Protein structure determination by electron diffraction of 3-dimensional protein microcrystals

Matthew Gregory Iadanza

A dissertation
submitted in partial fulfillment of the
requirements for the degree of:

Doctor of Philosophy

University of Washington
2014

Reading Committee:

Tamir Gonen, Chair
Alexey Merz
Ning Zheng

Program Authorized to Offer Degree:
Biochemistry

University of Washington

Abstract

Protein structure determination by electron diffraction of 3-dimensional protein microcrystals

Matthew Gregory Iadanza

Chairperson of the Supervisory Committee:
Tamir Gonen
Affiliated Associate Professor, Department of Biochemistry
Group Leader, HHMI Janelia Farm Research Campus

Crystallographic methods for protein structure determination are well established and have provided valuable insight into the workings of the biological world. X-ray crystallography of three dimensional crystals has benefited from the development of high throughput screening methods, but the method still requires the production of large well ordered crystals. Crystal screening and optimization is an extremely time consuming bottleneck and many proteins, often of high biological interest, are simply not able to be crystallized. Electron crystallography of two dimensional crystals is a much more limited method applicable only to extremely thin two-dimensional crystals, usually composed of protein and lipid. This method has provided insight into the structures and protein-lipid interactions of some integral membrane proteins but is extremely low throughput and not broadly applicable. This work presents a new method for protein structure determination which is essentially a hybrid of these two methods. A focused electron beam is diffracted by very small three dimensional protein microcrystals allowing for diffraction data to

be collected from crystals at least six orders of magnitude smaller than those used for traditional X-ray crystallography. This method allows data collection and subsequent structure determination using crystals that would be considered 'initial hits' in a crystallization screen and would normally require a great deal of optimization before yielding data for structure determination. As proof of principle the structure of hen egg white lysozyme determined to 3.0 Å resolution by microcrystal electron diffraction is presented. The procedures for data collection are detailed as well as an explanation of new software developed for processing the resulting data and possible future refinements and improvements to the technique. With further refinement microcrystal electron diffraction may prove useful for many proteins that are currently considered intractable because of their failure to form large well ordered three dimensional crystals .

Table of Contents

List of Figures

List of Tables

List of Appendices

Dedication

This work is dedicated to my parents Nicholas and Margret Iadanza without whose love and support it would have never been completed.

Acknowledgements

I must first and foremost acknowledge my advisor Dr. Tamir Gonen for all of his assistance and support during my graduate career. My graduate committee Drs. Alexey Merz, Dana Miller, Bill Zagotta, and Ning Zheng, who provided guidance and advice and showed infinite patience with helping me coordinate my last two years of graduate study from across the country.

All good science is collaborative, and this is no exception. All of this work was done in close collaboration with Dr. Dan Shi and Dr. Brent Nannenga. I am thankful for all of their assistance and much helpful discussion in both the theoretical and experimental aspects of the project. Their good humor and encouragement ("Easy math! See Easy math!" – Dan Shi Ph.D.) actually made crunching numbers fun.

Several members of the Gonen lab were instrumental in my development as a scientist. Dr. Goragot 'George' Wisedchaisri was my resident crystallography guru. Dr. Steve Reichow was instrumental in helping me both develop a critical scientific mind and build the skills to communicate effectively. Dr. Maen Sarhan provided intelligent debate, much input on experimental design, and encouragement to go to the gym.

Finally nothing would have been accomplished without the support staff at Janelia Farm, especially our lab coordinator James Martin, who dealt with all of my real world concerns so I could concentrate on science.

1. Introduction

… All bodies of sensible magnitude, whether liquid or solid, are constituted of a vast number of extremely small particles, or atoms of matter bound together by a force of attraction, which is more or less powerful according to circumstances, and which as it endeavors to prevent their separation… …it is one great importance of this work, to shew (sic) *the importance and advantage of ascertaining the relative weights of the ultimate particles, both of simple and compound bodies, the number of simple elementary particles which constitute one compound particle, and the number of less compound particles which enter into the formation of one more compound particle.*

John Dalton

A New System of Chemical Philosophy 1803

In his seminal 1803 work *A New System of Chemical Philosophy* John Dalton laid the groundwork for a new way of thinking about chemical structure. At that point in time concept of the atom, a small indivisible unit of which all matter is composed, had already existed for over 2000 years, since the time of Leucippus in the 5th century BCE. Early atomists assumed all substances were composed of one type of atom whose physical properties mirrored those of the macro scale substance. Dalton's insight was that chemical compounds are composed of different types of atoms, differentiated by atomic weight (today measured in Daltons, named

in his honor) and the ratios of different types atoms determined the chemical properties of the substance.

Dalton made no mention of another even more important factor in the current understanding of how the atomic makeup of a chemical substance affects its chemical properties; the arrangement of the atoms relative to each other in three dimensional space. The first actual visualization of this, and a powerful concrete demonstration of the atomic nature of matter was provided just over 100 years later in 1912 when William Henry Bragg and son William Lawrence Bragg demonstrate it was possible to back calculate the position of each the atoms in a crystal by observing how it diffracted a beam of X-rays.[1]

The Braggs' initial insights into the atomic structure of simple inorganic crystals such as sodium chloride soon yielded information about much more complicated and biologically interesting molecules. Fewer than 50 years after the Braggs' initial publication Kendrew et al. published the first structure of a protein to near atomic resolution.[2]

Today knowledge of three-dimensional structure of proteins gives us insight into their functions and powerful opportunities to understand how to modulate these functions. For proteins of an enzymatic nature the interaction between the protein and its substrate can be visualized with atomic level resolution. This gives us greater understanding of the biophysical process of enzymatic action has paved the way for the rational design of enzymes and other proteins with novel functions. Rational redesign of homing endonucleases has produced modified enzymes that cut novel DNA sequences[3] and modified enzymes have been created that catalyze

reactions on non-natural substrates.[4] In some not-so-distant future, researchers may simply be able to design an enzyme from scratch to catalyze a desired chemical reaction .

We are now able to visualize how and where various substrates, antagonists, and agonists bind their target proteins. As many drugs act by modulating protein function, the ability to visualize these interactions allows researchers to make targeted modifications to drug molecules. For example, the high resolution structure of influenza neuraminidase[5] gave insight that allowed for the improvement of know low affinity inhibitors, increasing both affinity and selectivity, eventually leading to the FDA approved drug Relenza (Zanamivir).[6]

Today, 55 years after the first protein structure was published, the protein data bank [7] contains over 90,000 entries.[8] Each one is a high resolution structure showing not only *"the number of simple elementary particles which constitute one compound particle*" which Dalton correctly surmised was "*of great importance*" but also how those elementary particles are arranged in three-dimensional space, which turns out to be equally important in determining their physical properties and biological function.

A variety of methods are currently available for protein structure determination but many proteins, including some of high biological interest, have remained intractable. Here I present electron diffraction of protein micro crystals (microED), a method for protein structure determination that draws from two established methods; X-ray crystallography and electron diffraction.

MicroED allows for structure determination from protein crystals several orders of magnitude smaller than those currently necessary for X-ray crystallography. This method may provide a means to determine high resolution structures of biologically important proteins that have so far proved intractable because of the inability to grow large, well-ordered crystals. This work presents a brief overview of current crystallographic methods for protein structure determination, their strengths and weaknesses, and the niche these methods currently leave open. Proof of principle is presented as the structure of hen egg white lysozyme, a common and well-studied protein, determined using the new microED method. I discuss the potential for technical improvements to the method, and the technique's broader applicability to novel unsolved protein structures and difficult targets that have so far eluded structure determination.

2. Current crystallographic techniques for protein structure determination

2.1 X-ray crystallography of 3-dimensional crystals

X-ray crystallography is a direct descendent of the Braggs' experiments at the turn of the 20th century and currently the most commonly used method for protein structure determination. Over 90% of the structures in the PDB were determined using this method.[8]

In 1958 Kendrew *et al*. determined the structure of sperm whale myoglobin to approximately 4 Å resolution.[2] Although their paper makes reference to Perutz *et al*.'s earlier determination of the structure of haemoglobin[9], this was the first published glimpse of the structure of a protein. They commented:

> *Perhaps the most remarkable features of the molecule are its complexity and its lack of symmetry. The arrangement seems to be almost totally lacking in the type of regularities which one instinctively anticipates and it is more complicated than has been predicted by any theory of protein structure.* [2]

The structure of myoglobin was quickly refined and solved to 2 Å resolution by the same group only two years later.[10] This level of resolution provided the first near atomic resolution, allowing the fitting of an atomic model to the observed structure.

Although the 2 Å structure of myoglobin is still quite respectable in modern terms, recent work has pushed the resolution limits of X-ray crystallography. A structure of the water channel aquaporin reached 0.88 Å resolution.[11] At this sub-angstrom resolution it is not only possible to observe the protein's structure, but also get insight into its function. The aquaporin channel allows passage of water molecules through its pore while excluding much smaller protons and hydroxide molecules. It was hypothesized that this is accomplished through several electrostatic interactions that repel the charged molecules and break the chain of hydrogen bonded water molecules.[12] With the 0.88 Å structure of aquaporin it was actually possible to visualize hydrogen bond interactions and the accurate positions of water molecules, which provided strong evidence supporting this hypothesis.[11]

2.1.1 Diffraction theory and the basis of X-ray crystallography

Any discussion of crystallography must begin with a most basic definition: what is a crystal? A material is considered crystalline when it is composed of multiple repeating units arranged in a regular manner. The spacing and arrangement of these repeating units can be described mathematically by only six measurements: three distances and three angles. These distances and angles (by convention a, b, c and α, β, γ) define the unit cell, the smallest repeating unit that can be used to recreate the entire crystal structure through only geometric translations. Some simple inorganic crystals, such as cesium chloride, contain only a single molecule, consisting of two atoms, in the unit cell. Protein crystals are

generally more complex, with unit cells containing protein (sometimes multiple copies), water molecules, and often other molecules such as ligands, metal cofactors, or other molecules present in the solvent during crystallization.

The interaction of waves with the repeating lattice is the basis of both X-ray and electron diffraction. When electromagnetic waves pass through matter they can be scattered, emerging with a modified trajectory. If a wave scatters with no loss of energy it is said to have been scattered elastically. If the wave transfers some of its energy to the matter it has interacted with it emerges with a different wavelength. This phenomenon is called inelastic scattering. Scattered waves also interact with each other. Scattering by an unstructured mass of matter leads to patterns of constructive and destructive interference too convoluted to easily interpret. When waves are elastically scattered by the repeating lattice of a crystal the interference results in predicable patterns of interference. Bragg's law describes the condition that leads to constructive interference:

$$n\lambda = 2dsin\theta$$

where d is the spacing of the repeating lattice, θ is the angle between the incident and diffracted waves, n is any positive integer, and λ is the wavelength. Bragg's law allows calculation of the lattice size necessary to diffract a given wavelength:

$$d = \frac{\lambda}{2}$$

For the example of red light with a of wavelength 700 nm the minimum required lattice spacing, and theoretical maximum resolvable resolution, is 350 nm. This is on the scale of cells but much too large to resolve the atomic structure of individual molecules. High quality X-ray radiation such as that from a modern synchrotron source has a wavelength of approximately 1 Å (0.01 nm) allowing for a maximum theoretical resolution of 0.5Å, approximately half the distance between a carbon-carbon bond. This makes X-ray radiation well suited for examining molecular structure at the atomic scale. Recent X-ray crystallography structures approach this resolution limit.[11]

When a recording device is used to measure the diffraction of electromagnetic waves by a crystal the constructive interference described by Bragg's law leads to a predictable pattern of diffraction spots or Bragg peaks. These patterns contain valuable information about the crystal.

2.1.2 From spots to structure

The Fourier transform de-convolutes a complicated waveform into the set of simple sinusoids that make it up, called a Fourier series. Any sinusoid can be described by three terms: wavelength, amplitude, and phase. A diffraction pattern is essentially a 2-dimensional slice of the 3-dimensional Fourier transform of an object, containing information about the wavelength and relative amplitude of the waves in the Fourier series. A series of diffraction patterns from a crystal taken at different angles, along with the waves' phases, which must be determined

experimentally or computationally, allows for the reconstruction of the entire three dimensional Fourier transform of the crystal unit cell. A inverse Fourier transform then allows the determination of the crystal's structure in real space.

Each spot in a diffraction pattern corresponds to a structure factor (*F*) which is contributed to by the scattering from all atoms in the crystal. The structure factor is the product of two terms, amplitude (*f*) and phase shift (Φ). The phase shift of an atom *a* at any given Miller index *hkl* is:

$$\phi_a = 2\pi(hx_a + ky_a + lz_a)$$

Where x_a, y_a, and z_a describe the position of the atom. Hence, the structure factor for any given atom *a* is:

$$F_a = f_a e^{i\phi_a} = f_a[\cos 2\pi(hx_a + ky_a + lz_a) + i \sin 2\pi(hx_a + ky_a + lz_a)]$$

Each spot on the diffraction pattern is the sum of the scattering from every atom in the crystal, so the structure factor for any spot *hkl* is simply the summation of the structure factors at that Miller index for every atom:

$$F_{hkl} = \sum_a f_a e^{i\phi_a} = \sum_a f_a[\cos 2\pi(hx_a + ky_a + lz_a) + i \sin 2\pi(hx_a + ky_a + lz_a)]$$

A Fourier transform of the above de-convolutes this mess into the contributions of each individual atom. With all of the measured diffraction spots taken into account, the electron density (ρ) position *xyz* in the crystal is estimated as:

$$\rho_{xyz} = \frac{1}{V} \sum_{hkl} F_{hkl} e^{-i2\pi(hx+ky+lz)}$$

where *V* is the volume of the unit cell.

The amplitude required for calculation of structure factors is obtained from the diffraction pattern by measuring the relative intensity of each diffraction spot. Phase information is not present the diffraction pattern and must be determined using other experimental approaches.

A huge body of work exists on methods for solving the 'phase problem' in crystallography.[13] Phasing methods broadly fall into two categories, *ab initio* or experimental phasing. Both methods produce initial phase estimates that provide a starting point for determining the actual phase angles. *Ab initio* phasing uses only data from the target crystal and derivatives while model based methods use a starting model predicted to be similar to the target structure.

Multiple Isomorphous Replacement[14] is a common *ab initio* phasing method that involves introducing a heavy atom such as mercury, lead, uranium, platinum or gold into a protein crystal. This can be accomplished by soaking preformed crystals in a solution containing heavy metal ions. The important prerequisite for this method is the crystal structure must not be affected by the introduction of the heavy

atoms. That is, it must be isomorphous to the original crystal. Comparison of diffraction data from the two isomorphous crystals allows the identification of the heavy atom. The structure factors of the native crystal, heavy atom derivative crystal, and heavy atom are interrelated, and because the intensity and phase of the heavy atom can be calculated based on its known position, this information can be used to determine two possible phase angles for the reflection. Repeating the process with a second isomorphic heavy atom crystal allows estimation of the phase.

The techniques of Multiple Anomalous Dispersion (MAD) and Single Anomalous Dispersion (SAD) phasing allow for the determination of phase when isomorphous heavy atom derivative crystals are not available. These techniques take advantage of X-ray absorption and emission by a subset of heavy atoms such as selenium, iron, or sulfur leading to what is called anomalous scattering. These atoms can be naturally present as cofactors in some proteins or can be introduced experimentally. A common method is by expressing the proteins in a selenomethionine rich medium which replaces the protein's methionine residues with selenium-containing selenomethionine. Anomalous scattering causes slight mismatches between the measured intensities of Friedel pairs. With very accurate intensity measurements these Friedel pair differences can be used to determine initial phase estimates.

An important consideration for both *ab initio* phasing methods is extremely accurate intensity measurements from the original diffraction patterns and proper scaling of partially recorded reflections.

The second method for phasing requires a starting idea of the structure of the protein which is used calculate an initial estimate of phase values through the technique of molecular replacement (MR).[15,16] The main source of starting models is homologous proteins. If the crystal of the homologous protein is isomorphous to the target crystal, theoretically possible when a ligand is soaked into an existing crystal or with two crystals of a protein with a minor mutation, the phases of the first model can be used as a starting estimate for the second. Usually though, the rotation and translation of the asymmetric unit in the unit cell must be determined through computational searching. Once these parameters are determined, the model can be used to obtain initial phase estimates.

Phasing by molecular replacement requires a search model similar to at least part of the unknown protein. For proteins of unknown structure, this may not be possible. Recent advances in the field of computational protein structure prediction have opened up the possibility of generating a MR search model using only the primary sequence of the target protein.[17] One study found that for diffraction data from a test set of 30 relatively short soluble proteins MR using a purely computational model was successful in 9 instances.[18]

Once initial phase estimates are obtained the model is iteratively refined. Diverse methods for refinement in both real space and reciprocal space are covered in a large body of literature. [19-21]

2.2 X-ray crystallography in practice

As a method that has enjoyed a great deal of attention and innovation over the last 50 years, X-ray crystallography is now highly automated, although the method is still quite labor intensive and has bottlenecks that keep structure determination from being completely routine.

2.2.1 Growing crystals

The most common method for growing crystals is the vapor diffusion method. Briefly, a protein solution is mixed with a precipitant solution and the drop sealed in a closed container separated from a larger volume of the precipitant solution, known as the reservoir, by air. Because the protein drop has a lower concentration of the precipitant solution the drops begin to equilibrate as water evaporates from the protein drop and partitions to the reservoir. This results in an increase of both protein and precipitant concentrations in the protein drop until equilibrium is reached. This slow increase in concentration can drive crystallization of the protein and/or buffer components.

A large variety of environmental variables influence the formation of protein crystals. The composition of the precipitant can have dramatic effects and must be determined experientially for each protein. The concentration difference between the two drops controls both the rate of diffusion and the final protein concentration in the drop. The rate of diffusion also depends on temperature, which in turn also influences the stability of the protein and precipitant. The starting concentration of the protein, along with the pH of the protein drop also affects the protein,

precipitant components, and their interactions. These variables, and innumerable others, all must be optimized to strike the balance that allows large well ordered crystals to form. The protein must be unstable enough in solution to allow the precipitation of a small number of molecules to nucleate crystals, yet soluble enough that crystal growth proceeds slowly, so as to yield large and well ordered crystals. A protein that is too insoluble will lead to too much nucleation, yielding many small crystals, or simply 'crash' into unordered precipitate. A protein that is too soluble will simply remain in solution. An additional confounding variable is that the conditions in the protein drop are not static. As protein molecules are removed from the solution, by joining a crystal or precipitating, the effective protein concentration in the drop decreases. Because the drop is not being physically disturbed, crystal growth, precipitation, or phase separation, can create multiple microenvironments, any one of which may be conducive to crystal growth.

Crystallization opens up a nearly infinite combinatorial space for screening possible conditions; the optimal conditions for crystal growth must be determined on a protein to protein basis by experimentation. This is can be accomplished through highly automated high-throughput screening using advanced liquid handling systems or many person-hours of labor. Tens (or hundreds) of thousands of drops are prepared and observed over time looking for signs of crystallization. If initial 'hits' are obtained, new screens are performed around the hit conditions until crystals suitable for diffraction are obtained. If no hits are obtained crystallographers generally begin to modify the protein itself through mutation, chemical modification, or choosing homologous proteins from other organisms.

These modification methods are sometimes effective in the process of optimizing initial hits.

2.2.2 Growing crystals in the lipidic cubic phase

Integral membrane proteins have proven especially resistant to crystallization. One technique that has shown promise for some of these difficult to crystallize targets has been crystallization in the lipid cubic phase (LCP). Crystallization *in meso* has proved successful for a small number of targets that were previously intractable using standard methods.[22-25]

In LCP crystallography protein is crystallized in the presence of hydrated lipids which are in a quasisolid cubic phase.[26] LCP crystallization is especially useful for integral membrane proteins because the cubic phase lipids can provide an approximation of the lipid bilayer where the proteins normally reside, increasing their stability and leading to the formation of small but highly ordered crystals.[27]

Although lipidic cubic phase crystallization has had some success, the technique has still not shown wide utility because of a variety of difficult logistical issues. Purpose built robots exist for automated setting of LCP screening trays[28] but the screening process is more difficult because of the added variable of the lipids, which show variation in their behavior due to temperature and interactions with many of the additives in common crystallization screening mixtures.[29] Even when LCP crystallization is successful it yields tiny crystals which pose additional challenges with detection, mounting, and diffraction due to their small size.[27]

2.2.2 Mounting and shooting

Once a suitable crystal is obtained the crystal is carefully removed from the drop using a thin wire loop. It is then flash frozen, usually in liquid nitrogen. The freezing process can sometimes be very disruptive to the crystal. This may require first exchanging the crystal drop buffer for a different buffer containing cryo-protectant compounds. The selection of proper freezing methods and cryo-protectants may require a great deal of optimization. This along with the skill required to successfully loop out very tiny crystals often results in the destruction of many valuable and hard to obtain crystals before a sample is ready for diffraction.

After a properly mounted crystal is looped and cryoprotected it can be exposed to X-ray radiation from an X-ray source such as a purpose-built X-ray generator or a synchrotron beamline. Only then will the dogged crystallographer know if the crystal is of sufficient quality to obtain high resolution diffraction data.

If, against all odds, high resolution diffraction is obtained, data collection begins. The X-ray beam is diffracted through the crystal as it is rotated over a small angle during the exposure. Multiple exposures are taken each starting with the crystals tilted at a different angle. The orientations of the initial diffraction patterns and the symmetry of the crystal are determined and this information used to develop a data collection strategy that covers enough data in reciprocal space for accurate structure determination. Often, the crystal is damaged or destroyed by the

X-ray beam before this is accomplished requiring the merging of data from several crystals before a suitably complete data set is obtained.

2.2.3 Data processing

A variety of software, both commercial and freeware, is available for the collection and processing of X-ray crystallography data. HKL-2000/3000[30] is a complete commercial package that interacts with a variety of X-ray detectors. The program contains utilities for designing data collection schemes, integration, merging, and scaling of the collected data, phasing by MAD, SAD, or MR, and structure determination.

The most commonly used free open source software for X-ray crystallography is the CCP4 suite.[21] The suite contains a large collection of programs that accomplish all the necessary tasks from intensity measurement to structure determination and refinement. The process begins with MOSFLM[31] which provides a graphical user interface for pattern indexing and intensity measurement. The utilities POINTLESS and SCALA determine the point and space group symmetries of the crystal and merge and scale symmetry related reflections respectively.[32] Experimental phasing can be performed using the program CRANK[33], or MOLREP[34] can be used for molecular replacement. Finally another suite, PHENIX[19], provides a front end to a large independent assortment of programs for structure refinement and model building.

Overall the CCP4 suite represents a large body of work which is constantly improved and updated by a large and active user group.[35]

2.3 Electron crystallography of 2-dimensional crystals

The coherent electron beam generated by a transmission electron microscope (TEM) also provides means for examining protein structure. The wavelength of electrons is orders of magnitude smaller than that of X-rays and the unit cell sizes of protein crystals are well within the resolutions achievable with TEM. This fact was recognized early, the first experiments in imaging protein crystals with TEM were concurrent with the original X-ray structure determinations.[36] The authors of these early studies highlighted one of the limitations of the electron beam: its very limited penetration into the crystal. This limitation was circumvented by either examining a replica of the crystals created by coating them with a layer of platinum, iridium, and carbon, a technique that only gives information about the surface, or by examining very thin crystals in transmission. These early experiments allowed for imaging of protein crystals and accurate calculation of unit cell dimensions[36], laying the groundwork for atomic resolution structure determination.

Near-atomic resolution protein structure determination with TEM was achieved in 1975 when Henderson & Unwin determined a 3-dimensional model of the purple membrane bacteriorhodopsin to 7 Å using electron diffraction.[37] This was also the first published structure of an integral membrane protein. The limited

penetration of the electron beam was circumvented by using crystals so thin they are considered 2-dimensional. The method used was superficially similar to X-ray crystallography. Diffraction patterns were recorded with crystals tilted at various angles and the 2-dimensional projections resulting from inverse Fourier transforms of each pattern were combined to determine the protein's 3-dimensional structure.

The diffraction patterns captured by Henderson & Unwin recorded the position and intensity of each Bragg peak but, like in X-ray diffraction, the phase information for each spot was lost. Here the advantages of diffraction in a TEM became apparent. The authors not only recorded diffraction patterns, they also imaged the crystals. The Fourier transform of images contains both phase and intensity information, so they were able to combine phase information from images with the more accurate intensity measurements from the diffraction patterns. [38] The inverse Fourier transforms of the combined intensities and phases were used to generate the 3-dimensional map of bacteriorhodopsin. Since this initial demonstration of the power of 2-dimensional electron crystallography, approximately 60 protein structures have been solved to higher than 10 Å resolution with the method [39] but only a very small number [40-42] have broken the 3 Å barrier to reach resolutions comparable to X-ray crystallography.

The few available high-resolution structures determined by electron crystallography of 2-dimensional crystals have given dramatic insights into the structure of integral membrane proteins. These structures have been solved using 2-dimensional crystals composed of protein and lipids. Although the packing of protein and lipid is ordered and crystalline, the lipids still form a biliayer providing

an approximation of the biological membrane. Integral membrane proteins have been shown to depend on lipid interactions for both oligomeric state [43] and function. [44] The highest resolution structures determined from 2-dimensional crystals allow the visualization of lipids and protein-lipid interactions.[40] Because 2-dimensional protein-lipid crystals can maintain proteins in a functional state even lower resolution structures and 2-dimensional projections derived from this method can provide insights. Some 2-dimensional crystals are sufficiently robust they can be transferred to different conditions and the resulting protein conformational changes can be observed. [45]

2.3.1 Formation of 2-dimensional crystals

Electron crystallography of 2-dimensional crystals begins with the formation of 2-dimensional crystals from purified protein and lipids. Most 2-dimensional crystals are produced using integral membrane proteins. Because of their hydrophobic nature these proteins, like the lipids that make up the membrane bilayer, are not highly soluble in water. This generally requires some sort of surfactant such as a chemical detergent to keep the protein in solution. Two-dimensional crystallization takes advantage of this property. Removal of detergent spurs the formation of lipid bilayers and the subsequent reconstitution of the protein in the newly formed bilayer. This reconstitution usually leads to a disordered membrane but, through careful control of a range variables, proteins and lipids can be coaxed into forming ordered arrays in 2-dimensions. The removal of

detergent can be achieved through a variety of methods such as dialysis, [46] simple dilution,[47] absorption to a substrate such as bio-beads,[48] or chelation by a reagent such as β-methyl-cyclodextran.[49] Detergent removal by dialysis has been the method of choice for most of the 2-dimensional crystals used for structure determinations. [39]

A number of other factors have been reported to affect the process of two dimensional crystallization. Two large considerations are the lipids and detergent used.

Two dimensional crystals suitable for electron diffraction have been generated using both native [50] and artificial lipids [46] as well as non-native mixed lipids. [51] Both the head groups and tails of membrane lipids have been shown to interact with integral membrane proteins, [40] so the proper lipid composition may be critical for 2-dimensional crystal formation. This must be empirically determined as there are no hard fast rules for the determination of which proteins and lipids to use together. Native lipids best represent the environment in which the protein normally resides, but may be sub-optimal for the production of well-ordered crystals or for logistical reasons because of difficulties in their extraction and purification. When artificial lipids are used the best results have been obtained using those that best approximate the of the natural lipid bilayer, which has a hydrophobic core thickness of ~35Å. Lipids with saturated or monounsaturated acyl chains are often preferred as they approximate the lipid bilayer and are less susceptible to oxidation.

The choice of detergent affects several variables in the crystallization process from the initial stability of the protein in solution to the speed at which lipid bilayers form. The detergent must be able to extract the protein from the native membrane but must be generally non-denaturing and able to maintain the protein stability in solution. The critical micelle concentration (CMC) of the detergent is an important factor in detergent choice. When detergent at sub-CMC concentrations the detergent molecules are distributed between micelles and free detergent molecules, above the CMC any additional detergent added increases the size of micelles. The goal of dilution and detergent removal processes is to drop the overall detergent concentration well below the CMC leading the lipids in the sample to form a bilayer. It may be very difficult to lower the detergent concentration below this threshold for extremely low CMC detergents, requiring extended dialysis or dilution beyond useful protein concentrations.

Additional considerations are temperature and composition of the crystallization buffer. Temperature affects both dialysis rate and the state of the lipids. Generally 2-dimensional crystallization is preformed at a temperature above the transition temperature of the lipids used in order to keep the lipids in a fluid state allowing the protein to more freely laterally diffuse within the bilayer. Individual crystallization mixes may be sensitive to the ionic strength of the dialysis buffer. Two dimensional crystals have been reported in solutions ranging from no salts to 600 mM NaCl [52] or 1 M KCl [53] The addition of divalent cations such as Mg^{2+}, Ca^{2+}, and Zn^{2+} has proved successful in some cases.

The large number of variables lends a high degree of combinatorial complexity to 2-dimensional crystallization. This is a major bottleneck in the process and becomes a major limitation of the utility of the technique. This is further compounded by the low throughput nature of screening crystal hits. Recently some advances have been made toward the automation of 2-dimensional crystal screening. A robot for high throughput screening described by Iacovache *et al* uses light scattering to assess the formation of larger lipid structures[54] and detergent concentration can be also be assessed by droplet contact angle measurements. [55] These techniques measure only the removal of detergent and formation of larger lipid structures in the crystallization mixture. They do not give information as to the protein's incorporation or degree of order in the newly formed lipid bilayer. This must be observed directly using an electron microscope. Despite some progress in automation [56] this process is largely manual and very time consuming.

Two dimensional crystals are generally screened by negative stain EM. Biological molecules scatter electrons very poorly and therefore generate images with very low contrast when unstained. Staining with a strongly scattering molecule such as uranyl formate, uranyl acetate, or sodium phosphotungstinate creates a higher contrast negative image where the biological molecule displaces the stain. This method offers the additional advantages of grids that can be prepared relatively rapidly and can be observed and imaged multiple times with little radiation damage to the sample.

EM grids are coated with a continuous layer of carbon or are purchased pre-coated. The grids are glow discharged to remove dust and impart a charge to the grid surface. The crystal mixture is applied to the gird, allowed to sit, and then blotted off. The grid is then washed with water or buffer followed by a negative stain solution. Negative stain procedures must be optimized for each individual sample. The optimum time the sample is left on the grid, amount of washing, polarity of negative discharge, and type of stain must be determined empirically for each sample.

Negatively stained EM grids are generally examined under low magnification to locate large lipid structures which are examined under higher magnification, usually 10,000x to 30,000x. In some 2d crystals the crystalline matrix is obvious and can be identified by eye. Less obvious crystals can be confirmed by the presence of spots on the Fourier transform of the image. An experienced user can prepare and survey a grid in approximately 15 minutes, an extremely low-throughput rate considering the combinatorial space that must be covered to determine appropriate crystallization conditions.

2.3.2 Preparation of cryo-grids

Once a sample with suitable 2-dimensional crystals has been identified a cryogrid must be prepared. As with most techniques in electron microscopy the most effective way to accomplish this depends on the sample and lab conditions and must be determined experimentally. A small amount of sample is applied to a holey

carbon coated EM grid, the sample is blotted to removed excess solvent, and it is frozen by either slow-freezing or fast freezing methods. The sample must be maintained In a frozen-hydrated state meaning it must be free of ice crystals, either in vitreous (non-crystalline) ice produced by the fast freezing method, or embedded in sugar (usually 3-20% glucose or trehalose)[57], which can be slow-frozen directly in the microscope. Fast-frozen grids are blotted and immediately plunged into liquid ethane or liquid nitrogen which lead to the formation of vitreous ice.[58] In either method blotting technique is crucial for determining the thickness of the ice on the grid and the quality of crystal embedding. Well embedded crystals will produce sharp clear spots, while diffraction patters from poorly embedded crystals will appear smeared or show no spots at all.[57] One of the challenges of producing cryogrids by hand is reproducibility.[58] Instruments such as the Vitrobot[59] improve the reproducibility of blotting by controlling both the blotting time and pressure, as well as the temperature and relative humidity of the chamber. Once frozen the grid must be stored under dry liquid nitrogen at all times while transferring it to the cryoholder and inserting it into the microscope.

2.3.3 Data collection

After a cryogrid with well-embedded high quality 2-dimensional crystals is obtained, data collection can begin. The grid is first surveyed in overfocused diffraction mode, which generates a high contrast low magnification image that allows for identification of the crystals while minimizing dosage and subsequent

radiation damage.[60] Once a suitable crystal is located, the microscope is focused in high magnification imaging mode and prepared for diffraction. The electron beam, beam stop and recording device must first be carefully aligned both to protect the recording device from overexposure and collect data to the maximum possible resolution. A selected area (SA) aperture is inserted which allows for illumination of only a small portion of the crystal, and diffraction data is collected by exposing the crystal to a dose of around 5-20 electrons per $Å^2$. [60] Alternatively, images may be collected instead of diffraction patterns by removal of the SA aperture. Either way this exposure destroys the crystal, so another crystal is then identified. The grid is then tilted, the microscope realigned and refocused, and a tilted diffraction pattern or image collected. This process is repeated until a dataset that covers as much angular space as possible is obtained. With sufficiently large crystals and a proper cryoholder this can be up to 70°.[57]

2.3.4 Data Processing

Several reviews give in depth coverage to data processing and structure determination from electron diffraction data[61-64] In a broad sense data processing for diffraction patterns generated by electron diffraction of 2-dimensional crystals is similar to that for X-ray crystallography, using the MRC suite of programs. The program XDP[43] provides a more user-friendly front end specially designed for using the MRC programs to process 2-dimensional crystal electron diffraction data. Alternatively the program 2dx[65,66] also serves as a front end to the MRC suite

allowing structure determination directly from images, with the added advantage that the Fourier transform of image data already contains phase information. Phasing of electron diffraction data can be accomplished through multiple isomorphous replacement or multiple anomalous dispersion methods [67] although initial phases for refinement are generally obtained through molecular replacement [68] or from images. Additional computational methods even enable phase extension, allowing the phase data from low resolution images to be used to calculate initial phases for high resolution diffraction data.[69]

3. A new method for protein structure determination: Microcrystal electron diffraction (microED).

This work presents a new method for protein structure determination that is essentially a hybrid of 2-dimensional electron crystallography and 3-dimensional X-ray crystallography methodologies described above. Using a TEM it is possible to collect electron diffraction from 3-dimensional crystals that are several orders of magnitude smaller than those currently used for X-ray crystallography. Extremely small crystals, on the order of 1 – 10 μM^3, are applied to an electron microscopy grid and frozen in vitreous ice. These crystals are exposed to the electron beam and diffraction patterns are collected. The crystals are then rotated on the microscope compustage and additional tilted diffraction patterns collected until a dataset covering the full angular space is collected. The strength of the beams is kept in what is considered an 'ultra low dose' regimen allowing the collection multiple patterns from each crystal and theoretically allowing a complete dataset to be collected from a single crystal before the dosage exceeds the critical threshold where the crystal degradation by radiation damage reaches a point where it negatively affects data quality. The diffraction patterns are then indexed and the intensities of the Bragg peaks measured. Once integrated intensities are obtained the structure can be phased using established crystallographic methods, and a map of the 3-dimensional structure can be calculated.

3.1 Rationale: why a new method is useful

Many proteins of significant biological interest are not amenable to crystallization. Despite researchers' best efforts these proteins cannot be induced to form large well ordered crystals usable for X-ray diffraction. The process of crystallization screening is time consuming and expensive, and even after promising conditions are found optimization of the initial condition may prove an even bigger challenge.

One example is the G-protein-coupled receptors (GPCRs). GPCRS are a diverse group of large integral membrane proteins with 7 transmembrane helices found in a wide variety of eukaryotic cell types. They play a large range of critical signal transduction roles in disparate processes such as neurotransmission, memory, and behavior[70] regulating prenatal[71] and postnatal development[72], nociception[73,74], vasoconstriction[75], and phototransduction in vision[76]. GPCRs are the targets of approximately 40% of drugs[77] and are understandably a major focus of drug discovery efforts in both academia and industry [78].

This family of proteins is notoriously difficult to crystalize. Despite intense interest the GPCRs eluded crystallization until 2000 when the first high-resolution structure of rhodopsin was published. [79] Brian Kolbilka's Nobel prize-winning work on the β_2 adrenergic GPCR (b_2AR) illustrates the lengths required to get very small crystals of these proteins and the difficulty of structure determination using these crystals. Despite an intense optimization regimen that included mutations, and Fab co-crystallization in the lipidic cubic phase, the largest crystals the group

was able to grow were approximately 300 x 30 x 10 μm. [80] These small crystals were extremely sensitive to radiation damage meaning only a 5 – 10 shots with the X-ray beam could be attempted before the crystals were destroyed. The final structure of b_2AR required merging data from over 20 crystals. The authors do not report how many crystals were required to get this final dataset, but this must have been a massive undertaking. The Kolbilka group's next GPCR structure, b_2AR in complex with the G protein Gs, required the merging of data collected from a similarly large number of small crystals. [24]

The examples above represent some of the smallest crystals that have yielded structures using traditional X-ray crystallography methods. Even crystals of this size are unattainable for many protein targets. During crystallization screening X-ray crystallographers generally get their first 'hits' as microcrystals. The conditions that yield these very tiny crystals, the largest of which are only barely visible under a light microscope, can sometimes be optimized to grow larger crystals. Often even after immense amounts of work to optimize the final crystals are still barely large enough for X-ray diffraction and often no amount of optimization moves structure determination efforts past the microcrystal stage.

The smallest of the GPCR crystals above have volumes of about 90,000 μm^3. X-ray microdiffraction experiments have gleaned useful data from crystals as small as 500 μm^3 but also suffer from the radiation damage issues further exacerbated by the small crystal size. [81,82] This is approaching the theoretical limit for crystal size in X-ray diffraction, which is estimated to be ~ 350 μm^3. [83] MicroED can be

performed on crystals with volumes down to 1μm^3, two orders of magnitude smaller.

3.1.1 Femtosecond Nanocrystallography attempts to fill this gap

Crystals this small may never be suitable for traditional X-ray diffraction because of their high susceptibility to radiation damage [84]. Any new method designed to utilize them must somehow get around the problem of radiation damage. One new experimental method uses femtosecond pulses of an X-ray Free Electron Laser (XFEL) to circumvent this problem. Femtosecond X-ray nanocrystallography (FXN) allows for collection of diffraction patterns from crystals down to 200 nm^3 and avoids radiation damage effects by "diffraction before destruction".[85]

In FXN a microcrystal slurry is passed through a small nozzle where individual microcrystals are irradiated by femtosecond pulses of the XFEL beam. These pulses are so fast diffraction patterns can be recorded on a time scale faster than that on which radiation damage occurs [86] A beam pulsing at up to 120 hz allows for collection of thousands of diffraction patterns per minute, each derived from different crystal at a random orientation. FXN is performed at non-cryogenic temperatures so the crystals are fully hydrated, an additional advantage beyond the ability to utilize nanocrystals.

The structure of lysozyme has been determined to 1.9 Å using FXN, an effort that required over 1.9 million individual diffraction patterns [87] and the

development of new software to reconstruct the Bragg peak intensities for the randomly oriented 'snapshot' diffraction patterns.[88]

Although very promising, FXN is still a technique in its infancy with its utility limited by several logistical issues. First and foremost the FXN requires access to an XFEL, which is prohibitive for most researchers. The current implementation of the technology also requires large volumes of concentrated protein solution (1mg/ml for determination of photosystem I [85]) an amount that would give most crystallographers pause, especially for difficult to purify and unstable proteins.

3.1.2 MicroED provides an accessible technique

Microcrystal electron diffraction attempts to fill a niche currently left by other methods. X-ray crystallography is high-throughput, but requires large crystals. Electron diffraction of 2-dimensional crystals can utilize small crystals, but is limited to extremely thin 2-d crystals, which are very difficult to produce and limited to a small subset of proteins. FXN can use very tiny crystals but is not yet a mature ,accessible technique, requiring large amounts of protein and equipment unavailable to most researchers. MicroED allows structure determination from small microcrystals produced using high-throughput methods. Diffraction data can be collected using electron microscopes that are standard in many academic institutions and data processing can be performed on any laptop computer. This technique may allow crystallographers to circumvent the time consuming work optimizing initial hits that generate microcrystals during screening and could prove

especially useful for previously intractable protein targets that have so far refused to form large crystals.

3. 2 Methods for microcrystal electron diffraction

3.2.1 Sample Preparation and Data collection

The mother liquor containing microcrystals is applied to an electron microscopy grid, blotted, and frozen in vitrified ice by plunging into liquid ethane. The grid is loaded into the TEM and surveyed in overfocused diffraction mode. In this mode the crystals generally appear as dark electron dense masses with well-defined edges. When a suitable crystal is identified the beam is focused and the stage is tilted to prepare for data collection. The crystal is exposed with an ultra-low dose, generally ~0.01 $e^-/Å^2$ per second, long enough to record a diffraction pattern. The stage is then tilted back toward 0° by a small increment and another diffraction pattern collected. This process is repeated until the crystal has received a critical dose of radiation (determination of this threshold is discussed in section 4.1). When a crystal has reached the critical dose it is discarded and another crystal located.

3.2.2 Data processing

The original intent of the project was to collect diffraction patterns by microED and process the data using the widely accepted crystallography software

packages MOSFLM [31], CCP4 [21], and/or PHENIX [19]. Although microED diffraction data are intrinsically no different than those collected by X-ray diffraction, some logistical hurdles made unit cell determination and indexing with MOSFLM intractable. Two major characteristics differentiate microED data from standard X-ray diffraction data: the short wavelength of electrons compared to X-rays and the lack of crystal precession during data collection. MOSFLM allows for the adjustment of the wavelength and angular range covered in procession, but the small electron wavelength and 0° precession angle, coupled with inaccuracy in the microscope's compustage angle measurements, caused the program to generate errors. Therefore, a suite of software was written to process the microED data for proof of concept studies.

