aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 1
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

ADVANCES IN IMAGING AND ELECTRON PHYSICS, VOL. 150

Ptychography and Related Diffractive Imaging Methods

J.M. RODENBURG[1]

[1] *Department of Electronic and Electrical Engineering, University of Sheffield, Mappin Street, S1 3JD, United Kingdom*

I. Introduction

Ptychography is a nonholographic solution of the phase problem. It is a method of calculating the phase relationships between different parts of a scattered wave disturbance in a situation where only the magnitude (intensity or flux) of the wave can be physically measured. Its usefulness lies in its ability (like holography) to obtain images without the use of lenses, and hence to lead to resolution improvements and access to properties of the scattering medium (such as the phase changes introduced by the object) that cannot be easily obtained from conventional imaging methods. Unlike holography, it does not

ISSN 1076-5670
DOI: 10.1016/S1076-5670(07)00003-1

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 2
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

need a stable reference wave; the only interferometer used in ptychography is diffraction interference occurring in the object itself.

The concept of ptychography was originated by the late Walter Hoppe between about 1968 and 1973 (Hoppe, 1969a, 1969b; Hoppe and Strube, 1969; Hegerl and Hoppe, 1970, 1972). In its original formulation it was difficult to contemplate how to implement an actual experimental proof of the idea at short (atomic scale) wavelengths, other than for a crystalline object in a scanning transmission electron microscope (STEM). Hoppe and Strube (1969) demonstrated the idea experimentally using visible light optics, but at the time that Hoppe was thinking about these ideas, STEMs were simply not sufficiently developed (particularly the detectors available) to undertake the necessary measurements. At that time, the benefits of the technique were ill defined and offered no clear improvements over existing imaging, diffraction, and holographic methods. In fact, when a proof of principle was eventually established by Nellist *et al.* (1995), it was discovered that ptychography did have one uniquely and profoundly important advantage over all other phase-retrieval or imaging techniques: it is not subject to the limitations of the coherence envelope (the "information limit"), which today is still regarded as the last hurdle in optimal electron imaging.

In view of the recent advances in iterative solution of the ptychographic data set (see Section IV.C) and their experimental implementation in visible light optics and hard X-ray imaging, a review of the neglected strengths of the ptychographical principle is of benefit. Recent years have seen a growing interest in iterative solutions of the phase problem. In most modern work, Hoppe's pioneering thoughts have been entirely neglected. This is unfortunate for two reasons. First is the scholarly misrepresentation of the precedence of certain key ideas in the history of the field of phase problems. Second, a true appreciation of what ptychography encompasses, especially with respect to "conventional" diffractive imaging methods, means that a great opportunity has yet to be fully exploited. This Chapter attempts to demonstrate these advantages are and why this Cinderella of the phase-retrieval story may be on the brink of revolutionizing short-wavelength imaging science.

II. Ptychography in Context

This section places ptychography in the context of other phase-retrieval methods in both historical and conceptual terms.

A. History

Recovery of the lost phase (or relative phase) by some calculation method is still considered a problem because when these issues were first addressed

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 3
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

(before World War II), it appeared that in certain scattering geometries (especially those most easily accessible experimentally) a solution could not exist, and if it did exist, it would be very ill conditioned mathematically. X-ray crystallography was the first experimental discipline to admit to the consequences of lost phase. In fact, given that large quantities of stereochemical *a priori* information is often available, the crystalline diffraction phase problem is much more tractable than the generalized phase problem. Indeed, it could be argued that Watson and Crick's discovery of the structure of DNA (Watson and Crick, 1953) was itself an ad hoc solution to a phase problem – they had access to the intensity of the scattered X-ray diffraction pattern from a crystal composed of DNA molecules and essentially found a "solution" (a physical distribution of atoms) that was consistent with that intensity and with the known stereochemistry constraints of the constituent amino acids. This use of *a priori* information is an important characteristic of all solutions to the phase problem.

The key advantage of phase-retrieval methods in the context of imaging is that image formation can be unshackled from the difficulties associated with the manufacture of high-quality lenses of large numerical aperture. In some fields, such as stellar interferometry (Michelson, 1890), it is only possible to record a property related to the intensity of the Fourier transform of the object. The creation of an image from these data is then a classic phase problem (in this case, rather easily solvable given that the *a priori* information that the object in question is finite, of positive intensity, and is situated on an empty background). In the field of visible wavelength optical microscopy, little need exists to solve the phase problem because very good lenses of very large numerical aperture can be relatively easily manufactured (although a phase problem still exists in the sense that the image itself is recorded only in intensity). For short-wavelength microscopies (using X-rays or high-energy electrons), good-quality lenses are not easily manufactured. For example, the field of electron imaging has recently seen great progress in the manufacture of non-round lenses correctors (for which spherical aberration need not be positive, as with a round magnetic lens), a scheme first proposed by Scherzer (1947) but which has only recently become feasible with the advent of powerful inexpensive computers that can be interfaced to very high-stability power supplies. However, the cost of such systems is high (typically of the order of $1000 per corrector), and the gains in resolution over existing round magnetic lenses are only of an order of a factor of 2 or so. In the case of X-ray microscopy, the zone plate remains the predominant lens technology, but this technology requires manufacturing precision of at least the final requisite resolution of the lens itself; making a sufficiently thick zone plate with very high lateral definition (< 50 nm) difficult, especially for hard X-rays. For example, compare the best resolution obtainable from such a lens

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 4
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

(also ∼50 nm) with the "resolution" of conventional X-ray diffraction using the same wavelength; notwithstanding the fact that the object of interest (the unit cell) must be amenable to crystallization, the effective resolution of the "image" (an estimate of that unit cell) is roughly the same as the wavelength: ∼0.1 nm.

In the case of aperiodic objects (unique structures), solution of the phase problem has evolved via several diverse routes in communities that have historically had little interaction. The most elegant direct solution – indeed, so direct that in its original form it does not require any mathematical computation at all – is holography. Gabor (1948) first formulated this concept in an amazingly concise and insightful short letter with the title "A New Microscopical Principle." His interest was in overcoming the rather profound theoretical limitations of round electron lenses, a fact that had been realized by Scherzer (1936) before the war and eloquently quantified by him shortly after the war (Scherzer, 1949). It was clear that aberrations intrinsic in the electron lens could not be overcome (although, as already discussed, later work on non-round lenses have circumvented these difficulties). Gabor's idea is briefly discussed here, but it should be noted that the use of a strong and known reference can record an analog of the phase of a wave field. Reconstruction of the actual electron image was envisaged as a two-step process, the second step performed using visible light optics that mimicked the aberrations present in the electron experiment. This is an example of a phase problem solution in conjunction with the use of a "poor" lens so as to improve the performance of that lens. In fact, the current popular perception of holography is as a method of generating the illusion of three-dimensional (3D) images; the 3D information in ptychography is discussed in Section IV.B.

The pure diffraction intensity (PDI) phase problem is defined as occurring in a situation where only one diffraction pattern can be measured in the Fraunhofer diffraction plane and is recorded without the presence of any lenses, physical apertures, or holographic reference wave. Even today many physicists unfamiliar with the field would declare that the PDI problem must be absolutely insolvable, given that the phase of a wave is known to encode most of the structural information within it. A common gimmick is to take two pictures of different people, Fourier transform both pictures, and then exchange the modulus of those transforms while preserving their phase. If these are then both transformed back into images, the two faces appear with reasonable clarity in the same image in which the Fourier domain phase was preserved. In other words, the modulus (or intensity) information in the Fourier domain is far less important in defining the structure an image than its phase. If the phase is completely lost, the chances of regaining it in some way are negligible. However, when the object is isolated and is not too

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 5
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

large relative to the wavelength and the scattering angles being processed, this problem is now known to be almost always tractable.

The provenance of the accumulated knowledge that led to this rather astonishing conclusion is complicated and merits its own lengthy review. In the context of X-ray crystallography, Bernal and his many eminent research students laid much of the groundwork for constructing 2D and 3D maps of the unit cell from diffraction patterns taken through many different angles. Patterson (1934) noted that the Fourier transform of the intensity of the diffraction pattern is the autocorrelation function of the unit cell. Absence of phase is therefore equivalent to having an estimate of the object correlated with the complex conjugate of itself. Coupled with further *a priori* information about related crystalline structures and stereochemistry, this and similar conceptual tools led to the systematic solution of a wide range of complicated organic molecules, exemplified by the work of Hodgkin, who determined the structure of penicillin and later vitamin B_{12} (Hodgkin *et al.*, 1956) using similar methods.

In the context of crystallography, the term *direct methods* (pioneered by workers such as Hauptman, Karle, Sayre, Wilson, Woolfson, Main, Sheldrick, and Giacovazzo; see for example the textbook by Giacovazzo, 1999) refers to computational methods that do not necessarily rely on the extremely detailed knowledge of stereochemistry necessary to propose candidate solutions that fit the diffraction data (or, equivalently, the autocorrelation (Patterson) function). Instead, the likelihood of the phase of certain reflections is weighted in view of how the corresponding Bragg planes must, in practice, fill space. Here the essential *a priori* knowledge is that the object is composed of atoms that cannot be arbitrarily close to one another. Direct methods can be used to solve routinely the X-ray diffraction problem for unit cells containing several hundreds of atoms (not counting hydrogen atoms), beyond which the probabilistic phase relationships become weak. A variety of other techniques are available for larger molecules – substituting a very heavy atom at a known site within the unit cell (this can be thought of acting as a sort of holographic reference for the scattered radiation) or changing the phase of the reflection of one or more atoms by obtaining two diffraction patterns at different incident energies, above and below an absorption edge.

Independent of the crystallographic problem, an entirely different approach to the recovery of phase was derived in the context of the mathematical structure of the Fourier transform itself, especially as it relates to analytic functions. In any one experiment, an X-ray diffraction pattern explores the surface of a sphere (the Ewald sphere) in a 3D reciprocal space, which is the Fourier transform of the electron density in real space. However, periodicity in real space confines the relevant Fourier data to exist on the reciprocal lattice. Classical X-ray diffraction is about Fourier *series*, not explicitly Fourier

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 6
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

transforms. In contrast, there are a variety of physical problems in which the continuous Fourier transform is the crucial element. One such class is the so-called half-plane problem, wherein it is known, for one reason or another, that the Fourier transform of a function only has finite magnitude at positive (or negative) frequencies, while negative (or positive) frequencies all have zero magnitude. An example would be a reactive electrical system with a complicated frequency response subjected to a step change in the input voltage (or current). Here the half-plane is in the time domain, but this places constraints on the phase of the allowable frequency components. Without further detail, depending on the form of the underlying differential equation that determines a response of a system, the solution could sometimes be expressed in terms of Laplace transforms (identical to the Fourier transform except for the lack of the imaginary number in the exponential). More generally, a solution could be composed of a Fourier integral with the "frequency" component in the exponential taking on any complex value (thus embracing both the Fourier and Laplace transforms). The theory of the resulting analytic function of the complex variable that arises in this class of situation has been extensively investigated from a purely mathematical standpoint (see, for example, Whittaker and Watson, 1950). This analytic approach leads to the Kramers–Kronig relations in dielectric theory (Kramers, **?**; Kronig, **?**) and a host of other results in a wide range of fields, including half-plane methods as a means of improving resolution in the electron microscope (Misell, 1978; Hohenstein, 1992). In the latter, the half-plane is a physical aperture covering half of the back focal plane of a conventional imaging lens. It is certainly true that Bates, who went on to contribute to variants of the ptychographical principle with Rodenburg (see Section IV.D) embraced this same "analytical solution" school of thought.

<ref:Kra1927?>
<ref:Kro1926?>

In an original and oblique approach to the phase problem, Hoppe proposed the first version of the particular solution to the phase problem explored in this Chapter. Following some earlier work (Hoppe, 1969a, 1969b; Hoppe and Strube, 1969), Hegerl and Hoppe (1970) introduced the word *ptychography* to suggest a solution to the phase problem using the convolution theorem, or rather the "folding" of diffraction orders into one another via the convolution of the Fourier transform of a localized aperture or illumination function in the object plane. *Ptycho* comes from the Greek word "πτυξ" (Latin transliteration "ptux") meaning "to fold" (as in a fold in a garment). We presume the authors were following the example set by Gabor (1949), who constructed the term *holography* from the Greek "ολο" (Latin transliteration "holo"), meaning "whole": the whole wave disturbance being both its modulus and phase over a surface in space. Like holography, ptychography aspires to reconstruct the entire wave field scattered from an object. Unlike holography, it does not require a reference beam; it derives its phase knowledge by recording intensity

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 7
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

only in the diffraction plane and then, via properties of the convolution theorem, solving directly for the structure of the object. Exactly how this can be achieved and put into practice is the primary focus of this Chapter.

B. The Relative Strength of Diffractive and Imaging Methods

Although not widely disseminated, it is now recognized that diffraction phase-retrieval microscopy has certain distinct advantages over conventional forms of imaging and holography. The key requirement is the coherent interference of different wave components. In the case of conventional transmission imaging, a lens is used to form an image. Beams that pass along different paths through the lens are required to arrive at a point in the image plane in such a way that their relative phase is well determined by a particular path difference. In the case of the short-wavelength microscopies (electron and X-ray) the stability and reproducibility of this path difference is difficult to achieve. Vibration, lens instability, energy spread in the illuminating beam, and other sources of laboratory-scale interference can easily sabotage the constancy of the pertinent path difference, leading to the incoherent superposition of waves. The establishment of the appropriate path length is itself problematic if good-quality lenses are unavailable. It should noted that even in the context of so-called incoherent imaging, where the object of interest is self-luminous or illuminated by an incoherent source of radiation, there is still a requirement that the waves emanating or being scattered from any point in the object interfere with themselves coherently when they arrive in the image plane. In the case of holography, where a reference wave is added to a wave field of interest (again, one that has been scattered from the object of interest), the requirements for mechanical stability and the absence of other sources of experimental error (arising, say, from ripple on the lens supplies in the case of electron holography) are extremely demanding.

In contrast to these conventional imaging methods, so-called diffractive imaging, wherein an estimate of the object is made indirectly from solving the phase problem in the diffraction pattern scattered by it, is now generally recognized as having certain experimental advantages (even though there are also a number of potentially grave disadvantages). In relation to the question of path length discussed previously, the great strength of diffraction is that the interferometer used to effect the encoding of structural information onto the wave field is (at least in the case of atomic-scale wavelength radiations) on the same dimension as the atoms themselves. The requirement to re-interfere nonparallel wave components that have traversed substantial distances (millimeters or more) through the laboratory environment is removed. This is the great experimental strength of diffraction – it is well known that rather crude

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 8
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

instrumentation can undertake reasonably accurate diffraction measurements. As an undergraduate, I was required to measure the spacing of the atomic planes in NaCl using a benchtop X-ray diffractometer during the course of a single afternoon. In contrast to conventional imaging or holography at atomic-scale wavelengths, diffraction is experimentally relatively trivial.

In the case of high-energy electrons, the experimental robustness of diffraction relative to imaging is observed simply by comparing the breadth of reciprocal space (scattering angle) in the selected area diffraction pattern, which has significant intensity with the corresponding width of the bright-field diffractogram (that is, the Fourier transform of the bright-field image). If a microscope was perfectly stable, then the diffractogram would be as wide in k-space (scattering angle) as the diffraction pattern. In fact, the former is quickly attenuated by the inevitable lens and high-tension instabilities. When these errors are coupled with strong aberrations in the imaging lens, the effective numerical aperture of an electron lens is severely attenuated by an envelope function (Frank, 1973). Wave components scattered through large angles (in fact, no more than a few degrees) are incapable of contributing usefully to the image. Since image resolution is proportional to the effective numerical aperture of the lens, then resolution is severely compromised.

Thus, if it is possible to dispense with the lens or any other form of macroscopic interferometry and instead rely on interference arising from waves scattered directly from the atoms themselves, it should be possible to obtain much more information about the structure of a specimen, without the expense and complication of a short-wavelength imaging lens. Certain experimental challenges are removed, albeit at the price of introducing a difficult computational problem. Ptychography is essentially a method of blending these strengths of diffraction with a further experimental variation (the collection of more than one diffraction pattern), but which thereby greatly reduces the complexity and difficultly of solving the phase problem while at the same time greatly improving the applicability of the diffractive imaging method, most importantly imaging objects of infinite size even in the presence of partial coherence in the illuminating beam.

C. The Definition of Ptychography

Referring directly to Hoppe's series of three papers in 1969, where these ideas were first suggested, might lead to the conclusion that all finite object solutions to the phase problem (including the so-called oversampling method of the PDI problem, Miao *et al.*, 1999) fall under the scope of ptychography. However, such a broad definition would not be usefully discriminating. Although Hoppe clearly understood that it should be possible to solve the

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 9
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

phase problem from a single diffraction pattern (assuming it is known that the object is finite), he pointed out that this would be exceedingly difficult, at least using the computational techniques available at that time. The main thrust of his argument in the 1969 articles was that the ambiguities that occur when only a single diffraction pattern is recorded can immediately be resolved if it is possible to either move the specimen relative to a defining aperture, or relative to some sort of illumination function, and record a second diffraction pattern. In a later article (Hegerl and Hoppe, 1972), the illumination function was not moved, but it was changed in its functional form.

For the purposes of this Chapter, the term *ptychography* is reserved to apply to a method that fulfills at least all of the following three characteristics:

1. The experimental arrangement comprises a transmission object that is illuminated by a localized field of radiation or is isolated in some way by an aperture mounted upstream of the object. Scattered radiation from this arrangement provides an interference pattern (usually, but not necessarily, a Fraunhofer diffraction plane) at a plane where only intensity can be recorded.
2. At least two such interference patterns are recorded with the illumination function (or aperture function) changed or shifted with respect to the object function by a known amount.
3. A calculation is performed using at least two of these patterns in order to construct an estimate of the phase of all diffraction plane reflections, or, equivalently, of the phase and amplitude changes that have been impressed on the incident wave in real space by the presence of the object.

As it stands, this definition certainly differentiates ptychography from many competing solutions to the phase problem. We will reserve the term *pure diffractive imaging* (PDI) to apply to a solution of the Fraunhofer phase problem in which only one diffraction pattern is recorded. This is quite distinct from ptychography, although in fact the solution of the PDI problem can be thought of as relying on the convolution theorem (see Section II.F), and hence being due to the *ptycho* element of ptychography. I would argue, however, that once the phase problem is formulated in terms of the convolution theorem, then the direct solution that follows via the Fourier shift theorem is so compelling that it would seem absurd to rely on only one diffraction pattern, provided of course that there is some way of achieving the shift of the object relative to the illumination function or aperture in some experimentally feasible way. For this reason, I contend that the recording of at least two interference patterns is a defining component of ptychography, quite independent of whether the Greek meaning of the word actually embraces this specification.

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 10
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

This definition thus also renders ptychography distinct from the through-focal-series (Misell, 1973) or transport of intensity (see for example Gureyev *et al.*, 1995) solutions to the phase problem. Although in both of these cases multiple interference patterns (in this case, Fresnel diffraction patterns) are collected, the illumination (usually a plane wave) always remains fixed and unchanged with respect to the object of interest. The same applies in the case of holography.

It could be argued that any method that alters the illumination condition, or the way in which the object is presented to the field of radiation, and subsequently records more than two intensity distribution as a means of solving for phase (e.g., McBride *et al.*, 2005) falls within the present definition. In fact, as will be shown in the following text, the main practical implementations of ptychography use only a lateral shift of the object or illumination function. Imposing this restriction would, ironically, discount one of few articles published by Hegerl and Hoppe (1972) demonstrating computationally what they had themselves called ptychography, although there is really no doubt that Hoppe had originally conceived of the scheme in terms of lateral shifts alone (Hoppe, 1982).

A further issue in its exact definition is whether ptychography relates only to crystalline objects. At one level this point is irrelevant in that any image, regardless of the size of its field of view, can always be represented computationally as one very big unit cell. Indeed, all PDI methods do exactly this, as does any image processing algorithm that relies on equally spaced samples in the reciprocal space; this includes any method that uses a fast Fourier transform. However, there is a common perception, fueled if only by Hoppe himself in a late review (Hoppe, 1982), that ptychography essentially relates only to physically periodic objects. That is to say, as the object, aperture or illumination field is shifted laterally, the next part of the object to be illuminated by radiation is identical to that part of it that has shifted out of the field of illumination. It is certainly true that the simplest computational implementation of ptychography does only relate to this very specific crystalline case. However, more recent work has shown that the crystalline restriction simply needs not apply. For the purposes of this Chapter, the three characteristics enumerated above, will be expanded with the fourth:

4. When the change of illumination condition is a simple lateral shift (or shift of the object), then ptychography allows a large number of interference patterns (as many as required) to be processed in order to obtain an image of a nonperiodic structure of unlimited size.

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 11
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

D. A Qualitative Description of Ptychography

A good starting point is to describe qualitatively the principle of ptychography (as defined above) using the very simplest manifestation of the method, as first described by Hoppe (1969a, 1969b) and Hoppe and Strube (1969). Detailed mathematical treatments, and a discussion of the principal approximations that are implicitly assumed in this section, are presented in later sections.

Consider a lens focusing a source of radiation onto a crystalline object of interest as shown in Fig. 1. (In this Chapter, optical components are drawn on a horizontal optic axis with radiation incident from the left.) Assume (for the time) being that the lens is aberration free, it has a sharp-edged circular aperture in its back focal plane, and the beam crossover at the object plane is perfectly focused. In other words, the illumination function in the plane of the entrance surface of the object is of the form of an Airy disk function (see, for example, Born and Wolf, 1999). Let us also assume that this illumination function interacts multiplicatively with the object function; the object function will be modeled as a (possibly strong) phase transmission function that

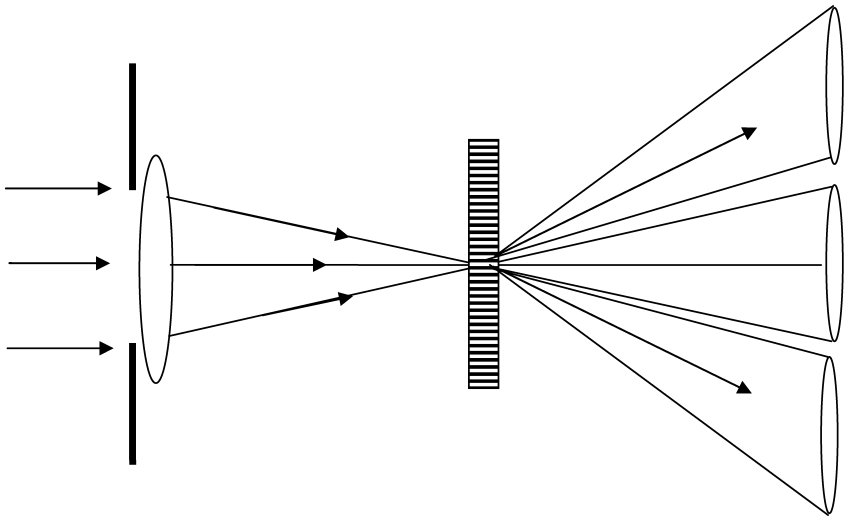


FIGURE 1. Elementary crystalline ptychographical experimental setup in the scanning transmission electron microscope (STEM) configuration. Radiation incident from the left impinges on a focusing lens that has a sharply defined diaphragm. The crystalline specimen lies in or near the plane of the beam crossover (this produces the illumination function incident on the specimen, often called a probe in the STEM literature). Downstream of the specimen a diffraction pattern evolves. Each diffraction order is spread into a disk corresponding to the range of incident illumination angles in the beam. We will assume that the distances between the lens and the object and the object and the detector plane are large: the detector plane lies in the Fraunhofer regime.

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 12
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

may also attenuate the wave as it passes through the object. Note that the transmission function is not the same as the atomic (or optical) potential function; we are not assuming the Born approximation (see Section III.C), but we are assuming that the object is quasi-2D (Cowley, 1995).

In the Fraunhofer diffraction plane, a very large distance downstream of the object, the far-field diffraction plane is observed. If this periodic object had been illuminated by a plane wave, the diffraction pattern would, as usual, consist of a series of spikes of intensity. In fact, because there is a range of incident vectors subtending from the pupil of the lens, each diffraction peak is spread out into the shape of a disk. In the absence of a specimen, there is only one disk – simply a shadow image projection of the aperture in the back focal plane of the lens. When the crystalline object is in place, it is possible to arrange for the aperture size to be such that the diffracted disks overlap one another as shown in Fig. 2(a). For what follows in this section, we require that there is a point between any two diffracted discs where only those two disks overlap with one another (i.e., there are no multiple overlaps). If the source of radiation is sufficiently small and monochromatic, then illumination over

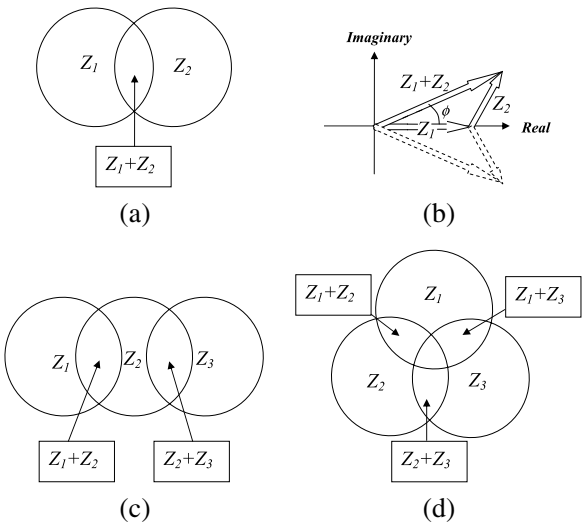


FIGURE 2. Schematic representation of the diffraction orders and phase relationships in the STEM ptychograph. (a) Two diffracted disks lying in the Fraunhofer diffraction plane (the right-hand side of Fig. 1). (b) Phase relationship of the underlying amplitudes of these two disks. The square roots of the measured intensities give the lengths of the arrows, but not their phase. However, the triangle of complex numbers must be closed, although there are two indistinguishable solutions. (c) For three linearly positioned interfering disks. (d) For 2D functions, ambiguity in this phase problem (like in all Fourier phase problems) is reduced because the ratio of measurements to unknowns increases.

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 13
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

the back focal plane of the lens will be effectively coherent. In the Fraunhofer plane, this means that the diffracted beams will interfere coherently within these regions of overlap (provided other sources of experimental instability, such as the lens, are negligible).

In the classic diffraction phase problem (where the illumination is a plane wave), the intensity (and only the intensity) of each of the diffracted beams can be measured: the phase is lost. In Fig. 2(a), we also measure intensity only, but now the intensity of the sum of two diffracted amplitudes also can be measured. Let us label these amplitudes by the complex numbers Z_1 and Z_2, each representing the real and imaginary components (or modulus and phase) of the wave filling each disk. We can therefore measure $|Z_1|^2$, $|Z_2|^2$, and, within their region of intersection, $|Z_1 + Z_2|^2$. Let the phase difference between Z_1 and Z_2 be an angle ϕ. In the complex plane, we see a relationship between amplitudes Z_1 and Z_2 and their sum $Z_1 + Z_2$ (shown in Fig. 2(b)). Because the moduli Z_1, Z_2 and $Z_1 + Z_2$ can be measured by taking the square root of the intensity measured in the two disks and their region of overlap, respectively, then there are only two values of ϕ (ϕ and $-\phi$) that will allow the triangle of amplitudes Z_1, Z_2 and $Z_1 + Z_2$ to connect with one another. In other words, by measuring these three intensities, and assuming they interfere with one another coherently, we can derive an estimate of the phase difference between Z_1 and Z_2, but not the sense (positive or negative) of this phase difference.

Clearly this simple experiment has greatly improved the chances of solving the phase problem; instead of total uncertainty in the phase of either Z_1 or Z_2, there are now only two candidate solutions for the phase difference between Z_1 and Z_2. In fact, without loss of generality, we will always assign one of these values (say, Z_1) as having an absolute phase of zero. This is because in any diffraction or holographic experiment only the relative phase between wave components can ever be possibly measured. Another way of saying this is that we are concerned with time-independent solutions to the wave equation, and so the absolute phase of the underlying waves (which are time dependent) is lost and is of no significance – in the case of elastic scattering, the relative phase of the scattered wave distribution has had all the available structural information about the object impressed upon it. If the disk of amplitude Z_1 is the straight-through (or zero order) beam, then in what follows we will always assign this as having zero phase: this particular two-beam phase problem (i.e., the relative phase of Z_2) has been completely solved apart from the ambiguity in Z_2 or Z_2^*, where Z_2^* is the complex conjugate of Z_2.

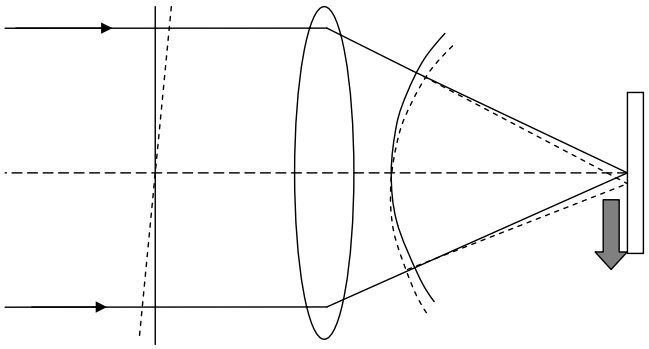
Now consider the interference of three beams lying in a row (Fig. 2(c)). The underlying amplitudes of these diffracted beams are labeled as Z_1, Z_2, and Z_3. We now measure five intensities: three in the disks and two in the

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 14
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

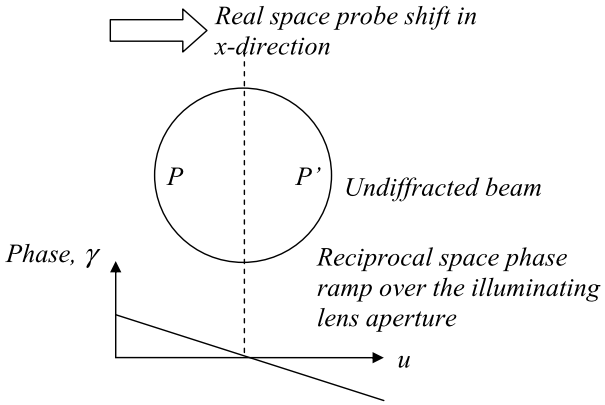
overlaps. These five numbers will not quite give us the six variables we need – the modulus and phase of Z_1, Z_2, and Z_3 – but they will give us $|Z_1|$, $|Z_2|$, $|Z_3|$ and two phase differences (each of which suffer from a positive/negative ambiguity). Assigning zero phase to Z_1, we can solve for the phase of all the diffracted beams (Z_2 and Z_3) to within four possible solutions, depending on the four combinations of the ambiguous phases. It is clear that as the number, N, of diffracted beams increases, then there will be 2^N possible solutions for the one diffraction pattern. In fact, this 2^N ambiguity is a completely general feature of the phase problem in one dimension (Burge *et al.*, 1976) when the object space is known to be finite (in the current case, the object and illumination function is of a particular form that gives a very clear manifestation of the phase ambiguity). Of course, if the three beams are not collinear, but form a triangle in a 2D diffraction pattern so that each beam overlaps with the other two (Fig. 2(d)), then the ratio of intensity measurements to unknowns can be increased, thereby limiting the number of ambiguous solutions to the entire pattern. Again, even though we are discussing a very specialized scattering geometry, the same argument applies in all diffractive imaging methods – as the dimension of the problem is increased, the likelihood of ambiguous solutions existing reduces (see for example the discussion by Hawkes and Kasper, 1996).

