

A direct method for inferring planar disorder and structure from X-ray diffraction studies

D. P. Varn,^{a,b,c,*†} G. S. Canright^{c,d,*‡} and J. P. Crutchfield^{b,e,*§}

^aMax-Planck-Institut für Physik komplexer Systeme, Nöthnitzer Straße 38, 01187 Dresden, Germany, ^bSanta Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA, ^cDepartment of Physics and Astronomy, University of Tennessee, 1408 Circle Drive, Knoxville, Tennessee 37996, USA, ^dTelenor Research and Development, 1331 Fornebu, Oslo, Norway, and ^eComputational 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

In a recent publication [*Phys. Rev. B* **66** 174110 (2002)] we introduced a new technique for discovering and describing planar disorder in close-packed structures directly from their diffraction spectra. Here we provide the theoretical development behind those results, adapting computational mechanics to describe one-dimensional structure in materials. We show that the resulting statistical model of the stacking structure—called the ϵ -machine—allows the calculation of measures of memory, structural complexity, and configurational entropy. By way of contrast, we give a detailed analysis of the current alternative approach, the fault model, and offer several criticisms. In the limit of weak faulting, we demonstrate how various architectural features of the ϵ -machine correspond to well-known structural faults in two common types of crystal, 2H and 3C. The methods developed here can be adapted to a wide range of experimental systems in which spectroscopic data is available.

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

1. Introduction

Fundamental to understanding the physical properties of a solid is a thorough description of its composition and the arrangement of its constituent parts. While chemical analysis provides information about composition, many complementary methods—such as X-ray diffraction, electron diffraction, high-resolution electron microscopy, optical microscopy, and X-ray diffraction tomography—are essential to discovering the large-scale structure of crystalline materials. For example, the placement of Bragg peaks in X-ray diffraction spectra has proved to be a particularly powerful source of structural information in ordered solids for nearly a century (von Laue, 1918). For solids that deviate from a strict periodic ordering of their constituent atoms, however, the diffraction spectrum typically shows a weakening and broadening of the Bragg peaks as well as the appearance of diffuse scattering. As the disorder becomes more pronounced, the Bragg peaks disappear altogether and leave a completely diffuse spectrum. The problem of inferring crystal structure for such disordered materials from X-ray diffraction has been addressed by many researchers (Frey, 1995; Jagodzinski, 1987; Proffen & Welberry, 1998; Welberry, 1985; Schulz, 1982; Guinier, 1963). In the most general case, however, the problem remains unsolved. Indeed, it is known that without the assumption of strict crystallinity, the problem has no unique solution (Woelfson, 1997).

Many kinds of disorder are present in solids (Hirth & Lothe, 1982; Ashcroft & Mermin, 1976; Kittel, 1996), such as Schottky defects, substitution impurities, screw and edge dislocations, and planar slips. Of these, only planar slips will be considered here. Planar defects occur in a crystal structure when one crystal plane is displaced from another by a non-Bravais lattice vector. These slips can occur during crystal growth or can result from some stress to the crystal, be it mechanical, thermal, or even through irradiation (Sebastian & Krishna, 1994). When an otherwise perfect crystal has some planar disorder, the portion of the structure that cannot be thought of as part of the crystal is called a *stacking fault*, or more simply, a *fault*. If there is a transition between two crystal structures, the interface between the two is also known as a stacking fault, even if each layer can be thought of as belonging to one of the crystal structures. Many different kinds of stacking faults have been postulated, including growth faults, deformation faults, and layer-displacement faults (Sebastian & Krishna, 1994; Frank, 1951*b*; Frey & Boysen, 1981).

Planar defects are surprisingly common in crystals, being especially prevalent in a broad class of materials known as *polytypes* (Verma & Krishna, 1966; Sebastian & Krishna, 1994; Trigunayat, 1991; Pandey & Krishna, 1982; Frevel *et al.*, 1992). First discovered in SiC (Baumhauer, 1912), polytypism has since been found in dozens of materials. Polytypism is the phe-

[†] Correspondence Author

[‡] Correspondence Author

[§] Correspondence Author

¹ This is only approximately true (Trigunayat, 1991).

international union of crystallography

nomenon of solids built up from identical layers,¹ called *modular layers* (MLs) (Varn & Canright, 2001), that differ only in the manner of the stacking. Typically, one finds that the *intra*-ML interactions are relatively strong as compared to the *inter*-ML interactions, so that disorder *within* a ML is rare. Energetic considerations usually restrict the allowed orientations of MLs to a discrete set, with only a small energy difference between two different stackings. Thus, the description of a polytype, ordered or disordered, formally reduces to a one-dimensional list—called the *stacking sequence*—that lists successive orientations encountered as one moves along the stacking direction.

The small energy difference between different stackings arises because the coordination of the nearest neighbor, next-nearest neighbor, and sometimes even higher neighbors is often the same regardless of the stacking arrangement. It is therefore possible to have many distinct stackings—some periodic and some not. For several of the most polytypic materials—e.g., SiC, ZnS, and CdI₂—there are about 150, 185, and 200 known crystalline structures, respectively. Remarkably, some have unit cells extending over 100 MLs (Sebastian & Krishna, 1994). The stacking period of many such polytypes is far in excess of the calculated inter-ML interactions, which are estimated to be, for example, ~ 1 ML in ZnS (Engel & Needs, 1990) and ~ 3 ML in SiC (Cheng *et al.*, 1987; Cheng *et al.*, 1988; Shaw & Heine, 1990; Cheng *et al.*, 1990). Nearly a dozen theories have been proposed (Sebastian & Krishna, 1994; Trigunayat, 1991; Frank, 1951a; Pandey & Krishna, 1982; Jagodzinski, 1954; Yeomans, 1988; Pandey, 1989), yet a satisfactory and systematic explanation is still lacking for the diversity and kinds of observed structure.

Much of the interest in polytypism has centered around the issue of long-range order and the existence of so many apparently stable structures. Reconciling the calculated range of interaction between MLs with the length scale over which organization appears has been the chief mystery of polytypism. Also of interest is the characterization of the solid-state transformations common in many of these materials. While the length scale on which spatial organization appears in these materials is easily found for crystal structures, the similar question for disordered structures has not so far been addressed. What is needed is a model that gives a statistical description of the observed stacking sequences from which characteristic length parameters are calculable. Finally, from a unified description of both crystalline and noncrystalline structures, a more comprehensive picture of polytypism should emerge and hopefully render polytypism more amenable to theoretical discussion and analysis.

A significant source of information about the structure of solids is derived from diffraction spectra. While it can be challenging to identify crystal structures with units cells over 100s of MLs, in fact most periodic structures have been identified (Sebastian & Krishna, 1994). In many polytypes, disordered sequences are also common, and a single crystal can contain regions of both ordered and disordered stackings. The main goal of this present work is to develop a technique for discovering and describing planar disorder in closed-packed structures (CPSs) directly from diffraction spectra. A further task is to

detail the connection between our model of disordered structures and physically relevant parameters derivable from it. We will then be in a position to treat similar questions about the possibility of long-range order in disordered structures (Varn *et al.*, 2005a).

Our development here is organized as follows: in §2 we discuss and offer several criticisms of the prevailing view of disorder in planar structures; in §3 we give a detailed account of our procedure for discovering and quantifying pattern and disorder in CPSs; in §4 we compare our approach to previous descriptions of disorder in CPSs; and in §5 we give our conclusions. In a pair of companion papers we demonstrate the use of our technique by applying it to several diffraction spectra from simulated stacking sequences (Varn *et al.*, 2005a), as well as to the analysis of experimental ZnS diffraction spectra (Varn *et al.*, 2005b).

2. The Fault Model

The problem of quantifying the effects of planar disorder on diffraction spectra has a long history (Sebastian & Krishna, 1994). Typically we can divide these approaches into two categories: *direct* or *indirect*. Direct methods make no assumption about the crystal structure or disorder that may be present in a specimen, but instead directly extract structural or correlation information from diffraction spectra. This contrasts to the more commonly used indirect methods which typically postulate a crystalline structure permeated by one or more fault structures. The effects of these faults on the observed diffraction spectrum are then used to fix the parameters of the model. We review significant efforts in each these areas in the next two subsections. We next offer our criticisms of these approaches in the third subsection and lastly discuss new theoretical techniques in statistical mechanics which render this inference problem more tractable.

2.1. Indirect Methods

Perhaps the first quantitative analysis was given by Landau and Lifschitz assuming no correlation between MLs (Landau, 1937; Lifschitz, 1937). Wilson provided an analysis of planar imperfections in hexagonal Co by considering the effect of the disorder on the Bragg peaks (Wilson, 1942). This approach is necessarily limited to the case of small amounts of faulting. Hendricks and Teller treated the problem rather generally, allowing for different form factors for the different kinds of MLs, variable spacing between MLs, and correlations between neighboring MLs (Hendricks & Teller, 1942). Since their method relies on extensive matrix calculations, it was found cumbersome and difficult to apply by early researchers.