The microED suite contains 8 programs that work together to accomplish essential data processing tasks: unit cell determination, spot prediction and indexing, measuring spot intensities, and merging symmetry related spots. An additional program with a more user-friendly Graphical User Interface (GUI) was later developed to simplify some of the more labor-intensive tasks.

3.2.3 Software philosophy

The microED suite is not intended as a long term solution for structure determination by microED. All of the tools necessary for indexing and intensity measurement should be available through modification of existing programs that are well established, tested, and very robust. Although the current indexing and

integration programs tested were unable to handle a microED dataset, the eventual goal of this work will be integration of microED techniques into existing crystallography software.

The microED suite programs are designed to be cross-platform and any modules, libraries, and/or outside programs therefore needed to be easily available. All of the software in microED suite is implemented in Python 2.7 using the standard modules numpy [89] and Python Image Library (PIL). [90] The GUI requires Tkinter, [91] also a common Python module. Two outside programs are required: Gnuplot [92] and imagemagick [93], both are standard on most UNIX installations. Any additional image processing necessary can be performed with FIJI [94], a standard program for scientific image manipulation.

3.2.4 Unit cell determination and refinement

Knowledge of the unit cell dimensions is the critical first step in crystallography. Although the lysozyme proof of concept was started with *a priori* knowledge of unit cell dimensions and angles for the crystals, crystallography of a novel protein with an unknown structure will not have this advantage. In order to make the technique widely applicable a method for *de novo* unit cell determination was required.

MOSFLM, currently the standard program for crystallographic indexing, uses the Rossman Fourier analysis method[95,96] which is a powerful and robust, although

computationally intensive, method for unit cell determination. For the microED suite a simplified method was employed.

Unit cell determination begins with the user picking 100 – 1000 spots from several diffraction patterns of various tilts. A vector $\boldsymbol{v}$ that defines the spot relative to Cartesian coordinates (0,0,0) is calculated for each spot:

$$\boldsymbol{v} = \langle x, y, z \rangle$$

$$x = x_i - x_c$$

$$y = cos\theta(y_i - y_c)$$

$$z = sin\theta(y_i - y_c) - a$$

where x_i, and y_i are the x and y coordinates on the diffraction pattern image, x_c and y_c are the x and y coordinates of the beam center, θ is the tilt angle, and a is a correction for the curvature of the Ewald sphere. Although Ewald sphere curvature also affects the x and y 3-D coordinates, it was determined that this difference was so small (~ 1 pixel at 2.0Å resolution) that it could be ignored. Effects of Ewald sphere curvature on the calculation of the z component of **v** is significant (~12 pixels at 2.0Å), so a is calculated as:

$$a = -b\left(1 - \frac{b}{\sqrt{(x^2 + y^2 + b^2)}}\right)$$

$$b = \frac{1}{\lambda}c$$

where λ is the electron wavelength, x and y are as calculated above, and c is the factor to convert Å to pixels.

After $\mathbf{v}$ is calculated for each spot the distance between spots d is calculated for every pair of spots. For spots defined by $\mathbf{v_a}$ and $\mathbf{v_b}$

$$d = \|\boldsymbol{v_a} - \boldsymbol{v_b}\|$$

The unit cell lengths can be estimated from the distribution of d for all spots. With two spots of adjacent Miller indices, d is equal to the unit cell dimension. This distance cannot be smaller than the smallest unit cell dimension, so the smallest peaks in the distribution of d represent the unit cell dimensions. This procedure is not exact and still requires some user intuition. For example $d_{(0,0,0)(1,0,0)}$ (between Miller indices (1,0,0) and (0,0,0)) equals the a unit cell dimension, but $d_{(1,1,0)(0,0,0)}$ might be smaller than $d_{(0,0,0)(0,0,1)}$ depending on unit cell dimensions. This, along with the possibility of multiple unit cell dimensions having the same length, or lengths that are close multiples of each other, means the user cannot simply pick the three shortest values of d as the unit cell dimensions. The general formula for these cross-unit cell vectors for a unit cell a,b,c with angles α,β,γ is:

$$d_{(n,n,n)(n+m,n+p,n)} = \sqrt{(ma + bcos\gamma)^2 + (pbsin\gamma)^2}$$

where *m* and *p* are integers. This allows for the calculation of the expected peaks in the distribution of *d*, which can be compared to the observed distribution and used to verify the correct unit cell dimensions have been chosen. Visual observation of diffraction patterns that hit on or near major planes of the crystal also allows rough measurements of the unit cell dimensions and angles which can be used to verify these findings.

After rough unit cell dimensions have ben determined the actual unit cell and its orientation in 3-D space can be calculated. This is initially accomplished by using the spots chosen by the user. First the vectors $\boldsymbol{d}$ between all of the chosen spots are calculated.

$$\boldsymbol{d}_{a,b} = \boldsymbol{v}_a - \boldsymbol{v}_b$$

All of the vectors are compared to the three unit cell dimensions and those within a user specified threshold are kept. The remaining vectors are then compared to four reference vectors: ⟨1,0,0⟩, ⟨0,1,0⟩, ⟨0,0,1⟩, and ⟨1,1,0⟩. The angle between the each vector and reference vector (σ) is calculated by

$$\sigma = acos\left(\frac{\boldsymbol{d}\cdot\boldsymbol{r}}{\|\boldsymbol{d}\|\|\boldsymbol{r}\|}\right)$$

where **d** is the difference vector and **r** is the reference vector. This allows the vectors to be divided into roughly parallel groups based on the angles between the vector and the four reference vectors. Four reference vectors are used to eliminate the very slim possibility that a unit cell vector is exactly equidistant for two of the reference vectors.

The orientations of the vectors in each group are determined by calculating the cross product of the vector and the ⟨1,0,0⟩ reference vector and appropriate vectors are flipped by multiplying by -1 so all vectors in each group point in the same direction. The vectors in each group are averaged, producing a list of candidates for the unit cell vectors. Each candidate is assigned a score based on how many vectors contributed to it. By examining the angles between the candidates and their scores the correct unit cell vectors can usually be chosen.

3.2.5 Initial spot prediction and unit cell vector refinement

Once the three vectors defining the unit cell are chosen they are used to predict the spots on each image. The unit cell vectors $\langle a_x, a_y, a_z\rangle$, $\langle b_x, b_y, b_z\rangle$, and $\langle c_x, c_y, c_z\rangle$ and are used to create a unit cell matrix

$$\begin{bmatrix} a_x & b_x & c_x \\ a_y & b_y & c_y \\ a_z & b_z & c_z \end{bmatrix}$$

Two reference spots are chosen from each image. These spots are chosen because they have strong intensity and are thought to represent complete intensities where the Ewald sphere passed directly through the center of the spot. The x,y,z coordinates of the reference spots are calculated as above and their Miller indices determined by multiplication with the inverse unit cell matrix

$$\begin{bmatrix} h \\ k \\ l \end{bmatrix} = \begin{bmatrix} a_x & b_x & c_x \\ a_y & b_y & c_y \\ a_z & b_z & c_z \end{bmatrix}^{-1} \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

A 'check plane' (**q**) normal to the plane containing the two reference points is calculated as

$$\boldsymbol{q} = \langle d, e, f \rangle = \langle h_1, k_1, l_1 \rangle \times \langle h_2, k_2, l_2 \rangle$$

Every Miller index is then checked against the check plane. For any given Miller index (h,k,l) the dot product of the Miller index and **q** can be used to determine if that Miller index lies on the same plane as the two reference spots. If any given Miller index *h,k,l* satisfies:

$$hd + ke + lf = 0$$

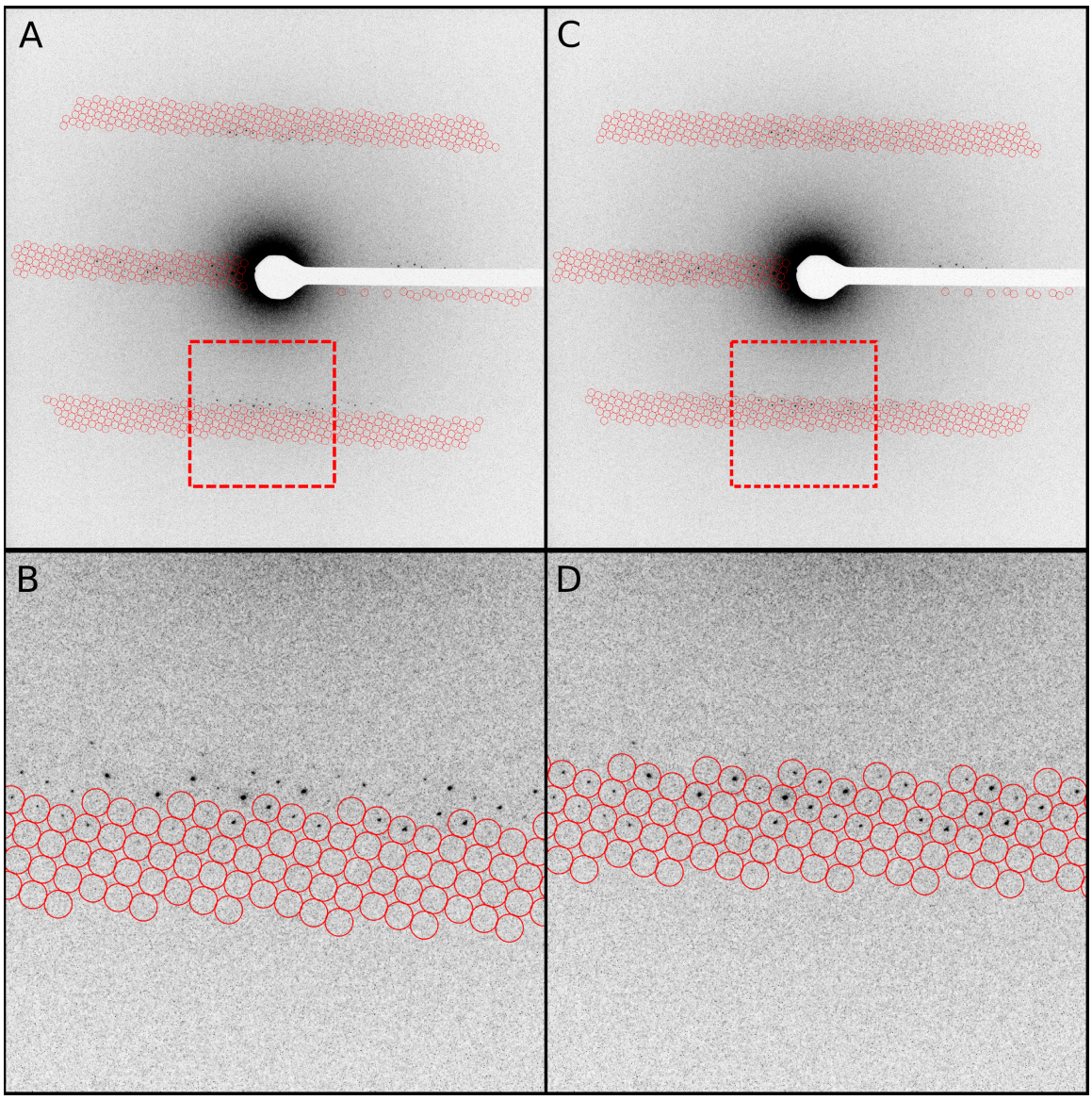
then the point lies on the check plane. The two reference points are known to exist because they are visible on the diffraction pattern so this can be used to predict the other spots that should appear on each given diffraction pattern. The quality of the reference points chosen is critically important. The spots must the be the user's best estimation of Bragg peaks that were perfectly bisected by the Ewald sphere. A good rule of thumb is to choose spots with adjacent spots on both sides and of high intensity. Figure 1 illustrates a diffraction pattern indexed with two different sets of reference points, demonstrating the effects of the reference set on the overall quality of the indexing.

The probability that the dot product of any given Miller index and the check plane is exactly zero is very small. The calculation of the check plane is based on the locations of the reference spots which introduces error as the measurement of these x,y coordinates will never be exactly perfect. To cope with this noise in the spot prediction the spots are instead held to a thresholded standard. A spot is considered to exist on an image if it satisfies the following condition:

$$|hd + ke + lf| < L$$

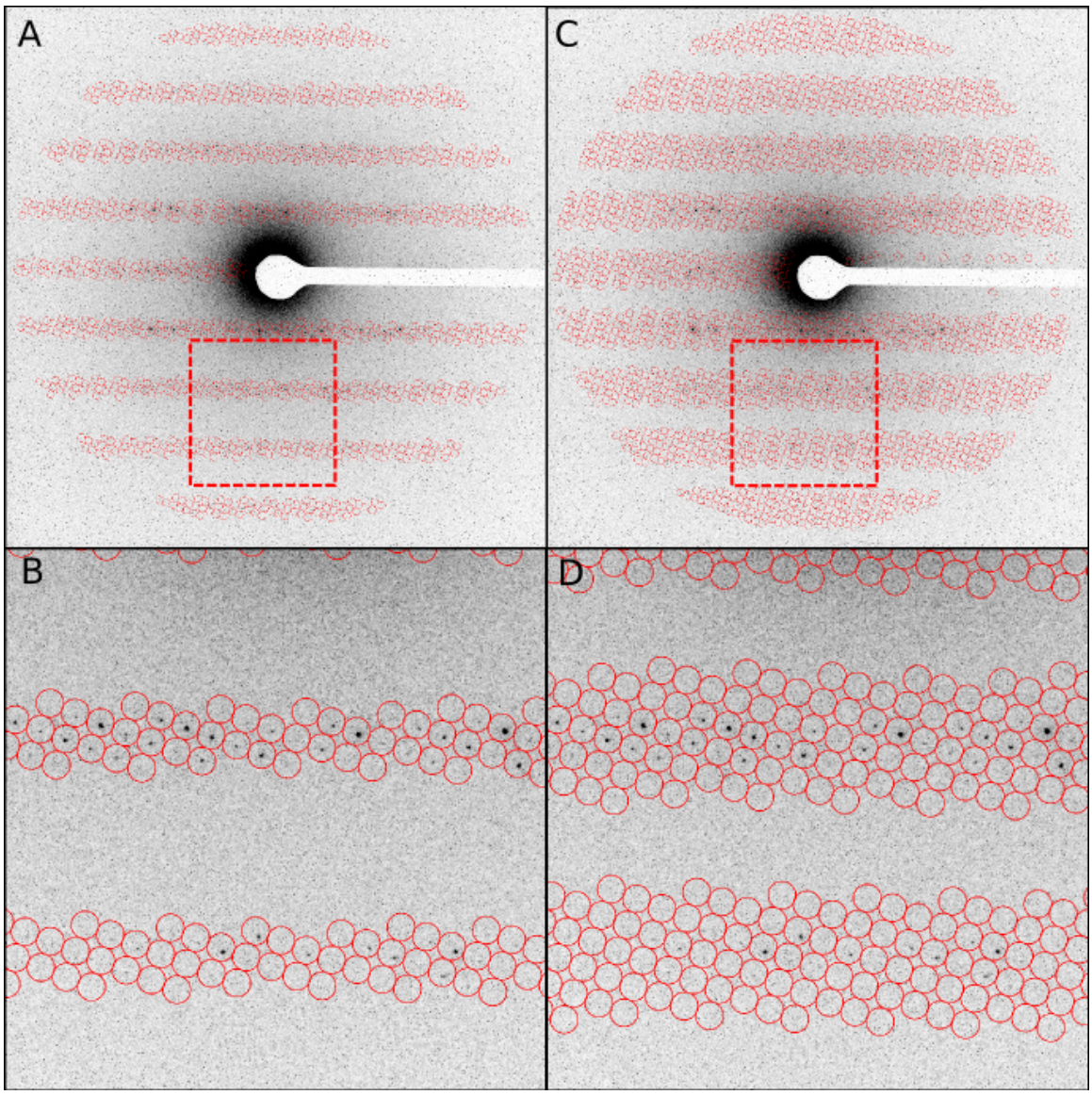
Where L is a 'Laue zone threshold' so named because its functional effect is to determine the widths of Laue zones in the spot predictions. Figure 2 shows the effects of varying L values on spot predictions. Increasing this threshold allows the

Figure 1.



Quality of reference points affects accuracy of spot prediction. **(A)** Lysozyme diffraction pattern indexed with poor quality reference points. **(B)** Zoomed in view of the region of panel **(A)** bounded by the dashed line. **(C)** The same diffraction pattern indexed with higher quality reference points. **(D)** Zoomed in view of the region bounded by the dashed line in **(C).**

Figure 2.



Effects of changing laue zone threshold on spot prediction. Predicted spots with 15% **(A** and **B)** and 30% **(C** and **D)** laue zone thresholds drawn on a lysozyme diffraction pattern.

recording of more Bragg peaks, but also increases the number of partial intensities recorded and 'false positives', indexing where no spot actually exists.

After a list of spots is created for each image their x,y,z coordinates calculated by

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} a_x & b_x & c_x \\ a_y & b_y & c_y \\ a_z & b_z & c_z \end{bmatrix} \begin{bmatrix} h \\ k \\ l \end{bmatrix}$$

The x,y,z coordinates are then used to calculate the coordinates of the spot on the 2 dimensional diffraction pattern (x', y'):

$$x' = x$$

$$y' = \frac{y}{cos\theta}$$

These coordinates are then used to draw circles around the predicted spots on the diffraction patterns. The user can inspect these images before continuing.

The initial spot predictions are dependent on the accuracy of the unit cell vectors, which were determined from a limited set of points picked by the user. A second iteration of the vector finding process allows the refinement of the vectors for more accurate spot prediction.

The predicted spots are first refined by mass centering. A square box is drawn around each spot and the pixel values put in a matrix. Each row and column of the matrix is summed and the maximum values of the rows and columns used to determine the actual center of mass for the spot. The box is moved to this center and the mass centering process repeated. If the second round of mass centering produces a large movement (more than one or two pixels in any given direction) the spot is discarded. This is to prevent the spot prediction from 'walking' between Miller indices.

After mass centering the intensity of the spot is compared to the background intensity. A square and a circle where the circle diameter is equal to the square edge length are drawn centered on the spot. The background intensity is defined as the mean pixel intensity of the area bounded by the square but outside the circle. The mean intensity of the area inside the circle is compared to the background intensity. Any spot with a low spot to background ratio is discarded. Because only intense spots which are cleanly bisected by the Ewald sphere are desired for unit cell determination this threshold is set high, usually around 10%.

The list of refined spots is then used to recalculate the unit cell vectors. Because this list contains more spots and their locations are more accurate the recalculated vectors produce better spot prediction and indexing. This process can be repeated iteratively until the unit cell vectors are stable and accurate.

3.2.6 Spot indexing, intensity measurement, and merging

Once satisfactory unit cell vectors are obtained the diffraction pattern image are indexed for a final time. The last set of spot indices is not mass centered. At this point the indexing should be accurate enough to capture all of the spots and mass centering raises the risk of a spot 'walking' to an adjacent Miller index which would lead to the intensity being attributed to the wrong reflection.

When the final indexing is complete the intensity of each spot is measured. A circle within a square is drawn and the mean background is calculated for each spot as above. The mean background is subtracted from each pixel within the circle and sum of the background subtracted pixel values recorded for that Miller index. The same mean spot intensity to background intensity comparison is then made as before, but a much lower threshold, usually ~0.5%, is used to capture weak spots.

After all of the images have been indexed and the intensities extracted, intensity measurements from symmetry-related Miller indices are merged. The symmetry relations of the different Miller indices are determined by the specific space group of the crystal. This proof of concept work took advantage of the *a priori* knowledge of the crystal space group to determine which spots to merge. Without this information the space group must be determined by examining the unit cell dimensions and angles, systematic absences, and trial and error. The merging program was written specifically for P422 symmetry but could be easily modified to handle other symmetries.

The lack of precession during data collection leads to issues with partially recorded reflections which must be addressed. In standard X-ray crystallography, the crystal is rotated during data collection allowing the Ewald sphere to move through the spot in three dimensions increasing the probability that any given reflection represents the full intensity. Comparison of multiple intensity measurements from the same Miller index allows for the calculation of r_{merge}, which is an indication of data quality.

$$r_{merge} = \frac{\sum|i - \bar{i}|}{\sum|i|}$$

where i is the observed intensity and $\bar{i}$ is the mean intensity for that reflection. Because the crystal is stationary during data collection in microED the probability of collecting partial reflections is much higher, leading to inaccurate intensity measurements if the partial reflections are not somehow scaled or excluded. To cope with this issue a strict cutoff was imposed. For any given reflection the largest recorded intensity was assumed to represent the complete reflection, any measurements for that Miller index with smaller intensities were discarded. This is an admittedly crude method for the merging of multiple intensity measurements, the ramifications of which will be discussed in more detail in section 4.2, along with proposed methods for improvement.

The final output of the merging program is a text file containing the Miller index, intensity, structure factor, sigF, and sigG for each recorded reflection.

Because each intensity measurement ultimately originated from a single observation, SigI and SigF values cannot be calculated and were replaced with the square root of the intensity and square root of the structure factor respectively. The output of the merging program can then be fed into the program COMBAT from the CCP4 suite [21] to generate a mtz file which can be used to solve the protein structure using standard molecular replacement phasing methods.

3.3 Determination of the structure of lysozyme to 3.0Å using microED - materials and methods

As a proof of concept for this new method the structure of hen egg white lysozyme (lysozyme hereafter) was determined using microED. Lysozyme is a very well studied and structurally characterized protein often used as a model for structure determination method development.[85,87] The protein's easy availability and rapid and reproducible formation of robust crystals makes it an especially useful model.

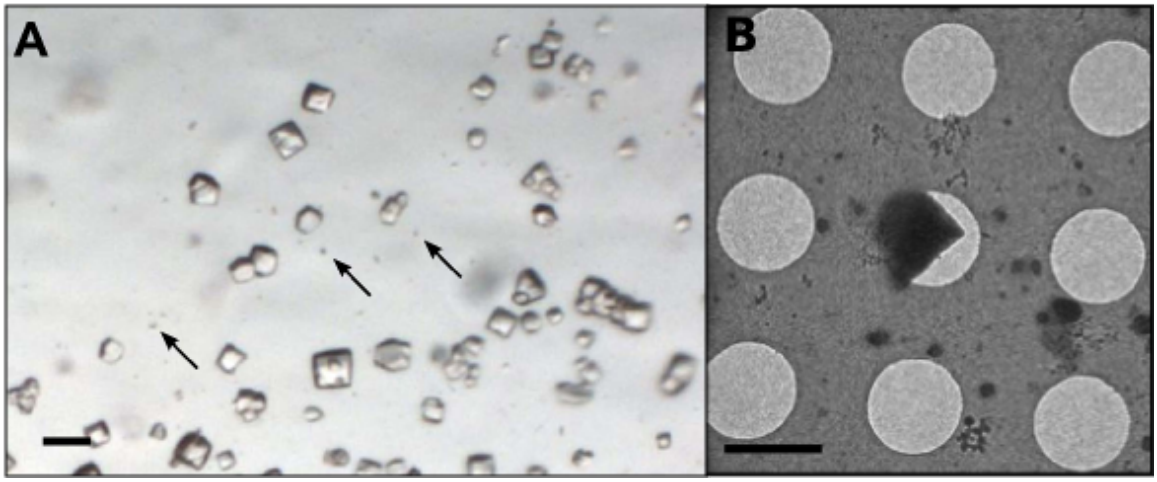
The term lysozyme was coined in 1922 by Sir Alexander Fleming, the discoverer of penicillin, to describe an unknown element with bacteriolytic activity present in a variety of animal tissues. [97] The protein was first purified from egg white and crystallized in the 1940s. [98] The structure of hen egg white lysozyme was determined by X-ray crystallography in 1967 [99] and was the first structure of an enzymatic protein. The first lysozyme structure to be deposited into the PDB was in

1974[100]. Subsequently, almost 400 high-resolution structures of this protein have been deposited.

3.3.1 Growth of lysozyme microcrystals and grid preparation

Lysozyme was purchased from Fisher Scientific and a 200 mg/ml solution was prepared in 50 mM Sodium Acetate pH 4.5. The lysozyme solution was mixed 1 to 1 with precipitant solution (3.5M Sodium chloride; 15% PEG 5,000; 50 mM Sodium acetate pH 4.5) and 2 µl drops were set for the hanging drop vapor diffusion crystallization. Crystals appeared after 1 day of incubation at room temperature. The drops contained crystals of varying size up to approximately 50 x 50 x 50 µm. Microcrystals appeared as barely visible specks under light microscopy (Figure 3).

Figure 3.



Images of lysozyme microrystals. (A) Light micrograph showing lysozyme microcrystals (three examples indicated by arrows) in comparison with larger crystals of the size normally used for X-ray crystallography. Scale bar is 50 µm. **(B)** Lysozyme microcrystals visualized in over-focused diffraction mode on the cryo-EM prior to data collection. Scale bar is 1 µm.

Following crystal formation the sample was diluted 3 to 5 times in 5% PEG 200. A 5µl drop of the crystal solution was applied to a quantifoil 2/2 holey carbon copper EM grid. The grid was then blotted by touching the back of grid with a Whatman filter paper and vitrified by plunging into liquid ethane using a Vitrobot Mark IV (FEI). The frozen-hydrated grid was loaded onto a Gatan 626 cryo-holder and transferred to a cryo-TEM.

3.3.2 Electron Diffraction

All electron microscopy was performed on a FEI Tecnai F20 TEM equipped with a field emission electron source (FEG) and operating at 200 kV. Electron diffraction pattern tilt series data were recorded with a bottom mount TVIPS F416 4k x 4k CMOS camera with pixel size 15.6 mm using built in series exposure mode. The electron dose was kept below 0.01 $e^-/Å^2$ per second and each frame of a data set was taken with an exposure time of up to 10 seconds per frame. The electron dosage was calibrated with the use of a Faraday cage as well as by calibrating the counts on the CMOS detector in bright field mode. The camera length was optimized for the desired resolution as described previously. [57]

Over 100 crystals were surveyed for potential diffraction with about 50% showing diffraction, the best to 1.7Å resolution. Data sets of varying quality were collected from approximately 60 individual microcrystals. Each data set consisted of

up to 90 still frames taken at 1° intervals with a maximum total dose of ~$9e^-/Å^2$ per crystal. Additional datasets were collected with 0.1° tilt intervals.

3.3.3 Data processing

Of the 60 data sets collected three were chosen for structure determination. These sets were chosen for a variety of factors. They had high quality spots extending well past 3.0 Å and were estimated to cover 100% of reciprocal space when symmetry related reflections were combined. Also one of the three data sets came close to intersecting a major plane, which although not necessary for structure determination, made the process easier as the unit cell dimensions could be estimated by measurement. The datasets are detailed in table 1.

Table 1: Datasets used for lysozyme structure determination

Dataset	Images	Angles covered	Increment
1	47	-13° to 33°	1°
2	40	0° to 39°	1°
3	56	-25° to 30°	1°

Each of the three data sets was processed with the programs written for the microED suite (Appendices A1.1 –A1.10):

`1_find_lengths.py`: rough determination of unit cell lengths.

`2_calc_ucvectors.py`: rough determination of the vectors that describe unit the unit cell in 3-dimensional space.

`3_spot_index.py`: indexing of the spots for finer unit cell vector determination

`4_refine_spots.py`: refinement of the spots for unit cell vectors determination

`5_UCR_index.py`: re-indexing of the images using the refined spots

`6_recalculate_vectors.py`: calculation of more accurate unit cell vectors from the new indexing.

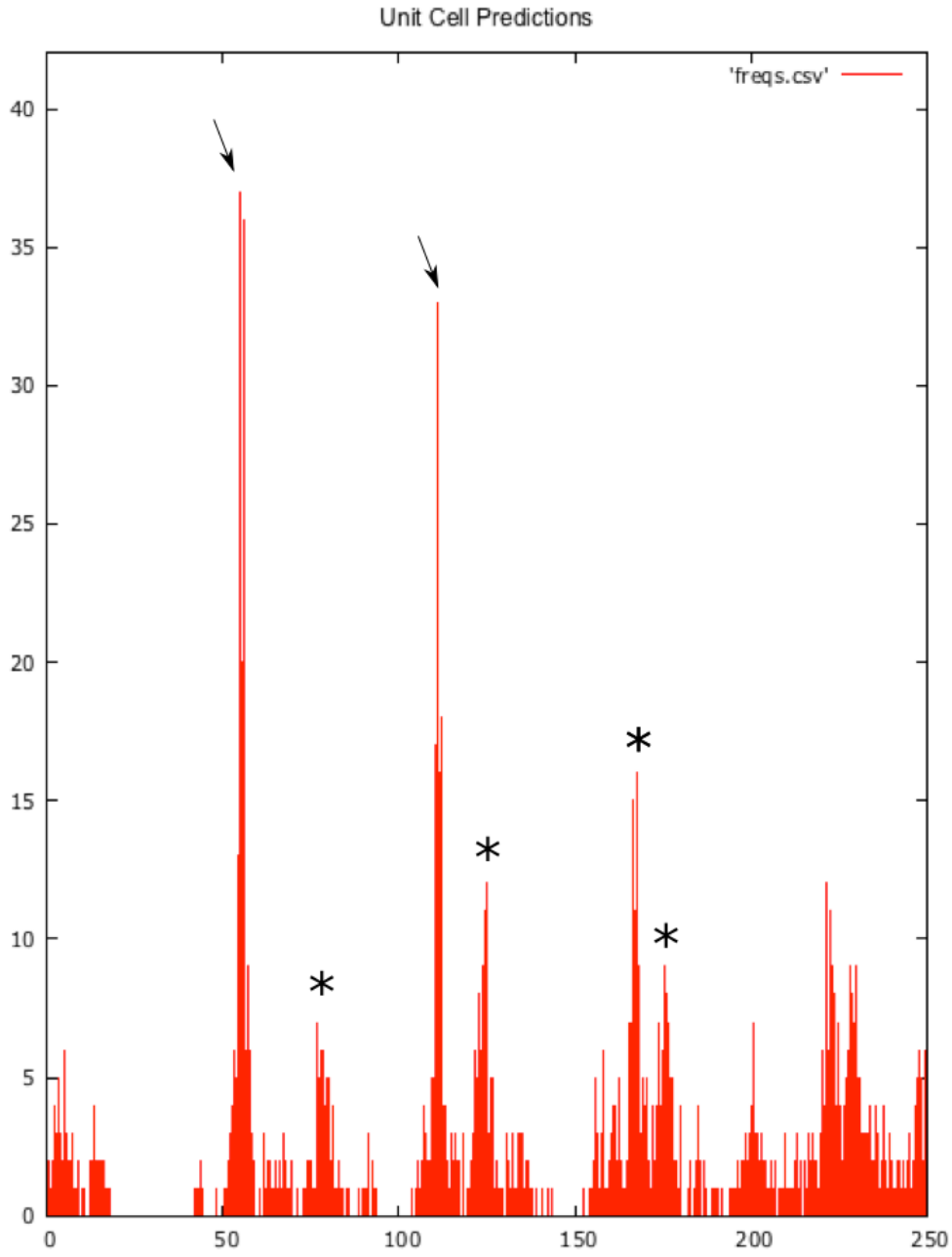
`7_measure_intensities.py`: measurement of background subtracted spot intensities

`merge_p422_maxonly.py`: merging of intensities using only the maximum intensity for each Miller index.

Although the unit cell dimensions of the crystals was already known, unit cell determination was performed on the first dataset to verify the accuracy of the programs. Approximately 900 spots were picked from 20 images from tilts of -20° to 20° and the program `1_find_lengths.py` was run. The output of `1_find_lengths.py` clearly shows the correct unit cell lengths of 55 pixels for a and b and 112 for c. (Figure 4) The *a priori* knowledge that this particular crystal has the dimensions, measured in pixels, of a = b = 55, c = 112 and $\alpha = \beta = \gamma = 90°$ eliminated the need for trial and error indexing that would be required to validate an unknown unit cell.

The unit cell dimensions were entered into `2_calcucvectors` and the resulting vectors used to index the diffraction spots using `3_spot_index.py`. For

Figure 4:



Unit cell predictions for lysozyme by 1_find_lengths.py Predictions were made with approximately 900 spots chosen from 20 images over -20° to 20° tilt. Peaks for the correct a and b (55 px) and c (112 px) unit cell lengths are denoted with arrows. Peaks for a→2b (65 px), 2(a → 2b) (130 px), a → 2c (156 px), and a→ 3b (173 px) are also apparent (marked with stars).

all three data sets the quality of spot indexing was very high so no additional refinement of the unit cell vectors was performed. The intensities of the indexed spots were measured using `7_measure_intensities.py`. The 'integration threshold' value in this program is extremely important: if it is set too low, many 'false positive' spots will be recorded. This occurs when a predicted spot is not present but the overall background in the integrated area is slightly higher than the mean background intensity leading to false assignment of an intensity value. Most of the false positive spots will be discarded during merging, but some false positives will remain, contributing noise and lowering overall data quality. To establish the best threshold that strikes a balance between recording faint spots and avoiding false positives the spot intensities were measured with several different integration thresholds (Table 2).

Table 2: Effect of intensity threshold on number of intensity measurements

Threshhold	Post-merge	Spots recorded			
		Dataset 1	Dataset 2	Dataset 3	Total
0	6169	46392	23742	36916	107050
0.05	6121	25333	14731	32910	72974
0.1	5460	12183	8109	16531	36821

The integrated intensity measurements from all three datasets were individually merged using `merge_p422_maxonly.py` and all three merged datasets taken forward for molecular replacement.

Molecular replacement was performed using Phaser [101] with the lysozyme structure 4AXT [102] as a search model. The highest TFZ score (14.7) was found for the 0.1 threshold dataset so it was taken forward for refinement. The model was

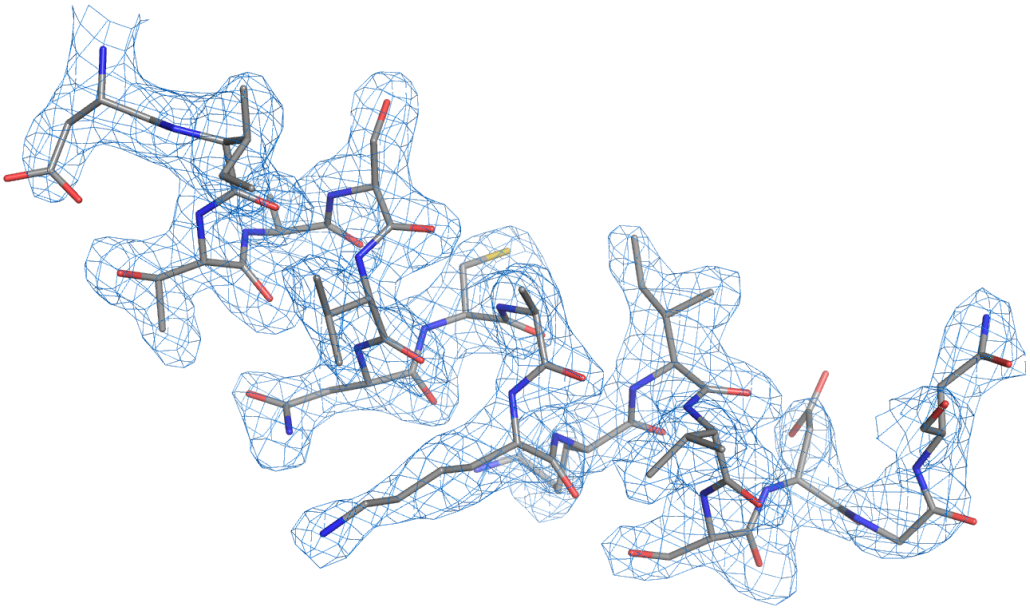
improved by multiple rounds of realspace, individual B-factor, rigid body refinement and simulated annealing using PHENIX. [19]

The final model was calculated to 3.0 Å resolution with acceptable statistics (table 3) and geometry when examined with procheck[103] (Figure 6). The model shows 0.369 Å RMSD (all atoms) and 0.313 Å (Cα only) from PDB 2EPE,[104] a published lysozyme structure at similar resolution but not the model used for molecular replacement

Table 3: Final model statistics

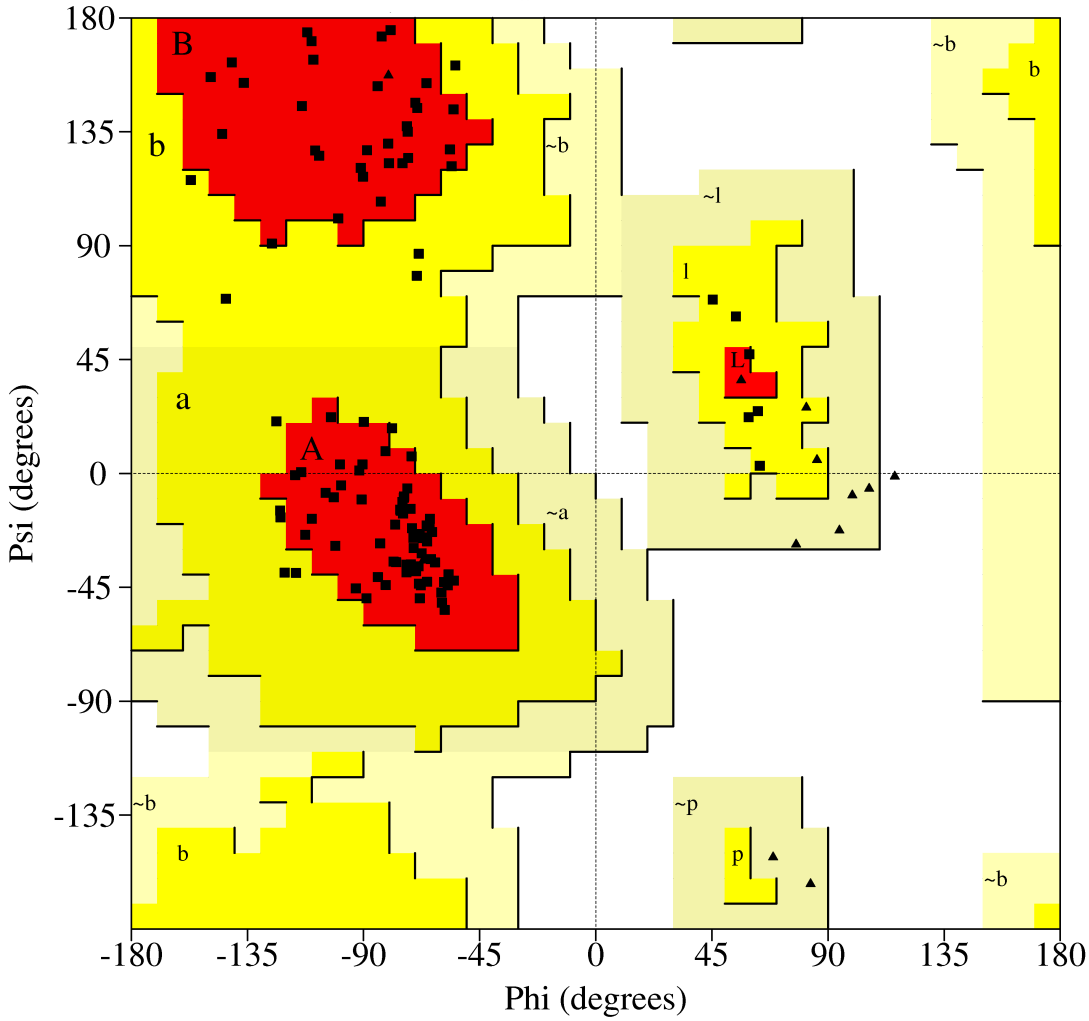
Data Collection	
Excitation voltage	200 kV
Electron Source	Field Emission Gun
Wavelength (Å)	~0.025
Total electron dose per crystal	< 6 $e^-/Å^2$
Number of patterns per crystal	40-60
No. crystals used	3
Total reflections to 2.9Å	36,821
Data Refinement	
Space group	$P4_32_12$
Unit cell dimensions	
a=b	77 Å
c	37 Å
$\alpha=\beta=\gamma$	90°
Resolution	3.0-20.0 Å
Total Unique Reflections	2182
Reflections in working set	1964
Reflections in test set	218
Completeness (3.0-3.7)	95% (89%)
R_{work}/R_{free} (%)	24.1/29.9
RMSD Bonds	0.009 Å
RMSD Angles	0.730°
Ramachandran (%)	
(Preferred, allowed, generous, disallowed)	85; 15; 0; 0

Figure 5.



Example of electron density from the final structure. Electron density for amino acids 87-103 of the 3.0 Å lysozyme structure with final model fit showing the quality of the determined density map.

Figure 6.



Ramachandran plot for the final model. 100% of non-proline/glycine amino acid geometries are allowed. 96 residues (85%) are in favored regions. 17 Residues (15%) are in allowed regions. No residue geometries were in generously allowed or disallowed regions.

4 Confronting the challenges of crystallography without large crystals

4.1 Crystal size and radiation damage

Large protein crystals are necessary because of the destructive effects of radiation damage. For every elastic scattering event that yields useful diffraction data there are on average 10 inelastic X-ray scattering events, each depositing approximately 8 KeV of energy into the crystal. [105] This adds up to approximately 80 KeV of energy being transferred to the crystal per useful elastic scattering event. The energy heats and damages the crystal; breaking covalent bonds, destroying unit cells, and degrading the order and regularity of the crystal lattice. As the crystal degrades and the number of unit cells declines, the recorded intensities are affected. The intensity of diffraction spots is determined by the number of unit cells available to diffract the incipient beam. Because unit cells are destroyed by radiation damage the intensity drops off rapidly during beam exposure. Because structure determination depends on accurate relative intensity measurements, radiation damage effects become a significant problem.