The key benefit of ptychography is that a complete solution of the phase, resolving the ambiguities discussed above, can be achieved by shifting either the object function or the illumination relative to one another by a small amount, and then recording a second set of intensity measurements. We can think of this via the Fourier shift theorem, but to begin we restrict ourselves to a qualitative description. In the scattering geometry of Fig. 1, a shift in the illumination function can be achieved by introducing a phase ramp across the illumination pupil. One way of thinking of this is via Fig. 3(a) (which is rather exaggerated). To produce a tilt in the wave coming out of the lens (which will consequently lead to a shift in the illumination function at the object plane), we must rotate slightly the hemisphere of constant phase that emerges from the lens and which results in the wavefronts being focused at the crossover at the object plane. If the path length is increased at the top of the hemisphere and decreased at bottom (as shown in Fig. 3(a)), then the effect is to move the position of the illumination downward relative to the object. The path difference changes linearly across the hemisphere. A linear phase ramp, where the phase is equal to 2π times the path difference divided by the wavelength, has the equivalent influence. (In what follows, we ignore the fact that this tilt in the beam will not only shift the illumination, but also shift the diffraction pattern in the far field; in the case of most situations of interest this latter shift is negligible for typical lens working distances.) In the Fraunhofer plane, one side of the zero order (unscattered beam) has therefore been advanced in

phase (marked as point P on Fig. 3(b)), say by a factor of $e^{i\gamma}$, and a point near the opposite side (P') has been retarded in phase, say by a factor of $e^{-i\gamma}$. If we label the underlying amplitude of the zero-order beam as Z_1, then the complex value of the wave at P is now $Z_1 e^{i\gamma}$. Suppose now that there is a diffracted beam of underlying amplitude Z_2 (shown separately on the top right



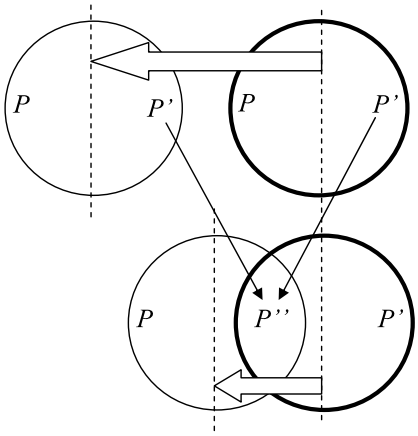
(a) In the STEM configuration, a shift in the probe (indicated by the large pointer) can be achieved by introducing a phase ramp in the plane waves (dotted lines) over the back focal plane of the lens (incident beams illuminating the lens are drawn parallel). This is equivalent to tilting slightly the hemisphere of wavefronts focusing the probe onto the object, consequently introducing the shift. This manifests itself as the phase change over $A(u, v)$.



(b) As the probe is shifted in real (object) space (top), a phase ramp is introduced over the undiffracted beam in reciprocal space (lying in the Fraunhofer diffraction plane).

FIGURE 3.

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 16
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV



(c) At the top we draw the undiffracted beam (heavy circle) and a diffracted beam. The probe movement has introduced phase ramps across both beams. When the spacing of the diffracted beams is smaller (bottom circles), the equivalent points P and P' meet up and interfere at P''.

FIGURE 3. (*continued*)

of Fig. 3(b)). At an equivalent point P' the actual complex value of the wave disturbance is now $Z_2 e^{-i\gamma}$. If these two disks overlap, then the equivalent points P and P' can meet up and interfere with one another, and the total amplitude where these points meet (P'' in Fig. 3(c)) will be Z_s, where

$$Z_s = Z_1 e^{i\gamma} + Z_2 e^{-i\gamma}. \tag{1}$$

Before we moved the illumination by introducing these phase changes, we measured Z_1, Z_2 and $Z_1 + Z_2$. With reference to the complex sum shown in Fig. 4, let Z_0 be equal to $Z_1 + Z_2$. In view of the complex conjugate ambiguity, $Z_1 + Z_2^* = Z_c$ was also a candidate solution. However, after shifting the illumination and recording Z_s, we can now discount this (wrong) solution because if the same phase changes had been applied to $Z_1 + Z_2^*$, the final sum would have been given by Z_w (see Fig. 4).

This rather naïve analysis actually summarizes the entire principle of ptychography. Later sections of the text will show that the same underlying principle can be applied to extended noncrystalline objects and the illumination function must not be of any particular form, as long as it is reasonably localized. The geometric and scattering theory approximations made in this section will also be considered.

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 17
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

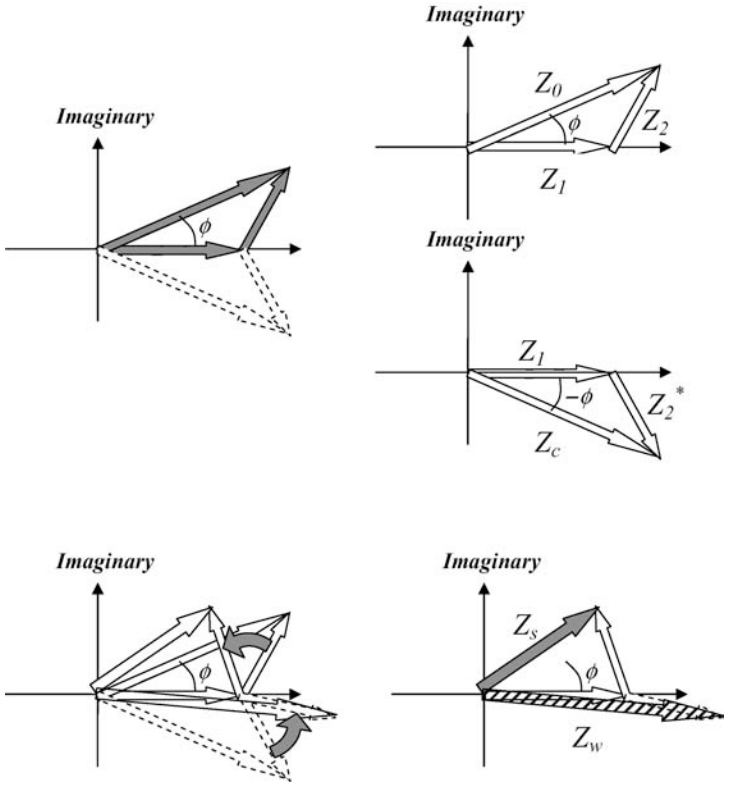


FIGURE 4. How ptychography unlocks the complex conjugate phase ambiguity. In this diagram, the lengths of filled pointers represent the square root of the intensity of our four measurements (two from the centers of the diffracted disks, two from the overlap region at different probe positions). Before moving the illumination (probe) the three initial measurements (top left) have two possible solutions (top and middle right). When the phase shifts (Fig. 3) are applied (bottom left), the wrong ($-\phi$) phase can be discounted: the hatched pointer (Z_w) is not the length of the measured modulus Z_s. All phases are plotted with respect to Z_1; in practice, Z_1 and Z_2 rotate equal angles in opposite directions, but the effect on the measured moduli is the same.

E. Mathematical Formulation of Crystalline STEM Ptychography

For the sake of simplicity, this section adopts an image processing-type of nomenclature and not an optical or electron scattering nomenclature. Section III will connect the relevant wave k-vectors with the implicit geometric approximations made here. This section simply defines a mathematical form

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 18
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

of the 2D Fourier transform as

$$\Im f(x, y) = \iint f(x, y) e^{i2\pi(ux+vy)}\, dx\, dy = F(u, v), \tag{2a}$$

and its inverse as

$$\Im^{-1} F(u, v) = \iint F(u, v) e^{-i2\pi(ux+vy)}\, dx\, dy = f(x, y), \tag{2b}$$

where $f(x, y)$ is a 2D image or transmission function that may be complex valued, described by Cartesian coordinates x and y; $F(u, v)$ is the Fourier transform of $f(x, y)$; u and v are coordinates in reciprocal space (that is, the Fourier transform space); and for the compactness of later equations we use the symbol $\Im$ to represent the whole Fourier integral. This definition of the Fourier transform differs from that used in some fields (including electron microscopy, although not crystallography) by having a positive exponential. The inclusion of the factor of 2π in the exponential renders the definition of the inverse much simpler by not requiring a scaling factor dependent on the dimensionality of the functions involved. This greatly simplifies some arithmetic manipulations that must be used in later sections.

As before, let us assume that the object can be represented by a thin complex transmission function. In other words, the presence of the object has the effect of changing the phase and intensity (or, rather, modulus) of the wave incident upon it. The exit wave, $\psi_e(x, y)$, immediately behind the object, will have the form

$$\psi_e(x, y) = a(x, y) \cdot q(x, y), \tag{3}$$

where x and y are coordinates lying in the plane of the (2D) object function, $a(x, y)$ is the complex illumination function falling on the object, and $q(x, y)$ is the complex transmission function of the object. As emphasized previously, $q(x, y)$ is *not* the potential or optical potential of the object; it is equal to the exit wave that would emerge from a thin (but possibly very strongly scattering) object when illuminated with a monochromatic plane wave. For a thick specimen, we might hope that

$$q(x, y) = e^{i \int \sigma V(x,y,z)\, dz}, \tag{4}$$

where $V(x, y, z)$ is the optical potential (or, in the case of high electrons, the atomic potential) of the object, and σ is a constant describing the strength of the scattering for the particular radiation involved. In this case the ptychographical image derived below can be related directly to the projection of the potential (by taking the logarithm of $q(x, y)$ – at least in the absence of phase wrapping), despite strong (dynamical) scattering. In fact, this is not the case for thick specimens because the integral with respect to z does not

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 19
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

account for propagation effects as the wave traverses and diffracts through the thickness of the object. This distinction is emphasized because it has led to many misunderstandings in the literature. In particular, because $q(x, y)$ can introduce very large phase changes, including more than 2π, it is quite wrong to suggest that the multiplicative approximation is a "weak phase" or kinematical approximation; it is really a "thin object" approximation.

Assuming this approximation, we can write via the convolution theorem that the amplitude in the far field, $M(u, v,)$, is given by

$$M(u, v) = \{\Im q(x, y)\} \oplus \{\Im a(x, y)\}, \tag{5}$$

where we note that in a real experiment u and v are must be scaled appropriately by the wavelength and the camera length (the distance between the object and the detector plane; see Section III.A). The convolution operator, $\oplus$, is defined for any two 2D functions, $g(x, y)$ and $h(x, y)$, as

$$g(x, y) \oplus h(x, y) = \iint g(X, Y)h(x - X, y - Y)\, dX\, dY. \tag{6}$$

Now let us relate this to our qualitative description of ptychography in the previous section. In the case of electrons, the optical setup shown in Fig. 1 is achieved in the STEM configuration. In this case, $a(x, y)$ is referred to as the "probe" because the very intense beam crossover at the specimen plane can be used to probe the specimen to obtain electron energy loss spectra and other analytical signals from the small volume of matter illuminated. $a(x, y)$ is itself the back Fourier transform of the aperture function, which we call $A(u, v)$, lying in the back focal plane of the lens. It should be noted that u and v are angular coordinates of the Fourier components that make up the incident radiation. Because beams cross over in the vicinity of the object, a diaphragm lying in the back focal plane appears reversed in the diffraction plane. $A(u, v)$ is therefore the complex amplitude of the illumination as it arrives in the diffraction plane. $A(u, v)$ can include phase components corresponding to the presence of aberrations; its modulus is usually of the form of a "top hat" function (although this is considerably modified when the illumination is partially coherent). The diffracted amplitude is now:

$$M(u, v) = \Im\left[\Im^{-1} A(u, v)\right] \oplus \Im q(x, y), \tag{7a}$$

or

$$M(u, v) = A(u, v) \oplus Q(u, v), \tag{7b}$$

where $Q(u, v)$ is the Fourier transform of the object transmission function. It is this convolution (folding) in the diffraction plane that leads to the term *ptychography*. Everything in this Chapter depends on the existence of

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 20
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

this convolution. The STEM configuration is particularly straightforward to understand when we assume a very simple form of $A(u, v)$ – a sharply defined circular aperture. In our simplest manifestation of ptychography, we make two measurements, intensities $I_1(u, v)$ and $I_2(u, v)$, each from different positions of the probe, where

$$I_1(u, v) = \left|A(u, v) \oplus Q(u, v)\right|^2 \tag{8a}$$

and

$$I_2(u, v) = \left|\left[A(u, v)e^{iu\Delta x}\right] \oplus Q(u, v)\right|^2 \tag{8b}$$

and where we have assumed the second measurement is made with the illumination function shifted with respect to the origin in real space by a distance Δx in the x-direction to give $a(x - \Delta x, y)$: The exponential derives from the Fourier shift theorem, namely, that

$$\begin{aligned}\Im g(x + \Delta x) &= \int g(x + \Delta x)e^{i2\pi ux}\,dx = \int g(X)e^{i2\pi u(X-\Delta x)}\,dX \\ &= G(u)e^{-iu2\pi\Delta x}\end{aligned} \tag{9}$$

for any function $g(x)$ whose Fourier transform is $G(u)$. In the simplest version of ptychography, we are interested in the case of $q(x, y)$ being a periodic 2D crystal. $Q(u, v)$ then only has values on an equivalent 2D reciprocal lattice. Let us suppose for simplicity that this lattice is rectilinear with a periodicity of Δu and Δv, in the u and v directions, respectively. We can say that

$$Q(u, v) = \sum_{n,m} Q_{n,m}\delta(u - n\Delta u, v - m\Delta v), \tag{10}$$

where $Q_{m,n}$ is a complex number associated with the amplitude of a reciprocal lattice point indexed by m and n, the sum is over the positive and negative integers for all m and n, and $\delta(u, v)$ is a Dirac delta function defined as

$$\delta(u - u_0, v - v_0) = \begin{cases} 0 & \text{for } u \neq u_0 \text{ or } v \neq v_0, \\ \infty & \text{for } u = u_0 \text{ and } v = v_0 \end{cases} \tag{11a}$$

and where

$$\int_{-\infty}^{\infty} \delta(u - u_0, v - v_0)\,du\,dv = 1. \tag{11b}$$

To achieve the overlapping disks we described in the previous section (Fig. 2(a)), the aperture function $A(u, v)$ must have a diameter of at least Δv or Δu, whichever is the larger. Suppose that $\Delta u > \alpha > \Delta u/2$, where α is radius of the aperture, as shown in Fig. 5. As before, we consider just two of

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 21
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

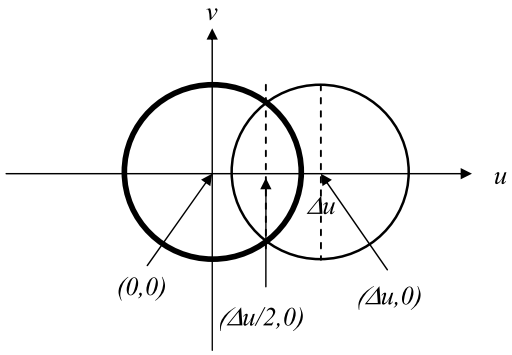


FIGURE 5. In the simplest STEM implementation of crystalline ptychography, intensity is measured at the centers of the zero-order and first-order diffraction disks and at the midpoint of their overlap region.

the disks in the far field: the zero-order beam ($n = m = 0$) and one diffracted beam ($n = 1, m = 0$), then

$$Q(u, v) = Q_{0,0}\delta(u, v) + Q_{1,0}\delta(u - \Delta u) \tag{12}$$

and so

$$M(u, v) = Q_{0,0}A(u, v) + Q_{1,0}A(u - \Delta u, v). \tag{13}$$

With reference to Fig. 5, let us measure the intensity at a point midway between the centers of the two interfering aperture functions, namely, at $u = \Delta u/2$ and $v = 0$. The wave amplitude at this point is

$$M\left(\frac{\Delta u}{2}, 0\right) = Q_{0,0}A\left(\frac{\Delta u}{2}, 0\right) + Q_{1,0}A\left(-\frac{\Delta u}{2}, 0\right) \tag{14}$$

and hence the intensity at this point in the diffraction plane with the illumination field (the probe) on axis, is

$$I_1\left(\frac{\Delta u}{2}, 0\right) = \left|Q_{0,0}A\left(\frac{\Delta u}{2}, 0\right) + Q_{1,0}A\left(-\frac{\Delta u}{2}, 0\right)\right|^2. \tag{15}$$

We now move the probe a distance Δx in the x-direction (parallel with the reciprocal coordinate u). From Eq. (8b) we obtain a second intensity measurement

$$I_2\left(\frac{\Delta u}{2}, 0\right) = \left|Q_{0,0}A\left(\frac{\alpha}{2}, 0\right)e^{i\frac{\Delta 2\pi u}{2}\Delta x} + Q_{1,0}A\left(-\frac{\alpha}{2}, 0\right)e^{-i\frac{\Delta 2\pi u}{2}\Delta x}\right|^2. \tag{16}$$

aiep v.2005/11/15 Prn:6/12/2007; 13:13
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

The amplitudes of these two equations are identical in form to the qualitative description derived in Section II.E [Eq. (1)], with the phase term γ replaced by $\pi \Delta u \Delta x$ and with Z_1 and Z_2 replaced by the underlying phases of the lattice reflections $Q_{0,0}$ and $Q_{1,0}$. In the previous section we assumed that the aperture function $A(u, v)$ did not have any aberrations (i.e., phase components) across it; this means the value of $A(u, v)$ is unity, and so this term does not appear in Eq. (1). In general, the aperture function *will* have phase changes across it, corresponding to defocus, spherical aberration, and so on, and so these must be accounted for in some way – indeed the probe movement (which leads to a linear phase ramp multiplying the aperture function, as in Eq. (8b) is itself a sort of aberration term.

It should be pointed out that in general it may not be experimentally convenient to use an aperture function that is small enough to guarantee that there exist points in the diffraction pattern where only two adjacent diffracted disks overlap. Clearly, as the unit cell increases in size, the diffraction disks become more and more closely spaced in the diffraction plane. In the limit of a nonperiodic object, it may appear as if the ptychographical principle is no longer applicable, because the pairs of interfering beams can no longer be cleanly separated. In fact, we will see in Section IV that this is not the case, provided we adopt more elaborate processing strategies.

In this very simplest arrangement with just two diffracted beams, we still need to measure a total of four intensities to derive their relative phase. As well as measurements made at the disk overlaps, we need the intensities of the diffracted beams themselves (lying outside the region of overlap, but within a single disk, as in Fig. 5). Our task is to solve for the phase difference between $Q_{0,0}$ and $Q_{1,0}$ given that

$$\begin{aligned} |Q_{0,0}|^2 &= I_1(0,0) = I_2(0,0), \\ |Q_{1,0}|^2 &= I_1(\Delta u, 0) = I_2(0,0), \\ |Q_{0,0} + Q_{1,0}|^2 &= I_1\left(\frac{\Delta u}{2}, 0\right), \\ \left|Q_{0,0}e^{i\gamma} + Q_{1,0}e^{-i\gamma}\right|^2 &= I_2\left(\frac{\Delta u}{2}, 0\right). \end{aligned} \tag{17}$$

Conventional crystallography is usually interested only in the *structure* of the object, not its absolute position in terms of x and y. A displacement ambiguity always occurs in the classic PDI phase problem: the intensity of the Fraunhofer plane is unaffected by lateral displacements of the object. In other words, normal diffraction is never sensitive to a linear-phase ramp applied across the diffraction plane. The usual crystallographic convention assumes that a point of symmetry is implicitly at the origin of real space. In contrast, in ptychography there is an explicit definition of absolute position in real space –

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 23
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

the known position of the confined illumination – but there is no reason to suppose that the crystal itself just happens to be aligned centrosymmetrically with this point in space. If the crystallographic phase of $Q(u, v)$ (arising from adopting a centrosymmetric convention for the phase of the diffracted beams) is $Q_0(u, v)$, then

$$Q(u, v) = Q_0(u, v)^{-i(u\Delta x + v\Delta y)}, \tag{18}$$

where Δx and Δy is the displacement of the crystal relative to the optic axis. As a result, the particular phases of $Q_{0,0}$ and $Q_{1,0}$ that are consistent with Eq. (17) relate to exact geometry of the lens, which defines the exact position of the illumination function $a(x, y)$. In practice, the value of γ, proportional to the probe movement distance, can be chosen at will. In almost all conceivable experimental implementations of crystalline STEM (electron) ptychography, the exact position of the probe relative to the object is not known accurately; only relative displacements can be measured (say by applying an offset to the probe scan coils in STEM). In this case, we could choose to scan the probe position until $I_1(\frac{\Delta v}{2}, 0)$ is at a maximum and simply assert the convention that the probe at this position is at $x = y = 0$. Relative to this definition, then $Q_{0,0}$ and $Q_{1,0}$ have zero phase difference between them (because in this condition they add to give maximum modulus). However, the phase of all other reflections will be altered accordingly via Eq. (18).

Equation (17) can be solved either geometrically (via the complex plane construction in Fig. 2(b)) or by direct substitution. Ill-defined solutions can still arise if we choose the value of γ unfavorably; for example, we find that $I_1 = I_2$ if either $\Delta x = n\pi/\Delta u$, where n is integer, or when $\gamma/2$ happens to be the same as the phase difference between $Q_{0,0}$ and $Q_{1,0}$. We note, however, that Eq. (17) only represents a tiny fraction of the number of measurements at our disposal. Obviously, the intensity at the midpoints of all pairs of reflections must be tracked to solve for all of the relative phases of all pairs of adjacent beams. Many diffraction patterns, from many probe positions, can also be recorded instead of just the two required to unlock the complex conjugate ambiguity. Achieving this in the most computationally favorable way has been the subject of much work and is discussed in Section IV.

F. Illumination of a "Top Hat" or Objects with Finite Support

The crystalline STEM implementation of ptychography is easy to understand intuitively because the illumination function (probe) in real space is of a particular form that conveniently provides a far-field diffraction pattern in which the interferences of the various pairs of beams are tidily separated. In general, the illumination function can be of any functional form. In the

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 24
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

previous section, the exact shape of $a(x, y)$ did not need to be known because it was itself a Fourier transform of the a neat shape that subsequently was convolved over the diffraction pattern.

In his original papers outlining the key ideas in ptychography, Hoppe (1969a, 1969b) first described the STEM implementation and then went on to consider the effect of sharply defined rectilinear illumination patches, achieved by apertures of various dimensions mounted right up against the specimen. He observed that the same broad concepts for solving the phase problem applied in the Fourier domain. We note that this physical implementation is identical to the concept of a finite support that arises in much of the theory of phase retrieval, although Hoppe's thoughts on this matter seem to exist in isolation, at least as far as the cited literature is concerned.

It is easiest to understand why a finite support leads to a ptychographical solution of the phase problem by thinking of a 1D object function, $q(x)$, illuminated by a top-hat function $a(x)$, of the forms shown in Fig. 6. Let us first discuss this qualitatively. If $q(x)$ is periodic, then its Fourier transform, $Q(u)$ consists of a series of spikes (crystalline reflections), the modulus and phase of which determine the Fourier series components of the structure lying within the unit cell. When the crystal is perfect and of infinite extent, then there is no scattered amplitude lying between these discrete reflections. When we isolate only a small region of the specimen by the function $a(x)$, then the diffracted amplitude $M(u)$ is, via Eq. (7), the convolution of the Fourier transform of $a(x)$ with $Q(u)$. Provided $a(x)$ has small physical extent (so that $A(u)$ has a width that is a significant proportion of the reciprocal lattice spacing), then the effect of the convolution will be to introduce substantial amplitude lying at points between the conventional lattice reflections; it is here that adjacent diffraction orders can interfere with one another, just as in the STEM configuration. Once again, by moving the aperture function laterally in the object plane, a phase ramp can be introduced across $A(u)$, altering the relative phase of the interfering crystalline reflections, and hence obtaining a direct estimate of their phase. Unfortunately, $A(u)$ is now of the form of a sinc function (see Eq. (20) below) and has infinite extent, so that the tails of many diffraction orders interfere with one another at points midway between the diffraction orders (illustrated in Fig. 6). This is in stark contrast to the STEM configuration where we can arrange for two and only two pairs of beams to interfere at the reciprocal lattice midpoints.

This argument can be stated mathematically as

$$a(x) = 1 \quad \text{for } x < \frac{b}{2} < -x \tag{19}$$

and $a(x) = 0$ elsewhere. The Fourier transform of $a(x)$ can be derived from direct evaluation of the 1D from of Eq. (2a), yielding the definition of the sinc

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 25
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

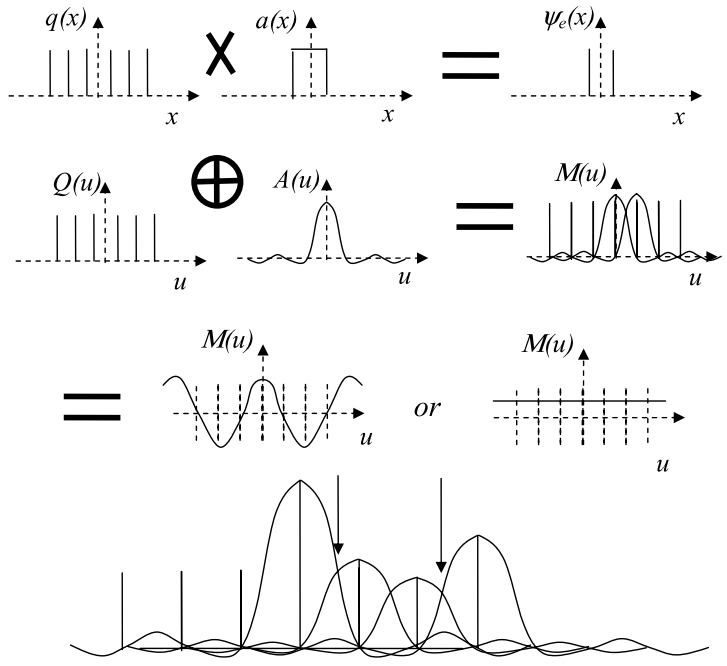


FIGURE 6. Illustrating ptychographical interference when the illumination function is of the form of a top hat (or there exists, equivalently, a support function). We assume the multiplicative approximation in real space (top), giving rise to an exit wave consisting of just two spikes of amplitude. In reciprocal space, we form the convolution, adding together a row of sinc functions. Rather than measuring points halfway between the reciprocal lattice points, the points where we infer that the amplitude of adjacent reflected moduli are equal (bottom diagram) give the most accurate ptychographic phase. In fact, this particular rather unrealistic example (a series of delta functions in the real and reciprocal space of $q(x)$) raises an interesting paradox. We know that, as a function of the position of the illumination, the resulting Young's slit intensity has either zero or unity intensity at the exact halfway point between the diffraction reflections (depending on the width of the top hat). This is because of the way the small ringing effects in phase-altered sinc functions add. When the unit cell has structure, and the top hat is the same size as the unit cell, we can infer the ptychographical phases, but the ringing effects will still corrupt the estimate (unlike in the STEM configuration where the interference terms are neatly separated).

function, such that

$$\Im a(x) = A(u) = b\frac{\sin \pi ub}{\pi ub} = b\,\mathrm{sinc}(\pi ub). \tag{20}$$

If the object of function is crystalline, it consists of a series of delta functions convolved with the structure of the unit cell that can be written as

$$q(x) = \left(\sum_n \delta(x - n\Delta x)\right) \oplus f_c(x). \tag{21}$$

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 26
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

The amplitude in the far field is

$$M(u) = \Im\big[a(x) \cdot q(x)\big] = A(x) \oplus Q(u). \tag{22}$$

Now the Fourier transform of a periodic array of delta functions of separation Δx is itself an array of delta functions of separation $\Delta u = 1/\Delta x$. Hence

$$Q(u) = \left(\sum_n \delta(u - n\Delta u)\right) \cdot \big(\Im f_c(x)\big). \tag{23}$$

In analogy with Eq. (10), this can be written as

$$Q(u, v) = \sum_n Q_n \delta(u - n\Delta u), \tag{24}$$

with the understanding that the Q_n are the Fourier coefficients (or structure factors) of the unit cell $f_c(x)$. Therefore

$$M(u) = \sum_n Q_n \delta(u - n\Delta u) \oplus b\,\mathrm{sinc}(\pi u b). \tag{25}$$

This convolution can be pictured as simply the addition of a series of sinc functions, each centered on a diffraction pattern reflection and multiplied by the modulus and phase of that reflection (Fig. 6); $\mathrm{sinc}(\pi u b)$ passes through zero whenever ub is a non-zero integer. So when $b = 1/\Delta u = \Delta x$, a sinc function centered on a particular reflection passes through zero at all other reflections. In other words, when the object plane aperture is the same width as the unit cell, the intensity measured at points in the diffraction pattern where the normal diffraction peaks would occur is unaffected by the presence of the aperture (except for a constant scaling factor). But in addition to these conventional diffraction data, the interference intensities between the conventional diffraction orders can be measured. Hoppe (1969a) suggested examining points illustrated at the bottom of Fig. 6 where, given the known moduli of the adjacent diffraction orders, one could infer the point at which the associated sinc functions would have equal moduli. The actual modulus (square root of intensity) measured at this exact point would be indicative of the phase difference between the two adjacent reflections, give or take the usual complex conjugate ambiguity (see Fig. 4), which, as before, can be resolved by shifting $a(x)$ by less than the width of the unit cell. Unlike with the STEM, we now have the problem that all other diffracted reflections will also add amplitude to points lying between any pair of reflections. In fact, these additions in the artificial example shown in Fig. 6 lead to a strange paradox (see caption). Nevertheless, Hoppe and Strube (1969) demonstrated in an optical experiment that in general the phase differences of adjacent diffraction orders could indeed be measured with reasonable accuracy using this method.

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 27
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

This process of interfering adjacent diffracted reflections is another way of thinking about where both the source of the phase information and the ambiguous solutions arise in conventional PDI. If we assume the object is in fact periodic (and in this discussion, one-dimensional), with a periodicity of its entire width, then the aperture function (support) is also the size of the unit cell. In reciprocal space, the crystalline lattice points must be convoluted with sinc functions, each one of which has zero value at all reflections except the one on which it is centered. This hypothetical aperture does not alter the relative intensity of the diffracted orders, but it does mean the diffracted orders interfere with one another at the "in-between" reciprocal lattice points. Hence, we have the classic phase problem with the 2^N complex conjugate ambiguity.

G. Two-Beam Ptychography

In referring back to the STEM implementation of ptychography, we note that a large part of the illumination disk is irrelevant as far as the phase determination process is concerned. In fact, all that is need are two pinholes, separated in the back focal plane of the lens by the exact distance between the lattice points. This is equivalent to illuminating the object by two coherent plane waves, inclined to one another by twice the Bragg angle (this method relates very closely to the Bragg–Brentano data set, discussed in Section IV.B). At the object, the waves interfere and give rise to a sinusoidal interference pattern of the same periodicity as that of the unit cell. In the far field we now have the coherent summation of two diffraction patterns, offset from one another by one reciprocal lattice vector. The amplitude of each of the reflections in the resulting pattern is what would have been obtained within the overlap region of each pair of disks in the STEM configuration. Once again, the complex-conjugate ambiguity can be resolved by introducing a shift in the illumination pattern of radiation which, as before, can be effected by introducing a change in phase in either or both the incident beams.

The article by Hegerl and Hoppe (1972) uses this condition to evaluate the ptychographical phase differences between reflections in a model calculation. In this manifestation of ptychography, the illumination function must be by a beam splitter or diffraction grating (as in the work of Berndt and Doll, 1976, 1978, 1983) and is therefore much harder to implement experimentally than the STEM configuration. However, it does have the distinct advantage that a crystalline object function can be very large, consisting of many thousands of unit cells, all of which are bathed in the illuminating beam. This is because there is no confined probe: the two incident plane waves extend over the entire object. In the far field, all the relevant intensity is channeled into exactly (and only) those points that give rise to the ptychographical

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 28
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

interference. In consequence, counting statistics are much more favorable and difficulties related to specimen damage due to a highly focused beam are reduced. On the hand, the angular separation of the two incident beams must be very accurately controlled so that the resulting illumination function has a periodicity that exactly matches that of the unit cell over moderately large distances (all the many unit cells which are illuminated). Two-beam ptychography may still find application in the future in protein crystallography where unit cells are large and low doses are required, as discussed by Hoppe (1982) in his retirement review. Section IV.B. discusses the Bragg–Brentano subset, which is a variant of the two-beam method, but where a range of angular separations in the incident beam are processed simultaneously.

H. Commentary on Nyquist Sampling

Some authors, primarily those with a background in X-ray crystallography, have introduced the term *oversampling* to describe the PDI method (Miao *et al.*, 1999, and the papers that cite this work). Note that here we are discussing a sampling periodicity in the diffraction plane, which is itself in the spatial frequency domain. The topic of discussion is the necessary pixel size in the Frauhofer diffraction plane, where that size is expressed as an angle subtended from the object (see Section III.B). In terms of the previous nomenclature, we are interested in the necessary sampling in the u coordinate to capture all the available information expressed in the experiment. We have seen that ptychography relies on sampling between the conventional diffraction orders, and so a discussion of how this "in-between the lattice point" sampling relates to that required for the PDI problem is of some interest in the present context.

