

**3D RECONSTRUCTION OF PROTEINS AND VIRUSES
FROM ANGULAR CORRELATIONS OF THE
SCATTERED INTENSITIES**

by

Fatemehsadat Jamalidinan

A Dissertation Submitted in
Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
in Engineering

at

The University of Wisconsin-Milwaukee

August 2017

ABSTRACT

3D RECONSTRUCTION OF PROTEINS AND VIRUSES FROM ANGULAR CORRELATIONS OF THE SCATTERED INTENSITIES

by

Fatemehsadat Jamalidinan

University of Wisconsin-Milwaukee, 2017
Under the Supervision of Professor George Hanson

There is a remarkable shortage of the detailed knowledge of membrane proteins at atomic resolution despite the fact that they are the targets of many of today's drugs. The reason is that membrane proteins tend to have large hydrophobic surfaces which ensure their correct positioning in a membrane. However, this seems to make crystallization difficult, and this makes traditional methods of structure determination by X-ray crystallography difficult. In this thesis, we take advantage of this very fact to suggest an alternative method for structure determination by X-ray scattering of the projected structures of membrane proteins in their natural environments. Although in such environments the proteins are not perfectly aligned as in a crystal, we find that the algorithm suggested by Kurta and Pedrini appears to promise structure determination, perhaps down to atomic resolution.

We also suggest and develop how the method may be extended to obtain general (non-symmetric) 3D structures by exploiting the curved nature of Ewald spheres at lower energy. The extension of the 2D idea into 3D is straightforward, since a curved Ewald sphere also consists of a set of rings (one expects the different rings have different q_z components). We can get the intensities in 3D reciprocal space as that is exactly what we need for 3D structure recovery via a phasing program. Of course, the construction of intensity data on a uniform grid in 3D reciprocal space (required by a typical phasing program) requires a gridding program. The registry between the

$I_m(q)$'s on different q 's can be found as before if one knows $B_m(q_1, q_2)$ from the experiment, as can be found from a set of diffraction patterns of the same energy. Of course, varying the energy then gives us the 3D reciprocal space for a range of q_z 's, just what we need for getting info about the 3D structure.

In the second part of this thesis, we have reconstructed icosahedral images of the Coliphage PR772 and Rice Dwarf (RDV) viruses from the angular correlations of experimental data. We calculate the correlations using the standard method that Hanbury, Brown and Twiss developed in astronomy. The pattern of dominant icosahedral angular momentum quantum numbers that results is a strong indication of the icosahedral nature of the capsid. Having first determined by objective means that the structure of the capsid has icosahedral symmetry, we then recover a dodecahedral diffraction volume from which we correctly reconstruct an icosahedral structure using our phasing algorithm. We quantify the quality of the reconstructed image using the Fourier shell correlation curve of two independent datasets. For PR772, the FSC curve stays above 0.5 throughout the range of experimental data, which suggests that the resolution is still determined by the limitations of the experimental data rather than by the reconstruction method. For RDV, the resolution is around 200 Å. We also calculated an R_{split} quantity that compares two randomly split diffraction patterns for PR772 and RDV data and, as expected, they remained low. In a nutshell, three most important things to come out of this work are:

- 1-We recover the 2D structure of an individual membrane proteins up to atomic resolution using our suggested 2D phasing algorithm.
- 2-We develop an idea for producing 3D images using 2D diffraction patterns by combining multi-wavelength data from a soft X-ray fluctuation scattering experiment on membrane proteins partially oriented in a membrane, for the first time.
- 3- We also determine the the three-dimensional structure of PR772 and RDV viruses from experimental data, using our new 3D method.

©Copyright by Fatemehsadat Jamalidinan, 2017
All Rights Reserved

To my parents

TABLE OF CONTENTS

Abstract	ii
List of Figures	viii
Acknowledgements	xi
I Determining the Structure of Membrane Proteins In Situ	1
1 Introduction	2
2 Calculated Diffraction Patterns	5
3 Theory	12
3.1 Two Dimensional Method	13
3.1.1 Phase Retrieval	17
3.1.1.1 Charge-Flipping	18
3.1.1.2 Shrink-Wrap	19
3.1.2 Result and Validation	20
3.2 Three Dimensional Method	23
3.2.0.1 Multi-Tiered Iterative Projections (M-TIP)	24
3.2.1 Result and Validation	25
4 Conclusions	28

II Reconstruction of the PR722 and RDV Viruses from Experimental Data	30
1 Introduction	31
2 Instrumentation and Data Collection	34
3 Theory	36
3.1 Icosahedral and Icosahedral Symmetry	36
3.2 Angular Correlation	39
3.3 Phase Retrieval	42
4 Validation	66
5 Conclusion	70
Bibliography	72
Curriculum Vitae	76

LIST OF FIGURES

1.1	Membrane proteins are proteins that interact with, or are part of, biological membranes. They include integral membrane proteins that are permanently anchored to the membrane and peripheral membrane proteins which are only temporarily attached to the lipid bilayer or to integral proteins (Johnson and Cornell, 1999).	3
1.2	Schematic diagram of an experiment to determine the projected structure of a black lipid membrane. The X-rays are incident from the right and detected on the left.	3
2.1	Diffraction pattern $ A(q_x, q_y) $ of the K-channel protein viewed from a direction parallel to the pore. The structure factor F is calculated as usual from the structural data in a PDB file together with the Cromer-Mann representation of the atomic form factors.	6
2.2	Calculated diffraction pattern from 10 randomly oriented particles in an experiment performed in the configuration of Fig. 1.2.	7
2.3	Flat Ewald Sphere vs. Curved Ewald Sphere. At moderate x-ray energies which are more common with XFELs, the shape of the Ewald sphere becomes more and more curved and approaches the shape of a saucer. As the x-ray energies rise, the diffraction pattern becomes better and better approximated by a plane in reciprocal space.	8
2.4	Simulated diffraction pattern from x-rays incident down the central pore of a single K-channel protein molecule in log-scale (Left). Simulated diffraction pattern from 10 randomly positioned and oriented model K-channel proteins in log-scale(Right).	10

2.5	Simulated diffraction pattern from x-rays incident down the central pore of a single lactose permease in log-scale (Left). Simulated diffraction pattern from 10 randomly positioned and oriented lactose permease in log-scale(Right).	11
3.1	Light Microscopy versus Diffraction. For structure determination using X-ray diffraction, one needs to use a mathematical lens, Fourier Transformation (FT), to focus the scattered radiation (Egli, 2010).	17
3.2	The flowchart of the charge flipping method	19
3.3	The object fits inside its autocorrelation.	20
3.4	The flowchart of the Hybrid Input-Output (HIO) method	21
3.5	Single particle diffraction pattern recovered from the multi-particle diffraction pattern of Fig. 2.2. Note the similarity with Fig. 2.1	22
3.6	Projected structure of the K-channel protein found from an iterative phasing algorithm applied to the reconstructed diffraction pattern of Fig. 3.5	22
3.7	Phase-retrieval transfer functions as a function of inverse resolution $1/d$	23
3.8	3D projected structure of the K-channel protein found from MTIP algorithm applied to the reconstructed diffraction patterns. Note the similarity with Fig. 3.9 (Generated by Jeffrey Donatelli).	25
3.9	3D projected structure of the K-channel protein found from PDB (Generated by Jeffrey Donatelli).	26
3.10	3D projected structure of the lactose permease found from MTIP algorithm applied to the reconstructed diffraction patterns. Note the similarity with Fig. 3.11 (Generated by Jeffrey Donatelli).	26
3.11	3D projected structure of the lactose permease found from PDB (Generated by Jeffrey Donatelli).	27
2.1	A typical snapshot of PR772. The blue horizontal line is due to the gap.	35
2.2	A typical snapshot of RDV. The blue vertical line is due to the gap.	35

3.1	Platonic solids	37
3.2	Angular momentum decomposition of the angular correlations, $B_l(q, q)$. Since the diagonal values ($q = q'$) of B depend on the value of q in addition to l this plot is for B vs. l for a particular value of q (Pr772).	53
3.3	Angular momentum decomposition of the angular correlations, $B_l(q, q)$. Since the diagonal values ($q = q'$) of B depend on the value of q in addition to l this plot is for B vs. l for a particular value of q (RDV).	63
3.4	Diffraction volume of the PR722 virus.	63
3.5	Electron density of the PR772 virus.	64
3.6	Diffraction volume of the RDV virus.	64
3.7	Electron density of the RDV.	65
4.1	Fourier shell correlation (Pr772).	67
4.2	Fourier shell correlation(RDV).	67
4.3	Rsplitted calculation (Pr772).	68
4.4	Rsplitted calculation (RDV).	68

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Dilano Saldin for the continuous support of my Ph.D study and related research, for his patience, motivation, and immense knowledge. His guidance helped me during the researching and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

Besides my advisor, I would like to thank the rest of my thesis committee, Prof. George Hanson, Prof. Eddy Tysoe, Prof. Zeyun Yu, Prof. Yi Hu and Prof. Abbas Ourmazd for their insightful comments and encouragement, but also for hard questions which incentivized me to widen my research from various perspectives.

Last but not the least, I would like to thank my family, my father, my sisters and to my fiancée' for supporting me spiritually throughout writing this thesis and my life in general.

Part I

Determining the Structure of Membrane Proteins In Situ

Chapter 1

Introduction

Membrane proteins Fig.1.1 are proteins situated at the membrane of the cell or at the lipid bilayer. Although membrane proteins are the targets of most of today's drugs, the protein data bank (PDB) (Berman et al., 2000) has a remarkably small percentage of the atomic-scale structures of such proteins. This is primarily due to the fact that the premier method of protein structure determination, namely X-ray crystallography, requires the formation of crystals. This is difficult with molecules such as membrane proteins, which have hydrophobic surfaces which are integral to their functions. This very feature is the one responsible for the fact that they are found in nature within lipid membranes. This project points out that the very feature responsible for the difficulty of forming crystals may be exploited for a different method of structure determination in situ in a membrane making use of the properties of the newly developed X-ray free electron laser (XFEL) together with the algorithm described here for structure recovery. In short, if it is difficult to crystallize a membrane protein, why not produce copies of the protein in environments in which they are found naturally where one can exploit the advantage of scattering by a large number of such molecules even if they are not ideally aligned as in a crystal. In the case of membrane proteins in situ, one can try to exploit the fact that although they are not precisely aligned with respect to three mutually perpendicular axes, they are in a state of single-axis alignment. Up to now, an accurate theory of reconstruction of the diffraction pattern of a single particle from the diffraction patterns of multiple particles randomly oriented about a single axis normal to the membrane has not been exploited, although an earlier attempt was made that had

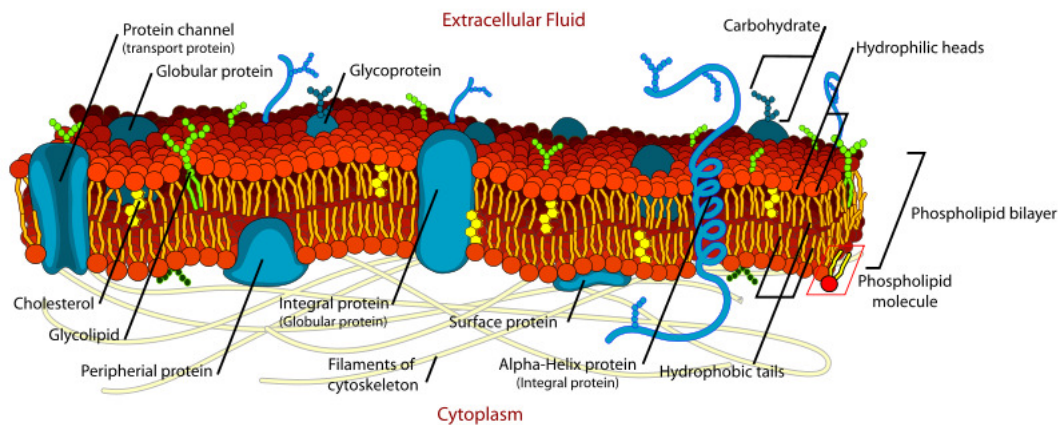


FIGURE 1.1: Membrane proteins are proteins that interact with, or are part of, biological membranes. They include integral membrane proteins that are permanently anchored to the membrane and peripheral membrane proteins which are only temporarily attached to the lipid bilayer or to integral proteins (Johnson and Cornell, 1999).

a much worse resolution (Saldin et al., 2010).

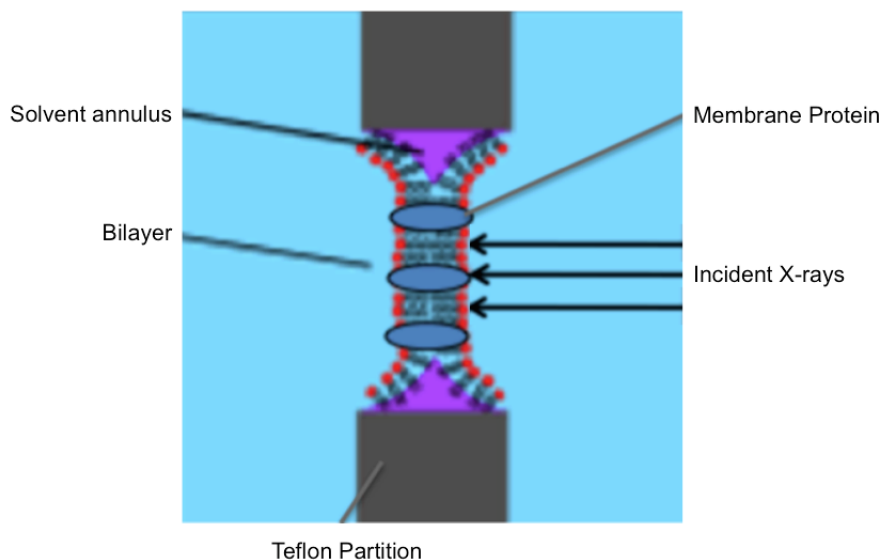


FIGURE 1.2: Schematic diagram of an experiment to determine the projected structure of a black lipid membrane. The X-rays are incident from the right and detected on the left.

A black lipid membrane is generally formed by painting a bilipid solution across a hole in a Teflon plate with a sable artist's brush. This looks like a soap film except in parts where it looks black due to the formation of a single bilipid membrane of width much smaller than a wavelength of light. This is a laboratory construction of a bilipid

membrane that is frequently found in nature. The function of an ion-channel protein is to transport ions between the inside and outside of a living cell through pores in the protein perpendicular to the membrane. The orientation of the protein about this axis is irrelevant for its function and is best considered random. Nevertheless, even in such a single-axis alignment, it is possible to increase the useful signal to determine the projected structure of a single protein by illuminating a large number of them simultaneously, as in a crystal, which we describe next. By generalizing the 2D theory involving the so-called 3-point triple correlations, we suggest and develop a fully 3D theory of recovering the structure of a particle in an x-ray free electron laser from the angular correlations of its scattered intensity, independent of its symmetry. We demonstrate this theory by recovering the 3D diffraction volume of symmetric K-channel protein (Alam and Jiang, 2009) and asymmetric lactose permease (Abramson et al., 2003) using the algorithm described in this part. We also calculated the structures directly from the PDB and compared the two. Remarkable agreement was found.

Chapter 2

Calculated Diffraction Patterns

The discovery of the structure and function of the K-channel membrane protein (Alam and Jiang, 2009) in the passage of ions between the inside and outside of a cell and its role in neurotransmission led to the 2003 Nobel Prize in chemistry, awarded to Roderick MacKinnon (MacKinnon et al., 1998). This is an ideal protein for our study since evolution has made the central pore perpendicular to the membrane, but there is no particular advantage to fixing the orientation of the molecule about the pore. Consequently, it should be regarded as being randomly oriented about this axis. This is ideal for studying randomness of particle orientations at a dimension lower to that in which they are found in a general SAXS experiment or in an experiment with an X-ray free electron laser, for example, in which the beam usually hits a molecule randomly oriented in $SO(3)$ which is the group of all rotations about the origin of three-dimensional Euclidean space R^3 under the operation of composition. This naturally gives rise to a form of single axis alignment ideal for the demonstration in this project.

We simulated the diffraction pattern due to a large number of molecules of random orientations by superposing single particle diffraction patterns. Strictly speaking for coherent radiation from an XFEL, one needs to add the scattered amplitudes from multiple proteins. Suppose the scattered amplitude (structure factor) of the j th protein at position \mathbf{r}_j is F_j . Then, the total scattered amplitude is $|A(\mathbf{q})|^2$ where

$$|A(\mathbf{q})|^2 = \left| \sum_j \mathbf{F}_j(\mathbf{q}) \exp(\mathbf{i}\mathbf{q}\cdot\mathbf{r}_j) \right|^2 = \sum_j |\mathbf{F}_j(\mathbf{q})|^2 + \sum_{j \neq k} \mathbf{F}_j^*(\mathbf{q}) \mathbf{F}_k(\mathbf{q}) \exp[\mathbf{i}\mathbf{q}\cdot(\mathbf{r}_j - \mathbf{r}_k)] \quad (2.1)$$

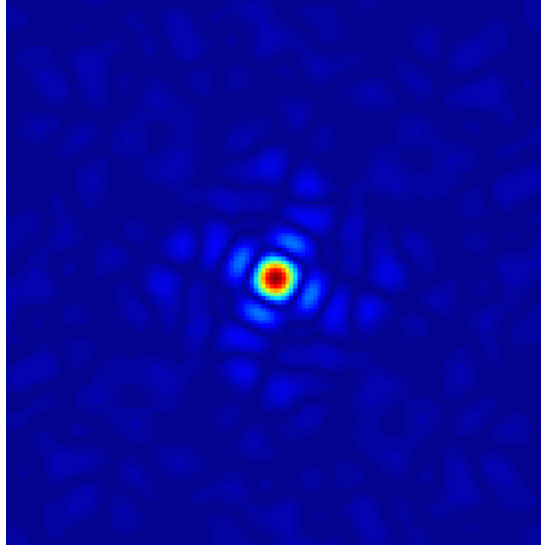


FIGURE 2.1: Diffraction pattern $|A(q_x, q_y)|$ of the K-channel protein viewed from a direction parallel to the pore. The structure factor F is calculated as usual from the structural data in a PDB file together with the Cromer-Mann representation of the atomic form factors.

The randomness of particle positions causes the second term above to become negligible compared to the first diagonal term over most of the q range. The sum in the first term is like a sum over the intensity contributions of the individual particles. Note that this is why it is generally unnecessary to consider the positions of multiple randomly positioned particles. The positions only correspond to random phases making contributions to only the cross terms amongst different particles, and which tend to become negligible for a large number of randomly positioned particles (as one would expect in the black-lipid membrane). The high, narrow coherent peaks in the correlations around the origin of reciprocal space (and its Friedel-related points) can usually be avoided on reconstruction. A plot of $|A(\mathbf{q})|$ vs \mathbf{q} is shown in Fig. 2.1 and may be thought of as constituting the diffraction pattern due to a single particle. Although the diffraction pattern from an individual protein may have structure (see Fig. 2.1) from which the projected structure of the protein may be recovered by an iterative phasing algorithm, the composite diffraction pattern is expected to look more like the set of concentric rings seen in Fig. 2.2. Nevertheless, as we see below, it is possible to reconstitute a diffraction pattern of the form of Fig. 2.1 from that of Fig. 2.2 using the angular correlation methods we describe in next section.

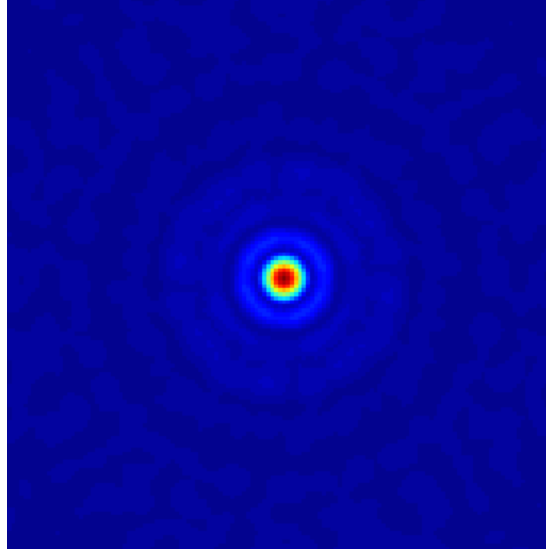


FIGURE 2.2: Calculated diffraction pattern from 10 randomly oriented particles in an experiment performed in the configuration of Fig. 1.2.

As we mentioned before, the structure of the K-channel protein (Alam and Jiang, 2009) is deposited in the Protein Data Bank, PDB, (Berman et al., 2000). We were thus able to calculate a diffraction pattern of a single K-channel protein (Alam and Jiang, 2009) for x-rays incident parallel to the pore. Such a pattern is, of course, diffuse, with no Bragg spots, and is shown in Fig. 2.1. In our previous work, we calculated this pattern to about 6 Å resolution. Due to the more ambitious aim of the present work, we simulated this diffraction pattern to 1.5 Å resolution.

In a 2D case, under assumption of a flat Ewald sphere (Fig. 2.3), the structure solution process will yield the projected density of the molecule in question up to an arbitrary orientation. While the determination of a 2D projection of a structure is of academic interest, relevant biological interpretations can only be obtained from 3D models. In this case, we obtain a 3D structure from partially aligned molecules using a multi-wavelength strategy because for a curved Ewald sphere (Fig. 2.3), q is always 3D. Each 2D image samples the 3D diffraction volume via the following formula

$$q = \text{sqrt}(q_x^2 + q_y^2 + q_z^2) \quad (2.2)$$

in which

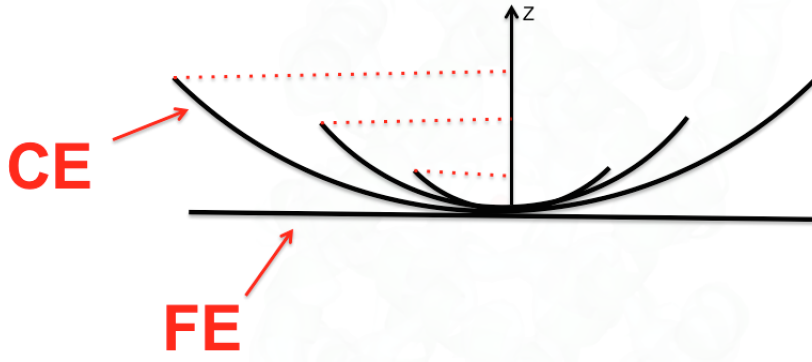
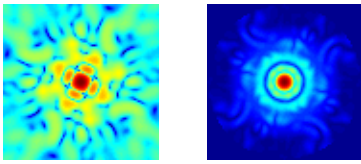


FIGURE 2.3: Flat Ewald Sphere vs. Curved Ewald Sphere. At moderate x-ray energies which are more common with XFELs, the shape of the Ewald sphere becomes more and more curved and approaches the shape of a saucer. As the x-ray energies rise, the diffraction pattern becomes better and better approximated by a plane in reciprocal space.

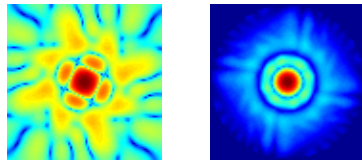
$$\begin{aligned}
 q_x &= 2\pi \times (x/\text{sqrt}(x^2 + y^2 + 1))/\lambda \\
 q_y &= 2\pi \times (y/\text{sqrt}(x^2 + y^2 + 1))/\lambda \\
 q_z &= 2\pi \times (1/\text{sqrt}(x^2 + y^2 + 1) - 1.0)/\lambda
 \end{aligned} \tag{2.3}$$

where x and y are the pixel coordinates, and λ is wavelength.

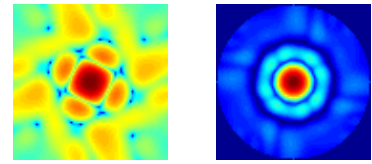
In order to test the proposed procedure, synthetic data was constructed from the symmetric K-channel protein (Alam and Jiang, 2009) and asymmetric lactose permease (Abramson et al., 2003). The assumed experimental geometry was an instrument capable of recording X-rays up to a scattering angle of 80 degrees. The data was computed down to 4 resolution. In total, 20 wavelengths between 0.2 to 2.5keV (0.2, 0.21, 0.23, 0.24, 0.25, 0.26, 0.28, 0.29, 0.31, 0.33, 0.36, 0.42, 0.45, 0.5, 0.56, 0.62, 0.71, 0.83, 1, 1.25, 1.67 and 2.50 keV.) were used in the simulation. The simulated diffraction patterns corresponding to the single particle and multi-particle of a K-channel protein (Alam and Jiang, 2009) and lactose permease (Abramson et al., 2003) are shown, respectively, in Fig. 2.4 and Fig. 2.5 .



