Detection and Visualization of Anomalous Structures in Molecular Dynamics Simulation Data

Sameep Mehta*
Computer Science and Engineering
The Ohio State University

Raghu Machiraju*
Computer Science and Engineering
The Ohio State University

Srini Parthasarathy *
Computer Science and Engineering
The Ohio State University

Kaden Hazzard †
Department of Physics
The Ohio State University

John Wilkins†
Department of Physics
The Ohio State University

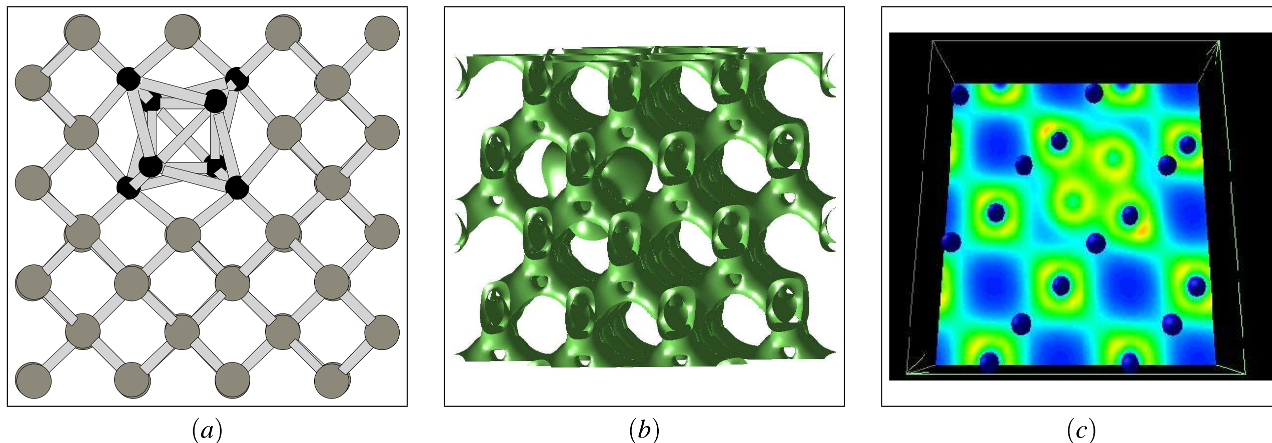


Figure 1: Si System with an interstitial defect (dataset I1) (a) Lattice with bulk and defect (b) Salient Iso-surface with ability to distinguish between bulk and defect (c) An arbitrary slice of electron density data showing the shape of defect

Abstract

In this article we explore techniques to detect and visualize features in data from molecular dynamics (MD) simulations. Although the techniques proposed are general, we focus on silicon (Si) atom systems . These systems are studied to understand the processes behind the formation of point defects. Point and extended defects (derived from point defects) have an impact on the electrical and mechanical properties of silicon. We examine datasets generated from *ab-initio* simulations of atom systems. The data has both the spatial location of atoms and additionally the electron density at sampled points in lattice.

The first set of methods use 3D location of atoms. Defects are detected and categorized using local operators and statistical modeling. Our second set of exploratory techniques employ electron density data. This data is visualized to glean the defects visually. We describe techniques to automatically detect the salient iso-values for iso-surface extraction and designing transfer functions.We compare and contrast the results obtained from both sources of data. Essentially, we find that the methods of defect (feature) detection are at least as robust those based on the exploration of electron density for Si systems.

Keywords: Feature Extraction, Scientific Data Visualization, Data Mining, iso-surface, Transfer Functions, Molecular Dynamics

*{mehtas,raghu,srini}@cis.ohio-state.edu
†{hazzard,wilkins}@pacific.mps.ohio-state.edu

1 Introduction

Scientific data analysis range from analyzing biological data to analyzing geophysical datasets, from analyzing fluid flows to analyzing astrophysical observations . In this article we focus our attention to datasets produced by Molecular Dynamics (MD) simulation. We seek to understand defect dynamics in Si lattices. In semiconductor devices, defects can alter electrical and material properties of the device dramatically. In laser diodes, for example, defects can lead to dark current which reduces device efficiency or even causes device failure. The effect of defects are also extremely important in device fabrication. In the front-end processing, extended defects dissolve to create small interstitial defect clusters, which enhance the diffusion of dopant such as boron by three orders of magnitude (not a desirable effect). Thus, presence of defects is one of the limiting factors in device fabrication. Therefore, to precisely control the distribution of a dopant it is important to understand of the extent and evolution of interstitial defects clusters.