Referring again to Fig. 6, consider an array of delta functions in real space, representing a set of periodic slits with separation Δx. Note that we have chosen to have the array shifted by $\Delta x/2$ relative to the origin; this is done so that every part of the amplitude of the diffraction pattern will be real (no imaginary components) in order to facilitate the graphical representation of what follows. Now let us introduce our illumination function $a(x)$, also centered on the origin of real space. If this aperture is narrow enough only to cover two slits in the periodic structure, then we have effectively set up a Young slits experiment. In the far field the *amplitude* of the scattered waves is of the form

$$M(u) = e^{-i2\pi\frac{\Delta x}{2}u} + e^{i2\pi\frac{\Delta x}{2}u} = 2\cos(\pi\,\Delta x u), \tag{26}$$

as shown in Fig. 7. Our task from the point of view of a crystallographer attempting to determine the structure of the underlying crystal (that is, before the introduction of the aperture) is to find the modulus and phase of the Bragg peaks, labeled with crosses in Fig. 7. These occur at a spacing

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 29
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

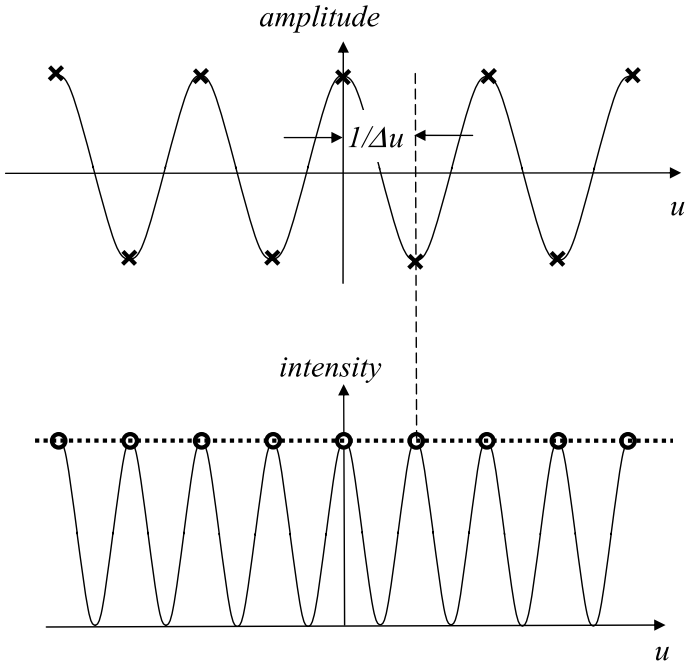


FIGURE 7. Sampling in reciprocal space. Crystallography samples at the diffracted beams lying at a spacing of $1/\Delta u$ (crosses in top diagram). In this case, the crystal is a centrosymmetric array of delta functions (as in Fig. 6). When isolated by an aperture, illumination function or support aperture (see also Fig. 6), the diffracted amplitude is a Young slits pattern. When recorded in intensity (lower diagram), sampling at $1/\Delta u$ (circles) gives radical aliasing, from which we would erroneously infer the function represented by the horizontal dotted line.

in reciprocal space of $\Delta u = 1/\Delta x$. In the case of the Young slits, that means we need to establish that the undiffracted beam has an amplitude of 2 (because $2\cos(0) = 2$), and that at the first diffraction order ($u = \Delta u$, so that $M(u) = 2\cos\pi = -2$) there is an amplitude of -2. In crystallographic terms we would say that the moduli of these orders are both 2, but their phases are 0 and π, respectively. Once we have discovered the phases, we have solved the structure of the unit cell. Of course, if the unit cell has elaborate structure, the phases of the various reflections could be very complicated. In the present context (infinitely narrow slits), then successive reflections simply have phases of either 0 or π.

However, we cannot measure *amplitude*, we can only measure *intensity*. We see that

$$\left|M(u)\right|^2 = 4\cos^2\left(\frac{\Delta x u}{2}\right) = 4\big(1 - \cos(\Delta x u)\big). \tag{27}$$

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 30
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

In other words, the periodicity of the diffracted intensity is now exactly the same as the periodicity of the reciprocal lattice, as shown in the lower half of Fig. 7. This is as would be expected. After all, if we were to slowly increase the width of the aperture, and hence let more and more pairs of real space delta functions contribute to the scattered amplitude, then the intensity would concentrate itself *only* at the reciprocal lattice points. However, from a signal-processing point of view, sampling *intensity* only at the reciprocal lattice points *under*samples the data by a factor of 2. In the simplest case of the Young's slits, if we only sample at the peaks of the intensity (shown as circles in Fig. 7), then from a computerized representation of the data we would conclude that the diffraction pattern has a numerical value of intensity of 4 at all points in the diffraction plane. In contrast, the Nyquist condition requires a function to be sampled at twice the frequency of the highest frequency component contributing to that function. In the case of intensity (as opposed to amplitude) the Nyquist sampling is $\Delta u/2$, not Δu.

This point is belabored here because I believe that the use of terms *double sampling* or *oversampling* in reference to diffracted intensity is a serious misnomer. The appropriate Nyquist sampling of the *intensity* of *any* diffraction pattern is incontrovertibly $\Delta u/2$. Of course, crystallographers only ever need to concern themselves the measurable intensity at the actual reciprocal lattice points. But from an information point of view, ignoring the gaps between diffraction orders introduces a massive quantity of *a priori* information: that for a perfect crystal all pixels lying between diffraction orders have zero intensity. Sayre's short paper in 1952 (Sayre, 1952) underlined this point, yet many workers in the field of PDI still refer to the need for oversampling (in the Nyquist sense) the diffraction *intensity* – a use that has sadly become widespread. Sampling intensity at half the reciprocal lattice periodicity exactly satisfies the Nyquist condition. On the contrary, conventional crystallography *undersamples* the diffracted intensity by a factor of 2: it can legitimately do this by assuming *a priori* that, for an infinite crystal, there is no intensity lying between the diffraction spots.

All noncrystallographic discussions of the phase problem infer this fact from the point of view of the width of the autocorrelation function or via simple consideration of the frequency-doubling effect of intensity relative to amplitude. It would seem, however, that Hoppe was the first author to explicitly state a direct method for the solution for the missing N phases from N diffraction orders via the 2^N numbers associated with the Nyquist frequency of intensity. In fact, he suggested that the ideal point to sample the intensity between the diffracted orders was at those points where the two adjacent sinc functions (arising from the ptychographical convolution) have identical modulus, as inferred from the form of the sinc functions when

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 31
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

scaled by the square root of the diffracted intensities of the crystallographic reflections (see Fig. 6).

I. Concluding Remarks on Classical Ptychography

We have seen that crystalline ptychography is a two-step process as described below:

1. A small region of the object is illuminated by a confined wave field, thus leading to the conventional crystallographic lattice points being convolved with the Fourier transform of the illumination function. As a result of this process, the amplitudes of adjacent lattice points are given the opportunity to interfere with one another. This means that if we now measure the diffracted intensity lying between the conventional diffraction spots, we can obtain an estimate of the phase difference between these amplitudes to within an uncertainty of the complex conjugate of one of these beams.
2. By shifting the illumination function (or an aperture placed in the object plane), the complex conjugate ambiguity can be resolved. This is because the act of shifting the illumination function leads to introduction of a phase ramp across the convolving function in reciprocal space. At points between the conventional diffraction spots, the net effect is to advance the phase of one of the interfering wave components and retard the other, thus leading to a new measurement that can determine which of the two possible solutions obtained in step 1 is correct.

In theory, the form of the illumination function can be engineered to facilitate the most efficacious measurements. If the illuminating function is of the form of an Airy disk (the Fourier transform of a circular top hat), then the convolution function is itself a circular top hat. This is very convenient because it means that by choosing an appropriate radius of the top-hat function, we can arrange for all the pairs of interferences to occur in well-defined regions of the diffraction plane, while at the same time allowing measurement of the intensity of each underlying beam, a configuration that naturally occurs in the STEM configuration (Fig. 1). For crystals with reciprocal lattice vectors that are not orthogonal and/or have magnitudes that are very different from one another, then the probe-forming aperture can be shaped appropriately to achieve this separation, as in the work of Hegerl and Hoppe (1970) and as discussed by Hoppe (1982).

These broad principles still apply if the illumination function is instead sharply defined in the object space, say via a rectangular aperture illuminated by a plane wave placed right against the object function. Most propitiously, if the aperture has the lateral dimensions of the unit cell, then the intensity of

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 32
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

the conventional diffraction orders are not altered at all. However, at points between the diffraction spots, amplitudes from more than two diffraction orders interfere with one another, because the Fourier transform of real-space aperture is of the form of a sinc function, which is extensive.

If we illuminate (an infinite) crystalline object with two coherent beams that are separated in reciprocal space by twice the Bragg angle, then what is measured in the diffraction plane is the coherent superposition of two diffraction patterns, one of which is displaced with respect to the other by one reciprocal lattice spacing. Each diffracted spot in this hybrid pattern now expresses the interference intensity of all pairs of diffracted beams separated by this reciprocal lattice displacement. This is essentially the same information as expressed in the STEM configuration, except that the illumination function consists not of a disk but of just the two points that interfere in the far field. Once again the complex conjugate ambiguity can be resolved by introducing a known phase shift into one of the coherent illumination beams. This phase change, of course, has the effect of shifting the (sinusoidal) pattern of illumination at the object plane in a way identical with all other flavors of ptychography. We will call an important variant on this data set the Bragg–Brentano subset (Section IV.B).

We note furthermore that Hoppe (1969a, 1969b) and Hoppe and Strube (1969) attempted to extend their first formulation of ptychography to non-crystalline objects but observed that the very simple interference phenomena that occur in crystalline ptychography no longer apply, and so an explicit solution algorithm for all the relevant phases did not at that time appear to be available. Even in his much later discussion of these matters, Hoppe (1982) seemed to conclude that the aperiodic ptychographic problem was intractable.

I refer to the above manifestations of ptychography as *classical ptychography* – the main characteristic being that we make use of intensity measurements occurring between the main diffracted peaks, and that these are measured at more than position of a confined illumination field incident upon the object. These classical manifestations can be mixed, refined, and extended. (I have concentrated on very confined subsets of the available data in order to reduce complexity of the description so far.) For example, Section II.E considered measuring intensity only at four points in the diffraction plane for two illumination beam positions. In fact, there is nothing to prevent measuring every single pixel in the diffraction plane for every single possible illumination function position in real space.

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 33
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

III. Coordinates, Nomenclature, and Scattering Approximations

This section establishes the basic geometry of the ptychographical data set for both small and large angle scattering and for 2D and 3D object functions. We also briefly explore the breakdown of various scattering approximations. This background is useful in discussions of the various experimental configurations of some principal variants of ptychography and how their associated methods of calculating the ptychographical image may or may not yield a useful representation of object function.

A. Geometry of the Two-Dimensional, Small-Angle Approximation

We begin with the very simplest statement of Fourier optics, which can found in any optics textbook; our purpose here is primarily to set up a coordinate system for later sections.

To be consistent with Eq. (2), we use

$$\psi_P(\boldsymbol{r}, t) = e^{i2\pi(\nu t - k\cdot r)} \tag{28}$$

as our definition of a traveling plane wave in free space of unity amplitude; where $\boldsymbol{r}$ is a 3D vector defining a point in free space, which can alternatively be expressed in Cartesian coordinates x, y, and z; t is time, and ν is the frequency of oscillation of the wave disturbance; and $\boldsymbol{k}$ is a 3D reciprocal vector, which can alternatively be expressed in reciprocal Cartesian coordinates u, v, and w; $\boldsymbol{k}$ points in the direction of travel of the wave, perpendicular to the planes of constant phase in $\psi_P(r)$, and is of length

$$|k| = \frac{1}{\lambda}, \tag{29}$$

where λ is the wavelength of the wave disturbance. Note that different traditions of nomenclature do not include the factor of 2π and/or use a $+$ or $-$ sign in the exponential term of the plane wave. This is not fundamental, merely conventional, but it does affect the sign and scaling of all reciprocal space quantities. We assume our source of radiation is monochromatic, so that we can ignore time evolution of the intensity of the wave disturbance.

Consider the experiment shown in Fig. 8 where the scattering angles are all small (a couple of degrees, as is common in electron microscopy), and the plane of the detector is described by coordinates (x', y'), parallel to (x, y) in the object plane. The z-axis lies along the optic axis perpendicular to these planes. Let the illumination function be just 2D and described by the stationary time-independent complex function $a(x, y)$. Let the object be

aiep y.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 34
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

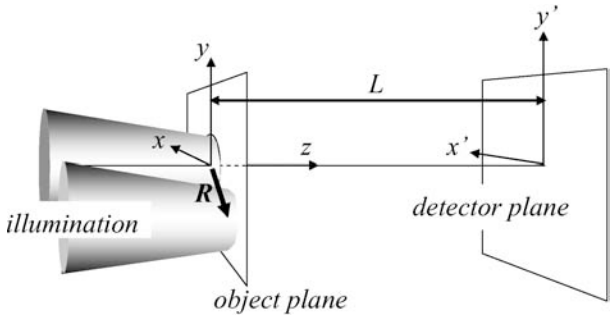


FIGURE 8. Definitions of coordinates. Illuminating radiation is incident from the left and can be shifted to a position $\boldsymbol{R}$ in the x–y plane. The distance between the object and detector plane is L. In this diagram, the size of the illumination relative to L has been grossly exaggerated.

described by a 2D transmission function $q(x, y)$, where

$$q(x, y) = q_M(x, y)e^{i\phi(x,y)}, \tag{30}$$

with $q_M(x, y)$ being a real number modulus between 0 and 1 representing the absorption of the object (0 being total absorption, 1 being total transparency), and $\phi(x, y)$ being a real number representing the phase change measured in radians induced into the incident wave as it passes through the transmission function. The exit wave immediately downstream of the object is then simply

$$\psi_e(x, y) = q(x, y) \cdot a(x, y), \tag{31}$$

which we will call the *multiplicative approximation*. In fact, in any real experiment, the object must have finite thickness, so this approximation is never formally satisfied.

Let the detector be a distance L downstream of the specimen plane (at $z = L$). The distance D between a point (x, y) in the object plane and a detector pixel at (x', y') is given by

$$D = \sqrt{L^2 + (x - x')^2 + (y - y')^2} = L\left(1 + \frac{(x' - x)^2 + (y' - y)^2}{L^2}\right)^{\frac{1}{2}}. \tag{32}$$

According to the usual approximations of the small-angle limit, we expand the square root to second order. If the total width, d, of $\psi(x, y)$ is small relative to L and to the dimensions of the detector (certainly always true with X-rays and electrons), we can further approximate that x^2 and y^2 are negligible relative to $x \cdot x'$ and $y \cdot y'$, so that

$$D \approx L + \frac{(x')^2}{2L^2} + \frac{(y')^2}{2L^2} - \frac{x \cdot x'}{L} - \frac{y \cdot y'}{L} + \cdots. \tag{33}$$

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 35
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

By Huygens' principle the propagation of the exit wave $\psi_e(x, y)$ to the plane of the detector through free space can be calculated by adding spherical waves, each one emanating from an elemental area $dx\,dy$ in the x–y plane and propagating to the x'–y' plane. The amplitudes of such waves decay as $1/D$ and they have a phase of $2\pi D/\lambda$. The total amplitude of the wave reaching the detector at a point (x', y'), which we will call $M(x', y')$, is simply a sum of such waves, so that

$$M(x', y') \approx \frac{\mu}{L} e^{-i\frac{2\pi}{\lambda}(L+\frac{(x')^2}{2L^2}+\frac{(y')^2}{2L^2})} \iint \psi_e(x, y) e^{\frac{i2\pi}{\lambda}(\frac{x\cdot x'}{L}+\frac{y\cdot y'}{L})}\,dx\,dy, \tag{34}$$

where μ is a constant. (Note that we do not account for obliquity factors in this discussion.) In fact, we cannot measure $M(x', y')$; we can only measure the intensity of the wave disturbance at the detector, namely

$$I(x', y') = M(x', y') \cdot M^*(x', y'). \tag{35}$$

Although this equation presents us with all the difficulties of the phase problem, it does at least have the advantage of obliterating the messy phase terms that precede the integral in Eq. (34).

The first factor $(1/L)$ in Eq. (34) is an approximation of $1/D$: true when L is large and the range of x' and y' is relatively small. The second factor is just the absolute phase of the wave as it would arrive at the center of the detector. The third factor accounts for the extra phase that is introduced at points off-axis in the detector plane, because the spherical wave must propagate further to reach these points; the spherical surface is here approximated by a parabola. On forming intensity we obtain

$$I(x', y') \approx \frac{\mu^2}{L^2} \left| \iint \psi(x, y) e^{i\frac{2\pi}{\lambda}(\frac{x\cdot x'}{L}+\frac{y\cdot y'}{L})}\,dx\,dy \right|^2. \tag{36}$$

The constant μ absorbs a number of theoretical factors, including a phase term to account properly for the Huygens' wavelet construction (see, for example, Born and Wolf, 1999). However, in any experiment these are of little concern relative to uncertainties in the efficiency of the detector and the absolute intensity of the illumination function $|a(r)|^2$. In practice, all ptychographical methods will give an output of arbitrary modulus because of these unknowns. We will call the output a *ptychograph* and label it Π_{sub}, where the subscript denotes the particular variant of the data processing algorithm we use (Section IV). Any structural information of interest is encoded in the relative magnitude and phase of Π_{sub} over the field of view. The absolute magnitude of Π_{sub} is of little interest and its phase is unaffected by the absolute of intensity. This is directly analogous to a conventional

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 36
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

microscope where the user has very little interest in the precise calibration of the brightness of the illumination, as long as it is sufficient to see the structures of interest. Henceforth we will therefore ignore the constant terms μ/L.

For the purposes of our theory, we can further remove the physical effects of the scaling of the diffraction pattern by the camera length L and the wavelength by changing the coordinates of the measurement plane to $u \approx x'/\lambda L$ and $v \approx y'/\lambda L$, and thus write

$$M(u, v) = \iint \psi(x, y) e^{i2\pi(x\cdot u + y\cdot v)}\, dx\, dy \tag{37}$$

or, more compactly,

$$M(\boldsymbol{u}) = \int \psi(\boldsymbol{r}) e^{i2\pi \boldsymbol{u}\cdot\boldsymbol{r}}\, d\boldsymbol{r} = \Im\psi(\boldsymbol{r}), \tag{38}$$

where $\boldsymbol{u}$ is a 2D detector coordinate vector $\boldsymbol{u} = u\hat{\boldsymbol{u}} + v\hat{\boldsymbol{v}}$, with $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ as unit vectors in reciprocal space; and $\boldsymbol{r}$ is a 2D vector in real space, $\boldsymbol{r} = x\hat{\boldsymbol{x}} + y\hat{\boldsymbol{y}}$.

We can move our illumination function to a position defined by a 2D vector $\boldsymbol{R}$, so that

$$\psi(\boldsymbol{r}) = a(\boldsymbol{r} - \boldsymbol{R}) \cdot q(\boldsymbol{r}). \tag{39}$$

The complete ptychographical data set for 2D functions a and q is then defined as the 4D intensity distribution

$$I(\boldsymbol{u}, \boldsymbol{R}) = \left|M(\boldsymbol{u}, \boldsymbol{R})\right|^2 = \left|\Im\big(a(\boldsymbol{r} - \boldsymbol{R}) \cdot q(\boldsymbol{r})\big)\right|^2, \tag{40}$$

which can be written as the ptychographical convolution

$$I(\boldsymbol{u}, \boldsymbol{R}) = \left|\big(A(\boldsymbol{u}) e^{i2\pi R\cdot u}\big) \oplus Q(\boldsymbol{u})\right|^2, \tag{41}$$

where $A(\boldsymbol{u}) = \Im a(\boldsymbol{r})$ and $Q(\boldsymbol{u}) = \Im a(\boldsymbol{r})$. We will call this the *fundamental ptychographic equation*.

These equations are the basis of all the theory of both the "direct" deconvolution methods developed by the Cambridge group in the 1990s (see Sections IV.A, IV.B, and IV.D) and the iterative solutions to the ptychographical data set (see Section IV.C). The nomenclature of the Cambridge work was borrowed from the image processing literature: $\boldsymbol{R}$ was named $\boldsymbol{\rho}$, $\boldsymbol{u}$ was named $\boldsymbol{r}'$, and the reciprocal of $\boldsymbol{\rho}$ was named $\boldsymbol{\rho}'$, which we will call $\boldsymbol{r}$. The biggest potential source of confusion is that what we will call $\boldsymbol{U}$ (the back Fourier transform coordinate of $\boldsymbol{u}$) was formerly named $\boldsymbol{r}$. The rationale of this altered nomenclature is to unify it with the reciprocal u–v coordinates widely used by Cowley and others (e.g., Cowley, 1995). Furthermore, the use of the dashed nomenclature for reciprocal coordinates becomes clumsy in thinking of 3D scattering in terms of the existing TEM nomenclature

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 37
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

(Section III.B). Note that the function $a(\boldsymbol{r})$ was in some earlier articles (e.g., Rodenburg and Bates, 1992) called $P(\boldsymbol{r})$ – equivalent to the "probe" function in STEM.

Perhaps a rather misleading nomenclature in this former schema was to call the object function $\psi(\boldsymbol{r})$ (which we have called here the transmission function $q(\boldsymbol{r})$). The author's intention was to suggest that the function we were aiming to solve for was the exit wave as it would appear under plane wave illumination perpendicular to the (infinitesimally thin) object. However, this led to some misunderstanding because many colleagues in the field of electron scattering supposed that this was also meant to represent the exit wave as it would appear (under plane wave illumination) from a *thick* object. Since the exit wave in electron microscopy arising from thick, strongly scattering objects bears no obvious relationship to the form of the scattering potential in the object (especially given a range of incident vectors necessary to form a localized probe), much of this work was dismissed as being based on a simply wrong assumption, or that the method could only possibly work for thin, weak, kinematically scattering objects. In fact, ptychography is remarkably robust to dynamical scattering and thickness effects even in the case of crystalline objects of small unit cell (Plamann and Rodenburg, 1998), for reasons discussed briefly in Section IV.A.

B. Three-Dimensional Geometry

Another way of thinking of the integral in Eq. (36) is via Fig. 9(a). The Fraunhofer approximation is equivalent to saying that only parallel beams scattered from the object interfere with one another. It is easy to show that this condition is met if

$$L \gg \frac{d^2}{\lambda}. \tag{42}$$

If we say that in the absence of the transmission function the phase of the Huygens' wavelet emanating from $x = y = z = 0$ is zero, then the extra phase introduced to a beam (or ray) emanating from a point (x, y) at $z = 0$ (that is, from within the 2D) is determined only by the two orthogonal angles β_x and β_y. β_x is the angle between the plane $x = 0$ and the plane containing both the y axis and the line of the ray in question (as shown in Fig. 9(b)); β_y is defined similarly with respect to the plane $y = 0$. If we redefine

$$u = \frac{\sin \beta_x}{\lambda} \tag{43a}$$

and

$$v = \frac{\sin \beta_y}{\lambda}, \tag{43b}$$

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 38
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

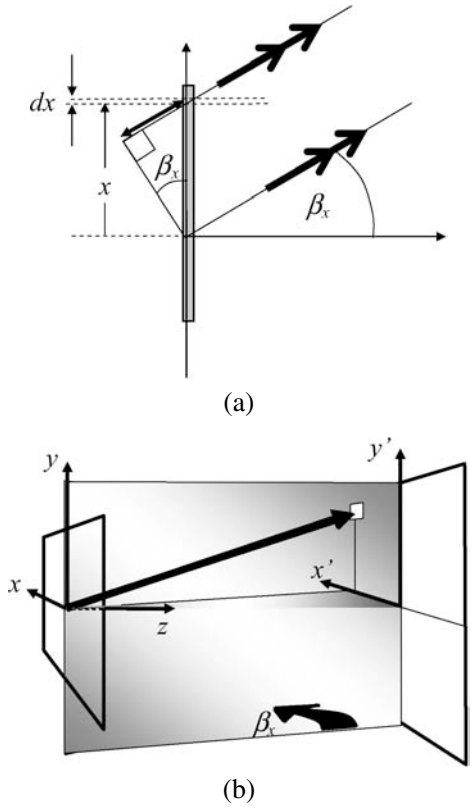


FIGURE 9. (a) Parallel beams scattered by the object will meet up and interfere in the Fraunhofer diffraction plane. For this 1D object, the path length of the beam emanating from the elemental length dx positioned at x is shorter than the beam scattered at the origin by an amount $x \cdot \sin \beta_x$, and consequently has an extra phase change $2\pi x \cdot \sin \beta_x / \lambda$, relative to the waves scattered from $x = 0$. By substitution from Eq. (43a), the total resultant amplitude is hence the Fourier transform $F(u) = \int f(x) e^{i2\pi ux}\, dx$, where $f(x)$ is the exit wave immediately downstream of the object. For a 3D object, similar arguments lead to Eq. (45). (b) Definition of β_x. The heavy arrow subtends from the specimen to a detector pixel. β_x is the angle between the plane $x = 0$ and the plane containing the heavy arrow and the y-axis.

then, with reference to Fig. 9(a), the resulting integral of all such beams over the exit wave is identical to Eq. (37). These rectilinear angular coordinates have the advantage of being correct at all (large) angles where 3D scattering effects dominate. We can think of the far-field scattered wave function as existing over a spherical hemisphere (Fig. 10(a)), of diameter L (satisfying Eq. (42)). We index the coordinates on this surface by the coordinates u and v. In Fig. 10(b) we draw the view of this hemisphere looking back at it from a

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 39
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

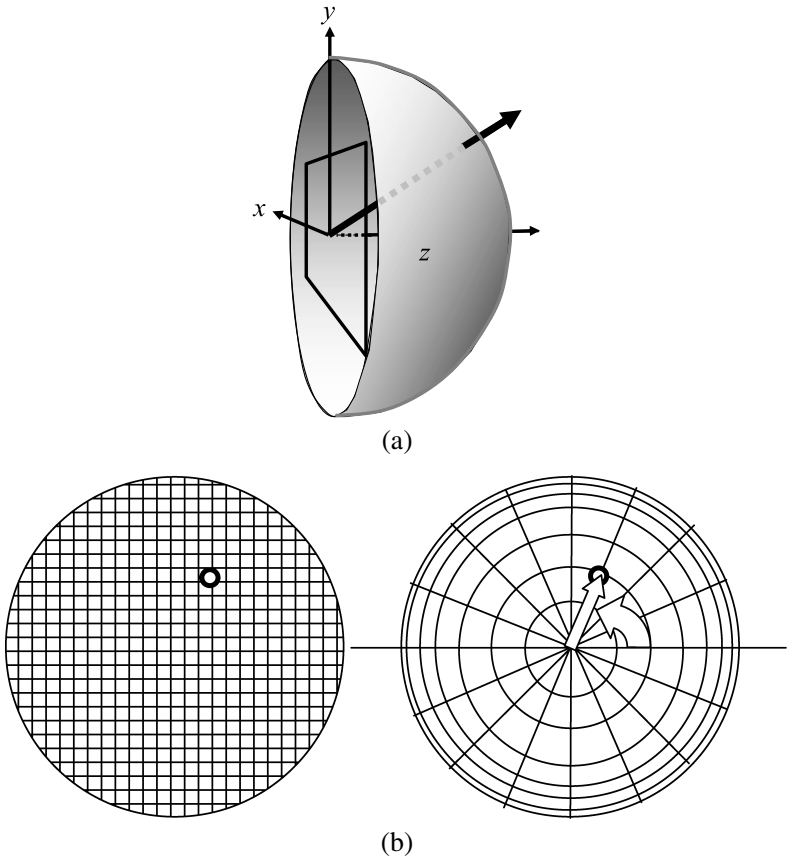


FIGURE 10. (a) Downstream of the object plane we can construct a spherical hemisphere. When the Fraunhofer condition is satisfied, it is the phase and modulus of the spherical wave that impinges on this surface that we aim to calculate to solve the scattering phase problem. This hemisphere is the real-space equivalent of the Ewald sphere in reciprocal space. (b) Comparison of rectilinear (u, v) and spherical coordinates (θ, ϕ), as seen from behind the hemisphere in part (a).

point a long way off on the z-axis. u and v project a square set of coordinates over this hemisphere; this is in contrast to conventional spherical coordinates (also illustrated schematically in Fig. 10(b)). Of course, any practical flat detector must have a further mapping function to convert its pixel coordinates into the angles β_x and β_y, and hence u and v. Other trigonometric factors must account for the varying angle incidence upon the outermost detector pixels. A further factor accounting for the $1/D$ decay of the spherical waves can scale the intensity of the signals recorded at these high angles of scatter. The important point is that when we think of solving the phase problem over

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 40
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

the entire range of scattering angles available, then the phase we are trying to solve manifests itself in the absolute phase of the spherical wave that impinges upon this hemispherical surface, and that when our processing involves a rectilinear detector and data manipulations involving fast Fourier transforms of functions sampled on equally spaced arrays, then it is very convenient to first map the detected intensities onto the coordinates u and v. We assume this mapping in all that follows.

Once we begin to deal with high scattering angles, then implicitly the 2D transmission object function also breaks down. Quite simply, if we look at a 3D object from a number of different angles, then the features on the top and bottom surfaces of the object appear to move with respect to one another. In this section, vector quantities $\boldsymbol{r}$ and $\boldsymbol{k}$ will always be 3D (in contrast to the previous section) and will sometimes be written in terms of coordinates (x, y, z) and $(u, v, w,)$ respectively.

Let the object in three dimensions be a function $V(\boldsymbol{r})$; its 3D Fourier transform is

$$F(\boldsymbol{k}) = c \int V(\boldsymbol{r}) e^{2\pi \boldsymbol{k}\cdot r}\, d\boldsymbol{r}, \tag{44}$$

where c is a constant. $\boldsymbol{k}$ is a 3D vector in reciprocal space that can be decomposed into unit vectors $\boldsymbol{k} = u\hat{\boldsymbol{u}} + v\hat{\boldsymbol{v}} + w\hat{\boldsymbol{w}}$. (Note that the vector $\boldsymbol{u}$ will be reserved for the 2D detector vector coordinate $\boldsymbol{u} = u\hat{\boldsymbol{u}} + v\hat{\boldsymbol{v}}$.) In the scalar theory of light optics, $V(\boldsymbol{r})$ is called the *optical potential*. In electron scattering theory, $V(\boldsymbol{r})$ is the spatial distribution of the atomic potential in the object, measured in volts. $V(\boldsymbol{r})$ can be complex, the imaginary components representing absorption of the radiation. The perfect imaging machine would aim to estimate accurately, and at very high resolution, $V(\boldsymbol{r})$.

$V(\boldsymbol{r})$ must be estimated from knowledge of the distribution of the incident radiation, which has k-vectors that will call $\boldsymbol{k}_i$, and the scattered radiation, which has k-vectors we will call $\boldsymbol{k}_s$. If the object is strongly scattering or thick, then the incident waves striking parts of it some distance from the entrance surface will be substantially different from what would have arrived at that region had the illumination function been allowed to propagate through free space. Under these circumstances the inference of $V(\boldsymbol{r})$ is in general nontrivial, even if the modulus and phase of the scattered radiation can be measured or calculated perfectly.

If the object scatters weakly (whether or not it is thick), then we can use the first Born approximation (Mott and Massey, 1965). That is to say, the intensity of the incident beam is not substantially attenuated as a result of passing through the object and the total intensity scattered away from the direction of the incident beam is relatively low. We will discuss this (so-called kinematical) approximation in the next section, as well as its relationship to

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 41
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

the weak phase grating approximation for a 2D object function. For a single incident plane wave of k-vector $\boldsymbol{k}_i$, the amplitude of the wave scattered into the far field in the direction of the k-vector $\boldsymbol{k}_s$, is given by

$$M_0(\boldsymbol{k}_s) = \delta(\boldsymbol{k}_i) - ic \int V(\boldsymbol{r}) e^{i2\pi(\boldsymbol{k}_s - \boldsymbol{k}_i)\cdot \boldsymbol{r}}\, d\boldsymbol{r}, \tag{45}$$

where the real constant c arises from various factors that scale the strength of the scattering to the voltage $V(\boldsymbol{r})$. The dot product in the exponential follows directly from considering the phase change induced from scattering at an elemental volume positioned at $\boldsymbol{r}$ in 3D space relative to the wave scattered from the volume at $\boldsymbol{r} = 0$. The subscript on M_0 is used to denote that in this coordinate system the center of the Ewald sphere lies at the origin of k-space. We also assume for simplicity that the incident wave has unity amplitude; the unattenuated transmitted beam that appears in the kinematical theory is represented by the delta function in Eq. (45). The scattered wave is $-\pi/2$ out of phase with the transmitted beam, this indicating that by passing through the potential well which is doing the scattering, the phase of the wave is advanced slightly, and hence it acquires a small negative (given our definition in Eq. (28)) imaginary component relative to the unscattered wave.

In addition to satisfying the Fraunhofer condition (Eq. (42)), we also require that we satisfy the Nyquist sampling frequency (Section II.H) in the detector plane. The minimum detector pixel size is therefore

$$\Delta\beta \approx \frac{\lambda}{2d}, \tag{46}$$

where $\Delta\beta$ is the notional pixel size over our hemispherical surface, indexed as before by our orthogonal angular coordinates β_x and β_y. Here d is not just the lateral width of the illuminated area of the object but is the approximate linear dimension of the illuminated volume in a direction perpendicular to the ray or beam corresponding to a particular angle (β_x, β_y). For example, in STEM, the aspect ratio of a focused electron probe is large (it is thin but long in the z direction), so actually the demands on the sampling condition even at relatively small angles (2 to 3 degrees) become demanding (smaller detector pixels are required at higher scattering angles). This is easily observed experimentally; if the object is amorphous, we see rings of speckled intensity becoming finer and finer as we move away from the optic axis (Rodenburg, 1988).

