SOLVING VIRUS STRUCTURES FROM XFEL DIFFRACTION PATTERNS

OF RANDOM PARTICLE ORIENTATIONS USING ANGULAR

CORRELATIONS OF INTENSITIES

by

Miraj Uddin

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Physics

at

The University of Wisconsin - Milwaukee

August 2013

ABSTRACT

SOLVING VIRUS STRUCTURES FROM XFEL DIFFRACTION PATTERNS OF RANDOM PARTICLE ORIENTATIONS USING ANGULAR CORRELATIONS OF INTENSITIES

by

Miraj Uddin

The University of Wisconsin-Milwaukee, 2013
Under the Supervision of Professor D. K. Saldin

The world's first x-ray free electron laser (XFEL), the Linac Coherent Light Source (LCLS) at the Stanford Linear Accelerator Center (SLAC) is now creating X-ray pulses not only of unprecendented brilliance; (a billion times brighter than the most powerful previous sources [8]) but also of extremely short duration. Amongst the promised capabilities of this fourth-generation x-ray sources is the ability to record diffraction patterns from individual bio-molecules. The very first XFEL "diffract and destroy" experiments are being performed on relatively large objects such as viruses. To quote from Caspar and Klug [2] "there are only a limited number of efficient designs possible for biological container which can be constructed from a large number of identical protein molecules-the two basic designs are helical tubes and icosahedral shells". Viruses have regular shapes since their protein coats are formed by the self assembly of identical protein subunits which are coded by their genetic material.

Here we develop a test based on the angular correlations of measured diffraction data to determine if the scattering is of an icosahedral particle. For a positive correlation test; an efficient algorithm can combine diffraction data from multiple shots of particles frozen in completely random orientations to generate a full 3-D image of the icosahedral particle. With this method it is expected to be possible to increase the concentration of particles in a solution beyond that of a single particle per snapshot thus allowing the possibility to get more signals from particles in the solvent. We sucessfully apply this method [3] to reconstruct 3-D images of satellite tobacco necrosis virus (STNV) whose atomic coordinates are given in Protein Data Bank entry 2BUK and of paramecium bursarium chlorella viruses (PBCV) from experimental data deposited at cxidb.org

Most of prior structural studies involve scattering by ensembles of biomolecules or viruses, often in the form of crystals. However the state of biomolecules or viruses could be altered by the crystallization process. The understanding of bio-functioning of those ultrasmall quantities could be greatly enhanced if the structural studies were performed on individual uncrystallized particles.

Fiber diffraction played a pioneering role for solving the structure of synthetic polypeptides [4], structure of deoxyribonucleic acid (DNA) [5] and the structure of helical viruses [6] to name only three of the most important. In a typical fiber diffraction experiment identical particles are all aligned along the fiber axis which give rise to layer lines. In this work we have shown that

fiber diffraction can be obtained from a single particle diffraction volume reconstructed from completely randomly oriented helical structures, thus obviating the need of single axis alignment done experimentally such as forming fibers, laser- or flow-alignment.

Acknowledgements

I find myself lucky to take the opportunity to thank many people (particular my committee members) who guided me for the past several years; an invaluable guidance that covers lot more rather than academic advising.

I would like to express my deep respect and sincere gratitude to my advisor, Dilano Saldin, without his advising I may not have found myself here with a completion of my dissertation. Professor Saldin is an endowed leader of a research group who managed to create a wonderfully relaxed research group where group members have the full liberty to work out, share and test their ideas. It is a great pleasure for me to quote "I had a throughly enjoyable graduate research experience while working under his supervision".

I would like to thank all current and former members of our group (to mention a few of them Valentin Shneerson (late), Hin-Cheuk Poon, Kanupriya Pande, Sung Soon Kim, Sandi Wibowo, Gabriel Ferns) and Ourmazd group (Peter Schwander, Russell Fung, Ahmad Hosseinizadeh). Special thanks to Marius Schmidt, Hin Chuck Poon and Peter Schwander for their collaboration and valuable suggestion.

I would like to thank my friends and colleagues here at Milwaukee, my middle school friends and teachers of K. M. Latif Institution (I was in company of a great school teacher S. Haq), my high school friends of Dhaka College and my undergrad friends of Dhaka University. Also I would like

to thank Dr. K. Kabir (an exceptional advisor and classrooom teacher) for motivating me for higher study while working with him at Dhaka University.

I dedicate my dissertation to my parents (my brothers and sisters) and to my family (my son Andalib and my wife Shohana) who gave me love and affection from the time I was born till today.

Contents

List of Tables

List of Figures

vi

Chapter 1

Introduction

Though X-ray crystallography is the primary method for solving structures of biomolecules or proteins, it requires large crystalline sample. A significant portion of biomolecules do not crystalize. We want to know protein structure similar to the way they function in life. In our body cell proteins are in an aquous environment. In a real XFEL (X-ray free electron laser) scattering experiment diffraction patterns are collected from biomolecules in a solvent droplet or biomolecules in a liquid capillary stream.

With a lower concentration (150 mg/mL), the incident X-ray beam illuminates a volume of the solvent which contain roughly 20 - 10,000 biomolecules depending on the size of the molecule [1]. Producing solvent droplet containing single particle or capilarry liquid jet stream with much lower concentration so that the illumination volume would contain a single particle is still very difficult. Besides the amount of useful scattered intensity from a single

particle droplet is relatively less than that from a droplet containing multiple particles.

Unlike X-ray scattering from a crystalline sample consisting of Bragg spots, XFEL scattering from uncrystallized samples with random orientation in a solvent droplet consists of diffuse distribution of intensity. Hence recovering the 3D real space image from diffraction shots of a solvent stream containing many uncrystallized particle with random orientations is a real theoretical and computational challenge. With our developed angular correlation mehod of recovering real space 3D electron density from an ensemble of N diffraction patterns of solvent droplet that could possibly contain single particle or many particles in random orientations is a huge advantage in terms of real experimental feasibility and the number of collected scattered photons on the pixelated detector.

The recent invention of the X-ray free electron laser (XFEL) at the cost of roughly a billion dollars apiece has opened a new era for scientists for studying structural and dynamic molecular biology. The SLAC (Stanford Linear Accelerator)'s two mile long linear accelerator (LINAC) has begun a new phase in its career with the creation of Linac Coherent Light Source (LCLS), the world's first hard X-ray free elctron laser. Stanford's LINAC has produced high-energy electrons for cutting-edge physics research for the past few decades. This tradition of discovery is being continued by using the LINAC to drive a new kind of laser creating X-ray pulses of unprecedented brilliance. Atomic scale snapshots captured by those ultrabright, ultrashort

pulses in quadrillionths of a second reveal never-before-seen structure and properties of matter.

Biomolecules are the building blocks of life. XFEL pulses can be used to investigate the structure and dynamics of biomolecules, proteins and viruses thus increasing the understanding of human life and diseases. Atoms of a molecule are constantly moving or vibrating. Conventional X-ray sources and synchrontons produce long pulses which yield blurred images of this motion. The LCLS is the first light source that produces X-rays of intense brilliance and delivers with femtosecond duration. The LCLS which was selected by Science magazine as one of the top-ten breakthrough innovations of 2012, takes a qualitative stride far beyond its predecessors by providing an exceedingly intense beam of X-rays, thus shedding light into the fundamental process of life.

An XFEL produces pulses which are a billion times brighter than previous light source [8]. XFEL pulses are extremely short so that scientists can take stop action pictures of atoms and molecules in motion.

Crystallization (and purification) alters the natural condition of the biomolecules functioning in nature and some of the biomolecules, particularly most membrane proteins, which are 70% of today's drug targets cannot be crystallized at all. Hence for understanding of the structure and functioning of biomolecules which cannot be crystallized, single particle scattering by an

XFEL is very important. Developing useful theory and algorithms for reconstructing images of single molecules as well as understanding their functioning is equally important.

As the XFEL user facility for single particle scattering has become available, the very first few experiments were performed on relatively large objects such as viruses and the data are now being made available to the public at cxidb.org. Viruses tend to have an efficient design to form a protein coat with a limited amount of genetic material, and then the protein coats are self assembled. Hence virus capsids are either icosahedral or helical according to Caspar and Klug [2].

We develop a theory to construct virus structure from XFEL data based on angular correlations of the scattered intensities and have applied it to two main types of viruses. Also a test based on angular correlation predicts whether the scattered intensity in XFEL diffract and destroy experiment is mainly from an icosahedral virus or not. We also applied this test on the data collected from chlorella virus (PBCV) which is available to public at cxidb.org and found that the scattered intensity is primarily from an icosahedral object and are working on that data set to reconstruct the 3D image of the chlorella virus.

Chapter 2

Basic Theory

2.1 Correlation theory

From a set of N diffraction patterns interpolated from a cartesian to a polar grid, the angular cross correlation is defined as [15] the sum of the product of intensities at different pixels $I(q, \varphi)I(q', \varphi')$ averaged over the set of all diffraction patterns. Since the correlation is obtained via averaging over all possible orientations, as the number of diffraction patterns for a dataset becomes larger, the correlation quantity

$$J(q, q', \Delta\varphi) = \frac{1}{N} \sum_{w=1}^{N} I^w(q, \varphi)I^w(q', \varphi + \Delta\varphi) \qquad (2.1)$$

becomes more accurate.

A $3D$ reciprocal space distribution of scattered intensity from a single molecule or an aggregate of many molecules may be constructed from an

ensemble of N diffraction patterns even if we do not know the relative orientation of each diffraction pattern. Each diffraction pattern either simulated or experimental is assumed to represent a $2D$ section of a reciprocal space intensity distribution and shell-by-shell spherical harmonic expansion coefficients of the intensity distribution may be determined from angular cross correlations [15] between intensities at different pixels averaged over the ensemble of measured diffraction patterns.

The intensity distribution of a particular resolution shell in the reciprocal space of a single molecule may be expanded as [15];

$$I(q, \theta, \varphi) = \sum_{LM} I_{LM}(q) Y_{LM}(\theta, \varphi) \tag{2.2}$$

where $Y_{LM}(\theta, \varphi)$'s are spherical harmonics and $I_{LM}(q)$'s are shell expansion coefficients.

Each diffraction pattern samples a $2D$ section through the reciprocal space of the molecule in a given orientation. Let us label this reciprocal space by Cartesian axes X, Y and Z (figure 2.1). Assuming X-ray incidence direction is antiparallel to Z axis; then the diffraction intensities are those that would lie on the portion of the Ewald sphere $S1$ of radius equal to wavenumber $\kappa = 2\pi/\lambda_{IR}$ where λ_{IR} is the wavelength of incident radiation. The view of the Ewald sphere from a direction antiparallel to X axis is shown in figure 2.2 and that of antiparallel to Z axis shown in figure 2.3 [15]. In that case the diffracted intensities from every pixel of a diffraction pattern
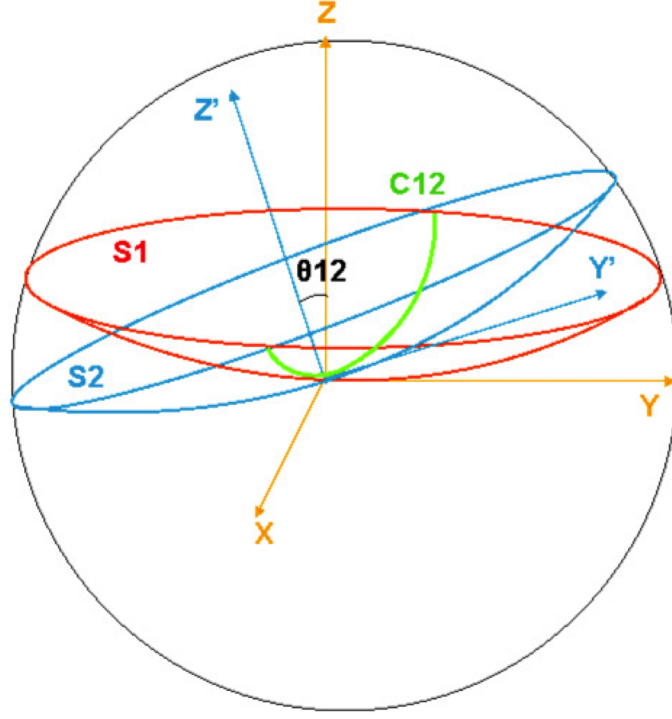
Figure 2.1: Construction of 3D diffraction volume from 2D section of Ewald spheres of random orientation. Ewald sphere $S1$ due to beam incidence antiparallel to the Z axis and $S2$ due to beam antiparallel to Z' axis. The orientation of each Ewald sphere is specified by a set of three Euler angles (α, β, γ).

may be specified by radial distance q, polar and azimuthal angle θ and φ in the frame of reference of the $3D$ diffraction volume [15] as

$$I_Z(q, \varphi) = \sum_{LM} I_{LM}(q) Y_{LM}(\theta(q), \varphi) \tag{2.3}$$

$$\theta(q) = \pi/2 - \sin^{-1}(q/2\kappa) \tag{2.4}$$
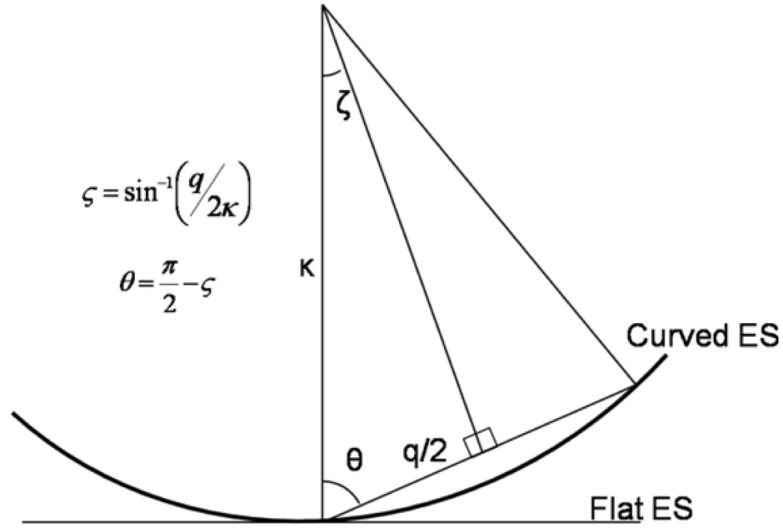
Figure 2.2: A section of the Ewald sphere $S1$ viewed antiparallel to Z axis. θ accounts for the curvature of the sphere for an incident X-ray wavenumber κ.

