

## Inferring planar disorder in close-packed structures via $\epsilon$ -machine spectral reconstruction theory: Examples from simulated diffraction spectra

D. P. Varn,<sup>a,b,c,\*†</sup> G. S. Canright<sup>c,d,\*‡</sup> and J. P. Crutchfield<sup>b,e,\*§</sup>

<sup>a</sup>Max-Planck-Institut für Physik komplexer Systeme, Nöthnitzer Straße 38, 01187 Dresden, Germany, <sup>b</sup>Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA, <sup>c</sup>Department of Physics and Astronomy, University of Tennessee, 1408 Circle Drive, Knoxville, Tennessee 37996, USA, <sup>d</sup>Telenor Research and Development, 1331 Fornebu, Oslo, Norway, and <sup>e</sup>Computational Science & Engineering Center & Physics Department, University of California, Davis, One Shields Avenue, Davis, California 95616, USA. Correspondence e-mail: dpvarn@pks.mpg.de, geoffrey.canright@telenor.com, chaos@cse.ucdavis.edu

Previously we detailed a novel algorithm,  $\epsilon$ -machine spectral reconstruction theory ( $\epsilon$ MSR), that infers pattern and disorder in planar-faulted, close-packed structures directly from X-ray diffraction spectra [Varn, Canright & Crutchfield, submitted to *Acta Crystallographica A*]. Here we apply  $\epsilon$ MSR to simulated diffraction spectra from five close-packed crystals. We find that for stacking structures with a memory length of three or less,  $\epsilon$ MSR reproduces the statistics of the stacking structure; the result being in the form of a directed graph called an  $\epsilon$ -machine. For stacking structures with a memory length larger than three,  $\epsilon$ MSR returns a model that captures many important features of the original stacking structure. These include multiple stacking faults and multiple crystal structures. Further, we find that  $\epsilon$ MSR 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, we define several length parameters calculable from the  $\epsilon$ -machine, and discuss their relevance.

**Keywords:** X-ray diffraction; diffuse scattering; one-dimensional disorder; polytypes; planar faults; computational mechanics.

### 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* also show considerable disorder and have been the subject of numerous theoretical and experimental investigations (Jagodziniski, 1949; Verma & Krishna, 1966; Pandey & Krishna, 1982; Trigunayat, 1991; Sebastian & Krishna, 1994). Polytypism is the phenomenon where a solid is built up by the stacking of identical layers, called *modular layers* (MLs) (Varn & Canright, 2001). Each ML is itself crystalline and the only possible 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 relative 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 150 known periodic stacking structures respectively. 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  $\sim 1$  ML in ZnS (Engel & Needs, 1990) and  $\sim 3$  MLs in SiC (Cheng *et al.*, 1987; Cheng *et al.*, 1988; Shaw & Heine, 1990; Cheng *et al.*, 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.

Significant simplifications in the analysis of X-ray diffraction spectra 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 inverting diffraction spectra 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,  $\epsilon$ -machine spectral reconstruction theory ( $\epsilon$ MSR or “emissary”), that does solve the problem of inferring planar disorder from

<sup>†</sup>Correspondence Author

<sup>‡</sup>Correspondence Author

<sup>§</sup>Correspondence Author

<sup>1</sup> We note that there are no inherent obstacles to applying  $\epsilon$ MSR to materials with more complicated MLs or stacking rules (Brindley, 1980; Thompson, 1981; Varn

diffraction spectra for the special case of close-packed structures (CPSs) (Varn *et al.*, 2002; Varn *et al.*, 2005a).<sup>1</sup> Although we do not find the particular stacking sequence that generated the experimental diffraction spectrum, we do find a unique, statistical expression for an ensemble of stacking sequences each of which could have produced the observed diffraction spectrum. This statistical description comes in the compact form of an  $\epsilon$ -machine (Crutchfield & Young, 1989; Shalizi & Crutchfield, 2001).

We claim in a companion paper (Varn *et al.*, 2005a) that  $\epsilon$ MSR has significant advantages over competing inference algorithms, particularly the fault model (FM).<sup>2</sup> These advantages include the following: (i)  $\epsilon$ MSR 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,  $\epsilon$ MSR can model crystals with multiple crystal or fault structures. (iii) Since  $\epsilon$ MSR doesn’t require a parent crystal structure, it can detect and quantify stacking structure in samples with even highly disordered stacking sequences. (iv)  $\epsilon$ MSR uses all of the information available from the diffraction spectrum, both Bragg and diffuse scattering. (v)  $\epsilon$ MSR results in a minimal and unique description of the stacking structure in the form of an  $\epsilon$ -machine. From knowledge of the  $\epsilon$ -machine, insight into the spacial 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  $\epsilon$ -machine.

Our purpose here is four-fold: (i) We wish to validate the above assertions concerning the efficacy of  $\epsilon$ MSR by demonstrating its application to the discovery of pattern and disorder in layered materials from their X-ray diffraction spectra. (ii) As developed in (Varn *et al.*, 2005a),  $\epsilon$ MSR can reconstruct processes up to 3<sup>rd</sup>-order Markovian. We wish to test the robustness of  $\epsilon$ MSR by analyzing diffraction spectra from stacking sequences not describable as 3<sup>rd</sup>-order Markovian. While we expect that  $\epsilon$ MSR 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  $\epsilon$ -machine and discuss their implication for finding long range order in polytypes. (iv) Lastly, we wish to demonstrate how the architecture of the  $\epsilon$ -machine provides an intuitive and quantitative understanding into the spacial organization of layered CPSs.

These goals are convincingly realized by analyzing diffraction spectra derived from simulated stacking sequences where there are no issues concerning experimental error. We are able to compare the  $\epsilon$ -machine reconstructed from spectral data with the  $\epsilon$ -machine that describes the original stacking structure, and thus we can explore how effectively  $\epsilon$ MSR 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  $\epsilon$ MSR.

Our development is organized as follows: in §2 we provide numerical details about the techniques we use to analyze the simulated diffraction spectra; in §3 we present our analysis of five simulated diffraction spectra using  $\epsilon$ MSR and contrast our results to those of the FM; in §4 we define several characteristic lengths calculable from a knowledge of the  $\epsilon$ -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  $\epsilon$ MSR to diffraction spectra obtained from single crystal X-ray diffraction experiments (Varn *et al.*, 2005b).

## 2. Methods

We use the same notational conventions and definitions introduced elsewhere (Varn *et al.*, 2005a). We examine five prototype processes in detail. Some of these processes are selected because they illustrate a particular feature of  $\epsilon$ MSR while others have a more physical motivation, *i.e.* they may represent real stacking structures in known polytypes. In Example A, we reconstruct an  $\epsilon$ -machine for a known memory length  $r_l = 3$  process and show that the technique works in this case and, indeed, for any process that has  $r_l \leq 3$ . Example A also nicely demonstrates how multiple crystal and fault structures can be simultaneously accommodated on the same  $\epsilon$ -machine. Examples B and E are selected because they may be similar to stacking structures in known polytypes. We also wish to test the effectiveness of  $\epsilon$ MSR on structures that require a memory length longer than  $r_l = 3$ . Example B needs  $r_l = 4$  and Examples D and E represent processes whose statistics can not be fully captured by any finite range process. This allows us to test  $\epsilon$ MSR on stacking structures we know that it can not completely detect and thus we can develop an intuition into the kinds of error one might expect. In order to illustrate the application of the equivalence relation, equation (11) of (Varn *et al.*, 2005a), to minimize the reconstructed  $\epsilon$ -machine (step 3c of Table 1 in (Varn *et al.*, 2005a)), we treat a process with  $r_l = 1$  in Example C. This shows that had we not terminated reconstruction at  $r = 1$ , the equivalence relation would require the merging of equivalent histories that would effectively find the  $r = 1$   $\epsilon$ -machine. Additionally, for each example we give a structural interpretation of the  $\epsilon$ -machine.

For each example we begin with a stacking structure as described by an  $\epsilon$ -machine. We generate a sample sequence from the  $\epsilon$ -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  $\approx 1\%$  of this value, which usually occurs for  $n \approx 25 - 100$ . We could, of course, find the CFs directly from spin-sequence probabilities, via equation (9) of (Varn *et al.*, 2005a).

---

& Canright, 2001). However the case of CPSs is by no means academic, since several important polytypes, such as SiC and ZnS, are describable as CPSs.

<sup>2</sup> We discuss the FM in detail in (Varn *et al.*, 2005a).

However, if one needed to calculate CFs for, say,  $n = 50$ , this would require finding the sequence probabilities for sequences of length 50. There are  $2^{50} \approx 10^{15}$  spin sequences for  $n = 50$ , so the sum implied by equation (9) of (Varn *et al.*, 2005a) is difficult to perform in practice. As an alternative, one changes representation and rewrites the  $\epsilon$ -machine in terms of the absolute stacking positions  $ABC$ . From this new representation the CFs are calculable from the *transition matrices*, equations (12) and (13) of (Varn *et al.*, 2005a). This has not been done here however. Additionally, it is possible to derive analytical expressions for the CFs in some cases (Varn, 2001).

We then calculate the *corrected diffracted intensity per ML*,  $l(l)$  (Varn *et al.*, 2005a), along the  $10.l$  row in increments of  $\Delta l = 0.001$  using equations (1) and (2) of (Varn *et al.*, 2005a) with a stacking sequence of 10 000 MLs. Throughout we refer to the corrected diffracted intensity per ML,  $l(l)$ , as simply the diffraction spectrum. We now now take this simulated diffraction spectrum as our “experimental” diffraction spectrum.

We apply  $\epsilon$ MSR (Table 1 of (Varn *et al.*, 2005a)) to each experimental diffraction spectrum. Since these are simulated diffraction spectra, we find that the figures-of-merit,  $\gamma$  and  $\beta$ , are equal to their theoretical values within numerical error over all unit  $l$  intervals. Therefore, we do not report  $\gamma$  and  $\beta$ , and instead perform  $\epsilon$ MSR over the interval  $0 \leq l \leq 1$ . Further, again since these are simulated spectra and hence have no error, we are not able to set an acceptable threshold error  $\Gamma$  in advance. Instead, each example, except for Example C, minimally requires the  $r = 3$  solutions. Thus we solve the *spectral equations* at  $r = 3$  (Appendix A.3 of (Varn *et al.*, 2005a)) via a Monte Carlo technique (Varn, 2001) to find sequence probabilities of length-4. We take the  $r = 3$   $\epsilon$ -machine given in Fig. 1 of (Varn *et al.*, 2005a) as our default or candidate  $\epsilon$ -machine. All *casual states* (CSs) and allowed transitions between CSs are initially assumed present. From the sequence probabilities we estimate the transition matrices,  $T_{S_i \rightarrow S_j}^{(s)}$ , for making a transition from a candidate CS  $S_i$  to a candidate CS  $S_j$  on seeing a spin  $s$ . We apply the equivalence relation, equation (11) of (Varn *et al.*, 2005a), to generate a final set of CSs. We refer to the resulting  $\epsilon$ -machine as the reconstructed or “theoretical”  $r = 3$   $\epsilon$ -machine for the spectrum. In the event that the reconstructed  $\epsilon$ -machine assigns to a CS an asymptotic state probability of less than 0.01, we take that CS to be nonexistent.