The first major advance in mitigating radiation damage was cryocrystallography. At liquid nitrogen temperature crystals were not only found to suffer significantly less radiation damage, but to yield higher resolution diffraction data.[106] Freezing crystals reduces disorder, which is especially advantageous. It can improve the intensities of high-resolution diffraction spots and can also stabilize

enzyme-substrate and protein-protein interactions allowing them to be visualized with X-ray crystallography. [107]

Electron crystallographers begin with an advantage in the radiation damage arena. The ratio of elastic to inelastic scattering for electrons is about one third that of X-rays and the amount of energy deposited by each inelastic event is approximately 0.075% of an inelastic X-ray scattering event. This leads to a total dose per elastic scattering event that is 1000-fold less then that of X-rays. [105] Despite this, high energy electrons still result in significant radiation damage, especially in the case of the thin 2-dimensional crystals used in electron crystallography. [108]

Microcrystal electron diffraction avoids the negative effects of radiation damage by taking advantage of the lower radiation damage caused by electrons over X-rays and collecting data with very low doses of electrons. We performed an experiment to determine the 'critical dosage threshold'; how much electron dosage a crystal could tolerate before radiation damage began to reduce the quality of diffraction data.

A single lysozyme microcrystal of dimensions approximately 1 x 5 x 0.5 μm was exposed to the electron beam for 10 seconds allowing for the collection of a usable diffraction pattern. The exposure subjected the crystal to a dose of approximately 0.01 $e^-/Å^2$ and was repeated 120 times for a cumulative dose of 12.0 $e^-/Å^2$. The intensities of three diffraction spots, with resolutions ranging from 2.9-4.6 Å, were measured in the 120 resulting diffraction patterns. A plot of the normalized spot intensities over the course of the experiment shows a distinct drop

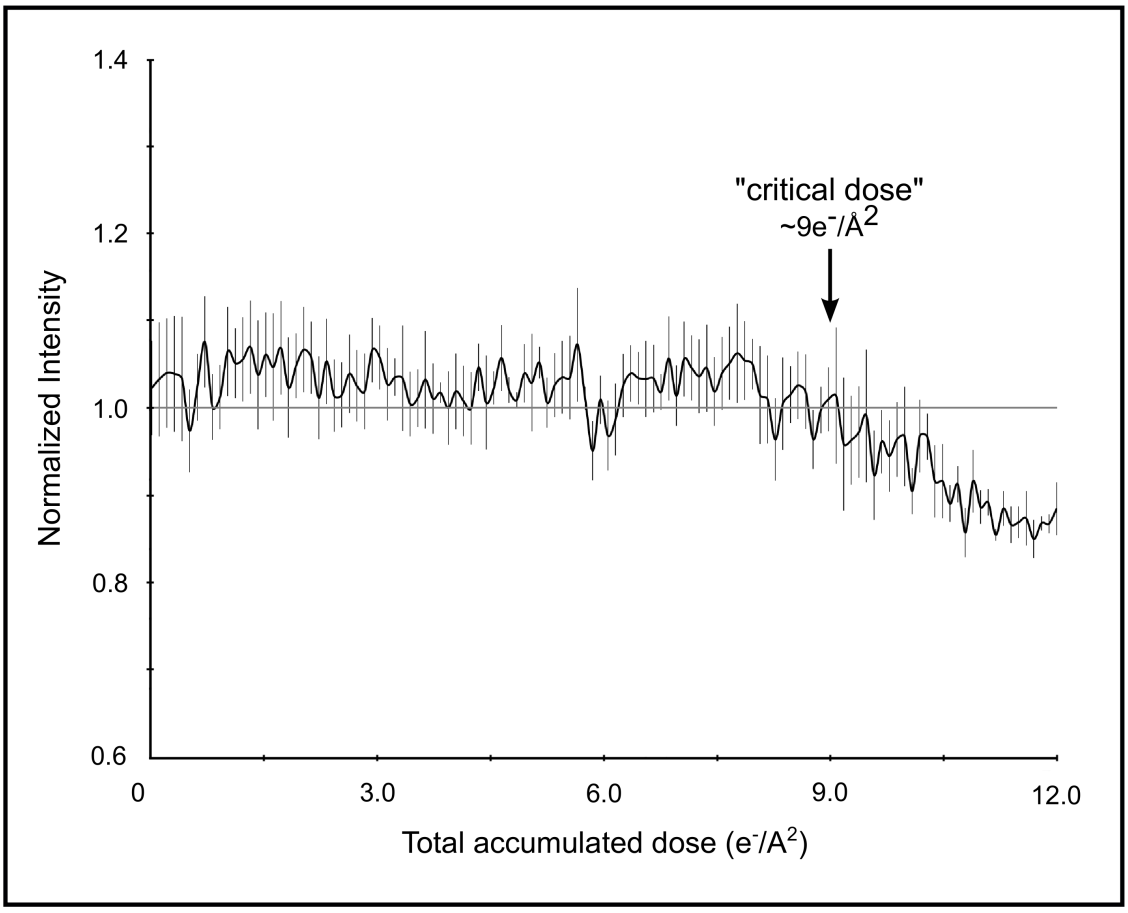
in the measured intensities at a cumulative dose of approximately 9 e-/Å^2. (Figure 7) This critical dose measurement suggests the current data collection procedures should be allow for the collection of a maximum of 90 diffraction patterns from this individual crystal before radiation damage becomes a consideration.

4.2 Measurement of partial intensities and merging

As stated in section 3.2.3.3 a great deal of data was discarded during the merging process. The overall multiplicity for the recorded Bragg peaks was 8.7, meaning that for each peak an average of 8.7 intensity measurements were made. Unfortunately the majority of measurements for any given peak do not record the full intensity. Each Bragg peak can be visualized as a sphere in reciprocal space. The intensity measurement is derived from the intersection of the plane of the Ewald sphere with Bragg peak sphere. If the Ewald sphere does not pass through the center of the reflection a partial intensity is recorded. In traditional X-ray crystallography the crystal is rotated during data collection allowing the ewald sphere to sweep through the Bragg peak, increasing the chance of recording each reflection's full intensity. In the current implementation of the microED technique the crystal is stationary, meaning that not only is probability of recording a partial intensity for any given reflection higher, but this probability increases as resolution increases.

Figure 8 shows the profile of a reflection at ~2.5 Å from 3 consecutive diffraction patterns over 0.2° of tilt. The spot is almost completely covered by this

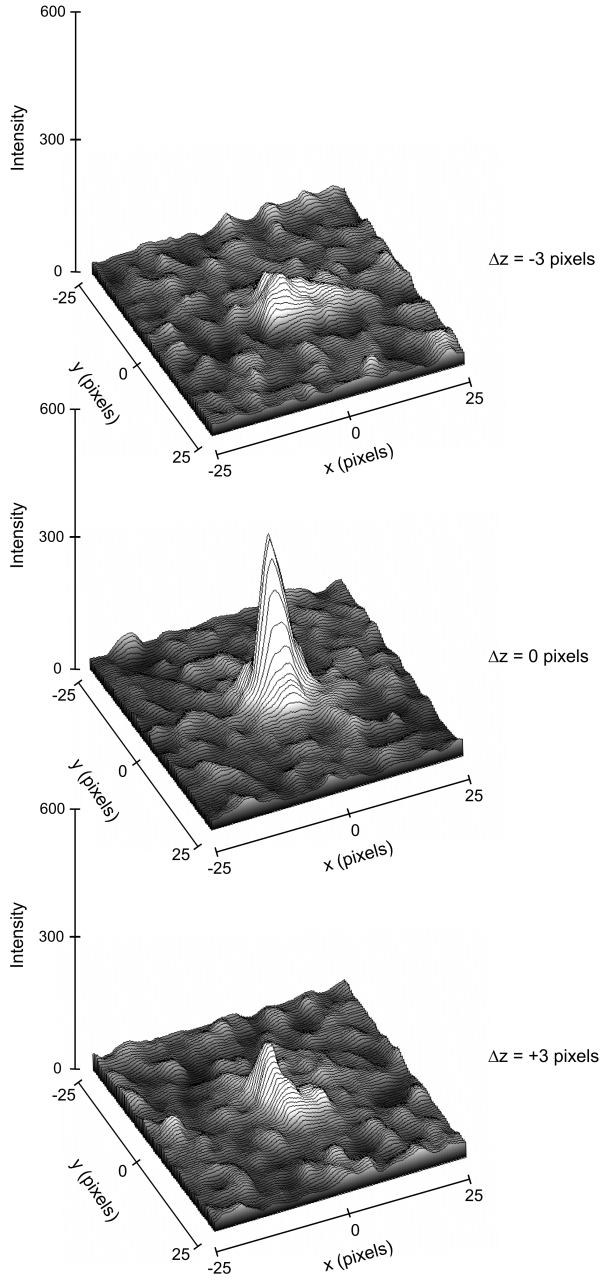
Figure 7.



Effects of cumulative electron dose on diffraction data quality. Normalized intensity versus total accumulated dose for three diffraction spots observed over 120 sequential exposures; each of a dose of ~0.1 $e^-/Å^2$ for a total accumulated dose of ~12 $e^-/Å^2$. A decrease in diffraction intensity becomes apparent at a dosage of ~9 $e^-/Å^2$ ("critical dose"). Bars indicate ± s.e.m.

small tilt increment. During data collection the crystal was tilted at 1° increments suggesting there would be a high probability of recording a partial intensity for this spot, if not missing it completely. This will have negative effects on the data set, which are compounded at high resolution. As resolution increases more of the recorded reflections will be partial intensities and data completeness will decrease

Figure 8.



Three-dimensional profiles of the intensity of a single reflection over three consecutive diffraction patterns. Patterns taken at −0.1°, 0°, and 0.1° degree relative tilts. The plots show the approximate dimensions of the full reflection with a width (full width at half maximum height) of 3–5 pixels in the x, y, and z direction.

as more reflections are missed completely. The probability of detecting any given spot can be crudely estimated by looking at how resolution affects the distance the Ewald sphere plane moves with each tilt increment. This 'travel' (t) can be described as:

$$t = \frac{L}{r} tan\ \theta$$

where L is the image edge length in pixels, r the resolution of the spot and θ the tilt angle. By this calculation the travel at 3.0 Å resolution in these datasets is approximately 23 pixels, approximately half of the smallest lysozyme unit cell.

For this first lysozyme structure the dataset was limited to 3.0 Å resolution to reach an acceptable level of fully recorded intensities. To obtain a dataset with high completeness and full reflections at higher resolutions the tilt increment must be decreased, meaning a larger number of exposures will be required to fully cover the reciprocal space. This would require merging data from a large number of crystals, which raises additional issues with scaling between crystals. More promising technical solutions to this problem are recording diffraction patterns from a continuously rotating crystal or beam precession, which will be discussed further in section 5.1. Beam precession simulates the crystal oscillation in X-ray crystallography by moving the electron beam relative to the crystal.

4.3 Multiple scattering

The vast majority of electromagnetic waves (electrons or X-rays) that interact with matter pass directly through without scattering . The emergent waves from the small percentage of scattered radiation have the possibility of being scattered again before exiting, this phenomenon is called multiple (or dynamical) scattering. Although this probability is extremely small, multiple scattering can have deleterious effects on the accuracy of Bragg peak intensity measurements. Because structure determination depends on accurate measurements of Bragg peak intensities it is important to account for the effects of multiple scattering.

The data collection method of X-ray crystallography eliminates the effects of multiple scattering. During diffraction the 3-dimensional crystal is not stationary, but slowly rotated. This reduces the intensity contributions from multiple scattering events. For multiple scattering to contribute to a diffraction spot, two lattice points must be in contact with the Ewald sphere simultaneously. If this condition is met for a static crystal the dynamic scattering contributes over the whole exposure time. In a moving crystal the amount of time when the two lattice points both share contact with the Ewald sphere is short relative to the total exposure. Although the multiply scattered waves are still measured, their intensities are smeared and contribute only background to the pattern.

Because the crystals in electron diffraction are stationary and electrons interact with matter more strongly, multiple scattering is a concern for electron diffraction. [109] The effects of multiple scattering on the intensities of 2-dimensional

crystal diffraction patterns have been well studied. Using low energy electrons (20 keV) the intensities from 2-dimensional bacteriorhodopsin crystals showed deviations of up to 40% due to multiple scattering.[110] Increasing the energy of the electron beam to 100 keV reduced multiple scattering to less then 10%, an acceptable level for structure determination.[110] The electron beam used for microED data collection had a 200 keV acceleration voltage which will further reduce the amount of multiple scattering. Multiple scattering is dependent on the thickness of the sample. Glaeser and Downing [111] estimated samples thicker than 20 nm would have too much multiple scattering for accurate intensity measurement. Because the lysozyme microcrystals used in these experiments are approximately 25 times thicker than Glaeser and Downing's proposed limit, multiple scattering must be addressed.

An experiment was performed to examine the contribution of multiple scattering to the intensities observed in the microcrystal electron diffraction. Multiply scattered waves contribute to the observed intensity of neighboring Miller indices making the full description of the intensity of any given Bragg peak:

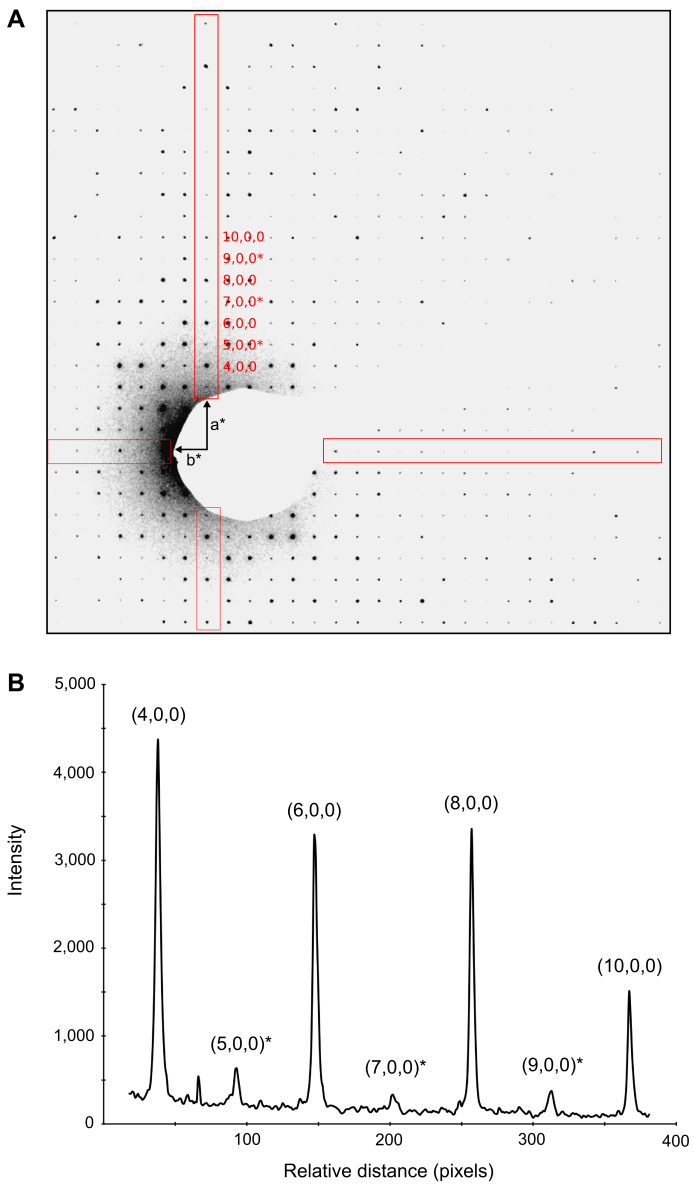
$$I_{total} = I - I_x + I_d$$

where I is the intensity from perfect kinematical scattering at the target index, I_x the intensity lost to multiple scattering at this index and I_d the intensity gained from multiple scattering at other indices.

The measured intensity is the sum of kinematic (I) and multiple scattering ($-I_x+I_d$) and there is no way to de-convolute sources of the final combined intensities. This experiment takes advantage of the symmetry relations of Bragg peaks in the $p4_32_12$ lysozyme crystal to observe only multiple scattering, an adaptation of the technique originally used to demonstrate multiple scattering phenomena in 1937.[112,113]. This symmetry is characterized by a series of systematic absences, Miller indices where, because of symmetry relations, no Bragg peaks are observed ($I = 0$). In the datasets collected for microED, small intensities were observed at the Miller indices expected to contain systematic absences, illustrated in figure 9. At these Miller indices, where I = 0, the observed intensity can be solely attributed to I_d and thus allows estimation of the amount of multiple scattering.

In $p4_32_12$ symmetry systematic absences are expected along the h = 0 and k = 0 Miller indices for Bragg peaks with Miller indices (2n + 1,0,0) and (0,2n + 1,0). The intensity observed at these spots is ascribed to multiple scattering at the four directly adjacent spots at Miller indices (2n ± 2,1,0) and (2n ± 2,-1,0) for the h plane and (1,2n ± 2,0) and (1,2n ± 2,-1,0) for the k plane. The intensities at these expected systematic absences were compared to each of the four adjacent spots that contributed through primary multiple scattering. On average the intensity in the systematic absences was found to be 4.9% of the total intensity of the adjacent spots (max 12.4%, standard deviation 2.7%, n = 17) suggesting that although not negligible, the effects of multiple scattering on the observed intensities are acceptably small.

Figure 9:



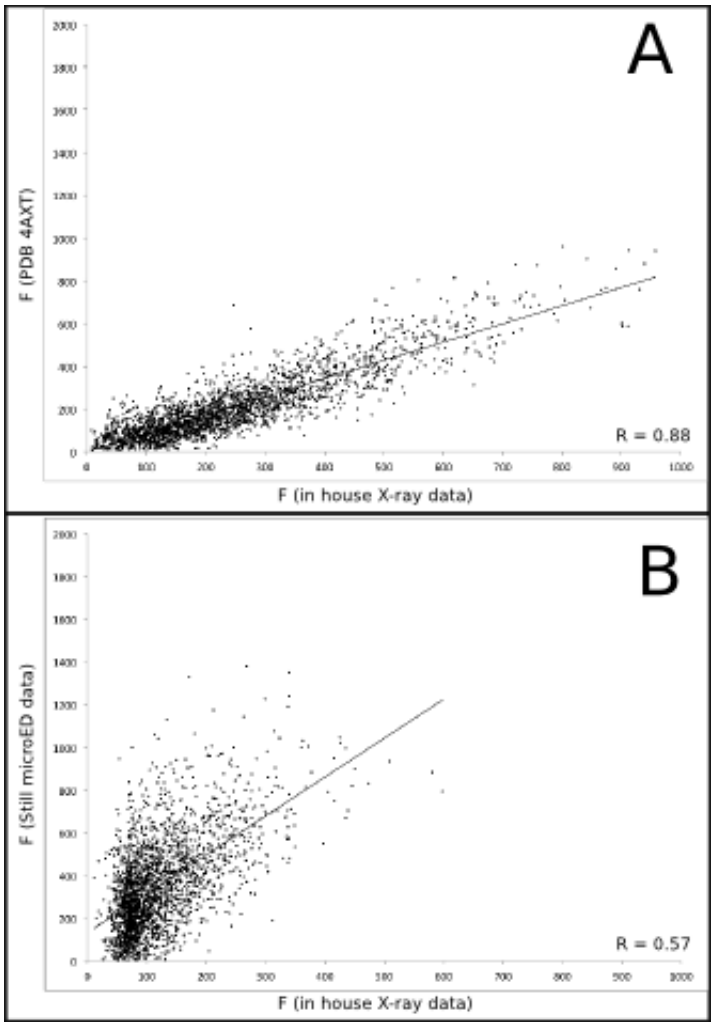
Dynamic scattering in lysozyme 3D crystals. Intensity measurements along the a* axis of a raw diffraction pattern illustrate the relatively small contributions due to dynamic scattering. (A) Diffraction pattern from the major plane of a lysozyme crystal with visible intensity in the (2n+1,0,0) and (0,2n+1,0) Miller indices (boxed in red). (B) (2n+1, 0, 0) reflections (starred) are expected to be systematically absent and observed intensities at these indices are assumed to be the result of dynamic scattering. Image contrast was enhanced for clarity using ImageJ.

4.4 Does molecular replacement work with microED data?

All of the operations for phasing by molecular replacement and model building after merging of symmetry related intensities are identical to those performed in standard X-diffraction. These methods are well-validated and when performed properly minimize the effects of model bias on the final structure.[114] The methods for spot integration and merging in this early incarnation of microED are crude and undoubtedly contribute error to the measured intensities. This raises the question of how much error is acceptable, and whether a solution could be found using what is essentially noise, a phenomenon which has been dramatically demonstrated with electron micrographs. [115] This is especially a concern for the test case using lysozyme, as the model used is essentially identical to the actual solution.

An initial comparison of the microED intensities with intensities obtained by X-ray diffraction of the same crystal form indicate that the intensity measurements follow a similar trend and are not dominated by noise. A complete X-ray diffraction data set was collected from larger crystals taken from the same batch that yielded the microcrystals used for microED. The intensity measurements for each Miller index were compared and showed an acceptable (R=0.57) cross-correlation. (Figure 10 panel B) This supports the hypothesis that the microED data set contains useful information and is not simply random noise.

Figure 10.



Comparisons of structure factors between datasets. Each point represents a single Miller index. In each case the slope of the fit represents the scaling factor between the two datasets. **(A)** Comparison of X-ray dataset collected in house with structure factors from PDB entry 4AXT. **(B)** Comparison of in-house X-ray dataset with microED structure factors from a static crystal with intensities extracted using the microED software suite.

A program was written (`ints-validation.py`, appendix 2.1) to generate multiple randomized datasets to test the robustness of the phasing and model building procedure. The test datasets were generated as follows:

1. All measured intensities were replaced with random numbers ranging between the minimum and maximum of the actual observed experimental values.

2. The experimental intensity values were retained but the Miller indices were randomized.

3. All experimental intensities were replaced with an actual intensity value that was measured by X-ray crystallography of an unrelated structure (Calmodulin, PDB 3SUI) [116]

4. Each experimental intensity was increased or decreased randomly by up to 35%.

In addition, the correct experimental dataset was also used and labeled as dataset "5". These five datasets were treated as "blind test cases" where the user did not know the identities of the test datasets. Each test dataset was subjected to molecular replacement against lysozyme [102] and two rounds of refinement in PHENIX [19] were performed. Only dataset 5, which contained the correct observed experimental intensities yielded a solution by molecular replacement that was able to be further refined to acceptable R_{work}, R_{free} and geometry. Datasets 1-3, which contained the random errors described above did not yield MR solutions. Dataset 4 yielded a lower quality MR solution that was not able to be refined to produce an acceptable structure (Table 4).

TABLE 4: Results of model validation with modified datasets

Data Set	Molecular Replacement Result	TFZ	Final R_{free} (%)*
1[a]	No solution	N/A	N/A
2[b]	Solution**	19.1	54.9
3[c]	No solution	N/A	N/A
4[d]	Solution	12.6	35.2
5[e]	Solution	14.7	29.9

[a]Random Intensities. [b]Shuffled Miller indicies. [c]Calmodulin replaced intensities. [d]Intensities ± up to 35%. [e]Original Data. *Final R_{free} after a minimum of 2 cycles of refinement. **Solution was found, however the space group was incorrect ($P4_12_12$)

A second test of the robustness of the MR procedure was performed by using a number of unrelated structures, chosen from the PDB for their similar unit cell dimensions and protein molecular weights, as search models against the microED experimental data. The unrelated structures were: T4 lysozyme [117], calmodulin [118], dodecin [119] , and αA crystallin [120], their relevant characteristics are listed in table 5. None of the test structures were able to give an acceptable solution with MR.

Together, these experiments indicate that the extracted intensities are sufficiently accurate to yield a reliable structure, and that model bias originating from MR did not skew the results.

Table 5: Models for molecular replacement validation

Protein	PDB ID	M.W. (kDa)	Symmetry	Unit cell dimensions	MR solution found
Hen Egg White Lysozyme	4AXT	14.3	$P4_32_12$	a = b = 78.24 Å c = 37.47 Å α = β = γ = 90°	Yes
T4 Lysozyme	2LZM	18.7	$P3_212$	a = b = 61.20 Å c = 96.80 Å α = β = 90° γ = 120°	No
Calmodulin	3CLN	16.7	P1	a = 29.71 Å, b = 53.79 Å, c = 24.99 Å α = 94.13° β = 97.57°, γ = 89.46°	No
Dodecin	4B2J	8.5	$F4_132$	a = b = c = 142.90 Å α = β = γ = 90°	No
αA Crystallin	3L1E	11.9	$P4_12_12$	a = b = 56.22 Å, c = 68.66 Å α = β = γ = 90°	No

5 Improvements to microED methods

This work is meant to serve as proof of concept that a high resolution protein structure can be determined using the microED method. This work utilized extremely robust crystals of a model protein and used some methodological shortcuts which resulted in a decrease in data accuracy. Additional method development will help to expand the utility of the method, make it accessible to general users, and allow for higher resolution structure determination.

5.1 More accurate intensity measurements

In traditional X-ray crystallography the crystal is rotated as data is collected. As discussed previously this serves two important purposes. First, it allows the Ewald sphere to move through each Bragg peak in reciprocal space resulting in a more complete measurement of the spot's intensity. Second, it spreads out multiple scattering reducing its relative contribution to each reflection. Because the data collection for the microED proof of concept was performed using a stationary crystal both partial intensities and multiple scattering contribute error to the intensity measurements, especially at high resolution. Implementing a rotation data collection method in microED will imporve the accuracy of the intensity measurements and allow better structure determination. This is especially important if the method is to be used to solve novel protein structures, as *ab initio*

phasing methods (as mentioned in section 2.1.2) are dependent on accurate intensity measurements.

5.1.1 Beam precession

The beam precession method developed by Vincent and Midgley[121] provides an effective method for reproducing the positive effects of sample rotation like that performed in X-ray crystallography. In precession the crystal is stationary and the electron beam is aimed so it hits the crystal at an angle. The beam is then rotated at a frequency of several Hz using the condenser lens deflector coils of the microscope. The rotation is in a conical geometry so as to keep the apex of the cone stationary on the sample. Counter-rotation by the image shift coils below the sample stabilizes the image resulting in a static image generated by a precessed beam. The technology for implementing beam precession is well-established in the materials science field and commercially available. [122]

5.2 Improvements to sample preparation

Additional experiments were performed in an attempt to collect data and possibly determine a structure for crystals of proteins other than lysozyme. These experiments demonstrate the sample preparation challenges that may be encountered when expending the technique to proteins that generate crystals that may be more fraglie and less robust that lysozyme.

For microcrystal electron diffraction to become a widely applicable method, sample preparation must be relatively straightforward and reproducible. Although lysozyme crystals readily adhered to the EM grids in these experiments, further screening has shown microcrystals of other proteins to be less cooperative. Attempts have been made to prepare girds with a variety of proteins; the transporters AlaT and XylE, thaumatin, green fluorescent protein (GFP), and glucose isomerase, with varying degrees of success, but no complete data sets have been collected from proteins other than lysozyme.

5.2.1 Cryonegative stain

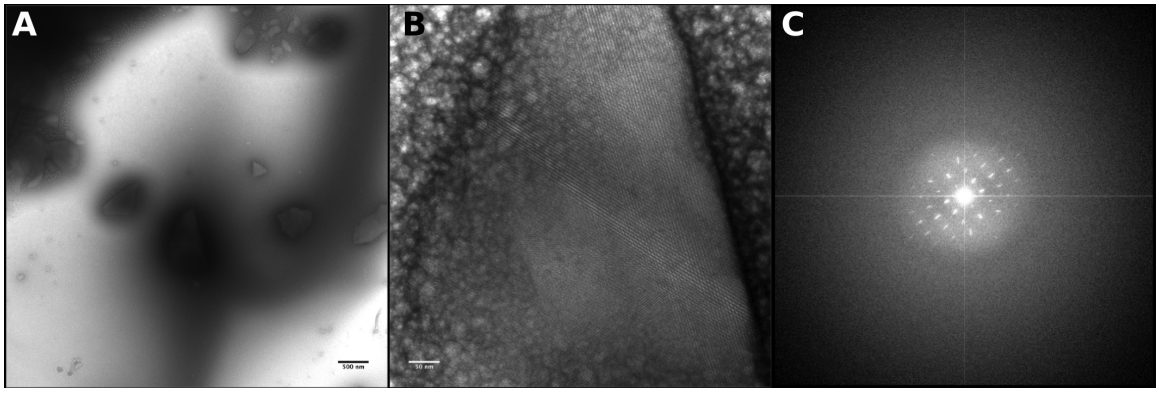
In initial experiments protein micro crystals were first screened by uranyl formate negative stain EM to determine if crystals were present. Negative stain EM is not a viable method for high resolution structure determination because its resolution is limited to 18 – 20 Å [123] and the process dehydrates samples, which would be especially deleterious to crystals containing high water content. Despite this, the speed and relative ease of grid preparation, as well as the high–contrast images generated, makes it a convenient method for higher throughput screening of candidates that can later be taken to cryoEM.

Crystallization drops prepared with the sodium-coupled alanine transporter AlaT were obtained from a collaborator. These drops contained large crystals that had previously shown X-ray diffraction to ~6 Å resolution. Under examination

under negative stain EM the drops showed large number of microcrystals with the crystalline lattice visible by eye. Upon preparing a cryogrid microcrystals were still visible but no diffraction could be obtained. A possible explanation for the lack of diffraction from these crystals is damage during the process of preparing the grid which involves pipetting, blotting, freezing, and subsequent handing of the grid.

A similar phenomenon was observed with Thaumatin, where microcrystals with readily observed under negative stain, showing multiple orders of reflections in their Fourier transform, but were absent on cryogrids. (Figure 11)

Figure 11.



Microcrystals of Thaumatin negatively stained with uranyl formate. **(A**) low magnification **(B)** high magnification, with a lattice visable to the eye. **(C)** Fourier transform of a region of image B, showing multiple reflections.

This raises the question of whether the negative staining procedure was improving the retention of the microcrystals on the EM grid, and if aspects of the technique can be harnessed to improve sample preparation. One technique that adds negative stain reagents to the procedure but still maintains hydration of the sample is cryo-negative staining. Cryonegative staining was originally developed to

combine the contrast enhancing effects of negative staining while preventing sample deformation caused by dehydration.[124] Although contrast enhancement is not necessary for microcrystal electron diffraction, cryo-negative staining may replicate some of the effects that allowed crystal retention under negative stain.

Under Cryo-negative staining samples are applied to EM grids with a relatively large amounts of ammonium molybdate compared to traditional negative staining then blotted and frozen as before. Particles under cryo-negative stain conditions have been observed to spontaneously form two-dimensional arrays, possibly due to the high ionic strength of the stain. [124] This raises the possibility that some effects of the staining process may encourage order in the sample and could be used to improve the microcrystal electron diffraction sample prep. Experiments have shown data collection from cryo-negative stained two-dimensional crystals with no loss of resolution [124] although the technique has never been applied to three-dimensional microcrystals.

Uranyl formate has also been used for cryo-negative stain using a novel carbon sandwich technique, and was also shown to preserve protein complexes in a hydrated state. [125] Modification of this method might also prove useful for microcrystal grid preparation.

5.2.2 Cryosectioning

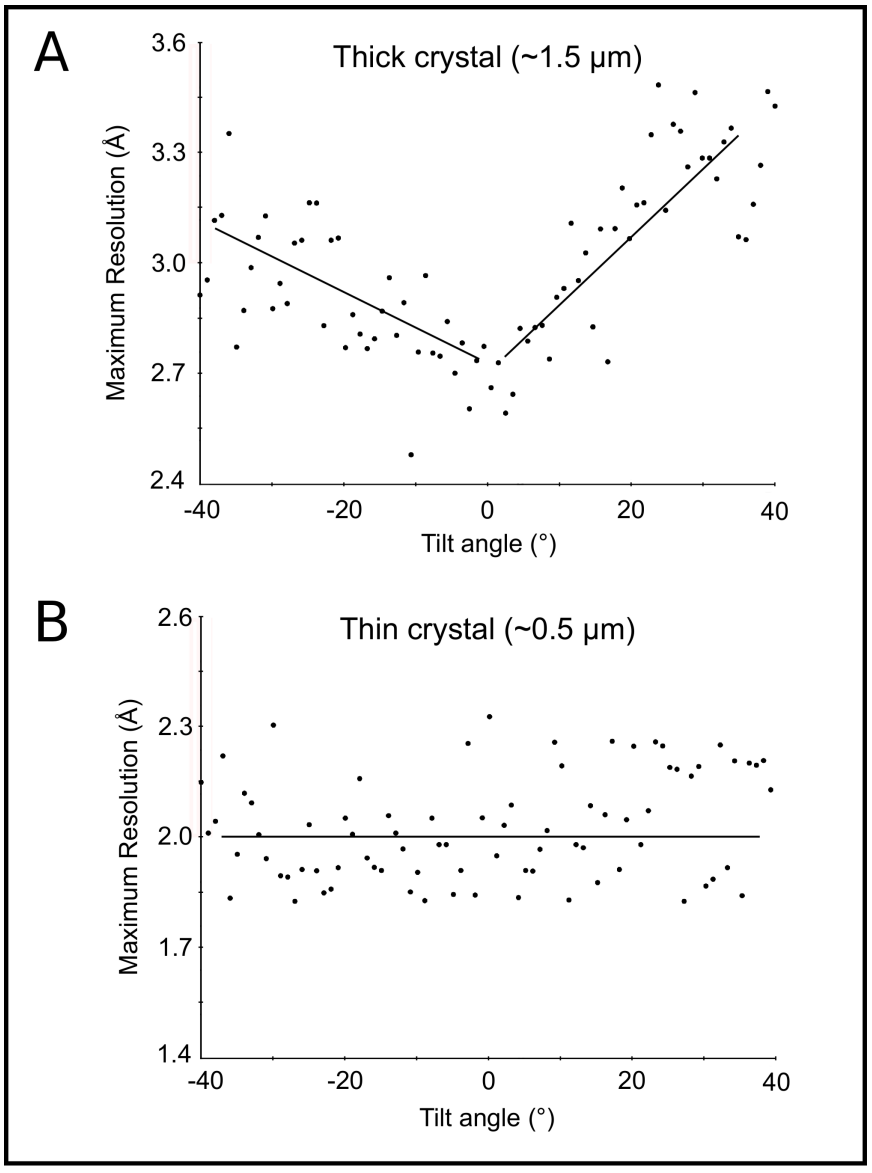
Microcrystal electron diffraction crystal preparations sometimes suffer from the opposite problem from x-ray crystallography; crystals are often too large. The

lysozyme crystal preparations contained crystals of varying size and thickness, and only the smallest and thinnest crystals were chosen for data collection. Even then, because of the limited electron beam penetrance, crystal thickness was an important factor in collection quality data. Crystals that appeared thick, estimated as >3μm, did not yield diffraction data because the electron beam could not penetrate the sample. Crystals that appeared slightly thinner, estimated at ~1.5μm, would provide diffraction, however the quality of diffraction would vary depending on the sample tilt. As the crystal was tilted the actual beam path through it became longer. This had significant effects on the attainable resolution as illustrated in figure 12. Therefore crystals of this thickness and size were not used for data collection. Only the thinnest of crystals, those ~0.5μm thick, were able to generate consistent diffraction resolutions throughout the tilt series.

The need to locate crystals of appropriate thickness may become a limitation especially if the target protein is prone to form crystals that are just slightly too large or if other factors reduce the retention of smaller crystals on the grid. One method that may bypass this problem and increase the high-throughput potential of microcrystal electron diffraction is cryoelectron microscopy of vitreous sections (CEMOVIS). CEMOVIS takes advantage of the fact vitreous ice will cut cleanly without fracturing like crystalline ice allowing for reproducible production of sections down to 50 nm thickness. [126] This might be very advantageous, not only insuring the ability to find crystals of the proper thinness but also allowing collection of diffraction data from crystal sections of uniform thickness. This would

eliminate variability in intensity measurements due to crystal thickness and simplify the merging of data from multiple crystals.

Figure 12.



Effects of crystal thickness on maximum attainable resolution. Maximum resolutions recorded from each frame of an 80 image tilt series with **(A)** thick and **(B)** thin crystals.

6 Conclusion

This work describes proof of principal for protein structure determination by electron diffraction of three-dimensional microcrystals (microED). The method allows for collection of diffraction data from crystals much smaller than those necessary for traditional X-ray diffraction and is more convenient and widely applicable than electron diffraction of two-dimensional protein-lipid crystals.

Currently the largest bottleneck in crystallography is the production of large well ordered crystals. Many proteins of biological interest will not readily form large crystals, although they will often generate very small microcrystals. Many a graduate student and postdoc's career has be spent of optimizing conditions that yield tiny crystals in order to improve the crystal size. It is hoped that with further development this method may allow researchers to streamline this painful optimization process.

This work shows the structure of hen egg white lysozyme, a model protein, determined, by microED, to a resolution comparable to that of X-ray crystallography. The next milestone for the technique is the determination of a novel high-resolution protein structure. This will require refinements and advances in several areas; sample preparation, collection and processing of high resolution data, and experimental phasing.

Another future challenge of this work will be to apply the technique to less tractable proteins. Well-behaved proteins such as lysozyme generally form large crystals very readily and their structure determination by X-ray crystallography has

become routine. Some proteins are only able to form microcrystals and may be prime candidates for this approach, but will require significant method development before mounting crystals to the grid, freezing, and data collection becomes routine.

Improvements in data collection and processing will allow for the collection of accurate intensities for high resolution Bragg peaks. Highly accurate intensity measurements will allow for experimental phasing, a key process for solving structures of proteins with no homologous known structures.

Finally, the integration of three-dimensional microcrystal electron diffraction data processing into existing software suites will make the process more user-friendly and widely applicable. To be truly successful microcrystal electron diffraction must not only be a method that fills gaps left by current techniques, but must be practical and accessible to academic and industrial researchers across the world.

References

1 Bragg, W. H. & Bragg, W. L. The Reflection of X-rays by Crystals. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **88**, 428-438, (1913).

2 Kendrew, J. C. *et al.* A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181**, 662-666, (1958).

3 Ashworth, J. *et al.* Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. *Nucleic Acids Research* **38**, 5601-5608, (2010).

4 Jiang, L. *et al.* De novo computational design of retro-aldol enzymes. *Science* **319**, 1387-1391, (2008).

5 Baker, A., Varghese, J., Laver, W., Air, G. & Colman, P. Three-Dimensional Structure of Neuraminidase of Subtype N9 From an Avian Influenza Virus. *Proteins: Structure, Function, and Genetics* **2**, 111-117, (1987).

6 Von Itzstein M *et al.* Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* **363**, 418-423, (1993).

7 Bernstein, F. C. *et al.* The Protein Data Bank: A Computer-based Archival File For Macromolecular Structures. *Journal of Molecular Biology* **112**, 535, (1977).

8 *Protein Data Bank*, <http://pdb.org/pdb/home/home.do> (2013).

9 Perutz, M. F., Rossmann, M. G., Cullis, A. F., Muirhead, H. & Will, G. Structure of Haemoglobin: A Three-Dimensional Fourier Synthesis at 5.5Å Resolution, Obtained by X-ray Analysis. *Nature* **185**, 416-422, (1960).

10 Kendrew, J. C. *et al.* Structure of Myoglobin: A Three Dimensional Fourier Synthesis at 2A Resolution. *Nature* **185**, 422-427, (1960).

11 Kosinska-Eriksson, U. *et al.* Subangstrom resolution X-ray structure details aquaporin-water interactions. *Science* **340**, 1346-1349, (2013).

12 de Groot, B. L., Frigato, T., Helms, V. & Grubmüller, H. The Mechanism of Proton Exclusion in the Aquaporin-1 Water Channel. *Journal of Molecular Biology* **333**, 279-293, (2003).

13 Taylor, G. The phase problem. *Acta Crystallographica Section D* **59**, 1881-1890, (2003).

14 Harker, D. The determination of the phases of the structure factors of noncentrosymmetric crystals by the method of double iomorphous replacement *Acta Crystallographia* **9**, 867-873, (1956).

15 Rossman, M. G. & Blow, D. Determination of Phases by the Conditions of Non-Crystallographic symmetry. *Acta Crystallographia* **16**, 39-45, (1963).

16 Rossman, M. G. & Blow, D. The Detection of Sub-Units Within the Crystallographic Asymmetric Unit. *Acta Crystallographia* **15**, 24-31, (1962).

17 Kaufmann, K. W., Lemmon, G. H., Deluca, S. L., Sheehan, J. H. & Meiler, J. Practically useful: what the Rosetta protein modeling suite can do for you. *Biochemistry* **49**, 2987-2998, (2010).

18 Das, R. & Baker, D. Prospects for de novo phasing with de novo protein models. *Acta Crystallogr D Biol Crystallogr* **65**, 169-175, (2009).

19 Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographia Section D* **D66**, 213-221, (2010).

20 Afonine, P. *et al.* Towards automated crystallographic structure refinement with phenix.refine. . *Acta Crystallographia Section D* **D68:**, 352-367, (2012).

21 Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallographia Section D* **D67**, 235-242, (2011).

22 Wu, B. *et al.* Structures of the CXCR4 chemokine GPCR with small-molecule and cyclic peptide antagonists. *Science* **330**, 1066-1071, (2010).