To satisfy this limitation on d, we must confine our illumination by having a range of incident plane waves over a range of $\boldsymbol{k}_i$; even if we use parallel illumination combined with an aperture (or support function) mounted co-incident with the specimen (as ptychography was originally envisaged by Hoppe), then this function itself diffracts the illuminating wave and hence

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 42
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

implies a corresponding range of incident vectors that form the ptychographic convolution in the diffraction plane (Eq. (41)). We must therefore perform a second integral over this range of $\boldsymbol{k}_i$. Let the modulus and phase of the incident plane waves be given by complex-valued function $A_0(\boldsymbol{k}_i)$, so that

$$M_0(\boldsymbol{k}_s) = A_0(\boldsymbol{k}_i) - ic \iint A_0(\boldsymbol{k}_i) V(\boldsymbol{r}) e^{-i2\pi(\boldsymbol{k}_s - \boldsymbol{k}_i)\cdot \boldsymbol{r}}\, d\boldsymbol{r}\, d\boldsymbol{k}_i. \quad (47)$$

Substituting from Eq. (44), the second term is of the form of a convolution:

$$\begin{aligned} M_0(\boldsymbol{k}_s) &= A_0(\boldsymbol{k}_i) - ic \int A_0(\boldsymbol{k}_i) F(\boldsymbol{k}_s - \boldsymbol{k}_i)\, d\boldsymbol{k}_i \\ &= A_0(\boldsymbol{k}_i) - icA_0(\boldsymbol{k}_i) \oplus F(\boldsymbol{k}_i). \end{aligned} \quad (48a)$$

Note that $\boldsymbol{k}_i$ and $\boldsymbol{k}_s$ describe vectors lying in the same (reciprocal) space, and so we could write this simply as

$$M_0(\boldsymbol{k}) = A_0(\boldsymbol{k}) - icA_0(\boldsymbol{k}) \oplus F(\boldsymbol{k}) \quad (48b)$$

but we retain the subscripts in order to emphasize the relationship to the conventional Ewald sphere construction.

Because we are dealing with monochromatic radiation

$$|\boldsymbol{k}_i| = |\boldsymbol{k}_s| = \frac{1}{\lambda}, \quad (49)$$

which means the functions $A_0(\boldsymbol{k}_i)$ and $M_0(\boldsymbol{k}_s)$ have zero value at all points in reciprocal space except over the spherical surface of radius $1/\lambda$. We define a vector $\boldsymbol{k}_0$ (also of length $1/\lambda$) that lies parallel with optic axis: that is, in the direction of w axis in reciprocal space and in the direction of z axis in real space.

Let the angle between $\boldsymbol{k}_0$ and $\boldsymbol{k}_i$ (or $\boldsymbol{k}_s$) be described by the two orthogonal angles β_u and β_v. β_u is the angle between the plane $u = 0$ in reciprocal space and the plane containing both the v axis and the vector $\boldsymbol{k}_i$ (or $\boldsymbol{k}_s$); β_v is defined similarly with respect to the plane $v = 0$. These angles are identical to the real-space angles β_x and β_y and map onto u and v as before:

$$u = \frac{\sin\beta_u}{\lambda} \quad \text{and} \quad v = \frac{\sin\beta_v}{\lambda}. \quad (50)$$

We further define

$$\boldsymbol{k}_a = \boldsymbol{k}_i - \boldsymbol{k}_0, \quad (51)$$

and

$$\boldsymbol{k}_m = \boldsymbol{k}_s - \boldsymbol{k}_0. \quad (52)$$

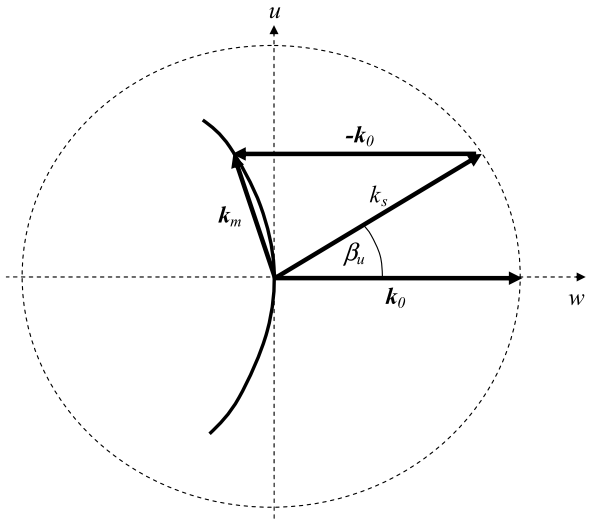


FIGURE 11. The k-vector relationships relative to reciprocal coordinates u and w (this diagram is in the plane $v = 0$), and the physical scattering angle $\beta_u = \beta_x$. The heavy circular line is the conventional Ewald sphere (passing through the origin of k-space).

The latter relationships are illustrated in Fig. 11. Let

$$A_a(\boldsymbol{k}_a) = A_0(\boldsymbol{k}_0 + \boldsymbol{k}_i) \tag{53}$$

and

$$M_m(\boldsymbol{k}_m) = M_0(\boldsymbol{k}_0 + \boldsymbol{k}_m), \tag{54}$$

which means the corresponding spherical surfaces of constant $\boldsymbol{k}_i$ or $\boldsymbol{k}_s$ pass through the origin of k-space, as with the conventional Ewald sphere.

Like A_0 and M_0, we can think of A_a and M_m as 3D functions of a single set of coordinates in k-space; as before, we do not write them both as a function of just $\boldsymbol{k}$ because the form of these functions depend on our particular definition of $\boldsymbol{k}_0$, and this in turn feeds into the scattering integral and the particular solution we obtain in the ptychographical reconstruction. It should be noted that if we think of ptychography as reconstructing the exit wave from the object (we will see it does not actually do this when the object is thick), then the plane of the reconstruction is perpendicular to $\boldsymbol{k}_0$. From an experimental point of view, it is usual to choose $\boldsymbol{k}_0$ to represent the average direction of the incident waves. A line (the optic axis) drawn between the object and the detector in the direction of $\boldsymbol{k}_0$ will arrive at a pixel that we define as the center of the u–v detector coordinate system.

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 44
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

Now both $A_a(\boldsymbol{k}_a)$ and $M_m(\boldsymbol{k}_m)$ only have a value over the Ewald sphere. We define a function $\Omega(u, v, w)$ which of the form of a "spherical delta function" in reciprocal space so that it satisfies (for any function $G(\boldsymbol{k})$),

$$\int G(\boldsymbol{k})\Omega(\boldsymbol{k}_w - \boldsymbol{k})\,d\boldsymbol{k} = \begin{cases} G(\boldsymbol{k}) & \text{for } \boldsymbol{k} = \boldsymbol{k}_w, \\ 0 & \text{for } \boldsymbol{k} \neq \boldsymbol{k}_w, \end{cases} \tag{55}$$

where, for given values of u and v, $\boldsymbol{k}_w$ is constrained to the locus of points

$$k_w = u\hat{\boldsymbol{u}} + v\hat{\boldsymbol{v}} + \frac{1}{\lambda}\Big(\big(1 - \lambda^2\big(u^2 + v^2\big)\big)^{\frac{1}{2}} - 1\Big)\hat{\boldsymbol{w}}. \tag{56}$$

(The unit vector $\hat{\boldsymbol{u}}$ should not be confused with the 2D detector vector $\boldsymbol{u}$.)

We can describe fully the 3D functions $A_a(\boldsymbol{k}_a)$ and $M_m(\boldsymbol{k}_m)$ by the 2D complex functions $A(\boldsymbol{u})$ and $M(\boldsymbol{u})$, where

$$A(\boldsymbol{u}) = A(u, v) = \Omega(u, v, w) \cdot A_a(\boldsymbol{k}_a) \tag{57}$$

and

$$M(\boldsymbol{u}) = M(u, v) = \Omega(u, v, w) \cdot M_m(\boldsymbol{k}_m), \tag{58}$$

where it must be remembered that u, v uniquely define w via Eq. (56). The scattering Eqs. (48a) and (48b) now become

$$M_m(\boldsymbol{k}_m) = A_a(\boldsymbol{k}_a) - ic\int A_a(\boldsymbol{k}_a)F(\boldsymbol{k}_m - \boldsymbol{k}_a)\,d\boldsymbol{k}_a \tag{59}$$

or

$$M(\boldsymbol{u}) = \Omega(\boldsymbol{k}) \cdot \big[A_a(\boldsymbol{k}) - icA_a(\boldsymbol{k}) \oplus F(\boldsymbol{k})\big]. \tag{60}$$

In the object space we have a 3D illumination function $a(\boldsymbol{r})$, which at $z = 0$ is simply the addition of the incident plane waves defined by the amplitudes $A(\boldsymbol{u})$, that is,

$$P(x, y, 0) = \iint A(u, v)e^{-i2\pi(u\cdot x + v\cdot y)}\,du\,dv = \Im^{-1}A(\boldsymbol{u}). \tag{61}$$

If the illumination is physically created by a lens (as in the STEM geometry), then the phase of $A(\boldsymbol{u})$ will depend on its aberrations.

Let us suppose that the lens is perfect. In the absence of a specimen, we simply have the transmitted beam in the detector plane, so that $I(\boldsymbol{u}) = M(\boldsymbol{u}) \cdot M^*(\boldsymbol{u}) = A(\boldsymbol{u})A^*(\boldsymbol{u})$ and so it is somewhat meaningless to worry about the exact z (or x, y) displacement of $a(\boldsymbol{r})$. However, as soon as we introduce a specimen, its position in the z-direction will strongly affect $M(\boldsymbol{u})$. How is this reflected in Eq. (60)? For kinematical scattering, in real space have $a(\boldsymbol{r})$ plus a term $-ic \cdot a(\boldsymbol{r}) \cdot V(\boldsymbol{r})$, the latter being a source of waves scattered by the potential, $\pi/2$ out of phase with the illuminating function.

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 45
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

Let us consider a single scattering center positioned at $x = y = z = 0$. If we shift the illuminating beam by a distance of z_0 along the optic axis, then we would expect to see a propagated illumination function $P(x, y, y - z_0)$ incident upon this scattering center. By the Fourier shift theorem, our 3D function representing the probe in reciprocal space now has a phase change introduced into it of $e^{i\phi}$, where $\phi = 2\pi z_0 w$. This means that when we form $A(\boldsymbol{u})$ via Eq. (60), the dependence of w on u and v will introduce a phase change in $A(\boldsymbol{u})$ equal to

$$\phi = z_0 \frac{2\pi}{\lambda}\left(\left(1 - \lambda^2\left(u^2 + v^2\right)\right)^{\frac{1}{2}} - 1\right), \tag{62}$$

which, for small values of u and v, approximates to the familiar defocus term

$$\phi \approx z_0 \lambda \pi \left(u^2 + v^2\right). \tag{63}$$

In other words, the phase change induced in reciprocal space as a result of a z-direction displacement of the specimen or probe function is equivalent to a defocus term when this shift is expressed in reciprocal space over the curvature of the Ewald sphere.

To clarify this scattering geometry in terms of the conventional Ewald sphere construction, we refer to Fig. 12. Once we have defined $\boldsymbol{k}_0$, then the

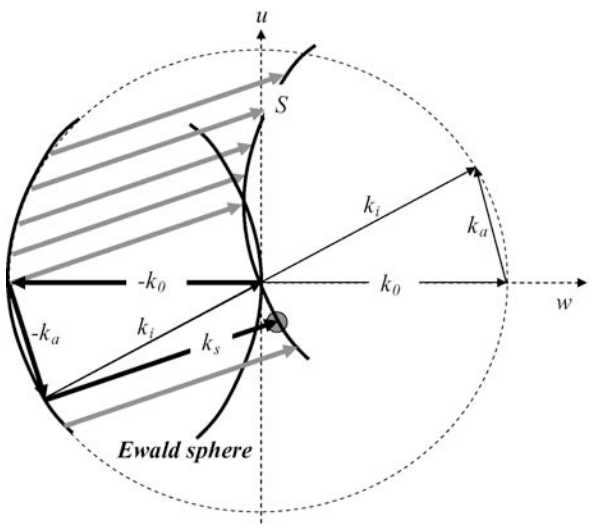


FIGURE 12. For a particular incident plane wave (e.g., $\boldsymbol{k}_i$ as shown) and scattering vector, $\boldsymbol{k}_s$, the amplitude of scattered wave is given by the value of $F(\boldsymbol{k})$ at the center of the small, gray circle. In fact, there is a range of incident vectors; hence the amplitude at any one scattering direction $\boldsymbol{k}_s$ is the integral over the surface S in $F(\boldsymbol{k})$.

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 46
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

scattering amplitude in the $\boldsymbol{k}_s$ direction is just the integral over the curved surface labeled S. The Ewald sphere is conventionally drawn with the end of the incident k-vector $\boldsymbol{k}_i$ hinged at the origin. It would probably be more logical to draw this vector pointing in the opposite sense ($-\boldsymbol{k}_s$), with the vector $-\boldsymbol{k}_a$ being defined as above (as in Fig. 12). The curvature of S is then obviously in the opposite sense to that of the Ewald sphere, because the scattering Eq. (47) dictates the form $\boldsymbol{k} = \boldsymbol{k}_s - \boldsymbol{k}_i = \boldsymbol{k}_s - (\boldsymbol{k}_0 + \boldsymbol{k}_a)$. We also observe that the phase change discussed in relation to the shift along the optic axis of the illumination function is equivalent to a shift of the specimen function (that is, a phase change in $F(\boldsymbol{k})$) in the opposite direction, again because of the opposite curvatures of $A_a(-\boldsymbol{k}_a)$ and the Ewald sphere.

C. Dynamical Scattering

In the various data-processing methods and algorithms discussed in Section IV, there is no restriction on the form of either of the complex quantities $q(\boldsymbol{r})$ or $a(\boldsymbol{r})$ as they appear in the fundamental ptychographical Eq. (41). If we assume the multiplicative approximation (Eq. (31), then all these various computational manipulations will render $q(\boldsymbol{r})$. The first test calculations on various ptychographical inversion methods (for example Bates and Rodenburg, 1989) often used rather unrealistic objects, for example isolated delta functions. In other words, $q(\boldsymbol{r})$ was modeled as being mostly opaque with a low proportion of pixels designated with unity transmission. Again, this followed from the image-processing tradition where it is well established that if you cannot get a phase-retrieval algorithm working for a series of delta functions, it is unlikely to work for more complicated, fully structured or textured objects. This sort of extreme object does not bear any relationship to the weak phase object approximation as commonly used in electron microscopy. In this, we assume the object can be described by

$$q(\boldsymbol{r}) = 1 - ic\vartheta(\boldsymbol{r}), \tag{64}$$

where c is a constant and

$$c\big|\vartheta(\boldsymbol{r})\big| \ll 1. \tag{65}$$

When we undertake a Fourier transform of a weak phase object $q(\boldsymbol{r})$, we obtain a diffraction pattern of the form of

$$Q(\boldsymbol{u}) = \Im q(\boldsymbol{r}) = \delta(\boldsymbol{u}) - icF(\boldsymbol{u}). \tag{66}$$

The delta function (which only has value at $\boldsymbol{u} = 0$) represents the unscattered beam. Now, in conventional electron imaging theory, this approximation has two important consequences. First, according to the Born approximation, the

aiep v.2005/11/15 Prn:6/12/2007; 13:13
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

second term $F(\boldsymbol{u})$ can be related directly to the scattering potential $V(\boldsymbol{r})$ in real space. Indeed, $F(\boldsymbol{u})$ is just the Fourier transform of $V(\boldsymbol{r})$, a fact already used in Section III.B (compare Eqs. (65) and (45)). The second use of this approximation is that the existence of a strong unscattered beam is essential to conventional electron imaging theory. We want the image to have contrast that is proportional to atomic potential. In fact, even if the object introduces large phase changes into the exit wave (in which case Eq. (65) no longer applies), then we will still see nothing in the electron image if the lens is perfect and the modulus of the object transmission grating is unity. To see contrast, we must introduce a Zernike-type phase plate into the back focal plane of the lens (achieved by a combination of defocus and spherical aberration) to make the second term in the above equation be rotated in the complex plane by $\pi/2$. In this way, information residing in the scattered waves $F(\boldsymbol{u})$ is expressed on the intensity of image. If the weak phase approximation breaks down, the conventional bright-field contrast is no longer a linear function of the atomic potential.

How does this scattering approximation affect ptychography? If the object is a truly infinitesimally thin transmission grating, the multiplicative approximation holds true. Within the limitations of the particular inversion method used, the ptychograph will be an accurate representation of the exit wave we would have obtained from the object had it been illuminated by a plane wave. In other words, for thin objects, ptychography puts no restrictions on the scattering strength of the object. This statement is true of any exit wave phase-retrieval method such as holography or through-focal series reconstruction. The point is that ptychography does not need the specimen to be weak as part of the image-generation process itself, as does the conventional bright-field image. In this sense, all exit wave methods are superior to the conventional image mechanism. If the object is thin (but possibly strong), we might hope that the phase of the exit wave field is roughly proportional to the projection of the atomic potential within this hypothetical object.

We can think of the equivalence of multiple scattering with the object being a strong phase grating as follows. When the phase change is strong, then the object is of the form

$$q(\boldsymbol{r}) = e^{i\phi(\boldsymbol{r})}, \tag{67}$$

where $\phi(r)$ can be large, even wrapping around several in the complex plane. We can regard the weak phase approximation as simply the first two terms in the expansion of $e^{i\phi(\boldsymbol{r})}$. Further terms are given by

$$q(\boldsymbol{r}) = e^{-i\phi(\boldsymbol{r})} = 1 - i\phi(r) - \frac{\phi^2(r)}{2} + i\frac{\varphi^3(r)}{6} + \cdots \tag{68}$$

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 48
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

so that means that the Fourier transform is (using the convolution theorem)

$$Q(\boldsymbol{u}) = \delta(\boldsymbol{u}) - i\Phi(\boldsymbol{r}) - \frac{1}{2}\Phi(\boldsymbol{r}) \oplus \Phi(\boldsymbol{r}) + \frac{i}{6}\Phi(\boldsymbol{r}) \oplus \Phi(\boldsymbol{r}) \oplus \Phi(\boldsymbol{r}) + \cdots. \tag{69}$$

The higher-order terms are the diffraction pattern convolved with itself; these terms represent the multiple scattering. All the beams in the diffraction pattern can act as a source of a new diffraction pattern centered on that beam.

Our problems therefore do not arise from dynamical scattering *per se*, but from the 3D scattering (as in Section III.B) *combined* with dynamical scattering. Think of this in terms of a multislice calculation. Suppose we simply switch off the propagation term introduced between each elemental weak phase grating. The result is just a product of all the weak phase changes introduced by each layer of the object (Fig. 13). This is simply

$$\psi_e(\boldsymbol{r}) = e^{i\varphi_1(\boldsymbol{r})}e^{i\varphi_2(\boldsymbol{r})}e^{i\varphi_3(\boldsymbol{r})}e^{i\varphi_4(\boldsymbol{r})}e^{i\varphi_5(\boldsymbol{r})}e^{i\varphi_6(\boldsymbol{r})}\cdots, \tag{70}$$

where the $\varphi_j(\boldsymbol{r})$ is the weak phase change induced by jth slice. A wave component of a high-energy electron traveling in one dimension in the z-direction over a potential well will accumulate phase relative to a wave component that does not pass over the same well. This is really the Wentzel–Kramers–Brillouin (WKB) approximation in quantum mechanics, which breaks down when there are sharp discontinuities in the potential or when the well is deep. If our infinitesimally thin, strong object physically existed, this approximation would indeed break down. The exact solution would require matching the amplitude and the gradient of the amplitude of the wave function at the entrance and exit surfaces, leading to backscattered waves, and we can no longer approximate that the k-vector of the high-energy electron linearly increases as function of the potential well depth.

It is important to emphasize that ptychography does *not* solve for the exit wave when the object is thick; in other words, a ptychograph will never provide the same modulus and phase as holography or through-focal series reconstruction, because it explores a different (indeed, more extensive) volume of reciprocal space. With reference to Fig. 14, a conventional image integrates the complex function lying over a circular region of the Ewald sphere, which is accommodated by the diaphragm of the imaging-forming lens. This integral is only identical to the ptychographic reconstruction when the object is so thin that all reciprocal features lying in the u–v plane are stretched out in the w-direction so much so that the integral over S is identical to the integral C over the Ewald sphere.

Once we allow propagation effects, understanding what is going to happen to the ptychograph is difficult. The theory of this has not been fully developed,

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 49
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

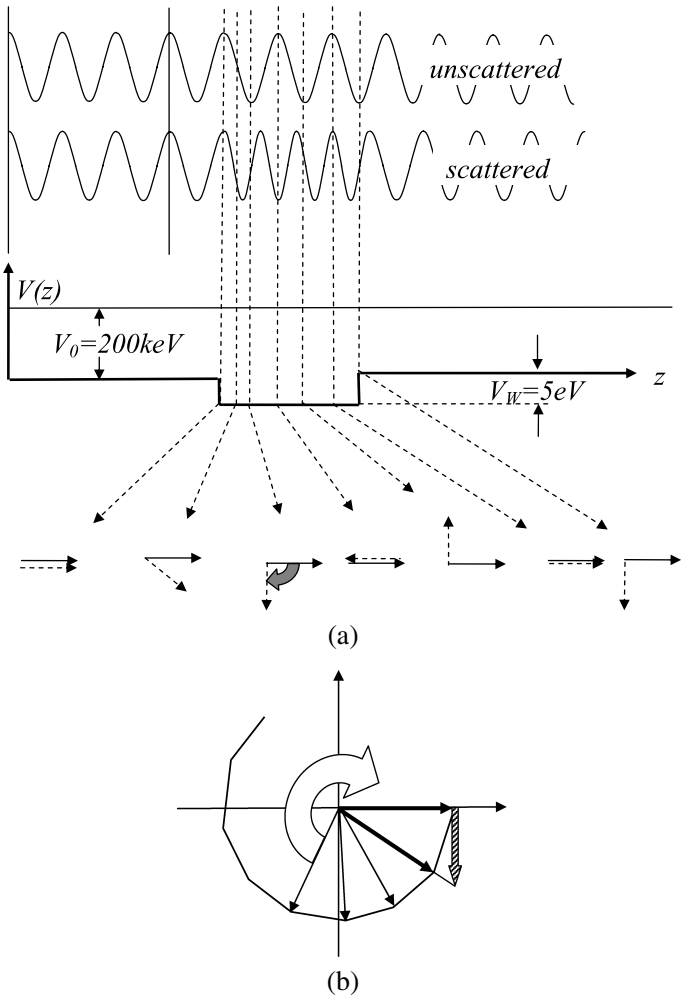


FIGURE 13. (a) Illustrating the phase progression of the unscattered and scattered wave components of an electron as it passes over a 1D potential well (i.e., no lateral propagation in x or y). When the electron is over the potential well, its kinetic energy is increased and its wavelength is shortened. Phasors in the complex plane are shown adjusted so that the unscattered beam has zero phase. In this highly exaggerated case, the scattered wave accumulates a phase change of $5\pi/2$ relative to the unscattered beam (very strong scattering). (b) The weak phase approximation that the expansion of $e^{i\phi}$ is $1 - i\phi$ (hashed pointer). With the propagation switched off, a multislice calculation accumulates phase as it progresses through a potential well, each layer introducing a weak phase perturbation. When allowed to propagate laterally through 3D space, there is no such simple relationship.

aiep v.2005/11/15 Prn:6/12/2007; 13:13
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

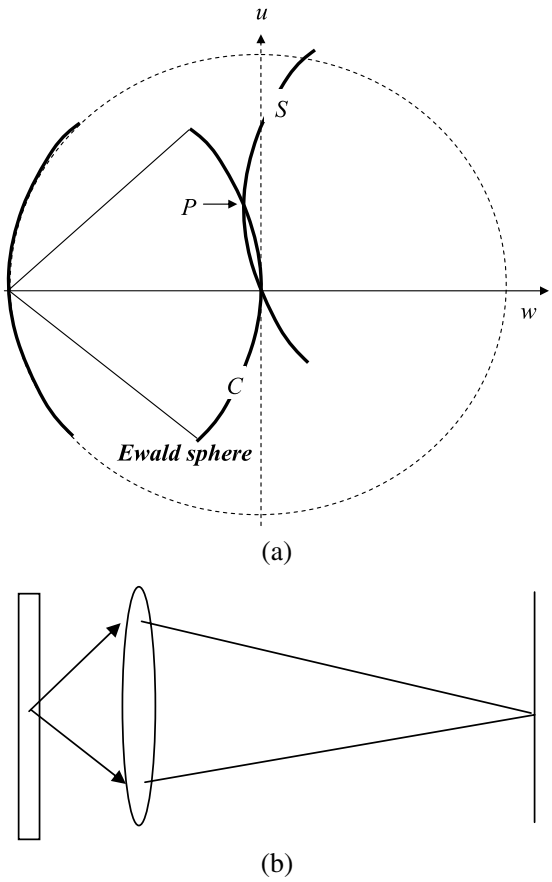


FIGURE 14. The integral over Ewald sphere, C, forms one image pixel in a conventional microscope. The lens acts to integrate the scattered waves falling within the lens diaphragm (lower diagram). Ptychography integrates over surface S (see Fig. 12). The various ptychographic reconstructions effectively deconvolve the surface C to derive a single amplitude at point P, and hence an estimate of S. Π_{lc} (Section IV.A) relies on S embracing only one or two reflections at any one position P. Π_p (Section IV.B) separates pairs of beams by Fourier transforming with respect to the probe position. Π_π (Section IV.C) performs a division in real space. Π_W (Section IV.D) deconvolves C via the Fourier transform of its intensity, H.

although Plamann and Rodenburg (1998) have compared model ptychographical calculations with second-order dynamical scattering theory. These show that even for strongly crystalline objects ptychography is in fact not as sensitive to thickness and dynamical effects as we might at first suppose. At 100 keV in silicon, the phase of the Π_{lc} ptychograph (see Section IV.A)

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 51
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

is a good representation of the real exit wave under plane wave illumination (when that exit wave is back-propagated to the mid plane of the object) up to a thickness of 10 nm (see Fig. 18). The Π_p ptychograph (see Section IV.B) is much more insensitive to thickness and multiple scattering. In similar calculations using crystalline silicon, a good representation of the projected potential is achieved up to a specimen thickness of 25 nm (the resulting images shown in Fig. 26 are explained in more detail in Section IV.B). This rather astonishing result (25 nm is certainly "thick" by the standards of conventional electron imaging) arises from the fact that Π_p uses only data from a plane in reciprocal space where all excitation errors are zero; under kinematical scattering, this yields a perfect projection of the atomic potential.

Whether a more comprehensive processing algorithm can be evolved to account more fully for dynamical effects remains an unanswered question; clearly the aim would be to recover the actual 3D potential, say via a combination of ptychography and tomography. Even in the crystalline Π_{lc} ptychograph (see Section IV.A), where only a few diffraction reflections are processed, there is wide choice of phase-closure paths, each of which will yield a different estimate of Π_{lc} when the object is thick and dynamically scattering, as observed experimentally by Nellist and Rodenburg (1998). In the case of the Wigner distribution deconvolution (WDDC) method (Section IV.D), a phase difference is obtained for every pair of pixels in the diffraction plane, each of which has explored a slightly different range of incident and scattered beams. There remains a great amount of possible research to determine exactly how to process this fantastic quantity of data most effectively.

Finally, it is noteworthy that, at least in the context of ptychography, it is not clear that the weak phase approximation is identical to the kinematical approximation. The former is usually thought of as a constraint on the exit wave in real space, the latter as a statement of scattering strength in reciprocal space. Consider a Gabor hologram. The main requirement for this method to work is that the object is "substantially" transparent so that the undiffracted beam can act as a reference wave for the scattered beams (i.e., the kinematical approximation holds). However, the object itself can be strong in the sense that, at any particular point it can be much stronger than a weak phase object, provided the regions over which it is strong are small compared with the areas of transparency.

Even in the conventional image, say of amorphous materials, these definitions are not orthogonal. An object can be thick and strong in the sense that the exit wave field is considerably changed in phase relative to an unscattered beam. However, because the average phase change of the exit wave over any area of the object is roughly constant, then the transmitted beam encoding this average (and possibly large) phase change acts itself like an unscattered beam. Only perturbations in the object potential (as seen in projection) deviating

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 52
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

from the average inner potential cause contrast. In this way, even quite thick amorphous specimens can appear to satisfy the weak phase approximation. Note that in order to obtain the crystalline projection ptychographs Π_p shown if Fig. 28, an average phase change was subtracted from the raw ptychograph to account for this accumulated average phase change. Furthermore, in the case of ptychography we might suppose that if the area of illumination is large, then the kinematical approximation will still hold, even though particular small features in the object are strong. This may account for the reason that the projection images discussed in Section IV.B, which are of objects consisting of totally opaque grid bars, though surrounded by large areas of empty space, produce ptychographs that are a good representation of the object.

IV. The Variants: Data, Data Processing, and Experimental Results

This section briefly describes the relationship between the main variants of ptychography and presents some examples of experimental results in the fields of visible light optical, hard X-ray, and high-energy electron radiation. As far as I am aware, there have been no experimental demonstrations of the "classical" forms of ptychography (as defined above) using subatomic wavelengths (high-energy electrons or X-rays) and employing only the minimal two diffraction patterns or illumination functions as originally envisaged by Hoppe. However, there have been many demonstrations of methods that use more than one diffraction pattern obtained from a localized area of an object and which exploit both the convolution theorem in reciprocal space and the Fourier shift theorem in order to reconstruct a ptychograph. Classical ptychography is difficult to put into practice for several reasons. For example, it would be hard to imagine constructing a real-space aperture of the exact size of an atomic-scale unit cell and moving it a distance that is a fraction of that unit cell size while recording diffraction patterns – one of the geometries described by Hoppe in the original papers. Even in the STEM configuration, knowing the exact displacement of the probe relative to a small unit cell is exceedingly demanding. Successful implementations have therefore involved much more extensive data in order obtain redundancy in the measurements acquired.

The most complete 4D data set is $I(\boldsymbol{u}, \boldsymbol{R})$, where the sampling in both $\boldsymbol{u}$ and $\boldsymbol{R}$ is "dense" is the sense that each probe movement is small and the detector has many pixels in it, as in a typical charged-couple device (CCD) camera. This involves, for example, $N_u \times N_v \times N_X \times N_Y = N_{\mathrm{TOTAL}}$ data pixels over u, v and $\boldsymbol{R} = (X, Y)$. Classical ptychography involves two diffraction

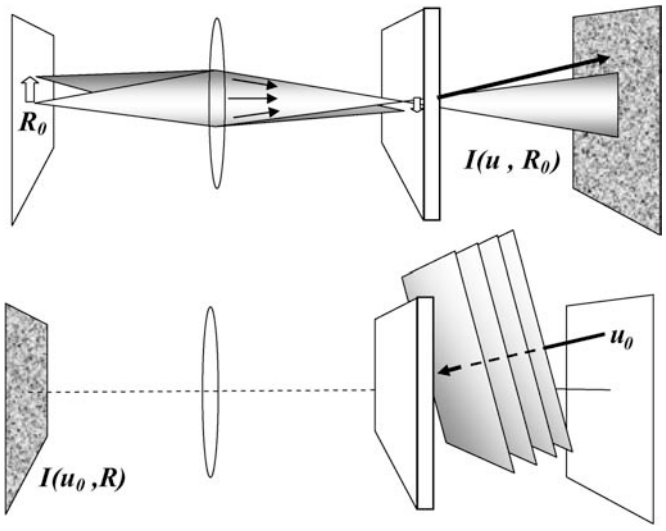


FIGURE 15. Equivalent ways of constructing the entire $I(\boldsymbol{u}, \boldsymbol{R})$ data set. In the STEM configuration (top), we collect a diffraction pattern for every probe position $\boldsymbol{R}_0$. In the conventional imaging configuration (bottom), we collect an image for every value of $\boldsymbol{u}_0$, the tilt in the plane wave illumination. The principle of reciprocity means that the intensity recorded at $\boldsymbol{u}_0$ when the source is in the image plane at $\boldsymbol{R}_0$ is the same as when a source of the same intensity is at $\boldsymbol{u}_0$ and we measure instead the intensity at $\boldsymbol{R}_0$. In the absence of incoherence, inelastic scattering or finite detector pixel size, the two (4D) data sets are identical.

patterns, the necessary minimum being $2(N_u \times N_v)$ measurements, where the product $N_u \times N_v = 2N_g - 1$, where N_g is the number of diffraction orders that we wish to phase.