Jagodzinski developed a theory of diffraction for planar disorder by considering nearest-neighbor correlations for general layered structures (Jagodzinski, 1949a) and for next-nearest neighbor correlations of CPSs (Jagodzinski, 1949b). By noting the direction, length, and intensity of non-Laue streaks in X-ray diffraction spectra of Cu-Si alloys, Barrett was able to estimate the kind and approximate amount of stacking disorder present (Barrett, 1950). Paterson considered the effect of deformation faults on face-centered cubic crystals (fcc or

3C²) and demonstrated how one can calculate the fraction of faulted layers from the widths and displacements of the Bragg peaks (Paterson, 1952). Gevers demonstrated how to calculate the effects of growth faults with an n -layer influence on the diffraction spectra of CPSs (Gevers, 1954b). He also demonstrated how to treat both growth and deformation faults randomly inserted into hexagonal (hcp or 2H) and cubic crystals, as well as deformation faults randomly distributed into 4H and 6H crystals (Gevers, 1954a). Johnson examined the effects on the diffraction pattern of random extrinsic faulting (insertion of a ML) in the 3C structure and compared this to the effects of intrinsic faulting (deletion of a ML) on the Bragg peaks in the limit of small fault probabilities (Johnson, 1963). Prasad and Lele considered the effects of faulting on the diffraction spectra of 4H crystals containing up to nine kinds of randomly distributed faults (Prasad & Lele, 1970). Pandey and Krishna treated the similar case of the 6H close-packed structure containing a random distribution of 14 distinct intrinsic faults (Pandey & Krishna, 1976). They later derived an expression for the intensity of diffracted radiation from a 2H crystal containing any amount of randomly placed deformation and growth faults (Pandey & Krishna, 1977). By measuring the broadening of the diffraction maxima of a SiC crystal they were able to determine the amount of each kind of faulting. Michalski developed a general theory for the random distribution of single stacking faults for an arbitrary periodic structure (Michalski, 1988) and applied it to several hexagonal and rhombohedral structures (Michalski *et al.*, 1988).

The theory of *nonrandom faulting* was developed in an effort to understand experimental data concerning solid-state transformations from the 2H to the 6H structure in annealed SiC crystals (Pandey *et al.*, 1980a; Pandey *et al.*, 1980b; Pandey *et al.*, 1980c). The idea is that the presence of a fault in a structure affects the probability of finding another fault in close proximity. By postulating two possible faulting mechanisms—deformation faulting and layer-displacement faults—Pandey *et al.* were able to understand the stacking in SiC within this model. Similar work was done on the 2H to 3C transformation in ZnS crystals (Sebastian, 1988; Sebastian & Krishna, 1984; Sebastian & Krishna, 1987b; Sebastian & Krishna, 1987c; Sebastian *et al.*, 1982; Sebastian & Krishna, 1994; Frey *et al.*, 1986; Pandey & Lele, 1986a; Pandey & Lele, 1986b; Jagodzinski, 1972).

All of the above methods center on finding a relationship, preferably an analytical one, that relates the observed diffracted intensity to fault types and probabilities. In order to make a quantitative estimate of the faulting, typically the placement, broadening, shape, and symmetry of the Bragg peaks are compared with that expected for a crystal containing a particular kind of fault. Then the full width at half-maximum (FWHM) of one or several of the Bragg peaks is used to determine the

fraction of faulting.

Thus, all of these approaches are limited to small amounts of disorder that preserve the integrity of the Bragg peaks.³ Should the disorder become sufficiently large, the Bragg peaks become too broad and do not stand out sufficiently from the diffuse, background scattering. We call these kinds of approaches *indirect*, because one begins with a set of postulated faults and then sorts through them, searching for one or several that best fit the data.

2.2. Direct Methods

Perhaps the first efforts at a *direct* method for determining the structure of disordered close-packed crystals were given by Dornberger-Schiff (Dornberger-Schiff, 1972) and Farkas-Jahnke (Farkas-Jahnke, 1973b; Farkas-Jahnke, 1973a). Dornberger-Schiff gave an algorithm for relating Patterson values, *i.e.* fractions of faulted layers from Bragg peaks, to sequence probabilities but, to our knowledge, did not follow up with a method for finding the Patterson values from spectra showing diffuse scattering. Farkas-Jahnke used Patterson-like functions to estimate the frequency of occurrence of layer sequences up to length five. He was not able to obtain a complete set of equations and this forced the use of inequalities derived from symmetry arguments that do not generally hold in disordered crystals. Recently it has been demonstrated that correlation information can be extracted from powder diffraction spectra to measure characteristic length parameters in a faulted crystal (Estevez-Rams *et al.*, 2001a; Estevez-Rams *et al.*, 2001b; Estevez-Rams *et al.*, 2003).

Unlike the previously considered techniques, these methods are direct since they make no assumption about underlying crystal structure or the faults it may contain. To our knowledge, since their introduction three decades ago, neither of the first two methods have been used to discover stacking structure in real materials. Hence we do not treat them further here.

2.3. The Fault Model Defined

We refer to the indirect approaches—analyzing a crystal structure assuming it contains a distribution of stacking errors or faults—as the *fault model* (FM) (Varn *et al.*, 2002). To date, it has been the dominant way in which planar disorder in crystals has been viewed. However, we find a number of difficulties with the FM, many of which have been recognized by previous researchers (Gosk, 2000; Proffen & Welberry, 1998; Palosz & Przedmojski, 1976; Farkas-Jahnke, 1973b; Treacy *et al.*, 1991; Michalski, 1988). Our objections to the FM and the way it has been used to discover structural information are severalfold.

(i) *The FM assumes a parent crystal.* For the fault model to make sense, it is necessary to assume some crystal structure in which to introduce faulting. This may be satisfactory for

² For close-packed crystal structures it is common to use the Ramsdell notation, nX , where n gives the total number of MLs in the unit cell and X specifies the crystal symmetry. We use H to denote hexagonal symmetry, C for cubic, R for rhombohedral and L if the symmetry is unknown. This notation is not necessarily unique for some longer period polytypes. We will use the Ramsdell notation to specify crystal structures. The equivalents in the Hägg notation are given in Table 2. For a more complete discussion, see (Sebastian & Krishna, 1994).

³ An exception to this is the *disorder model* (Jagodzinski, 1949b). This approach has several features in common with our own and can in fact be thought of as a constrained case of our approach.

international union of crystallography

weakly faulted crystals, but for those with significant disorder or those undergoing a solid-state phase transition to another crystal structure (Frey *et al.*, 1986; Jagodzinski, 1972; Kabra & Pandey, 1995; Krishna & Marshall, 1971a; Krishna & Marshall, 1971b; Pandey & Lele, 1986a; Pandey & Lele, 1986b; Pandey *et al.*, 1980a; Pandey *et al.*, 1980b; Pandey *et al.*, 1980c; Sebastian *et al.*, 1982; Sebastian & Krishna, 1987b; Sebastian & Krishna, 1987c; Sebastian & Krishna, 1987a; Sebastian *et al.*, 1987; Shrestha & Pandey, 1996a; Shrestha & Pandey, 1996b; Shrestha *et al.*, 1996; Shrestha & Pandey, 1997), this picture is untenable.

(ii) *Each parent crystal must be treated separately.* Since the FM introduces stacking “mistakes” into a parent crystal, each kind of parent crystal must be treated individually. There has been significant work on only two CPSs; namely, the 2H and 3C. In polytypism hundreds of other crystalline structures are known to exist. In the FM, each must be analyzed separately by postulating appropriate crystal-specific kinds of defects. Given this degree of complication, it is desirable to find a theoretical framework that unites the description of the various kinds of fault and parent structure into a single, coherent picture.

(iii) *In practice, the FM treats only the Bragg peaks quantitatively, effectively ignoring the diffuse scattering.* Most researchers use formulæ that give the FWHM of Bragg peaks in terms of the fraction of certain postulated defects to find the amount of faulting. Some do acknowledge that diffuse scattering is important, but to our knowledge none use the diffuse scattering to quantitatively measure crystal structure.⁴

(iv) *The FM is unable to capture the variety of naturally occurring stacking sequences.* By assuming a small set of possible ways that a parent structure can deviate from crystallinity, the FM necessarily assumes that there are stacking sequences that do not occur. It is desirable to have an approach that considers as many candidate structures as possible with as few *a priori* restrictions as possible.

(v) *The FM’s description of the disorder is not unique.* It is possible to give two different faulting schemes that describe the same weakly faulted material (Varn *et al.*, 2002). This is readily seen by noting that the layer-displacement fault in a 2H crystal can be viewed as two adjacent, but oppositely oriented deformation faults.

2.4. Computational Mechanics

Our method of discovery and quantification of planar structure and disorder in crystals overcomes all of these difficulties for the special but important case of CPSs. We do not assume an underlying crystalline structure. Indeed, we make no assumptions at all about either the crystal or fault structure that may be present. Instead, we find the frequency of occurrence of all possible stacking sequences up to a given length and use this to construct a model that captures the statistics of the stacking sequence. In this sense, we directly determine the stacking structure. Our scheme for describing planar disorder unites

both fault and crystal structure into a single framework. There is no need to treat each crystal structure or faulting scheme separately. Our method treats any amount and kind of planar disorder present. Finally, we quantitatively use all of the information contained in the diffraction spectra, both in Bragg peaks and in diffuse scattering, to build a unique model of the stacking structure. This model does *not* find the particular stacking sequence of the specimen that generated the diffraction pattern—this is not possible from diffraction spectra alone—but rather finds the statistical regularities across an ensemble of stacking sequences that could have given rise to the observed spectra. This is the best that can be done, in principle.

The history of discovering planar disorder in crystals is one of consistent, incremental progress over nearly seventy years. There are two factors, however, that have hindered progress in this area. The first is calculational. Much of the early work centered on finding analytical expressions for the diffracted intensity of a given crystalline structure permeated with a particular fault type. With the advent of modern numerical and symbolic calculational methods and the concomitant ability to estimate diffraction patterns from any arbitrarily layered structure (Treacy *et al.*, 1991), much of this early work has been superseded. The second hindrance to progress has been a lack of fundamental understanding of structure and disorder in one-dimensional sequences generated by nonlinear dynamical systems. Recently, however, a unifying framework has been introduced in the theory of *computational mechanics* (Crutchfield & Young, 1989; Crutchfield & Feldman, 1997; Crutchfield & Feldman, 2003; Feldman & Crutchfield, 1998; Shalizi & Crutchfield, 2001; Young, 1991; Hansen, 1993; Feldman, 1998).