23 Shimamura, T. *et al.* Structure of the human histamine H1 receptor complex with doxepin. *Nature* **475**, 65-70, (2011).

24 Rasmussen, S. G. *et al.* Crystal structure of the beta2 adrenergic receptor-Gs protein complex. *Nature* **477**, 549-555, (2011).

25 Chien, E. Y. *et al.* Structure of the human dopamine D3 receptor in complex with a D2/D3 selective antagonist. *Science* **330**, 1091-1095, (2010).

26 Landau, E. & Rosenbusch, J. Lipidic cubic phases: A novel concept for the crystallization of membrane proteins. *Proceedings of the National Acadamy of Sciences USA* **93**, 14532-14535, (1996).

27 Cherezov, V. Lipidic cubic phase technologies for membrane protein structural studies. *Curr Opin Struct Biol* **21**, 559-566, (2011).

28 Li, D., Boland, C., Walsh, K. & Caffrey, M. Use of a Robot for High-throughput Crystallization of Membrane Proteins in Lipidic Mesophases. *Journal of Visualized Experiments* **67**, e4000, (2012).

29 Cherezov, V., Fersi, H. & Caffrey, M. Crystallization Screens: Compatibility with the Lipidic Cubic Phase for in Meso Crystallization of Membrane Proteins. *Biophysical Journal* **81**, 225-424, (2001).

30 Otwinowski, Z. & Minor, W. Processing of X-ray Diffraction Data Collected in Oscillation Mode. *Methods in Enzymology* **276**, 307-326, (1997).

31 Leslie, A. G. W. & Powell, H. R. Processing Diffraction Data with Mosflm. in: *Evolving Methods for Macromolecular Crystallography* Vol. 245 (eds Randy J. Read & Joel L. Sussman) 41-51 (Springer, 2007).

32 Evans, P. R. An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallographia Section D* **D67**, 282-292, (2011).

33 Ness, S. R., de Graaff, R. A., Abrahams, J. P. & Pannu, N. S. CRANK: new methods for automated macromolecular crystal structure solution. *Structure* **12**, 1753-1761, (2004).

34 Vagin, A. & Teplyakoz, A. MOLREP: an Automated Program for Molecular Replacement. *Journal of Applied Crystallography* **30**, 1022-1025, (1997).

35 Cowtan, K., Emsley, P. & Wilson, K. S. From crystal to structure with CCP4. *Acta Crystallogr D Biol Crystallogr* **67**, 233-234, (2011).

36 Bradley, D. E. Electron-Microscopic Study of Finback Whale Myoglobin Crystals. *Nature* **183**, 941-943, (1959).

37 Henderson, R. & Unwin, P. N. T. Three-dimsensional model of purple membrane obtained by electron microscopy. *Nature* **257**, 28-32, (1975).

38 Unwin, P. N. T. & Henderson, R. Molecular Structure Determination by Electron Microscopy of Unstained Crystalline Specimens. *Journal of Molecular Biology* **94**, 425-440, (1975).

39 Abeyrathne, P. D. *et al.* Electron Microscopy Analysis of 2D Crystals of Membrane Proteins. in: *Comprehensive Biophysics* Vol. 1.19 277-310 (Academic Press, 2011).

40 Gonen, T. *et al.* Lipid-protein interactions in double-layered two-dimensional AQP0 crystals. *Nature* **438**, 633-638, (2005).

41 Tani, K. *et al.* Mechanism of aquaporin-4's fast and highly selective water conduction and proton exclusion. *Journal of Molecular Biology* **389**, 694-706, (2009).

42 Hite, R. K., Li, Z. & Walz, T. Principles of membrane protein interactions with annular lipids deduced from aquaporin-0 2D crystals. *EMBO Journal* **29**, 1652-1658, (2010).

43 Mitsuoka, K. *et al.* The Structure of Bacteriorhodopsin at 3.0 Å Resolution Based on Electron Crystallography: Implication of the Charge Distribution. *Journal of Molecular Biology* **286**, 861-882, (1999).

44 Hunte, C. & Richers, S. Lipids and membrane protein structures. *Current Opinions in Structural Biology* **18**, 406-411, (2008).

45 Appel, M., Hizlan, D., Vinothkumar, K. R., Ziegler, C. & Kuhlbrandt, W. Conformations of NhaA, the Na/H exchanger from Escherichia coli, in the pH-activated and ion-translocating states. *Journal of Molecular Biology* **386**, 351-365, (2009).

46 Gonen, T., Sliz, P., Kistler, J., Cheng, Y. & Walz, T. Aquaporin-0 membrane junctions reveal the structure of a closed water pore. *Nature* **429**, 193-197, (2004).

47 Rémigy, H. W. *et al.* Membrane protein reconstitution and crystallization by controlled dilution. *FEBS Letters* **555**, 160-169, (2003).

48 Rigaud, J. L. *et al.* Bio-Beads: An Efficient Strategy for Two-Dimensional Crystallization of Membrane Proteins. *Journal of Structural Biology* **118**, 226-235, (1997).

49 Signorell, G. A., Kaufmann, T. C., Kukulski, W., Engel, A. & Remigy, H. W. Controlled 2D crystallization of membrane proteins using methyl-beta-cyclodextrin. *Journal of Structural Biololgy* **157**, 321-328, (2007).

50 Unwin, N. Refined structure of the nicotinic acetylcholine receptor at 4Å resolution. *Journal of Molecular Biology* **346**, 967-989, (2005).

51 Murata, K. *et al.* Structural determinants of water permeation through aquaporin-1. *Nature* **407**, 599-605, (2000).

52 Sazanov, L. A. & Walker, J. E. Cryo-electron Crystallography of Two Sub-complexes of Bovine Complex I Reveals the Relationship between the Membrane and Peripheral Arms. *Journal of Molecular Biology* **302**, 455-464, (2000).

53 Frey, T. G., Chan, S. H. P. & Schatz, G. Structure and Orientation of Cytochrome c Oxidase in Crystalline Membranes. *Journal of Biological Chemistry* **253**, 4389-4395, (1978).

54 Iacovache, I. *et al.* The 2DX robot: a membrane protein 2D crystallization Swiss Army knife. *Journal of Structural Biology* **169**, 370-378, (2010).

55 Kaufmann, T. C., Engel, A. & Remigy, H. W. A novel method for detergent concentration determination. *Biophysical Journal* **90**, 310-317, (2006).

56 Coudray, N. *et al.* Automated screening of 2D crystallization trials using transmission electron microscopy: a high-throughput tool-chain for sample preparation and microscopic analysis. *Journal of Structural Biology* **173**, 365-374, (2011).

57 Gonen, T. The collection of high-resolution electron diffraction data. in: *Electron crystallography of soluable and membrane proteins: Methods and protocols* Vol. 955 (eds Ingeborg Schmidt-Krey & Yifan Cheng) 153-169 (Methods in molecular biology, 2013).

58 Grassucci, R. A., Taylor, D. J. & Frank, J. Preparation of macromolecular complexes for cryo-electron microscopy. *Nature Protocols* **2**, 3239-3246, (2007).

59 Iancu, C. *et al.* Electron cryotomography sample preparation using the Vitrobot. *Nature Protocols* **1**, 2813-2819, (2006).

60 Goldie, K. N. *et al.* Cryo-electron Microscopy of Membrane Proteins. *Methods Molecular Biology* **1117**, 325-341, (2014).

61 Schenk, A. *et al.* 3D reconstruction from 2D crystal image and diffraction data. *Methods in Enzymology* **482**, 101-129, (2010).

62 Arheit, M. *et al.* Merging of image data in electron crystallography. *Methods Molecular Biology* **955**, 195-209, (2013).

63 Arheit, M., Castaño-Diéz, D., Thierry, R., Gipson, B. & Stahlberg, H. Automation of image processing in electron crystallography. *Methods Molecular Biology* **955**, 313-330, (2013).

64 Arheit, M. *et al.* Image processing of 2D crystal images. *Methods Molecular Biology* **955**, 171-194, (2013).

65 Gipson, B., Zeng, X., Zhang, Z. Y. & Stahlberg, H. 2dx--user-friendly image processing for 2D crystals. *Journal of Structural Biology* **157**, 64-72, (2007).

66 Gipson, B., Zeng, X. & Stahlberg, H. 2dx_merge: data management and merging for 2D crystal images. *Journal of Structural Biology* **160**, 375-384, (2007).

67 Burmester, C. & Schröder, R. R. Solving the phase probem in protein electron crystallogrpahy: Multiple isomorphous replacement and anamolous dispersion as alternatives to imaging. *Scanning Microscopy* **11**, 323-334, (1997).

68 Wisedchaisri, G. & Gonen, T. Phasing electron diffraction data by molecular replacement: Strategy for structure determination and refinement. in: *Electron Crystallography of Soluable and Membrane Proteins: Methods and Protocols.* Vol. 955 (eds Ingeborg Schmidt-Krey & Yifan Cheng) Ch. 14, 243-272 (Springer Science, 243).

69 Wisedchaisri, G. & Gonen, T. Fragment-based Phase Extension for Three-Dimensional Structure Determination of Membrane Proteins by Electron Crystallography. *Structure* **19**, 976-987, (2011).

70 Fogaça, M. V., Reis, F. M. C. V., Campos, A. C. & Guimarães, F. S. Effects of intra-prelimbic prefrontal cortex injection of cannabidiol on anxiety-like behavior: Involvement of 5HT1A receptors and previous stressful experience. *European Neuropsychopharmacology*, (2013).

71 Sasaki, K. & Norwitz, E. R. Gonadotropin-releasing hormone/gonadotropin-releasing hormone receptor signaling in the placenta. *Current Opinion in Endocrinology & Diabetes and Obesity* **18**, 401-408, (2011).

72 Noel, S. D. & Kaiser, U. B. G protein-coupled receptors involved in GnRH regulation: Molecular insights from human disease. *Molecular and Cellular Endocrinology* **346**, 91-101, (2011).

73 Milan-Lobo, L., Enquist, J., van Rijn, R. M. & Whistler, J. L. Anti-analgesic effect of the mu/delta opioid receptor heteromer revealed by ligand-biased antagonism. *PLoS One* **8**, e58362, (2013).

74 Chen, C. C. *et al.* A role for ASIC3 in the modulation of high-intensity pain stimuli. *Proceedings of the National Acadamy of Sciences USA* **99**, 8992-8997, (2002).

75 Jain, S. *et al.* A haplotype of Angiotensin receptor type 1 associated with human hypertension increases blood pressure in transgenic mice. *Journal of Biological Chemistry* **288**, 37048-37056, (2013).

76 Shichida, Y. & Matsuyama, T. Evolution of opsins and phototransduction. *Philosophical Transacions of the Royal Society Series B Biological Sciences* **364**, 2881-2895, (2009).

77 Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nature Reviews Drug Discovery* **5**, 993-996, (2006).

78 Salon, J. A., Lodowski, D. T. & Palczewski, K. The significance of G protein-coupled receptor crystallography for drug discovery. *Pharmacological Reviews* **63**, 901-937, (2011).

79 Palczewski, K. Crystal Structure of Rhodopsin: A G Protein-Coupled Receptor. *Science* **289**, 739-745, (2000).

80 Rasmussen, S. G. *et al.* Crystal structure of the human beta2 adrenergic G-protein-coupled receptor. *Nature* **450**, 383-387, (2007).

81 Riekel, C., Burghammer, M. & Schertler, G. Protein crystallography microdiffraction. *Current Opinions in Structural Biology* **15**, 556-562, (2005).

82 Nelson, R. *et al.* Structure of the cross- β spine of amyloid-like fibrils. *Nature* **435**, 773-778, (2005).

83 Glaeser, R. *et al.* Characterization of Conditions Required for X-Ray Diffraction Experiments with Protein Microcrystals. *Biophysical Journal* **78**, 3178-3185, (2000).

84 Teng, T.-y. & Moffat, K. Primary radiation damage of protein crystals by an intense synchrotron X-ray beam. *Journal of Synchrotron Radiation* **7**, 313-317, (2000).

85 Chapman, H. N. *et al.* Femtosecond X-ray protein nanocrystallography. *Nature* **470**, 73-77, (2011).

86 Richard Neutze, Remco Wouts, David van der Spoel, Edgar Weckert & Hajdu, J. Potential for biomolecular imaging with femtosecond X-ray pulses. *Nature* **406**, 752-757, (2000).

87 Sébastien Boutet *et al.* High-Resolution Protein Structure Determination by Serial Femtosecond Crystallography. *Science* **337**, 362-364, (2012).

88 White, T. A. *et al.* CrystFEL: a software suite for snapshot serial crystallography. *Journal of Applied Crystallography* **45**, 335-341, (2012).

89 Oliphant, T. E. Python for scientific computing. *Computing in Science and Engineering* **9**, 10-20, (2007).

90 *Python Imaging Library (PIL)*, <http://www.pythonware.com/products/pil/> (2013).

91 *24.1. Tkinter — Python interface to Tcl/Tk*, <http://docs.python.org/2/library/tkinter.html> (2013).

92 Williams, T. & Kelley, C. *Gnuplot 4.5: an interactive plotting program*, <http://gnuplot.info> (2011).

93 *Imagemagick*, <http://www.imagemagick.org/script/index.php> (2013).

94 Schindelin, J. *et al.* Fiji: an open-source platform for biological-image analysis. *Nature Methods* **9**, 676-682, (2012).

95 Steller, I., Bolotovsky, R. & Rossmann, M. G. An Algorithm for Automatic Indexing of Oscillation Images using Fourier Analysis. *Journal of Applied Crystallography* **30**, 1036-1040, (1997).

96 Powell, H. R. The Rossmann Fourier autoindexing algorithm in MOSFLM. *Acta Crystallographia Section D* **D55**, 1690-1695, (1999).

97 Fleming, A. On a Remarkable Bacteriolytic Element Found in Tissues and Secretions. *Proceedings of the Royal Society B: Biological Sciences* **93**, 306-317, (1922).

98 Alderton, G. & Fevold, H. L. Direct crystallization of lysozyme from egg white and some crystalline salts of lysozyme. *Journal of Biological Chemistry* **165**, 1-5, (1946).

99 Blake, C. C. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. On the Conformation of the Hen Egg-White Lysozyme Molecule. *Proceedings of the Royal Society B: Biological Sciences* **167**, 365-377, (1967).

100 Diamond, B. Real-space Refinement of the Structure of Hen Egg-white Lysozyme. *Journal of Molecular Biology* **82**, 371-391, (1974).

101 McCoy, A. J. *et al.* Phaser crystallographic software. *Journal of Applied Crystallography* **40**, 658-674, (2007).

102 Cipriani, F. *et al.* CrystalDirect: a new method for automated crystal harvesting based on laser-induced photoablation of thin films. *Acta Crystallographia Section D* **D68**, 1393-1399, (2012).

103 Laskowski, R., MacArthur, M., Moss, D. & Thornton, J. PROCHECK - a program to check the stereochemical quality of protein structures. . *Journal of Applied Crystallography* **26**, 283-291, (1993).

104 Naresh, M. D. *et al. Crystal structure analysis of Hen egg white lysozyme grown by capillary method*, <http://www.pdb.org/pdb/explore/explore.do?structureId=2epe> (2007).

105 Henderson, R. The potential and limitations of neutrons, electrons and X-rays for atomic resolution microscopy of unstained biological molecules. *Quarterly Reviews of Biophysics* **28**, 171-193, (1995).

106 Petsko, G. A. Protein crystallography at sub-zero temperatures: Cryo-protective mother liquors for protein crystals. *Journal of Molecular Biology* **97**, 381-392, (1975).

107 Douzou, P. & Bon Hoa, G. H. Protein crystallography at sub-zero temperatures: Lysozyme-substrate complexes in cooled mixed solvents. *Journal of Molecular Biology* **96**, 367-380, (1975).

108 Glaeser, R. N. Limitations to Significant Information in Biological Electron Microscopy as a Result of Radiation Damage *Journal of Ultrastructure Research* **36**, 466-482, (1971).

109 Zhang, D., Oleynikov, P., Hovmöller, S. & Zou, X. Collecting 3D electron diffraction data by the rotation method. *Zeitschrift für Kristallographie* **225**, 94-102, (2010).

110 Glaeser, R. M. High-voltage electron diffraction from bacteriorhodopsin (purple membrane) is measurably dynamical. *Acta Crystallographia Section A* **A45**, 620-628, (1989).

111 Glaeser, R. M. & Downing, K. H. High-resolution electron crystallography of protein molecules. *Ultramicroscopy* **52**, 478-486, (1993).

112 Renninger, M. "Umweganregung" eine bisher unbeachtete Wechselwirkungserscheinung bei Raumgitterinterferenzen. *Zeitschrift für Physik* **106**, 141-176, (1937).

113 Rossmanith, E. UMWEG-98: a program for calculation and graphical representation of multiple diffraction patterns. *Journal of Applied Crystallography* **32**, 355-361, (1999).

114 Rossman, M. G. The Molecular Replacement Method. *Acta Crystallographia Section A* **A46**, 73-82, (1990).

115 Shatsky, M., Hall, R. J., Brenner, S. E. & Glaeser, R. M. A method for the alignment of heterogeneous macromolecules from electron microscopy. *Journal of Structural Biology* **166**, 67-78, (2009).

116 Lau, S. Y., Procko, E. & Gaudet, R. Distinct properties of Ca2+-calmodulin binding to N- and C-terminal regulatory regions of the TRPV1 channel. *Journal of General Physiology* **140**, 541-555, (2012).

117 Weaver, L. H. & Matthews, B. W. Structure of bacteriophage T4 lysozyme refined at 1.7 Å resolution. *Journal of Molecular Biology* **193**, 189-199. , (1987).

118 Babu, Y. S., Bugg, C. E. & Cook, W. J. Structure of calmodulin refined at 2.2 Å resolution. . *Journal of Molecular Biology* **204**, 191-204. , (1988).

119 Staudt, H. *et al.* Directed manipulation of a flavoprotein photocycle. . *Angewandte Chemie* **52**, 8463-8466, (2013).

120 Laganowsky, A. *et al.* Crystal structures of truncated alphaA and alphaB crystallins reveal structural mechanisms of polydispersity important for eye lens function. *Protein Science* **19**, 1031-1043, (2010).

121 Vincent, R. & Midgley, P. Double conical rocking-beam system for measurement of integrated electron diffraction intensities. *Ultramicroscopy* **53**, 271-282, (1994).

122 *NanoMegas Advanced Tools for Electron Diffraction*, <http://www.nanomegas.com/> (2014).

123 Robin Harris, J. & Horne, R. W. Negative staining: A brief assessment of current technical benefits, limitations and future possibilities. *Micron* **25**, 5-13, (1994).

124 Adrian, M., DuBochet, J., Fuller, S. D. & Harris, J. R. Cryo-negative Staining. *Micron* **29**, 145-160, (1998).

125 Golas, M. M., Sander, B., Will, C. L., Luhrmann, R. & Stark, H. Molecular architecture of the multiprotein splicing factor SF3b. *Science* **300**, 980-984, (2003).

126 Al-Amoudi, A. *et al.* Cryo-electron microscopy of vitreous sections. *EMBO J* **23**, 3583-3588, (2004).

Appendix 1: Source code of programs in the microED suite.
All programs are distributed under the GNU General Public License (GPL) and available at http://github.com/gonenlab/2013UED.git.

A1.1: find lengths.py

Charts the distribution of the shortest vectors between user chosen spots allowing for estimation of unit cell dimensions. Output a histogram and .csv file containing number of observations for each length bin.

```
#! /usr/bin/env python
## imports

## Finds the average unit cell vector using find cell_params.json which is created by
user or from imageJ files and i2cfiles.py

# Matt Iadanza 2013-07-05

import json
import os
import math
import numpy
from collections import Counter
import datetime

# get some variables:
data = json.load(open('cataspot.json'))
output = open("magfreqs.csv", "w")
drawplot = open("plot", "w")
params = json.load(open('params.json'))


### prepare the logfile:
logout = open("logfile.txt", "w")
logout.write("** MxED logfile **\nusing cataspot.json created: %s\nuser name:
%s\n\nPrograms run and results\n-------------------------\n" %
(data["metadata"]["date"],data["metadata"]["username"]))
logout.write("Initial Parameters\n------------------\nimage size:\t%sx%s\nedge
res:\t%s\nh,k,l range:\t%s,%s,%s\nres cutoff:\t%s\n\n" %
(params["imgsize"],params["imgsize"],params["imgmaxres"],params["hrange"],params["krange"
],params["lrange"],params["reslimit"]))
now = datetime.datetime.now()
logout.write("1_find_lengths \t%s\n" % now.strftime("%Y-%m-%d %H:%M"))

os.system('clear')

#### version check
vers = 1
catspotvers = data["metadata"]["file version"]
print "** Rough Unit Cell Determination vers %s **" % round(vers,2)
if vers != catspotvers:
    print "datafile version mismatch %s/%s - may cause errors" %
(round(vers,2),round(catspotvers,2))
else:
    print "version check passed"
##### do it

print ""
print "make the spolist dictionarys"
spotlist= []   # input raw spot number returns numpy array with vector
xydic = {}  # original x,y coords to image#,spot#
xyzdic = {} # calculated x,y,z coords to image#,spot#

with open('imagelist.txt') as f:
    images2process = f.read().splitlines()

spotrange = []
for eachimage in images2process:
    theta = float(data["data"]["images"][eachimage]["tiltangle"])
```

```
    for i in data["data"]["images"][eachimage]["spots"]:
        ox = i[0]
        oy = i[1]
        x = i[0] - data["data"]["images"][eachimage]["beamcenter"][0][0]
        y = -(oy -
data["data"]["images"][eachimage]["beamcenter"][0][1])*math.cos(theta*math.pi/180.0)
        z = -(oy -
data["data"]["images"][eachimage]["beamcenter"][0][1])*math.sin(theta*math.pi/180.0)
        spotrange.append(numpy.array([x,y,z]))

print "build spotlist: %s spots picked" % len(spotrange)

#### calculate the difference vectors for all combinations of spots

print"subtract every vector from every other"
diffvectors = []
for n in spotrange:
     for i in spotrange:
        diff = numpy.subtract(n,i)
        if diff[0] and diff[1] and diff[2] != 0:
            diffvectors.append(diff)

print "%s difference vectors found" % len(diffvectors)

### count the distribution of the difference vectors

count = Counter()
for i in diffvectors:
    mag = round(numpy.linalg.norm(i),1)
    if mag != 0:
        count[mag] += 1

ymax = float(max(count.values()))/2+5

for i in count:
    if 0 < i < 250:
        output.write("%s %s\n" % (i,count[i]/2))


########  make the plot
os.system("touch outgraph.eps")
print "drawing plot"
drawplot.write("set terminal postscript eps enhanced color font 'Verdana,10'\n")
drawplot.write("set yrange [0:%s]\n" % ymax)
drawplot.write('set title "Unit Cell Predictions" font "Arial,12"\n')
drawplot.write("set output 'outgraph.eps'\n")
drawplot.write("set style line 1 lc rgb '#0060ad' lt 1 lw 2 pt 7 ps 1.5\n")
drawplot.write("plot 'magfreqs.csv' with impulses\n")

os.system("gnuplot plot")
os.system("open outgraph.eps")
```

A1.2: calc_ucvectors.py

Calculates up to 12 candidate unit cell vectors based on predicted unit cell lengths from 1_find_lengths.py. Outputs the candidate vectors and a matrix of all the angles between candidate vectors.

```
#! /usr/bin/env python
## imports

## Finds the average unit cell vector using cataspot.json which is created by cataspot

# Matt Iadanza 20-07-05

import json
import os
import math
import numpy
import datetime
```

```
#### user input variables


ea = 55.0       # expected a unit cell length
athresh = 1
eb = 55.0       # expected b unit cell length
bthresh = 1
ec = 112.0      # expected c dimension
cthresh = 1


#### get yer data:
data = json.load(open('cataspot.json'))
numberofimages = len(data["data"]["images"])
output = open("output_cellfind.txt", "w")
params = json.load(open('params.json'))
imgsize = params['imgsize']
maxres = params['imgmaxres']

### open logfile
logout = open("logfile.txt", "a")
now = datetime.datetime.now()
logout.write("\n2_calc_uc vectors \t%s\n" % now.strftime("%Y-%m-%d %H:%M"))
logout.write("vector lengths(a,b,c): %s,%s,%s    thresholds(a,b,c): %s,%s,%s\n" %
(ea,eb,ec,athresh,bthresh,cthresh))

#### init and version check
os.system('clear')
vers = 1
catspotvers = data["metadata"]["file version"]
print "** Unit Cell Vector Determination vers %s **" % round(vers,2)
if vers != catspotvers:
    print "datafile version mismatch %s/%s - may cause errors" %
(round(vers,2),round(catspotvers,2))
else:
    print "version check -- passed"

##### do it

print ""
print "make the spolist dictionarys"
with open('imagelist.txt') as f:
    images2process = f.read().splitlines()

### calculate 3-D coordinates for each spot

##------- ewald sphere correction (z dimension) function----------
def ewaldcorr(xdim,ydim):
    wavelength = 0.025
    oneoverlambda = 1/(wavelength * (1/(0.5 * imgsize * maxres)))
    a = oneoverlambda/math.sqrt(xdim**2+ydim**2+oneoverlambda**2)
    deltaz = (1-a)*oneoverlambda
    return deltaz
##-------------------------------------------------------------------

spotlist = []
for eachimage in images2process:
    theta = float(data["data"]["images"][eachimage]["tiltangle"])
    for i in data["data"]["images"][eachimage]["spots"]:
        ox = i[0]
        oy = i[1]
        x = i[0] - data["data"]["images"][eachimage]["beamcenter"][0][0]
        y = -(oy -
data["data"]["images"][eachimage]["beamcenter"][0][1])*math.cos(theta*math.pi/180.0)
        z = -(oy -
data["data"]["images"][eachimage]["beamcenter"][0][1])*math.sin(theta*math.pi/180.0) -
ewaldcorr(x,y)
        spotlist.append(numpy.array([x,y,z]))

print "build spotlist: %s spots picked" % len(spotlist)
logout.write("%s spots\n" % len(spotlist))
```

```
#### calculate difference vectors and sort out those that could be unit cells
print"subtract every vector from every other and determine the magnitude of results -
keep possible unit cell vectors\n"

poucvs = []
for n in spotlist:
    for i in spotlist:
        diffvec = numpy.subtract(n,i)
        sub = numpy.linalg.norm(diffvec)
        if i[0] != n[0] and i[1] != n[1] and i[2] != n[2] and ((ea-athresh < sub <
ea+athresh)or(eb-bthresh < sub < eb+bthresh)or(ec-cthresh < sub < ec+cthresh)):
            poucvs.append(diffvec)

print "%s possible unit cell vectors found" % len(poucvs)

### sort the unit cell vectors into parallel groups
### compare all vectors to xy normal test vectors
### use results to sort into roughly (within threshold) parallel groups

print "sorting unit cell vectors into parallel groups"

testvector1 = numpy.array([0,0,10000])
testvector2= numpy.array([0,10000,0])
testvector3= numpy.array([10000,0,0])
testvector4 = numpy.array([5000,5000,5000])
oset1 =  []
oset2 = []
oset3 = []


##----------- function to calculate the angle betwen two vectors ----------

def calcang(a,b):
    v12 = numpy.dot(a,b)
    v1mag = numpy.linalg.norm(a)
    v2mag = numpy.linalg.norm(b)
    cosphi = abs((v12)/(v1mag*v2mag))
    return round((180/math.pi)*math.acos(round(cosphi,12)),0)
##--------------------------------------------------------------------------

##----------- function to calculate the angle betwen two vectors - nonabsoulte ----------

def calcangnonabs(a,b):
    v12 = numpy.dot(a,b)
    v1mag = numpy.linalg.norm(a)
    v2mag = numpy.linalg.norm(b)
    cosphi = ((v12)/(v1mag*v2mag))
    return round((180/math.pi)*math.acos(round(cosphi,12)),0)
##--------------------------------------------------------------------------

### initial grouping for parallel vectors########
group01 = []
group02 = []
group03 = []
group10 = []
group12 = []
group13 = []
group20 = []
group21 = []
group23 = []
group30 = []
group31 = []
group32 = []
################################################

## identify which reference vectro each is closest to and furthest from (in terms of
angles) use to classify into roughly parallel groups
orientations = {}
for i in poucvs:
```

```
    orientations[(i[0],i[1],i[2])] =
[calcang(i,testvector1),calcang(i,testvector2),calcang(i,testvector3),calcang(i,testvecto
r4)]

for i in orientations:
    scores = [orientations[i][0],
orientations[i][1],orientations[i][2],orientations[i][3]]
    maxscore = scores.index(max(scores))
    minscore = scores.index(min(scores))
    i = numpy.array(i)
    if sum(numpy.cross(i,testvector1)) < 0:
            i = numpy.multiply(-1,i)
    globals()['group'+str(maxscore)+str(minscore)].append(i)



##### refine the groups - only keep vectors within 1 STD of eth mean for each of th four
reference vectors
#### goodgroups
good01 = []
good02 = []
good03 = []
good10 = []
good12 = []
good13 = []
good20 = []
good21 = []
good23 = []
good30 = []
good31 = []
good32 = []

### ---------group calc function-------------------------------------------------->>>>>
def groupcalc(x,y):

    ang1 = []
    ang2 = []
    ang3 = []
    ang4 = []



    for i in x:
            ang1.append(calcangnonabs(i,testvector1))
            ang2.append(calcangnonabs(i,testvector2))
            ang3.append(calcangnonabs(i,testvector3))
            ang4.append(calcangnonabs(i,testvector4))



    for i in x:
        score1 = int(abs(numpy.mean(ang1)-
calcangnonabs(i,testvector1))/(numpy.std(ang1)+.00001))
        score2 = int(abs(numpy.mean(ang2)-
calcangnonabs(i,testvector2))/(numpy.std(ang2)+.00001))
        score3 = int(abs(numpy.mean(ang3)-
calcangnonabs(i,testvector3))/(numpy.std(ang3)+.00001))
        score4 = int(abs(numpy.mean(ang4)-
calcangnonabs(i,testvector4))/(numpy.std(ang4)+.00001))


        if score1+score2+score3+score4 < 1:
            globals()['good'+str(y)].append(i)

###-------------------------------------------------------------------------------->>>>>


scaled01 = []
scaled02 = []
scaled03 = []
```

```
scaled10 = []
scaled12 = []
scaled13 = []
scaled20 = []
scaled21 = []
scaled23 = []
scaled30 = []
scaled31 = []
scaled32 = []




###### output everything to screen ####
## make raw output file later?

## rough groups
for i in ('01','02','03','10','12','13','20','21','23','30','31','32'):
    if len(globals()['group'+i]) > 0:
        groupcalc(globals()['group'+i],i)

####### scale the vectors


##----calc minimum funct---------------
def calcmin(x):
    allmags = []
    for vector in x:
        allmags.append(numpy.linalg.norm(vector))
    return min(allmags)
##--------------------------------------

for i in ('01','02','03','10','12','13','20','21','23','30','31','32'):
    for x in globals()['good'+i]:
        if numpy.linalg.norm(x) > 2*calcmin(globals()['good'+i]):
            globals()['scaled'+i].append(x/2)
        else:
            globals()['scaled'+i].append(x)

## -- list printing function------
def printlist(x):
    print "-----"
    for i in x:
        print i
##-----------------------------

##--------- funct to calculate mean vectors -----------------
def vecmean(x):
    if len(x) > 0:
        iss = []
        js = []
        ks = []
        for i in x:
            iss.append(i[0])
            js.append(i[1])
            ks.append(i[2])
        mean = numpy.array([numpy.mean(iss),numpy.mean(js),numpy.mean(ks)])
        return [mean, numpy.linalg.norm(mean)]
    else:
        return ["no vector","n/a"]
##-------------------------------------------------------


candidates = []
## final mean vectors
print '\n------- Candidate vectors -------'
for i in ('01','02','03','10','12','13','20','21','23','30','31','32'):
    if len(globals()['good'+i]) > 2:
        print"%s\t%s  \tmag: %s  \tscore: %s" %
(i,vecmean(globals()['scaled'+i])[0],vecmean(globals()['scaled'+i])[1],
len(globals()['scaled'+i]))
```

```
        logout.write("%s\t%s  \tmag: %s  \tscore: %s\n" %
(i,vecmean(globals()['scaled'+i])[0],vecmean(globals()['scaled'+i])[1],
len(globals()['scaled'+i])))
        candidates.append([i,vecmean(globals()['scaled'+i])[0]])
print"\n------- Angles Between Candidates -------"
print "\t",
logout.write("\t")
for i in candidates:
    print str(i[0])+"\t",
    logout.write(str(i[0])+"\t")
print ""
logout.write("\n")
for i in candidates:
    anglist = []
    for n in candidates:
        anglist.append(str(calcangnonabs(i[1],n[1])))
    print "%s\t%s" % (i[0],'\t'.join(anglist))
    logout.write("%s\t%s\n" % (i[0],'\t'.join(anglist)))
```

A1.3: spots_index.py

Predicts diffraction spots on all images based on eth vectors calculated bu 2_calc_ucvectors. Draws illustrations showing the positions of predicted spots for review.

```
#! /usr/bin/env python

## based off ts2_mod.py - modified to use Cataspot and new global parameters file.
# Matt Iadanza 2013-09-16

# predict spots in all images (including non major plain images) based on the unit cell
vectors produced by findcell.py and refined using findparallel.py
# outputs batch#.csv that is used by revcoords_tf.py to produce an illustration of the
spots.  Uses Brent's new (more accurate) matrix method
#
# Matt Iadanza 2013-07-08

#### Imports:

import numpy
import math
import os
import json
import datetime


########### Variables

params = json.load(open('params.json'))
imgsize = params["imgsize"]
boxsize =  params["circrad"]
imgmaxres = params["imgmaxres"]
reslimit = params["reslimit"]
hvals = range(-params["hrange"], params["hrange"]+1)
kvals = range(-params["krange"], params["krange"]+1)
lvals = range(-params["lrange"], params["lrange"]+1)
a = numpy.array(params["aucvec"])
b = numpy.array(params["bucvec"])
c = numpy.array(params["cucvec"])
maxresrad = (imgmaxres/reslimit)*(0.5*imgsize)
circrad =  params["circrad"]

### open logfile
logout = open("logfile.txt", "a")
now = datetime.datetime.now()
logout.write("\n3_spot_index \t%s\n" % now.strftime("%Y-%m-%d %H:%M"))
logout.write("vectors:\t%s\n\t\t%s\n\t\t%s\n" % (a,b,c))


################ build the list of all the matrix of hk coords
```

```
hklindices = []
for eachhval in kvals:
    for eachkval in hvals:
        for eachlval in lvals:
            hklindices.append((eachhval, eachkval, eachlval))


###############  Get the images to process from the list

with open('imagelist.txt') as pfile:
    imagestoprocess = pfile.read().splitlines()
print "processing "+str(len(imagestoprocess))+" images\n"


data = json.load(open("cataspot.json"))
for eachimage in imagestoprocess:
    spots2draw = []
## get the image specific parametrs

    imgsplit = eachimage.split(".")
    imagename = imgsplit[0]+".gif"
    output = file(imgsplit[0]+"_spots.csv", "w")

###### calculate the vectors for the reference spots and their dot product:

    xcenter = data["data"]["images"][eachimage]["beamcenter"][0][0]
    ycenter = data["data"]["images"][eachimage]["beamcenter"][0][1]
    theta =float(data["data"]["images"][eachimage]["tiltangle"])
    refpoints = data["data"]["images"][eachimage]["references"]
    bscentx = data["data"]["images"][eachimage]["beamstopcenter"][0][0]
    bscenty = data["data"]["images"][eachimage]["beamstopcenter"][0][1]
    bsrtopy = bscenty-100
    bsrboty = bscenty+100
    bsrad = params["beamstoprad"]
    imagename = imgsplit[0]+".gif"

    refvectors = []
    for i in refpoints:
        x = i[0] - xcenter
        y = (ycenter - i[1])*math.cos(theta*math.pi/180)
        z = (ycenter - i[1])*math.sin(theta*math.pi/180)
        refvectors.append(numpy.array([x,y,z]).reshape(3,1))

## make the unit matrix

    unitmatrix = numpy.array([[a[0],b[0],c[0]],[a[1],b[1],c[1]],[a[2],b[2],c[2]]])
    uinv =  numpy.linalg.inv(unitmatrix)
    r1 = numpy.dot(uinv, refvectors[0])
    r2 = numpy.dot(uinv, refvectors[1])

    r1round =
numpy.array([int(round(r1.item((0,0)),0)),int(round(r1.item((1,0)),0)),int(round(r1.item(
(2,0)),0))])
    r2round =
numpy.array([int(round(r2.item((0,0)),0)),int(round(r2.item((1,0)),0)),int(round(r2.item(
(2,0)),0))])

    checkplane = numpy.cross(r1round,r2round)
    print "%s\t" % eachimage,

## test every hkl index
    spots = []
    for i in hklindices:
        hkltest = numpy.array([i[0]*checkplane[0],i[1]*checkplane[1],i[2]*checkplane[2]])

        if  -0.25*numpy.linalg.norm(checkplane) <
int(hkltest[0])+int(hkltest[1])+int(hkltest[2]) < numpy.linalg.norm(checkplane)*0.25:
            spots.append((i[0],i[1],i[2]))
######## calculate the xyz coordinates of all of miller indices
    xydic = {}          # miller index returns point xy
    xyzvectors = {}        # miller index returns x,y,z vector
```

```
    for i in spots:
        x = (i[0]*a[0])+(i[1]*b[0])+(i[2]*c[0])
        y = (i[0]*a[1])+(i[1]*b[1])+(i[2]*c[1])
        z = (i[0]*a[2])+(i[1]*b[2])+(i[2]*c[2])
        xyzvectors[i] = numpy.array([x,y,z])

        xproj = x + xcenter
        yproj = ycenter - (y/(math.cos(theta*(math.pi/180))))
        xydic[i] = (round(xproj,1),round(yproj,1))
        if xydic[i][1] > 0 and xydic[i][0] > 0:
            if math.sqrt((math.ceil(xydic[i][0]) - xcenter)**2 + (math.ceil(xydic[i][1])
- ycenter)**2) < maxresrad:
                if math.ceil(xydic[i][0]) < bscentx+bsrad or (math.ceil(xydic[i][0]) >
bscentx+bsrad and (bsrtopy-circrad > math.ceil(xydic[i][1]) or math.ceil(xydic[i][1]) >
bsrboty+circrad)):
                    if (math.ceil(xydic[i][0]) - bscentx)**2 + (math.ceil(xydic[i][1])-
bscenty)**2 > (bsrad**2)+circrad:
                        output.write("%s,%s,%s\t%s,%s\t%s\n" %
(i[0],i[1],i[2],xydic[i][0],xydic[i][1],math.ceil(xydic[i][0])+(4096*math.ceil(xydic[i][1
])))))
                        spots2draw.append((xydic[i][0],xydic[i][1]))
### draw the images:

    imagename = imgsplit[0]+".gif"
    csvfile = imgsplit[0]+"_spots.csv"
    spotlist = []
    for i in spots2draw:
        spotlist.append("circle %s,%s %s,%s" % (i[0],i[1],float(i[0])+circrad,
float(i[1])+circrad))

    print "\t%s spots" % len(spotlist)
    command  = ('convert -size 4096x4096 xc:Black -font Arial -pointsize 10 -stroke Red -
strokewidth 2 -fill none -transparent black -draw "%s" +compress points_px.tif' % '
'.join(spotlist))
    os.system('%s' % command)
    os.system("convert %s %s -gravity center -compose over -composite %s" % (imagename,
"points_px.tif",imgsplit[0]+"_spots.gif"))
```