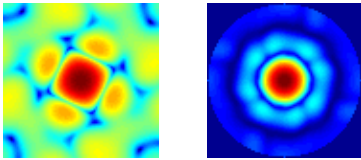
(A) E=0.2 Kev



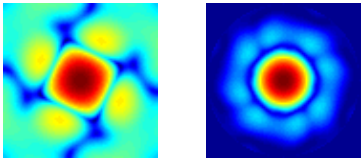
(B) E=0.21 Kev



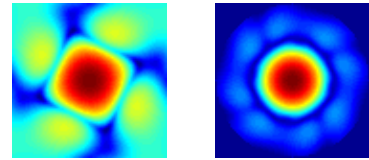
(C) E=0.23 Kev



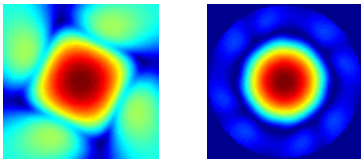
(D) E=0.24 Kev



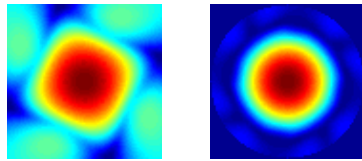
(E) E=0.25 Kev



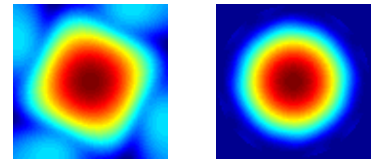
(F) E=0.26 Kev



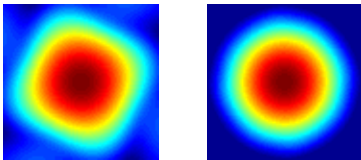
(G) E=0.28 Kev



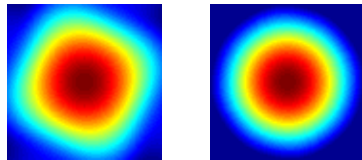
(H) E=0.29 Kev



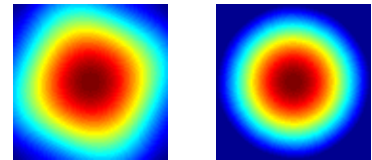
(I) E=0.31 Kev



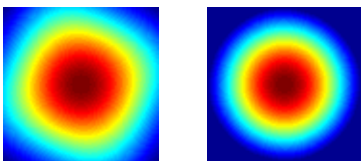
(J) E=0.33 Kev



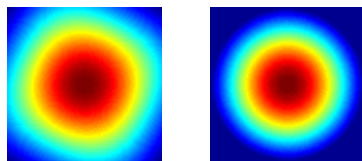
(K) E=0.36 Kev



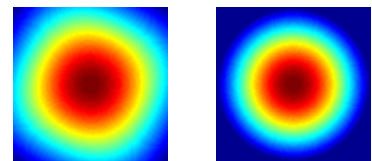
(L) E=0.42 Kev



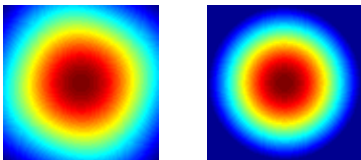
(M) E=0.45 Kev



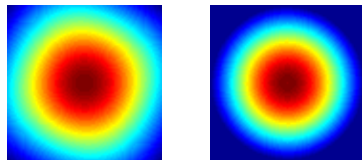
(N) E=0.5 Kev



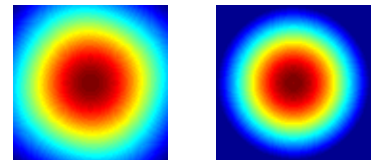
(O) E=0.56 Kev



(P) E=0.62 Kev



(Q) E=0.71 Kev



(R) E=0.83 Kev

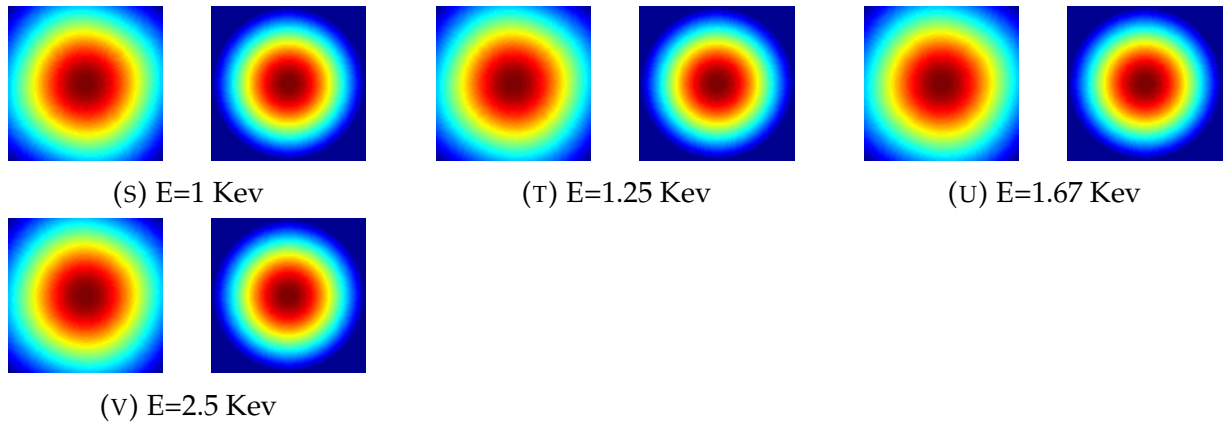
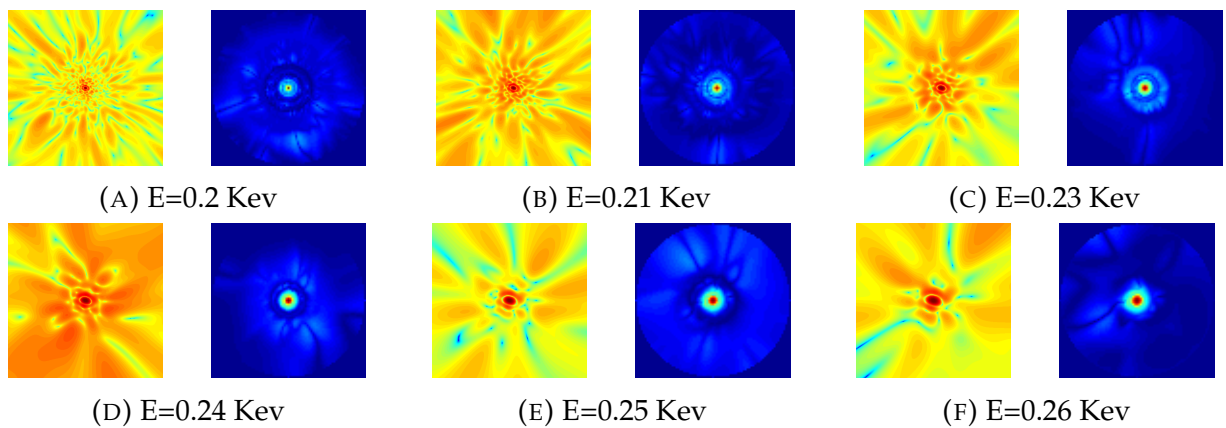


FIGURE 2.4: Simulated diffraction pattern from x-rays incident down the central pore of a single K-channel protein molecule in log-scale (Left). Simulated diffraction pattern from 10 randomly positioned and oriented model K-channel proteins in log-scale(Right).



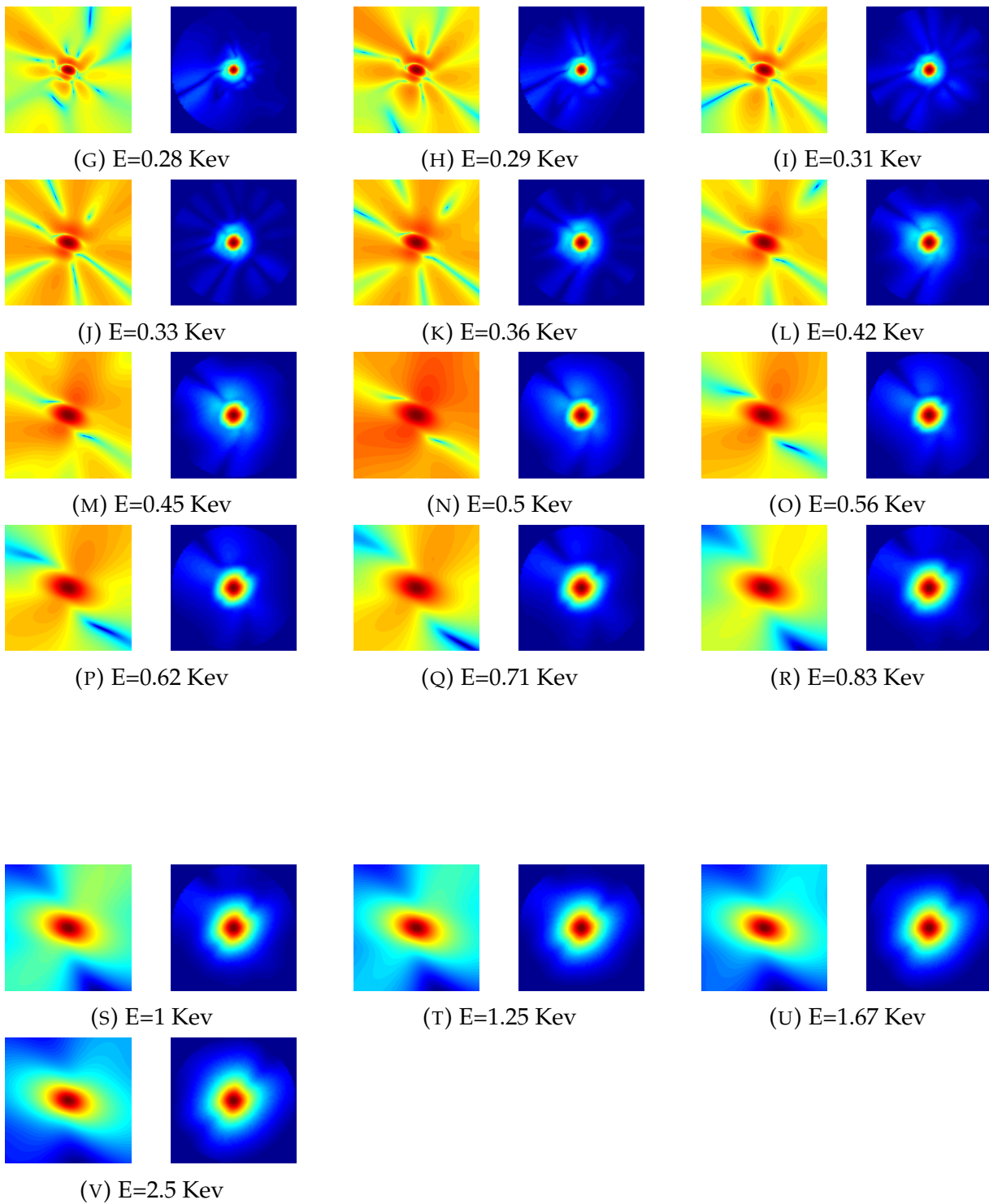


FIGURE 2.5: Simulated diffraction pattern from x-rays incident down the central pore of a single lactose permease in log-scale (Left). Simulated diffraction pattern from 10 randomly positioned and oriented lactose permease in log-scale(Right).

Chapter 3

Theory

We have earlier reported (Saldin et al., 2010) that it is possible to recover a single particle diffraction pattern like the one above from multiple proteins randomly oriented about an axis by the use of a phasing algorithm that constrains the solution to measured angular correlations and a positivity condition on the reconstructed intensities. An iterative phasing algorithm was then able to find from this single particle diffraction pattern the projected density of the protein. The new development since (Saldin et al., 2010) is the proposal of a closed form expression for both the amplitudes and phases of the circular harmonic coefficients to determine a single particle diffraction pattern from the so-called three-point triple correlations by Kurta et al. (Kurta et al., 2013) and Pedrini et al. (Pedrini et al., 2013). Of course, the use of three-point triple correlations requires simultaneous non-zero intensities in three pixels which may be unlikely in low-intensity diffraction patterns. However since the pair correlations in question are the result of averaging over many azimuthal angles and over many diffraction patterns, the method may be feasible in practice, especially at powerful XFELs like the European XFEL due to come online in 2017.

The design specification for the focal diameter of X-rays from the Linac Coherent Light Source (LCLS), the world's first XFEL, which commenced operation a few years ago, is ~ 100 nm. A K-channel protein has a width of about 5 nm, a factor of some 20 times smaller. Consequently, one could increase the measured signal by scattering a large number of K-channel proteins, which in principle, may be found within the focus of an XFEL beam. The problem is that, unlike in a crystal, the proteins are not likely in

the same rotational state about the normal to the membrane. We show in this research that this is of no consequence, as the rotational average overall diffraction patterns of the angular correlations is the same for particles of random orientations.

In X-ray crystallography, one exploits scattering from a large number of molecules to enhance the weak scattering of X-rays for structure determination. In that application, all protein molecules have to be identically oriented. We point out that through the evaluation of angular correlations, perfect alignment is unnecessary to exploit the advantage of increased scattering from an ensemble of identical particles.

3.1 Two Dimensional Method

For a set of experimental diffraction patterns, the average pair correlation function may be defined by

$$C_2(q_1, q_2; \Delta\phi) = \langle J(q_1, \phi) J(q_2, \phi + \Delta\phi) \rangle_\phi \quad (3.1)$$

Where q_1 and q_2 refer to two resolution shells. J refers to intensity. The $\langle .. \rangle_\phi$ symbol means averages are taken over all values of ϕ .

Where the molecules are randomly oriented about an axis normal to the membrane (and parallel to an incident X-ray beam) each particle can be thought of as giving rise to identical diffraction patterns, just randomly oriented about the surface normal. If one adds together many such randomly oriented but otherwise identical diffraction patterns, the resulting pattern looks something like Fig. 2.2, namely a set of rings of intensity. If the aim is a determination of the projected structure of a single particle, a first step would be a determination from this multi-particle diffraction pattern of a single particle diffraction pattern. Although we have shown this in principle before (Saldin et al., 2010) the particular method of structure recovery that we used involved two consecutive iterative phasing algorithms (the first to determine the single particle diffraction pattern from the multiple particles, and the second to determine the projected structure

from the single particle diffraction pattern). The use of two consecutive phasing algorithms introduces errors in the single particle diffraction pattern which magnify errors in the second step, namely the determination of the projected structure of the protein from the reconstructed single particle diffraction pattern. The subsequent work of Kurta et al. (Kurta et al., 2013) and Pedrini et al. (Pedrini et al., 2013) showed how it is possible to determine the single particle diffraction pattern essentially analytically from the angular correlations. This gives rise to much fewer numerical errors in reconstructing the protein in real-space, as we now demonstrate. Indeed the resolution of such an image is great enough that one can essentially see the projection of atomic resolution features. This also shows that despite the fact that there are multiple particles, in this case, there does not seem to be a fundamental resolution limit of such correlation methods. Likewise one can define, as did Kurta et al. (Kurta et al., 2013) and Pedrini et al. (Pedrini et al., 2013), a three-point triple-correlation function by:

$$C_3(q_1, q_2, q_3, \Delta\phi_1, \Delta\phi_2) = \langle J(q_1, \phi)J(q_2, \phi + \Delta\phi_1)J(q_3, \phi + \Delta\phi_2) \rangle_\phi \quad (3.2)$$

The polar coordinates of the three points in question are given by the arguments of the intensities I , and the symbol $\langle \dots \rangle_\phi$ stands for averaging all values of ϕ .

It should be noted that the angular correlations themselves (3.1) are between intensities on the same diffraction patterns, and thus insensitive to shot-to-shot intensity variations, a problem with an X-ray laser. However, an average over the same diffraction patterns is much less sensitive to shot-to-shot variations than the correlation between intensities of different diffraction patterns. All that is required is a reasonable statistical distribution of the shot-to-shot intensity variations. It should perhaps be pointed out that this method of determining a diffraction volume from random particle orientations works from averages over all particles of different orientations and at no stage involves the determination of the relative orientations of the particles contributing to individual

diffraction patterns. From the quantities C_2 one may compute

$$B_m(q_1, q_2) = \int C_2(q_1, q_2, \Delta\phi) \exp(-im\Delta\phi) d(\Delta\phi) \quad (3.3)$$

Note that B_m is independent of $\Delta\phi$, and can be shown to be equivalent to

$$\langle B_m(q_1, q_2) \rangle_{DP} = I_m^*(q_1) I_m(q_2) N \quad (3.4)$$

in which N is the number of particles and the symbol $\langle \dots \rangle_{DP}$ stands for averaging over a large number of diffraction patterns and m is quantum number. Also define

$$T_{m_1, m_2}(q_1, q_2, q_3) = \int C_3(q_1, q_2, q_3, \Delta\phi_1, \Delta\phi_2) \exp(-im_1\Delta\phi_1) \exp(-im_2\Delta\phi_2) d(\Delta\phi_1) d(\Delta\phi_2) \quad (3.5)$$

Likewise, it can be shown that

$$\langle T_{m_1, m_2}(q_1, q_2, q_3) \rangle_{DP} = I_{m_1+m_2}^*(q_1) I_{m_2}(q_3) I_{m_1}(q_2) \langle A_{m_1, m_2} \rangle_{DP} \quad (3.6)$$

where

$$A_{m_1, m_2} = \sum_{i, j, k}^N \exp[i(m_1 + m_2)\psi_i - m_1\psi_j - m_2\psi_k] \quad (3.7)$$

The statistical average of $\langle A_{m_1, m_2} \rangle_{DP}$ in (3.7) tends toward N (number of particles) for a sufficiently large number of diffraction patterns (Kurta et al., 2013).

It will be noted that both B_m and T_{m_1, m_2} are functions of the circular harmonic expansion coefficients $I_m(q)$ of the expansion of a single particle diffraction pattern, namely

$$I(q, \phi) = \sum_m I_m(q) \exp(im\phi) \quad (3.8)$$

Thus reconstructing a single particle diffraction pattern is tantamount to finding the amplitudes and phases of these coefficients. Note that the same principle applies to the determination of a single particle diffraction volume from a set of diffraction patterns of 3D random particle orientations in an X-ray free electron laser (XFEL), except that in that

case one seeks to determine the coefficients of a spherical harmonic expansion, which in turn determines the intensities on a shell rather than a circle in reciprocal space.

Remarkably, the above equations show that the measurements of C_2 and C_3 from the ensemble of diffraction patterns of random particle orientations is sufficient to determine both the amplitudes and phases of the $I_m(q)$ coefficients. As we will further show, this is sufficient to determine the projected structure of a particle to atomic resolution from a diffraction pattern consisting of circular rings as in Fig. 2.2

$$|I_m(q)| = \sqrt{B_m(q, q)/N} \quad (3.9)$$

This has been known for a while. What the recent work of Kurta et al. (Kurta et al., 2013) and Pedrini et al. (Pedrini et al., 2013) showed is that it is also possible to find the phases of the $I_m(q)$ coefficients from the B 's and the T 's. It will be seen from the expression of the T 's (3.9) that if any two of the expressions for I are known in amplitude and phase, along with the value of T itself, it is possible to find the third from Eq. (3.9).

The determination of such coefficients can, therefore, proceed as follows: first, choose a reference circle of radius q_{ref} . In the case of the 4-fold symmetric K-channel protein, one only needs values of the expansion coefficients with m modulo 4. Suppose the phase of $I_m(q)$ of the azimuthal quantum number m and the ring of radius q is $\gamma(m, q)$, then we can start with the fact that $\gamma(4, q)$ only determines the arbitrary orientation of the reconstructed image and $\gamma(0, q) = 0$. One then makes use of the equation that

$$\arg(T_m(q, q)) = \gamma(4, q) + \gamma(m, q) - \gamma(m + 4, q) \quad (3.10)$$

To find the phases for arbitrary m , one can then propagate the solution to all resolution rings using the properties of the pair correlations. Since the phases of the LHS are known from the experiment, the phases of the remaining γ 's of the RHS may be found with an obvious recursion relation for the reference shell q_{ref} . This will determine the value of the coefficient with $m = 8$. In the next iteration, one can make $m_2 = 8$, while

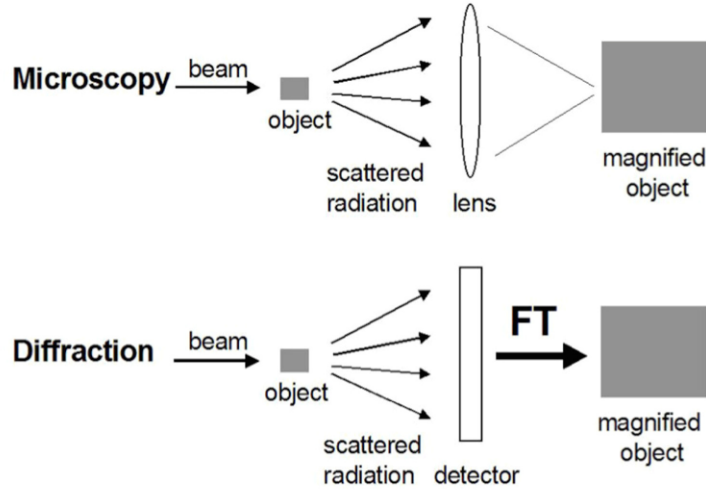


FIGURE 3.1: Light Microscopy versus Diffraction. For structure determination using X-ray diffraction, one needs to use a mathematical lens, Fourier Transformation (FT), to focus the scattered radiation (Egli, 2010).

keeping the $m_1 = 4$. This will give the coefficients for $m= 12$ and others. By these means one can determine the phases of all the relevant $I_m(q_{ref})$'s of the reference ring q_{ref} . Then the values of γ for the same m but different q 's can be found from Eq. (3.8). This way one can determine the phases of all the resolution rings and all values of m and q . Since the magnitudes are known, one then knows everything about all the $I_m(q)$ coefficients needed to reconstruct a single particle diffraction pattern (Fig. 3.5) from the angular correlations that can be found even from the multi-particle diffraction pattern (Fig. 2.2).

3.1.1 Phase Retrieval

Unlike light or electron beams, X-rays cannot be focused. This is the main difference between optical or electron microscopy and X-ray diffraction (Fig. 3.1). Therefore, the X-ray crystallographic and diffraction visualization of a particle requires a mathematical lens (Fourier transformation) that generates a 3D structure from the amplitudes of the scattered radiation (the structure factors) and the phases (Egli, 2010). We find the amplitude of scattered radiation via angular correlation technique but the phase information is lost in the diffraction experiment. Fortunately, several methods allow one to recover the phases. The most famous and widely used iterative techniques for solving the phase

problem are Error Reduction (ER) (Fienup, 1982) , Hybrid Input-Output (HIO) (Fienup, 1982) , Solvent Flipping (SF) (Bauschke, Combettes, and Luke, 2002) , Averaged Successive Reflection (ASR) (Elser, 2003) , Random Averaged Alternating Reflector (RAAR) (Luke, 2004), Charge Flipping (Oszlányi and Sütő, 2004), Shrink-Wrap (Marchesini et al., 2003). Here, we recovered the projected electron density using the combination of the charge-flipping (Oszlányi and Sütő, 2004) and shrink-wrap algorithm (Marchesini et al., 2003).

3.1.1.1 Charge-Flipping

This method was proposed by (Oszlányi and Sütő, 2004). Charge flipping operation is the key part of the charge flipping algorithm. In this method, all scattering density pixels with density lower than a small positive threshold δ are multiplied by -1 (flipped) and the rest of the pixels are unchanged. Fig. 3.2 shows the flowchart of the charge-flipping method. In the first step, the user should assign random starting phases to all experimental amplitudes and make all unobserved amplitudes equal to zero (Kolb et al., 2012).

$$F^{(0)}(h) = \begin{cases} |F^{(obs)}(h)| \exp(i\phi_{rand}(h)) & \text{if } |F^{(obs)}(h)| \text{ is known} \\ 0 & \text{otherwise} \end{cases} \quad (3.11)$$

The iteration then proceeds as follows:

- 1-Using the inverse Fourier transform of $F^{(n)}$, one can calculate the electron density $\rho^{(n)}$.
- 2-By flipping the density of all pixels with density values below a small positive threshold δ and keeping the rest of the pixels unchanged, one can obtain the modified density $g^{(n)}$.

$$g_i^{(n)} = \begin{cases} \rho_i^{(n)} & \text{if } \rho_i^{(n)} > \delta \\ -\rho_i^{(n)} & \text{otherwise} \end{cases} \quad (3.12)$$

- 3- Using Fourier transform of $g^{(n)}$, one can calculate the temporary structure factors.