Computational mechanics is an approach to discovering, describing, and quantifying patterns. It provides for the construction of the minimal and unique model for a process that is optimally predictive; this model is called an ϵ -machine. A process’s ϵ -machine is minimal in the sense of requiring the fewest model components to represent the process’s structures and disorder; it is optimal in the sense that no alternative representation is more accurate; and it is unique in the sense that any alternative which is both minimal and optimally predictive is isomorphic to it. An ϵ -machine’s algebraic structure captures a process’s symmetries and approximate symmetries. From an ϵ -machine measures of a process’s memory, entropy production, and structural complexity can be found. We demonstrate elsewhere (Varn *et al.*, 2005b) that knowledge of the ϵ -machine and the energy coupling between MLs allows one to calculate the average stacking energy for a disordered polytype.

Before being adapted to the present application of disorder in crystals, computational mechanics had been used to analyze structural complexity in a wide range of nonlinear processes, such as cellular automata (Hansen, 1993; Hanson & Crutchfield, 1997; Hordijk *et al.*, 2001), the logistic map (Crutchfield & Young, 1989; Young, 1991), and the one-dimensional Ising model (Feldman, 1998; Crutchfield & Feldman, 1997), as well

⁴ There are other techniques for discovering structure in solids that do use the diffuse scattering quantitatively, such as reverse Monte Carlo (RMC) simulation (Keen & McGreevy, 1990). The application of RMC to the discovery of planar disorder in CPSs is a current topic of research.

as to experimental physical systems, such as the dripping faucet (Gonçalves *et al.*, 1998), atmospheric turbulence (Palmer *et al.*, 2000), and geomagnetic data (Clarke *et al.*, 2003). Additionally, information theoretic ideas have been applied to the problem of polytypism (Estevez-Rams *et al.*, 2003) and hidden Markov models have been used to treat the problem of electron transport in mesoscopic systems (Kanter *et al.*, 2005). In the present setting, this latter example can be thought of as a special case of ϵ -machines. Finally, computational mechanics has also been used to quantify structure and self-organization in two dimensional patterns (Feldman & Crutchfield, 2003), and to distinguish between healthy and diseased brain function (Young *et al.*, 2005).

3. ϵ -Machine Spectral Reconstruction

Previous techniques of ϵ -machine reconstruction have used a sequence of data produced by the process (Crutchfield & Young, 1989; Hansen, 1993; Crutchfield, 1994; Shalizi *et al.*, 2002). Here, the experimental signal comes in the form of a power spectrum, and we need to develop a technique to infer the ϵ -machine from this type of data. We call this new class of inference algorithms *ϵ -machine spectral reconstruction*—abbreviated ϵ MSR and pronounced “emissary”. We emphasize that our goal remains unchanged—to find the process’s underlying description. It is only the inference procedure that is changed. In this section we give a detailed account of ϵ MSR as applied to the problem of discovering pattern and disorder in CPSs.

We divide ϵ MSR into five steps. First, we extract correlation information from a diffraction spectrum. Second, we use this to estimate stacking-sequence probabilities of a given length. Third, we reconstruct an ϵ -machine from this distribution. Fourth, we generate a diffraction spectrum from the ϵ -machine. And, finally, we compare this ϵ -machine spectrum to the original. If there is insufficient agreement, we repeat the second through fourth steps, estimating stacking-sequence probabilities at a longer length, building a new ϵ -machine, and again comparing with the original spectrum. In the final two subsections, we give relations that can be used to determine the quality of experimental data and briefly review several information- and computation-theoretic quantities of physical import that can be directly estimated from the reconstructed ϵ -machine.

3.1. Correlation Factors from Diffraction Spectra

We start with the conventional assumptions concerning polytypism in CPSs. Namely, we assume that

- the MLs themselves are undefected and free of any distortions;
- the spacing between MLs does not depend on the local stacking arrangement;
- each ML has the same scattering power; and
- the faults extend completely across the crystal.

We make the additional assumption that the probability of finding a given stacking sequence in the crystal remains constant

through the crystal. [In statistics parlance, we assume that the process is *stationary*.]

Let N be the number of hexagonal, close-packed MLs, with each ML occupying one of three orientations, denoted A , B , or C (Ashcroft & Mermin, 1976; Kittel, 1996; Sebastian & Krishna, 1994). We introduce three statistical quantities, $Q_c(n)$, $Q_a(n)$ and $Q_s(n)$: the two-layer *correlation functions* (CFs), where c , a , and s stand for *cyclic*, *anti-cyclic*, and *same*, respectively (Yi & Canright, 1996). $Q_c(n)$ is defined as the probability that any two MLs at a separation of n are cyclically related. By cyclic, we mean that if the i^{th} ML is in orientation A (B , C), say, then the $(i+n)^{\text{th}}$ ML is in orientation B (C , A). $Q_a(n)$ and $Q_s(n)$ are defined in a similar fashion. Since these are probabilities, $0 \leq Q_\alpha(n) \leq 1$, where $\alpha \in \{c, a, s\}$. Additionally, at each n it is clear that $\sum_\alpha Q_\alpha(n) = 1$.

With these assumptions and definitions in place, the *total diffracted intensity* along the $10.l$ can be written as (Guinier, 1963; Berliner & Werner, 1986; Yi & Canright, 1996)

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

where l is a continuous variable that indexes the magnitude of the perpendicular component of the diffracted wave, $k = 2\pi l/c$, and c is the spacing between adjacent MLs.⁵ $\psi^2(l)$ is a function that accounts for atomic scattering factors, the structure factor, dispersion factors, or any other effects for which the experimentally obtained diffraction spectra may need to be corrected (Hahn *et al.*, 1992; Woolfson, 1997; Milburn, 1973).

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

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

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

The form of Eqs. (1) and (2) suggests that the CFs can be found from the diffraction pattern by Fourier analysis (Varn, 2001; Varn *et al.*, 2002; Estevez-Rams *et al.*, 2003; Estevez-Rams *et al.*, 2005). Let us define $X(n)$ and $Y(n)$ as

$$X(n) = \oint I(l) \cos(2\pi n l) dl \quad (3)$$

and

$$Y(n) = \oint I(l) \sin(2\pi n l) dl, \quad (4)$$

⁵ We use the standard notation conventions here. See (Guinier, 1963) and references therein for a discussion of typical geometries and notations. We set $H = h_1 = 1$ and $K = h_2 = 0$ as is typically reported from experiment. The stacking direction is perpendicular to the faulted planes. Note that our definition of l (see text) differs from that of many authors (Sebastian & Krishna, 1994).

international union of crystallography

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

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

and

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

Thus, the CFs can be found by Fourier analysis of the diffraction spectrum.

3.2. Estimating the Stacking-Sequence Distribution

In the second part of our approach, we estimate the distribution of stacking sequences from the two-layer CFs. First, though, we must consider what kind of information the CFs contain about stacking sequences. Therefore let us define the *stacking process* as the effective stochastic process induced by scanning the stacking sequence along the stacking direction. It is convenient to represent the stacking sequence in terms of the Hägg notation (Sebastian & Krishna, 1994), where one replaces the set of allowed orientations $\{A, B, C\}$ of a ML with a binary alphabet $\mathcal{A} = \{0, 1\}$. On moving from the i^{th} to the $(i + 1)^{\text{th}}$ ML, we label each inter-ML transition or *spin* (Varn & Canright, 2001) as “1” if the two MLs are cyclically related $[A \rightarrow B \rightarrow C \rightarrow A]$ and “0” if the two MLs are anti-cyclically related $[A \rightarrow C \rightarrow B \rightarrow A]$. Thus, the stacking constraint that no two adjacent MLs may have the same orientation $[A, B \text{ or } C]$ is built into the notation. There is a one-to-one mapping between the stacking orientation sequence and the spin sequence, up to an overall rotation of the crystal; and we use them interchangeably.

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

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

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

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

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

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

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

where $\mathcal{A}_\alpha^n$ is the subset of length- n sequences that generate a cyclic ($\alpha = c$) or an anti-cyclic ($\alpha = a$) rotation between MLs

at separation n . A sequence generates a cyclic (anti-cyclic) rotation between MLs at separation n if $2m - n = 1 \pmod{3}$, where m is the number of 1s (0s) in the sequence. We take as many of the relations in Eq. (9) as necessary to form a complete set of equations to solve for $P(\omega)$. For $r = 1$ and $r = 2$ the sets of equations are linear and admit analytical solutions. At $r = 3$, the first nonlinearities appear due to the necessity of using CFs at $n = 5$ to obtain a complete set of equations. We rewrite the conditional probabilities at $n = 5$ in terms of those at $n = 4$ via relations of the form

$$\begin{aligned} P(s_0 s_1 s_2 s_3 s_4) &= P(s_0 s_1 s_2 s_3) P(s_4 | s_0 s_1 s_2 s_3) \\ &\approx P(s_0 s_1 s_2 s_3) P(s_4 | s_1 s_2 s_3) \\ &= \frac{P(s_0 s_1 s_2 s_3) P(s_1 s_2 s_3 s_4)}{P(s_1 s_2 s_3 0) + P(s_1 s_2 s_3 1)} , \end{aligned} \quad (10)$$

where, in the second line, the approximation is invoked (Badii & Politi, 1997). We refer to this approximation as *memory-length reduction*, as it effectively limits the memory that we consider in order to obtain a complete set of equations. At fixed r , the set of equations describes the stacking sequence as an r^{th} -order Markov process.

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

3.3. ϵ -Machine Reconstruction from the Stacking Process

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

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

The effective states or *causal states* (CSs) of the stacking process are defined as the *sets* of pasts $\overleftarrow{\omega}$ that lead to statistically equivalent futures:

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

for all futures $\overrightarrow{\omega}$, where $P(\overrightarrow{\omega} | \overleftarrow{\omega}_i)$ is the conditional probability of seeing $\overrightarrow{\omega}$ having just seen $\overleftarrow{\omega}_i$ (Crutchfield & Young, 1989; Crutchfield, 1994; Shalizi & Crutchfield, 2001).