A1.4: refine_spots.py

Refines spots by mass centering and draws illustrations showing the positions of the original and refined spots.

```
#! /usr/bin/env python

### use mass centering to find centers of spots

## Matt Iadanza 2013-07-23

#### imports:
import json
from PIL import Image
import os
import csv
import math
import numpy
import json
import datetime

### get yer data

with open('imagelist.txt') as pfile:
    imagestoprocess = pfile.read().splitlines()
print "processing "+str(len(imagestoprocess))+" images"

data = json.load(open('cataspot.json'))
params = json.load(open('params.json'))

boxsize = params["circrad"]
fullbox = 2*boxsize
drawlist = []
```

```
drawlist2 = []

### open logfile
logout = open("logfile.txt", "a")
now = datetime.datetime.now()
logout.write("\n4_refine_spots \t%s\n" % now.strftime("%Y-%m-%d %H:%M"))


for eachimage in imagestoprocess:
    print eachimage
    imgsplit = eachimage.split('.')
    spots = []
    split = []
    imageroot = imgsplit[0]
    pic = Image.open(eachimage)
    values = csv.reader(open(imageroot+'_spots.csv', 'rb'), delimiter='\t')
    output = open(imageroot+"_spots_r.csv", "w")
    rawout = open(imageroot+"_spotrefinement_raw.txt", "w")
    rawout.write("spot\t\toriginal xy\tnewxy\t\t\tdisplacement\n")
    for i in values:
        split = i[1].split(",")
        spots.append([float(split[0]),float(split[1]),i[0]])
    rootname = eachimage.split('.')

## for each spot make a spot sub array
    for eachspot in spots:
        spotbox = (int(float(eachspot[0])-boxsize),int(float(eachspot[1])-
boxsize),int(float(eachspot[0])+boxsize),int(float(eachspot[1])+boxsize))
        spotregion = pic.crop(spotbox)
        spot = numpy.array(spotregion.getdata()).reshape(spotregion.size[0],
spotregion.size[1])

## calculate the background mask
        zeros = numpy.zeros((fullbox,fullbox), numpy.int)


        maskcoords = []
        maskrange = range(0,boxsize+1)
        center = [boxsize,boxsize]
        for i in maskrange:
            for n in maskrange:
                if math.sqrt((boxsize-i)**2 + (boxsize-n)**2) > boxsize:
                    maskcoords.append([i,n])
                    maskcoords.append([(2*boxsize-1)-i,n])
                    maskcoords.append([i,(2*boxsize-1)-n])
                    maskcoords.append([(2*boxsize-1)-i,(2*boxsize-1)-n])
        for i in maskcoords:
            zeros[i[0],i[1]] = 1
        bgmask = numpy.multiply(zeros,spot)
        bgmean = numpy.mean(bgmask[numpy.nonzero(bgmask)])

### calculate the spot mask
        zeros = numpy.zeros((fullbox,fullbox), numpy.int)


        maskcoords = []
        maskrange = range(0,boxsize+1)
        center = [boxsize,boxsize]
        for i in maskrange:
            for n in maskrange:
                if math.sqrt((boxsize-i)**2 + (boxsize-n)**2) < boxsize:
                    maskcoords.append([i,n])
                    maskcoords.append([(2*boxsize-1)-i,n])
                    maskcoords.append([i,(2*boxsize-1)-n])
                    maskcoords.append([(2*boxsize-1)-i,(2*boxsize-1)-n])
        for i in maskcoords:
            zeros[i[0],i[1]] = 1
            spotmask = numpy.multiply(zeros,spot)

## calculate center score
        rows = spot.sum(axis = 0)
        cols = spot.sum(axis = 1)
        displacement = (rows.argmax(axis = 0),cols.argmax(axis = 0))
```

```
        dispvalues = (float(displacement[0]-boxsize), float(displacement[1]-boxsize))

## calculate the new xy
        newxy = []
        newxy = [eachspot[0]+dispvalues[0],eachspot[1]+dispvalues[1]]
        drawlist.append("circle %s,%s %s,%s" %
(float(eachspot[0]),float(eachspot[1]),float(eachspot[0]),float(eachspot[1])+boxsize))
        drawlist2.append("circle %s,%s %s,%s" %
(float(newxy[0]),float(newxy[1]),float(newxy[0]),float(newxy[1])+boxsize))

        rawout.write("%11s\t%s,%s\t%s  \t%s\n" %
(eachspot[2],eachspot[0],eachspot[1],newxy,dispvalues))

        output.write("%s\t%s,%s\n" % (eachspot[2],newxy[0],newxy[1]))

#### draw the images
    command1  = ('convert -size 4096x4096 xc:Black -stroke Blue -strokewidth 1 -fill none
-transparent black -draw "%s" points_old.gif' % ' '.join(drawlist))
    command2 = ('convert -size 4096x4096 xc:Black -stroke Yellow -strokewidth 1 -fill
none -transparent black -draw "%s" points_new.gif' % ' '.join(drawlist2))
    os.system('%s' % command1)
    os.system('%s' % command2)
    drawlist = []
    drawlist2 = []

    os.system("convert %s %s -gravity center -compose over -composite %s" %
(imageroot+".gif", "points_old.gif",imageroot+"_refine_spots.gif"))
    os.system("convert %s %s -gravity center -compose over -composite %s" %
(imageroot+"_refine_spots.gif", "points_new.gif",imageroot+"_refine_spots.gif"))

## cleanup

os.system('rm points_old.gif')
os.system('rm points_new.gif')
```

A1.5: UCR_index.py

Indexes all diffraction patterns using a high threshold for spot intensity. Generates a spot list used by 6_recalculate_vectors.py for unit cell vector calculation and draws illustrations showing the chosen spots for review.

```
#!/usr/bin/env python

###################


## refines unit cell vectors based on spot centering...
## matt iadanza 2013-09-26

## modifications of integrate2_batch.py so it can handle data from cataspot.
## Matt Iadanza 2013-09-16

# findsw intensity or spots specified in a csv file.  Version 2, trying to get more
accurate intensities, wih improved
# background subtraction and some sor of thresholding

# Matt Iadanza 2013-07-22

################### Variables

sqthresh = 1.15     # spot quality threshold % of mean background
centererrorthresh = 3


###################imports
from PIL import Image
import os
import csv
import math
import numpy
import json
```

```
import datetime

## get global parameters from parfile.

data = json.load(open('cataspot.json'))
params = json.load(open('params.json'))
imgsize = params["imgsize"]
boxsize =  params["circrad"]
radius = range(1,boxsize+1)
fullbox = 2*boxsize
output = open("refined.txt","w")


### open logfile
logout = open("logfile.txt", "a")
now = datetime.datetime.now()
logout.write("\n5_UCR_index \t%s\n" % now.strftime("%Y-%m-%d %H:%M"))
logout.write("Spot intensity threshold: %sx background\tCenter error threshold: %s px\n"
% (sqthresh,centererrorthresh))


#### init and version check
os.system('clear')
vers = 1
catspotvers = data["metadata"]["file version"]
print "** Indexing for Unit Cell Vector Determination vers %s **" % round(vers,2)
if vers != catspotvers:
    print "datafile version mismatch %s/%s - may cause errors" %
(round(vers,2),round(catspotvers,2))
else:
    print "version check -- passed"



## get list of images to process

with open('imagelist.txt') as pfile:
    imagestoprocess = pfile.read().splitlines()
print "processing "+str(len(imagestoprocess))+" images"

################### Start processing images

for eachimage in imagestoprocess:
    qualityspots = {}       # input x,y return intensity
    theta =float(data["data"]["images"][eachimage]["tiltangle"])

## get the image specific parametrs

    imgsplit = eachimage.split(".")
    spotthreshold = params["integrationthresh"]
    rawoutput = open(imgsplit[0]+"_rawint.txt", "w")

## open the image:

    pic = Image.open(eachimage)

## prepare to process each spot

    values = csv.reader(open(imgsplit[0]+'_spots_r.csv', 'rb'), delimiter='\t')
    indexdic = {}   # input xy return miller index
    coordsdic = {}  # input miller index return xy
    for row in values:
        indexdic[row[1]] = row[0]
        coordsdic[row[0]] = row[1]

## for each spot make a spot sub array
    xy = []
    for eachspot in indexdic:
        xy = eachspot.split(",")
        spotbox = (int(float(xy[0])-boxsize),int(float(xy[1])-
boxsize),int(float(xy[0])+boxsize),int(float(xy[1])+boxsize))
        spotregion = pic.crop(spotbox)
```

```
        spot = numpy.array(spotregion.getdata()).reshape(spotregion.size[0],
spotregion.size[1])
        rawoutput.write( "\nRaw spot %s\n" % indexdic[eachspot])
        rawoutput.write( '\n'.join('\t'.join(str(cell) for cell in row) for row in spot))

## calculate the background mask
        zeros = numpy.zeros((fullbox,fullbox), numpy.int)

        rawoutput.write( "\n-bgmatrix-\n")

        maskcoords = []
        maskrange = range(0,boxsize+1)
        center = [boxsize,boxsize]
        for i in maskrange:
            for n in maskrange:
                if math.sqrt((boxsize-i)**2 + (boxsize-n)**2) > boxsize:
                    maskcoords.append([i,n])
                    maskcoords.append([(2*boxsize-1)-i,n])
                    maskcoords.append([i,(2*boxsize-1)-n])
                    maskcoords.append([(2*boxsize-1)-i,(2*boxsize-1)-n])
        for i in maskcoords:
            zeros[i[0],i[1]] = 1
        bgmask = numpy.multiply(zeros,spot)
        rawoutput.write( '\n'.join('\t'.join(str(cell) for cell in row) for row in
bgmask))
        bgmean = numpy.mean(bgmask[numpy.nonzero(bgmask)])
        rawoutput.write( "\nmean BG intensity: %s\n" % bgmean)

### calculate the spot mask and quality test spots
        zeros = numpy.zeros((fullbox,fullbox), numpy.int)

        rawoutput.write("\n-raw spotmatrix-\n")

        maskcoords = []
        maskrange = range(0,boxsize+1)
        center = [boxsize,boxsize]
        for i in maskrange:
            for n in maskrange:
                if math.sqrt((boxsize-i)**2 + (boxsize-n)**2) < boxsize:
                    maskcoords.append([i,n])
                    maskcoords.append([(2*boxsize-1)-i,n])
                    maskcoords.append([i,(2*boxsize-1)-n])
                    maskcoords.append([(2*boxsize-1)-i,(2*boxsize-1)-n])
        for i in maskcoords:
            zeros[i[0],i[1]] = 1
        spotmask = numpy.multiply(zeros,spot)
        rawoutput.write( '\n'.join('\t'.join(str(cell) for cell in row) for row in
spotmask))

## calculate center score
        rows = spot.sum(axis = 0)
        cols = spot.sum(axis = 1)
        displacement = (rows.argmax(axis = 0),cols.argmax(axis = 0))
        dispvalues = (displacement[0]-boxsize, displacement[1]-boxsize)
        dispscore = abs(dispvalues[0])+abs(dispvalues[1])

### background subtract
        bgsubmatrix = numpy.ndarray(shape=(2*boxsize,2*boxsize), dtype=float)
        bgsubmatrix.fill(bgmean)
        rawoutput.write( "\n-BG subtracted spotmatrix-\n")
        subtracted = numpy.subtract(spot,bgsubmatrix)
        subtractedmasked = numpy.multiply(subtracted,zeros)
        rawoutput.write( '\n'.join('\t'.join(str(cell) for cell in row) for row in
subtractedmasked.round(0)))
        rawoutput.write("\nspot: %s\t i: %s quality: %s\n\n" %
(indexdic[eachspot],numpy.sum(subtractedmasked),numpy.mean(spotmask[numpy.nonzero(spotmas
k)])/bgmean))
        if numpy.mean(spotmask[numpy.nonzero(spotmask)])/bgmean >= sqthresh and dispscore
<= centererrorthresh:
            qualityspots[eachspot] = numpy.sum(subtractedmasked)
            hkl = indexdic[eachspot].split(",")
```

```
            print "index:%10s\t intensity: %14s \tquality: %14s center error: %s : %s" %
(indexdic[eachspot],numpy.sum(subtractedmasked),numpy.mean(spotmask[numpy.nonzero(spotmas
k)])/bgmean,dispvalues,dispscore)

## print image summary, write data to output,and draw the picture

    rawoutput.write( "- %s image summary -\n" % eachimage)
    rawoutput.write( "%s/%s spots kept - threshold: %s\n" %
(len(qualityspots),len(indexdic),sqthresh))

    print ""
    print "- %s image summary -" % eachimage
    print "%s/%s spots kept - threshold: %s" % (len(qualityspots),len(indexdic),sqthresh)

    drawlist = []
    for spot in qualityspots:
        output.write("%s\t%s\t%s\t%s\n" % (xy[0],xy[1],theta,eachimage))
        xy = []
        xy = spot.split(',')
        drawlist.append("circle %s,%s %s,%s" %
(float(xy[0]),float(xy[1]),float(xy[0]),float(xy[1])+boxsize))



    command  = ('convert -size 4096x4096 xc:Black -stroke Yellow -strokewidth 2 -fill
none -transparent black -draw "%s" points_px.gif' % ' '.join(drawlist))
    os.system('%s' % command)
    os.system("convert %s %s -gravity center -compose over -composite %s" %
(imgsplit[0]+".gif", "points_px.gif",imgsplit[0]+"_ucr_spots.gif"))

## cleanup

    os.system('rm points_px.gif')
```

A1.6: recalculate_vectors.py

Calculates unit cell vectors based on complete indexing of all diffraction patterns performed by 5_UCR_index.py

```
#! /usr/bin/env python
## imports

## Finds the average unit cell vector using eth refined spots generated by the first four
setps
## in these programs.

# Matt Iadanza 20-07-05

import os
import math
import numpy
import json
import datetime

#### user input variables



####  some user entered varibales

ea = 55.0       # expected a unit cell length
athresh = 1
eb = 55.0       # expected b unit cell length
bthresh = 1
ec = 112.0      # expected c dimension
cthresh = 1


#### get yer data:
data = json.load(open('cataspot.json'))
```

```
numberofimages = len(data["data"]["images"])
params = json.load(open('params.json'))
imgsize = params['imgsize']
maxres = params['imgmaxres']

### open logfile
logout = open("logfile.txt", "a")
now = datetime.datetime.now()
logout.write("\n6_recalculate_vectors \t%s\n" % now.strftime("%Y-%m-%d %H:%M"))
logout.write("vector lengths(a,b,c): %s,%s,%s    thresholds(a,b,c): %s,%s,%s\n" %
(ea,eb,ec,athresh,bthresh,cthresh))


#### init and version check
os.system('clear')
vers = 1
catspotvers = data["metadata"]["file version"]
print "** Unit Cell Vector Determination vers %s **" % round(vers,2)
if vers != catspotvers:
    print "datafile version mismatch %s/%s - may cause errors" %
(round(vers,2),round(catspotvers,2))
else:
    print "version check -- passed"

##### do it

print ""
print "make the spolist dictionarys"
with open('imagelist.txt') as f:
    images2process = f.read().splitlines()
with open('refined.txt') as spotfile:
    allspots = spotfile.read().splitlines()

### calculate 3-D coordinates for each spot

##------- ewald sphere correction (z dimension) function----------
def ewaldcorr(xdim,ydim):
    wavelength = 0.025
    oneoverlambda = 1/(wavelength * (1/(0.5 * imgsize * maxres)))
    a = oneoverlambda/math.sqrt(xdim**2+ydim**2+oneoverlambda**2)
    deltaz = (1-a)*oneoverlambda
    return deltaz
##--------------------------------------------------------------------


spotlist = []
for eachspot in allspots:
    i = eachspot.split('\t')
    theta = float(i[2])
    ox = float(i[0])
    oy = float(i[1])
    x = ox - data["data"]["images"][i[3]]["beamcenter"][0][0]
    y = -(oy -
data["data"]["images"][i[3]]["beamcenter"][0][1])*math.cos(theta*math.pi/180.0)
    z = -(oy -
data["data"]["images"][i[3]]["beamcenter"][0][1])*math.sin(theta*math.pi/180.0) -
ewaldcorr(x,y)
    spotlist.append(numpy.array([x,y,z]))


print "build spotlist: %s spots picked" % len(spotlist)
logout.write("%s spots\n" % len(spotlist))

#### calculate difference vectors and sort out those that could be unit cells
print"subtract every vector from every other and determine the magnitude of results -
keep possible unit cell vectors\n"

poucvs = []
for n in spotlist:
    for i in spotlist:
        diffvec = numpy.subtract(n,i)
        sub = numpy.linalg.norm(diffvec)
```

```
        if i[0] != n[0] and i[1] != n[1] and i[2] != n[2] and ((ea-athresh < sub <
ea+athresh)or(eb-bthresh < sub < eb+bthresh)or(ec-cthresh < sub < ec+cthresh)):
            poucvs.append(diffvec)

print "%s possible unit cell vectors found" % len(poucvs)

### sort the unit cell vectors into parallel groups
### compare all vectors to xy normal test vectors
### use results to sort into roughly (within threshold) parallel groups

print "sorting unit cell vectors into parallel groups"

testvector1 = numpy.array([0,0,10000])
testvector2= numpy.array([0,10000,0])
testvector3= numpy.array([10000,0,0])
testvector4 = numpy.array([5000,5000,5000])
oset1 =  []
oset2 = []
oset3 = []


##----------- function to calculate the angle betwen two vectors ----------

def calcang(a,b):
    v12 = numpy.dot(a,b)
    v1mag = numpy.linalg.norm(a)
    v2mag = numpy.linalg.norm(b)
    cosphi = abs((v12)/(v1mag*v2mag))
    return round((180/math.pi)*math.acos(round(cosphi,12)),0)
##-------------------------------------------------------------------------

##----------- function to calculate the angle betwen two vectors - nonabsoulte ----------

def calcangnonabs(a,b):
    v12 = numpy.dot(a,b)
    v1mag = numpy.linalg.norm(a)
    v2mag = numpy.linalg.norm(b)
    cosphi = ((v12)/(v1mag*v2mag))
    return round((180/math.pi)*math.acos(round(cosphi,12)),0)
##-------------------------------------------------------------------------



### initial grouping for parallel vectors########
group01 = []
group02 = []
group03 = []
group10 = []
group12 = []
group13 = []
group20 = []
group21 = []
group23 = []
group30 = []
group31 = []
group32 = []
################################################

## identify which reference vectro each is closest to and furthest from (in terms of
angles) use to classify into roughly parallel groups
orientations = {}
for i in poucvs:

    orientations[(i[0],i[1],i[2])] =
[calcang(i,testvector1),calcang(i,testvector2),calcang(i,testvector3),calcang(i,testvecto
r4)]

for i in orientations:
    scores = [orientations[i][0],
orientations[i][1],orientations[i][2],orientations[i][3]]
    maxscore = scores.index(max(scores))
    minscore = scores.index(min(scores))
    i = numpy.array(i)
```

```
    if sum(numpy.cross(i,testvector1)) < 0:
            i = numpy.multiply(-1,i)
    globals()['group'+str(maxscore)+str(minscore)].append(i)



##### refine the groups - only keep vectors within 1 STD of eth mean for each of th four
reference vectors
#### goodgroups
good01 = []
good02 = []
good03 = []
good10 = []
good12 = []
good13 = []
good20 = []
good21 = []
good23 = []
good30 = []
good31 = []
good32 = []

### ---------group calc function-------------------------------------------------->>>>>
def groupcalc(x,y):

    ang1 = []
    ang2 = []
    ang3 = []
    ang4 = []



    for i in x:
            ang1.append(calcangnonabs(i,testvector1))
            ang2.append(calcangnonabs(i,testvector2))
            ang3.append(calcangnonabs(i,testvector3))
            ang4.append(calcangnonabs(i,testvector4))



    for i in x:
        score1 = int(abs(numpy.mean(ang1)-
calcangnonabs(i,testvector1))/(numpy.std(ang1)+.00001))
        score2 = int(abs(numpy.mean(ang2)-
calcangnonabs(i,testvector2))/(numpy.std(ang2)+.00001))
        score3 = int(abs(numpy.mean(ang3)-
calcangnonabs(i,testvector3))/(numpy.std(ang3)+.00001))
        score4 = int(abs(numpy.mean(ang4)-
calcangnonabs(i,testvector4))/(numpy.std(ang4)+.00001))


        if score1+score2+score3+score4 < 1:
            globals()['good'+str(y)].append(i)

###-------------------------------------------------------------------------------->>>>>


scaled01 = []
scaled02 = []
scaled03 = []
scaled10 = []
scaled12 = []
scaled13 = []
scaled20 = []
scaled21 = []
scaled23 = []
scaled30 = []
scaled31 = []
scaled32 = []
```

```
###### output everything to screen ####
## make raw output file later?

## rough groups
for i in ('01','02','03','10','12','13','20','21','23','30','31','32'):
    if len(globals()['group'+i]) > 0:
        groupcalc(globals()['group'+i],i)

####### scale the vectors


##----calc minimum funct---------------
def calcmin(x):
    allmags = []
    for vector in x:
        allmags.append(numpy.linalg.norm(vector))
    return min(allmags)
##--------------------------------------

for i in ('01','02','03','10','12','13','20','21','23','30','31','32'):
    for x in globals()['good'+i]:
        if numpy.linalg.norm(x) > 2*calcmin(globals()['good'+i]):
            globals()['scaled'+i].append(x/2)
        else:
            globals()['scaled'+i].append(x)

## -- list printing function------
def printlist(x):
    print "-----"
    for i in x:
        print i
##-----------------------------


##--------- funct to calculate mean vectors -----------------
def vecmean(x):
    if len(x) > 0:
        iss = []
        js = []
        ks = []
        for i in x:
            iss.append(i[0])
            js.append(i[1])
            ks.append(i[2])
        mean = numpy.array([numpy.mean(iss),numpy.mean(js),numpy.mean(ks)])
        return [mean, numpy.linalg.norm(mean)]
    else:
        return ["no vector","n/a"]
##------------------------------------------------------------


candidates = []
## final mean vectors
print '\n------- Candidate vectors -------'
for i in ('01','02','03','10','12','13','20','21','23','30','31','32'):
    if len(globals()['good'+i]) > 2:
        print"%s\t%s  \tmag: %s  \tscore: %s" %
(i,vecmean(globals()['scaled'+i])[0],vecmean(globals()['scaled'+i])[1],
len(globals()['scaled'+i]))
        logout.write("%s\t%s  \tmag: %s  \tscore: %s\n" %
(i,vecmean(globals()['scaled'+i])[0],vecmean(globals()['scaled'+i])[1],
len(globals()['scaled'+i])))
        candidates.append([i,vecmean(globals()['scaled'+i])[0]])
print"\n------- Angles Between Candidates -------"
print "\t",
logout.write("\t")
for i in candidates:
    print str(i[0])+"\t",
```

```
    logout.write(str(i[0])+"\t")
print ""
logout.write("\n")
for i in candidates:
    anglist = []
    for n in candidates:
        anglist.append(str(calcangnonabs(i[1],n[1])))
    print "%s\t%s" % (i[0],'\t'.join(anglist))
    logout.write("%s\t%s\n" % (i[0],'\t'.join(anglist)))
```

A1.7: merge_p422_maxonly.py

Merges measured intensities accoding to p422 symmetry keeping only the maximum recorded intensity for each spot.

```
#! /usr/bin/env python

## imports:

import numpy
import math

# Takes file of all combined intensities makes lists of intensities for corresponding
symmetry mates.

# assumes the largest value represents the closest to a full intensity measurement -
keeps only that one.
# rmerge is not callculated becasue it is always 0

# Matt Iadanza 2013-07-18

############
# some variables
############

hrange = range(0,31)                    #
krange = range(0,31)                    # the maximum possible values for h,k and l
lrange = range(0,16)                    #
rawoutput = open("rawoutput_merge2.txt", "w")
output = open("f_mergedint.txt", "w")
vers = 2

### open the combined intensities files
with open('combint.txt') as pfile:
    lines = pfile.read().splitlines()


# make dictionary structure to contain the merged spots

spotint = {}    # input spot return all intensities for it

for h in hrange:
    for k in krange:
        for l in lrange:
            if h >= k:
                spotint[(h,k,l)] = []


## put intensity values into the spotint dict

for each in lines:
    data = each.split("\t")
    h,k,l = int(data[0]),int(data[1]),int(data[2])

# first deal with all nonzero miller indices
## hkl 'root' spots (+++) and their freidel pairs (---) -h -k -l

    if h > 0 and k > 0 and l >= 0 and  h >= k:
        spotint[(h,k,l)].append(float(data[5]))
    if h < 0 and k < 0 and l < 0 and  h < k:
        spotint[(-h,-k,-l)].append(float(data[5]))
```

```
## khl symmetry mates  (+++) and their freidel pairs (---) -k -h -l
    if h > 0 and k > 0 and l >= 0 and  h < k:
        spotint[(k,h,l)].append(float(data[5]))
    if h < 0 and k < 0 and l < 0 and  h >= k:
        spotint[(-k,-h,-l)].append(float(data[5]))

## h k -l (++-) symmetry mates and freidel pairs -h -k l (--+)
    if h > 0 and k > 0 and l < 0 and h >= k:
        spotint[(h,k,-l)].append(float(data[5]))
    if h < 0 and k < 0 and l >= 0 and h < k:
        spotint[(-h,-k,l)].append(float(data[5]))

## k h -l (++-) symmetry mates and freidel pairs  -k -h l  (--+)
    if h > 0 and k > 0 and l < 0 and h < k:
        spotint[(k,h,-l)].append(float(data[5]))
    if h < 0 and k < 0 and l >= 0 and h >= k:
        spotint[(-k,-h,l)].append(float(data[5]))

## h -k l (+-+) symmetry maes and freidel pairs -h k -l (-+-)
    if h > 0 and k < 0 and l >= 0 and h >= -k:
        spotint[(h,-k,l)].append(float(data[5]))
    if h < 0 and k > 0 and l < 0 and -h >= k:
        spotint[(-h,k,-l)].append(float(data[5]))

## h -k -l (+--) symmetry mates and freidel pairs -h k l (-++)
    if h > 0 and k < 0 and l < 0 and h >= -k:
        spotint[(h,-k,-l)].append(float(data[5]))
    if h < 0 and k > 0 and l >= 0 and -h >= k:
        spotint[(-h,k,l)].append(float(data[5]))

## k -h -l (+--) symmetry mates and freidel pairs -k h l (-++)
    if h < 0 and k > 0 and l < 0 and -h < k:
        spotint[(k,-h,-l)].append(float(data[5]))
    if h > 0 and k < 0 and l >= 0 and h < -k:
        spotint[(-k,h,l)].append(float(data[5]))

## k -h l (-+-)symmetry mates and freidel pairs -k h -l
    if h < 0 and k > 0 and l >= 0 and -h < k:
        spotint[(k,-h,l)].append(float(data[5]))
    if h > 0 and k < 0 and l < 0 and h < -k:
        spotint[(-k,h,-l)].append(float(data[5]))

## next deal with the special-case zero miller indices
    if h == 0 and k != 0:
        if k < 0:
            k = k*-1
        if l < 0:
            l = l*-1
        spotint[(k,h,l)].append(data[5])
    if k == 0 and h != 0:
        if h < 0:
            h = h*-1
        if l < 0:
            l = l*-1
        spotint[(h,k,l)].append(data[5])
    if k == 0 and h == 0:
        if l < 0:
            l = l *-1
        spotint[(h,k,l)].append(data[5])

## go over spotint dict and remove any entries with no intensity values

for i in spotint.keys():
    if len(spotint[i]) == 0:
        del(spotint[i])

count = 0
for i in spotint:
    count = count+len(spotint[i])

## go over spotint dict and keep only the maximum value:
```

```
keepvals = {}            # input spot(root spot) and return all intensity values that
meet the threshold
finalvalues = {}
vals = []

for i in spotint:
    maxval = max(spotint[i])

### Calculate the statistcs

    intensity = float(maxval)
    stdi = math.sqrt(intensity)
    sigi = stdi
    sigf = math.sqrt(sigi)
    finalvalues[i] = (sigi,sigf,intensity,stdi)
    rawoutput.write("***** %s: %s\n--------\n%s\n\n\n" % (i, spotint[i],maxval))
    output.write("%2i %2i %2i %12.4f %12.4f %12.4f %12.4f\n" %
(i[0],i[1],i[2],finalvalues[i][0],finalvalues[i][1],finalvalues[i][2],finalvalues[i][3]))


print "** max-only merge p422 vers: %2.1f **" % vers
print "%s intensities" % count
print "%s merged intensities" % len(finalvalues)
```

A1.8: merge_p422_thresh.py:

Merges measured intensities accoding to p422 symmetry. Artificially constrains r_{merge} by throwing away any values less than a specific threshold under the maximum recorded value for each intensity.

```
#! /usr/bin/env python

## imports:

import numpy
import math

# Takes file of all combined intensities makes lists of intensities for corresponding
symmetry mates in p422 symmetry.
# chooses which ones to sum based on a threshold (tfactor), sums them, and calculates
Rmerge for the data set. version 2

# Matt Iadanza 2013-07-18

############
# some variables
###########

tfactor = 0.6                           # threshold for keeping spots.  MaxRmerge = 1-t
hrange = range(0,31)                    #
krange = range(0,31)                    # the maximum possible values for h,k and l
lrange = range(0,31)                    #
rawoutput = open("rawoutput_merge2.txt", "w")
output = open("f_mergedint.txt", "w")
vers = 2

### open the combined intensities files
with open('combint.txt') as pfile:
    lines = pfile.read().splitlines()


# make dictionary structure to contain the merged spots

spotint = {}    # input spot return all intensities for it

for h in hrange:
    for k in krange:
        for l in lrange:
            if h >= k:
```

```
                spotint[(h,k,l)] = []


## put intensity values into the spotint dict

for each in lines:
    data = each.split("\t")
    h,k,l = int(data[0]),int(data[1]),int(data[2])

# first deal with all nonzero miller indices
## hkl 'root' spots (+++) and their freidel pairs (---) -h -k -l

    if h > 0 and k > 0 and l >= 0 and  h >= k:
        spotint[(h,k,l)].append(float(data[5]))
    if h < 0 and k < 0 and l < 0 and  h < k:
        spotint[(-h,-k,-l)].append(float(data[5]))

## khl symmetry mates  (+++) and their freidel pairs (---) -k -h -l
    if h > 0 and k > 0 and l >= 0 and  h < k:
        spotint[(k,h,l)].append(float(data[5]))
    if h < 0 and k < 0 and l < 0 and  h >= k:
        spotint[(-k,-h,-l)].append(float(data[5]))

## h k -l (++-) symmetry mates and freidel pairs -h -k l (--+)
    if h > 0 and k > 0 and l < 0 and h >= k:
        spotint[(h,k,-l)].append(float(data[5]))
    if h < 0 and k < 0 and l >= 0 and h < k:
        spotint[(-h,-k,l)].append(float(data[5]))

## k h -l (++-) symmetry mates and freidel pairs  -k -h l  (--+)
    if h > 0 and k > 0 and l < 0 and h < k:
        spotint[(k,h,-l)].append(float(data[5]))
    if h < 0 and k < 0 and l >= 0 and h >= k:
        spotint[(-k,-h,l)].append(float(data[5]))

## h -k l (+-+) symmetry maes and freidel pairs -h k -l (-+-)
    if h > 0 and k < 0 and l >= 0 and h >= -k:
        spotint[(h,-k,l)].append(float(data[5]))
    if h < 0 and k > 0 and l < 0 and -h >= k:
        spotint[(-h,k,-l)].append(float(data[5]))

## h -k -l (+--) symmetry mates and freidel pairs -h k l (-++)
    if h > 0 and k < 0 and l < 0 and h >= -k:
        spotint[(h,-k,-l)].append(float(data[5]))
    if h < 0 and k > 0 and l >= 0 and -h >= k:
        spotint[(-h,k,l)].append(float(data[5]))

## k -h -l (+--) symmetry mates and freidel pairs -k h l (-++)
    if h < 0 and k > 0 and l < 0 and -h < k:
        spotint[(k,-h,-l)].append(float(data[5]))
    if h > 0 and k < 0 and l >= 0 and h < -k:
        spotint[(-k,h,l)].append(float(data[5]))

## k -h l (-+-)symmetry mates and freidel pairs -k h -l
    if h < 0 and k > 0 and l >= 0 and -h < k:
        spotint[(k,-h,l)].append(float(data[5]))
    if h > 0 and k < 0 and l < 0 and h < -k:
        spotint[(-k,h,-l)].append(float(data[5]))

## next deal with the special-case zero miller indices
    if h == 0 and k != 0:
        if k < 0:
            k = k*-1
        if l < 0:
            l = l*-1
        spotint[(k,h,l)].append(data[5])
    if k == 0 and h != 0:
        if h < 0:
            h = h*-1
        if l < 0:
            l = l*-1
        spotint[(h,k,l)].append(data[5])
    if k == 0 and h == 0:
```

```
        if l < 0:
            l = l *-1
        spotint[(h,k,l)].append(data[5])

## go over spotint dict and remove any entries with no intensity values

for i in spotint.keys():
    if len(spotint[i]) == 0:
        del(spotint[i])

count = 0
for i in spotint:
    count = count+len(spotint[i])

## go over spotint dict and remove any intensities that are outside the threshold of the
max value:

keepvals = {}            # input spot(root spot) and return all intensity values that
meet the threshold
finalvalues = {}
vals = []

for i in spotint:
    maxval = max(spotint[i])
    keepvals[i] = []
    for n in spotint[i]:
        if n >= tfactor*float(maxval):
            keepvals[i].append(n)

### Calculate the statistcs

rmergerunning = []
rmergecount = 0

for i in keepvals:
    nums = []
    for n in keepvals[i]:
        nums.append(float(n))
    mean = numpy.mean(nums)
    rmergecount = rmergecount + len(nums)
    for n in nums:
            val = abs(n-mean)/mean
            rmergerunning.append(float(val))
    if len(keepvals[i]) == 1:
        intensity = float(nums[0])
        stdi = math.sqrt(intensity)
        sigi = stdi
        sigf = math.sqrt(stdi)
    if len(keepvals[i]) == 2:
        intensity = sum(nums)/len(nums)
        stdi = math.sqrt(intensity)
        sigi = stdi
        sigf = math.sqrt(stdi)
    if len(keepvals[i]) > 2:
        intensity = sum(nums)/len(nums)
        stdi = numpy.std(nums)
        sigi = math.sqrt(intensity)
        sigf = math.sqrt(stdi)
    finalvalues[i] = (sigi,sigf,intensity,stdi)
    rawoutput.write("***** %s: %s\n--------\n%s\n\n" % (i, spotint[i],nums))
    output.write("%2i %2i %2i %12.4f %12.4f %12.4f %12.4f\n" %
(i[0],i[1],i[2],finalvalues[i][0],finalvalues[i][1],finalvalues[i][2],finalvalues[i][3]))



print "** thresholded merge p422 vers: %2.1f **" % vers
print "max allowable Rmerge: %s" % (1-tfactor)
print "%s intensities" % count
print "%s merged intensities" % len(finalvalues)
print "Rmerge: %s" % (sum(rmergerunning)/rmergecount)
```

A1.9 params.json

an example of the universal parameters file used by all programs.

```
{
    "imgsize" : 4096,
    "circrad" : 20,
    "imgmaxres" : 2.01,
    "reslimit" : 3.0,
    "hrange" : 30,
    "krange" : 30,
    "lrange" : 30,
    "aucvec" : [28.10672237,47.005545,-7.72205285],
    "bucvec" : [-44.36131083,28.49151531,17.70802328],
    "cucvec" : [-38.933103206,13.71934825,-103.90264167],
    "beamstoprad" : 100,
    "integrationthresh":0
}
```

A1.10: cataspot.json

An example of the format for the datafile containing information about each image, drawn on by all programs. An actual data file would contain up to 90 image entrys with 10-20 sets of "spots" coordinates each. This file is generated by the program cataspot.py or manually using measurements performed in imagej.

```
{
  "data": {
    "directory": "/Users/matti/projects/3dem/test_dataset",
    "images": {
      "062613_0001.tif": {
        "spots": [
          [ 1308.7734448792335, 2764.512751315612],
          [2588.729123314937, 2452.709852528725],
          [2469.8207290922087, 2678.3273766976413],
          [2788.596800234845, 2519.7868780273006 ],
          [2644.0289418846955, 1256.57698541329 ]
        ],
"beamstopcenter": [
          [2058, 2037]
        ],
        "tiltangle": "-40.0",
        "references": [
          [2046, 1060],
          [1906, 2964]
        ],
        "beamcenter": [
          [2040.175, 2057.062]
        ],
        "centers": [
          [1955, 2773],
          [1928, 1283],
]
      },
},

  "metadata": {
    "date": "Thu Sep 26 12:24:49 2013",
    "username": "matti",
    "description": "cataspot data file",
    "file version": 1
  }
```

Appendix 2: Additional programs written as tools for data validation.

A2.1: ints-validation.

Creates testfiles 1-4 used for model bias testing in section 3.5.

```
#! /usr/bin/env python

#### making intensity files for validation
#

###### IMPORTS

import random
import math


###### first test: randomized intensities between min and max values of the actual
intensities
randoutput = open("ints-rando.txt", "w")

with open('combint.txt') as pfile:
    lines = pfile.read().splitlines()
minmaxcalc = []
for i in lines:
    splitted = i.split()
    minmaxcalc.append(float(splitted[5]))

randomized = [0,0,0,0,0,0,0]
randminmaxcalc = []

for i in lines:
    single = i.split()
    randomized[0] = int(single[0])
    randomized[1] = int(single[1])
    randomized[2] = int(single[2])
    randomized[5] = round(random.uniform(min(minmaxcalc), max(minmaxcalc)),4)
    randomized[6] = round(math.sqrt(randomized[5]),4)
    randomized[3] = round(math.sqrt(randomized[5]),4)
    randomized[4] = round(math.sqrt(randomized[3]),4)
    randminmaxcalc.append(randomized[5])
    randoutput.write("%2i %2i %2i %12.4f %12.4f %12.4f %12.4f\n" % (randomized[0],
randomized[1], randomized[2], randomized[3], randomized[4], randomized[5],
randomized[6]))

print "testfile 1: completely randomized intensities bewteen actual min and max"
print "actual min:\t%s    \tactual max:\t%s" %  (min(minmaxcalc),max(minmaxcalc))
print "randomized min:\t%s   \trandomized max:\t%s" %
(min(randminmaxcalc),max(randminmaxcalc))


########  second test: actual intensities all independently and randomly +/- 0 to 35%:
plusminusoutput = open("ints-plusminus.txt", "w")

plusminus = [0,0,0,0,0,0,0]
for i in lines:
    single = i.split()
    plusminus[0] = int(single[0])
    plusminus[1] = int(single[1])
    plusminus[2] = int(single[2])
    plusminus[5] = round(float(single[5])*random.uniform(0.65,1.35),4)
    plusminus[6] = round(math.sqrt(plusminus[5]),4)
    plusminus[3] = round(math.sqrt(plusminus[5]),4)
    plusminus[4] = round(math.sqrt(plusminus[3]),4)
    plusminusoutput.write("%2i %2i %2i %12.4f %12.4f %12.4f %12.4f\n" % (plusminus[0],
plusminus[1], plusminus[2], plusminus[3], plusminus[4], plusminus[5], plusminus[6]))
print ""
print "made +/- randomized 0 - 35% test file"
```

```
########## third test shuffle the values = all intensity values are actual but their
miller indicies are shuffled:
shuffledoutput = open("ints-shuffled.txt", "w")

valstoshuff = []
for i in lines:
    single = i.split()
    valstoshuff.append(single[5])

random.shuffle(valstoshuff)


shuffled = [0,0,0,0,0,0,0]
for i in lines:
    single = i.split()
    shuffled[0] = int(single[0])
    shuffled[1] = int(single[1])
    shuffled[2] = int(single[2])
    shuffled[5] = float(valstoshuff[lines.index(i)])
    shuffled[6] = round(math.sqrt(shuffled[5]),4)
    shuffled[3] = round(math.sqrt(shuffled[5]),4)
    shuffled[4] = round(math.sqrt(shuffled[3]),4)
    shuffledoutput.write("%2i %2i %2i %12.4f %12.4f %12.4f %12.4f\n" % (shuffled[0],
shuffled[1], shuffled[2], shuffled[3], shuffled[4], shuffled[5], shuffled[6]))

print ""
print "made shuffled test file"

######### test 4: intensities froma completely unrelated protein structure
subbedoutput = open("ints-subbed.txt", "w")

## make dict from reference file
indexref = {}
with open('3sui.txt') as reffile:
    reflines = reffile.read().splitlines()
    for i in reflines:
        splitreflines = i.split()
        indexref[(splitreflines[3],splitreflines[4],splitreflines[5])] =
(splitreflines[7],splitreflines[8])

### make the ints test file:
###

subbed = [0,0,0,0,0,0,0]
with open('combint.txt') as pfile:
    lines = pfile.read().splitlines()
for i in lines:
        single = i.split()
        if (single[0],single[1],single[2]) in indexref.keys():
            subbed[0] = single[0]
            subbed[1] = single[1]
            subbed[2] = single[2]
            subbed[5] = float(indexref[(single[0],single[1],single[2])][0])**2
            subbed[6] = float(indexref[(single[0],single[1],single[2])][1])**2
            subbed[3] = float(indexref[(single[0],single[1],single[2])][0])
            subbed[4] = float(indexref[(single[0],single[1],single[2])][1])
            subbedoutput.write("%2i %2i %2i %12.4f %12.4f %12.4f %12.4f\n" %
(int(subbed[0]), int(subbed[1]), int(subbed[2]), subbed[3], subbed[4], subbed[5],
subbed[6]))
```

A2.2: EM-MTZdump-comp.py.

Example of the program that compares structure factors from a uED suite merged intensity file and structure factors extracted from a MTZ file using MTZ.dump from PHENIX. Slight modifications allow this program to be used with a variety of combinations of file formats.

```
#! /usr/bin/env python

### compare structure factors from MTZdumped mtzfiles to MicroED suite intensity file
```

```
### prepare the mtz dump file using the following shell command
###     phenix.mtz.dump -f s -c [filename] > [output_filename]

################################################
## User edited variables
emfile = "microED_dataset.txt"
dumpfile = "mosflm_rot.txt"
fcol = 3    # which column in MTZdump file contains F values 1st column = 0
output = open("EM-MTZ.csv","w")
################################################


## open MTZdumped cif file and make dictionaries

dumpdatadic = {}
with open(dumpfile) as reffile:
    reflines = reffile.read().splitlines()
    for i in reflines:
        single = i.split(',')
        if single[fcol] == "":
            single[fcol] = "0"
        dumpdatadic[(single[0],single[1],single[2])] = single[fcol]


for i in dumpdatadic:
    print "%s\t%s" % (i,dumpdatadic[i])

## open EM data file and make dictionaries

emdatadic = {}
with open(emfile) as expfile:
    explines = expfile.read().splitlines()
    for i in explines:
        single = i.split()
        emdatadic[(single[0],single[1],single[2])] = single[3]

output.write("h,k,l,%s,%s,ratio\n"% (emfile,dumpfile))
for i in emdatadic:
    if emdatadic[i] != "?":
        if i in dumpdatadic:
            if dumpdatadic[i] != "?":
                if float(dumpdatadic[i]) != 0:
                    print (i, emdatadic[i],dumpdatadic[i],
float(emdatadic[i])/float(dumpdatadic[i]))
                    output.write("%s,%s,%s,%s,%s,%s\n" %
(i[0],i[1],i[2],float(emdatadic[i]),float(dumpdatadic[i]),float(emdatadic[i])/float(dumpd
atadic[i])))
```

Appendix 3: Additional publications with abstracts.