Diffracted intensity from another snapshot from a different orientation of that molecule may be expressed in the frame of the first one through a Wigner D - matrix rotation of Euler angles of the above quantities of equation (2.3) in the following way:

$$I^{(w)}(q, \varphi) = \sum_{LMM'} D^{(w)}_{LMM'}(\alpha, \beta, \gamma) I_{LM'}(q) Y_{LM}(\theta(q), \varphi) \qquad (2.5)$$

Hence the cross-correlation $J(q, q', \Delta\varphi)$ among all the N diffraction patterns can be written as,

$$J(q, q', \Delta\varphi) = \frac{1}{N} \sum_{LMM'} D^{(w)*}_{LMM'}(\alpha, \beta, \gamma) I^*_{LM'}(q) Y^*_{LM}(\theta(q), \varphi)$$
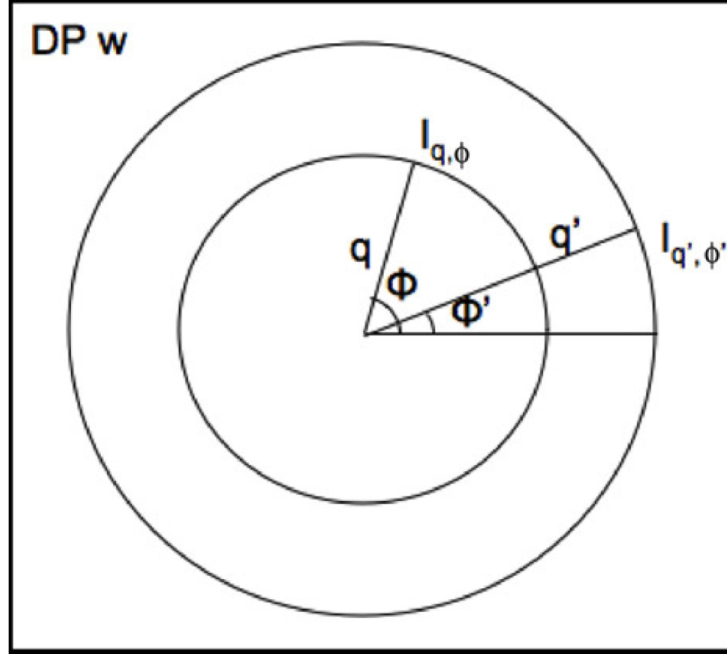
Figure 2.3: Each pixel of a single diffraction pattern for a particular orientation of the molecule may be labelled by the magnitude of the scattering vector q and azimuthal angle ϕ in the frame of reference attached to the difafrction pattern [15]. Intensity cross-correlations may be obtained by multiplying the intensities $I_{q,\phi}$ and $I_{q',\phi'}$ on each diffraction pattern and then averaging over all diffraction patterns.

$$\times \sum_{L'M''M'''} D^{(w)}_{L'M''M'''}(\alpha, \beta, \gamma) I_{L'M'''}(q') Y_{L'M''}(\theta'(q'), \varphi') \quad (2.6)$$

Each diffraction pattern has a unique unknown orientation specified by a given set of Euler angles (α, β, γ); a group element of the representation of the full rotation or $SO(3)$ Lie group. Here we assume that for a sufficiently large number of diffraction pattern all the sets of Euler angles span the entire $SO(3)$ space uniformly. Now the summation over w which is effectively the summation over the space of all elements of the $SO(3)$ space only involves

the D function. Using the great orthogonality theorem (Tinkham 2003) we can write;

$$\frac{1}{N}\sum_{w} D^{(w)*}_{LMM'}(\alpha, \beta, \gamma) D^{(w)*}_{L'M''M'''}(\alpha, \beta, \gamma) = \frac{1}{2L+1}\delta_{LL'}\delta_{MM''}\delta_{M'M'''} \quad (2.7)$$

This orthogonality relationship has huge impact over further simplication of the cross-correlation function. Let us define the following function [15] of Legendre polynomial P_L by doing the M summation over the spherical harmonics.

$$
\begin{aligned}
F_L(qq'; \Delta\varphi) &= \frac{1}{2L+1}\sum_{M} Y^*_{LM}(\theta(q), \varphi)Y_{LM}(\theta'(q'), \varphi') \\
&= \frac{1}{4\pi}P_L[\cos\theta(q)\cos\theta'(q') + \sin\theta(q) \\
&\times \quad \sin\theta'(q')\cos(\varphi - \varphi')]
\end{aligned}
\quad (2.8)
$$

Where $\Delta\varphi = \varphi - \varphi'$

Similarly we define another function $B_L(q, q')$ by doing the M summation over the I_{LM} coefficients;

$$B_L(q, q') = \sum_{M} I^*_{LM}(q)I_{LM}(q') \quad (2.9)$$

Now the intensity cross correlation can be written as [15];

$$J(qq'; \Delta\varphi) = \sum_L F_L(qq'; \Delta\varphi) B_L(q, q') \qquad (2.10)$$

Since indices q and q' are common on both sides of the above equation hence for a particular pair of q and q' the above equation may be written as [15];

$$J_{\Delta\varphi} = \sum_L F_{\Delta\varphi, L} B_L \qquad (2.11)$$

Here $F_L(qq'; \Delta\varphi)$ is a standard mathematical function of a real valued Legendre polynomial; hence $B_L(q, q')$ may be obtained by a matrix inversion [15].

$$B_L = \sum_{\Delta\varphi} \{F^{-1}\}_{L, \Delta\varphi} \, J_{\Delta\varphi} \qquad (2.12)$$

The above equations imply that $B_L(q, q')$ is an experimental XFEL quantity which can be recovered from the average crosscorrelation of the scattered intensities of the diffraction patterns.

2.2 Numerical tests

Equation 2.1 takes simple form for a correlation on the same resolution shell i.e.; $q = q'$; a particular case which we call autocorrelation.

$$J(qq; \Delta\varphi) = \frac{1}{N} \sum_{w=1}^{N} I^w(q, \varphi) I^w(q, \varphi + \Delta\varphi) \qquad (2.13)$$

For autocorrelation equation 2.8 takes relatively simple form
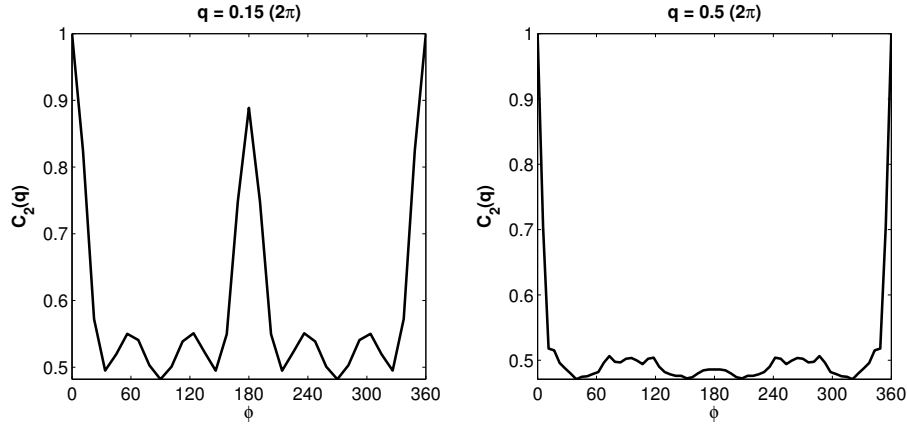
Figure 2.4: Plot of intensity autocorrelation $C_2(\varphi, \varphi')$ vs φ for $\varphi' = 0$ from a set of 5000 simulated diffraction patterns (left plot resolution 6.6 A^0, right plot resolution 2 A^0) for protein chignolin using PDB entry 1UAO. Each plot is a linear combination of Legendre polynomials (as predicted by theory [15] entangled with structural information of the biomolecule which resides in $B_l(q, q)$.

$$F_L(qq; \Delta\varphi) = \frac{1}{4\pi} P_L[\cos^2\theta(q) + \sin^2\theta(q)\cos(\Delta\varphi)] \qquad (2.14)$$

Hence the autocorrelation function can be rewritten in terms of Legendre polynomials.

$$C_2(qq; \Delta\varphi) = \frac{1}{4\pi} \sum_L P_L[\cos^2\theta(q) + \sin^2\theta(q)\cos(\Delta\varphi)] \times B_L(q,q) \qquad (2.15)$$

For a particular resolution shell $C_2(qq; \Delta\varphi)$ is a function of $\Delta\varphi$ only. C_2 is a real quantity since we defined it as the sum of the product of the intensities at different pixels averaged over the set of all diffraction pattern; hence the quantity $B_L(q,q)$ is real. Equation 2.11 has been verified by evaluating $C_2(qq; \Delta\varphi)$ over 5000 simulated diffraction patterns of random orientations for small protein chignolin.

Friedel's rule;

$$I(\vec{q}) = I(-\vec{q}) \qquad (2.16)$$

imposes further restriction on the allowed values of angular momentum L; i.e.;

$$\begin{aligned}
\sum_{LM} I_{LM}(q)Y_{LM}(\hat{q}) &= \sum_{LM} I_{LM}(q)Y_{LM}(-\hat{q}) \\
&= \sum_{LM} I_{LM}(q)\,(-1)^L Y_{LM}(\hat{q}) \qquad (2.17)
\end{aligned}$$

The above equation shows that spherical harmonic expansion coefficients I_{LM}'s vanish for the odd values of angular momentum L_{odd}. Hence the sum over the angular momentum quantum number L in equation 2.15 is just for Legendre polynomials of even orders such as; $L = 0, 2, 4, 6, 8, 10, ...$
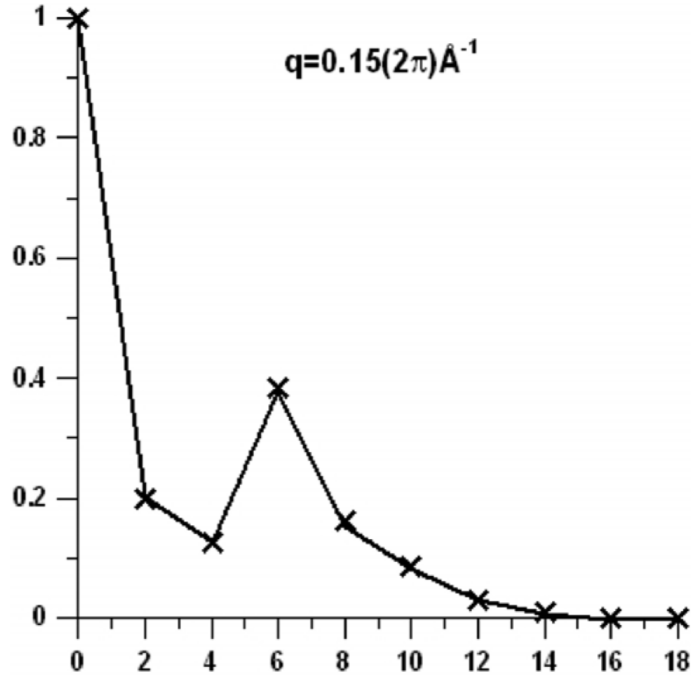
Figure 2.5: Plot of $B_L(q,q)$ versus L for $q = 0.15 \times 2\pi \ A^{-1}$ shows the ratio of the contribution to angular momenta to auto correlation. The crosses represents the value of $B_L(q,q)$ obtained using auto correlation (equation 2.13) in comparison to the same quantity obtained using spherical harmonic expansion coefficients $I_{LM}(q)$ of the 3D distribution of scattered intensity computed using PDB entry 1UAO [15]. The near perfect agreement shows the correctness of the theory which implies $I_{LM}(q)$ coefficients may be extracted from the intensity cross correlation of the measured diffraction patterns.

Careful comparisons of the two $B_L(q,q)$ plots (figure 2.5 and 2.6) reveal meaningful information regarding scattering between the boson field (photon) and fermion field (mainly electron charge cloud). For q scattering field (figure 2.5), the contribution of angular momentum to the diagonal term of $B_L(q,q)$ dies away for $L_{max} > 14$ which corresponds to the physical dimension of the biomolecule $\mathcal{L}$ whose spatial electronic charge distribution causes
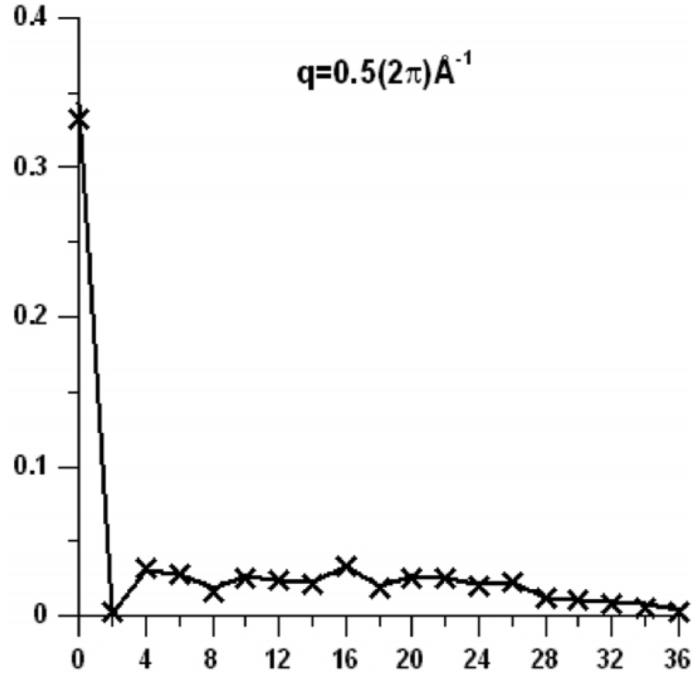
Figure 2.6: Same plot of $B_L(q, q)$ versus L [15] as shown in figure 2.5; except this one is for $q = 0.5 \times 2\pi \ A^{-1}$.

the incoming photon field to be scattered. As per conventional wisdom angular momentum can be related to the magnitude of scattering vector in the following way;

$$\mathcal{L} \times q = L_{max} \tag{2.18}$$

As per the above equation the physical length factor is $\mathcal{L}_{(q=0.15)} = 14 \ A^0$ which is roughly comparable to the physical size of the protein chignolin. Similarly for a higher value of $q = 0.5 \times 2\pi$ we expect to see higher value of L_{max} to contribute the intensity cross correlation. As per equation 2.18;

$L_{max} \sim 42$ which agrees with the vanishing trend of $B_L(q,q)$ in the high L region of figure 2.6. This lead to a crucial discussion addressing the following issue. While the incoming photon field scatters off the electric field of the electronic charge distribution of the $3D$ molecule; a classical definition of angular momentum can be thought with respect to some axis of the physical molecule based on the distribution of the atoms containing electronic charges within the molecule. However; as per conventional wisdom, we can say the maximum allowed values of angular momentum required for the spherical harmonic expansion of the $3D$ diffraction volume may be related to the size of the molecule in the following way [3],

$$\mathcal{R} \times q = L_{max} \qquad (2.19)$$

where $\mathcal{R}$ is radius of the molecule.

2.3   Correlation convergence criterion

A diffraction pattern is a 2D Fourier projection of a 3D object for an unknown orientation. With a point of view of constructing a 3D Fourier map of the object we define the average angular correlation [15] under the assumption that a sufficient large number of diffraction patterns fill the SO(3) space uniformly so that constructed 3D diffraction volume from the calculated orientation-independent radial coefficients can be used as an input for

recovering the real space electron density of the particle via a standard phasing algorithm. One way to test the number of diffraction patterns required to fill the SO(3) space is to check the number of diffraction patterns required for correlation convergence.
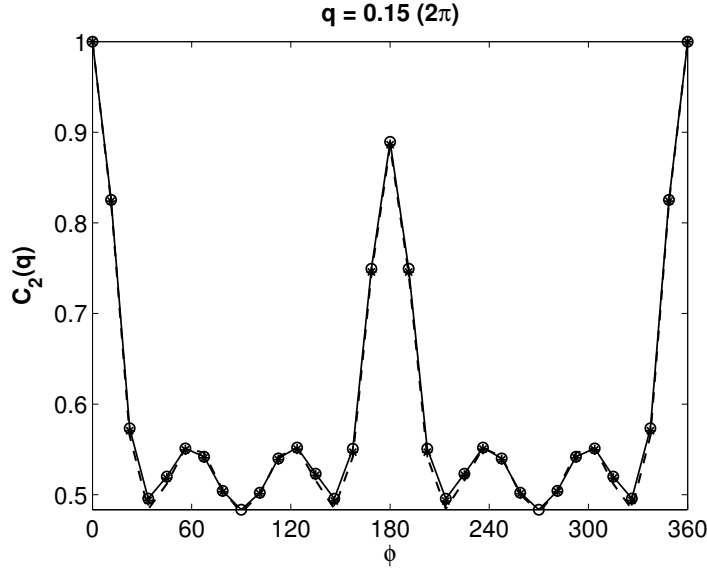
Figure 2.7: Correlation convergence tested for uniform orientation for small protein chignolin via PDB entry 1UAO (www.pdb.org) for various number of diffraction patterns (circles 20,000 diffraction patterns, star 500 diffraction patterns, dashed line 300 diffraction patterns). Comparison of the three sets of data shows correlation converges for 500 diffraction patterns with an R factor of 0.0042

An alternative time-saving approach could be to test the number of DP's required for convergence so that a 2D resolution shell can be constructed via calculating the I_{lm} coefficients from the angular correlations. Figure 2.7 shows a correlation plot for a small protein chignolin using various number of simulated diffraction patterns obtained via uniform cuts across a precalcu-

lated 3D diffraction volume. It turns out that $C_2(q,q)$ for chignolin converges for 500 diffraction patterns with an R factor 0.0042. This result is remarkable since it indicates that a 3D electron density reconstruction of a particle may be possible from an ensemble of some 500 diffraction patterns in contarst to more than a million diffraction patterns [15]. While analyzing the experimental diffraction pattern of chlorella (open source data from www.cxidb.org); we were able to approximately recover the $B_L(q,q)$ selection rule that proves that the data is primarily from an icosahedral object (please see chapter five for detail).