This entire data set can also be considered via the imaging capabilities of our illumination function, according to Fig. 15, which is drawn in the STEM configuration. Everything discussed so far involves moving the illumination (probe) function, which is equivalent to moving the source of radiation so that its image in the object plane moves laterally. For each such probe position we collect a diffraction pattern in the far field. The usual imaging configuration in STEM is to scan the probe rapidly to all such object-plane positions while collecting scattered parts of the intensity in the far field. To form a STEM bright-field image we place a small aperture on the optic axis and display the intensity measured there synchronously with the position of the probe. By reciprocity (time reversal of the traveling waves), this is equivalent to illuminating the specimen from the position of the detector and recording the image in the plane of the source of radiation. This latter plane is simply the

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 54
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

conventional image data: it is a map of $I(\boldsymbol{u}_0, \boldsymbol{R})$ for just one value of $\boldsymbol{u}$ over all values of $\boldsymbol{R}$. We then have $N_u \times N_v = 1$, so that $N_{\mathrm{TOTAL}} = 1 \times N_X \times N_Y$.

It would be rather absurd to encompass all of conventional imaging theory under the auspices of ptychography. However, it is true that there is a continuous spectrum of subsections of the complete $I(\boldsymbol{u}, \boldsymbol{R})$, many of which embrace classical imaging concepts. For example, $I(\boldsymbol{u}, \boldsymbol{R}_0)$ is the far-field diffraction plane mapped over all $\boldsymbol{u}$ for a single probe position; if $a(\boldsymbol{r})$ represents a highly defocused lens-generated illumination function and $q(\boldsymbol{r})$ is substantially transparent or mostly empty, then these data are a Gabor hologram (Gabor, 1948, 1949). Is Gabor holography a form of ptychography? The answer has to be no; that is partly why we adopted the definition in Section II.C, which stipulated that ptychography requires an illumination or object shift to unlock the phase ambiguity. In this context, it is interesting to note that Lin and Cowley (1986) pointed out that a lateral shift in the illumination function in Gabor holography would resolve the defocus ambiguity that occurs in that method, because the direction of shift of the shadow-image hologram would be in opposite senses for the two possible reconstructions. This can, obliquely, be thought of a ptychographical resolution of the complex conjugate phase ambiguity. We will not ponder any further exactly where to draw the boundary between ptychography and conventional holography and imaging; instead, the methods and experimental data presented in this section are arranged, approximately, by their decreasing relevance to the classical definition of ptychography developed in Section II.

A. The Line Scan Subset

The first short-wavelength experimental proof-of-principle that came closest to classical STEM ptychography was by Nellist *et al.* (1995). Rather than processing just two diffraction patterns, the probe was scanned along the length of a line, a diffraction pattern being collected at a number of positions (about 64), each spaced evenly by a distance less than the size of the unit cell. The dimension of the data cube recorded was therefore $N_{\mathrm{TOTAL}} = 2(N_g - 1) \times N_X \times 1$. We will call this type of ptychograph Π_{lc}, being calculated from a line scan set of crystalline diffraction patterns. The sample was the old favorite for TEM resolution tests – crystalline silicon orientated onto the [110] zone axis. By taking a Fourier transform of data collected in the diffraction pattern at the central points between diffraction orders as a function of the probe position, it was possible to obtain the relative phase of 13 diffraction orders (i.e., $N_g = 13$). Putting the complex values so obtained (the magnitudes measured from the non-overlapped areas of the diffracted disks) a near-perfect image of the so-called dumbbells (separation 0.136 nm) was obtained in both modulus and phase (shown in Fig. 16).

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 55
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

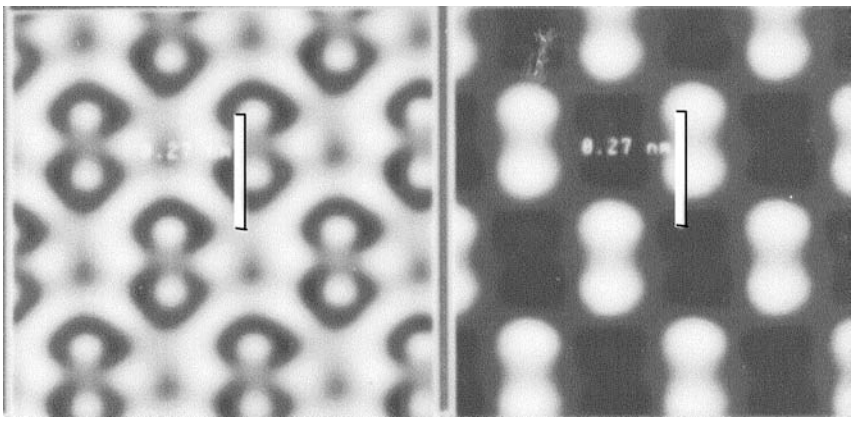


FIGURE 16. Line scan ptychograhs Π_{lc} reconstructed from 13 diffraction reflections from crystalline silicon oriented on the [110] zone axis. The left image is the modulus and the right image is the phase of the reconstruction. The data were obtained with incident beam energy of 100 keV in a VG HB501 STEM with intrinsic lens resolution of 0.3–0.4 nm. The scale bar is 0.27 nm in length, the separation of dumbbells being 0.136 nm, although they appear slightly more widely spaced for reasons discussed by Nellist and Rodenburg (1994, 1998). This image clearly demonstrates the resolution enhancement possible with ptychography, even when the coherence envelope of the probe-forming lens is relatively small.

The purpose of the experiment was to demonstrate that ptychography is not compromised by the overall coherence envelope of the microscope (arising from a combination of lens and accelerating voltage instability). The model HB501 microscope (Vacuum Generators, Hastings, UK) had a nominal point-to-point resolution of approximately 0.42 nm. Its coherence envelope was equivalent to a resolution of ∼0.34 nm. In other words, by achieving a coherent resolution of 0.136 nm, all the conventional resolution limitations of the microscope had been surpassed.

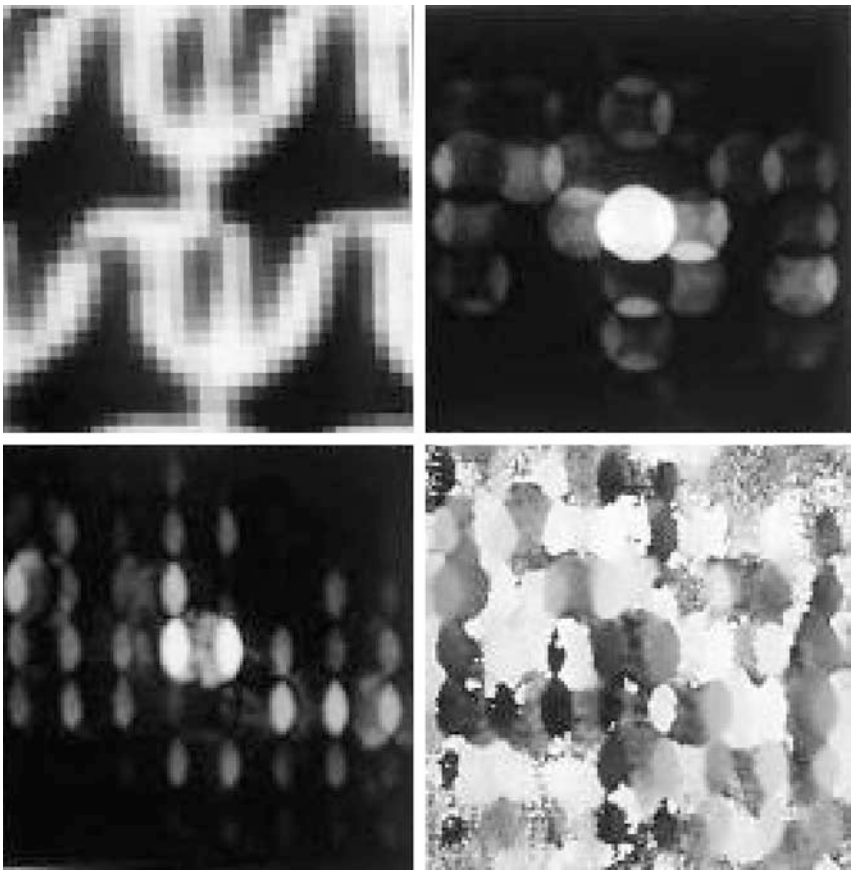
Since then, 300-keV and aberration-corrected STEMs have been routinely able to achieve the same resolution via the incoherent annular dark field (ADF) image, but it must be remembered that the incoherent image has intrinsically higher resolution than the coherent image (the impulse response function is much sharper because the incoherent probe is the square of the amplitude of the coherent response function). Obtaining a complex image in modulus and phase is also not particularly spectacular; similar data can be obtained via a through-focal series reconstruction or holography (e.g., Coene *et al.*, 1992; Lichte, 1986), the latter obtaining a representation of the exit wave, $\psi_e(\boldsymbol{R})$. However, both these competing techniques require a very good lens and very good coherence. Note that the incoherent nature of ADF imaging derives from the size of the detector, giving an image which,

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 56
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

via reciprocity, can be thought of as a conventional image for which the illumination is wide-angled and hence incoherent. The probe-forming optics must still be coherent in the sense that the source must be small and bright, the back focal plane of the lens must be illuminated coherently, and the defocus lens instabilities must be negligible; in other words, all the usual limitations of electron optics apply to the ADF image. On the contrary, the silicon ptychography experiment demonstrated that partial coherence does not present a fundamental resolution limit.

A visible light implementation of this same scanned probe version of ptychography was demonstrated by McCallum and Rodenburg (1993b). A laser printer was used to print a periodic array of the Greek letter ψ (see Fig. 17). This was then reduced in size onto a transparent negative by conventional photography. The resulting negative image (the ψs were transparent) was placed into an optical bench that emulated the STEM configuration. The cube of data so formed (a diffraction pattern for every probe position) was then Fourier transformed with respect to the probe coordinate, X. In the resulting cube of data there is a plane – plotted as a function of u, v at one particular Fourier component of the probe movement coordinate – which exactly picks out all the relative ptychographic phase relationships (bottom right, Fig. 17(a)). Why is this? Think of it as follows. As the probe is moved, the intensity in the region of overlap between the adjacent circular diffraction orders varies periodically as the vector shown as Z_2 in Fig. 4 cycles around in the complex plane. The intensity can be plotted at this particular point as a function of the probe position, which gives a periodic sinusoid. The same method was used to construct the electron ptychograph of silicon (Fig. 16). Each of these sinusoids has the ptychographic phase difference between the corresponding diffraction orders, and so in this method all such crystalline phase differences are captured in a single picture.

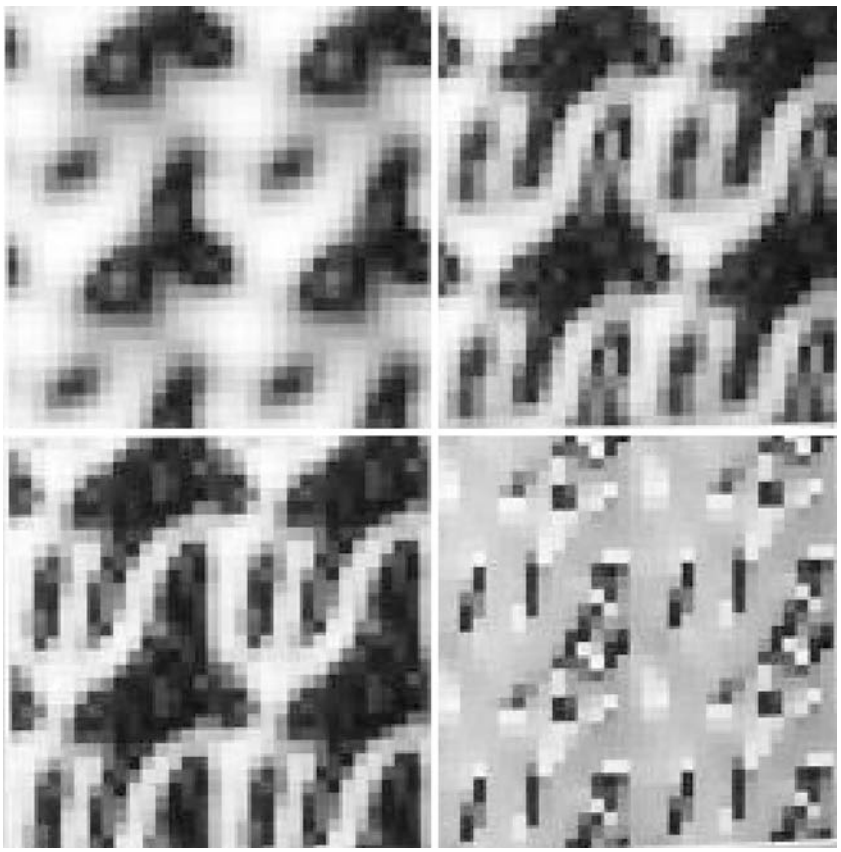
Extracting these phases and putting together the resulting crystalline Fourier components yields a high-resolution image of the object (shown in Fig. 17(b)). All the high-angle beams rely on obtaining their phase from beams that lie closer to the undiffracted beam; recall that in this configuration we can only measure phase *differences* between *adjacent* beams. The process of successively phasing higher-order beams is called *stepping out*. Errors would be expected to accumulate as we step out to higher and higher-order reflections (successively larger numbers of beams are used in Fig. 17(b)). However, there are a correspondingly larger number of permutations of stepping routes that can be taken as the number of reflections is increased, and so it is possible to adopt strategies that minimize the accumulation of error. The important result in the present case was to demonstrate that 81 reflections could be phased in this way, giving a good representation of the object.



(a)

FIGURE 17. (a) Visible light optical analog of crystalline line scan ptychography. Top left: Conventional image of the object acquired using a lens of large numerical aperture. Top right: Diffracted intensity using a probe-forming lens with a very small aperture. No diffraction peaks interfere at the central point of the diffraction pattern, so a lens of this numerical aperture would not be able to see any structure in the conventional image. Bottom left (modulus) and right (phase) of the plane of the data cube Fourier transformed with respect to the probe position (see text). Only regions of strong ptychographical interference are picked out in this plane. The right-hand image is a map of all the relevant ptychographical phases (the grayscale being proportional to phase plotted from 0 to 2π). The phase of all the beams can be established by stepping out through each diffraction order using the ptychographical phase differences between each pair of beams. (b) Ptychographic reconstruction of Π_{lc} for the data shown in (a). As more of the diffracted beams are factored into the image, the resolution increases. The top left (modulus) of Π_{lc} has used the nearest 5×5 diffracted reflections seen in Fig. 16(a), while the top right image (modulus) has used 7×7 reflections. The two lower images shown in modulus (left) and phase have used 9×9 reflections (a total of 81 ptychographic phases); the phase is flat, as expected, except where the modulus is small and so phase is ill determined.

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 58
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV



(b)

FIGURE 17. (*continued*)

The work on silicon was extended in two later papers that contain some interesting further results on the question of the effects of dynamical scattering on Π_{lc} (Plamann and Rodenburg, 1998) and on phase closure and other experimental issues (Nellist and Rodenburg, 1998). Multislice calculations were performed to compare the experimental images with theory and to estimate the thickness of specimen at which Π_{lc} ceases to relate meaningfully to the underlying atomic potential. Images calculated assuming Eq. (4), that is, assuming the 2D transmission function approximation Eq. (3), were compared with the multislice calculation of the actual exit wave. In the 2D theory we must assume the illumination is in some fixed plane perpendicular to the z direction. Surprisingly, even in rather thin specimens, the illumination

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 59
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

defocus does slightly affect the calculated image. With the small range of k-vectors in the incident radiation we would expect changes in the probe as a function of z to be small. The radius of the aperture used in these experiments in the back focal plane of the lens was only a little larger than half the angle of the diffraction pattern g-vectors. This means it was too small to form conventional bright-field image fringes at the periodicity of the unit cell. Even a high-resolution conventional image of silicon (where the scattering angles processed by the lens are much larger) produces effectively a projection of the object; defocusing the lens within the depth of the object produces negligible change in the image. It must be remembered, however (Section III.B) that a ptychograph integrates a rather larger volume of reciprocal space relative to the conventional image (Fig. 14). The latter is derived as a result of the lens integrating the amplitude over a central disk lying on the Ewald sphere.

In fact, the dynamical calculations showed that Π_{lc} is remarkably similar to the real (strongly scattered) exit wave function in silicon provided it is back-propagated to a plane in the center of the object. Figure 18 shows the modulus phase of the calculated exit wave, the modulus and phase of that same wave back-propagated to the mid-plane of the object, and Π_{lc}. Π_{lc} is independent of illumination propagation if the relative ptychographical phases are measured

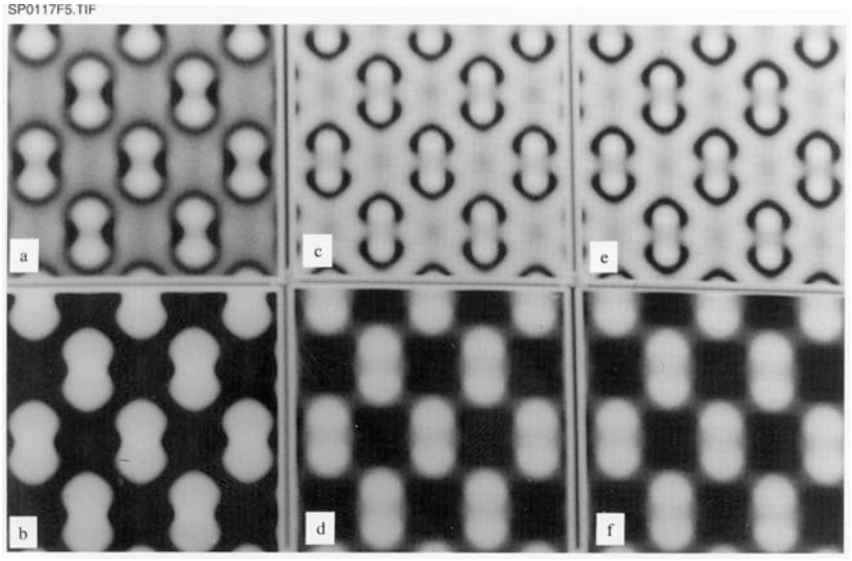


FIGURE 18. Images a and b are the modulus and phase, respectively, of the exit wave calculated for [110] silicon with incident energy of 100 keV and sample thickness of 7 nm. This wave is then back-propagated by 3.5 nm to the plane in the middle of the object (images c and d). The Π_{lc} ptychograph (images e and f) accurately corresponds to this back-propagated wave.

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 60
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

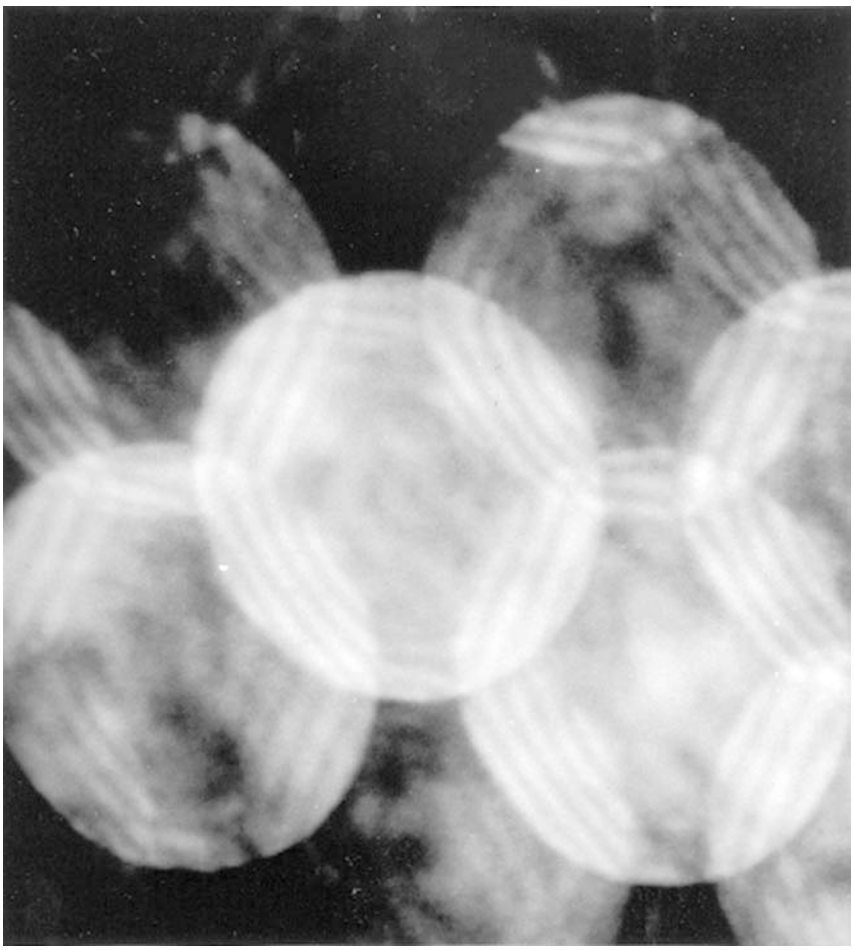


FIGURE 19. Interfering disks in the STEM electron diffraction pattern. The specimen is silicon oriented onto the [110] zone axis. The fringes result from the parabolic defocus term over the aperture function (this pattern was recorded with large defocus in the STEM probe). When recorded in intensity, the defocus phases from adjacent beams contrive to form straight fringes. When the aperture is very large, the fringes cross one another and hence build up the shadow image of the periodic object, albeit dominated by coherent effects like in a Gabor hologram.

from exactly the midpoints of the diffracted disk overlaps (provided the probe-forming optics are symmetric) as first discussed by Spence and Cowley (1978). Figure 19 shows raw data of the diffracted disks from silicon interfering with one another when substantial defocus has been applied to the probe-forming lens: an image that graphically illustrates ptychographic interference.

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 61
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

Nellist and Rodenburg (1998) measured phase closure between the various routes through these diffraction orders (Fig. 20). The very large phase-closure errors induced by taking triangular routes arise primarily from astigmatism present in the lens. A parallelogram route allows such asymmetries to cancel one another and so leads to much smaller phase-closure error. However, even in the absence of such errors, discrepancies would be expected to arise from dynamical scattering. Plamann and Rodenburg (1998) calculated these for the same scattering geometry (Fig. 21). It is noteworthy that for Π_{lc} the phase-closure errors are relatively small for thicknesses of silicon less than 10 nm (at 100-keV incident energy) and that low-order reflections (corresponding to low-resolution information) are reasonably accurate up to 30-nm thickness. (The Bragg–Brentano data set, detailed in the next section, yields much more reliable reconstructions.) In short, even in a fully dynamical model calculation of a crystalline object where all the usual channeling and propagation effects are strong, the 2D ptychographic approximation does not begin to break down until quite substantial specimen thickness. Spence (1978) performed some similar calculations years ago and reached a similar conclusion: that the relative phase between diffracted beams at their midpoint of STEM disk interference is relative insensitive to dynamical scattering effects. He also pointed out that as a function of thickness, a phase jump occurs at a thickness equal to half the extinction distance, in the same way as a contrast reversal occurs in the bright-field image.

In conclusion, the results of this first experimental demonstration of electron ptychography and the associated dynamical calculations showed three important things. First, although the Π_{lc} ptychographic reconstruction does not attempt to make any account for the breakdown of 3D geometric and propagation effects, the resulting image is still a very good estimate of object structure seen in projection. In fact, as will be seen in Section IV.B, it is possible to measure a different subset of $I(\boldsymbol{u}, \boldsymbol{R})$ that explores rather more of the 3D information extant in the entire 4D data set. Proper consideration of this may well suggest ways of further accounting for dynamical effects in order to calculate ptychographs that are even more representative of the scattering potential. Second, perhaps the most significant result from this proof-of-principle is that the constraints of finite coherence (and hence the so-called information limit) that occur in lens imaging and are generally acknowledged as the final resolution limit in electron microscopy (even when undertaken by aberration-corrected lenses) can be overcome by several factors using ptychography. Third, although this proof-of-principle was applied to a periodic and very simple specimen, the same benefits in relation to partial coherence and dynamical scattering apply to many of the noncrystalline variants of ptychography. All of these factors bode rather well for the relative imaging strengths of ptychography.

aiep y.2005/11/15 Prn:6/12/2007; 13:13
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

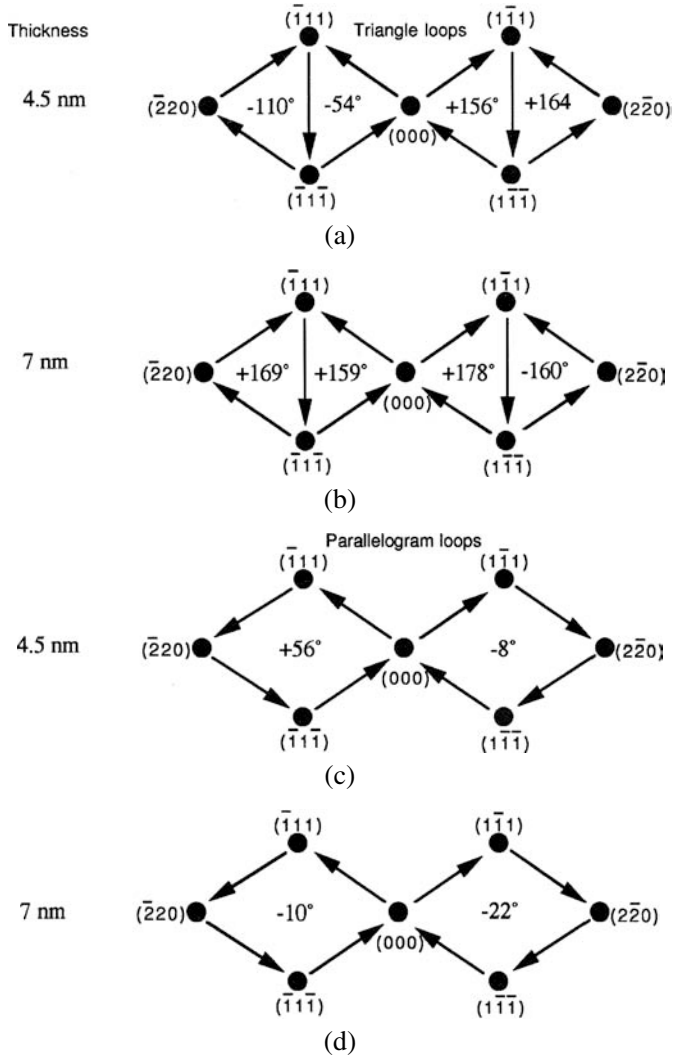


FIGURE 20. Phase closure loops in the experimental silicon Π_{lc}. Errors in the triangular routes are large, mostly because of astigmatism in the probe-forming lens. Parallelogram routes use pairs of symmetric interferences (relative to the astigmatism), and hence the phase error cancel. Even perfect measurements will give phase errors around closure loops because of thickness and dynamical scattering effects.

aiep v.2005/11/15 Prn:6/12/2007; 13:13
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

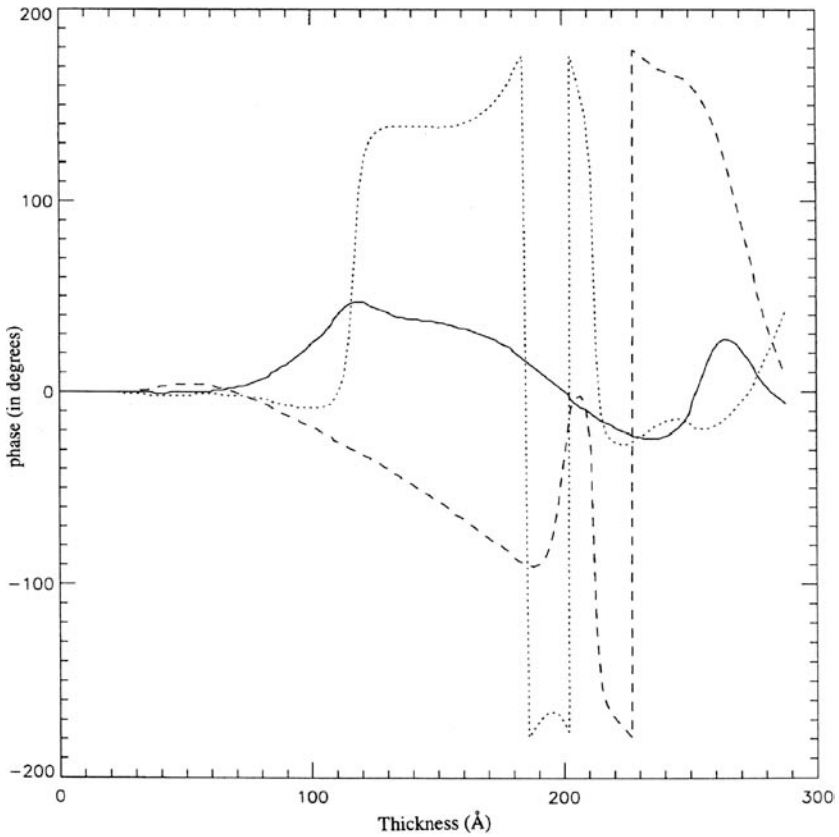


FIGURE 21. Calculation of phase closure as a function of specimen thickness in Ångstroms (specimen is silicon on [110] zone axis, accelerating voltage 100 keV) in the presence of dynamical effects. Solid line: stepping route (000) → (1$\bar{1}$1) → (002) → (000). Dashed line: stepping route (1$\bar{1}$1) → (002) → (1$\bar{1}$3) → (1$\bar{1}$1). Dotted line: stepping route (002) → (1$\bar{1}$3) → (004) → (002).

B. The Bragg–Brentano Subset: Projection Achromatic Imaging

The line scan data described previously can be extended to form a type of ptychograph that we will term Π_p and which has truly astonishing properties. Π_p can account for at least some of the 3D geometric effects discussed in Section III.C. The method produces a near-perfect projection image of the object (which need not be crystalline) at a resolution unlimited by the information limit or any symmetric aberrations in the lens used to focus the illumination function (Rodenburg *et al.*, 1993; Plamann and Rodenburg, 1994). It is also immune to lens-current instability and chromatic spread in the

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 64
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

beam. It is also relatively immune to dynamical scattering and yields a good representation of the object function even for strong, thick crystals.

The experimental procedure is to scan a 2D array of closely separated real-space illumination positions and then select a particular Fourier component of the 2D scan at associated detector pixels in the diffraction pattern, each of which corresponds to the Bragg–Brentano diffraction condition for that same real-space (probe-position) periodicity. Some lengthy explanation is required, but first note that as far as the amount of measured data that are required, this method uses the entire 4D data set: $N_{\mathrm{TOTAL}} = N_u \times N_v \times N_X \times N_Y$, where all the sampling is dense. However, unlike the G- and H-sets that are discussed below, we need not undertake entire Fourier transforms (with respect to the illumination position) for every single detector pixel. Instead, we only need a single Fourier component picked at each detector pixel.

Let the Fourier transform of the 4D intensity data set $I(\boldsymbol{u}, \boldsymbol{R})$ with respect to the 2D $\boldsymbol{R}$ coordinate be

$$G(\boldsymbol{u}, \boldsymbol{U}) = \int I(\boldsymbol{u}, \boldsymbol{R}) e^{i2\pi \boldsymbol{R}\cdot\boldsymbol{U}}\, d\boldsymbol{R}. \tag{71}$$

The Bragg–Brentano data set then corresponds to a 2D plane in this 4D data cube given by

$$B(\boldsymbol{U}) = G\left(\frac{\boldsymbol{U}}{2}, \boldsymbol{U}\right). \tag{72}$$

We can think of this as follows. To form $B(\boldsymbol{U})$, data are collected in the diffraction plane, each pixel having coordinates given by the 2D vector $\boldsymbol{u}$. We collect diffraction patterns for all the possible (densely sampled) illumination (probe) positions $\boldsymbol{R}$. Consider a pixel in the detector plane lying at $\boldsymbol{u}_0$. We form a 2D image as a function of $\boldsymbol{R}$ by observing the intensity that arrives at this pixel. This is just a conventional image we could call, say, $f(\boldsymbol{R})$. We take a Fourier transform of this to create a diffractogram, $F(\boldsymbol{U})$. We then extract from $F(\boldsymbol{U})$ the (complex) value of the single pixel lying at $\boldsymbol{U}_0 = 2\boldsymbol{u}_0$ and put this into the array $B(\boldsymbol{U})$ at the point $\boldsymbol{U}_0$. We do this for all $\boldsymbol{u} = \boldsymbol{u}_0$ and $\boldsymbol{U} = 2\boldsymbol{u}_0$ and thus construct the 2D reciprocal-space function $B(\boldsymbol{U})$. The back Fourier transform of $B(\boldsymbol{U})$ is the ptychograph $\Pi_p(\boldsymbol{R})$.