A suite of software for processing MicroED data
Iadanza MG and Gonen T
Journal of Applied Crystallography (In press)

Electron diffraction of extremely small 3-dimensional crystals (MicroED) allows for structure determination from crystals orders of magnitude smaller than those used for X-ray crystallography. MicroED diffraction patterns, which are collected in a transmission electron microscope, were initially not amenable to indexing and intensity extraction by standard software which necessitated development of a suite of programs for data processing. The MicroED suite was developed to accomplish the tasks of unit cell determination, indexing, background subtraction, intensity measurement, and merging, resulting in data which can be carried foreword to molecular replacement and structure determination. The suite is written in Python and the source code available under a GNU General Public License.

Proton-coupled sugar transport in the prototypical major facilitator superfamily protein XylE

Wisedchaisri G*, Park M*, Iadanza MG, Zheng H & Gonen T
Nature Communications (Submitted)

The major facilitator superfamily of membrane proteins is the largest collection of structurally related membrane proteins that transport a wide array of substrates. The proton-coupled sugar transporter XylE is the first member of the MFS that has been structurally characterized in multiple transporting conformations including both the outward and inward facing states. Here we report the crystal structure of XylE in a new inward-facing open conformation. Structural comparison of XylE in this conformation with its outward-facing partially occluded conformation suggested how this transporter functions through a non-symmetrical rocker switch movement of the N-domain as a rigid body and the C-domain as a flexible body. Molecular dynamics simulations were employed to help describe how XylE transitions in a lipid membrane to facilitate sugar transport. Asp27 and Arg133 are highlighted as potential residues involved in proton coupling. In our simulations, the interaction between Asp27 and Arg133 appear to suppress transporter function and mutation of these residues severely diminished transporter function in our uptake studies. Protonation of Asp27 disrupts its interaction with Arg133 and imparts flexibility in XylE to facilitate the transition from outward to inward facing for sugar transport. These findings provide important and new insights into the mechanism of transport employed by this large and ubiquitous family of secondary transporters in health and disease.

Three-dimensional electron crystallography of protein microcrystals
Shi D*, Nannenga BL*, Iadanza MG*, Gonen T.
elife 2013;2:e01345 DOI: 10.7554/eLife.01345
(* co-first authors)

We demonstrate that it is feasible to determine high-resolution protein structures by electron crystallography of three-dimensional crystals in an electron cryo-microscope (CryoEM). Lysozyme microcrystals were frozen on an electron microscopy grid, and electron diffraction data collected to 1.7Å resolution. A new data collection protocol was developed to collect a full-tilt series in electron diffraction to atomic resolution. A single tilt series contains up to 90 individual diffraction patterns collected from a single crystal with tilt angle increment of 0.1 - 1° and a total accumulated electron dose less than 10 electrons per angstrom squared. New algorithms were developed for indexing, and data originating from 3 crystals were used for structure determination of lysozyme by molecular replacement followed by crystallographic refinement to 2.9Å resolution. This proof of principle paves the way for the implementation of a new technique that may have wide applicability in structural biology.

Overview of electron crystallography of membrane proteins: crystallization and screening strategies using negative stain electron microscopy.
Nannenga BL, Iadanza MG, Vollmar BS, Gonen T.
Curr Protoc Protein Sci. 2013;Chapter 17:Unit17.15.
DOI: 10.1002/0471140864.ps1715s72.

Electron cryomicroscopy, or cryoEM, is an emerging technique for studying the three-dimensional structures of proteins and large macromolecular machines. Electron crystallography is a branch of cryoEM in which structures of proteins can be studied at resolutions that rival those achieved by X-ray crystallography. Electron crystallography employs two-dimensional crystals of a membrane protein embedded within a lipid bilayer. The key to a successful electron crystallographic experiment is the crystallization, or reconstitution, of the protein of interest. This unit describes ways in which protein can be expressed, purified, and reconstituted into well-ordered two-dimensional crystals. A protocol is also provided for negative stain electron microscopy as a tool for screening crystallization trials. When large and well-ordered crystals are obtained, the structures of both protein and its surrounding membrane can be determined to atomic resolution.

The structure of purified kinetochores reveals multiple microtubule-attachment sites.

Gonen S,* Akiyoshi B*, Iadanza MG*, Shi D, Duggan N, Biggins S, Gonen T.
Nat Struct Mol Biol. 2012 Sep;19(9):925-9.
DOI: 10.1038/nsmb.2358. Epub 2012 Aug 12.
(* co-first authors)

Chromosomes must be accurately partitioned to daughter cells to prevent aneuploidy, a hallmark of many tumors and birth defects. Kinetochores are the macromolecular machines that segregate chromosomes by maintaining load-bearing attachments to the dynamic tips of microtubules. Here, we present the structure of isolated budding-yeast kinetochore particles, as visualized by EM and electron tomography of negatively stained preparations. The kinetochore appears as an ~126-nm particle containing a large central hub surrounded by multiple outer globular domains. In the presence of microtubules, some particles also have a ring that encircles the microtubule. Our data, showing that kinetochores bind to microtubules via multivalent attachments, lay the foundation to uncover the key mechanical and regulatory mechanisms by which kinetochores control chromosome segregation and cell division.

Appendix 4: Curriculm vitae.

Matthew Gregory Iadanza
Curriculum vitae

Howard Hughes Medical Institute
Janelia Farm Research Campus
44103 Gala Circle
Ashburn, VA 20147
+1 970-545-0986
miadanza@uw.edu
Nationality: US citizen

Education

Ph.D in Biochemistry. University of Washington, Seattle, WA. 2014. (expected)
Area of specialization: Structural Biology.

BS in Biology. Beloit College, Beloit, WI. 2000.

Doctoral Research Experience

Howard Hughes Medical Institute Janelia Farm Research Campus, Ashburn, VA. 2012-2014.
University of Washington Department of Biochemistry, Seattle, WA. 2009 – 2014.
Supervisor: Dr. Tamir Gonen

Dissertation Title: Protein structure determination by electron diffraction of 3-dimensional protein microcrystals.

Research focus: Structure determination of integral membrane proteins with cryo-electron microscopy.
- Electron crystallography of 2-dimensional crystals and single particle reconstruction.
- Method development for electron crystallography of 3-dimensional microcrystals.
- Functional studies on homologues of human glucose transporters using radiolabelled substrates.
- X-ray crystallography of integral membrane proteins.

Professional Experience

Medicinal Chemistry Research Associate. Aileron Therapeutics, Cambridge, MA. 2007-2009.
- Synthetic development for structure based design of peptide based drugs.
- Synthesis of stabilized α-helical peptides.
- Small molecule synthetic organic chemistry.

Director of Research and Development. Global Peptide Services, Fort Collins, CO. 2001-2007.
- Method development for synthesis of peptides and peptidomimetics.
- Independent lab management, training and supervision of four technicians.
- Solid phase synthesis of peptides and peptidomimetics.
- Synthetic organic chemistry: small molecule synthesis, peptide crosslinking.

Predoctoral Laboratory Rotations

University of Washington, Seattle, WA. 2010. Supervisor: Dr. Alexey Merz
- Structural and biochemical characterization of Atg9 autophagy associated integral membrane protein.
- Cloning and expression, analysis by single particle EM and imaging, crystallization screening.

University of Washington, Seattle, WA. 2009. Supervisor: Dr. Gabriele Varani
- Fragment screening for small molecule inhibitors of human phosphodiesterase-5.
- Cloning and expression, fragment library design

Undergraduate Research Experience

University of North Carolina, Chapel Hill, NC. Supervisor: Dr. Jason Reed
- Yeast two-hybrid screening for interactions between auxin response factors in *Arabidopsis thaliana*.

Teaching Experience

Teaching Assistant: Biochemistry 426: Basic Techniques in Biochemistry, University of Washington. 2011.
Undergraduate practical laboratory course for 3 sections each with approximately 30 students. Prepared and delivered lectures, wrote and graded written and practical exams, supervised students in the laboratory.

Teaching Assistant: Biochemistry 405/406: Introduction to Biochemistry, University of Washington. 2010.
Introductory undergraduate biochemistry course with over 600 students. Held weekly discussion sections, moderated message boards on a daily basis, proctored exams.

Awards

University of Washington, NIH Molecular Biophysics Training Grant. 2011-2012.
University of Washington Schultz Travel Fellowship. 2011.
Beloit College Presidential Scholarship. 1996-2000.
Magna cum laude, Beloit College. 2000.
Departmental Honors, Beloit College Department of Biology. 2000.
Susan Fulton Welty Award, Beloit College Department of Biology. 2000.
J.N. "Ding" Darling Award, Beloit College Department of Biology. 1998.

Publications in referred journals

Shi D*, Nannenga, BL*, **Iadanza MG***, Gonen T. (2013) Three-dimensional electron crystallography of protein microcrystals. *eLife*. 2:e01345 (* co-first authors)

Gonen S*, Akiyoshi B*, **Iadanza MG***, Shi D, Duggan N, Biggins S, Gonen T. (2012) The structure of purified kinetochores reveals multiple microtubule-attachment sites. *Nature Structural & Molecular Biology* 19:925–929. doi:10.1038/nsmb.2358 (* co-first authors)

Nannenga BL, **Iadanza MG**, Vollmar BS, Gonen T. (2012) Overview of Electron Crystallography of Membrane Proteins: Crystallization and Screening Strategies Using Negative Stain Electron Microscopy. *Current Protocols in Protein Science* 17.15. 1-17.15. 11. doi: 10.1002/0471140864.ps1715s72

Iadanza MG & Gonen T. (2014) A suite of software for processing microcrystal electron diffraction (MicroED) data. *Journal of Applied Crystallogrpahy.* In Press

Publications - submitted

Wisedchisri G*, Park, M*, **Iadanza MG***, Zheng H, Gonen T. (2013) Proton-coupled sugar transport in the prototypical major facilitator superfamily protein XylE.

Patents

Kapeller-Libermann R, Nash HM, Sawyer TK, Kawahata N, Guerlevais V**, Iadanza M**. Biologically Active Peptidomimetic Macrocycles. US Patent 20,130,023,646 (2013); EP Patent 2,310,407(2011); WO Patent 2,010,034,032 (2010)

Nash HM, **Iadanza M**, Leitheiser C, Kawahata N. Methods for Preparing Purified Polypeptide Compositions. US Patent 20,120,264,674 (2012); EP Patent 2,342,221 (2011); WO Patent 2,010,034,032 (2010)

Posters

Shi D, Nannenga BL, **Iadanza MG**, Gonen T. (2013) Electron crystallography of 3-dimensional protein microcrystals. 2013 HHMI Science Meeting. Janelia Farm Research Campus, Ashburn VA.

Ramanathan R, Mahadevan R, **Iadanza M**, Woodrow K. (2012) Biophysical characterization of hydrogel-core, lipid-shell nanolipogels for HIV chemoprophylaxis. 2012 International Microbicides Conference, Sydney, Australia

Kapeller R, Han J, Sun K, Gangurde P, Kawahata N, **Iadanza M**, Guerlevais V, Horstick J, Noehre J, Annis A, Licklider L, Nash HM, Kung AL, Sawyer TK. (2008) Stapled Peptides: Leveraging the BH3 α-Helix to Create a New Class of Drugs to Treat Hematological Malignancies. American Society of Hematology 50th Annual Meeting, San Francisco, CA, USA. Poster Abstract No. 2929.

Invited talks

2013 Junior Scientists Seminar Series, Howard Hughes Medical Institute

Appendix 5: Reprints of Additional Publications

Three-dimensional electron crystallography of protein microcrystals
Shi D*, Nannenga BL*, Iadanza MG*, Gonen T. *elife* 2013;2:e01345 DOI: 10.7554/eLife.01345

The structure of purified kinetochores reveals multiple microtubule-attachment sites.
Gonen S,* Akiyoshi B*, Iadanza MG*, Shi D, Duggan N, Biggins S, Gonen T. *Nat Struct Mol Biol.* 2012 Sep;19(9):925-9. DOI: 10.1038/nsmb.2358. Epub 2012 Aug 12.

eLIFE
elife.elifesciences.org

RESEARCH ARTICLE

Three-dimensional electron crystallography of protein microcrystals

Dan Shi†, Brent L Nannenga†, Matthew G Iadanza†, Tamir Gonen*

Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, United States

Abstract We demonstrate that it is feasible to determine high-resolution protein structures by electron crystallography of three-dimensional crystals in an electron cryo-microscope (CryoEM). Lysozyme microcrystals were frozen on an electron microscopy grid, and electron diffraction data collected to 1.7 Å resolution. We developed a data collection protocol to collect a full-tilt series in electron diffraction to atomic resolution. A single tilt series contains up to 90 individual diffraction patterns collected from a single crystal with tilt angle increment of 0.1–1° and a total accumulated electron dose less than 10 electrons per angstrom squared. We indexed the data from three crystals and used them for structure determination of lysozyme by molecular replacement followed by crystallographic refinement to 2.9 Å resolution. This proof of principle paves the way for the implementation of a new technique, which we name 'MicroED', that may have wide applicability in structural biology.
DOI: 10.7554/eLife.01345.001

***For correspondence:** gonent@janelia.hhmi.org

†These authors contributed equally to this work

Competing interests: The authors declare that no competing interests exist.

Funding: See page 15

Received: 07 August 2013
Accepted: 05 October 2013
Published: 19 November 2013

Reviewing editor: Stephen C Harrison, Harvard Medical School, United States

Introduction

X-ray crystallography depends on large and well-ordered crystals for diffraction studies. Crystals are solids composed of repeated structural motifs in a three-dimensional lattice (hereafter called '3D crystals'). The periodic structure of the crystalline solid acts as a diffraction grating to scatter the X-rays. For every elastic scattering event that contributes to a diffraction pattern there are ~10 inelastic events that cause beam damage (***Henderson, 1995***). Therefore, large crystals are required to withstand the high levels of radiation damage received during data collection (***Henderson, 1995***). Despite the development of highly sophisticated robotics for crystal growth assays and the implementation of microfocus beamlines (***Moukhametzianov et al., 2008***), this important step remains a critical bottleneck. In an attempt to alleviate this problem, researchers have turned to femtosecond X-ray crystallography (***Chapman et al., 2011***; ***Boutet et al., 2012***), in which a very intense pulse of X-rays yields coherent signal in a time shorter than the destructive response to deposited energy. While this technique shows great promise, the current implementation of the technology requires an extremely large number of crystals (millions) and access to sources is still in developmental stages.

Electron crystallography is a bona fide method for determining protein structure from crystalline material but with important differences. The crystals that are used must be very thin (***Henderson and Unwin, 1975***; ***Henderson et al., 1990***; ***Kuhlbrandt et al., 1994***; ***Kimura et al., 1997***). Because electrons interact with materials more strongly than X-rays (***Henderson, 1995***), electrons can yield meaningful data from relatively small and thin crystals. This technique has been used successfully to determine the structures of several proteins from thin two-dimensional crystals (2D crystals) (***Wisedchaisri et al., 2011***). High energy electrons result in a large amount of radiation damage to the sample, leading to loss in resolution and destruction of the crystalline material (***Glaeser, 1971***). As each crystal can usually yield only a single diffraction pattern, structure determination is only possible by merging data originating from hundreds of individual crystals. For example, electron diffraction data from more than 200 individual crystals were merged to generate a data set for aquaporin-0 at 1.9 Å resolution (***Gonen et al., 2005***).

eLife digest X-ray crystallography has been used to work out the atomic structure of a large number of proteins. In a typical X-ray crystallography experiment, a beam of X-rays is directed at a protein crystal, which scatters some of the X-ray photons to produce a diffraction pattern. The crystal is then rotated through a small angle and another diffraction pattern is recorded. Finally, after this process has been repeated enough times, it is possible to work backwards from the diffraction patterns to figure out the structure of the protein.

The crystals used for X-ray crystallography must be large to withstand the damage caused by repeated exposure to the X-ray beam. However, some proteins do not form crystals at all, and others only form small crystals. It is possible to overcome this problem by using extremely short pulses of X-rays, but this requires a very large number of small crystals and ultrashort X-ray pulses are only available at a handful of research centers around the world. There is, therefore, a need for other approaches that can determine the structure of proteins that only form small crystals.

Electron crystallography is similar to X-ray crystallography in that a protein crystal scatters a beam to produce a diffraction pattern. However, the interactions between the electrons in the beam and the crystal are much stronger than those between the X-ray photons and the crystal. This means that meaningful amounts of data can be collected from much smaller crystals. However, it is normally only possible to collect one diffraction pattern from each crystal because of beam induced damage. Researchers have developed methods to merge the diffraction patterns produced by hundreds of small crystals, but to date these techniques have only worked with very thin two-dimensional crystals that contain only one layer of the protein of interest.

Now Shi et al. report a new approach to electron crystallography that works with very small three-dimensional crystals. Called MicroED, this technique involves placing the crystal in a transmission electron cryo-microscope, which is a fairly standard piece of equipment in many laboratories. The normal 'low-dose' electron beam in one of these microscopes would normally damage the crystal after a single diffraction pattern had been collected. However, Shi et al. realized that it was possible to obtain diffraction patterns without severely damaging the crystal if they dramatically reduced the normal low-dose electron beam. By reducing the electron dose by a factor of 200, it was possible to collect up to 90 diffraction patterns from the same, very small, three-dimensional crystal, and then—similar to what happens in X-ray crystallography—work backwards to figure out the structure of the protein. Shi et al. demonstrated the feasibility of the MicroED approach by using it to determine the structure of lysozyme, which is widely used as a test protein in crystallography, with a resolution of 2.9 Å. This proof-of principle study paves the way for crystallographers to study protein that cannot be studied with existing techniques.

DOI: 10.7554/eLife.01345.002

While electron crystallography has been successful with 2D crystals, previous attempts at using electron diffraction for structure determination from protein 3D crystals were not successful. A number of studies detail the difficulties associated with data collection and processing of diffraction data that originates from several hundreds of 3D crystals, limiting the ability to integrate and merge the data in order to determine a structure in such a way (***Shi et al., 1998***; ***Jiang et al., 2011***).

We show here that atomic resolution diffraction data can be collected from crystals with volumes up to six orders of magnitude smaller than those typically used for X-ray crystallography. The technique, which we call 'MicroED', uses equipment standard in most cryo-EM laboratories and facilities. We developed a strategy for data collection with extremely low electron dose and procedures for indexing and integrating reflections. We processed the diffraction data and determined the structure of lysozyme at 2.9 Å resolution. Thus, a high-resolution protein structure can be determined from electron diffraction of three-dimensional protein crystals in an electron microscope.

Results

Sample preparation and data collection

Lysozyme was chosen as a model protein because it is a well-behaved and well-characterized protein that readily forms well-ordered crystals. From the time its structure was first analyzed (***Blake et al.,***

1962, *1965*), lysozyme has been a well-studied protein and the model protein of choice for many new methods in crystallography (*Boutet et al., 2012*; *Cipriani et al., 2012*; *Nederlof et al., 2013*). Small microcrystals of lysozyme were grown by slightly modifying the crystal growth conditions as detailed in the 'Materials and methods' section. *Figure 1A* shows a typical crystallization drop containing microcrystals, which appear as barely visible specks (arrows) alongside the larger crystals that are typically used for X-ray crystallography. These specks are up to 6 orders of magnitude smaller in volume than the larger crystals in the drop. The solution containing these microcrystals was applied to an electron microscopy holey-carbon grid with a pipette and plunged into liquid ethane. The grids were then imaged using a 200 kV TEM under cryogenic conditions (*Figure 1B*). More than 100 microcrystals were typically observed per grid preparation, and these ranged in size from several microns to sub micron. The crystals typically appeared as electron dense rectangular or triangular forms with very sharp edges.

Electron diffraction was used to assess the quality of the cryo-preparations. Crystals that appeared thick (estimated as >3 μm) did not yield diffraction data because the electron beam could not penetrate the sample. Crystals that appeared slightly thinner, estimated at ~1.5 μm, did show diffraction, but because the quality of the pattern varied depending on the sample tilt (*Figure 2A*), we did not use crystals of this thickness and size for data collection. Approximately 50% of the crystals in our preparations appeared much thinner, estimated at ~0.5 μm, and showed a distribution of attainable resolutions with the best diffracting to ~1.7 Å resolution (*Figure 2B,C*). Generally, we were only able to obtain high quality diffraction data from the very thin crystals, ~0.5–1 μm thick and 1–6 μm long and wide. While these crystals are exceptionally small, they still contain approximately 55 × 10^6 unit cells. Moreover, we found that for such thin crystals the tilt had no significant adverse affect on the diffraction quality (*Figure 2D*).

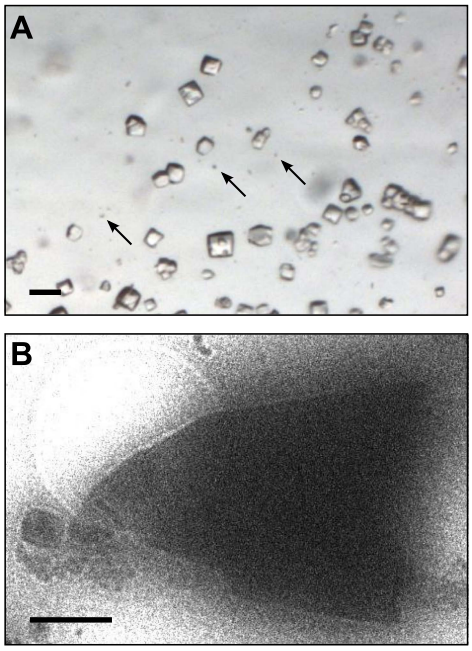


Figure 1. Images of lysozyme microrystals. (**A**) Light micrograph showing lysozyme microcrystals (three examples indicated by arrows) in comparison with larger crystals of the size normally used for X-ray crystallography. Scale bar is 50 μm. (**B**) Lysozyme microcrystals visualized in over-focused diffraction mode on the cryo-EM prior to data collection. The length and width of the crystals varied from 2 to 6 μm with an estimated thickness of ~0.5–1 μm. Scale bar is 1 μm.

DOI: 10.7554/eLife.01345.003

For 2D electron crystallography, the electron dose that is typically used in diffraction causes significant radiation damage to the sample, leading to a rapid loss in resolution and destruction of the crystal (*Glaeser, 1971*; *Unwin and Henderson, 1975*; *Taylor and Glaeser, 1976*). As a result, each crystal exposed to high dose usually only yields a single diffraction pattern, and structure determination requires the merging of data originating from a large number of individual crystals. However, 3D crystals can deliver electron diffraction data to atomic resolution with very low doses. A recent study documents ~3 Å resolution diffraction data from catalase 3D crystals after a single exposure of less than 10e^-/Å^2 (*Baker et al., 2010*).

We reasoned that one way to overcome the difficulties of indexing and merging data from hundreds of crystals is to collect a complete diffraction data set from a single crystal while keeping the total dose below ~10e^-/Å^2. Because all the data would originate from a single crystal, indexing, integration and merging should be straightforward and structure determination possible. We used a sensitive CMOS based detector (Tietz Video and Image Processing Systems GmbH), previously shown to be beneficial for electron diffraction studies (*Tani et al., 2009*) and modified our data collection procedure. We found that even with extremely low electron dose of <0.01 e^-/Å^2 per second, we could record diffraction data from lysozyme microcrystals showing strong and sharp diffraction spots extending well beyond the 2 Å resolution mark (*Figure 2C*).

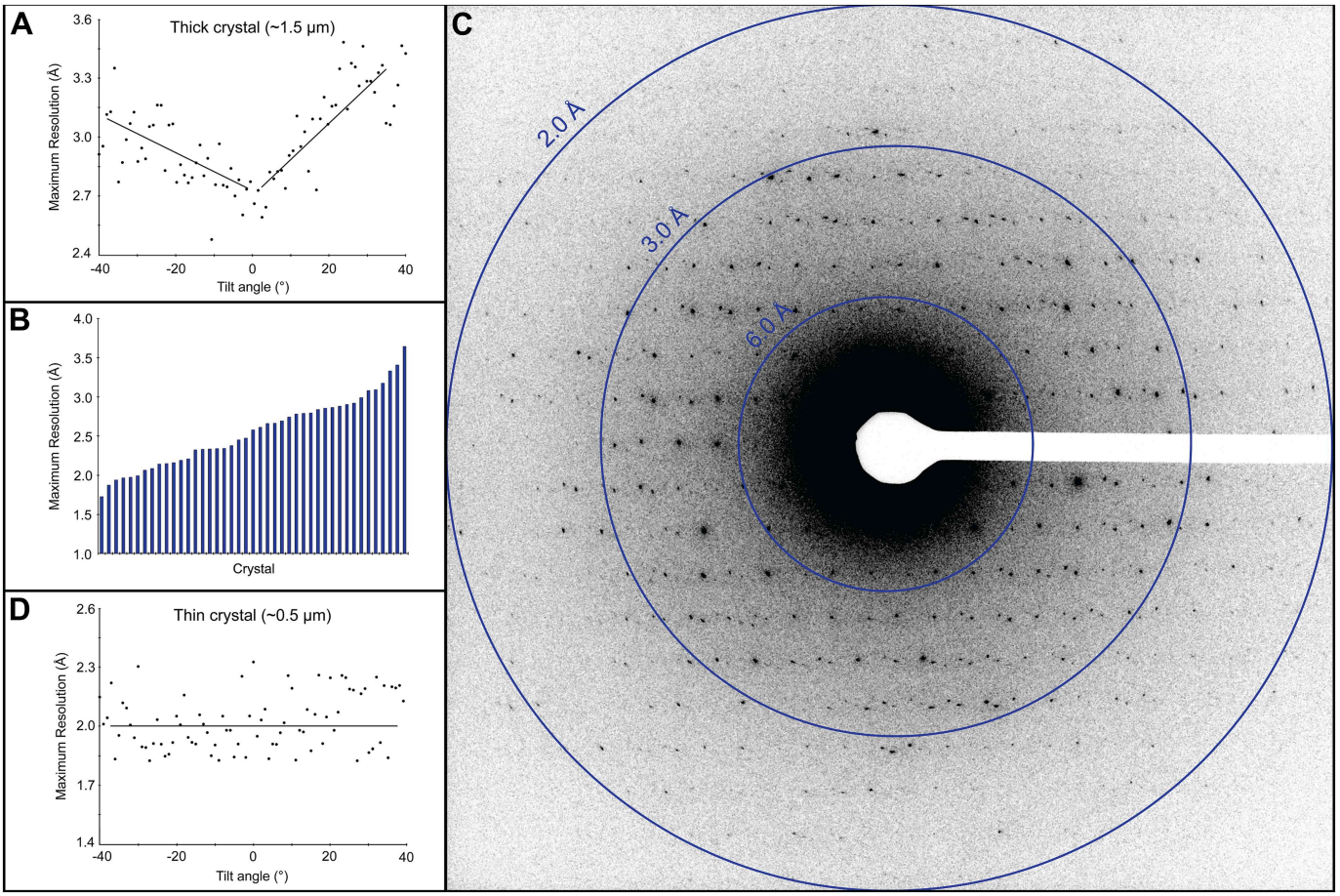


Figure 2. Resolution and data quality of lysozyme microcrystals. (**A**) Analysis of the effects of crystal thickness on maximum resolution of observed reflections from thick crystals. The analysis shows adverse effects of crystal thickness on the obtainable resolution as large crystals are tilted. (**B**) For assessing the quality of our cryo preparations, diffraction data were obtained from 100 lysozyme microcrystals. 43/100 were thin crystals that showed reflections in the 2–4 Å range, with the best crystal in this set yielding data to ~1.7 Å resolution. (**C**) An example of lysozyme diffraction data collected at 0.01e^{-}/Å^2/second and a 10 s exposure. The pattern shows strong and sharp spots surpassing 2 Å resolution. This diffraction pattern was processed with ImageJ and despeckled for ease of viewing. (**D**) Analysis of the effects of crystal thickness on maximum resolution of observed reflections from thin crystals. The small crystal shows a relatively constant maximum resolution that does not appear to be affected by crystal tilt.
DOI: 10.7554/eLife.01345.004

As a dataset containing multiple exposures from a single crystal is collected, energy transferred by inelastic scattering will damage the crystalline matrix, negatively affecting both the resolution limit and intensities of observed reflections. Although the overall damage from electron scattering is much lower than that for X-rays (approximately 60 eV deposited per elastic scattering event vs 80 keV per elastic X-ray scattering event [***Henderson, 1995***]), accumulating radiation damage will eventually contribute significant error to the recorded intensities.

We performed an experiment to quantify the effects of increasing electron dosage on recorded intensities (***Figure 3***). A single protein microcrystal was subjected to sequential 10 s exposures, each delivering ~0.1 e^{-}/Å^2, until a total accumulated dose of ~12 e^{-}/Å^2 was reached. The intensities of three diffraction spots, ranging from resolutions of 2.9 to 4.6 Å, were measured on each of the 120 resulting diffraction patterns and compared. There were no observable adverse effects on resolution (***Figure 2D***) or intensity until the accumulated dose had reached ~9 e^{-}/Å^2 (***Figure 3***). We therefore optimized the data collection protocol to keep the total accumulated electron dosage below this critical value.

By using such a low dose, we could limit the radiation damage to the crystal, allowing us to collect multiple diffraction patterns from a single crystal instead of just a single pattern. Using this modified

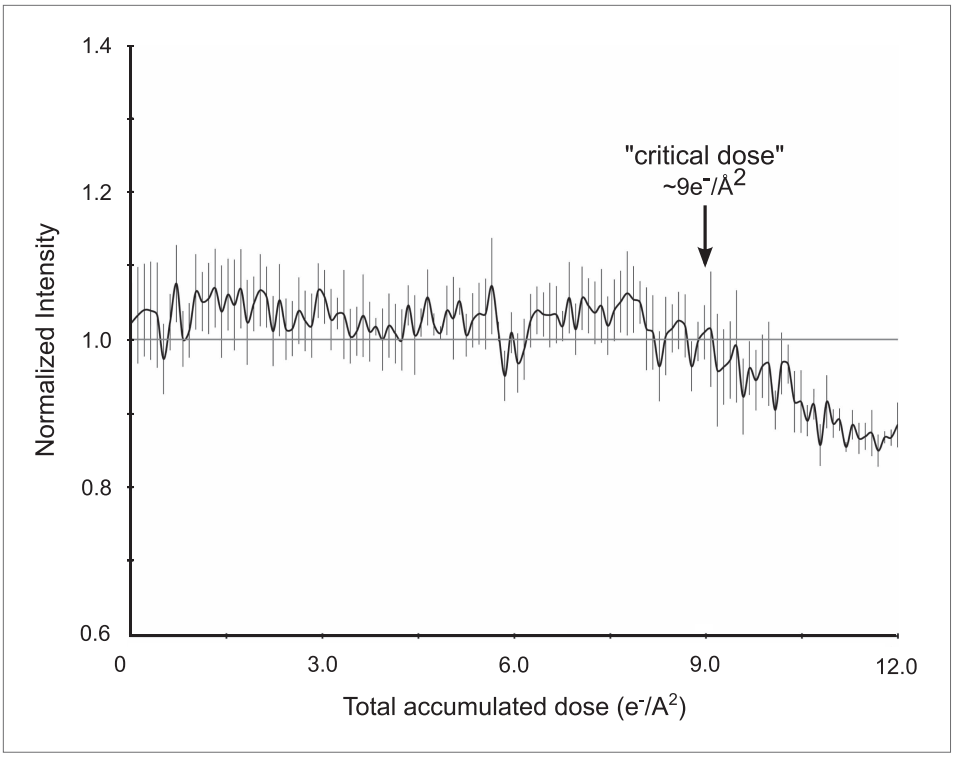
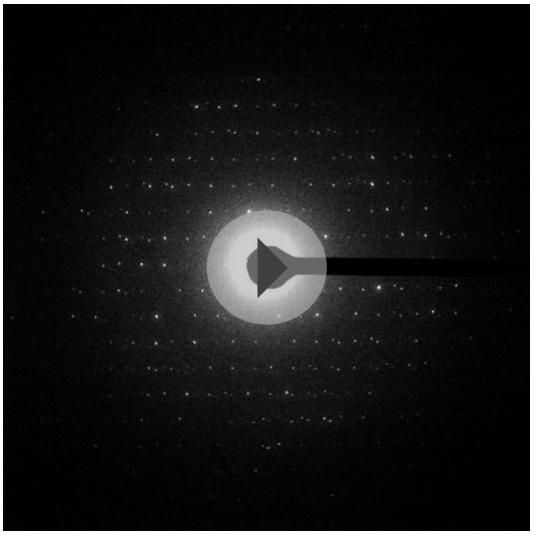


Figure 3. Effects of cumulative electron dose on diffraction data quality. A single lysozyme microcrystal was subjected to 120 sequential exposures without tilting, each of a dose of ~0.1 $e^-/Å^2$ for a total accumulated dose of ~12 $e^-/Å^2$. Normalized intensity vs total accumulated dose for three diffraction spots observed over all 120 sequential frames was plotted. A decrease in diffraction intensity becomes apparent at a dosage of ~9 $e^-/Å^2$ ('critical dose'). Bars represent standard error of the mean.

DOI: 10.7554/eLife.01345.005



Video 1. An example of a complete three-dimensional electron diffraction data set from a single lysozyme microcrystal. In this example, diffraction patterns were recorded at 1° intervals from a single crystal, tilted over 47°. Cumulative dose was ~5 $e^-/Å^2$ in this example.

DOI: 10.7554/eLife.01345.006

procedure, we were able to collect up to 90 individual diffraction patterns from a single crystal (***Video 1***). Each pattern was recorded following a 1° tilt to cover ~40–90° (begin with the stage tilted at −45° and proceed to collect data to +45° in order to cover a 90° wedge). 0.1 and 0.2 degree increments were also applied to sample the reciprocal space at higher resolution. Each exposure lasted up to 10 s at a dosage of approximately 0.01 $e^-/Å^2$ per second, for a cumulative dose of no more than ~9 $e^-/Å^2$ per data set.

Data processing and structure determination

The lattice parameters were determined and the lattice indexed with software based on previous studies (***Shi et al., 1998***). By collecting multiple frames from the same crystal, it was possible to determine the orientation and magnitude of the reciprocal unit cell vectors a*, b*, and c* as described in the 'Materials and methods' section. These vectors were calculated for each data set, allowing the prediction of the position of the reflections in each diffraction pattern (***Figure 4***;

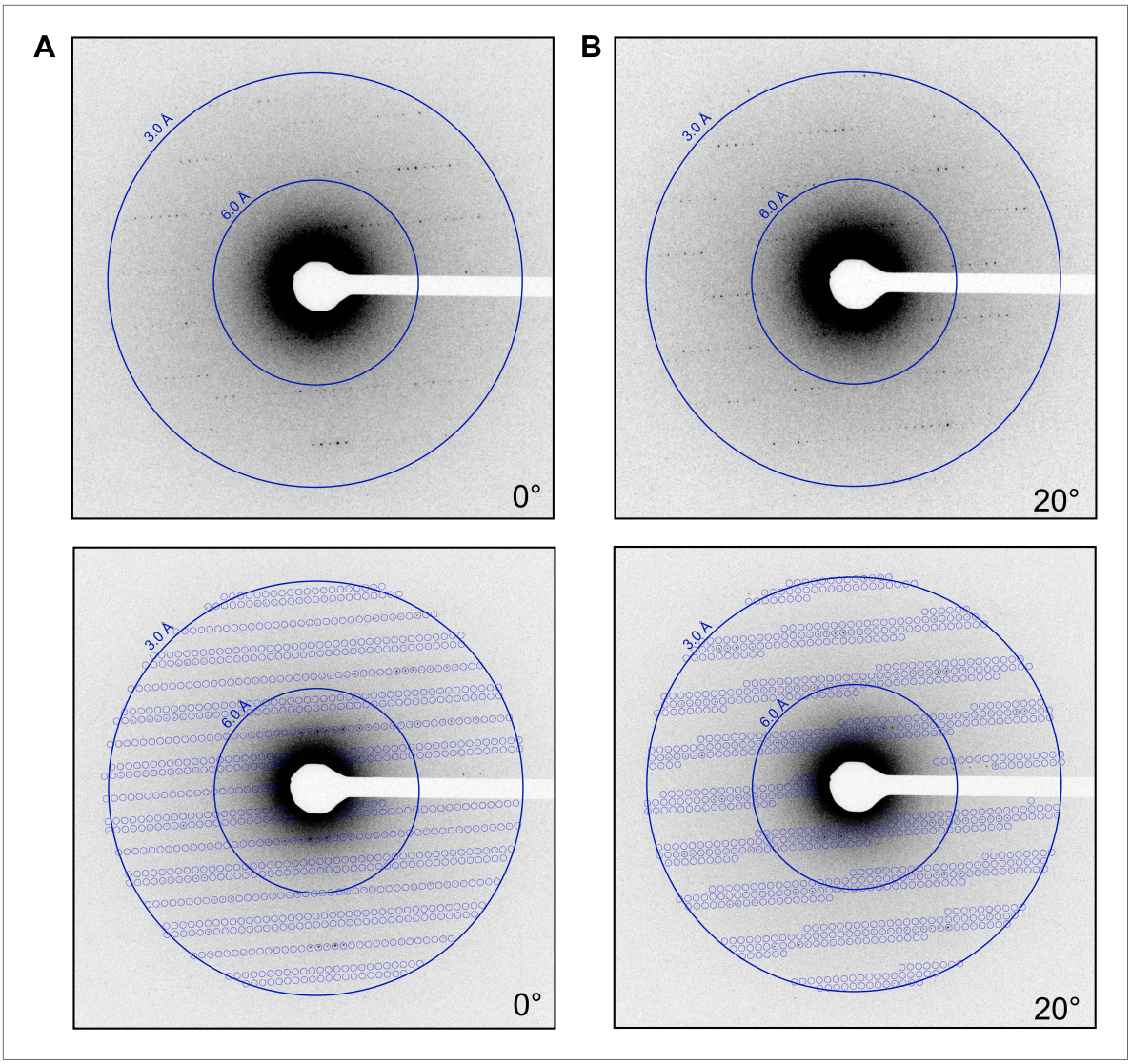
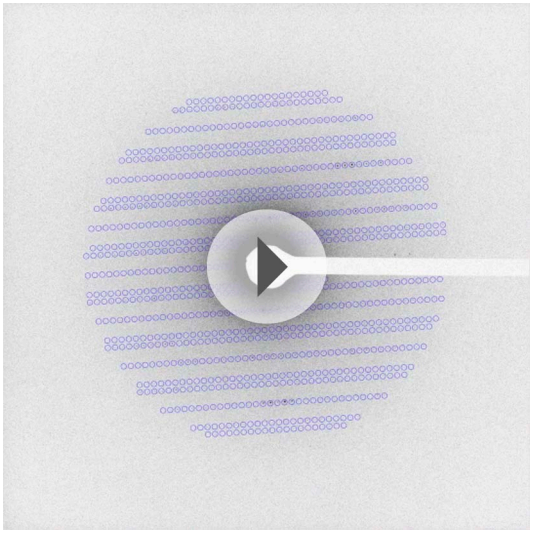


Figure 4. Prediction of reflections and indexing in the diffraction patterns. (**A** and **B**) Two examples of diffraction patterns obtained from a single crystal at tilt angles of 0° and 20° respectively. Locations indicated by circles were predicted to contain diffraction spots by our spot prediction algorithm. Additional examples from the same crystal are presented in ***Video 2***. The resolution limit was set at 2.9 Å resolution for this study.
DOI: 10.7554/eLife.01345.007

Video 2) and indexing of the entire data set. The unit cell dimensions were calculated as a = b = 77 Å, c = 37 Å, α = β = γ = 90° and $P4_32_12$ symmetry. This space group symmetry and unit cell dimensions are consistent with previous lysozyme X-ray diffraction data (***Diamond, 1974***; ***Sauter et al., 2001***; ***Cipriani et al., 2012***).

The electron diffraction data were collected on a microscope operating at 200 kV and equipped with a field emission gun (FEG) electron source. The FEG can generate a very coherent beam with an energy-spread function of <1 eV at 200 kV acceleration voltage. The electron beam wavelength is 0.025 Å at 200 kV compared with ~1 Å for X-rays. Under such conditions, the Ewald sphere in our experiments is nearly flat (the sphere is off the reciprocal plane by only 0.003 $Å^{-1}$ at 2 Å resolution) even in the high-resolution range. Measurements of the full width at half maximum intensity for the strongest reflections indicate that the reflections in our experiments are very tight, spreading less than a 6 pixel sphere that corresponds to ~1/1000 Å. (***Figure 5***) In our experiments, the shortest unit cell dimension for lysozyme in reciprocal space is a* = b* = 1/77 Å. Therefore, without beam oscillation or mechanical oscillation of the crystal (microscope compustage), the lattice points on a single projection that are not exactly at the Ewald sphere surface will give partial intensities.



Video 2. An example of spot prediction in diffraction data from a single crystal. Reflections predicted on representative diffraction patterns obtained from a single crystal tilted over 39° sampled every 2° in this video. Predictions were made to 2.9 Å resolution using our spot prediction algorithm.
DOI: 10.7554/eLife.01345.008

Because we densely sampled the reciprocal space, we recorded multiple observations for every lattice point (***Table 1***). Therefore, we could sample the observed intensity values for each reflection multiple times (multiplicity value = 34), and we made the assumption that the strongest intensity roughly approximated the complete intensity. Therefore, we kept only the maximum intensity and treated it as a unique reflection in the final structure factor file. All other recorded intensities were presumed to be partial reflections and were therefore discarded. The merging of data in P422 symmetry from three separate crystals processed in this manner resulted in a final data set with 2490 unique reflections with ~92% cumulative completeness at 2.9 Å resolution (***Table 1***, ***Video 3***). The measured intensities were converted to amplitudes by assuming $I_{hkl} \approx |F_{hkl}|^2$ (***Drenth, 1994***) and an mtz file generated.