To find the predicted CFs for each  $\epsilon$ -machine, we again take a sample spin sequence generated by the  $\epsilon$ -machine of length 400 000 and find the CFs by directly scanning the resulting stacking sequence. The diffraction spectrum along the  $10.l$  row is again calculated from equations (1) and (2) of (Varn *et al.*, 2005a) using a sample of 10 000 MLs and we compare this with the diffraction spectrum for the original process.

We also calculate the information-theoretic quantities described in §2.6 of (Varn *et al.*, 2005a) for each example and the reconstructed  $\epsilon$ -machine.

### 3. Analysis

<sup>3</sup> Here and elsewhere we use the Ramsdell notation to specify crystalline stacking structures in CPSs. Recall that the  $\epsilon$ -machine gives stacking sequences in terms of the Hagg notation. For a discussion of nomenclature and faulting structure as revealed by the CS architecture of  $\epsilon$ -machines, see (Varn *et al.*, 2005a).

<sup>4</sup> We will denote a CSC by giving the sequence of CSs that compose the cycle in square brackets [].

#### 3.1. Example A

We begin with the sample process given in Fig. 1. This process can approximately be decomposed into FM structural components using equation (24) of (Varn *et al.*, 2005a) in the following way:

2H	54%
3C <sup>+</sup>	24%
Deformation fault	16%
Growth fault	6%

where the “+” on 3C indicates that only the positive chirality (...1111...) structure is present. The faulting is given with reference to the 2H crystal.<sup>3</sup> The diffraction spectrum from this process is shown in Fig. 2. The experienced crystallographer has little difficulty guessing the underlying crystal structure: the peaks at  $l \approx 0.50$  and at  $l \approx 1.00$  are indicative of the 2H structure; while the peak at  $l \approx 0.33$  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 diffraction spectrum of the 2H structure, the difference being 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  $l \approx 0.33$ , 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 clearly not the case here. In fact, the two peaks associated with the 2H structure at  $l \approx 0.50$  and 1.00 are shifted by  $\Delta l \approx 0.006$  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.

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 ( $\Delta l \approx 0.002$ ) for the  $l \approx 0.33$  peak and the broadening seems (arguably) symmetric, so one is tempted to guess that deformation faulting is important here. Indeed, the *causal state cycle* (CSC)  $[S_7 S_6 S_5 S_3]$ <sup>4</sup> 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  $\epsilon$ -machine description does better. The reconstructed  $\epsilon$ -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  $S_4$  to  $S_1$ ,  $T_{S_4 \rightarrow S_1}^{(1)} = 0.33$ , which was 1% too small. Not surprisingly,

the process shown in Fig. 1 is the reconstructed  $\epsilon$ -machine and so we do not repeat the Fig..

The two-layer CFs  $Q_s(n)$  versus  $n$  from the process and from the reconstructed  $\epsilon$ -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.*, 2005a) to compare the experimental spectrum (Example A) to the theoretical spectrum (reconstructed  $\epsilon$ -machine) and find a value of  $\mathcal{R} = 2\%$ . If we generate several spectra 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 spectra is shown in Fig. 2. This kind of agreement is typical of  $\epsilon$ MSR from any process that can be represented as a  $r = 3$   $\epsilon$ -machine (Varn, 2001).

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

Since the original process was represented as an  $r = 3$   $\epsilon$ -machine, this first example is largely a consistency check on  $\epsilon$ MSR. In the next example, we treat an  $r > 3$  process not represented by the  $r = 3$   $\epsilon$ -machines that we reconstruct.

### 3.2. Example B

Upon annealing, a solid-state transformation in ZnS from the 2H structure to either the 3C or 6H structures is possible, sometimes both occurring in different parts of the same crystal (Sebastian & Krishna, 1994). However, two crystal structures represented with an  $\epsilon$ -machine cannot share a CS, as discussed in §3.1.2 of (Varn *et al.*, 2005a). On an  $r = 3$   $\epsilon$ -machine, for example, both the CSCs associated with the 3C and the 6H structures share  $\mathcal{S}_7$  and  $\mathcal{S}_0$ , so a crystal containing both structures cannot be properly modeled at  $r = 3$ . In fact, it is necessary to use an  $r = 4$   $\epsilon$ -machine to encompass both structures. So, to see how well  $\epsilon$ MSR works at  $r = 3$  for an  $r = 4$  process, we consider the process shown in Fig. 4.  $[\mathcal{R}_1\mathcal{R}_3\mathcal{R}_7\mathcal{R}_{14}\mathcal{R}_{12}\mathcal{R}_8]$  would give rise to 6H structure if it were a strong CSC, but we find that the causal state cycle probability  $P_{CSC}(6H) = 0.25$  (Varn *et al.*, 2005a). We say then that this is mild 6H structure.  $[\mathcal{R}_0]$  and  $[\mathcal{R}_{15}]$  give the twinned 3C structures.

Employing spectral reconstruction, we find the  $r = 3$   $\epsilon$ -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  $\epsilon$ -machine is given in Fig. 6. The agreement is remarkably good. It seems that the  $r = 3$   $\epsilon$ -machine picks up most of the structure in the original process.

There is similar, though not as good, agreement in the diffraction spectra, as Fig. 7 shows. The most notable discrepancies

are in the small rises at  $l \approx 0.17$  and  $l \approx 0.83$ . We calculate a profile  $\mathcal{R}$ -factor of  $\mathcal{R} = 12\%$  between the diffraction spectra for Example B and the reconstructed  $\epsilon$ -machine. The  $r = 3$   $\epsilon$ -machine has difficulty reproducing the 6H structure in the presence of 3C structure, as expected.

Given the good agreement between the CFs and the spectra generated by Example B and the  $r = 3$   $\epsilon$ -machine, we are led to ask what the differences between the two are. In Table 1 we 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  $r = 3$   $\epsilon$ -machine allows with a small probability ( $\sim 0.03$ ). At the level of length-3 sequences, the  $\epsilon$ -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 2. There are more striking differences here. The frequencies of the two most common length-4 sequences in Example B,  $P(1111) = P(0000) = 0.227$ , are overestimated by the  $r = 3$   $\epsilon$ -machine, which assigns them a probability of  $\sim 0.30$  each. Similarly, sequences forbidden by Example B—1101, 1011, 1010, 1001, 0110, 0101, 0100, 0010—are not necessarily forbidden by the  $r = 3$   $\epsilon$ -machine. In fact, the  $r = 3$   $\epsilon$ -machine forbids only two of them, 0101 and 1010. This implies that  $r = 3$   $\epsilon$ -machine can find spurious sequences that are not in the original stacking sequence. This is to be expected. But the  $r = 3$   $\epsilon$ -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.)

One can also attempt to decompose the  $r = 3$   $\epsilon$ -machine into a sum of CSCs and interpret this as crystal and fault structure. However, as is typically the case, there is no unique decomposition and so therefore such an exercise is of questionable validity. With the exception of the sequences 1111 and 0000, the other twelve non-vanishing sequences all appear with a small, but rather constant probability in the range 0.024 - 0.052. One possible interpretation is to say that  $[\mathcal{S}_0]$  and  $[\mathcal{S}_7]$  contribute to 3C structure with a weight of 0.58. We could further interpret  $[\mathcal{S}_7\mathcal{S}_6\mathcal{S}_5\mathcal{S}_3]$  and  $[\mathcal{S}_0\mathcal{S}_1\mathcal{S}_2\mathcal{S}_4]$  as deformation faulting of the 3C structure giving a combined weight of 0.24. And finally, we could associate  $[\mathcal{S}_1\mathcal{S}_3\mathcal{S}_6\mathcal{S}_4]$  with 4H structure. This last interpretation of  $[\mathcal{S}_1\mathcal{S}_3\mathcal{S}_6\mathcal{S}_4]$  with any crystal structure is troublesome as the  $P_{CSC}([\mathcal{S}_1\mathcal{S}_3\mathcal{S}_6\mathcal{S}_4]) \ll 1$ . Another possible decomposition would be to again assign  $[\mathcal{S}_0]$  and  $[\mathcal{S}_7]$  to the 3C structure with a weight of 0.58, to interpret the paths  $\mathcal{S}_7\mathcal{S}_6\mathcal{S}_4\mathcal{S}_0$  and  $\mathcal{S}_0\mathcal{S}_1\mathcal{S}_3\mathcal{S}_7$  as twin faulting with a probability weight of 0.18, treat  $[\mathcal{S}_1\mathcal{S}_3\mathcal{S}_6\mathcal{S}_4]$  as 4H structure, and finally to interpret  $[\mathcal{S}_1\mathcal{S}_2\mathcal{S}_4]$  and  $[\mathcal{S}_3\mathcal{S}_6\mathcal{S}_5]$  as 9R structures. These two descriptions are clearly rather different and, arguably, have no use in any account, other than serving to illustrate the ambiguity of FM-like structural interpretations.

In addition to the non-uniqueness difficulties, by simply listing the probability density of the various crystals and fault

structures, we say nothing about how one crystal converts into another as one scans the stacking sequence. This exercise demonstrates the impoverished view of crystal structure inherent in the FM. In short, the stacking sequence implied by the  $\epsilon$ -machine in Fig. 5 comes from a physical structure that is not describable in terms of the FM.

We find by direct calculation that the Example B process has a configurational entropy of  $h_\mu = 0.51$  bits/spin, 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/spin, a statistical complexity of  $C_\mu = 2.44$  bits, and an excess entropy of  $\mathbf{E} = 0.83$  bits.

### 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  $\epsilon$ -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 to only 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  $\phi = (1 + \sqrt{5})/2$ .

