UNIVERSIDAD SAN PABLO - CEU

ESCUELA POLITÉCNICA SUPERIOR

INGENIERÍA DE TELECOMUNICACIÓN

PROYECTO FINAL DE CARRERA

DETECTION OF MICROSCOPE ABERRATIONS IN ELECTRON MICROSCOPY

Autor: Lucía Gallego Quevedo
Directores: Dr. Carlos Óscar Sánchez Sorzano
Dra. Slavica Jonic

Junio de 2011

Calificación del Proyecto Fin de Carrera

Datos personales del alumno

D.N.I.	
APELLIDOS	NOMBRE

Directores

Director 1 (tantos como sean los directores)

D/Dª

Tribunal calificador

Presidente

D/Dª	FIRMA

Secretario

D/Dª	FIRMA

Vocal

D/Dª	FIRMA

Fecha de calificación

Calificación

RESUMEN

El microscopio electrónico de transmisión (TEM) es un dispositivo muy útil para adquirir información estructural de los complejos macromoleculares en las células vivas. TEM presenta aberraciones ópticas que suelen ser modeladas en el espacio de Fourier por la función de transferencia de contraste (CTF). La determinación exacta de la CTF es crucial para su posterior corrección. Por otra parte, la estimación de la CTF debe ser rápida y robusta si se pretenden llevar a cabo estudios de microscopía electrónica de alto rendimiento en tres dimensiones (3DEM).

En este proyecto se presentan varias metodologías con el propósito de mejorar el algoritmo que realiza la estimación de la CTF ajustando un modelo de la densidad del espectro de potencia (PSD) medida en una micrografía específica. Primero nos centramos en mejorar el método existente de la estimación de la CTF. Luego estudiaremos la detección the estimaciones incorrectas de la CTF y PSDs de baja calidad mediante la generación de clasificaciones con varios criterios desarrollados teniendo en cuenta distintos parámetros de la PSD. Finalmente, analizamos la posibilidad de volver a calcular la CTF de las micrografías cuyas PSD eran de alta calidad pero cuyas estimaciones de CTF se calcularon de manera errónea.

La metodología desarrollada está implementada en C++ y Java usando el entorno de desarrollo Eclipse en un sistema operativo Linux. Esta nueva metodología se presenta como parte del software de código abierto de imagen digital XMIPP (X-windows based microscopy image processing package). XMIPP es un programa de continuo desarrollo del Centro Nacional de Biotecnología (CSIC), y está orientado a la transformación completa de las partículas de EM, desde la adquisición de la imagen hasta la reconstrucción 3D.

ABSTRACT

Transmission electron microscope (TEM) is a very useful device to acquire structural information about macromolecular complexes within living cells. TEM introduces optical aberrations that are usually modeled in Fourier space by the so-called contrast transfer function (CTF). Accurate determination of the CTF is crucial for its posterior correction. Furthermore, the CTF estimation must be fast and robust if high-throughput three-dimensional electron microscopy (3DEM) studies are to be carried out.

In this project we present several methodologies that improve the algorithm that estimates the CTF by fitting a model of the power spectrum density (PSD) with its measure on a specific micrograph. We first focus on correcting an existing CTF estimation. Then we study the detection of wrong CTF estimations and low-quality PSDs by generating classifications with several criteria developed taking into account different parameters of the PSD. At the end, we explore the recalculation of the CTF estimations on images for which high-quality PSD was detected previously.

The developed methodology is implemented in C++ and Java using Eclipse development environment on a Linux operating system. This new methodology is currently a part of the open-source digital image processing software XMIPP (X-windows based microscopy image processing package). XMIPP is continiously developed by the National Center of Biotechnology (CSIC) and is oriented to the full processing of EM single particles images in structural biology, from image acquisition to 3D reconstruction.

AGRADECIMIENTOS

Este trabajo no se habría podido realizar sin la colaboración de muchas personas que me han brindado su ayuda, sus conocimientos y su apoyo. Quiero agradecerles a todos ellos cuanto han hecho por mí para que este trabajo saliera delante de la mejor manera posible.

Quedo especialmente agradecida con mis dos directores de proyecto. El Dr. Carlos Óscar Sánchez Sorzano, profesor de la Universidad San Pablo CEU, me brindó la gran oportunidad de poder realizar este proyecto en París. Le agradezco su ayuda durante la adquisición de datos en todos los estudios de campo que se presentan en esta tesis. Ha corregido minuciosamente este trabajo y me ha dado la posibilidad de mejorarlo. Sus comentarios, direcciones, sugerecias y las correciones han sido de gran ayuda para poder elaborar una adecuada memoria de todo el trabajo realizado durante este año. La Dra. Slavica Jonic, del Instituto IMPMC de París, me acogió en su grupo de trabajo y le agradezco sinceramente su confianza y todo el apoyo, consejos y ayuda. Gracias a mis dos directores he podido recibir una formación que me ha permitido acometer este trabajo.

Agradezco a la Univesidad San Pablo CEU y especialmente a nuestros profesores, que compartieron sus conocimientos y nos guiaron en nuestra formación como ingenieros.

No puedo olvidar a mis compañeros y amigos con los cuales he compartido incontables horas de trabajo y estudio. Gracias por todos por los momentos vividos dentro y fuera de la universidad.

Finalmente me gustaría expresar un profundo agradecimiento hacia mi familia, por el esfuerzo y cariño que me han dado durante esta etapa. Soy afortunada por contar siempre con su amor, comprensión y ejemplo.

INDEX

1. INTRODUCTION

Biology is now one of the scientific fields that integrate multi-scale knowledge due to cellular, molecular and genetic discoveries that are taking place. The analyses of these large amounts of data are particularly powerful when they can be interpreted graphically by images representing how the different events occur in cellular and molecular levels. This leads us to high-resolution (< 1nm) structures from transmission electron microscopy (TEM) images. There are various data acquisition techniques, each time more powerful, but those images must be processed in order to highlight the most relevant part of the information. It is at this point where telecommunication engineers play a fundamental role, since image processing is in fact a two-dimensional signal processing and signal processing is a common topic in our field.

Structural biology is a key tool to fully understand the function of macromolecular complexes within living cells. TEM is a very useful technique to acquire structural information about these complexes. However, the electron microscope distorts the structural information by changing amplitudes and phases in recorded images. This is due to the aberrations that exist in the microscope as in any imaging device, and to the particular nature of the propagation of electron waves. These distortions can be modeled in Fourier space by a multiplication of the Fourier transform of an ideal two-dimensional projection of a three-dimensional object with the microscope transfer function called in the field Contrast Transfer Function (CTF).

The final step in structural biology by TEM is three-dimensional reconstruction from TEM images of macromolecular complexes. The CTF estimation and correction has to be carried out as it severely limits the achievable resolution in the three-dimensional reconstruction.

Automated methods for data collection increase the data quantity that can be collected during a single EM. These methods combined with techniques for automated particle picking can generate a three-dimensional reconstruction at sub-nanometer resolution within 24 hours after inserting the specimen grid into the microscope. However, for a high-resolution reconstruction, data quality is as much important as data quantity.

This master thesis is about the CTF estimation from TEM images. The main objective is to implement control checks to detect wrongly estimated CTF and design robust algorithms for the improvement of CTF estimation. Outstanding overall data quality will produce a more precise three-dimensional reconstruction.

2. ELECTRON MICROSCOPY PRINCIPLES

2.1. Electron Microscope

The electron microscope is a device that uses highly accelerated electrons, focused with electromagnetic "lenses", to obtain images of the specimen under study (see Fig. 1). The source of illumination is a filament (cathode) that emits the electrons. Since electrons are scattered by air molecules, the air must be removed by creating a high vacuum. The electrons are accelerated from the cathode to a nearby anode (electric potentials in the order of 200 kV or higher are typically used). Magnetic coils act as lenses and focus the electron beam crossing the specimen. The outcoming electron beam is recorded by a photographic plate or a CCD array. Most of the electrons never interact with the specimen and only contribute to form a background noise. A few electrons will interact elastically (without changing their energy) with the specimen and, finally, a negligible amount will interact strongly (inelastic scattering).

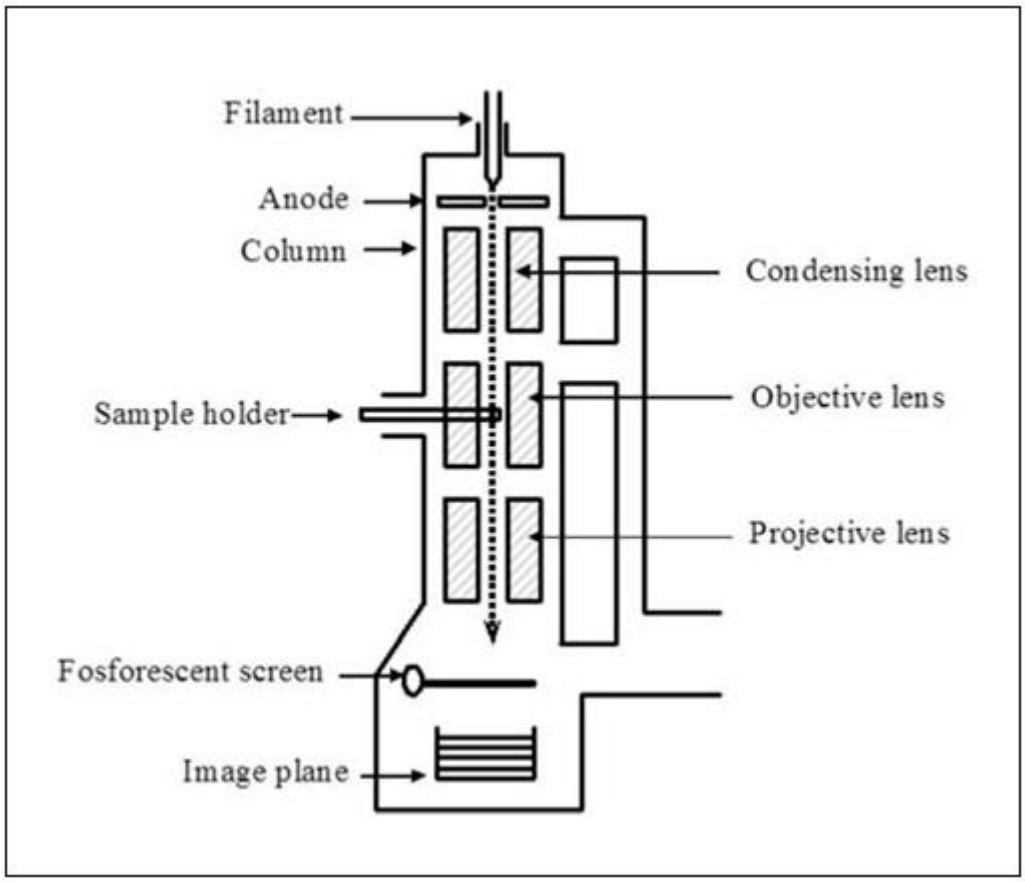

Fig. 1 Shematic representation of an electron microscope

As any other imaging device, the electron microscope introduces some distortion in the acquired images. This distortion is usually modeled in a first order approximation by the convolution with a Point Spread Function (PSF). Its representation in the Fourier space is called the Contrast Transfer Function (CTF).

The CTF looks like a damped two-dimensional sine function. The effect of the CTF is twofold: it introduces zones of alternate contrast (some components are projected as white on a black background, while others are projected as black on a white background) and it introduces band pass filtration.

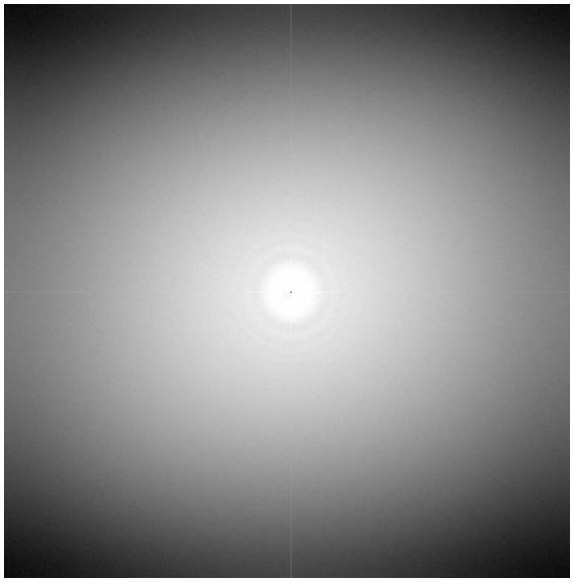

Fig. 2 Experimental PSD

The CTF severely limits the achievable resolution in the three-dimensional reconstruction. In particular, it filters both the high and the low frequencies, introduces zones of alternate contrast and eliminates all information at certain frequencies. It is, therefore, desirable to replace the reconstruction obtained by a 'real' microscope by a reconstruction that would be obtained from images that would be produced by an ideal, aberration-free microscope.

The biological sample

Before taking into account the reconstruction problem itself, we should discuss the kind of object to be reconstructed and its behavior during the recording process. Biological macromolecules are small. Their size ranges from 100 to 10,000 Angtroms. This small size implies that a direct manipulation is extremely difficult, if at all possible, and can only be performed under rather restrained conditions, which represents an obstacle for their characterization.

The conditions inside the electron microscope, high vacuum and high electron radiation level, are very deleterious for the specimens, which should therefore be protected somehow (for example by embedding the sample in ice). This protection has as a side effect in that it decreases the signal-to-noise ratio (SNR). In addition, the problem of beam induced damage is by no means negligible. Electron radiation induces intense ionization of the sample with the formation of free radicals and ions that produce important alterations of the structure. In order to minimize this damage, very low electron doses are used, which in turn produce images with extremely low SNR. Typically observed SNRs can be as low as 1/10.

The solution devised for improving the poor SNR in the micrographs has been to "average" over many (thousands) of identical copies of the specimen. This can be done directly in the case of 2D crystals, where particles are *a priori* ordered (a crystal is a structure made by an object that repeats itself following a regular pattern), or in the case of single particles (i.e., identical copies of a molecule that are recorded in random orientations inside the electron microscope) only after translational and rotational alignment.

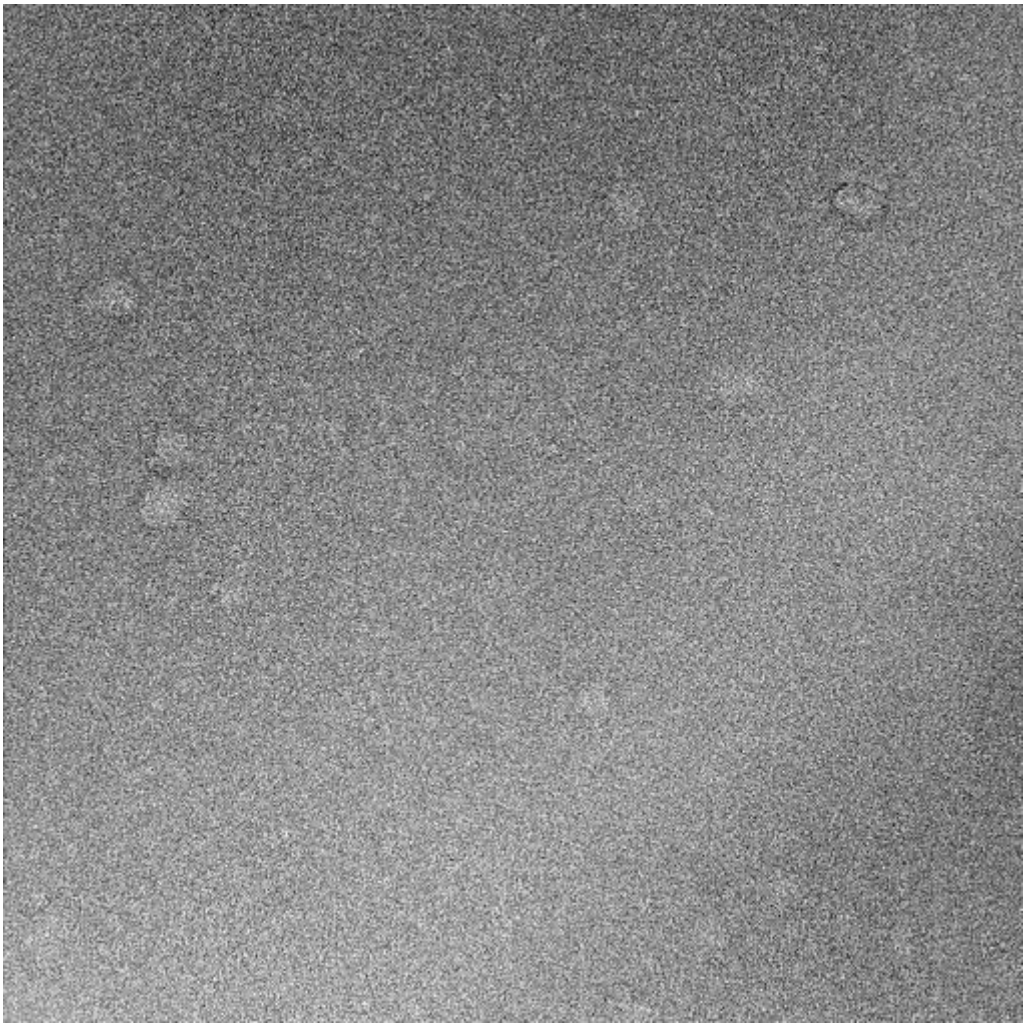

Fig. 3 A micrograph

Noise

As already introduced, the SNR in EM image processing is very low. Noise is generated by many sources. Among others the low, and possibly varying, electron dose, the random nature of the electron emission, the interaction of the electrons with the sample holder, the granular composition of the film where the image is recorded, the electronic noise of the scanner used to digitize the image, etc. The resulting noise has been shown to be additive and normally distributed. This helps simplifying the mathematical formulation of many of the optimization problems involved from the image acquisition step towards the 3D reconstruction of the macromolecule.

3D Reconstruction

Different approaches have been devised to reconstruct 3D structures from their EM projections. These approaches can be classified depending on the kind of data they work with, more specifically on the kind of symmetry that the imaged particle exhibits. In the case of helical filaments, a single view carries enough information to reconstruct the specimen up to certain resolution. Other types of symmetry that are typically encountered for biological macromolecules are: 2D-crystals and icosahedral viruses. For the general case, however, we cannot count on symmetry. In the rest of this project we will focus on the latter case, which is termed single particle reconstruction (Fig. 4).

The process followed to obtain a 3D-reconstruction for single particles can be briefly described as follows (only those steps related to the digital image processing will be enumerated):

1. Images containing many identical copies of the specimen are recorded in the electron microscope and converted to digital form.
2. Micrographs may be preprocessed:
    (i) aberrations introduced by the microscope (CTF) are estimated and corrected,
    (ii) images are denoised.
3. Particle projections are identified and extracted from the micrographs.
4. Projections are normalized, aligned and classified (the particles are classified to distinguish possible structural variability, different projection directions or contaminating particles). This is an iterative process, the better the particles are aligned the better they may be classified, and vice-versa.
5. Finally, when a structurally homogeneous and aligned set of particles has been obtained, it can be combined to obtain a volume.

The whole procedure is iterative, since a first rough reconstruction helps to better identify, classify and align the 2D projections. The newly aligned projections are then used to build a finer reconstruction which in turn is again used to align the 2D projections. This process is iterated until convergence (usually defined as no significant change of the projection alignment, or no significant improvement of the resolution achieved.)

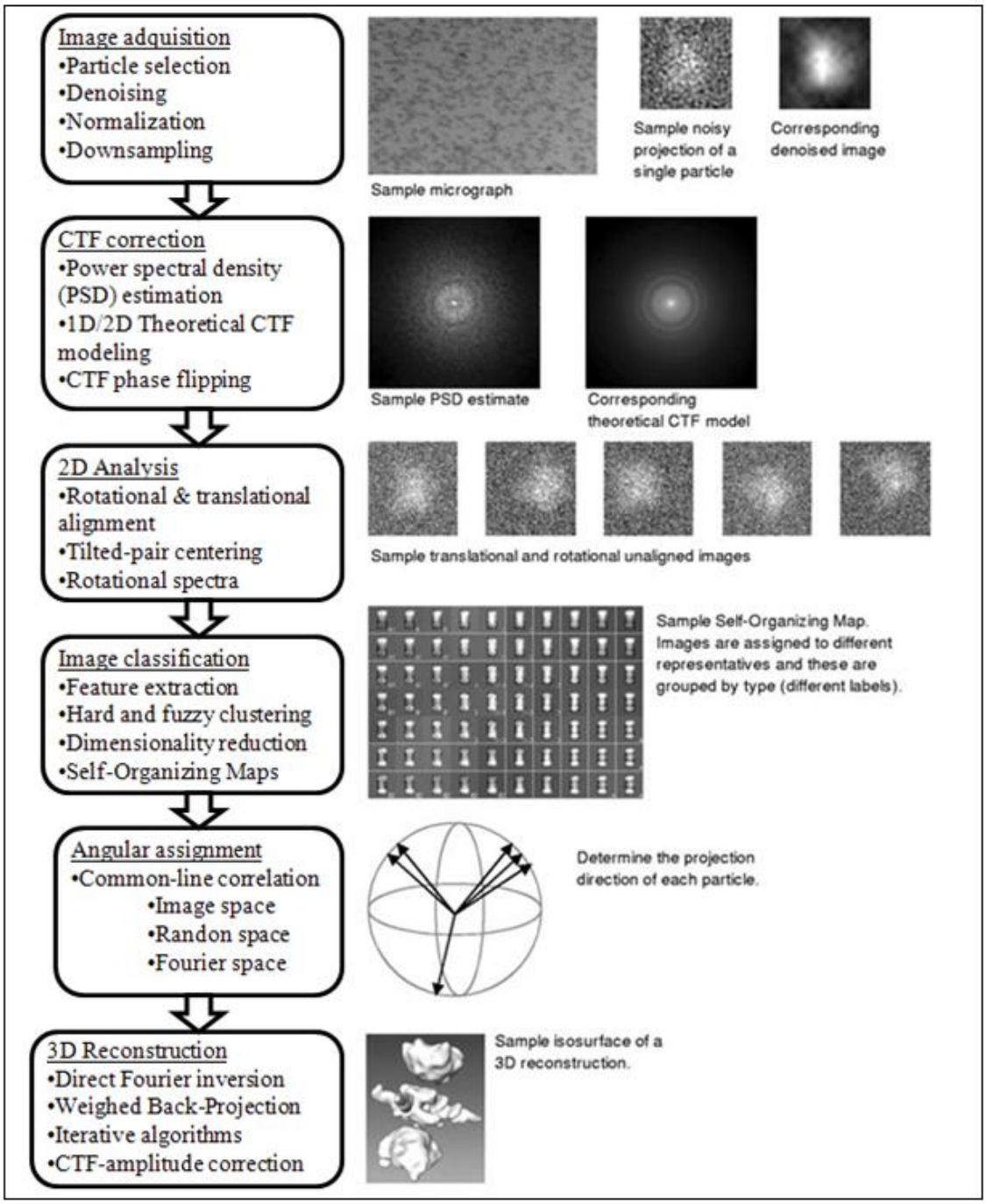

Fig. 4 Schematic work-flow of EM image processing analysis

2.2. Mathematical basis of the determination of the contrast transfer function

We first present the algorithm that fits a theoretical power spectrum density (PSD) based on a CTF model to the PSD measured on a specific micrograph.

The estimation of CTF parameters is usually performed in two steps:

- Estimation of the power spectrum density. The PSD determines the amount of energy present at each spectral frequency. Considering the CTF as a transfer function of a linear system that takes an input (unknown) image and transforms it into an output (experimentally observed) image, and without taking into account noise, the PSD of the output image is the PSD of the input image multiplied by the square modulus of the CTF. Therefore, the PSD of experimental images is directly related to the CTF.

- Estimation of CTF parameters. Once the PSD has been computed, CTF parameters corresponding to the experimental PSD are estimated. This is usually done by minimizing some measure of dissimilarity between the experimental PSD and the theoretical PSD determined by the CTF parameters.

First, we are going to describe the PSD model that is fitted and then we will explain the algorithm that has been developed for fitting the PSD parameters.

PSD Model

We assume that the model of the image formed in the electron microscope is

$$p_{experimental}(r) = h(r) * \left(p_{ideal}(r) + p_{n_b}(r)\right) + p_{n_a}(r) \qquad (1)$$

Where $r \in \mathbb{R}^2$ is a spatial location, p_{ideal} is the ideal projection of the 3D object studied, h is the point spread function (PSF) of the microscope, and p_{n_b} and p_{n_a} represent noise terms added before and after the convolution with the PSF.

Given this image formation model, the corresponding PSD is

$$PSD_{experimental}(R) = |H(R)|^2\left(PSD_{ideal}(R) + PSD_{n_b}(R)\right) + PSD_{n_a}(R) \quad (2)$$

Where $R \in \mathbb{R}^2$ is a spatial frequency, $H(R)$ is the Fourier transform of the PSF (i.e. the CTF), and $PSD_{n_b}(R)$ and $PSD_{n_a}(R)$ are the power spectrum density of the noise terms.

We assume $PSD_{ideal}(R) = 0$, which is not so far from the truth since the noise power is much more important than the signal power in a typical electron micrograph. Moreover, we will assume that the noise before the CTF is white $PSD_{n_b}(R) = K^2$. Under these two hypotheses, the model simplifies to

$$PSD_{theoretical}(R) = K^2|H(R)|^2 + PSD_{n_a}(R) \quad (3)$$

The structure of this PSD is formed by two terms. The first one is the PSD of the noise colored by the CTF. The second one is the PSD of the noise after CTF and is referred to as ‘‘background’’ PSD. Models for these two terms are described in the next sections.

CTF model

A typical microscope has a frequency response approximated by

$$H_{ideal}(R) = -\left((\sin\left(\chi(R)\right) + Q(R)\cos\left(\chi(R)\right)\right) \quad (4)$$

where $Q(R)$ is the fraction of electrons being scattered at each frequency (in our model it is assumed to be constant, Q_o) and

$$\chi(R) = |\Delta f(R)||R|^2 + \frac{1}{2}\,C_s\,|R|^4\,\lambda^2 \quad (5)$$

C_s represents the spherical aberration coefficient, and $\Delta f(R)$ is the defocus vector given by

$$\Delta f(R) = (\Delta f_M \cos(\angle R - \theta), \Delta f_M \cos(\angle R - \theta)) \quad (6)$$

$\angle R$ is the angle of the 2D frequency R. The defocus vector describes an ellipse with major and minor semi-axes Δf_M respectively. The angle of the major semi-axis with respect to the horizontal axis is θ. λ is the electron wavelength which is computed as

$$\lambda = \frac{12.3}{\sqrt{V + 10^{-6} V^2}} \quad (7)$$

where V is the acceleration voltage of the microscope.

A real microscope has a frequency response similar to the ideal one except for a damping envelope $E(R)$, which results in a low-pass filtering of the ideally projected 3D object. Thus, the frequency response of a real microscope is

$$H(R) = E(R)\, H_{ideal}(R) \quad (8)$$

We consider three different effects hindering the maximum achievable resolution: the beam energy spread, the beam coherence, and the sample drift. The three effects combine into a single envelope function as

$$E(R) = E_{spread}(R)\, E_{coherence}(R)\, E_{drift}(R) \quad (9)$$

The beam energy spread envelope is computed as

$$E_{spread}(R) = exp\left(-\frac{\left(\left(\frac{\pi}{4} C_a \lambda\right)\left(\frac{\Delta V}{V} + 2\frac{\Delta I}{I}\right)^2\right)}{\log(2)} |R|^4 \right) \quad (10)$$

where C_a is the chromatic aberration coefficient, $\frac{\Delta V}{V}$ is the energy spread of the emitted electrons represented as a fraction of the nominal acceleration voltage, and $\frac{\Delta I}{I}$ is the lens current instability expressed as a fraction of the nominal current.

We compute the beam coherence envelope as

$$E_{coherence} = exp(-\pi^2\,\alpha^2\,(C_s\,\lambda^2\,|R|^3 + \,|\Delta f(R)|\,|R|)^2) \tag{11}$$

where α is the semi-angle of aperture.

Finally, assuming the mechanical displacement perpendicular to the focal plane ΔF and the displacement in the focal plane (drift) ΔR, the envelope due to sample shift is modeled as

$$E_{drift}(R) = J_0(\pi\,\Delta F\,\lambda\,|R|^2)sinc\,(|R|\Delta R) \tag{12}$$

The envelope model E can be well approximated by a Gaussian function if $\Delta F = \Delta R = \frac{\Delta V}{V} = \frac{\Delta I}{I} = 0$ and $C_s\,\lambda^2\,|R|^3 \ll \Delta f(R)|R|$. However, our model is not simplified in this way and keeps all envelope terms modeling the microscope physics.

Summarizing, the parameters required to fully specify the CTF in our model are

$$\left(K, V, C_s, \Delta f_M, \Delta f_m\,, \theta, Q_0, C_a, \frac{\Delta V}{V}, \frac{\Delta I}{I}, \alpha, \Delta F, \Delta R\right) \tag{13}$$

We assume that V and C_s are fixed for a given microscope and known by the user. The rest of the parameters, 11 in total, will be searched by our algorithm.

