Campus Internacional excelencia UAM CSIC+

UAM UNIVERSIDAD AUTONOMA DE MADRID

Máster en

Física de la materia condensada y los sistemas biológicos

Integration and analysis of the results of three methods to estimate particle local defocus in cryo-EM using Scipion

Estrella Fernández Giménez

FACULTAD DE CIENCIAS

Director: Carlos Óscar Sánchez Sorzano
Tutor: David Gómez Míguez
Lugar de realización: Centro Nacional de Biotecnología (CNB-CSIC) Unidad de Biocomputación

Trabajo Fin de Máster. Curso 2018-2019

Universidad Autónoma de Madrid

Abstract

Cryo-electron microscopy (cryo-EM) is becoming the technique of choice for determining macro-molecular structures. Single particle analysis (SPA) is a cryo-EM methodology which is able to acquire multiple projections (particles) from identical copies of a macromolecule and reconstruct its structure. To obtain such volume, it is necessary to process the acquired images, named micrographs. There are several challenges in this process. One of them, remains in knowing the local defocus of the particles acquired. In this project, we have analyzed and compared three methods (Xmipp, Relion and Gctf) to estimate the local defocus of the particles in the software platform for cryo-EM image processing Scipion. As the results of the different methods present several discrepancies, we have proposed a consensus solution for the local defocus of the particles.

1. Introduction

Cryo-electron microscopy (cryo-EM) is one of the principal methods used in the determination of the three-dimensional atomic structure of macromolecules, which generates a high-resolution macromolecule density map, with details lesser than 3Å.

Single particle analysis (SPA) is a form of cryo-EM which allows to capture simultaneously snapshots of numerous molecular views of the same protein in its native state in a single sample. The sample is assumed to be of multiples copies of a purified macromolecule. Then, it is imaged with a transmission EM obtaining a set of micrographs, each of them containing several copies (projections) of the macromolecule in different orientations. All these projections will be processed and combined to obtain a reconstructed volume which will lead to a three-dimensional density map of the protein.

This methodology is based on two hypotheses. The first one consists on considering the sample homogeneous: all specimens (particles) are the same macromolecule in the same conformation but with different orientations. The second hypothesis is the projection assumption which considers that every image (micrograph) is a projection of the sample under a given magnification of the microscope [1].

The problem now is to assign angles to the projections and then, reconstruct the volume with those projections. The reconstruction task can be done by using the Central slice theorem, which postulates that the two-dimensional Fourier transform (FT) of a projection corresponds to a FT slice of the three-dimensional FT volume (Fig. 1) [1].

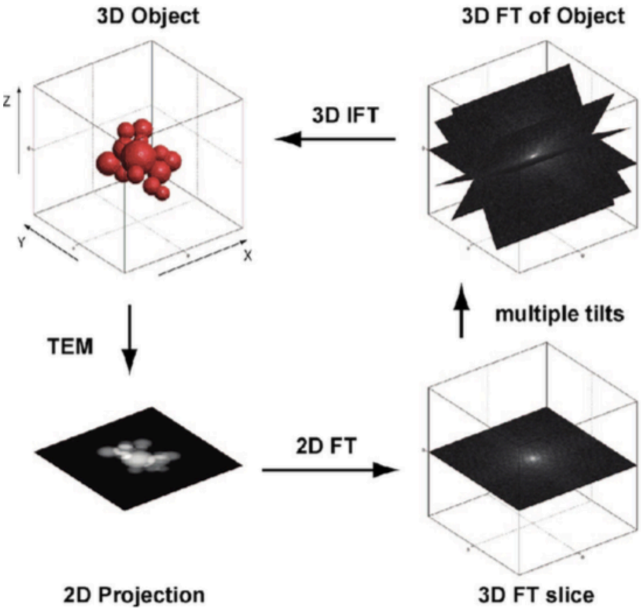


Figure 1. *Graphical explanation of the Central Slice Theorem [1].*

However, the acquired image by the microscope is not the ideal, but convolved with the Point Spread Function, which is the inverse Fourier transform of the Contrast Transfer Function

(CTF) of the microscope and models the aberrations and defocus (1):

$$Irecorded = Iideal * FT^{-1}\,[CTF]. \qquad (1)$$

Thus, accurate estimation of the CTF is a critical step to acquire a high-resolution reconstructed volume in cryo-EM [2]. CTF estimation allows to correct the aberrations and defocus and increase the signal to noise ratio (SNR), because the CTF basically models the performance of the microscope. CTF is affected by the acquiring conditions, and thus, this function must be determined for each micrograph during its processing [1].

Based on the global CTF determination for each micrograph, there are methods that estimates the local defocus for each particle that appears in that micrograph. The estimation of the defocus for near-atomic resolution should be accurate. Nevertheless, stage tilt, uneven ice and other phenomena can lead to a defocus variation among particles in the same micrograph (Fig 2).

Thus, it is significant to refine the global defocus, to achieve accurate local defocus for each particle [2]. It is important to remark that the defocus may be different in the two directions of the micrograph (X and Y) due to the possible astigmatism of the lenses of the microscope. Thus, the defocus is measured by two parameters instead of just one (defocusU and defocusV, respectively).

There are three principal methods used to estimate the local defocus of the particles in cryo-EM, which are the ones developed by Gctf [2], Relion [4] and Xmipp [5]. However, the ground truth about local defocus for each particle cannot be known. Though, in this project we will compare the results obtained for the local defocus of each particle by these three methods and propose a consensus solution.

To carry out such comparison we have implemented a workflow (described in section 2.1) in Scipion [6], a software platform for cryo-EM image processing and analysis which integrates different cryo-EM image processing software packages. The Gctf and Relion methods to estimate local defocus were already integrated as protocols in Scipion, but the Xmipp method has been integrated as a Scipion protocol in this project (described in section 2.2). We have also developed two Scipion protocols to analyze and compare all local defocus estimations (described in section 2.3).

