $r = 3$ ε-machine *does* detect important features of the original process. It finds that this is a twinned 3C structure. It also finds that 2H structure plays no role in the stacking process. (We see this by the absence of transitions between $\mathcal{S}_2$ and $\mathcal{S}_5$ in Fig. 5.)

We can also compare the probability that each ε-machine assigns to seeing a 111000 sequence, the sequence that generates the 6H structure. For Example B, direct calculation from the ε-machine, Fig. 4, gives a probability of $\Pr^{(\mathrm{Ex})}(111000) = 0.09$. In contrast, a similar calculation from the theoretical ε-machine, Fig. 5, gives a much lower value, $\Pr^{(\mathrm{Th})}(111000) = 0.006$. This discrepancy is directly related to the fact that 3C and 6H share the $\mathcal{S}_0$ and $\mathcal{S}_7$ CSs and, thus, the $r = 3$ ε-machine cannot simultaneously model both 3C and 6H structure.

We find by direct calculation that the Example B process has a configurational entropy of $h_\mu = 0.51$ bits/ML, a statistical complexity of $C_\mu = 2.86$ bits and an excess entropy of $\mathbf{E} = 0.82$ bits. The reconstructed process gives similar results with a configurational entropy $h_\mu = 0.54$ bits/ML, a statistical complexity of $C_\mu = 2.44$ bits and an excess entropy of $\mathbf{E} = 0.83$ bits.

Example B illustrates several points. (i) It is possible, at least in some cases, to model an $r = 4$ process on an $r = 3$ ε-machine. It is likely, though, that εMSR at $r = 3$ will fail for some processes not describable by an $r = 3$ ε-machine. We discuss this in more detail in §3.5. (ii) Even though εMSR does not detect the true process here, it does reveal important structural features of that process. (iii) The sequence probabilities found from solving the SEs can differ from those of the true process and εMSR can assign small probabilities to sequences forbidden by the true process.

Table 4
The frequencies of length-4 sequences obtained from Example B and the ε-machine reconstructed at $r = 3$.

Sequence	Example B	εMSR
1111	0.227	0.300
1110	0.091	0.024
1101	0.000	0.029
1100	0.091	0.052
1011	0.000	0.027
1010	0.000	0.000
1001	0.000	0.049
1000	0.091	0.027
0111	0.091	0.025
0110	0.000	0.045
0101	0.000	0.000
0100	0.000	0.026
0011	0.091	0.046
0010	0.000	0.030
0001	0.091	0.026
0000	0.227	0.296

3.3. Example C

We treat this next system, Example C, to contrast it with the last and to demonstrate how pasts with equivalent futures are merged to form CSs. The ε-machine for this system is shown in Fig. 8 and is known as the *Golden Mean Process*. The rule for generating the Golden Mean Process is simply stated: a 0 or 1 are allowed with equal probability unless the previous spin was a 0, in which case the next spin is a 1. Clearly then, this process needs only to remember the previous spin and, hence, it has a memory length of $r = 1$. It forbids the sequence 00 and all sequences that contain this as a subsequence. The process is so-named because the total number of allowed sequences grows with sequence length at a rate given by the golden mean $\varphi = (1 + \sqrt{5})/2$.

We employ the εMSR algorithm and find the ε-machine given (again) in Fig. 8 at $r = 1$. A comparison of the CFs from Example C and the Golden Mean Process are given in Fig. 9. The differences are too small to be seen. We next compare the DPs and these are shown in Fig. 10. We find excellent agreement and calculate a profile $\mathcal{R}$ factor of $\mathcal{R} = 2\%$. At this point

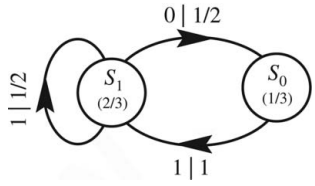

Figure 8
The recurrent portion of the ε-machine for the Golden Mean Process, Example C. The process has a memory length of $r = 1$, and so we label each CS by the last spin seen.

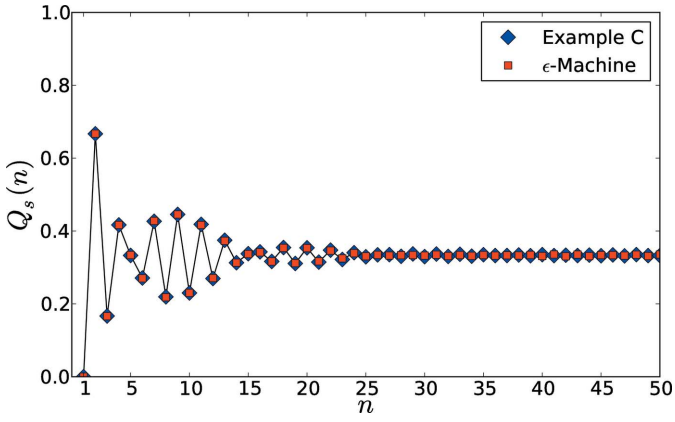

Figure 9
A comparison of the CFs $Q_s(n)$ generated by the $r = 1$ reconstructed ε-machine (red squares) and the Golden Mean Process of Example C (blue diamonds). The CFs decay quickly to their asymptotic value of 1/3.

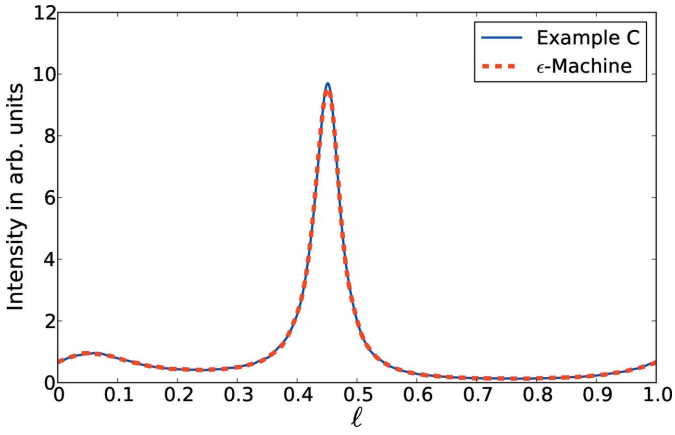

Figure 10
A comparison of the DPs for Example C (solid blue line) and the reconstructed $r = 1$ ε-machine (dashed red line). The agreement is excellent. One finds a profile $\mathcal{R}$ factor of 2% between the experimental DP, Example C and the theoretical DP calculated from the reconstructed ε-machine.