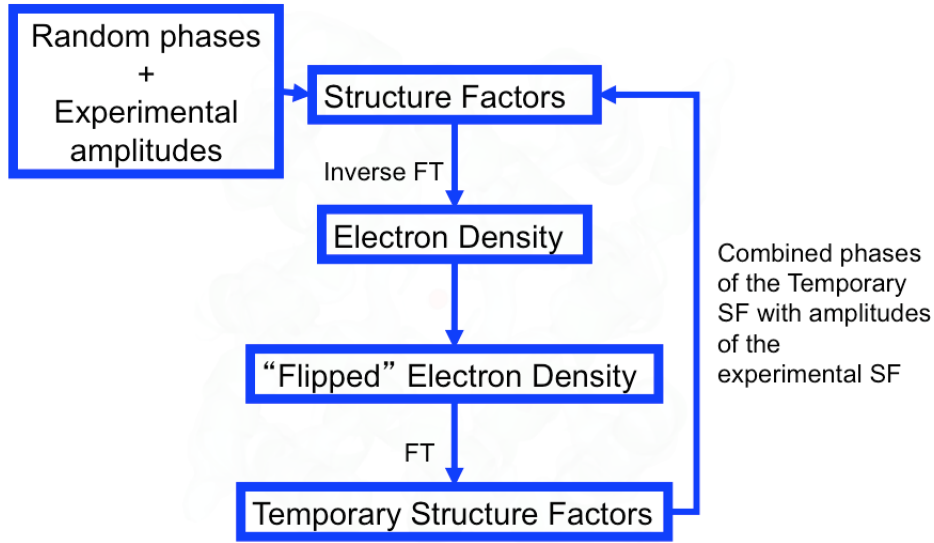


FIGURE 3.2: The flowchart of the charge flipping method .

$$G^{(n)}(h) = |G^{(n)}(h)| \exp(i\phi_G(h)) \quad (3.13)$$

4- New structure factors $F^{(n+1)}$ are obtained by combining the experimental amplitudes with the phases ϕ_G and setting all non-measured structure factors to zero.

$$F^{(n+1)}(h) = \begin{cases} |F^{(obs)}(h)| \exp(i\phi_G(h)) & \text{if } |F^{(obs)}(h)| \text{ is known} \\ 0 & \text{otherwise} \end{cases} \quad (3.14)$$

5-One can go to the next cycle of iteration using the modified structure factors.

3.1.1.2 Shrink-Wrap

This method was proposed by (Fienup, 1982). It tries to find support dynamically during iterations. One can obtain the autocorrelation function of the object by inverse Fourier transforming the diffraction pattern. Since the object must fit within the autocorrelation function (Fig. 3.3), the mask obtained from this function is a nice first estimation of the support. Then one can apply the HIO algorithm (Fienup, 1982) with

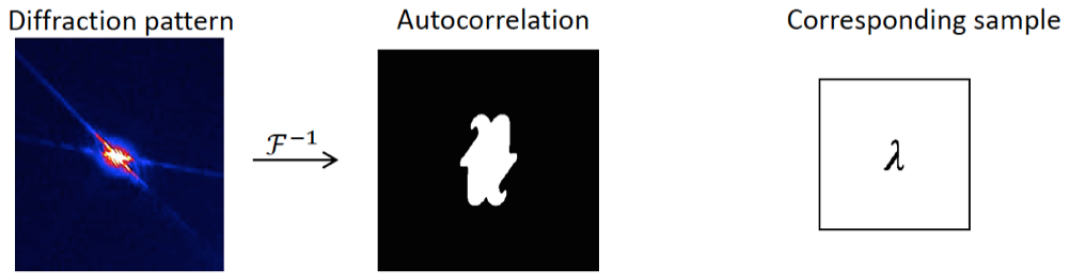


FIGURE 3.3: The object fits inside its autocorrelation.

relaxation parameter $\beta = 0.9$. To find the new support mask, every 20 iterations, the reconstructed image is convolved with a Gaussian of width $FWHM = 2.3548$. The mask is then obtained by applying a threshold at 20% of its maximum.

3.1.1.2.1 Hybrid Input-Output (HIO)

This method was proposed by (Fienup, 1982). It is a modification of the error reduction algorithm for retrieving the phases. Fig.3.4 shows the flowchart of the charge-flipping method. The procedure is the same as charge flipping; with this difference, here one can use some kind of support (mathematics) to iteratively calculate the phases. Assume $\rho(n)$ is a trial scattering density in the n th iteration cycle. One can obtain $\dot{\rho}(n)$ from $\rho(n)$ applying Fourier-transforming $\rho(n)$. Now, one can replace all Fourier amplitudes by the experimentally observed amplitudes, and apply inverse Fourier transform. Then the density $\rho(n+1)$ is defined pixel-wise by the following formula:

$$\rho_i^{(n+1)} = \begin{cases} \dot{\rho}_i^{(n)} & \text{if pixel } i \text{ belongs to the support} \\ \rho_i^{(n)} - \beta \dot{\rho}_i^{(n)} & \text{otherwise} \end{cases} \quad (3.15)$$

in which β is a fixed real relaxation parameter.

3.1.2 Result and Validation

In this project, we used a value $\beta = 0.9$. The progress of the iterations can be monitored by use of the R factor, calculated from the modulus of the Fourier transform of the current estimate of image and known Fourier modulus.

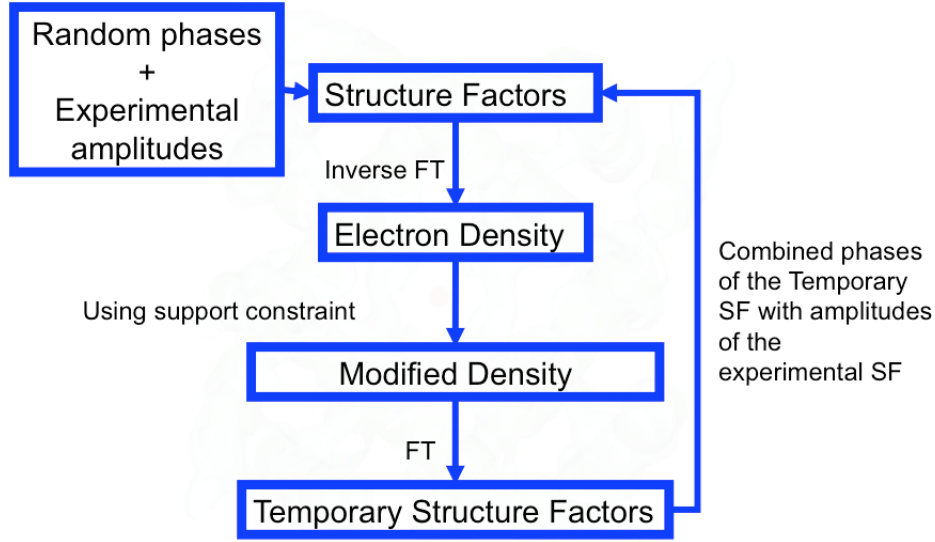


FIGURE 3.4: The flowchart of the Hybrid Input-Output (HIO) method .

$$R = \frac{\sum_{u,v} |F_{cal}(u, v) - F_{obs}(u, v)|}{\sum_i |F_{obs}(u, v)|} \quad (3.16)$$

where $F_{cal}(u, v)$ and $F_{obs}(u, v)$ refer to calculated and observed Fourier modulus, and (u, v) run through all Fourier space coordinates.

A sharp drop in R versus iteration number is a clear sign of convergence of the algorithm. In this project, we recovered the projected electron density of an individual protein molecule Fig. 3.6 from the reconstructed diffraction pattern of Fig. 3.5 after 1000 iterations of the combination of the charge-flipping (Oszlányi and Sütő, 2004) and shrink-wrap algorithm (Marchesini et al., 2003).

We further quantified the quality of our reconstructions with the phase-retrieval transfer function (PRTF). Phase-retrieval transfer function (PRTF) (Chapman et al., 2006) compares the magnitude of the complex-valued average of patterns phased with different starting guesses to the square root of the measured diffraction pattern.

$$PRTF(u, v) = | \langle \exp(i\phi(u, v)) \rangle | = | \langle \frac{F_{cal}(u, v)}{|F_{cal}(u, v)|} \rangle | \quad (3.17)$$

where $F_{cal}(u, v)$ refers to calculated Fourier modulus and $\langle .. \rangle$ denotes an average over independent reconstructions.

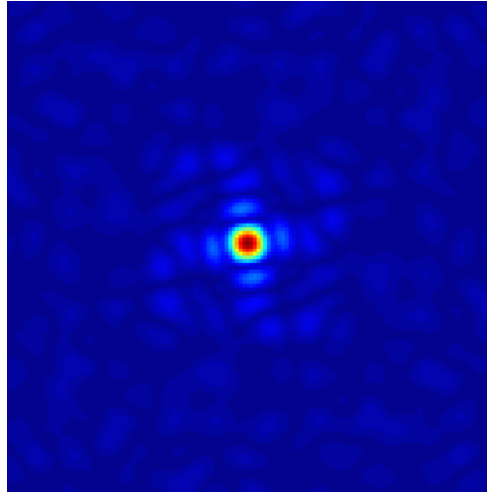


FIGURE 3.5: Single particle diffraction pattern recovered from the multi-particle diffraction pattern of Fig. 2.2. Note the similarity with Fig. 2.1

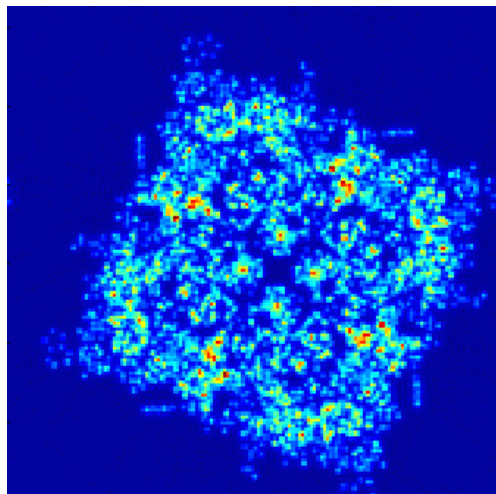


FIGURE 3.6: Projected structure of the K-channel protein found from an iterative phasing algorithm applied to the reconstructed diffraction pattern of Fig. 3.5

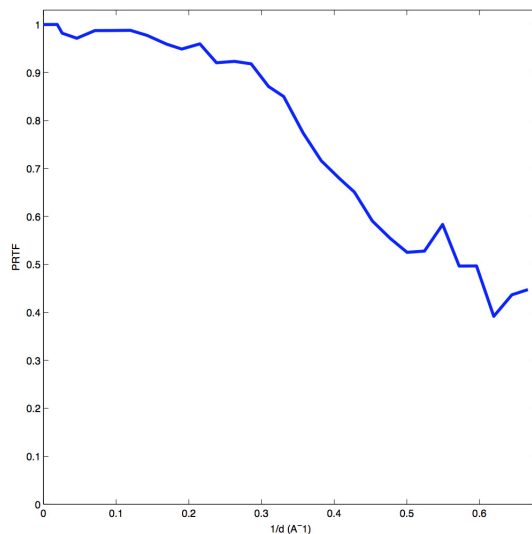


FIGURE 3.7: Phase-retrieval transfer functions as a function of inverse resolution $1/d$.

At a particular pixel of the diffraction pattern, if the phases are consistently reconstructed, the PRTF will be unity because the sum over N patterns will give a magnitude N times higher than the measured magnitude. If the phases are random, then this sum will tend to zero. The PRTF is shown in Fig. 3.7, which maintains values of greater than 0.5 to resolutions better than 2 \AA .

3.2 Three Dimensional Method

A possibility offered by a technique for 2D structure determination is that it may form the basis of a technique for the 3D structure determination of membrane proteins from the angular correlations to high resolution without an assumption of symmetry. Although at higher X-ray energies, as in the example here, the Ewald sphere could be considered flat, this is no longer true as the X-ray energy falls into the range that is commonly produced by XFELs. The combination of data from several energies thus explores 3D reciprocal space.

The wavelength dependent 2D phased intensity expansion coefficient is:

$$I_m^{\lambda_j}(q) = \sum_{l=0}^{\infty} I_{lm}(q) P_l^m(q\lambda_i/2) \quad (3.18)$$

where $I_{lm}(q)$ is the intensity expansion coefficient of the 3D diffraction volume.

The 3D diffraction volume can be expressed as a linear combination of spherical harmonics:

$$I(q, \theta_q, \phi_q) = \sum_{l=0}^{\infty} \sum_{m=-l}^l I_{lm}(q) Y_l^m(\theta_q, \phi_q) \quad (3.19)$$

Notice that the latter can be written as:

$$I(q, \theta_q, \phi_q) = \sum_{l=0}^{\infty} \sum_{m=-l}^l I_{lm}(q) P_l^m(\cos(\theta)) \exp(im\phi) = \sum_{m=-l}^l I_m^{\lambda_j} \exp(im\phi) \quad (3.20)$$

Diffraction geometry dictates:

$$\cos(\theta_q) = q\lambda/2 \quad (3.21)$$

where λ is the wavelength of the incident X-ray. Using 2D methods we can uniquely determine all phases of the full set of $I_m^{\lambda_j}(q)$ values, up to an arbitrary rotation. Note that expression (16) depends on the wavelength of the incident X-rays! So, each set of energy specific $I_m^{\lambda_j}(q)$ values can be phased analytically, yielding wavelength-dependent, curved slices of the diffraction volume. If the wavelength dependent slices can be oriented with respect to each other, and if the missing wedges of intensities that lie in between the measured $I_m^{\lambda_j}(q)$ slices do not prohibit subsequent phasing attempts, a 3D structure can be obtained from such data.

Using the the M-TIP reconstruction algorithm one is able to reconstruct the 3D electron density .

3.2.0.1 Multi-Tiered Iterative Projections (M-TIP)

This method was proposed by (Donatelli, Zwart, and Sethian, 2015). The iteration proceeds as follows:

1-Compute its intensity function, given a model density (it should be initially just random noise).

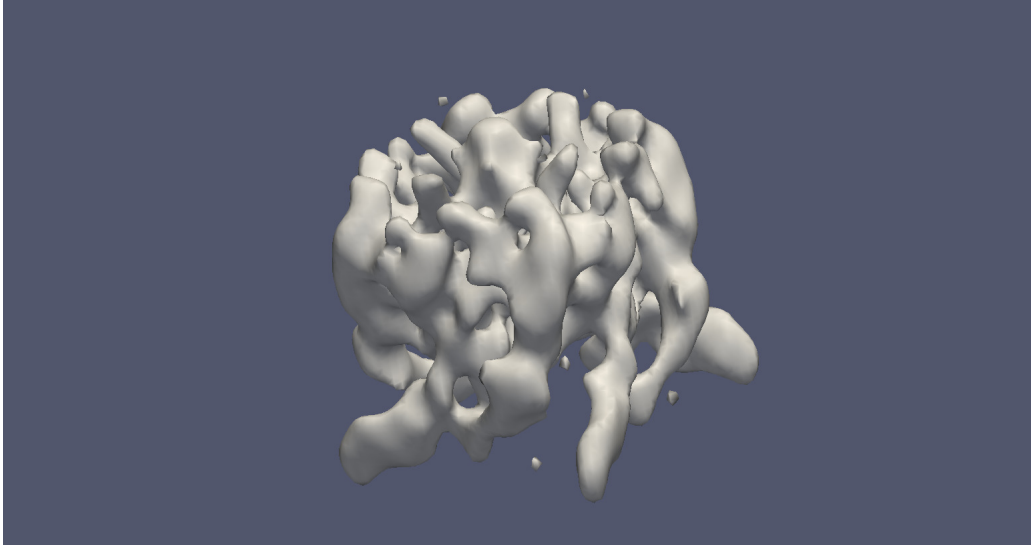


FIGURE 3.8: 3D projected structure of the K-channel protein found from MTIP algorithm applied to the reconstructed diffraction patterns. Note the similarity with Fig. 3.9 (Generated by Jeffrey Donatelli).

2-Align each 2D image to the intensity function.

3-Project the intensity function to be consistent with the aligned image. This basically fills in the diffraction volume in the un-sampled region with the intensity of the previous density model.

4-Take the inverse Fourier transform to get a new density model.

5-Enforce real-space constraints on the new density model through either the ER or HIO schemes.

Repeat 1-5 until convergence.

3.2.1 Result and Validation

We reconstructed the 3D electron density of the symmetric K-channel protein (Alam and Jiang, 2009) and asymmetric lactose permease (Abramson et al., 2003). The results are shown in (Fig. 3.8 and Fig. 3.10). For the evaluation of this method we reconstructed the projected structure found from PDB (Fig. 3.9 and Fig. 3.11).

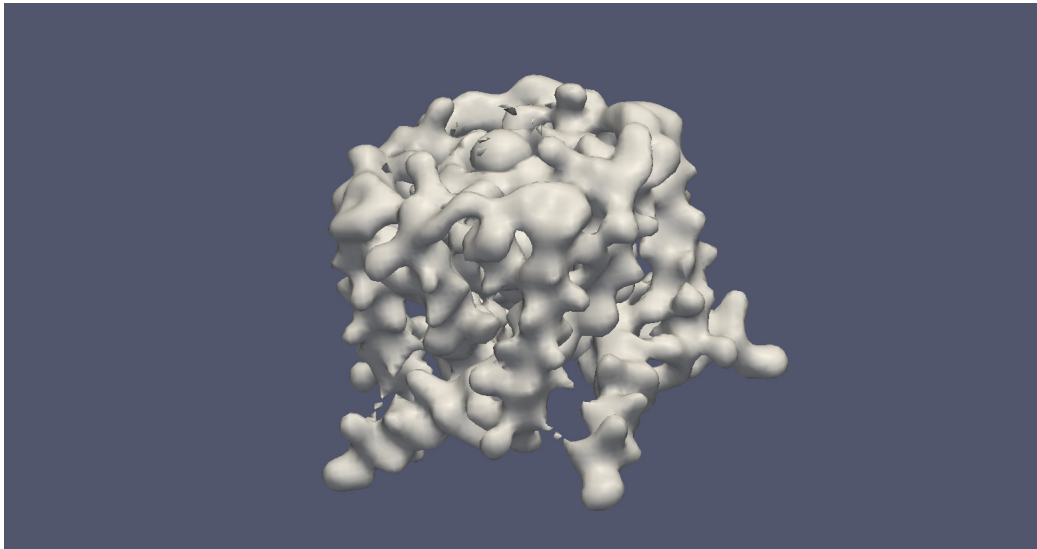


FIGURE 3.9: 3D projected structure of the K-channel protein found from PDB (Generated by Jeffrey Donatelli).

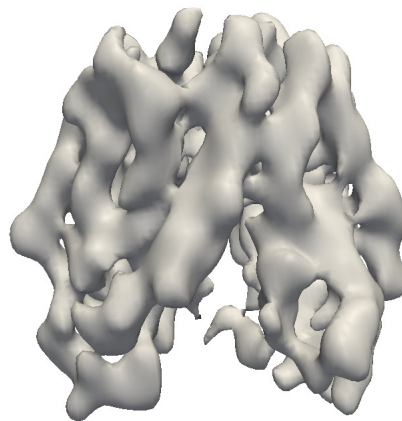


FIGURE 3.10: 3D projected structure of the lactose permease found from MTIP algorithm applied to the reconstructed diffraction patterns. Note the similarity with Fig. 3.11 (Generated by Jeffrey Donatelli).

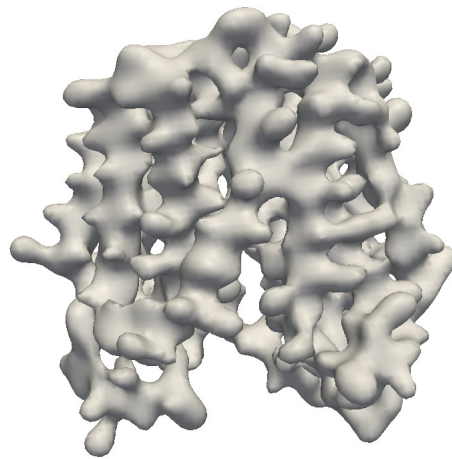


FIGURE 3.11: 3D projected structure of the lactose permease found from PDB (Generated by Jeffrey Donatelli).

Chapter 4

Conclusions

The difficulty of crystallizing a membrane protein means that there is limited information about this important class of proteins. Instead of trying to crystallize membrane proteins we suggest an alternative way of getting an enhanced signal from an ensemble of molecules even if they are not perfectly aligned as in a crystal. This enables a reconstruction of a single particle diffraction pattern from which an iterative phasing method reconstructs the projected structure of the particle. The experiment can be performed on a black lipid membrane used as a model of a membrane in the laboratory. In an XFEL, since the pulses arrive at the sample at about a 100 Hz frequency, and they are destructive for the sample, it is important that some kind of stepper motor arrangement be devised whereby the beam (or sample) is moved across the plate containing the holes (the support) so that each beam hits a different hole. A similar arrangement has been used in fixed target methods of XFEL work (Roedig et al., 2016).

What is extraordinary about Fig. 3.6 is that atomic resolution details are visible in the projected image. This image rivals the best atomic resolution images obtained from simulated diffraction patterns by single particle methods. This is indicative of the fact that if single axis assignment is perfect, the correlations can be found accurately from the experiment, and there is no obvious limit to resolution. A possibility offered by a technique for 2D structure determination is that it may form the basis of a technique for the 3D structure determination of membrane proteins from the angular correlations to high resolution without an assumption of symmetry. Although at higher X-ray energies, as in the example here, the Ewald sphere could be considered flat, this is no longer true

as the X-ray energy falls into the range that is commonly produced by XFELs. The combination of data from several energies thus explores 3D reciprocal space. We showed that the combination of data from different energies will be combined to reconstruct a fully 3D image without any recourse to symmetry information.

Although the simulated data is a simplification of what is obtained during a real experiment, the results obtained here serve as proof of principle for a new class of experiments that can be conducted at modern light sources, such as an ultra bright diffraction limited (soft) X-ray light source (Eriksson, Veen, and Quitmann, 2014). The sample environment to do these measurements at such a source are very similar to a soft X-ray ptychographic experiment (Shapiro et al., 2013), making this experiment a practical reality in the near future.

