1.3 Image formation in electron microscopy

This section will summarize the process of image formation in the electron microscope. It will be shown how the signal conveyed by the electron beam is subject to contrast inversion artifacts, information envelope functions, and noise corruption. A well rounded understanding of the image formation process is of fundamental importance in SPA.

1.3.1 The weak phase/weak amplitude approximation

A sound understanding of image formation in the electron microscope requires an appreciation of wave-optics, lens systems and quantum mechanics. Detailed summaries have been presented previously (Angert et al., 2000; De Graef, 2003; Erickson and Klug, 1970; Frank, 1996; Zhu et al., 1997). Here the prevailing theory of image formation will be summarized in concise form.

In transmission electron microscopy the incoming electron beam is described as a plane wave $\psi(r) = e^{irz}$ where z is the direction of the electron microscope column and r is the 2D coordinate vector in orthogonal directions denoted by x and y. This plane wave is a solution of the Schrödinger equation (Tipler, 1991). In quantum mechanics the quantity $P(r) \equiv \psi(r) \bullet \psi^*(r)$ is the probability distribution function of observing the wave $\psi(r)$ at point r. The probability distribution $P(r)$ of the electron beam at the point of its interaction with the image-recording device in the back focal plane of the electron microscope objective lens is the image formed and used as the basis of all image processing in SPA.

The interaction of the incoming electron wave, denoted by $\psi_i(r)$, with the macromolecules is described mathematically as a phase shift which is proportional to the three-dimensional Coulomb potential distribution within the sample. This interaction between the beam and specimen can be understood mathematically as

$$\psi_e(r) = \psi_i(r)\exp\left(-i\pi\lambda\int U(r,z)dz\right). \tag{1.1}$$

The wave transmission function at the exit surface of the specimen described in Eq. 1.1 is denoted by $\psi_e(r)$. In Eq. 1.1 λ is the electron and U is the electrostatic potential of the specimen (Fig. 1.1A2) It is convenient to write $\phi(r) = -\pi\lambda\int U(r,z)dz$ allowing the simplifying terminology, $\psi_e(r) = \psi_i(r)\exp\left(i\phi(r)\right)$. In the well known *weak phase approximation* (WPA) the term $\exp\left(i\phi(r)\right)$ is approximated using a Taylor series expansion (Stewart, 1996) yielding

$$\exp\left(i\phi(r)\right) = 1 + i\phi(r) - \frac{1}{2}\phi(r)^2 \ldots \approx 1 + i\phi(r). \tag{1.2}$$

This truncation is based on the assumption that $\phi(r) << 1$. The influence of second (and higher) order effects is mostly negligible in most experimental situations (Erickson and Klug, 1970). Eq. 1.1 can therefore be stated more simply as

$$\psi_e(r) = \psi_i(r)\left(1 + i\phi(r)\right). \tag{1.3}$$

Interestingly, Eq. 1.3 can be interpreted as the superposition of two waves, namely the unscattered or background wave and the scattered wave induced by the interaction with the Coulomb potential of the particle. To accommodate for absorption phenomena attributed to inelastic and multiple electron scattering, scattering outside the objective aperture and other effects (Wade, 1992) an amplitude attenuation term $\mu(r)$ (which is understood to be less than or equal to zero) is modeled into Eq. 1.3 which yields the modified wave description of the electron planar wave

$$\psi_e(r) = \psi_i(r)\left(1 + i\phi(r) + \mu(r)\right). \tag{1.4}$$

The accuracy of this model of the electron beam wave structure has been experimentally confirmed in cryo-EM and negative stain conditions. (Angert et al., 2000; Toyoshima and Yonekura, 1993). We can normalize the incident (complex) plane wave $\psi_i(r)$ with no loss of generality. This allows Eq. 1.4 to be written as

$$\psi_e(r) = 1 + i\phi(r) + \mu(r). \tag{1.5}$$

This result is known as the weak phase/weak amplitude approximation (Misell, 1978). Eq. 1.5 summarizes a quantum description of wave interference in the electron microscope due to specimen interaction and amplitude attenuation. The corresponding image at the exit stage surface of the specimen is revealed by measuring the associated probability distribution which is

$$\psi_e(r)\psi_e^*(r) = 1 + \phi^2(r) + \mu^2(r) + 2\mu. \tag{1.6}$$

Note that Eq. 1.6 is a theoretical description of the intensity distribution of the electron beam image at the A3 stage of illumination in Fig. 1.4A.

1.3.2 Imaging defects associated with defocus and spherical aberration

The spherical aberration of the objective lens and the degree of defocus each introduce significant aberrations to the electron plane wave traveling through the electron microscope. The phase shift attributed to defocus and lens aberrations will be denoted $\gamma(s,\theta)$, which is a function of 2D spatial frequency polar coordinates s and θ. The specific form of the lens aberration function will be discussed below in Section 1.3.3. For now we will focus on how these aberrations induce a frequency-dependent phase shift in the Fourier transform (Fourier transforms explained in Appendix IV and Appendix V) of the exit plane wave as described in Eq. 1.5, which is stated mathematically as

$$\Psi_\gamma(s,\theta) = F\{\psi_e(r)\}\exp(i\gamma(s,\theta)), \tag{1.7}$$

$$\Rightarrow \Psi_\gamma(s,\theta) = (\delta(s,\theta) + i\Phi(s,\theta) + M(s,\theta))\exp(i\gamma(s,\theta)). \tag{1.8}$$

Here F denotes the Fourier transform and $\delta(s,\theta)$ (the Dirac delta function), $\Phi(s,\theta)$ and $M(s,\theta)$ are the Fourier transforms of 1, $\phi(r)$ and $\mu(r)$ respectively. Note that $\Phi(s,\theta)$ is known as the structure factor of the particle. Similarly, Ψ_γ is the Fourier transform of the planar electron wave traveling through the microscope. We are interested in the effects of these aberrations on the final image, i.e. we wish to evaluate the probability distribution of the exit plane wave, defined as

$$\psi_\gamma(r)\psi_\gamma^*(r), \tag{1.9}$$

where $\psi_\gamma(r)$ is the inverse Fourier transform of Ψ_γ, (i.e. the quantum description of the electron plane wave after the convoluting influences of defocus and lens aberration have been accounted for). To evaluate this expression and gain a deeper insight into the measured image we take the inverse Fourier transform (denoted F^{-1}) of Eq. 1.7 and make use of the convolution theorem to elucidate that

$$\psi_\gamma(r) = F^{-1}\{\Psi_\gamma(s,\theta)\} = \psi_e \otimes F^{-1}\{\exp(i\gamma(s,\theta))\}, \tag{1.10}$$

$$= \psi_e \otimes F^{-1}\{\cos\gamma(s,\theta)\} + i\psi_e \otimes F^{-1}\{\sin\gamma(s,\theta)\}. \tag{1.11}$$

Here the $\otimes$ symbol denotes the convolution operation. Using the description supplied by Eq. 1.11 to evaluate Eq. 1.9 and disregarding non linear terms (Misell, 1978) yields

$$\psi_\gamma(r)\psi_\gamma^*(r) = 1 + 2\phi(r) \otimes F^{-1}\{\sin\gamma(s,\theta)\} - 2\mu(r) \otimes F^{-1}\{\cos\gamma(s,\theta)\}. \tag{1.12}$$

Eq. 1.12 is a simplified description of the image formed in the back projection plane of the electron microscope, and describes the effects of the microscope aberration function. These effects are made more transparent in terms of the equivalent Fourier image, which is described as

$$F\{\psi_\gamma(r)\psi_\gamma^*(r)\} = F\{I_\gamma(r)\} = \delta(s,\theta) + 2\Phi(s,\theta)\sin\gamma(s,\theta) - 2M(s,\theta)\cos\gamma(s,\theta) \qquad (1.13)$$

.

In Eq. 1.13 $I_\gamma(r)$ is introduced to simplify terminology and represents the image measured at the appropriate illumination stage in the electron microscope (A4, Fig. 1.4A). Eq. 1.13 can be simplified by removing the $\delta(s,\theta)$ term, which only contributes for $[s, \theta] = [0,0]$ (i.e. the DC component of the Fourier transform) and represents the Fourier transform of the (normalized) unscattered component of the incoming wave $\psi_i(\boldsymbol{r})$. Also, the constant factor of 2 in Eq. 1.13 can be disregarded since, in practice, scaling factors are arbitrary. On the basis of these assumptions, Eq. 1.13 simplifies to

$$F\{I_\gamma(r)\} = \Phi(s,\theta)\sin\gamma(s,\theta) - M(s,\theta)\cos\gamma(s,\theta). \qquad (1.14)$$

By introducing the term $W(s,\theta) = M(s,\theta) / \Phi(s,\theta)$ and the commonly made assumption that W(s) is constant for all spatial frequencies, i.e. that $W(s,\theta) \approx W$ (Angert et al., 2000; Zhu et al., 1997) we arrive at a simplified form of Eq. 1.14,

$$F\{I_\gamma(r)\} = \Phi(s,\theta)\left(\sin\gamma(s,\theta) - W\cos\gamma(s,\theta)\right) = \Phi(s,\theta)C(s,\theta). \qquad (1.15)$$

The factor W is referred to as the amplitude contrast ratio which is of the order of 5-7% for biological samples imaged in cryogenic conditions (Angert et al., 2000; Toyoshima and Yonekura, 1993; Wade, 1992) and is of the order 35% for negative stain specimens (Wade, 1992). $C(s,\underline{\theta})$ is known as the CTF of the electron microscope. Eq. 1.15 is therefore a summary of the weak phase/weak amplitude approximation in its most concise form that is representative of the state of the electron beam at the A4 stage of

illumination in Fig. 1.5A. The equivalent intensity distribution or image at this stage of illumination can be interpreted in real space as

$$I_\gamma(r) = \phi(r) \otimes c(r)\,, \tag{1.16}$$

where $c(r)$ is the inverse Fourier transform of the $C(s,\underline{\theta})$, which is the known as the point spread function of the CTF in the electron microscope. Importantly, the framework presented in Eq. 1.15 and Eq. 1.16 does not destroy the linear relationship between the Fourier transform of the image and the projected potential of the specimen $\Phi(s,\underline{\theta})$, which implies that the signal of the specimen is recoverable. This is a fundamental principle of SPA.