2.4   Orientation independent characterstic quantity and symmetry of the object

A key point of this work is to develop a theory and efficient algorithm for solving virus structure from XFEL data and one of the main challenges for that is to reconstruct the 3D reciprocal space volume of scattered intensity from 2D reciprocal space slices which constitute an ensemble of a large number of diffraction patterns. While doing so we calculate a quantity called average angular correlation of the scattered intensity (as described in detail in the previous section; equation 2.1) from the 2D reciprocal snapshots of an identical copy of the particle which is totally random in its orientation under XFEL diffract and destroy experiments. The beauty of this technique is that though an experimental data set consisting of a large number (anywhere

from a thousand to a million) of diffraction patterns (each pattern might be 1000×1000 pixels or even higher; a typical count); but the reduced data set consists of a few Δq steps (say 60) times the azimuthal steps (say 64). In fact while averaging over a large number of the measured diffraction patterns of the angular correlation of their intensities [15]; we average out the angular dependence of the measured 2D reciprocal space constructions leaving behind the radial dependence as a quadratic function of the spherical harmonic expansion coefficients. Solving for those expansion coefficients I_{LM}'s from the quadratic functions $B_L(q, q')$'s is a major bottleneck; particularly if the quadratic function involves a summation over azimuthal quantum number M on the right hand side as shown here (equation 2.11);

$$B_L(q, q') = \sum_M I_{LM}^*(q) I_{LM}(q')$$

However; if the object under XFEL diffraction experiment has symmetry (and this is almost true for most viruses); we developed an alternative approach for solving it. As per Caspar and Klug [2] virus capsids tend to be either helical or icosahedral in shape. This symmetry of the capsid in real space is translated into reciprocal space in the 3D reconstruction of the diffraction volume via the 2D experimental diffraction patterns of the scattered intensities obtained from the identical copies of the viruses (totally random in its orientations) under diffraction experiments. The next two

paragraphs will summarize two alternative approaches for solving icosahedral and helical virus structures.

Due to icosahedral point group symmetry the scattered intensity distribution from an icosahedral particle may be expanded in terms of icosahedral harmonics [3] which is a restricted sum of real spherical harmonics along with its azimuthal and angular momentum dependent coefficients (details discussed in section 3.1; equation 3.24).

$$I(q, \theta, \phi) = \sum_L g_L(q) \mathcal{I}_L(\theta, \phi)$$

$$\mathcal{I}_L(\theta, \phi) = \sum_{M \geq 0} a_{LM} \mathcal{Y}_{LM}(\theta, \phi)$$

The azimuthal summation of the square of the coefficients a_{LM} is unity for any L (equation 3.7); The characterstic radial quantity is now only a quadratic function of expansion coefficients $g_L(q)$'s; no M summation on the RHS (equation 3.21). The remaining issue regarding the sign determination of the expansion coefficients will be discussed in detail in the relevant chapter (chapter three).

$$B_L(q, q) = g_L(q) g_L(q)$$

For the case of a helical viruses though we expand the scattered intensity distribution in terms of spherical harmonics (a very natural selection for expansion) however the summation over the azimuthal quantum number M of the characterstic structural quadratic radial function may be avoided [9] up to a certain maximum resolution as a consequence of 49_3 helical symmetry of 2TMV (section 4.3; equation 4.20). Here I should mention that the TMV (tobacco mosaic virus) protein coat consists of roughly 45 subunits of c repeat unit; each of 69 A^0 length. Each subunit consists of 3 helical turns with a total of 49 proteins [2].

Chapter 3

Reconstruction of an Icosahedral Virus from a Single-Particle Diffraction Experiment

3.1 Icosahedral harmonics

The intensity distribution of any resolution shell may be expanded in terms of spherical harmonics in the reciprocal space of a single particle [15].

$$I(q, \theta, \phi) = \sum_{LM} I_{LM}(q) Y_{LM}(\theta, \phi) \tag{3.1}$$

Because of the inversion (or Friedel) symmetry of Y_{LM} i.e.;

$$Y_{LM}(\pi - \theta, -\pi + \phi) = (-1)^L Y_{LM}(\theta, \phi) \tag{3.2}$$

a spherical harmonic expansion of intensity distribution contains only even values of angular momentum quantum number L. Since intensity distribution is real, this applies further restriction on the expansion i.e.; intensity distribution may be expanded [3] in terms of real spherical harmonics (RSHs) $\mathcal{Y}_{LM}(\theta, \phi)$ which are defined as a linear combination of spherical harmonics (equation 3.3).

$$\mathcal{Y}_{LM}(\theta, \phi) = \begin{cases} \dfrac{1}{\sqrt{2}}[Y_{LM}(\theta, \phi) + (-1)^M Y_{LM}(\theta, \phi)] & M > 0 \quad (3.3a) \\[2mm] Y_L^0(\theta, \phi) & M = 0 \quad (3.3b) \\[2mm] \dfrac{1}{i\sqrt{2}}[Y_{LM}(\theta, \phi) - (-1)^M Y_{LM}(\theta, \phi)] & M < 0 \quad (3.3c) \end{cases}$$

On the RHS of the above equation for $M > 0$ the ϕ dependence is of the form of $\cos(\phi)$ and for $M < 0$ the ϕ dependence is of the form of $\sin(\phi)$. If the reconstructed intensity has a mirror plane (we may choose the $x - z$ plane where $\phi = 0$), then equation (3) can be replaced by summation over the subset of RHS [3] for which $M \geq 0$ namely;

$$I(q, \theta, \phi) = \sum_{L, M \geq 0} R_{LM}(q) \mathcal{Y}_{LM}(\theta, \phi) \tag{3.4}$$

Since intensities are real and they are expanded in terms of real spheri-
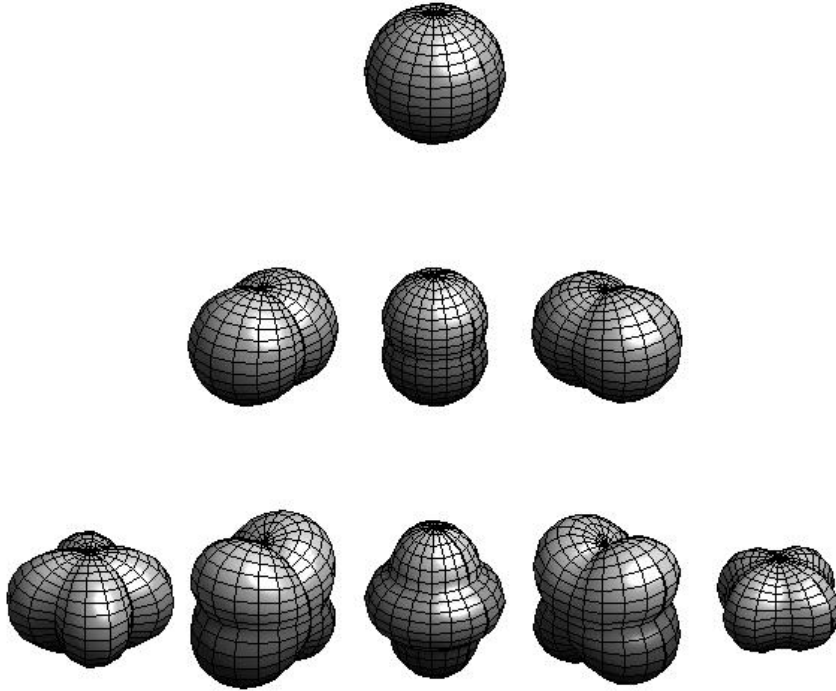
Figure 3.1: Visualization of real spherical harmonics (RSH) for angular momentun $L = 0, 1$ and 2. Plots are made with MATLAB by an adoption of code by Denise L. Chan

cal harmonics; so the expansion coefficients R_{LM} may be considered as real quantities.

Icosahedral point group symmetry imposes further restriction on the expansion of $I(q, \theta, \phi)$; namely $I(q, \theta, \phi)$ can be expanded in terms of icosahedral harmonics

$$I(q, \theta, \phi) = \sum_L g_L(q) \mathcal{I}_L(\theta, \phi) \tag{3.5}$$

Icosahedral harmonics are certain linear combination of real spherical harmonics defined as
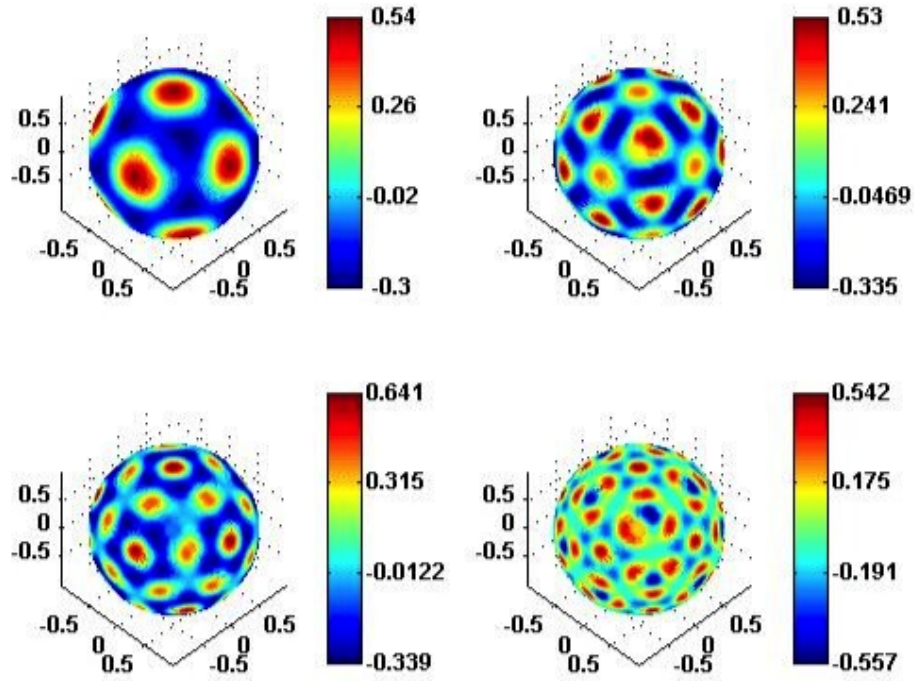
Figure 3.2: Visualization of icosahedral harmonics (RSH) for angular momentun L =6 (upper left), 10 (upper right), 12 (lower left) and 16 (lower right) where the 2 fold, 3 fold and the 5 fold symmetry axis can be seen clearly. Plots are made using MATLAB code written by Haiguang Liu

$$
\mathcal{I}_L(\theta, \phi) =
\begin{cases}
\displaystyle\sum_M a_{LM} P_{L,M} \cos(\theta) \cos(M\phi) & M > 0 & (3.6a)\\[2ex]
Y_L^0(\theta, \phi) & M = 0 & (3.6b)\\[2ex]
\displaystyle\sum_M a_{LM} P_{L|M|} \cos(\theta) \sin(|M|\phi) & M < 0 & (3.6c)
\end{cases}
$$

where the a_{LM} coefficients are real numbers for normalized real spherical harmonics tabulated as for example in Jack and Harrison [16]. Since the real spherical harmonics are orthogonal with respect to integration over spherical shell, the icosahedral harmonics $\mathcal{I}_L(\theta, \phi)$ are orthogonal with respect to the same integration provided

$$
\sum_M a_{LM}^2 = 1 \tag{3.7}
$$

Comparing equation (3.4) with equation (3.5) and (3.6) we can write

$$
R_{LM}(q) = g_L(q) a_{LM} \tag{3.8}
$$

Following equation (3.4) we define the orientation independent quantiy $B_L(q, q)$ as

$$
B_L(q, q) = \sum_{M \geq 0} R_{LM} R_{LM} \tag{3.9}
$$

Substituting equation (3.8) into (3.9) we get

$$B_L(q, q) = g_L(q)g_L(q) \sum_M a_{LM}^2 \tag{3.10}$$

Using equation (3.7) we can write

$$B_L(q, q) = g_L(q)g_L(q) \tag{3.11}$$

3.2   Reconstruction of the diffraction volume

The average angular correlation C_2 [15] among the diffraction patterns of the measured diffraction patterns contains information of the 3D diffraction volume of a single particle. We define C_2 by;

$$C_2(q, q', \Delta\phi) = \frac{1}{N_p} \frac{1}{N} \sum_p \sum_{pix=0}^{N-1} I^p(q, \phi_{pix}) I^p(q', \phi_{pix} + \Delta\phi) \tag{3.12}$$

where I_p is the intensity on diffraction pattern p, N_p is the number of diffraction patterns in random orientation and ϕ_n is the nth discrete value of ϕ.

The angular Fourier transform of a resolution ring q for the pth diffraction pattern can be written as [3]

$$I_{n'}(q) = \sum_{n=0}^{N-1} I_p(q, \phi_n) exp(in'\phi_n) \tag{3.13}$$

Using the cross-correlation theorem and by performing an angular Fourier transform on each diffraction pattern of random orientations, taking the
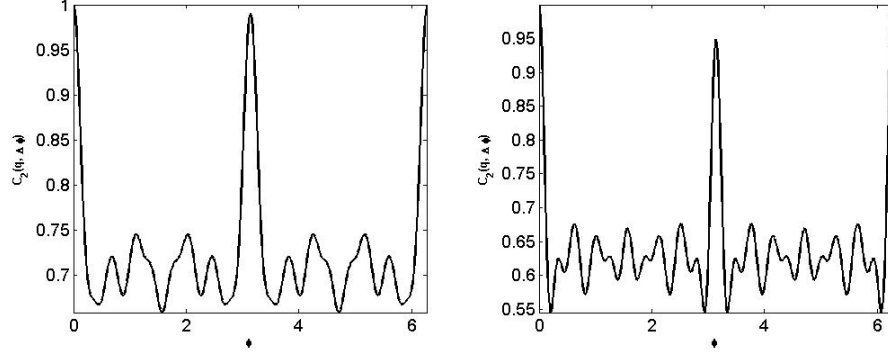
Figure 3.3: Plot of the diagonal term of intensity cross correlation for $q = 0.18A^{-1}$ and $q = 0.25A^{-1}$ from the simulated diffraction pattern (data provided by Peter Schwander) of randomly oriented satellite tobacco necrosis virus (PDB entry: 2BUK).

product of the Fourier transform and its complex conjugate followed by the inverse transform; is the fastest way [3] to calculate the average angular correlation $C_2(q, q', \Delta\phi)$.

The intensity auto-correlation $C_2(q, q', \Delta\phi)$ and the orientation independent characterstic quantity $B_L(q, q')$ are related by [15]

$$C_2(q, q', \Delta\phi) = \sum_L F_L(q, q', \Delta\phi) B_L(q, q') \tag{3.14}$$

where

$$F_L(q, q', \Delta\phi) = \frac{1}{4\pi} P_L[\cos\theta(q)\cos\theta'(q') + \sin\theta(q)\sin\theta'(q')\cos(\Delta\phi)] \tag{3.15}$$

where P_L is the Legendre polynomial of order L and

$$\theta(q) = \pi/2 - \sin^{-1}(q/2\kappa) \tag{3.16}$$

where κ is the wavevector of the incident beam. Following equation (3.16) $B_L(q, q)$ can be obtained by matrix inversion of $F_L(q, q, \Delta\phi)$; i.e.;

$$B_L(q, q) = C_2(q, q, \Delta\phi)F_L^{-1}(q, q, \Delta\phi) \tag{3.17}$$

Note that this quantity, derivable from experiment, does not depend on the azimuthal quantum number M. The construction of the diffraction volume via

$$I(\vec{q}) = \sum_{LM} I_{LM}(q)Y_{LM}(\theta, \phi) \tag{3.18}$$

requires coefficients $I_{LM}(q)$ which depend on M. However the diffraction volume of an icosahedral particle may be expanded as

$$I(\vec{q}) = \sum_{L} g_L(q)\mathcal{I}_L(\theta, \phi) \tag{3.19}$$

where $\mathcal{I}_L(\theta, \phi)$ are icosahedral harmonics obtained via M summation of real spherical harmonics $\mathcal{Y}_{LM}(\theta, \phi)$ for only certain angular momentum quantum number L as permitted by icosahedral symmetry of the object (equation 3.6).