Datasets produced by MD simulation are often very large which impedes easy understanding. Systematic study of defects can produce huge amount of data. In typical silicon defect simulations, more than 120 million time steps are generated to study the evolution of single- or di-interstitial in a lattice [7]. Manual analysis to seek point defects is cumbersome and error-prone. The other factor which limits human capabilities to deduce useful information is the presence of thermal noise and the resulting uncertainty. Uncer-

tainty is inherent in almost all MD simulation data given round-off and associated measurement errors. This is especially true of methods which estimate locations of atoms in a defect ensemble.

Retrieving useful information and drawing conclusions from such a large scale simulation requires efficient and reliable feature mining methods to search and verify defects generated in the simulation. We describe some of those methods in this paper. These feature mining techniques can not only uncover fundamental defect nucleation and growth processes but also provide essential parameters towards the modeling of macroscopic properties of materials. This need is well recognized in the semiconductor industry as evinced in its silicon road map that identifies the short- and long-range problems necessary to continually pack more transistors on a chip.

In this article we present two techniques to explore features in Si lattices. The features we are interested are defects. The data we examine is derived from a single simulation exercise based on *ab-initio* calculations for various systems. The data is composed of two parts namely i) the spatial location of atoms and ii) electron density in regularly partitioned lattices. We deploy two sets of methods to gain understanding about the inherent defects in the Si systems under scrutiny.

The first method relies on the domain knowledge and statistical modeling to locate atoms which constitute the defects. Local operators are proposed after analyzing distribution of bond angles and bond lengths. These measurements are replete with uncertainty. Hence, there is a need to verify or validate the results. Often physicists use only the density data to visualize the atoms and anomalies in the bulk. Whereas, in reality they really need the location and the configuration of atoms. *The premise of this paper is therefore to demonstrate the utility of visualization techniques in validating the feature detection exercise*. However as we shall soon show, visualization without suitable analytic tools will not suffice. We describe appropriate analysis of data to glean useful information. We expect to see the defect in same spatial position using both techniques. Essentially we wish to gain confidence in the feature detection exercise and hence resort to visualizing the results. Figure 1 shows the lattice, iso-surface and slicing results. We explain these images later in the article.

The contributions of this application case study are the following

1. **Detection of defects based on domain knowledge and statistical modeling.**
2. **Use of visualization tools for verification of defect detection process.**
3. **Determination of salient iso-values that best describe the defect.**

We use three datasets depicting the presence of three distinct defects. Each dataset has a 67-atom lattice configuration. The lattice is partitioned in 112*112*112 regular grid. Electron density is calculated at each grid point using the VASP suite [12]. The first and second dataset each have a single tri-interstitial defect, we refer to them as **I1** and **I2** respectively. The defects are of different type and have different shapes. The third dataset has two defects in it, referred as **D1**.

The paper is structured as follows. Related work is discussed in Section 2. Section 3 provides an overview of MD simulations, while Section 4 explains the generation of above mentioned local operators. Visualization techniques developed for MD simulations are described in Section 5. Finally, in Section 6 we summarize our findings and describe our plans for the future.

2 Previous Work

Traditionally physicists have used ground state energy and electrostatic potential to find defects in lattice. [18, 6] use ab-initio methods to locate interstitial defects in silicon lattice. These methods exploit anomalies in the energy/potential fields available at all points in the lattice . The calculation and analysis of these energies/potential is very time consuming. However the most relevant work is embodied in an approach called common neighbor analysis (CNA) [5, 10]. CNA strives to understand the crystallization structures in lattices. CNA as the name implies takes into account the number of neighbors of each atom and analyzes the data. However the effectiveness of this approach is limited by the fact that number of neighbors alone cannot capture all (geometrical) properties especially at high temperatures Moreover, the CNA approach does not account for noise effects in such data. Machiraju et.al. [14, 15] presented a framework for feature detection and classification of data from simulations. We use the same framework to process MD simulation data. The framework emphasizes on the use of shape and structure of the features (defects) for classification and tracking. Traditional visualization techniques have relied on generating video clips that depict the animated movement of atoms in a given system. [3] proposed an MPEG-based method to visualize and generate the animations for MD simulation. Levoy proposed the use of special transfer functions to visualize molecular ensembles [13]. Recently newer techniques that do not necessarily rely on animation alone have been proposed to visualize the 3D atomic data. Notable work includes [2, 1, 4]. Other efforts have targeted specific atomic and molecular systems. [17] proposed a method to visualize biochemical data. Additionally, several software application toolkits also have been offered for use by researchers. Visual Molecular Dynamics (VMD) [9, 16][1] is one such software package that allows the manipulation and visualization of atoms in real time. However visualization alone can not uncover important features. Hence, there is a strong need to couple visualization techniques with data analysis.