1.3.3 The contrast transfer function

In the absence of lens astigmatism the CTF, denoted $C(s,\theta)$ in Eq. 1.15, is isotropic. Thus the CTF can be expressed more simply as

$$C(s,\theta) = C(s) = \sin\gamma(s) - W\cos\gamma(s)\,. \tag{1.17}$$

That is, given that astigmatism has been studiously avoided by careful calibration of the lenses in the electron microscope, the effects of the CTF are solely dependent on the Fourier radial coordinate. This simplification will suffice for the discussion presented here, but it should be noted that as higher resolution reconstructions are sought in SPA the influence of lens astigmatism may become increasingly important, particularly if the degree of allowable astigmatism is small. For a more detailed description of the CTF in the presence of lens astigmatism see (De Graef, 2003; Frank, 1996). It can be shown using optics, geometry and diffraction theory (De Graef, 2003) that the phase shift attributed to defocus and lens aberrations is given by

$$\gamma(s) = 2\pi\left(\frac{C_s\lambda^3 s^4}{4} + \frac{\Delta Z\lambda s^2}{2}\right), \tag{1.18}$$

where C_s is the spherical aberration constant of the objective lens in the electron microscope (usually of the order ~10^{-3} m), ΔZ is the defocus of the microscope (usually of the order ~10^{-6} m) and λ is the relativistic wavelength defined as

$$\lambda = \frac{1266.39}{\sqrt{E + 9.7845\times10^{-5}E^2}}. \tag{1.19}$$

In Eq. 1.19 E is used to describe the operating voltage of the electron microscope (in volts) and the resulting wavelength is given in picometers. The parameter of Eq. 1.18 most varied by the human operator is the defocus. This is because the spherical aberration is a physical parameter of the microscope that cannot be changed and the operating voltage (which changes the electron wavelength) is usually kept constant. As the defocus of the microscope is increased, the oscillations of the CTF become more rapid as shown in Fig. 1.6A, which depicts the CTF incurred for two commonly used defocus values Fig.1.x panels C and D). The effect of the CTF is literally to scale Fourier components by a number on the interval [-1,1]. Scaling by a negative number in complex space is equivalent to phase shifting by 180°, which is referred to as the phase-flipping effect. As the inverse Fourier transform reveals that a real function is a sum of (cosine) sinusoidal signals, phase flipping can be interpreted as real space contrast inversion, and can be ultimately understood as the inversion of cosine waves. Hence the CTF induces both contrast inversion and amplitude attenuation artifacts.

The amplitude attenuation constant W (Eq. 1.17) can be understood as a phase shift of the CTF and examples are shown in Fig. 1.6B for different amplitude constants. Note that a non-zero, positive amplitude contrast value will cause phase flipping in the lowest frequency domain. It is therefore imperative in practice that the amplitude contrast be known or approximated. The CTF is one of the principal causes of Coulomb projection

corruption in the electron microscope and occurs schematically at the A4 stage, as indicated in Fig. 1.4A.

The amplitude attenuation constant *W* (Eq. 1.17) can be understood as a phase shift of the

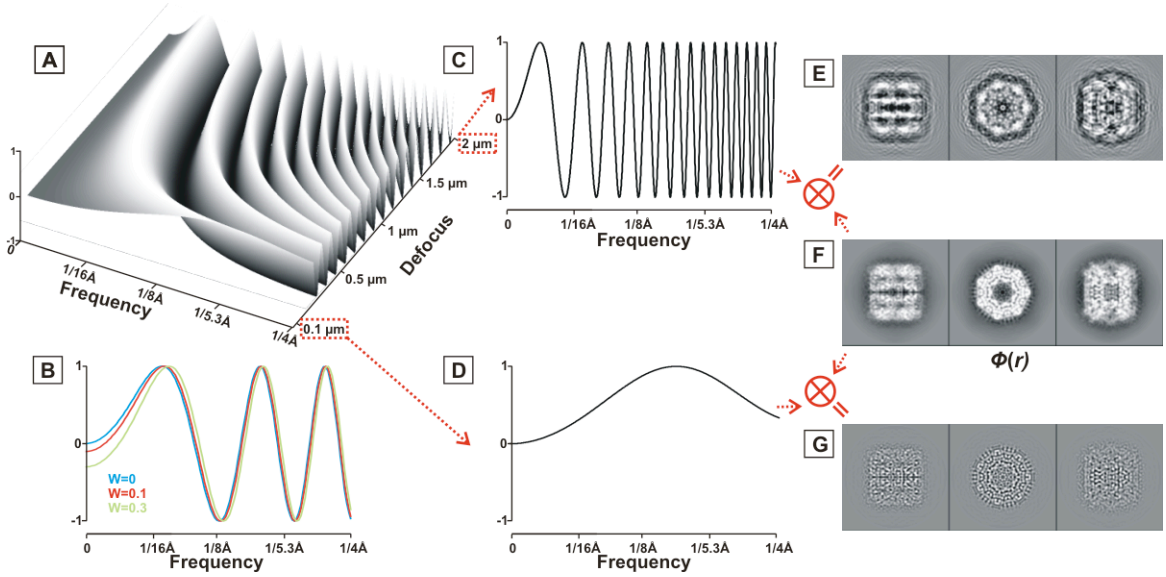

Figure 1. 1. The influence of the CTF on image formation: A. The CTF plotted in surface representation to show its dependence on the defocus. In this figure the electron wavelength corresponds to an operating voltage of 200 kV and a spherical aberration of 2.2 mm is used. **B**. Increasing the amplitude attenuation term (*W*) essentially shifts the CTF and incurs additional phase shifting in the low frequency range. **C**. The CTF profile resulting from a defocus setting of 2 µm. **D**. The CTF profile resulting from a defocus setting of 0.1 µm. **E**. The result of applying the CTF depicted in C to the pure projection data in F. **F**. Pure projection data of the GroEL model, generated from the published 6 Å structure (Ludtke et al., 2004). **G**. The result of applying the CTF depicted in D to the pure projection data in F.

CTF and examples are shown in Fig. 1.6B for different amplitude constants. Note that a non-zero, positive amplitude contrast value will cause phase flipping in the lowest frequency domain. It is therefore imperative in practice that the amplitude contrast be known or approximated. The CTF is one of the principal causes of Coulomb projection corruption in the electron microscope and occurs schematically at the A4 stage, as indicated in Fig. 1.4A.

Operating an electron microscope with a spherical aberration of 2.0 mm at an operating voltage of 200 kV and a defocus of 2 µm yields the CTF profile shown in Fig. 1.6C. Changing the defocus to 0.1 µm yields the CTF profile in Fig. 1.6D. The main difference between these two profiles is that a defocus of 2 µm incurs many more CTF oscillations,

but also has higher low frequency amplitude values which assist in particle detection. Indeed, the nature of the CTF dictates that its oscillation become more compressed towards the origin as the defocus is increased (Fig. 1.6A). This has important implications. For example, while the CTF has remained completely positive for the defocus setting of 0.1 μm, the transfer of information at low spatial frequencies is highly retarded. This is observed directly in Fig. 1.6D and is shown by example in Fig. 1.6G, which reveals the convolution of the pure projection data in Fig. 1.6F with the frequency-scaling profile in Fig. 1.6D. In Fig. 1.6G observe that mostly high spatial-frequency information remains and that low frequency information is largely unobservable. Note also that the differences between the raw (Fig. 1.6F) and the convoluted image (Fig. 1.6G) are attributed solely to amplitude scaling effects, as described in Fig. 1.6D (no phase flipping occurs). It can therefore be understood that, on its own, the amplitude corruption aspect of the CTF produces significant mass displacements. As an example of the consequence of using a greater defocus, the convolution of the 2 μm defocus profile (Fig. 1.6C) with the data in Fig. 1.6F is shown in Fig. 1.6E. While it can be readily observed that low frequency spatial information is more clearly preserved for the higher defocus setting (Fig. 1.6E), it is also apparent that the signal has undergone significant corruption (Fig. 1.6E), which is related to both CTF phase flipping and CTF amplitude attenuation (Fig. 1.6C) Even though a greater amount of defocus incurs more CTF oscillations, in practice this can be the preferred option because stronger transmission in the low frequency domain makes particle selection easier. Note also in Fig. 1.6B,C & D that the CTF intersects the frequency axis implying zero information transfer. This causes some frequency information to be completely removed from the transmitted signal. To accommodate this in practice it is imperative that images are recorded at varying degrees of defocus.