Part II

Reconstruction of the PR722 and RDV

Viruses from Experimental Data

Chapter 1

Introduction

The method of angular correlations for molecular structure determination has been known since the time of Kam (Kam, Shore, and Feher, 1978) for reconstructing a 3D image from Small Angle X-ray Scattering (SAXS) diffraction patterns. The method also has been used to address the problem of reconstructing 3D images from the snapshot diffraction patterns measured with X-ray Free-Electron Laser (XFEL) following the work of Saldin et al. (Saldin et al., 2009). The first experimental structure determination from X-ray diffraction patterns was reported by Starodub et al. (Starodub et al., 2012). The next application to experimental data is the method described in this thesis to the Coliphage PR772 virus (Lute et al., 2004) and Rice Dwarf Virus (RDV) (Fukushi, Shikata, and Kimura, 1962). Our present work is a part of the Single Particle Initiative (SPI) at the Linac Coherent Light Source (LCLS) in Menlo Park, California, which involves work aimed at reconstructing just such particles (this includes experimental work and the development of theoretical algorithms designed to reconstruct the particles at atomic or near-atomic resolution). Unlike the crystallographic problem in which the orientation of the particle may be determined by indexing techniques that depend on the properties of crystals, in the present case the imaging is done on single uncrystallized particles, where even that method of determining the orientation does not exist. Consequently our approach is to consider the angular correlation of the intensities, which are exactly the same, independent of the particle orientations. Thus any errors in finding the correlations from any one particle will be compensated by another. All one needs is a method

of reconstructing a particle from the correlations. Since all correlations are the same, independent of orientations, the actual orientations of the particle may be reconstructed in an orientation chosen for the investigator's convenience. Angular correlations were calculated from the diffraction patterns classified as single hits. We describe a method in which the particles are assumed randomly oriented. If this assumption is made (not unreasonable in the context of XFEL experiments) the particle's structure may be determined, especially if the diffraction volume is assumed to be symmetrical. We realize a symmetry assumption may not be justifiable for an entirely unknown structure. We posit that the structure determination of a virus is not a completely unknown problem. People have published work on the physical principles of the construction of virus capsids. In fact, Caspar and Klug (Caspar and Klug, 1962) have suggested that most regular viruses have particular symmetries, namely to be icosahedral or helical, due to their methods of formation. We acknowledge a lot of water had passed under the bridge since the work of Caspar and Klug (Caspar and Klug, 1962), but it is undeniably true that many common viruses are either nearly icosahedral, as in the present case, or the recently-discovered zika virus (Sirohi et al., 2016) or approximately helical in the case of the tobacco mosaic virus. Even if other virus shapes have been determined in the meanwhile, it would be necessary to explain so striking a shape as an icosahedron. Caspar and Klug (Caspar and Klug, 1962) offered an elegant explanation that remains valid today. The inside of a capsid contains all the genetic material for the coding of the capsid itself. The problem is there is not much space in the inside of a capsid, so the shape chosen by nature is likely to be one that has the largest volume enclosed to accommodate the genetic material for the part of the capsid that needs to be coded. It is true that a sphere encloses more volume per given surface area, but it's difficult to see how one can get by coding only for part of a sphere. An icosahedral shape allows the genetic material to code for just the 20 constituent triangles with strong chemical bonds on the outside. In the case of an icosahedron, all that needs to be coded is the component triangles, i.e. $1/20$ of the entire surface area of the capsid. The triangles self-assemble to form the whole capsid as may be confirmed by a model in which a set of plastic

equilateral triangles with magnetic forces representing unsatisfied chemical bonds are put in a box and shaken. The triangles self-assemble into the shape of an icosahedral capsid. Since we can independently confirm that the capsid is icosahedral, it is not unreasonable to exploit such symmetry, as is quite routinely done in crystallography. We report here on the reconstruction of the structure of the icosahedral virus Coliphage PR772 (Lute et al., 2004) and Rice Dwarf Virus (RDV) (Fukushi, Shikata, and Kimura, 1962) from experimental data of the angular correlations of their intensities. The method consists first of the reconstruction of the intensity distribution in 3D (the diffraction volume). The reliability and resolution of the reconstructed diffraction volume were tested by R-split. Having found the diffraction volume, the next step is the reconstruction of the electron density by a phasing algorithm, applied to these intensities. Traditionally, this is followed by the reconstruction of a real space image by an iterative phasing algorithm. We found that not all phasing algorithms are equal to the task. We suggest and develop a new method which worked for us well. It is alternate cycles of charge-flipping (Oszlanyi, 2004) and error reduction (Fienup, 1982) with a spherical real-space support constraint. The reliability and resolution of the reconstructed electron density were tested by comparing the resulting 3D electron densities by FSC.

Chapter 2

Instrumentation and Data Collection

Measurements were carried out at the Atomic Molecular Optics (AMO) beamline (Bozek, 2009) of LCLS (Emma et al., 2010). To introduce virus (70 *nm* diameter) into the focused XFEL beam with photon energy $E = 1.6 \text{ keV}$, they employed the aerodynamic lens stack system (Bogan et al., 2008) with a gas dynamic virtual nozzle (DePonte et al., 2008). The distance between sample and detector is 581 *mm*.

A pair of pnCCD detectors with a resolution of 11.6 *nm* at the detector edge measured the data that we used in this project. Assembled detector dimensions in pixels were 1040 *by* 1028. Each half of the pnCCD is 1024 *by* 512 pixels and there is a gap of 12 pixels. Pixel size is 75 μm . More details about the experimental setup and sample preparation are available in ref(Aquila, 2017). They performed hit finding using LCLS software called psocake (Damiani et al., 2016) to throw away the blank images. The hits were then classified into single hits and multiple hits using unsupervised machine learning similar to that described in (Yoon et al., 2011). The original diffraction patterns have been down sampled with 4 *by* 4 binning method.

The resulting data used in this analysis contained 211 single particle diffraction patterns (260 *by* 257) from Rice Dwarf Virus (RDV) particles 2.2 and 10000 patterns (260 *by* 257) from Coliphage PR772 particles 2.1.

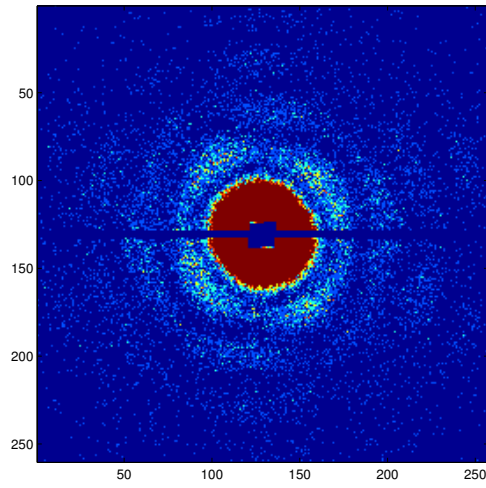


FIGURE 2.1: A typical snapshot of PR772. The blue horizontal line is due to the gap.

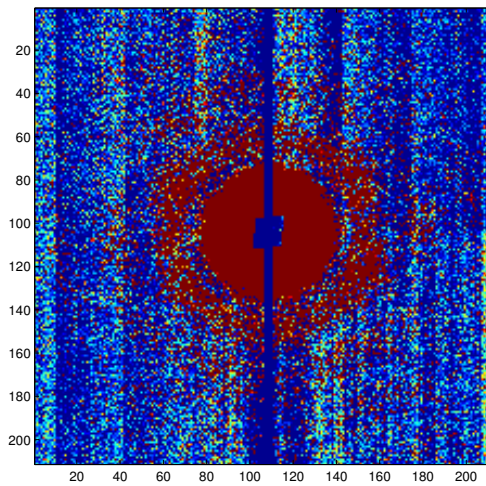


FIGURE 2.2: A typical snapshot of RDV. The blue vertical line is due to the gap.

Chapter 3

Theory

This thesis represents the first step in the reconstruction of an image of an entirely non-symmetric structure. We first find the structure that is close to the actual structure. Then one can determine the exact structure by a perturbation theory (Pande et al., 2014). Perturbation theory only requires an initial structure close to the actual structure and does not assume symmetry. If we have a sufficient number of the same type (or confirmation) of 2D diffraction patterns, we are able to reconstruct a 3D structure. The symmetry we will exploit here is icosahedral symmetry, so a few words seem appropriate at the outset about icosahedral and icosahedral symmetry.

3.1 Icosahedral and Icosahedral Symmetry

In Euclidean geometry, a Platonic solid is a regular, convex polyhedron. Such a solid is constructed by an equal number of congruent, regular, polygonal faces meeting at each vertex. In fact, five platonic solids (a tetrahedron, cube, octahedron, dodecahedron, or icosahedron) meet those criteria Fig.3.1. The icosahedron has 12 polyhedron vertices, 30 polyhedron edges, and 20 equivalent equilateral triangle faces and it is represented by the Schlegel symbol $3, 5$.

The symmetry group of a polyhedron in R^3 is the set of rotations and reflections which leave the polyhedron unchanged, in the sense that after the operation, each vertex is located where it or another vertex was previously, and similarly for the edges. A regular icosahedron has a symmetry group of order 120 including transformations



FIGURE 3.1: Platonic solids

that combine a reflection and a rotation, out of which 60 are orientation-preserving or rotational symmetries (Cohan, 1958).

In the polar coordinate, the values of the intensity I are specified q, θ, ϕ . For the same value of q , the diffraction volume will be the same in directions of symmetry (θ_k, ϕ_k) for a set of values of k . As we mentioned above, for icosahedral symmetry, there are 60 different values of k (Cohan, 1958).

The icosahedral harmonics are represented by the linear combinations of spherical harmonics.

$$\mathcal{I}_l^{(j)}(\theta, \phi) = \sum_m b_{lm}^{(j)} Y_{lm}(\theta, \phi) \quad (3.1)$$

where l is the quantum number of the orbital angular momentum and m the magnetic quantum number. The coefficients $b_{lm}^{(j)}$ are given by Zheng et al. (Zheng, Doerschuk, and Johnson, 1995) for arbitrary j, l , and m .

For a particular order l , let Nl be the number of icosahedral harmonics. In particular, $Nl = 0$ or $Nl = 1$ for $l < 30$. The corresponding weights j drop out of the problem. Prior to Zheng et al. (Zheng, Doerschuk, and Johnson, 1995), Jack and Harrison (Jack and Harrison, 1975) gave the same coefficients for non-degenerate icosahedral harmonics that exist for $l < 30$ and called them a_{lm} . $Nl = 1$ or 2 for $30 \leq l < 60$, so we need to keep the corresponding weights j .

The diffraction volume is given as a linear combination of icosahedral harmonics

$$I(\mathbf{q}) = \sum_{l,j} g_l^{(j)}(q) \mathcal{I}_l^{(j)}(\theta, \phi) \quad (3.2)$$

The only thing that needs to be determined from the experimental data are the coefficients $g_i^{(j)}(q)$. In general, as l becomes greater we have a greater chance of forming icosahedral harmonics. In fact, from $l = 30$ to $l = 60$, there are only three values of l which are degenerate with two-fold degeneracy. By modeling the electron density with icosahedral harmonics, Zheng et al. (Zheng, Doerschuk, and Johnson, 1995) showed that it is possible to get more detailed information about the virus structure using information present in just solution scattering a fraction of the data available through an XFEL experiment. It should be noted that Zheng et al. (Zheng, Doerschuk, and Johnson, 1995) do not try to recover the structure purely from the experimental data but use a different methodology of seeing which model gives intensities that agree best with the measured intensities.

Of course, it is impossible to form linear combinations of some spherical harmonics to make icosahedral symmetry. The values of l that allow this are $l=0, 6, 10, 12, 16$ and every even value after that, and so are appropriate values of l for icosahedral harmonics, which also satisfies Friedel (inversion) symmetry. These are called icosahedral harmonics

$$\mathcal{I}_l(\theta, \phi) = \sum_m b_{lm} Y_{lm}(\theta, \phi) \quad (3.3)$$

It is, therefore, arguable that one can construct a diffraction volume of these values of l and no others for a particle satisfying icosahedral symmetry.

Why this emphasis on icosahedral symmetry? Due to their methods of formation, Caspar and Klug (Caspar and Klug, 1962) once suggested that regular viruses have either icosahedral or helical symmetries. In both cases we have found solutions using angular correlations (Saldin et al., 2011), (Poon et al., 2013), despite using only the diagonal correlations $q = q'$. It will be understood that the method of angular correlation uses only spherical coordinates. We find this is an appropriate representation of the use of XFEL radiation as in this case particles are incident in random orientations in addition to random positions. If one is dealing with the scattered intensities, all particle positions give rise to the same intensity (the positions being sensitive only to unmeasured phases

while angular rotations may be dealt with using rotation matrices).

3.2 Angular Correlation

The first advantage in the case of symmetrical particles is the ability to quickly analyze the symmetry of the particles even before reconstructing them. The fundamental quantity reconstructed from the data is the angular pair correlations $C_2(q, q'; \Delta\theta)$ and triple correlations $C_3(q, q'; \Delta\phi)$ and then averaging them over all measured diffraction patterns. Let us denote the averages by $\langle C_2(q, q'; \Delta\phi) \rangle_{DP}$ and $\langle C_3(q, q'; \Delta\phi) \rangle_{DP}$, respectively. In fact, the quantities calculated from the intensities in each diffraction pattern (scattered intensities) are

$$C_2(q, q'; \Delta\phi) = \int I(q, \phi)I(q', \phi + \Delta\theta)d\phi \quad (3.4)$$

and

$$C_3(q, q'; \Delta\phi) = \int I^2(q, \phi)I(q', \theta + \Delta\phi)d\phi \quad (3.5)$$

They are then averaged over all diffraction patterns. We then calculate

$$B_l(q, q') = \int \langle C_2(q, q'; \Delta\phi) \rangle_{DP} P_l(\cos(\Delta\phi))d\Delta\phi \quad (3.6)$$

where $P_l(\cos(\Delta\phi))$ is a Legendre polynomial. Likewise,

$$T_l(q, q') = \int \langle C_3(q, q'; \Delta\phi) \rangle_{DP} P_l(\cos(\Delta\phi))d\Delta\phi \quad (3.7)$$

Although this is how the $B_l(q, q')$ are calculable from the experimental data, in relating these quantities to the diffraction volume, we try to reconstruct it. We use the independently derived expressions

$$B_l(q, q') = g_l(q)g'_l(q') \quad (3.8)$$

$$T_l(q, q') = \sum_{l', l''} b(l', l, l'')g_{l'}(q')g_l(q)g_{l''}(q) \quad (3.9)$$

where

$$b(l', l, l'') = \sum_{m', m, m''} a_{l'm'} a_{lm} a_{l''m''} \quad (3.10)$$

where the a_{lm} are the Harrison coefficients (Jack and Harrison, 1975) to relate $T_l(q, q')$ to the diffraction volume and $g_l(q)$ are the coefficients that relate to the final volume using icosahedral harmonics via

$$I(\mathbf{q}) = \sum_{\mathbf{l}} \mathbf{g}_{\mathbf{l}}(\mathbf{q}) \mathcal{I}_{\mathbf{l}}(\theta, \phi) \quad (3.11)$$

An advantage of the use of icosahedral harmonics is that since each harmonic has icosahedral symmetry, any linear combination will have it, too. Suppose we use the usual expression instead (Saldin et al., 2009). The diffraction volume is expanded instead as a function of spherical harmonics in a conventional way, as

$$I(\mathbf{q}) = \sum_{lm} I_{lm}(q) Y_{lm}(\theta, \phi) \quad (3.12)$$

The two expressions will be equivalent if

$$I_{lm}(q) = a_{lm} g_l(q) \quad (3.13)$$

In this expression the a_{lm} , are the Harrison coefficients of icosahedral harmonics (Jack and Harrison, 1975), and the g_l are coefficients in the expansion of the diffraction volume with icosahedral harmonics.

In the expression (3.11) the only uncertainty is the sign of $g_l(q)$ (it is real). The signs may be found from an exhaustive search through all signs of $g_l(q)$ because the LHS may be obtained by experiment and the $|g_l(q)|$'s are known. It should be noted that the number of signs in the icosahedral values of l up to 18 is six, so the number of combinations is not astronomical and not difficult for a computer. That way we can search for the best agreement with the experimentally determined LHS. It will be noted that $B_l(q, q')$ and $T_l(q, q')$ are calculated by not assuming icosahedral symmetry. The particular pattern of l values is, however, characteristic of icosahedral symmetry that many regular viruses are known to have (Caspar and Klug, 1962). On the assumption

of icosahedral symmetry, however, the fact that the $B_l(q, q')$'s are only functions of l amongst the angular momentum quantum numbers is no problem with the use of the Harrison coefficients (Jack and Harrison, 1975). Then one may evaluate the diffraction volume from (3.11). We can find the magnitudes of $g_l(q)$ from

$$|g_l(q)| = \sqrt{B_l(q, q)} \quad (3.14)$$

and then their signs are found through an exhaustive search through all possible signs of $g_l(q)$ in (3.9).

Thus one can overcome the problem that $B_l(q, q)$ is a function of only l amongst the angular momentum quantum numbers by allowing the a_{lm} coefficients to define the function of the magnetic quantum numbers m . In other words, if one can assume icosahedral symmetry, one may instead find the diffraction volume via (3.11).

We first determine the symmetry without any assumptions. Having done this, we are free then to exploit the symmetry to reconstruct the details of the diffraction volume by using the known icosahedral symmetry to apportion the coefficients characterized by the value of m in addition to l and use the icosahedral harmonics. The details of the procedure are found in the paper by Poon and Saldin (Poon and Saldin, 2015). This approach uses only the diagonal terms $q = q'$. Note that a further assumption of this symmetry is rather like a denoising process, which neglects non-icosahedral components (Saldin et al., 2011). This is a bit like ignoring the intensities of a manifold in manifold-embedding methods (Hosseinizadeh et al., 2015). In the approach pioneered by Donatelli et al. (Donatelli, Zwart, and Sethian, 2015) a phasing algorithm is used to go directly from the correlations to the electron density of the sample. This method allows the reconstruction of non-symmetric reproducible structures. As we mentioned before, for an icosahedral scatterer, the values of l are $l = 0, 6, 10, 12, 16$ and every even value after that. Additionally, they are the only ones which can satisfy both Friedel's Law and can form icosahedral harmonics. Since this first application uses the non-degenerate icosahedral harmonics, it uses angular momentum values only up to $l_{max} = 28$. Hence

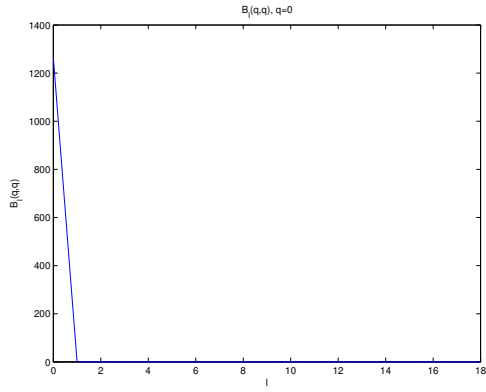
$q_{max}=l_{max}/R=28/350 \text{ \AA}^{-1}= 0.08 \text{ \AA}^{-1}$ (in which R is the radius of virus) or a resolution of 78.5 \AA , which is quite adequate at the present experimental resolution of about 116 \AA (Aquila, 2017). The angular momentum decomposition of the $B_l(q, q)$ are shown respectively for Pr772 and RDV in Fig. 3.2 and Fig. 3.3. These figures reveal the characteristic icosahedral pattern at least up to $l=16$. Right now there are plans to increase the resolution to 5 \AA , with new experiments at the LCLS, which will require, in our case, the use of the degenerate icosahedral harmonics beyond $l = 28$. However, we believe this is possible, as demonstrated by Zheng et al. (Zheng, Doerschuk, and Johnson, 1995) for SAXS. If we average the data in all our measured diffraction patterns, we recover an SAXS diffraction pattern. So, in fact, the SAXS forms a fraction of XFEL data, and thus there is no reason we may get more accurate images with the method of Zheng et al. (Zheng, Doerschuk, and Johnson, 1995).

3.3 Phase Retrieval

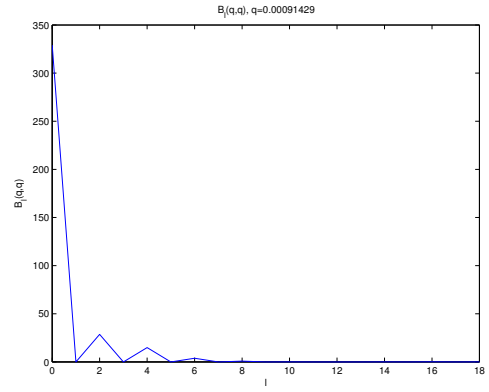
The present limitations of the program to non-degenerate harmonics $l_{max} = 28$ is unimportant in the current application as the maximum angular momentum for the level of resolution of the experimental data is $l_{max} = 18$. Since icosahedral symmetry is an angular symmetry, it is always possible to expand the diffraction volume in shells of constant q . This way one can reconstruct a diffraction volume that has perfect icosahedral symmetry (Fig. 3.4 and Fig. 3.6). In fact, when the real space is icosahedral (20 triangles) the diffraction volume is dodecahedral (12 pentagons) which has icosahedral symmetry. The behavior of different shapes of real and reciprocal space is not uncommon. It is well known that the reciprocal space of an FCC lattice is BCC and vice versa.

Having found the diffraction volume, the next step is the reconstruction of the electron density by a phasing algorithm, applied to these intensities. We found that not all phasing algorithms are equal to the task, but one which worked well for us is alternate cycles of charge-flipping (Oszlányi and Sütő, 2004) and error reduction (Fienup, 1982) with a spherical real-space support constraint. In each phasing iteration, a spherical

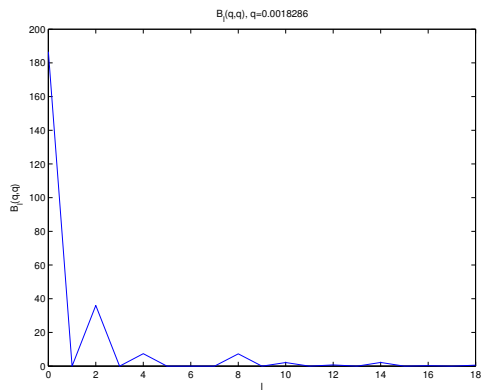
support is updated. The support is then obtained by applying a threshold at 20% of its maximum. In particular, we found this combination worked much better than an algorithm based on charge-flipping (Oszlányi and Sütő, 2004) alone. The results are shown in Fig. 3.5 and Fig. 3.7. which reveals a distinct icosahedral appearance.