Background PSD model

We assume the noise after the CTF to be colored by the film/scanner or CCD frequency response. Physically modeling the corresponding PSD as the output of a linear system, although possible, is out of the scope of this work. Instead, we will model the background PSD as a linear combination of exponential functions. The background PSD depends on R mainly as $e^{-\sqrt{|R|}}$. This term has to be corrected at low resolution with a couple of Gaussians, a positive and a negative one. The formal model for the background noise PSD used in this work is

$$\begin{aligned} PSD_{n_a}(R) = b + & K_s \exp\left(-|s(R)|\sqrt{|R|}\right) \\ & + K_G \exp(-|G(R)|\,(|R| - |C(R)|)^2) \\ & - K_G \exp(-|g(R)|\,(|R| - |c(R)|)^2) \end{aligned} \quad (14)$$

where

$$\begin{aligned} s(R) &= (s_M \cos(\angle R - \theta_s),\ s_m \sin(\angle R - \theta_s)), \\ C(R) &= (C_M \cos(\angle R - \theta_G),\ C_m \sin(\angle R - \theta_G)), \\ G(R) &= (G_M \cos(\angle R - \theta_G),\ G_m \sin(\angle R - \theta_G)), \\ c(R) &= (c_M \cos(\angle R - \theta_g),\ c_m \sin(\angle R - \theta_g)), \\ g(R) &= (g_M \cos(\angle R - \theta_g),\ g_m \sin(\angle R - \theta_g)). \end{aligned} \quad (15)$$

The first term provides a constant baseline; the second term is a decaying exponential representing the background PSD behavior; the third and fourth terms of the model are intended to provide more flexibility in the PSD modeling process. The second term will be referred to as the $\sqrt{}$ - exponential term because of its dependence on the square root of the frequency. The third and fourth terms will be referred to as the positive and negative Gaussians, respectively. To simplify the writing of the equations we will use the following notation for the background PSD

$$\begin{aligned} PSD_{n_a}(R) &= b + PSD_{\sqrt{}}(R) + PSD_G(R) - PSD_g(R) \\ &= PSD_{lower}(R) - PSD_g(R) \end{aligned} \quad (16)$$

All terms are assumed to be elliptically symmetric accounting for a possible anisotropy of the noise after convolution with the CTF.

Summarizing, there are 17 parameters defining the background PSD, namely

$$\left(b, K, s_M, s_m, \theta_s, K_G, G_M, G_m, C_M, C_m, \theta_G, K_g, g_M, g_m, c_M, c_m, \theta_g\right). \quad (17)$$

CTF determination algorithm

The CTF determination algorithm searches automatically for the values of the 28 unknown parameters (11 for the CTF and 17 for the background noise) determining the estimated PSD that best fits the experimental PSD.

The flow diagram below shows the scheme of the CTF determination model. We first obtain the experimental PSD from the micrograph. The enhancement of the PSD is a filtering method for improving visibility of diffraction rings in the experimental PSD of micrographs. We will use the experimental PSD and the enhanced versión of the PSD to determine the CTF parameters based on a theoretical PSD model. The output will be the estimated PSD.

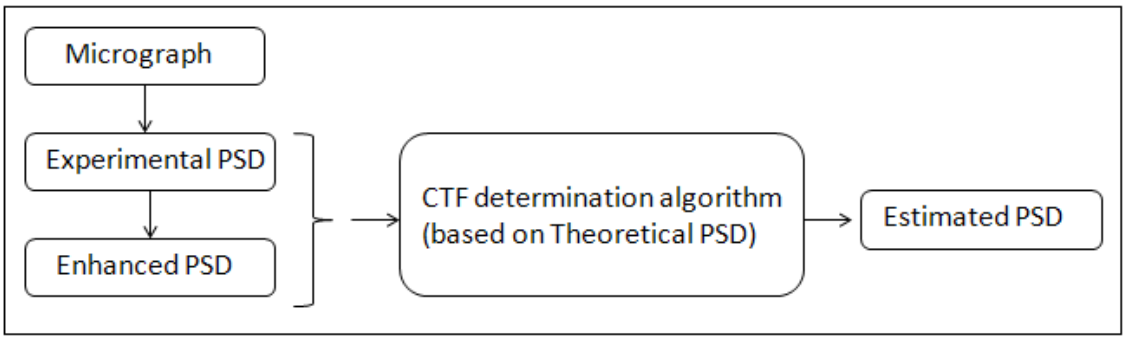

Fig. 5 Schematic diagram of CTF determination model

This fit is evaluated as a fit of two 2D images. The determined CTF is therefore also a 2D image. Attempting to look simultaneously for all 28 parameters without any guidance is a formidable task for any optimization algorithm. Hence, the optimization problem is divided into smaller sub problems that can be easily solved either because there is an analytical solution or because they involve the adjustment of a few parameters with respect to the values of parameters found in a previous step of the algorithm.

Therefore, the parameters of the CTF as well as those of the background PSD are determined in the following four steps:

- Step 1: Determination of the theoretical PSD lower bound.
- Step 2: Determination of the theoretical PSD upper bound.
- Step 3: Defocus determination.
- Step 4: Final model adjustment.

As will be further explained, the fitting is always done by minimizing a given measure of error between a 2D experimental PSD and a 2D theoretical PSD computed for the values of parameters known at the stage (parameters are progressively estimated; thus, in the first substeps only a few of them are known). In our algorithm, the dissimilarity between two 2D images is usually computed based on the l_1- norm of the error vector.

This is so since computing the absolute value of a given quantity is much faster than performing a multiplication (related to the more popular l_2- norm). The employed optimizer is the Powell's conjugate gradient algorithm which is known for a fast local convergence without the need of explicit derivatives of the goal function. However, there are situations in which the problem structure is simple enough so that a solution of the weighted l_2- norm optimization problem can be analytically computed. In these cases, we first compute the analytical solution of the corresponding weighted l_2- norm optimization problem, and then input it to Powell's algorithm as the initial solution of the l_1- norm related problem.

Except in step 4c, all optimizations are performed by considering a coarse regular grid of frequencies. That is, we do not compare all possible frequencies since this will result in a much slower algorithm. In the last optimization step, the coarse regular grid is made finer and finer until all available frequencies are used for the fitting.

To understand better the adjustment of parameters of the CTF, we are going to analyze each step with an example of a well estimated PSD of a random micrograph (Fig. 6). In the left hand side of the image we can see the left half plane of the enhanced PSD. The right hand side shows the estimated PSD. The idea is to observe if both sides match correctly to determine whether the estimation has been done correctly or not.

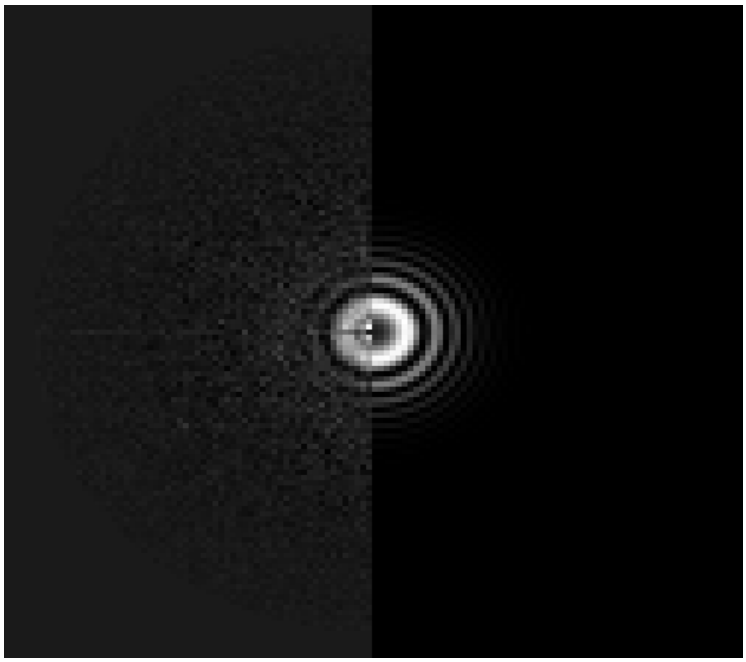

Fig. 6 Enhanced PSD vs estimated PSD

The fitting is done in 2 dimensions, over the X axis and the Y axis, because we are processing 2D images. However, the CTF parameters are fitted imposing a radial symmetric background, so we are going to observe the graphs of the radial average of the enhanced PSD and the estimated PSD.

Step 1: Determination of the theoretical PSD lower bound.

The estimation of the theoretical PSD lower bound is performed in four substeps:

- Steps 1a and 1b: adjustment of initial values of the $\sqrt{}$ - exponential parameters and of the baseline.
- Steps 1c and 1d: adjustment of initial values of the positive Gaussian parameters.

Steps 1a and 1b: adjustment of initial values of the $\sqrt{}$ - exponential parameters and of the baseline

In this step, we compute rough estimates of parameters b, K_s, s_M, s_m and θ_s. First, an initial guess with $s_M = s_m$ and $\theta_s = 0$ is found so that the l_2- norm of the error between the experimental PSD and the theoretical PSD is minimized. Second, this solution is refined now letting $s_M \neq s_m$ and $\theta_s \neq 0$ so that it optimizes the error in the error in the l_1 sense. Finally, the theoretical PSD is further refined by optimization of a penalized l_1-error measure. This penalization moves down the estimated $PSD_{theoretical}$ so that it becomes a lower bound of the experimental PSD.

Step 1a: parameters K_s, s_M, s_m and θ_s are sought with the constraints $s_M = s_m$ and $\theta_s = 0$ so that the l_2 - norm of the error between the experimental PSD and the theoretical PSD is minimized. This is achieved by the weighted least-squares solution of the equation system

$$\log\left(PSD_{experimental}(R)\right) = \log(K_s) - s_M\sqrt{|R|} \tag{18}$$

where we have one equation for each R in a regular grid $\Omega \subset \mathbb{R}^2$, the region in the frequency space where the two PSDs (experimental and theoretical) are being compared. In practice, Ω is an annular region defined by the inner and outer radii specified by the user. It is important to judiciously select this region since at very low frequencies the approximation $PSD_{ideal}(R) = 0$ is not valid. The weight of each one of these equations is

$$w(R) = 1 + \max_{R' \in \Omega}|R'| - |R| \tag{19}$$

That is, the goal function to be minimized is

$$L = \sum_{R_i \epsilon \Omega} w(R_i) \left(\log\left(PSD_{experimental}(R_i) \right) - \log(K_s) + s_M \sqrt{|R_i|} \right)^2 \quad (20)$$

Below there is an illustration of this first step. From now on, we will only show the enhanced PSD to compare it to the estimated PSD. The red line is the enhanced PSD and the green line is the estimated PSD. In this step, we first give an upper bound which adjusts fairly well to the final part of the curve.

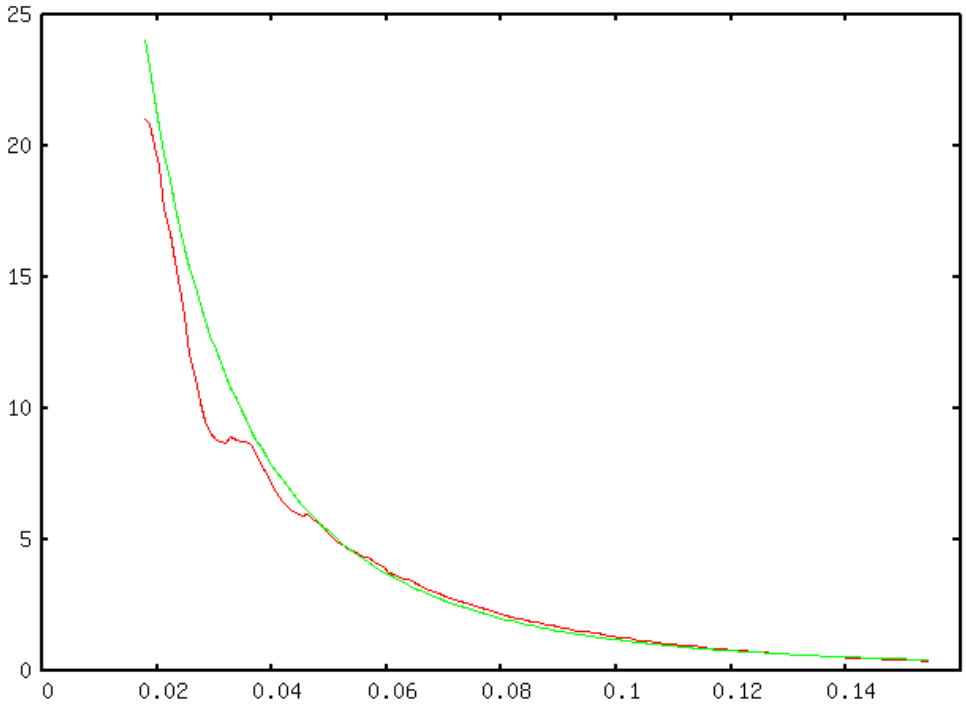

Fig. 7 Radial average of enhanced PSD (red) and estimated PSD (green) after step 1a

Step1b: The first guess of the $\sqrt{}$ - exponential term obtained in the previous step is refined and pushed down in this step. To this goal, the two constraints $s_M = s_m$ and $\theta_s = 0$ are removed, the parameter b (whose initial value is 0) is also estimated, and the error is penalized at frequencies where the theoretical PSD is above the experimental PSD.

Thus, the functional to be minimized in this step is

$$L = \sum_{R_i \epsilon \Omega} |PSD_{experimental}(R_i) - \left(b + PSD_{\surd}(R_i)\right)|$$

$$\left(1 + WI_{PSD_{experimental} < PSD_{theoretical}}(R_i)\right) \quad (21)$$

where $I_A(x)$ denotes the indicator function (this function is one if x belongs to the set A, and is 0 otherwise), W is the penalization weight and follows the sequence 0, 2, 4, 8, 16, and 32. For each W, Powell's conjugate gradient algorithm is used to minimize the penalized functional starting from the solution obtained for the previous value of W.

An illustration of this step is shown

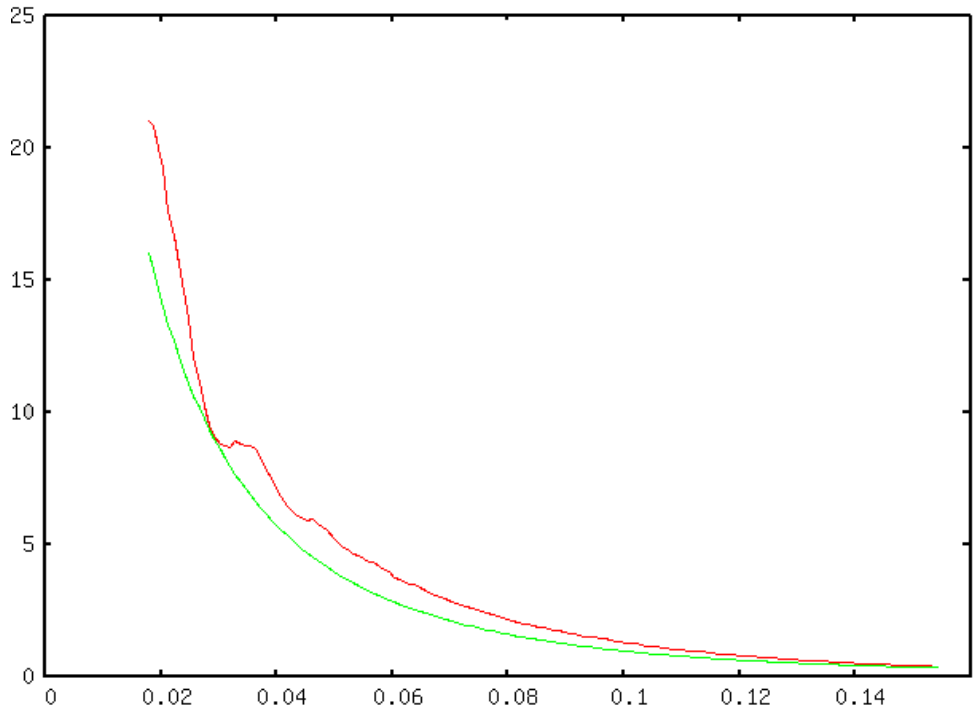

Fig. 8 Radial average of enhanced PSD (red) and estimated PSD (green) after step 1b

Steps 1c and 1d: adjustment of initial values of the positive Gaussian parameters.
In this step, we compute rough estimates of parameters K_G, C_M, C_m, and θ_G and refine these newly introduced parameters together with the ones of the baseline and the $\sqrt{}$ - exponential term (b, K_s, s_M, s_m and θ_s). As in Steps 1a and 1b, the positive Gaussian term of the background PSD is computed in two steps: first, a constrained l_2 – error optimization is performed on a low frequency region of the experimental PSD previously radially symmetrized. This produces a radially symmetric Gaussian that helps the $\sqrt{}$ - exponential term to reproduce the low frequency spectrum of the experimental PSD; second, the condition of radial symmetry of the Gaussian is removed, and a penalized l_1 – error optimization is performed.

Step 1c: the experimental PSD is radially symmetrized as well as the penalized the $\sqrt{}$ - exponential term and baseline found in Step 1.Starting from the lowest frequency, we look for the first frequency at which the two curves are closer. This frequency called R_{min} determines the end of the low frequency region. Within this region, we look for the frequency at which the two curves are more separated, R_{max}. Parameters c_M and c_m are set to $c_M = c_m = R_{max}$. θ_G is set to 0, and G_M is constrained to be equal to G_m. Therefore, only parameters K_G and $G_M$are left. They are chosen so that they minimize the weighted l_2 – norm of the error between the experimental PSD and the theoretical PSD. This is achieved by the weighted least-squares solution of the equation system,

$$\log\left(PSD_{experimental}(R) - \left(b + PSD_{\sqrt{}}(R)\right)\right) = \log K_G - G_M(|R| - C_M)^2 \qquad (22)$$

We evaluate this equation in the same spectral grid points as the one of the $\sqrt{}$ - exponential and we give the same weight as in Eq. (19).

The radial average of the enhanced PSD after this step is shown below.

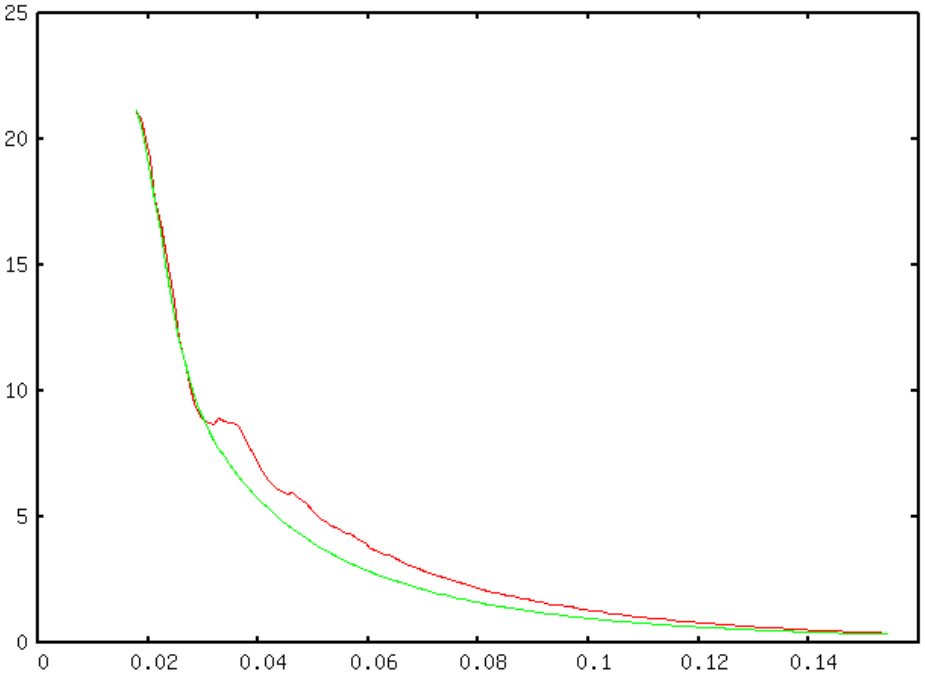

Fig. 9 Radial average of enhanced PSD (red) and estimated PSD (green) after step 1c

Step 1d: this curve is also pushed down so that it really is a background support. The pushing down is done by minimization of

$$L = \sum_{R_i \epsilon \Omega} |PSD_{experimental}(R_i) - PSD_{lower}(R_i)| \left(1 + WI_{PSD_{experimental} < PSD_{theoretical}}(R_i)\right) \quad (23)$$

With respect to all parameters estimated estimated so far. The weight W follows the sequence 0, 2, 4, 8, 16 and 32. The output of this Step 1d is called the theoretical PSD lower bound. It is a fully 2D lower bound although for clarity purposes we only represent its radial average.

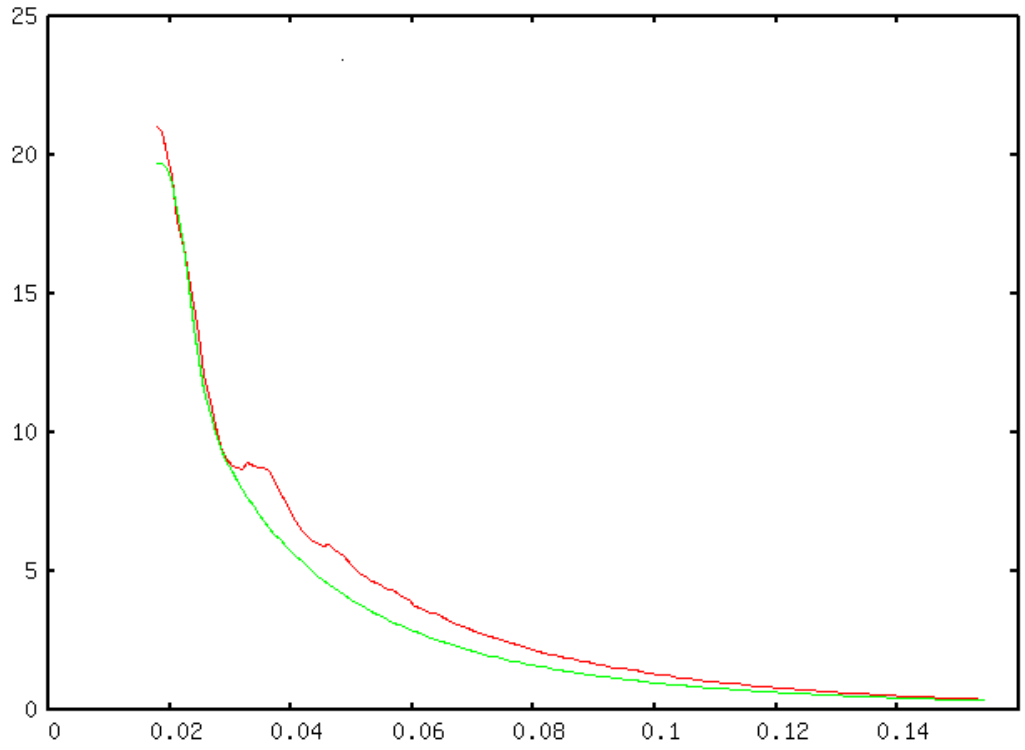

Fig. 10 Radial average of enhanced PSD (red) and estimated PSD (green) after step 1d

Step 2: Determination of the theoretical PSD upper bound.

In this step, we search for the following parameters of the envelope: $K, C_a, \frac{\Delta V}{V}, \frac{\Delta I}{I}, \Delta F,$ and ΔR. The other two unknown parameters of the envelope (Δf and α) are coupled in the term $E_{coherence}$. They will be determined in Step 3 when searching for defocus parameters (Δf_M, $\Delta f_m, \theta$) on which this term depends. Therefore, we assume that $\Delta f(R) = (0,0)$ and $\alpha = 0$ (i.e., $E_{coherence}(R) = 1$ at this point. We set $PSD_{n_a}(R)$ to its lower bound found in Step 1d and we only look for the envelope parameters. As in Step 1d, the search for the upper bound of the PSD is performed by minimizing a penalized goal function. The goal function used in this step is

$$L = \sum_{R_i \in \Omega} |PSD_{experimental}(R_i) - \left(PSD_{lower}(R_i) + K^2 E^2(R_i)\right)| \cdot \left(1 + WI_{PSD_{experimental} > PSD_{theoretical}}(R_i)\right) \quad (24)$$

The initial values of the unknown parameters in this optimization step are

$$\left(K, C_a, \frac{\Delta V}{V}, \frac{\Delta I}{I}, \Delta F, \Delta R\right) = \left(1, {C_a}^{(0)}, 0, 0, 0, 0\right) \tag{25}$$

where ${C_a}^{(0)}$ is an initial chromatic aberration coefficient that can be supplied by the user (by default, its value is 0). The penalization W follows the sequence 0, 2, 4, 8, 16 and 32.

The output of this step is referred to as the theoretical PSD lower and upper bounds. There is no graphic representation of this step, but the idea is the same as in step 1.

Step 3: Defocus determination.

In this step, we determine the defocus parameters ($\Delta f_M,\ \Delta f_m, \theta$) and the aperture semi-angle α. First, we compute a rough estimate of the defocus values making use of the estimated lower and upper bounds. Then, we refine all parameters determined until that point.

One of the problems encountered when fitting a PSD model is that the fitting errors committed at high frequencies are of little importance because of the PSD amplitude damping (the PSD amplitude is very small at these high frequencies). Here is where the lower and upper bounds of the PSD come into play to help us define an error measure that is less dependent on the frequency. Given the lower bound $PSD_{lower}(R)$ and the upper bound $PSD_{lower}(R) + K^2\,E^2\,(R)$, each PSD used in this step is normalized as follows

$$\widetilde{PSD}\,(R) = \frac{PSD(R) - PSD_{lower}(R)}{PSD_{lower}(R) + K^2\,E^2\,(R) - PSD_{lower}(R)}$$

$$= \frac{PSD(R) - PSD_{lower}(R)}{K^2\,E^2\,(R)} \qquad (26)$$

This normalization guarantees that any PSD within the lower and upper bounds will be mapped between 0 and 1, and therefore all frequencies will similarly contribute to the PSD fitting error as long as the lower and upper bounds are accurately computed.

The goal function to be minimized at this stage is

$$L = \frac{1}{|\Omega|}\sum_{R_i \in\, \Omega} \left|\widetilde{PSD}_{\ experimental}(R_i) - \widetilde{PSD}_{\ defocus}(R_i)\right|$$

$$- \rho\left(\widetilde{PSD}_{\ enhanced}(R_i),\ H_{ideal}(R_i)\,E(R_i)\right) \qquad (27)$$

where $|\Omega|$ is the number of spectral grid points in the set Ω, and $\rho(x,y)$ is the correlation coefficient between signals x and y defined as

$$\rho(x,y) = \frac{E\{xy\}}{\sqrt{E\{x^2\}\,E\{y^2\}}} \tag{28}$$

where $E\{.\}$ is the expectation operator. $PSD_{enhanced}$ is a filtered version of the experimental PSD. $PSD_{defocus}$ is computed as follows

$$PSD_{defocus}(R) = PSD_{lower}(R) + K^2\,|H(R)|^2 \tag{29}$$

Again, this step is divided in the following two substeps.

Step 3a: a first estimate of the defocus values is obtained by exhaustive search of the three parameters ($\Delta f_{M,}\,\Delta f_m\,,\theta$) on a regular grid. Each explored point is used as the initial solution of Powell's conjugate gradient optimizer. This algorithm computes values of the three parameters by minimizing the goal function and it is quite "fast" despite the initial exhaustive search. The best fitting parameters computed at this step are used as the initial solution for Step 3b.

In the graph below we can appreciate how we model the function with the upper and lower bounds estimated before, in order to adjust the zeros and poles of the enhanced PSD.

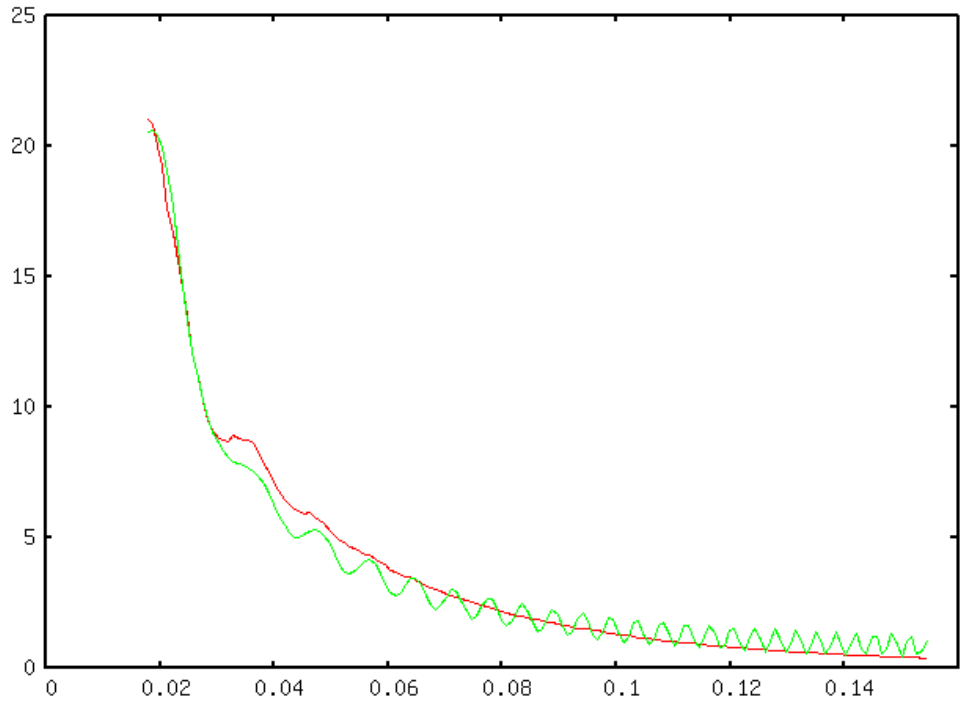

Fig. 11 Radial average of enhanced PSD (red) and estimated PSD (green) step 3a

At this point it is interesting to see the adjustment of the estimated PSD to the enhanced PSD.

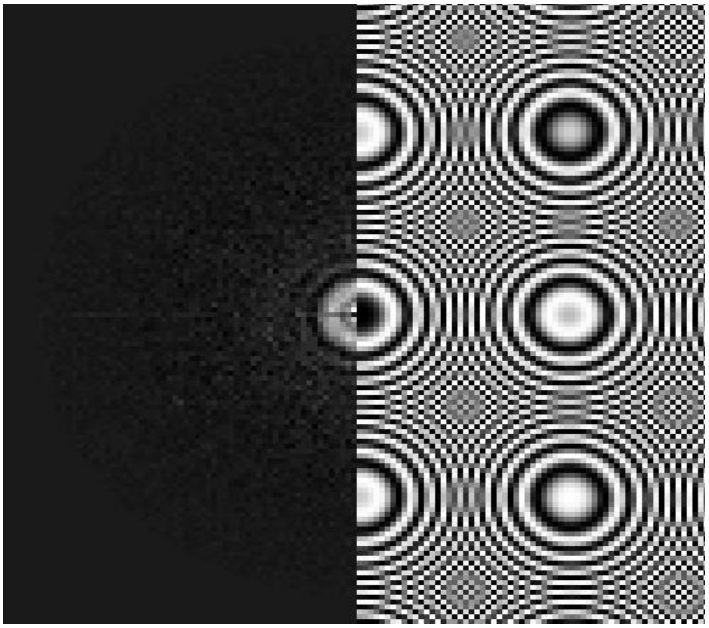

Fig. 12 Half plane of enhanced PSD (left) and estimated PSD (right) after step 3a

Step 3b: we refine all parameters found so far (23 parameters) by minimizing the goal function. The only parameters that have not been found yet are those of the negative background Gaussian (PSD_g). They will be determined in Step 4.