3 Background

The key complexity of real materials for commercial applications is not that they are defected in the trivial sense of being imperfect or impure, but rather that their material properties depend critically on their nonideality. As an example, the enhanced diffusion of dopants in the presence of extended $\{311\}$ defects[2] in silicon is a limiting factor in the fabrication of shallow junction devices [8]. Our objective is to mine such datasets to aid in the discovery of rules that govern nucleation and defect growth. To do so we must effectively detect and visualize defects. We now define some relevant aspects of MD simulations.

Lattice: A lattice is an arrangement of points or particles or objects in a regular periodic pattern in 3 dimensions. Consider the simple silicon lattice in Figure 2a. The "atoms", are denoted by circles, stabilized by "bonds" denoted by cylinders connecting the atoms. The bonds strive to preserve both the lattice spacing of 2.36 Angstroms between atoms and the dihedral angles of 109.28 degrees. It should be noted that only a portion of the "infinite" lattice is shown. A perfect or ideal lattice is defined to be composed of "bulk" atoms.

[1] http://www.ks.uiuc.edu/Research/vmd/

[2] Point defect which evolves to extend through the lattice.

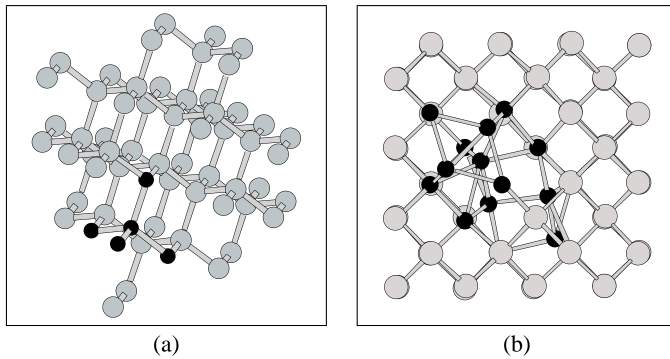


Figure 2: (a) Original lattice with repeating structure marked (b) Lattice with one tri-interstitial defect

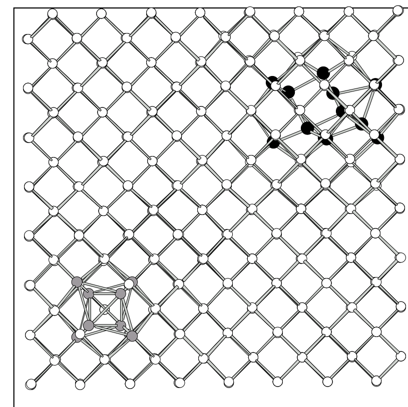


Figure 3: Two separate defects in same lattice

Repeating Structure: A repeating structure is set of atoms which are repeated in preferred directions to form the bulk lattice. This structure for Si system is composed of 5 atoms organized as a tetrahedron, with 4 atoms connected to a single central atom with bond length of 2.36 Angstroms and dihedral angle of 109.28 degrees. Figure 2a shows a bulk lattice with a specific structure shaded differently.

Defect: Consider adding single atom to the lattice. The bonds between the extra atom and other atoms strive to satisfy the lattice parameters as described above. Figure 2b shows one such lattice with one extra atom. Note that the dark shaded atoms do not follow the regular tetrahedral structure. While this defect no longer has the symmetry of the original simple square lattice, we can upon visual inspection of, recognize the atoms that seem to be displaced as “defect” atoms.

Multi-defects lattice: The real challenge arise with multiple defects in a lattice that can then form even more extended defects. Consider for simplicity two extra interstitial defects added to the crystal that form disconnected defects. Figure 3 illustrates two possible defects: in the lower left and upper right corners respectively of a 512-atom lattice. The different shades again represent separate and distinct defects.

Electron Density: The location of an electron is not fixed, but is instead described by a probability density function. The sum of the probability densities of all the electrons in a region is the electron density in that region. The density function describes the probability of finding an electron around atom. This data is generated by using **V**ienna **A**b-intio **S**imulation **P**ackage (VASP) [12, 11] [3]. VASP is a package for performing ab-initio quantum-mechanical molecular dynamics (MD) through the solution of integral equations. The lattice is partitioned to a regular grid and electron density is calculated at each grid point.

4 Defect Detection using Local Operators

In this section we describe defect detection based upon statistical modeling. The idea is intuitive and simple: first identify the bulk atoms; the rest are defect atoms. Identification of bulk atom can be done by checking the bond lengths and bond angles it forms with other atoms. From existing literature in material sciences, it is easy to obtain these rules. These simple rules will work in the case of noise-free lattices. However this is not the case when simulation is conducted at higher temperatures. In the next section we describe the methods to model the noise in discovery of rules.