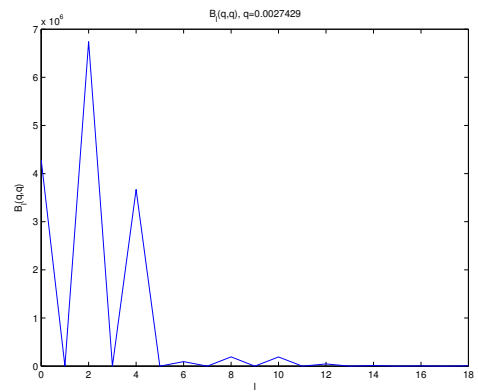
(A) *Pr*772, $B_l(q, q), q = 0$



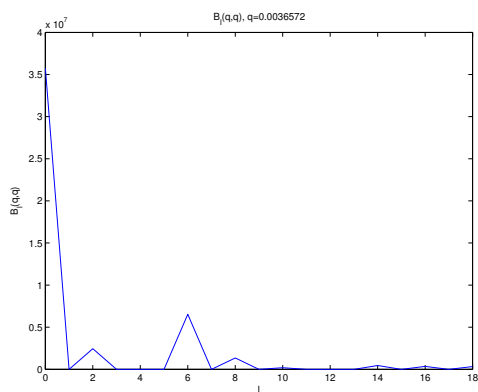
(B) *Pr*772, $B_l(q, q), q = 0.0009$



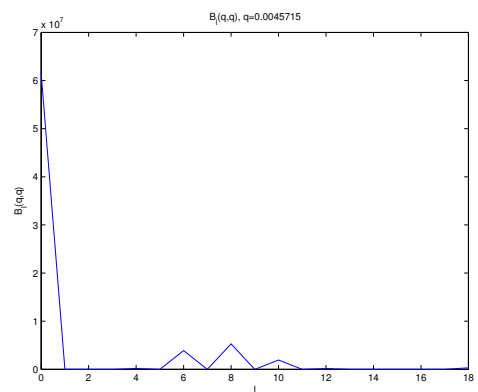
(C) *Pr*772, $B_l(q, q), q = 0.0018$



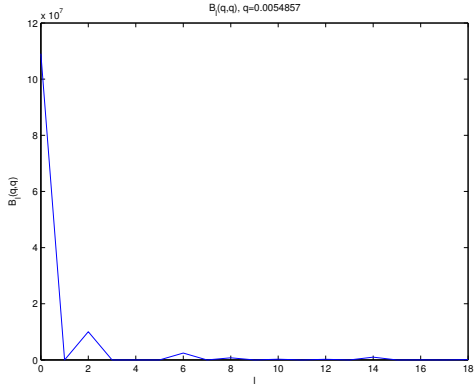
(D) *Pr*772, $B_l(q, q), q = 0.0027$



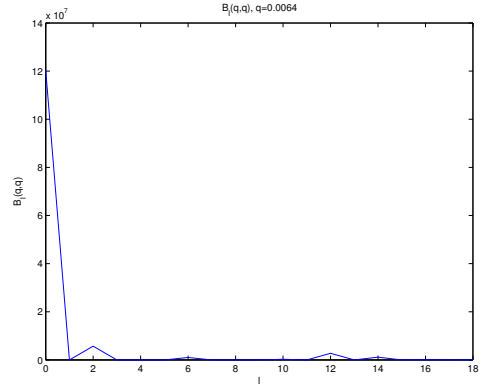
(E) *Pr*772, $B_l(q, q), q = 0.0037$



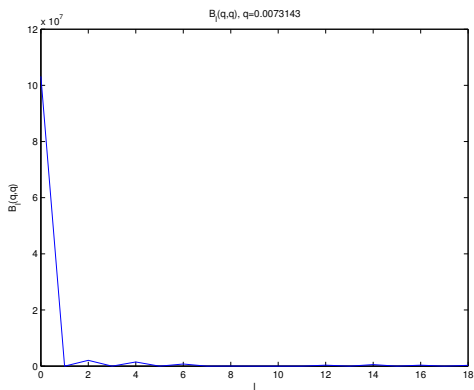
(F) *Pr*772, $B_l(q, q), q = 0.0046$



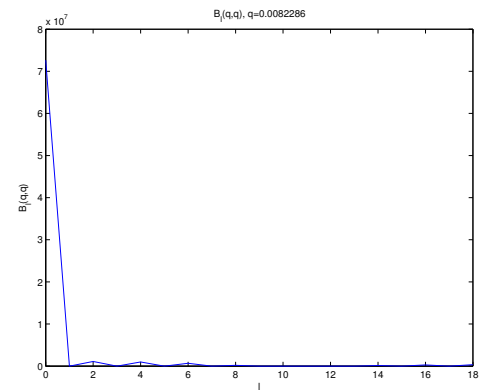
(G) *Pr772*, $B_l(q, q)$, $q = 0.0055e$



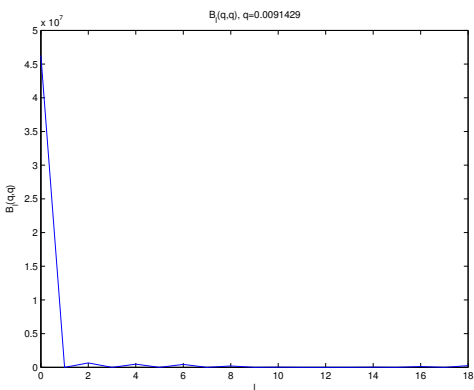
(H) *Pr772*, $B_l(q, q)$, $q = 0.0064$



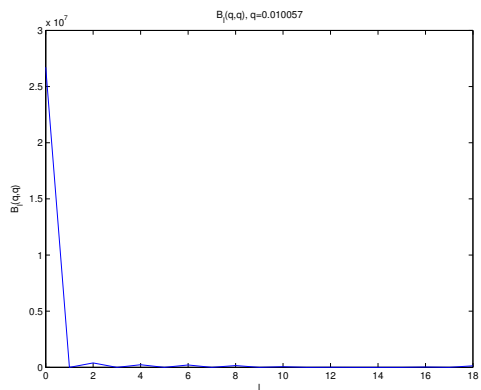
(I) *Pr772*, $B_l(q, q)$, $q = 0.0073$



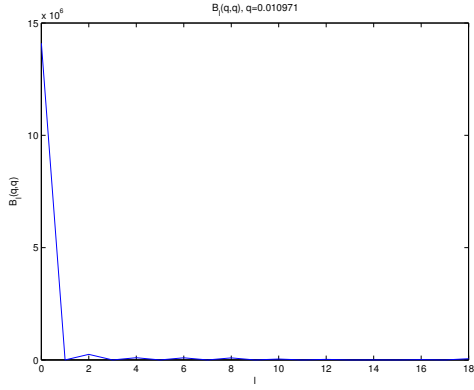
(J) *Pr772*, $B_l(q, q)$, $q = 0.0082$



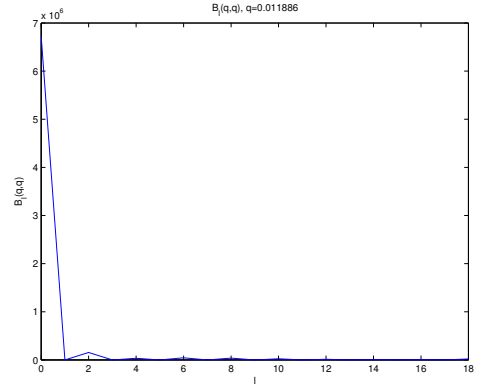
(K) *Pr772*, $B_l(q, q)$, $q = 0.0091$



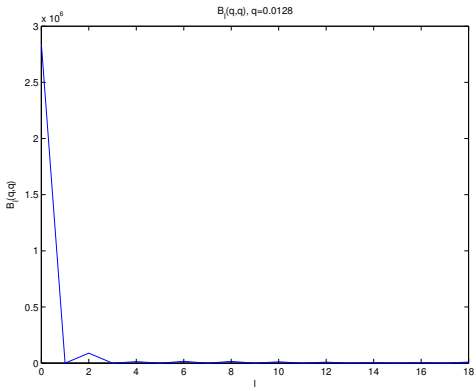
(L) *Pr772*, $B_l(q, q)$, $q = 0.0101$



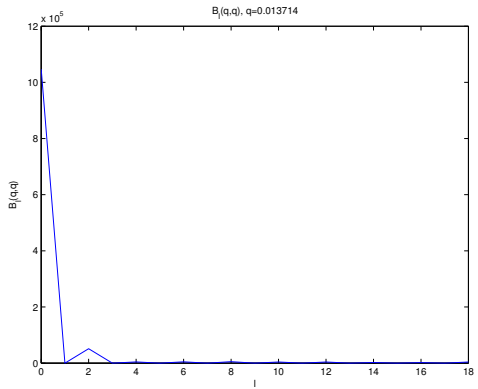
(M) $Pr772, B_l(q, q), q = 0.0110$



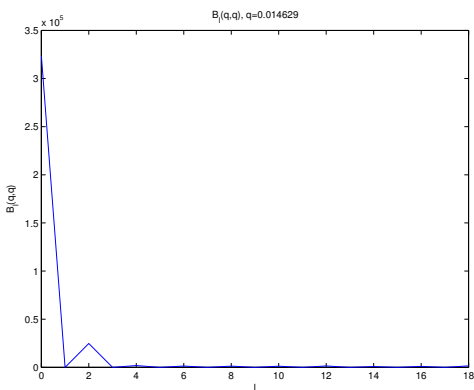
(N) $Pr772, B_l(q, q), q = 0.0119$



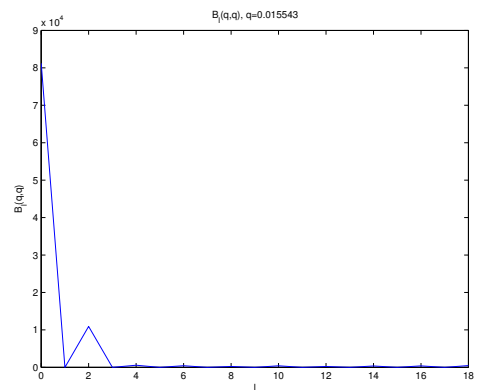
(O) $Pr772, B_l(q, q), q = 0.0128$



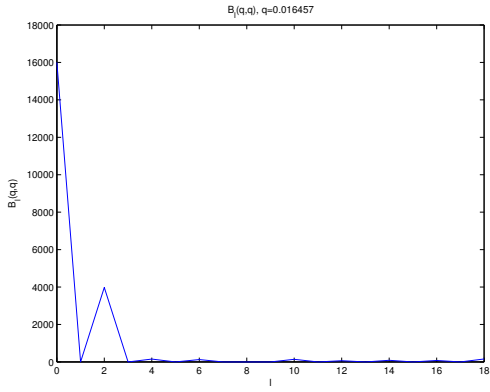
(P) $Pr772, B_l(q, q), q = 0.0137$



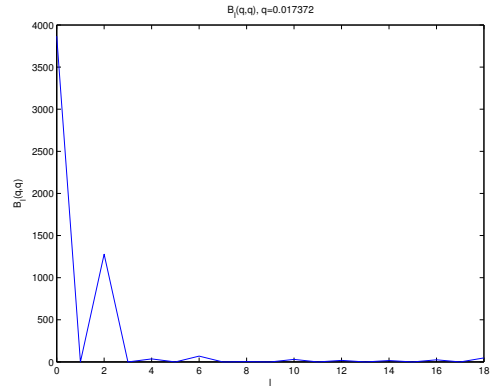
(Q) $Pr772, B_l(q, q), q = 0.0146$



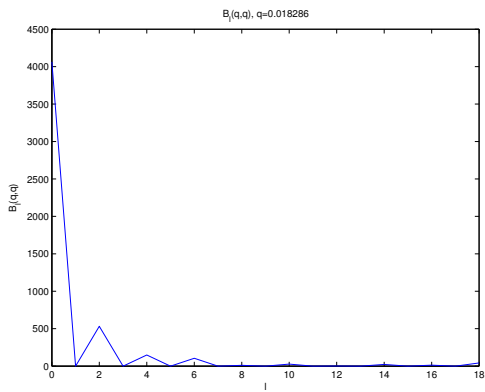
(R) $Pr772, B_l(q, q), q = 0.0155$



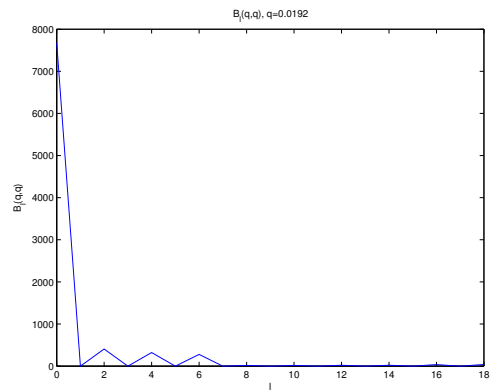
(s) $Pr772, B_l(q, q), q = 0.0165$



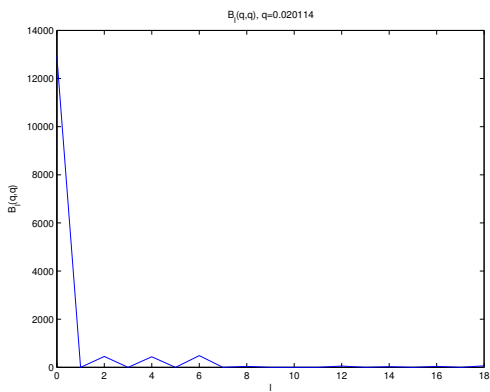
(t) $Pr772, B_l(q, q), q = 0.0174$



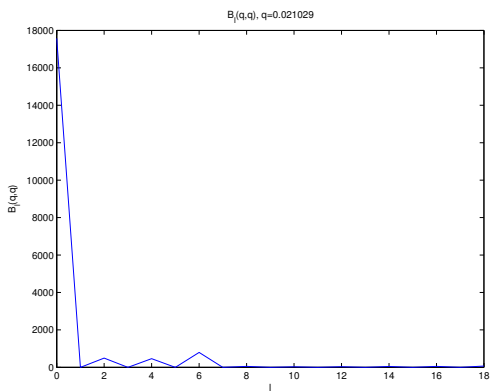
(u) $Pr772, B_l(q, q), q = 0.0183$



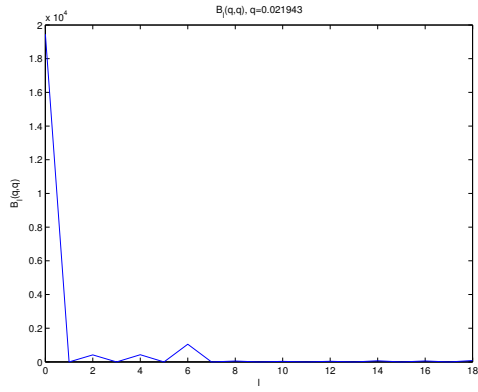
(v) $Pr772, B_l(q, q), q = 0.0192$



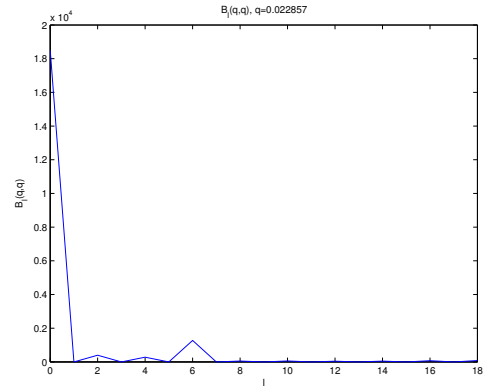
(w) $Pr772, B_l(q, q), q = 0.0201$



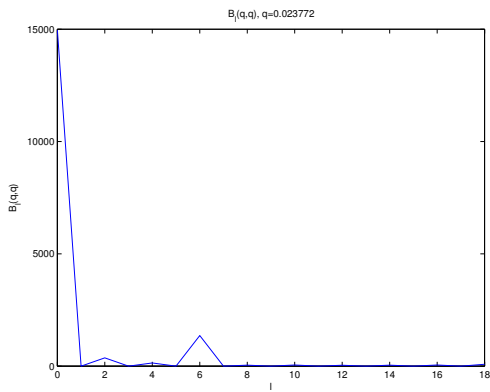
(x) $Pr772, B_l(q, q), q = 0.0210$



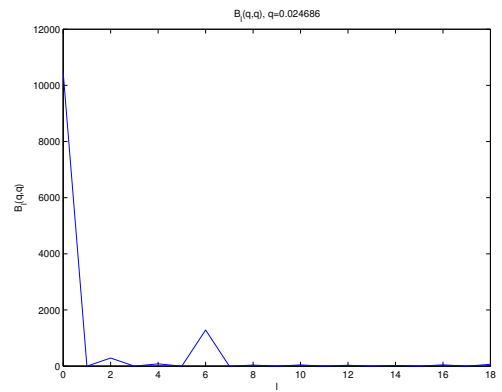
(Y) $Pr772, B_l(q, q), q = 0.0219$



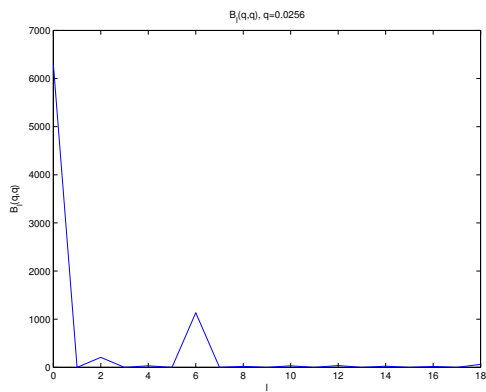
(Z) $Pr772, B_l(q, q), q = 0.0229$



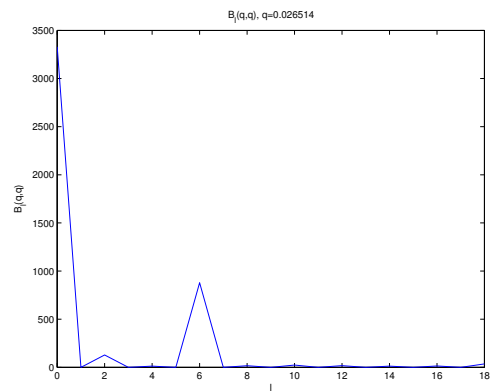
(AA) $Pr772, B_l(q, q), q = 0.0238$



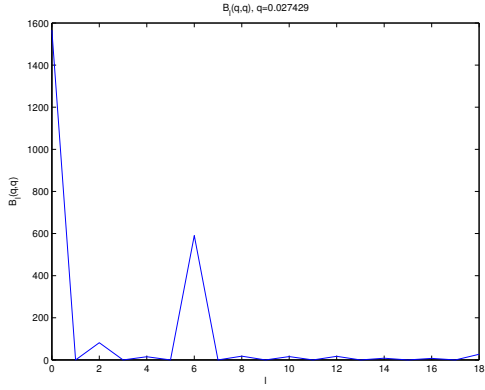
(AB) $Pr772, B_l(q, q), q = 0.0247$



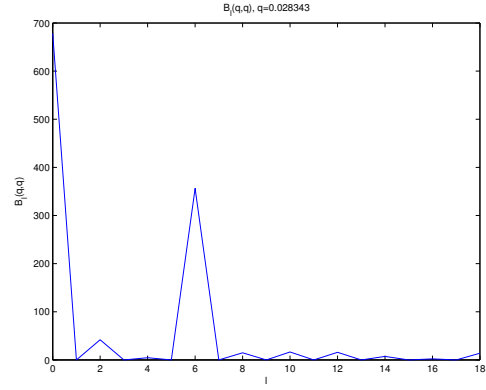
(AC) $Pr772, B_l(q, q), q = 0.0256$



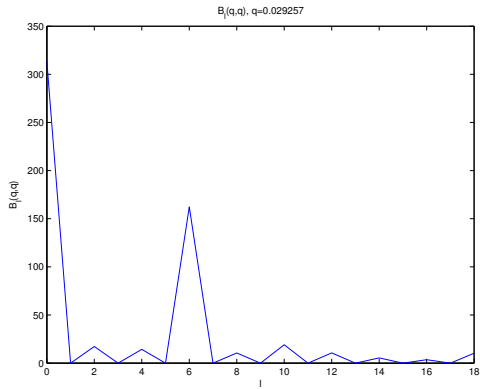
(AD) $Pr772, B_l(q, q), q = 0.0265$