We employ the  $\epsilon$ MSR algorithm and find the  $\epsilon$ -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 diffraction spectra, 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  $\epsilon$ MSR should terminate, as we have found satisfactory agreement (to within the numerical error of our technique) between “experiment”, Example C, and “theory”, the reconstructed  $\epsilon$ -machine.

Let us suppose that instead, we increment  $r$  and follow the  $\epsilon$ MSR algorithm as if the agreement at  $r = 1$  had been unsatisfactory. In this case, we would have generated the “ $\epsilon$ -machine” shown in Fig. 11 at the end of step 3b (Table 1 of (Varn *et al.*, 2005a)). We have yet to apply the equivalence relation equation (11) of (Varn *et al.*, 2005a) and so let us call this the *non-minimal*  $\epsilon$ -machine. That is, we have not yet combined pasts with equivalent futures to form CSs, step 3c (Table 1 of (Varn *et al.*, 2005a)). 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  $P(1|\mathcal{S}_1) = P(1|\mathcal{S}_3) = 1/2$  and  $P(0|\mathcal{S}_1) = P(0|\mathcal{S}_3) = 1/2$  and, thus, these states share the same probability of seeing futures of length-1. More formally, we can write

$$\mathbb{T}_{01 \rightarrow 1s}^{(s)} = \mathbb{T}_{11 \rightarrow 1s}^{(s)}. \quad (1)$$

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 thus  $\mathcal{S}_1$  and  $\mathcal{S}_3$  can be merged to form a single CS. The result is the  $\epsilon$ -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, because 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

$$\mathbb{T}_{11 \rightarrow 1s}^{(s)} = \mathbb{T}_{10 \rightarrow 0s}^{(s)} \quad (2)$$

and

$$\mathbb{T}_{1s \rightarrow ss'}^{(s')} = \mathbb{T}_{0s \rightarrow ss'}^{(s')} \quad (3)$$

for all  $s, s'$ .

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

### 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  $\epsilon$ MSR 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  $\epsilon$ -machine for this process is shown in Fig. 12. Its causal-state transition structure is equivalent to that in the  $\epsilon$ -machine for the golden mean process. They differ only in the *spins* emit-

ted 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  $\epsilon$ -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  $r = 3$   $\epsilon$ -machine. One finds that the sequence 010—forbidden by the process—is also forbidden by the reconstructed  $\epsilon$ -machine. This occurs because  $\mathcal{S}_2$  is missing.<sup>5</sup> We do notice that the reconstructed  $\epsilon$ -machine has much more “structure” than the original process. We now examine the source of this additional structure.

Let us first contrast differences between  $\epsilon$ MSR and other  $\epsilon$ -machine reconstruction techniques, taking the subtree-merging method (SMM) of Crutchfield and Young (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 diffraction spectra and not a symbol sequence, we find it necessary to invoke the memory-length reduction approximation (Varn *et al.*, 2005a) 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 3. The agreement is excellent. All sequence frequencies are within  $\pm 0.01$  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  $\epsilon$ -machines generated from each method given the same “exact” or “correct” length-4 sequence probabilities. Doing so, SMM gives the  $\epsilon$ -machine for the even process shown in Fig. 12.  $\epsilon$ MSR gives a different result. After merging pasts with equivalent futures, one finds the  $\epsilon$ -machine shown in Fig. 15. For clarity, we explicitly show the length-3 sequence histories associated with each CS, but do not write out the asymptotic state probabilities.

The  $\epsilon$ -machine generated by  $\epsilon$ MSR 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  $\epsilon$ MSR, 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 (0.92 bits as compared to 1.92 bits), can be found. So the notion of minimality is violated. That is,  $\epsilon$ MSR 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  $\epsilon$ MSR 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 spectral equations, we would have found (within some error) the

$\epsilon$ -machine given in Fig. 12.

We find, then, that there are two separate consequences to applying  $\epsilon$ MSR that affect the reconstructed  $\epsilon$ -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 (non-sofic) 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  $\mathcal{S}_{\text{even}}$ . Similarly, the histories 001 and 101 always leave the process in CS  $\mathcal{S}_{\text{odd}}$ . 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.

Then why do we not find sequence probabilities by solving the spectral equations and then use SMM to reconstruct the  $\epsilon$ -machine? There are two reasons. The first is that in general one must know sequence probabilities for longer sequences than is necessary for  $\epsilon$ MSR. Solving the spectral equations for these longer sequence frequencies is onerous. The second is that error in the sequence probabilities found from solving the spectral equations 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  $\epsilon$ MSR and SSM, we now return to a comparison between CFs and diffraction spectrum generated by the  $\epsilon$ MSR and the even process. The CFs for the even process and the reconstructed  $\epsilon$ -machine are given in Fig. 14. We see that both decay quite quickly to their asymptotic values of  $1/3$ . There is good agreement, except in the region between  $5 \leq n \leq 10$ . Examining the diffraction spectra in Fig. 16, we see that there is likewise good agreement except in the region  $0.7 < l < 0.9$ . We calculate the profile  $\mathcal{R}$ -factor between the theoretical and experimental spectra to be  $\mathcal{R} = 8\%$ .

There is a curious isolated zero in the process’s spectrum at  $l \approx 0.83$ . The other interesting feature is the broad peak at  $l \approx 0.33$ . One might guess that this originates from some  $3C^+$  structure and, indeed, glancing at the reconstructed  $\epsilon$ -machine of Fig. 13 shows that  $[\mathcal{S}_7]$  is strongly represented. The faulting is less clear. We would expect, though, that the presence of  $[\mathcal{S}_7\mathcal{S}_6\mathcal{S}_4\mathcal{S}_0\mathcal{S}_1\mathcal{S}_3]$  would indicate layer-displacement faulting of the  $3C^+$  structure and  $[\mathcal{S}_7\mathcal{S}_6\mathcal{S}_5\mathcal{S}_3]$  is characteristic of deformation faulting of the  $3C^+$  structure. But given that most non-vanishing transitions between CSs occur with a probability near  $\sim 0.5$ , such an identification is questionable.

We find by direct calculation from the even process that it has a configurational entropy of  $h_\mu = 0.67$  bits/spin, a statis-

<sup>5</sup> We do note that the solution of the spectral equations at  $r = 3$  assigns the sequences 0100 and 0010 a small probability,  $P(0100) \approx P(0010) \approx 0.005$ , which implies that the sequence 010 is also present with a small probability,  $P(010) < 0.01$ . 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 spectral equations at  $r = 3$  are difficult to satisfy with purely real probabilities. We also note that the solution of the spectral equations at  $r = 2$  *does* forbid the 010 sequence. For additional discussion, see (Varn, 2001).

tical complexity of  $C_\mu = 0.92$  bits, and an excess entropy of  $\mathbf{E} = 0.91$  bits. The reconstructed  $\epsilon$ -machine gives information-theoretic quantities that are rather different. We find a configurational entropy  $h_\mu = 0.79$  bits/spin, a statistical complexity of  $C_\mu = 2.58$  bits, and an excess entropy of  $\mathbf{E} = 0.21$  bits. Thus we find the reconstructed  $\epsilon$ -machine is *more complex* than the original process, ( $C_\mu(\text{theory}) = 2.58$  bits as compared to  $C_\mu(\text{experiment}) = 0.92$  bits) but *less predictive* ( $h_\mu(\text{theory}) = 0.79$  bits/spin as compared to  $h_\mu(\text{experiment}) = 0.67$  bits/spin).

One reason that the reconstructed  $\epsilon$ -machine gives CFs and diffraction spectra 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 diffraction spectra to the frequencies of individual long sequences: equation (9) of (Varn *et al.*, 2005a) 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.  $\epsilon$ MSR 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  $\epsilon$ -machine.

### 3.5. Example E

ZnS is believed to have only two stable phases, the high-temperature phase, 2H and the low-temperature phase, 3C. Crystals can be grown at high temperatures (above 1024 C) in the 2H phase and then cooled to a temperature range where the 3C phase becomes stable. The crystal then transforms enantiotropically from the former into the latter predominantly via deformation faulting (Roth, 1960; Sebastian & Krishna, 1984; Sebastian, 1988). This transformation can be arrested at any point by cooling the crystal further to a temperature range where the MLs lack the necessary thermal activation energy to slip. Thus it is possible to experimentally study partially transformed crystals.

This martensitic transformation can be modeled in a straightforward fashion (Varn & Crutchfield, 2004). We note that an undefected 2H crystal can be represented by the antiferromagnetic phase of a linear chain of Ising spins and a 3C crystal is just the ferromagnetic phase. Let us make four assumptions: (i) Deformation faulting is the primary mode of transformation. In terms of spins, this corresponds to flipping a single spin, *i.e.* Glauber dynamics (Glauber, 1963). (ii) Only interactions between neighboring spins are important. (iii) A spin can flip only if it is energetically favorable to do so. (iv) The transformation happens slowly. Putting this all together, let us begin with an antiferromagnetic chain. We visit a spin randomly (but never more than once) and flip this spin only if it is antiparallel to *both* of its neighbors. We call the fraction of spins so visited the *faulting parameter*  $f$ . Due to its formal similarity to elementary cellular automaton rule 232, except that here the update rule is applied asynchronously to only a fraction of spins, this model is called ACA 232. While much simpler than other models of solid-state transformations (Kabra & Pandey, 1988; Engel, 1990; Shrestha & Pandey, 1996a; Shrestha *et al.*, 1996; Shrestha

& Pandey, 1996b), ACA 232 nonetheless reproduces many of the significant features seen in experimental diffraction spectra of annealed ZnS crystals.

Real transformations in ZnS crystals are undoubtedly much more complex than this. However, despite its simplicity, the  $\epsilon$ -machine that describes the stacking process for a crystal transformed under ACA 232 has an infinite memory length, *i.e.* it is sofic. The physical origin of this soficity is not difficult to understand. Note that the original unfaulted crystal has only odd spin domains. (Indeed each spin domain in the unfaulted 2H crystal is exactly one spin long.) A spin flip (deformation fault) has the effect of joining two such odd spin domains by flipping the single spin that separates them. Thus the resulting larger spin domain must also have an odd number of spins. It follows then that a perfect 2H crystal undergoing this transform can never have even spin domains. Just as for the even system, Example D, one must remember the oddness (evenness) of the previous like spins scanned to determine the admissibility of the next spin. So in general the description of this process requires one to remember an indefinitely long history of spins. An important consequence of soficity is that no finite-order Markov process can fully reproduce the statistics. Thus it is reasonable to ask how much of the stacking structure  $\epsilon$ MSR can capture.