[3] http://cms.mpi.univie.ac.at/vasp

As noted above, these precise rules cannot be directly used for formulating rules to define the defect given the noise. To glean the rules, we generate a histogram of the bond angles and bond lengths of several silicon lattices. As shown in Figure 4 the distribution follows the Normal Distribution with a mean very close to the ideal values of bond angle and bond length as described in Section 3. Since the defect atoms are relatively rare, we can consider them to be outliers. A simple method for detecting outliers within a normal distribution is to to use the 95% two sided confidence interval of the distribution. Under normal distribution 95% of data lies between $\pm 2 * \sigma$, therefore we obtain the following two relaxed rules for silicon [14]:

1. **R1**: All bond angles with neighbors should lie in the interval 90-130 degrees.
2. **R2**: Each atom should form exactly 4 bonds with a bond length $\leq$ 2.6 Angstrom. Unlike **R1**, **R2** has only upper limit because two atoms cannot get much closer to each other due to presence of electrostatic forces.

These rules are applied locally to each atom in every frame generated by the simulation data. All the atoms which fails either one of the rules is labeled a defect atom.

4.1 Segmentation of defects

The rules described above are local operators which only mark the defect atoms. However since there can be several defects in a lattice, an additional step is needed to group defect atoms in one or more connected substructures (defects). Input to this stage is a list of atoms along with their locations. We start with one atom at random from this list and identify any neighbors to this atom within a distance of εAngstrom from this atom. Each of these neighbors repeatedly identify its neighbors until there are no additions to the connected substructure. This defect substructure and the list of atoms it is composed of are deleted from the original list of atoms, and is labeled a defect. If there are any more atoms in the original list, this process is repeated until all defects are found.

We use the upper bound of Rule **R2** to determine the value of the parameter ε. The final result of this step does not depend on the choice of the initial atom. Figure 3 shows two detected defects embedded in 512-atom lattice. The different shades again represent

separate and distinct defects. However there is a strong need to verify the results.

5 Defect Verification using Visualization

A simple visualization using standard iso-surface or transfer function techniques is not helpful. It is not clear what the iso-value should be. Also, it is not clear what transfer functions are best. Note the results in Figure 5. We use the linear ramp as the opacity transfer function to render this volume. The position of the defect structure cannot be easily ascertained by visually inspecting the image. Some analysis is required. We now analyze the electron density data to better understand the behavior of this scalar field and visualize the defect structure in a meaningful manner.

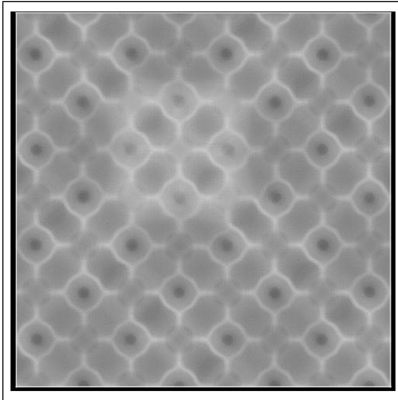


Figure 5: Dataset I1 volume rendered using the common ramp Transfer Function

Levoy [13] proposed the use of transfer functions that vary with distance from the grid points. Let us first examine the variation of electron density to see if there is any trend we can exploit. The electron density is maximum at the actual spatial position of atom and decreases as we move away from the atom towards another atom. At some point the density again starts to increase and reaches maximum at spatial location of other atom. This behavior is well exhibited by any two bulk atoms. However, the electron density around the defect atoms do not follow these rules. This anomalous behavior can be explained considering the fact that density is not a local property between two atoms. It is affected by the presence of neighboring atoms. Since in a defect the positions of atoms does not follow a regular geometry, the density function deviates from the normal behavior. Figure 6a shows the change in density between two bulk atoms and Figure 6b shows the behavior between defect atoms. One can capture this variation of electron density by designing transfer functions that change in both data space and the embedded euclidean space. However, we choose to use a simpler yet effective method for this effort. Our intent is to provide simple yet effective tools that a physicist can use in a tangible manner. Therefore, we choose to determine the salient iso-value that can discriminate between defect and bulk. We now describe an analytic method to detect the defect.

5.1 Iso-value Analysis

Iso-surfacing is a common technique to visualize the data. However the most difficult part is to find the correct iso-value so that the transition point is well captured. The salient iso-value if detected correctly should depict the defect. Once this iso-value is determined, we can use it to extract significant iso-surfaces and even use it for constructing transfer functions. Next we present a method to automatically find the correct iso-value for this problem.