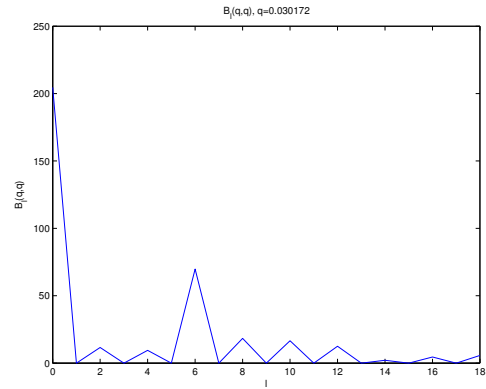
(AE) $Pr772, B_l(q, q), q = 0.0274$



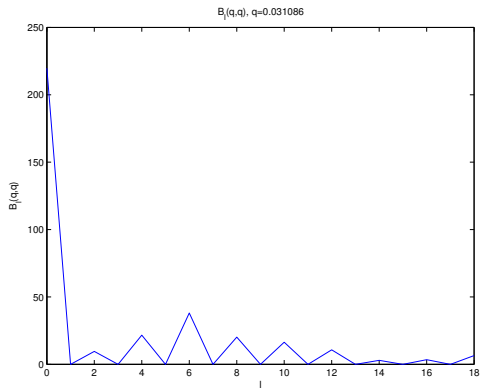
(AF) $Pr772, B_l(q, q), q = 0.0283$



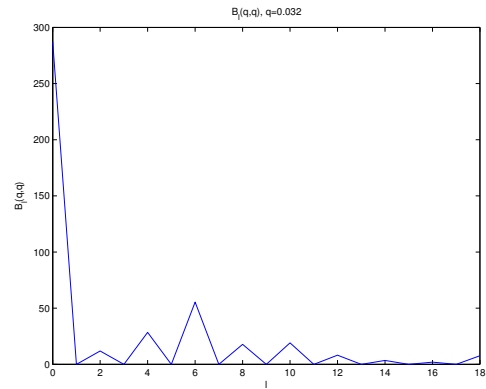
(AG) $Pr772, B_l(q, q), q = 0.0293$



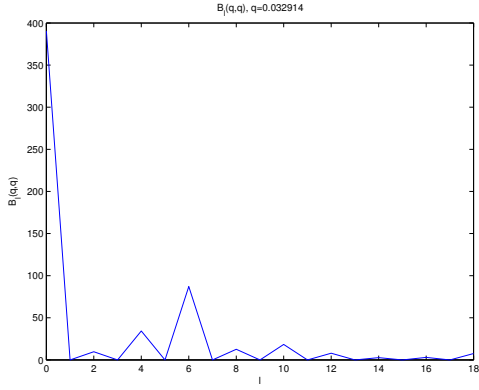
(AH) $Pr772, B_l(q, q), q = 0.0302$



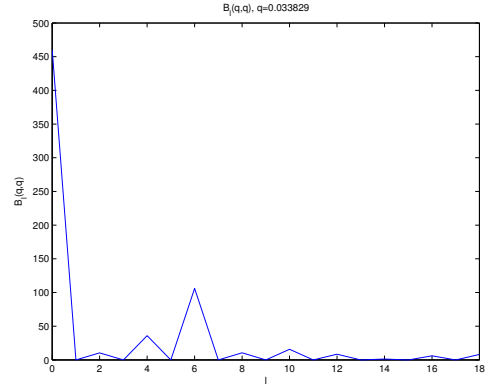
(AI) $Pr772, B_l(q, q), q = 0.0311$



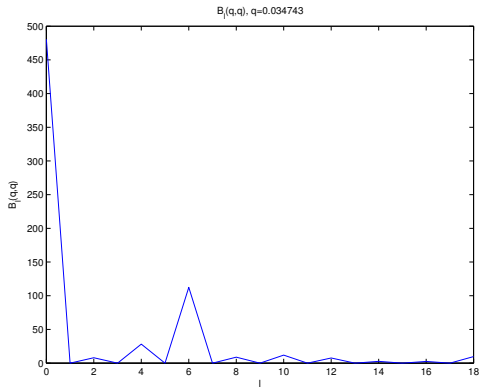
(AJ) $Pr772, B_l(q, q), q = 0.0320$



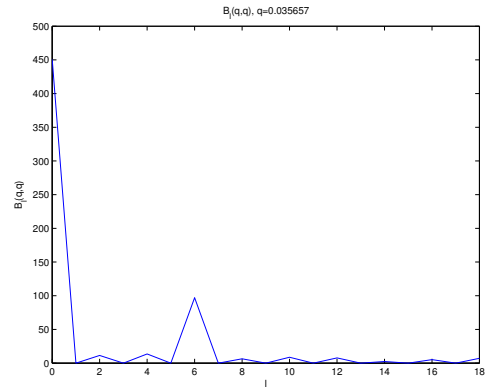
(AK) $Pr772, B_l(q, q), q = 0.0329$



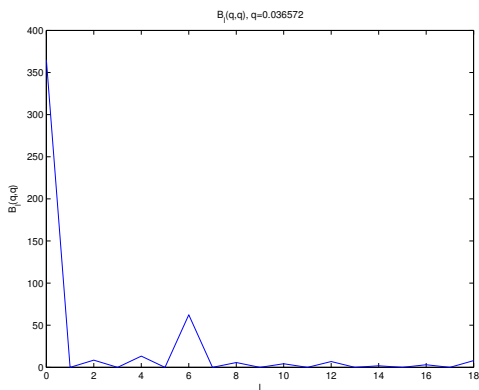
(AL) $Pr772, B_l(q, q), q = 0.0338$



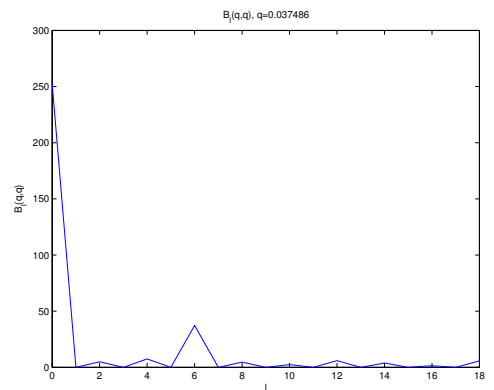
(AM) $Pr772, B_l(q, q), q = 0.0347$



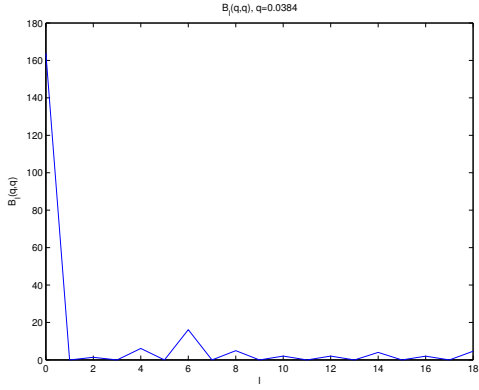
(AN) $Pr772, B_l(q, q), q = 0.0357$



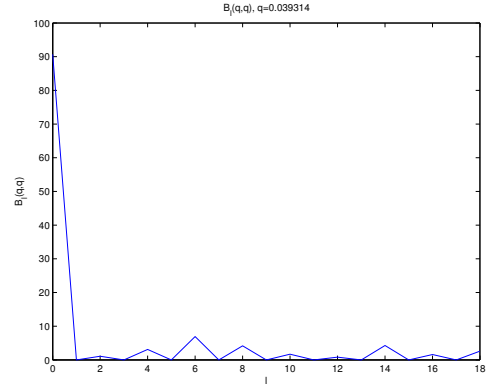
(AO) $Pr772, B_l(q, q), q = 0.0366$



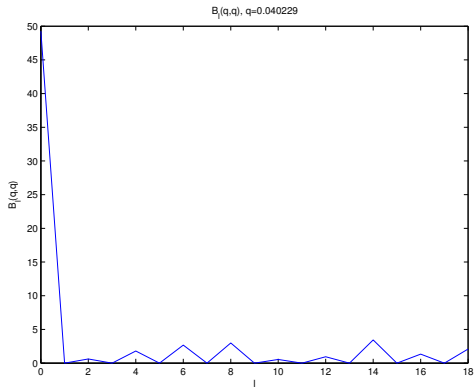
(AP) $Pr772, B_l(q, q), q = 0.0375$



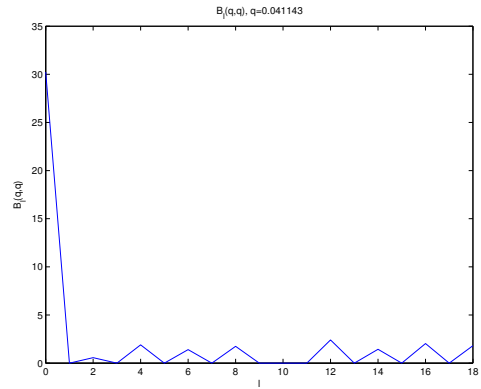
(AQ) $Pr772, B_l(q, q), q = 0.0384$



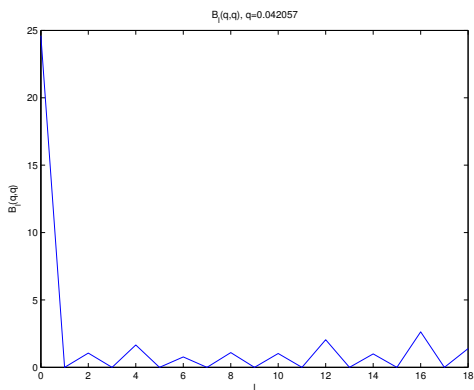
(AR) $Pr772, B_l(q, q), q = 0.0393$



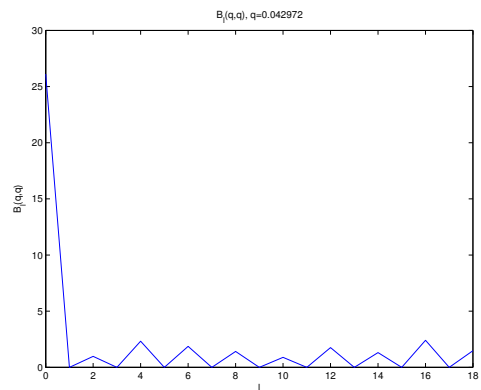
(AS) $Pr772, B_l(q, q), q = 0.0402$



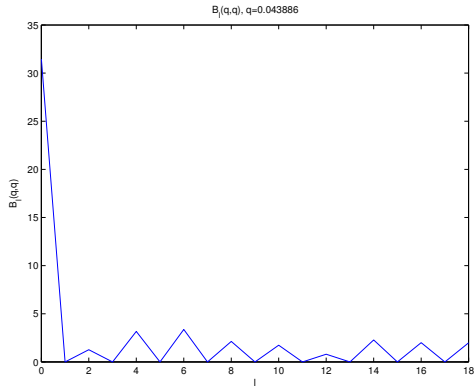
(AT) $Pr772, B_l(q, q), q = 0.0411$



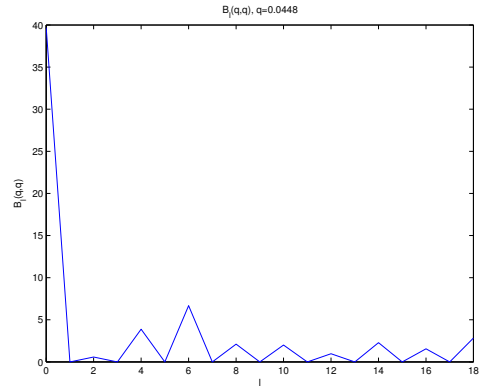
(AU) $Pr772, B_l(q, q), q = 0.0421$



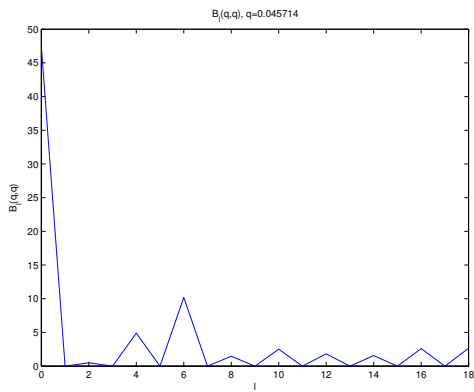
(AV) $Pr772, B_l(q, q), q = 0.0430$



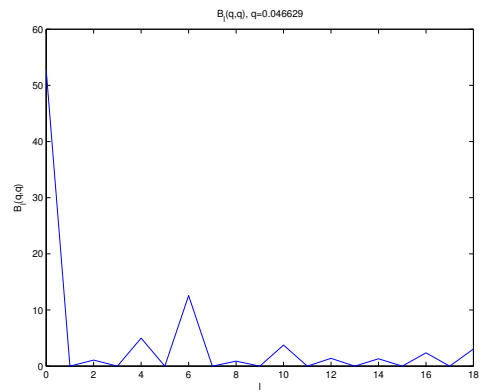
(AW) $Pr772, B_l(q, q), q = 0.0439$



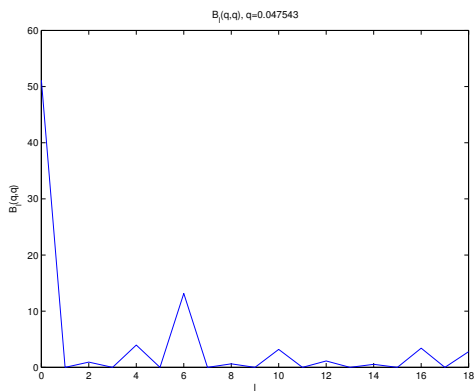
(AX) $Pr772, B_l(q, q), q = 0.0448$



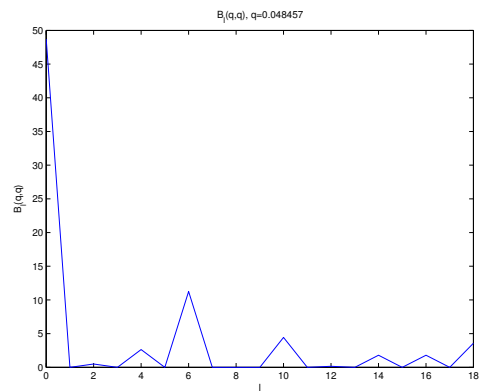
(AY) $Pr772, B_l(q, q), q = 0.0457$



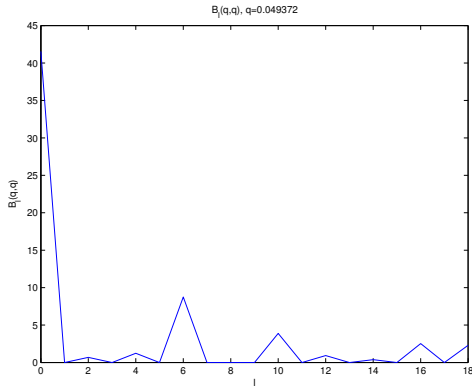
(AZ) $Pr772, B_l(q, q), q = 0.0466$



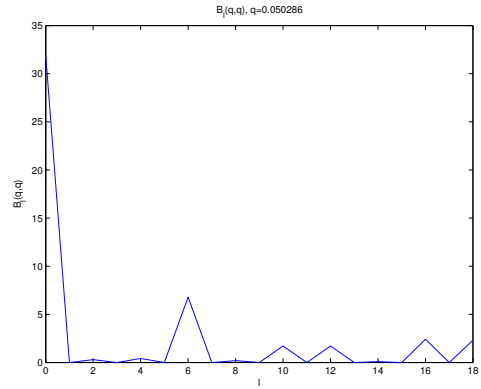
(BA) $Pr772, B_l(q, q), q = 0.0475$



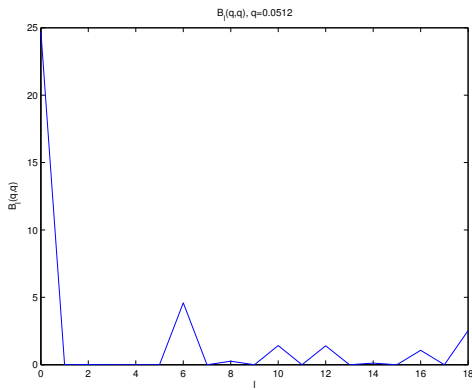
(BB) $Pr772, B_l(q, q), q = 0.0485$



(BC) *Pr772*, $B_l(q, q)$, $q = 0.0494$

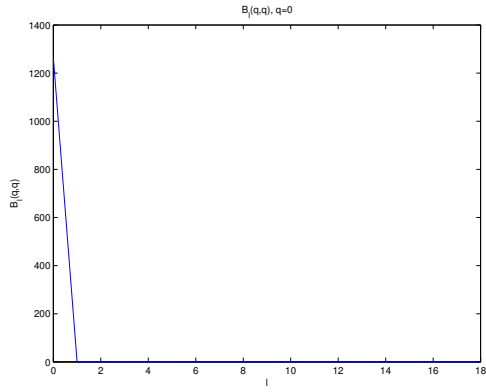


(BD) *Pr772*, $B_l(q, q)$, $q = 0.0503$

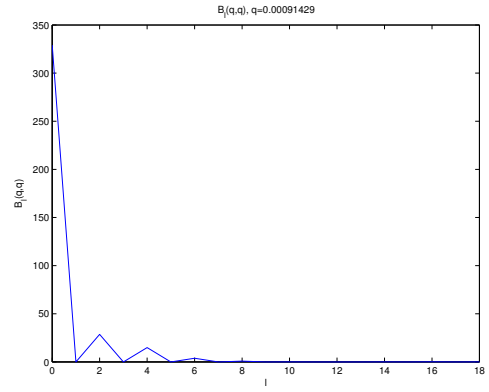


(BE) *Pr772*, $B_l(q, q)$, $q = 0.0512$

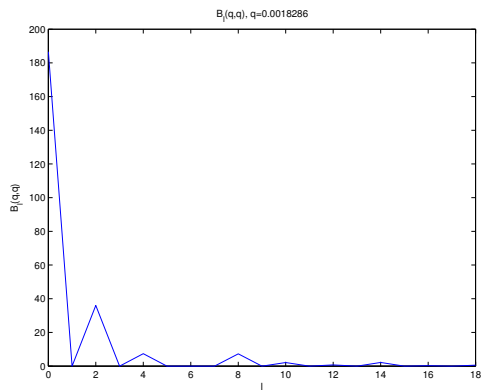
FIGURE 3.2: Angular momentum decomposition of the angular correlations, $B_l(q, q)$. Since the diagonal values ($q = q'$) of B depend on the value of q in addition to l this plot is for B vs. l for a particular value of q (*Pr772*).