Now we can appreciate the first zero in the theoretical PSD, as well as good values of the curves at the beginning and at the end.

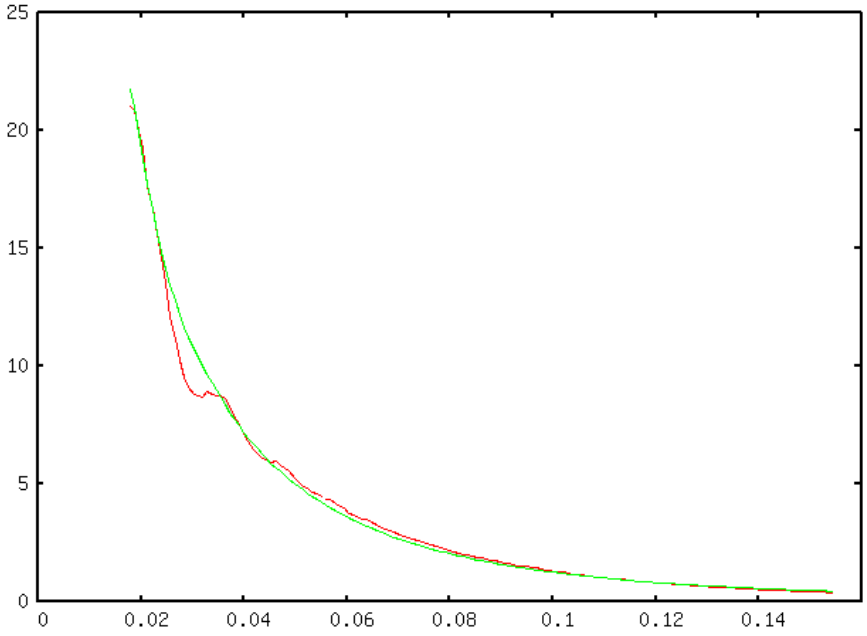

Fig. 13 Radial average of enhanced PSD (red) and estimated PSD (green)

Step 4: Final model adjustment.

In this step, we estimate first the parameters of PSD_g (Step 4a). Then, we refine all parameters of the model using a coarse grid (the same grid as in all previous steps) (Step 4b). Finally, we refine all parameters using a fine evaluation grid (Step 4c). The output of this step is the output of the CTF determination procedure.

Step 4a: we compute a first estimate of the PSD_g similarly as in Step 1c. We assume the term to be circularly symmetric and therefore $c_M = c_m$, $g_M = g_m$ and $\theta_g = 0$. Thus, this first guess can be found as a weighted least-squares solution of the equation system

$$\log\left(PSD_{lower}(R) + K^2\,|H(R)|^2 - PSD_{experimental}(R)\right)$$
$$= \log K_g - g_M(|R| - c_M)^2 \qquad (30)$$

There is one equation for each 2D frequency for which the theoretical PSD estimated in Step 3b is larger than the experimental PSD. We do this because at these frequencies some negative term is needed in order to compensate the difference between the experimental and the theoretical PSD.

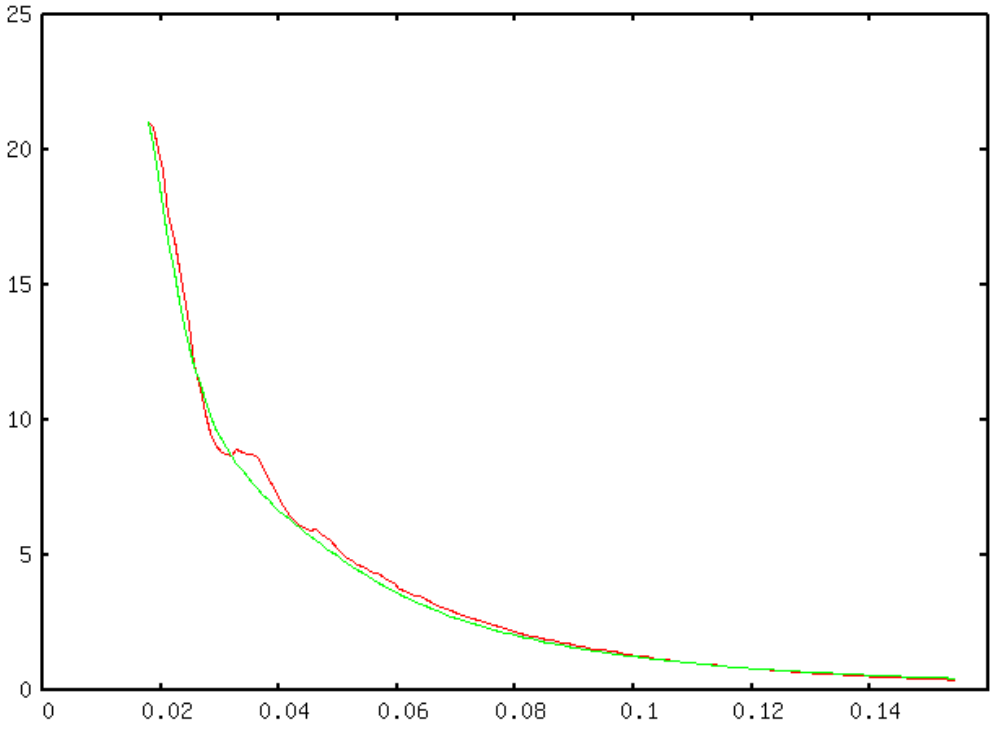

Fig. 14 Radial average of the enhanced PSD (red) and the estimated PSD (green) after step 4a

Step 4b: all model parameters are refined on a coarse frequency grid. By default, the coarse grid is defined by taking 1 frequency sample out of 4 consecutive ones in each direction. The goal function to be optimized is show in Eq. (27).

At this point we can actually see how the radial average of the estimated PSD follows the curve of the enhanced PSD very well.

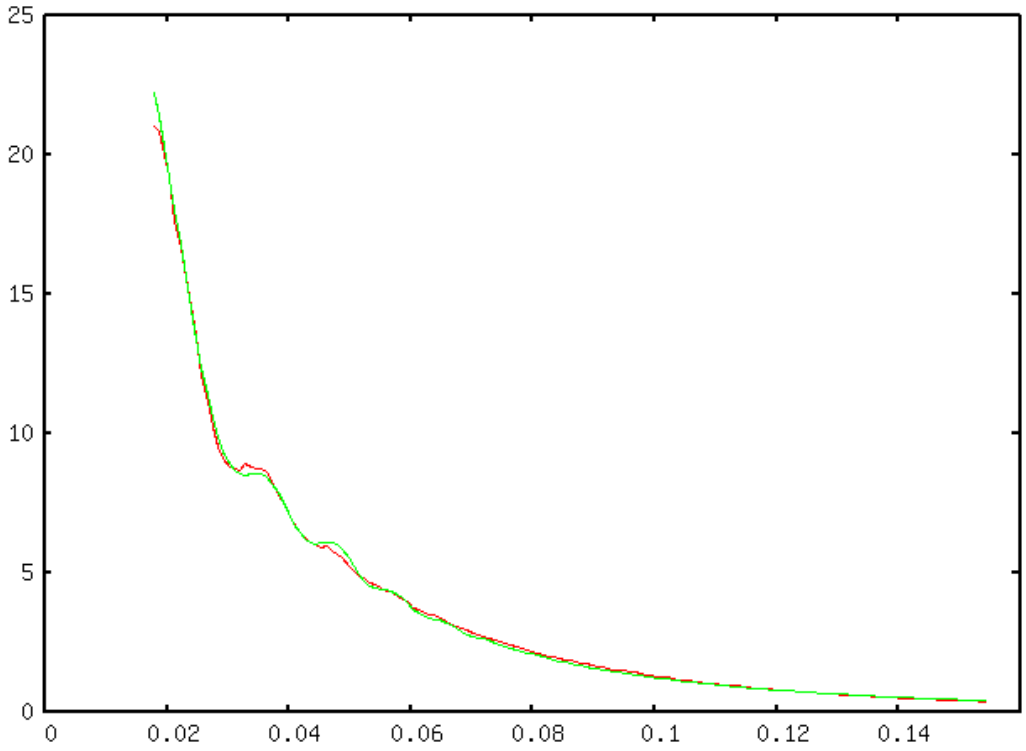

Fig. 15 Radial average of enhanced PSD (red) and estimated PSD (green) after step 4b

Step 4c: the frequency grid is made finer and finer until all available frequencies are used. Thus, the grid is made finer by dividing by 2 the grid spacing until this value is 1. For each grid spacing, the model parameters are refined again using the same function and optimization algorithm as in Step 4b.

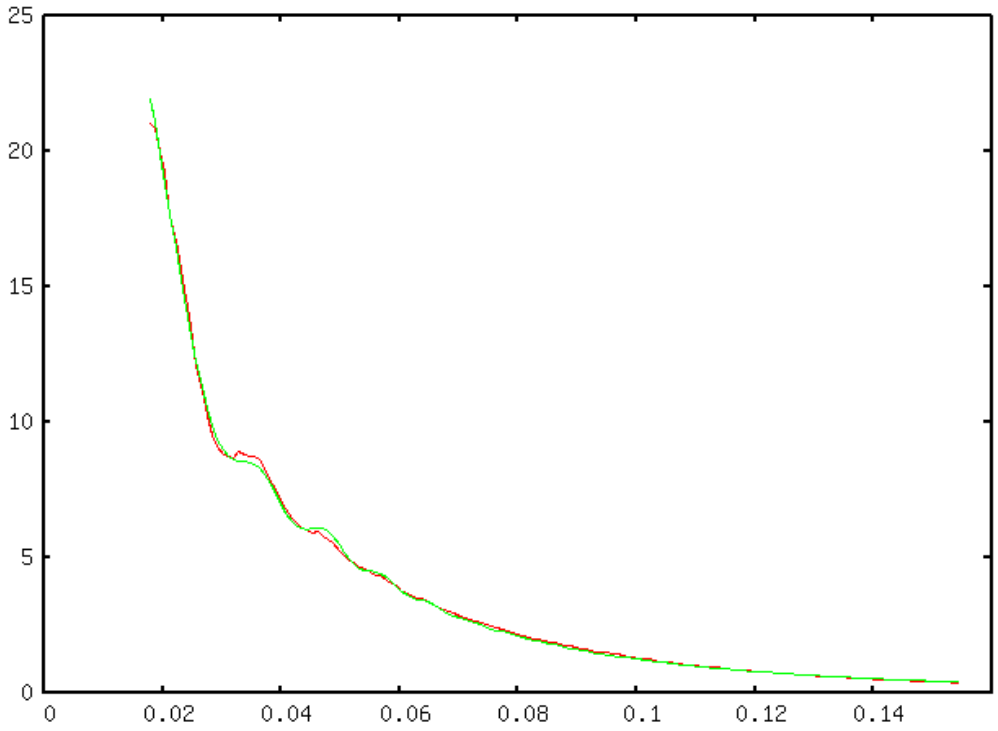

Fig. 16 Radial average of enhanced PSD (red) and estimated PSD (green) after step 4c

We cannot forget that these graphs represent the radial average of the enhanced PSD and the estimated PSD. Radial plots along two perpendicular axes show that the two backgrounds are different (see Fig. 17 and Fig. 18), that is, the background noise level effectively depends on the direction. This experiment shows that assuming radial symmetric backgrounds may result in inaccurate estimates of the defocus.

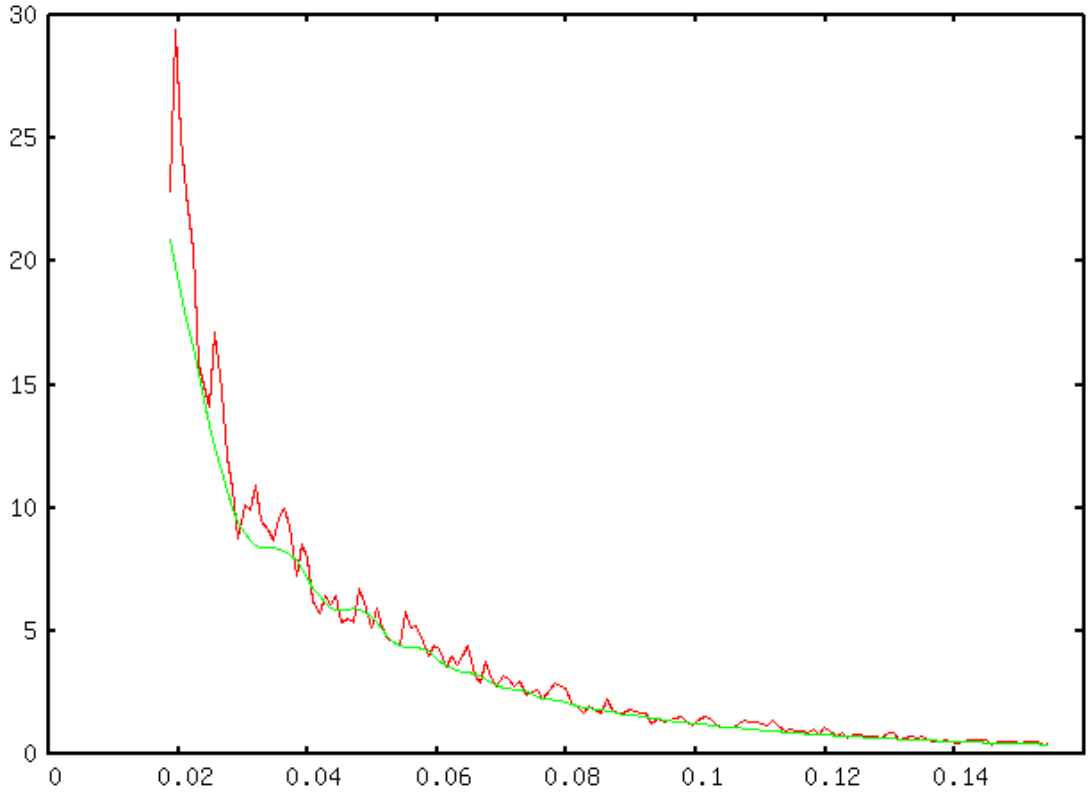

Fig. 17 Fit along x-axis of the enhanced PSD (red) and the estimated PSD (green)

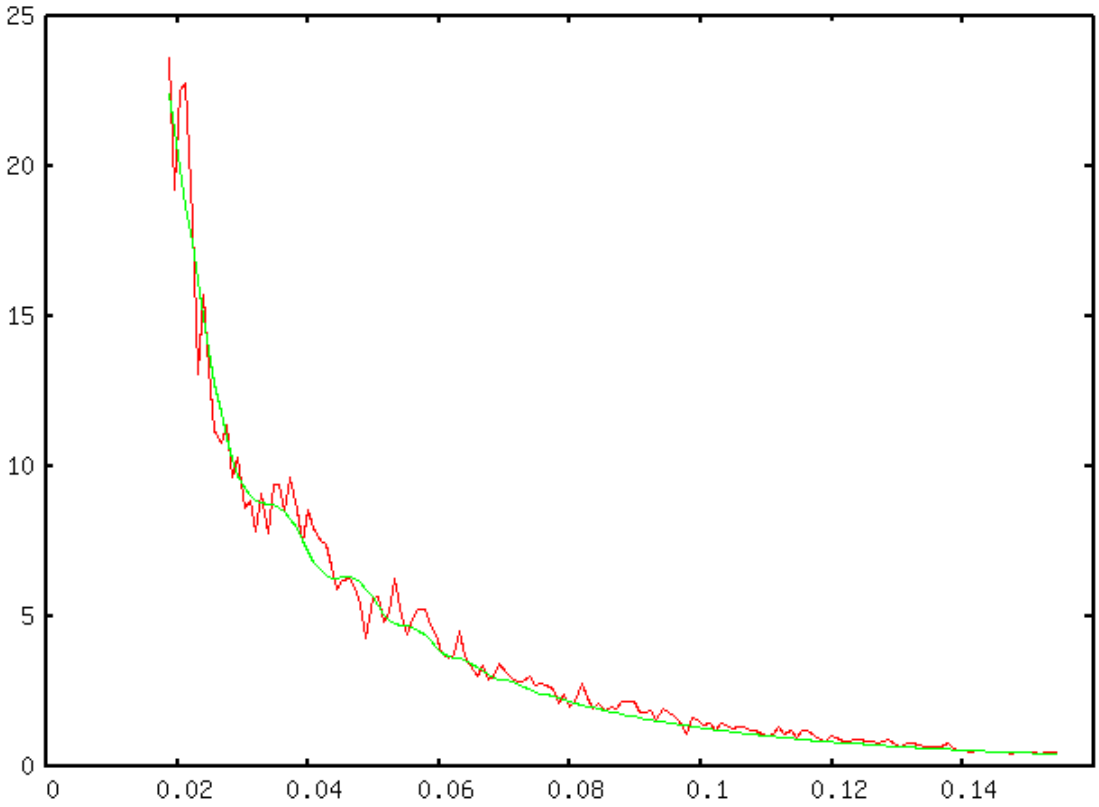

Fig. 18 Fit along y-axis of the enhanced PSD (red) and the estimated PSD (green)

The final estimation of the PSD is shown in Fig. 19. The estimated PSD clearly improves the visibility of the rings of the enhanced PSD and facilitate the quality assessment of micrograph.

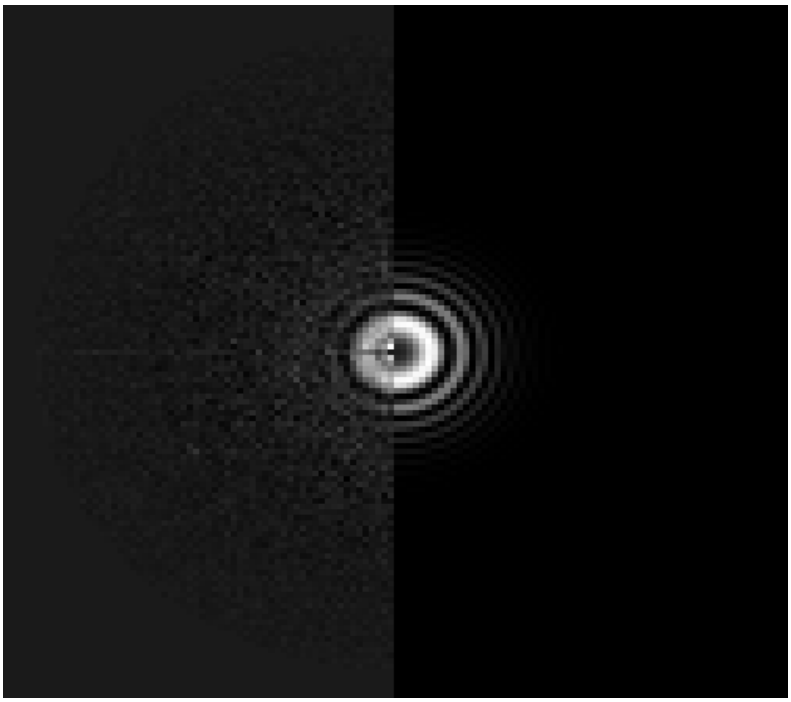

Fig. 19 Final result of the model

3. IMPROVED CTF ESTIMATION AND PSD/CTF CLASSIFICATION

The main objective of this project is to detect and correct wrongly estimated CTF. To this goal, we worked with a set of micrographs to analyze the calculated CTFs and develop several methods to improve results. The new methodology was implemented in C++ and Java ona a Linux operating system, and integrated into the open-source digital image processing software XMIPP (X-windows based microscopy image processing package) developed by the National Center of Biotechnology (CSIC).

We first classified the micrographs into 3 categories:

- Micrographs with good-quality enhanced experimental PSDs and correctly estimated PSDs.
- Micrographs with good-quality enhanced experimental PSDs and incorrectly estimated PSDs.
- Micrographs with bad-quality enhanced experimental PSDs.

The formula used to assess the correctness of the CTF estimation program is the following

$$\eta = \frac{GG}{GG + GB} \quad \% \ success \tag{31}$$

where GG refers to the number of good enhanced PSDs giving correctly estimated theoretical PSDs and GB refers to the number of good enhanced PSDs producing badly estimated theoretical PSDs. Although it is not necessary in the formula for the correctness η, we will also introduce the term BB which will correspond to micrographs with a bad-quality enhanced PSDs.

The formula represents the percentage of good experimental PSD which resulted in a good estimation of the theoretical PSD out of the total number of good experimental PSDs. This way, we do not take into account the bad-quality experimental PSDs as with that data we cannot make a good CTF estimation.

We will apply this formula before and after we introduce any modifications to the program to see in which grade they improve the CTF estimation.

4. METHODOLOGY AND RESULTS

4.1. Introduction

To assess the actual correctness of the calculation of the CTF we will work with a set of 753 micrographs. This set of micrographs is subdivided into several subsets which correspond to samples taken by different electron microscopes and will therefore have different parameter values.

We can distinguish a bad estimation of a PSD from a good estimation by observing the enhanced version of the experimental PSD half plane versus the estimated PSD half plane (Fig. 19). Now we need to specify what we refer to when we talk about a good-quality experimental PSD and a bad-quality experimental PSD in order to determine which micrographs should be discarded.

The good-quality PSD comes from a good-quality micrograph, which typically present multiple concentric rings, extending from the image center toward its edges. Bad micrographs may lack any rings or only have very few rings that hardly extend from the image center. Other reasons to discard micrographs may be the presence of strongly asymmetric rings (astigmatism) or rings that fade in a particular direction (drift). Some examples that illustrate the micrograph selection based on their PSDs are shown in Fig. 20.

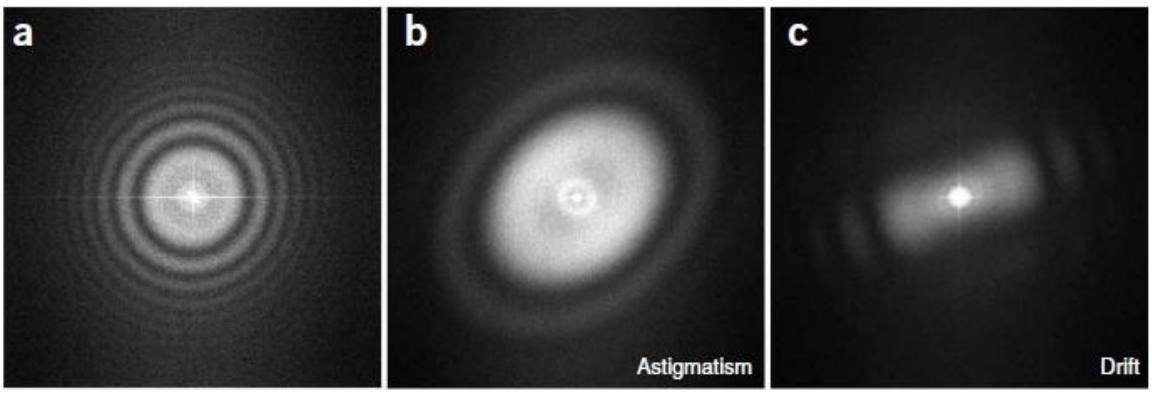

Fig. 20 (a) A suitable PSD has several rotationally symmetric rings. PSDs should be discarded if they present astigmatism, that is, (b) rotationally asymmetric, or drift, that is, (c) fading in a particular direction.

The CTF calculation for 753 micrographs produced the following results:

- Micrographs GG – 486 micrographs
- Micrographs GB – 194 micrographs
- Micrographs BB – 73 micrographs

The computed correctness of CTF estimation is $\eta = 71$ %. It is a fairly good number of correct estimations to start with. In the next section we will develop methods to rise this percentage.

4.2. Improvement of the goal function

4.2.1. Methodology

The estimation of the CTF parameters is done between the minimum digital frequency (which should be a little lower than the digital frequency of the first CTF zero) and the maximum digital frequency (which should be higher than the last zero of the CTF). At higher frequencies there is more correlation with noise, so if we focus on calculating the parameters only between the first and the third zero, where noise affects less, we will probably make a better estimation of the parameters.

In the mathematical basis of the determination of the CTF we presented the defocus determination in Step 3. The goal function to be minimized at this stage is shown in Eq. (27). In this equation $|\Omega|$ is defined as the number of spectral grid points in the set Ω. This set is defined as the region in the frequency space where the two PSDs (experimental and theoretical) are being compared. Set Ω is an annular region defined by the inner and outer radii specified by the user which corresponds to the minimum digital frequency and the maximum digital frequency. We will modify this by denoting that set Ω at this point will be a region defined by the first and the third zero of the PSD.

Therefore, we define a mask for this region and we will work over it instead of working over the whole range of frequencies of the PSD.

The mask is the light red region traced in the figure below.

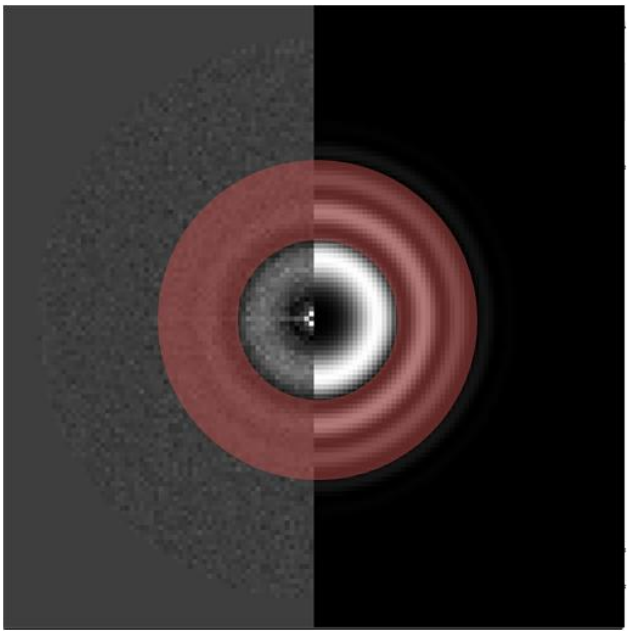

Fig. 21 Mask over the PSD of the micrograph

4.2.2. Results

After implementing this method we could see that many good experimental PSDs that had badly estimated PSDs result new in correctly estimated PSDs. For example, see the following figure.

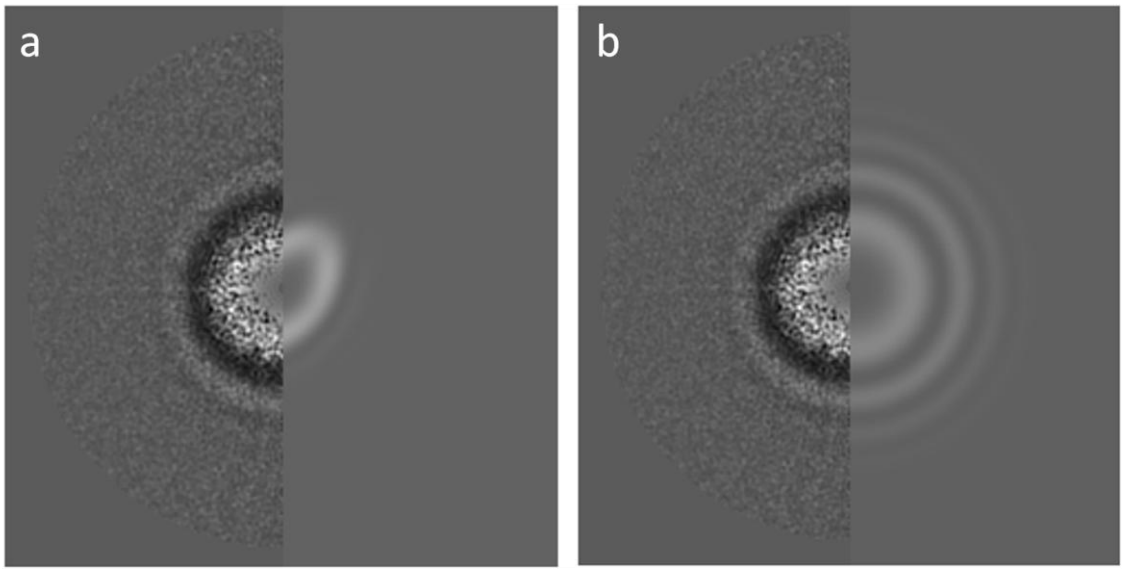

Fig. 22 (a) Initial calculation of estimated PSD, (b) calculation with new methodology

This improvement was introduced in the class *adjust_ctf.cpp* where we calculate the parameters of the CTF.

The correctness η will show evaluate the improvement quantitatively. The CTF calculation on 753 micrographs resulted in:

- Micrographs GG – 574 micrographs
- Micrographs GB – 106 micrographs
- Micrographs BB – 73 micrographs

If we compare these results with the initial ones, we can see that 88 micrographs that belonged to Micrographs GB now belong to Micrographs GG.

The correctness consequently improved form $\eta = 71\%$ to $\eta = 84\%$.

4.3. Semi automatic classification

In this section we are going to develop several criteria to sort the set of micrographs. The aim of this methodology is not to correct badly estimated PSDs but to detect them. If we manage to find a method which sorts the micrographs from better estimations to worst estimations, badly estimated PSDs could be removed from the set of data or even be recalculated. Moreover, bad experimental PSDs could also be removed from our set of images.

4.3.1. Individual criterion

The goal was to implement in XMIPP several different criteria to classify our set of micrographs depending on different properties of its experimental PSD and its estimated PSD. Different criteria give the user the possibility to choose among different classifications. In the following sub-sections we will present the criteria we tested.