Since there are two surfaces namely bulk and defect present in same volume, there should exist some scalar value where both the surfaces can be seen. Our method tries to find that "special" value by analyzing the distribution(histograms) of the electron density scalar field. We divide the scalar field into N bins. The histogram is first smoothed using an Gaussian kernel of appropriate width. Figure 8(a) shows the original histogram while Figure 8(b) shows smoother version. The smoothed histogram S_H is then transformed into the frequency domain using the fast Fourier transform (FFT) to obtain F_H. Since we wish to retain the high frequency components of the histogram, we construct the following exponential function that serves well as a band-pass filter.

$$G(i) = \exp(-2s^2)/i^2$$

where $i \in [1 \dots N]$ and s is constant scaling factor in frequency domain

F_H is then convolved with G to obtain C_H. This convolution amplifies the high frequency component. An inverse Fourier transform (IFFT) is then applied to C_H to obtained a highly enhanced histogram. Figure 8(c) shows the histogram after the inverse FFT.

One should notice the dramatic change in the shape of the histogram. We believe that the values of electron density in the bins spanning the large change include the salient iso-value. The bins selected for inspection are those where the curvature of the histogram is large. Finally, the values in the bins are averaged to get single iso-value.

Please note that the change occurs in bins 15 and 16 and that the average iso-value in these bins is around 450. Figure 9 shows the iso-surface for iso-values less than, equal to and greater than the automatically determined iso-value. It is clear from the rendered images for iso-values less the salient iso-value the defect surfaces are hard to distinguish. Same is true for iso-values greater than the transitioning salient iso-value. However at the thresholded salient iso-value, the defect structure is well separated and easily distinguishable.

5.2 Transfer Functions for Volume Rendering

We now use the derived iso-value to construct appropriate opacity transfer functions. The intuition is that since at some particular iso-value bulk and defect can be distinguished, all the points at that iso-value should have highest opacity and other data values should be assigned lower values of opacity. Therefore, in effect the opacity should be highest at iso-value and then gradually decrease. The relevant question the is 'how should the opacity decrease in the data space ?'. While calculating the iso-value we average all scalar values which lie in bins across which the transition takes place. We assign the opacity for all other grid points to a small constant value. For points corresponding to iso-values in bin(s) our transfer function is a Gaussian with μ = iso-value and σ = standard deviation of all scalar values in the bin(s) spanning the transition. Figure 7a shows the transfer function and Figure 7b shows the volume rendered using these transfer function.

It should be noted that although we did not incorporate any distance-based variation in the transfer functions, the resulting images are quite telling.

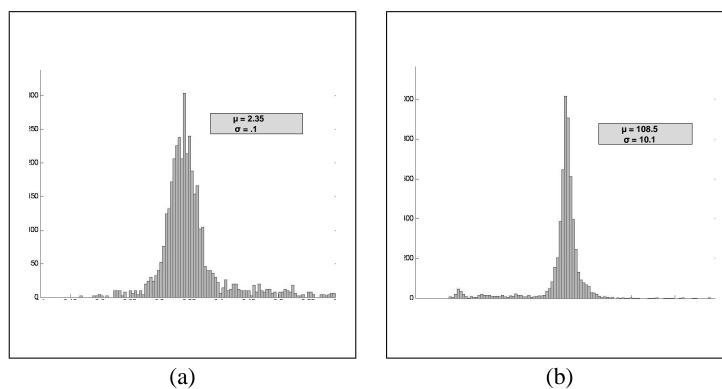


Figure 4: (a) Distribution of bond length (b) Distribution of bond angles

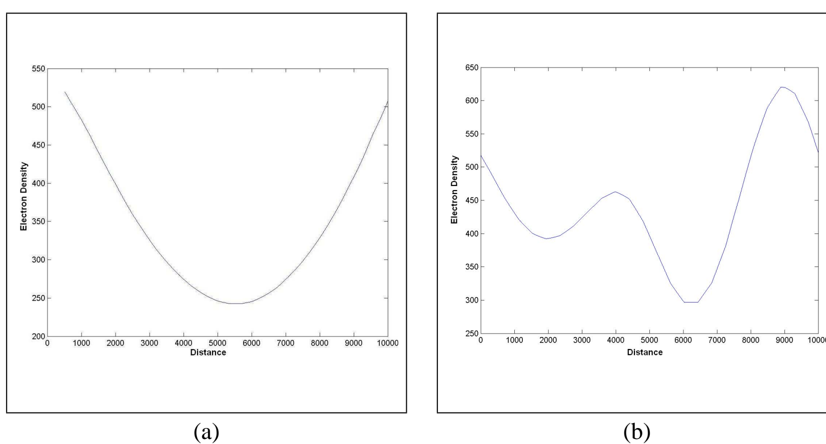


Figure 6: (a) Density behavior between two bulk atoms - First atom is at distance 0 and second atom is at distance 10,000 (b) Density behavior between two defect atoms