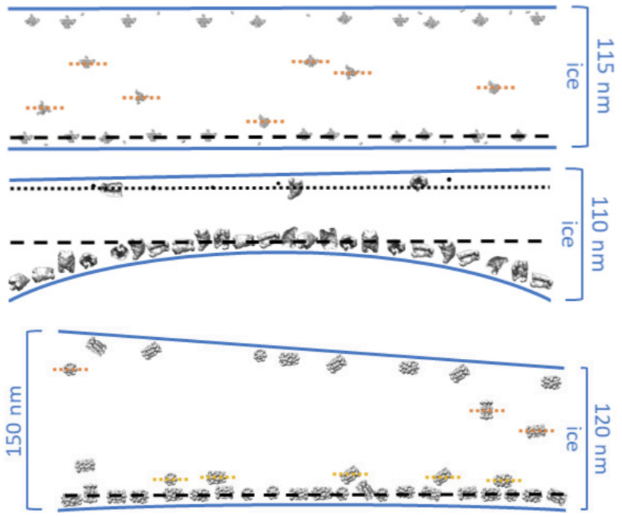


Figure 2. *Cross-sectional diagrams showing locations of particles embedded in ice in a cryo-EM sample [3].*

2. Methods

2.1. The general SPA workflow

To execute the three different protocols to estimate the local defocus of the particles, it is necessary to do a general SPA processing of the micrographs (Fig. 3).

The workflow has been done in Scipion and it starts with the importation of the data. These data can be movies (several micrographs taken with a smaller exposure time) or micrographs. In the case of movies, it is necessary to do a *movie* *alignment* to obtain the micrographs, because the exposure of the sample to the electron beam causes a slight misalignment in movie frames. Once we have the micrographs, the CTF of the microscope can be estimated for each micrograph and it will be used to correct aberrations and defocus. Besides, from each of the micrographs it is necessary to obtain the particles contained in it. To do that, it is necessary to *pick* all the particles in each micrograph with manual, semiautomatic or fully automatic methods.

Once the particles are picked, the coordinates of each particle in the micrograph are known, so the particles can be *extracted* as smaller images. When the particles are obtained, the local defocus estimation by Gctf could be computed using also the CTF estimation. However, the other two estimations, Relion and Xmipp, need not only the particles, but also the reconstructed volume.

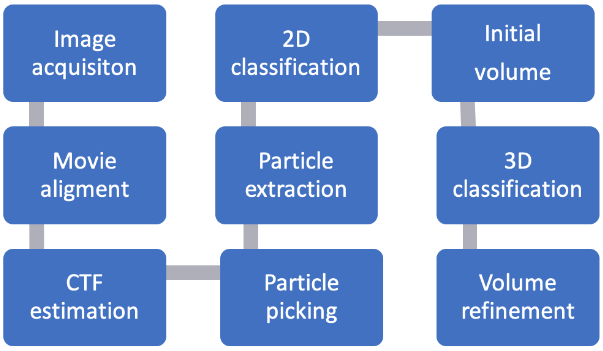


Figure 3. *SPA general workflow [1].*

To obtain a reconstructed volume, it is necessary to classify the particles (*2D classification*) depending on their orientations, such that each class will contain all the particles which are projections under close directions of the macromolecule. Before the 2D classification step, it is recommendable to perform a *screening* of the particles, to evaluate the quality of the particles via different measurements, as Zscore (which evaluates the similarity of the particles with an average particle) and the SNR. This permits to remove the worst particles (high Zscore and low SNR) to improve the quality of the reconstruction.

With the 2D classification, an *initial volume* is determined, obtaining a coarse estimation of the macromolecule. Once we have an initial volume, a *3D classification* of the particles is performed to try to guess the existence of different conformations and then, to look for better angular assignments, obtaining a refined volume [1]. The final volume and the extracted set of particles will be the input of the Relion and Xmipp local defocus estimation methods.

2.2. Integration of Xmipp method for estimating local defocus in Scipion

As explained before, Xmipp package had a method to compute an estimation of local defocus of the particles, but it was not integrated in Scipion yet. Thus, it was necessary to develop a Scipion protocol to execute it. The protocol is basically a wrapper which executes the call to the corresponding Xmipp program ("*xmipp_angular _continuous_assign2*") with the adequate parameters.

The input of this protocol is the set of particles of interest together with the high resolution or *refined* volume. Moreover, the user can introduce other parameters as the *maximum defocus change* allowed for the particles, which indicates the maximum distance that defocus of the particle can be corrected with respect to the original global defocus.

Following the input step, the protocol runs the Xmipp program for local defocus, which returns the refined defocus for each particle and the difference with the original, which is the global one. After that, the final step of the protocol consists on creating a valid output that can be processed and analyzed by other Scipion protocols. This output displays to the user all the information obtained for each particle by the previous protocols together with the new information computed about local defocus by Xmipp. This protocol is publicly available in the *scipion-em-xmipp* repository[1].

2.3. Protocols for analysis and comparison

For the analysis and comparison of all the defocus estimations computed by different methods, two Scipion protocols where implemented. On one hand, there is an analysis protocol which can examine the results obtained by each of the local defocus estimators individually. On the other hand, we have developed a comparison protocol that allows to join all the local defocus estimations computed by different methods and gives the user a consensus solution. The concept of consensus solution means that the protocol gives a single final local defocus estimation for each of the particles under study, considering all the estimations computed.

As said before, ideally all the particles in a micrograph will have the same defocus, so they should be in the same plane. Thus, the analysis protocol that we have developed considers the particles in each one of the micrographs and tries to adjust, for each of them, the estimated local defocus of the particles contained in it into a linear polynomial (2):

$$\overline{\boldsymbol{Z}} = c\overline{\boldsymbol{Y}} + b\overline{\boldsymbol{X}} + a \qquad (2)$$

being $\bar{X}$ and $\bar{Y}$ vectors that contain the x and y coordinates respectively of all the particles in the micrograph. The vector $\bar{Z}$ contains the estimated local defocus for each of the particles in the micrograph, which can be understood as the height of the particle in the sample (Fig. 2). It was mention that the defocus is different in the two directions of the micrograph, and thus, the defocus is characterized by two numbers. Hence, we have considered the mean of both defocus (defocusU and defocusV) as the final defocus, for simplification in the adjustment. Lastly, a, b and c are the adjustment coefficients.

We have now an adjustment plane, which can be considered as the expected height for the particles in the micrograph. We have also the estimated height (local defocus) for each particle. Thus, we could compute the difference between the estimated value and the expected value for each particle, i.e., the residual analysis of the particle. Moreover, we have computed the adjustment quality coefficient (R^2) for the expected data (the plane) and the estimated data for each of the micrographs, which are showed to the user.

We have also included in this protocol a three-dimensional viewer that allows to visualize the adjustment plane and the estimated localization of the particles in each micrograph, with the estimated local defocus as the z coordinate of the particle (Fig. 7). This protocol could take as input any of the local defocus estimation protocols output. The protocol and its viewer are publicly available in the *scipion-em-xmipp* repository[2].