To understand how each criterion classifies, the best thing is to select one image that represents a good estimation of the PSD and another one that represents a bad estimation of the PSD. We will not assess these sorting methods with the correctness η because we are not correcting bad estimations; we are detecting bad estimations and bad experimental PSDs. In the annex (A.1) we show a partial view of the classification for each criterion.

4.3.1.1. *Damping*

4.3.1.1.1. Methodology

The damping is the envelope value at the border of the PSD. Micrographs with a high envelope value at border are either wrongly estimated strongly or under sampled.

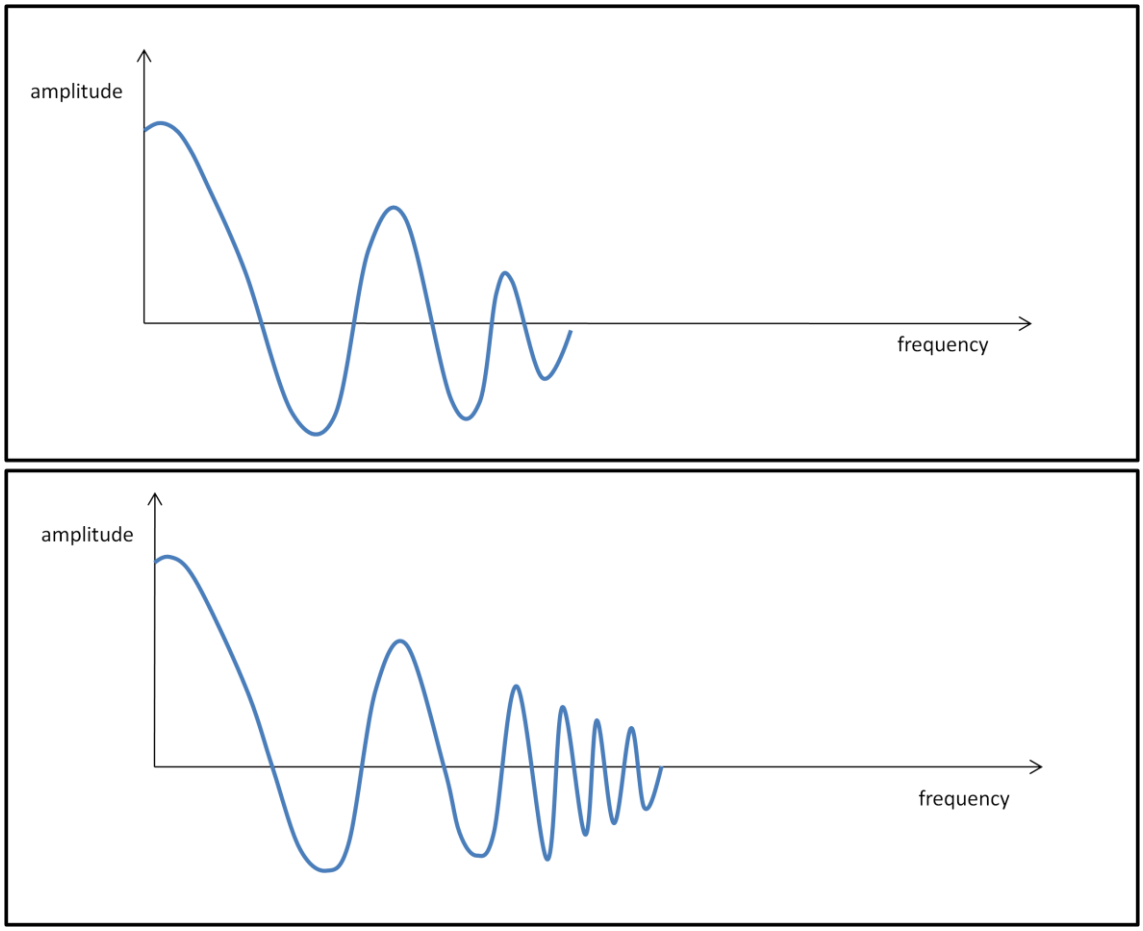

Fig. 23 (1) Radial average of a PSD with a low envelope value, and (2) radial average of a PSD with a high envelope value

The damping envelope is called $E(R)$ and it is shown in Eq. (9). The formula of this criterion is shown below

$$damping = E(R_{border}) \tag{32}$$

where R_{border} is the spatial frequency corresponding to the border of the PSD and it is defined as $R_{border} = \frac{N}{2}$, where N is the size of the image in pixels.

4.3.1.1.2. Results

As we said before, this method evaluates the envelope value at the border of the PSD. In the image below we can appreciate how the envelope of Fig. 24 (b) extends more to the border than the envelope of Fig. 24 (a). Image (a) will appear somewhere at the beginning of the classification and image (b) somewhere near the end.

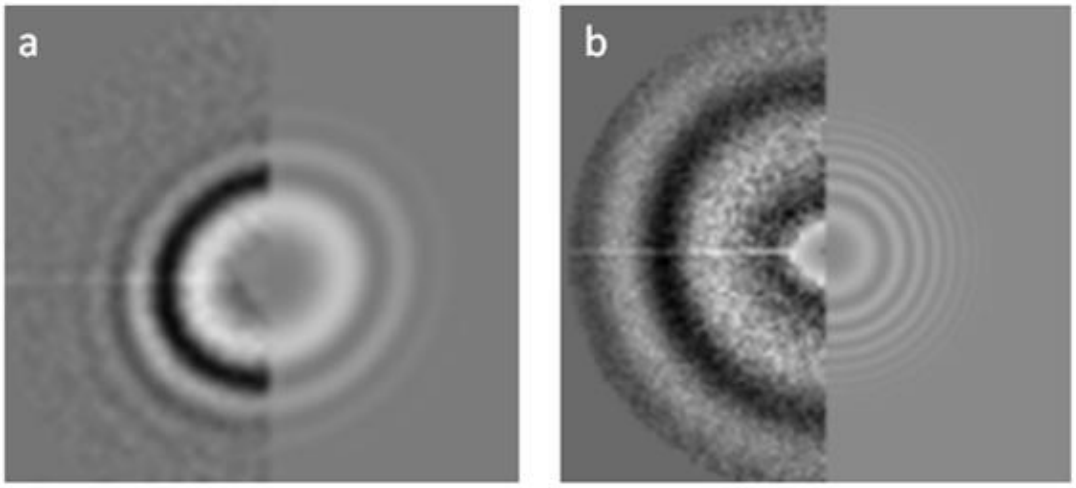

Fig. 24 (a) A PSD with a low envelope value, (b) a PSD with a high envelope value

It is obvious that PSD (b) is badly estimated. In fact, we can demonstrate that the bad estimation of this PSD is due to an error when introducing the value of the sampling rate of the micrograph.

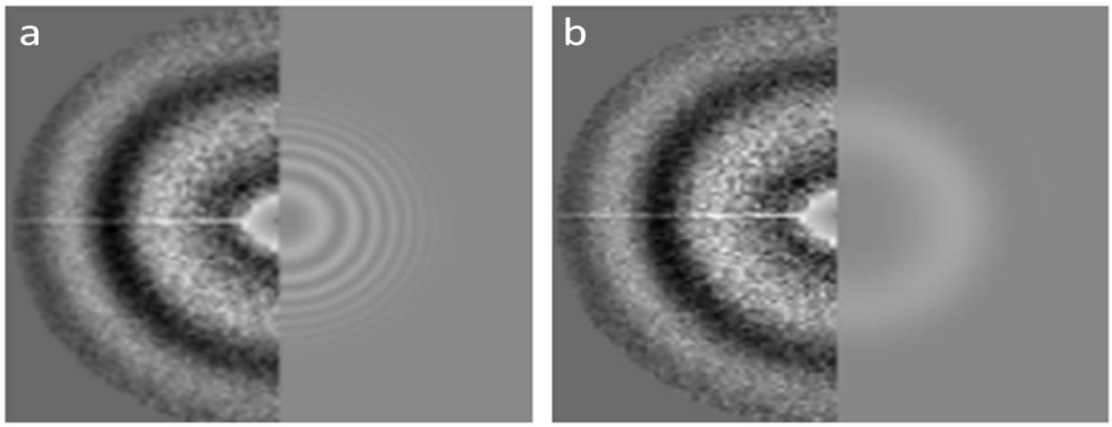

Fig. 25 (a) PSD (Fig. 24 (b)) evaluated with sampling rate of 0.84 Angstroms, (b) PSD (Fig. 24 (b)) evaluated with sampling rate of 2.52 Angstroms

We solved the problem by multiplying the sampling rate by three (Fig. 25). This problem is caused by an error of the user when introducing the sampling rate, so there is no way we can correct this, but we can detect this kind of mistakes with this criterion.

4.3.1.2. *First zero average*

4.3.1.2.1. Methodology

In practice, the frequency of the first zero of the PSD is between 0.1 and 0.25 in normalized units (normalized frequency is between 0 and 0.5), which corresponds to the values between 10x and 4x the sampling rate in Angstroms. PSDs with the first zero out of this range will be penalized. The diagram of Fig. 26 shows the range of values in which the first zero should be according to experimental results.

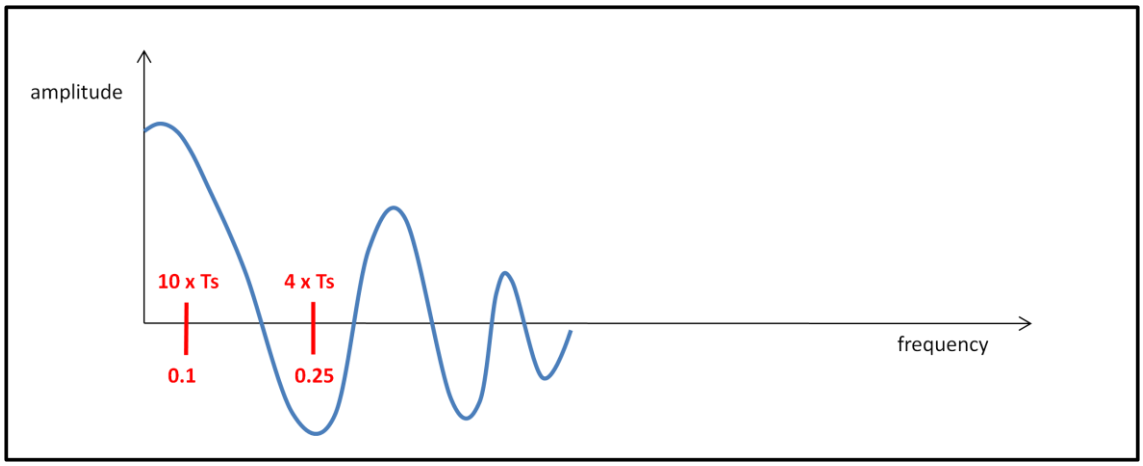

Fig. 26 Radial average of a PSD with the normalized frequency margins marked

In the first place, we calculate the radial integral around the first zero

$$first\ zero\ average = \left| \sum_{\alpha=0^{\circ}}^{360^{\circ}} PSD_{enhanced}(A_{\alpha}R_1) \right| \tag{33}$$

where α is the radial angle $\alpha = [0^{\circ}, 360^{\circ}]$, R_1 is the spatial frequency corresponding to the first zero, and A_{α} is defined below

$$A_{\alpha} = \begin{bmatrix} \cos\alpha & sin\alpha \\ -sin\alpha & cos\alpha \end{bmatrix} \tag{34}$$

The formula of this criterion (Eq. (33)) will give a numerical result which corresponds to the normalized frequency of the first zero of the corresponding PSD. If this result is not between 0.1 and 0.25, the resulting value of the criterion will be a high value.

4.3.1.2.2. Results

When we saw the classification results, we discovered that this method is very useful to identify CTFs which were not estimated at all because they are highly penalized.

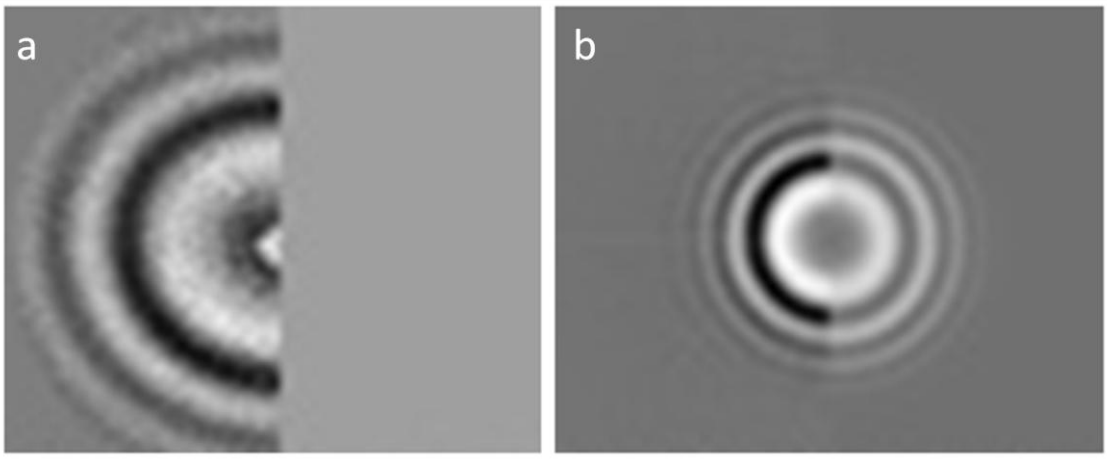

Fig. 27 (a) PSD at the beginning of the list, and (b) PSD at the end of the list

The enhanced PSD of Fig. 27 (a) was not calculated, therefore the radial integral of the first zero will not satisfy the premises of this criterion. This means that this PSD will be highly penalized and will be situated at the beginning of the list. Enhanced PSDs with the first zero of the radial integral that are between the defined experimental range will have a small value for this criterion and will be situated near the end of the list (Fig. 27 (b).

4.3.1.3. *First zero ratio*

4.3.1.3.1. Methodology

This measures the astigmatism of the PSD by computing the ratio between the largest and smallest axes of the first zero ellipse. Ratios close to 1 indicate no astigmatism.

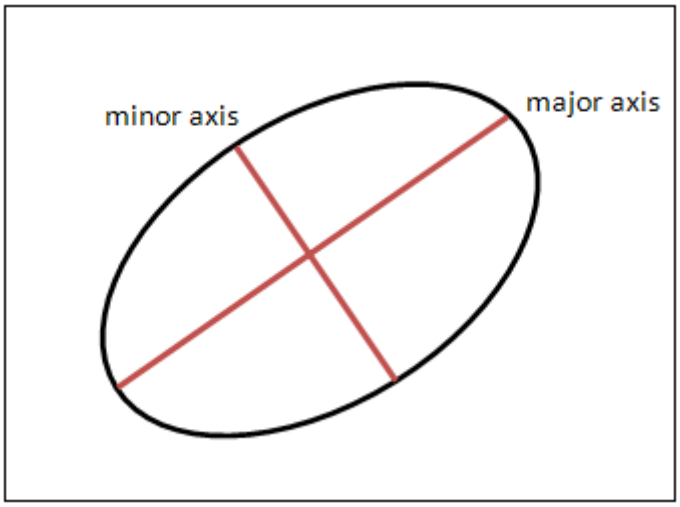

Fig. 28 Major and minor axis of the first zero ellipse

The formula of this criterion is defined as

$$first\ zero\ ratio\ = \frac{\sum_{R=0}^{R_1} PSD(\bar{U}R)}{\sum_{R=0}^{R_1} PSD(\bar{V}R)} \tag{35}$$

where $\bar{U}$ and $\bar{V}$ are unitary vectors corresponding to the directions of axis U and axis V respectively, and R_1 is the spatial frequency corresponding to the first zero.

4.3.1.3.2. Results

At the beginning of the list we will have CTFs with no astigmatism and at the end of the list we will have the more astigmatic CTFs.

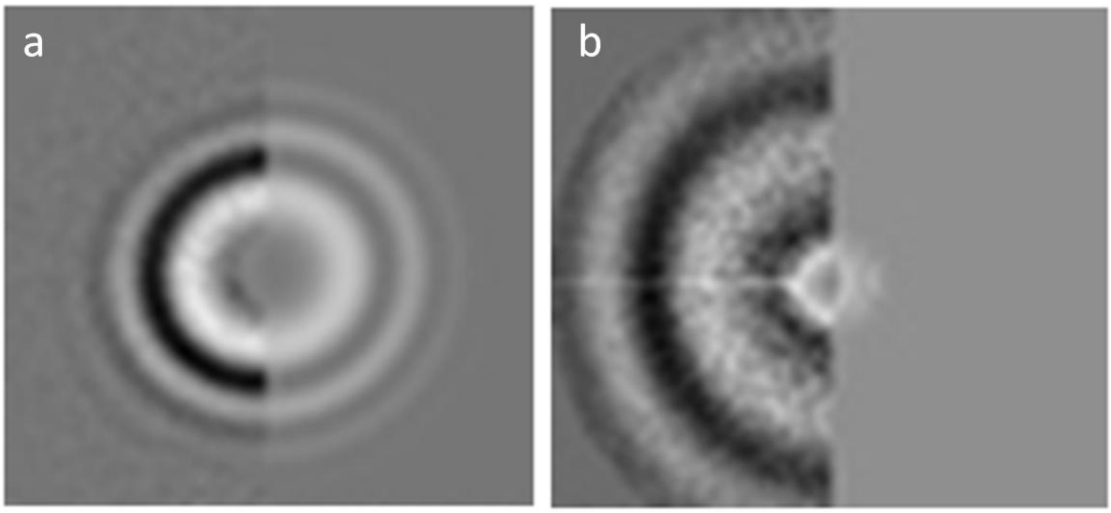

Fig. 29 (a) PSD at the beginning of the list, and (b) PSD near the end of the list

In Fig. 29 (a) we see that the eccentricity of the ellipse (the ratio of the distance between the two foci to the length of the major axis) is small. For this reason, this PSD will be located in the beginning of the list. The PSD of Fig. 29 (b) has a large eccentricity so it will be placed near the end of the list.

4.3.1.4. *Fitting score*

4.3.1.4.1. Methodology

The CTF is computed by fitting a theoretical model to the experimentally observed PSD. This criterion is the fitting score. Smaller scores correspond to better fits.

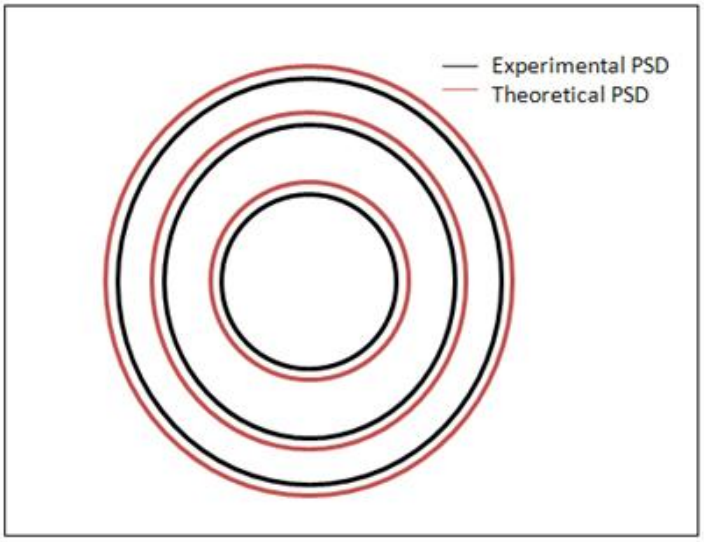

Fig. 30 Zeros of experimental PSD vs theoretical PSD

The formula corresponding to this criterion is

$$fitting\ score = \sum_{R=0}^{R_{max}} |PSD_{theoretical}(R) - PSD_{experimental}(R)|^2 \qquad (36)$$

where R_{max} is the maximum value of the spatial frequency.

4.3.1.4.2. Results

In the figure below we can clearly observe that the PSD at the beginning of the list has a better fitting of its theoretical model with its enhanced experimental PSD than the PSD at the end of the list.

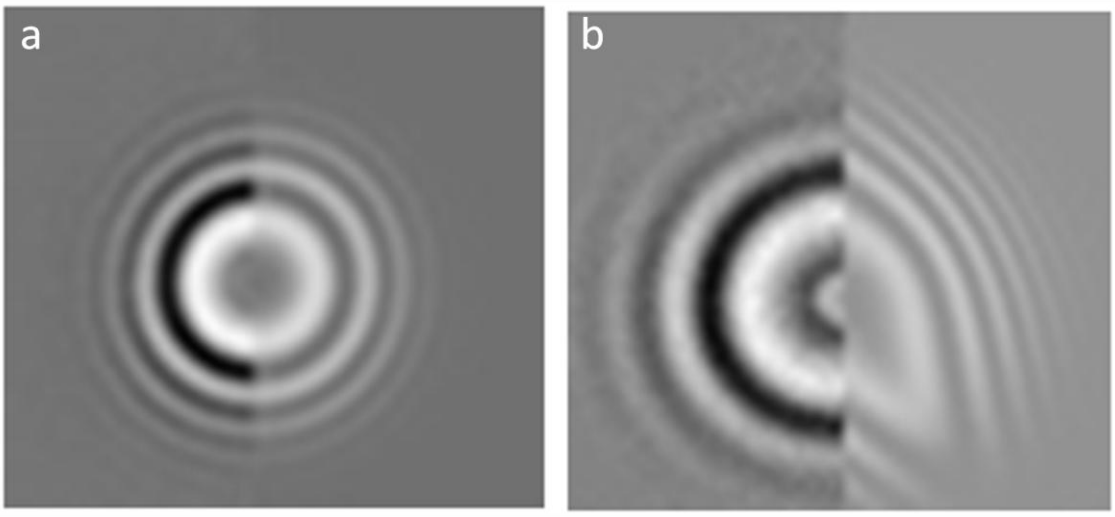

Fig. 31 (a) PSD at the beginning of the list, and (b) PSD at the end of the list

Fig. 31 (b) will have very high values for this criterion because we can clearly see that the theoretical PSD and the experimental PSD do not fit properly. This PSD will appear near the end of the list. The PSD in Fig. 31 (a) has a very good fitting so it will be placed at the beginning of the list.

4.3.1.5. *Fitting correlation between zeros 1 and 3*

4.3.1.5.1. Methodology

The region between the first and third zeroes is particularly important since it is where the Thon rings are most visible. This criterion reports the correlation between the experimental and theoretical estimation of PSDs within this region. High correlations indicate good fits.

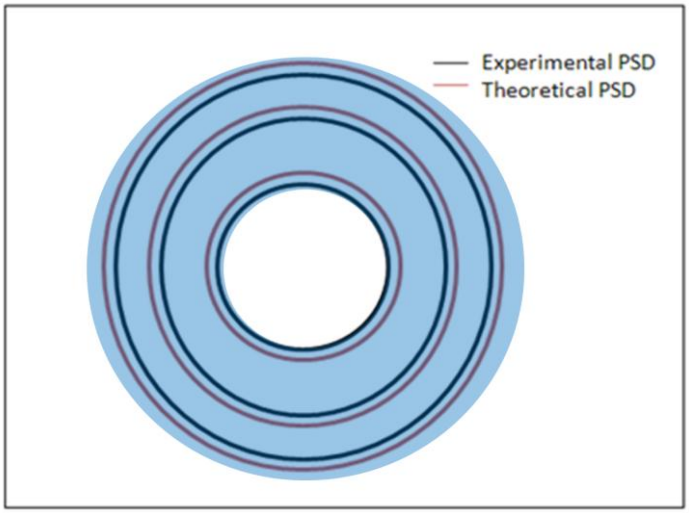

Fig. 32 Mask between zeros 1 and 3

The formula for this criterion will be the same formula as the *fitting score* criterion shown in Eq.36, but the range of the spatial frequency will be different.

$$corr13 = \sum_{R=R_1}^{R_3} |PSD_{theoretical}(R) - PSD_{experimental}(R)|^2 \quad (37)$$

where R_1 and R_3 are the spatial frequencies of the first and the third zero respectively.

4.3.1.5.2. Results

This criterion is a modified version of the *fitting score* criterion that gives very good results. At the beginning of the list will be the worst estimated CTFs.

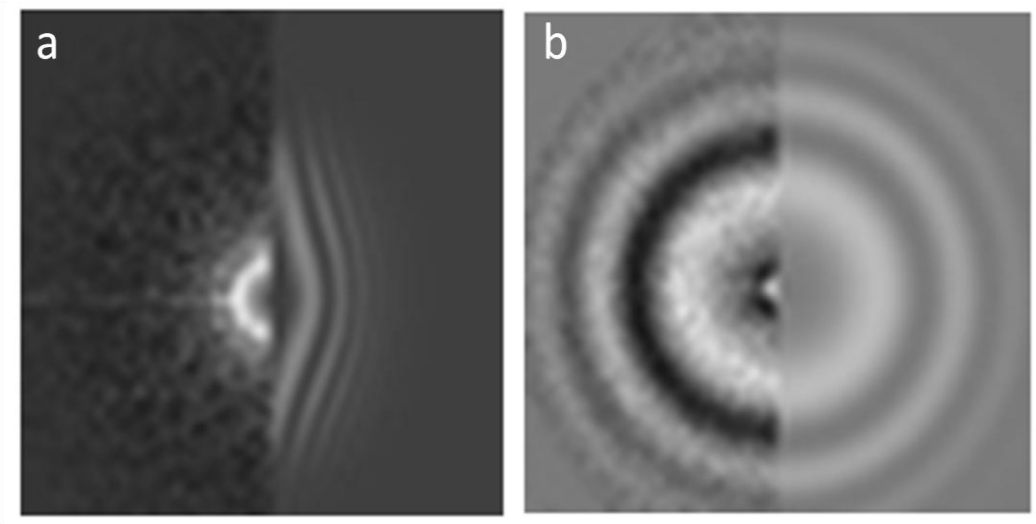

Fig. 33 (a) PSD at the beginning of the list, and (b) PSD at the end of the list

In Fig 33 (a), the theoretical PSD does not model well the rings 1 to 3. The theoretical PSD of Fig. 33 (b) does it correctly.

4.3.1.6. *PSD correlation at 90 degrees*

4.3.1.6.1. Methodology

The PSD of non-astigmatic micrographs correlate well with its rotated version by 90 degrees. This is so because non-astigmatic PSDs are circularly symmetrical, while astigmatic micrographs are elliptically symmetrical. High correlation when rotating 90 degrees is an indicator of non-astigmatism. This criterion is computed on the enhanced PSD.

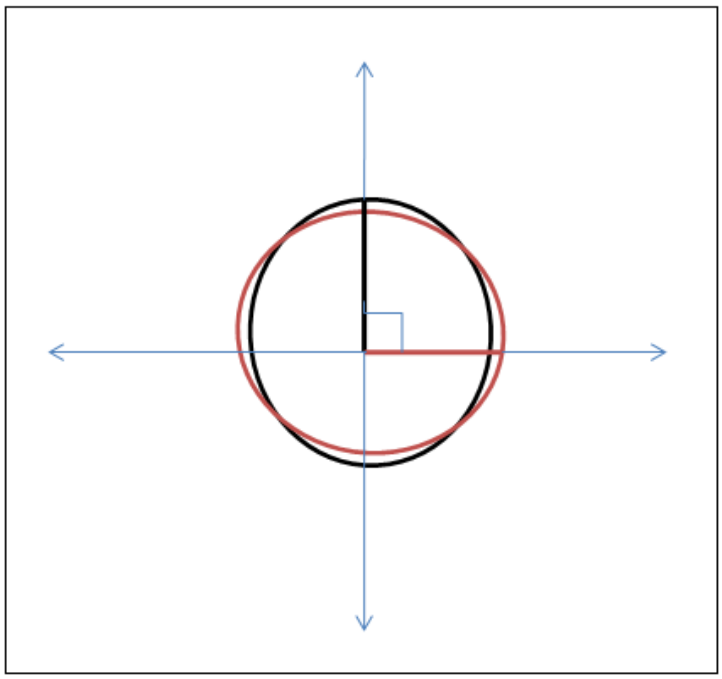

Fig. 34 The first ring of PSD vs the first ring of the same PSD rotated 90º

To perform this pair-wise comparison, we computed the normalized cross correlation (NCC) between these two images, which is a simple and fast computation. The NCC is defined as

$$NCC = \max_j \frac{\sum_k (f_k - \bar{f})(g_{k-j} - \bar{g})}{\sqrt{\sum_k (f_k - \bar{f})^2}\sqrt{\sum_k (g_k - \bar{g})^2}} \quad (38)$$

where f_k and g_k are the samples of two images at the pixel coordinate k, and $\bar{f}$ and $\bar{g}$ are the mean values of the corresponding images. The denominator in the equation

serves to normalize correlation coefficients such that $-1 \leq NCC \leq 1$, $NCC = 1$ indicating maximum correlation (here, ideally circular diffraction rings), $NCC = 0$ no correlation, $NCC = -1$ meaning that one image is the inverse of the other, and $-1 < NCC < 0$ meaning that one image has small values in the same part where the other image has large values. In the ideal case of perfectly circular rings whithout noise, the NCC depends neither on the number of rings nor on the contrast in the spectrum ($NCC = 1$ for any number of rings and for any contrast). In reality, however, noise in the spectra and imperfect circularity of the rings lead to different NCC values below 1.

4.3.1.6.2. Results

The PSD at the beginning of the list (Fig. 35 (a)) is visibly astigmatic whilst the PSD at the end of the list (Fig. 35 (b)) is circularly symmetrical.