εMSR should terminate, as we have found satisfactory agreement (to within the numerical error of our technique) between 'experiment', Example C, and 'theory', the reconstructed ε-machine.

Let us suppose that instead, we increment r and follow the εMSR algorithm as if the agreement at $r = 1$ had been unsatisfactory. In this case, we would have generated the 'ε-machine' shown in Fig. 11 at the end of step 3(b) [Table 1 of Varn *et al.* (2013)]. We have yet to apply the equivalence relation, equation (11) of Varn *et al.* (2013), and so let us call this the *nonminimal* ε-machine. That is, we have not yet combined pasts with equivalent futures to form CSs, step 3(c) [Table 1 of Varn *et al.* (2013)]. Let us do that now.

We observe that the state $\mathcal{S}_2$ is different from the other two, $\mathcal{S}_1$ and $\mathcal{S}_3$, in that one can only see the spin 1 upon leaving this state. Therefore, it cannot possibly share the same futures as $\mathcal{S}_1$ and $\mathcal{S}_3$, so no equivalence between them is possible. However, we do see that $\Pr(1|\mathcal{S}_1) = \Pr(1|\mathcal{S}_3) = 1/2$ and $\Pr(0|\mathcal{S}_1) = \Pr(0|\mathcal{S}_3) = 1/2$ and, thus, these states share the

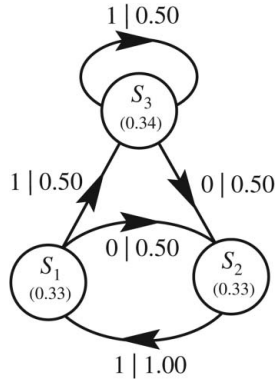

Figure 11
The $r = 2$ reconstructed nonminimal ε-machine for the Golden Mean Process, Example C. Applying the equivalence relation, equation (11) of Varn *et al.* (2013), we find that $\mathcal{S}_1$ and $\mathcal{S}_3$ have the same futures, and thus should be collapsed into a single CS. Doing so gives the ε-machine in Fig. 8.

same probability of seeing futures of length 1. More formally, we can write

$$\mathsf{T}^{(s)}_{01\to 1s} = \mathsf{T}^{(s)}_{11\to 1s}. \tag{1}$$

Since we are labeling the states by the last two symbols seen at $r = 2$, within our approximation they do have the same futures and so $\mathcal{S}_1$ and $\mathcal{S}_3$ can be merged to form a single CS. The result is the ε-machine shown in Fig. 8.

In general, in order to merge two histories, we check that each has an equivalent future up to the memory length r. In this example, we need only check futures up to length 1 since, after the addition of one spin (s), each is labeled by the same past: namely, 1s. Had we tried to merge the pasts 11 and 10, we would need to check all possible futures after the addition of *two* spins, after which the states would have the same futures (by assumption). That is, we would require

$$\mathsf{T}^{(s)}_{11\to 1s} = \mathsf{T}^{(s)}_{10\to 0s} \tag{2}$$

and

$$\mathsf{T}^{(s')}_{1s\to ss'} = \mathsf{T}^{(s')}_{0s\to ss'} \tag{3}$$

for all s and s'.

We find by direct calculation from the ε-machine that both Example C and the reconstructed process have a configurational entropy of $h_\mu = 0.67$ bits/ML, a statistical complexity of $C_\mu = 0.92$ bits and an excess entropy of $\mathbf{E} = 0.25$ bits.

Example C illustrates several important aspects of εMSR. (i) We explicitly demonstrate the merging of pasts that have equivalent futures. Thus, εMSR builds a model (within the space of Markov processes) that invokes the least complexity to describe the DP. (ii) With $h_\mu = 0.67$ bits/ML,[6] Example C has significant disorder. Nonetheless, εMSR has no difficulty finding the true process. (iii) εMSR needs no *a priori* information about the underlying crystal structure. Indeed, Example C really does not have any underlying

[6] For comparison, a completely random stacking of MLs for CPSs would have $h_\mu = 1$ bit/ML.

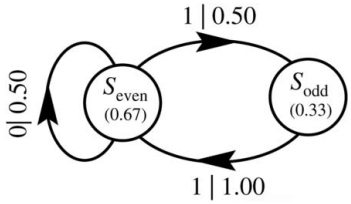

Figure 12
The recurrent portion of the ε-machine for the Even Process, Example D. Since the CSs cannot be specified by a finite history of previous spins, we have labeled them $\mathcal{S}_{\text{even}}$ and $\mathcal{S}_{\text{odd}}$. We find that this ε-machine has a statistical complexity of $C_\mu = 0.92$ bits.

crystal structure. (iv) As with Example A, the fact that this process does not have spin inversion symmetry presents no difficulties.

3.4. Example D

We now consider a simple finite-state process that cannot be represented by a finite-order Markov process, called the *Even Process* (Crutchfield & Feldman, 2003; Crutchfield, 1992), as the previous examples could. The *Even Language* (Hopcroft & Ullman, 1979; Badii & Politi, 1997) consists of sequences such that between any two 0s either there are no 1s or an even number of 1s. In a sequence, therefore, if the immediately preceding spin was a 1, then the admissibility of the next spin requires remembering the *evenness* of the number of previous consecutive 1s, since seeing the last 0. In the most general instance, this requires an indefinitely long memory and so the Even Process cannot be represented by any finite-order Markov chain.

