Detection and Visualization of Anomalous Structures in Molecular Dynamics Simulation Data

Sameep Mehta*
Computer Science and Engineering
The Ohio State University

Kaden Hazzard †
Department of Physics
The Ohio State University

Raghu Machiraju*
Computer Science and Engineering
The Ohio State University

Srinivasan Parthasarathy *
Computer Science and Engineering
The Ohio State University

John Wilkins†
Department of Physics
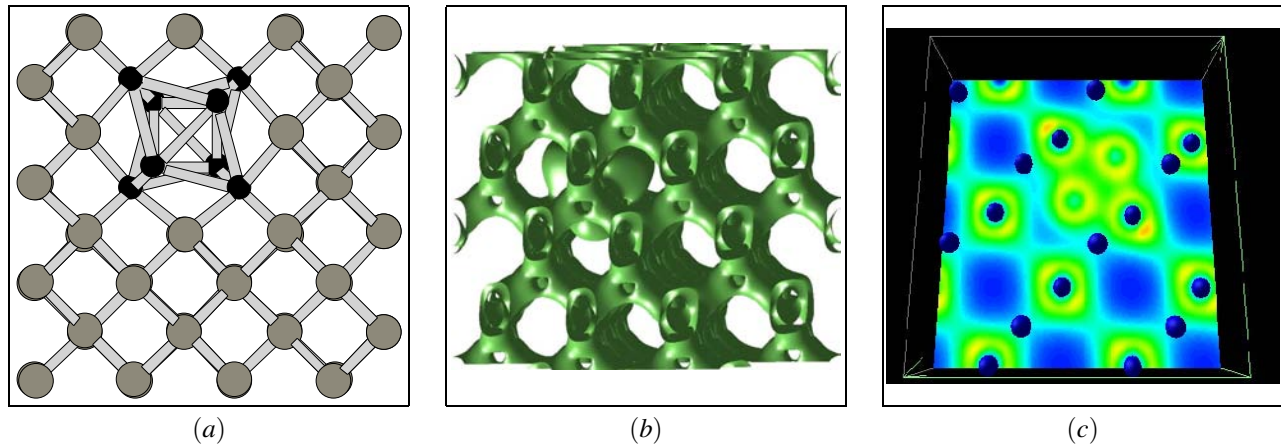The Ohio State University

(a)   (b)   (c)

Figure 1: Si **system** with a **tri-**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) atomic systems. 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. 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 as those based on the exploration of electron density for Si systems.

CR Categories: J.2 [Computer Applications]: Physical Sciences And Engineering —Physics

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

---

*{mehtas,raghu,srini}@cse.ohio-state.edu

†{hazzard,wilkins}@pacific.mps.ohio-state.edu

1 INTRODUCTION

Scientific data analysis ranges from analyzing biological data to analyzing geophysical datasets, from analyzing fluid flows to analyzing astrophysical observations. Direct numerical simulations are also being increasingly used to generate the datasets. These datasets are analyzed to understand the domain. In this article we focus our attention to datasets produced by Molecular Dynamics (MD) simulations. We seek to understand defect dynamics in Si lattices. In semi-conductor devices, defects can alter electrical and material properties of the device dramatically. In laser diodes, for example, defects can lead to dark currents which reduce device efficiency and can even cause device failure. Knowledge of defects and their impact is 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 simulations are often very large which impedes easy understanding. Systematic study of defects can produce a 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 defects in a lattice [8]. 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 uncer-

465

tainty. Uncertainty is inherent in almost all MD simulations 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 a $112 \times 112 \times 112$ regular grid. Electron density is calculated at each grid point using the **V**ienna **A**b-intio **S**imulation **P**ackage (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 in each of the two datasets are of different types and possess very different shapes. On the other hand, the third dataset, **D1**, includes two distinct defect structures.