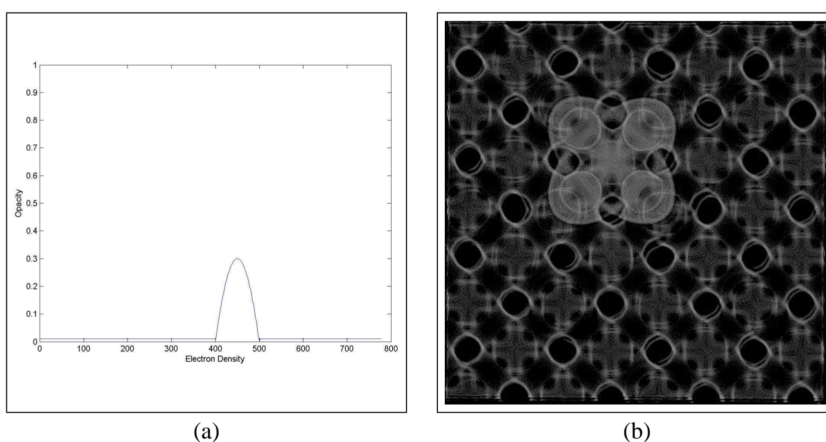


Figure 7: Dataset I1 (a) Transfer Function derived using iso-value (b) Volume Rendered Image using the transfer function

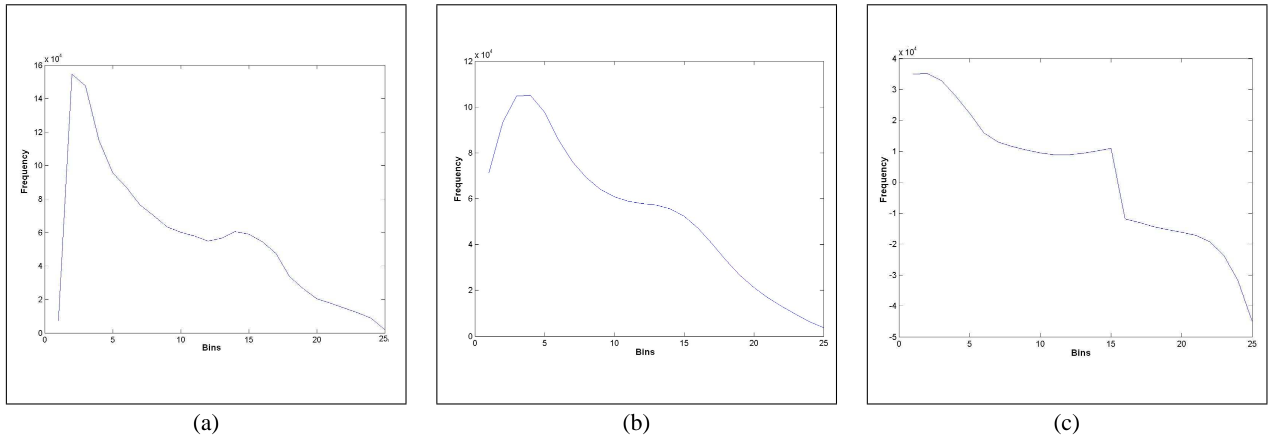


Figure 8: (a) Original Distribution (b) Smoothed Distribution (c) Band Pass Distribution

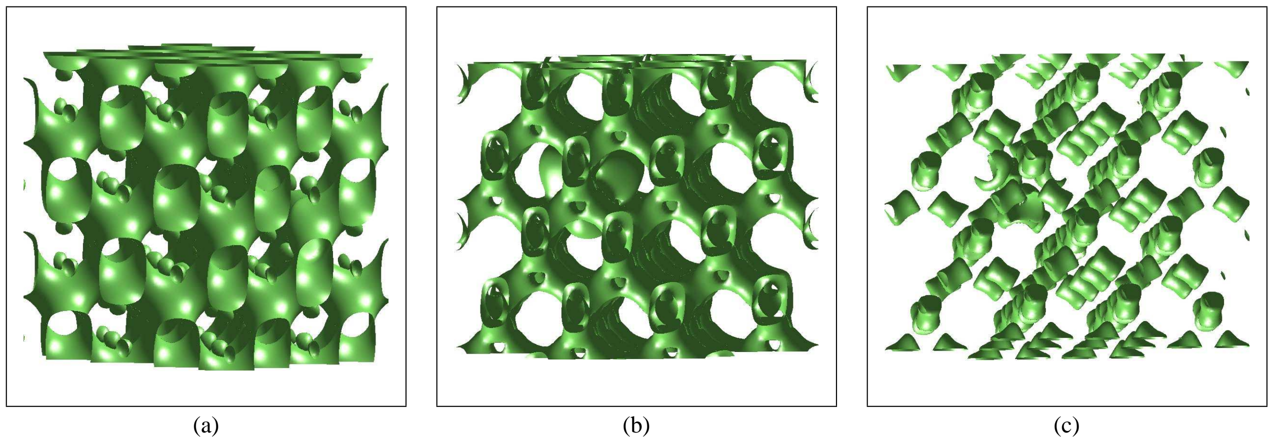


Figure 9: Dataset I1 (a) iso-surface before the transition point (b) At the transition point (c) After the transition Point