We define the Even Process as follows: if a 0 or an even number of consecutive 1s were the last spin(s) seen, then the next spin is either 1 or 0 with equal probability; otherwise the next spin is 1. While this might seem somewhat artificial for the stacking of simple polytypes, one cannot exclude this class of (so-called *sofic*) structures on physical grounds. Indeed, such long-range memories may be induced in solid-state phase transformations between two crystal structures (Kabra & Pandey, 1988; Varn & Crutchfield, 2004). It is instructive, therefore, to explore the results of our procedure on processes with such structures.

Additionally, analyzing a sofic process provides a valuable test of εMSR as practiced here. Specifically, we invoke a finite-order Markov approximation for the solution of the $r = 3$ equations and we shall determine how closely this approximates the Even Process with its effectively infinite range.

The ε-machine for this process is shown in Fig. 12. Its causal-state transition structure is equivalent to that in the ε-machine for the Golden Mean Process. They differ only in the *spins* emitted upon transitions out of the $\mathcal{S}_1$ ($\mathcal{S}_{\text{even}}$) CS. It seems, then, that this process should be easy to detect.

The result of ε-machine reconstruction at $r = 3$ is shown in Fig. 13. Again, it is interesting to see if the sequences forbidden by the Even Process are also forbidden by the $r = 3$ ε-machine. One finds that the sequence 010 – forbidden by the process – is also forbidden by the reconstructed ε-machine. This occurs because $\mathcal{S}_2$ is missing.[7] We do notice that the reconstructed ε-machine has much more 'structure' than the original process. We now examine the source of this additional structure.

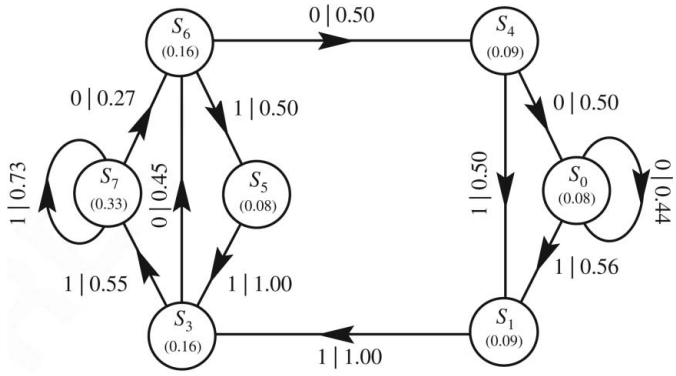

Figure 13
The $r = 3$ reconstructed ε-machine for the Even Process of Example D. Since the Even Process forbids the sequences $\{01^{2k+1}0, k = 0, 1, 2, \ldots\}$ and all sequences containing them, it is satisfying to see that 010 is forbidden by the reconstructed ε-machine, as shown by the missing $\mathcal{S}_2$ CS. We find that $C_\mu = 2.58$ bits.

Let us first contrast differences between εMSR and other ε-machine reconstruction techniques, taking the subtree-merging method (SMM) (Crutchfield & Young, 1989; Hansen, 1993; Crutchfield, 1994) as the alternative prototype. There are two major differences. First, since here we estimate sequence probabilities from the DPs and not a symbol sequence, we find it necessary to invoke the memory-length reduction approximation (Varn *et al.*, 2013) at $r \geq 3$ to obtain a complete set of equations. Specifically, we assume that (i) only histories up to range r are needed to make an optimal prediction of the next spin and (ii) we can label CSs by their length-r history.

We can test these assumptions in the following way. For (i), we compare the frequencies of length-4 sequences obtained from each method. This is shown in Table 5. The agreement is excellent. All sequence frequencies are within ± 0.01 of the correct values. The small differences are due to the memory-length reduction approximation. So this does have an effect, but it is small here.

To test (ii), we can compare the ε-machines generated from each method given the same 'exact' or 'correct' length-4 sequence probabilities. Doing so, SMM gives the ε-machine for the Even Process shown in Fig. 12. εMSR gives a different result. After merging pasts with equivalent futures, one finds the ε-machine shown in Fig. 14. For clarity, we explicitly show the length-3 sequence histories associated with each CS, but do not write out the asymptotic state probabilities.

[7] We do note that the solution of the SEs at $r = 3$ assigns the sequences 0100 and 0010 a small probability, $\Pr^{(\text{Th})}(0100) \simeq \Pr^{(\text{Th})}(0010) \simeq 0.005$, which implies that the sequence 010 is also present with a small probability, $\Pr^{(\text{Th})}(010) < 0.01$. Since this falls below our threshold, we take this CS as being nonexistent. For this example, probabilities of this small magnitude are not meaningful, as the SEs at $r = 3$ are difficult to satisfy with purely real probabilities. We also note that the solution of the SEs at $r = 2$ *does* forbid the 010 sequence. For additional discussion, see Varn (2001).

Table 5
The frequencies of length-4 sequences obtained from εMSR and SMM for the Even Process, Example D.

At most, they differ by ± 0.01.

Sequence	εMSR	SMM
1111	0.24	0.25
1110	0.09	0.08
1101	0.09	0.08
1100	0.08	0.08
1011	0.08	0.08
1010	0.00	0.00
1001	0.04	0.04
1000	0.04	0.04
0111	0.09	0.08
0110	0.07	0.08
0101	0.00	0.00
0100	<0.01	0.00
0011	0.08	0.08
0010	<0.01	0.00
0001	0.05	0.04
0000	0.04	0.04

The ε-machine generated by εMSR is in some respects as good as that generated by SMM. Both reproduce the sequence probabilities up to length 4 from which they were estimated. The difference is that for εMSR, our insistence that histories be labeled by the last r spins forces the representation to be Markovian of range r. Here, a simpler model for the process, as measured by the smaller statistical complexity ($C_\mu = 0.92$ bits as compared to 1.92 bits), can be found. So, the notion of minimality is violated. That is, εMSR searches only a subset of the space from which processes can belong. Should the true process lie outside this subset (Markovian processes of range r), then εMSR returns an approximation to the true process. The approximation may be both more complex and less predictive than the true process. It is interesting to note that had we given SMM the sequence probabilities found from the solutions of the SEs, we would have found (within some error) the ε-machine given in Fig. 12.