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 a lattice. In [6, 18] ab-initio methods were used to locate interstitial defects in a silicon lattice. These methods exploit anomalies in the energy/potential fields available at all points in the lattice. The calculation and analysis of electrostatic energies and potential is very time consuming. The most relevant work is embodied in an approach called common neighbor analysis(CNA) [5, 10]. CNA strives to glean the crystallization structures in lattices. As the name CNA implies, the method takes into account the number of neighbors of each atom and analyzes the field 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. In [14, 15], we presented a framework for detection and classification of features in data from fluid and MD simulations. The framework emphasizes on the use of shape and structure of the features (defects) for classification and tracking.

Visualization techniques have relied on generating video clips that depict the animated movement of atoms in a given system. In [3] a MPEG-based method was proposed to visualize and generate the animations for MD simulations. Recently, new techniques that do not necessarily rely on animation alone have been proposed to visualize the 3D atomic data. Varshney and his associates were one of the first to exploit visualization techniques [1, 2] towards understanding the structure and form of molecular structures. Smooth surfaces representing force fields were computed and visualized using contouring methods. On the other hand, Levoy proposed the use of special transfer functions to volume render molecular ensembles [13]. Other efforts have targeted specific atomic and molecular systems. In [17] a method was proposed 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. In [4] statistical analysis of data was conducted to glean various sub-structures. Our work reported here is similar in essence.

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 [7]. 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 2.36Å between atoms and the dihedral angles of 109.28° It should be noted that only a portion of the "infinite" lattice is shown. A perfect or ideal lattice is composed of only "bulk" atoms.

---

[1]http://www.ks.uiuc.edu/Research/vmd/
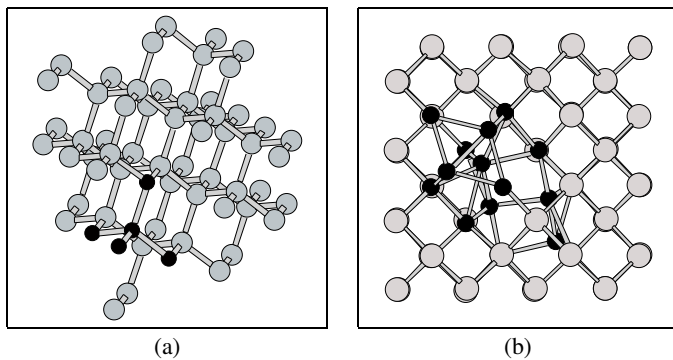[2]Point defect which evolves to extend through the lattice.

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 2.36Å and dihedral angle of 109.28°. Figure 2a shows a bulk lattice with a specific repeating structure shaded differently.

Defect: Consider adding a 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: Real challenges 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 nucleus. This data is generated by using **V**ienna **A**b-intio **S**imulation **P**ackage (VASP) [11, 12] [3]. VASP is a package for performing *ab-initio* quantum-mechanical molecular dynamics (MD) through the solution of integral equations. The lattice is partitioned into 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 rules will work in the case of noise-free lattices. However this is not the case when simulations are conducted at higher temperatures. Next we describe the methods to model the noise in discovery of rules.
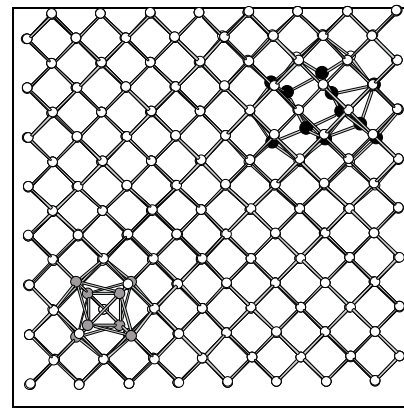
---

[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 with mean values very close to the ideal values of bond angle and length as listed 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 use the 95% two-sided confidence interval of the distribution. Under the 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 $90° - 130°$.

2. **R2**: Each atom should form exactly 4 bonds with a bond length ≤ 2.6Å.

   Unlike **R1**, rule **R2** only imposes an upper bound because two atoms cannot get much closer to each other due to the presence of electrostatic forces.