We consider a partially transformed crystal with a faulting parameter  $f = 0.10$ . For a crystal only weakly faulted by the ACA 232 process, as is the case here, the  $\epsilon$ -machine shown in Fig. 17 gives an excellent representation of the stacking structure and we take this to be our experimental  $\epsilon$ -machine. The concomitant diffraction spectrum is shown in Fig. 18. From the Bragg-like reflections at  $l \approx 0.50$  and  $l \approx 1.00$ , it is clear that the structure of this crystal is predominantly 2H.

Since the faulting is weak, we are able to perform a FM analysis. We find that the Bragg peaks are broadened symmetrically and any shifting in their placement is negligible. We further find that the FWHM is 0.059 for the integer- $l$  peaks and 0.058 for the half-integer- $l$  peaks. All of this is consistent with deformation faulting of the 2H structure.

Employing spectral reconstruction, we find the  $r = 3$   $\epsilon$ -machine shown in Fig. 19. We notice that the CS architecture between the two  $\epsilon$ -machines *appears* to be rather different. We compare the theoretical and experimental CFs in Fig. 20. The agreement is excellent. There is, however, some discrepancy in the range  $10 \leq n \leq 30$ , where the theoretical CFs have slightly stronger oscillations. Similarly, we compare the diffraction spectra in Fig. 18. Here we also find excellent agreement as evidenced by the  $\mathcal{R}$ -factor between the two spectra of  $\mathcal{R} = 8\%$ . Given such good agreement between the theoretical and experimental diffraction spectra and CFs, we are led to ask how this is possible when their respective  $\epsilon$ -machines seem to be so different. We find however, the differences are indeed more apparent than real.

Let us follow the same kind of analysis as we performed earlier (Varn & Crutchfield, 2004), where one begins in one of the CSs that is part of the 2H crystal structure and then follows a path of CSs associated with faulting. First we note that the both  $\epsilon$ -machines have CSCs that generate the 2H stacking structure:

[AB] in the experimental  $\epsilon$ -machine and  $[\mathcal{S}_2\mathcal{S}_5]$  for the theoretical one. They even have nearly the same CS transition probabilities connecting them:  $T_{A \rightarrow B}^{(0)} = 0.90 \approx T_{\mathcal{S}_5 \rightarrow \mathcal{S}_2}^{(0)} = 0.92$  and  $T_{B \rightarrow A}^{(1)} = 0.90 \approx T_{\mathcal{S}_2 \rightarrow \mathcal{S}_5}^{(1)} = 0.88$ . Thus these two CSCs perform equivalent functions on their respective  $\epsilon$ -machines. For small faulting as is the case here, the remainder of the CSs on each  $\epsilon$ -machine describe deviations from this crystal structure. As noted elsewhere (Varn & Crutchfield, 2004), the three spin sequence 100 necessarily places the experimental  $\epsilon$ -machine in D. Thus  $\mathcal{S}_4$  in the theoretical  $\epsilon$ -machine (which by definition assumes the three spin history of 100) is analogous to D in the experimental  $\epsilon$ -machine (at least for length-3 spin histories). We find that transitions out of these two CSs are identical:  $P(0|D) = P(0|\mathcal{S}_4) = 1$  and  $P(1|D) = P(1|\mathcal{S}_4) = 0$ . This demonstrates that the theoretical  $\epsilon$ -machine also prohibits the 1001 stacking sequence just as the experimental  $\epsilon$ -machine does. After the sequence history 1000 the experimental  $\epsilon$ -machine is in F and the theoretical  $\epsilon$ -machine is in  $\mathcal{S}_0$ . We find that transitions out of these two CSs are equal (to within the numerical accuracy of solving the spectral equations):  $P(0|F) = 0.096 \approx P(0|\mathcal{S}_0) = 0.10$  and  $P(1|F) = 0.904 \approx P(1|\mathcal{S}_0) = 0.90$ . However, the destination CSs after these latter transitions do not appear to be analogous on the two  $\epsilon$ -machines. A 1 on the experimental  $\epsilon$ -machine returns the  $\epsilon$ -machine to A, *i.e.* it has now returned to [AB] or the 2H structure. The theoretical  $\epsilon$ -machine however advances to  $\mathcal{S}_1$ , rather than  $\mathcal{S}_5$ , the CS analogous to A on the theoretical  $\epsilon$ -machine. The transition probabilities for the next spin are a little different for the two  $\epsilon$ -machines:  $P(0|A) = 0.90 \neq P(0|\mathcal{S}_1) = 1.00$ . But if we do follow this transition on 0, we will find each  $\epsilon$ -machine back into [AB] or  $[\mathcal{S}_2\mathcal{S}_5]$  associated with 2H structure. We find then that for [BDF A] on the experimental  $\epsilon$ -machine we have an analogous CSC,  $[\mathcal{S}_2\mathcal{S}_4\mathcal{S}_0\mathcal{S}_1]$ , on the theoretical  $\epsilon$ -machine, *if* we allow  $\mathcal{S}_1$  to play a similar role to  $\mathcal{S}_5$ . In fact,  $\mathcal{S}_1$  and  $\mathcal{S}_5$  have nearly identical futures. Each transitions to  $\mathcal{S}_2$  on a 0, and had the spectral equations not found a vanishing probability for the sequence 0011,  $\mathcal{S}_1$  would transition to  $\mathcal{S}_3$  on a 1, just as  $\mathcal{S}_5$  does. Indeed, had the conditional probabilities out of  $\mathcal{S}_1$  and  $\mathcal{S}_5$  been equal, the equivalence relation, equation (11) of (Varn *et al.*, 2005a), would have required the merger of these two CSs to form a single CS. The spin sequence associated with [BDF A] and  $[\mathcal{S}_2\mathcal{S}_4\mathcal{S}_0\mathcal{S}_1]$  is just 0100010, where the first three spins can be inferred as necessary to fix each  $\epsilon$ -machine into B or  $\mathcal{S}_2$ . The interpretation is clear: these two CSCs represent a single, isolated deformation fault of the 2H structure. This exercise strengthens the interpretation of the CS structure of weak deformation faulting on an  $r = 3$   $\epsilon$ -machine in a 2H crystal given in §3.2.2 of (Varn *et al.*, 2005a).

We can further demonstrate the similarity between the experimental and theoretical  $\epsilon$ -machines with the following exercise. Since the  $\mathcal{S}_1$  and  $\mathcal{S}_5$  do have nearly identical futures, let us merge them in to a single CS, and call it  $\mathcal{S}_1\mathcal{S}_5$ . Similarly, let us also merge  $\mathcal{S}_2$  and  $\mathcal{S}_6$  and label the resulting CS  $\mathcal{S}_2\mathcal{S}_6$ . To find the transition probabilities for these new states, we just take a weighted average of the transition probabilities for the old states. Further, we can rearrange the CSs on the theoretical  $\epsilon$ -machine so that the CSs occupy the same position as their

analogous states on the experimental  $\epsilon$ -machine. We call this the “reduced” theoretical  $\epsilon$ -machine and it is shown in Fig. 21. The similarity between the reduced  $\epsilon$ -machine and the experimental one, Fig. 17, is striking. The CS architectures are nearly identical, the only difference being that the  $\mathcal{S}_0$  and  $\mathcal{S}_7$  on the reduced theoretical  $\epsilon$ -machine have a self-state transition on a 0 and 1 respectively, whereas on the experimental  $\epsilon$ -machine F and E transition to different CSs on a 0 and 1 respectively. Further, the transition probabilities between CSs and the asymptotic CS probabilities are nearly identical.

Given that the theoretical  $\epsilon$ -machine and the experimental one are indeed so similar, we can ask why  $\epsilon$ MSR didn’t find the experimental  $\epsilon$ -machine. As with Example D, we can trace the reasons to two difficulties: (i) errors in sequence probabilities as found by solving the spectral equations, and (ii) the state labeling scheme. We compare the probabilities for length-4 sequences in Table 4. The spectral equations reproduce the sequence probabilities from the experimental  $\epsilon$ -machine reasonably well. For sequences appearing only rarely, however, there are some relatively large deviations. Notably, the sequences 1100 and 0011 each occur with a frequency of 0.004 in the experimental  $\epsilon$ -machine, but the theoretical  $\epsilon$ -machine assigns them probabilities of 0.014 and 0.000 respectively. This, along with the error in the probabilities for the 1101 and 0010 sequences, gives transition probabilities out of  $\mathcal{S}_1$  and  $\mathcal{S}_6$  that prevent these CSs from being merged with  $\mathcal{S}_5$  and  $\mathcal{S}_2$ , respectively. Thus the theoretical  $\epsilon$ -machine makes distinctions about pasts that the experimental one does not. The second difficulty lies with the state labeling scheme. Since each state is initially labeled by the last three spins seen,  $\mathcal{S}_0$  and  $\mathcal{S}_7$  necessarily have self-state transitions. So the kind of CS architecture on the experimental  $\epsilon$ -machine that generates the infinite range memory—that bouncing between CSs, such as that between F and D that prohibits even spin domains while allowing odd spin domains of any size—can never be realized if states are labeled by finite histories.

Returning our attention to the theoretical  $\epsilon$ -machine in Fig. 19, we examine how its CS architecture reveals information about the stacking structure. As previously noted, the large asymptotic state probabilities for  $\mathcal{S}_2$  and  $\mathcal{S}_5$ ,  $\Pr(\mathcal{S}_2) = \Pr(\mathcal{S}_5) = 0.37$ , as well as their large casual state cycle probability,  $P_{CSC}([\mathcal{S}_2\mathcal{S}_5]) = 0.81$ , indicate that this crystal is predominantly 2H. The remaining CSCs give the faulting structure. Since there is no CS transition from either  $\mathcal{S}_3$  to  $\mathcal{S}_6$  or from  $\mathcal{S}_1$  to  $\mathcal{S}_4$ , we see that stacking structure associated with both growth and layer displacement faults is absent. Further, the self-state transition probabilities for  $\mathcal{S}_0$  and  $\mathcal{S}_7$  are likewise small, so we conclude that there are no large regions where the crystal has transformed to the 3C structure.

This  $\epsilon$ -machine can be approximately broken down into FM structural components using equation (24) of (Varn *et al.*, 2005a) and we find:

2H	66%
Deformation fault	31%
Other	3%.