As a default set of CSs, we initially assume that each history of length r forms a unique CS. So, for ϵ MSR at r , we begin with 2^r CSs, each labeled by its unique length- r history. We refer to this set of CSs as *candidate causal states*, as they may not be the true CSs that describe the stacking process. We now estimate the state-to-state transition probabilities between candidate CSs as follows. Define the *transition matrices* $T_{S_i \rightarrow S_j}^{(s)}$ as the probability of making a transition from a candidate CS S_i to a candidate CS S_j on seeing spin s . If we label each past by the last r spins

seen, then this implies that only transitions of the form $s_0v \rightarrow vs$ are allowed, where $v \in \mathcal{A}^{r-1}$. All other transitions are taken to be zero. Then we can write the transition matrix as

$$T_{S_i \rightarrow S_j}^{(s)} = T_{s_0v \rightarrow vs}^{(s)} \quad (12)$$

We estimate these transition probabilities from the conditional probabilities,

$$\begin{aligned} T_{s_0v \rightarrow vs}^{(s)} &\approx P(s|s_0v) \\ &= \frac{P(s_0vs)}{P(s_0v)}. \end{aligned} \quad (13)$$

We now apply the equivalence relation, Eq. (11), to merge histories with equivalent futures. The set of resulting CSs, along with the transitions between states, defines the process's ϵ -machine. This is the minimal, unique description of the stacking process that optimally produces the stacking distribution $P(\omega)$. At this point, we should refer to this as the *candidate* ϵ -machine, as it will reproduce the CFs used to find it, but it may fail to reproduce CFs at larger n satisfactorily. We address this issue of agreement between theory and experiment in §3.5.

3.4. Correlation Functions and Diffraction Spectra from the Reconstructed ϵ -machine

In the fourth part, we use the reconstructed ϵ -machine to generate CFs and diffraction spectra. This is most simply accomplished by using the reconstructed ϵ -machine to generate a sample spin sequence M spins long in the Hägg representation. One can then change representations by mapping this spin sequence to a stacking-orientation sequence in the *ABC* notation. CFs can be found directly by scanning the stacking-orientation sequence.⁶ The diffraction spectrum is readily calculated from Eqs. (1) and (2). It has been shown that for sufficiently large M , the diffraction spectrum for diffuse scattering scales as M (Varn, 2001), so that the number of MLs used to calculate the diffraction spectrum is not important, if M is sufficiently large [say, 10 000]. To reduce the error due to fluctuations, it is desirable to use as long a sequence as possible to find the CFs. In this way, the ϵ -machine's predicted CFs and diffraction spectrum can be calculated.

It is worth repeating that our method of ϵ -machine reconstruction is novel in the sense that we do not estimate sequence probabilities from a long string of symbols generated by the process, as has been done previously (Crutchfield & Young, 1989; Hansen, 1993; Crutchfield, 1994; Shalizi *et al.*, 2002). Rather we use the two-layer CFs obtained from Fourier analysis of the diffraction spectra. In this way, ϵ MSR is accomplished purely from spectral information.

3.5. Comparing Original with ϵ -Machine Diffraction Spectra

In the fifth and final part, we compare the CFs and diffraction spectrum predicted by the ϵ -machine to those of the original

spectrum. If there is not sufficient agreement, we increment r and repeat the reconstruction and comparison.

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

$$\mathcal{R} = \frac{\oint |l_{\epsilon M}(l) - l_{exp}(l)| dl}{\oint l_{\epsilon M} dl} \times 100\%, \quad (14)$$

where $l_{\epsilon M}(l)$ is the ϵ -machine diffraction spectrum and $l_{exp}(l)$ is the experimental diffraction spectrum. Notice that the denominator is unity due to normalization.

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

$$\mathcal{R}_{err} = \frac{\oint |\delta l_{exp}(l)| dl}{\oint l_{exp} dl} \times 100\%. \quad (15)$$

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

3.6. Figures-of-Merit for Spectral Data

An issue we have so far neglected is the CFs' independence. In order to solve the spectral equations, part 3 in ϵ MSR (§3.3), we need 2^{r+1} independent constraints. It is therefore important to identify and avoid using any redundancies inherent in the CFs to solve the spectral equations. Rather than finding this a hindrance, any relations that CFs obey can be exploited to assess the quality of experimental data over a given l -interval. We find that, as a result of stacking constraints and conservation of probability, there are two equalities that the CFs must satisfy. We develop and define these measures here.

We find the first by observing that, at $n = 1$, due to stacking constraints, $Q_c(1) + Q_a(1) = 1$. Adding Eqs. (5) and (6) with $n = 1$ immediately gives $X(1) = -1/2$. This suggests that we define a *figure-of-merit* γ as

$$\gamma = \oint l(l) \cos(2\pi l) dl. \quad (16)$$

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

⁶ There are other techniques one can use to find CFs from ϵ -machines. See (Varn *et al.*, 2005a).

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

international union of crystallography

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

$$P(00) + 2P(1) - P(11) = 1. \quad (17)$$

Making the identification from Eq. (9) that $Q_c(1) = P(1)$, $Q_a(2) = P(11)$, and $Q_c(2) = P(00)$ gives

$$2Q_c(1) + Q_c(2) - Q_a(2) = 1. \quad (18)$$

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

$$\beta = 2Q_c(1) + Q_c(2) - Q_a(2). \quad (19)$$

β should be unity for error-free data. This can also be used to evaluate the quality of the experimental data over a given unit l -interval. Together, γ and β are the figures-of-merit over a unit l -interval for a diffraction spectrum. Therefore, in the first part of ϵ MSR (§3.1) we evaluate each over candidate l -intervals and choose an interval for ϵ -machine reconstruction that gives figures-of-merit best in agreement with the theoretical values. These two constraints on the CFs imply that only two out of the first four correlation functions, $Q_c(1)$, $Q_a(1)$, $Q_c(2)$, and $Q_a(2)$ are independent. We choose to take the $n = 2$ terms as the independent parameters in the spectral equations.

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

3.7. Measures of Structure and Intrinsic Computation

There are a number of different quantities in computational mechanics that describe the way information is processed and stored. (Crutchfield & Feldman, 2003; Shalizi & Crutchfield, 2001). We consider only the following.

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

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

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

where $\mathcal{S}$ is the set of CSs for the process and $P(\sigma)$ is the asymptotic probability of CS σ . The latter is the left eigenvector, normalized in probability, of the state-to-state transition matrix $T = \sum_{s \in \mathcal{A}} T^{(s)}$. Physically, the statistical complexity is related to the *average* number of previous spins one needs to observe on scanning the spin sequence to make an optimal prediction of the next spin. The statistical complexity is also related to a generalization of the stacking period for non-periodic processes (Varn et al., 2005a).

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

$$h_\mu = -\sum_{\sigma \in \mathcal{S}} P(\sigma) \sum_{s \in \mathcal{A}} T_{\sigma \rightarrow \sigma'}^{(s)} \log_2 T_{\sigma \rightarrow \sigma'}^{(s)}, \quad (21)$$

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

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

$$E = \sum_{\omega} P(\omega) \log_2 \frac{P(\omega)}{P(\bar{\omega})P(\omega)}. \quad (22)$$

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

$$C_\mu = E + rh_\mu. \quad (23)$$

For general nonfinite-range Markov processes, at present all that can be said is that the statistical complexity upper bounds the excess entropy: $E \leq C_\mu$ (Shalizi & Crutchfield, 2001).

4. ϵ -Machine and Fault Model Structural Analyses

Now that a statistical description of the stacking process has been found in the form of an ϵ -machine, it is desirable to give an intuitive notion of what the structure of the ϵ -machine tells us about the patterns and disorder in a stacking process. In this section, we define and discuss architectural features of $r = 3$ ϵ -machines and their relation to the FM. Specifically, we detail the form that growth, deformation, and layer-displacement faulting for the 2H and 3C structures assume on an $r = 3$ ϵ -machine. With this connection in place, we then address the general interpretation of ϵ -machines as related to the stacking of CPSs.

4.1. Causal-State Cycles

4.1.1. Definitions Since the ϵ -machine reconstructed at r can distinguish at most only 2^r pasts, it can have no more than 2^r CSs. The most general reconstructed ϵ -machine of memory length r is topologically equivalent to a *de Bruijn graph* (Teubner, 1990) of order r . By “most general” we mean that all length- r pasts are distinguished and all allowed transitions between CSs exist. Under these assumptions, the most general binary $r = 3$ ϵ -machine [which has $2^3 = 8$ CSs and $2^{3+1} = 16$ transitions] is shown in Fig. 1. It is known that de Bruijn graphs can be broken into a finite number of closed, nonself-intersecting loops called *simple cycles* (SCs) (Canright & Watson, 1996).

By analogy, we define a *causal-state cycle* (CSC) as a finite, closed, nonself-intersecting, symbol-specific path on an ϵ -machine. We denote a CSC by the sequence of CSs visited in square brackets []. The states themselves are labeled with a number (in binary notation) that gives the sequence of the last three spins leading to that CS. For example, for an $r = 3$ reconstructed ϵ -machine, CS S_3 means that 011 were the last three

spins observed before reaching that CS. The *period* of the CSC is the number of CSs that comprise it.

We further divide CSCs into *strong* and *weak* depending on the strengths of the transitions between the CSs that make up the CSC. The *causal-state cycle probability* P_{CSC} is defined as the cumulative probability to complete one loop of a CSC, beginning on the CSC. We identify CSCs with large P_{CSC} as strong CSCs and all others as weak CSCs.

4.1.2. Structural Interpretations We begin by noting that a purely crystalline structure is simply the repetition of a sequence of MLs. This is realized on an ϵ -machine as a CSC with a $P_{CSC} = 1$. That is, an ϵ -machine consisting of a single CSC repeats the same state sequence endlessly, giving a periodic stacking sequence, which physically is some crystal structure. It is therefore useful to catalog all of the possible CSCs on an $r = 3$ ϵ -machine, and this is done in Table 2. There are 19 CSCs on an $r = 3$ ϵ -machine, and each can be thought of as a crystal structure if that CSC is strongly represented. [These should be verified by tracing them out on Fig. 1.]