The very special properties of $B(\boldsymbol{U})$ depend on the specimen being reasonably weakly scattering. (I say "reasonably" because it is still very much an open question exactly how robust this data set is to dynamical scattering.) In the domain of the limited calculations that were performed by Plamann and Rodenburg (1998), and the visible light proof-of-principle performed by the same authors (Plamann and Rodenburg, 1994), it would appear as if Π_p is

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 65
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

indeed a very good estimate of the object, even in the presence of quite strong scattering and for thick specimens.

Following the derivation in Rodenburg *et al.* (1993), we first assume the object is a 2D transmission weak phase grating of the form

$$q(\boldsymbol{r}) = 1 - ic\vartheta(\boldsymbol{r}), \tag{73a}$$

where c is a constant and

$$c\big|\vartheta(\boldsymbol{r})\big| \ll 1, \tag{73b}$$

then the Fourier transform (diffraction pattern) of $q(\boldsymbol{r})$ is given by

$$Q(\boldsymbol{u}) = \Im q(\boldsymbol{r}) = \delta(\boldsymbol{u}) - icF(\boldsymbol{u}) \tag{74}$$

where

$$F(\boldsymbol{u}) = \Im\vartheta(\boldsymbol{u}). \tag{75}$$

Now according to the fundamental ptychographical Eq. (41)

$$I(\boldsymbol{u}, \boldsymbol{R}) = \big|\big(A(\boldsymbol{u})e^{i2\pi R\cdot u}\big) \oplus Q(\boldsymbol{u})\big|^2, \tag{76}$$

which can be expanded as

$$I(\boldsymbol{u}, \boldsymbol{R}) = \iint A(\boldsymbol{u}_a)Q(\boldsymbol{u}-\boldsymbol{u}_a)A^*(\boldsymbol{u}_b)Q^*(\boldsymbol{u}-\boldsymbol{u}_b)e^{i2\pi\boldsymbol{R}\cdot(\boldsymbol{u}_a-\boldsymbol{u}_b)}\,d\boldsymbol{u}_a\,d\boldsymbol{u}_b, \tag{77}$$

where $\boldsymbol{u}_a$ and $\boldsymbol{u}_a$ are dummy variables. Arithmetic manipulations of $I(\boldsymbol{u}, \boldsymbol{R})$ are key in developing the theory of the entire ptychographical data set. Therefore the required steps in this first example are spelled out in some detail. We use the technique of collapsing integrals via the Dirac delta function; this can easily be extended to all the results that are simply stated in later sections. In order to form $G(\boldsymbol{u}, \boldsymbol{U})$, and hence $B(\boldsymbol{U})$, we must substitute Eq. (77) into Eq. (71) to derive

$$G(\boldsymbol{u}, \boldsymbol{U}) = \iiint A(\boldsymbol{u}_a)Q(\boldsymbol{u}-\boldsymbol{u}_a)A^*(\boldsymbol{u}_b)Q^*(\boldsymbol{u}-\boldsymbol{u}_b)e^{i2\pi\boldsymbol{R}\cdot(\boldsymbol{u}_a-\boldsymbol{u}_b+\boldsymbol{U})}d\boldsymbol{u}_a\,d\boldsymbol{u}_b\,d\boldsymbol{R}. \tag{78}$$

Note that the integral of $e^{i2\pi x\cdot u}$ over infinite limits in u is zero unless x is zero. When x is zero, the integral is infinite. Between finite limits – which will always occur in a real experiment because the span of the probe position $\boldsymbol{R}$ and detector range $\boldsymbol{u}$ is finite – the resulting integral as a function of x is

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 66
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

of the form of $\mathrm{sinc}(x)$ in one dimension and has a value of unity. This means that in the limit of large limits of integration, $\mathrm{sinc}(x)$ has the same properties of the Dirac δ function (Eq. (11)). A and Q in the above have no dependence on $\boldsymbol{R}$, so we can integrate over it to obtain

$$G(\boldsymbol{u},\boldsymbol{U}) = \iint A(\boldsymbol{u}_a)Q(\boldsymbol{u}-\boldsymbol{u}_a)A^*(\boldsymbol{u}_b)Q^*(\boldsymbol{u}-\boldsymbol{u}_b)\delta(\boldsymbol{u}_a-\boldsymbol{u}_b+\boldsymbol{U})\,d\boldsymbol{u}_a\,d\boldsymbol{u}_b. \tag{79}$$

We can next choose to integrate over either $\boldsymbol{u}_a$ or $\boldsymbol{u}_b$. The choice is not essential (but it can reduce the need for subsequent variable substitution). By the time coherence functions and 3D scattering considerations are introduced, such integrals can be over many dimensions (e.g., up to 20 in Rodenburg and Bates, 1992). Integrating here over $\boldsymbol{u}_b$, the delta function only has value at $\boldsymbol{u}_b = \boldsymbol{u}_a + \boldsymbol{U}$, so

$$G(\boldsymbol{u},\boldsymbol{U}) = \int A(\boldsymbol{u}_a)Q(\boldsymbol{u}-\boldsymbol{u}_a)A^*(\boldsymbol{u}_a+\boldsymbol{U})Q^*(\boldsymbol{u}-\boldsymbol{u}_a-\boldsymbol{U})\,d\boldsymbol{u}_a. \tag{80}$$

We then put

$$\boldsymbol{u}_c = \boldsymbol{u}-\boldsymbol{u}_a \tag{81}$$

and reorder to obtain

$$G(\boldsymbol{u},\boldsymbol{U}) = \int Q(\boldsymbol{u}_c)Q^*(\boldsymbol{u}_c-\boldsymbol{U})A^*(\boldsymbol{u}-\boldsymbol{u}_c+\boldsymbol{U})A(\boldsymbol{u}-\boldsymbol{u}_c)\,d\boldsymbol{u}_c. \tag{82}$$

From Eq. (74)

$$Q(\boldsymbol{u}_c)Q^*(\boldsymbol{u}_c-\boldsymbol{U}) = \big(\delta(\boldsymbol{u}_c)-icF(\boldsymbol{u}_c)\big)\cdot\big(\delta(\boldsymbol{u}_c-U)-icF(\boldsymbol{u}_c-U)\big)^* \tag{83}$$

so

$$\begin{aligned} G(\boldsymbol{u},\boldsymbol{U}) = {} & \int \delta(\boldsymbol{u}_c)\delta(\boldsymbol{u}_c-\boldsymbol{U})A(\boldsymbol{u}-\boldsymbol{u}_c)A^*(\boldsymbol{u}-\boldsymbol{u}_c+\boldsymbol{U})\,d\boldsymbol{u}_c \\ & - ic\int \delta(\boldsymbol{u}_c-\boldsymbol{U})F(\boldsymbol{u}_c)A(\boldsymbol{u}-\boldsymbol{u}_c)A^*(\boldsymbol{u}-\boldsymbol{u}_c+\boldsymbol{U})\,d\boldsymbol{u}_c \\ & - ic\int \delta(\boldsymbol{u}_c)F(\boldsymbol{u}_c-\boldsymbol{U})^*A(\boldsymbol{u}-\boldsymbol{u}_c)A^*(\boldsymbol{u}-\boldsymbol{u}_c+\boldsymbol{U})\,d\boldsymbol{u}_c+\cdots. \end{aligned} \tag{84}$$

The first integral only has a value when $\boldsymbol{u}_c = 0$ and $\boldsymbol{u}_c = \boldsymbol{U}$. The second integral only has value when $\boldsymbol{u}_c = \boldsymbol{U}$; the third integral only has value when

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 67
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

$\boldsymbol{u}_c = 0$. We ignore the F^2 term because the object is weak, so now

$$G(\boldsymbol{u},\boldsymbol{U}) = \delta(\boldsymbol{U})\big|A(\boldsymbol{U})\big|^2 - icF(\boldsymbol{U})A(\boldsymbol{u}-\boldsymbol{U})A^*(\boldsymbol{u}) - icF^*(-\boldsymbol{U})A(\boldsymbol{u})A^*(\boldsymbol{u}+\boldsymbol{U}) + \cdots. \tag{85}$$

Forming $B(\boldsymbol{U})$ we find

$$B(\boldsymbol{U}) = \delta(\boldsymbol{U})\big|A(\boldsymbol{U})\big|^2 - icF(\boldsymbol{U})A\left(-\frac{\boldsymbol{U}}{2}\right)A^*\left(\frac{\boldsymbol{U}}{2}\right) - icF(-\boldsymbol{U})A\left(\frac{\boldsymbol{U}}{2}\right)A^*\left(\frac{3\boldsymbol{U}}{2}\right). \tag{86a}$$

Now if the aperture function (the aberration function of the lens) is centrosymmetric, then

$$A\left(-\frac{\boldsymbol{U}}{2}\right) = A\left(\frac{\boldsymbol{U}}{2}\right), \tag{86b}$$

and thus whatever the phase of the lens aberrations, the second term in Eq. (??) will just be the intensity of the aperture function. If we further assume that the back focal plane of the probe-forming lens is evenly illuminated with an intensity of radiation $|A|^2$, but has a diaphragm of radius of $|\boldsymbol{U}| = \alpha$, then this second term will have intensity $|A|^2$ provided $|\boldsymbol{U}| < 2\alpha$, otherwise being zero. The third term will also have intensity $|A|^2$, but only if $|\boldsymbol{U}| < 2\alpha/3$; for $|\boldsymbol{U}| > 2\alpha/3$ the third term is zero. In other words, for the high spatial frequency region $2\alpha > |\boldsymbol{U}| > 2\alpha/3$, we can write simply that

$$B(\boldsymbol{U}) = |A|^2\big(\delta(\boldsymbol{U}) - icF(\boldsymbol{U})\big), \tag{87}$$

and the back Fourier transform is defined as the ptychograph

$$\Pi_p(\boldsymbol{r}) \approx |A|^2\big(1 - ic\vartheta(\boldsymbol{r})\big) = |A|^2 q(\boldsymbol{r}), \tag{88}$$

the subscript p designating that $\Pi_p(\boldsymbol{r})$ is a projection image of the object (see below). Amazingly, the transmission function of the object has been reconstructed, scaled by the intensity of the illumination. If fact, we have made two approximations; we have ignored the second-order F^2 term and the lower third of the frequency spectrum of the image. However, both of these issues arise in conventional imaging, but the conventional image is additionally subject to all the usual constraints of the form of the particular transfer function, including its zeros, whereas $\Pi_p(\boldsymbol{r})$ has unity transfer at all spatial frequencies as emphasized by Cowley (2001). Furthermore, its transfer function is *twice* the width of the imaging aperture, and in theory this aperture

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 68
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

can be as large as desired because aberrations and defocus instabilities do not appear to express themselves in $\Pi_p(\boldsymbol{r})$.

That is half the good news. Before considering the (very) bad news, I first explain perhaps an even more interesting property of $\Pi_p(\boldsymbol{r})$. It is called a *projection image* because $\Pi_p(\boldsymbol{r})$ only reconstructs information from a set of perfect Bragg reflections that explore a plane of reciprocal space (not the curved surface of the Ewald sphere) (Fig. 22). By reciprocity, we illuminate an object in a conventional TEM by radiation incident at a particular angle β_0, where $u_0 = \sin\beta_0/\lambda$. If the object is reasonably weak, then the image will have an intensity component arising from the interference of this unscattered wave with waves scattered through the objective aperture. It is well known (especially by electron microscopes salesmen) that the highest-resolution interference fringes in any TEM are obtained by illuminating a crystal tilted off the zone axis into a strongly excited two-beam condition with a g-vector of, say, $\boldsymbol{g}$. If the beam tilt is now put at $-\boldsymbol{g}/2$, then the two beams (transmitted and diffracted) will travel through the objective lens at equal and opposite angles to the optic axis as shown in Fig. 22. The interference fringes obtained in the image plane will have a spatial frequency of $\boldsymbol{U} = \boldsymbol{g}$. However, as far as demonstrating the best-possible fringe resolution of a microscope, this condition is very convenient because the two beams lie

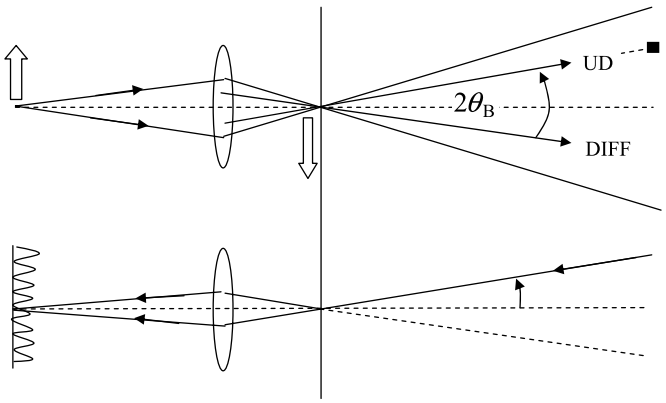


FIGURE 22. The top image shows the STEM configuration; the bottom image shows the equivalent TEM experiment via reciprocity (compare Fig. 15). The Bragg–Brentano data set interferes an undiffracted beam (labeled UD) with a beam that set off from within the lens in the direction of beam DIFF. The latter is diffracted through twice the Bragg angle and hence interferes with UD at the detector pixel (marked with a black box). By reciprocity, this is the same as illuminating the object with a beam tilted by θ_B from the optic axis. The diffracted beam now lies at an equal but opposite point in the back focal plane. In the image plane, we have lattice fringes. Π_π takes a Fourier transform of these fringes (i.e., a Fourier transform with respect to the probe position coordinate $\boldsymbol{R}$), and hence extracts only the information relating to these symmetric Bragg beams.

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 69
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

on the same achromatic ring of the object lens. In other words, variations in lens excitation or energy spread (chromatic spread) in the beam induce identical phase changes in the two beams. The fringe resolution is therefore only predicated on the effective source size in the illuminating beam, which leads to a substantially better resolution (in the fringes) than can be obtained when the illumination is parallel with the optic axis, in which case all the image-forming beams are not on the same achromatic ring as the transmitted beam. Now in the case of $B(\boldsymbol{U})$, all such pairs of achromatic beams are processed simultaneously. This is possible because in the STEM configuration each detector pixel is equivalent (by reciprocity) to each possible illumination beam angle in TEM. The sets of interference fringes that would be measured in TEM are separated from one another by the Fourier transform of $I(\boldsymbol{u}, \boldsymbol{R})$ we undertake with respect to $\boldsymbol{R}$. (Remember that $\boldsymbol{R}$ is the TEM image pixel coordinate.)

Now consider how these symmetric beams correspond to one another in 3D reciprocal space. As the angle of illumination (in the TEM view of this experiment) is increased to $\boldsymbol{u}$, we pick out the periodicity $2\boldsymbol{U}$ in the image plane; that is to say, the periodicity that arises from the interference of the incident beam with the beam that is scattered to $-\boldsymbol{u}$. Therefore for all values of u we are in a symmetric Bragg scattering condition, with $u = \sin\theta_B/\lambda$. This is exactly the same as in the X-ray Bragg–Brentano condition, except there the goniometer of the specimen is rotated at half the angular speed of the detector. With an X-ray diffractometer the principal advantage is that a large specimen diffracts the same Bragg beam into the detector provided both are equidistant from the object, thus optimizing counting statistics. In terms of the description in Section III.C, the negative curvature of the 3D surface S compensates for the curvature of the Ewald sphere, thus solving for a plane in 3D diffraction space (Fig. 23). A plane in reciprocal space corresponds to projection in real space. This is simply a restatement of the fact that $\Pi_p(\boldsymbol{r})$ is independent of defocus.

The projection property of $\Pi_p(\boldsymbol{r})$ was demonstrated by Plamann and Rodenburg (1994) using visible light optics. Two TEM copper grids were mounted separated by a distance of 14 mm (Fig. 24). The two square grids were rotated at an angle to each other and illuminated by a convergent beam of laser light emulating a STEM probe. The lens diameter and working distance were chosen such that the intrinsic resolution of a conventional image was $\sim$120 μm. This was just sufficient to resolve the grid bars (spacing 250 μm, each of width 50 μm) but was large enough so that the illuminating probe spread significantly through the depth of the entire object. In other words, by adjusting the z-displacement of the object, either grid could be brought noticeably into better focus than the other. Data were collected on a CCD camera as a function of all specimen shifts, $-\boldsymbol{R}$, equivalent to shifting the

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 70
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

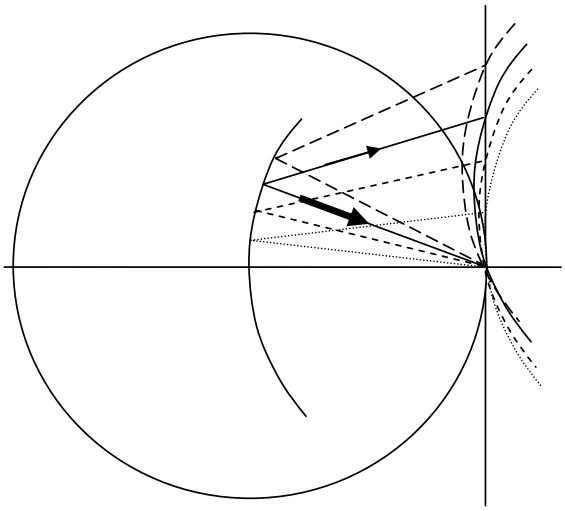


FIGURE 23. In the Bragg–Brentano data set, the incident beam is symmetrically tilted relative to the optic axis with respect to the scattered beam. In this way, we solve for a plane in reciprocal space, not the amplitude lying over the Ewald sphere as in the conventional image (compare Figs. 12 and 14).

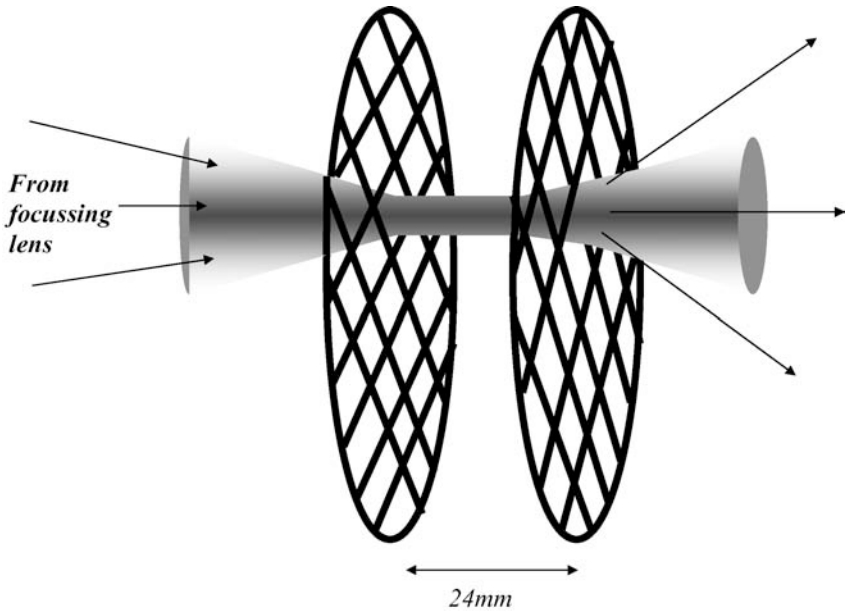


FIGURE 24. Schematic of the specimen used to obtain the reconstructions in Fig. 25.

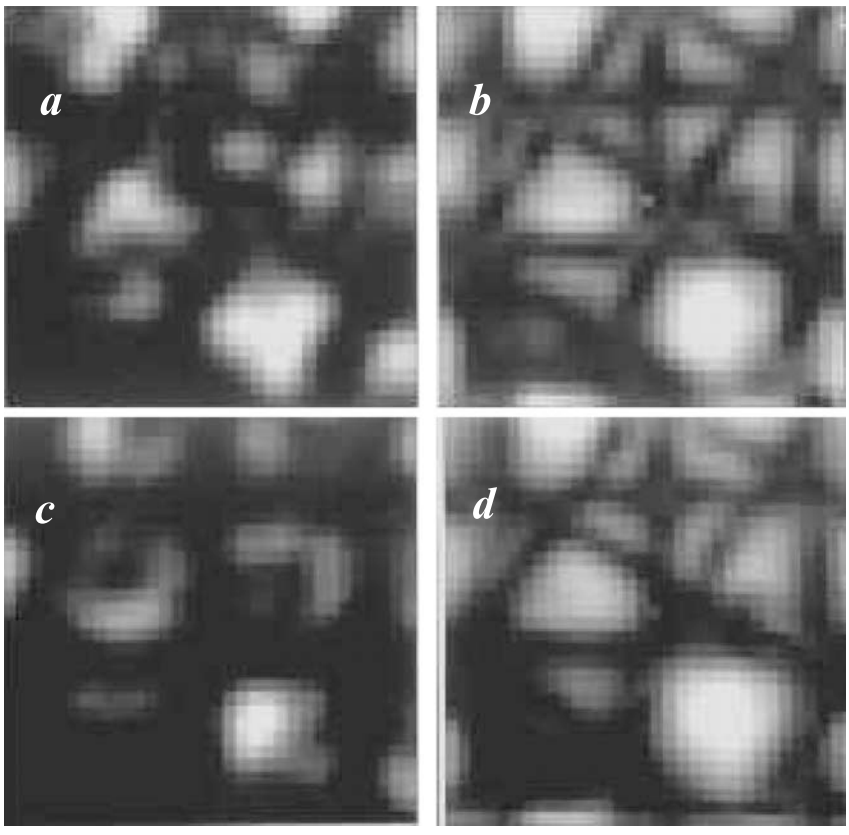


FIGURE 25. Images a and c are the conventional visible light analog STEM images of the object shown in Fig. 24. Image a is with the beam focused on the diagonally mounted grid, and image c is on the squarely-mounted grid. Images b and d are their respective Π_p ptychographs (all these images are in modulus only). The Π_p ptychographs produce a good representation of the projection of the object at double resolution.

illumination through $\boldsymbol{R}$. In the conventional bright-field images obtained with the beam crossover focused on each grid, poor images are produced, each showing the respective grids at different angles to one another (Fig. 25, a and c). In the corresponding $\Pi_p(\boldsymbol{r})$ constructed from the same experimental data (Fig. 25, b and d), but using all $\boldsymbol{u}$ and the respective Fourier components of $\boldsymbol{U}$, we obtain a double-resolution image of the two grids superimposed on each another – a projection of the object. The reason the image has higher resolution can be perceived as arising from the fact that an incident beam lying at the edge of the aperture function is scattered through twice the angle of the aperture radius to meet up with the beam that set off from the opposite

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 72
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

side of the aperture. The conventional image interferes the on-axis beam with beams lying up to only the radius of the aperture function, not its diameter.

Plamann and Rodenburg (1998) undertook some electron dynamical scattering calculations for the same scattering geometry as the Bragg–Brentano subset for the silicon [110] zone-axis dumbbell Π_p (called *reconstruction method II* in the cited paper). The results are shown in Fig. 26. The lower image in each image pair is the phase of Π_p, which approximately preserves the overall structure of the silicon dumbbells up to a total specimen thickness of almost 30 nm (these calculations were performed for 300-keV electrons). In looking at the phase and amplitude of the Bragg beams of which Π_p is composed (Fig. 27), all the beams accumulate phase in an approximately linear function of thickness. By multiplying Π_p by an appropriate phase

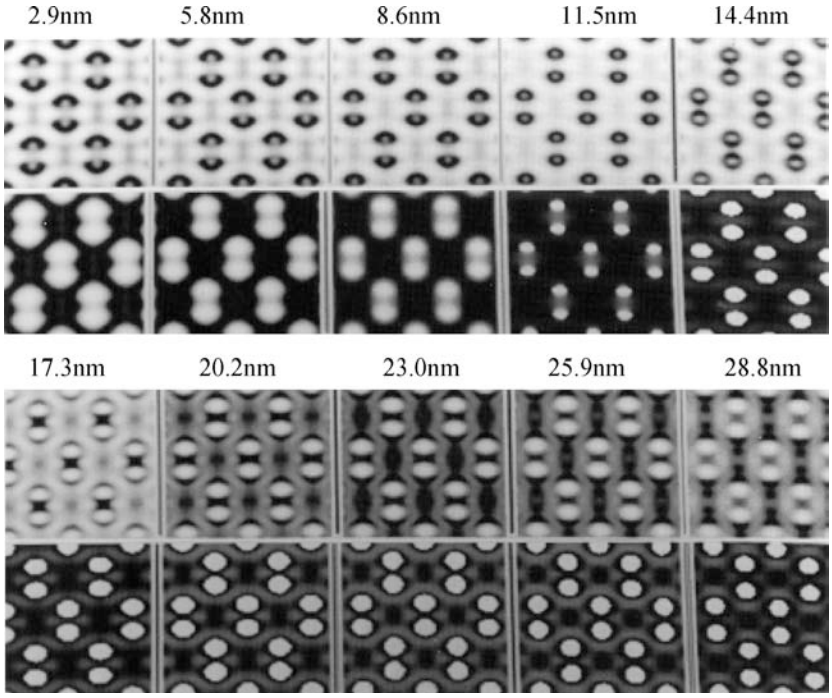
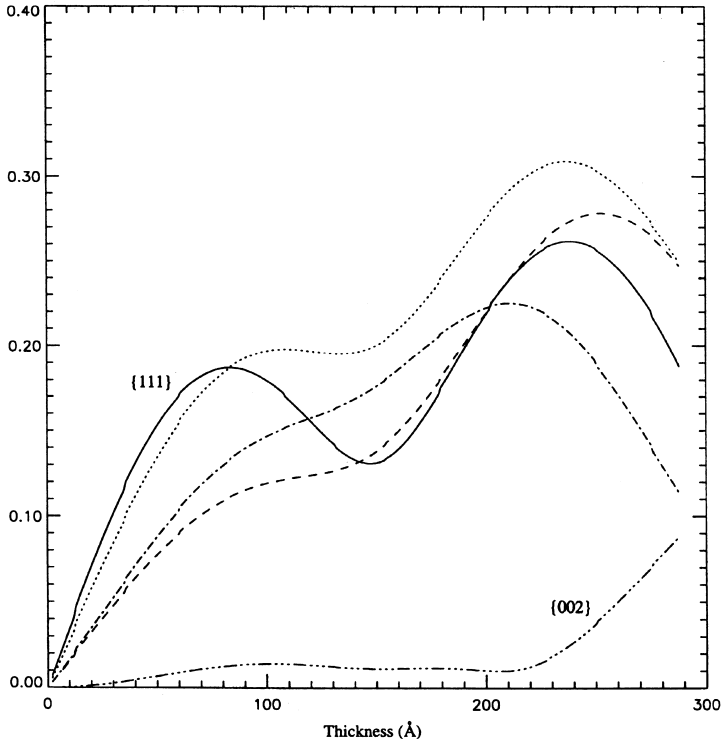


FIGURE 26. Modulus (top rows) and phase (bottom rows) of Π_p, simulated by multislice calculations, for various specimen thicknesses of silicon (see text). Thickness ranges from 2.9 nm (top left) to 28.8 nm (bottom right). The phase image is a reasonable estimate of the structure of the dumbbells. A better representation can be obtained by subtracting the phase ramp (shown in Fig. 27) and plotting the imaginary part of Π_p (Fig. 28).

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 73
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

offset, the imaginary part of the result is almost a perfect projection of the potential up to a thickness of ∼20 nm (Fig. 28).

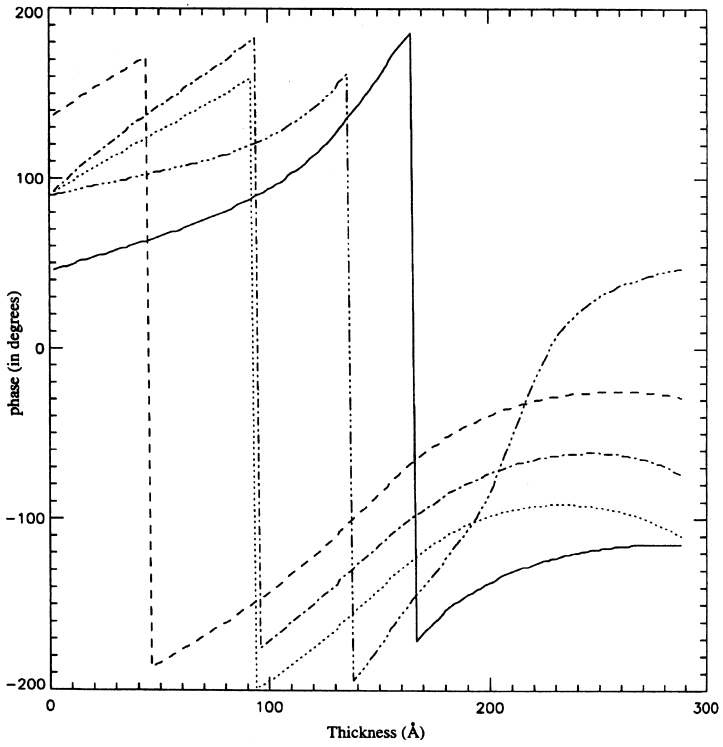
After enumerating the very beneficial properties of Π_p, there is some moderately "bad news." So far the theory above has ignored the assumption that pixels in $\boldsymbol{u}$ and $\boldsymbol{R}$ are infinitesimally small and that the positions and angles spanned by $\boldsymbol{R}$ and $\boldsymbol{u}$ are infinite. If our Fourier transforms are not undertaken over infinite limits, the convenience of the Dirac delta function no longer applies. This is not intractable, but it must be accounted for by using a softly varying filter function applied at the experimental limits of the data (McCallum and Rodenburg, 1992). A related problem that severely affects



(a)

FIGURE 27. The modulus (a) and phase (b) of various calculated diffracted beams (see text) as a function of specimen thickness: these are the beams that compose Π_p. Solid line: (111). Lowest amplitude line is the forbidden reflection (002); in this configuration it remains virtually unexcited until a thickness of 20 nm. Other beams shown are (220), (004), and (113). The main inference is that all the beams acquire a (roughly) linear phase change as a function of thickness. This is why the phase-offset ptychograph (Fig. 28) is more closely representative of the atomic structure.

aiep v.2005/11/15 Prn:6/12/2007; 13:13
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV



(b)

FIGURE 27. (*continued*)

the experimental feasibility of the method is that these integrals must indeed be performed over reasonably large fields of view. In other words, a great deal of data must be collected before any processing can occur – unlike the PIE method (the next section), there is no possibility of constructing a real-time image of the object regardless of how much computing power is available. If the detector pixels are finite in size – which they must be to get any counts at all – then the aberration insensitivity is lost, at least to first order. This is because when the aberrations are large, the intensities fringes in the Ronchigram (the central disk, the shadow image of the probe-forming aperture) are on a very small scale. When a finite pixel integrates over an area larger than a fringe spacing, the required signal is obliterated. We can picture this via reciprocity as the illuminating beam in TEM subtending a significant range of incoherent beams at the specimen plane – an important factor which affects the usable information limit. The form of the effective transfer function given these parameters was examined by Nellist and Rodenburg (1994).

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 75
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

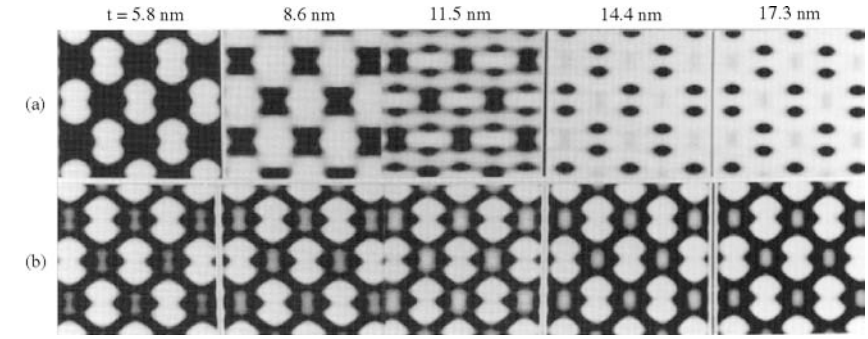


FIGURE 28. Π_p calculated for crystalline silicon on the [110] zone axis, but with a phase offset (corresponding to the inner potential) subtracted from the ptychograph. The bottom row of images shows the imaginary part of this phase-offset image (the top row is the real part of the same image). The phase very closely represents the projection of the atomic potential in the z-direction, even up to rather large thicknesses (17.3 nm on extreme right). Refer to Plamann and Rodenburg (1998) for further details.

Now for the very bad news. It should be noted here that there are some substantial experimental complications with actually collecting and processing these data, even if the mathematics is so "neat." (This issue is rather close to my own personal experience.) After the initial rush of optimism on realizing the potential benefits of processing $I(\boldsymbol{u}, \boldsymbol{R})$ in various ways, and after much fanfare at conferences declaring that this must be the way ahead for TEM, experimental realities began to gnaw at the "good news." Yes, the data seem to be robust to so many of the usual constraints of the transfer function and partial coherence, but there is a correspondingly large complement of different experimental difficulties.