The structure of lysozyme was solved at 2.9 Å resolution by molecular replacement (MR) using the lysozyme PDB 4AXT (***Cipriani et al., 2012***) as a search model. The initial MR $2F_{obs}-F_{calc}$ map prior to refinement is presented in ***Figure 6***. The map shows well-defined density around the model, indicating high quality phases from MR (***Figure 6A,B***). Likewise, a composite-omit map that was calculated by omitting 5% at a time showed good agreement with the original map obtained by MR (***Figure 6C***). When a poly-alanine (polyA) model of lysozyme was used for MR, the resulting map showed significant density beyond the alanine side chains (indicated by arrows in ***Figure 6E,F***), into which the correct side chains could be built. These results indicated that our solution from MR was not dominated by model bias.

Following refinement that included the use of electron scattering factors, rigid body, simulated annealing, and B-factor refinement, a solution was found with acceptable statistics (R_{work}/R_{free} = 25.5%/27.8%) and good geometry at 2.9 Å resolution (***Table 1***). The density map obtained by electron diffraction shows good agreement with the refined model (***Figure 7A***, ***Video 4***). The $F_{obs}-F_{calc}$ difference map shows no interpretable features (***Figure 7B***). Additionally, the final structure has a very low RMSD (0.475 Å for Cα, 0.575 Å for all atoms) when compared to the previously published high-resolution structure of lysozyme (***Cipriani et al., 2012***).

Model validation and bias tests

To further validate the method and test for model bias, we performed a number of tests on the data to check whether a good solution could be obtained from random noise as has been demonstrated for electron micrographs (***Shatsky et al., 2009***). We created multiple randomized datasets to test the robustness of the phasing and model building procedure. The test datasets were generated as follows:

1. All measured intensities were replaced with random numbers ranging between the minimum and maximum of the actual observed experimental values.
2. The experimental intensity values were kept but the Miller indices were randomized.
3. All experimental intensities were replaced with an actual intensity value that was measured by X-ray crystallography of an unrelated structure (Calmodulin PDB ID:3SUI [***Lau et al., 2012***]).
4. Each experimental intensity was increased or decreased randomly by up to 35%.

In addition, the correct experimental dataset was also used and labeled as dataset '5'. These five datasets were treated as 'blind test cases', in which the user did not know the identities of the various test datasets. Each test dataset was used for molecular replacement with the lysozyme model (***Cipriani et al., 2012***), followed by a single round of refinement in PHENIX (***Adams et al.,***

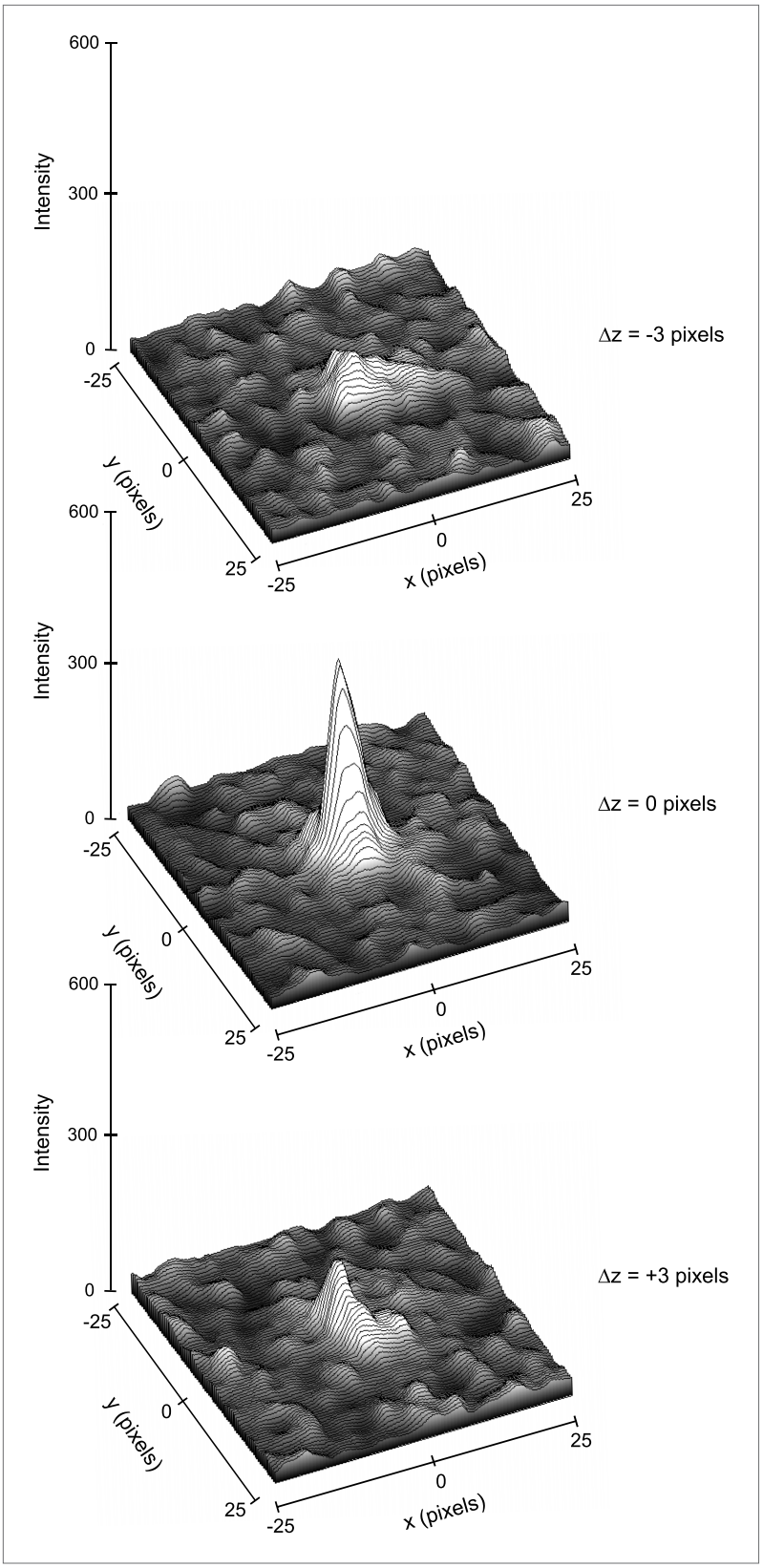


Figure 5. Three-dimensional profiles of the intensity of a single reflection over three consecutive diffraction patterns at −0.1°, 0°, and 0.1° degree tilts. The plots show the approximate dimensions of the full reflection with a width (full width at half maximum height) of 3–5 pixels in the x, y, and z direction.

DOI: 10.7554/eLife.01345.009

Table 1. MicroED crystallographic data

Data collection	
Excitation voltage	200 kV
Electron source	Field emission gun
Wavelength (Å)	0.025
Total electron dose per crystal	~9 $e^-/Å^2$
Number of patterns per crystal	40–90
No. crystals used	3
Total reflections to 2.9 Å	84,889
Data refinement	
Space group	$P4_32_12$
Unit cell dimensions	
a = b	77 Å
c	37 Å
α = β = γ	90°
Resolution	2.9–20.0 Å
Total unique reflections	2490
Reflections in working set	2240
Reflections in test set	250
Multiplicity*	34
Completeness (2.9–3.1)	92% (57%)
R_{work}/R_{free} (%)	25.5/27.8
RMSD bonds	0.051 Å
RMSD angles	1.587°
Ramachandran (%)† (allowed, generous, disallowed)	99.1; 0.9; 0

*Multiplicity is defined as total measured reflections divided by number of unique reflections.
†Statistics given by PROCHECK (***Laskowski et al., 1993***).
DOI: 10.7554/eLife.01345.010

2010). Only dataset 5, which contained the correct observed experimental intensities, yielded a solution that could be further refined to acceptable R_{work}/R_{free} and geometry. Datasets 1–4, which contained the random errors described above, either did not yield MR solutions or would not allow refinement to produce an acceptable structure (***Table 2***).

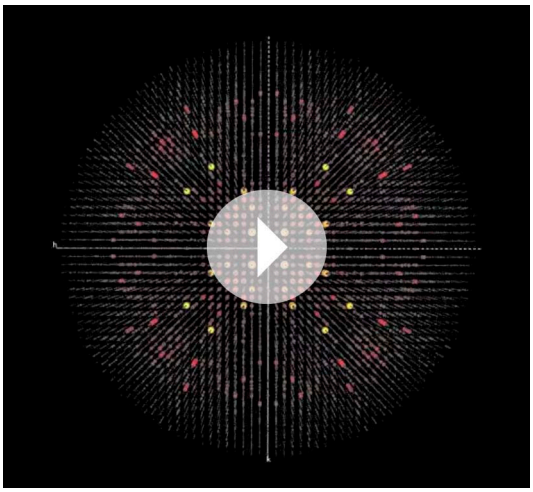
We also tested the robustness of the MR procedure by using a number of unrelated structures, chosen from the PDB for their similar unit cell dimensions and protein molecular weights, as search models against our experimental data. The unrelated structures were: T4 lysozyme, calmodulin, dodecin, and αA crystallin (***Table 3***). None of these structures gave an acceptable MR solution.

Together, these experiments indicate that the extracted intensities are accurate enough to yield a reliable structure and that model bias originating from MR did not skew our results.

Completeness and accuracy of the measured intensities in electron diffraction

Data sets collected in electron crystallography of 2D crystals suffer from a missing cone due to the limitation of the maximum achievable tilt angle in the TEM. Previous reports estimate that with tilt angles up to 60°, the missing cone is roughly 13% (***Glaeser et al., 1989***), and the resolution in plane is typically higher than the resolution perpendicular to the tilt axis (z*). In our experiments, because the data from 3 crystals were eventually used, and the orientation of each crystal on the grid varied, we could cover the full reciprocal space (***Video 3***).

Dynamic scattering likely introduces inaccuracies in the electron diffraction data. In electron diffraction, dynamic scattering (multi scattering events) could redistribute primary reflection intensities, reducing the accuracy of the intensity measurements by randomly contributing to the observed intensities (***Grigorieff et al., 1996***). The lysozyme crystals have $P4_32_12$, symmetry and systematic absences are expected at (2n+1,0,0). However, very weak reflections were observed at the positions where absences were expected (***Figure 8***). It is likely that these reflections originate from dynamic scattering events. We plotted the intensities along the a* and b* axes and compared the intensity values. The intensities of Miller indices (2n+1,0,0) and (0,2n+1,0) were measured and compared to the intensities of the four immediately adjacent reflections (2n+2,1,0), (2n+2,−1,0), (2n−2,1,0), and (2n−2,−1,0). On average, the intensity in the systematic absences was found to be 4.9% of the total intensity of the adjacent spots. (Standard deviation 2.7%, Max 12.4%, n = 17). Moreover, comparison of our experimental intensities with intensities obtained by X-ray diffraction of lysozyme of the same crystal form indicates that our data follow a similar trend and are not dominated by intensity randomness. A Pearson correlation coefficient between the two data sets was 0.63 from 6.0 to 13.5 Å (0.56 from 2.0–13.5 Å), indicating conservation of reflection hierarchy—strong intensities remain strong and weak intensities remain weak. Together, our analyses suggest that multiple scattering contributes at maximum roughly 10% to the intensity value and that at least for structure determination at 2.9 Å resolution such an error in intensity appears to be tolerable. It is possible that dynamic scattering will become a significant source of error at higher resolutions and some correction algorithm will then have to be developed.



Video 3. Three-dimensional representation of merged intensity values. 2490 total unique reflections are present for an overall completeness of 92% at 2.9 Å resolution. Video begins with a* axis horizontal, b* axis vertical, and the c* axis normal to the image plane.
DOI: 10.7554/eLife.01345.011

Discussion

We present a method, 'MicroED', for structure determination by electron crystallography. It should be widely applicable to both soluble and membrane proteins as long as small, well-ordered crystals can be obtained. We have shown that diffraction data at atomic resolution can be collected and a structure determined from crystals that are up to 6 orders of magnitude smaller in volume than those typically used for X-ray crystallography.

For difficult targets such as membrane proteins and multi-protein complexes, screening often produces microcrystals that require a great deal of optimization before reaching the size required for X-ray crystallography. Sometimes such size optimization becomes an impassable barrier. Electron diffraction of microcrystals as described here offers an alternative, allowing this roadblock to be bypassed and data to be collected directly from the initial crystallization hits.

While our proof of principle is an important first step, further optimization of the method is required. Better programs need to be developed for accurately determining lattice parameters, indexing all reflections, extracting the intensities and correcting for incomplete intensities, dynamic scattering, and Ewald sphere curvature. Specifically, developing procedures for postrefinement (unit cell refinement, estimating the mosaic spread, rocking curve, etc) should allow for the proper correction and scaling of partially recorded reflections, leading to improved estimation of full intensities. Relatively minor modifications to existing programs such as MOSFLM (***Leslie and Powell, 2007***) should allow the handling of electron diffraction data from 3D crystals and take advantage of the large body of work already dedicated to processing X-ray diffraction data.

The accuracy of the microscope compustage can be improved and procedures for crystal or beam oscillation implemented. Our method of using the maximum intensity measurement as an approximation of the full intensity of any given spot is admittedly crude, as it depends on the intersection of the Ewald sphere through the center of each spot at some point in the tilt series. As the resolution increases, this event becomes increasingly unlikely. Crystal oscillation or related methods such as precession of the electron beam (***Gjønnes et al., 1998***) would allow more accurate determination of spot intensities, especially at very high resolutions.

Further development of various methods for phasing the diffraction data are also required and could possibly include heavy metal phasing. Such phasing methods are standard in X-ray crystallography and rely on differences in intensity values between a native data set and heavy metal derivative data sets. It is possible that in electron crystallography dynamic scattering could hinder phasing by such methods, and new algorithms will need to be developed to make this possible. Phase extension from projection maps or from low-resolution density maps can also be used for direct phasing (***Gipson et al., 2011***; ***Wisedchaisri and Gonen, 2011***). It is also possible that single particle cryo-EM could be used for direct phasing as previously demonstrated where a low-resolution single particle map was used to phase X-ray diffraction data (***Speir et al., 1995***; ***Dodson, 2001***; ***Xiong, 2008***). Moreover, a double tilt cryo holder as well as newly developed goniometer-based grid holders could be used to cover more of the Fourier space. Finally this method could benefit from automation in data collection.

This first study serves as a proof of principle that three-dimensional electron diffraction can yield an accurate protein structure from microcrystals. As additional protocols and programs are developed, MicroED promises to advance the field of structural biology and open the door to many exciting new studies.

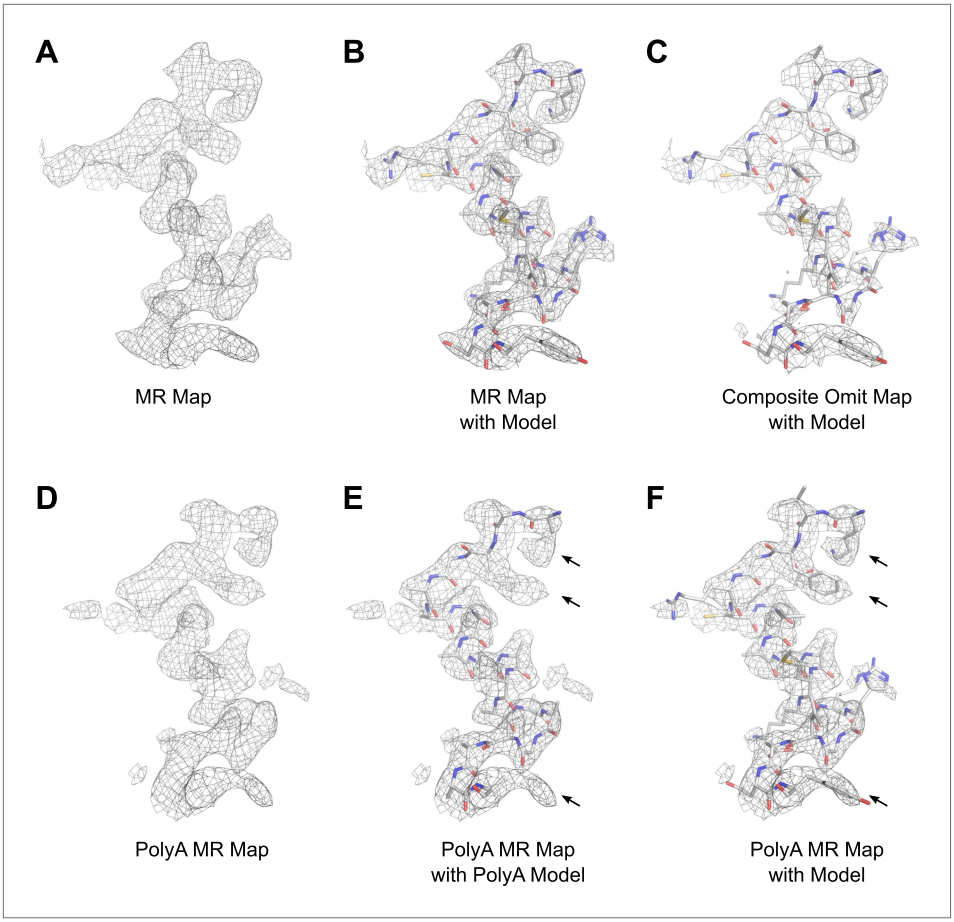


Figure 6. Results of phasing by molecular replacement prior to crystallographic refinement. Molecular replacement was performed with both the full model of lysozyme (PDB 4AXT, top panels) as well as a poly-alanine model (bottom panels) and the resulting $2F_{obs}-F_{calc}$ maps around residues 1–20 are shown. (**A** and **B**) The phases following molecular replacement with the full model were of good quality demonstrated by how well the density surrounding the model fits, even before any refinement is performed. (**C**) A composite-omit map calculated by omitting 5% at a time showed good agreement with the unrefined structure indicating the phases were not dominated by model bias. (**D**–**F**) As an additional test of model bias, phasing was done with a poly-alanine homology search model of lysozyme. The resulting $2F_{obs}-F_{calc}$ map is of good quality (**D**) and shows density extending beyond the poly-alanine model (**E** and **F**, arrows). (**F**) The same density map as **E** but with the structure of lysozyme fit. Arrows in **D** and **E** show examples of clear side chain density from the poly-alanine map. All maps are contoured at 1.0σ.
DOI: 10.7554/eLife.01345.012

Materials and methods

Lysozyme crystallization and sample preparation

Lysozyme was purchased from Fisher Scientific and a 200 mg/ml solution was prepared in 50 mM sodium acetate pH 4.5. Lysozyme solution was mixed 1 to 1 with precipitant solution (3.5M sodium chloride; 15% PEG 5,000; 50 mM sodium acetate pH 4.5) and crystals were grown by the hanging drop method. Following the crystal formation, the sample was diluted three to five times in 5% PEG 200. A 5 µl drop of the crystal solution was applied to a quantifoil 2/2 holey-carbon copper EM grid. The grid was then blotted and vitrified by plunging into liquid ethane using a Vitrobot Mark IV (FEI). The frozen-hydrated grid was loaded onto a Gatan 626 cryo-holder and transferred to a cryo-TEM.

Electron diffraction

All electron microscopy was performed on a FEI Tecnai F20 TEM equipped with a field emission electron source (FEG) and operating at 200 kV. Electron diffraction pattern tilt series data were recorded

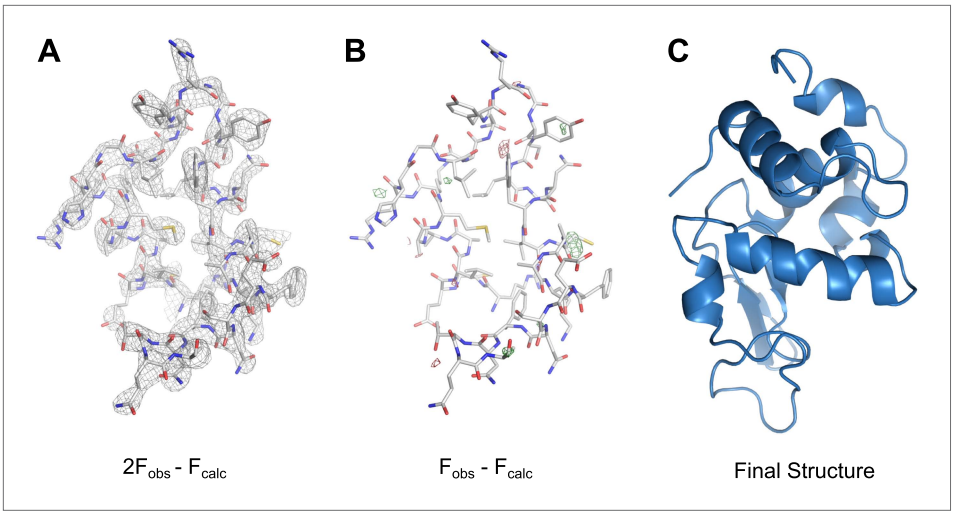
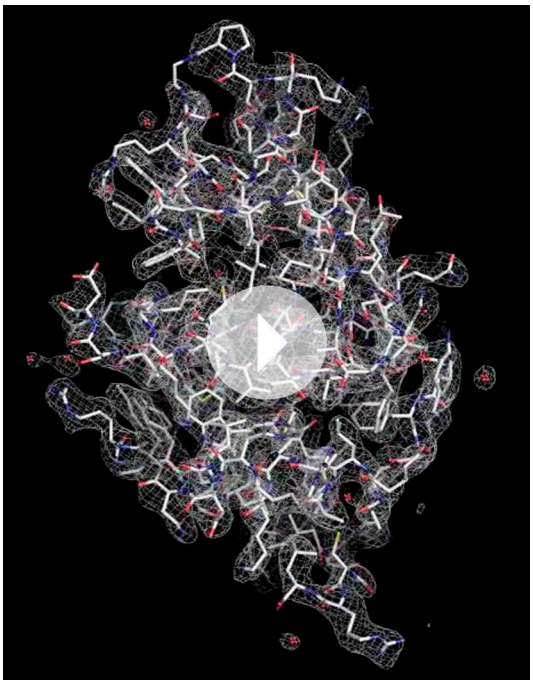


Figure 7. MicroED structure of lysozyme at 2.9 Å resolution. (**A**) The $2F_{obs}-F_{calc}$ (contoured at 1.5σ) map covers protein residues 5–45 of lysozyme. (**B**) $F_{obs}-F_{calc}$ difference map contoured at +3.0σ (green) and −3.0σ (red) for the same protein region. The map (**A**) shows well-defined density around the vast majority of side chains and the difference map (**B**) shows no large discrepancies between the observed data (F_{obs}) and the model (F_{calc}). The final structure of lysozyme is shown in panel **C** and the complete three-dimensional map is presented in ***Video 3***.
DOI: 10.7554/eLife.01345.013

with a bottom mount TVIPS F416 4 k × 4 k CMOS camera with pixel size 15.6 μm using built in series exposure mode. The electron dose was kept below 0.01 $e^{-}/Å^2$ per second, and each frame of a data set was taken with an exposure time of up to 10 s per frame. The electron dosage was calibrated with the use of a Faraday cage as well as by calibrating the counts on the CMOS detector in bright field mode. Each data set consisted of up to 90 still frames taken at 0.1–1° intervals with a maximum total dose of ~$9e^{-}/Å^2$ per crystal. The camera length was optimized for the desired resolution as described previously (***Gonen, 2013***).



Video 4. $2F_{obs}-F_{calc}$ density around the complete lysozyme model at 2.9 Å resolution (contoured at 1.5σ).
DOI: 10.7554/eLife.01345.014

Data processing

Although our original intent was to perform all data analysis with existing X-ray crystallography software various incompatibilities and logistical roadblocks necessitated the development of some additional tools. Diffraction patterns were indexed and background subtracted intensities extracted and merged with in-house developed software implemented in python using methods adapted from those developed by ***Shi et al. (1998)***.

Briefly, measurements were made on images identified as major planes of the crystal with ImageJ and used to determine the approximate magnitudes of the unit cell vectors a*, b*, and c* and the angles between them (α, β, and γ). Subsequently, 100 to 350 spots were chosen across several images from each set of diffraction patterns. Vectors in reciprocal space were calculated for all of the selected spots. Difference vectors between spot vectors were calculated allowing vectors approximating the estimated unit cell lengths to be identified. The angles between potential unit cell vectors were calculated

Table 2. Results of model validation and bias tests

Data set	Molecular replacement result	TFZ	Final R_{free} (%)¶
1*	No solution	N/A	N/A
2†	Solution**	19.1	54.9
3‡	No solution	N/A	N/A
4§	Solution	12.6	35.2
5#	Solution	14.7	27.8

*Random intensities.
†Shuffled Miller indices.
‡Calmodulin replaced intensities.
§Intensities ± 35%.
#Original data.
¶Final R_{free} after a minimum of two cycles of refinement.
**Solution was found; however, the space group was incorrect ($P4_12_1$).
DOI: 10.7554/eLife.01345.015

and 'orthogonal triplets' identified. Orthogonal triplets are defined as sets of vectors that contain a predicted a*, b*, and c*, which are all 90° from each other (α = β = γ = 90° for this crystal). All sets of the orthogonal triplets were averaged to yield estimated a*, b*, and c* vectors. The estimated a*, b*, and c* vectors were then refined by identifying parallel difference vectors derived from the original selected spots with lengths that were multiples of the unit cell lengths.

The calculated unit cell vectors were then used to predict the spots in each diffraction pattern. Two reference spots were chosen for each image and their Miller indices calculated using the previously

Table 3. Models for molecular replacement validation

Protein	PDB ID	Molecular weight (kDa)	Symmetry	Unit cell dimensions	MR solution
Hen Egg White Lysozyme*	4AXT	14.3	$P4_32_12$	a = b = 78.24 Å	Yes
				c = 37.47 Å	
				α = β = γ = 90°	
T4 Lysozyme†	2LZM	18.7	$P3_212$	a = b = 61.20 Å	No
				c = 96.80 Å	
				α = β = 90°	
				γ = 120°	
Calmodulin‡	3CLN	16.7	P1	a = 29.71 Å,	No
				b = 53.79 Å,	
				c = 24.99 Å	
				α = 94.13°,	
				β = 97.57°,	
				γ = 89.46°	
Dodecin§	4B2J	8.5	$F4_132$	a = b = c = 142.90 Å	No
				α = β = γ = 90°	
αA Crystallin#	3L1E	11.9	$P4_12_12$	a = b = 56.22 Å,	No
				c = 68.66 Å	
				α = β = γ = 90°	

*__Cipriani et al. (2012)__.
†__Weaver and Matthews (1987)__.
‡__Babu et al. (1988)__.
§__Staudt et al. (2013)__.
#__Laganowsky et al. (2010).__
DOI: 10.7554/eLife.01345.016

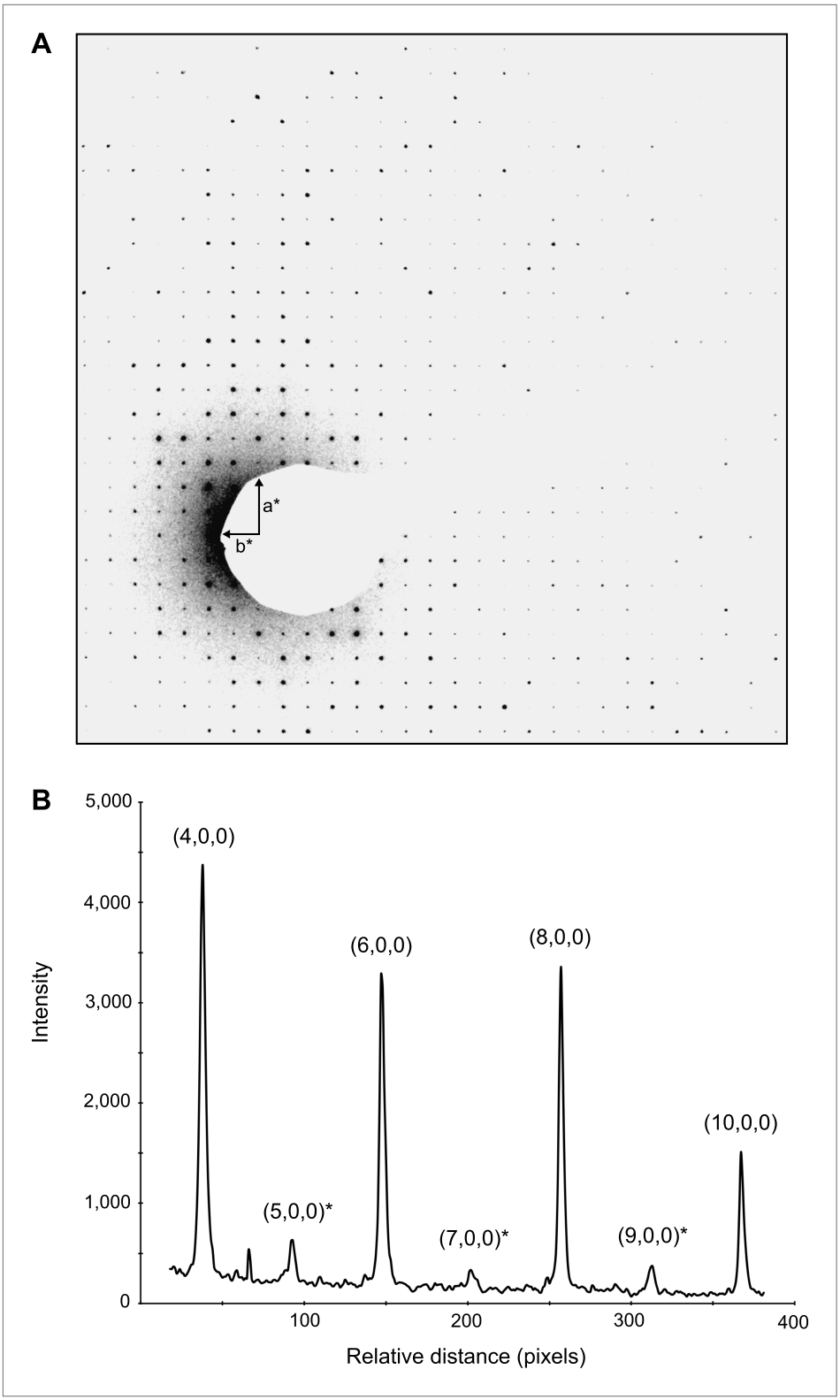


Figure 8. Dynamic scattering in lysozyme 3D crystals. Intensity measurement along the a* axis of a raw diffraction pattern illustrating the relatively small contributions due to dynamic scattering. (**A**) Diffraction pattern from the major plane of a lysozyme crystal with visible intensity in the (2n+1,0,0) and (0,2n+1,0) Miller indices. (**B**) (2n+1, 0, 0) reflections (starred) are expected to be systematically absent and observed intensities at these indices are assumed to be the result of dynamic scattering. Image contrast was enhanced for clarity using ImageJ.

DOI: 10.7554/eLife.01345.017

determined unit cell vectors. For every diffraction pattern, the vector normal to the detector plane was calculated as:

$$\mathbf{r}_1 \times \mathbf{r}_2 = \mathbf{n}$$

where $\mathbf{r}_1$ and $\mathbf{r}_2$ are the vectors defined by the Miller indices from reference spots one and two, respectively, and $\mathbf{n}$ is the resulting vector normal to the detector plane. Any reflection that appears on a given diffraction pattern will satisfy:

$$\mathbf{n} \cdot \mathbf{v} = 0$$

where $\mathbf{v}$ is any set of Miller indices. For any h, k, l that satisfied the above equation, within a defined threshold, that particular reflection was predicted to appear on the diffraction image, and its x, y detector coordinates on the diffraction pattern image were calculated.

Intensities for each predicted reflection were integrated by first drawing both a square and a circular mask centered on the reflection, with the diameter of the circle identical to the length of the square. The mean pixel intensity outside the circle but within the square was calculated yielding the mean background intensity. The mean background was then subtracted from each pixel within the circle, and the resulting pixel intensities were summed. All related intensities from three data sets were grouped based on P422 symmetry. The maximum value for each group of equivalent reflections was assumed to best approximate the full intensity and was used for that reflection in the final data set. Because each intensity measurement ultimately originated from a single observation, SigI and SigF values were estimated as the square root of the intensity and square root of the structure factor, respectively. The final mtz file contains columns h, k, l, F, SIGF, I, SIGI. The final data set contained 2490 unique reflections from 2.9–20 Å with cumulative completeness of 92% (***Table 1***).

Structure refinement

Phaser (***McCoy et al., 2007***) was used to obtain phases with lysozyme structure 4AXT (***Cipriani et al., 2012***) as a MR search model (LLG = 372 and TFZ = 14.7). The structure was then refined using CNS (***Brünger et al., 1998***) and PHENIX (***Adams et al., 2010***) by rounds of rigid body, simulated annealing, and B-factor refinement. The R_{free} data set represented 10% of the total data set. The data were subjected to twinning analysis; however, twinning with this symmetry group is forbidden and therefore we ruled out twinning in our crystals. Electron scattering factors (***Gonen et al., 2005***) were used during refinement.

Data deposition and software availability

The structure factors and coordinates of the final model were deposited in the Protein Data Bank with accession code 3J4G. The in house developed program that was used for processing the MicroED data is available for download at http://www.github.com/gonenlab/microED.git.

Acknowledgements

The authors wish to thank members of the Gonen lab and Goragot Wisedchaisri for helpful discussions. We also would like to thank Nikolaus Grigorieff (HHMI, Janelia Farm Research Campus) for helpful discussions and suggestions and for critically reading this manuscript. Shane Gonen (JFRC, UW) assisted with the preparation of ***Video 3***. This work is dedicated to the memory of Prof K H Kuo.

Additional information

Funding

Funder	Author
Howard Hughes Medical Institute	Dan Shi, Brent L Nannenga, Matthew G Iadanza, Tamir Gonen

The funder had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

DS, BLN, MGI, TG, Conception and design, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article

Additional files

Major dataset

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset ID and/or URL	Database, license, and accessibility information
Shi D, Nannenga B.L, Iadanza M.G, Gonen T	2013	Structure of lysozyme solved by MicroED to 2.9Å	3J4G; http://www.rcsb.org/pdb/explore/explore.do?structureId=3j4g	Publicly available at RCSB Protein Data Bank (http://www.rcsb.org).

References

Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW, Echols N, et al. 2010. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr D Biol Crystallogr* **66**:213–21. doi: 10.1107/S0907444909052925.

Babu YS, Bugg CE, Cook WJ. 1988. Structure of calmodulin refined at 2.2 A resolution. *J Mol Biol* **204**:191–204. doi: 10.1016/0022-2836(88)90608-0.

Baker LA, Smith EA, Bueler SA, Rubinstein JL. 2010. The resolution dependence of optimal exposures in liquid nitrogen temperature electron cryomicroscopy of catalase crystals. *J Struct Biol* **169**:431–7. doi: 10.1016/j.jsb.2009.11.014.

Blake CC, Fenn RH, North AC, Phillips DC, Poljak RJ. 1962. Structure of lysozyme. A Fourier map of the electron density at 6 angstrom resolution obtained by x-ray diffraction. *Nature* **196**:1173–6. doi: 10.1038/1961173a0.

Blake CC, Koenig DF, Mair GA, North AC, Phillips DC, Sarma VR. 1965. Structure of hen egg-white lysozyme. A three-dimensional Fourier synthesis at 2 Angstrom resolution. *Nature* **206**:757–61. doi: 10.1038/206757a0.

Boutet S, Lomb L, Williams GJ, Barends TR, Aquila A, Doak RB, et al. 2012. High-resolution protein structure determination by serial femtosecond crystallography. *Science* **337**:362–4. doi: 10.1126/science.1217737.

Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, et al. 1998. Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr D Biol Crystallogr* **54**:905–21. doi: 10.1107/S0907444998003254.

Chapman HN, Fromme P, Barty A, White TA, Kirian RA, Aquila A, et al. 2011. Femtosecond X-ray protein nanocrystallography. *Nature* **470**:73–7. doi: 10.1038/nature09750.

Cipriani F, Rower M, Landret C, Zander U, Felisaz F, Marquez JA. 2012. CrystalDirect: a new method for automated crystal harvesting based on laser-induced photoablation of thin films. *Acta Crystallogr D Biol Crystallogr* **68**:1393–9. doi: 10.1107/S0907444912031459.

Diamond R. 1974. Real-space refinement of the structure of hen egg-white lysozyme. *J Mol Biol* **82**:371–91. doi: 10.1016/0022-2836(74)90598-1.

Dodson EJ. 2001. Using electron-microscopy images as a model for molecular replacement. *Acta Crystallogr D Biol Crystallogr* **57**:1405–9. doi: 10.1107/S0907444901013415.

Drenth J. 1994. *Principles of protein X-ray crystallography*. New York: Springer-Verlag.

Gipson BR, Masiel DJ, Browning ND, Spence J, Mitsuoka K, Stahlberg H. 2011. Automatic recovery of missing amplitudes and phases in tilt-limited electron crystallography of two-dimensional crystals. *Phys Rev E Stat Nonlin Soft Matter Phys* **84**:011916. doi: 10.1103/PhysRevE.84.011916.

Gjønnes J, Hansen V, Berg BS, Runde P, Cheng YF, Gjønnes K, et al. 1998. Structure model for the phase AlmFe derived from three-dimensional electron diffraction intensity data collected by a precession technique. Comparison with convergent-beam diffraction. *Acta Crystallogr A* **54**:306–19. doi: 10.1107/S0108767397017030.

Glaeser RM. 1971. Limitations to significant information in biological electron microscopy as a result of radiation damage. *J Ultrastruct Res* **36**:466–82. doi: 10.1016/S0022-5320(71)80118-1.

Glaeser RM, Tong L, Kim SH. 1989. Three-dimensional reconstructions from incomplete data: interpretability of density maps at "atomic" resolution. *Ultramicroscopy* **27**:307–18. doi: 10.1016/0304-3991(89)90021-1.

Gonen T. 2013. The collection of high-resolution electron diffraction data. *Methods Mol Biol* **955**:153–69. doi: 10.1007/978-1-62703-176-9_9.

Gonen T, Cheng Y, Sliz P, Hiroaki Y, Fujiyoshi Y, Harrison SC, et al. 2005. Lipid-protein interactions in double-layered two-dimensional AQP0 crystals. *Nature* **438**:633–8. doi: 10.1038/nature04321.

Grigorieff N, Ceska TA, Downing KH, Baldwin JM, Henderson R. 1996. Electron-crystallographic refinement of the structure of bacteriorhodopsin. *J Mol Biol* **259**:393–421. doi: 10.1006/jmbi.1996.0328.

Henderson R. 1995. The potential and limitations of neutrons, electrons and x-rays for atomic-resolution microscopy of unstained biological molecules. *Q Rev Biophys* **28**:171–93. doi: 10.1017/S003358350000305X.

Henderson R, Baldwin JM, Ceska TA, Zemlin F, Beckmann E, Downing KH. 1990. Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J Mol Biol* **213**:899–929. doi: 10.1016/S0022-2836(05)80271-2.

Henderson R, Unwin PN. 1975. Three-dimensional model of purple membrane obtained by electron microscopy. *Nature* **257**:28–32. doi: 10.1038/257028a0.

Jiang LH, Georgieva D, Nederlof I, Liu ZF, Abrahams JP. 2011. Image processing and lattice determination for three-dimensional nanocrystals. *Microsc Microanal* **17**:879–85. doi: 10.1017/S1431927611012244.