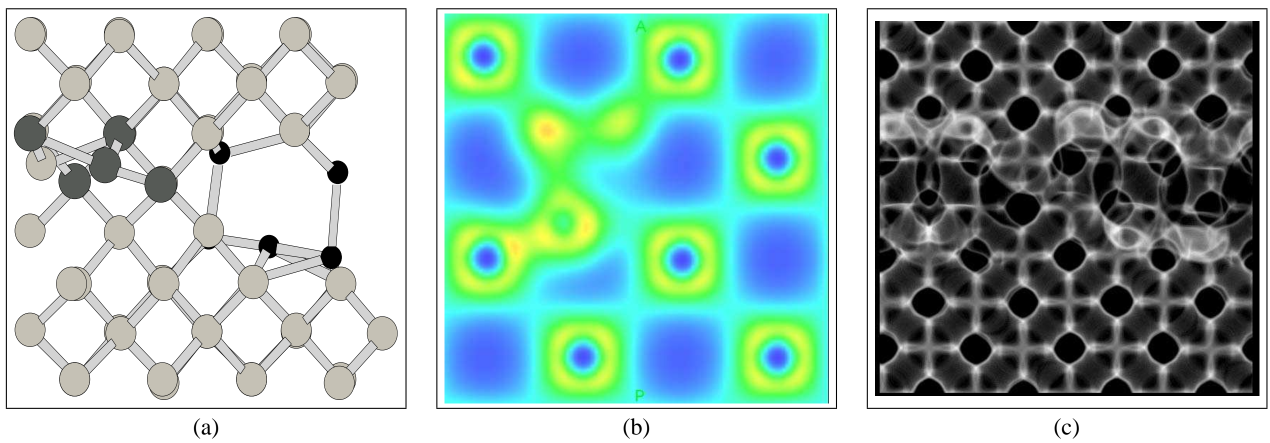
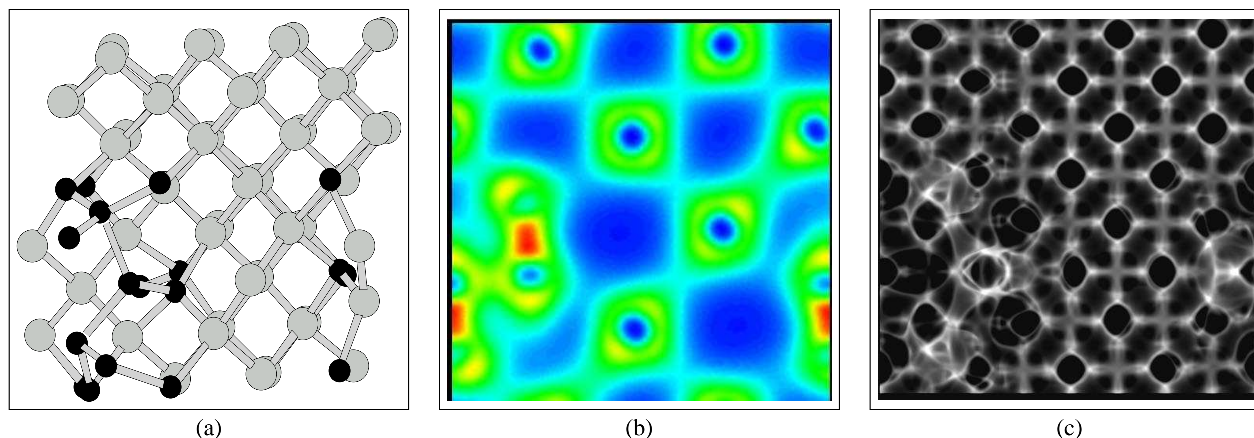


Figure 10: Dataset D1 (a) Original Marked Lattice with two defects (b) Slicing showing shape of one defect (c) volume rendering showing both defects.



(a) (b) (c)

Figure 11: Dataset I2 (a) Original marked lattice (b) Slicing (c) Volume rendering

6 Discussion

Figure 1a shows the detected defect in dataset I1. Figure 1b and c shows the iso-surface at the salient iso-value and an arbitrary slice for our dataset I1. Figure 7 shows the transfer function and volume rendered using that transfer function for the same dataset. Figure 11 shows same three results for our second dataset I2. Please note that the defect structure is split at boundary in I2. Our approach takes care of boundary conditions by wrapping the defect, however we can-not wrap the electron density data. Therefore for visualization purposes we are not taking the boundaries in account.