In early attempts to reconstruct single particles, contrast inversion artifacts were accounted for by low-pass filtering the data to remove all frequency-flipped information beyond the first zero crossing of the CTF. This approach limited the resolvability of the structure to the realm of 20 Å. In modern SPA, the phase flipping nature of the CTF is corrected by the simple operation of multiplying the affected regions in Fourier space by

negative one. This is equivalent to undoing the contrast inversion induced by the CTF and is referred to as phase CTF correction. To achieve this, the power spectrums of the raw micrographs (or cropped, averaged subfiles) are calculated and the CTF zero crossing estimated to determine the parameters of Eq. 1.15. While there has been some progress toward the automation of CTF parameter determination (Huang et al., 2003; Mallick et al., 2005; Mindell and Grigorieff, 2003; Sander et al., 2003) it is still common to perform manual fitting. In addition to phase correction, compensation for the amplitude attenuation of the CTF can also be made (amplitude CTF correction). This usually occurs in conjunction with a Wiener filtering operations but depends on the software package being used (Grigorieff, 2007; Ludtke et al., 1999; Zhu et al., 1997).

1.3.4 Imaging defects attributed to the microscope envelope functions

In this section the known envelope functions of the electron microscope will be summarized, the combined effects of which can be reasonably approximated as a single envelope function, denoted $e(r)$ in real space and $E(s,\theta)$ in Fourier space. These imperfections of the electron microscope imaging system convolute the signal conveyed in the electron beam and blur spatial information.

Envelope effects can be understood in Fourier space as a modified form of Eq. 1.15 which is summarized as

$$F\{I_{\gamma,e}(r)\} = \Phi(s,\theta)C(s,\theta)E(s,\theta)\,. \tag{1.20}$$

Here $I_{\gamma,e}(r)$ denotes the intensity distribution of the electron beam, as influenced by the contrast transfer and envelope functions of the electron microscope (recall $\Phi(s,\theta)$ is the structure factor, $C(s,\theta)$ is the CTF and $E(s,\theta)$ is the newly modeled envelope function). Three separate envelope functions of the electron microscope have been theorized and experimentally confirmed. These will be summarized here based on previously presented work (De Graef, 2003; Frank, 1996; Zhu et al., 1997). The first envelope function relates

to the partial coherence of the electron beam (Frank, 1973) and is defined formally in Fourier polar coordinates as

$$E_{pc}(s,\theta) = E_{pc}(s) = \exp\left(-\pi^2 q_o^2 \left(C_s s^3 \lambda^3 - \Delta Z s \lambda\right)^2\right) \approx \exp\left(-\pi^2 q_o^2 \Delta Z^2 s^2 \lambda^2\right) \qquad (1.21)$$

.

In Eq. 1.21 s and λ are as previously defined and the constant q_o is related to the source size associated with the illumination system and is generally of the order 1 x 10^8 $m^{1/2}$ (De Graef, 2003; Frank, 1976). The simplifying approximation made in Eq. 1.21 is valid when $s\lambda << 1$, which is true for operating voltages and sampling rates typical of current cryo-EM (100-300 kV). The lack of dependence on the Fourier coordinate angle θ dictates that this envelope function is isotropic in Fourier space. Eq. 1.21 reveals the dependence of E_{pc} on both the operating voltage (which affects the electron wavelength λ) and the defocus (ΔZ) of the instrument. In particular, Eq. 1.21 predicts that the fall-off of information transfer in the high frequency domain will be less pronounced if the defocus is decreased or operating voltage is increased, as depicted in Fig. 1.7A and Fig. 1.7B respectively. This demonstrates why higher operating voltages are often preferred in practice, because the resulting boost in information transfer in the high frequency domain improves the chances of recovering high resolution data in the final 3D reconstruction. Even though Fig. 1.7B predicts that increased defocus settings reduce high frequency coherence in the electron beam, in practice higher defocus settings are often preferred because they ensure a strong transmission of low frequency information. This is based on the dependence of the CTF on the defocus as depicted in Fig. 1.6.

In addition to the partial coherence of the beam, the information content of the signal is affected by an energy spread envelope function, defined in polar coordinates as

$$E_{es}(s,\theta) = \exp\left(-\pi^2 \delta z^2 s^4 \lambda^2\right) \qquad (1.22)$$

In this equation δz is a parameter relating to the defocus variation associated with the energy spread of the electron microscope and is independent of the actual defocus used

(De Graef, 2003; Frank, 1976; Zhu et al., 1997). Similar to the partial coherence envelope function as described in Eq. 1.21, the energy-spread function is strongly dependent of the operation voltage and will exhibit characteristics similar to those shown in Fig. 1.7A.

Finally, it is possible to summarize the effects of specimen drift, vibration, and multiple inelastic-elastic scattering events in a single envelope function (Kenney et al., 1992) defined as

$$E_{dv}(s,\theta) = \exp\left(-\pi^2 u^2 s^2\right). \tag{1.23}$$

Here u is a constant relating to the vector amplitude of the sample vibration and drift and is generally of the order $1\text{x}10^{-10}$ m (i.e. ~ 1 Å) . The convoluting influence of sample drift and/or vibration is independent of the operating voltage. Instead, it reflects the stability of the specimen in the microscope. In addition, drift can occur in specific directions in which case the isotropic assumption in Eq. 1.23 fails. Although theory can accommodate for direction specific drift in a straight forward fashion, the simplifications embedded in Eq. 1.23 will suffice for the discussion presented here.

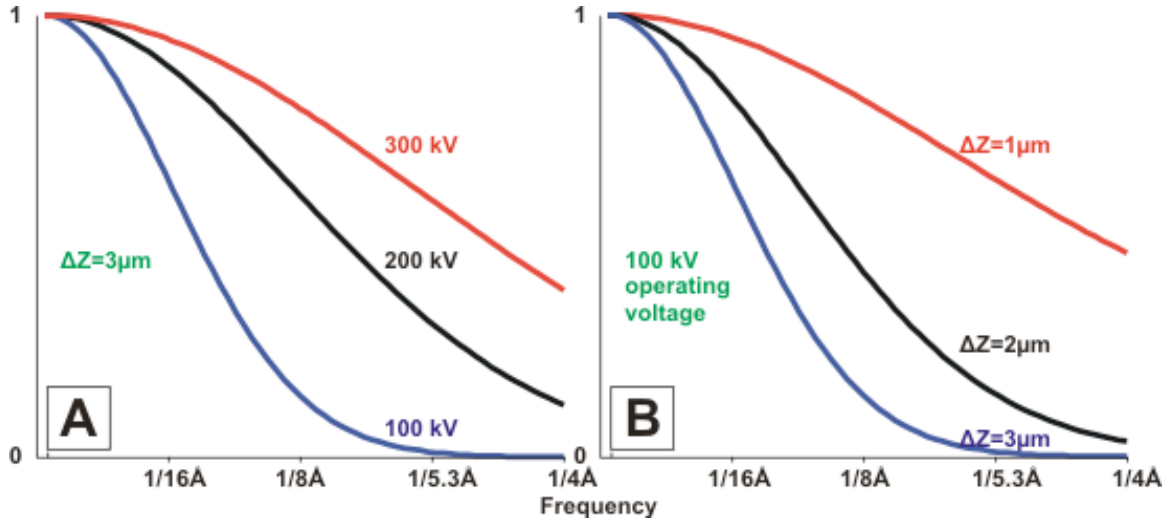

Figure 1. 2. The partial coherence envelope function. This illustrates the behavior of the partial coherence envelope function of the microscope as the defocus (A) and operating voltage (B) changes. For these figures the constant q_o was specified as $5\text{x}10^7$ and the full form of the partial coherence envelope function was used, as defined in Eq. 1.21.

In Fourier space the effects of the three envelope functions mentioned above will be cumulative. Thus we can gain an expression for the all encompassing envelope function E_{total} describing the experimental conditions in the electron microscope by evaluating Fourier products, i.e.,

$$E_{total}(s,\theta) = E_{pc}E_{ex}E_{dv} \approx \exp\left(-\pi^2 q_o^2 \Delta Z^2 s^2 \lambda^2 - \pi^2 \delta z^2 s^4 \lambda^2 - \pi^2 u^2 s^2\right)$$
$$\approx \exp\left(-\pi^2\left(q_o^2 \Delta Z^2 \lambda^2 - u^2\right)s^2\right) \approx \exp\left(-Bs^2\right). \quad (1.24)$$

In Eq. 1.24 the constant value $\pi^2\left(q_o^2 \Delta Z^2 \lambda^2 - u^2\right)$ has been replaced by B (which has the simple units of Å^2) and the assumption that λ « 1 has been used to simplify the expression. Importantly, Eq. 1.24 reveals that, in the absence of drift and/or specimen vibrations, E_{total} is dominated by the partial coherence envelope function (Eq. 1.21) and implies that operating voltage and defocus will influence the image-recording process precisely as shown in Fig. 1.7. In addition, Eq. 1.24 demonstrates that the combined effects of the individual envelope functions can be adequately described as a single envelope having the functional form of a Gaussian (Huang and Penczek, 2004; Ludtke et al., 1999).