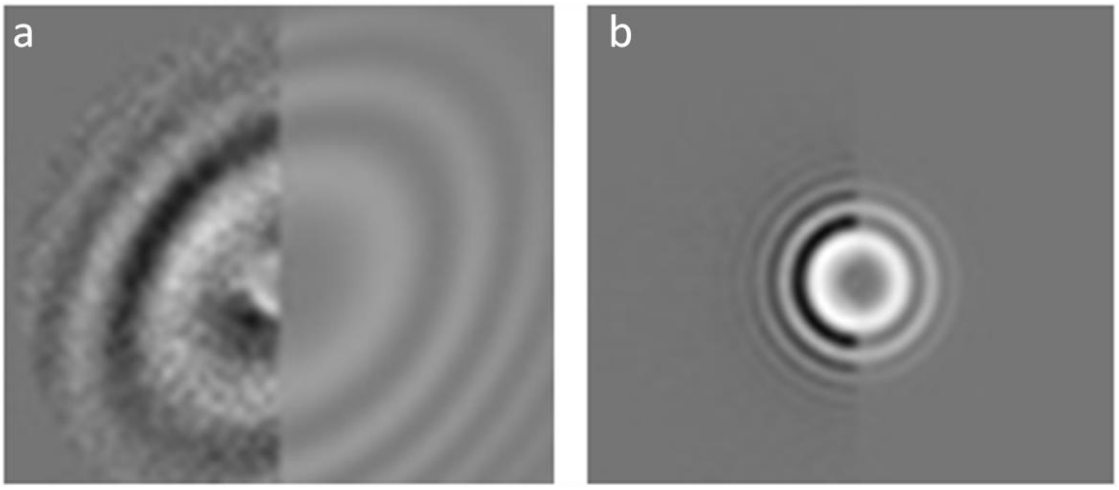

Fig. 35 (a) PSD at the beginning of the list, and (b) PSD at the end of the list

4.3.1.7. *PSD radial integral*

4.3.1.7.1. Methodology

This criterion reports the integral of the radially symmetrized PSD. This criterion can highlight differences among the background noises of PSD. This criterion is computed on the enhanced PSD.

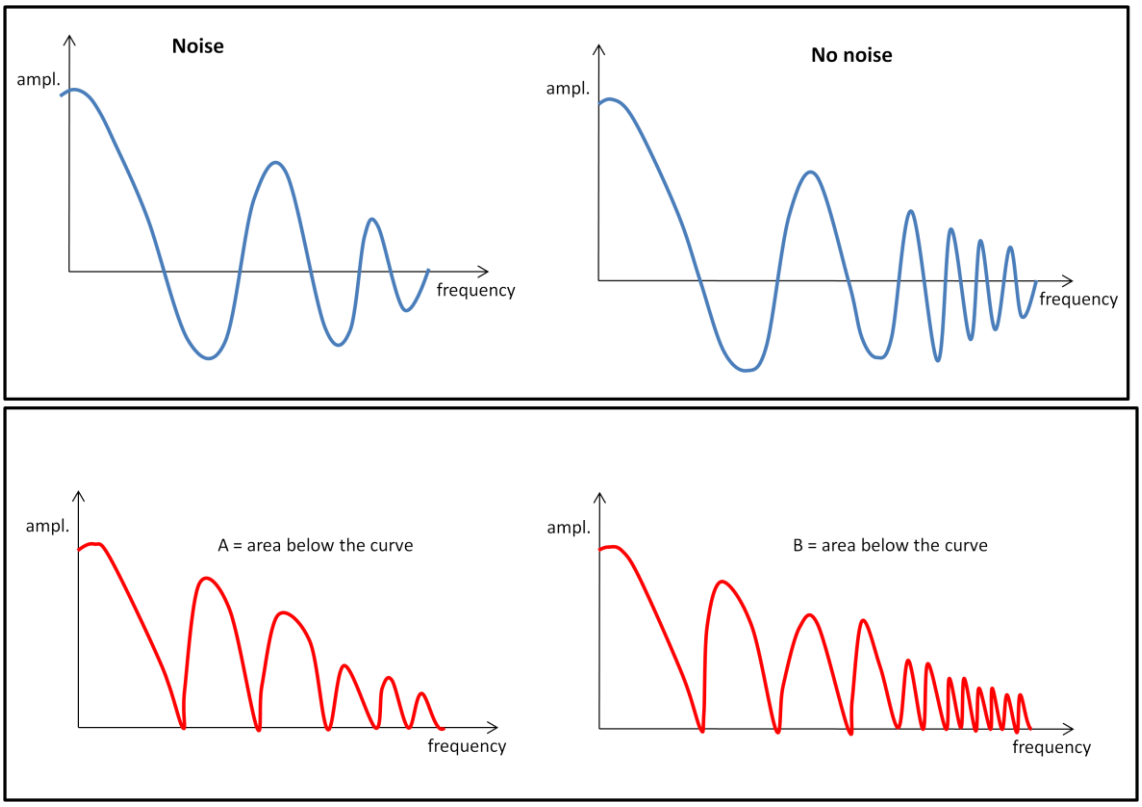

Fig. 36 PSD with no noise vs PSD with noise

In the diagram of Fig. 36 there is a graphical sketch (first two diagrams in blue) of the calculation of this criterion. The PSD with no noise (this is obviously an approximation as the PSD will always have some noise), will have a higher envelope than the PSD with noise. This is because at higher frequencies there is more correlation with noise, and if there is a lot of noise the values of the envelope at those frequencies won't be as visible as if there is no noise. If we sum the values of the absolute values of the envelope for each case we will obtain the area below the curve (named A for the PSD with no noise and B for the PSD with noise). We can observe that A > B, so this criterion will penalize noisy PSD as it is probable that they won't have as good estimation as a less noisy PSD.

This criterion is calculated the following way

$$radial\ integral = \sum_{R=0}^{R_{max}} \frac{1}{N_R} \left| \sum_{\alpha=0º}^{360º} PSD(A_\alpha R) \right| \tag{39}$$

Where α is the radial angle $= [0º, 360º]$, N_R is the total number of pixels in the image, and A_α is defined below

$$A_\alpha = \begin{bmatrix} \cos\alpha & sin\alpha \\ -sin\alpha & cos\alpha \end{bmatrix} \tag{40}$$

To compute the radial integral for all spatial frequencies, we need to define the maximum frequency value of axis *U* and axis *V*.

$$R_{max} = \begin{bmatrix} \frac{N}{2} \\ \frac{N}{2} \end{bmatrix} \tag{41}$$

4.3.1.7.2. Results

The bad visibility of the rings at high frequencies in Fig 37 (a) is due to background noise. This means that this PSD will be at the beginning of the list. The rings have a better definition in Fig. 37 (b). In this case, this criterion suggests that this PSD will be better adjusted to the theoretical model, therefore it will be placed near the end of the list.

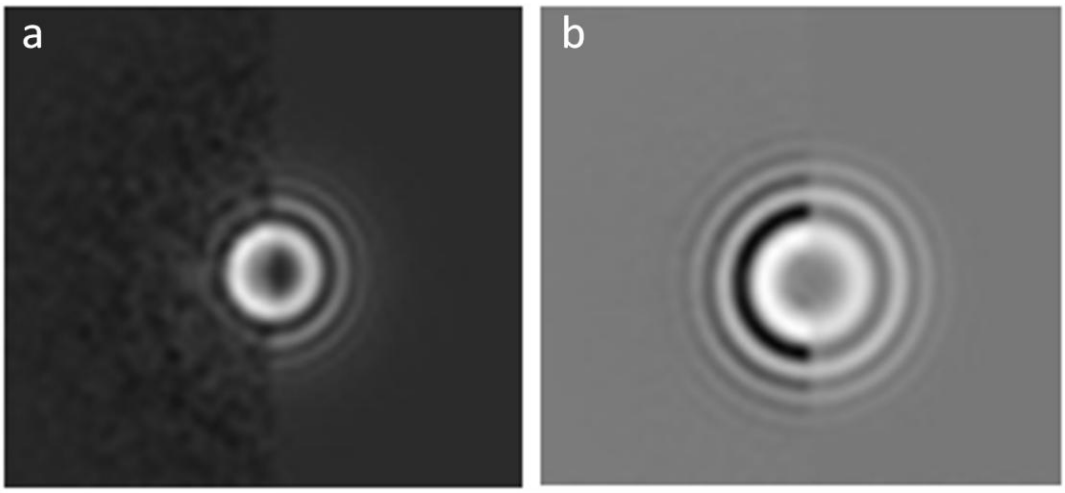

Fig. 37 (a) PSD at the beginning of the list, and (b) PSD near the end of the list

4.3.1.8. *PSD variance*

4.3.1.8.1. Methodology

The PSD is estimated by averaging different PSD local estimates in small regions of the micrograph. This criterion measures the variance of the different PSD local estimates. Untilted micrographs have equal defocus all over the micrograph, and therefore, the variance is due only to noise. However, tilted micrographs have an increased PSD variance since different regions of the micrograph have different defocus. Low variance of the PSD is indicative of non-tilted micrographs.

The formula of this criterion is similar to the formula of the variance.

$$variance = \frac{1}{N_{pieces}} \sum_{i=1}^{N_{pieces}} |PSD_i(R) - \overline{PSD}(R)|^2 \quad (42)$$

where i iterates through all the small regions of the PSD. N_{pieces} is therefore the number of regions in which we divide the PSD.

$\overline{PSD}(R)$ is the mean of the small regions of the PSD and is defined as

$$\overline{PSD}(R) = \frac{1}{N_{pieces}} \sum_{i=1}^{N_{pieces}} PSD_i(R) \quad (43)$$

4.3.1.8.2. Results

The PSD at the beginning of the list (Fig. 38 (a)) would correspond to a tilted micrograph and the PSD at the end of the list (Fig. 38 (b)) would correspond to a non-tilted micrograph.

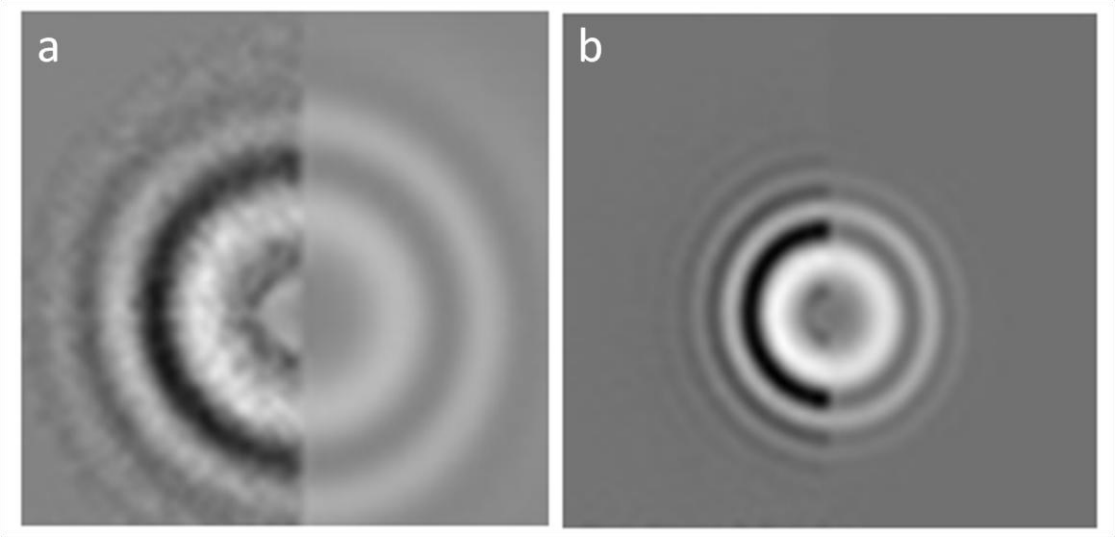

Fig. 38 (a) PSD at the beginning of the list, and (b) PSD at the end of the list

4.3.1.9. *PSD PCA Runs test*

4.3.1.9.1. Methodology

When computing the projections onto the first principal component, as discussed in the previous criterion, one might expect that the sign of the projection is random for untilted micrographs. Micrographs with a marked non-random pattern of projections are indicative of tilted micrographs. The larger the value of this criterion, the less random the pattern is.

For this criterion we will use the Principal component analysis (PCA) which is a mathematical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of uncorrelated variables called principal components. PCA is useful when you have obtained data on a number of variables (possibly a large number of variables), and believe that there is some redundancy in those variables. In this case, redundancy means that some of the variables are correlated with one another, possibly because they are measuring the same construct. Because of this redundancy, you believe that it should be possible to reduce the observed variables into a smaller number of principal components (artificial variables) that will account for most of the variance in the observed variables.

To characterize the trends exhibited by the set of data, PCA extracts directions where the cloud is more extended. For instance, if the cloud is shaped like an ellipse, the main direction of the data would be a midline or axis along the length of the ellipse. This is called the first component, or the principal component. PCA will then look for the next direction, orthogonal to the first one, reducing the multidimensional cloud into a two-dimensional space. The second component would be the axis along the ellipse width.

This transformation is defined in such a way that the first principal component has as high a variance as possible (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible. Principal components are guaranteed to be independent only if the data set is jointly normally distributed.

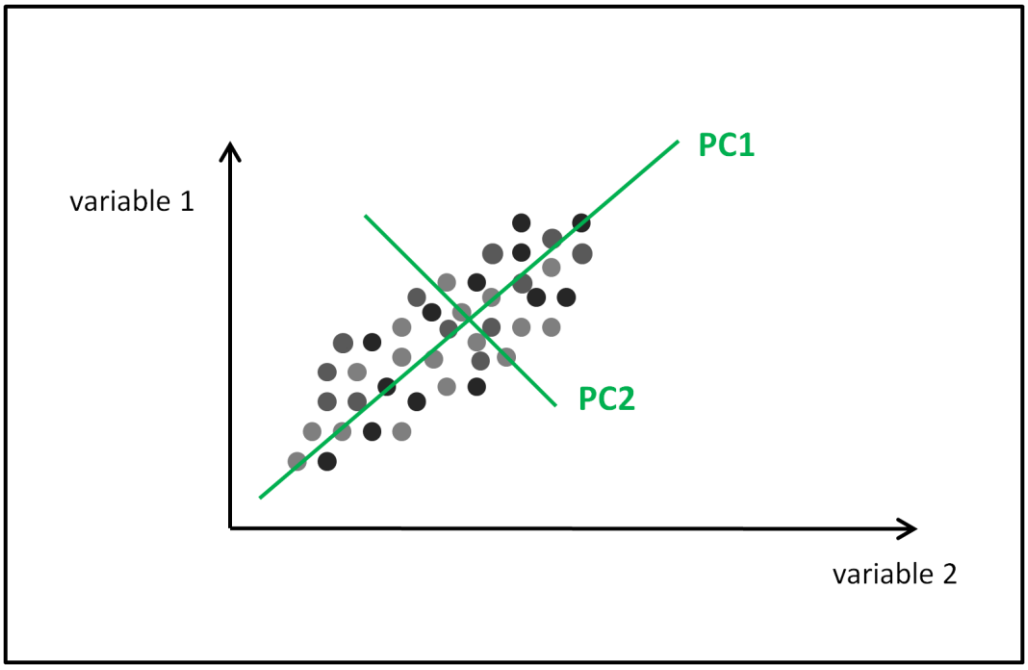

Fig. 39 Example of data set with two main components

This criterion is computed using the formula of the covariance to measure how much the first and the second components (PC1 and PC2 respectively, shown in Fig. 39) change together.

$$PCA\ Runs = \frac{1}{(R_{max} - 1)} \sum_{R=0}^{R_{max}} (|PSD_i(\overline{U}R) - \overline{PSD}(\overline{U}R)|^2)\ (|PSD_i(\overline{V}R) - \overline{PSD}(\overline{V}R)|^2) \quad (44)$$

where R_{max} is the maximum value of the spatial frequency along axis U (which is the large axis). $\overline{PSD}(R)$ is the mean of the small regions of the PSD and is defined in Eq. (43). $\overline{U}$ and $\overline{V}$ are unitary vectors corresponding to the directions of axis U and axis V respectively.

The covariance will provide a measure of the strength of the correlation between the two sets of random variables. For uncorrelated variables, the covariance is zero. However, if the variables are correlated, their covariance will be non-zero. The larger the value of this formula, traduces to a less random pattern in the micrograph and therefore there is a bigger possibility that the micrograph is tilted.

4.3.1.9.2. Results

The PSD at the beginning of the list (Fig. 40 (a)) would correspond to a tilted micrograph (micrographs with a marked non-random pattern) and the PSD at the end of the list (Fig. 40 (b)) would correspond to a non-tilted micrograph.

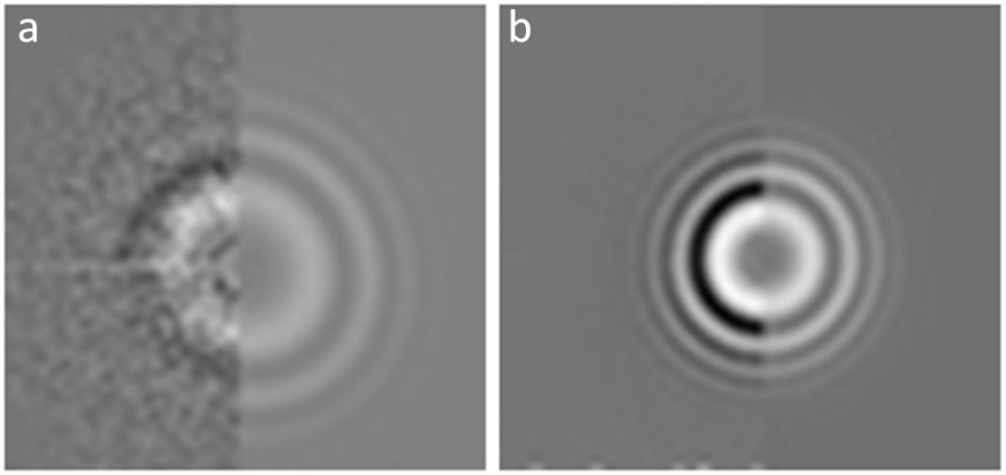

Fig. 40 (a) PSD at the beginning of the list, and (b) PSD at the end of the list

4.3.1.10. *Normality*

4.3.1.10.1. Methodology

There are some micrographs which present areas with a lot of very dark or very light tones due to ice imperfections, dust or simply because they are badly scanned. This fact can be used to design a new classification criterion. This criterion is going to be the normality test. Normality tests are used to determine whether a data set is well-modeled by a normal distribution or not. An informal approach to testing normality is to compare a histogram (gray-level quantization) of the sample data to a Gaussian function. The empirical distribution of the data (the histogram) should be bell-shaped and resemble the normal distribution.

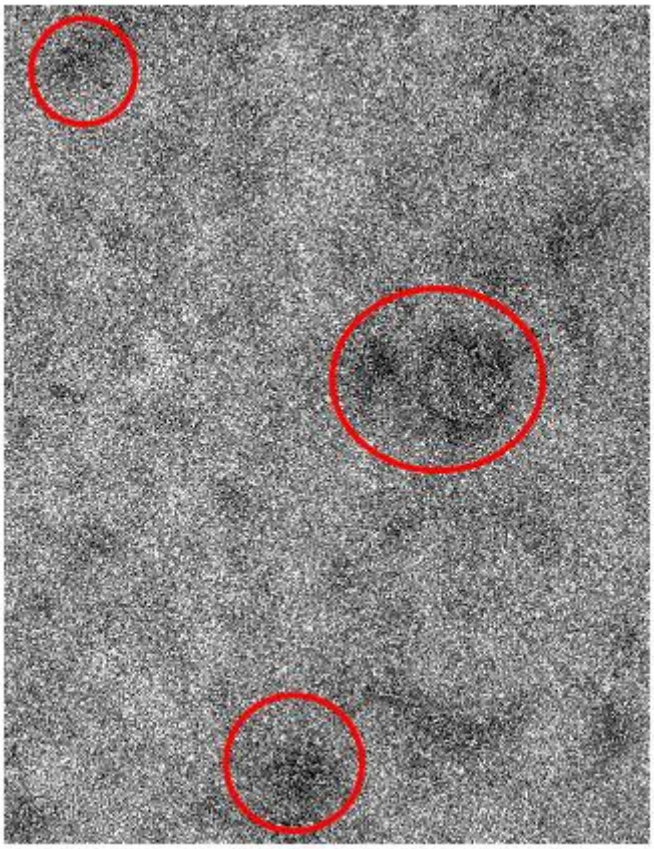

Fig. 41 Micrograph with darker tones outlined in red

In the case of the micrograph with a lot of very dark or very light tones, the histogram is non-Gaussian-like because it shows many gray levels concentrated near zero or many gray levels concentrated near the maximum value (255 for images scanned with 8 bits per pixel). This may typically result in 0.25% gray levels that are very dark or very light (Fig. 41).

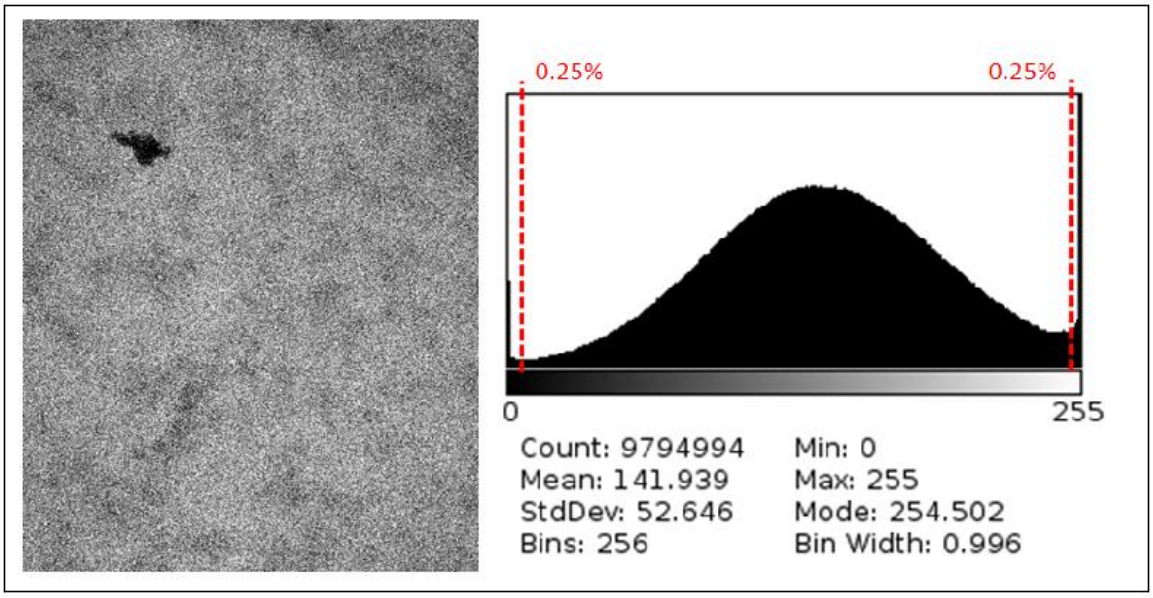

Fig. 42 A micrograph and its corresponding histogram

In the case of easily detectable very bright or dark regions in the micrograph, we may remove the regions because the particles are very low contrasted in such regions, which mean that around 0.25% of darkest gray values and 0.25% of brightest values in the histrogram would be replaced by the black value and the white value respectively. An equivalent approach would be to replace the values of the pixels in the micrograph without cutting it: the pixels with 0.25% of darkest gray values in the histogram would be replaced by the black pixel and the pixels with 0.25% of brightest values in the histogram would be replaced by the white value. This second method was implemented in the work here.

If we observe the two micrographs inf Fig. 43, we see how the one on the left hand side is more homogeneous in terms of its scale of grays, whilst the micrograph on the right hand side presents lighter tones in the image making it less homogeneous. This contrast will produce that the histogram of the micrograph will be less normal than the one of the other micrograph and therefore the PSD estimation will be worse.

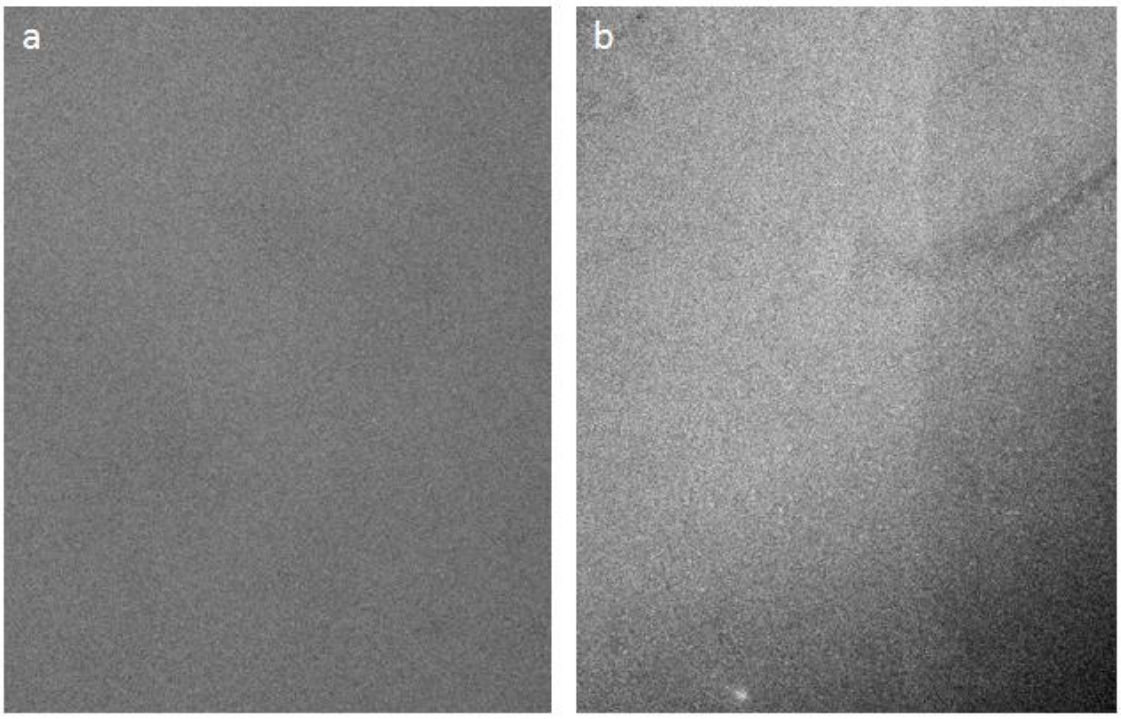

Fig. 43 (a) A homogeneous micrograph, (b) a micrograph with a less homogeneous aspect

The comparison of the histogram to the normal, Gaussian curve was used here for the classification of the PSDs. We will use the Kullback-Leibler divergence theory to define the normality test.

Kullback-Leibler divergence

For probability distributions P and Q of a discrete random variable their K–L divergence is defined to be

$$D_{KL}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)} \tag{45}$$

In words, it is the average of the logarithmic difference between the probabilities P and Q, where the average is taken using the probabilities P. The K-L divergence is only defined if P and Q both sum to 1 and if $Q(i) > 0$ for any i such that $P(i) > 0$. If the quantity $0log0$ appears in the formula, it is interpreted as zero.

One might be tempted to call it a "distance metric" on the space of probability distributions, but this would not be correct as the K-L divergence is not symmetric nor does it satisfy the triangle inequality. Still, being a premetric, it generates a topology on the space of generalized probability distributions, of which probability distributions proper are a special case.

To obtain a measure of the Kullback-Leibler divergence

$$dist(Q, P) = \frac{1}{2}\left(D_{KL}(P||Q) + D_{KL}(Q||P)\right) \tag{46}$$

In our specific problem, P is the histogram of the micrograph and Q is the probability density function of a Gaussian distribution of the same mean and variance as the micrograph.

As the K-L divergence is not symmetric, we can calculate the average between $D_{KL}(P||Q)$ and $D_{KL}(Q||P)$. If the K-L divergence is symmetric the average will be the same and this will mean that the micrograph has a Gaussian distribution. In the case in which the micrograph presents imperfections (such as dark regions) the divergence will not be symmetric and therefore the distance (i.e. our criterion) will be a large value.

4.3.1.10.2. Results

To show the results we will compare the two micrographs in Fig. 43. The histograms of the micrographs have different normality.

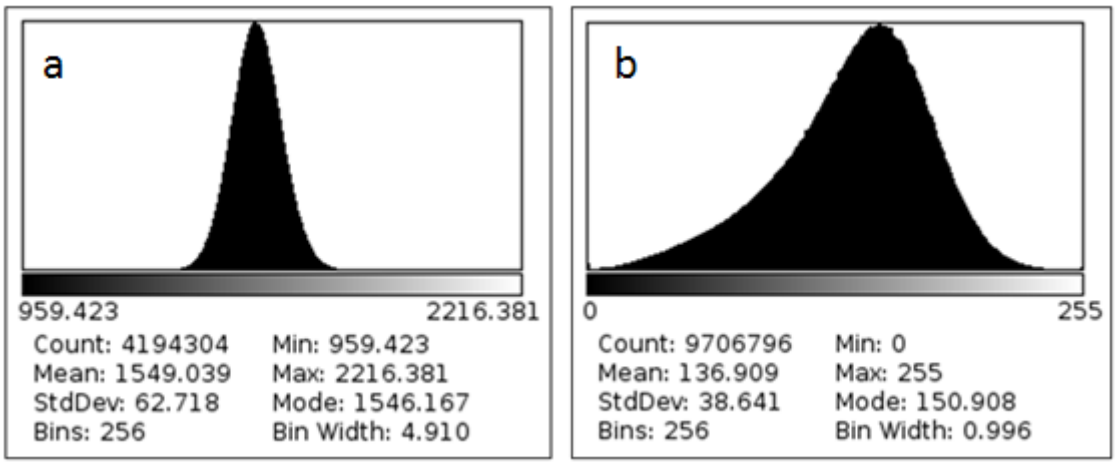

Fig. 44 (a) Histogram of micrograph in Fig. 43 (a), (b) histogram of micrograph in Fig. 43 (b)

The histogram of the more homogeneous micrograph (Fig. 43 (a)), which is shown in Fig. 44 (a), presents a more normal distribution than the other one. This means that when comparing the histogram to the normal in the normality test, the micrograph of Fig. 43 (a) will give better results.

The PSD of Fig. 45 (a) corresponds to the more homogeneous micrograph. We can observe that this PSD has been better fitting. We can also see that its enhanced PSD has better quality than the enhanced PSD of Fig. 45 (b), which corresponds to the less homogeneous micrograph. The high quality of the experimental PSD contributes to make better estimation of the PSD.