The magnitudes of the coefficients $g_L(q)$ can be directly found from the
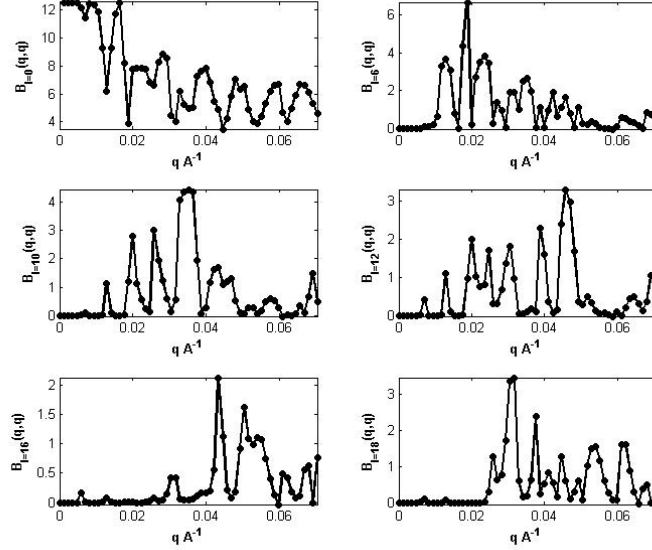
Figure 3.4: $B_L(q, q)$ vs q plot for $L = 0, 6, 10, 12, 16$ and 18. Plots for $L = 1, 2, 3, 4, 5, 7$ etc are absent since those components are zero as per icosahedral selection rule. Plots are obtained from some 10 thousand simulated diffraction pattern (data provided by Peter Schwander) of satellite tobacco necrosis virus (STNV) whose atomic coordinates are deposited on pdb.org as 2BUK

.

equation

$$B_L(q, q) = g_L(q)g_L(q) \tag{3.20}$$

and the $B_L(q, q)$ coefficients can be obtained from the intensity auto correlation on a resolution ring q. Thus

$$|g_L(q)| = \sqrt{B_L(q, q)} \tag{3.21}$$

Now the only remaining task to construct the diffraction volume is to get

the correct *sign* of the $g_L(q)$ coefficients. Once we know the *sign* we can construct

$$I(q, \theta, \phi) = \sum_L sign(g_L(q)) \, |g_L(q)| \, \mathcal{I}_L(\theta, \phi) \tag{3.22}$$

Since the only permitted values of angular momentum quantum number for an icosahedral harmonic expansion of the diffraction volume are $L = 0, 6, 10, 12, 16, 18, 20, 22, 26, 28$ and 30, The correct *sign* combination for a single q resolution shell was determined by an exhaustive search over $2^{12} \simeq 4000$ *sign* commbinations that minimizes the $(-)$ ve values of I for a fixed resolution shell.
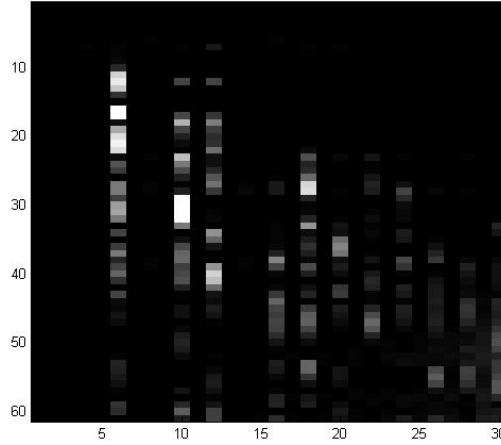
Figure 3.5: $B_L(q, q)$'s (plotted along the Y axis) vs L (plotted along the X axis) for STNV obtained from intensity auto correlation from randomly oriented simulated diffraction patterns (data provided by Peter Schwander) clearly shows that for an icosahedral objects the only nonvanishing $B_L(q, q)$'s are for $L = 0, 6,\ 10, 12, 16, 20, 22, 24, 26$ and 30. $L = 0$ was omitted from the plot so that other can be visualized properly.

$$\sum_{\theta,\phi} |I_-^q(\theta,\phi)| \tag{3.23}$$

Once we know the coefficient $g_L(q)$ for a particular shell q, we can get the coefficients for other shells q' via the following simple quotient;

$$g_L(q') = B_L(q,q')/g_L(q) \tag{3.24}$$

It is worth mentioning here that the numerator on the right hand side of the above equation $B_L(q,q')$ can be directly calculated from the average intensity cross correlation between different diffraction patterns on the measured diffraction pattern. Note that an exhaustive search over the signs of the coefficients was done for only one ring.

Once we know all signs and magnitudes of expansion coefficients for all the resolution shells a 3D diffraction volume may be reconstructed via

$$I(q,\theta,\phi) = \sum_L g_L(q)\mathcal{I}_L(\theta,\phi) \tag{3.25}$$

This intensity distribution can be interpolated onto an oversampled 3D Cartesian reciprocal space grid (q_x, q_y, q_z) and an iterative phasing algorithm may be applied to reconstruct the 3D electron density of the particle [21].

3.3   Numerical tests

The key point of this work is that the scattered intensity from an icosahedral particle may be represented as a sum of icosahedral harmonics. This proposition was verified first by calculating the spherical harmonic expansion coefficients of a simple icosahedral particle via the expression

$$A_{lm}(q) = i^l \sum_j f_j(q) j_l(qr_j) Y_{lm}(\hat{r}_j) \tag{3.26}$$

where $f_j(q)$ is the form factor of the jth atom whose coordinate is $\vec{r}_j$, j_l is a spherical Bessel function and Y_{lm} is a spherical harmonic.

As a first verification the simulation was done for an artificial regular icosahedral molecule of edge length 2 A^o having identical atoms at the vertices whose coordinates in Cartesian frame given by [17];

$$(0, \pm 1, \pm \Phi)$$

$$(\pm 1, \pm \Phi, 0)$$

$$(\pm \Phi, 0, \pm 1)$$

where Φ is the golden ratio $(1 + \sqrt{5})/2$.

Calculated values of A_{lm} coefficients for possible values of l's and m's are listed in Fig. 3.7. Clearly we see the calculated coefficients are real and coefficients from non icosahedral components are zero; i.e.; $A_{lm}(q)$ coefficients
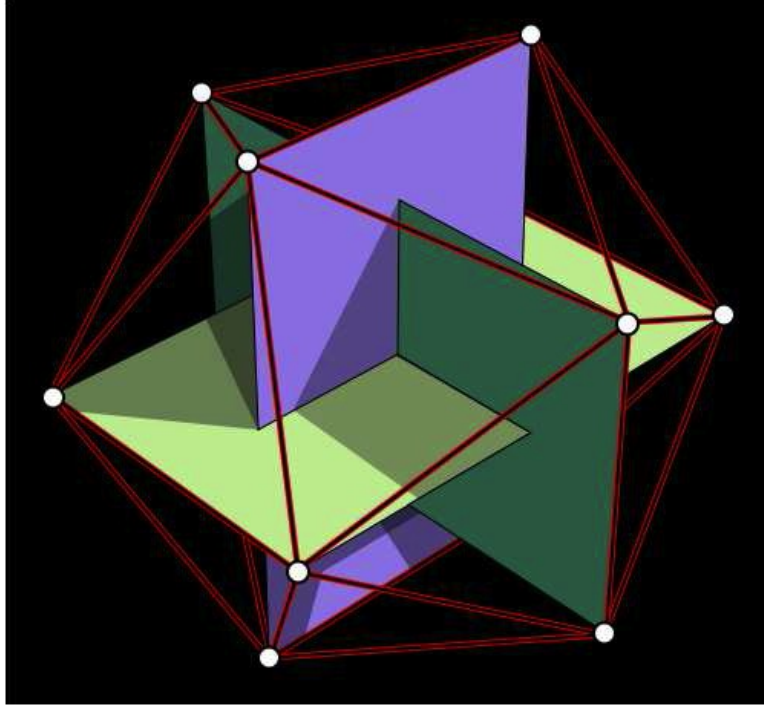
Figure 3.6: A regular icosahedron with twelve identical atoms at the vertices [17].

for $l = 2, 4, 8, 14$ are zero. These results hold for any orientation of the icosahedral particle since Wigner D matrix rotation only mixes amplitudes of different magnetic quantum number corresponding to the same angular momentum L [3]. It is of greater interest that this method allows the calculation of the $I_{LM}(q)$ coefficients, since the $I_{LM}(q)$'s are spherical harmonic expansion coefficients of the scattered intensity;

$$I(\hat{q}) = \sum_{LM} I_{LM}(q) Y_{LM}(\hat{q}) \qquad (3.27)$$

Also $I(\vec{q})$ can be written as the mod square of the amplitude $A(\vec{q})$;

```
0 0 3.385138 0.000000          5 -3 -0.000000 0.000000
1 -1 -0.000000 0.000000        5 -2 -0.000000 0.000000
1 0 0.000000 0.000000          5 -1 0.000000 0.000000
1 1 -0.000000 0.000000         5 0 0.000000 -0.000000
2 -2 -0.000000 -0.000000       5 1 0.000000 0.000000
2 -1 0.000000 0.000000         5 2 -0.000000 0.000000
2 0 -0.000000 0.000000         5 3 -0.000000 0.000000
2 1 0.000000 0.000000          5 4 0.000000 -0.000000
2 2 -0.000000 -0.000000        5 5 -0.000000 -0.000000
3 -3 0.000000 -0.000000        6 -6 -2.592501 -0.000000
3 -2 0.000000 0.000000         6 -5 0.000000 0.000000
3 -1 -0.000000 0.000000        6 -4 -3.139675 -0.000000
3 0 0.000000 0.000000          6 -3 0.000000 -0.000000
3 1 -0.000000 0.000000         6 -2 3.845301 0.000000
3 2 0.000000 0.000000          6 -1 0.000000 0.000000
3 3 0.000000 -0.000000         6 0 1.678227 0.000000
4 -4 -0.000000 0.000000        6 1 0.000000 0.000000
4 -3 0.000000 0.000000         6 2 3.845301 0.000000
4 -2 0.000000 0.000000         6 3 0.000000 -0.000000
4 -1 0.000000 0.000000         6 4 -3.139675 -0.000000
4 0 0.000000 0.000000          6 5 0.000000 0.000000
4 1 0.000000 0.000000          6 6 -2.592501 -0.000000
4 2 0.000000 0.000000          7 -7 0.000000 -0.000000
4 3 0.000000 0.000000          7 -6 -0.000000 0.000000
4 4 -0.000000 0.000000         7 -5 0.00000 0.000000
5 -5 -0.000000 -0.000000       7 -4 0.00000 0.000000
5 -4 0.000000 -0.000000        7 -3 0.00000 0.000000
```

Figure 3.7: A_{lm} coefficients calculated assuming 12 identical atoms at the vertices of a regular icosahedron ($f_j = 1 \; \forall j$). The four entries in each column are respectively l, m, $\mathrm{Re}[A_{lm}(q)]$ and $\mathrm{Im}[A_{lm}(q)]$.

$$I(\vec{q}) = |A(\vec{q})|^2 \qquad\qquad (3.28)$$

It is obvious from the above two equations that if $A(\vec{q})$ has icosahedral symmetry; $I(\vec{q})$ must have icosahedral symmetry as well. However it is not entirely obvious that the $I_{LM}(q)$'s possesses icosahedral symmetry. A straight calculation of $I_{LM}(q)$ coefficients via equation (3.29) (figure 3.8) using $A(\vec{q})$ as in figure (3.7) reveals that $I_{LM}(q)$ coefficients possess icosahedral symmetry,

so do $A_{lm}(q)$ coefficients.

$$
\begin{aligned}
I_{LM}(q) &= \sum_{lm;l'm'} A_{lm}(q)A^*_{l'm'}(q) \int Y_{lm}(\hat{q})Y^*_{l'm'}(\hat{q})Y^*_{LM}(\hat{q})d\hat{q} \\
&= \sum_{lm;l'm'} A_{lm}(q)A^*_{l'm'}(q) \int Y^*_{lm}(\hat{q})Y_{l'm'}(\hat{q})Y_{LM}(\hat{q})d\hat{q} \\
&= \sum_{lm;l'm'} A_{lm}(q)A^*_{l'm'}(q)\sqrt{(\frac{(2l'+1)(2L+1)}{4\pi(2l+1)})}C^{l0}_{l'0L0}C^{lm}_{l'm'LM} \quad (3.29)
\end{aligned}
$$

where $C^{lm}_{l'm'LM}$'s are Clebsch-Gordan coefficients [19]; L ranges from $|l - l'|$

0 0 29.638449 0.000000	5 -3 -0.000000 0.000000
1 -1 -0.000000 0.000000	5 -2 -0.000000 0.000000
1 0 0.000000 0.000000	5 -1 0.000000 0.000000
1 1 -0.000000 0.000000	5 0 0.000000 -0.000000
2 -2 -0.000000 -0.000000	5 1 0.000000 0.000000
2 -1 0.000000 0.000000	5 2 -0.000000 0.000000
2 0 -0.000000 0.000000	5 3 -0.000000 0.000000
2 1 0.000000 0.000000	5 4 0.000000 -0.000000
2 2 -0.000000 -0.000000	5 5 -0.000000 -0.000000
3 -3 0.000000 -0.000000	6 -6 -30.398445 -0.000000
3 -2 0.000000 0.000000	6 -5 0.000000 0.000000
3 -1 -0.000000 0.000000	6 -4 -36.814346 -0.000000
3 0 0.000000 0.000000	6 -3 0.000000 -0.000000
3 1 -0.000000 0.000000	6 -2 45.088181 0.000000
3 2 0.000000 0.000000	6 -1 0.000000 0.000000
3 3 0.000000 -0.000000	6 0 19.678095 0.000000
4 -4 -0.000000 0.000000	6 1 0.000000 0.000000
4 -3 0.000000 0.000000	6 2 45.088181 0.000000
4 -2 0.000000 0.000000	6 3 0.000000 -0.000000
4 -1 0.000000 0.000000	6 4 -36.814346 -0.000000
4 0 0.000000 0.000000	6 5 0.000000 0.000000
4 1 0.000000 0.000000	6 6 -30.398445 -0.000000
4 2 0.000000 0.000000	7 -7 0.000000 -0.000000
4 3 0.000000 0.000000	7 -6 -0.000000 0.000000
4 4 -0.000000 0.000000	7 -5 0.00000 0.000000
5 -5 -0.000000 -0.000000	7 -4 0.00000 0.000000
5 -4 0.000000 -0.000000	7 -3 0.00000 0.000000

Figure 3.8: I_{LM} coefficients evaluated [3] for an artificial icosahedral molecule using equation (3.30) (values of A_{lm} coefficients obtained from figure (3.7)) shows vanishing trends for non icosahedral components for $L = 1, 2, 3, 4, 5$ and 7.

to $|l + l'|$ as per the standard rule for the addition of angular momentum.

We also tested our model with a realistic icosahedral virus (satellite tobacco necrosis virus (STNV)) whose atomic coordinates are deposited in protein data bank (www.pdb.org) as 2BUK (figure 3.9). As usual $A(\vec{q})$ is defined by summing the form factor for every atom in the molecule along with its phase factor;
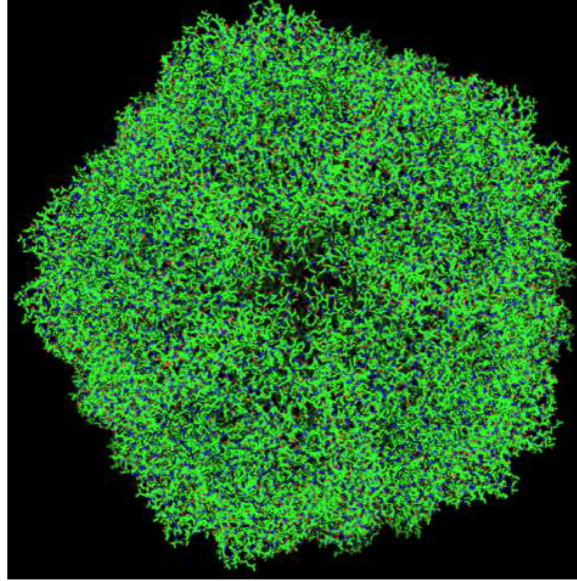
Figure 3.9: Satellite tobacco necrosis virus (STNV) viewed down along the 5-fold axis. Image obtained using UCSF Chimera.

$$A(\vec{q}) = \sum_j f_j(q) \exp(i\vec{q}.\vec{r_j}) \qquad (3.30)$$

Spherical harmonic expansion coefficients $I_{LM}(q)$'s of the 3D diffraction volume of STNV was calculated [3] by integrating over spherical shells of $I(\vec{q})$;

$$I_{LM}(q) = \int I(\vec{q})Y_{LM}(\hat{q})d\hat{q} \qquad (3.31)$$
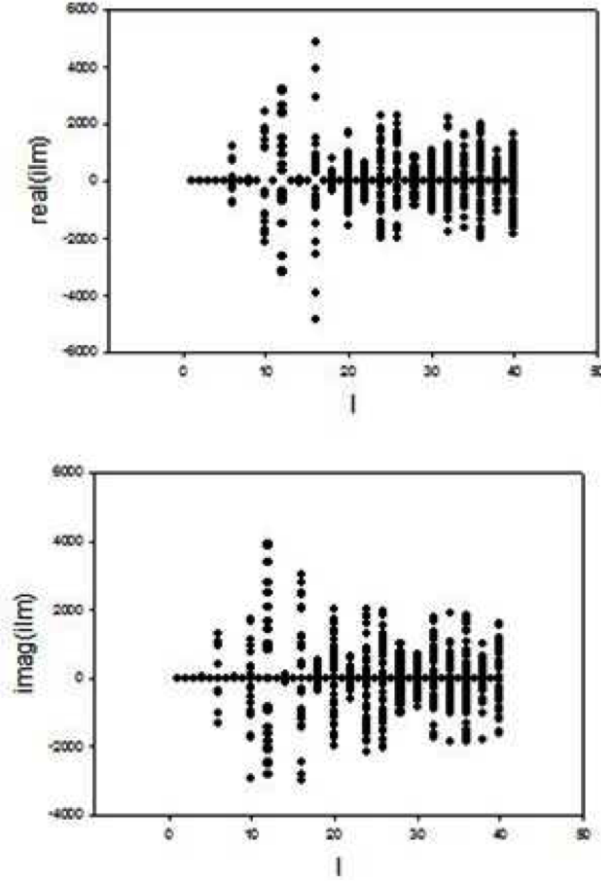
Figure 3.10: Calculated $I_{LM}(q)$ coefficients (real and imaginary parts) from the simulated diffraction volume of STNV. Each dot on the plots represents an LM pair either for real or imaginary part of $I_{LM}(q)$.

where the integration is preferably done by Gaussian quadrature [20].