Furthermore, we have the comparison protocol. In this case, the protocol takes as input all the estimations computed by different protocols for the local defocus. With these estimations, it constructs a matrix with all of them for each particle, which will be referred as the defocus

[1] Available at: https://github.com/I2PC/scipion-em-xmipp/blob/ef_localCTF/xmipp3/protocols/protocol_local_ctf.py

[2] Available at: https://github.com/I2PC/scipion-em-xmipp/blob/ef_localCTF/xmipp3/protocols/protocol_analyze_local_ctf.py
Viewer: https://github.com/I2PC/scipion-em-xmipp/blob/ef_localCTF/xmipp3/viewers/viewer_analyze_local_ctf.py

matrix. Moreover, the protocol computes a correlation matrix from the defocus matrix to realize the degree of similarity between the approximations. These matrices are stored as extra information for the user to analyze the results if interested.

However, the output of the protocol is directly the consensus solution for the local defocus for each one of the particles. As the estimates could differ considerably between them, we have decided that it is necessary to offer a robust consensus solution for each particle. This consensus solution will be the median of the estimations obtained.

The median is chosen because it is more robust if any of the estimations strongly disagree with the rest, and thus, it does not bias the consensus solution as much as the mean. The protocol returns also the "residual" between the estimations. To do that, the median absolute deviation is computed for each particle (3). This measure is used instead of just the deviation for being, again, more robust in the case that the estimations strongly differ.

$$\mathrm{MAD} = \mathit{median}(|\mathrm{X}_i - \tilde{X}|). \qquad (3)$$

It is important to remark that during the processing of the micrograph previous to compute the local defocus, some particles could be removed by different protocols because their low quality. Thus, the input set of particles for each of the local defocus estimation protocols could be slightly different (some of them will have less particles than others, depending on how many previous processing steps they do need).

However, all the particles are identified with a unique id and therefore, the protocol is able to know if there are particles without any of the estimations. If this is the case, the protocol considers just the estimations that have been obtained for each particle. This protocol is also publicly available at *scipion-em-xmipp* repository[3].

2.4. Project workflow

As it was said in Section 2.1, there is an established general workflow for processing cryo-EM micrographs. However, the steps of the workflow slightly vary depending on the dataset which is being processing. The final workflow for this project is showed in Fig. 4.

In this workflow, the most relevant part that has changed is the refinement of the volume. Firstly, we have refined the volume in two different ways, one using the Xmipp package and the other using the Relion package. This has been done in order to provide to each estimate local defocus protocol a volume refined with the same software package (except Gctf, which does not require a volume for the estimation), to improve their performance.

Furthermore, each of the paths for refinement of the workflow are composed by several iterations and additional protocols (create a subset of the best particles in the 3D classification in Xmipp and create and apply a mask in Relion). All these extra steps are included to improve the quality of the final reconstruction volume. There are also some crop and resize steps in between the workflow to adjust the size and resolution of the images to improve the performance of the following protocols.

As stated in section 1, we have compared three methods to estimate local defocus from three software packages (Gctf, Xmipp and Relion). Though in the case of Relion, the protocol can be executed in three different ways, regarding on

[3] Available at: https://github.com/I2PC/scipion-em-xmipp/blob/ef_localCTF/xmipp3/protocols/protocol_compare_local_ctf.py

how to manage the astigmatism of the microscope. One option considers for each particle the astigmatism estimated for the micrograph which contains that particle. The second option estimates the astigmatism for each of the particles. The last option does not consider the astigmatism. The three options have been executed and analyzed as individual estimations, and thus, we have five different estimations for the local defocus of each particle: Gctf, Xmipp, Relion no astigmatism, Relion astigmatism per micrograph and Relion astigmatism per particle.