In practice, it is difficult to characterize the parameters of all of the envelope functions contributing in a given cryo-EM experiment and therefore it is common to make use of simplifying approximations as in Eq. 1.24, where the attenuation of high resolution information transfer can be described as a single numerical B-factor, which is generally in the range of 50-400 A^2 (Ludtke et al., 2001; Saad et al., 2001). This simplified approach also provides a convenient means of comparing SPA data derived from different electron microscopes. The application of several envelope functions using two different B-factors is shown Fig. 1.8, in combination with a CTF corresponding to a 1 μm defocus. The projection images in Fig. 1.6 reflect the state of the transmitted projection data in varying experimental conditions, as described explicitly in Eq. 1.20, and show how an increased B-factor corresponds to an increase in the overall blurring. It is therefore critically important that the envelope functions of a given electron microscope

be studied and described accurately, as the exponential decay of information transfer can significantly reduce the expected resolution in the final 3D reconstructions, as will be shown in sections to come. In reference to Fig. 1.4, the combined effects of the CTF and envelope functions as described in Eq. 1.20 can be conceptually understood to describe the state of the projection data at the A4 stage of illumination.

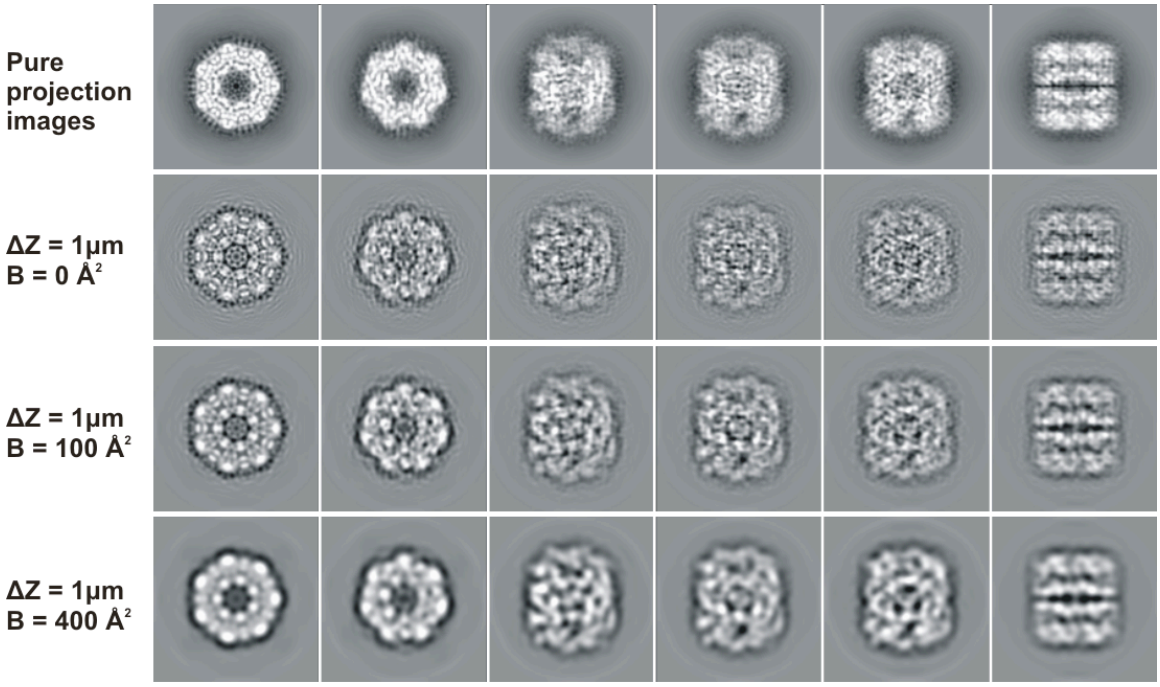

Figure 1. 3. Illustration of the combined effects of CTF and envelope functions. The top row depicts pure projection of the GroEL structure published at 6 Å (Ludtke et al., 2004) . The second row depicts how these projections change when the CTF of the microscope, imaged at 1 μm defocus, is applied. The third row depicts how the CTF-corrupted images in the second row are in turn blurred by the envelope functions when the experimental *B*-factor is 100. The last row is equivalent to the third row except the *B*-factor is 400, which causes greater blurring.

At this point it should also be noted that the modulation transfer functions (MTFs) of scanning devices (Roseman and Neumann, 2003; Typke et al., 2005), electron microscopy film (Zhu et al., 1997) and CCD cameras (Booth et al., 2004; Sander et al., 2005) have been well characterized and are known to cumulatively blur SPA data in a manner identical to the nature of envelope functions. However, in the context of the microscope envelope functions their influence is comparatively modest. For the purposes of most SPA software packages the effects of the cumulative MTFs of the recording devices are simply included in one global, modeled form of the overall experimental

envelope function, i.e. as embodied in Eq. 1.24 and described by heuristic parameters such as the *B*-factor.

1.3.5 The influence of noise

Noise in the electron microscope is attributed to random statistical variation in the number of incident electrons on the image recording medium, inherent noise related to the use of film which has non-uniform grain sizes, shot noise, and quantum noise related to elastic and inelastic scattering (Brink and Chiu, 1994; Downing and Hendrickson, 1999; Huang et al., 2003; Zhu et al., 1997). The contribution of noise in a given projection image obtained in an electron microscope is difficult to directly quantify due to its random, poorly characterized nature (Pantelic et al., 2006). In real space the noise in an individual pixel is approximated well by a (random) Gaussian distribution. In frequency space the noise profile, as revealed by the power spectrum, is much stronger at low spatial frequencies and diminishes as the frequency is increased. Therefore, in Fourier space the amplitude of the noise component is related to the Fourier radial coordinate and will presumably be evenly distributed about the mean radial noise value (in polar coordinate representation). Due to the random nature of noise in real space, the Fourier noise component will have completely random phase. Thus in real space and in Fourier space, with sufficient averaging the influence of the noise tends towards zero.

An important assumption made in image formation in electron microscopy is that the interference induced by the specimen by its interaction with the electron beam does not incur any directly related noise events as it travels through the electron microscope column. That is, noise recorded in the final image can be treated independently from the image formed by the weak phase/weak amplitude approximation (i.e., it is an additive superposition). This is embodied as a modification of the Fourier image described in Eq. 1.20 which now becomes

$$F\left\{I_{\gamma,e,n}(r)\right\} = \Phi(s,\theta)C(s,\theta)E(s,\theta) + N(s,\theta)\,, \tag{1.25}$$

where $N(s,\theta)$ is the newly modeled noise function and $I_{\gamma,e,n}(r)$ is the description of the measured image formed as a result of the interaction of electrons with the specimen, which is convoluted by the CTF and envelope functions and influenced (now) by the additive background noise. Random pixel fluctuations attributed to noise are independent of the signal of the specimen and therefore contribute incoherently to the measured image. Due to its inherent variation, it is simpler to characterize the noise function in terms of its power spectrum profile (Section 1.4.1), which is estimated using polar integrals (rotational averaging) and denoted $N^2(s)$. The functional form employed by many groups (Huang et al., 2003; Ludtke et al., 1999; Mallick et al., 2005) contains four heuristic parameters and is of the form similar to, or identical to,

$$N^2(s) = n_1 \exp\left(n_2 s + n_3 s^2 + n_4 \sqrt{s}\right). \tag{1.26}$$

This empirical description of the noise profile has been robust enough to approximate the noise contribution for most data sets in general (Ludtke et al., 1999; Mallick et al., 2005; Zhu et al., 1997). In practice, the parameters ($n_{1\text{-}4}$) of this (or a similar) equations are approximated interactively using graphical user interfaces or by the application of some automated algorithm (Ludtke et al., 1999; Mallick et al., 2005; Mindell and Grigorieff, 2003; Saad et al., 2001; Sander et al., 2003; Zhu et al., 1997). As a demonstration of the accuracy of the model described in Eq. 1.25, a noise profile typical of cryo-EM is applied to the projection data, corrupted by CTF and envelope artifacts as in the left term in Eq. 1.25 (top row of Fig. 1.9), the results of which are shown in the second row (Fig. 1.9). The images shown in the middle row of Fig. 1.9 therefore summarize the image formation process as described in Eq. 1.26 and reflect the image measured at the A5 (final) stage in Fig. 1.4. Also shown are real cryo-EM data (bottom row, Fig. 1.9) for qualitative comparison with modeled data (middle row). In essence, details of the original projection image are entirely obscured by the additive noise component, making the direct interpretation of structural information from cryo-EM data practically impossible. This motivates the subsequent use of averaging procedures.