We find, then, that there are two separate consequences to applying εMSR that affect the reconstructed ε-machine. The first is that for $r \geq 3$, the memory-length reduction approximation must be invoked to obtain a complete set of equations. This approximation limits the histories treated and can affect the values estimated for the sequence probabilities. The second is the state-labeling scheme. Only for Markovian (nonsofic) processes can CSs be labeled by a unique finite history. Making this assumption effectively limits the class of processes one can detect to those that are block-r Markovian. To see this more clearly, we can catalog the possible histories that lead to the two CSs in Fig. 12. In doing so, we find that the histories 000, 011, 110, 100 and 100 always leave the process in CS $\mathcal{S}_{\rm even}$. Similarly, the histories 001 and 101 always leave the process in CS $\mathcal{S}_{\rm odd}$. But having seen the history 111 does not specify the CS as one can arrive in both CSs from this history. So, the labeling of CSs by histories of a finite length fails here.

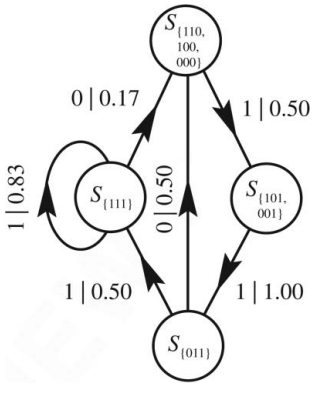

Figure 14
The ε-machine inferred from the exact sequence frequencies for the Even Process of Example D. The CSs are labeled with the (possibly several) length-3 histories that can lead to them. We find that $C_\mu = 1.92$ bits.

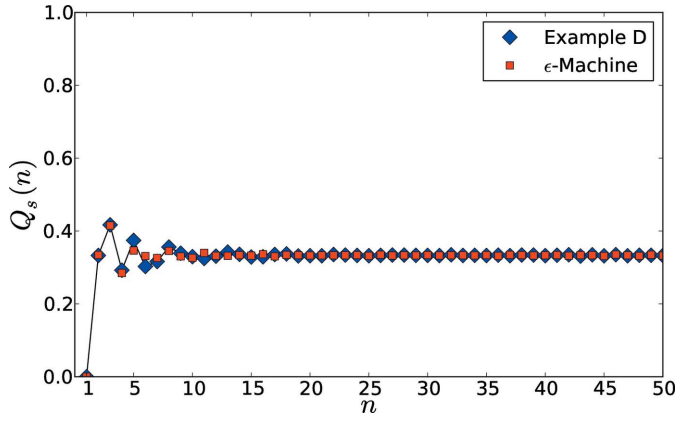

Figure 15
A comparison of the CFs $Q_s(n)$ generated by the $r = 3$ reconstructed ε-machine (red squares) and the Even Process of Example D (blue diamonds). The CFs decay quickly to their asymptotic value of 1/3.

Then, why do we not find sequence probabilities by solving the SEs, using SMM to reconstruct the ε-machine? There are two reasons. The first is that in general one must know sequence probabilities for longer sequences than is necessary for εMSR. Solving the SEs for these longer sequence frequencies is onerous. The second is that error in the sequence probabilities found from solving the SEs for these longer sequences makes identifying equivalent pasts almost impossible. The Even Process is an exception here, since one needs to consider only futures of length 1. This is certainly not the case in general.

Having explored the differences between εMSR and SSM, we now return to a comparison between CFs and DPs generated by the εMSR and the Even Process. The CFs for the Even Process and the reconstructed ε-machine are given in Fig. 15. We see that both decay quite quickly to their asymptotic values of 1/3. There is good agreement, except in the region between $5 \lesssim n \lesssim 10$. Examining the DPs in Fig. 16, we see that there is likewise good agreement except in the region $0.7 \lesssim \ell \lesssim 0.9$. We calculate the profile $\mathcal{R}$ factor between the theoretical and experimental DPs to be $\mathcal{R} = 9\%$.

We find by direct calculation from the Even Process that it has a configurational entropy of $h_\mu = 0.67$ bits/ML, a statistical complexity of $C_\mu = 0.92$ bits and an excess entropy of $\mathbf{E} = 0.91$ bits. The reconstructed ε-machine gives information-

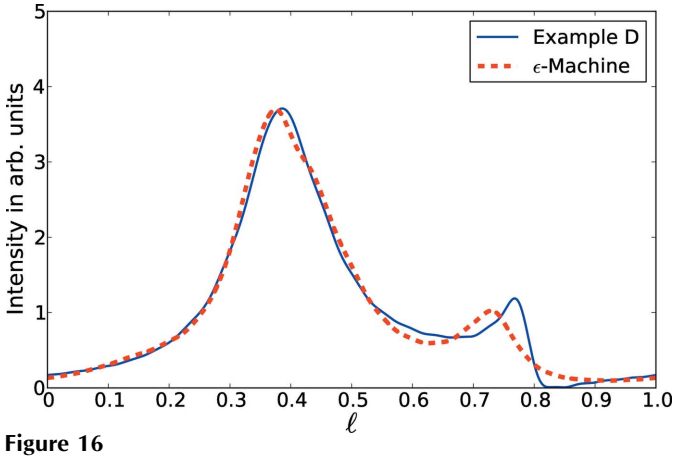

Figure 16
A comparison between the DPs $\mathrm{I}(\ell)$ generated by the $r = 3$ reconstructed ε-machine (red dashed line) and by the Even Process of Example D (solid blue line). The agreement is good ($\mathcal{R}$ = 9%) except in the region $0.7 \lesssim \ell \lesssim 0.9$. Notably, the DP for the Even Process has an isolated zero at $\ell = 5/6$.

theoretic quantities that are rather different. We find a configurational entropy of $h_\mu = 0.79$ bits/ML, a statistical complexity of C_μ = 2.58 bits and an excess entropy of $\mathbf{E}$ = 0.21 bits. Table 6 summarizes these properties along with those, for comparison, of the previous examples.