This then is consistent with 2H crystal that has been weakly

deformation faulted. We also note that *fault probability* (Varn *et al.*, 2005a), *i.e.* the probability that upon scanning the crystal one finds a particular kind of fault, can also be approximated directly from the theoretical  $\epsilon$ -machine. For this weakly faulted crystal we take  $[\mathcal{S}_2\mathcal{S}_5]$  as the parent crystal structure. The probability of leaving  $[\mathcal{S}_2\mathcal{S}_5]$  averaged over the CSC is just  $(1/2)\{P(1|\mathcal{S}_5) + P(0|\mathcal{S}_2)\} = (1/2)\{0.12 + 0.08\} = 0.10$ . This is the quantity usually reported in the literature. Since there is but a single parent structure with one kind of fault, finding the fault probability here is unambiguous. For multiple crystalline and fault structures, this kind of simple analysis may not be possible.

We find by direct calculation from ACA 232 that it has a configurational entropy of  $h_\mu = 0.42$  bits/spin, a statistical complexity of  $C_\mu = 1.86$  bits, and an excess entropy of  $\mathbf{E} = 1.01$  bits. The reconstructed  $\epsilon$ -machine gives similar information-theoretic quantities. We find a configurational entropy  $h_\mu = 0.42$  bits/spin, a statistical complexity of  $C_\mu = 2.26$  bits, and an excess entropy of  $\mathbf{E} = 0.99$  bits. Thus we find the reconstructed  $\epsilon$ -machine is *more complex* than the original process, ( $C_\mu(\text{theory}) = 2.26$  bits as compared to  $C_\mu(\text{experiment}) = 1.86$  bits) but equally predictive, ( $h_\mu(\text{theory}) = 0.42$  bits/spin as compared to  $h_\mu(\text{experiment}) = 0.42$  bits/spin).

For comparison we list each example's information-theoretic properties in Table 5.

### 3.6. Anticipated Difficulties with Applying $\epsilon$ MSR

We have considered five examples that demonstrate successful applications of  $\epsilon$ MSR. We have found instances, however, when the  $\epsilon$ MSR has difficulties converging to a satisfactory result. We now analyze each step in  $\epsilon$ MSR as given in Table 1 of (Varn *et al.*, 2005a) and discuss possible problems that may be encountered.

*Step 1.* Several problems can arise here. One is that the figures-of-merit,  $\beta$  and  $\gamma$ , are sufficiently different from their theoretical values over all possible  $l$ -intervals that  $\epsilon$ MSR should not even be attempted. Even if one does find an interval such that they indicate satisfactory spectral data, 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  $\Gamma$ .

*Step 2.* The  $P(\omega^r)$  solutions to the spectral equations 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 spectrum.

*Step 3.* Given  $P(\omega^r)$  that satisfy the elementary conditions of probability (*i.e.*, there is no difficulty at step 2), step 3 will return a machine that generates  $P(\omega^r)$ . It is possible, however, that the resulting CSs are not *strongly connected*, and thus the result may not be interpreted as a single  $\epsilon$ -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 diffraction spectrum, (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.*, 2005a)) are not met, then since equation (1) of (Varn *et al.*, 2005a) is no longer valid, the CFs found by Fourier analysis will not reflect the actual stacking probabilities. This will likely be interpreted as poor figures-of-merit, and  $\epsilon$ MSR will terminate at step 1.

Of the four possible difficulties only (ii) and (iii) should be considered to be inherent to  $\epsilon$ MSR. It is satisfying that  $\epsilon$ MSR can detect errors in the diffraction spectrum 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  $\epsilon$ -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  $\epsilon$ -machines.

(i) *Correlation Length*,  $\lambda_c$ . From statistical mechanics, we have the notion of a correlation length, (Binney *et al.*, 1993; Yeomans, 1992) which is simply the characteristic length scale over which "structures" are found. The correlation functions  $Q_c(n)$ ,  $Q_a(n)$ , and  $Q_s(n)$  are known to decay to 1/3 for many disordered stackings.<sup>6</sup> For the disordered cases considered here, exponential decay to 1/3 seems to be the rule. We therefore define the *correlation length*  $\lambda_c$  as the characteristic length over which correlation information is lost with increasing separation  $n$ . More precisely, let us define  $\Psi_q(n)$  as

$$\Psi_q(n) = \sum_\alpha \left| Q_\alpha(n) - \frac{1}{3} \right|, \quad (4)$$

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

$$\Psi_q(n) = F(n) \times 2^{-n/\lambda_c}, \quad (5)$$

where  $F(n)$  is some function of  $n$ .

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,<sup>7</sup> and

<sup>6</sup> There are some exceptions to this. See Kabra and Pandey (1988), Yi and Canright (1996), and Varn (2001) for some examples.

<sup>7</sup> The exponential decay of correlations is discussed by Crutchfield & Feldman (2003).

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 and Pandey (Shrestha & Pandey, 1996a; Shrestha & Pandey, 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:

$$\mathcal{P} \equiv \prod_{S_i \in \mathcal{S}} T_i^{p_i}, \quad (6)$$

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

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

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  $\epsilon$ -machine returns to a CS. Notice that this is defined as the average recurrence length *in the Hagg 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 §3.7 of (Varn *et al.*, 2005a) 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  $r^{\text{th}}$ -order Markov process this is  $r$ .

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

We calculated the  $\lambda_c$ ,  $\mathcal{P}$ , and  $r_l$  (in units of MLs) for Examples A-E as well as for three crystal structures. The results are displayed in Table 6. We see that each captures a different aspect of the system. The correlation length  $\lambda_c$  sets a scale over which a process is coherent. For crystals, as shown in Table 6, 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 system (Example D) and ACA 232 (Example E). 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  $\epsilon$ -machine for a process, then we can directly calculate  $\lambda_c$ ,  $\mathcal{P}$ , and  $r_l$ . How, then, do these relate to the interaction length  $r_l$ ? Infinite correlation lengths can be achieved with very small  $r_l$ , 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  $r_l \geq \log_2 \mathcal{P}$ —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  $r_l$  and a quantity generated from the  $\epsilon$ -machine is  $r_l$ . Indeed  $r_l$  sets a lower bound on  $r_l$ , *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, and, consequently one does not know what the relation between  $r_l$  and  $r_l$  is. It is conceivable, especially in the midst of a solid-state phase transition, that small  $r_l$  could generate large  $r_l$  (Varn & Crutchfield, 2004). While an  $\epsilon$ -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  $r_l$ . This reflects the different ways in which a Hamiltonian and an  $\epsilon$ -machine describe a material.

## 5. Conclusions

We have demonstrated the feasibility and accuracy of  $\epsilon$ -machine spectral reconstruction by applying it to five simulated diffraction spectra. In each case, we find that  $\epsilon$ MSR either reproduces the statistics of the stacking structure, as for Examples A and C, or finds a close approximation to it. Elsewhere we apply the same procedures to the analysis of experimental diffraction spectra from single-crystal planar faulted ZnS, focusing on the novel physical and material properties that can be discovered with this technique (Varn *et al.*, 2005b).

It is worthwhile to return one final time to how  $\epsilon$ MSR differs from other spectral inference algorithms—particularly the FM—and discuss how  $\epsilon$ MSR gives an improved framework in which to discover and understand disorder and structure in planar faulted crystals. (i)  $\epsilon$ MSR makes no assumptions about either the crystal or faulting structures that may be present. Instead, using correlation information as input,  $\epsilon$ MSR constructs a model of the stacking structure—in the form of an  $\epsilon$ -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,  $\epsilon$ MSR is able to detect and describe stacking structures that contain multiple crystal and fault structures. Indeed, Example A represented a crystal that was predominately 2H, but also had significant portions of 3C crystal structure. Additionally, two faulting structures, growth and deformation faults, were identified. (iii) Since  $\epsilon$ MSR doesn't need to assume any underlying crystal structure, it can detect and describe even highly disordered structures. Example C has significant disorder ( $h_\mu = 0.67$  bits/ML<sup>8</sup>) and doesn't contain any readily identifiable crystal structure. Nevertheless,  $\epsilon$ MSR is capable of finding and describing the statistics of even such highly disordered stacking structures. (iv) In contrast to many other techniques,  $\epsilon$ MSR uses all of the information available in diffraction spectrum. By integrating the diffraction spectrum over a unit interval in reciprocal space to find the CFs,  $\epsilon$ MSR makes no distinction between diffuse scattering and Bragg-like peaks. Each is treated equally. Indeed, even though Example B shows both Bragg-like peaks as well as considerable diffuse scattering between peaks,  $\epsilon$ MSR naturally captures the information contained in both by integrating over the entire spectrum. (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",  $\epsilon$ MSR builds the smallest possible model that reproduces the experimentally observed diffraction spectrum without over-fitting the data. Example C shows how  $\epsilon$ MSR is able to find this minimal expression for the stacking structure. (vi) Finally, the resulting expression of the stacking structure, the process's  $\epsilon$ -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.*, 2005b), we show how the average stacking energy and hexagonality may be calculated from the  $\epsilon$ -machine.

Additionally, we have identified three length parameters that are calculable from the  $\epsilon$ -machine: the correlation length,  $\lambda_c$ ; 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 a ML can carry nonredundant information about the orientation of another ML. This is most closely related to the  $r_l$ . If the assumption of equilibrium can be made for polytypes,  $r_l$  places a lower bound on  $r_l$ . But the assumption of equilibrium is critical, and not likely met by many polytypes.

Even with these advantages, however,  $\epsilon$ MSR as practiced here is not without its shortcomings. Perhaps most restrictive

is that  $\epsilon$ MSR is limited to Markov processes, and has only been worked out for 3<sup>rd</sup>-order Markov processes. Since the maximum number of terms in the spectral equations grows as the exponential of an exponential in the memory length, the task of writing out the higher order spectral equations quickly becomes prohibitively difficult. We believe that the  $r = 4$  case is almost certainly tractable, but the case of  $r \geq 5$  is probably not. Although  $r = 3$   $\epsilon$ -machines certainly identify much of the structure in higher order processes, we found two difficulties. (i) Approximations made in the derivation of the spectral equations can result in sequence probabilities that differ from those of the true process. As was shown in Example E this could interfere with the identification of stacking histories that have equivalent futures. (ii) The state labeling scheme imposes a CS architecture on the reconstructed  $\epsilon$ -machine that may be too restrictive. The  $\epsilon$ -machines in Examples D and E both belonged to a class of processes, formally known as sofic processes, that have a special kind of infinite range memory. The CSs on the  $\epsilon$ -machines that describe these processes can not be specified by any finite history. So the scheme of labeling states by the last  $r$ -spins seen, as is done here, is inadequate. Since the range of interaction between MLs in some materials, e.g. SiC (Cheng *et al.*, 1987; Cheng *et al.*, 1988; Shaw & Heine, 1990; Cheng *et al.*, 1990), is calculated to be  $r = 3$  and numerical simulations of martensitic transformations in ZnS suggest that the effective memory length is infinite (Varn & Crutchfield, 2004), alternate methods of inferring such long range structure from spectral data are needed. Reverse Monte Carlo techniques (Keen & McGreevy, 1990) have been applied to a wide range of disordered materials, and may be useful here. This is a current subject of research. Additionally, we are investigating alternative techniques to the direct solution of the spectral equations.