Kimura Y, Vassylyev DG, Miyazawa A, Kidera A, Matsushima M, Mitsuoka K, et al. 1997. Surface of bacteriorhodopsin revealed by high-resolution electron crystallography. *Nature* **389**:206–11. doi: 10.1038/38323.
Kuhlbrandt W, Wang DN, Fujiyoshi Y. 1994. Atomic model of plant light-harvesting complex by electron crystallography. *Nature* **367**:614–21. doi: 10.1038/367614a0.
Laganowsky A, Benesch JL, Landau M, Ding L, Sawaya MR, Cascio D, et al. 2010. Crystal structures of truncated alphaA and alphaB crystallins reveal structural mechanisms of polydispersity important for eye lens function. *Protein Sci* **19**:1031–43. doi: 10.1002/pro.380.
Laskowski RA, Macarthur MW, Moss DS, Thornton JM. 1993. *PROCHECK* - a program to check the stereochemical quality of protein structures. *J Appl Crystallogr* **26**:283–91. doi: 10.1107/S0021889892009944.
Lau SY, Procko E, Gaudet R. 2012. Distinct properties of Ca2+-calmodulin binding to N- and C-terminal regulatory regions of the TRPV1 channel. *J Gen Physiol* **140**:541–55. doi: 10.1085/jgp.201210810.
Leslie AGW, Powell HR. 2007. Processing diffraction data with MOSFLM. *Nato Sci Ser Ii Math* **245**:41–51. doi: 10.1007/978-1-4020-6316-9_4.
McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. 2007. *Phaser* crystallographic software. *J Appl Crystallogr* **40**:658–74. doi: 10.1107/S0021889807021206.
Moukhametzianov R, Burghammer M, Edwards PC, Petitdemange S, Popov D, Fransen M, et al. 2008. Protein crystallography with a micrometre-sized synchrotron-radiation beam. *Acta Crystallogr D Biol Crystallogr* **64**:158–66. doi: 10.1107/S090744490705812X.
Nederlof I, van Genderen E, Li YW, Abrahams JP. 2013. A Medipix quantum area detector allows rotation electron diffraction data collection from submicrometre three-dimensional protein crystals. *Acta Crystallogr D Biol Crystallogr* **69**:1223–30. doi: 10.1107/S0907444913009700.
Sauter C, Otalora F, Gavira JA, Vidal O, Giege R, Garcia-Ruiz JM. 2001. Structure of tetragonal hen egg-white lysozyme at 0.94 A from crystals grown by the counter-diffusion method. *Acta Crystallogr D Biol Crystallogr* **57**:1119–26. doi: 10.1107/S0907444901008873.
Shatsky M, Hall RJ, Brenner SE, Glaeser RM. 2009. A method for the alignment of heterogeneous macromolecules from electron microscopy. *J Struct Biol* **166**:67–78. doi: 10.1016/j.jsb.2008.12.008.
Shi D, Lewis MR, Young HS, Stokes DL. 1998. Three-dimensional crystals of Ca2+-ATPase from sarcoplasmic reticulum: merging electron diffraction tilt series and imaging the (h, k, 0) projection. *J Mol Biol* **284**:1547–64. doi: 10.1006/jmbi.1998.2283.
Speir JA, Munshi S, Wang G, Baker TS, Johnson JE. 1995. Structures of the native and swollen forms of cowpea chlorotic mottle virus determined by X-ray crystallography and cryo-electron microscopy. *Structure* **3**:63–78. doi: 10.1016/S0969-2126(01)00135-6.
Staudt H, Hoesl MG, Dreuw A, Serdjukow S, Oesterhelt D, Budisa N, et al. 2013. Directed manipulation of a flavoprotein photocycle. *Angew Chem Int Ed Engl* **52**:8463–6. doi: 10.1002/anie.201302334.
Tani K, Mitsuma T, Hiroaki Y, Kamegawa A, Nishikawa K, Tanimura Y, et al. 2009. Mechanism of aquaporin-4's fast and highly selective water conduction and proton exclusion. *J Mol Biol* **389**:694–706. doi: 10.1016/j.jmb.2009.04.049.
Taylor KA, Glaeser RM. 1976. Electron microscopy of frozen hydrated biological specimens. *J Ultrastruct Res* **55**:448–56. doi: 10.1016/S0022-5320(76)80099-8.
Unwin PN, Henderson R. 1975. Molecular structure determination by electron microscopy of unstained crystalline specimens. *J Mol Biol* **94**:425–40. doi: 10.1016/0022-2836(75)90212-0.
Weaver LH, Matthews BW. 1987. Structure of bacteriophage T4 lysozyme refined at 1.7 A resolution. *J Mol Biol* **193**:189–99. doi: 10.1016/0022-2836(87)90636-X.
Wisedchaisri G, Gonen T. 2011. Fragment-based phase extension for three-dimensional structure determination of membrane proteins by electron crystallography. *Structure* **19**:976–87. doi: 10.1016/j.str.2011.04.008.
Wisedchaisri G, Reichow SL, Gonen T. 2011. Advances in structural and functional analysis of membrane proteins by electron crystallography. *Structure* **19**:1381–93. doi: 10.1016/j.str.2011.09.001.
Xiong Y. 2008. From electron microscopy to X-ray crystallography: molecular-replacement case studies. *Acta Crystallogr D Biol Crystallogr* **64**:76–82. doi: 10.1107/S090744490705398X.

The structure of purified kinetochores reveals multiple microtubule-attachment sites

Shane Gonen[1,2,6], Bungo Akiyoshi[2,3,5,6], Matthew G Iadanza[1,4,6], Dan Shi[4], Nicole Duggan[2], Sue Biggins[2] & Tamir Gonen[1,4]

Chromosomes must be accurately partitioned to daughter cells to prevent aneuploidy, a hallmark of many tumors and birth defects. Kinetochores are the macromolecular machines that segregate chromosomes by maintaining load-bearing attachments to the dynamic tips of microtubules. Here, we present the structure of isolated budding-yeast kinetochore particles, as visualized by EM and electron tomography of negatively stained preparations. The kinetochore appears as an ~126-nm particle containing a large central hub surrounded by multiple outer globular domains. In the presence of microtubules, some particles also have a ring that encircles the microtubule. Our data, showing that kinetochores bind to microtubules via multivalent attachments, lay the foundation to uncover the key mechanical and regulatory mechanisms by which kinetochores control chromosome segregation and cell division.

The generation and survival of all organisms requires the precise partitioning of duplicated chromosomes to daughter cells. Defects in segregation lead to aneuploidy, the state in which entire chromosomes are gained or lost. Aneuploidy is a hallmark of tumor cells and has been postulated to be a major factor in the evolution of cancer, and it is also the leading cause of spontaneous miscarriages and hereditary birth defects[1,2]. It is therefore critical to understand the mechanisms that ensure accurate chromosome segregation and thus maintain genomic stability.

Chromosome segregation requires forces, generated by spindle microtubules, that are translated into chromosome movement through interactions with kinetochores, highly conserved structures assembled from distinct subcomplexes[3–6]. The simplest kinetochore is in budding yeast, where 38 core structural proteins assemble onto centromeric DNA to form a single microtubule-binding site (**Fig. 1a**)[7]. Because most subcomplexes are present in multiple copies, the simplest kinetochore contains more than 250 core proteins as well as additional regulatory proteins. The majority of yeast kinetochore proteins are conserved, and it is thought that, in multicellular eukaryotes, kinetochores that bind to multiple microtubules may simply contain repeat units of the budding-yeast kinetochore[3,8,9]. The inner kinetochore contains subcomplexes that directly bind to centromeric DNA, whereas the outer kinetochore is composed of subcomplexes that mediate microtubule attachment. The major microtubule-binding activity of the kinetochore is mediated by KMN, an assembly of the KNL-1, Mis12 and Ndc80 subcomplexes, which attaches to microtubules cooperatively[10]. The yeast-specific Dam1 complex also exhibits microtubule-binding activity, and it has been proposed that the vertebrate Ska1 complex may be an ortholog of this[11–14].

Although a number of models have been proposed[15–21], the structure of the kinetochore and the mechanism by which it attaches to microtubules is still not clear. Elegant fluorescence-microscopy studies have shown that the overall positioning and stoichiometry of kinetochore components is highly conserved[9,17,18,22], leading to a proposal for overall kinetochore architecture (**Fig. 1a**). However, it has been difficult to obtain higher-resolution information about complete kinetochores. The prevailing picture from EM studies on vertebrate cells reveals that the kinetochore is a three-tiered structure[23–26]. More recent studies have visualized an outer-kinetochore network connected to microtubules, supporting a model whereby multiple weak attachment sites mediate coupling activity[21]. In one study, peeling microtubule protofilaments could be seen attached to fibrils at the inner kinetochore, which led to the proposal that these fibrils could couple chromosome movement to microtubule depolymerization[20]. The Dam1 complex forms rings around microtubules *in vitro* at high concentrations, supporting proposals that suggest rings as the major coupling mechanism[27,28].

Visualization of the attachment state of kinetochores requires the isolation of large kinetochore assemblies that can be imaged at higher resolution. Although progress has been made in elucidating the structure of recombinant kinetochore subcomplexes, the subcomplexes have not yet been reconstituted into larger assemblies suitable for structural work. We previously developed an assay to purify native budding-yeast kinetochore particles that contain the majority of core structural components and, under force, can maintain attachments to microtubules[29]. Here we set out to analyze these assemblies by EM, in both the presence and absence of microtubules, to obtain information about their structure.

[1]Howard Hughes Medical Institute, Department of Biochemistry, University of Washington, Seattle, Washington, USA. [2]Division of Basic Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. [3]Molecular and Cellular Biology Program, University of Washington, Seattle, Washington, USA. [4]Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, Virginia, USA. [5]Present address: Sir William Dunn School of Pathology, University of Oxford, Oxford, UK. [6]These authors contributed equally to this work. Correspondence should be addressed to S.B. (sbiggins@fhcrc.org) or T.G. (gonent@janelia.hhmi.org).

Received 8 June; accepted 9 July; published online 12 August 2012; doi:10.1038/nsmb.2358

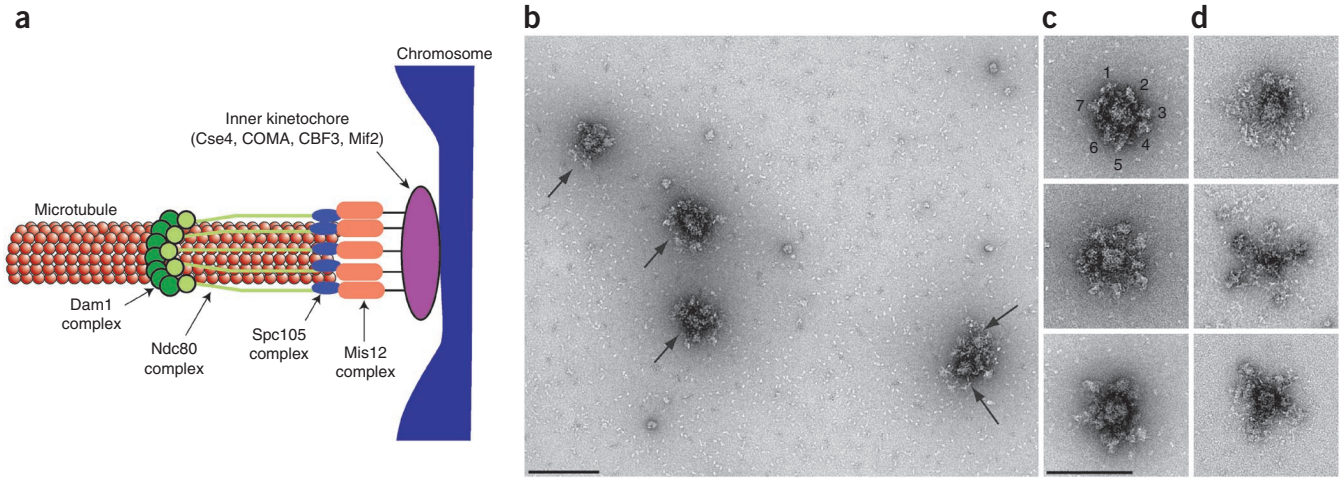


Figure 1 Kinetochore particles contain a central hub surrounded by a number of globular domains. (**a**) A model for the budding-yeast kinetochore shows that multiple copies of the Dam1, Ndc80, KNL-1 (Spc105) and Mis12 kinetochore subcomplexes mediate binding of the chromosome (blue) to the microtubule. The inner kinetochore contains one or more copies of the Cse4, COMA, CBF3 and Mif2 subcomplexes. (**b**) A field of kinetochore particles in microtubule-polymerization buffer was visualized by EM of negatively stained preparations. Five particles (arrows) and other small material are apparent. Note that two particles are touching. Scale bar, 200 nm. (**c**) Images of representative compact kinetochore particles in microtubule-polymerization buffer with lower salt. The globular domains on a single particle in the top panel are numbered. Scale bar, 150 nm. (**d**) The particles are more extended in higher-salt buffer used for purification. Additional particles are presented in **Supplementary Figure 1**.

RESULTS

Kinetochores contain a hub surrounded by globular domains

Native kinetochore particles were isolated by affinity capture of the kinetochore component Dsn1-Flag on beads and were eluted with a Flag peptide[29]. To avoid potential cell-cycle variability, kinetochores were purified from cells arrested in mitosis by the addition of the microtubule-depolymerizing drug benomyl. Eluates were negatively stained and imaged on an electron microscope. The largest structures visible on the grids were approximately 126 ± 13 nm in length, end to end (n = 88; **Fig. 1b**, arrows). Smaller particles (that may or may not represent kinetochore subcomplexes) were also visible in the background because of the low-stringency purification conditions needed to maintain functional and intact kinetochores[29,30] (**Fig. 1b**). As a negative control, we purified Dsn1-Flag from *ndc80-1* mutant cells with abolished kinetochore function[31]. We previously found that these particles cannot bind to microtubules and lack most of the outer kinetochore, as assayed by silver-stained SDS-PAGE[29]. Consistent with this, large particles were not detected in these eluates, and we instead observed material of variable size and shape on the grids for the *ndc80-1* samples (**Supplementary Fig. 1**). Additional controls were prepared from cells lacking a Flag epitope–tagged protein or expressing the Alk1-Flag protein that does not associate with kinetochores. The large particles were

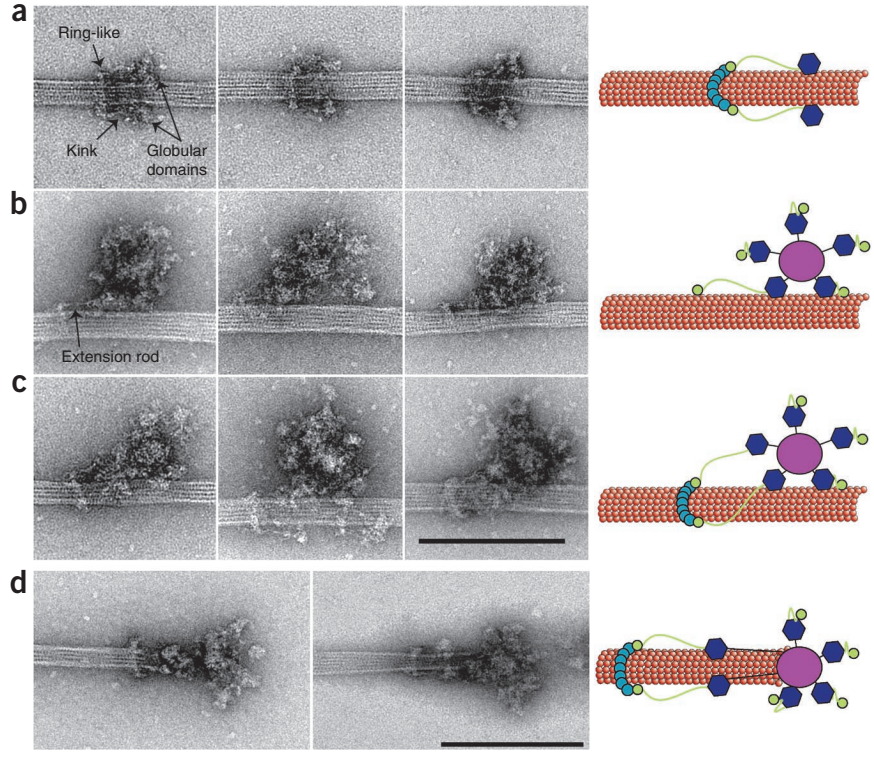


Figure 2 Kinetochore particles bound to taxol-stabilized microtubules. (**a**) Representative images of fragments of kinetochore particles (56 nm long) bound to taxol-stabilized microtubules reveal a rod with a kink connected to a ring on one end and a globular domain on the opposite end. (**b**) Large kinetochore particles (126 nm long) bind to microtubules through globular domains and an additional extension rod (arrow) that emanates from one of the globular domains. (**c**) Large kinetochore particles bind to microtubules through multiple globular domains and contain an extension that connects to a ring. Scale bar, 200 nm. (**d**) Two selected images of kinetochores at the tip of taxol-stabilized microtubules. Globular domains extending 50 nm from the central hub bind to the microtubules and are connected to a distal ring 50 nm away. Scale bar, 200 nm. Cartoons at right schematize the key features of the images.

Figure 3 Three-dimensional structures of two types of kinetochore particles bound to a microtubule. (**a**,**b**) Projection images of the two types of kinetochore assemblies on taxol-stabilized microtubules selected for tomographic reconstruction. Representative slices through each particle are presented with the corresponding segmentation analysis in color underneath. Finally, three-dimensional reconstruction of this complex by electron tomography is presented on the right. All slices are arranged from left to right, traversing through the *z* axis from bottom to top, with the bottom defined at the grid surface (carbon layer). Both structures contain a ring-like structure (blue arrowheads) and multiple rods (white arrowheads). Scale bars, 100 nm (all panels). **Supplementary Movies 1 and 2** are presented in the online version of this paper and show the reconstructed 3D density.

missing from these controls, consistent with the initial identification of the large structures as kinetochore particles (data not shown).

The kinetochores contained a 37 ± 3–nm central hub (n = 72) surrounded by a number of globular domains of variable shape and with an average diameter of 21 ± 2 nm (n = 97, **Fig. 1b–d**, **Supplementary Fig. 1**). We analyzed the appearance of the kinetochores in two different buffer conditions and found that they appear more compact when incubated in a lower-salt buffer compatible with microtubule polymerization (**Fig. 1c**). The majority of kinetochores (n = 54) contained five globular domains, although particles with as many as seven globular domains (n = 7) were also seen (**Fig. 1c**,**d**, **Supplementary Table 1**). Particles visualized in buffer containing a higher salt concentration appeared more extended and had a maximum of five globular domains radiating (**Fig. 1d**). These data suggest that the particles either lose structural integrity at high salt concentrations or are structurally flexible and can undergo large conformational rearrangements. All of the measurements reported in the manuscript were therefore performed on particles that had been incubated in the lower-salt microtubule-polymerization buffer.

Kinetochore assemblies bound to microtubules

We next visualized the kinetochore particles bound to microtubules. When taxol-stabilized microtubules were incubated with Flag eluate, distinct assemblies became enriched on the microtubules (**Fig. 2** and **Supplementary Fig. 2**). One observed assembly contained a rod-shaped structure oriented parallel to the microtubule, with globular domains on one end and a ring-like structure oriented orthogonally to the microtubule at the opposite end (**Fig. 2a** and **Supplementary Table 2**). The average length of the complex from the ring-like structure to each globular domain was 56 ± 4 nm (n = 128). The ring-like structure had an average outer diameter of 50 ± 3 nm (n = 99; **Supplementary Table 2**). There was a kink in the rod an average of 25 ± 2 nm (n = 67) away from the ring (**Fig. 2a**, arrow), and the globular domains often appeared to contact the microtubule.

The other assemblies detected on microtubules contained the 126-nm structures bound to microtubules in either the absence (**Fig. 2b**) or presence (**Fig. 2c**) of the ring-like structure shown in **Figure 2a**. In both cases, at least two of the globular domains that radiated from the central hub appeared to contact the microtubule, suggesting that these domains contain microtubule-binding elements of the kinetochore. The contacts between the globular domains and the microtubules were more apparent in the three-dimensional tomographic reconstructions (**Fig. 3** and see below). In sharp contrast, the central hub never appeared to contact the microtubule directly. When the ring-like structure was missing, a rod structure that appeared similar to the rod in **Figure 2a** could be seen extending from the 126-nm particle (**Fig. 2b**, arrow). When a ring-like structure was present, this rod touched it in a similar fashion as in the assemblies shown in **Figure 2a**. It is therefore most likely that the assemblies in **Figure 2a** result from a portion of the larger particle falling off the microtubule or else represent a subset of smaller microtubule-binding kinetochore assemblies present in the eluate.

We also observed a few kinetochores at microtubule tips (**Fig. 2d**). In these cases, binding was only observed in the presence of a ring-like structure (n = 8). The structure was similar to the lateral attachments shown in **Figure 2c** because they also contained a ring connected to a rod with a large globular domain. However, the particle appeared to extend an additional ~50 nm from the globular domains to the tip of the microtubule. This suggests that the linkage between globular

domains and the central hub is flexible, and globular domains can extend outward from the central hub, consistent with our observation that the particles may possess flexible elements (**Fig. 1c**,**d**).

Particles purified from mutants lack kinetochore structure

We next attempted to assign elements of the kinetochore to specific kinetochore subcomplexes. Repeated attempts at defining components by immunogold labeling and negative-stain EM have so far not been successful (data not shown), so we used temperature-sensitive kinetochore-protein mutants to analyze the corresponding changes in appearance, as an alternative approach. In eluates from wild-type cells, 13% of the large particles ($n = 410$) bound to microtubules also contained a ring-like structure; in contrast, although particles purified from *dad1-1* mutant cells that are defective in the Dam1 complex[32] were visually indistinguishable from wild-type particles in the absence of microtubules, they completely lacked ring-like structures when bound to microtubules ($n = 175$, **Supplementary Table 3** and **Supplementary Fig. 1**). The lack of the ring-like structure in the *dad1-1* mutant particles bound to microtubules is statistically significant ($P = 0.001$) and strongly suggests that the rings correspond to the Dam1 complex. Consistent with this, previous studies have determined that the Dam1 complex is not required for lateral attachments to microtubules *in vivo*[33] or for kinetochore particles to bind to microtubules *in vitro*[29]. Eluates from *ndc80-1* mutant cells lacked all of the microtubule-binding complexes that were visualized with wild-type particles in **Figure 2** (**Supplementary Fig. 1**). Whereas wild-type eluates showed 0.68 particles per micrometer of microtubule (335 particles on a total of 493 μm observed), there were virtually none found for *ndc80-1* cells (three particles on a total of 585 μm observed), a difference that is highly statistically significant ($P < 0.0001$) and consistent with the role of the outer kinetochore in making microtubule attachments.

Tomography reveals rings encircling the microtubule

We next used electron tomography and image processing to calculate the three-dimensional reconstruction of two particles representative of those shown (**Fig. 2a**,**c**, **Fig. 3** and **Supplementary Movies 1** and **2**). Slices through these data identify at least two prominent features. First, a ring encircles the microtubule. Although it is not clear whether all the ring-like structures we observe by EM fully encircle the microtubule, these particles at least contain complete rings around the microtubule (**Fig. 3**, blue arrowheads). Second, although only two rod-like structures were observed in projection (**Fig. 3a**,**b**), multiple rod-like structures were connected to the ring-like structure in the tomogram. Although the stain flattens the assemblies so volumes for each component cannot be calculated accurately, the rods appear to be relatively evenly spaced around the ring-like structure.

DISCUSSION

Here, we describe the structure of isolated kinetochore particles from budding yeast. The general architecture reveals a central hub surrounded by globular domains. When microtubules are present, at least one globular domain and an extension emanating from it make contact with the microtubule lattice. The extension is not seen in the absence of microtubules, which is consistent with a dynamic change in kinetochore structure in the presence of microtubules.

We propose the following model to describe the kinetochore particles we visualized by EM (**Fig. 4**). Recombinant Dam1 complex forms rings with an outer diameter of 50 nm[27,28] around microtubules, similar to the ring-like structures we observe in the presence of microtubules. These ring-like structures were never seen on the particles in the absence of microtubules, consistent with the substoichiometric amounts of

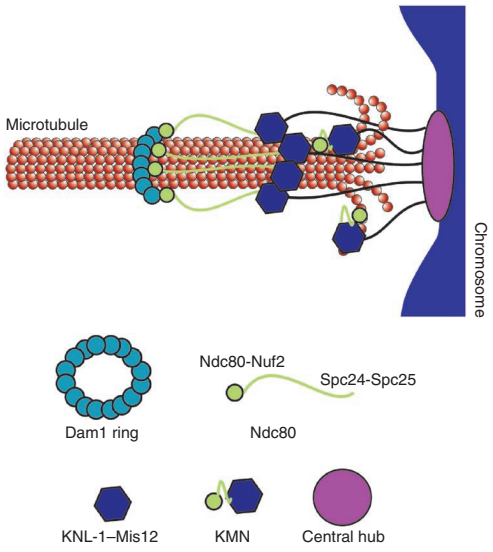


Figure 4 Schematic of the proposed model of kinetochore architecture. The central globular domain binds to the centromeric locus of the chromosome, and globular domains containing the KMN complex extend to attach to the microtubule. The Ndc80 subcomplex makes an additional extension to contact a distal ring composed of the Dam1 subcomplex.

Dam1-complex proteins in our kinetochore preparations and the requirement for microtubules in loading Dam1 on kinetochores[12,29]. It is likely that there is either soluble or kinetochore-bound Dam1 that nucleates ring-like structures around microtubules at very low concentrations in the presence of other kinetochore components.

Given that Dam1 facilitates the function of the conserved Ndc80 complex on microtubules[34,35], we propose that the extended rod-shaped structure connected to the ring is Ndc80. Consistent with this, the ends of the Ndc80 complex extend away from each other in the presence of microtubules *in vivo*[18]. The average length of the rod we observed is 56 nm, and the rod contains a kink, similar to what was observed for recombinant Ndc80 complex[36,37]. Although the kink we observed is in a different average position than for recombinant Ndc80 complexes, our measurements were made in the presence of microtubules in the context of larger assemblies. When the orientation *in vivo* of the Ndc80 complex relative to the Dam1 ring is taken into account[17], the Nuf2-Ndc80 head of the Ndc80 complex would be positioned at the ring. At this time, the resolution is not high enough for us to determine whether the head is interior or exterior to the ring, although our tomographic reconstructions suggest that the latter may be true. Our data are consistent with the possibility that the Ndc80 CH domain, rather than the Ndc80 loop, interacts with Dam1 (ref. 38). This orientation means that the large globular entities at the opposite end of the rods would contain the Spc24 and Spc25 proteins; however, additional proteins must also be present to account for the size of the globular domain. We therefore propose that components of the Mis12 and/or KNL-1 subcomplexes that are known to bind to the Ndc80 complex to form KMN[10], which contains the core microtubule-binding activity of the kinetochore, are located in these large globular domains. In this model, KMN would contain two microtubule-binding sites, one composed of the Ndc80-Nuf2 head that can be seen extending from a subset of the globular domains to touch the microtubule, and the other containing Spc24 and Spc25 bound to Mis12–KNL-1 complexes. This model is consistent with the cooperative microtubule-binding behavior of KMN[10]. Given that the central hub never appears to associate with the microtubule, we propose that the chromosomal

attachment site is mediated through this region. To date, we have not visualized nucleic acid in the particles by EM. At this time, it is not possible to estimate the stoichiometry of components within the particles, but we favor the possibility that each globular domain represents a single KMN complex. In the future, it will be important to identify the position of each kinetochore protein within the particles to determine the precise architecture of the structures.

In summary, we have visualized isolated kinetochore particles and found that they appear as 126-nm particles containing a central hub with multiple outer globular domains. Our images exhibit some similarity to tomograms of laterally attached vertebrate kinetochores *in vivo*[21]. Because two or more of the globular domains bind to the microtubule, kinetochores appear to interact with microtubules via multivalent attachments that are flexible to move along microtubules, consistent with a biased diffusion mechanism[39]. We propose that when the particle is bound to a chromosome and the tip of a dynamic microtubule, a greater number of globular domains might bind to the microtubule to stabilize the interaction and maintain larger forces (**Fig. 4**). In this case, the distance from the central hub to the outer region of the kinetochore would be consistent with measurements of tip-attached kinetochores *in vivo*[17,18]. The filaments observed at the ends of kinetochore microtubules *in vivo*[20] could correspond to the rod-like structures we have observed along the microtubules or to connections between the globular domains and the central hub. Because yeast kinetochores bind to a single microtubule, a ring-based mechanism likely ensures processivity but is not required for direct attachment[40]. Together, these studies lay the foundation for future high-resolution structural and mechanistic studies aimed at understanding how the kinetochore ensures accurate chromosome segregation during cell division.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Structure deposited to EMDB reference number EMD-2154 and EMD-2153.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We are grateful to members of the Biggins and Gonen laboratories for valuable discussions and for comments on the manuscript. We are also grateful to C. Asbury, A. Powers, B. Stoddard and J. Al-Bassam for discussion and comments on the manuscript. This work was supported by US National Institutes of Health grants (GM078079 and GM064386 to S.B.), a US National Cancer Institute Cancer Center Support grant (CA015704 to S.B.) and the Howard Hughes Medical Institute (T.G.).

AUTHOR CONTRIBUTIONS

All authors contributed to designing the research. B.A., N.D. and S.B. performed the kinetochore purifications. S.G., M.G.I. and D.S. collected the EM data and did the EM analysis. S.B. and S.G. performed the microtubule-binding experiments. T.G. and S.B. analyzed the data and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/doifinder/10.1038/nsmb.2358.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Pfau, S.J. & Amon, A. Chromosomal instability and aneuploidy in cancer: from yeast to man. *EMBO Rep.* **13**, 515–527 (2012).
2. Compton, D.A. Mechanisms of aneuploidy. *Curr. Opin. Cell Biol.* **23**, 109–113 (2011).
3. Santaguida, S. & Musacchio, A. The life and miracles of kinetochores. *EMBO J.* **28**, 2511–2531 (2009).
4. Cheeseman, I.M. & Desai, A. Molecular architecture of the kinetochore-microtubule interface. *Nat. Rev. Mol. Cell Biol.* **9**, 33–46 (2008).
5. Takeuchi, K. & Fukagawa, T. Molecular architecture of vertebrate kinetochores. *Exp. Cell Res.* **318**, 1367–1374 (2012).
6. DeLuca, J.G. & Musacchio, A. Structural organization of the kinetochore-microtubule interface. *Curr. Opin. Cell Biol.* **24**, 48–56 (2012).
7. Westermann, S., Drubin, D.G. & Barnes, G. Structures and functions of yeast kinetochore complexes. *Annu. Rev. Biochem.* **76**, 563–591 (2007).
8. Zinkowski, R.P., Meyne, J. & Brinkley, B.R. The centromere-kinetochore complex: a repeat subunit model. *J. Cell Biol.* **113**, 1091–1110 (1991).
9. Joglekar, A.P. *et al.* Molecular architecture of the kinetochore-microtubule attachment site is conserved between point and regional centromeres. *J. Cell Biol.* **181**, 587–594 (2008).
10. Cheeseman, I.M., Chappie, J.S., Wilson-Kubalek, E.M. & Desai, A. The conserved KMN network constitutes the core microtubule-binding site of the kinetochore. *Cell* **127**, 983–997 (2006).
11. Hofmann, C. *et al. Saccharomyces cerevisiae* Duo1p and Dam1p, novel proteins involved in mitotic spindle function. *J. Cell Biol.* **143**, 1029–1040 (1998).
12. Li, Y. *et al.* The mitotic spindle is required for loading of the DASH complex onto the kinetochore. *Genes Dev.* **16**, 183–197 (2002).
13. Welburn, J.P. *et al.* The human kinetochore Ska1 complex facilitates microtubule depolymerization-coupled motility. *Dev. Cell* **16**, 374–385 (2009).
14. Hanisch, A., Sillje, H.H. & Nigg, E.A. Timely anaphase onset requires a novel spindle and kinetochore complex comprising Ska1 and Ska2. *EMBO J.* **25**, 5504–5515 (2006).
15. Asbury, C.L., Tien, J.F. & Davis, T.N. Kinetochores' gripping feat: conformational wave or biased diffusion? *Trends Cell Biol.* **21**, 38–46 (2011).
16. Schittenhelm, R.B. *et al.* Spatial organization of a ubiquitous eukaryotic kinetochore protein network in *Drosophila* chromosomes. *Chromosoma* **116**, 385–402 (2007).
17. Joglekar, A.P., Bloom, K. & Salmon, E.D. *In vivo* protein architecture of the eukaryotic kinetochore with nanometer scale accuracy. *Curr. Biol.* **19**, 694–699 (2009).
18. Wan, X. *et al.* Protein architecture of the human kinetochore microtubule attachment site. *Cell* **137**, 672–684 (2009).
19. Welburn, J.P. & Cheeseman, I.M. Toward a molecular structure of the eukaryotic kinetochore. *Dev. Cell* **15**, 645–655 (2008).
20. McIntosh, J.R. *et al.* Fibrils connect microtubule tips with kinetochores: a mechanism to couple tubulin dynamics to chromosome motion. *Cell* **135**, 322–333 (2008).
21. Dong, Y., Vanden Beldt, K.J., Meng, X., Khodjakov, A. & McEwen, B.F. The outer plate in vertebrate kinetochores is a flexible network with multiple microtubule interactions. *Nat. Cell Biol.* **9**, 516–522 (2007).
22. Johnston, K. *et al.* Vertebrate kinetochore protein architecture: protein copy number. *J. Cell Biol.* **189**, 937–943 (2010).
23. Brinkley, B.R. & Stubblefield, E. The fine structure of the kinetochore of a mammalian cell in vitro. *Chromosoma* **19**, 28–43 (1966).
24. Jokelainen, P.T. The ultrastructure and spatial organization of the metaphase kinetochore in mitotic rat cells. *J. Ultrastruct. Res.* **19**, 19–44 (1967).
25. Roos, U.P. Light and electron microscopy of rat kangaroo cells in mitosis. II. Kinetochore structure and function. *Chromosoma* **41**, 195–220 (1973).
26. Rieder, C.L. The structure of cold stable kinetochore microtubules in metaphase PtK1 cells. *Chromosoma* **84**, 145–158 (1981).
27. Miranda, J.J., De Wulf, P., Sorger, P.K. & Harrison, S.C. The yeast DASH complex forms closed rings on microtubules. *Nat. Struct. Mol. Biol.* **12**, 138–143 (2005).
28. Westermann, S. *et al.* Formation of a dynamic kinetochore- microtubule interface through assembly of the Dam1 ring complex. *Mol. Cell* **17**, 277–290 (2005).
29. Akiyoshi, B. *et al.* Tension directly stabilizes reconstituted kinetochore-microtubule attachments. *Nature* **468**, 576–579 (2010).
30. Akiyoshi, B., Nelson, C.R., Ranish, J.A. & Biggins, S. Quantitative proteomic analysis of purified yeast kinetochores identifies a PP1 regulatory subunit. *Genes Dev.* **23**, 2887–2899 (2009).
31. Wigge, P.A. *et al.* Analysis of the *Saccharomyces* spindle pole by matrix-assisted laser desorption/ionization (MALDI) mass spectrometry. *J. Cell Biol.* **141**, 967–977 (1998).
32. Enquist-Newman, M. *et al.* Dad1p, third component of the Duo1p/Dam1p complex involved in kinetochore function and mitotic spindle integrity. *Mol. Biol. Cell* **12**, 2601–2613 (2001).
33. Tanaka, K. *et al.* Molecular mechanisms of kinetochore capture by spindle microtubules. *Nature* **434**, 987–994 (2005).
34. Tien, J.F. *et al.* Cooperation of the Dam1 and Ndc80 kinetochore complexes enhances microtubule coupling and is regulated by aurora B. *J. Cell Biol.* **189**, 713–723 (2010).
35. Lampert, F., Hornung, P. & Westermann, S. The Dam1 complex confers microtubule plus end-tracking activity to the Ndc80 kinetochore complex. *J. Cell Biol.* **189**, 641–649 (2010).
36. Wei, R.R., Sorger, P.K. & Harrison, S.C. Molecular organization of the Ndc80 complex, an essential kinetochore component. *Proc. Natl. Acad. Sci. USA* **102**, 5363–5367 (2005).
37. Wang, H.W. *et al.* Architecture and flexibility of the yeast Ndc80 kinetochore complex. *J. Mol. Biol.* **383**, 894–903 (2008).
38. Alushin, G.M. *et al.* The Ndc80 kinetochore complex forms oligomeric arrays along microtubules. *Nature* **467**, 805–810 (2010).
39. Hill, T.L. Theoretical problems related to the attachment of microtubules to kinetochores. *Proc. Natl. Acad. Sci. USA* **82**, 4404–4408 (1985).
40. Westermann, S. *et al.* The Dam1 kinetochore ring complex moves processively on depolymerizing microtubule ends. *Nature* **440**, 565–569 (2006).

ONLINE METHODS

Yeast strains, plasmids and microbial techniques. Media and genetic and microbial techniques were performed essentially as described[41]. Mitotic cultures were prepared with benomyl as described[30]. For temperature-sensitive mutants, cells were shifted to 37 °C for 3 h. Yeast strains and plasmids used in this study are listed in **Supplementary Table 4**. The 3× Flag epitope–tag strains were made using a PCR-based integration system and were confirmed by PCR[42–44]. The 6× His–3× Flag epitope tagging of the endogenous *DSN1* gene was performed with a PCR-based integration system using primers SB2434 and SB2435 and plasmid pSB1590 as a template[29]. All tagged strains we constructed are functional *in vivo* and do not cause any detectable growth defects or temperature sensitivity. Specific primer sequences are available upon request.

Isolation of kinetochore particles. Native kinetochore particles were isolated from budding yeast as described[29]. Briefly, 2 liters of yeast cells (SBY8253 or relevant strain) expressing Dsn1-Flag or Dsn1-His-Flag were arrested in mitosis with 60 µg/ml of the microtubule-depolymerizing drug benomyl for 3 h and harvested. Cells were lysed in Buffer H (25 mM HEPES, pH 8.0, 2 mM $MgCl_2$, 0.1 mM EDTA, 0.5 mM EGTA, 150 mM KCl, 15% glycerol and 0.1% NP-40) supplemented with protease and phosphatase inhibitors, and kinetochore particles were captured with anti-Flag antibodies and eluted with 40 µl of Buffer H containing 0.5 mg/ml Flag peptide. The eluted material was used directly for negative-stain EM studies as described below (note that this buffer resulted in less-compact structures, as shown in **Fig. 1d**). Alternatively, the eluate was prepared for microtubule-binding experiments as described below (note that these kinetochores appeared more compact, as shown in **Fig. 1c**). All measurements in the paper were performed on particles that had been incubated for microtubule-binding experiments.

EM and image processing. All samples were prepared for negative-stain EM as previously described[45], with the following modifications. A 3-µl drop of Flag eluate was applied to a negatively charged carbon-coated copper grid (Gilder 400 or 200 mesh) for 20 s and washed with a single drop of water, followed by two drops of freshly prepared 0.75% uranyl formate. Samples containing microtubules were treated similarly but were applied to positively charged copper grids. Specimens were screened on either a 100-kV transmission electron microscope (TEM) (Morgagni, FEI) or a 120 kV TEM (Spirit T12, FEI). Images were recorded using a Gatan slow-scan bottom-mount charge-coupled device (CCD) camera at a nominal magnification of 18,000× at the specimen level. Measurements were taken in either the Digital Micrograph suite (Gatan, v. 1.71.38) or ImageJ64 (v 1.43).

Electron tomography. Negatively stained samples prepared as above were coated with a second layer of carbon by evaporation, and gold-conjugated anti–mouse IgG (5–10 nm) (Sigma-Aldrich) were added as fiducial markers. Tilt series were collected using a Spirit T12 120 kV transmission electron microscope (FEI Company). Images were recorded using a Gatan slow-scan 4,000 × 4,000 bottom-mount CCD with a pixel size of 4.3 Å at the sample level (52,000×). Tilt series were recorded from −70° to +70° with an increment of 2° at 2 µm defocus. Three-dimensional reconstructions were calculated using Amira (v 5.3.1)[46] and IMOD (v. 4.1.9)[47] software.

Microtubule-binding experiments. Taxol-stabilized microtubules were prepared freshly as described[48]. A 200-µl aliquot was centrifuged at 58,000 r.p.m. (Beckman TLA 100.1 rotor) at 37 °C for 10 min. The supernatant was decanted, and the pellet was used for microtubule-kinetochore binding experiments as follows. Two µl of Dsn1-Flag eluate was mixed with 7 µl of the microtubule pellet and incubated at RT for 10 min. The sample was then diluted with 200 µl warm BTAX (80 mM PIPES, pH 6.9, 1 mM $MgCl_2$, 1 mM EGTA and 10 µM Taxol, 37 °C). Grids for EM were prepared as described above. Note that excess tubulin dimers could be seen on the grids, owing to the microtubule polymerization buffer. Images were recorded on a CCD camera using either the 100-kV TEM or the 120-kV TEM at a nominal magnification range of 14,000×–36,000× at the specimen level. Measurements were taken either in the Digital Micrograph suite (Gatan, v 1.71.38) or ImageJ64 (v 1.43).

Quantification of microtubule binding. A total of 100 microtubules (for the *dad1-1* mutant) or 200 microtubules (for wild type and *ndc80-1*), ranging in size from 0.5 to 9 microns, were assayed for the number of large particles bound in the presence or absence of a ring (**Supplementary Table 3**). Eluates from SBY9047 or SBY7441 were used for all microtubule-binding experiments with wild-type kinetochore particles.

41. Rose, M.D., Winston, F. & Heiter, P. *Methods in Yeast Genetics*, 198 (Cold Spring Harbor Laboratory Press, 1990).
42. Longtine, M.S. *et al.* Additional modules for versatile and economical PCR-based gene deletion and modification in *Saccharomyces cerevisiae*. *Yeast* **14**, 953–961 (1998).
43. Gelbart, M.E., Rechsteiner, T., Richmond, T.J. & Tsukiyama, T. Interactions of Isw2 chromatin remodeling complex with nucleosomal arrays: analyses using recombinant yeast histones and immobilized templates. *Mol. Cell. Biol.* **21**, 2098–2106 (2001).
44. Sikorski, R.S. & Hieter, P. A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in *Saccharomyces cerevisiae*. *Genetics* **122**, 19–27 (1989).
45. Ohi, M., Li, Y., Cheng, Y. & Walz, T. Negative Staining and Image Classification - Powerful Tools in Modern Electron Microscopy. *Biol. Proced. Online* **6**, 23–34 (2004).
46. Stalling, D., Westerhoff, M. & Hege, H.-C. *The Visualization Handbook* (Elsevier, 2005).
47. Kremer, J.R., Mastronarde, D.N. & McIntosh, J.R. Computer visualization of three-dimensional image data using IMOD. *J. Struct. Biol.* **116**, 71–76 (1996).
48. Franck, A.D. *et al.* Tension applied through the Dam1 complex promotes microtubule elongation providing a direct mechanism for length control in mitosis. *Nat. Cell Biol.* **9**, 832–837 (2007).

doi:10.1038/nsmb.2358