However, if a nearly perfect crystal has a few randomly inserted stacking errors, these “mistakes” are physically an interruption of the regular ordering of MLs. That is, some error occurs, but after a relatively short distance the crystal returns to its regular stacking rule, thus restoring the crystalline structure. This is realized on an ϵ -machine as a CSC with $P_{CSC}(\text{crystal}) \approx 1$ and another weakly represented CSC with $P_{CSC}(\text{fault}) \ll 1$. In this way, we interpret weakly represented CSCs as faults.

With this understanding in place, we note that an ϵ -machine can quite naturally accommodate more than one crystal structure. Each such CSC must have a $P_{CSC}(\text{crystal}) \approx 1$, but they can be “connected” through a weak CSC, [$P_{CSC}(\text{fault}) \ll 1$]. However, to interpret two CSCs as crystalline structure, each must have $P_{CSC}(\text{crystal}) \approx 1$, and therefore necessarily they *do not* share a CS. [If they did, at least one CSC could not have a $P_{CSC}(\text{crystal}) \approx 1$.] Similarly, ϵ -machines can accommodate more than one faulting structure.

4.2. Faulting Structures on ϵ -Machines

As we did with crystal structures on an $r = 3$ ϵ -machine, it is instructive to identify some of the more common faults on the most general $r = 3$ ϵ -machine. We will consider only 2H and 3C structures with growth, deformation, and layer-displacement faults; but the extension to other crystal and fault structures is straightforward. We will only give the faulting structure on an ϵ -machine to first order in the faulting probability, so that the basic graphical structure is clear. Thus, the connection with the FM is valid only for weak faulting; which is consistent with the FM’s domain of applicability. The ϵ -machine, however, is valid for any degree of disorder, it is only the connection to the FM that is limited to weak faulting.

We also note that the faults on the ϵ -machine in this interpretation are such that the occurrence of two adjacent faults is *suppressed*. If the probability of encountering a fault on a ML is (say) p , then the probability of two adjacent faults is p^2 . That is, in our attempt to use the FM to interpret the structures captured

by an ϵ -machine, we ignore these higher-order terms. Thus, the issue of random versus nonrandom faulting in polytypism is not addressed here. But, again, we note that the ϵ -machine description quite naturally describes random, nonrandom, and periodic faulting structures (Sebastian & Krishna, 1994).

4.2.1. Growth Faults Crystal growth often proceeds by a layer-addition process. Suppose a ML is added that cannot be thought of as a continuation of the previous crystal structure, but the MLs added subsequent to that ML return to the original stacking rule. Such a ML inserted into the sequence is called a *growth fault*. For the 2H structure, the rule is that the added ML has the same orientation as the next-to-last ML. For example, imagine an unfaulted 2H crystal, consisting of A and B MLs, is ...ABABAB.... Then a growth fault in this structure is a B ML followed by a C ML. The remaining MLs continue to follow the 2H stacking rule, giving an overall stacking sequence such as

$$\dots ABABAB \underline{C} BCB CB \dots,$$

where underlining indicates the fault plane.

Notice that the original crystal is composed of alternating A and B MLs, while after the fault it becomes a sequence of alternating B and C MLs. In terms of the Hägg notation, a growth fault for the 2H crystal corresponds to the insertion of a single 0 or 1 into the spin sequence. For example, ...01010101... becomes ...01010101... upon insertion of a 1, where the underlining indicates the inserted spin.

This can be demonstrated on the ϵ -machine shown in Fig. 2 with small faulting probabilities a and b . This ϵ -machine implies that [S_2S_5] is dominant, which is simply the 2H structure. With small faulting probabilities a or b , a 0 or a 1, respectively, is randomly inserted into the 2H crystalline structure. [We do note that it is possible to express growth faults of the 2H structure on a $r = 1$ ϵ -machine. But since most of the samples of practical interest have more structure than growth faults, and hence need a $r = 3$ ϵ -machine, we detail their structure on this larger ϵ -machine.]

In the 3C structure, the stacking rule is that the added ML is different from the previous two MLs. There are, of course, two distinct, symmetry-related 3C structures; one being the ...ABCABC... and the other its spatial inversion ...CBACBA.... The spin sequences for these are ...1111... and ...0000..., respectively. A growth fault for this crystal gives a stacking sequence such as

$$\dots ABCAB \underline{C} BACBA \dots,$$

where underlining again indicates the fault plane. It is conventional to take the indicated ML as the fault plane since it is the only ML in the cubic stacking sequence that is hexagonally related to its neighbors. In terms of Hägg notation, the sequence is ...1111|0000..., where the vertical line indicates the fault plane. The effect of a growth fault in a 3C structure is then to switch from 3C structure of one chirality to another or to flip all spins after the fault plane. This fault is also known as a *twin fault* of the 3C structure, because it produces a crystal containing both kinds of 3C sequences. This growth fault is

international union of crystallography

demonstrated in Fig. 3 with small faulting probabilities a and b . $[S_7]$ and $[S_0]$ correspond to the two twinned 3C structures, with the transition sequences connecting them having a small total probability. [Again, it is possible to express growth faults of the 3C structure on a $r = 1$ ϵ -machine.]

4.2.2. Deformation Faults Other faults can occur after a crystal structure has been formed. Caused by external stresses or inhomogeneous temperature distributions within the crystal, *deformation faults* are the result of one plane in the crystal slipping past another in a direction transverse to the stacking. An example of deformation faulting in the 2H structure is the following:

$$\dots A B A B \mid C A C A C A \dots$$

The vertical bar indicates the plane across which the slip occurred. In the Hägg notation a deformation fault in the 2H structure is realized by flipping a spin. In this example, the unfaulted sequence $\dots 10101010 \dots$ transforms to $\dots 10111010 \dots$, where again the underlined spin demarcates the one flipped. The ϵ -machine representation of this fault is shown in Fig. 4.

In the 3C structure, deformation faults appear much the same. An example of a deformation fault in a 3C structure is

$$\dots A B C A B C \mid B C A B C A \dots$$

The vertical bar again indicates the slip plane. Expressed in spin notation, the unfaulted 3C crystal $\dots 11111111 \dots$ becomes $\dots 11110111 \dots$, a single spin flip. The two corresponding ϵ -machines are shown in Fig. 5.

4.2.3. Layer-Displacement Faults Characterized by a shifting of one or two MLs in the crystal while leaving the remainder of the crystal undisturbed, *layer-displacement faults* do not disrupt the long-range order present in a structure. They are thought to be introduced at high temperatures by diffusion of the atoms through the crystal (Sebastian & Krishna, 1994). In the 2H structure, an example of a layer-displacement fault is:

$$\dots A B A B \underline{C} B A B A \dots,$$

where the underlined ML is the faulted layer. Written as spins, $\dots 10101010 \dots$ becomes $\dots 10110010 \dots$, the underlined spins indicating those that have flipped. A layer-displacement fault in the 2H structure is shown in Fig. 6.

Layer-displacement faults in 3C structures are more difficult to realize, since each ML is sandwiched between two unlike MLs and changing its orientation violates the stacking constraints. It is therefore necessary for two adjacent MLs to shift. Consequently, one expects that they are rarer. An example of layer-displacement in the 3C structure is the following:

$$\dots A B C A B C \underline{B A} C A B C A B \dots,$$

where the underlined layers are faulted. The spin sequence changes from $\dots 11111111 \dots$ to one where three consecutive spins have been flipped to 0: $\dots 11000111 \dots$. A layer-displacement fault on the 3C structure is shown in Fig. 7.

These common faulting structures for the 2H and 3C crystals are given in Table 3 along with the CSCs associated with them.

4.3. ϵ -Machine Decomposition and General Interpretation

The previous discussion has emphasized the important rôle that CSCs play in reflecting stacking structures—crystalline and fault. We have found that CSCs directly correspond to crystal and fault structures. The question then arises, can any arbitrary $r = 3$ ϵ -machine be decomposed into crystal and fault structures? We first note that the *only* difference between fault and crystal structure is the magnitude of the P_{CSC} associated with each CSC. It seems reasonable, then, to break an ϵ -machine into a sum of CSCs. So we formally write,

$$\mathcal{E} \sim \sum_i \eta_i (CSC_i), \quad (24)$$

where $\mathcal{E}$ is the ϵ -machine, ν_i is the fraction of the ϵ -machine that can be attributed to the i^{th} CSC, and CSC_i is the i^{th} CSC. The most general binary $r = 3$ ϵ -machine can be specified by eight variables. It is known, however, that there are 19 CSCs on such an ϵ -machine (Teubner, 1990). So, unless there is a fortuitous vanishing of either CSs or ϵ -machine transitions, or the imposition of additional constraints, the decomposition in Eq. (24) is *not* unique and therefore of questionable use. We note this situation is not expected to improve with larger r . The number of parameters on the $r = r_l$ ϵ -machine grows exponentially in r , while the number of CSCs appears to increase as the exponential of an exponential in r (Teubner, 1990). Thus, in general, it is not possible to decompose an ϵ -machine into CSCs uniquely.

The main purpose of such a decomposition is to provide intuition into the structure present and possibly insight into the physical mechanisms that may have led to a particular structure. In this limited capacity Eq. (24) may be helpful. We stress that Eq. (24) has no other use than this and certainly cannot be used to calculate physical quantities. Only the entire ϵ -machine is suitable for such calculations.

How, then do we intuitively understand structure on an ϵ -machine? In one picture—the weak faulting limit—we view the ϵ -machine as a collection of CSCs. The decomposition given by Eq. (24), while not unique, may be sensible. In this case, we have the same interpretation as the FM. However, the ϵ -machine has a broader range of applicability. It can, for instance, accommodate more than one crystal structure and detail how the stacking alternates between the two. The FM, to our knowledge, admits no such multicrystalline structures. The ϵ -machine provides a more detailed account of multiple faulting structures. The FM, too, can reflect more than one kind of faulting structure, but the description is a stochastic one. The ϵ -machine has no difficulty in reflecting any stacking structures, even closely spaced faulting structures. Thus, we retain the structural interpretations of §4.1.2 for the case of weakly faulted structures.