These rules are applied locally to each atom in every frame generated by the simulation data. All 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 defect 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 εÅ from this atom. Each of these neighbors repeatedly identifies 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** for 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.
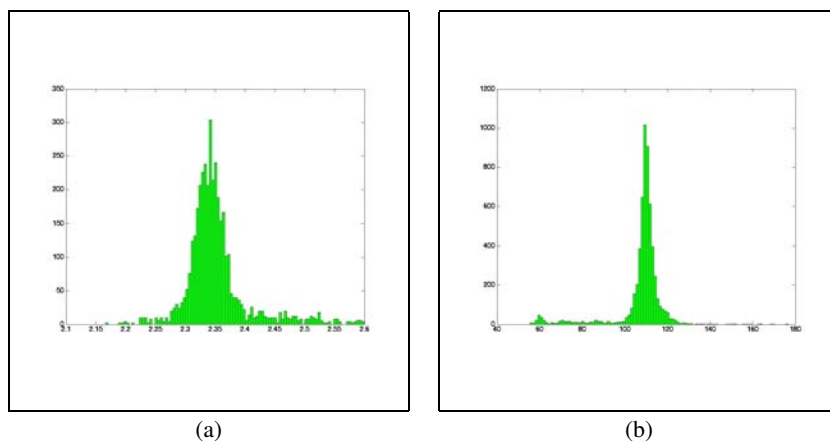
Figure 4: (a) Distribution of bond length (b) Distribution of bond angles

5 DEFECT VERIFICATION USING VISUALIZATION

A simple visualization using standard iso-surface or transfer function techniques is not helpful without some knowledge of the properties of the system being visualized. Note the ambiguous results in Figure 5. We use the standard linear ramp used as opacity transfer function to render this volume. The position of the defect structure cannot be easily ascertained by visually inspecting the image. How does one pick an appropriate iso-value and an appropriate transfer function? We can learn from the data and domain knowledge available to us to better understand the behavior of this scalar field We plan to analyze the electron density to discover the meaningful iso-value(s) and develop the transfer functions for better visualization.
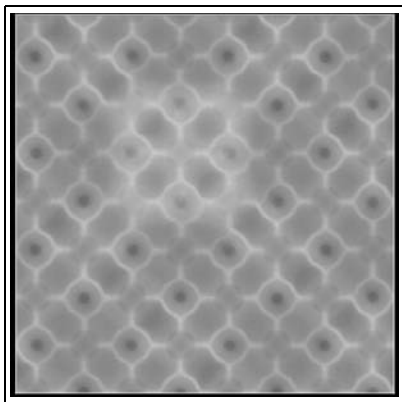

Figure 5: Dataset I1 volume rendered using the common ramp Transfer Function

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 neigh-

boring atoms. Since in a defect the positions of atoms do 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 the 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 7(a) shows the original histogram, while Figure 7(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 $i \in [1 \ldots N]$ 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 is then applied to C_H to obtained a highly enhanced histogram. Figure 7(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 a single iso-value.

Please note that the change occurs in bins 15 and 16 respectively 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 than the salient iso-value that 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 this iso-value and then gradually decrease. The relevant question 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 these bins our transfer function is a Gaussian with μ = iso-value and σ = standard deviation of all scalar values in the bin spanning the transition. Figure 8a shows the transfer function and Figure 8b shows the volume rendered using the 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.

6 Discussion

Figure 1a shows the detected defect in dataset **I1**. Figure 1b and Figure 1c show the iso-surface at the salient iso-value and an arbitrary slice for our dataset **I1**. Figure 8 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 handles the boundary by assuming that the lattice under scrutiny extends periodically in the directions of the Principal Cartesian axes.

Figure 10a shows the original dataset **D1** lattice with defect atoms marked. Figure 10b and Figure 10c depict the results from slicing and volume rendering the dataset **D1** respectively. The slice shows only one defect, the other defect is visible at another slicing angle. All these visualizations are conducted for electron density data of the same lattice. The defects can be seen very easily in volume rendering. Once again, 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. Note the shape of the defects. The shape of the detected defects is very similar to the shapes gleaned from images obtained from a slicing operation and volume rendering . The similarity is extremely high for our first dataset(Figure 1). The defect in our first dataset is very compact and well connected and hence the closeness of the two results. In the other two cases, the defects are larger. However, the shape is well captured by all methods of visualization. Further, the defects are also located at the same spatial locations.