Finally, we stress that there is a difference between structure and mechanism in disordered stacking sequences. The  $\epsilon$ -machine describes the structure, but has little to say about how the material came to be stacked in this fashion. While it is possible to formally identify CSCs with "faulting structures" as we have done here, this can be misleading. It is certainly possible that the cumulative effects of repeated faulting by a particular mechanism may lead to a structure that is different from a crystal simply permeated with that kind of fault. That is, for high fault densities, adjacent faults may be produced in the same way, but the close proximity of the faults may cause us to interpret the structure differently—*e.g.*, as a small segment of complex crystal.

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  $\epsilon$ -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  $\epsilon$ -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

<sup>8</sup> For comparison, a completely random stacking of MLs for CPSS would have  $h_\mu = 1$  bit/ML.

# international union of crystallography

a crystal (Kabra & Pandey, 1988; Engel, 1990; Shrestha & Pandey, 1996a; Shrestha & Pandey, 1997; Gosk, 2000; Gosk, 2001; Gosk, 2003; Varn & Crutchfield, 2004). We note that in such simulations, the  $\epsilon$ -machine can be directly calculated from the sequence to high accuracy. Some experimental work on solid-state phase transitions has been done (Sebastian & Krishna, 1994), but we hope that this improved theoretical framework will stimulate additional effort in this direction.

## Acknowledgements

We thank L. J. Biven, D. P. Feldman, R. Haslinger, C. Moore, C. R. Shalizi and E. Smith for helpful conversations. 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*, 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. (1993). *The Theory of Critical Phenomena*. Clarendon Press.

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

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

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). In *Santa Fe Studies in the Sciences of Complexity*, edited by M. Casdagli & S. Eubanks, vol. XII. Reading, Massachusetts: 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.

Engel, G. E. (1990). *J. Phys. Cond. Mat.* **2**, 6905–6919.

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

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

Glauber, R. J. (1963). *J. Math. Phys.* **4**, 294–306.

Gosk, J. B. (2000). *Crys. Res. Tech.* **35**, 101–116.

Gosk, J. B. (2001). *Crys. Res. Tech.* **36**, 197–213.

Gosk, J. B. (2003). *Crys. Res. Tech.* **38**, 160–173.

Hansen, J. E. (1993). *Computational Mechanics of Cellular Automata*. Ph.D. thesis, University of California, Berkeley.

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

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

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

Keen, D. A. & McGreevy, R. L. (1990). *Nature*, **344**, 423–425.

Pandey, D. & Krishna, P. (1982). In *Current Topics in Materials Science*, edited by E. Kaldis. North-Holland.

Roth, W. L. (1960). *Faulting in ZnS*. Tech. Rep. 60-RL-2563M. General Electric Research, Schenectady, New York.

Sebastian, M. T. (1988). *J. Mat. Sci.* **23**, 2014–2020.

Sebastian, M. T. & Krishna, P. (1984). *Philos. Mag. A*, **49**, 809–821.

Sebastian, M. T. & Krishna, P. (1994). *Random, Non-Random and Periodic Faulting in Crystals*. 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. Cond. Mat.* **2**, 4351–4361.

Shrestha, S. P. & Pandey, D. (1996a). *Europhys. Lett.* **34**(4), 269–274.

Shrestha, S. P. & Pandey, D. (1996b). *Acta Mater.* **44**, 4949–4960.

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

Shrestha, S. P., Tripathi, V., Kabra, V. K. & Pandey, D. (1996). *Acta Mater.* **44**, 4937–4947.

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

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

Varn, D. P. (2001). *Language Extraction from ZnS*. Ph.D. thesis, University of Tennessee, Knoxville.

Varn, D. P. & Canright, G. S. (2001). *Acta Crystallogr. Sec. 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., (2005a). submitted to *Acta Crystallogr. Sec. A*.

Varn, D. P., Canright, G. S. & Crutchfield, J. P., (2005b). submitted to *Acta Crystallogr. Sec. B*.

Varn, D. P. & Crutchfield, J. P. (2004). *Phys. Lett. A*, **324**(4), 299–307.

Verma, A. R. & Krishna, P. (1966). *Polymorphism and Polytypism in Crystals*. John Wiley & Sons.

Yeomans, J. (1988). *Solid State Physics*, **41**, 151–200.

Yeomans, J. (1992). *Statistical Mechanics of Phase Transitions*. Clarendon Press.

Yi, J. & Canright, G. S. (1996). *Phys. Rev. B*, **53**(9), 5198–5210.

**Table 1**

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

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

**Table 2**

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

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

**Table 3**

The frequencies of length-4 sequences obtained from  $\epsilon$ MSR and SMM for the even process, Example D. At most, they differ by  $\pm 0.01$ .

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

**Table 4**

The frequencies of length-4 sequences obtained from Example E (ACA 232) and the  $\epsilon$ -machine reconstructed at  $r = 3$ .

Sequence	Example E	$\epsilon$ MSR	Sequence	Example E	$\epsilon$ MSR
1111	0.009	0.005	0111	0.041	0.043
1110	0.041	0.043	0110	0.000	0.000
1101	0.037	0.029	0101	0.331	0.336
1100	0.004	0.014	0100	0.037	0.029
1011	0.037	0.043	0011	0.004	0.000
1010	0.331	0.322	0010	0.037	0.043
1001	0.000	0.000	0001	0.041	0.043
1000	0.041	0.043	0000	0.009	0.005

**Table 5**

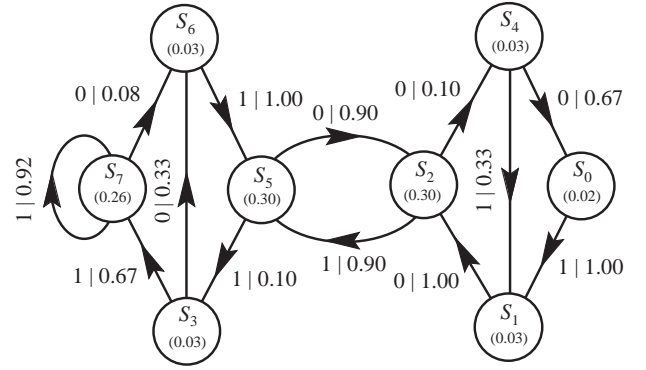
Measures of intrinsic computation calculated from the processes of Examples A, B, C, D and E and their ( $r = 3$ ) reconstructed  $\epsilon$ -machines. For Examples A, B, C and E the reconstructed  $\epsilon$ -machines give good agreement. For Example D, however, the reconstructed  $\epsilon$ -machine requires more memory and still has an entropy density  $h_\mu$  significantly higher than that of the even process. The last column gives  $\Delta = C_\mu - E - rh_\mu$  as a consistency check derived from Eq. (23) of (Varn *et al.*, 2005a), which describes order- $r$  Markov processes. Recall that the even process of Example D and ACA 232 of Example E are not a finite- $r$  processes and so Eq. (23) of (Varn *et al.*, 2005a) does not hold. All one can say is that  $E \leq C_\mu$ , which is the case for both Examples D and E.

System	Range	$h_\mu$ [bits/ML]	$C_\mu$ [bits]	E [bits]	$\Delta$
Example A	3	0.44	2.27	0.95	0.00
$\epsilon$ -machine	3	0.44	2.27	0.95	0.00
Example B	4	0.51	2.86	0.82	0.00
$\epsilon$ -machine	3	0.54	2.44	0.83	-0.01
Example C	1	0.67	0.92	0.25	0.00
$\epsilon$ -machine	1	0.67	0.92	0.25	0.00
Example D	$\infty$	0.67	0.92	0.91	
$\epsilon$ -machine	3	0.79	2.58	0.21	0.00
Example E	$\infty$	0.42	1.86	1.01	
$\epsilon$ -machine	3	0.42	2.26	0.99	0.01

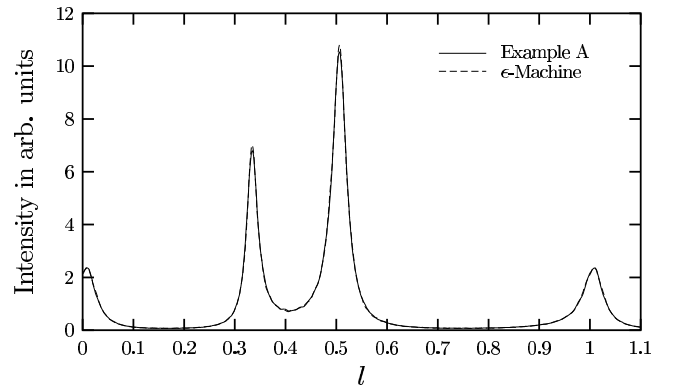
**Table 6**

The three characteristic lengths that one can calculate from knowledge of the  $\epsilon$ -machine: the correlation length  $\lambda_c$ , the recurrence length  $\mathcal{P}$ , and the memory length  $r_l$ . For comparison, we also give these quantities for several crystalline structures.