In other circumstances, when a sensible decomposition of the ϵ -machine into crystal and faulting structures is not possible, it still gives insight into important stacking sequences and their spatial relations. Although it is no longer as advantageous to view the ϵ -machine as a collection of CSCs, we note that stacking sequence probabilities are readily observed on the ϵ -machine either through direct calculation—the ϵ -machine specifies sequence frequencies of any length—or, more simply for

shorter sequences, by asymptotic CS probabilities. The likelihood of seeing two sequences in close proximity can be found by tracing the appropriate path through the ϵ -machine. Since the ϵ -machine is valid for *any* degree of disorder, we can find the relative importance of stacking sequences for even heavily faulted crystals or for crystals in which no regular stacking structures exist. The architecture of the ϵ -machine—*i.e.*, the number, arrangement, and connections between the CSs, then provides an intuitive interpretation for the complexity and organization of the structure. One sees how the various stacking structures are related and how one blends into another upon scanning the crystal.

Thus, in addition to providing a formidable calculational tool, the ϵ -machine provides a new way of viewing structure in layered materials; it is not tethered to the assumption of a parent crystal permeated with weak faults. It gives a generalized way to view and compare the structure of different crystals, even when they have different—or no—parent structures. This should prove especially helpful in understanding solid-state transformations in layered materials (Varn & Crutchfield, 2004).

Finally, we note that these are interpretations of convenience, not necessity. The ϵ -machine is a unique description of the stacking process, and thus any quantities that depend directly on a statistical description of the stacking are amenable to calculation.

For those instances where a sensible decomposition of the ϵ -machine is possible—*i.e.*, the weak faulting limits—we employ Eq. (24) for the limited purpose of providing an intuitive understanding of the disordered structure. We will call η_i either the *fraction* of crystal structure or, for weak CSCs, the *fault density*. This is, of course, different from the *fault probability* generally used in the literature. The fault probability is the frequency, upon scanning the stacking sequence, that one finds a particular fault in the sequence.

5. Summary

We offer here a new theoretical framework for the treatment planar disorder in CPSs, which we call ϵ MSR. In two companion papers we demonstrate its application to diffraction spectra from simulated processes (Varn *et al.*, 2005a) as well diffraction spectra from single crystal ZnS (Varn *et al.*, 2005b). With a minimum of assumptions and using only correlation information between MLs, ϵ MSR infers a statistical description of the stacking structure—in the form of an ϵ -machine. Our description is necessarily statistical, in that we do not find the specific stacking sequence that gave rise to the diffraction pattern, but rather a minimal description of the ensemble of stackings that could have generated the diffraction spectrum. We contend that this statistical description is the most useful form in which to express the structure of the crystal. Indeed, could we have found a specific stacking sequence millions of MLs in length, one still would search for some way to compress this information into a useful form. In short, one would want to find its ϵ -machine.

ϵ MSR offers a number of significant advantages over previous spectral inference techniques. (i) There is no need to

assume an underlying crystal structure. Indeed, ϵ MSR makes no assumption at all about what crystal or fault structure may be present. Unlike competing models, ϵ MSR does not impose a preselected architecture on the model to describe the stacking structure. [Although it is limited to Markov models.] (ii) ϵ MSR is *not* limited to structures that contain only weak faulting. That is, ϵ MSR can be used to describe the stacking for any amount and kind of ordered and disordered sequence that a material may contain. (iii) Thus, ϵ MSR can treat crystals that have more than one crystal or fault structure present. (iv) ϵ MSR uses all of the information in the diffraction spectrum—Bragg and diffuse—instead of considering only the effect disorder has on the Bragg peaks alone. (v) ϵ MSR defines two figures-of-merit— β and γ —that can be used to evaluate the error in experimental diffraction spectra. (vi) ϵ MSR results in the minimal and unique statistical expression of the stacking sequence—the ϵ -machine. (vii) And finally, from the reconstructed ϵ -machine, parameters of physical interest such as the entropy per ML, the statistical complexity, various length parameters and the average stacking-fault energy for disordered stacking sequences are directly calculable (Varn *et al.*, 2005a; Varn *et al.*, 2005b).

We have also examined the structure of the ϵ -machine in the weak faulting limit. We have shown that closed paths on the reconstructed ϵ -machine, the CSCs, correspond to well-known crystalline and fault structures. Specifically we have shown the causal architecture associated with growth, deformation and layer displacement faulting on 2H and 3C crystals.

While we have addressed only CPSs here, the extension to other layered structures is straightforward. A first theoretical task in this is to find an expression for the diffracted intensity in terms of suitable CFs and to relate these CFs to the sequence distribution [and thence to an ϵ -machine]. Significant progress has already been made in this area (Estevez-Rams *et al.*, 2001a). While such an ϵ -machine may draw from an alphabet larger than two for more complicated polytypes, such as micas and kaolins (Varn & Canright, 2001), there are in principle no theoretical obstacles to applying ϵ MSR to more complicated polytypic structures.

Further, ϵ MSR applied to polytypism of CPSs shows the kind of detailed analysis and new insights possible from a computational mechanics treatment of experimental data. It should be possible to adapt ϵ MSR to other physical systems where the measured signal is in the form of a power spectrum. We expect that there will continue to be additional applications of computational mechanics to those areas of physics and crystallography in which one seeks to detect and analyze structural complexity.

Additionally, ϵ MSR also contributes to the machine-learning side of computational mechanics. ϵ MSR is novel, in that we use a power spectrum to reconstruct the ϵ -machine instead of a temporal data sequence, as prior algorithms have. We see this as a prelude to the question of how one infers an ϵ -machine from general spectral data and are continuing research along these lines.

There are, however, some limitations to ϵ MSR, as presented here. We only attempted ϵ -machine reconstruction up to $r = 3$. It has recently been shown that a model for a simple solid

international union of crystallography

state transformation from the 2H to the 3C structure in CPSs results in stacking sequences that imply an infinite memory length (Varn & Crutchfield, 2004). While in principle one can attempt ϵ MSR for any r , there are computational complexity difficulties. In the most general case, the number of variables one needs to solve for is exponential in r , and many of the equations are nonlinear. More seriously, the maximum number of terms in any equation grows as an exponential of an exponential in r . For $r = 3$, there were 11 terms in two of the equations. At $r = 4$, two of the equations have 171 terms, all of them nonlinear. For $r = 5$, this grows to 43690 terms (Varn, 2001). These terms are all additive, so a fortuitous cancellation is not possible. It is possible, however, that physical insight into the relative importance of sequences may allow one to neglect a number of terms in these equations. We feel that the general case of $r = 4$ is tractable, and this is a subject of current research. We also suspect that there are alternative algorithms that will greatly reduce the computational complexity of finding solutions.

Appendix A The Spectral Equations

A.1. $r = 1$

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

$$\begin{aligned} P(11) &= Q_r(2), \\ P(01) &= p(10) = \frac{1}{2}[1 - Q_c(2) - Q_r(2)], \\ P(00) &= Q_c(2). \end{aligned}$$

A.2. $r = 2$

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

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

A.3. $r = 3$

At $r = 3$, we require 16 relations to constrain the length-4 binary-sequence probabilities. Now, however, we encounter nonlinearities, and by necessity the spectral equations are solved numerically. We write them out here.

At $r = 3$, Eq. (7) implies the following seven equations.

$$\begin{aligned} P(0111) &= P(1110), \\ P(0001) &= P(1000), \\ P(0011) + P(1011) &= P(0111) + P(0110), \\ P(0101) + P(1101) &= P(1011) + P(1010), \\ P(0010) + P(1010) &= P(0101) + P(0100), \\ P(0001) + P(1001) &= P(0011) + P(0010), \\ P(0100) + P(1100) &= P(1001) + P(1000). \end{aligned}$$

Equation (8) provides for normalization, providing one additional constraint. Finally, the remaining 8 relations by relating sequence probabilities to CFs as prescribed by Eq. (9). We further reduce the last two relations which involve sequence probabilities of length-5 to those of length-4 via relations of the form given by Eq. (10). We find,