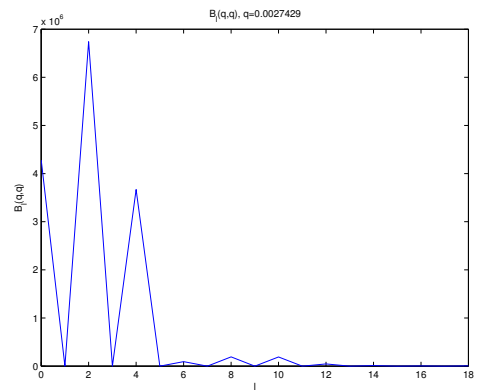
(A) $RDV, B_l(q, q), q = 0$



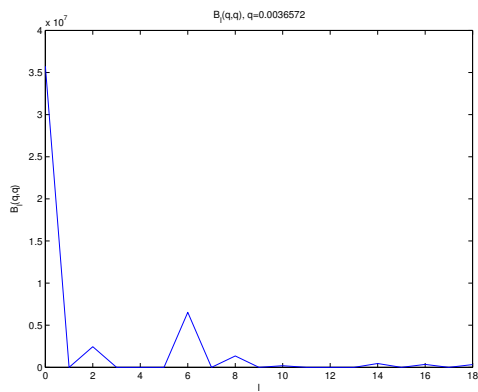
(B) $RDV, B_l(q, q), q = 0.0009$



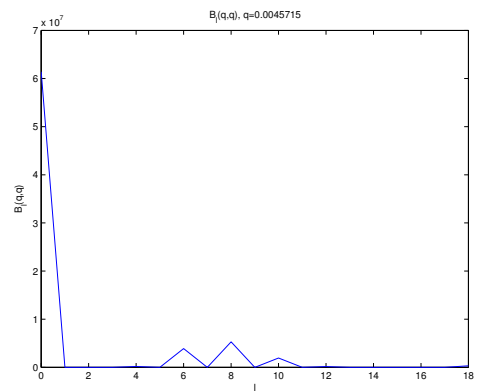
(C) $RDV, B_l(q, q), q = 0.0018$



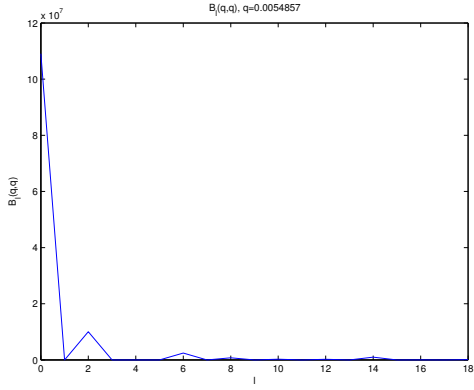
(D) $RDV, B_l(q, q), q = 0.0027$



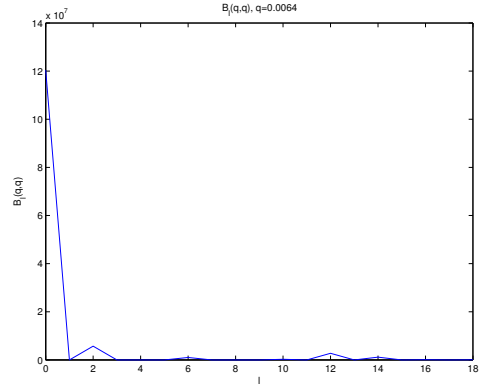
(E) $RDV, B_l(q, q), q = 0.0037$



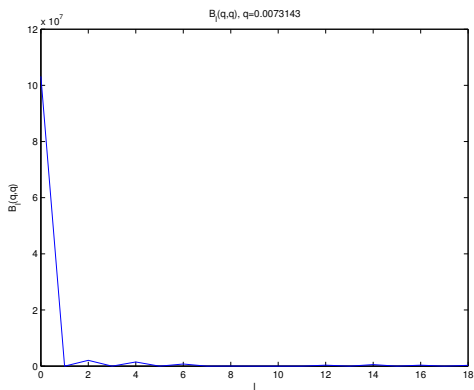
(F) $RDV, B_l(q, q), q = 0.0046$



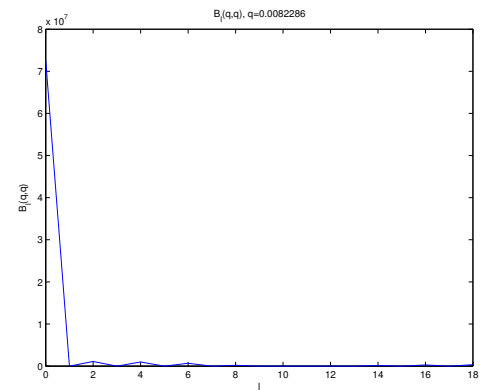
(G) $RDV, B_l(q, q), q = 0.0055e$



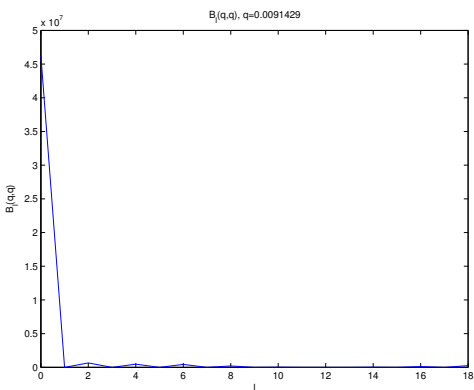
(H) $RDV, B_l(q, q), q = 0.0064$



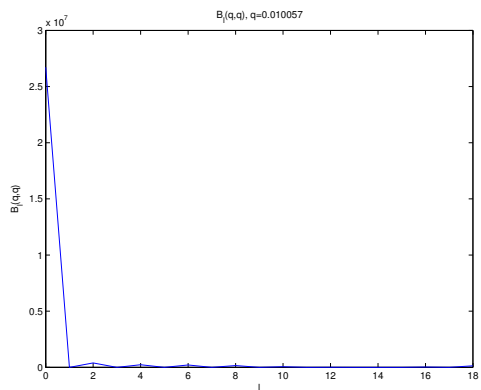
(I) $RDV, B_l(q, q), q = 0.0073$



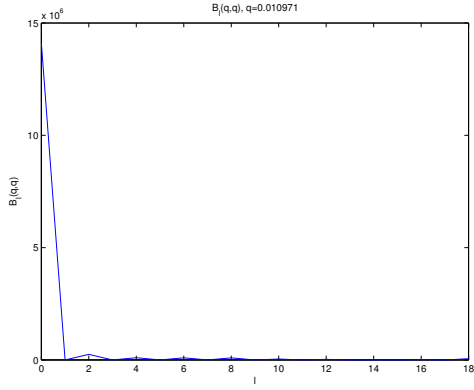
(J) $RDV, B_l(q, q), q = 0.0082$



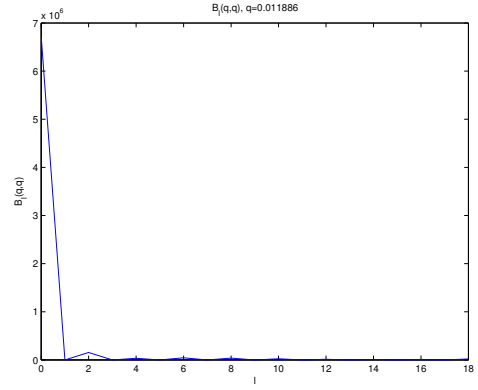
(K) $RDV, B_l(q, q), q = 0.0091$



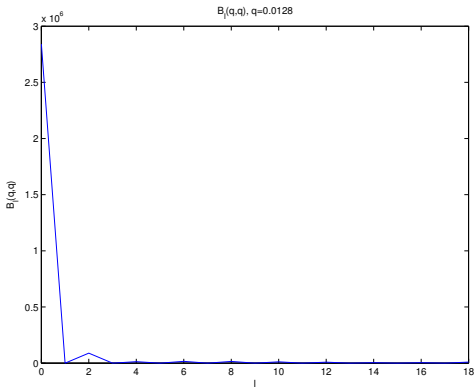
(L) $RDV, B_l(q, q), q = 0.0101$



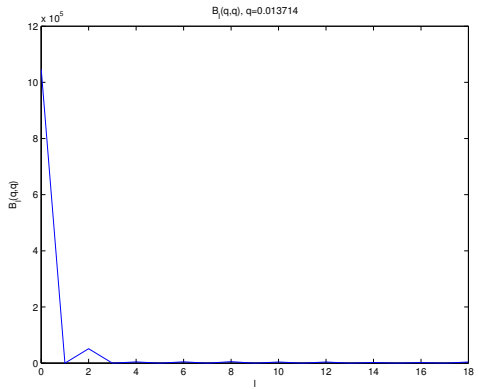
(M) $RDV, B_l(q, q), q = 0.0110$



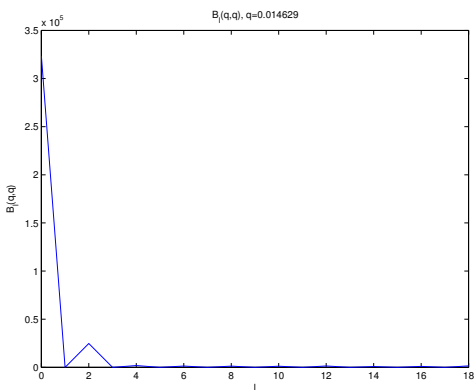
(N) $RDV, B_l(q, q), q = 0.0119$



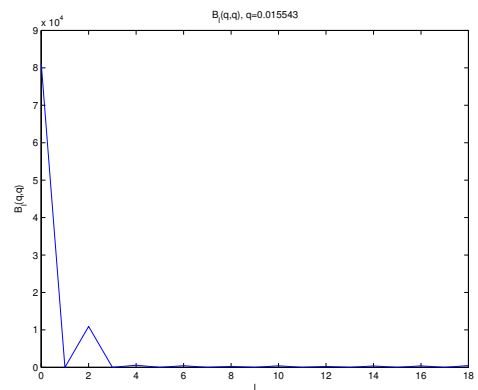
(O) $RDV, B_l(q, q), q = 0.0128$



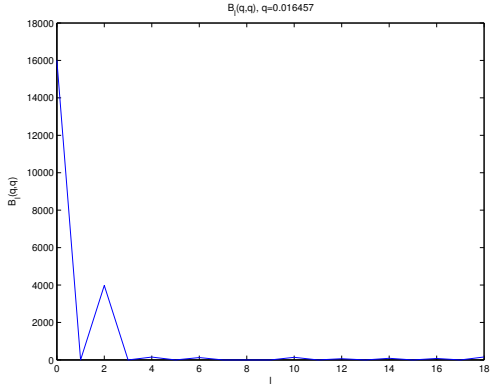
(P) $RDV, B_l(q, q), q = 0.0137$



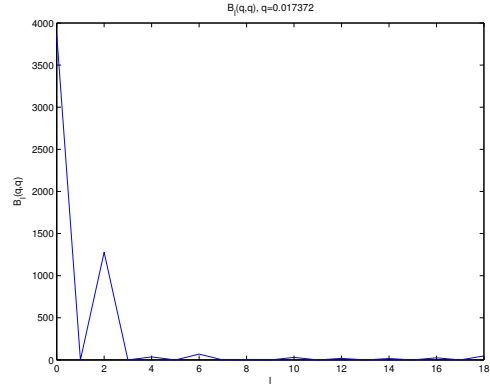
(Q) $RDV, B_l(q, q), q = 0.0146$



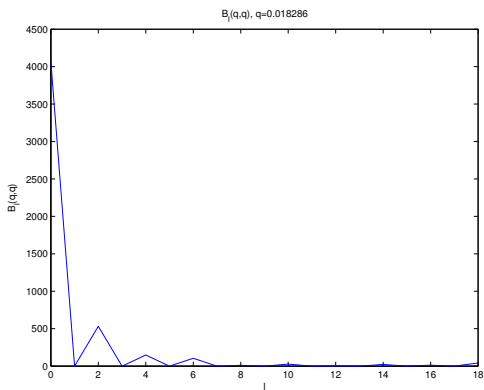
(R) $RDV, B_l(q, q), q = 0.0155$



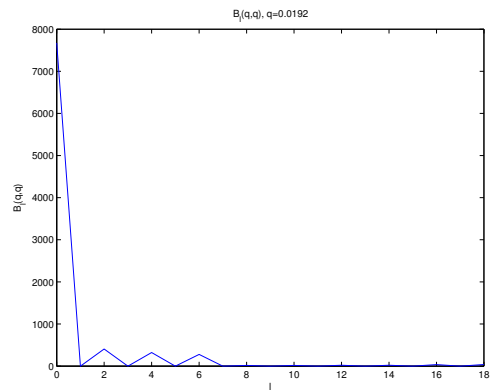
(s) $RDV, B_l(q, q), q = 0.0165$



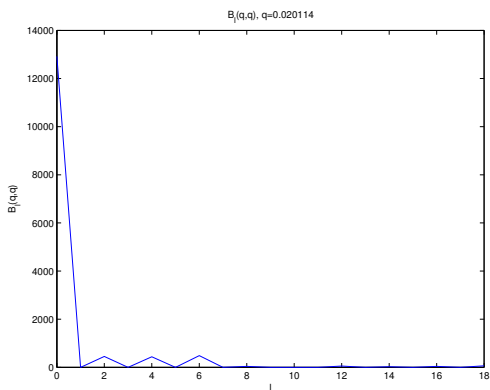
(t) $RDV, B_l(q, q), q = 0.0174$



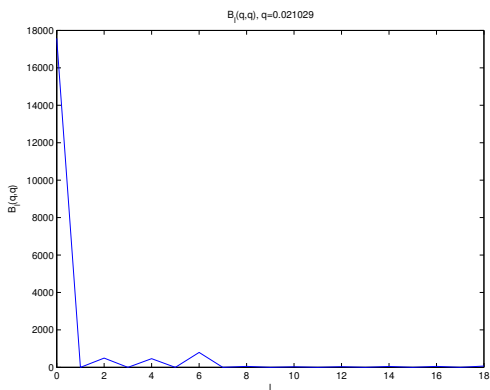
(u) $RDV, B_l(q, q), q = 0.0183$



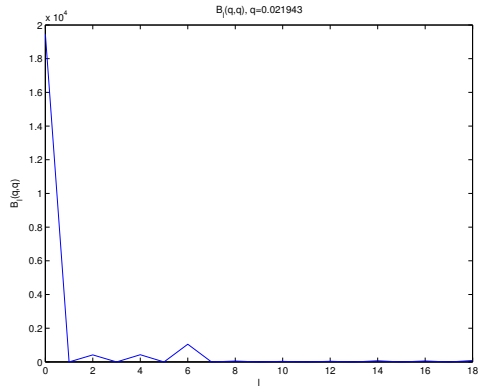
(v) $RDV, B_l(q, q), q = 0.0192$



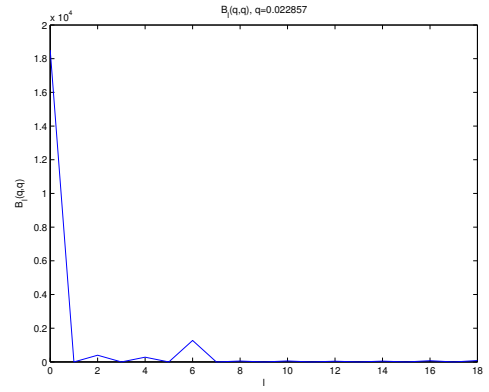
(w) $RDV, B_l(q, q), q = 0.0201$



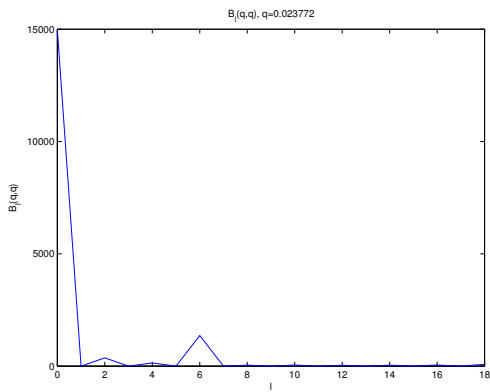
(x) $RDV, B_l(q, q), q = 0.0210$



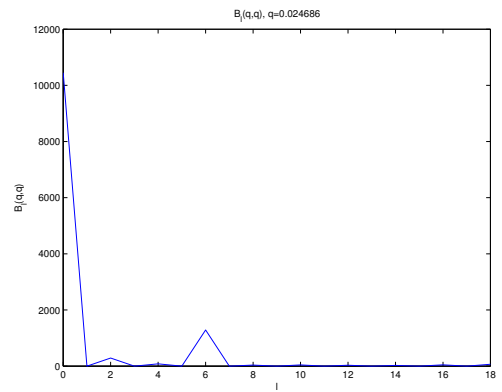
(Y) $RDV, B_l(q, q), q = 0.0219$



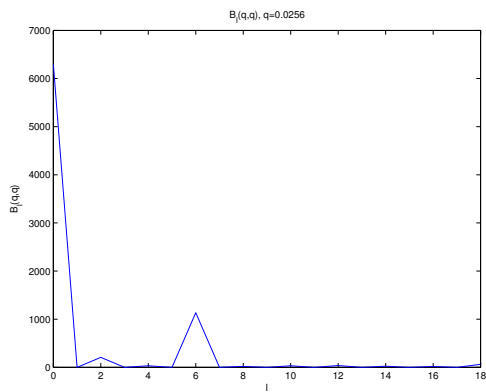
(Z) $RDV, B_l(q, q), q = 0.0229$



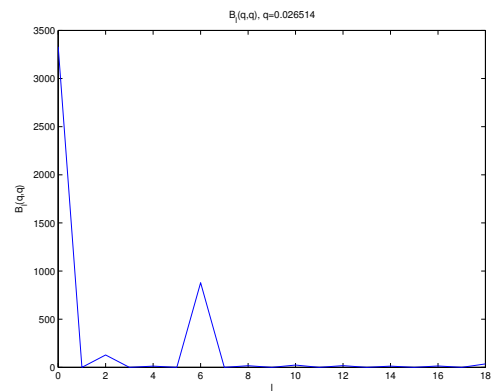
(AA) $RDV, B_l(q, q), q = 0.0238$



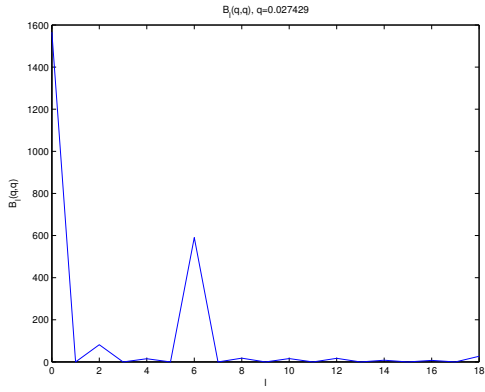
(AB) $RDV, B_l(q, q), q = 0.0247$



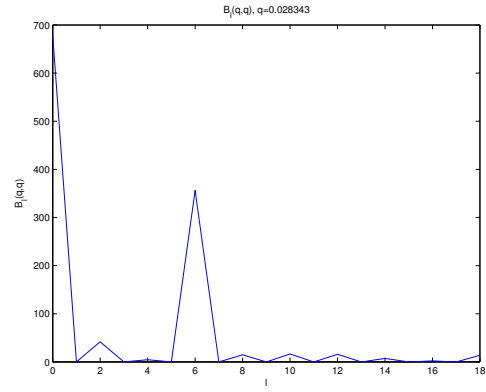
(AC) $RDV, B_l(q, q), q = 0.0256$



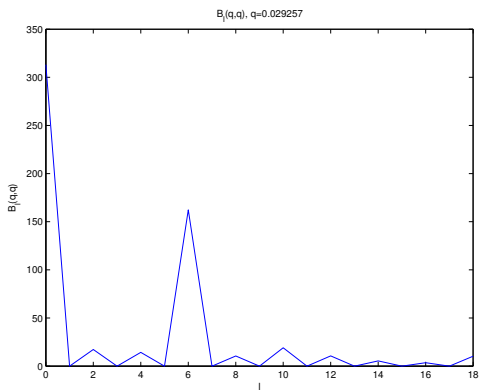
(AD) $RDV, B_l(q, q), q = 0.0265$



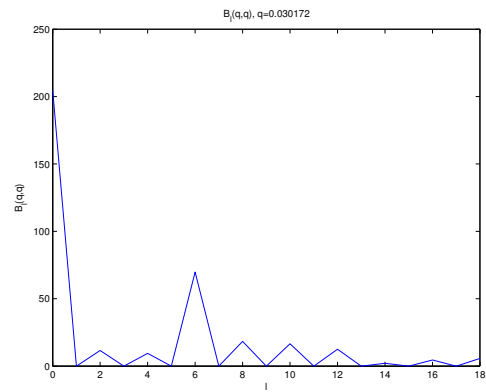
(AE) $RDV, B_l(q, q), q = 0.0274$



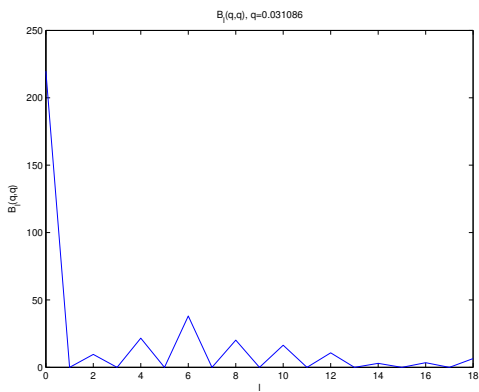
(AF) $RDV, B_l(q, q), q = 0.0283$



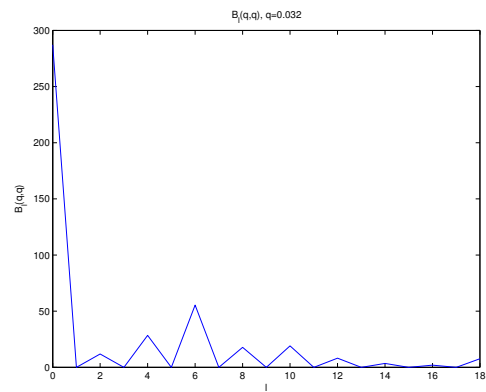
(AG) $RDV, B_l(q, q), q = 0.0293$



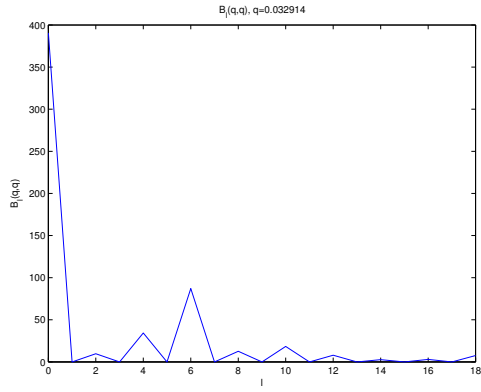
(AH) $RDV, B_l(q, q), q = 0.0302$



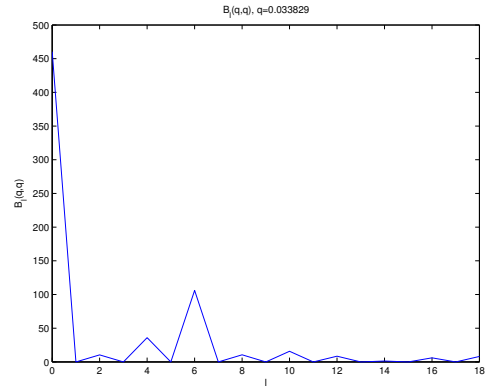
(AI) $RDV, B_l(q, q), q = 0.0311$



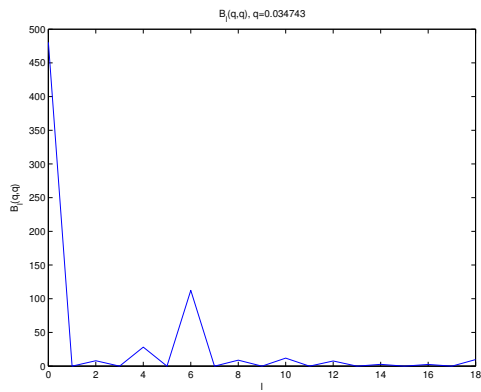
(AJ) $RDV, B_l(q, q), q = 0.0320$



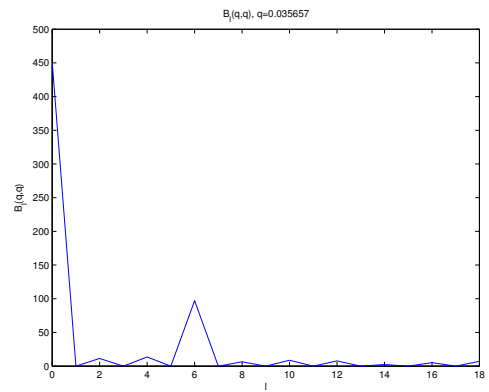
(AK) $RDV, B_l(q, q), q = 0.0329$



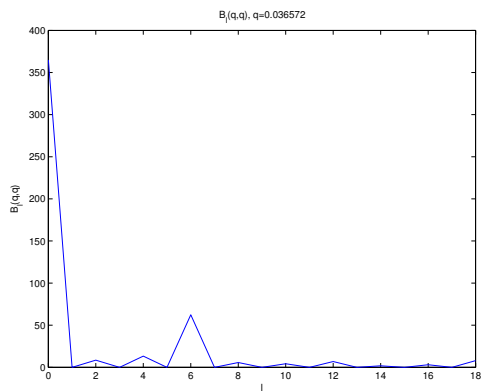
(AL) $RDV, B_l(q, q), q = 0.0338$



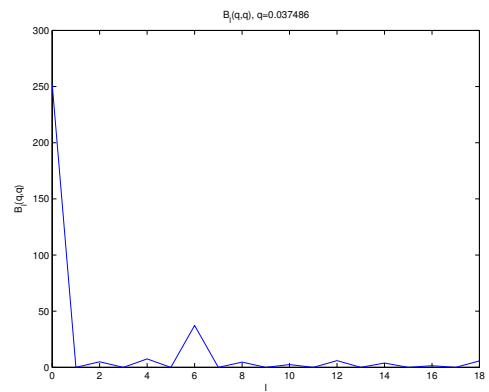
(AM) $RDV, B_l(q, q), q = 0.0347$



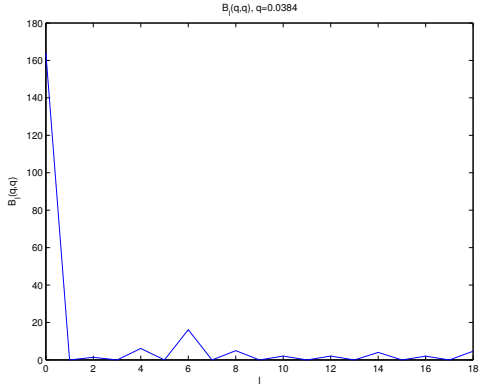
(AN) $RDV, B_l(q, q), q = 0.0357$



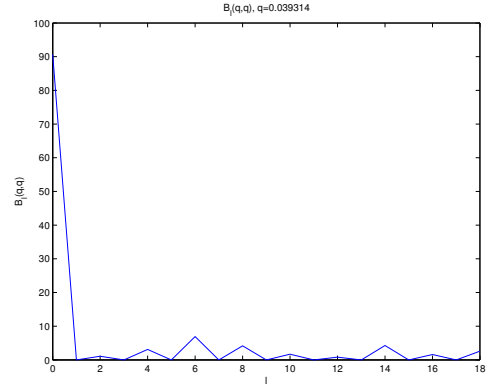
(AO) $RDV, B_l(q, q), q = 0.0366$



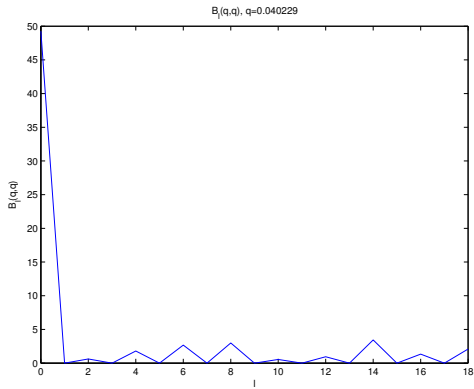
(AP) $RDV, B_l(q, q), q = 0.0375$



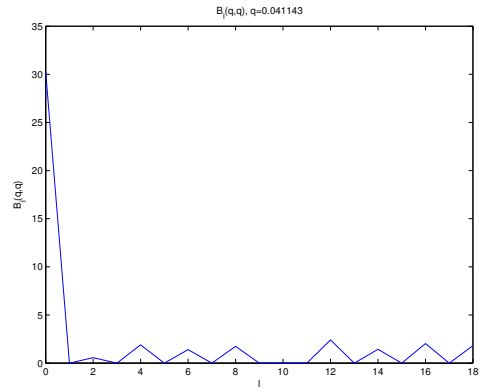
(AQ) RDV, $B_l(q, q), q = 0.0384$



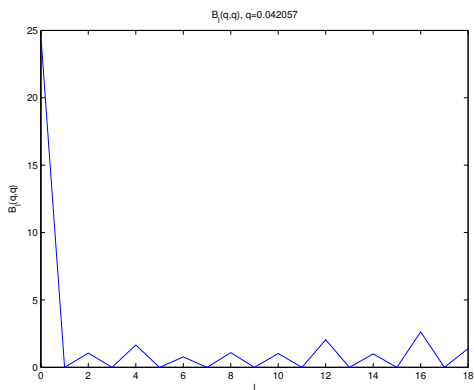
(AR) RDV, $B_l(q, q), q = 0.0393$



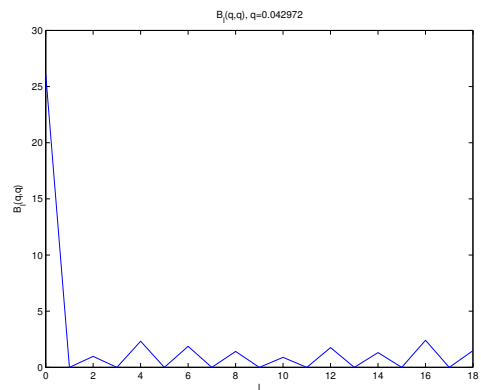
(AS) RDV, $B_l(q, q), q = 0.0402$



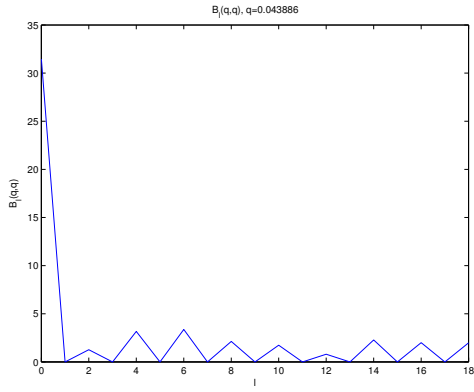
(AT) RDV, $B_l(q, q), q = 0.0411$



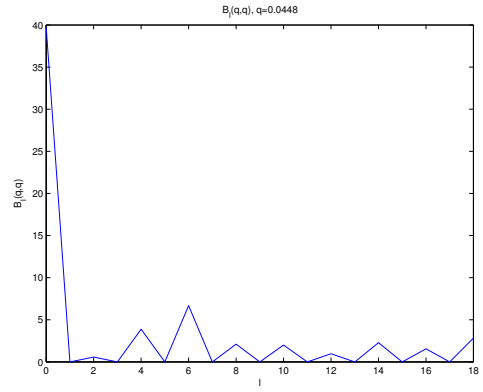
(AU) RDV, $B_l(q, q), q = 0.0421$



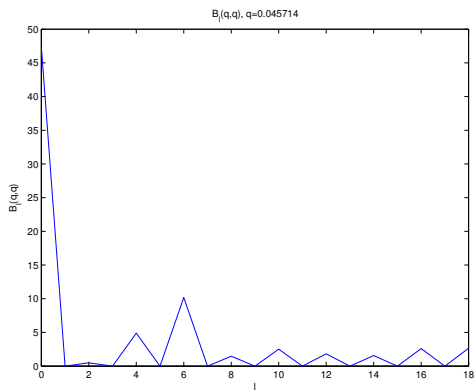
(AV) RDV, $B_l(q, q), q = 0.0430$



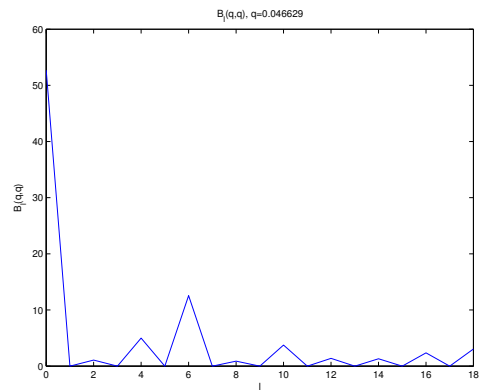
(AW) $RDV, B_l(q, q), q = 0.0439$



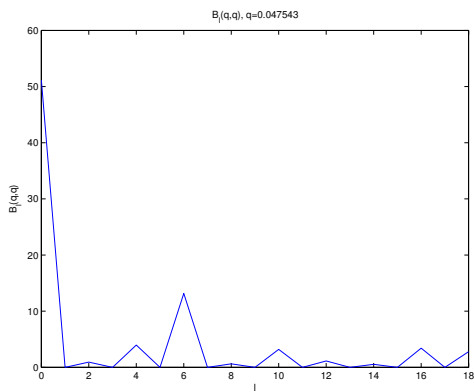
(AX) $RDV, B_l(q, q), q = 0.0448$



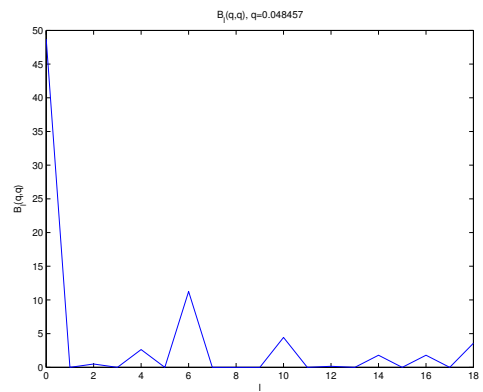
(AY) $RDV, B_l(q, q), q = 0.0457$



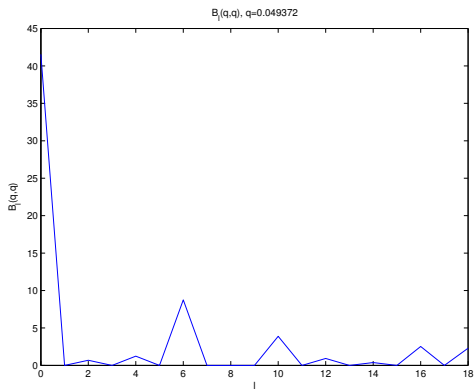
(AZ) $RDV, B_l(q, q), q = 0.0466$



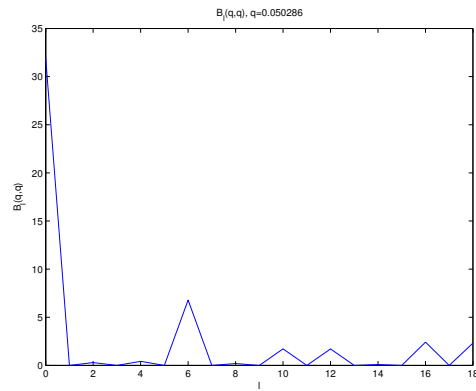
(BA) $RDV, B_l(q, q), q = 0.0475$



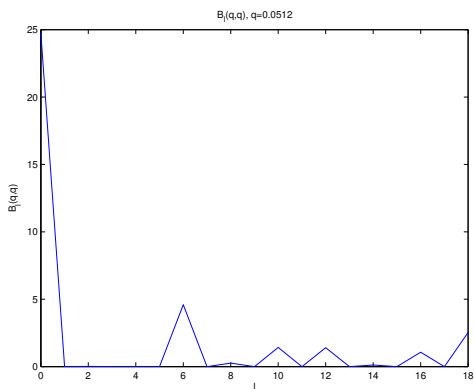
(BB) $RDV, B_l(q, q), q = 0.0485$



(BC) *RDV*, $B_l(q, q)$, $q = 0.0494$



(BD) *RDV*, $B_l(q, q)$, $q = 0.0503$



(BE) *RDV*, $B_l(q, q)$, $q = 0.0512$

FIGURE 3.3: Angular momentum decomposition of the angular correlations, $B_l(q, q)$. Since the diagonal values ($q = q'$) of B depend on the value of q in addition to l this plot is for B vs. l for a particular value of q (*RDV*).

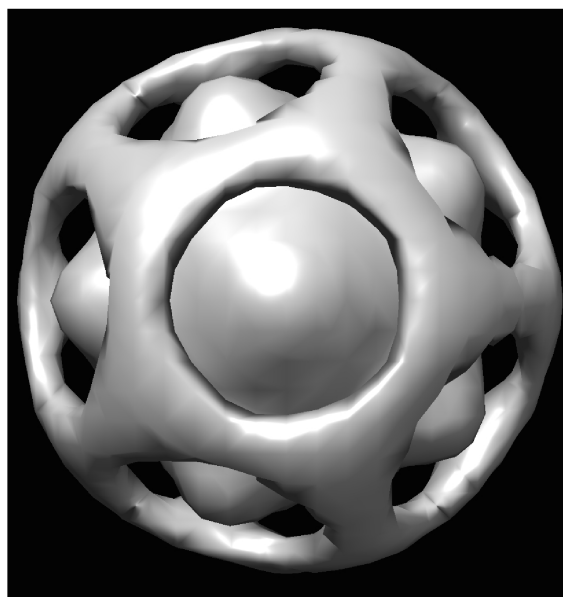


FIGURE 3.4: Diffraction volume of the PR722 virus.