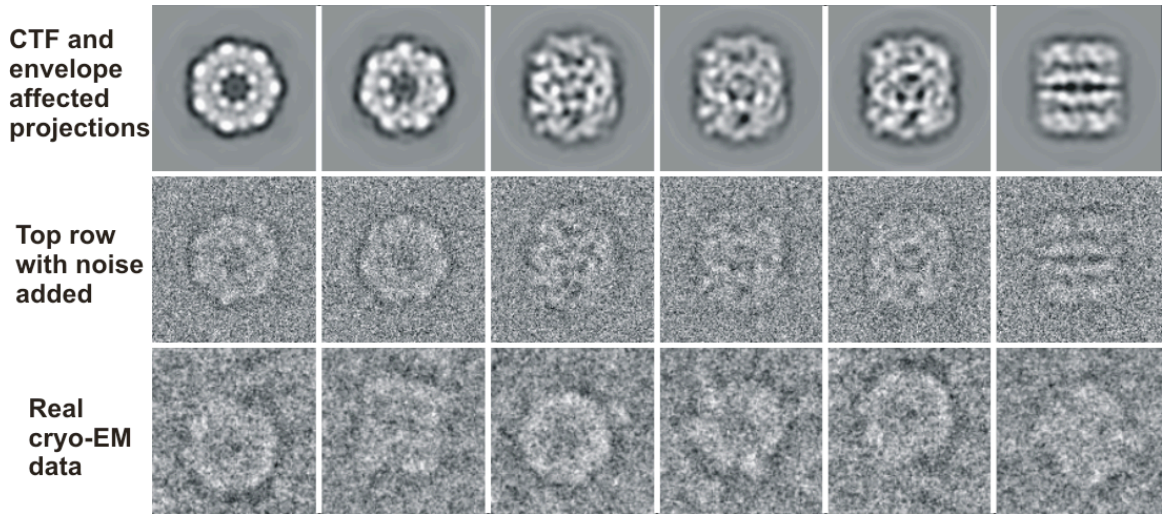

Figure 1. 4. Noise in the electron microscope. This figure shows how noise is modeled in the image formation process. Noise in the electron microscope is understood to contribute to the final image (middle row) independently of the signal which is corrupted by the CTF and envelope functions (top row), Hence random noise is literally added to the projection data in the top row, based using a typical noise profile of an electron microscope as described in Eq. 1.26, as shown in the middle row. The bottom row shows a selection of randomly selected (real) GroEL particles, cropped out from data distributed by the Scripps Research Institute (http://ami.scripps.edu/prtl_data/) and shows how models of image formation agree well with real experimental data.

Because SPA is most clearly understood in Fourier space, it is difficult to justify the use of denoising procedures that have no direct interpretation in Fourier space. As an example, one of the simplest denoising techniques, namely averaging, can be understood directly in Fourier space. Let I^v denote the vth projection image in a group of n particles classified based on similarity, i.e. in a class. Proceeding on the assumption that the particles in the class are translationally and rotationally aligned, we generate the Fourier class average $F\{CA\}$ by evaluating

$$F\{CA\} = \frac{1}{n}\sum_n F\{I^v_{\gamma,e,n}(r)\} = \frac{\Phi(s,\theta)}{n}\sum_n C^v(s,\theta)E^v(s,\theta) + \frac{1}{n}\sum_n N^v(s,\theta). \qquad (1.27)$$

Here the signal of the particle ($\Phi(s,\theta)$) in the orientation of the class average is identical. Because the phase of the noise is random and its amplitude is evenly distributed, with sufficient averaging the right term becomes zero, leaving the signal-boosted Fourier class average

$$F\{CA\} \approx \frac{\Phi(s,\theta)}{n} \sum_{n} C^{v}(s,\theta) E^{v}(s,\theta) . \tag{1.28}$$

This demonstrates that basic averaging can be understood to cancel the effects of noise in Fourier space, thereby facilitating the recovery of the true signal, $\Phi(s,\theta)$. The convoluting influence of variable CTF and envelope functions as seen in Eq. 1.28 must be accounted for by other means, which will be elaborated upon later.

This concludes the theory of image formation in electron microscopy. The most significant result presented in this section has been Eq. 1.25, which is the most fundamental description of SPA data, and will be the basis for various image processing tools introduced in the following sections.

1.4 Important image processing concepts in single particle analysis

1.4.1 The rotationally averaged power spectrum and the contrast transfer, envelope, and noise functions

Generating the rotationally averaged power spectrum is a widely used technique that facilitates estimation of parameters for the CTF (C(s)), the envelope function (E(s)) and the noise component ($N(s,\theta)$) of the image as defined above in Eq. 1.25 above. To proceed we first introduce the simplifying notation $M(s,\theta) = F\{I_{\gamma,e,n}(r)\}$ to denote the Fourier transform of the measured image on the objective lens back focal plane of the electron microscope (Fig. 1.4A5). The rotationally averaged power spectrum of electron microscopy image data, denoted $M^2(s)$, is formulated as a rotational integral and defined as

$$M^2(s) = \frac{1}{2\pi} \int_0^{2\pi} \left| M(s,\theta) \right|^2 d\theta , \tag{1.29}$$

$$= \frac{1}{2\pi}\int_0^{2\pi} \left| C(s)E(s)\Phi(s,\theta) + N(s,\theta) \right|^2 d\theta \,. \tag{1.30}$$

Here the |.| notation denotes the amplitude of the complex number, which is obtained by taking the square root of the product of the complex number and its conjugate. The amplitude is thus a real (non complex) number. The squared amplitude is by definition the power of the Fourier component. To simplify Eq. 1.30 we first expand the integral to write

$$M^2(s) = \frac{1}{2\pi}\int_0^{2\pi} C^2(s)E^2(s)\left|\Phi(s,\theta)\right|^2 + \left|N(s,\theta)\right|^2 + C(s)E(s)\left|\Phi(s,\theta)N^*(s,\theta) + \Phi^*(s,\theta)N(s,\theta)\right|^2 d\theta \tag{1.31}$$

Here the $*$ symbol denotes complex conjugate. It can be assumed that the complex vector cross terms between the noise component ($N(s,\theta)$) and the (power) profile of the particle structure factor $\Phi(s,\theta)$ go to zero with sufficient averaging. This is based on the assumption that noise in the electron microscope contributes independently of true signal, or in other words, is incoherent with respect to the beam's interaction with the specimen. Thus Eq. 1.29 becomes

$$M^2(s) = \frac{1}{2\pi}\int_0^{2\pi} C^2(s)E^2(s)\left|\Phi(s,\theta)\right|^2 + \left|N(s,\theta)\right|^2 d\theta \,, \tag{1.32}$$

or more simply,

$$M^2(s) = C^2(s)E^2(s)\Phi^2(s) + N^2(s) \,. \tag{1.33}$$

Eq. 1.33 makes use of the simplifying notation $\Phi^2(s) = \int_0^{2\pi} \left|\Phi(s,\theta)\right|^2 d\theta$ which describes the (power) profile of the particle structure factor ($\Phi(s,\theta)$) and $N^2(s) = \int_0^{2\pi} \left|N(s,\theta)\right|^2 d\theta$

which is the rotational of average of the noise power spectrum. The contribution of the noise is difficult to characterize and is modeled in practice using functions similar to that presented in Eq. 1.26. Also, the arbitrary scaling factor $1/2\pi$ has been removed for simplicity.

Rotationally averaged power spectra are utilized extensively by procedures designed to facilitate the estimation of the parameters of the C(s), E(s) and $N^2(s)$ functions (Mallick et al., 2005; Mindell and Grigorieff, 2003; Zhu et al., 1997), regardless of whether they are manual or automated. However, on their own, the power spectra of individual (or cropped portions of) images generally contain insufficient information to elucidate any useful information. This is demonstrated in Fig. 1.10A&B, which depict rotationally averaged power spectra generated from the inset small boxed images of GroEL. Both of the images in Fig. 1.10A&B are taken from the same micrograph and therefore, in the absence of specimen tilt and variable ice thickness, may be assumed to be imaged under identical conditions. Therefore, both power spectra (Fig. 1.10A&B) should reveal the same functional form of the CTF, envelope and noise function. However due to the inherent lack of detail in these spectra, the estimation of the associated imaging parameters is practically impossible and alternative methods are required.

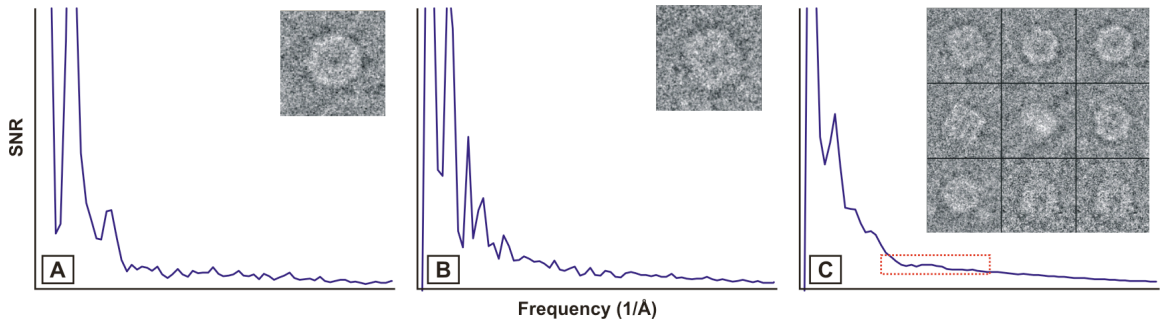

Figure 1. 5. The rotationally averaged power spectrum. A & B. Rotationally averaged power spectrum of inset images. **C**. Averaged rotationally averaged power of a large (100 particle) data set (example images shown inset. The dashed red box shows the region of the averaged power spectrum where the first CTF maximum can be observed.