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

Acknowledgements

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

References

- Ashcroft, N. W. & Mermin, N. D. (1976). *Solid State Physics*. Saunders College Publishing.
- Badii, R. & Politi, A. (1997). *Complexity: Hierarchical Structures and Scaling and Physics*, vol. 6 of *Cambridge Nonlinear Science Series*. Cambridge University Press.
- Barrett, C. S. (1950). *Trans. Metall. Soc. AIME*, **188**, 123–135.
- Baumhauer, H. (1912). *Z. Kristallogr.* **50**, 33–39.
- Berliner, R. & Werner, S. (1986). *Phys. Rev. B*, **34**, 3586–3603.
- 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.
- Clarke, R. W., Freeman, M. P. & Watkins, N. W. (2003). *Phys. Rev. E*, **67**, 016203.
- Crutchfield, J. P. (1994). *Physica D*, **75**, 11–54.
- Crutchfield, J. P. & Feldman, D. P. (1997). *Phys. Rev. E*, **55**, R1239–R1242.
- Crutchfield, J. P. & Feldman, D. P. (2003). *Chaos*, **13**, 25–54.
- Crutchfield, J. P. & Young, K. (1989). *Phys. Rev. Lett.* **63**, 105–108.
- Dornberger-Schiff, K. (1972). *Sov. Phys. Crystallogr.* **16**, 1091–1095.
- Engel, G. E. & Needs, R. J. (1990). *J. Phys. Cond. Mat.* **2**, 367–376.
- Estevez-Rams, E., Aragon-Fernandez, B., Fuess, H. & Penton-Madrigal, A. (2003). *Phys. Rev. B*, **68**, 064111.
- Estevez-Rams, E., Martinez, J., Penton-Madrigal, A. & Lora-Serrano, R. (2001a). *Phys. Rev. B*, **63**, 054109.
- Estevez-Rams, E., Penton, A., Martinez-Garcia, J. & Fuess, H. (2005). *Crys. Res. Tech.* **40**, 166–176.
- Estevez-Rams, E., Penton-Madrigal, A. & Martinez-Garcia, R. L.-S. J. (2001b). *J. App. Crystall.* **34**, 730–736.
- Farkas-Jahnke, M. (1973a). *Acta Crystallogr., Sec. B*, **29**, 413–420.
- Farkas-Jahnke, M. (1973b). *Acta Crystallogr., Sec. B*, **29**, 407–413.
- Feldman, D. P. (1998). *Computational Mechanics of Classical Spin Systems*. Ph.D. thesis, University of California, Davis.
- Feldman, D. P. & Crutchfield, J. P. (1998). *Santa Fe Institute Working Paper 98-04-026*.
- Feldman, D. P. & Crutchfield, J. P. (2003). *Phys. Rev. E*, **67**, 051104.
- Frank, F. C. (1951a). *Philos. Mag.* **42**, 1014.
- Frank, F. C. (1951b). *Philos. Mag.* **42**, 809–819.
- Frevel, L. K., Petersen, D. R. & Saha, C. K. (1992). *J. Mater. Sci.* **27**, 1913–1925.
- Frey, F. (1995). *Acta Crystallogr., Sec. B*, **51**, 592–603.
- Frey, F. & Boysen, H. (1981). *Acta Crystallogr., Sec. A*, **37**, 819–826.
- Frey, F., Jagodzinski, H. & Steger, G. (1986). *Bull. Min.* **109**, 117–129.
- Gevers, R. (1954a). *Acta Crystallogr.* **7**, 337–343.
- Gevers, R. (1954b). *Acta Crystallogr.* **7**, 492–494.
- Gonçalves, W. M., Pinto, R. D., Sartorelli, J. C. & de Oliveira, M. J. (1998). *Physica A*, **257**, 385–389.
- Gosk, J. B. (2000). *Crys. Res. Tech.* **35**, 101–116.
- Guinier, A. (1963). *X-Ray Diffraction in Crystals, Imperfect Crystals, and Amorphous Bodies*. W.H. Freeman and Company.
- Hahn, T., Wilson, A. J. C. & Shmueli, U. (eds.) (1992). *International Tables for Crystallography*, 3rd, revised edition. Kluwer Academic publishers.
- Hansen, J. E. (1993). *Computational Mechanics of Cellular Automata*. Ph.D. thesis, University of California, Berkeley.
- Hanson, J. E. & Crutchfield, J. P. (1997). *Physica D*, **103**, 169–189.
- Hendricks, S. & Teller, E. (1942). *J. Chem. Phys.* **10**, 147–167.
- Hirth, J. P. & Lothe, J. (1982). *Theory of Dislocations*. John Wiley, 2nd ed.
- Hordijk, W., Crutchfield, J. P. & Shalizi, C. R. (2001). *Physica D*, **154**, 240–258.
- Jagodzinski, H. (1949a). *Acta Crystallogr.* **2**, 201–207.
- Jagodzinski, H. (1949b). *Acta Crystallogr.* **2**, 208–214.
- Jagodzinski, H. (1954). *Acta Crystallogr.* **7**, 300.
- Jagodzinski, H. (1972). *Sov. Phys. Crystallogr.* **16**, 1081–1090.
- Jagodzinski, H. (1987). *Progress in Crystal Growth and Characterization*, **14**, 47–102.
- Johnson, C. A. (1963). *Acta Crystallogr.* **16**, 490–497.
- Kabra, V. K. & Pandey, D. (1995). *Acta Crystallogr. Sec. A*, **51**, 329–335.
- Kanter, I., Frydman, A. & Ater, A. (2005). *Europhys. Lett.* **69**, 798–804.
- Keen, D. A. & McGreevy, R. L. (1990). *Nature*, **344**, 423–425.
- Kittel, C. (1996). *Introduction to Solid State Physics*. John Wiley & Sons, 7th ed.
- Krishna, P. & Marshall, R. C. (1971a). *J. Cryst. Growth*, **11**, 147–150.
- Krishna, P. & Marshall, R. C. (1971b). *J. Cryst. Growth*, **9**, 319–325.
- Landau, L. (1937). *Phys. Z. Sowjetunion*, **12**, 579.
- von Laue, M. (1918). *Ann. Phys., Lpz.* **56**, 497–506.
- Lifschitz, I. M. (1937). *Phys. Z. Sowjetunion*, **12**, 623.
- Michalski, E. (1988). *Acta Crystallogr. A*, **44**, 640–649.
- Michalski, E., Kaczmarek, S. & Demianiuk, M. (1988). *Acta Crystallogr. A*, **44**, 650–657.
- Milburn, G. H. W. (1973). *X-ray Crystallography: An Introduction to the Theory and Practice of Single-crystal Structure Analysis*. Butterworth & Company.
- Palmer, A. J., Fairall, C. W. & Brewer, W. A. (2000). *IEEE Trans. Geosci. Remote Sens.* **38**, 2056–2063.
- Palosz, B. & Przedmojski, J. (1976). *Acta Crystallogr., Sec. A*, **32**, 409–411.
- Pandey, D. (1989). *Phase Transitions*, **16/17**, 247–261.
- Pandey, D. & Krishna, P. (1976). *Acta Crystallogr., Sec. A*, **32**, 488–492.
- Pandey, D. & Krishna, P. (1977). *J. Phys. D*, **10**, 2057–2068.
- Pandey, D. & Krishna, P. (1982). In *Current Topics in Materials Science*, edited by E. Kaldis. North-Holland.
- Pandey, D. & Lele, S. (1986a). *Acta Metall.* **34**, 405–413.
- Pandey, D. & Lele, S. (1986b). *Acta Metall.* **34**, 415–424.
- Pandey, D., Lele, S. & Krishna, P. (1980a). *Proc. R. Soc. London Ser. A*, **369**, 435–449.
- Pandey, D., Lele, S. & Krishna, P. (1980b). *Proc. R. Soc. London Ser. A*, **369**, 451–461.
- Pandey, D., Lele, S. & Krishna, P. (1980c). *Proc. R. Soc. London Ser. A*, **369**, 463–477.
- Pandey, D., Prasad, L., Lele, S. & Gauthier, J. P. (1987). *J. Appl. Crystallogr.* **34**, 415–424.
- Paterson, M. S. (1952). *J. Appl. Phys.* **23**(8), 805–811.
- Prasad, B. & Lele, S. (1970). *Acta Crystallogr., Sec. A*, **26**, 54–64.
- Proffen, T. & Welberry, T. R. (1998). *Phase Transitions*, **67**, 373–397.
- Schulz, H. (1982). In *Current Topics in Materials Science*, edited by E. Kaldis, vol. 8. Amsterdam: North-Holland.
- 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. (1987a). *Phys. Stat. Sol. A*, **101**, 329–337.
- Sebastian, M. T. & Krishna, P. (1987b). *Crys. Res. Tech.* **22**, 929–941.

international union of crystallography

Sebastian, M. T. & Krishna, P. (1987c). *Crys. Res. Tech.*, **22**, 1063–1072.

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

Sebastian, M. T., Narayanan, K. & Krishna, P. (1987). *Phys. Stat. Sol. A*, **102**, 241–249.

Sebastian, M. T., Pandey, D. & Krishna, P. (1982). *Phys. Stat. Sol. A*, **71**, 633–640.

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

Shalizi, C. R., Shalizi, K. L. & Crutchfield, J. P. (2002). *Santa Fe Institute Working Paper 02-10-060*.

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.

Teubner, M. (1990). *Physica A*, **169**, 407–420.

Treacy, M. M. J., Newsam, J. M. & Deem, M. W. (1991). *Proc. R. Soc. London, Ser. A*, **433**, 499–520.

Trigunayat, G. C. (1991). *Solid State Ionics*, **48**(1/2), 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. A*.

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.

Welberry, T. R. (1985). *Rep. Prog. Phys.* **48**, 1543–1593.

Wilson, A. J. C. (1942). *Proc. R. Soc. Ser. A*, **180**, 277–285.

Woolfson, M. M. (1997). *An Introduction to X-ray Crystallography*. Cambridge.

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

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

Young, K. (1991). *The Grammar and Statistical Mechanics of Complex Physical Systems*. Ph.D. thesis, University of California, Santa Cruz.

Young, K., Chen, Y., Kornak, J., Matson, G. B. & Schuff, N. (2005). *Phys. Rev. Lett.* **94**, 098701.

Table 1

The ϵ MSR algorithm. Here ω^r signifies the set of length- r sequences.

1. Find the CFs from the diffraction spectrum.
 - 1a. Correct the spectrum for any experimental factors.
 - 1b. Calculate the figures-of-merit (§3.6) over possible l -intervals to find an interval suitable for ϵ -machine reconstruction.
 - 1c. Find the CFs over this interval.
 - 1d. Estimate the spectral error $\mathcal{R}_{err}$ from the diffraction spectrum.
2. Estimate stacking distribution $P(\omega^r)$ from CFs.
 - 2a. Set $r = 1$.
 - 2b. Solve the spectral equations for $P(\omega^r)$.
3. Reconstruct the ϵ -machine from the $P(\omega^r)$.
 - 3a. Label candidate CSs by their length- r histories.
 - 3b. Estimate transition probabilities between states from sequence probabilities.
 - 3c. Merge histories with equivalent futures to form CSs.
4. Generate CFs and diffraction spectrum from the ϵ -machine.
5. Calculate the error $\Gamma(r) = |\mathcal{R} - \mathcal{R}_{err}|$ between the original and ϵ -machine spectra:
 - 5a. If $\Gamma(r) \geq \Gamma$, replace r with $r + 1$ and go to step 2b;
 - 5b. Otherwise, stop.