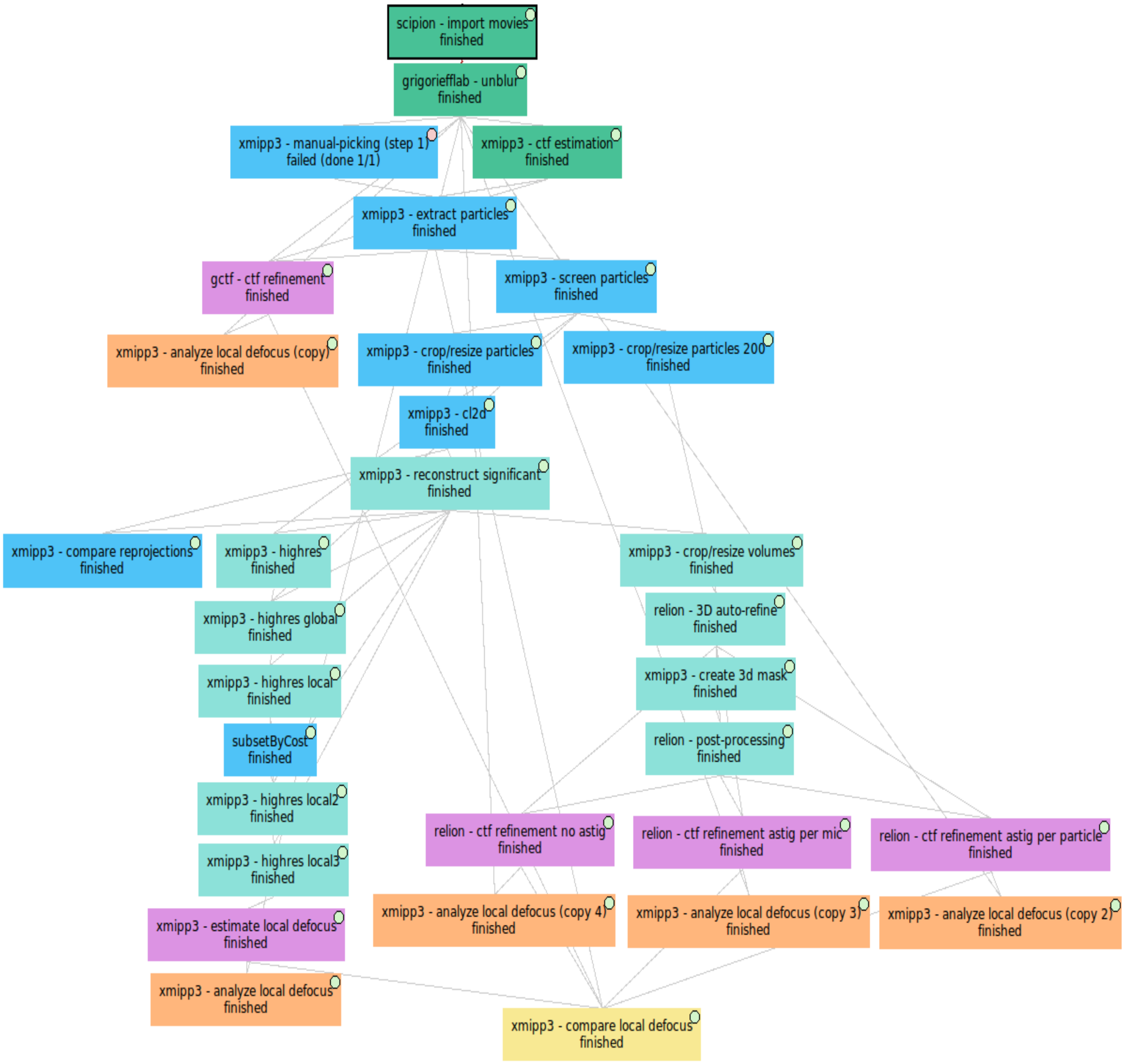

scipion - import movies
finished
grigoriefflab - unblur
finished
xmipp3 - manual-picking (step 1)
failed (done 1/1)
xmipp3 - ctf estimation
finished
xmipp3 - extract particles
finished
gctf - ctf refinement
finished
xmipp3 - screen particles
finished
xmipp3 - analyze local defocus (copy)
finished
xmipp3 - crop/resize particles
finished
xmipp3 - crop/resize particles 200
finished
xmipp3 - cl2d
finished
xmipp3 - reconstruct significant
finished
xmipp3 - compare reprojections
finished
xmipp3 - highres
finished
xmipp3 - crop/resize volumes
finished
xmipp3 - highres global
finished
relion - 3D auto-refine
finished
xmipp3 - highres local
finished
xmipp3 - create 3d mask
finished
subsetByCost
finished
relion - post-processing
finished
xmipp3 - highres local2
finished
relion - ctf refinement no astig
finished
relion - ctf refinement astig per mic
finished
relion - ctf refinement astig per particle
finished
xmipp3 - highres local3
finished
xmipp3 - estimate local defocus
finished
xmipp3 - analyze local defocus (copy 4)
finished
xmipp3 - analyze local defocus (copy 3)
finished
xmipp3 - analyze local defocus (copy 2)
finished
xmipp3 - analyze local defocus
finished
xmipp3 - compare local defocus
finished

Figure 4. *Project workflow in Scipion. The protocols which process micrographs are in green. Blue boxes represent the protocols which process particles. Light blue protocols process volumes. In magenta, the protocols which estimates local defocus, following by the analysis protocol (orange). In yellow, the comparison protocol. The connection lines represent from which protocol become the input set.*

3. Results and discussion

The degree of adjustment of the particles to the plane has been measured through the R^2 coefficient for each micrograph and compared for the five estimations considered in this project (Fig. 5). It can be appreciated that in general Gctf has a better adjustment (an order of magnitude more than the rest of the estimations in most the cases). However, the rest of estimations (Relion in all its ways and Xmipp) presents the same low degree of adjustment in all the micrographs. Thus, what this comparison reflects is that Gctf estimates smaller differences between the global and the local defocus than the rest of the methods, having the majority of the particles in each micrograph close to the "ideal" plane.

There is also much more variability of R^2 between the micrographs for the Gctf method than for the others.

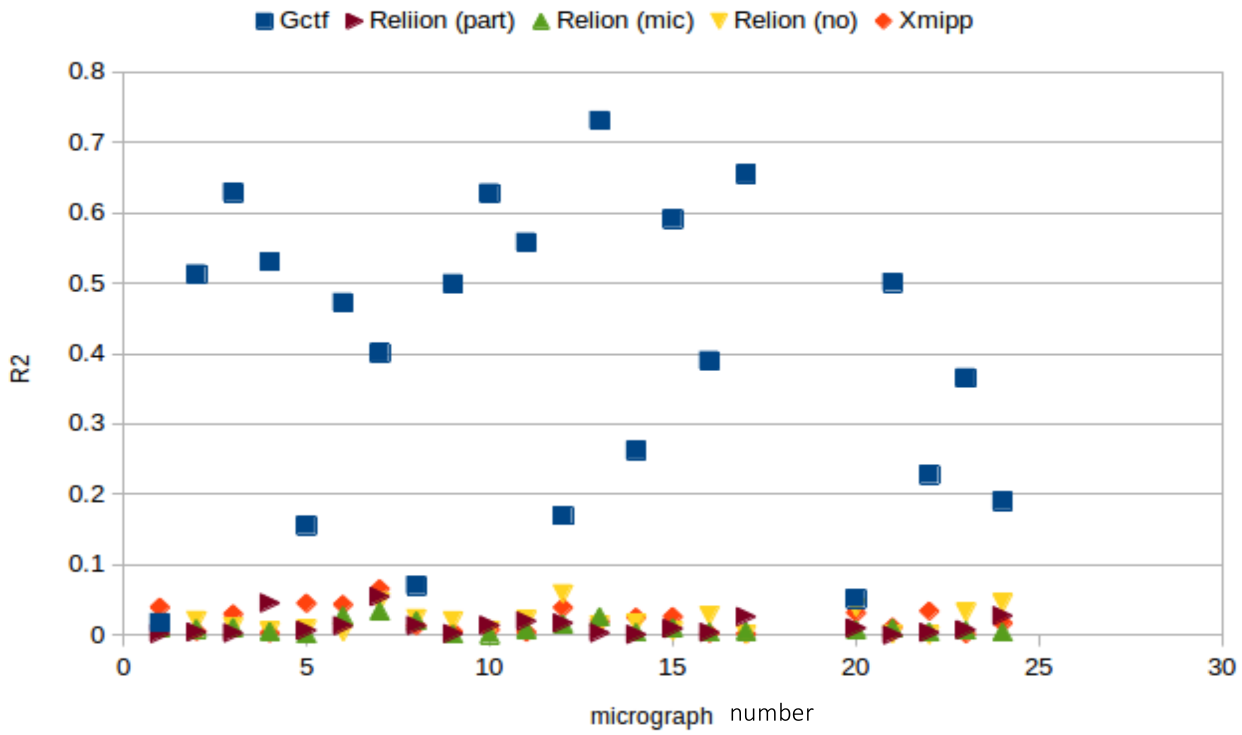
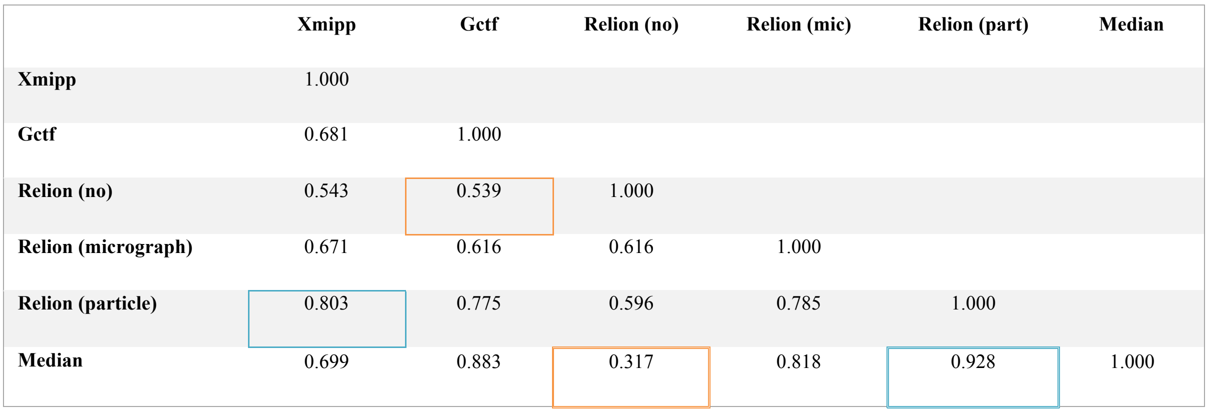