One reason that the reconstructed ε-machine gives CFs and DPs in such good agreement with the Even Process in spite of the fact that the information-theoretic quantities are different is the insensitivity of the CFs and DPs to the frequencies of individual long sequences: equation (9) of Varn *et al.* (2013) sums sequence probabilities to find CFs. The fact that the Even Process has such a long memory is masked by this. However, information-theoretic quantities are sensitive to the structure of long sequences. εMSR at $r = 4$ should prove interesting, in this light, since the Even Process picks up another forbidden sequence – 01110 – and this additional structure would be reflected in the reconstructed ε-machine.

Example D illustrates several significant features of εMSR. (i) Most importantly, *εMSR is limited to discovering Markov processes only*. (ii) εMSR can result in an ε-machine that is *more complex* than the original process (2.58 *versus* 0.92 bits) but *less predictive* (0.79 *versus* 0.67 bits/ML). Nonetheless, the resultant ε-machine does approximate the true process. (iii) As in Example C, we find that εMSR is applicable for DPs where there is only broadband scattering without Bragg or Bragg-like peaks and, thus, high disorder and no discernible underlying crystal structure.

3.5. Challenges when applying εMSR

We have considered four examples that demonstrate successful applications of εMSR. We have found instances, however, when εMSR has difficulties converging to a satisfactory result. We now analyze each step in εMSR as given in Table 1 of Varn *et al.* (2013) and discuss possible problems that may be encountered. We concentrate here on issues that arise primarily in the application of εMSR. There are of course many experimental considerations that frustrate the analysis of DPs in layered materials, some of which are discussed in, *e.g.*, Velterop *et al.* (2000).

Table 6
Measures of intrinsic computation calculated from the processes of Examples A, B, C and D, and their ($r = 3$) reconstructed ε-machines.

For Examples A, B and C the reconstructed ε-machines give good agreement. For Example D, however, the reconstructed ε-machine requires more memory and still has an entropy density h_μ significantly higher than that of the Even Process. The last column gives $\Delta = C_\mu - \mathbf{E} - rh_\mu$ as a consistency check derived from equation (23) of Varn *et al.* (2013), which describes order-r Markov processes. Recall that the Even Process of Example D is not a finite-r process and so equation (23) of Varn *et al.* (2013) does not hold. All one can say is that $\mathbf{E} \leq C_\mu$ (Shalizi & Crutchfield, 2001), which is the case for Example D.

System	Range	h_μ (bits/ML)	C_μ (bits)	$\mathbf{E}$ (bits)	Δ (bits)
Example A	3	0.44	2.27	0.95	0.00
ε-machine	3	0.44	2.27	0.95	0.00
Example B	4	0.51	2.86	0.82	0.00
ε-machine	3	0.54	2.44	0.83	−0.01
Example C	1	0.67	0.92	0.25	0.00
ε-machine	1	0.67	0.92	0.25	0.00
Example D	∞	0.67	0.92	0.91	
ε-machine	3	0.79	2.58	0.21	0.00

Step 1. Several problems can arise here due to data quality. One is that the figures-of-merit, β and γ, are sufficiently different from their theoretical values over all possible ℓ intervals that εMSR should not even be attempted. Even if one does find an interval such that they indicate satisfactory DPs, it is possible that the CFs extracted over this interval are unphysical. That is, there is no guarantee that all of the CFs are both positive and less than unity. In such a case, no stacking of MLs can reproduce these CFs. Finally, if error ranges have not been reported with the experimental data, it may not be possible to set the error threshold Γ.

Step 2. The $\Pr(\omega^r)$ solutions to the SEs are not guaranteed to be either real or positive for $r \geq 3$. If this is so, then no physical stacking of MLs can reproduce the CFs from the DP.

Step 3. Given $\Pr(\omega^r)$ that satisfy the elementary conditions of probability (*i.e.*, there is no difficulty at step 2), step 3 will return a machine that generates $\Pr(\omega^r)$. It is possible, however, that the resulting CSs are not *strongly connected*, and thus the result may not be interpreted as a single ε-machine.

Step 4. There are no difficulties here.

Step 5. It is possible that one is required to go to an r that is cumbersome to calculate. In this case, one terminates the procedure through practicality.

We find that the roots of these difficulties can be ultimately traced to four problems: (i) excessive error in the DP, (ii) the process has statistics that are too complex to be captured by a finite-range Markov process, (iii) the memory-length approximation is not satisfied and (iv) the initial assumptions of polytypism are violated. We are likely to discover (i) in step 1. For (ii) and (iii), we find no difficulties at step 1, but rather at steps 2, 3 and 5. For (iv), we have not examined this case in detail. However, we expect that if the assumptions of the

stacking of MLs [see §2.1 of Varn *et al.* (2013)] are not met then, since equation (1) of Varn *et al.* (2013) is no longer valid, the CFs found by Fourier analysis will not reflect the actual correlations between MLs. This will likely be interpreted as poor figures-of-merit and εMSR will terminate at step 1.

Of the four possible difficulties only (ii) and (iii) should be considered to be inherent to εMSR. It is satisfying that εMSR can detect errors in the DP and then stop, so that it does not generate an invalid representation that simply describes 'error' or 'noise'.

4. Characteristic lengths in CPSs

We now return to one of the mysteries of polytypism, namely that of the long-range order which they seem to possess. It is of interest, then, to ask what, if anything, the spectrally reconstructed ε-machine indicates about the range of interactions between MLs. In this section, we discuss and quantify several characteristic lengths that can be estimated from reconstructed ε-machines.