To construct $\Pi_p(\boldsymbol{r})$, we may only need a 2D plane in $I(\boldsymbol{u}, \boldsymbol{R})$, but we still must collect the whole of $I(\boldsymbol{u}, \boldsymbol{R})$ in order to extract it. Between 1991 and 1996, our small Cambridge group struggled, with increasingly marginal returns, to put this (and the WDDC method described in Section IV.D) onto a practical experimental footing with high-energy electrons. Reconstructions of $\Pi_p(\boldsymbol{r})$ were obtained, but their quality and scientific relevance were exceedingly low. An example, the best obtained after several years' work, is shown in Fig. 29. It is on a grid of 32×32 pixels in $\boldsymbol{R}$. That meant that more than 900 diffraction patterns needed to be collected. Even if captured at TV rates (40 ms/frame), this implies that we needed a total data capture time of 36 s. To confirm that the reconstruction was at least reproducible, the data were collected from the same region of specimen twice to check that the reconstructions from independent data but from the same area of object were the same. For a number of irritating experimental reasons related

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 76
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

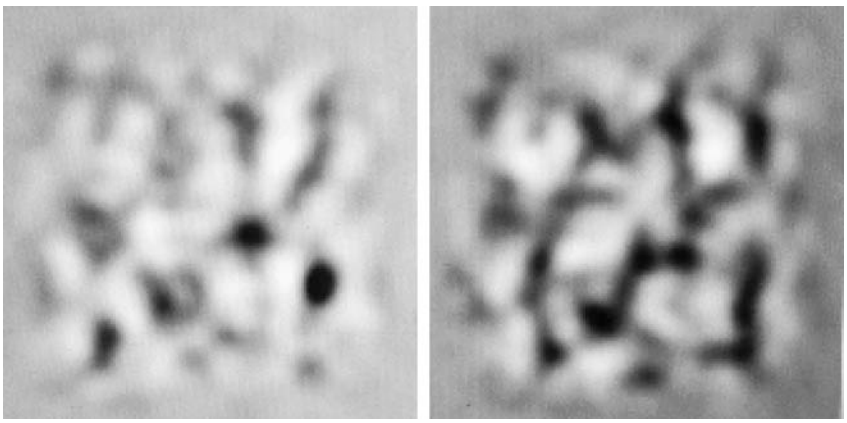


FIGURE 29. The phase of two Π_p double-resolution electron ptychographs of amorphous carbon. The two images are of the same area of the specimen (field of view 2.5 nm), but using entirely different data collected one after the other. The accelerating voltage was 100 keV. The objective aperture (in STEM mode) was stopped down so that the intrinsic resolution of the imaging lens was 0.8 nm. The reconstruction doubles this resolution and is ∼0.4 nm. Certain features (e.g., the bright V shape (top right) and the dark V shape (top left), are clearly reproduced. However, the reconstructions are not very impressive given the effort undertaken to obtain the necessary data (see text and Rodenburg *et al.*, 1993 for further details).

to simultaneously controlling the microscope and synchronizing the data capture, TV rates could not in fact be achieved in practice. The data for this tiny image therefore took about 80 s to collect. This may not sound long until it is remembered that all the foregoing theory assumes that the data are collected from a set of equally spaced points in $\boldsymbol{R}$, each separated at the Ångstrom level, with no distortion in the scan. Lateral wobble in the probe position (the source of the ptychographic phase) is fatal.

During the course of his doctoral work, Peter Nellist slowly removed more and more of the wires connecting the console of the microscope (a VG HB501-dedicated STEM) to the column in an attempt to remove sources of electronic instability in the probe position. This, combined with the other usual difficulties of TEM (such as specimen contamination, specimen damage, and those phenomena that many microscopists secretly recognize but do not very often talk about openly – like the occasional proclivity for the specimen to jump laterally every few 10 s of seconds, presumably because of charging – further hampered our efforts. Other irritations included charging of the objective aperture, inducing non-round aberrations (in a dedicated STEM, the objective aperture is the probe-forming aperture); hysteresis in the scan coils; earthing loops; imperfections and blooming in the detector (we used an image intensifier); the general paucity of electron counts that we could obtain

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 77
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

when the source was sufficiently demagnified to achieve an appropriate (even if rather low) degree of coherence; vibration in the field-emission tip (the VG design had the tip mounted on a thin bent wire, which had several vibrational modes with which we became all too familiar); and, perhaps worst of all, the need to reduce thermal drift to a usable level of about 1 or 2 Å/min. Even quite simple things, like focusing the probe at the specimen plane, were difficult tasks when the machine was switched to collecting diffraction patterns. (In short, this was very much a living nightmare, made bearable only by the group's persistent sense of humor.) Cowley and Winterton (2001) did eventually implement the scheme successfully with electrons for a 1D scan of probe positions across the edge of a nanotube. Perhaps now that detector and data transfer/storage technology is so much more advanced, it may be worth revisiting the technique for 2D objects.

C. *Ptychographical Iterative (Pi) Phase-Retrieval Reconstruction*

Conventional iterative phase-retrieval methods collect a set of data, whether a single diffraction pattern from an isolated object, and then process this offline, applying object size constraints in real space. However, as a microscopic imaging method this leaves much to be desired. Microscopists need the ability to scan their specimens over large fields of view in order to identify areas of interest. Once identified, the user must be able to zoom in to a feature of interest, observing it at higher and higher resolution. Isolated object PDI methods do not allow for this sort of essential microscope functionality.

This sections discusses a recent development in the iterative phase-retrieval method that actually owes much more to the ptychographical direct solution: multiple diffraction patterns are recorded as an illumination field is shifted with respect to the object of interest (or *vice versa*). A modified iterative method is used to construct the resulting image – a ptychograph that we will call Π_π, or a "pi-type" reconstruction. Here we examine in detail a particular variant of pi-type reconstruction called the *PIE algorithm*.

Iterative methods of phase retrieval were first proposed by Gerchberg and Saxton (1972) and later developed significantly by Fienup (1978, 1982). Early work was performed on 1D objects, if only because of computational constraints at that time. John Chapman (as differentiated from Henry Chapman, who has also contributed significantly to this field) was the first to consider the practical issues surrounding an experimental implementation of the technique in the electron microscope for the 1D image-diffraction plane iterative phase-retrieval (Gerchberg–Saxton) method (Chapman, 1975). The development of the personal computer greatly facilitated testing the efficacy of iterative methods in two dimensions and determining the results of different feedback

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 78
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

parameters immediately. In two or more dimensions, the method is much more robust because of the usual phase problem constraints (discussed in Section II.D). Rather than requiring both image and diffraction intensity data (and hence a good-quality lens, as in the Gerchberg–Saxton algorithm), it was found (Fienup, 1978) that simply knowing that the object is isolated, effectively embedded in an empty unit cell that has roughly as many (known) zero-valued pixels as pixels in the object image, then the 2D phase problem can be very effectively solved without significant ambiguity. There are still stagnation issues resulting from the competition of the inherently ambiguous solutions (such as $q(\boldsymbol{r}) = q^*(-\boldsymbol{r})$, these having the same diffraction pattern intensity). Fully complex images are more difficult to solve for (Fienup, 1987), but with some refinements, this can sometimes be achieved.

Following a rather stimulating conference organized by Spence *et al.* (2001) on solution of the nonperiodic phase problem, I was encouraged to revisit the ptychographical data set, but with the goal to try to incorporate into it all the advantages of the iterative technique. The possibility of doing the PDI experiment in the electron microscope had been discussed as part of a review of ptychography (Rodenburg, 1989), but in view of the detector technology available at that time had been dismissed. Using image-plate technology, Zuo *et al.* (2003) obtained a very impressive reconstruction of a carbon nanotube from a single electron diffraction pattern in a single, PDI data set. Given that it was possible to achieve an iteratively reconstructed image from a single diffraction pattern, and given that it is known that it is possible to solve the phase problem directly from two or more diffraction patterns via ptychography, it seemed logical to suppose that a hybrid method existed that could exploit an optimized combination of these methods.

Landauer (1996) had briefly considered processing a number of images collected under different illumination conditions in the electron microscope to reconstruct the modulus and phase of the Fraunhofer diffraction using the known constraint of the size and shape of the objective aperture lying in the back focal plane of the object lens. This was far from lenses diffractive imaging; the work was purely theoretical and had to assume the presence of a good lens. However, it did indicate that multiple applications of the Fienup hybrid input–output method could be "sewn" together to extend the plane of interest (in this case, not the image but the diffraction pattern). Faulkner and Rodenburg (2004) used an aperture constraint alone, presumed to be coincident with the object plane, to show in model calculations that the iterative method could indeed be applied to extended real-space objects. Preliminary investigations into the possibility of manufacturing an aperture via focused ion beam lithography that was small enough to perform this type of reconstruction in the electron microscope were fraught with difficulties. The aspect ratio of the hole required in order to block off the unwanted

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 79
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

radiation was large (say a diameter of 10 nm with a depth of 200 nm). Such apertures contaminate quickly in the microscope and introduce channeling and propagation artifacts.

It became obvious that in any practical implementation of iterative phase-retrieval ptychography, at least in the electron microscope, it would be necessary to use a focused spot of illumination of radiation, as in the STEM configuration. Unfortunately, such patches of illumination are necessarily "soft" in the sense that there are always low-amplitude ringing effects that extend to large distances in the object plane. Because the lens has a limited extent in reciprocal space, its real-space impulse-response function is always smooth. There are therefore no analogous support-type constraints in real space. Furthermore, any aberration or defocus in the lens introduces phase changes into the illuminating beam. A method was required that would feed into the phase-retrieval loop the essence of ptychography: when the illumination function is moved, its shape remains constant; equivalently, the diffraction pattern is a convolution, but a convolution that can be recorded for a series of phase ramps applied across the Fourier transform of the illumination function, according to the fundamental ptychographical Eq. (41).

The breakthrough came with the development of the so-called PIE update function (Rodenburg and Faulkner, 2004a, 2004b). The use of the measured modulus constraint in the Fraunhofer diffraction plane is identical in PIE to that used in all conventional iterative phase-retrieval methods (Marchesini, 2007, provides an excellent review). The difference with PIE that is there is no simple support constraint in real-space. Instead, a current estimate of the object function is changed incrementally according to the intensity information that has been fed in from the diffraction plane.

The iteration proceeds as follows. We seed the algorithm with an initial estimate of Π_π, which we will call $\Pi_{e,1}(\boldsymbol{r})$, this being our first ($n = 1$) such estimate, labeled (A) in Fig. 30. It is usual to put $\Pi_{e,1}(\boldsymbol{r}) = 1$; that is, we start by supposing that $\Pi_{e,1}(\boldsymbol{r})$ is of the form of a transmission grating function of unity transmission (free space) at all points in $\boldsymbol{r}$. Suppose we have a current nth estimate of the ptychograph $\Pi_{e,n}(\boldsymbol{r})$. Then in the next $(n+1)$th iteration, a modified estimate, will be given by $\Pi_{e,n+1}(\boldsymbol{r})$ according to the equation

$$\Pi_{e,n+1}(\boldsymbol{r}) = \Pi_{e,n}(\boldsymbol{r}) + U(\boldsymbol{r})\big[\psi_{c,n}(\boldsymbol{r}, \boldsymbol{R}_j) - \big(a(\boldsymbol{r} - \boldsymbol{R}_j) \cdot \Pi_{e,n}(\boldsymbol{r})\big)\big], \tag{89}$$

where $U(\boldsymbol{r})$ is an update function, given by

$$U(\boldsymbol{r}) = \frac{|a(\boldsymbol{r} - \boldsymbol{R}_j)|}{|a_{\max}(\boldsymbol{r} - \boldsymbol{R}_j)|} \frac{a^*(\boldsymbol{r} - \boldsymbol{R}_j)}{(|a(\boldsymbol{r} - \boldsymbol{R}_j)|^2 + \varepsilon)}. \tag{90}$$

In Eq. (89) we multiply our current estimate of the ptychograph by our illumination function (which we presume is known accurately) when it is

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 80
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

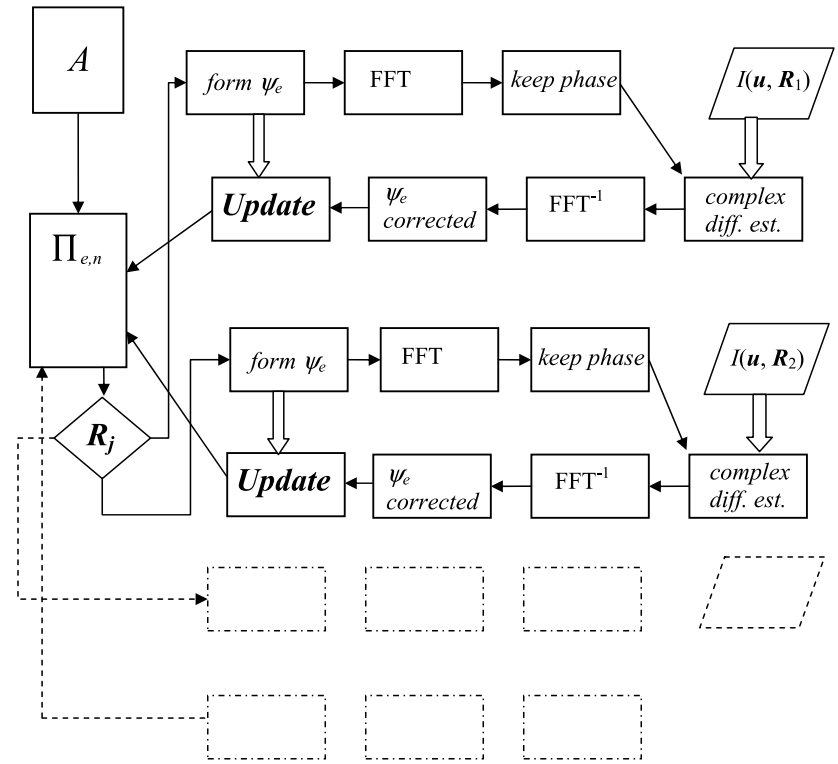


FIGURE 30. The PIE flow diagram. At 'A' we start with an initial guess of the object function, $\Pi_{e,n}(\boldsymbol{r})$, which usually is perfectly transparent (i.e., of unity modulus everywhere and zero phase). This estimate is then continuously modified at each iteration via the update function. An iteration consists of choosing a particular illumination position, $\boldsymbol{R}_j$, and then forming $\psi_e = a(\boldsymbol{r}-\boldsymbol{R}_j)\cdot\Pi_{e,n}(\boldsymbol{r})$. We then propagate this to the far field (in fact, any propagator can be used), preserve the phase of the resulting diffraction pattern amplitude but substitute the modulus calculated from the square root of the intensity of the diffraction pattern obtained from that probe position $\boldsymbol{R}_j$ to give an estimate of the complex value of the diffraction pattern. The difference between ψ_e and the corrected back-propagated exit wave, ψ_e, is used to update $\Pi_{e,n}(\boldsymbol{r})$. Any number of diffracted patterns can be used (three are shown, one dotted). Hollow pointers represent the transfer of data, line arrows show the flow of control.

at a particular position $\boldsymbol{R}j$ with respect to the object function. Assuming the multiplicative approximation (Eq. (3)), we now form what would be an estimate of the exit wave if the object was truly a 2D grating, represented in Eq. (89) by the last term in the square brackets, that is, $a(\boldsymbol{r} - \boldsymbol{R}_j) \cdot \Pi_{e,n}(\boldsymbol{r})$. This wave is then propagated to the detector plane where, as usual in iterative methods, we preserve the phase of the resulting calculated diffraction pattern but replace the modulus with $\sqrt{I(\boldsymbol{u}, \boldsymbol{R}_j)}$. We then transform back to the

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 81
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

object plane to find a "corrected" estimate of the same exit wave function (corrected in the sense that measured intensity data in the far-field has now modified our original estimate of the exit wave), appearing as the first term in the square brackets of Eq. (89), which we call $\psi_{c,n}(\boldsymbol{r}, \boldsymbol{R}_j)$. We subtract (point-by-point) the estimated exit wave from the corrected exit wave. The resulting difference function is then fed back into our running estimate of the ptychograph. The amount we change the current estimate $\Pi_{e,n}(\boldsymbol{r})$ is different at different parts of the image, weighted point by point as a function of the probe amplitude when it is positioned at $\boldsymbol{R}j$. In Eq. (90) $a_{\max}$ is the maximum modulus value of the illumination function, and ε is a constant determined by the total number of counts in a typical diffraction pattern. This term is like a Wiener filter as used in conventional deconvolution methods to suppress the effects of noise in the data, especially in the regions where $a(\boldsymbol{r})$ has low or zero amplitude. It is essentially a division by $a(\boldsymbol{r})$, effecting the deconvolution of the ptychographic folding (Eq. (41)) of the far-field intensity.

Although the PIE can be run with some success using only one diffraction pattern (one value of $\boldsymbol{R}j$), its real advantages arise when data from different probe positions are used concurrently. If the area of the illumination function has an approximate diameter d, convergence is rapidly increased if data are collected from a grid of points separated in $\boldsymbol{R}$ by about $d/2$. Thus, for example, we could collect four diffraction patterns from $\boldsymbol{R}_1 = (0, 0)$, $\boldsymbol{R}_2 = (d/2, 0)$, $\boldsymbol{R}_3 = (0, d/2)$, and $\boldsymbol{R}_4 = (d/2, d/2)$. The first iteration could use experimental data $I(\boldsymbol{u}, \boldsymbol{R}_1)$, the second $I(\boldsymbol{u}, \boldsymbol{R}_2)$, and so on, returning at $n = 5$ to reuse data $I(\boldsymbol{u}, \boldsymbol{R}_1)$ and so on, reusing each diffraction pattern in a cyclically repeating order. However, many other permutations are possible. The jth pattern could be used for the first hundred iterations before moving on to the $(j + 1)$th pattern. Clearly, this latter strategy results in rather slower convergence because the ptychographic phase data are only being exploited once in every hundred iterations.

In general, processing a different pattern for each subsequent iteration is the most effective strategy. Indeed, this can produce a reasonable quality image after iterating with each diffraction pattern just once (Rodenburg *et al.*, 2007a) In conventional iterative methods, it routinely takes thousands of iterations before a reasonable image can be obtained. The method is also fantastically immune to noise. Faulkner and Rodenburg (2005) modeled a variety of experimental sources of noise in a STEM configuration and found that recognizable images could be obtained when Poisson noise corresponding to very low beam currents was added. Up to 50% of random noise added to each detector pixel reading also produced interpretable images. Conventional iterative methods often cease to converge sensibly at noise levels of more a than few percent.

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 82
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

It should be remembered that the data collection process is completely decoupled from the iterative calculation itself. The reconstruction can be processed offline, after all the data have been collected, or in real time, concurrently as further data are being collected. Because the PIE is completely indifferent to the position of the next $\boldsymbol{R}_j$, the field of view can extended continuously, in real time, according to whatever part of the specimen the user wishes to image. This is one of the defining properties of PIE. The field of view is independent of the size of the detector. Clearly, the sampling condition must still be satisfied, but this depends only on the width of the illumination function – it is nothing to do with the size of the final field of view. The forward and back propagations necessary for each iteration can be conducted on a reciprocal space sampling of say, $\Delta \boldsymbol{u}_s$, which defines a value of $d \approx 2/\Delta u_s$. However, any particular value of $\boldsymbol{R}j$ is unconstrained by d. We could use data from $\boldsymbol{R}_1 = (100d, 50d)$ and $\boldsymbol{R}_2 = ((100 \cdot 5)d, (50 \cdot 5)d)$ and use the update function to alter only the area of ptychograph spanned by the illumination function when it is centered on either of these coordinates. There is nothing to stop Π_π spanning a total image area of, say, $1000 \times 1000d^2$. If the detector has 1000×1000 pixels, then in this situation the entire ptychograph could contain $10^6 \times 10^6$ pixels: unlike conventional iterative methods, the size of the reconstruction can be as large as desired.

Figure 31 shows a visible light experimental demonstration of the PIE method (from Rodenburg *et al.*, 2007a). The illumination function was generated simply by an aperture somewhat upstream of the object (a 3-mm long ant sectioned and mounted between a microscope slide and cover slip). The aperture was 0.8 mm in diameter, with $\lambda = 630$ nm. The object was moved over a grid of 10×10 illumination positions, each separated by slightly less than half the aperture width. An example of a typical diffraction pattern is shown in Fig. 32, as is the recorded intensity of the illumination function at the plane where it impinges on the object. As can be seen, the disc of the real-space aperture has been blurred out by propagation, causing Fresnel fringe-like features to develop around the edges of it. This function is therefore "soft" and does not present us with the sort of sharp support function required for conventional iterative phase retrieval. Figure 31 shows intensity and phase reconstructions of Π_π. The resolution of the ptychrograph is $\sim$17 μm; that is, about 50 times smaller than the diameter of the aperture.

Note that in the phase of the reconstruction we can discern a phase ramp across the entire image. This is due to the cover slip not being perfectly parallel to the underlying slide. More recent unpublished work has shown slowly changing phase distribution over very wide fields of view can be accurately reconstructed. The same sort of information can be extracted from holograms, but here we do not have any interferometric optical component

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 83
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

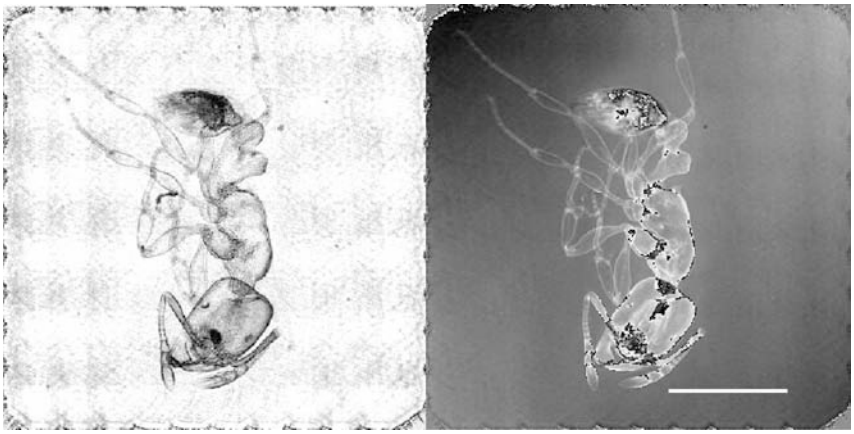


FIGURE 31. Visible light Π_π reconstruction (i.e., $\Pi_{e,n}$, where in this case $n = 25$). Residual structure in the modulus (left) arises from slight detector misalignment and misestimate of the illumination function: the 10×10 scan of probe positions is hence visible. Note that in the phase image (right), a slowly varying phase ramp is discernible – such large-scale phase variations (in this case due to a curved cover slip) are lost in conventional Zernike and electron microscopic images. The ant is ∼3 mm in length, the scale bar is 1 mm. For more details, see Rodenburg *et al.* (2007a).

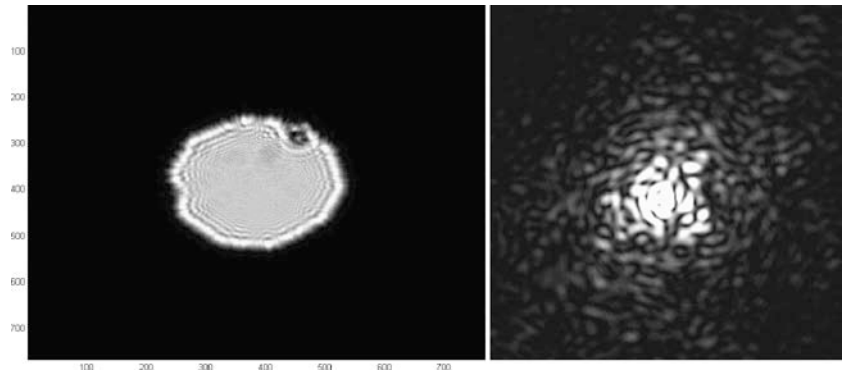


FIGURE 32. Raw data of ant reconstruction shown in Fig. 31. The intensity of the illumination function (left) is measured at the plane of the object. Note that propagation between the aperture and the object has led to the development of Fresnel fringes around its edges. A typical diffraction pattern is shown on the right.

or lens, except the diffraction itself occurring at the object plane. This is a potentially important imaging capability of pi-type ptychography.

A much more scientifically significant result is shown in Fig. 33 (from Rodenburg *et al.*, 2007b) where an exactly analogous experimental configuration

aiep v.2005/11/15 Prn:6/12/2007; 13:13
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

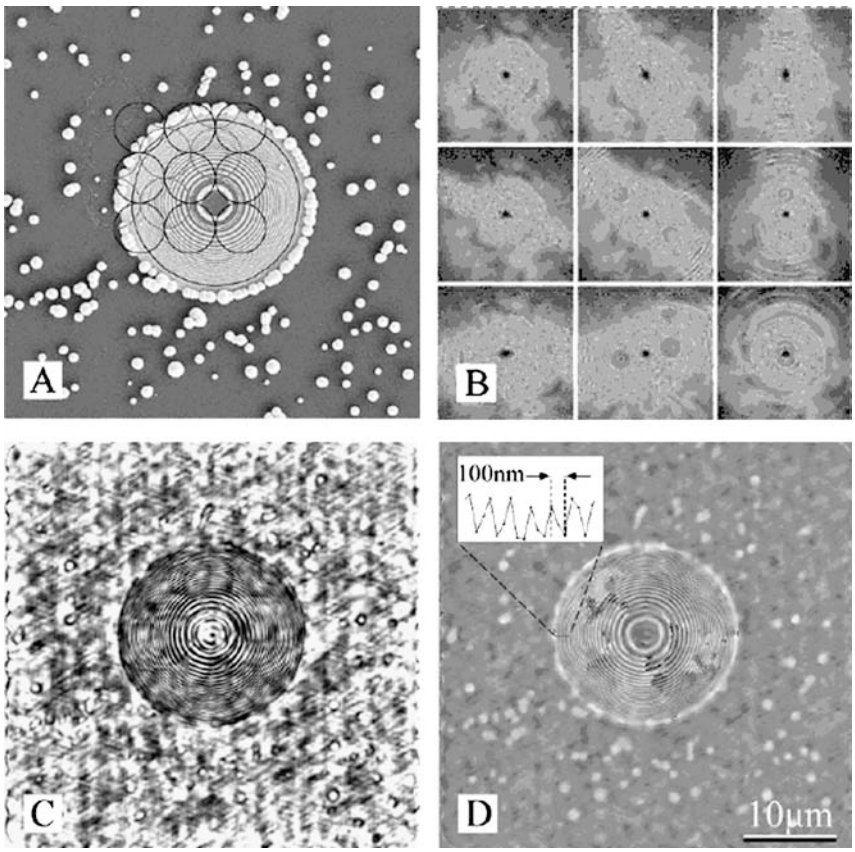


FIGURE 33. Hard X-ray PIE reconstructions from Rodenburg *et al.* (2007b). Image A: High-resolution SEM image of the object. B: Examples of nine of the diffraction patterns recorded from the illumination at the positions circled on A. C: Modulus and D phase of the Π_π ptychograph. Aliasing effects in the reconstructions are due to the sampling in the image file; these features are not visible in the raw data – see the plot across the outer rings of the zone plate in D.

was used with hard (8-keV) X-rays. In visible light optics there is very little need to resort to ptychography to obtain good-quality images. However, any sort of imaging with X-rays is difficult, and especially difficult with hard X-rays because of the thickness required in the zone plate to introduce the requisite phase change in the incident beam. The round object in the center of Fig. 33 is in fact such a zone plate, although here it is being used as a resolution test object. The experiment was performed at the Swiss synchrotron at the Paul Scherrer Institute in collaboration with the group of Franz Pfeiffer. The only difference in the experimental setup is the source of radiation

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 85
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

(a substantially coherent third-generation synchrotron) and the scale of the aperture, which in this case had a diameter 5 μm. The aperture to object distance was a few millimeters. It was not possible to image the intensity of the incident illumination field, and so this had to be estimated by assuming the aperture was a perfect disk illuminated evenly by a perfectly coherent plane wave and propagating this computationally to the object plane. The object was stepped through a square array of 17×17 positions, each separated from one another by 2 μm. The (mechanical) stepper motors used were not of a particularly accurate: piezoelectric actuators could certainly improve on the result. The specimen includes a random distribution of gold particles scattered around the zone plate.

The reconstruction (especially the intensity component) is far from perfect, but this was the first attempt at this experiment; the data were collected over a single weekend on a beamline that was scheduled for maintenance. Given that the outer zones of the zone plate are resolved, the ptychograph has clearly achieved resolution that is as good as that which could be obtained from this (reasonably state-of-the art) zone plate when used as an imaging lens. With further development, it would seem that pi imaging might provide routine hard X-ray micrographs at potentially much higher resolution than currently available technology.

Pi is certainly the most compact and elegant solution to the calculation of $\Pi(\boldsymbol{r})$ from the fundamental ptychographic Eq. (41). The deconvolution of the diffraction plane amplitude (Fig. 14) is achieved in real-space by the division implicit in Eq. (41) of $a(\boldsymbol{r})$. The method allows for immediate real-time output of ptychographs as data are concurrently being collected. The solution so obtained can be refined in one part of the field of view while further data collected from new areas of the object. There is no limit to the size of the field of view. The computational method has been called an "engine" (Faulkner and Rodenburg, 2005) – a "ptychographic iterative engine" (PIE) – because, unlike all the preceding iterative phase algorithms, the measured data arrays can be continuously expanded while the algorithm progresses through the data. In principle, the user can define in real time which parts of the objects should be reconstructed.

The pi data set is of dimension $N_u \times N_v \times N_{1-\infty} \times N_{1-\infty}$. In this sense, it really is a diffractive method, most of the information being captured in the $\boldsymbol{u}$ coordinate, only a few sparse points in the $\boldsymbol{R}$ plane being required. It is both extremely flexible in that only as much data as are needed for a particular object size need to be collected (the number of illumination positions is variable between 1 and infinity). Since each area of the object is generally illuminated about twice, it is exceedingly efficient relative to the Bragg–Brentano or Wigner distribution deconvolution (WDDC) methods (Sections IV.B and IV.D). Indeed, the method will work even if most of the

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 86
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

object is only illuminated once, although increased redundancy helps with the convergence of the algorithm.

D. The Wigner Distribution Deconvolution Method

In Section IV.B we formed a function $G(\boldsymbol{u}, \boldsymbol{U})$, calculated by taking the Fourier transform of $I(\boldsymbol{u}, \boldsymbol{R})$ with respect to the $\boldsymbol{R}$ coordinate. From this we could extract $B(\boldsymbol{U})$ and hence the projection pytochograph. However, $B(\boldsymbol{U})$ is just a 2D subset of an intrinsically 4D data set. By picking only one particular Fourier component in $\boldsymbol{U}$ for each location in $\boldsymbol{u}$, a huge amount of data are wasted. In fact, the Bragg–Brentano subset was only developed much later as a result of a different line of reasoning that had been sparked in the autumn of 1988 when Owen Saxton very helpfully introduced me to Professor Richard Bates, who was passing through Cambridge on his way from New Zealand to a conference in Europe. Our meeting only lasted 20 minutes and was mostly spent brewing and drinking tea. Richard was severely jet lagged. His long white hair was characteristically disheveled; he struck me as the archetypal mad professor. I had an imminent teaching commitment that afternoon and so had to bring the social niceties to an abrupt end. In the last two minutes before racing off, I scribbled down Eq. (41) and said that I would be very grateful to receive any suggestions on how to obtain a comprehensive solution for the object function $q(\boldsymbol{r})$, given only the intensity $I(\boldsymbol{u}, \boldsymbol{R})$ and a knowledge of $a(\boldsymbol{r})$. I did not hear from him at all until I received a telex 6 months later: it said, "I have solved your problem. Please send fax number." The stream of equations that subsequently came out of the fax machine at first completely confused me. A day or two later I was first enlightened and then delighted. Two weeks later I was frantically writing a research grant proposal. Without that telex, this Chapter, and most of what is in it, would not exist. The brief collaboration that followed was truly fruitful. I was a cynical experimentalist; Richard was an inspired theoretician who passionately loved the phase problem and had worked on it for many years from a wide range of perspectives. I could relate many stories at this point: sometimes sparks would fly; Richard was a true original but also, at times, incredibly stubborn. He died of cancer 2 years after our first meeting, just as a vast horizon of future work was expanding before us. I was impressed that he remained absolutely focused on productive work right until the very end, never at any point being aggrieved that fate had determined that he would die much younger than today's average span.