Figure 5. *Adjustment coefficient (R^2) for the five local defocus estimations considered in this project.*

The correlation matrix (Table 1) measures the degree of similarity between the different estimations and it also measures the similarity between each estimation and the consensus solution (the median) for all the particles. This matrix reveals that the most similar estimations are Xmipp and Relion measuring the astigmatism per particle, with a correlation value of 0.803.

On the other hand, the estimations that differ the most are Gctf and Relion without estimation with a correlation value of 0.539. In comparison with the consensus solution, the most similar estimation to the median is Relion per particle, with a correlation value of 0.928, and the least similar is Relion without astigmatism, with a correlation value of 0.317.



	Xmipp	Gctf	Relion (no)	Relion (mic)	Relion (part)	Median
Xmipp	1.000					
Gctf	0.681	1.000				
Relion (no)	0.543	0.539	1.000			
Relion (micrograph)	0.671	0.616	0.616	1.000		
Relion (particle)	0.803	0.775	0.596	0.785	1.000	
Median	0.699	0.883	0.317	0.818	0.928	1.000

Table 1. *Correlation values for the five local defocus estimations considered in this project and the consensus solution (median). In blue the best correlation values and in orange the lowest ones, between the estimations (single line) and in comparison with the median (double line).*

The residual for each particle returned by the comparison protocol indicates how much the local defocus consensus solution has change with respect to the original micrograph defocus estimation. In Fig. 6, the residual of each particle, grouped by micrograph are shown. The gap in the graph indicates that the micrographs 18 and 19 have been discarded by a previous protocol because their low quality does not allow to obtain information from them.

It can be appreciated that most of the particles have a residual lower than 1500 nm. However, in micrograph 14 there are many particles with a residual higher than 1500, which indicates that in micrograph 14 there are many particles with local defocus which strongly differ from the global defocus previously obtained. Micrograph 22 seems to be the micrograph with the particles which differ the least from the global defocus.

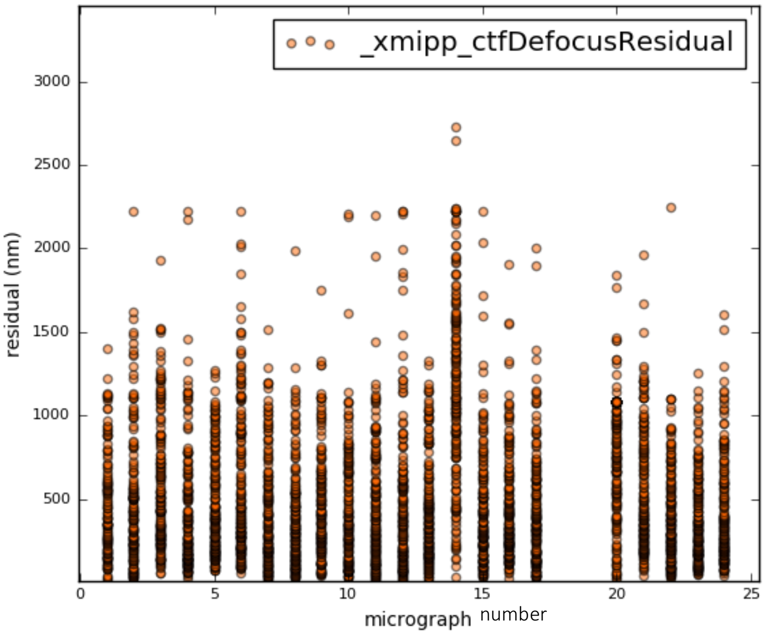
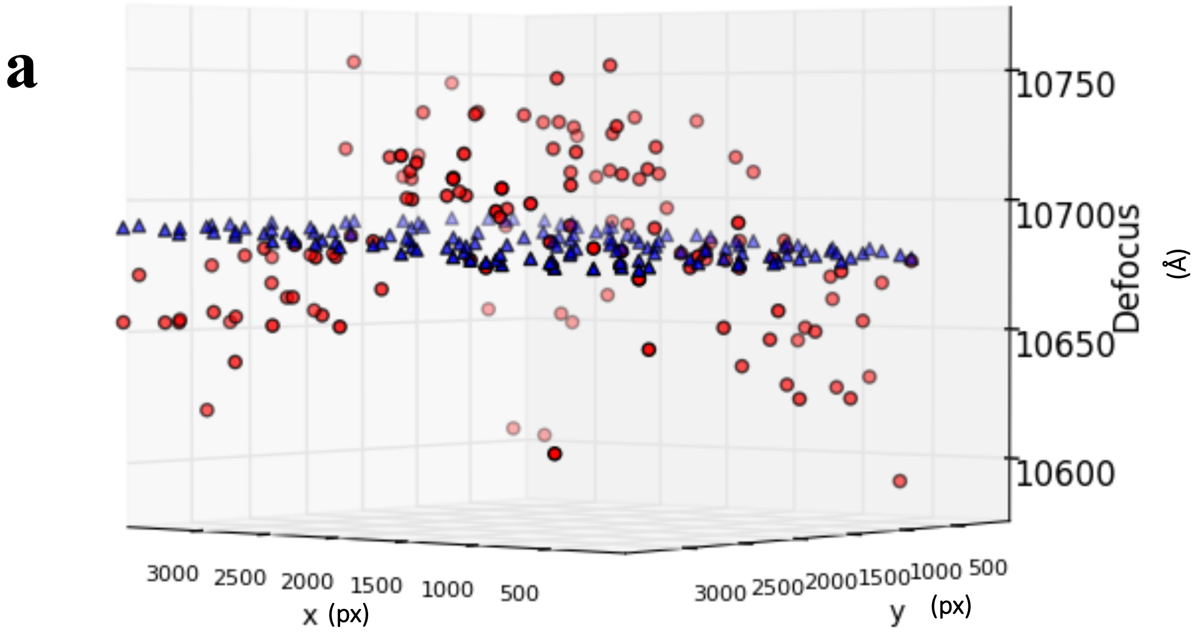