These observations validate our premise that our segmentation approach marks the correct atoms as defects. The MD simulation data offers many advantages to both the physicist and the visualization professional. Local operators are easy to apply and further the time and space-complexity of the necessary operations is small. Consider this - a lattice has 67 atoms with each atom represented by x, y and z coordinates. Thus, there exists 67×3 floating points. However, for the electron density data the size of a collection measures to $112 \times 112 \times 112$ floats. Also, local operators directly provide the defect atoms and not an image. Other data sources require pre-processing (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

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* , pages 19–25, 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*, pages 33–38, 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* , pages 80–90, 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. B56,47*, 1997.

[7] N. E. B. Cowern, M. J. J. Theunissen, F. Roozeboom, and J. G. M. van Berkum. Boride-enhanced diffusion in silicon: Bulk and surface layers. *Applied Physics Letter, 75(2) 181*, 1999.

[8] D. A. Richie, J.N. Kim and J.W. Wilkins . Applications of real-time multiresolution analysis for molecular dynamics simulations of infrequent events. In *Materials Research Society Symposia Proceedings 677, AA5.1-7* , 2001.

[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. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B 54 pages 11169-11186*, 1996.

[12] G. Kresse and J. Hafner. Ab initio molecular dynamics for liquid metals. *Physical Review B 47, 558*, 1993.

[13] M. Levoy. Display of surfaces from volume data. *IEEE Computer Graphics and Application Vol 22 No. 8*, pages 19–37, 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
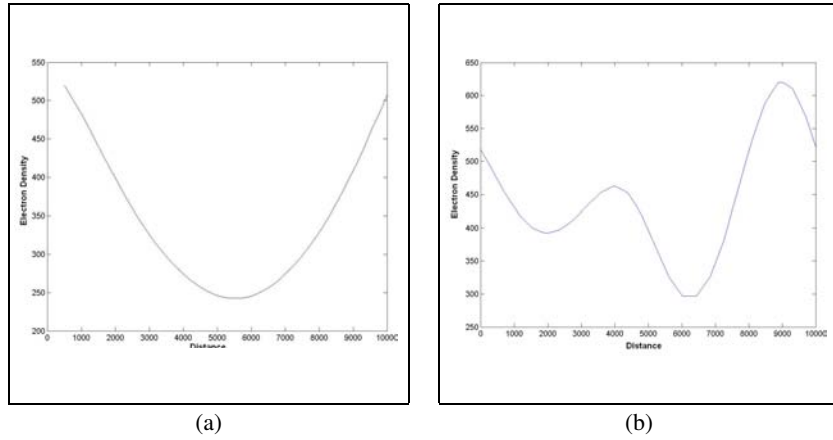
Figure 6: (a) Density behavior between two bulk atoms - First atom is at 0 and second atom is at $10,000$ (b) Density behavior between two defect atoms
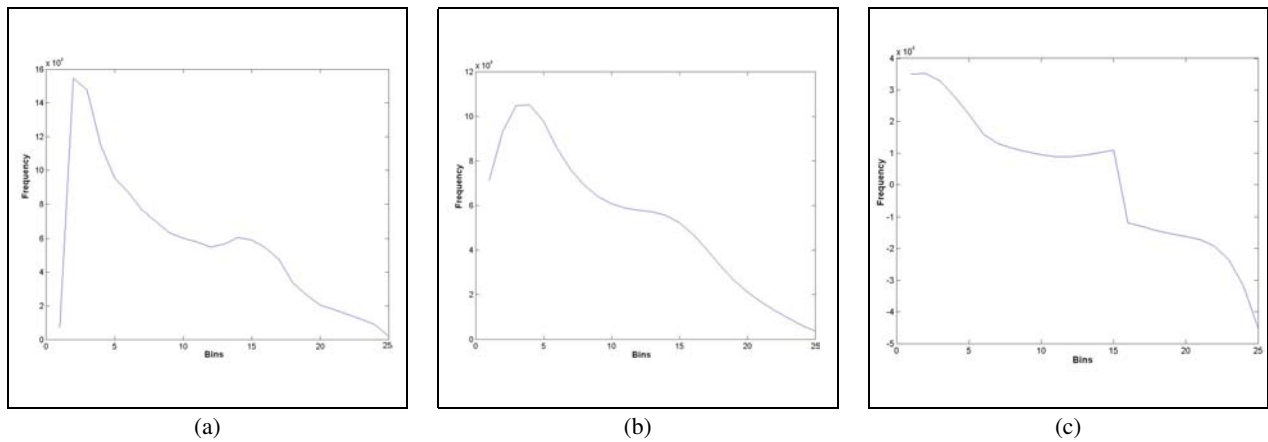