It is well known that coherently averaging the power spectra ($M^2(s)$) generated from many individual subregions of the same micrograph (i.e., many boxed particles or consecutively boxed image subregions) will simplify the process of obtaining estimates

of the parameters governing the $C(s)$, $E(s)$ and $N^2(s)$ functions. The averaging of many different profiles, resulting in the average profile $M_{ave}^2(s)$, can be described mathematical as

$$M_{ave}^2(s) = \frac{1}{2\pi n} \sum_{i=1}^{n} \int_0^{2\pi} \left| M_i(s,\theta) \right|^2 d\theta \,. \tag{1.34}$$

Here $M_i(s,\theta)$ is understood to mean i^{th} subfiled or cropped portion of an electron micrograph (M). This is precisely the approach used in popular SPA software such as FREALIGN (Grigorieff, 2007) , SPIDER (Frank et al., 1996) and EMAN (Ludtke et al., 1999), which utilize Eq. 1.34 for the purposes of generating accurate power spectra that facilitate the determination of parameters governing $C(s)$, $E(s)$ and $N^2(s)$. Automated algorithms that achieve these ends are often based on the fact that the zero crossings of the CTF in the averaged rotational power spectrum must coincide with the background noise profile (Huang et al., 2003; Mallick et al., 2005; Mindell and Grigorieff, 2003). An example of an averaged rotational power spectrum is shown in Fig. 1.10C, which is generated from the average of the power spectra of approximately 100 particle images taken from the same micrograph used in Fig. 1.10A&B. Data shown in Fig. 1.10C is typical of cryo-EM and in the context of Fig. 1.10A&B, where the form of the CTF is clearly ambiguous, which demonstrates the utility of rotational power spectrum averaging concept.

In practice, the approach used for the determination of these parameters is strongly dependent the SPA software being used. Furthermore, the parameters determined are generally not transferable between separate software packages. This is a consequence of the unique SPA reconstruction approaches embodied in the competing software packages currently used, but which may be overcome in the near future if common conventions are adopted (Heymann et al., 2005) and/or unifying software projects are initiated (Hohn et al., 2007).

1.4.2 The absolute signal to noise ratio and its effect on projected resolution

In addition to providing the basis for CTF, envelope and noise parameter determination, the power spectrum provides a means to generate the absolute SNR metric, which in turn is used to estimate the maximum recoverable resolution from a given particle data set. This directly influences the resolution of the final 3D reconstruction and provides a guide for the user to estimate maximal attainable resolution. The SNR is a function of spatial frequency, defined in terms of the rotationally averaged power spectrum and can described as

$$SNR(s) = \frac{C^2(s)E^2(s)\Phi^2(s)}{N^2(s)} = \frac{M^2(s) - N^2(s)}{N^2(s)}. \tag{1.35}$$

In accordance with the concept of SNR, the definition supplied in Eq. 1.35 reflects the ratio of the strength of the signal (numerator) to the strength of the noise (denominator). This expression can be evaluated in two ways: If the parameters of the functions C, E, N^2 and Φ^2 are known or approximated; or if the noise profile N^2 has been modeled it can be subtracted directly from the measured power spectrum of the data (M^2) (right side, Eq. 1.35) to obtain a direct estimate of the power of the signal. This latter approach is the simplest and most direct method for estimated the SNR. In either case the parameters of C, E, N^2, or M^2 are approximated using power spectrum techniques as described above. The unknown structure factor Φ can be difficult to approximate. The process of doing so is generally based on the fact that the power spectrum is dominated in the low frequency domain by the structure factor itself, as revealed in Fig. 1.10 and the fact that the approximating structure factor profile $\Phi^2_{approx}(s)$ can be separated from the power spectrum defined in Eq. 1.33 using simple algebra, as expressed in Eq. 1.36.

$$\Phi^2_{approx}(s) = \frac{M^2(s) - N^2(s)}{C^2(s)E^2(s)}. \tag{1.36}$$

It should be understood that any inaccuracies in defining the CTF, envelope and noise functions will reduce the accuracy of the approximated structure factor. To improve accuracy, the structure factor profile of the macromolecule being studied should be extrapolated and averaged using at least several micrographs imaged at varying degrees of defocus, based on Eq. 1.36. This technique will yield a structure factor estimate sufficiently accurate on the Fourier interval 1/0 Å (the Fourier origin which has no frequency) to 1/25 Å. Because the high resolution structure information is difficult to resolve from noise in the cryo-EM data, the extrapolated structure factor profile generated is generally merged with the high resolution component of a previously solved protein structure at the nominal low resolution cut off (i.e. 1/25 Å). This is based on the loose assumption that, when comparing two independent structure factors from two different molecules, most variation will be observed in the low frequency domain. The merged structure factor can then be used as the basis for the SNR estimation, the Wiener filtering operation (which requires a structure factor estimate, see above), and has been used in the reconstructions of various models at subnanometer resolutions using SPA (Booth et al., 2004; Ludtke et al., 2005; Menetret et al., 2005).

Similar to the approach espoused in Section 1.4.1, an improved estimate of the SNR in a given micrograph can be arrived at using the incoherently averaged power spectra of many cropped particles, as described in Eq. 1.34. This leads to an adapted form of the SNR given by the formula

$$SNR(s) = \frac{M_{ave}^2(s) - N^2(s)}{N^2(s)}, \tag{1.37}$$

where M_{ave}^2 is determined according to Eq. 1.34. This is a more effective means of determining the SNR of a data set as the averaging procedure reduces the error of the estimate. As an example, the power spectrum of the data shown in Fig. 1.10C is shown in Fig. 1.11A, but with the noise profile $N^2(s)$ modeled (red line). The corresponding SNR profile is shown in Fig. 1.11B which is calculated using Eq. 1.37, and shows how the

SNR profile is relatively strong at low frequencies but rapidly approaches values near to zero.

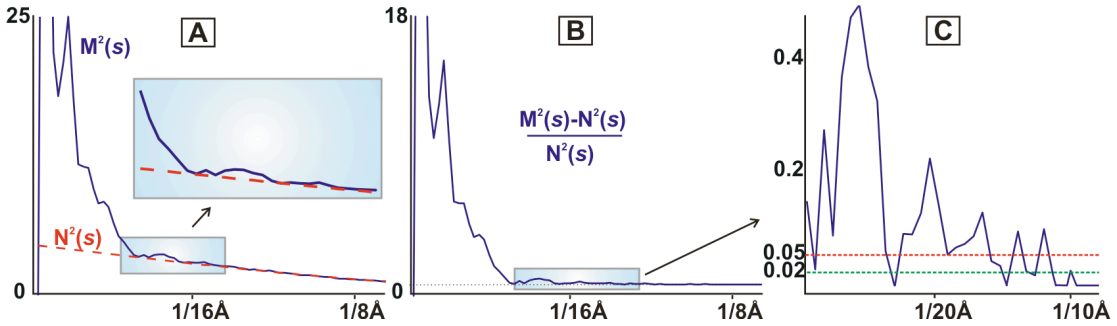

Figure 1. 6. Interpreting the rotationally averaged power spectrum. A. Rotationally averaged power spectrum as shown in Fig. 1.10C but with modeled noise profile and inset zoom in on subregion where the first CTF maximum can be observed. **B**. The SNR profile of the data generated from A. **C**. Zoom in focusing on the SNR profile in B so that the point of intersection with the 0,05 and 0.02 cutoffs can be observed.

The absolute SNR can be used to estimate the achievable resolution in a given single particle experiment (Saad et al., 2001; Thuman-Commike et al., 1999). At present the nominal SNR recoverable by SPA software packages is approximately 0.05. This was confirmed in the context of the EMAN software package (Ludtke et al., 1999) that reconstructed an artificially generated test data set to a resolution of 1/4 Å (Ludtke et al., 1999), at which point the SNR was approximately 0.05. The article also claimed that, very broadly, an SNR of 0.02 is recoverable in the context of the EMAN package (Ludtke et al., 1999). Given these parameters it is reasonable to estimate that the GroEL data (shown pictorially in Fig. 1.11C) contains sufficient signal to reveal molecular detail to 1/11 Å using the 0.05 SNR cutoff, as shown in Fig. 1.11C. If this cutoff is relaxed to 0.02 the same data could potentially yield structural information at 1/10 Å (green dashed line, Fig. 1.11C).

At high resolution the most significant influences on SNR are the envelope function $E(s)$ and the structure factor $\Phi(s)$ (Eq. 1.36). If the envelope function is overly dampening, for instance as a result of relative large defocus settings or low microscope operating voltage, the SNR will diminish rapidly and the projected resolution of the final 3D model will be limited. It is also important to note the influence of the structure factor itself. For

example, if this macromolecular structure itself contains comparatively little information at high resolution the problem of revealing fine detail will be further exacerbated. This problem is compounded by that fact that in most SPA experimental conditions the structure factor is unknown to begin with. Indeed, recovery of the structure factor is precisely the aim of the SPA experiment. In either of these potentially resolution limited scenarios, and in the absence of any significant algorithmic advances, the simplest strategy is to increase the total number particle images used in the experiment.