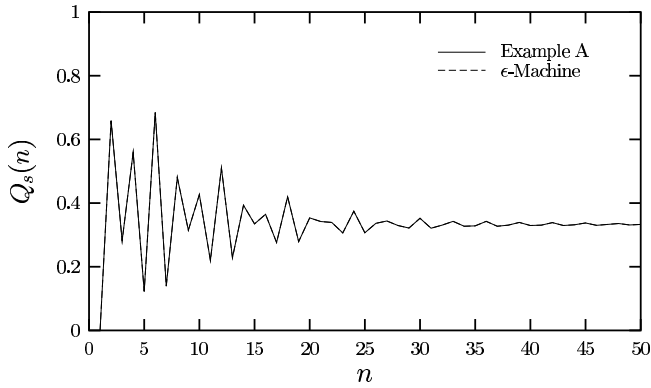
System	$\lambda_c$	$\mathcal{P}$	$r_l$
Example A, $r = 3$	$\sim 7.4$	4.8	3
Example B, $r = 4$	$\sim 7.8$	7.3	4
Example C, Golden Mean	$\sim 4.5$	1.9	1
Example D, Even Process	$\sim 1.7$	1.9	$\infty$
Example E, ACA 232, $f = 0.10$	$\sim 3.9$	3.6	$\infty$
3C	$\infty$	1	0
2H	$\infty$	2	1
6H	$\infty$	6	3


**Figure 1**

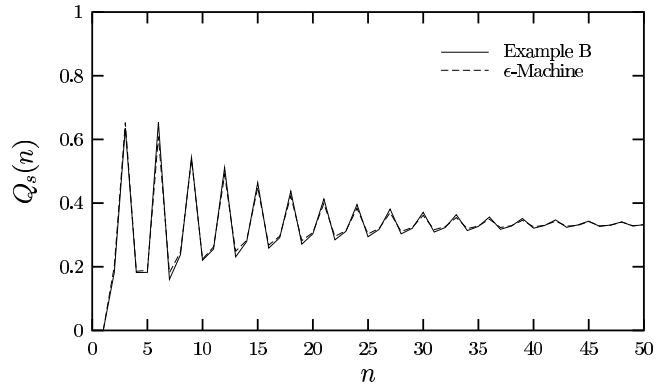
The  $r = 3$  theoretical and experimental  $\epsilon$ -machine for the Example A process. The nodes represent CSCs and the directed arcs are transitions between them. The edge labels  $s|p$  indicate that a transition occurs between the two CSCs on symbol  $s$  with probability  $p$ . The asymptotic probabilities for each CSC are given in parentheses. The large CSC probabilities for the  $[S_7]$  CSC ( $P_{CSC}([S_7]) = 0.92$ ) and the  $[S_2S_5]$  CSC ( $P_{CSC}[S_2S_5] = 0.81$ ) suggest that one think of these cycles as crystal structure and everything else as faulting.


**Figure 2**

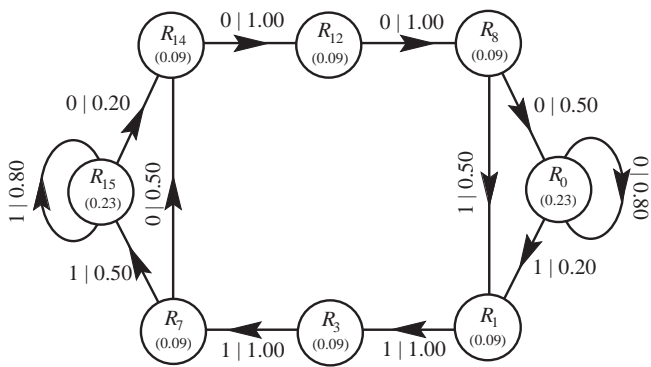
A comparison between the diffraction spectra  $I(l)$  generated by Example A and by the  $r = 3$  spectrally reconstructed  $\epsilon$ -machine. The differences between the diffraction spectra for Example A and the  $r = 3$  reconstructed  $\epsilon$ -machine are too small to be seen. We calculate  $\mathcal{R} = 2\%$ , but this is largely due to numerical error. (See text.) The peak at  $l \approx 1/3$  corresponds to the 3C structure and the two peaks at  $l \approx 1/2$  and  $l \approx 1$  to the 2H structure.



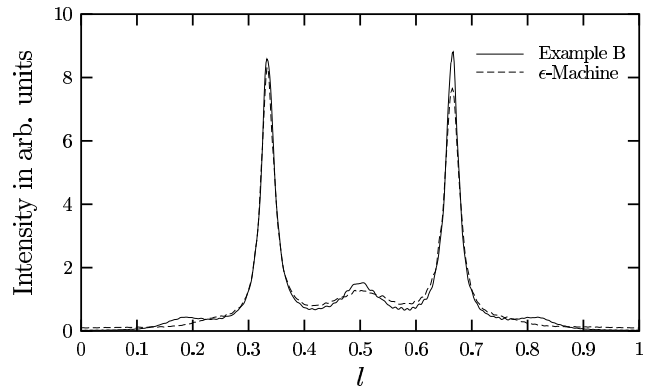
**Figure 3**  
A comparison of the CFs  $Q_s(n)$  between the Example A process and the  $r = 3$  reconstructed  $\epsilon$ -machine. As with the diffraction spectra, the differences are too small to be seen on the graph. As an aid to the eye, here and in other graphs showing CFs, we connect the the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



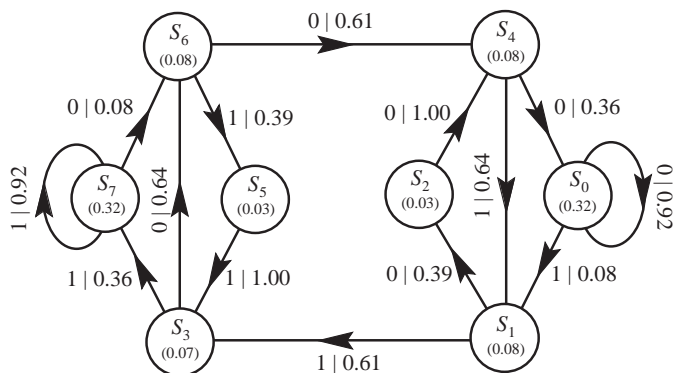
**Figure 6**  
A comparison of the CFs  $Q_s(n)$  generated by the  $r = 3$  reconstructed  $\epsilon$ -machine (dashed line) and generated by Example B (solid line). The agreement is excellent.



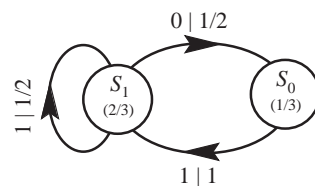
**Figure 4**  
The experimental  $\epsilon$ -machine for Example B. Since it has a memory of  $r_l = 4$ , we label the states with the last four spins observed: i.e.,  $\mathcal{R}_{12}$  means that 1100 were the last four spins. The CSCs  $[\mathcal{R}_{15}]$  and  $[\mathcal{R}_0]$  give rise to 3C 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.



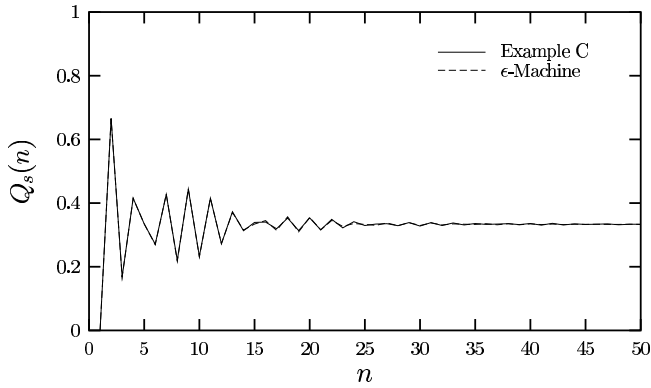
**Figure 7**  
A comparison of the diffraction spectra  $I(l)$  between  $r = 3$  reconstructed  $\epsilon$ -machine and the process of Example B. The agreement is surprisingly good; we calculate a profile  $\mathcal{R}$ -factor of  $\mathcal{R} = 12\%$ . The small peaks at  $l \approx 1/6$  and  $l \approx 5/6$  correspond to the 6H structure. The  $r = 3$   $\epsilon$ -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  $\epsilon$ -machine reconstructed at  $r = 4$  to properly disambiguate them.



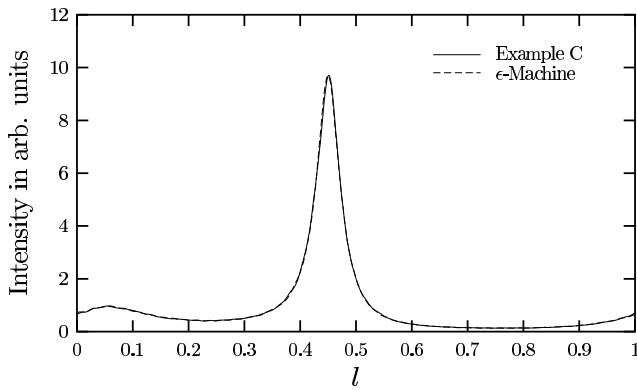
**Figure 5**  
The reconstructed (theoretical)  $\epsilon$ -machine at  $r = 3$  for Example B.



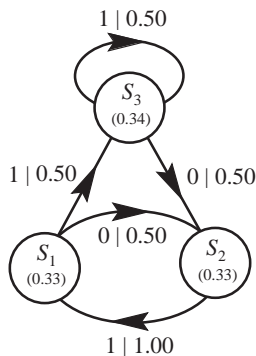
**Figure 8**  
The recurrent portion of the  $\epsilon$ -machine for the golden mean process, Example C. The process has a memory length of  $r = 1$ , and so we label each CS by the last spin seen.



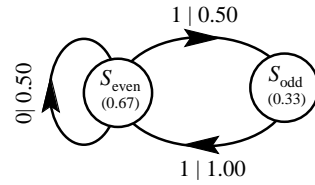
**Figure 9**  
A comparison of the CFs  $Q_s(n)$  generated by the  $r = 1$  reconstructed  $\epsilon$ -machine and the golden mean process of Example C. The CFs decay quickly to their asymptotic value of  $1/3$ .



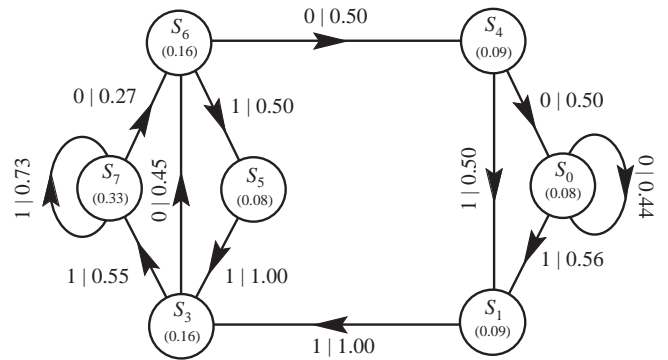
**Figure 10**  
A comparison of the diffraction spectra for Example C and the reconstructed  $r = 1$   $\epsilon$ -machine. The agreement is excellent. One finds a profile  $\mathcal{R}$ -factor of 2% between the experimental spectrum, Example C, and the theoretical spectrum calculated from the reconstructed  $\epsilon$ -machine.