Table 2

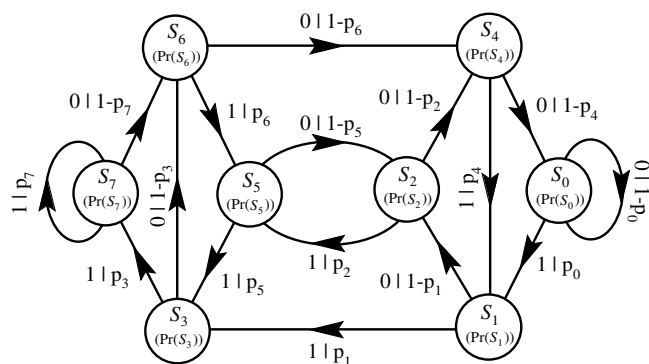
The 19 CSCs on an $r = 3$ ϵ -machine. In the first column, we give the CSC and in the second we show the stacking sequence in the Hägg notation implied by this CSC. If these CSCs are strongly represented on the ϵ -machine, then we can interpret them as crystal structure. The corresponding crystal structures are shown in the third column in the Ramsdell notation. Some CSCs come in pairs related by spin-inversion symmetry, (Varn & Canright, 2001) *i.e.* $[S_0]$ and $[S_7]$ are both 3C structure, differing only in chirality. In cases where the Ramsdell notation is identical for different structures, we have attached a subscript to distinguish them. We list the period-8 hexagonal structures with a subscript to differentiate them from the more common 8H structure (00001111). One must perform ϵ MSR at $r = 4$ to discover this latter 8H structure.

$[S_0]$	(0)*	3C
$[S_7]$	(1)*	3C
$[S_2S_5]$	(01)*	2H
$[S_1S_3S_6S_4]$	(0011)*	4H
$[S_1S_3S_7S_6S_4S_0]$	(000111)*	6H
$[S_5S_2S_4S_1S_3S_7]$	(001101)*	6H _a
$[S_2S_5S_3S_7S_4S_1]$	(110010)*	6H _a
$[S_5S_2S_4S_0S_1S_3S_7S_6]$	(00011101)*	8H _a
$[S_2S_5S_3S_7S_6S_4S_0S_1]$	(11100010)*	8H _a
$[S_3S_6S_5]$	(011)*	9R
$[S_4S_1S_2]$	(100)*	9R
$[S_7S_6S_5S_3]$	(0111)*	12R
$[S_0S_1S_2S_4]$	(1000)*	12R
$[S_3S_6S_4S_0S_1]$	(00011)*	15R
$[S_4S_1S_3S_7S_6]$	(11100)*	15R
$[S_5S_2S_4S_0S_1S_3S_6]$	(0001101)*	21R _a
$[S_2S_5S_3S_7S_6S_4S_1]$	(1110010)*	21R _a
$[S_3S_6S_4S_0S_1S_2S_5]$	(0001011)*	21R _b
$[S_4S_1S_3S_7S_6S_5S_2]$	(1110100)*	21R _b

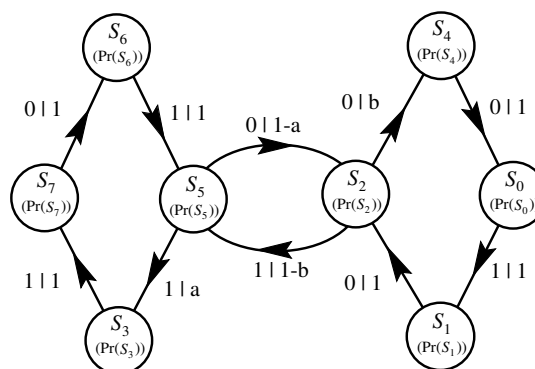
The more common fault structures for the 2H and 3C structures on an $r = 3$ ϵ -machine. We make the following interpretation: If there is one parent structure (2H or 3C) that is strongly represented and a single additional CSC is associated with it as shown above, then we say that that crystal has the structure of the parent crystal with a certain amount of the given faulting. We should be clear here, however, not to confuse structure with mechanism. In this interpretation, the ϵ -machine gives the structure that a crystal would have *if* it experienced a small amount of the faulting given. Structure does not necessarily imply mechanism.

		S_5	S_6	S_7	S_8	S_9	S_{10}
2H	Growth fault	$S_5 S_3 S_6$					
		$S_2 S_4 S_1$					
	Deformation fault	$S_5 S_3 S_7 S_6$					
		$S_2 S_4 S_0 S_1$					
	Layer-displacement fault	$S_5 S_3 S_6 S_4 S_1 S_2$					
		$S_2 S_4 S_1 S_3 S_6 S_5$					
3C	Growth fault	$S_7 S_6 S_4 S_0$					
		$S_0 S_1 S_3 S_7$					
	Deformation fault	$S_7 S_6 S_5 S_3$					
		$S_0 S_1 S_2 S_4$					
	Layer-displacement fault	$S_7 S_6 S_4 S_0 S_1 S_3$					
		$S_0 S_1 S_3 S_7 S_6 S_4$					

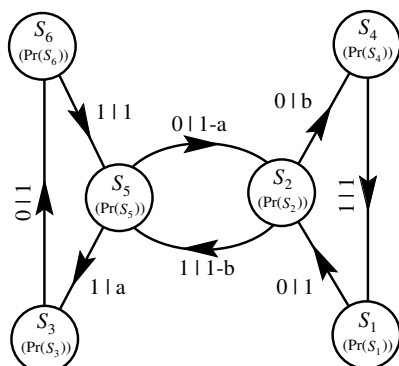
Growth faults for the 3C structure on a $r = 3$ ϵ -machine with small faulting probabilities a and b . $[S_7]$ and $[S_0]$ give the twinned 3C structure, while the paths $S_7S_6S_4S_0$ and $S_0S_1S_3S_7$ give the faulting. Here we have an example of a faulting structure that induces a transition between two crystal structures.



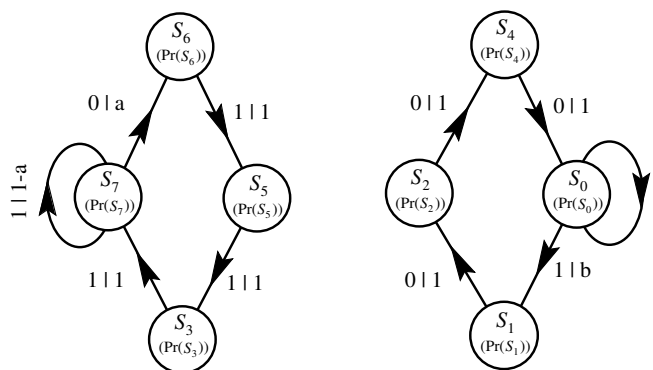
The most general $r = 3$ ϵ -machine. We show only the recurrent portion of the ϵ -machine, as the transient part is not physically relevant (at this stage). The CSs are labeled by the last three spins seen, *i.e.* S_5 means that 101 were the last three spins seen. The numbers in parentheses are the asymptotic CS probabilities. The edge label $s|p$ indicates a transition on spin s with probability p .



Deformation faulting for the 2H structure on a $r = 3$ ϵ -machine for small fault probabilities a and b . A deformation fault is represented by a single spin flip.



Growth faults for the 2H structure as they appear on a $r = 3$ ϵ -machine for small faulting probabilities a and b . $\{S_2S_5\}$ is the 2H structure and $\{S_5S_3S_6\}$ and $\{S_2S_4S_1\}$ give the faulting. This interpretation is only valid for small faulting probabilities.



The two possible ϵ -machines for deformation faulting in the 3C structure with small faulting probabilities a and b . There are two ϵ -machines, one for faulting from each of the 3C structures, $[S_7]$ and $[S_0]$. They are disconnected, and hence faulting in the 3C structure of one chirality cannot cause the crystal to switch to another chirality. Hence this faulting mechanism does not cause twinning.

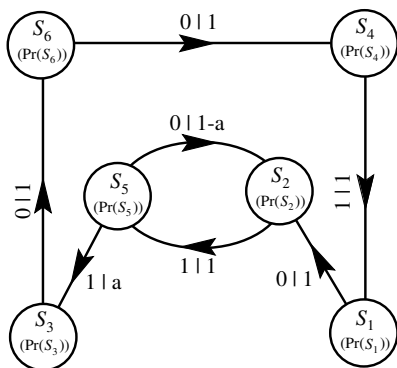


Figure 6

Layer-displacement faults for the 2H structure on a $r = 3$ ϵ -machine with small faulting probability a . Here, for the sake of clarity, we show only faulting initiating from S_5 , although a similar faulting structure can begin from S_2 .

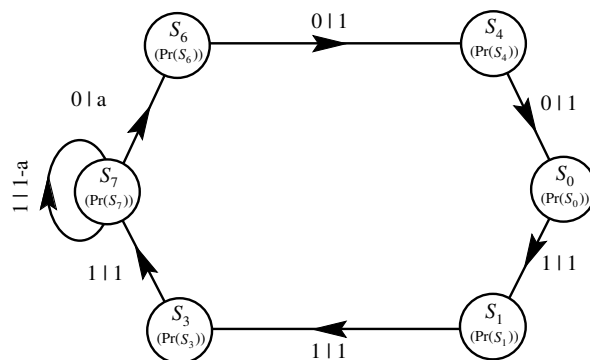


Figure 7

Layer-displacement faults for the 3C structure on a $r = 3$ ϵ -machine with small faulting probability a . Again, we show only faulting from S_7 , corresponding to the ...1111... spin sequence. A similar fault structure could be drawn for faulting from S_0 . A layer-displacement fault for the 3C structure is achieved by three consecutive spin flips. The difference between the fault structure here and that of a growth fault for the 3C structure is that growth faults produce twinning, whereas here the fault returns to the original crystal structure.