Clearly we see the vanishing trend for non icosahedral components of $I_{LM}(q)$'s

for $L = 2, 4, 8, 14$ (for both the real and imaginary part). Calculated coefficients also show the precise condition:

$$I_{LM}(q) = I_{L,-M}(q)(-1)^M \qquad (3.32)$$

to ensure the reality of the real spherical harmonic expansion $R_{LM}(q)$ coefficients (equation 3.4) as well as the icosahedral harmonic expansion coefficients $g_L(q)$'s via (3.11). The symmetry of the real icosahedral object (assuming the protein coat of the icosahedral virus is the dominant scatterer) is translated into reciprocal space and this result is a consequence of the icosahedral symmetry of the diffraction volume $I(\bar{q})$. This follows from equation (3.33) that $B_L(q, q')$ coefficients computed from the data of diffraction patterns of icosahedral viruses with completely random orientations must have vanishing values for $L = 1, 2, 3, 4, 5, 7, 8, 9, 11, 13, 14, 15, ...,$. Hence this provides a very important characterstic test to tell whether the particle subject to XFEL diffract and destroy experiment is an icosahedral object. However this is only approximately true in a real experiment since the so called icosahedral viruses may have appendages which break the icosahedral symmetry of the protein coat. Also the genetic material inside the protein coat may not have this symmetry since X-rays penetrate more inside the virus; hence the collected diffraction pattern may contain information about the inside genetic materials as well as the outer protein coat [3].

3.4 Reconstruction of STNV from simulated diffraction pattern

Simulated diffraction patterns of satellite tobacco mosaic virus (STNV) were calculated for direction of incidence on a single particle for a uniform angular distribution in SO(3) [15] using data for the biological assembly of STNV from PDB entry 2BUK. Instead of calculating diffraction patterns for various orientations we took slices through a precalculated 3D diffraction volume since the biological assembly of 2BUK consists of roughly $100,000$ of atoms. Average angular correlations for those (10,000) simulated diffraction patterns were obtained using equation 3.12 and $B_L(q, q')$ coefficients were obtained via a matrix inversion as given in equation 3.18.

Since the $B_L(q, q')$ coefficients are related to the real spherical harmonic expansion coefficients $R_{LM}(q)$; it was found that $R_{LM}(q)$ coefficients for $L = 0, 6, 10, 12, 16, 18, 20, 22, 24, 26, 28, 30$; obtained from the simulated diffraction volume of STNV are dominant [3]. However for some larger viruses which are of basic icosahedral shape might have appendages like the unique vertex of the chlorella virus [28] or "hair" of mimivirus [27]. Extracting those coefficients from the experimental single particle diffraction patterns via calculating the average correlation of the intensities would be an excellent test to see whether the icosahedral components of the coefficients are dominantly larger than the non icosahedral components. In fact $B_L(q, q)$ coefficients extracted from

some 800 experimental diffraction patterns of PBCV approximately satisfies icosahedral selection rule as discussed in chapter six.

Finally a $61 \times 61 \times 61$ pixels of 3D diffraction volume of scattered intensities $I(q_x, q_y, q_z)$ was constructed using the steps as described in section 3.2; which was the input of a standard iterative phasing algorithm ("charge flipping" algorithm of Oszlanyi and Suto [29] and [30]) for the recovery of the 3D electron density of the particle.
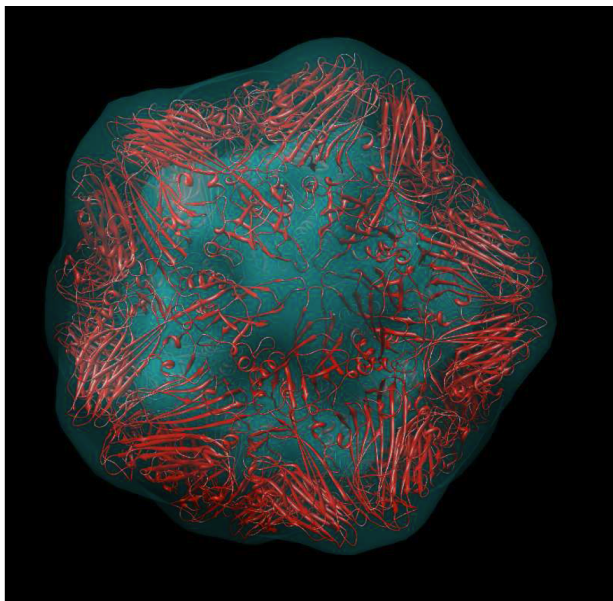
Figure 3.11: Reconstructed image of STNV (which is about 20 mm in diameter) directly from structure factor calculation. The figure depicts a view of the icosahedral virus close down along its 5 -fold rotation axis. The resolution of the reconstruction is about 1.3 nm. A ribbon diagram of the structure factor in PDB entry of 2BUK seems to fit the capsid (plot produced by Peter Schwander).

In order to judge the accuracy of our reconstruction; first the 3D diffrac-

tion volume of scattered intensity was calculated directly using a structure factor and then by using the expansion coefficients recovered from the average angular correlations of the diffraction patterns of scattered intensities. The recovered 3D electron density for the two cases are shown in figure 3.11 and figure 3.12 for comparing the correctness of our reconstruction. The similar-
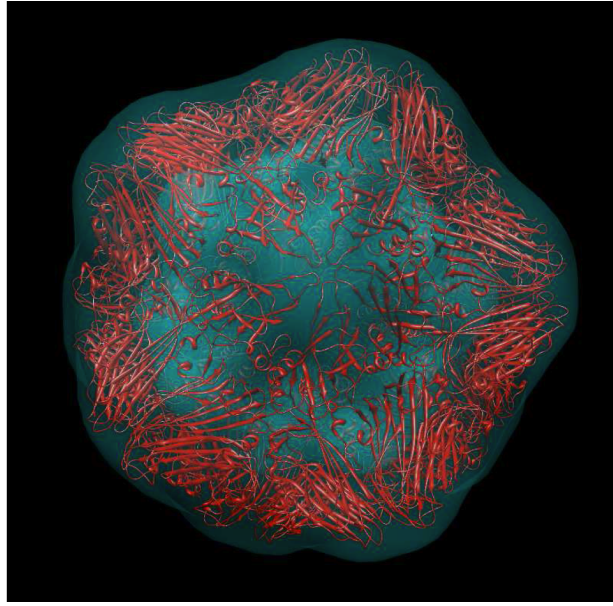
Figure 3.12: Same as 3.12 except the diffraction volume was reconstructed from the average angular correlations of $10,000$ simulated diffraction pattern of STNV [3] uniformly distributed over directions of SO(3) space. Reconstructed electron density seems to be remarkably similar to that of figure 3.11 (plot produced by Peter Schwander).

ity of the reconstructed images of figure 3.11 and 3.12 proves the correctness of our method of reconstructing the diffraction volume from the $B_L(q, q')$ coefficients obtained from the avearge angular correlation of the diffraction pattern of scattered intensities.
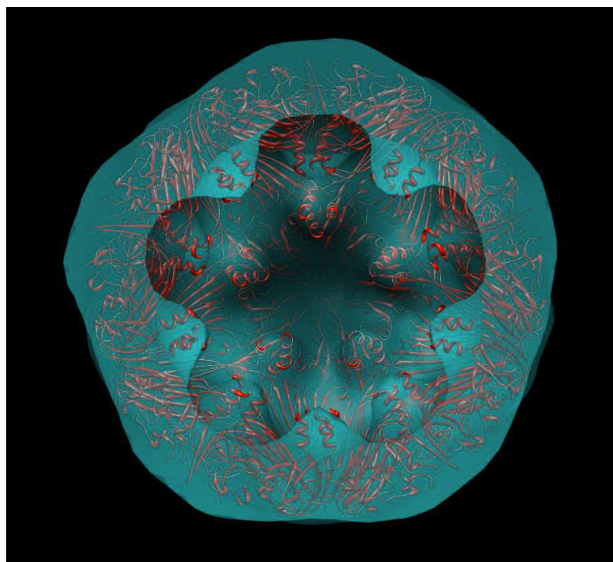
Figure 3.13: Same as 3.12 except a portion of the reconstructed image of STNV has been cut perpendiculr to the 5-fold axis. The 5-fold symmetry of both the internal surface and external capsid can be seen clearly across the cut (plot produced by Peter Schwander).

In fact the reconstructed image consists of a thin protein shell [3] as shown in figure 3.13; obtained by slicing figure 3.12 perpendicular to the 5-fold rotation axis. All those reconstructions were fitted to ribbon diagrams obtained from the PDB entry of the biological assembly of STNV to prove the excellence of the reconstruction. The reconstructed STNV structure consists of a thin shell as obtained by the iterative phasing algorithm without any prior assumptions on our part [3].

3.5 Beyond the icosahedral approximation

Regular viruses are mainly of two basic shape [2], icosahedral and helical. Satellite tobacco virus (STNV) is an example of icosahedral virus that has perfect icosahedral symmetry [32]. Many of the larger viruses are predominantly of icosahedral shape, except they might have appendages such as a neck sticking out of the coat which is used to inject the genetic material into a host cell to produce more identical biological assemblies to be combined to make another virus.

An ultimate reconstruction algorithm should be able to reconstruct those nonicosahedral parts as well as the dominant icosahedral bulk volume. The above procedure has determined the icosahedral harmonic expansion coefficients $g_L(q)$ that best fit the measured quantities $B_L(q, q')$. Any deviations from those are due to the non-icosahedral part of the structure which may be written as [15];

$$\delta B_L(q, q') = \sum_M a_{LM}[g_L(q)\delta R_{LM}(q') + \delta R_{LM}(q)g_L(q')] + \delta R_{LM}(q)R_{LM}(q')$$

$$(3.33)$$

The additional term on the RHS out of the square bracket of the above equation is due to deviation in structure from icosahedral symmetry. Note that for (L, M) combination not associated with icosahedral harmonics; the a_{LM} coefficients will be zero and only the quadratic term in R_{LM} will survive. Determination of the $\delta R_{LM}(q)$ coefficients which optimize the agreement of

theoretical expression (equation 3.33) and measured values might allow us to do a reconstruction of bulk icosahedral particle with the presence of their appendages via the construction of diffraction volume as [3];

$$I(\vec{q}) = \sum_{LM}[g_L(q)a_{LM} + \delta R_{LM}(q)]\mathcal{Y}_{LM}(\hat{q}) \qquad (3.34)$$

Note that the L values for the correction term $\delta R_{LM}(q)$ has no symmetry restriction except the Friedel symmetry.

Chapter 4

Fiber Diffraction Without Fibers

4.1 Introduction

Fiber diffraction by helical structures are of great interest since the early ninteen fifties starting from the work of Cochran, Crick and Vand (denoted as CCV; 1952) for solving the structure of synthetic polypeptides [4]. The work of CCV contains the full diffraction theory which is regarded as the seminal work in this field. Other best known fiber diffraction work done in the fifties and sixties are the following; structure of deoxyribonucleic acid (DNA) [5], tobacco mosaic virus (TMV) [6], to name only two of the most important. In a typical fiber diffraction experiment diffracting particles are all randomly oriented along the fiber axis. For example; tobacco mosaic

viruses (TMV) which are rod shaped are drawn into a capillary tube in a gel to line up their long axis parallel to the capillary axis.

In this work we have shown that fiber diffraction patterns can be recovered from the single particle diffraction patterns of completely randomly oriented helical particles as measured in so called "diffract-and-destroy" experiments with an x-ray free electron laser (XFEL). This computatinal alignment technique of postprocessed diffraction patterns of totally randomly oriented helical particles obviates the need of single axis alignment done experimentally such as forming fibers, laser- or flow-alignment.

4.2 Fourier transform of a helix with continuous structure factor

A Fourier transform of (x, y, z) at a reciprocal space point (ξ, η, ζ) can be written as [4]

$$T(\xi, \eta, \zeta) = \int \exp[2\pi i(x\xi + y\eta + z\zeta)]dV \qquad (4.1)$$

A uniform helix of radius r, axis spacing $\mathcal{C}$ is defined as

$$
\begin{aligned}
x &= r\cos(2\pi z/\mathcal{C}) & (4.2) \\
y &= r\sin(2\pi z/\mathcal{C}) & (4.3) \\
z &= z & (4.4)
\end{aligned}
$$

For a uniform helix the volume element dV is proportional to dz, hence the transform function can be written as

$$T(\xi, \eta, \zeta) = \int_0^P \exp[2\pi i(r\xi \cos 2\pi \frac{z}{P} + r\eta \sin 2\pi \frac{z}{P} + z\eta)]dz \qquad (4.5)$$

For a helical system, the natural choice of the coordinate is cylindrical. In cylindrical reciprocal space coordinates (R, φ, ζ) the helix transform function can be written as

$$T(R, \varphi, \zeta) = \int_0^P \exp[2\pi i\{Rr \cos(2\pi \frac{z}{P} - \varphi) + z\zeta\}]dz \qquad (4.6)$$

where $R^2 = \xi^2 + \eta^2$ and $\tan \varphi = \eta/\xi$

(4.6) vanishes unless $\zeta = \lambda/\mathcal{C}$, where λ is an integer $((+)$ve or $(-)$ve). This corresponds to the fact that X-ray scattering from a helix which has an exact repeat along the vertical distance λ, is confined to layer lines at height $\zeta = \lambda/\mathcal{C}$ in reciprocal space [4] where $\lambda = 0, \pm 1, \pm 2, \pm 3, ...$

Hence (4.6) can be written as

$$T(R, \psi, \frac{\lambda}{\mathcal{C}}) = \int_0^P \exp[2\pi i\{Rr \cos(2\pi \frac{z}{P} - \varphi) + \frac{\lambda z}{\mathcal{C}}\}]dz \qquad (4.7)$$

Taking $X = 2\pi Rr$ and $\phi = 2\pi z/\mathcal{C}$ the above integral can be solved using the identity [4].

$$\int_0^{2\pi} exp(iX \cos \phi) \; exp(i\lambda\phi)d\phi = 2\pi i^\lambda J_\lambda(X) \qquad (4.8)$$

where J_λ denotes the λth-order Bessel function.

Hence the transform function can be written as

$$T(R, \varphi, \frac{\lambda}{C}) = J_\lambda(2\pi Rr) \; exp[i\lambda(\varphi + \frac{\pi}{2})] \qquad (4.9)$$

Clearly we see that the amplitude of the Fourier transform of a continuous helix does not depend on φ i.e.; it has cylindrical symmetry.

4.3 Helix selection rule

The structure factor of a c repeat unit of TMV consisting of u number of poteins whose atoms are specified by cylindrical coordinates (r_k, ψ_k, z_k) may be written as

$$
\begin{aligned}
F_c(\mathbf{q}) &= \sum_{k \, \epsilon \, \mathcal{H}} f_k \; exp(2\pi i \mathbf{q}.\mathbf{r}) \\
&= \sum_{k \, \epsilon \, \mathcal{H}} f_k \; exp(2\pi i[q_x x_k + q_y y_k + q_z z_k]) \\
&= \sum_{k \, \epsilon \, \mathcal{H}} f_k \; exp(2\pi i[\zeta z_k + Rr_k(\cos\psi_k \cos\varphi + \sin\psi_k \sin\varphi)]) \\
&= \sum_{k \, \epsilon \, \mathcal{H}} f_k \; exp\{2\pi i[\zeta z_k + Rr_k \; \cos(\varphi - \psi_k)]\} \qquad (4.10)
\end{aligned}
$$

where $\mathcal{H}$ denotes the set of all atoms in one c repeat unit (69 A^0 for 2TMV). (4.10) can be expanded in terms of Bessel function using Jacobi-Anger expansion (4.11)

$$exp(iZ\cos\theta) = \sum_{n \, \epsilon \, \mathcal{I}} i^n J_n(Z)\,exp(in\theta) \tag{4.11}$$

where $\mathcal{I}$ denotes the set of all integers.