What he wrote in that fax (disguised somewhat in his inimitable nomenclature and penchant for massive equations) was that if

$$H(\boldsymbol{U}, \boldsymbol{r}) = \iint I(\boldsymbol{u}, \boldsymbol{R}) e^{i2\pi(\boldsymbol{R}\cdot\boldsymbol{r} - \boldsymbol{u}\cdot\boldsymbol{U})}\, d\boldsymbol{u}\, d\boldsymbol{R}, \tag{91}$$

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 87
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

then

$$H(\boldsymbol{U},\boldsymbol{r}) = \chi_q(\boldsymbol{U},\boldsymbol{r}) \cdot \chi_a(\boldsymbol{U},-\boldsymbol{r}), \tag{92}$$

where for some general function, $f(\boldsymbol{r})$,

$$\chi_f(\boldsymbol{U},\boldsymbol{r}) = \int f(\boldsymbol{r}+\boldsymbol{r}_a) f^*(\boldsymbol{r}_a) e^{i2\pi \boldsymbol{U}\cdot\boldsymbol{r}_a}\, d\boldsymbol{r}_a. \tag{93}$$

Similar equations occur in the radar literature, where Eq. (73) defines the "ambiguity function," although it that field the measured data are phase sensitive insofar as the time of arrival of a reflection is known. This class of function is also called a *Wigner distribution function* (Johnston, 1989) – it arises as an expression of the uncertainty principle in quantum mechanics, reciprocal and real-space coordinates.

Why is this important? First, it served as a very useful mathematical tool to analyze the whole of the "entire" data set $I(\boldsymbol{u},\boldsymbol{R})$. Second, it showed that $I(\boldsymbol{u},\boldsymbol{R})$ could be decomposed in this mixed real and reciprocal space into a product of two functions, one of which depends only on the specimen transmission function $q(\boldsymbol{r})$, the other depending only on the illumination function $a(\boldsymbol{r})$. It follows that $I(\boldsymbol{u},\boldsymbol{R})$ can in principle be separated via deconvolution into a 4D function that depends only on the object function and has all the properties of the probe forming optics, whatever they may be, removed from it. Third, once this separation has been accomplished, one further Fourier transform (over two of the four coordinates) of $\chi_q(\boldsymbol{U},\boldsymbol{r})$, can yield one of two representations of all the relative phase relationships in the scattered wavefield. In real space we have

$$L(\boldsymbol{U},\boldsymbol{R}) = q(\boldsymbol{R}+\boldsymbol{U}) q^*(\boldsymbol{R}), \tag{94}$$

and in reciprocal space we have

$$D(\boldsymbol{u},\boldsymbol{r}) = Q^*(\boldsymbol{u}-\boldsymbol{r}) Q(\boldsymbol{u}). \tag{95}$$

It is perhaps not immediately obvious that either one of the two equations above gives the phase of the entire scattered wavefield (or at least, the phase and amplitude of the resulting ptychograph). In fact, it does so with massive overredundancy. Of course, in any phase problem we are never going to able to solve the absolute phase of the scattered waves because this changes at very high frequency as a function of time and so is meaningless in the context of the time-independent wave equation. This is expressed in, say, Eq. (94), by the fact that only the difference in the complex value of q between a point at $\boldsymbol{R}$ and $\boldsymbol{R}+\boldsymbol{U}$ is expressed in $L(\boldsymbol{U},\boldsymbol{R})$. We therefore start by agreeing to assign zero phase to one pixel in $q(\boldsymbol{R})$, for example, by putting

$$q(0) = \sqrt{L(0,0)}. \tag{96}$$

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 88
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

The WDDC ptychograph is then given by

$$\Pi_W(\boldsymbol{U}) = \frac{L(\boldsymbol{U}, 0)}{\sqrt{L(0,0)}}. \tag{97}$$

However, we can choose to assign any phase to any pixel in $q(\boldsymbol{R})$, and form a different ptychograph by dividing $L(\boldsymbol{U}, \boldsymbol{R})$ by that pixel. In other words, we can form a whole family of ptychographs

$$\Pi_W(\boldsymbol{U}, \boldsymbol{R}_U) = \frac{L(\boldsymbol{U}, \boldsymbol{R}_u)}{|\sqrt{L(0, \boldsymbol{R}_u)}|}. \tag{98}$$

It was quickly discovered (Rodenburg, 1989) that though the WDDC could be made relatively immune to noise by performing the deconvolution step using a Wiener filter, fatal errors arose when the illumination function was "soft" in being bandwidth limited in the sense that there is a sharp aperture in the back focal plane (e.g., as in the case when $a(\boldsymbol{r})$ is of the form of a STEM probe generated by a lens with a diaphragm). This is akin to the difficulties that arise in iterative phase retrieval in the same configuration when there is no definite and sharp support function in the object plane. The calculation presented in Bates and Rodenburg (1989) had to be performed using a truncated probe (i.e., a finite probe cut off with a sharp edge) in order to make the method work at all. Although this was alluded to in the last paragraph in Section 3 of that paper, it was of great concern to the author. Clearly, if a sharp-edged feature is a requirement of the method, then it is also true that a lens of very wide numerical aperture is required, defeating the whole purpose of ptychography.

The reciprocal version of the method (Eq. (96)) was found much later to overcome this latter difficulty (Rodenburg and Bates, 1992). In reciprocal space, the aperture function is sharp, and so its Fourier transform has value out to very large values of $\boldsymbol{u}$, meaning that the deconvolution step does not encounter divide-by-zeros except at a few points; these can be avoided by using other volumes of the data set (which is extremely redundant). Once again, we can construct a whole family of ptychographs. The fact that the aperture is finite may seem to imply that one such ptychograph, for example

$$\Im\big(\Pi_W(-\boldsymbol{r})\big) = \frac{D^*(0, \boldsymbol{r})}{\sqrt{D(0,0)}} \tag{99}$$

would be limited in $\boldsymbol{r}$, and hence there would be no gain in resolution. But this is not the case. We use the term *stepping out* (see Section IV.A and Fig. 17(a)) to suggest that even if $D(\boldsymbol{u}, \boldsymbol{r})$ is of limited extent in $\boldsymbol{r}$, we can use the phase reconstructed from one plane, say from $D(\boldsymbol{0}, \boldsymbol{r})$, to obtain an estimate of $\Im(\Pi_W(\boldsymbol{r}))$ up to the aperture diameter 2α. Once we have the phase at $\Im(\Pi_W(2\alpha))$, we now use a new plane of $D(\boldsymbol{u}, \boldsymbol{r})$, namely, $D(2\alpha, \boldsymbol{r})$, to

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 89
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

increase the extent of $\Im(\Pi_W(\boldsymbol{r}))$, via

$$\Im\big(\Pi_W(-2\alpha-\boldsymbol{r})\big)=\frac{D^*(2\alpha,\boldsymbol{r})}{\Im(\Pi_W(-2\alpha))}. \tag{100}$$

The procedure can be repeated many times. There is also an equivalent but much richer set of possible phase-closure paths (as in the crystalline case; see Fig. 20). $D(\boldsymbol{u},\boldsymbol{r})$ can be considered as a continuous set of *all* relative phase differences between *every* pair of Fourier components of Π_W.

This leads to an important – perhaps the most important – theoretical result derived from the WDDC analysis. In practice, no electron or X-ray experiment is perfectly coherent. Even the Bragg–Brentano subset is subject to the coherence envelope if, as must be the case, the detector pixels are of finite size. Consider the question of spatial coherence deriving from the fact that any practical source is also of finite size. In the STEM configuration this means the back focal is not illuminated perfectly coherently. In the diffraction plane we can no longer write Eq. (41), but we must moderate the strength of the interference between every pair of beams that interfere with one another by a coherence function. The entire data set becomes

$$\begin{aligned} I(\boldsymbol{u},\boldsymbol{R})=\iint\Gamma(\boldsymbol{u}-\boldsymbol{u}_a)A(\boldsymbol{u}_a)Q(\boldsymbol{u}-\boldsymbol{u}_a)A^*(\boldsymbol{u}_b)Q^*\\ \times(\boldsymbol{u}-\boldsymbol{u}_b)e^{i2\pi\boldsymbol{R}\cdot(\boldsymbol{u}_a-\boldsymbol{u}_b)}\,d\boldsymbol{u}_a\,d\boldsymbol{u}_b, \end{aligned} \tag{101}$$

where $\Gamma(\boldsymbol{u}-\boldsymbol{u}_a)$ is called *complex degree of coherence*: assuming that the propagator between the source and the back focal plane of the probe-forming lens satisfies the Fraunhofer approximation, then, via the van Cittert–Zernike theorem (see, for example, Born and Wolf, 1999), $\Gamma(\boldsymbol{u}-\boldsymbol{u}_a)$ is given by the Fourier transform of the intensity profile of the source function. We simply state the result (derived in Rodenburg and Bates, 1992) that under these circumstances

$$D(\boldsymbol{u},\boldsymbol{r})=\Gamma(\boldsymbol{r})Q^*(\boldsymbol{u}-\boldsymbol{r})Q(\boldsymbol{u}). \tag{102}$$

In other words, if the transfer function is limited by partial coherence effectively reducing the usable aperture size, then this appears just as an attenuation factor in $D(\boldsymbol{u},\boldsymbol{r})$, but only in the $\boldsymbol{r}$-directions. The stepping out process is not affected by finite aperture size, requiring instead only adjacent beams to interfere with one another; the quality and resolution of the reconstruction obtained via WDDC are not limited by partial coherence. This is equivalent to the observations made in Section IV.A in the context of crystalline ptychography. Indeed, the whole point of the silicon line-data reconstruction was to demonstrate that ptychography could extract image data at a resolution not limited by the coherence width of the lens. Historically,

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 90
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

the realization of the independence of the ptychographical data set on partial coherence came earlier (by about 4 years) than the very limited proof of this fact achieved using crystalline line-scan data set (described in Section IV.A). It was the magnitude of the experimental difficulties in collecting the entire data set (see Section IV.B) that meant this (perhaps the most crucial) property of the WDDC data set could not be demonstrated on nonperiodic objects. Sadly, the order of publication has implied that the only ptychographical data set that is immune to partial coherence is the crystalline data set.

A 1D visible light demonstration of the WDDC principle was first achieved by Friedman (Friedman and Rodenburg, 1992) as his master's project. McCallum and Rodenburg (1992) demonstrated a 2D imaging version of the technique, also using visible light. Pictures of the reconstructions are shown in Fig. 34. Again, these are not particularly impressive, but it must be remembered that even the largest laboratory-scale computer workstations

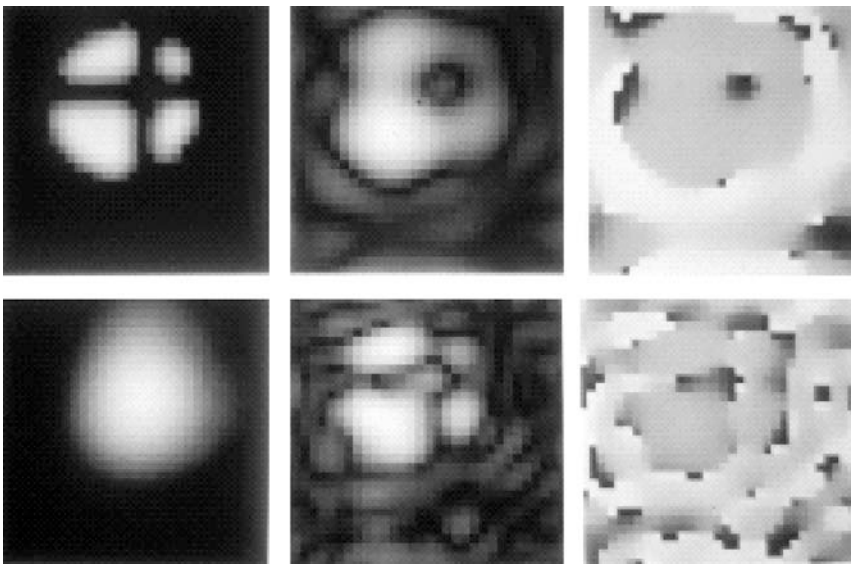


FIGURE 34. First visible light optical demonstration of the WDDC method. The top left image is the intensity of a good-resolution conventional image, taken with a lens of relatively large numerical aperture. The bottom left image is the conventional image obtained with the aperture stopped down so that all the image features are no longer resolvable. All the other images are reconstruction from ptychographical data collected using this latter (small numerical aperture) lens configuration. The top middle and right images are the modulus and phase of the Π_W ptychograph reconstructed at double resolution. The corresponding lower pair of images are at four times improved resolution (compare bottom left), thus demonstrating the so-called stepping-out procedure for nonperiodic objects. See McCallum and Rodenburg (1992) for experimental details.

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 91
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

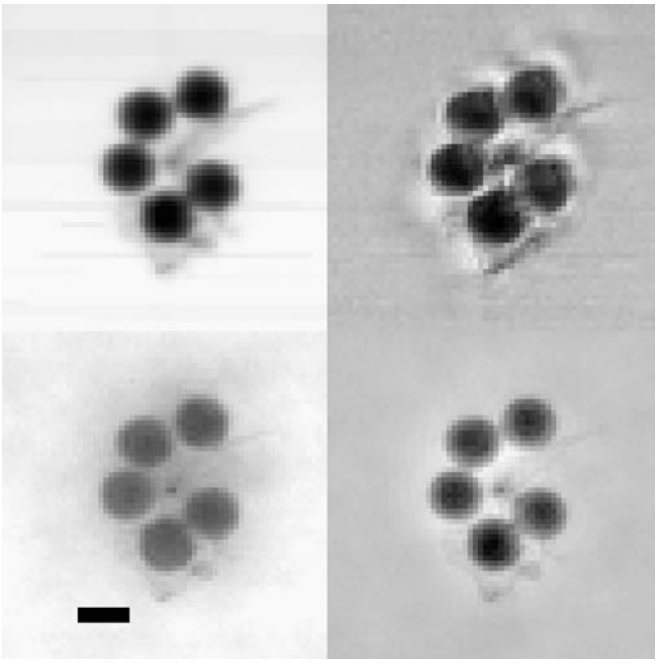


FIGURE 35. First demonstration of the WDDC method using X-rays by Chapman (1996). Top left: the incoherent bright-field image of the object (a group of latex particles). Top right: the conventional coherent bright-field image taken from an offset pixel in the diffraction plane. Lower images are the modulus (left) and phase (right) of the Π_W ptychograph. The scale bar is of length 0.5 μm.

at that time only had about 64 Mbyte of RAM. Given that we performed the deconvolution in double-precision arrays, it becomes clear why images on a scale larger than 32×32 pixels (i.e., 900 diffraction patterns) could not be processed. In an astonishing solo effort, Henry Chapman demonstrated the technique using soft X-rays (Chapman, 1996), as shown in Fig. 35. These results were much more impressive than anything obtained in the electron microscope, but the gains in resolution relative to the incoherent image were nevertheless not entirely compelling given the extreme quantities of data and beam time required for the technique. Nellist and Rodenburg (1994) explored how other aspects of partial coherence in the electron microscope – for example, lens instability, detector pixel size (which can be thought of via reciprocity as form of incoherence function), and source voltage instability – would impact the WDDC method. McCallum and Rodenburg (1993a) showed that the huge redundancy in the entire WDDC set means that it is possible to solve for both the object function and the

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 92
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

illumination function simultaneously, provided it is at least known that the illumination function is well localized. For those who prefer not to deal with the mathematics, Rodenburg (2001) has written a rather simplified review using pictorial representations of how WDDC relates to Fourier holography, tilt-series reconstruction, and other methods.

It is possible that it may be appropriate to revisit the WDDC method now Moore's law has made computers so much larger, faster, and more powerful. Certainly there are many unanswered questions about how best to optimize the technique, especially in the context of recent success of the PIE method. The latter is far more practical and much easier to implement. However, melding or extending PIE with WDDC or the Bragg–Brentano data in some way should, I think, be the subject of further work.

V. Conclusions

The scope of this Chapter has been limited to a particular definition of the rarely used term *ptychography*. I believe the definition chosen is in the spirit of that intended of its inventor, Walter Hoppe. However, there is no doubt that I have stretched this definition to embrace some techniques that are more comprehensive than the original concepts outlined in Hoppe's 1969 papers. Indeed, my chosen formal definition excludes one of the techniques (the two-beam technique) published under the designation "ptychography" by Hoppe and his then research student, Hegerl (Hegerl and Hoppe, 1972). Some years ago I had dinner with Hegerl (who is now retired). I got the distinct impression that he did not very much enjoy working on his doctorate. Even in the *Ultramicroscopy* issue celebrating Hoppe's retirement, Hoppe (1982) himself wrote about the subject in such a way as to suggest that this was an obscure but interesting idea that was never properly pursued and would probably be exceedingly difficult to put into practice. Hegerl certainly seemed to be genuinely surprised that anyone should be bothered to want to know more details about the work that he had undertaken during a period of his life that he was grateful to have put behind him. Indeed, his initial reaction seemed to be that anyone pursuing this line of reasoning must themselves be heading towards disillusion and despair. (During the 1990s, I did temporarily reach a similar state of mind for the reasons discussed in Section IV.B.)

Since I am so close to the subject, it is possible that I have overinterpreted Hoppe's insight into the phase problem, unconsciously dressing it up with the wisdom of hindsight. His original thoughts certainly appear to reside in isolation from the rest of the existing phase-problem literature. I suppose the concept was simply too far ahead of its time. In 1968 there were huge barriers to putting the idea into practice at atomic wavelengths. The first was that any

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 93
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

sort of computing was extraordinarily more difficult than it is today. I imagine that Hegerl spent all of his research coding the necessary equations to solve for the ptychographical phases. The same work today can be done in an afternoon using MATLAB. Indeed, solution of the phase problem in any field progressed so much faster once computers were powerful enough to provide real-time graphical output of images and functions. Such facilities only became widely available in the 1980s, and even then they were not entirely easy to use. Most of the progress made in the early 1990s with the WDDC and Bragg–Brentano methods derived from the access to a dedicated network of high-powered workstations (which were grossly underpowered relative to a modern laptop), all of which were equipped with graphical interfaces. The 4D data sets we used could be processed by the image processing program IMPROC, originally developed by Owen Saxton. The version we used had been greatly changed first by Richard Bates' New Zealand group and then by Bruce McCallum, who had been a doctoral student with Bates and who rewrote it into a 4D version. Once we had immediate access to graphical output from slices taken through the entire 4D set and its Fourier transform, many new ideas came to mind and could be put into practice.

Aside from computers, the two most important experimental issues that affect ptychography relate to the degree of coherence in the illuminating beam and the detector efficiency and dynamic range. In the case of X-rays, third-generation sources with high coherence have only relatively recently become available. Although a suitable degree of coherence could be achieved in the STEM about the time of Hoppe's work (Crewe was at that time obtaining the first atomic-resolution STEM results; Crewe *et al.*, 1968), this technique was far from routine and vacuum constraints in STEM meant that photographic film (the highest-efficiency 2D detector technology at that time) could not be used because it would outgas into the ultrahigh vacuum. The work undertaken in the early 1990s on the method was therefore really the first time it could ever have been undertaken. Even then the detectors and computing facilities were only marginally capable of doing justice to the technique. Now that these technological difficulties are tractable, perhaps the age of ptychography has arrived at last.

Various flavors of ptychography are discussed in Section IV. Unlike holography or through-focal reconstruction, the method does not recover the exit wave unless the object is thin. However, the image it provides can, if processed appropriately, give a good representation of the object structure, even in the presence of dynamical scattering. Furthermore, the scattering geometry of ptychography (Fig. 14) gives access to a rather larger volume of reciprocal space than that of the conventional image. Section IV.B showed why one subset of these data can provide a perfect projection of the object, suggesting that ptychography may be extended to 3D imaging via conventional tomographic

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 94
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

techniques. A further key advantage of the technique is that it is not subject to the usual limitations of the coherence envelope and hence there is, in theory, no limit to the resolution of the ptychography other than the wavelength. What has not been discussed here is the fundamental limitations imposed by thermal diffuse scattering and the fall-off of the elastically scattered signal at high angles, at least in the case of electrons. This will certainly define an ultimate physical limitation to the resolution of electron ptychographic reconstructions (which, short of using atoms or protons, will provide the highest imaging resolution of the solid state). However, these limits can in principle be reached via ptychography without expensive lenses or very high coherence. Perhaps the biggest potential benefit of ptychography will therefore be to dispose of the main costs of the electron microscope (the lenses), substituting them for a good detector and a large (but now inexpensive) computer. No other phase-retrieval technique can offer this prospect while at the same time delivering the usual imaging functionality of a microscope.

REFERENCES

Bates, R.H.T., Rodenburg, J.M. (1989). Sub-Angstrom transmission microscopy: A Fourier transform algorithm for microdiffraction plane intensity. *Inform. Ultramicrosc.* **31**, 303–308.

Berndt, H., Doll, R. (1976). Method for direct phase determination in electron microscopic diffraction patterns. *Optik* **46**, 309–332.

Berndt, H., Doll, R. (1978). Electron interferences for evaluation of phases in electron-diffraction patterns. *Optik* **51**, 93–96.

Berndt, H., Doll, R. (1983). An electron interferometer for phase determination in diffraction patterns of natural crystals. *Optik* **64**, 349–366.

Born, M., Wolf, E. (1999). *Principles of Optics*, seventh ed. Cambridge Univ. Press, Cambridge.

Burge, R.E., Fiddy, M.A., Greenaway, A.H. (1976). Phase problem. *Proc. Roy. Soc. (London) Series A* **350**, 191–212.

Chapman, J.N. (1975). Application of iterative techniques to investigation of strong phase objects in electron-microscope. 1. Test calculations. *Phil. Mag.* **32**, 527–552.

Chapman, H.N. (1996). Phase-retrieval X-ray microscopy by Wigner-distribution deconvolution. *Ultramicroscopy* **66**, 153–172.

Coene, W., Janssen, G., deBeeck, M.O., van Dyck, D. (1992). Phase retrieval through focus variation for ultraresolution in field-emission transmission electron microscopy. *Phys. Rev. Lett.* **69**, 3743–3746.

Cowley, J.M. (1995). *Diffraction Physics*. North-Holland, Amsterdam.

Cowley, J.M. (2001). Comments on ultra-high-resolution STEM. *Ultramicroscopy* **87**, 1–4.

Cowley, J.M., Winterton, J. (2001). Ultra-high-resolution electron microscopy of carbon nanotube walls. *Phys. Rev. Lett.* **87**. art. No. 016101.

Crewe, A.V., Wall, J., Welter, L.M. (1968). A high-resolution scanning transmission electron microscope. *J. Appl. Phys.* **39**, 5861–5868.

Faulkner, H.M.L., Rodenburg, J.M. (2004). Moveable aperture lensless transmission microscopy: A novel phase retrieval algorithm. *Phys. Rev. Lett.* **93**. Article No. 023903.

Faulkner, H.M.L., Rodenburg, J.M. (2005). Error tolerance of an iterative phase retrieval algorithm for moveable illumination microscopy. *Ultramicroscopy* **103**, 153–164.

Fienup, J.R. (1978). Reconstruction of an object from the modulus of its Fourier transform. *Opt. Lett.* **3**, 27–29.

Fienup, J.R. (1982). Phase retrieval algorithms – A comparison. *Appl. Opt.* **21**, 2758–2769.

Fienup, J.R. (1987). Reconstruction of a complex-valued object from the modulus of its Fourier-transform using a support constraint. *J. Opt. Soc. Am. A* **4**, 118–123.

Frank, J. (1973). Envelope of electron-microscopic transfer functions for partially coherent illumination. *Optik* **38**, 519–536.

Friedman, S.L., Rodenburg, J.M. (1992). Optical demonstration of a new principle of far-field microscopy. *J. Phys. D. Appl. Phys.* **25**, 147–154.

Gabor, D. (1948). A new microscopic principle. *Nature* **161**, 777–778.

Gabor, D. (1949). Microscopy by reconstructed wave-fronts. *Proc. Roy. Soc. (London) A* **197**, 454–487.

Gerchberg, R.W., Saxton, W.O. (1972). Practical algorithm for determination of phase from image and diffraction plane pictures. *Optik* **35**, 237–246.

Giacovazzo, C. (1999). Direct phasing in crystallography: Fundamentals and applications. In: *International Union of Crystallography Monographs on Crystallography*. In: *Oxford Science Publications*, vol. **8**. Oxford.

Gureyev, T.E., Roberts, A., Nugent, K.A. (1995). Partially coherent fields, the transport-of-intensity equation, and phase uniqueness. *J. Opt. Soc. Am. A* **12**, 1942–1946.

Hawkes, P.W., Kasper, E. (1996). *Principles of Electron Optics*. Academic Press. [see pp. 1641–1649].

Hegerl, R., Hoppe, W. (1970). Dynamische Theorie der Kristallstrukturanalyse durch Elektronenbeugung im inhomogen Primarstrahlwellenfeld. *Ber. Bunsenges. Physik. Chemie* **74**, 1148–1154.

Hegerl, R., Hoppe, W. (1972). Phase evaluation in generalized diffraction (ptychography). In: *Proceeding of the 5th European Congress on Electron Microscopy*, pp. 628–629.

Hodgkin, D.C., Kamper, J., Mackay, M., Pickworth, J., Trueblood, K.N., White, J.G. (1956). The structure of vitamin-B_{12}. *Nature* **178**, 64–66.

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 96
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

Hohenstein, M. (1992). Single side-band imaging in high-resolution electron-microscopy. *Appl. Phys. A* **54**, 485–492.

Hoppe, W. (1969a). Diffraction in inhomogeneous primary wave fields. 1. Principle of phase determination from electron diffraction interference. *Acta Crystallogr. A* **25**, 495–501.

Hoppe, W. (1969b). Diffraction in inhomogeneous primary wave fields. 3. Amplitude and phase determination for nonperiodic objects. *Acta Crystallogr. A* **25**, 508–515.

Hoppe, W. (1982). Trace structure analysis, ptychography, phase tomography. *Ultramicroscopy* **10**, 187–198.

Hoppe, W., Strube, G. (1969). Diffraction in inhomogeneous primary wave fields. 2. Optical experiments for phase determination of lattice interferences. *Acta Crystallogr. A* **25**, 502–507.

Johnston, J.A. (1989). Wigner distribution and FM radar signal design. *IEE Proc. F* **136**, 81–87.

Landauer, M.N. (1996). PhD thesis, University of Cambridge.

Lichte, H. (1986). Electron holography approaching atomic resolution. *Ultramicroscopy* **20**, 293–304.

Lin, J.A., Cowley, J.M. (1986). Reconstruction from in-line electron holograms by digital processing. *Ultramicroscopy* **19**, 179–190.

Marchesini, S. (2007). A unified evaluation of iterative projection algorithms for phase retrieval. *Rev. Sci. Instrum.* **78**. Article No. 011301.

McBride, W., O'Leary, N.L., Nugent, K.A., Allen, L.J. (2005). Astigmatic electron diffraction imaging: A novel mode for structure determination. *Acta Crystallogr. A* **61**, 321–324.

McCallum, B.C., Rodenburg, J.M. (1992). 2-Dimensional optical demonstration of Wigner phase retrieval microscopy in the STEM configuration. *Ultramicroscopy* **45**, 371–380.

McCallum, B.C., Rodenburg, J.M. (1993a). Simultaneous reconstruction of object and aperture functions from multiple far-field intensity measurements. *J. Opt. Soc. Am. A* **93**, 231–239.

McCallum, B.C., Rodenburg, J.M. (1993b). An error analysis of crystalline ptychography in the STEM mode. *Ultramicroscopy* **52**, 85–99.

Miao, J., Charalambous, P., Kirz, J., Sayre, D. (1999). Extending the methodology of X-ray crystallography to allow imaging of micrometer-sized noncrystalline specimens. *Nature* **400**, 342–344.

Michelson, A.A. (1890). On the application of interference methods to astronomical measurements. *Phil. Mag.* **30**, 1–21.

Misell, D.L. (1973). A method for the solution of the phase problem in electron microscopy. *J. Phys. D* **6**, L6–L9.

Misell, D.L. (1978). The phase problem in electron microscopy. *Adv. Opt. Electron Microsc.* **7**, 185–279.

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 97
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

Mott, N.F., Massey, E.C. (1965). *The Theory of Atomic Collisions*, third ed. Oxford Univ. Press, New York.

Nellist, P.D., Rodenburg, J.M. (1994). Beyond the conventional information limit: The relevant coherence function. *Ultramicroscopy* **54**, 61–74.

Nellist, P.D., Rodenburg, J.M. (1998). Electron ptychography I: Experimental demonstration beyond the conventional resolution limits. *Acta Crystallogr. A* **54**, 49.

Nellist, P.D., McCallum, B.C., Rodenburg, J.M. (1995). Resolution beyond the "information limit" in transmission electron microscopy. *Nature* **374**, 630.

Patterson, A.L. (1934). A Fourier series method for the determination of the components of interatomic distances in crystals. *Phys. Rev.* **46**, 372–376.

Plamann, T., Rodenburg, J.M. (1994). Double resolution imaging with infinite depth of focus in single lens scanning microscopy. *Optik* **96**, 31–36.

Plamann, T., Rodenburg, J.M. (1998). Electron ptychography II: Theory of three-dimensional scattering effects. *Acta. Crystallogr. A* **54**, 61–73.

Rodenburg, J.M. (1988). Properties of electron microdiffraction patterns from amorphous materials. *Ultramicroscopy* **25**, 329–343.

Rodenburg, J.M. (1989). The phase problem, microdiffraction and wavelength-limited resolution – A discussion. *Ultramicroscopy* **27**, 413–422.

Rodenburg, J.M. (2001). A simple model of holography and some enhanced resolution methods in electron microscopy. *Ultramicroscopy* **87**, 105–121.

Rodenburg, J.M., Bates, R.H.T. (1992). The theory of super-resolution electron microscopy via Wigner distribution deconvolution. *Phil. Trans. A* **339**, 521–553.

Rodenburg, J.M., Faulkner, H.M.L. (2004a). High resolution imaging. UK patent application PCT/GB2005/001464.

Rodenburg, J.M., Faulkner, H.M.L. (2004b). A phase retrieval algorithm for shifting illumination. *Appl. Phys. Lett.* **85**, 4795–4797.

Rodenburg, J.M., McCallum, B.C., Nellist, P.D. (1993). Experimental tests on double resolution coherent imaging via STEM. *Ultramicroscopy* **48**, 304–314.

Rodenburg, J.M., Hurst, A.C., Cullis, A.G. (2007a). Transmission microscopy without lenses for objects of unlimited size. *Ultramicroscopy* **107**, 227–231.

Rodenburg, J.M., Hurst, A.C., Cullis, A.G., Dobson, B.R., Pfeiffer, F., Bunk, O., David, C., Jefimovs, K., Johnson, I. (2007b). Hard X-ray lensless imaging of extended objects. *Phys. Rev. Lett.* **93**. Article No. 034801.

Sayre, D. (1952). Some implications of a theorem due to Shannon. *Acta. Crystallogr.* **5**, 843.

aiep v.2005/11/15 Prn:6/12/2007; 13:13 F:aiep1503.tex; VTEX/JOL p. 98
aid: 3 pii: S1076-5670(07)00003-1 docsubty: REV

Sayre, D. (1980). In: Schlenker, M., et al. (Eds.), *Image Processing and Coherence in Physics*. In: *Springer Lecture Notes in Physics*, vol. **112**. Springer, Berlin, pp. 229–235.
Scherzer, O. (1936). Some defects of electron lenses. *Z. Phys.* **101**, 593–603.
Scherzer, O. (1947). Sphärische und chromatische Korrektur von Elektronen-Linsen. *Optik* **2**, 114–132.
Scherzer, O. (1949). The theoretical resolution limit of the electron microscope. *J. Appl. Phys.* **20**, 20–29.
Spence, J.C.H. (1978). Practical phase determination of inner dynamical reflections in STEM. *Scan. Electron Microsc.* **1**, 61–68.
Spence, J.C.H., Cowley, J.M. (1978). Lattice imaging in STEM. *Optik* **50**, 129–142.
Spence, J.C.H., Howells, M., Marks, L.D., Miao, J. (2001). Lensless imaging: A workshop on new approaches to the phase problem for non-periodic objects. *Ultramicroscopy* **90**, 1–6.
Watson, J.D., Crick, F.H.C. (1953). Molecular structure of nucleic acids – A structure for deoxyribose nucleic acid. *Nature* **171**, 737–738.
Whittaker, E.T., Watson, G.N. (1950). *A Course of Modern Analysis*. Cambridge Univ. Press, London and New York.
Zuo, J.M., Vartanyants, I., Gao, M., Zhang, R., Nagahara, L.A. (2003). Atomic resolution imaging of a carbon nanotube from diffraction intensities. *Science* **300**, 1419–1421.