(i) *Correlation length,* λ_c. From statistical mechanics, we have the notion of a correlation length (Binney *et al.*, 1992; Yeomans, 1992), which is simply the characteristic length scale over which 'structures' are found. The CFs $Q_c(n)$, $Q_a(n)$ and $Q_s(n)$ are known to decay exponentially to 1/3 for many disordered stackings (Estevez-Rams *et al.*, 2003).[8] We therefore define the *correlation length*, λ_c, as the characteristic length over which correlation information is lost with increasing separation n. More precisely, let us define $\Psi_q(n)$ as

$$\Psi_q(n) = \sum_{\alpha} |Q_\alpha(n) - \tfrac{1}{3}|, \tag{4}$$

so that $\Psi_q(n)$ gives a measure of the deviation of the CFs from their asymptotic value. Then we say that

$$\Psi_q(n) \propto \text{'oscillating term'} \times 2^{-n/\lambda_c}. \tag{5}$$

For those cases where the CFs do not decay to 1/3, we say that the correlation length is infinite. We find that exponential decay is not always obeyed, but it seems to be common,[9] and the correlation length thus defined gives a useful measure of the rate of coherence loss as n increases. Our definition of correlation length is similar to the *characteristic length L* defined by Shrestha & Pandey (1996, 1997).

(ii) *Recurrence length*, $\mathcal{P}$. For an exactly periodic process, the period gives the length over which a template pattern repeats itself. We can generalize this for arbitrary, aperiodic processes in the following way. Let us take the *recurrence length* $\mathcal{P}$ as the geometric mean of the distances between visits to each CS weighted by the probability to visit that CS:

$$\mathcal{P} \equiv \prod_{\mathcal{S}_i \in \mathcal{S}} T_i^{p_i}, \tag{6}$$

[8] There are some exceptions to this. See Kabra & Pandey (1988), Yi & Canright (1996) and Varn (2001) for examples.

[9] The exponential decay of correlations is discussed by Crutchfield & Feldman (2003).

Table 7
The three characteristic lengths that one can calculate from knowledge of the ε-machine: the correlation length λ_c, the recurrence length $\mathcal{P}$ and the memory length r_ℓ.

For comparison, we also give these quantities for several common crystalline structures as well as a completely random process.

System	λ_c	$\mathcal{P}$	r_ℓ
Example A, $r = 3$	∼7.4	4.8	3
Example B, $r = 4$	∼7.8	7.3	4
Example C, Golden Mean	∼4.5	1.9	1
Example D, Even Process	∼1.7	1.9	∞
3C	∞	1	0
2H	∞	2	1
6H	∞	6	3
Completely random	1	1	0

where T_i is the average distance between visits to a CS and p_i is the probability of visiting that CS. Then,

$$\begin{aligned}\mathcal{P} &= \prod_{\mathcal{S}_i \in \mathcal{S}} (2^{\log_2 T_i})^{p_i} \\ &= \prod_{\mathcal{S}_i \in \mathcal{S}} 2^{-p_i \log_2 p_i} \\ &= 2^{-\sum_{\mathcal{S}_i \in \mathcal{S}} p_i \log_2 p_i} \\ &= 2^{C_\mu},\end{aligned} \tag{7}$$

where we have used the relation $T_i = 1/p_i$.

For periodic processes, $C_\mu = \log_2 \mathcal{P}$ and so $\mathcal{P}$ is simply a process's period. For aperiodic processes $\mathcal{P}$ gives a measure of the average distance over which the ε-machine returns to a CS. Notice that this is defined as the average recurrence length *in the Hägg notation*. For cubic and rhombohedral structures, for example, this is one-third of the physical repeat distance in the absolute stacking sequence.

(iii) *Memory length*, r_ℓ. Recall from §3.1 of Varn *et al.* (2013) that the *memory length* is an integer which specifies the maximum number of previous spins that one must know in the worst case to make an optimal prediction of the next spin. For an rth-order Markov process this is r.

(iv) *Interaction length*, r_I. The *interaction length* is an integer that gives the maximum range over which spin–spin interactions appear in the Hamiltonian.

We calculated the λ_c, $\mathcal{P}$ and r_ℓ (in units of MLs) for Examples A–D as well as for three crystal structures and a completely random stacking of MLs (that still obeys the stacking constraints, however). The results are displayed in Table 7. We see that each captures a different aspect of the system. The correlation length λ_c sets a scale over which a process is coherent. For crystals, as shown in Table 7, this length is infinite. For more disordered systems, this value decreases. The generalized period $\mathcal{P}$ is a measure of the scale over which the pattern produced by the process repeats. The memory length r_ℓ is most closely related to what we might think as the maximum range of 'influence' of a spin. That is, it is the maximum distance over which one might need to look to obtain information to predict a spin's value.

For periodic, infinitely correlated systems, spins at large separation carry information about each other, as seen in crystals. But this information is redundant. Outside a small neighborhood one gets no additional information by knowing the orientation a spin assumes. Notice that one can have an infinite memory length with a relatively small correlation length, as seen for the Even Process (Example D). That is, even though on *average* the knowledge one has about a spin may decay, there are still configurations in which distantly separated spins carry information about each other that is not stored in the intervening spins.

If we know the ε-machine for a process, then we can directly calculate λ_c, $\mathcal{P}$ and r_ℓ. How, then, do these relate to the interaction length r_I? Infinite correlation lengths can be achieved with very small r_I, as in the case of simple crystals. So correlation lengths alone imply little about the range of interactions. For a periodic system in the ground state, the configuration's period puts a lower bound on the interaction length *via* $r_I \geq \log_2 \mathcal{P}$, barring fine tuning of parameters, such as found at the multiphase boundaries in the ANNNI model (Yeomans, 1988) or those imposed by symmetry considerations (Canright & Watson, 1996; Yi & Canright, 1996; Varn & Canright, 2001). The most likely candidate for a useful relation between r_I and a quantity generated from the ε-machine is r_ℓ. Indeed, r_ℓ sets a lower bound on r_I, *if* the system is in equilibrium. For polytypes, the multitude of observed structures suggests that most are not in equilibrium but rather trapped in nonequilibrium metastable states. Consequently, one does not know what the relation between r_I and r_ℓ is. It is conceivable, especially in the midst of a solid-state phase transition, that small r_I could generate large r_ℓ (Varn & Crutchfield, 2004). While an ε-machine is a complete description of the underlying stacking process, one must additionally require that the material is in equilibrium in order to make inferences concerning r_I. This reflects the different ways in which a Hamiltonian and an ε-machine describe a material.