1.4.3 Wiener filtering

Wiener filtering is a Fourier based image processing technique for recovering signal from noisy data. In an ideal SPA situation (i.e. in the absence of noise) the original signal can be recovered from the measured signal by performing a simple deconvolution in Fourier space, which can be expressed as

$$\Phi_{recoverd}(s,\theta) = \frac{C(s)E(s)\Phi(s,\theta)}{C(s)E(s)}, \quad C(s) \neq 0. \tag{1.38}$$

Here $\Phi_{recovered}(s)$ denotes the deconvoluted (recovered) image. Note that in the presence of zeros in the CTF function this deconvolution process will fail to yield the true projection image $\Phi(s)$, due to the loss of information at these radial frequency ranges. In the presence of noise, the operations defined in Eq. 1.38 will cause noise amplification, especially at high spatial resolutions where the signal is weak. This motivates the derivation of the Wiener filter, a Fourier based filter designed specifically for data recovery from noise corrupted signals. We seek an optimal deconvolution operation of the form

$$\overline{\Phi}(s,\theta) = W(s,\theta)\big(C(s)E(s)\Phi(s,\theta) + N(s,\theta)\big), \tag{1.39}$$

where $W(s,\theta)$ is the unknown Wiener filter and $\overline{\Phi}(s,\theta)$ is the filtered, output approximation of the true signal, such that

$$\left|\Phi(s,\theta)-\overline{\Phi}(s,\theta)\right|^2=\left(\Phi(s,\theta)-\overline{\Phi}(s,\theta)\right)\left(\Phi(s,\theta)-\overline{\Phi}(s,\theta)\right)^* \tag{1.40}$$

is minimized in the least squares sense.

The orthogonality principle dictates that the error vector corresponding to the optimum vector $\overline{\Phi}(s,\theta)$ should be perpendicular to $\Phi(s,\theta)$, that is

$$\left(\Phi(s,\theta)-\overline{\Phi}(s,\theta)\right)\bullet\Phi(s,\theta)=0. \tag{1.41}$$

Evaluating Eq. 1.41 in detail reveals the functional form of the optimal Wiener filter which is may be stated as

$$W(s,\theta)=\frac{C(s)E(s)\left|\Phi(s,\theta)\right|^2}{C^2(s)E^2(s)\left|\Phi(s,\theta)\right|^2+\left|N(s,\theta)\right|^2}, \tag{1.42}$$

$$=\frac{C(s)E(s)}{C^2(s)E^2(s)+1/SNR(s,\theta)}. \tag{1.43}$$

Here we have introduced the polar (absolute) SNR defined as

$$SNR(s,\theta)=\frac{\left|\Phi(s,\theta)\right|^2}{\left|N(s,\theta)\right|^2}. \tag{1.44}$$

Eq. 1.43 is a concise description of the optimal Wiener filter which minimizes the difference between the estimated and true signal, based on the assumption that the SNR can be approximated. Wiener filters, in various forms, are applied extensively in SPA

software packages such as SPIDER (Penczek et al., 1997), FREALIGN (Grigorieff, 2007) and EMAN (Ludtke and Chiu, 2002). The main difficulty in applying a Wiener filter in signal processing is that, in general, the SNR (defined in Eq. 1.44) can not be estimated because the structure factor $\Phi(s,\theta)$ is unknown and the noise function $N(s,\theta)$ is random. In practice this makes Wieners filters dependent on estimated SNR.

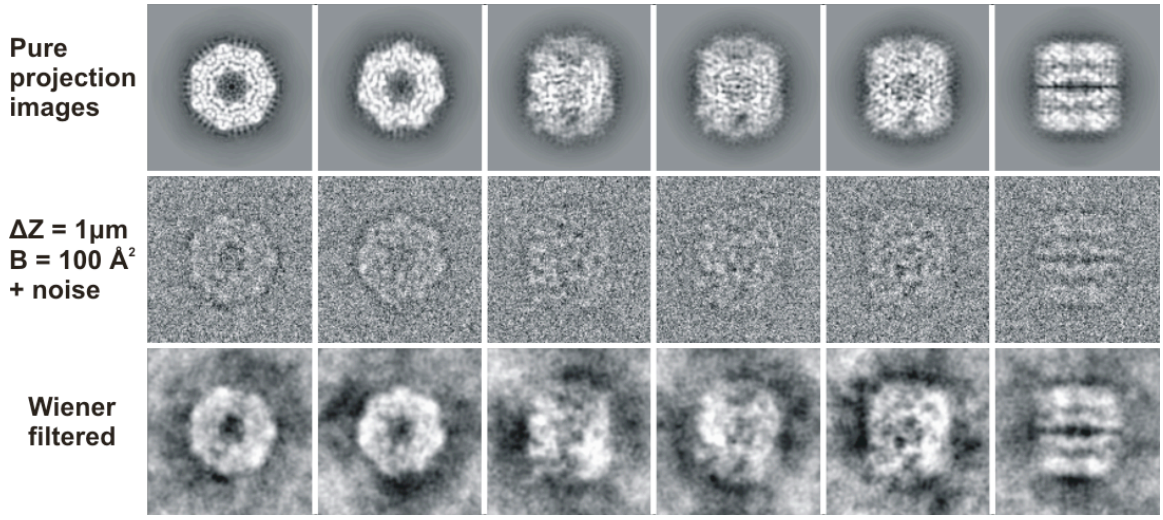

Figure 1. 7. Wiener filtering. This figure depicts typical results obtained by the Wiener filtering operation. The results shown here are applied to synthetic data. The top row depicts pure projection images as yet uncorrupted by the electron microscope imaging system. The middle row depicts the results of applying typical CTF, envelope distortions and adding background noise to the projections shown in the top row (parameters shown). The bottom row depicts the output of the Wiener filtering option as defined in Eq. 1.42, determined by using estimates of structure factor amplitudes from x-ray measurements in a beam line synchrotron.

It is however possible to obtain estimates of the average 2D profiles $\Phi^2(s)$ and $N^2(s)$ using the rotationally averaged power spectrum of many raw data sets, as described in Section 1.4.2. Thus, the SNR defined in Eq. 1.44 can be approximated using the formula

$$SNR(s,\theta) \approx SNR(s) = \frac{\Phi^2(s)}{N^2(s)}. \tag{1.45}$$

In Eq. 1.45 the noise profile is defined according to Eq. 1.26 and the (power) structure factor can estimated using techniques described in Section 1.4.2, or by measuring the structure factor directly for example by using a synchrotron beamline. An example of Wiener filtering applied to synthesized data using the SNR approximation embodied in

Eq. 1.45 is shown in Fig. 1.12, using the known GroEL structure factor (Ludtke et al., 2004).

The use of structure factor estimates from x-ray measurements using a synchrotron has been proposed for use in conjunction with SPA software for many years now (Ludtke et al., 2001; Saad et al., 2001; Thuman-Commike et al., 1999; Zhou and Chiu, 2003). Indeed, such an approach was utilized to reconstruct GroEL at the benchmark resolution of 6 Å (Ludtke et al., 2004).

1.4.4 The singular value decomposition in single particle analysis

The Singular Value Decomposition (SVD) is a useful mathematical concept that allows the analysis of intrinsic, mutual variation in an arbitrary data set, such as a set of boxed particle images used in SPA. This intrinsic variation is identified based on a mathematical technique for generating a set of orthogonal basis vectors/images that literally point in the direction of the most significant underlying data variation of the image data, in a high-dimensional space. This section will explore the mathematics of the SVD, and describe how it is utilized in by SPA software packages.

Let the data set being studied consist of n boxed particle images, each of l x k pixel dimensions, and consisting of a total of l x k = m pixels. The SVD theorem states that, given A, a real n by m matrix, then there exists orthogonal matrices $U \in \Re^{m \times m}$ and $V \in \Re^{n \times n}$ such that

$$A = UDV^T, \tag{1.46}$$

where

$$D = diag(\sigma_1, \sigma_2, ..., \sigma_\rho),\ \rho = \min(m, n),\ D \in \Re^{m \times n} \tag{1.47}$$

and

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_\rho = 0 . \quad (1.48)$$

Here the values σ_i are referred to as the singular values of A, which are positive and non zero for all i. In the special case where $n = m$ (and matrix A is non singular), the SVD is precisely the eigenvalue decomposition. The SVD reveals a great deal about the underlying nature of the data in the matrix A. For instance, the number r is the rank of A. In addition the span of the column vectors $[u_1,\ldots,u_r]$ is the range of A, and the span of the column vectors $[v_{r+1},\ldots,v_n]$ is the null space of A (Golub and van Loan, 1996).

In image processing applications the data are most often placed into the columns of A. That is, the data in each 2D image is instead treated as a one dimensional (1D) vector of length m and arranged side by side in the matrix A. This concept is depicted in Fig. 1.13. Indeed, Subsequently the SVD is calculated using standard numerical techniques (Golub and van Loan, 1996), which directly generate the matrices U, D and V (Golub and van Loan, 1996). In image processing the columns of the matrix U have particular significance, and are referred to as the eigenimages of that data set. SVD-based analysis can be utilized as the basis of particle classification schemes at various stages in 3D reconstruction procedures in different software packages (Chen et al., 2006; Frank et al., 1996; van Heel et al., 2000).