Figure 7: (a) Original Distribution (b) Smoothed Distribution (c) Band Pass Distribution
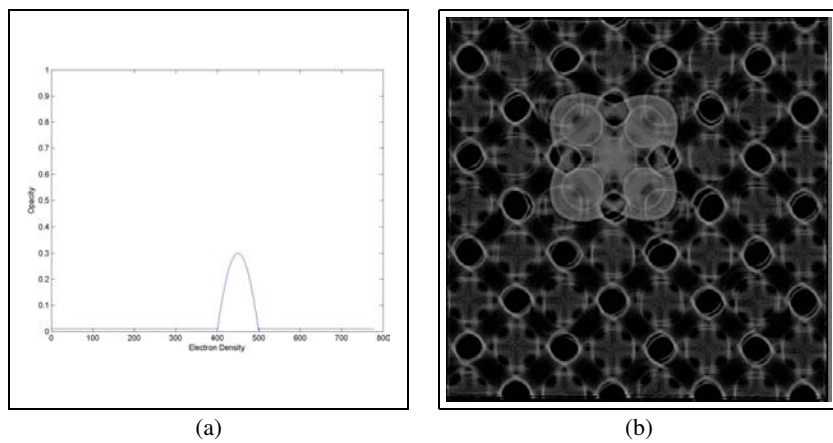

Figure 8: Dataset I1 (a) Transfer Function derived using iso-value (b) Volume Rendered Image using the transfer function
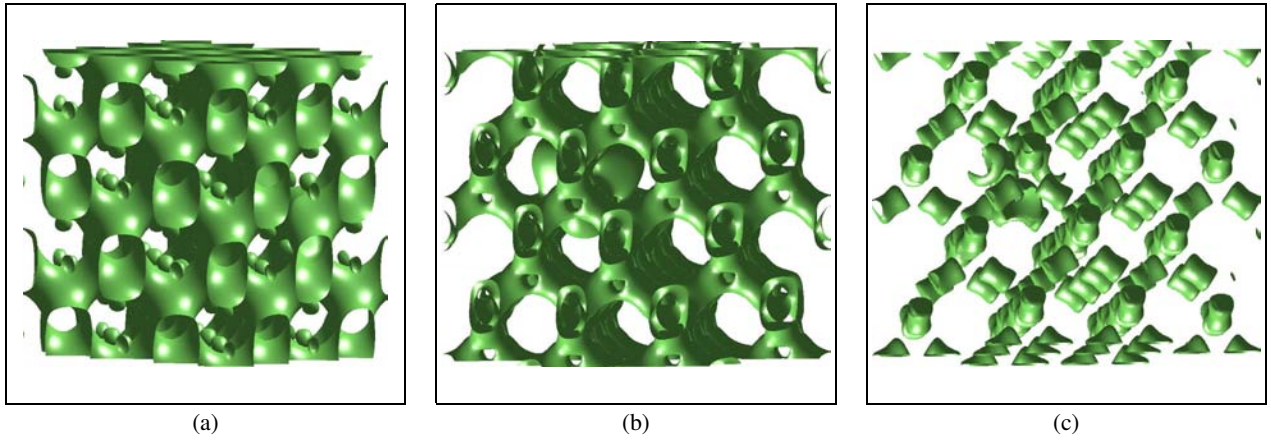
Figure 9: Dataset I1 (a) iso-surface before the transition point (b) At the transition point (c) After the transition Point
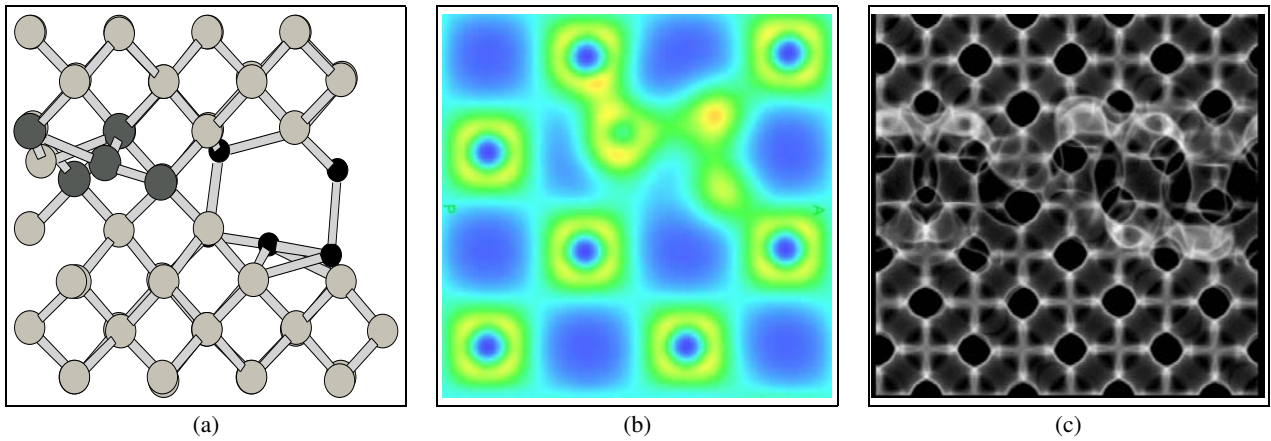