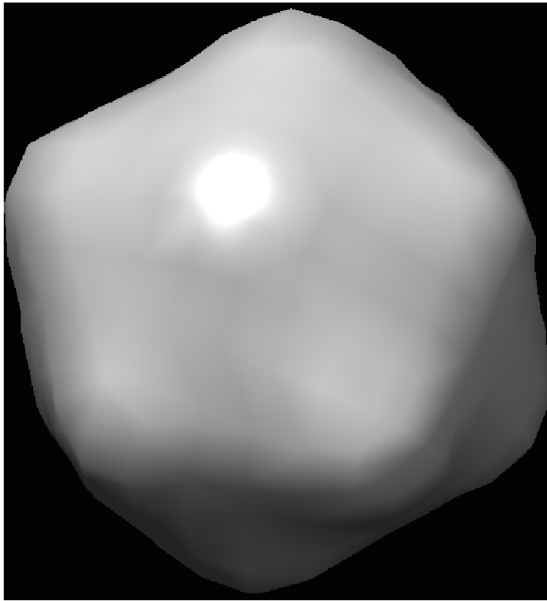


FIGURE 3.5: Electron density of the PR772 virus.

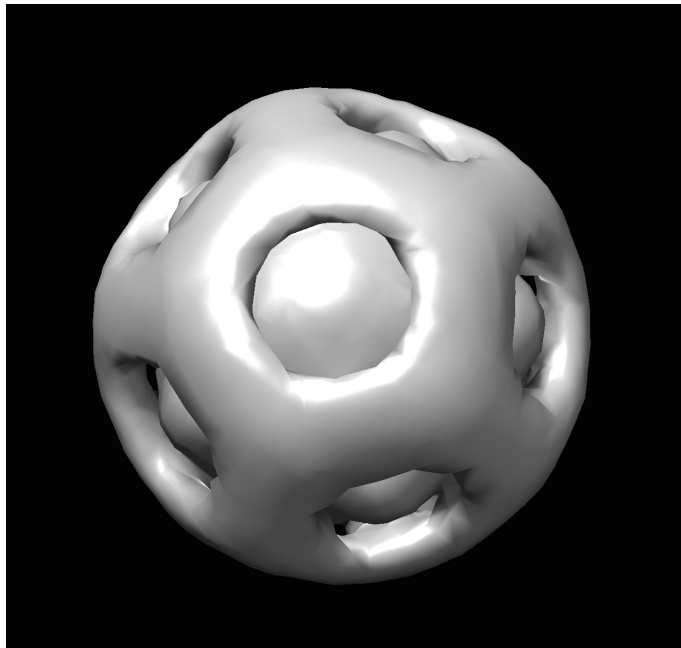


FIGURE 3.6: Diffraction volume of the RDV virus.

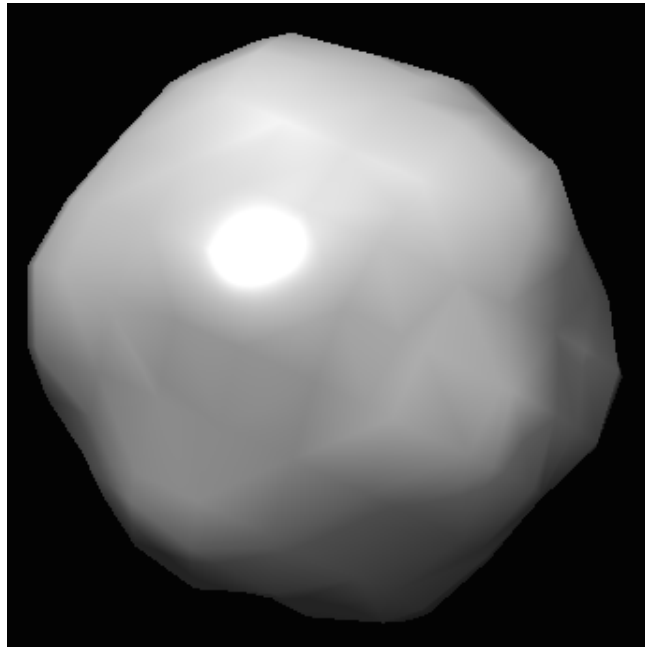


FIGURE 3.7: Electron density of the RDV.

Chapter 4

Validation

A useful measure of resolution has come to us from cryo-EM (Penczek, 2010). The Fourier shell correlation (FSC) is defined by

$$FSC(q) = \sum_{q_i \in q} F_1(q_i) \cdot F_2(q_i)^* / \sqrt{\sum_{q_i \in q} |F_1(q_i)|^2 \cdot \sum_{q_i \in q} |F_2(q_i)|^2} \quad (4.1)$$

Thus the q_i represents a set of data points (on a Cartesian grid) within the shell representing the radius q . Note that despite its appearance, this result is real since the sum involving the imaginary part is zero (Penczek, 2010). Here F_1 and F_2 are the Fourier transforms of the electron densities calculated from two randomly selected subsets of data. q is the magnitude of the spatial frequencies and q_i 's are specific voxels corresponding to this spatial frequency. Since the electron densities are real, Friedel's Law holds. We find that the FSC's fall off with q as would be expected. In general when the $FSC = 0.5$ the corresponding value of q is a measure of the resolution of the data. $(2\pi/q)$ is the resolution. For PR772, the FSC does not become 0.5 until the end of the diffraction pattern is reached (Fig. 4.1) and thus the resolution is limited by the experimental data, whose resolution is about 116 Å at the center of a side of the measured diffraction data (Aquila, 2017). For RDV the resolution is about 200 Å (Fig. 4.2).

Whereas the FSC is sensitive to phases since it involves taking the scalar product of structure factors, a quantity that is much more commonly used in X-ray diffraction is the R-factor that does not include phases. In fact, we also calculated a reliability factor R-split(q), defined by (White et al., 2012):

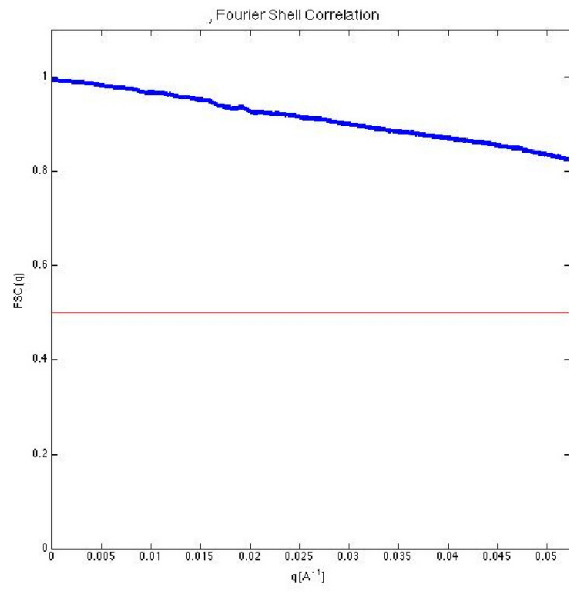


FIGURE 4.1: Fourier shell correlation (Pr772).

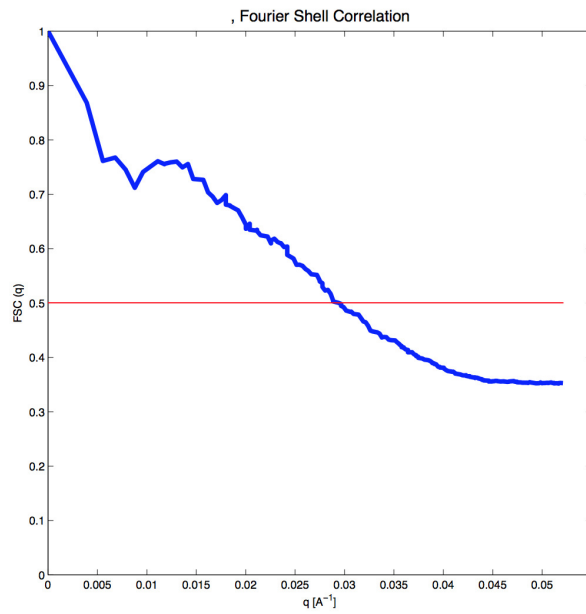


FIGURE 4.2: Fourier shell correlation(RDV).

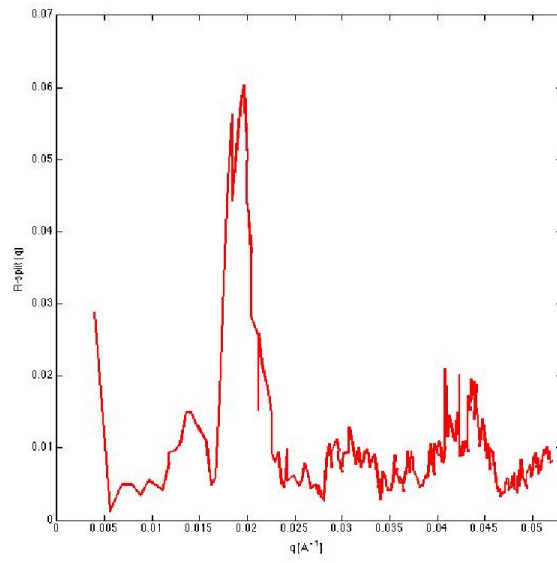


FIGURE 4.3: Rsplit calculation (Pr772).

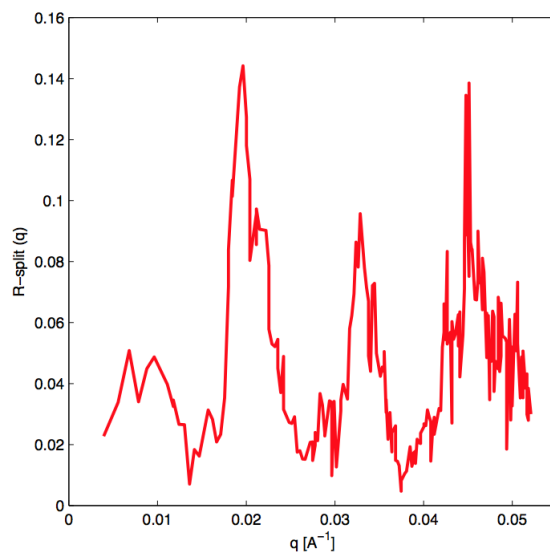


FIGURE 4.4: Rsplit calculation (RDV).

$$R_{split}(q) = \sqrt{2} \sum_{q_i \in q} [I_1(q) - I_2(q)] / \sum_{q_i \in q} [I_1(q) + I_2(q)] \quad (4.2)$$

Of course, since this involves only experimentally measured values, this is closer to quantities obtained only by experiment. In this case the smaller the R_{split} the greater the consistency of the results. As may be seen from (Fig. 4.3 and Fig. 4.4), R_{split} is nowhere greater than about 0.06 (for PR772) and 0.14 (for RDV) in the range of experimental data.

Chapter 5

Conclusion

We emphasize that the calculation of the correlations follows the method pioneered by Hanbury, Brown and Twiss (Brown and Twiss, 1956) in astronomy and does not assume icosahedral symmetry. This symmetry is only assumed in the reconstruction of the diffraction volume having first determined the symmetry by objective means. Despite our careful justification of the use of the icosahedral approximation, it might be argued that it is only the virus capsid that is icosahedral and that one should use a method that is not tied to any particular symmetry. Fortunately, we have developed such a method during our work on time-resolved structures. We propose using the icosahedral approximation only to get an initial estimate $\rho(\mathbf{r})$ of the electron density. If the correct electron density of $\rho_c(\mathbf{r})$ is associated with the experimental correlations $B_{cl}(q, q')$, the perturbation in the electron density $\Delta\rho(\mathbf{r}) = \rho_c(\mathbf{r}) - \rho(\mathbf{r})$ is directly related to $\Delta B_l(q, q') = B_{cl}(q, q') - B_l(q, q')$ (the latter from the icosahedral approximation). The current thinking is that the internal genetic material is probably disordered. This means that unless there are sufficient numbers of the same structure in all orientations, even the method proposed for dealing with the simultaneous occurrences of variations in both conformations and orientations (Schwander et al., 2010) is questionable. There is very little that is known about the structure of the internal genetic material, even whether it is reproducible. If the point about reproducibility is important, especially as what are trying to do is to find a 3D structure from its 2D diffraction patterns, it's difficult to see how such a reconstruction would be possible if variations of structure for different

orientations are random. One might ask which of the many structures is being reconstructed in 3D. Perhaps the only thing possible under such circumstances is to make the reproducibility assumption in at least two different algorithms and to see whether a consistent picture emerges. For example, if at least two different methods find the same structure, including the internal genetic material, this structure is quite likely correct. We believe that this research is crucial as the first step to a symmetry-independent structure determination. Even within the icosahedral structure, it is possible (as we have argued above) that we can use the degenerate icosahedral terms from $l=30$ to about $l=60$. This algorithm has been demonstrated for solution scattering by Zheng et al's (Zheng, Doerschuk, and Johnson, 1995) using a fraction of the data available to an XFEL. Our next step is to see how far we can go with the icosahedral approximation. The idea is then to use the icosahedral approximation only in the first step of determining a diffraction volume, which is undertaken only to get close to the correct structure. In the final stage, which was developed originally for time-resolved structure determination, the aim is to find a perturbed structure starting from a nearby unperturbed one (Pande et al., 2014). Then there would be no symmetry assumption whatsoever.

Bibliography

- Abramson, Jeff et al. (2003). "Structure and mechanism of the lactose permease of *Escherichia coli*". In: *Science* 301.5633, pp. 610–615.
- Alam, Amer and Youxing Jiang (2009). "High-resolution structure of the open NaK channel". In: *Nature structural & molecular biology* 16.1, pp. 30–34.
- Aquila, Andrew (2017). *Coherent Soft X-ray Diffraction Imaging of Coliphage PR772 at the Linac Coherent Light Source*. Tech. rep. SLAC National Accelerator Laboratory (SLAC).
- Bauschke, Heinz H, Patrick L Combettes, and D Russell Luke (2002). "Phase retrieval, error reduction algorithm, and Fienup variants: a view from convex optimization". In: *JOSA A* 19.7, pp. 1334–1345.
- Berman, Helen M et al. (2000). "The protein data bank". In: *Nucleic acids research* 28.1, pp. 235–242.
- Bogan, Michael J et al. (2008). "Single particle X-ray diffractive imaging". In: *Nano letters* 8.1, pp. 310–316.
- Bozek, John D (2009). "AMO instrumentation for the LCLS X-ray FEL". In: *The European Physical Journal-Special Topics* 169.1, pp. 129–132.
- Brown, R Hanbury, Richard Q Twiss, et al. (1956). "Correlation between photons in two coherent beams of light". In: *Nature* 177.4497, pp. 27–29.
- Caspar, Donald LD and Aaron Klug (1962). "Physical principles in the construction of regular viruses". In: *Cold Spring Harbor symposia on quantitative biology*. Vol. 27. Cold Spring Harbor Laboratory Press, pp. 1–24.
- Chapman, Henry N et al. (2006). "High-resolution ab initio three-dimensional x-ray diffraction microscopy". In: *JOSA A* 23.5, pp. 1179–1200.

- Cohan, Norah V (1958). "The spherical harmonics with the symmetry of the icosahedral group". In: *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 54. 01. Cambridge Univ Press, pp. 28–38.
- Damiani, D et al. (2016). "Linac Coherent Light Source data analysis using psana". In: *Journal of Applied Crystallography* 49.2, pp. 672–679.
- DePonte, DP et al. (2008). "Gas dynamic virtual nozzle for generation of microscopic droplet streams". In: *Journal of Physics D: Applied Physics* 41.19, p. 195505.
- Donatelli, Jeffrey J, Peter H Zwart, and James A Sethian (2015). "Iterative phasing for fluctuation X-ray scattering". In: *Proceedings of the National Academy of Sciences* 112.33, pp. 10286–10291.
- Egli, Martin (2010). "Diffraction techniques in structural biology". In: *Current protocols in nucleic acid chemistry*, pp. 7–13.
- Elser, Veit (2003). "Solution of the crystallographic phase problem by iterated projections". In: *Acta Crystallographica Section A: Foundations of Crystallography* 59.3, pp. 201–209.
- Emma, Paul et al. (2010). "First lasing and operation of an ångstrom-wavelength free-electron laser". In: *nature photonics* 4.9, pp. 641–647.
- Eriksson, Mikael, J Friso van der Veen, and Christoph Quitmann (2014). "Diffraction-limited storage rings—a window to the science of tomorrow". In: *Journal of synchrotron radiation* 21.5, pp. 837–842.
- Fienup, James R (1982). "Phase retrieval algorithms: a comparison". In: *Applied optics* 21.15, pp. 2758–2769.
- Fukushi, Teikichi, Eishiro Shikata, and Ikuo Kimura (1962). "Some morphological characters of rice dwarf virus". In: *Virology* 18.2, pp. 192–205.
- Hosseinizadeh, A et al. (2015). "Single-particle structure determination by X-ray free-electron lasers: Possibilities and challenges". In: *Structural Dynamics* 2.4, p. 041601.
- Jack, A and SC Harrison (1975). "On the interpretation of small-angle x-ray solution scattering from spherical viruses". In: *Journal of molecular biology* 99.1, pp. 15–25.

- Johnson, Joanne E and Rosemary B Cornell (1999). "Amphitropic proteins: regulation by reversible membrane interactions (review)". In: *Molecular membrane biology* 16.3, pp. 217–235.
- Kam, Z, HB Shore, and G Feher (1978). "On the crystallization of proteins". In: *Journal of molecular biology* 123.4, pp. 539–555.
- Kolb, Ute et al. (2012). *Uniting Electron Crystallography and Powder Diffraction*. Springer.
- Kurta, RP et al. (2013). "Solution of the phase problem for coherent scattering from a disordered system of identical particles". In: *New journal of physics* 15.1, p. 013059.
- Luke, D Russell (2004). "Relaxed averaged alternating reflections for diffraction imaging". In: *Inverse Problems* 21.1, p. 37.
- Lute, Scott et al. (2004). "Characterization of coliphage PR772 and evaluation of its use for virus filter performance testing". In: *Applied and environmental microbiology* 70.8, pp. 4864–4871.
- MacKinnon, Roderick et al. (1998). "Structural conservation in prokaryotic and eukaryotic potassium channels". In: *Science* 280.5360, pp. 106–109.
- Marchesini, Stefano et al. (2003). "X-ray image reconstruction from a diffraction pattern alone". In: *Physical Review B* 68.14, p. 140101.
- Oszlányi, Gábor and Andras Sütő (2004). "Ab initio structure solution by charge flipping". In: *Acta Crystallographica Section A: Foundations of Crystallography* 60.2, pp. 134–141.
- Pande, K et al. (2014). "Deducing fast electron density changes in randomly orientated uncrystallized biomolecules in a pump–probe experiment". In: *Phil. Trans. R. Soc. B* 369.1647, p. 20130332.
- Pedrini, B et al. (2013). "Two-dimensional structure from random multiparticle X-ray scattering images using cross-correlations". In: *Nature communications* 4, p. 1647.
- Penczek, Pawel A (2010). "Chapter one-fundamentals of three-dimensional reconstruction from projections". In: *Methods in enzymology* 482, pp. 1–33.
- Poon, H-C et al. (2013). "Fiber diffraction without fibers". In: *Physical review letters* 110.26, p. 265505.

- Poon, HC and DK Saldin (2015). "Use of triple correlations for the sign determinations of expansion coefficients of symmetric approximations to the diffraction volumes of regular viruses". In: *Structural Dynamics* 2.4, p. 041716.
- Roedig, Philip et al. (2016). "Room-temperature macromolecular crystallography using a micro-patterned silicon chip with minimal background scattering". In: *Journal of applied crystallography* 49.3.
- Saldin, DK et al. (2009). "Structure of isolated biomolecules obtained from ultrashort x-ray pulses: exploiting the symmetry of random orientations". In: *Journal of Physics: Condensed Matter* 21.13, p. 134014.
- Saldin, DK et al. (2010). "Beyond small-angle x-ray scattering: Exploiting angular correlations". In: *Physical Review B* 81.17, p. 174105.
- Saldin, DK et al. (2011). "Reconstructing an icosahedral virus from single-particle diffraction experiments". In: *Optics express* 19.18, pp. 17318–17335.
- Schwander, P et al. (2010). "Mapping the conformations of biological assemblies". In: *New Journal of Physics* 12.3, p. 035007.
- Shapiro, D et al. (2013). "Development of coherent scattering and diffractive imaging and the COSMIC facility at the Advanced Light Source". In: *Journal of Physics: Conference Series*. Vol. 425. 19. IOP Publishing, p. 192011.
- Sirohi, Devika et al. (2016). "The 3.8 Å resolution cryo-EM structure of Zika virus". In: *Science* 352.6284, pp. 467–470.
- Starodub, Dmitri et al. (2012). "Single-particle structure determination by correlations of snapshot X-ray diffraction patterns". In: *Nature communications* 3, p. 1276.
- White, Thomas A et al. (2012). "CrystFEL: a software suite for snapshot serial crystallography". In: *Journal of Applied Crystallography* 45.2, pp. 335–341.
- Yoon, Chun Hong et al. (2011). "Unsupervised classification of single-particle X-ray diffraction snapshots by spectral clustering". In: *Optics express* 19.17, pp. 16542–16549.
- Zheng, Yibin, Peter C Doerschuk, and John E Johnson (1995). "Determination of three-dimensional low-resolution viral structure from solution x-ray scattering data". In: *Biophysical journal* 69.2, pp. 619–639.

Curriculum Vitae

FATEMEHSADAT JAMALIDINAN

Education

**DOCTOR OF PHILOSOPHY| UNIVERSITY OF WISCONSIN - MILWAUKEE
|USA|2014- 2017**

- Major: Electrical Engineering
- Thesis: 3D RECONSTRUCTION OF PROTEINS AND VIRUSES FROM ANGULAR CORRELATIONS OF THE SCATTERED INTENSITIES

PATENT

- **F. Jamalidinan, A.R.Taheri**, “Secure watermarking in image with lossless compression”, *Invention Declaration. No (13915014000302811), Patent. No (77124), Department of Registration of Company and Industrial Ownership, Tehran, Iran, 2012.*

Selected Publications

Books

- **F. Jamalidinan, M.Fateh**, “Common Types of Video Files and Applications”, in Persian, RCISP, Iran.

Book Chapter

- **F. Jamalidinan, P. Mosayebi, H. AbrishamiMoghaddam, M. giti, S. Kermani**, “A Fully 3D system for cardiac wall deformation analysis in MRI data”, Book chapter in *Functional Imaging and Modeling of the Heart, Lecture Notes in Computer Science LNCS 4466, Springer, 12-21, 2007.*

Journal Papers

- **F. Jamalidinan, P. Mosayebi, H. AbrishamiMoghaddam, M. Giti**, “A dynamic surface model for segmentation of cardiac MRI images and its usage in cardiac wall motion tracking”, *Measurement Science Review, 7(2), no. 1, 5-11, 2007.*
- **M.Beygzadeh, F. Jamalidinan, M.Rezaiee**, “Jpeg image Steganalysis based on classification of statistical features and two stage decision making”, *Journal of Signal and Data Processing, Iran, Iran, no.1, 53-70, 2009.*
- **F. Jamalidinan, M.Rezaiee, M.Beygzadeh**, “Review of JPEG steganography methods and their security analysis”, *Accepted for publication in Journal of Signal and Data Processing, Iran, 2010.*
- **F. Jamalidinan, M. Uddin, X. Lordu, H. De Mirci, H. K. N. Reddy, R. G. Sierra, M. Svenda, C. H. Yoon, A. Aquila, and D. K. Saldin**, “Reconstruction of icosahedral images of the PR722 virus from experimental data of the LCLS' Single Particle Initiative”, submitted.

Conference Papers

- **F. Jamalidinan**, P. Mosayebi, H. AbrishamiMoghaddam, M. giti, “A dynamic surface model for segmentation of cardiac MRI images and its usage in cardiac wall motion tracking”, Proc. 6th International Conference on Measurement (Measurement 2007), Slovakia, May 2007.
- **F. Jamalidinan**, P. Mosayebi, H. AbrishamiMoghaddam, M. Giti, S. Kermani, “Enhancement of the algorithm for tracking wall motion in 3D cardiac MRI sequences using active mesh model”, Proc. 15th Iranian Conference on Electrical Engineering (ICEE2007), Tehran, Iran, May 2007.
- **F. Jamalidinan**, P. Mosayebi, H. AbrishamiMoghaddam, M. giti, S. Kermani, “Tracking of cardiac motion in 3D MRI sequences using active mesh model”, Proc. 13th Iranian Conference on Biomedical Engineering, Tehran, Iran, Feb. 2007.