5. Conclusions

We demonstrated the feasibility and accuracy of ε-machine spectral reconstruction by applying it to four simulated DPs. In each case, we find that εMSR either reproduces the statistics of the stacking structure, as for Examples A and C, or finds a close approximation to it. Elsewhere, we applied the same procedures to the analysis of experimental DPs from single-crystal planar faulted ZnS, focusing on the novel physical and material properties that can be discovered with this technique (Varn *et al.*, 2007).

It is worthwhile to return one final time to some of the important features of εMSR. (i) εMSR makes no assumptions about either the crystal or faulting structures that may be present. Instead, using correlation information as input, εMSR constructs a model of the stacking structure – in the form of an ε-machine – that reproduces the observed correlations. Therefore, the algorithm need not rely on the experience or ingenuity of the researcher to make *a priori* postulates about crystal or fault structure. (ii) As the analysis of Example A shows, εMSR is able to detect and describe stacking structures that contain multiple crystal and fault structures. Indeed, Example A represented a specimen that was predominantly 2H, but also had significant portions of 3C crystal structure. Additionally, two faulting structures, growth and deformation faults, were identified. (iii) Since εMSR does not need to assume any underlying crystal structure, it can detect and describe even highly disordered structures. Example C has significant disorder (h_μ = 0.67 bits/ML) and does not contain any readily identifiable crystal structure. Nevertheless, εMSR is capable of finding and describing the statistics of even such highly disordered stacking structures. (iv) εMSR uses all of the information available in a DP. By integrating the DP over a unit interval in reciprocal space to find the CFs, εMSR makes no distinction between broadband scattering and Bragg-like peaks. Each is treated equally. Indeed, even though Example B shows both Bragg-like peaks as well as considerable broadband scattering between peaks, εMSR naturally captures the information contained in both by integrating over the entire DP. (v) It is advantageous not to invoke a more complicated explanation than is necessary to understand experimental data. By initially assuming a small memory length and incrementing this as needed to improve agreement between theory and experiment, as well as merging stacking 'histories' with equivalent 'futures', εMSR builds the smallest possible model that reproduces the experimentally observed DP without over-fitting the data. Example C shows how εMSR is able to find this minimal expression for the stacking structure. (vi) Finally, the resulting expression of the stacking structure, the process's ε-machine, allows for the calculation of parameters of physical interest. For each example, we were able to find the configurational entropy associated with the stacking process and the statistical complexity of the stacking structure. In a companion paper (Varn *et al.*, 2007), we show how the average stacking (fault) energy and hexagonality may be calculated from the ε-machine.

Additionally, we have identified three length parameters that are calculable from the ε-machine: the correlation length, λ_c; the recurrence length, $\mathcal{P}$; and the memory length, r_ℓ. Each measures a different length scale over which structural organization appears. New to this work is $\mathcal{P}$, which is a generalization of the period of a periodic process. $\mathcal{P}$ is a measure of the average length between visits to each CS. As such it quantifies the average distance over which the pattern repeats itself. Thus, both periodic and aperiodic patterns have a characteristic length scale after which they begin to repeat. The last length parameter we identified is r_ℓ, the distance over which an ML can carry nonredundant information about the orientation of another ML. This is most closely related to the r_I. If the assumption of equilibrium can be made for polytypes, r_ℓ places a lower bound on r_I. But the assumption of equilibrium is critical and not likely met by many polytypes.

Even with these advantages, however, εMSR as practiced here is not without its shortcomings. Perhaps most restrictive is that εMSR is limited to Markov processes and has only been worked out for third-order Markov processes. Since the

maximum number of terms in the SEs grows as the exponential of an exponential in the memory length, the task of writing out the higher-order SEs quickly becomes prohibitively difficult. While the $r = 4$ case is almost certainly tractable, the $r = 5$ is questionable and $r \geq 6$ is probably not (with current methods). Although $r = 3$ ε-machines identify much of the structure in higher-order processes, we found two difficulties. (i) Approximations made in the derivation of the SEs can result in sequence probabilities that differ from those of the true process, as seen in Example B. (ii) The state-labeling scheme imposes a CS architecture on the reconstructed ε-machine that may be too restrictive. The ε-machine in Example D belonged to a class of processes, formally known as sofic processes, that have a special kind of infinite-range memory. The CSs on the ε-machines that describe these processes cannot be specified by any finite history. So, the scheme of labeling states by the last r spins seen, as is done here, represents a serious drawback to εMSR.

Lastly, we note the εMSR can help give a detailed account of how a crystal becomes disordered. In order to determine the mechanism of faulting in, say, an annealed crystal undergoing a solid-state phase transition, it is desirable to begin with many (identical) crystals and arrest the solid-state transformation at various stages. By reconstructing the ε-machine after different annealing times, the route to disorder can be made plain. The result is a picture of how structure (as captured by intermediate ε-machines) changes during annealing. This change in structure should give direct insight into the structure-forming mechanisms. This should be compared with the numerical simulation of faulting in a crystal (Kabra & Pandey, 1988; Engel, 1990; Shrestha & Pandey, 1996, 1997; Gosk, 2000, 2001, 2003; Varn & Crutchfield, 2004). We note that, in such simulations, the ε-machine can be directly calculated from the stacking sequence to high accuracy. Some experimental work on solid-state phase transitions has been done (Sebastian & Krishna, 1994; Boulle *et al.*, 2010; Dompoint *et al.*, 2012), but we hope that this improved theoretical framework will stimulate additional efforts in this direction.