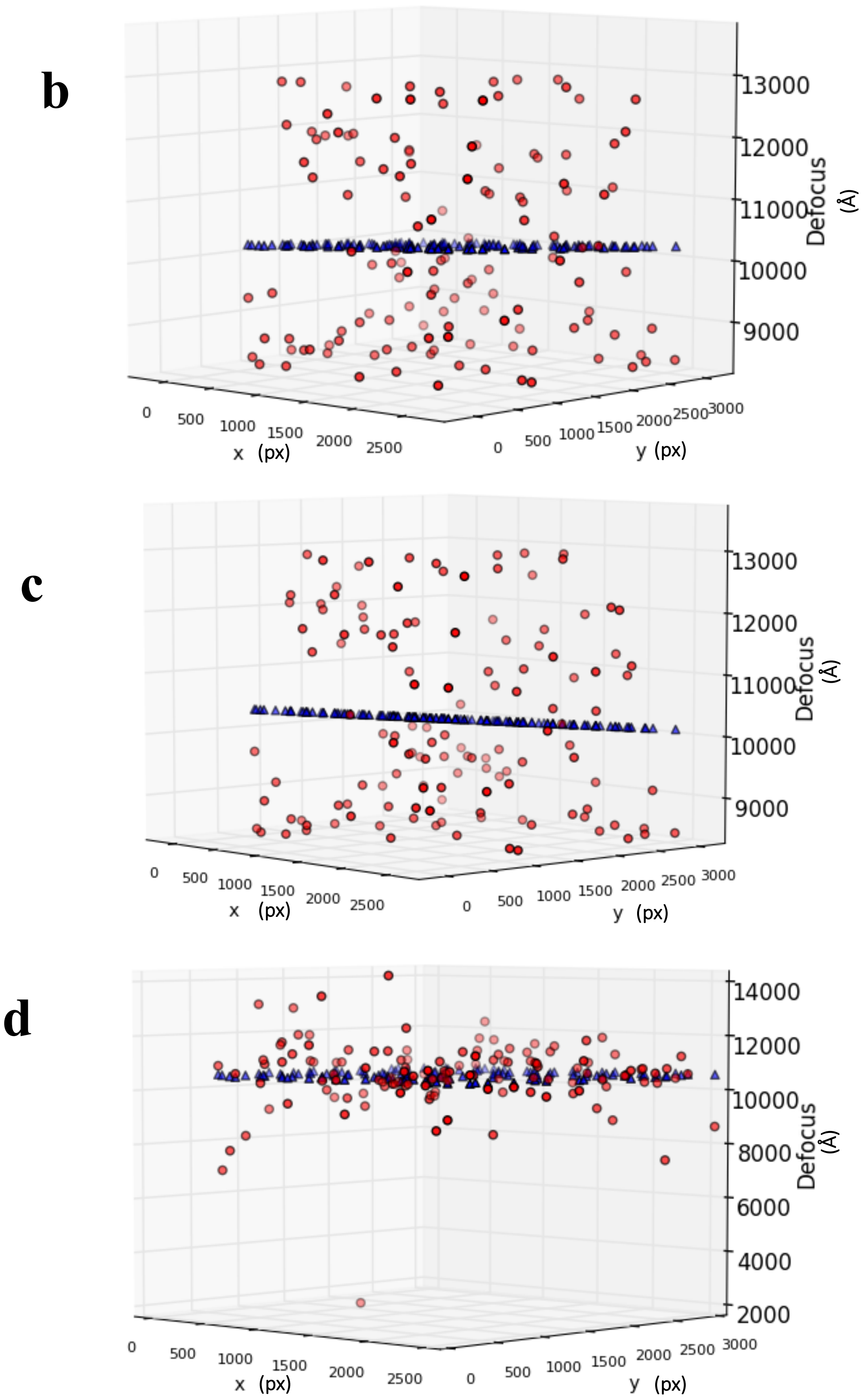
Figure 6. *Particle residuals returned by the comparison protocol grouped by micrograph. The gap indicates that micrographs 18 and 19 have been discarded by a previous protocol.*

In Fig. 6, it can be also appreciated that in all the micrographs there are few outlier particles, which means that they are in a very different height of the sample in comparison with the rest of the particles in the micrograph. Hence, it seems that those particles can be discarded to improve the performance of the following protocols, improving the reconstruction volume. The rejection of the particles with a high residual could be applied automatically to the output of the comparison protocol through the Scipion interface.

In Fig. 7, the analyze protocol viewer is displayed for all the estimations performed on micrograph 1. The adjusted defocus plane is showed in blue and the estimate localization for each of the particles in the micrograph is showed in red for the five estimations. It can be appreciated that the adjustment plane is at different height and has a different inclination for each of the estimations. However, it is similar between the three Relion estimations.