With $Z = Rr$ and $\theta = \varphi - \psi$ structure factor expression can be written as

$$\begin{aligned} F_c(\mathbf{q}) &= \sum_{n \, \epsilon \, \mathcal{I}} \sum_{k \, \epsilon \, \mathcal{H}} f_k\,i^n\,exp\{2\pi i(\zeta z_k)\}\,exp\{in(\varphi - \psi_k)\}\,J_n(2\pi R r_k) \\ &= \sum_{n \, \epsilon \, \mathcal{I}} G'_{n,\lambda}(R)\,exp(in\varphi) \tag{4.12} \end{aligned}$$

With $\zeta = \frac{\lambda}{c}$; $G'_{n,\lambda}(R)$ is defined as ;

$$G'_{n,\lambda}(R) = \sum_{k \, \epsilon \, \mathcal{H}} f_k\,i^n\,exp[2\pi i(\zeta z_k)]\,exp(-in\psi_k)\,J_n(2\pi R r_k) \tag{4.13}$$

In the absence of helix symmetry the diffracted intensity for λ-th layer line is given by

$$I_\lambda(R) = \sum_{n,n' \, \epsilon \, \mathcal{I}} G'_{n,\lambda}(R)G'^{*}_{n',\lambda}(R) \tag{4.14}$$

In the presence of p_h helix symmetry (let us assume a c repeat unit consists of h turns with a total of p protein; for TMV $h = 3$ and $p = 49$); structure factor of a c repeat unit F_c can be obtained from the structure factor of one protein F_p via a coordinate transformation of each helix repeating unit.

Following the same steps of (4.12), the structure factor of one repeating

unit of the helix F_p can be written as

$$F_p(\mathbf{q}) = \sum_{n \, \epsilon \, \mathcal{I}} G_{n,\lambda}(R) \, exp(in\varphi) \tag{4.15}$$

where $G_{n,\lambda}(R)$ is defined as

$$G_{n,\lambda}(R) = \sum_{k \, \epsilon \, \mathcal{P}} f_k \, i^n \, exp[2\pi i(\zeta z_k)] \, exp(-in\psi_k) \, J_n(2\pi R r_k) \tag{4.16}$$

where $\mathcal{P}$ denotes the set of all atoms in one helix repeating unit (one protein molecule).

With p_h helix symmetry the k-th atom in J-th protein with coordinate (r_k, ψ_k, z_k) is accompanied by the k-th atom of the $(J+1)$-th protein according to the coordinate transformation [7].

$$(r_k, \psi_k, z_k) \rightarrow (r_k, \psi_k + 2\pi hJ/p, z_k + cJ/p) \tag{4.17}$$

Under this transformation $G'_{n,\lambda}(R)$ of (4.13) can be written as

$$
\begin{aligned}
G'_{n,\lambda}(R) &= \sum_{J=0}^{u-1} \sum_{k \, \epsilon \, \mathcal{P}} f_k \, i^n \, exp[2\pi i\{\zeta(z_k + cJ/p)\})] \, exp[-in(\psi_k + 2\pi hJ/p)] \, J_n(2\pi R r_k) \\
&= G_{n,\lambda}(R) \sum_{J=0}^{u-1} exp[2\pi iJ(\lambda - nh)/p] \tag{4.18}
\end{aligned}
$$

The sum in the above equation 4.18 is non zero only if $(\lambda - nh)$ is integer

multiple of p [7]; i.e.;

$$\lambda = nh + mp \tag{4.19}$$

For a 49_3 helix symmetry of TMV, equation (4.19) can be written as

$$\lambda = 3n + 49m \tag{4.20}$$

With 49_3 helix symmetry diffracted intensity for λ-th layer line is redefined as

$$I_\lambda(R) = \sum_{n,n' \,\epsilon\, \mathcal{S}} G'_{n,\lambda}(R) G'^{*}_{n,\lambda}(R) \tag{4.21}$$

Where $\mathcal{S}$ is set of all allowed integers as permitted by the selection rule.

Table 4.1: Helix selection rule of TMV 49_3 helix. Layer lines index λ shown along Y in *bold* and the permitted Bessel function order n corresponding to each layer line shown along X.

λ/n	-5	-4	-3	-2	-1	0	1	2	3	4	5
−5	–	–	–	31	–	–	−18	–	–	−67	–
−4	–	64	–	–	15	–	–	−34	–	–	–
−3	–	–	48	–	–	−1	–	–	−50	–	–
−2	–	–	–	32	–	–	−17	–	–	−66	–
−1	–	65	–	–	16	–	–	−33	–	–	–
0	–	–	49	–	–	0	–	–	−49	–	–
1	–	–	–	33	–	–	−16	–	–	−65	–
2	–	66	–	–	17	–	–	−32	–	–	–
3	–	–	50	–	–	1	–	–	−48	–	–
4	–	–	–	34	–	–	−15	–	–	−64	–
5	–	67	–	–	18	–	–	−31	–	–	–

Clearly we see that because of 49_3 helical symmetry not every Bessel function are allowed on every layer line. Also we see that for each layer line

the order of the allowed Bessel functions are separated by 49.

4.4 Diffraction from identical molecules of random azimuthal orientation of helical viruses

Fiber diffraction [4] assumes objects periodic along the direction of the fiber axis give rise to layer lines. This section describes the relationship between the electron density of each particle and the composite diffraction pattern produced by the incoherent superposition of diffraction patterns of multiple copies of the particle differing only in their azimuthal orientations. The structure factor of a molecule can be expressed in terms of cylindrical reciprocal space coordinates (R, φ, ζ) [13] as

$$F(R, \varphi, \zeta) = \sum_n G_n(R, \zeta) exp(in\varphi) \tag{4.22}$$

Where $G_n(R, \zeta)$ is a cylindrical harmonic expansion coefficient defined as

$$G_n(R, \zeta) = \sum_k f_k \, i^n \, exp[2\pi i(\zeta z_k)] \, exp(-in\psi_k) \, J_n(2\pi R r_k) \tag{4.23}$$

The structure factor of the j-th molecule on the helix rotated relative to the first one by an angle ψ_j can be expressed as (reference equation (4.10) and (4.12));

$$F_j(R, \varphi, \zeta) = \sum_n G_n(R, \zeta) exp[in(\varphi - \psi_j)] \tag{4.24}$$

The total scattered intensity can be obtained by summing over the contribution of all the molecules in the ensemble. Hence,

$$
\begin{aligned}
I(R,\zeta) &= I_0(R,\zeta) \\
&= \sum_j I_j(R,\varphi,\zeta) \\
&= \sum_n |G_n(R,\zeta)|^2 + \sum_{n,n'} G_n(R,\zeta)G_{n'}^*(R,\zeta) \\
&\qquad\qquad exp[i(n-n')(\psi-\psi_j')] \qquad (4.25)
\end{aligned}
$$

The off diagonal terms in the above expression sum to zero due to the randomness of the azimuthal orientations ψ_j of different molecules. Since the diagonal terms do not depend on φ, the total scattered intensity from the ensemble may be written as

$$
\begin{aligned}
I(R,\zeta) &= I_0(R,\zeta) \\
&= N \sum_n |G_n(R,\zeta)|^2 \qquad (4.26)
\end{aligned}
$$

where N is the number of molecules in the ensemble.

The corresponding intensity $I_\lambda(R)$ of layer line λ in fiber diffraction is given in equation (4.21). The essential difference between the usual fiber diffraction equations (4.12) and (4.21) and (4.23) and (4.26) is that the discrete layer line index λ in the former two equations is replaced by the continuous coordinate ζ parallel to the alignment axis in (4.23) and (4.26).

4.5 Fiber diffraction in XFEL"diffract-and-destroy" experiment

Structure determination from a "diffract-and-destroy" experiment with X-ray free electron laser (XFEL) [22] is at least superficially quite distinct, where reproducible particles of completely random 3D orientations are injected into the XFEL beam and single particle diffraction patterns are collected for random orientations of that particle. The natural choice of a reciprocal space coordinate system for such an ensemble of collected diffraction patterns from a single particle or an aggregate of many particles completely in 3D random orientations is spherical rather than cylindrical irrespective of the size, shape and symmetry of the particle under study. One of the possible approaches for structure determination from an ensemble of such a large number of diffraction patterns is to average over the diffraction patterns the angular correlations of their intensities as described in chapter two (equation 2.1) [15]. An orientation-independent quadratic function of the spherical harmonic expansion coefficients $I_{LM}(q)$ may be extracted from the average over all measured diffraction pattern of angular autocorrelations of the measured intensiites for a resolution shell q.

For the case of helical viruses like 2TMV the azimuthal periodicity of the proteins along the helices or more elaborately 49_3 helical symmetry (as discussed in section 4.3) imposes selection criterion on the value of angular momentum quantum number L and hence the summation over azimuthal

quantum number M may be overcome if we limit the resolution of our reconstruction upto 12 A^0. As in equation 4.12;

$$\begin{aligned} I(\bar{q}) &= F(R, \psi, \zeta) F^*(R, \psi, \zeta) \\ &= \sum_{n,n'} G_n(R, \zeta) G_{n'}^*(R, \zeta) exp[i(n - n')\psi] \end{aligned} \qquad (4.27)$$

according to helix selection rule $n - n' = M$ can have only values $0, \pm49, \pm98$. If we limit oursalves for $M = 0$ then the value of L can only be upto $L = 48$ since the value value of azimuthal quantum number M must be equal or less than the angular momentum quantum number L. Hence within our resolution-restricted model the diffraction volume of intensity (not the amplitude !) of 2TMV is azimuthally symmetric (i.e.; the only permitted value of azimuthal quantum number $M = 0$). Hence equation (4.27) can be simplified as;

$$I(\bar{q}) = \sum_{n \epsilon \mathcal{S}} G_n^2(R, \zeta) \qquad (4.28)$$

where $\mathcal{S}$ is the set of all allowed integers as permitted by the selection rule.

The 3D reconstruction of reciprocal space intensity of a helical object may be found via calculating the spherical harmonic expansion coefficients of a single c repeat unit (TMV consists of roughly 45 c repeat unit each of 69 A^0 length). According to conventional wisdom the reconstruction of diffraction volume of scattered intensity [3] upto a reciprocal-space radius q_{max} of a macromolecule or virus of radius R may be obtained by using

angular momentum quantum number up to L_{max} where

$$q_{max} * R = L_{max} \qquad (4.29)$$

For TMV $R = 90A$; the limitation of $L_{max} < 49$ is approximately true for $q_{max} < 0.5\ A^{-1}$; i.e.; upto $12\ A^0$ Resolution.

4.6   "One-Term" Model of Cylindrical Harmonics for 2TMV

Cylindrical harmonics $G_n(R, \zeta)$ for 2TMV fall off rapidly with increasing n for a fixed layer line. The "one-term" model is valid at low resolution; $Res < 12A^0$ because the higher-order Bessel functions peak at high value of R as shown in figure (4.2) where we see the contribution of intensity $I(R, \zeta_\lambda)$ for various layer line from the first and the second Bessel term. Hence the "one-term" model [14] uses $G_{n_{lowest}}(R, \frac{\lambda}{c})$ for the intensity calculation of different layer lines. Figure (4.3) shows comparison of "One-Term" model with many term calculation via $I(R, \frac{\lambda}{c})$ plot for layer lines from $\lambda = 0\ to$ $\lambda = 5$. Clearly we see the validity of the one-term approximations with many term calculations except for $\lambda = 2$ which shows some deviation around $q_{max} = 0.45A^{-1}$

The scattered intensity $I(\bar{q})$ can be expanded in terms of spherical har-
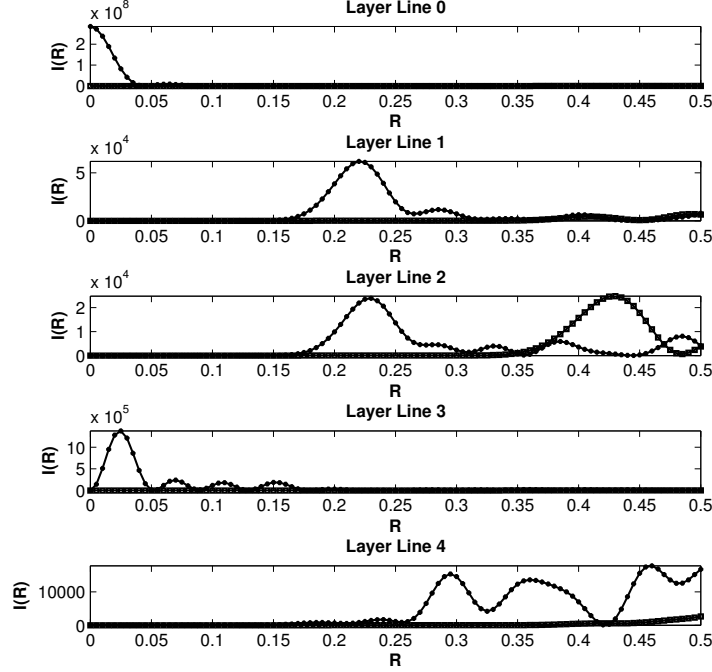
Figure 4.1: Intensity plot for various layer lines where contribution towards intensity from the maximally dominant order Bessel term (as per selection rule, table 4.1) shown by asterisk ($*$). Contribution from the next dominant order Bessel term is insignificant (except layer line 2) shown by square for $q < q_{max}$ corresponding to resolution less than $12\ A^0$.

monics

$$I(\bar{q}) = \sum_{L,M} I_{L,M}(q) Y_{L,M}(\theta, \varphi) \tag{4.30}$$

where $Y_{L,M}(\theta, \varphi)$'s are spherical harmonics. For our theory we are considering only $M = 0$ term and L ranges from 0 to 48. Hence the above equation can be written as

$$I(\bar{q}) = \sum_{L=0}^{48} \sqrt{\frac{2L+1}{4\pi}} I_{L,0}(q) P_{L,0}(cos\theta) \tag{4.31}$$
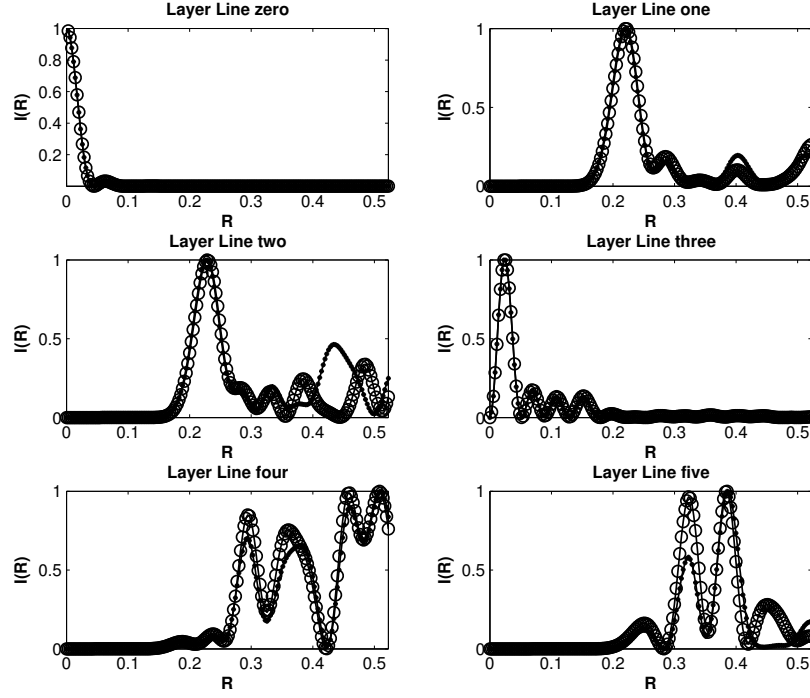
Figure 4.2: Comparison of calculated intensities for 2TMV for various layer lines using One-Term approximation (thick line) with many term calculation (thin line) via $I(R, \frac{l}{c})$ vs R plot.

where P_L is Legendre polynomial of order L.