We thank L. J. Biven, D. P. Feldman and E. Smith for helpful conversations; and P. M. Riechers and three anonymous referees for comments that improved 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, DARPA Agreement F30602-00-2-0583 and ARO grant W911NF-12-1-0234. DPV's visit to SFI was partially supported by the NSF.

References

Badii, R. & Politi, A. (1997). *Complexity: Hierarchical Structures and Scaling in Physics*, Vol. 6 of *Cambridge Nonlinear Science Series.* Cambridge University Press.

Bataronov, I. L., Posmet'yev, V. V. & Barmin, Y. V. (2004). *Ferroelectrics*, **307**, 191–197.

Binney, J. J., Dowrick, N. J., Fisher, A. J. & Newman, M. E. J. (1992). *The Theory of Critical Phenomena.* Oxford: Clarendon Press.

Boulle, A., Dompoint, D., Galben-Sandulache, I. & Chaussende, D. (2010). *J. Appl. Cryst.* **43**, 867–875.

Brindley, G. W. (1980). In *Crystal Structures of Clay Minerals and their X-ray Identification*, edited by G. W. Brindley & G. Brown, ch. II. London: Mineralogical Society.

Canright, G. S. & Watson, G. (1996). *J. Stat. Phys.* **84**, 1095–1131.

Cartwright, J. H. & Mackay, A. L. (2012). *Philos. Trans. R. Soc. London Ser. A*, **370**, 2807–2822.

Cheng, C., Heine, V. & Jones, I. L. (1990). *J. Phys. Condens. Matter*, **2**, 5097–5113.

Cheng, C., Needs, R. J. & Heine, V. (1988). *J. Phys. C: Solid State Phys.* **21**, 1049–1063.

Cheng, C., Needs, R. J., Heine, V. & Churcher, N. (1987). *Europhys. Lett.* **3**, 475–479.

Crutchfield, J. P. (1992). *Santa Fe Studies in the Sciences of Complexity*, edited by M. Casdagli & S. Eubanks, Vol. XII, pp. 317–359. Reading: Addison-Wesley.

Crutchfield, J. P. (1994). *Physica D*, **75**, 11–54.

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. & Chaussende, D. (2012). *Nucl. Instrum. Methods Phys. Res. B*, **284**, 19–22.

Engel, G. E. (1990). *J. Phys. Condens. Matter*, **2**, 6905–6919.

Engel, G. E. & Needs, R. J. (1990). *J. Phys. Condens. Matter*, **2**, 367–376.

Erenburg, A., Gartstein, E. & Landau, M. (2005). *J. Phys. Chem. Solids*, **66**, 81–90.

Estevez-Rams, E., Aragon-Fernandez, B., Fuess, H. & Penton-Madrigal, A. (2003). *Phys. Rev. B*, **68**, 064111.

Ferraris, G., Makovicky, E. & Merlino, S. (2008). *Crystallography of Modular Materials*, Vol. 15. USA: Oxford University Press.

Gosk, J. B. (2000). *Cryst. Res. Technol.* **35**, 101–116.

Gosk, J. B. (2001). *Cryst. Res. Technol.* **36**, 197–213.

Gosk, J. B. (2003). *Cryst. Res. Technol.* **38**, 160–173.

Hansen, J. E. (1993). PhD thesis, University of California, Berkeley, USA.

Hopcroft, J. E. & Ullman, J. D. (1979). *Introduction to Automata Theory, Languages, and Computation.* Reading: Addison-Wesley.

Jagodzinski, H. (1949). *Acta Cryst.* **2**, 201–207.

Kabra, V. K. & Pandey, D. (1988). *Phys. Rev. Lett.* **61**, 1493–1496.

Mardix, S. (1986). *Phys. Rev. B*, **33**, 8677–8684.

Mélinon, P., Masenelli, B., Tournus, F. & Perez, A. (2007). *Nat. Mater.* **6**, 479–490.

Nasir, N., Shah, C., Leech, P., Reeves, G., Pirogova, E., Istivan, T., Tanner, P. & Holland, A. (2012). *Biomedical Engineering (ICoBE)*, 2012 International Conference, pp. 589–593. Piscataway: IEEE Conference Publication Operations.

Price, G. D. (1983). *Phys. Chem. Miner.* **10**, 77–83.

Sebastian, M. T. & Krishna, P. (1994). *Random, Non-Random and Periodic Faulting in Crystals.* New York: Gordon and Breach.

Shalizi, C. R. & Crutchfield, J. P. (2001). *J. Stat. Phys.* **104**, 817–881.

Shaw, J. J. A. & Heine, V. (1990). *J. Phys. Condens. Matter*, **2**, 4351–4361.

Shrestha, S. P. & Pandey, D. (1996). *Europhys. Lett.* **34**, 269–274.

Shrestha, S. P. & Pandey, D. (1997). *Proc. R. Soc. London Ser. A*, **453**, 1311–1330.

Thompson, J. B. (1981). In *Structure and Bonding in Crystals II*, edited by M. O'Keeffe & A. Navrotsky, ch. 22. New York: Academic Press.

Trigunayat, G. C. (1991). *Solid State Ionics*, **48**, 3–70.

Varn, D. P. (2001). 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**69**, 197–206.
Varn, D. P. & Crutchfield, J. P. (2004). *Phys. Lett. A*, **324**, 299–307.
Velterop, L., Delhez, R., de Keijser, Th. H., Mittemeijer, E. J. & Reefman, D. (2000). *J. Appl. Cryst.* **33**, 296–306.
Wang, Z., Wang, S., Zhang, C. & Li, J. (2011). *J. Nanopart. Res.* **13**, 185–191.
Yeomans, J. (1988). *Solid State Phys.* **41**, 151–200.
Yeomans, J. (1992). *Statistical Mechanics of Phase Transitions.* Oxford: Clarendon Press.
Yi, J. & Canright, G. S. (1996). *Phys. Rev. B*, **53**, 5198–5210.