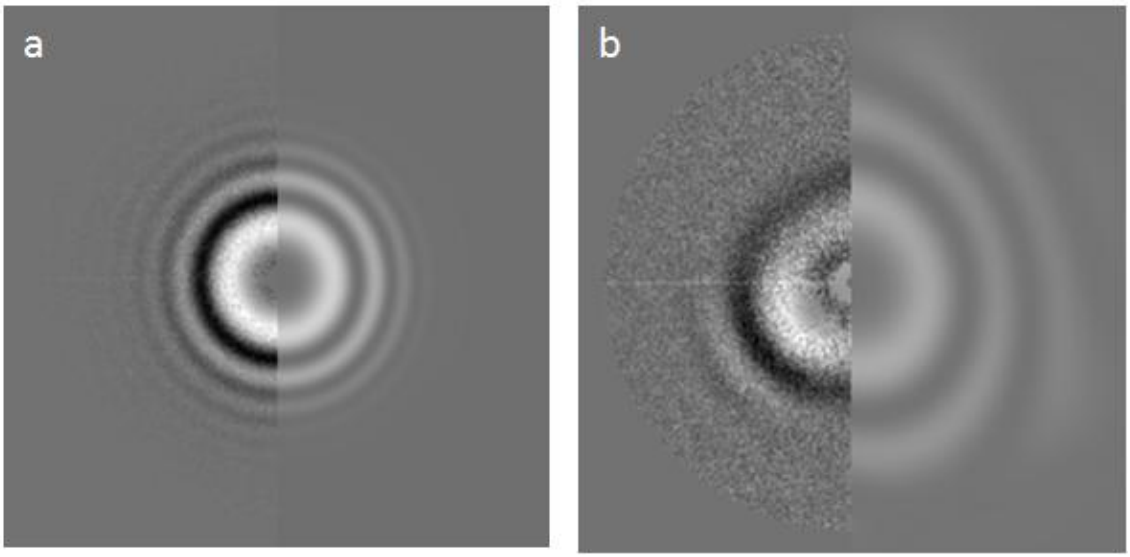

Fig. 45 (a) PSD of micrograph Fig. 44 (a), (b) PSD of micrograph Fig. 44 (b)

The PSD of Fig. 45 (b) is drifted (its rings fade in a particular direction Fig. 20 (c)). This is due to a default when producing the micrograph. Drifted micrographs are non-stationary signals. A signal can be considered stationary in the wide sense, if the two following criteria are met:

1. The mean values or expectations of the signal are constant for any shift in time.
2. The autocorrelation function is also constant over an arbitrary time shift.

This means that when calculating the histogram, there will be different histograms over time shift.

In the same PSD we can also see that the damping is not properly estimated (we observe that the fluctuation of zeros continue more than it should). This is because when we calculate the damping it is supposed that micrograph is a stationary signal, and this one, as we said before, is not.

In the annex (A.14) there is a partial representation of some PSDs sorted with this criterion. We concluded that the classification was good and the criterion was therefore implemented. This new criterion was added to the class *ctf_sort_psds.cpp*.

4.3.1.11. Conclusions

Analyzing the classifications of all criteria, we take as the best sorting criteria:

- *First zero average*: it produces a fairly good sorting.
- *Fitting correlation between zeros 1 and 3*: also generates a good sorting.
- *PSD correlation at 90 degrees*: this criterion is the best one. The sorting actually makes it possible for the user to select the PSD estimates from the point where he/she considers that the estimations are beginning to be acceptable. Of course, this is done visually and it is an approximation, but it is definitely a very useful tool.

Finally, we conclude that although there are better sorting methods than others, all criteria are valid and can be useful to process micrographs, therefore all criteria will be introduced in the class *ctf_sort_psds.cpp*.

4.3.2. Combined criterion

4.3.2.1. Methodology

We will now make another classification based on the three criteria that had the best sorting. We are going to work only with the good micrographs sorted with *PSD correlation at 90 degrees*. We remove the badly estimated PSDs (we determine visually at which point the PSDs start to be badly estimated). We will work with 243 micrographs out of our set of 753. A representative extract of our set of data is show in the annex (A.10).

Our first attempt was to sort these micrographs according to f*irst zero average* and *fitting correlation between zeros 1 and 3* criteria to see if we can improve our classification but there were no decisive results.

PSD correlation at 90 degrees is a good start, but the problem with this criterion is that it only takes into account the experimental part of the micrographs (it rotates de experimental PSD 90 degrees and compares de image result with the original version).This is actually measuring the astigmatism, which is definitely a good measure to select a threshold to identify good images, but we somehow also need to take into account a test for the actual achievement of the theoretical part of the micrograph.

We classified the whole set of 753 micrographs (that is including the bad ones also) the following way:

- Descending order of the average of the scores obtained by sorting using *PSD correlation at 90 degrees*, *first zero average* and *fitting correlation between zeros 1 and 3*: first we sort the set of micrographs according to the value obtained by each criterion in descending order. We will have three classifications. Then we calculate for each micrograph the average position it holds in each list. The resulting number can be real or integer. In the end, all micrographs will have a number assigned. The set of micrographs will be sorted according to the descending order of those values.

- Descending order of the standard deviation of the scores obtained by sorting using *PSD correlation at 90 degrees*, *first zero average* and *fitting correlation between zeros 1 and 3*: this method follows the same steps as the above, but instead of calculating the average it will calculate the standard deviation of the positions the micrograph has in each list.

For this task we used GNU Octave, a program for solving numerical problems similar to Matlab. In the code, our input data is *firstZero*, *corr13* and *psdcorr90* (the three classifications), and the outputs are *avg_comb* (average combination) and *std_comb* (standard deviation).

```
load firstZero.txt
load corr13.txt
load psdcorr90.txt

crit1 = firstZero;
crit2 = corr13;
crit3 = psdcorr90;
critnum = 3;

imgnum = length(crit1);

[sort1,ind1] = sort(crit1);
[sort2,ind2] = sort(crit2);
[sort3,ind3] = sort(crit3);

n = (1:1:imgnum)';
temp1=[ind1 n];
temp2=[ind2 n];
temp3=[ind3 n];

[stemp1,itemp1] = sort(temp1(:,1));
[stemp2,itemp2] = sort(temp2(:,1));
[stemp3,itemp3] = sort(temp3(:,1));

mat1 = temp1(itemp1,:);
mat2 = temp2(itemp2,:);
mat3 = temp3(itemp3,:);

avg_comb = (mat1(:,2) + mat2(:,2) + mat3(:,2))/critnum;
std_comb = sqrt(((mat1(:,2)-avg_comb).^2 + (mat2(:,2)-avg_comb).^2
+ (mat3(:,2)-avg_comb).^2)./(critnum-1));
```

Fig. 46 GNU Octave code for average and standard deviation sorting

After comparing the results obtained with two methods, we concluded that the best method was the average average between *PSD correlation at 90 degrees*, *first zero average* and *fitting correlation between zeros 1 and 3*. The results of only that method will be showed.

4.3.2.2 Results

As we said before, the descending order of average between *PSD correlation at 90 degrees*, *first zero average* and *fitting correlation between zeros 1 and 3* is the best sorting out of all possibilities. In the Fig. 47 we see the best PSD estimations in the top row and the worst PSD estimations in the bottom row obtained by this method.

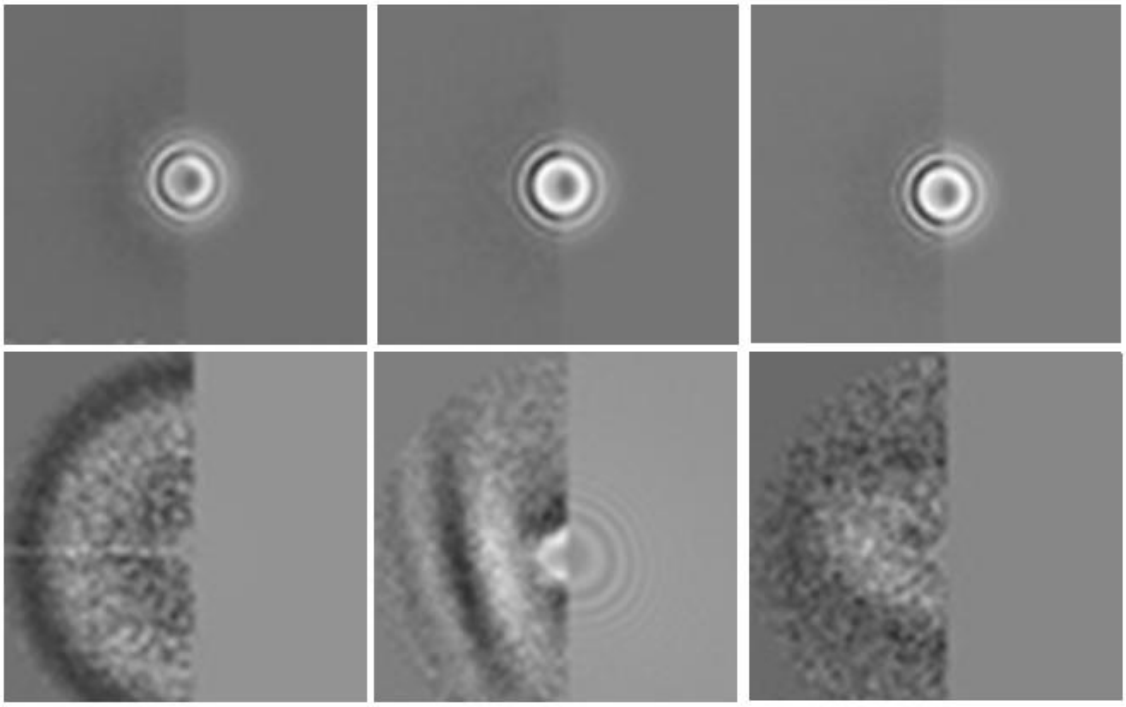

Fig. 47 Top row corresponds to first 3 PSDs of the list and bottom row corresponds to last 3 PSDs of the list

An extract of this classification is shown in the annex (A.11). This sorting criterion is very good, in fact, we can determine a threshold visually when we see that the CTF estimation starts to fail, and there will be very few incorrect CTF estimations above this threshold.

After deciding that the average is a good method for classifying, we tried to use it with other criteria to see if this way we get better results. We determined a new threshold visually and made a new file with only good images. An extract of only the good images is shown in the annex (A.12).

Our initial criterion is the average between *PSD correlation at 90 degrees, first zero average* and *fitting correlation between zeros 1 and 3*. The next classifications that we did and their results are described below:

- Average of *PSD correlation at 90 degrees* and *fitting correlation between zeros 1 and 3*: this sorting was almost the same as the initial criterion.
- Average of *fitting correlation between zeros 1 and 3* and *first zero ratio*: it does not provide good sorting.
- Average of *PSD correlation at 90 degrees, fitting correlation between zeros 1 and 3* and *first zero ratio*: it seems it works although it can be used only if we want to keep astigmatic images because astigmatic and non-astigmatic images are mixed.

We also tried *PSD correlation at 90 degrees* and *fitting correlation between zeros 1 and 3* with other criteria as follows:

- Average of *PSD correlation at 90 degrees, fitting correlation between zeros 1 and 3* and *damping*: it does not provide a good sorting; it gives a mixture of PSD.
- Average of *PSD correlation at 90 degrees, fitting correlation between zeros 1 and 3* and *psdint*: it does not provide a good sorting; it gives a mixture of PSD.
- Average of *PSD correlation at 90 degrees, fitting correlation between zeros 1 and 3* and *PSD variance*: it does not provide a good sorting; it gives a mixture of PSD.
- Average of *PSD correlation at 90 degrees, fitting correlation between zeros 1 and 3* and *PSD PCA Runs test*: it does not provide a good sorting; it gives a mixture of PSD.

After this exhaustive search for the best combination of sorting criteria, we decided to implement in C++ the combination of *PSD correlation at 90 degrees* and *fitting correlation between zeros 1 and 3.* To this goal, we translated the Octave code into C++ and added it to the class *ctf_sort_psds.cpp*.

4.3.3. The graphical interface

A graphical interface was developed using ImageJ (a Java based image processing program of public domain) to sort micrographs according to all criteria described in section 4.3. The idea is to show the PSD and the CTF estimation of all the set of micrographs in a table which will also contain the values of each criterion for the micrographs. An extract of the interface is shown in Fig. 48.

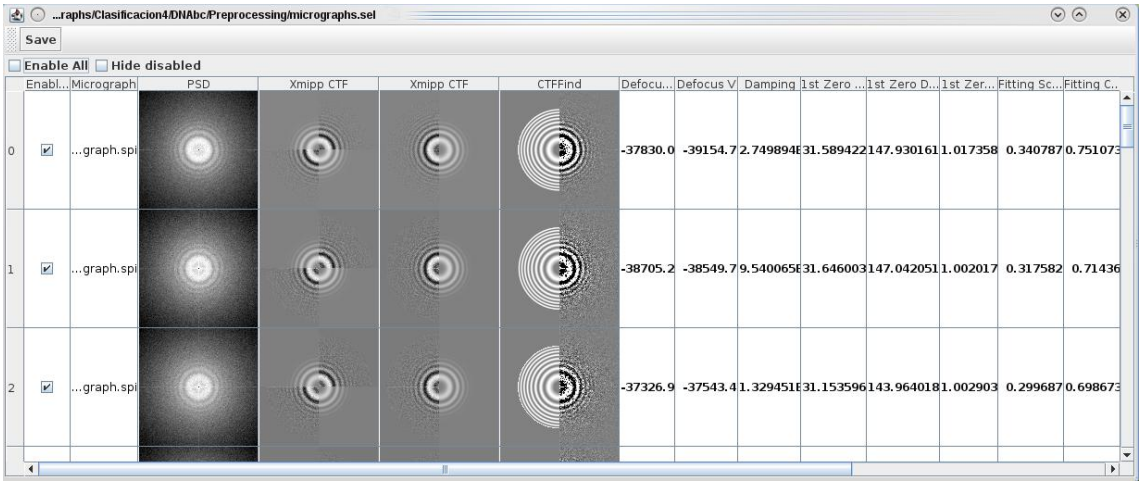

Fig. 48 Graphical interface

The user can now sort the micrographs according to one of the criteria by clicking on the top of the column of that criterion. There is also a possibility to discard or select micrographs.

An example of how well this interface works is now described. We start with our set of 753 micrographs. We sort them by the combination criterion. We determine a threshold visually and discard the micrographs below that threshold (badly estimated PSDs). We continue working only with 174 good images (only perfect circular symmetric PSDs). In the annex (A.12) there is an extract of these good images. Now we sort this set of micrographs according to each individual criterion to see its effect: *damping*, *first zero average*, *first zero ratio*, *PSD radial integral* and *PSD PCA Runs test*. There is no good or bad sorting but perhaps we could say that *damping* and *PSD radial integral* criteria are very similar and give good sorting in the sense that one can visualize the rings of the CTF going from more to less concentrated around the center. In the annex (A.13) there is an extract of this sorting.

Regarding the results, we can conclude that it is very useful to have a flexible interface in which the user can decide the criterion to classify the set of micrographs, and also be able to discard the ones that he/she does not want to keep.

4.4. Manual initialization

Once we processed all the micrographs, we observed that when sorting with the *first zero average* criterion, shown in annex (A.2), at the beginning of the list there are several PSDs that are not calculated at all, i.e. there is no image at all in the right hand side of the half plane. These PSDs should have been calculated because their experimental PSDs are good.

As we can see in the image below, the experimental PSD is fairly good but somehow the estimated PSD was not calculated at all.

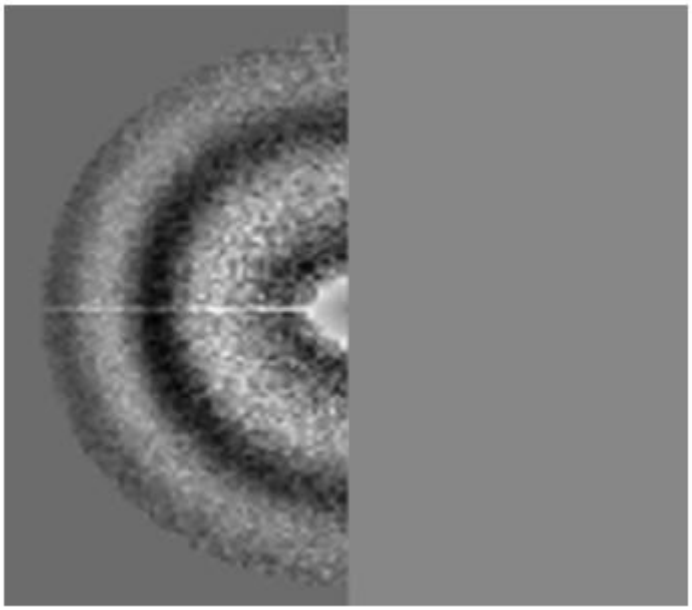

Fig. 49 Enhanced PSD vs estimated PSD of micrograph

The table shown in Fig. 50 is the parameters of the CTF calculation that failed (Fig. 49).

```
# XMIPP_STAR_1 *
#
data_
 _CTF_Sampling_rate 4.2
 _CTF_Voltage 100
 _CTF_Defocus_U -4299.8
 _CTF_Defocus_V -5783.5
 _CTF_Defocus_angle 75.0246
 _CTF_Spherical_aberration 5.6
 _CTF_Chromatic_aberration 3
 _CTF_Energy_loss 0.0204536
 _CTF_Lens_stability 0
 _CTF_Convergence_cone 4.72499e-12
 _CTF_Longitudinal_displacement 0
 _CTF_Transversal_displacement 0
 _CTF_Q0 -0.115235
 _CTF_K 1.28453
 _CTFBG_Gaussian_K 2.2801
 _CTFBG_Gaussian_SigmaU 100000
 _CTFBG_Gaussian_SigmaV 25000
 _CTFBG_Gaussian_CU 0.00709218
 _CTFBG_Gaussian_CV 0.00701788
 _CTFBG_Gaussian_Angle 85.4425
 _CTFBG_Sqrt_K 15.4447
 _CTFBG_Sqrt_U 14.3453
 _CTFBG_Sqrt_V 13.0065
```

Fig. 50 File of parameters of micrograph

We can see that the *defocus_U* is around 4000Å but the estimation is incorrect (bad correspondence with enhanced PSD), so we are going to re-estimate the PSD giving an initial defocus value 5000 Å. We can observe that the results are now good (Fig 51).

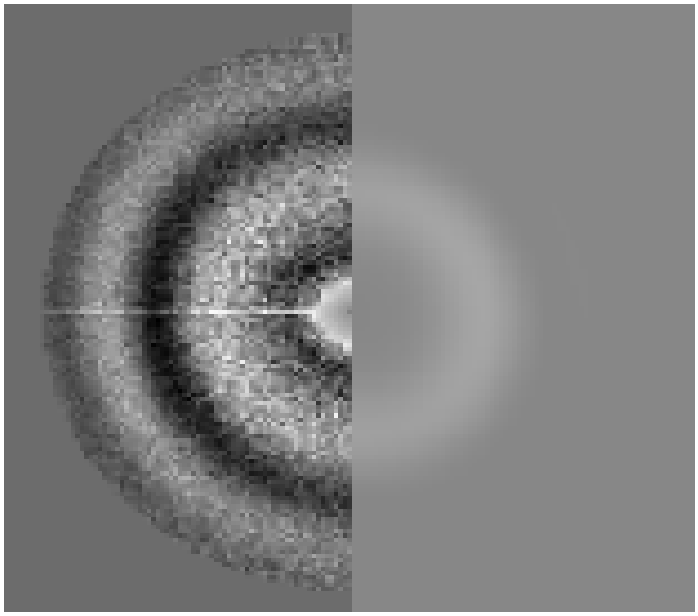

Fig. 51 Enhanced PSD vs estimated PSD of micrograph with initial defocus value of 5000Å

The conclusion wass to introduce in the CTF classification graphical interface the refinement of the CTF parameters strating from some initial defocus value specified by the user in order to reduce the number of CTF estimation failures.

4.4.1. Methodology

A good initial estimation for the defocus is the first zero of the CTF. The problem is that the value of the first zero will be given in pixels and the defocus is measured in Angstroms. We will therefore need to make a conversion of units.

For this we need to look at the mathematical formula which contains the defocus.

$$\chi(R) = \pi\,\lambda\left(|\Delta f(R)||R|^2 + \frac{1}{2}\,C_s\,|R|^4\,\lambda^2\right) \qquad (47)$$

C_s represents the spherical aberration coefficient, R ϵ $\mathbb{R}^2$ is a given spatial frequency, $\Delta f(R)$ is the defocus vector, and λ is the electron wavelength which is computed as in Eq. (7).

A typical microscope has a frequency response H_{ideal} detailed in Eq. (4).
We make $H_{ideal}(R) = 0$ and obtain the equation for $\chi(R)$ (Eq. (5)).

$$-\big((\sin\big(\chi(R)\big) + Q(R)\cos\big(\chi(R)\big)\big) = 0 \tag{48}$$

$$Q(R) = -\frac{\cos\big(\chi(R)\big)}{\sin\big(\chi(R)\big)} \tag{49}$$

$$\chi(R) = \operatorname{atan}\big(-Q(R)\big) \tag{50}$$

When we tested the equation we realized we needed to add π to $\chi(R)$ to get the value of the periodic function that matches the solution. The period of trigonometric functions is usually 2π, but in the case of the tangent the period is π, so we will adjust it now

$$\chi(R) = \operatorname{atan}\big(-Q(R)\big) - \pi \tag{51}$$

We will now obtain the equation of the defocus from the other formula and substitute $\chi(R)$

$$|\Delta f(R)| = \frac{\chi(R)}{\pi\lambda\,|R|^2} - \left(\frac{1}{2}\,C_s\,|R|^2\,\lambda^2\right) \tag{52}$$

$$|\Delta f(R)| = \frac{\operatorname{atan}\big(-Q(R)\big)}{\pi\lambda\,|R|^2} - \left(\frac{1}{\lambda\,|R|^2}\right) - \left(\frac{1}{2}\,C_s\,|R|^2\,\lambda^2\right) \tag{53}$$

This is the formula for the conversion of units. The method will consist in measuring the first zero in pixels and applying the conversion formula to obtain the initial estimation value of the defocus.

4.4.2. Results

To validate this method we will use the formula for the conversion of units to calculate the defocus of a PSD of a micrograph and check if it is correctly calculated by comparing it with the value calculated by the CTF estimation program.

This is the PSD we are going to use to check our equation. The theoretical PSD has to be well calculated so that we are able to compare our result with the correct value.

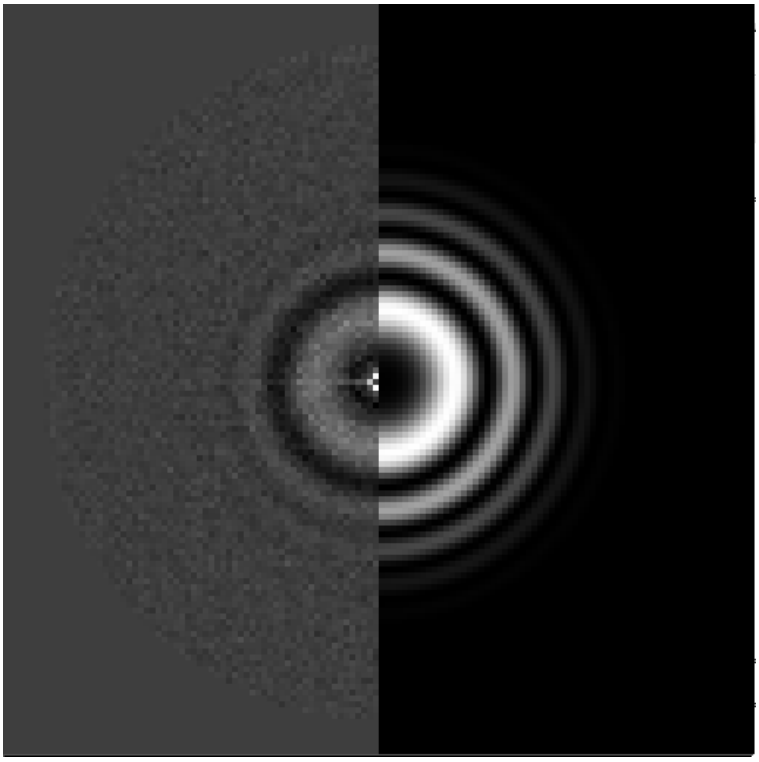

Fig. 52 PSD of micrograph with defocus value of 4000Å

To obtain the parameters we need for our formula we will check the file of the CTF parameters calculated by the program.

```
# XMIPP_STAR_1 *
#
data_
 _CTF_Sampling_rate 1.4
 _CTF_Voltage 200
 _CTF_Defocus_U -3831.98
 _CTF_Defocus_V -3979.66
 _CTF_Defocus_angle -25.3597
 _CTF_Spherical_aberration 2.26
 _CTF_Chromatic_aberration 3
 _CTF_Energy_loss 0.00762972
 _CTF_Lens_stability 0
 _CTF_Convergence_cone 3.21106e-10
 _CTF_Longitudinal_displacement 0
 _CTF_Transversal_displacement 0
 _CTF_Q0 -0.07
 _CTF_K 0.647649
 _CTFBG_Gaussian_K 1.39655
 _CTFBG_Gaussian_SigmaU 1457.08
 _CTFBG_Gaussian_SigmaV 2933.94
 _CTFBG_Gaussian_CU 0.0569474
 _CTFBG_Gaussian_CV 0.0589674
 _CTFBG_Gaussian_Angle 1.06951e-10
 _CTFBG_Sqrt_K 14.0317
 _CTFBG_Sqrt_U 11.1941
 _CTFBG_Sqrt_V 9.57745
```

Fig. 53 File of parameters of micrograph

From this information we obtain the values of the parameters we need.

$$|\Delta f(R)| = \frac{\operatorname{atan}\left(-Q(R)\right)}{\pi\lambda\,|R|^2} - \left(\frac{1}{\lambda\,|R|^2}\right) - \left(\frac{1}{2}\,C_s\,|R|^2\,\lambda^2\right) \qquad (54)$$

$Q(R) = -0.07$

$V = 200\ kV$

$C_s = 2.26$

$T_s = 1.4$

$$R = \frac{|r(x,y)|}{|w(x,y)|\,T_s} \tag{55}$$

where $|r(x,y)|$ is the estimation of the first zero given in pixels, $|w(x,y)|$ is the length of the window of the PSD given in pixels, and T_s is the sampling period.

We need to calculate the estimation of the first zero by measuring it in the PSD.

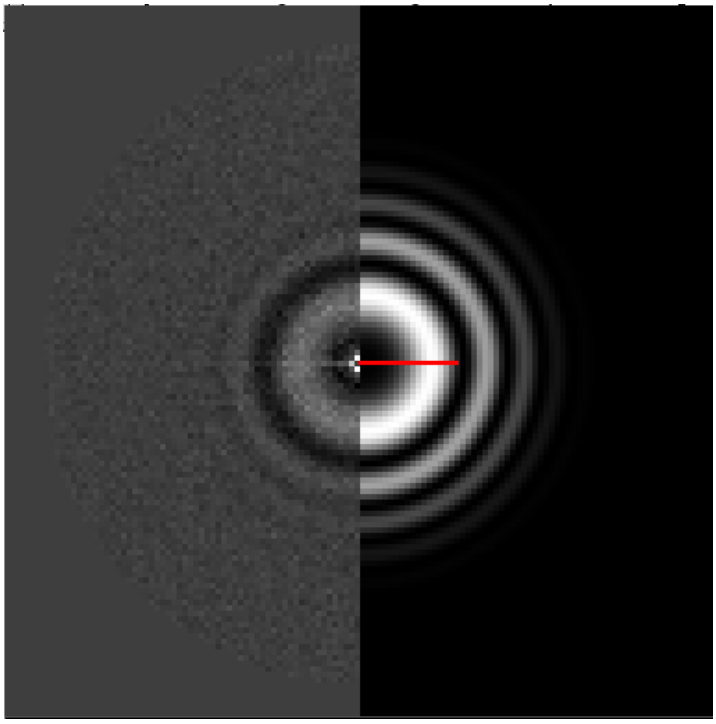

Fig. 54 Measure of the first zero of the PSD

We obtain that $|r(x,y)| = 18\,pixels$ and we also measure the window size, which is 128 x 128, so $|w(x,y)| = 128$.

$$R = \frac{|r(x,y)|}{|w(x,y)|\,T_s} = \frac{18}{128\,x\;1.4} = 0.1004$$

We also know the parameters to calculate λ

$$\lambda = \frac{12.3}{\sqrt{V + 10^{-6}\,V^2}} = \frac{12.3}{\sqrt{200\,10^3 + 10^{-6}\,(200\,10^3)^2}} = 0.025107$$

Substituting the values in Eq. (54) we obtain the following results

$$|\Delta f(R)| = \frac{\operatorname{atan}(0.07)}{\pi\,(0.025107)\,(0.1004)^2} - \left(\frac{1}{(0.025107)\,(0.1004)^2}\right) - \left(\frac{1}{2}\,(2.26)\,(0.1004)^2\,(0.025107)^2\right)$$

$$= -3.8598\,10^3 \text{ Angstroms}$$

If we compare the value obtained for the defocus with the *defocus_U* in the parameter file we can check that our calculations are similar so we can conclude that is works correctly.

$|\Delta f(R)| = -3859.8$ Angstroms
defocus_U $= -3831.98$ Angstroms

The evaluation criterion will show how good our implementation is. When we execute the CTF calculation of the 753 micrographs we now have:

- Micrographs GG – 655 micrographs
- Micrographs GB – 25 micrographs
- Micrographs BB – 73 micrographs

When we apply the formula we obtain the correctness of η = 96%. This is a great improvement.

We conclude that manual initialization of the CTF estimation is very useful in order to recalculate PSDs that have not been calculated at all. Its implementation is in the class *EllipseCTF.java*.