In Gctf, (Fig. 7a) the particles are considerably less scattered in height than in the rest of estimations. Moreover, the particles in this case seem to be grouped in three clusters, the one in the middle above the plane and the other two below the plane. In the case of Relion without astigmatism (Fig. 7b) and Relion with astigmatism per micrograph (Fig. 7c), the particles are distributed in a homogenically way in a range of 4000 nm approximately. In the plots for Relion per particle (Fig. 7d) and Xmipp (Fig. 7e), the distribution of the particles is more adjusted to the plane in both cases, except for several outliers, which seems to be the most coherent distributions according with the theory.




b
13000
12000
11000
10000
9000
Defocus (Å)
0 500 1000 1500 2000 2500
x (px)
0 500 1000 1500 2000 2500 3000
y (px)
c
13000
12000
11000
10000
9000
Defocus (Å)
0 500 1000 1500 2000 2500
x (px)
0 500 1000 1500 2000 2500 3000
y (px)
d
14000
12000
10000
8000
6000
4000
2000
Defocus (Å)
0 500 1000 1500 2000 2500
x (px)
0 500 1000 1500 2000 2500 3000
y (px)

e

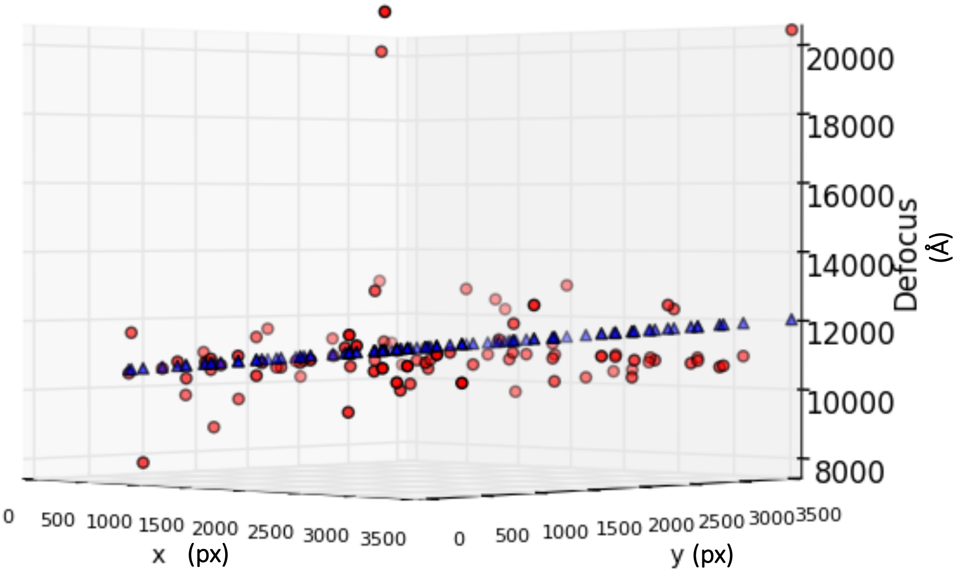
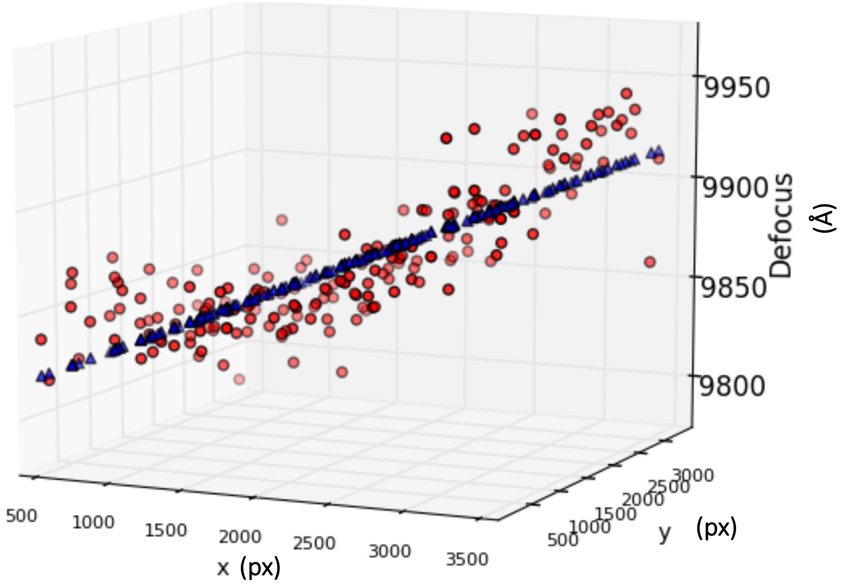


Figure 7. *Analyze protocol viewer for different estimations of micrograph 1. a) Gctf, b) Relion without astigmatism, c) Relion with astigmatism per micrograph, d) Relion with astigmatism per particle, e) Xmipp.*

In Fig. 8 the estimations of Gctf (a), Xmipp (b) and Relion per particle (c) are showed for micrograph 13. The distribution of the particles for each estimation is similar to the ones showed for micrograph 1 (Fig. 7). In Gctf it seems to be three clusters, the one in the middle below the plane in this case and the other two above the plane. Xmipp and Relion per particle show distributions more adjusted to the plane, except for several outliers. The range of height in Gctf is again significantly smaller (one order of magnitude) than in the other two approximations.

In the plot of Gctf it is appreciated how not only the adjustment plane, but also the particles are considerably inclined. A slight inclination could be seen in this case also in Xmipp and Relion, but not as marked as in Gctf due to the larger range of height. The inclination of the particles and the adjustment plane seems to indicate a tilt of the sample during the acquisition process.