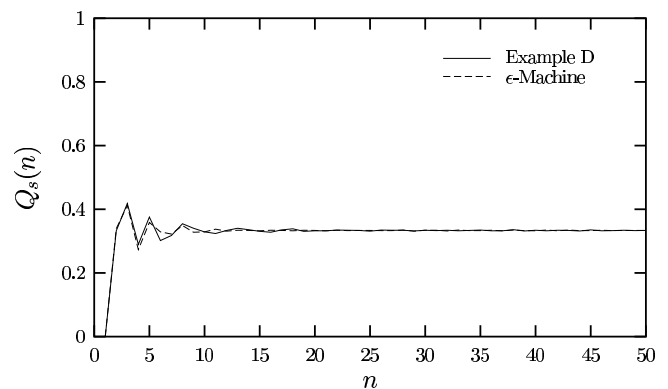
**Figure 11**  
The  $r = 2$  reconstructed non-minimal  $\epsilon$ -machine for the golden mean process, Example C. Applying the equivalence relation, equation (11) of (Varn *et al.*, 2005a), we find that  $S_1$  and  $S_3$  have the same futures, and thus should be collapsed into a single CS. Doing so gives the  $\epsilon$ -machine in figure 8.



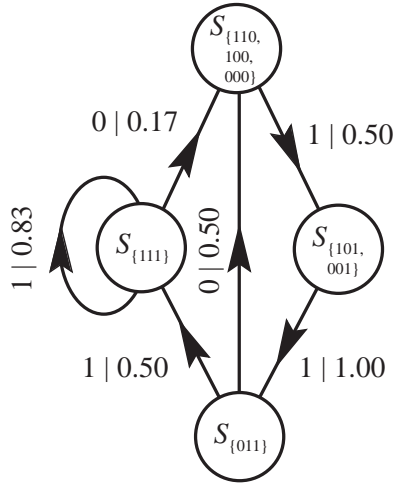
**Figure 12**  
The recurrent portion of the  $\epsilon$ -machine for the even process, Example D. Since the CSs cannot be specified by a finite history of previous spins, we have labeled them  $S_{\text{even}}$  and  $S_{\text{odd}}$ . We find that this  $\epsilon$ -machine has a statistical complexity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



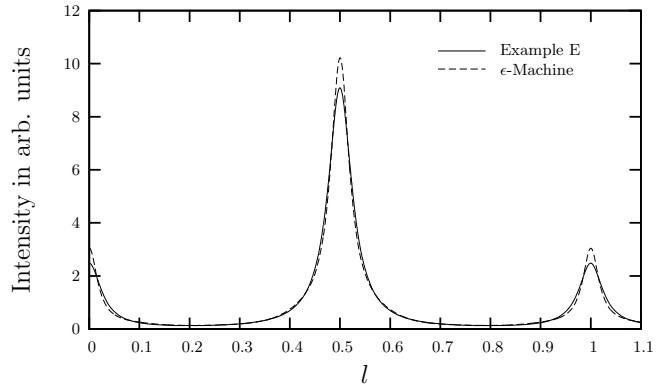
**Figure 13**  
The  $r = 3$  reconstructed  $\epsilon$ -machine for the even process of Example D. Since the even process forbids the sequences  $\{01^{2k+1}0, k = 0, 1, 2, \dots\}$  and all sequences containing them, it is satisfying to see that 010 is forbidden by the reconstructed  $\epsilon$ -machine, as evidenced by the missing  $S_2$  CS. We find that  $C_\mu = 2.58$  bits.



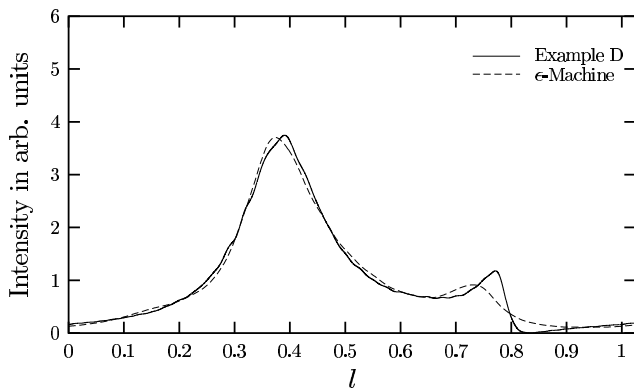
**Figure 14**  
A comparison of the CFs  $Q_s(n)$  generated by the  $r = 3$  reconstructed  $\epsilon$ -machine and the even process of Example D. The CFs decay quickly to their asymptotic value of  $1/3$ .



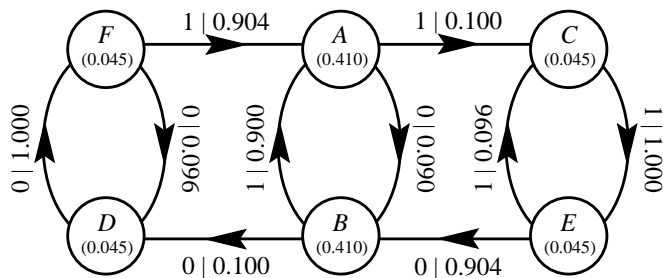
**Figure 15**  
The  $\epsilon$ -machine inferred from the exact sequence frequencies. The causal states are labeled with the (possibly several) histories that can lead to them. We find that  $C_\mu = 1.92$  bits.



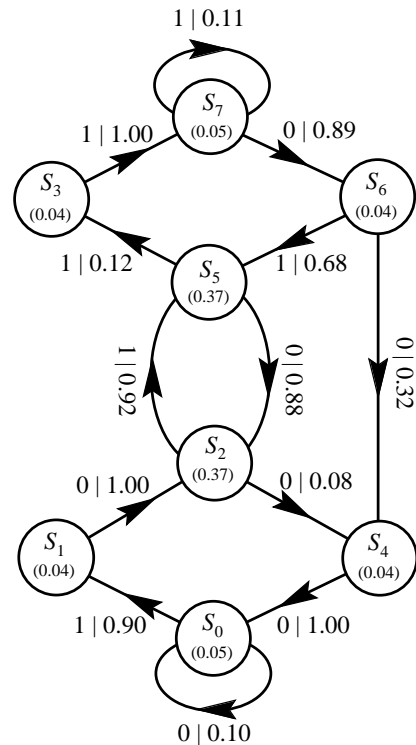
**Figure 18**  
A comparison between the diffraction spectra  $I(l)$  generated by the  $r = 3$  reconstructed  $\epsilon$ -machine and by Example E, ACA 232, with  $f = 0.10$ . We calculate a profile  $\mathcal{R}$ -factor of  $\mathcal{R} = 8\%$  between the experimental and theoretical diffraction spectra.



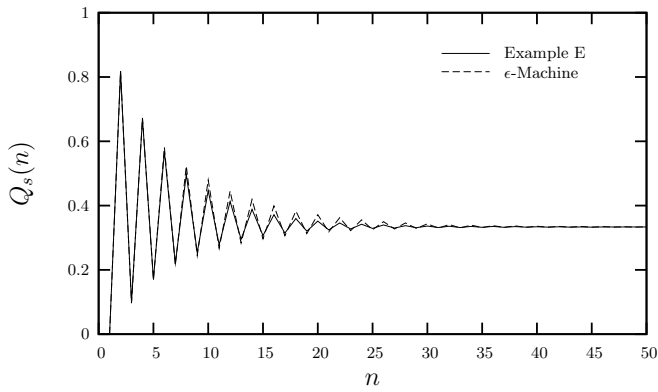
**Figure 16**  
A comparison between the diffraction spectra  $I(l)$  generated by the  $r = 3$  reconstructed  $\epsilon$ -machine and by the even process of Example D. The agreement is good ( $\mathcal{R} = 8\%$ ) except in the region  $0.7 < l < 0.9$ . Notably, the diffraction spectra for the even process has an isolated zero at  $l = 5/6$ .



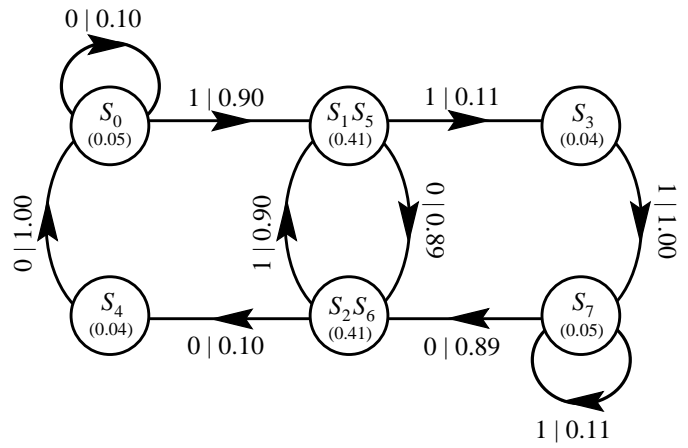
**Figure 17**  
The recurrent portion of the  $\epsilon$ -machine for Example E, ACA 232, with  $f = 0.10$ . This  $\epsilon$ -machine is sofic, as it prohibits spin domains with an even number of spins. This  $\epsilon$ -machine should be compared to the 10 state  $\epsilon$ -machine that describes ACA 232 for an arbitrary amount of faulting, figure 2 of (Varn & Crutchfield, 2004). For small amounts of faulting, we find that the CSs I, J, G and H of this latter  $\epsilon$ -machine collapse in to the CSs D, F, C and E respectively.



**Figure 19**  
The  $r = 3$  reconstructed  $\epsilon$ -machine for Example E, ACA 232, with  $f = 0.10$ . The large asymptotic state probabilities for  $S_2$  and  $S_5$ , as well as their large casual state cycle probability,  $P_{CSC}([S_2S_5]) = 0.81$ , indicate that this crystal is predominantly 2H.  $[S_2S_4S_0S_1]$  and  $[S_5S_3S_7S_6]$  are characteristic of deformation faulting of the 2H crystal structure.



**Figure 20**  
A comparison of the CFs  $Q_s(n)$  generated by the  $r = 3$  reconstructed  $\epsilon$ -machine and Example E, ACA 232.



**Figure 21**  
The reduced theoretical  $\epsilon$ -machine for Example E. This  $\epsilon$ -machine should be compared to the experimental  $\epsilon$ -machine given in figure 17. The CS architecture is nearly identical as are the CS probabilities and transitions between CSs.