As revealed by Eq. 1.46 and Fig. 1.13, each individual image in A is a linear combination of the so called eigenimages (the columns of $\underline{U}$). The eigenimages are scaled linearly by the singular values of the matrix D and the corresponding coordinates values in the (orthogonal, coordinate) rows of V to yield each unique input image in the columns of the matrix A. More precisely, the image a_j in the j^{th} column of the input matrix A may be expressed as

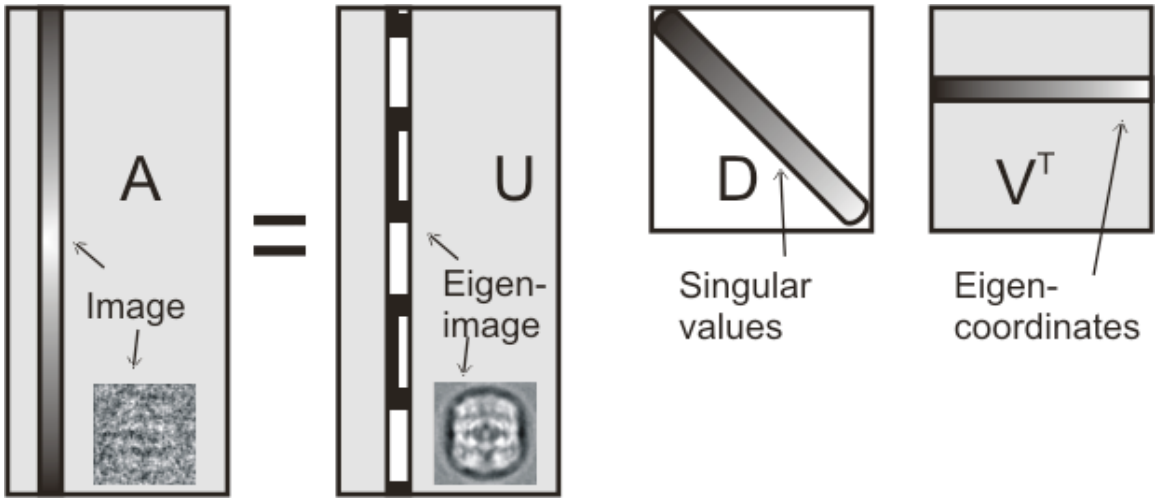

Figure 1. 8. The SVD applied to particle data analysis. This figure depicts how Eq. 1.46 can be interpreted in terms of particle data set as generated by SPA. The matrix *A* consists of the particle images arranged into columns (shown). After application the SVD the matrices *U*, *D* and *V* will be generated. The matrix U will contain the so called eigenimages which reflect the intrinsic variation of the data set. The matrices *D* and *V* having different but equally important interpretations (see text for details).

$$a_j = \sum_{k=1}^{r} v_{jk}\sigma_k u_k \,. \tag{1.49}$$

This concept is described graphically in Fig.1.14. As the matrix *V* is orthogonal each v_j must be precisely of length 1. Therefore, the singular values σ_i and the eigenimages u_j uniquely define a solution bounding hyperellipsoid, the surface of which contains all possible permutations of the matrix vector product *Ax*, for all permutations of x of (vector) length 1. All particle images stored in the matrix *A* will coincide precisely with this hyperelliptic surface.

In SPA the SVD is often used in conjunction with k-means classification (Frank, 1996) or the so called hierarchical ascendant classification (HAC) scheme (van Heel et al., 2000), both of which use clustering algorithms that group particles according to their spatial distribution in eigenimage 'correlation-score' space (or similar). More specifically, SVD can be performed using an entire (potentially massive) particle data set to yield basis (eigen) images as shown by example in Fig. 1.14 (denoted u_j). Following this, the similarity exhibiting between each raw particle image and each basis image can then be evaluated (for instance, using correlation techniques) yielding a set of similarity metrics for each raw particle. According to the (set of) similarities exhibited by each image to the

bases, particles can then be grouped in self similar classes and averaged to produce signal boosted projection image reflecting the intrinsic variation in the data set. Because variation is strongly dependent on Euler orientation, this process will generally yield projection views of the particle in various orientations, as shown in Fig. 1.15. Such approaches can be used to for generating the initial model in the iterative 3D reconstruction procedure or alternatively for evaluating data set heterogeneity (Chen et al., 2006).

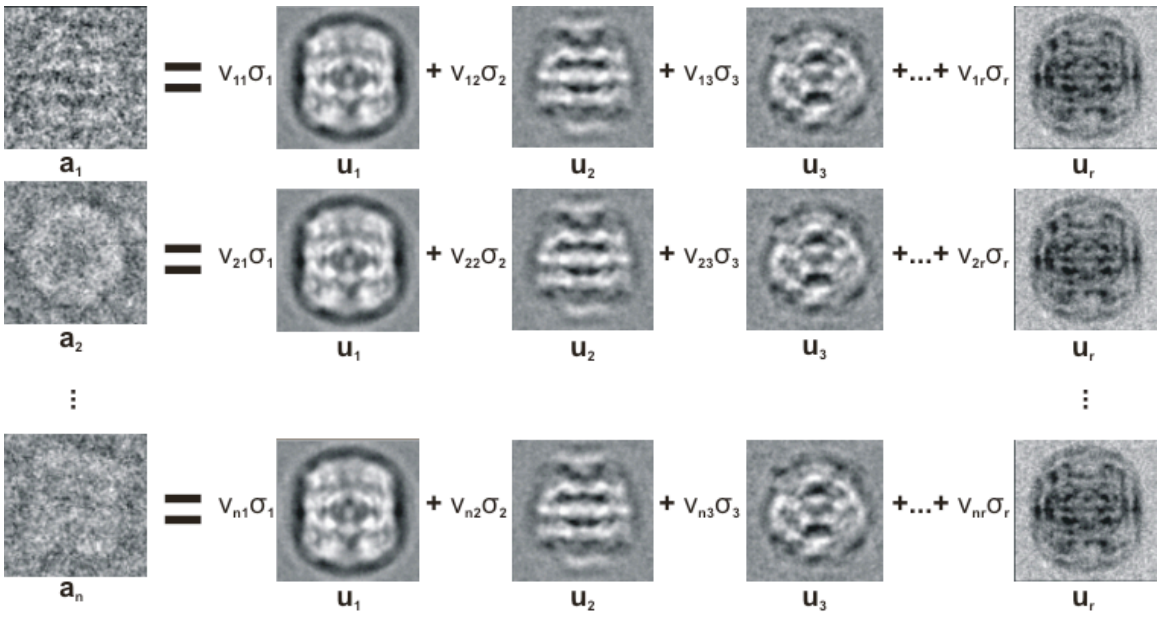

Figure 1. 9. Interpreting the SVD. Each of the particles in the SPA data set (arranged in the columns of A) can be interpreted as a linear combination of the eigenimages. The way in the which the eigenimages sum to give the specific particle image is dependent on the contents of the row vectors of *V*, which are sometimes called the *eigencoordinates* and which are, by definition, of length 1. The extent to which any eigenimage can contribute towards the photometric densities in a given particle image is limited by the diagonal entries of *D*, which conceptually define the boundaries of a hyperellipsoid in eigenspace where the axes of this hyperellipsoid are the eigenimages themselves, and all particles images are incident with its surface.

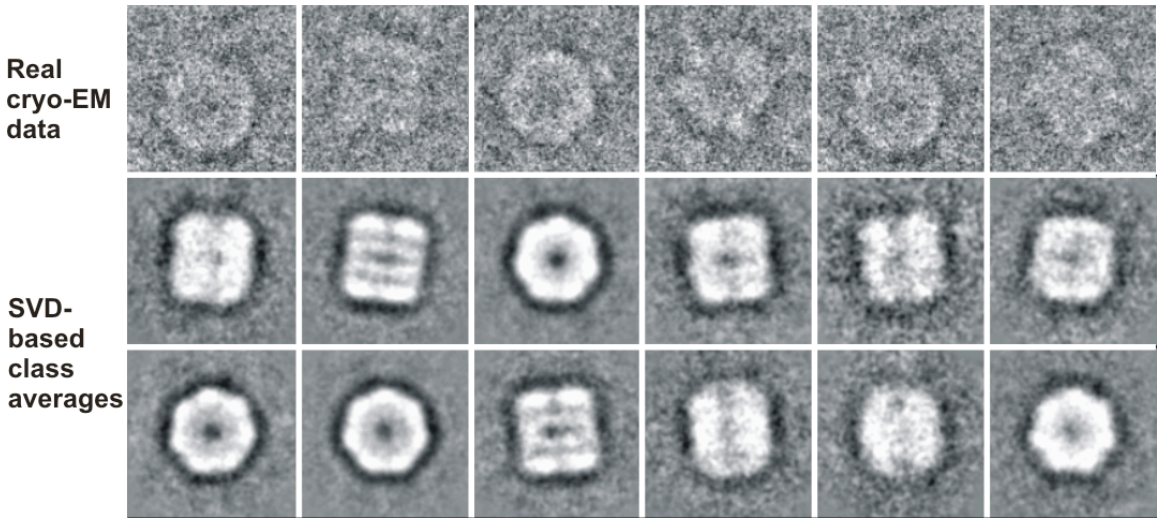

Figure 1. 10. SVD-based class average generation. Classes are generated using k-means classification based on exhibited similarities to eigenimage basis, which are produced by the application of the SVD.