Figure 10: Dataset D1 (a) Original Marked Lattice with two defects (b) Slicing showing shape of one defect (c) Volume rendering showing both defects.
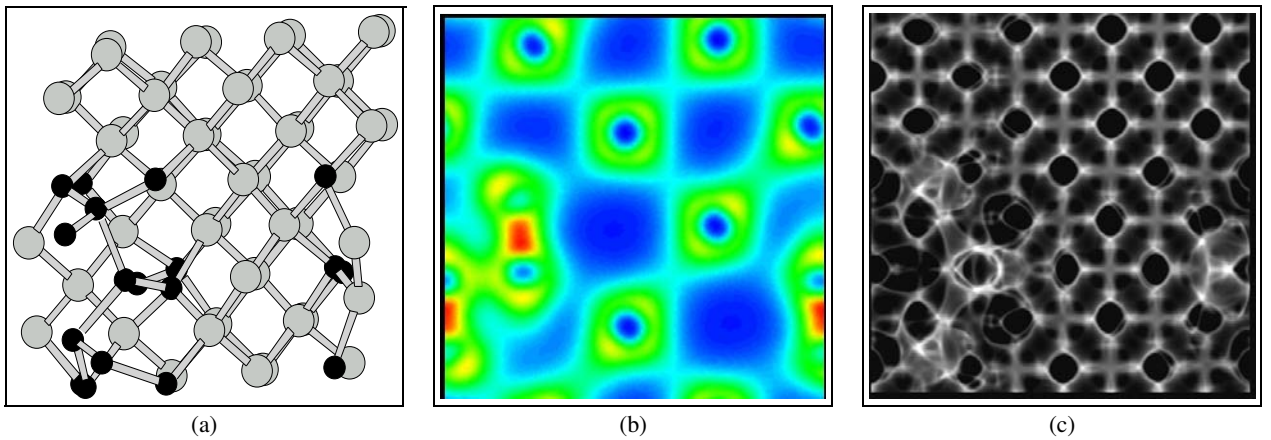

Figure 11: Dataset I2 (a) Original marked lattice (b) Slicing (c) Volume rendering

B. Gatlin. Feature Mining Paradigms for Scientific Data . In *SIAM*, pages 13–24, 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*, pages 29–36, 2002.

[18] R.J. Needs W.K. Leung and G. Rajagopal. Calculation of silicon self interstitial defects. *Physical Review Letters vol. 83, 2351*, 1999.