Using (4.28) we can say more precisely

$$|G|^2(R, \zeta_\lambda) = \sum_{L=0}^{48} \sqrt{\frac{2L+1}{4\pi}} I_{L,0}(q) P_{L,0}(\frac{\zeta_\lambda}{q}) \tag{4.32}$$

Using the above equation $I_{L,0}(q)$ can be obtained by matrix inversion which is purely a orientation independent quantity as measured in so called "diffract-

and-destroy" XFEL experiment.

$$I_{L,0}(q) = |G|^2(R, \zeta_\lambda)[P'_{L,0}(\frac{\zeta_\lambda}{q})]^{-1} \tag{4.33}$$

where

$$P'_{L,0}(\frac{\zeta_\lambda}{q}) = \sqrt{\frac{2L+1}{4\pi}} P_{L,0}(\frac{\zeta_\lambda}{q}) \tag{4.34}$$

4.7 Recovering a fiber diffraction pattern from an XFEL "diffract-and-destroy" experiment

Diffraction patterns of completely randomly oriented helical molecules as measured in "diffract-and-destry" XFEL experiment [22] can be post processed to yield a "fiber diffraction" pattern via computational alignment. We recover the fiber diffraction pattern from the calculated $B_L(q)$ coefficients; an orientation-independent radial quantity which can be obtained from an experimental XFEL "diffract and destroy" experiment using correlation theory [15]. In a "diffract and destroy" experiment with an X-ray free electron laser, particles are injected into the beam in completely random orientations and hence one would not expect to see the layer line structure of the oriented helix in those diffraction patterns. Here we have shown that appropriate computational alignment of the ensemble of such diffraction patterns enable
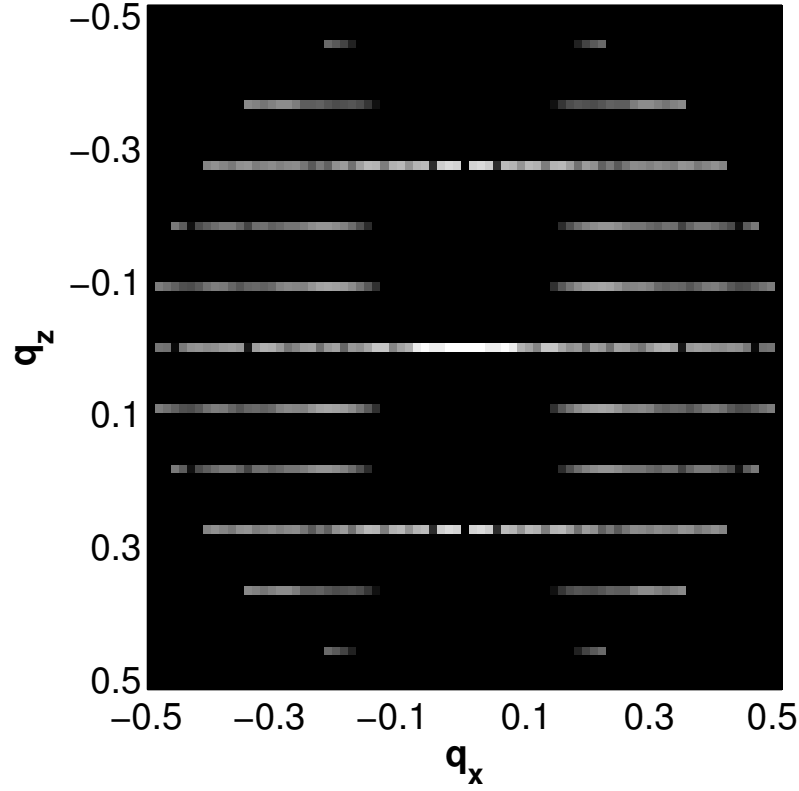
Figure 4.3: Fiber diffraction pattern of TMV obtained solely from the quantity expected to be measured from XFEL "diffract and destroy" experiment of single viruses completely in 3D random orientations

us to reconstruct such a fiber diffraction pattern as shown in Figure 5.4; thus

obviating the need to do it by experimental means such as forming fibres,

laser- or flow-alignment. Clearly we see the intensities along the layer lines;

everywhere else it is zero.

In order to judge the accuracy of the recovery of our fiber diffraction

pattern (figure 4.3); a comparison of maximum layer line intensities (scaled

to layer line 3) was made with a standard fiber diffraction pattern (figure 4.4) [11] and we found fairly close agreement (within $10 - 12\%$ variation) for all the layer lines except the first layer line (16% variation).

This variation in intensities for different layer lines is okay since our diffraction pattern is limited by a resolution of 12 A^0 whereas Namba's construction has a resolution 2.9 A^0. In contrast to the resolution comparison; a remarkable achievement in our computational construction is that this can be obtained from XFEL scattering without the need of uniaxial alignment such as forming fibers. It is worth mentioning here that until the discovery of the computational alignment technique [9] fiber diffraction was the only method for solving helical bio-structure such as DNA or TMV.
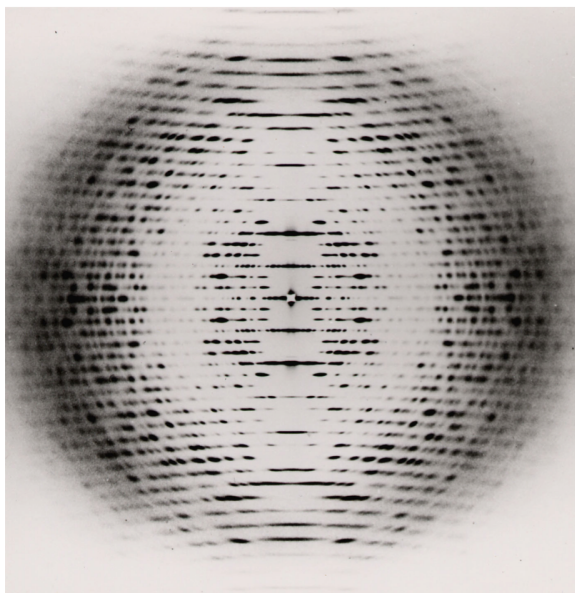
Figure 4.4: Fiber diffraction pattern from an oriented solution of tobacco mosaic virus [11].

4.8 Reconstruction of TMV diffraction volume

If $M = 0$ is the only term included in the expansion of the 3D intensity distribution of the scattered intensity of 2TMV; the diffraction volume up to the resolution limit can be written as

$$I(\vec{q}) = \sum_L I_{L0} Y_{L0}(\theta, \varphi) \tag{4.35}$$

where the magnitudes of the real expansion coefficients $I_{L0}(q)$ are determined as

$$|I_{L0}(q)| = \sqrt{B_L(q,q)} \tag{4.36}$$

and their signs be determined using the ring triple correlation function [26]. Once an oversampled diffraction volume is constructed from the $I_{L0}(q)$ coefficient; the 3D real-space electron density is obtained by using an iterative phasing algorithm. The magnitude of the expansion coefficients can be obtained from the average angular correlation of the measured diffraction pattern of their intensities and their signs from the ring triple correlation (RTC) function

$$C_3(q, \Delta\phi) = \int T_L(q) P_L(cos(\Delta\phi)) \tag{4.37}$$

where

$$T_L(q) = \sum_{L1L2} G(L_1 0; L_2 0; L0) I_{L_1 0}(q) I_{L_2 0}(q) I_{L_0}(q) \tag{4.38}$$
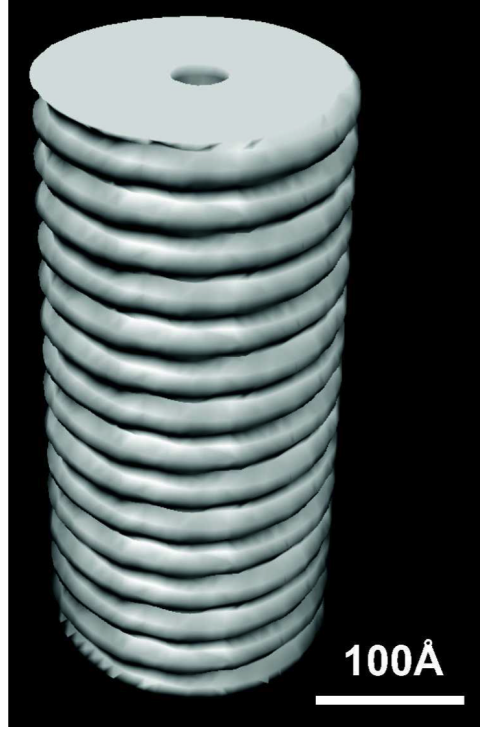
Figure 4.5: Real space image of a portion of tobacco mosaic virus (TMV) recovered from an oversampled low resolution 3D diffraction volume constructed only using the $M = 0$ term of the spherical harmonic expansion coefficient; a quantity expected to be recovered from an ensemble of XFEL diffraction patterns from random orientations of the virus.

where G is a Gaunt coefficient. Assuming $B_L(q)$ and $T_L(q)$ are the only quantities recovered from the XFEL "diffract and destroy" experiment, the spherical harmonic expansion coefficients $I_{L0}(q)$ can be recovered from (4.36).

From equation (4.38) it is clear that the magnitude of $T_{L0}(q)$ depends on the sign of $I_{L0}(q)$ since for each L, $T_{L0}(q)$ can be calculated for all possible sign combinations of $I_{L_1 0}(q)$ and $I_{L_2 0}(q)$ since the summation on the RHS of

equation (4.38) runs over L_1 and L_2 where

$$L_1 + L_2 > L > |L_1 - L_2| \qquad (4.39)$$

as per standard rule for the addition of angular momenta. Consequently those signs may be recovered by optimization of $T_L(q)$. The recovered $I_L(0)$ coefficients can be used to generate a low resolution (about 12 A^0) 3D diffraction volume as per (4.36). A section of this diffraction volume passing through the origin parallel to ζ axis is expected to be identical to a fiber diffraction pattern consisting of layer lines as shown in figure (4.3).

This result is remarkable. Using fiber diffraction TMV has been determined upto a resolution of 2.9 A^0 [31]. However attempts to align particles by all other ways always encountered the obstacle of the entropic tendency to disorder them at any finite temperature [10]. Here we have demonstrated our ability to produce a near-perfect fiber diffraction pattern [9] of 2TMV up to 12 A^0 resolution by post processing the diffraction patterns of completely randomly oriented particles in $3D$! [9].

In order to judge the feasibility of recovering the real-space image of helical particles via the recovery of a fiber diffraction pattern from an XFEL "diffract-and-destroy" experiment using the phasing technique used by fiber diffraction experts, we sent our diffraction pattern (figure 4.3) to Gerald Stubbs (Vanderbilt University); a well known fiber diffraction expert. Using our data as an input aided by the insertion of a heavy atom derivative;

the reconstructed real space image of 2TMV is shown in figure (4.6), where we see the enhancement of resolution due to post processing our data via insertion the heavy atom derivative.
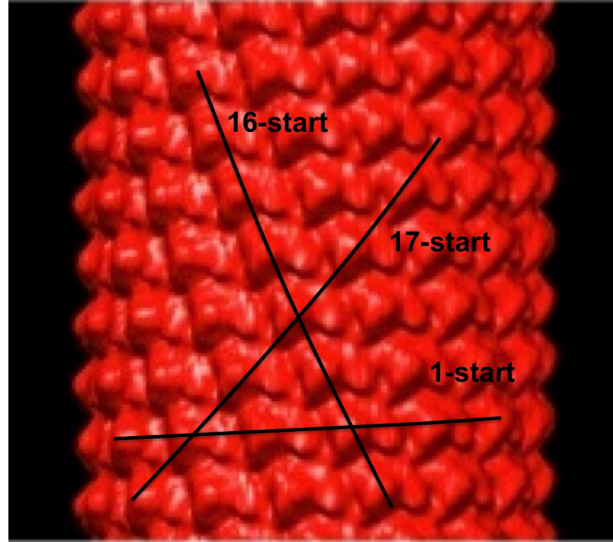
Figure 4.6: A portion of the reconstructed real space image of 2TMV using our recovered fiber diffraction pattern (figure 4.3) as an input aided by the simulated insertion of heavy atom derivative. Image reconstruction done by Gerald Stubbs (vanderbilt university).

Chapter 5

Reconstruction of PBCV from XFEL Ultrashort Pulses

5.1   Introduction

The discovery of X-rays by Wilhelm Rontgen at the end of 19th century may be regarded as one of the most important discoveries for the advancement in science. Following Max von Laue's discovery of X-ray diffraction by a crystal, William and Lawrence Bragg made it possible to calculate the positions of the atoms within a crystal from x-ray scattering data. However states of the biomolecules could be affected by the crystallization process and a significant fraction of biomolecules (as for example membrane proteins) cannot be crystalized at all. Understanding the functioning of those ultrasmall quantities may be greatly enhanced if structural studies were performed on individual

uncrystallized particles such as viruses. However most X-ray sources including synchrontons do not offer enough brilliance so that a single particle can be detected from scattering.

Other techniques for structural study of single nanoparticles like SAXS, cryo EM have their own drawbacks. SAXS aims to reconstruct the real space image from angularly averaged intensity scattered from a large number of biomolecules in solution through the variation of one radial parameter. On the other hand; electrons suffer multiple scattering and might not penetrate sufficiently to reveal the internal structure of a large virus. Hence the XFEL which delivers an intense beam which is roughly 10 billion times brighter [23] than a conventional synchronton source allows us the unique possibility of understanding the structure of the capsid as well as the internal genetic materials of the bio-molecule.

XFEL pulses are so intense that they could have destroyed the biomolecules and turned them into plasma. However simulation shows [22] that the extremely short duration of the pulse ($<$ 50 femtosecond; time in which light travels the distance of a tiny fraction of the width of human hair) can outrun radiation damage in the so called "diffract and destroy" experiment. In this work we report the 3D structure determination of the *paramecium bursarium chlorella virus* (PBCV) from some 800 diffraction patterns from the highly intense ultrashort pulses from the world's first x-ray free electron laser at LCLS (Stanford, California). This is the first XFEL

benchmark data set freely available to the public through the coherent x-ray imaging data bank (www.cxidb.org) for testing the theoretical approach, as well as algorithm development for 3D image reconstruction.

An aerosolized solution of PBCV viruses (which is roughly of icosahedral shape except the spike at one of the vertices of 5-fold symmetry axis along Z [6] were injected into the XFEL interaction region using an aerodynamic lens stack [25] and diffraction patterns [23] were collected from the back detector (roughly 740 mm from the interaction region). There are several challenges while constructing a real space image from an ensemble of a 2D diffraction patterns. Firstly; each diffraction shot is obtained from an individual identical particle with completely random orientation which is unkown. Unlike scattering from a periodic objects, the diffraction pattern from single particle scattering consists of a diffuse intensity distribution rather than intensity being concentrated on Bragg spots. To maintain similarity to the natural environment of the viruses in which they function, a solution of PBCV was injected into the LCLS atomic, molecular and optical science (AMO) beamline in the form of a tiny droplet of a buffer solution (mainly water) using the virtual gas dynamic nozzle developed by De Ponte et al [24]. Hence a significant amount of scattering is from the water droplet. The deposited set of diffraction data on coherent x-ray imaging data bank does not guarantee that the diffraction patterns arises from the scattering by a single virus. The set of roughly 800 diffraction patterns include a single particle diffraction pattern

as well as diffraction patterns from solvent droplets and sample aggregates of multiple viruses and water droplets [23]. Despite all those challenges here we are reporting the approximate recovery of icosahedral selection rule from experimental data.

5.2   Data classification and smoothing

Data as obtained from www.cxidb.org consists of roughly 800 diffraction pattern

Figure 5.1: Comparison of a typical single shot PBCV diffraction pattern (top: raw data as obtained from cxidb.org; botton: smoothed data).

One of the main challenges for constructing a real space image from single particle experimental diffraction pattern is the high signal to noise ratio. However, this can be overcome with a unique smoothing technique called "Contour Gaussian Smoothing" where each pixel of the diffraction pattern is weight avearged with a gaussian kernel by its neighboring pixels. Out of 800 diffraction pattern roughly 600 diffraction pattern were manually classified as diffraction patterns from PBCV and multiple aggregates of PBCV as per [23]. The original set of 800 diffraction patterns consists of patterns from

solvent droplets, sample aggregates or multiple particles including PBCV
and nanorice garins.
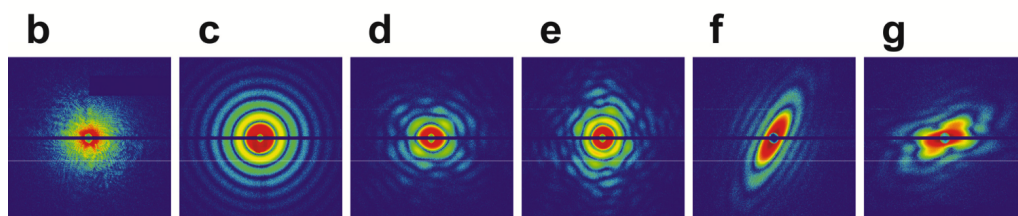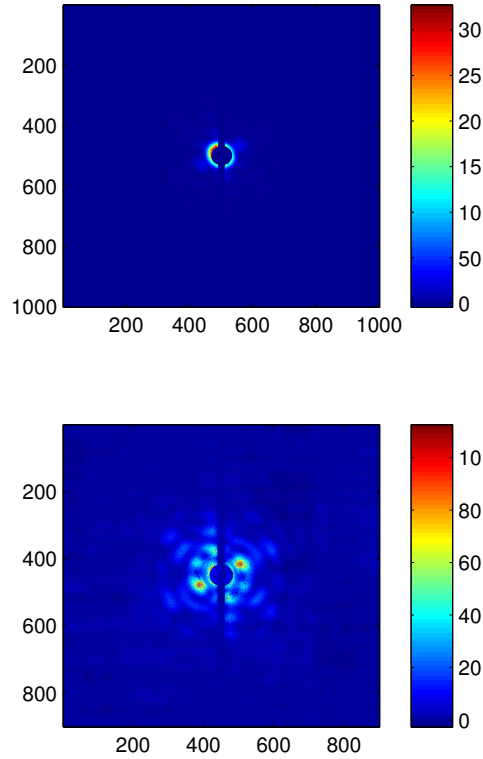

Figure 5.2: Comparison of a typical single shot PBCV diffraction pattern
(top: raw data as obtained from cxidb.org; botton: smoothed data). The
spike-like features which arise due to the flat side (hedra) of the virus in
water droplet. Smoothing recovers the features shaded by noise during XFEL
diffract and destroy experiment. Two images are shown on the scale.