4.4.3. The graphical interface

It is not easy to estimate a good initial defocus for the user, so we will let the user watch the PSD and trace an ellipse around the first zero. The class *EllipseFitter.java* will generate the best fitting ellipse to our tracing. What constitutes the best fitting ellipse? First, it should have the same area as the tracing. In statistics, the measure that attempts to characterize some two-dimensional distribution of data points is the 'ellipse of concentration' (see Cramer, Mathematical Methods of Statistics, Princeton Univ. Press, 945, page 283). This measure equates the second order central moments of the ellipse to those of the distribution, and thereby effectively defines both the shape and size of the ellipse. This class will return the parameters to define the ellipse: length of major axis, length of minor axis, center of ellipse and angle between the x-axis and the major axis.

We normally calculate the defocus in two radial directions (although the defocus can be calculated in any radial direction). The major semi axis of the ellipse will be used to estimate the *defocus_V* and the minor semi axis of the ellipse will be used to estimate the *defocus_U* (Fig. 55). We chose an ellipse instead of a circle to determine the first zero because the rings of the PSD always have its *defocus_V* and *defocus_U* slightly different (if they have a lot of difference the PSD would be astigmatic).

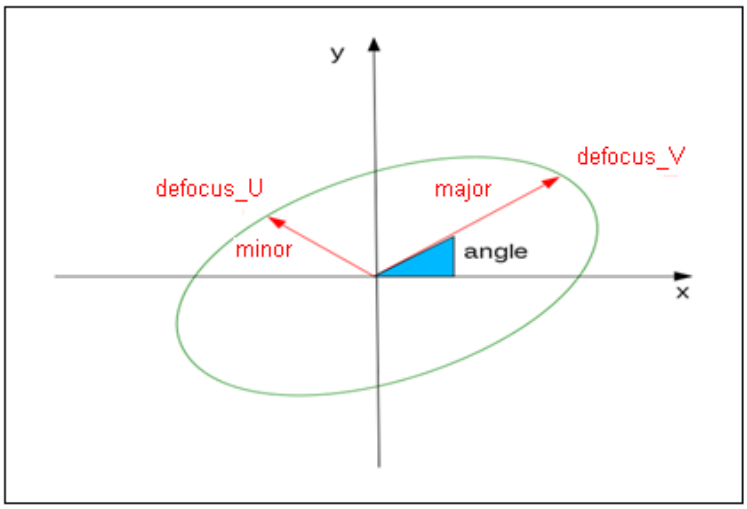

Fig. 55 Diagram of ellipse with *defocus_V* and *defocus_U*

The program will calculate *defocus_V* and *defocus_U* and and use them to recalculate the CTF parameters. In Fig. 56 we can see how the ellipse is fitted to give an initial value of the defocus to recalculate the CTF parameters of a micrograph which was not calculated.

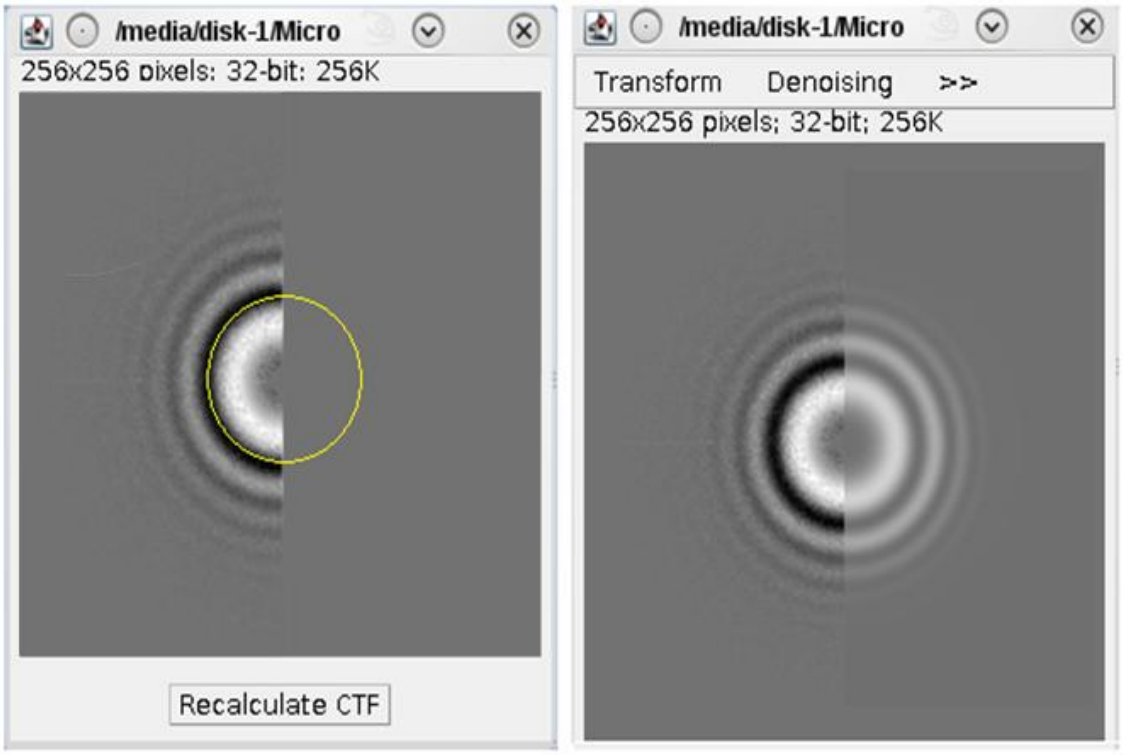

Fig. 56 Recalculate CTF of micrograph

5. CONCLUSION

In this study we analized the methodology of the estimation of CTF parameters in micrographs used in XMIPP and explained in article [2]. The aim was to detect and correct badly estimated PSDs so that the process is fast and robust.

We worked with a set of 753 micrographs to make a first evaluation of the CTF estimation of parameters. To assess the program we used a formula to evaluate for how many micrographs with good enhanced experimental PSDs we had correct estimations of theoretical PSDs. After obtaining a 71% we developed three types of methodologies to rise this percentage.

The first method is focused on improving the CTF estimation and it managed to rise the percentage up to 84%. The second method was introduced to detect badly estimated PSDs by generating classifications with several criteria and their combinations. The third method intends to recalculate badly estimated PSDs and it resulted in the correctness evaluation of 96%. The conclusion is that the methodology developed produced great results in detecting and correcting CTF estimation parameters. A graphical interface was therefore developed for using these tools, which was made publicly available as part of the open-source XMIPP software specialized for 3D electron microscopy in structural biology.

6. BIBLIOGRAPHY

[1] C.O.S. Sorzano, J. M. Carazo. *Avances en el procesamiento de imágenes biológicas a nivel microscópico: Segmentación*. Bit (Revista del Colegio de Ingenieros de Telecomunicación), 172: 71-74 (2008).

[2] C.O.S. Sorzano, S. Jonic, R. Núñez, N. Boisset, J.M. Carazo. *Fast, robust and accurate determination of transmission electron microscopy contrast transfer function*. Journal Structural Biology 160: 249-262 (2007).

[3] C.O.S.Sorzano, R. Marabini, A. Pascual-Montano, S.H.W. Scheres, J.M. Carazo. *Optimization problems in electron microscopy of single particles*. Annals of Operations Research, 148: 133-165 (2006).

[4] S. H. W. Scheres, R. Núñez-Ramírez, C.O.S. Sorzano, J.M. Carazo, R. Marabini. *Image processing for electron microscopy single-particle analysis using Xmipp*. Nature protocols, 3: 977-990 (2008).

[5] Jonic, S., Sorzano, C. O., Cottevieille, M., Larquet, E., and Boisset, N. *A novel method for improvement of visualization of power spectra for sorting cryo-electron micrographs and their local areas*. J Struct Biol *157*, 156-167 (2007).

[6] Mindell, J. A., and Grigorieff, N. *Accurate determination of local defocus and specimen tilt in electron microscopy*. J Struct Biol *142*, 334-347 (2003).

[7] P. W. Hawkes, J. C. H. Spence. *Science of Microscopy*. Volume 1. Ed. Springer (2007).

[8] J. Frank. *Three - Dimensional Electron Microscopy of Macromolecular Assemblies*. Ed. Academic Press (2006).