a



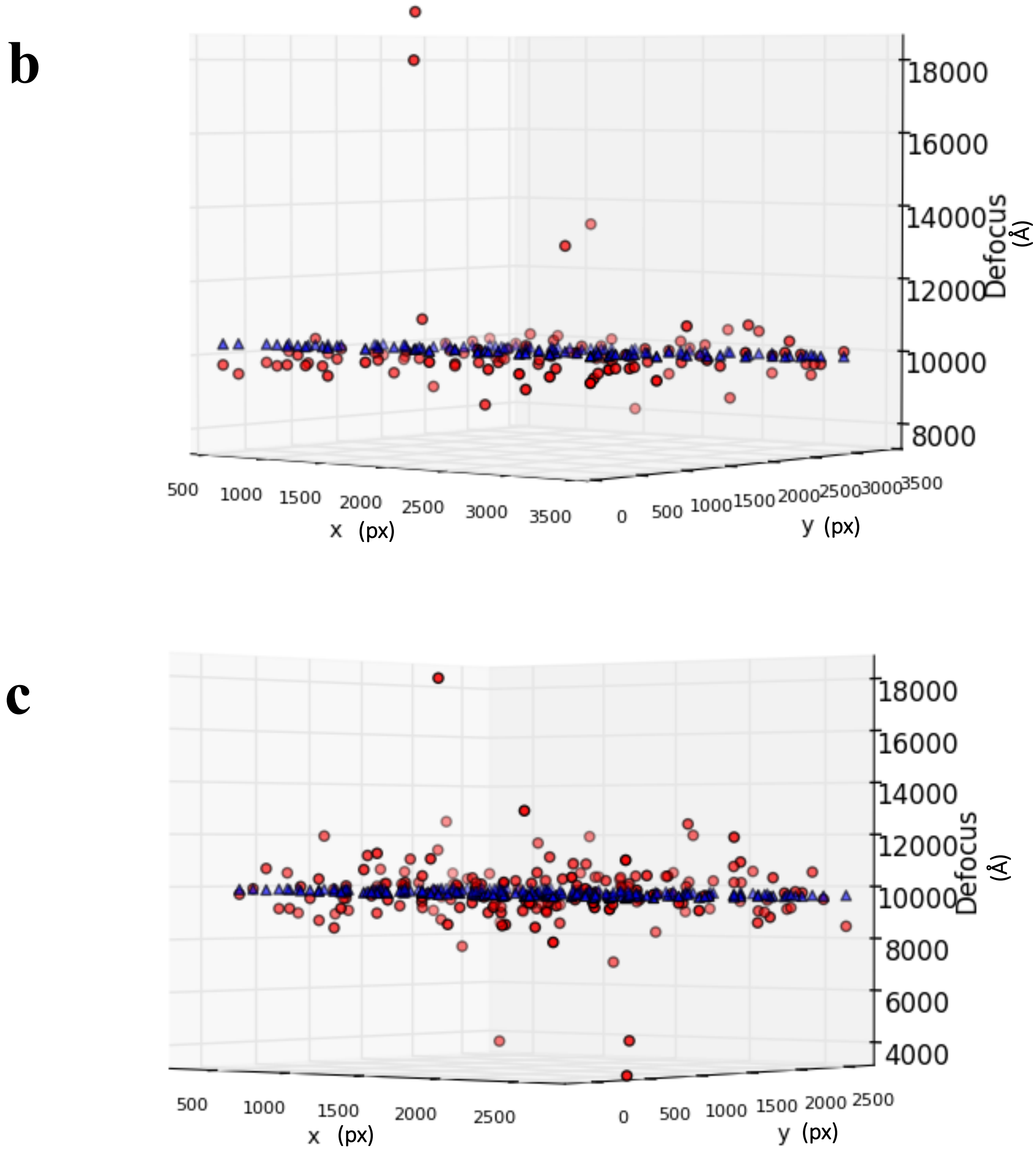


Figure 8. *Analyze protocol viewer for different estimations of micrograph 13. a) Gctf, b) Xmipp, c) Relion with astigmatism per particle. The inclination of the particles and the adjustment planes seems to indicate a tilt of the sample.*

4. Conclusions

The results obtained in this project show that discordance exists between the estimations on local defocus computed by the five different methods under study.

One of the most significant discordance is that Gctf returns small changes in the local defocus estimations, while the rest of the methods return changes greater than one order of magnitude. This discrepancy could be related with the fact that Gctf does not use a reconstructed volume to compute the estimations, while the rest of the methods take into account that volume.

However, the ground truth about the local defocus could not be known and thus, we can not know for the moment which estimations are better than the others. For that reason, we propose in this project a consensus solution between all the estimations performed, the median, due to its greater robustness to the outliers than the mean. We also provide in the developed protocols different tools and measurements to evaluate the differences and degree of discrepancy of the different local defocus estimations for each one of the particles under study.

We consider that being able to analyze the local defocus of the particles instead of just take the global defocus of each micrograph could help in the evaluation of the quality of each one of the particles. Hence, the local defocus of the particles could be considered now another parameter to take into account when discarding particles during the workflow, in order to increase the performance of the classification and reconstruction protocols. This improvement will be eventually translated into a higher quality of the reconstructed final volume.

Nevertheless, further work should be done to strongly confirm the discrepancies between the different local defocus estimation methods. Thus, we are now executing a similar workflow with a larger dataset to see if the same behavior is reproduced.

Furthermore, we have implemented a protocol that is able to locate and display the localization of the particles in each micrograph in the three-dimensional space of the sample. This allows to know the height in which each particle is in the thickness of the sample, without the need of acquiring a tomography of the sample (Fig. 2). Besides, the obtained plot of the particles for each micrograph and the adjustment plane could reveal if the sample has been tilted (Fig. 8).

Finally, this project has contributed with three protocols to the Xmipp package for Scipion. One incorporates a Xmipp functionality in Scipion (estimate local defocus) and the other two allows to integrate, analyze and compare results of different protocols from different packages in Scipion. This integration facilitates the obtaining of a consensus solution from different estimations for a problem without a known ground truth.

References

[1] Vilas Prieto, J. L. (2019). Local quality assessment of cryoem reconstructions and its applications, From local resolution to local sharpening. Facultad de Ciencias. Universidad Autónoma de Madrid.

[2] Zhang, K. (2016). Gctf: Real-time CTF determination and correction. Journal of Structural Biology, 193 (1), pp. 1-12.

[3] Noble, A., Dandey, V., Wei, H., Brasch, J., Chase, J., Acharya, P., Tan, Y., Zhang, Z., Kim, L., Scapin, G., Rapp, M., Eng, E., Rice, W., Cheng, A., Negro, C., Shapiro, L., Kwong, P., Jeruzalmi, D., des Georges, A., Potter, C. and Carragher, B. (2018). Routine single particle CryoEM sample and grid characterization by tomography. eLife, 7.

[4] Zivanov, J., Nakane, T., Forsberg, B., Kimanius, D., Hagen, W., Lindahl, E. and Scheres, S. (2018). New tools for automated high-resolution cryo-EM structure determination in RELION-3. eLife, 7.

[5] de la Rosa-Trevín, J., Otón, J., Marabini, R., Zaldívar, A., Vargas, J., Carazo, J. and Sorzano, C. (2013). Xmipp 3.0: An improved software suite for image processing in electron microscopy. Journal of Structural Biology, 184(2), pp.321-328.

[6] de la Rosa-Trevín, J., Quintana, A., del Cano, L., Zaldívar, A., Foche, I., Gutiérrez, J., Gómez-Blanco, J., Burguet-Castell, J., Cuenca-Alba, J., Abrishami, V., Vargas, J., Otón, J., Sharov, G., Vilas, J., Navas, J., Conesa, P., Kazemi, M., Marabini, R., Sorzano, C. and Carazo, J. (2016). Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy. Journal of Structural Biology, 195(1), pp.93-99.