5.3 Icosahedral selection rule

According to Caspar and Klug [2]; viruses are mainly of two basic shapes; icosahedral or helical. PBCV is mainly of icosahedral shape; except the spike along the 5-fold symmetry axis [6]; which breaks the icosahedral symmetry. Hence $B_L(q,q)$ obtained from experimental diffraction shots on PBCV which was in a water droplet (PBCV has a radius of 1000 A^0 and avearge estimate of the radius of the water droplet is roughly 0.5μ); we do not expect to see a perfect agreement of icosahedral selection criterion on the extracted $B_L(q,q)$ coefficients.
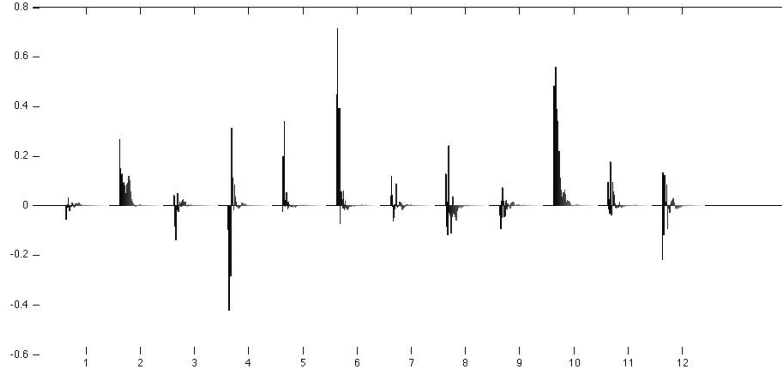
Figure 5.3: Plot of $B_L(q,q)$ vs L obtained from the average angular correlation of the diffraction pattern (roughly 800) of scattered intensities of paramecium bursarium chlorella virus (PBCV), which is totally random in its orientations exposed to XFEL single particle diffraction shot. Plot shows that the icosahedral component ($L = 0$ component is not shown on the plot; so that other components can be compared with each other) of the square of the expansion coefficients are dominant; i.e.; $B_L(q,q)$ for $L = 6$ is dominant than that of $L = 2, 4, 8, ..$ and $B_L(q,q)$ for $L = 10$ is dominant than that of $L = 8, 12, ...$ In the plot the $B_L(q,q)$'s for all the q values for a fixed L are plotted side by side.

As shown in figure (5.2); we see the non-icosahedral components ($B_L(q, q)$ for $L = 2, 4, 8, ..$) do not vanish completely; however the icosahedral components dominate the non-icosahedral components. Since the icosahedral selection rule is approximately satisfied in the $B_L(q, q)$ coefficients extracted from the experimental diffraction patterns; so we may predict that the scattered intensity is primarily from an icosahedral object. This prediction is also verified if we look at the $B_L(q)$ vs q plot for different L values. In figure (5.4)
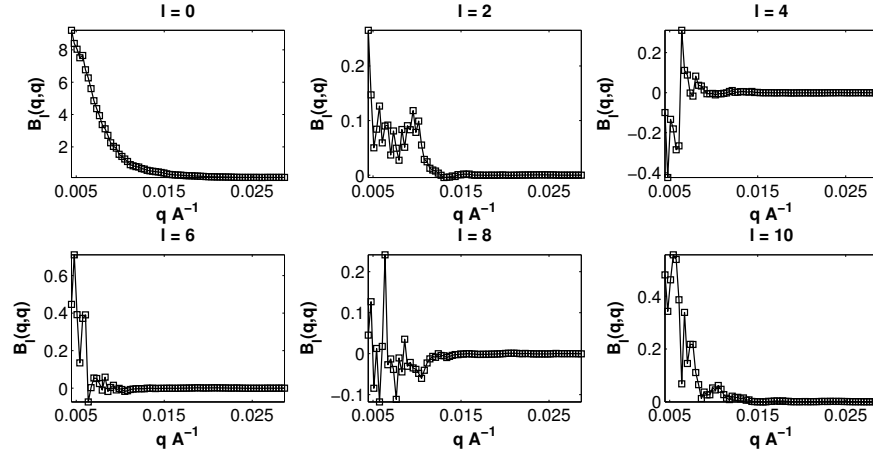
Figure 5.4: $B_l(q, q)$ vs q plot for $l = 0, 2, 4, 6, 8, 10$ obtained from the average angular correlation of the experimental diffraction pattern of scattered intensities

we see that $B_L(q, q)$ for $L = 6$ and $L = 10$ dominate compared to $L = 2, 4$ and $L = 8$.

Chapter 6

Discussion and Conclusion

Single particle "diffract and destroy" experiments with ultrashort and ultra-bright pulses of XFEL radiation are an important means of determining the structures of uncrystalized biomolecules. In contrast, X-ray crystallography, the leading technique for structure determination, requires large, high quality single crystalline samples. However, not all biomolecules (about 40%) can be crystalized [15]. Some membrane proteins or certain viruses are difficult to crystalize though they are a significant portion of today's drug targets. Although half a million proteins have been sequenced; structure of only about 10% have been determined so far (www.pdb.org). Hence our ability to determine the structure of certain biomolecules or viruses as well as the way they function in nature may be determined via uncrystalized particle scattering. This will constitute a breakthrough.

As the first XFEL user facilities are being made available, the very first

few experiments being done on relatively large objects such as viruses which produce low resolution diffraction patterns. In this work we developed a theory and computational method for solving virus structure by XFEL scattering from uncrystalized particles. With this method an ensemble of a large number of diffraction patterns obtained from the new generation of powerful x-ray free-electron lasers from a single particle or particles, completely random in their orientation, can be combined together to reconstruct a $3D$ Fourier map of the particle and then a phasing algorithm can be used to recover its full $3D$ image.

With our theory we can extract a quadratic function of the spherical harmonic expansion coefficients I_{LM}'s of the diffraction volume from a set of diffraction patterns of single molecule or molecular aggregates completely random in their orientation while XFEL radiation is diffracted from them and the intensity is recorded before they are destroyed by the radiation damage of the XFEL. The main challenge is to recover the I_{LM}'s coefficients from the quadratic function which includes a summation over the azimuthal quantum number M of the intensity distribution for a single particle for each angular momentum quantum number L. The maximum allowed values of angular momentum for our theory is limited by the size of the particle and the resolution of the real space image reconstruction, as discussed in chapter two. However if the particle under study has symmetry, so does the diffraction volume; we showed our ability to reconstruct $3D$ diffraction volumes and to recover the real space $3D$ electron density distribution of the molecule.

We mainly applied our method for solving virus structure since most of the virus structures have known symmetries. Viruses tend to have efficient designs to form protein coats with limited amount of genetic materials so that protein coats can be self assembled. As per Caspar and Klug [2] viruses tend to be primarily icosahedral (satellite tobacco mosaic virus (STNV) or chlorella virus (PBCV)) or helical (tobacco mosaic virus (TMV)).

We developed a test based on angular correlation to predict whether a virus is icosahedral or not. For icosahedral virus, the extracted quadratic function of expansion coefficients only survives for certain L values due to the point group symmetry of icosahedral particle. This selection rule in the radial quadratic function is a test to tell whether the set of diffraction patterns collected from a diffract-and-destroy XFEL experiment is primarily from an icosahedral molecule.

The intensity distribution of an icosahedral particle may be expanded in terms of icosahedral harmonics which is an azimuthal quantum number weighted linear combination of real spherical harmonics (details in chapter three) rather than expanding with usual spherical harmonics. Since for each value of the angular momentum quantum number L, the azimuthal sum of the square of the weights is unity (equation 3.7). So the azimuthal summation disappears in the radial quadratic function $B_L(q, q)$ of expansion coefficients. Hence the magnitudes of the expansion coefficients can simply be obtained by taking square roots of $B_L(q, q)$'s and sign can be determined via a minimization method. It is important to mention here that icosahedral harmonics

are non degenearte upto $L_{max} = 28$ which limits our resolution with our current method for real space image reconstruction in addition to the size of the particle. With all those contraints, we reconstructed the real space image of satellite tobacco necrosis virus (STNV) using $10,000$ simulated diffraction patterns [3] upto a resolution of 13 A^0 and the reconstructed image fits nicely with the model from the Protein Data Bank (PDB) coordinates as shown in figure (3.13).

So far fiber diffraction is the primary method for solving some of the helical structures in nature such as DNA or some of the helical viruses (such as 2TMV) where the helices need to be lined up along a uniaxial direction. However we have shown that a fiber diffraction pattern can be recovered from single particle XFEL diffract and destroy experiments [9] as highlighted as a Synopsis in Physics [18]: "No Need to Line Up". 2TMV consists of roughly 44 periodic c repeat units along the Z direction. X-ray scattering from an ensemble of such a periodic 2TMV uniaxial bundle (or fiber) gives rise layer lines in fiber diffraction pattern.

The natural selection of coordinate system for the expansion of layer line intensity should be cylindrical using cylindrical harmonics [13] as in the CCV theory [5]. In this work (chapter four) we have demonstrated that intensity distribution of 2TMV for q-th resolution ring may be expanded in terms of spherical harmonics and an orientationally independent quadratic function of expansion coefficients may be recovered from XFEL diffract and destroy scattering experiment from uncrystalized particles from completely randomly

oriented 2TMV helices thus obviating the need of uniaxial alignment. It is important to mention here that for our reconstruction of the diffraction volume the only allowed values of azimuthal quantum number $M = 0$ [9] provided if we limit the value of angular momentum quantum number upto $L_{max} = 48$. This is a consequence of the 49_3 helix of 2TMV which gives rise to the helix selection rule (equation 4.20; table 4.1). Once the radial quadratic function of expansion coefficients is known (since $M = 0$ is the only azimuthal quantum number; no summation over M), the magnitudes of the expansion coefficients can be obtained by taking the square root of $B_L(q)$s and their phases can be recovered by the ring triple correlation function [26], [9].

If the $M = 0$ term is the only allowed azimuthal quantum number and $L_{max} = 48$ for the reconstruction of diffraction volume and radius equals half the length of c repeat unit (34.5 A^0) along c axis and 100 A^0 along the perpendicular direction of c axis; the resolution for our real space reconstruction is 12 A^0 (equation 4.29).

Bibliography

[1] Poon, H. -C., Schmidt M. & Saldin D. K., (2013) *Advances in Condensed Matter Physics*, ID **750371**

[2] Caspar, D. L. D. & Klug A., (1962) *Cold Spring Herb Symposium; Quant Biol.*, A**27**, 1-24

[3] Saldin, D.K., Poon H.C., Schwander P. Uddin M. & Schmidt M. (2011) *Optics Express.*, A**19**, 17318-17335

[4] Cochran, W., Crick, F.H. & Vand, V. (1952). *Acta Cryst.* **5**, 581-586.

[5] Watson, J. D., & Crick F. H. C. (1953). *Nature* **171**, 738-740

[6] Namba K. & Stubbs G. (1986). *Science* **231**, 1401

[7] Millane R. P. (1991) *Acta Cryst* **A47**, 449-451

[8] *http://www6.slac.stanford.edu/news/2012-08-13-worlds-most-powerful-xray.aspx*

[9] Poon H. C., Schwander P., Uddin M., & Saldin D. K. (2013) *Phys. Rev. Lett.* **110**, 265505

[10] Spence J. C. H. *et. al.* (2005) *Acta. Cryst.* **A61**, 237

[11] Namba K. Pattanayek R. & Stubbs G. (1989) *J. Mol. Biol* **208**, 2, 307-325

[12] Millane, R.P., & Dorset, D.L. (2001). *International Tables for Crystallography*, Vol. B, 2nd ed., pp. 466-485. Dordrecht: Kluwer Academic Publishers.

[13] Saldin, D.K., , Starodub D. & Spence C.H. (2010) *Acta Cryst.*, A**66**, 32-37

[14] Barrett, A.N., Barrington Leigh J, Holmes K. C., Leberman R., Mandelkow E., & Von Sengbusch P. (1972) *Cold Spring Herb Symposium; Quant Biol.*, A**36**, 433-448

[15] Saldin, D.K., Shneerson V.L., Fung R & Ourmazd M. (2009) *J. Phys.: Condens. Matter*, A**21**, 17318-17335

[16] Jack, A., Harrison S.C. (1975) *J. Mol. Biol.*,**99**, 15-25

[17] the Wikipedia entry *http://en.wikipedia.org/wiki/Icosahedron*

[18] the APS Physics entry *http://physics.aps.org/browse?facets[arttype][]=synopsis*

[19] Varshalovich, D. A., Moskalev, A. N., & Khersonskii, V. K. (1988) *Quantum Theory of Angular Momentum*, **World Scientific**, p.148

[20] Davis & P. J., Rabinowitz, P.(1984) *Methods of Numerical Integration*, 2nd ed. **Dover Publications**, p.148

[21] Marchesini, S., (2007) *Rev. Sci. Instrum*, **78** p. 049901

[22] Neutze, R., Wouts, R., van der Spoel, D., Weckert, E. & Hajdu, J. (2000) *Nature*, **406**, 752-757

[23] Kassemeyer *et al.* (2012) *Optics Express*, **Vol. 20, Issue 4**, 4149-4158

[24] DePonte D. P., Weierstall, U., Schmidt, K., Warner, J., Starodub, D., Spence, J. C. H., & Doak R. B. (2008) *J. Phys. D: Appl. Phys.*, **41**, 195505

[25] Bogan *et. al.*; (2008) *Nano Letters*, **Vol. 8, No. 1**, 310-316

[26] Kam Z. (1980) *J. Theor. Biol.*, **82**, 15-39

[27] Xiao *et al.* (2005) *J. Mol. Biol.*, **353**, 493-496

[28] Cherrier *et al.* (2009) *PNAS*, **106**, 11085-11089

[29] Oszlayi *et al.* (2004) *Acta Cryst*, **A60**, 134-141

[30] Oszlayi *et al.* (2005) *Acta Cryst*, **A61**, 147-152

[31] Namba *et al.* (1989) *J. Mol. Biol.*, **208**, 307

[32] A database for icosahedral virus capsid structure *http://viperdb.scripps.edu*

Curriculum Vitae

Miraj Uddin

Place of Birth: Bangladesh

Education:

Ph.D., University of Wisconsin-Milwaukee, Aug 2013. Major: Physics

M.Sc., University of Kentucky, Lexington, Aug 2007. Major: Physics

M.Sc., University of Dhaka, Dhaka, Aug 2000. Major: Physics

B.Sc., University of Dhaka, Dhaka, Aug 1998. Major: Physics

Dissertation Title: Solving Virus Structures from XFEL Diffraction Patterns of Random Particle Orientations Using Angular Correlations of Intensities.

Publications:

Fiber diffraction without fibers; H.-C. Poon, P. Schwander, M. Uddin and D. K. Saldin; Physical Review Letter 110, 265505(2013).

Reconstructing an icosahedral virus from single particle diffraction experiments; D. K. Saldin, H. C. Poon, P. Schwander, M. Uddin and M.

Schmidt; Optics Express 19, 17318-17335(2011).

Electro-optic search for critical divergence of the charge-density-wave diffusion constant at the onset of depinning; L Ladino, EG Bittle, M. Uddin, JW Brill Physical Review B 76 (15), 153104(2007).

Dynamics of the electro-optic response of blue bronze $K_{0.3}MoO_3$
L Ladino, JW Brill, M. Freamat, M. Uddin, D Dominko; Physical Review B 74 (11), 115104(2006).

Dynamics of the electro-optic response of charge-density-wave conductors; JW Brill, L Ladino, M. Freamat, M. Uddin, RC Rai; Journal de Physique IV (Proceedings) 131, 135-138(2005).