7. ANNEX

A. Classifications

A.1. *Damping*

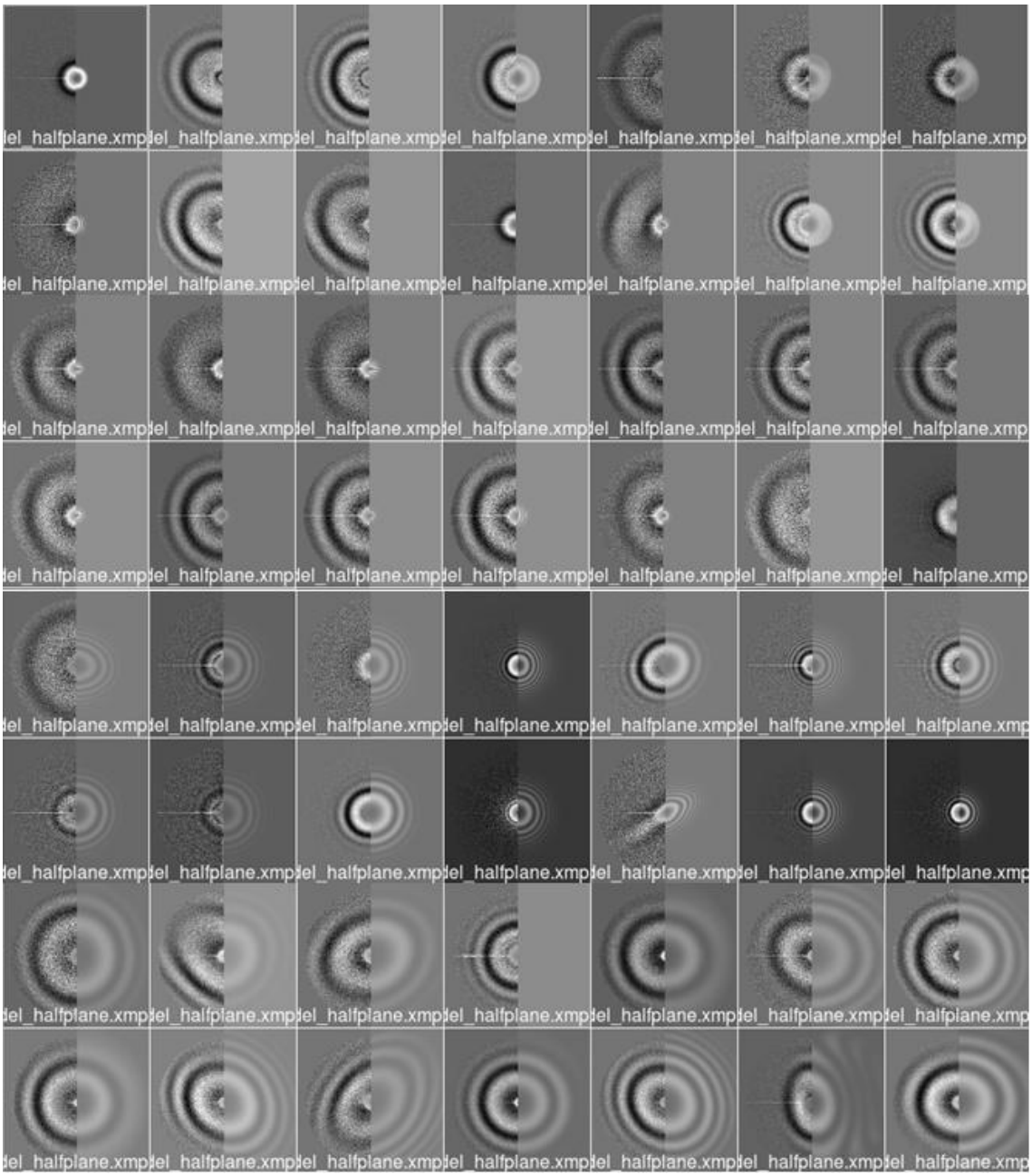

Fig. 57 Representative extract of sorting with *damping* criterion

A.2. *First zero average*

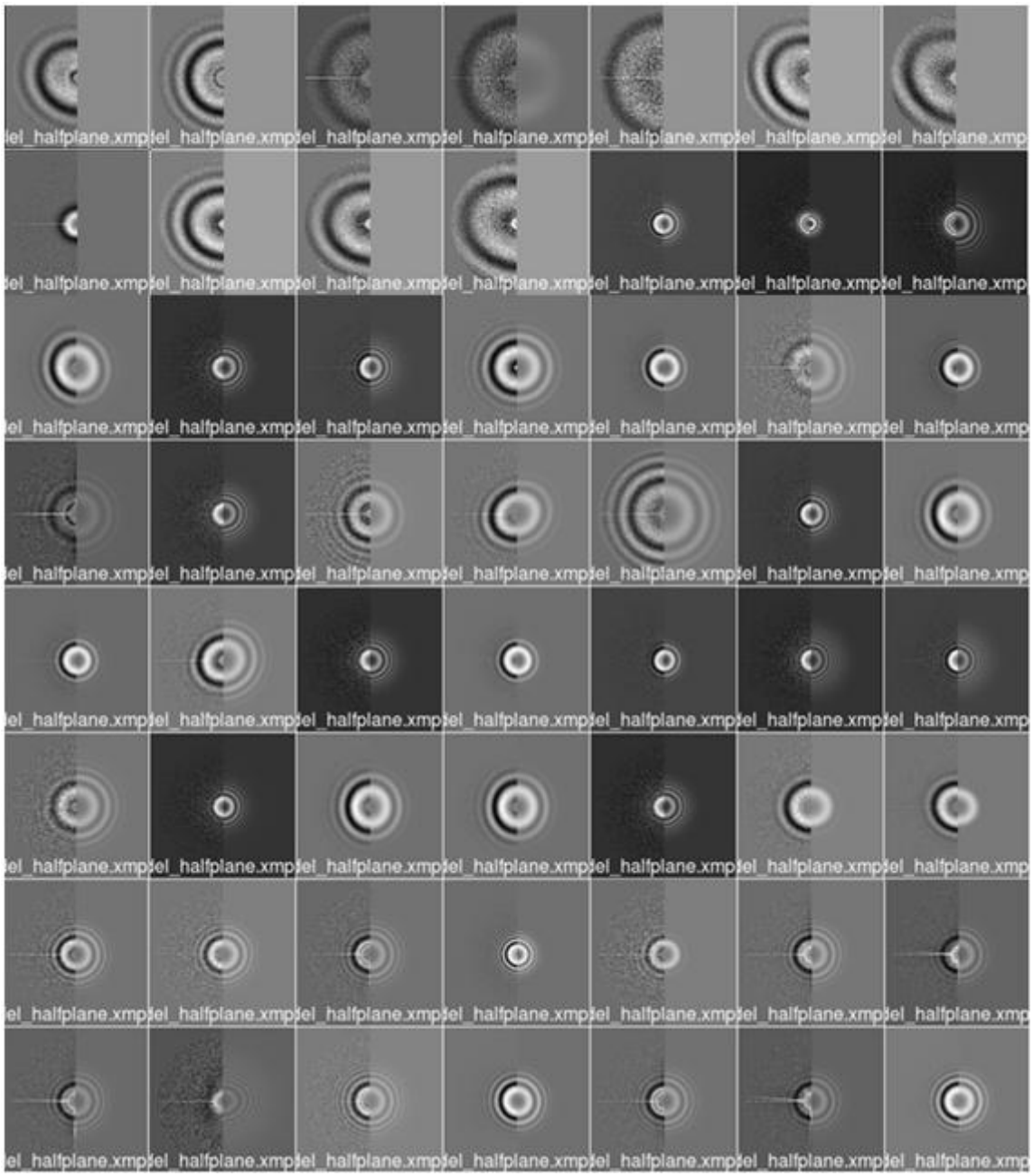

Fig. 58 Representative extract of sorting with *first zero* criterion

A.3. *First zero ratio*

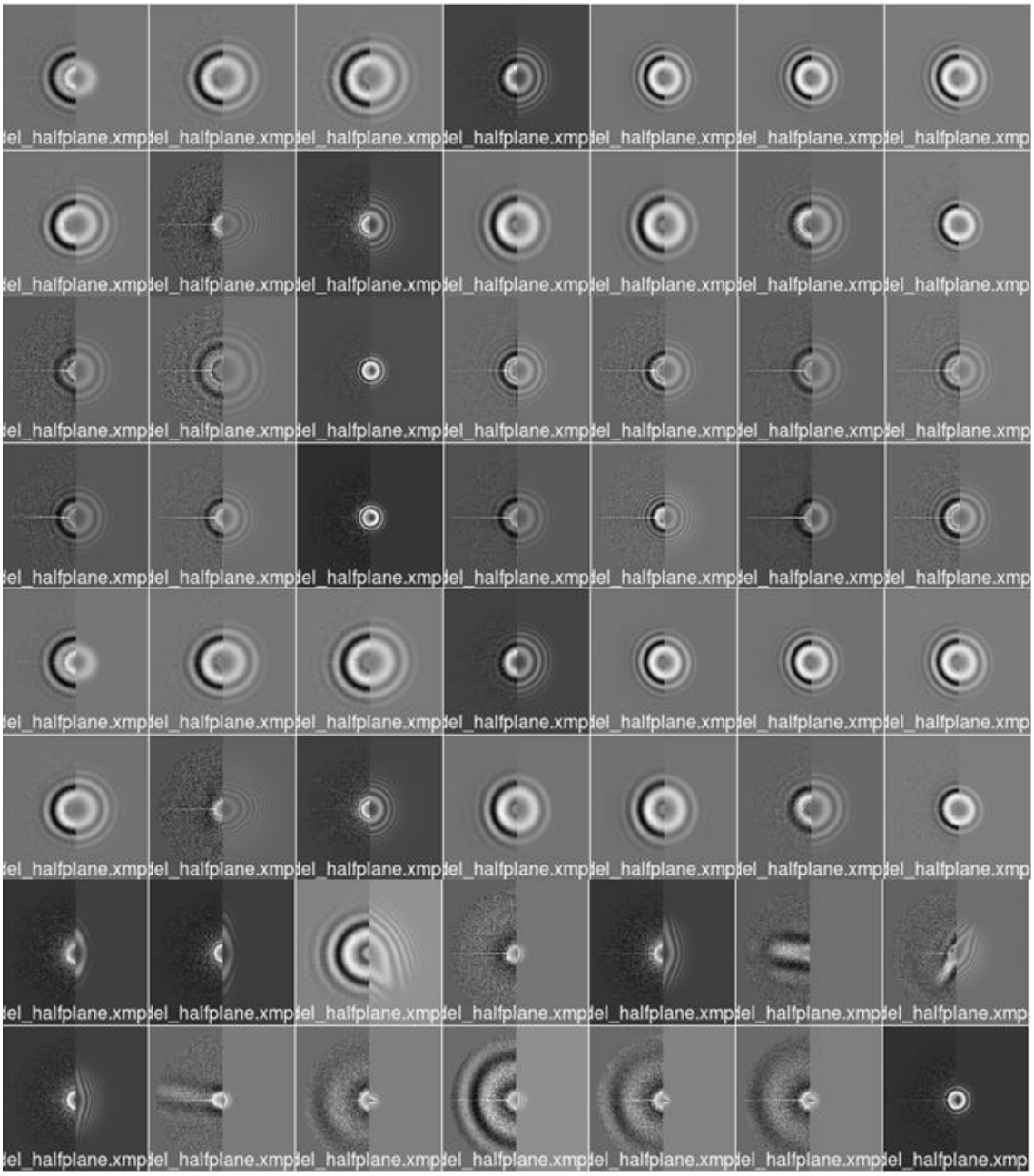

Fig. 59 Representative extract of sorting with *first zero ratio* criterion

A.4. *Fitting score*

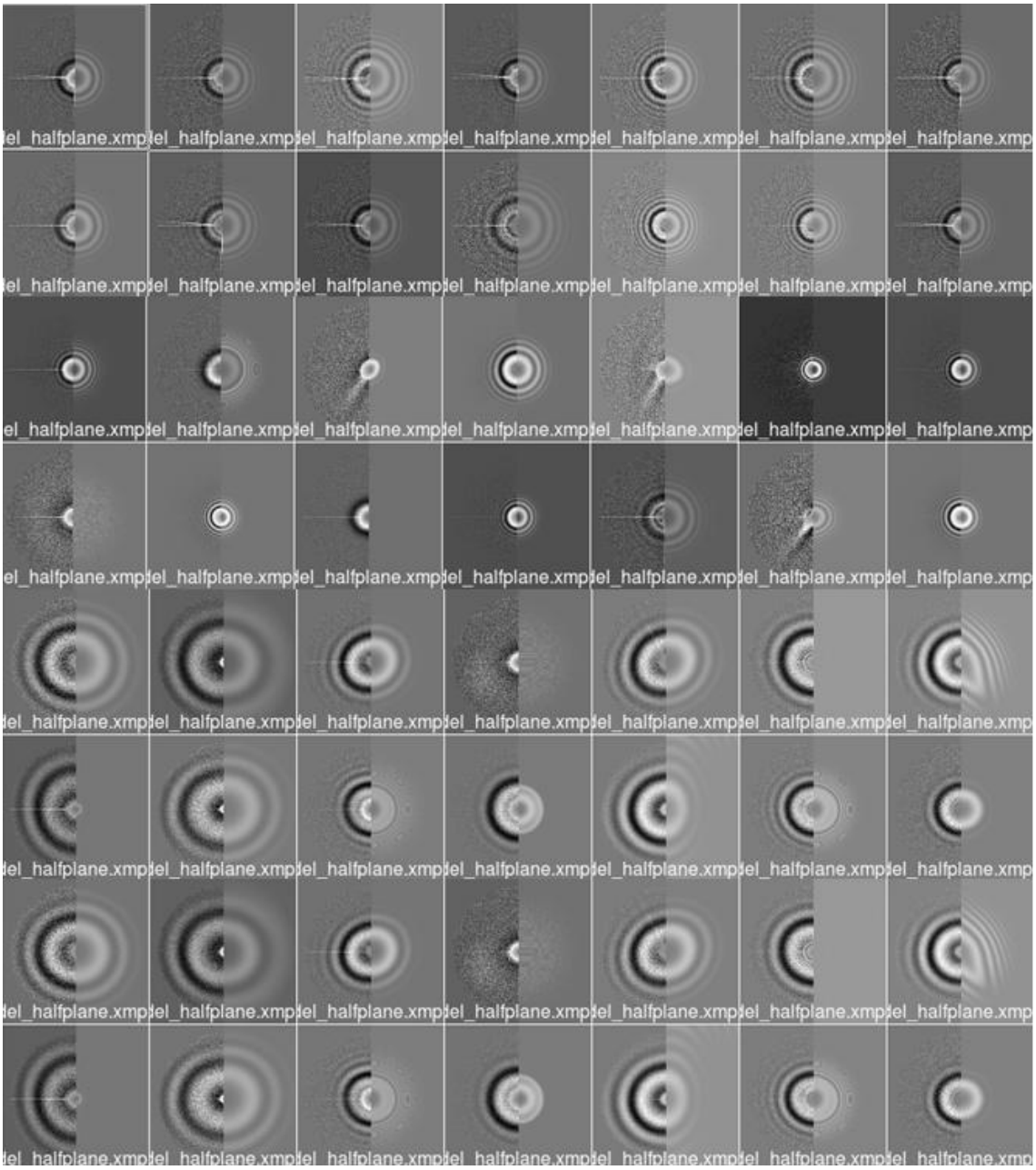

Fig. 60 Representative extract of sorting with *fitting* criterion

A.5. *Fitting correlation between zeros 1 and 3*

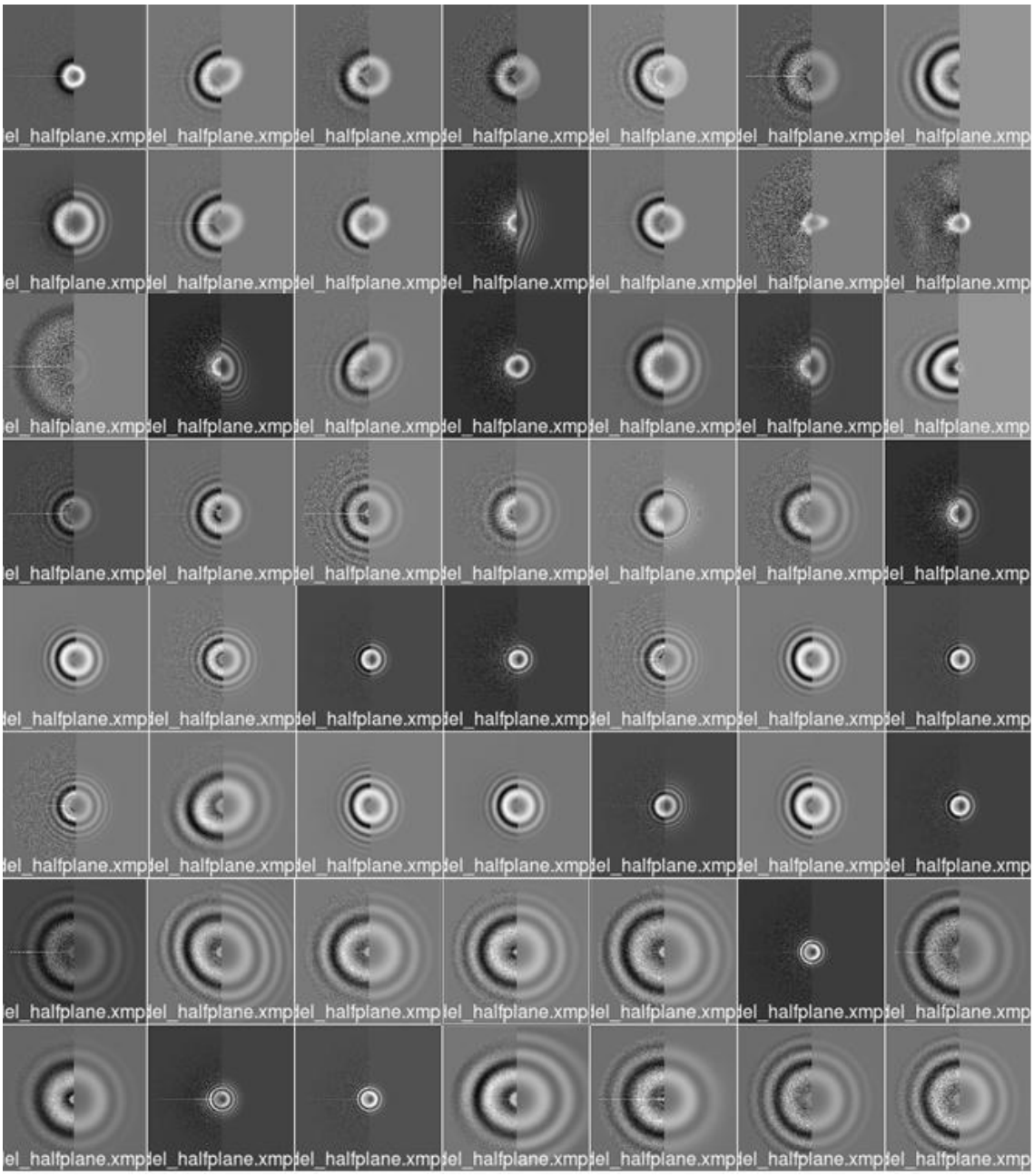

Fig. 61 Representative extract of sorting with *corr13* criterion

A.6. *PSD correlation at 90 degrees*

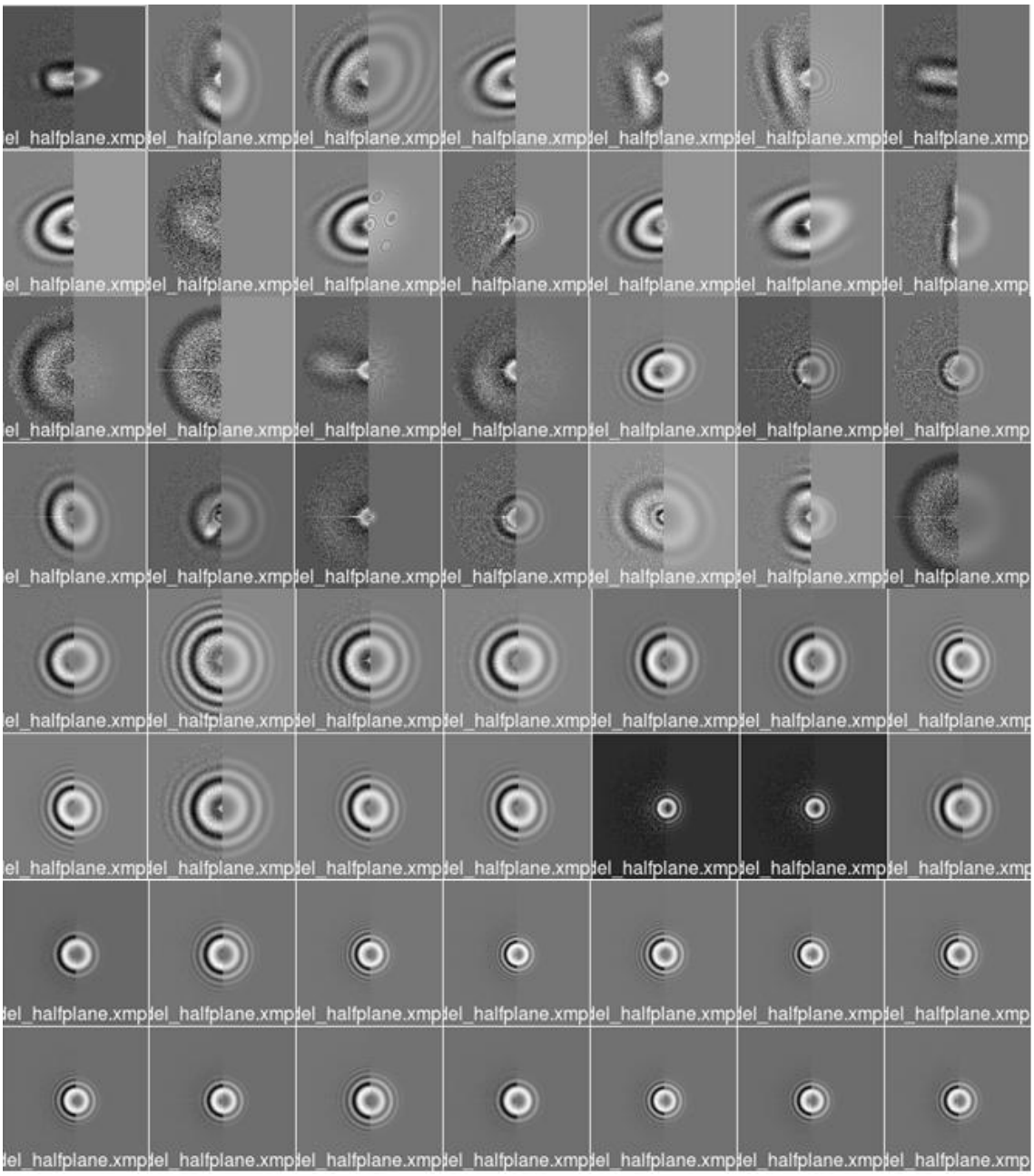

Fig. 62 Representative extract of sorting with *psdcorr90* criterion

A.7. *PSD radial integral*

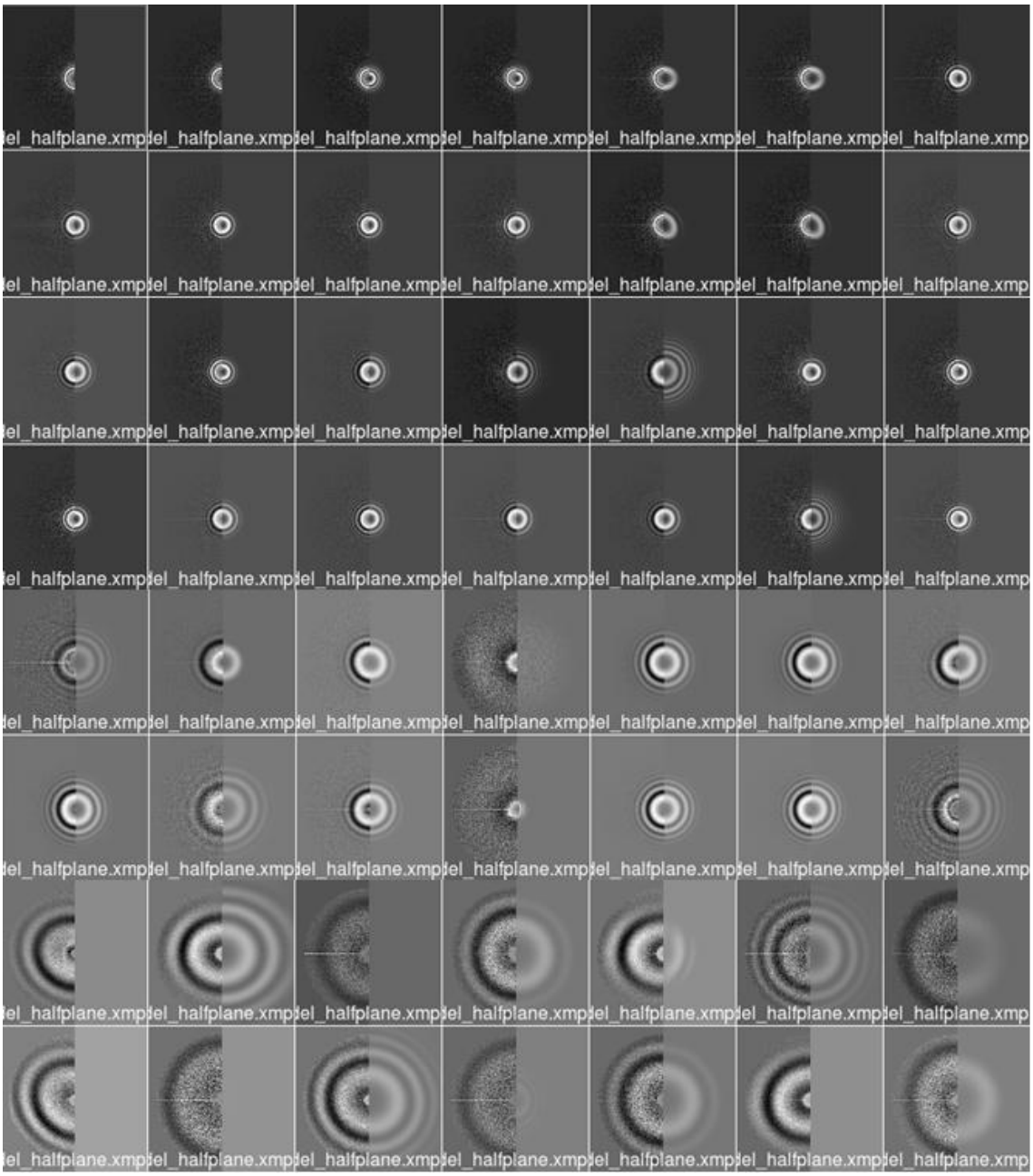

Fig. 63 Representative extract of sorting with *psdint* criterion

A.8. *PSD variance*

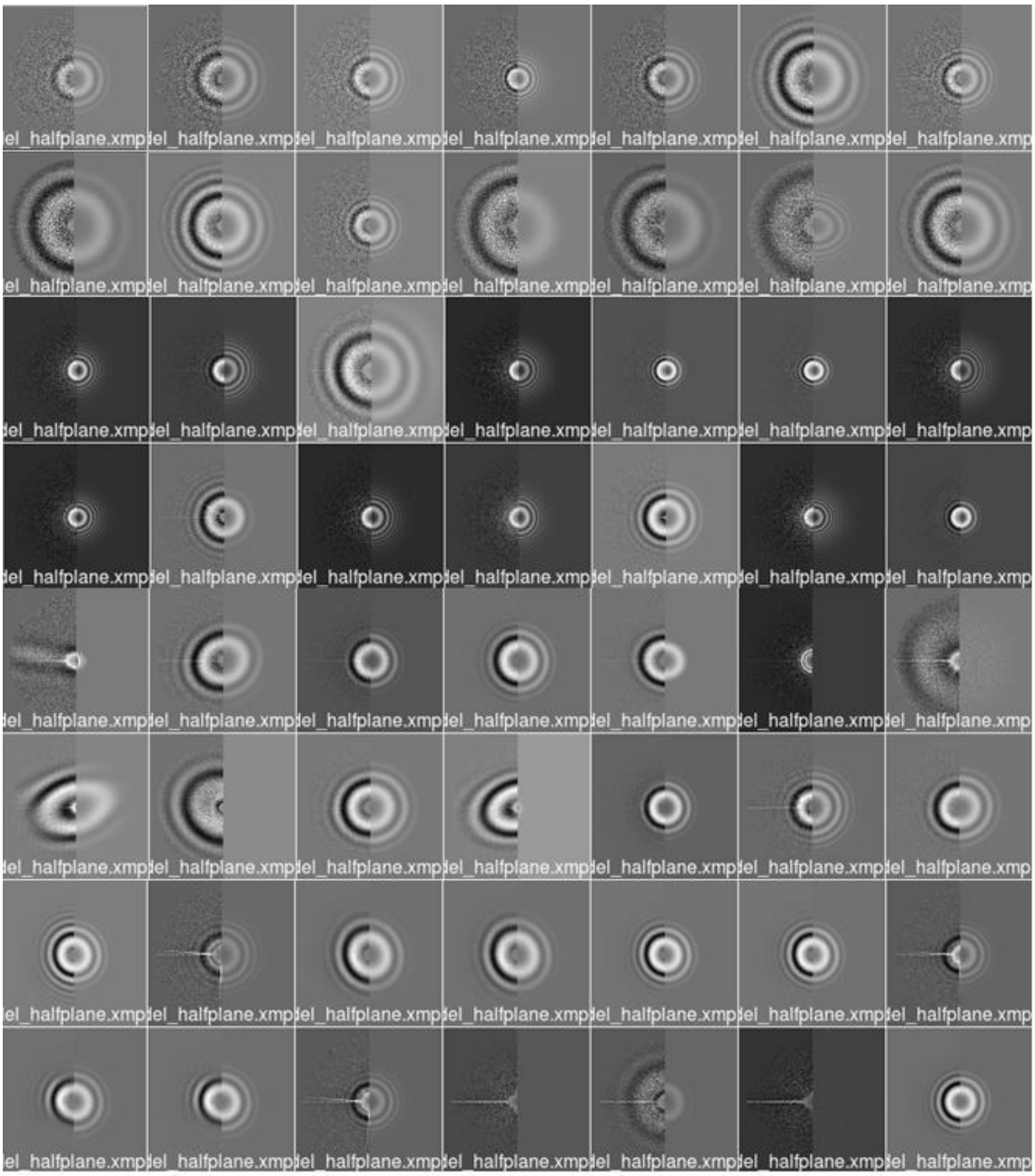

Fig. 64 Representative extract of sorting with *PSDstdQ* criterion

A.9. *PSD PCA Runs test*

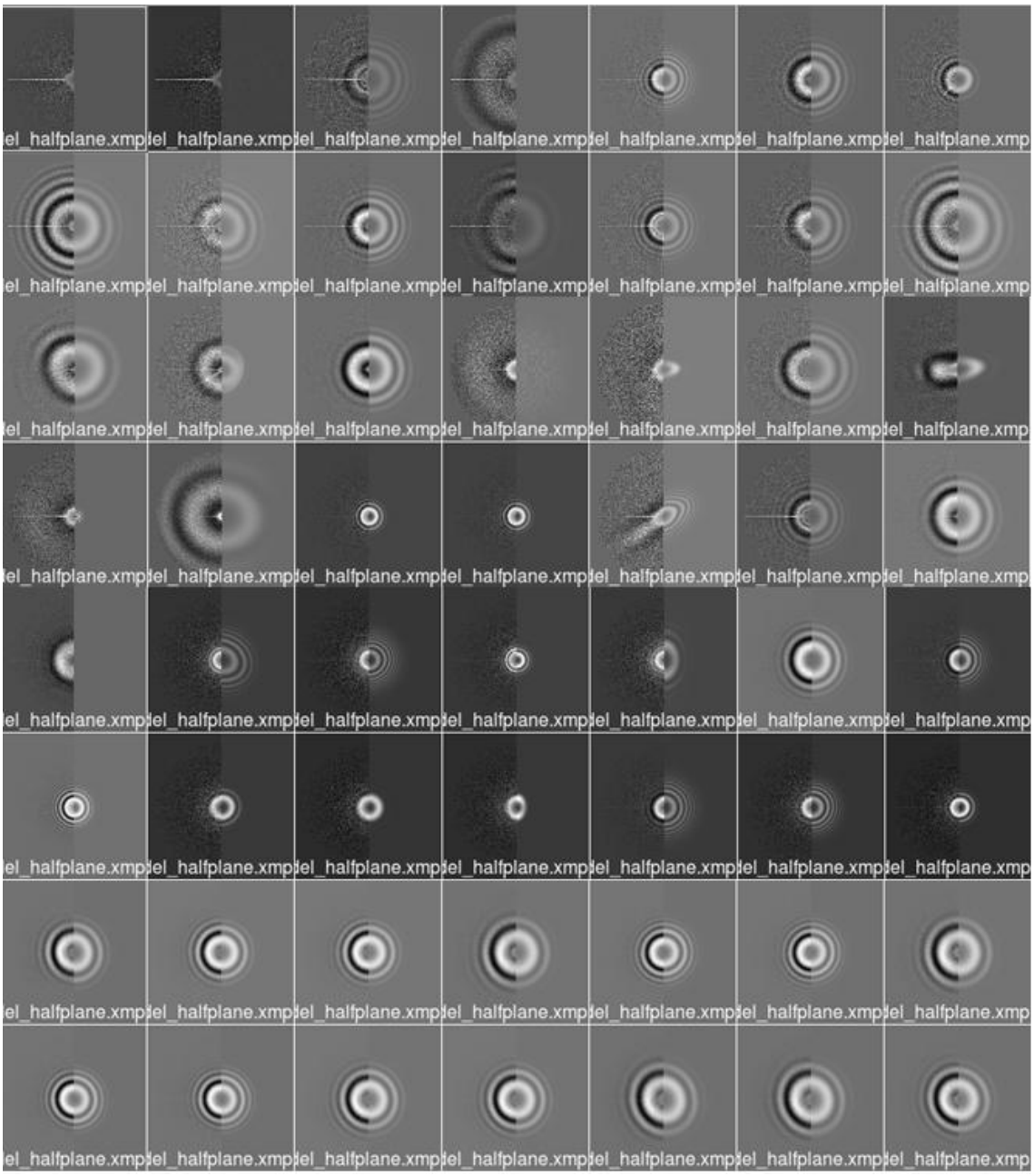

Fig. 65 Representative extract of sorting with *PSDPCARuns* criterion

A.10. Good images sorted with *PSD correlation at 90 degrees*

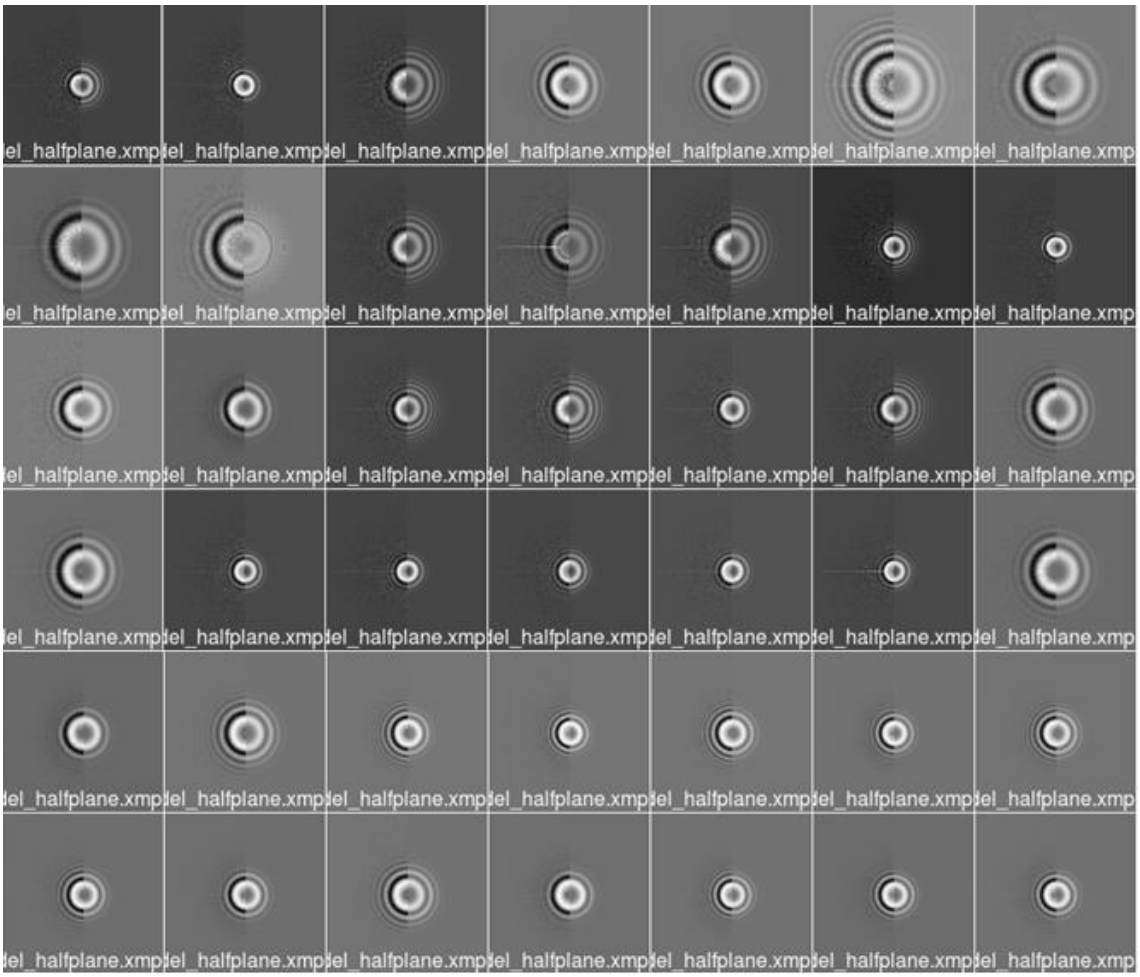

Fig. 66 Representative extract of good images sorted with *psdcorr90* criterion

.

A.11. Sorting with average between *PSD correlation at 90 degrees*, *fitting correlation between zeros 1 and 3* and *first zero average*.

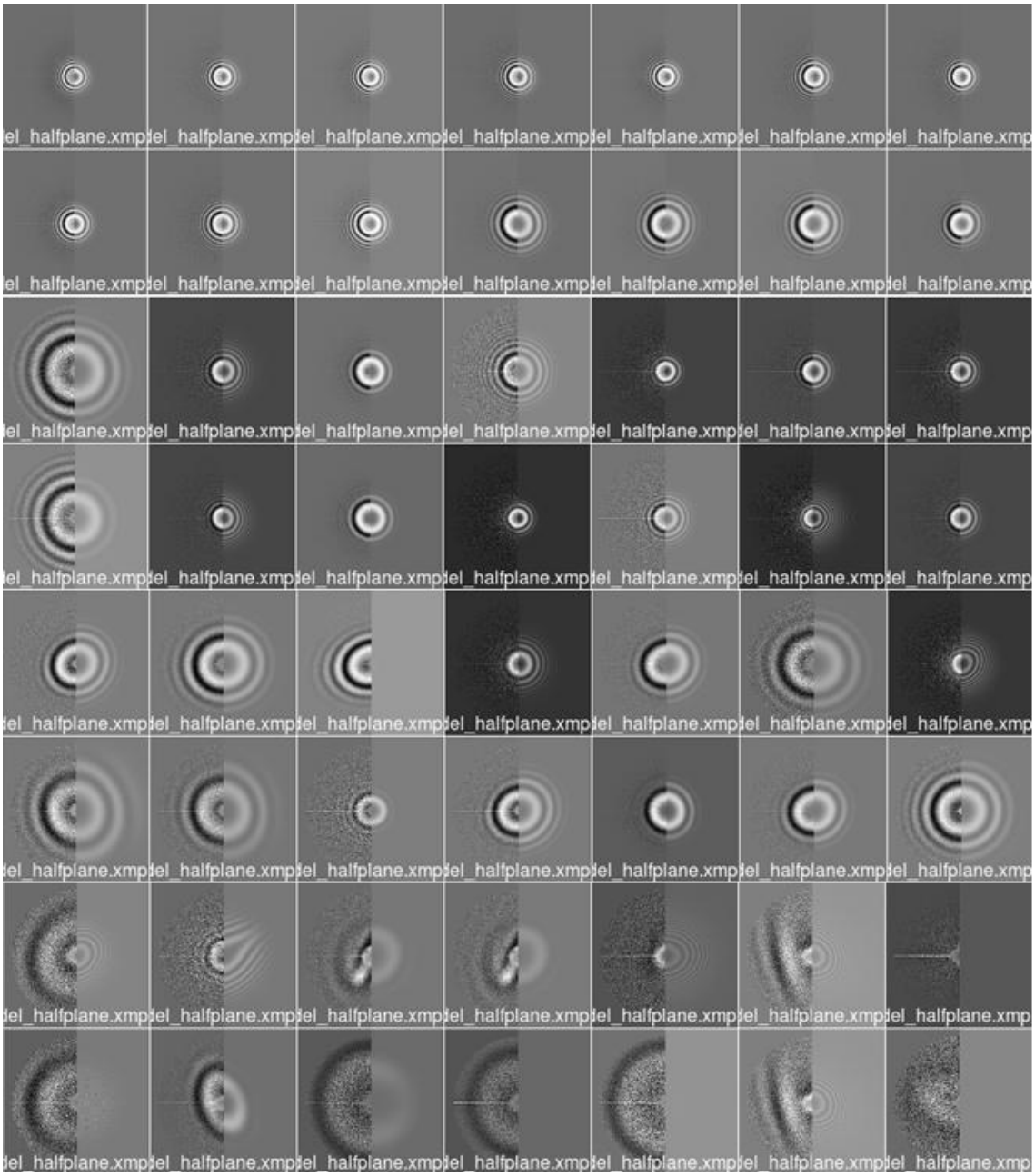

Fig. 67 Representative extract of sorting with average between *psdcorr90*, *corr13* and *first zero*

A.12. Sorting only good images with average between *PSD correlation at 90 degrees* and *fitting correlation between zeros 1 and 3* and *first zero average*.

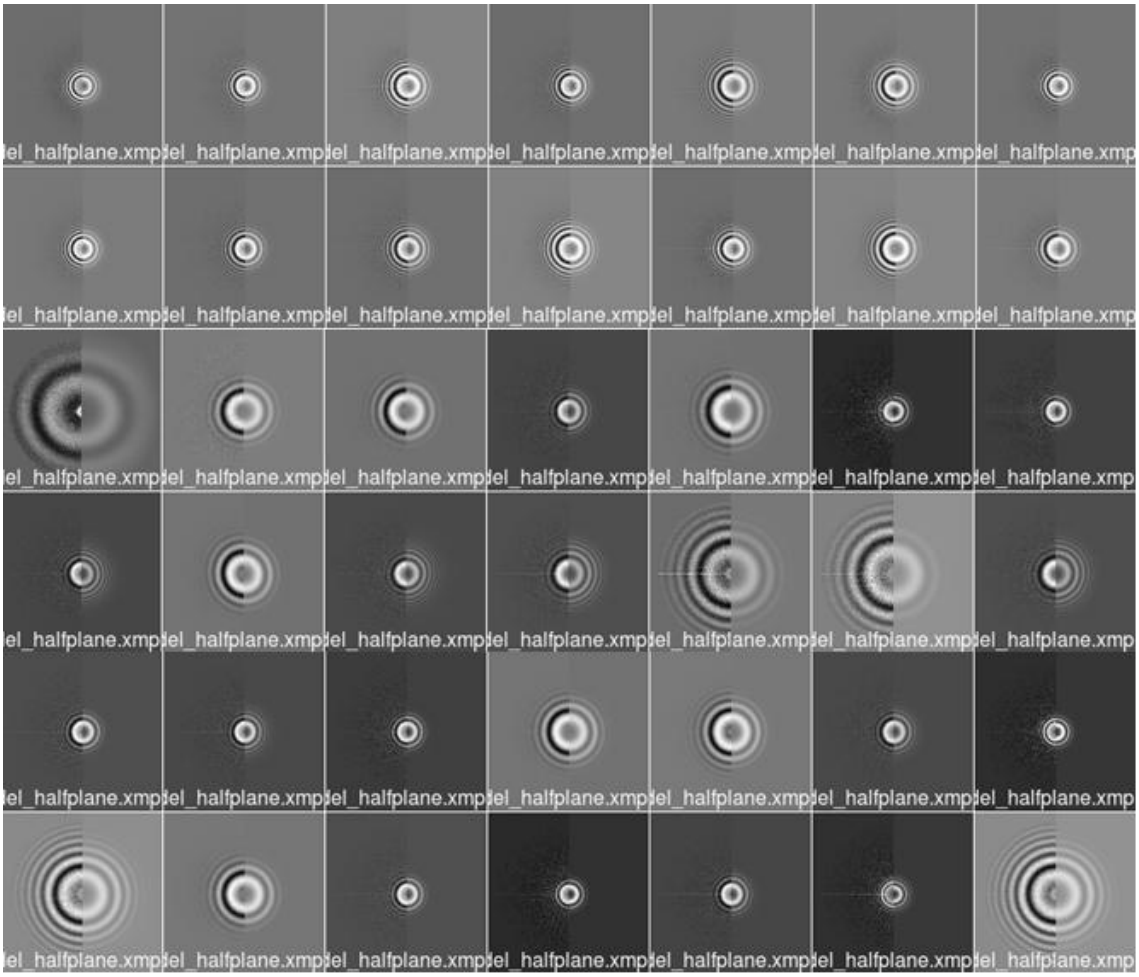

Fig. 68 Representative extract of sorting only good images with average between *psdcorr90* and *corr13*

A.13. Sorting only good images with *damping*

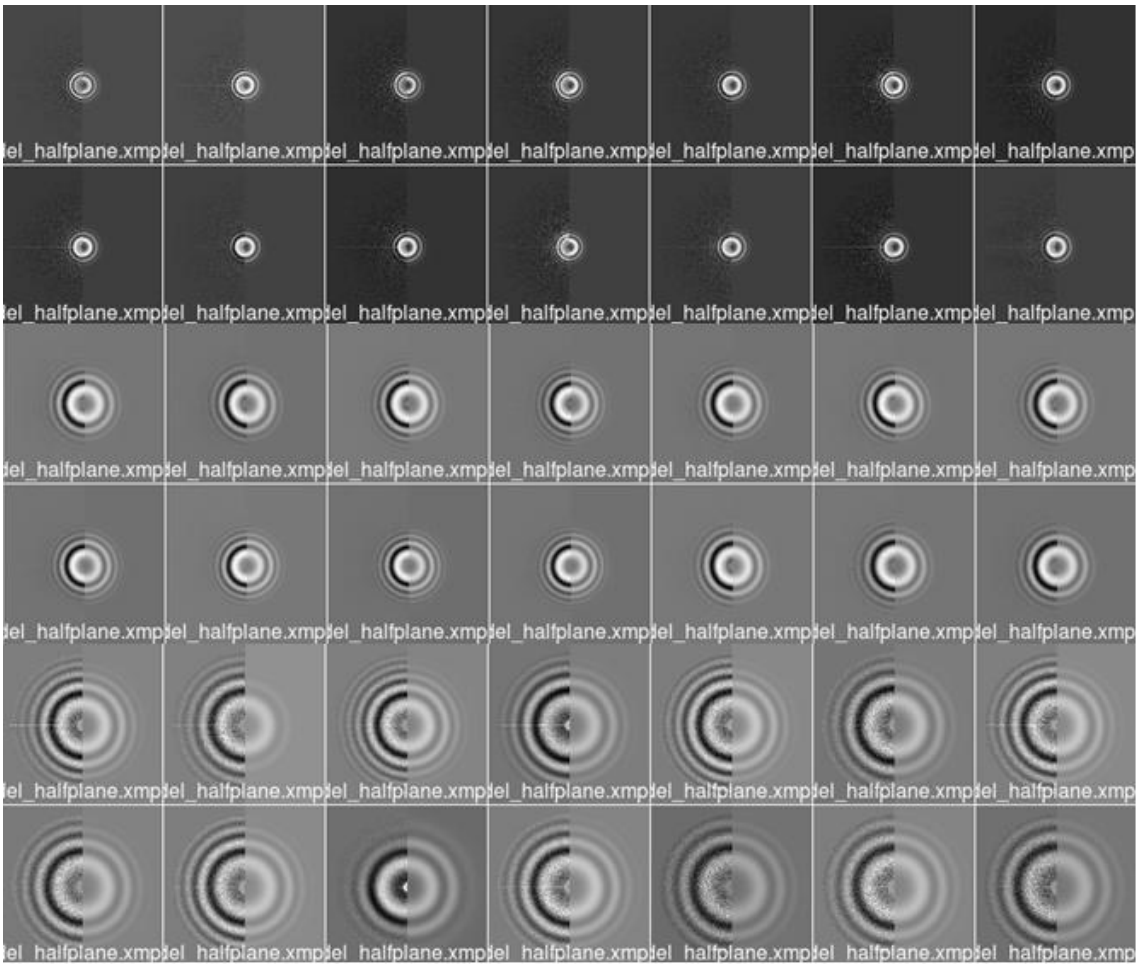

Fig. 69 Representative extract of sorting only good images with *damping*

A.14. *Normality*

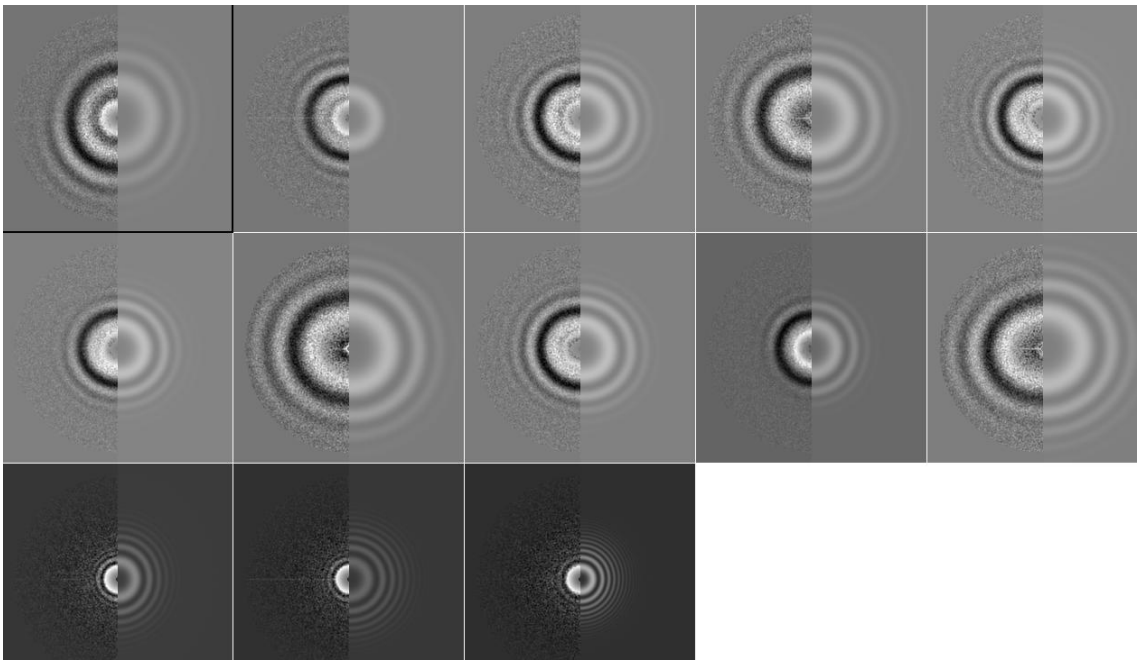

Fig. 70 Representative extract of *normality* criterion