Figure 10a shows the original dataset D1 lattice with defects atoms marked. Figure 10b and c show the slicing and volume rendering results for D1 respectively. The slice only shows one defect, the other defect is visible at other slicing angle. All these visualizations are done for electron density data of same lattice. The defects can be seen at very easily in volume rendering. In this case also the transfer function is constructed based on the derived iso-value.

In all three datasets our local operators are able to correctly locate defects. Also multiple defects are correctly segmented. Also note the shape of the defects. The shape of the detected defects is very similar to the shapes gleaned on images obtained from a slicing operation and volume rendering. The similarity is extremely high for our first dataset. This can be explained by fact that the defect in our first data is very compact and well connected. In other two cases defects are bigger but still the shape is well captured. The defects are also located at same spatial locations.

These observations validate that our approach mark correct atoms as defects. Local operators are easy to apply and also time and space complexity is less. A lattice has 67 atoms with each atom represented by x, y and z coordinates. Thus we have 67×3 floating points. However for electron density the size of datasets is $112\times\times 112\times 112$ floats. Also local operators directly give us the defect atoms. Other data sources require pare-processing l(e.g. finding iso-values or the correct slice orientation). However these techniques provide a solid and reliable way for verifying our approach.

For the future we will consider larger Si systems. Moreover, we plan to study evolution of defects in Titanium alloy systems. The unit cell is dramatically different and new defect rules have to be discovered. Additionally, we plan to further investigate the design of transfer functions that vary with distance.

7 Acknowledgments

NSF , David Richie, Kim, Alex, Steve, MCC

References

[1] A. Varshney, F.P. Brooks, Jr., and W.V. Wright. Linearly Scalable Computation of Smooth Molecular Surfaces. In *IEEE Computer Graphics and Applications Vol 14*, 1994.

[2] A. Varshney, F.P. Brooks, Jr., D.C. Richardson, W.V. Wright, and D. Manocha. Defining, Computing, and Visualizing Molecular Interfaces . In *IEEE Visualization*, 1995.

[3] V.L. Bulatov and R.W. Grimes. Visualization of molecular dynamics simulations. *Eurographics UK Chapter*, 1996.

[4] C.H. Lee and A. Varshney. Representing Thermal Vibrations and Uncertainty in Molecular Surfaces. In *SPIE Conference on Visualization and Data Analysis* , 2002.

[5] A.S. Clark and H. Janssen. Structural changes accompanying densification of hard sphere packing. *Physical Review E vol.74 pages 3975-3984*, 1993.

[6] S.J. Clark and G.J. Ackland. Ab initio calculation of the self interstitial in silicon. *Physical Review Letters vol. 56*, 1997.

[7] D.A. Richie, J. Kim, and J.W. Wilkins. Real-time multiresolution analysis for accelerated molecular dynamics simulations. In *American Physical Society March Meeting*, 2001.

[8] N.B. Cowern et.al. *Physical Review Letters 82 4460*, 1999.

[9] W. Humphrey, A. Dalke, and K. Schulten. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics*, 14:33–38, 1996.

[10] H. Janssen and H.C. Anderson. Icosahedral ordering in lennard-zones liquid and glasses. *Physical Review Letters vol.60 pages 2295-2298*, 1988.

[11] G. Kresse and J. Furthmuller. *Physical Review B 54 pages 2295-2298*, 1996.

[12] G. Kresse and J. Hafner. *Physical Review B 47, 558*, 1993.

[13] M. Levoy. Display of surfaces from volume data. *IEEE Computer Graphics and Application Vol 22 No. 8*, 1988.

[14] M. Jiang, T.-S. Choy, S. Mehta, M. Coatney, S. Barr, K. Hazzard, D. Richie, S. Parthasarathy, R. Machiraju, D. Thompson, J. Wilkins, and B. Gatlin. Feature Mining Algorithms for Scientific Data . In *SIAM*, 2003.

[15] R. Machiraju, S. Parthasarathy, J. Wilkins, D. Thompson, B. Gatlin, D. Richie, T. Choy, M. Jiang, S. Mehta, M. Coatney, and S. Barr. Mining of Complex Evolutionary Phenomena, Next Generation Data Mining. In *NGDM*, 2003.

[16] R. Sharma, M. Zeller, V.I. Pavlovic, T.S. Huang, Z. Lo, S. Chu, Y. Zhao, J.C. Phillips and K. Schulten. Speech/gesture interface to a

visual-computing environment. *IEEE Computer Graphics and Application*, 20:29–37, 2000.
[17] S. Prohaskaa and H-C. Hege. Fast Visualization of Plane-Like Structures in Voxel Data. In *IEEE Visualization*, 2002.
[18] R.J. Needs W.K. Leung and G. Rajagopal. Calculation of silicon self interstitial defects. *Physical Review Letters vol. 83*, 1999.