

Machine learning needs and trends for Single Particle Analysis

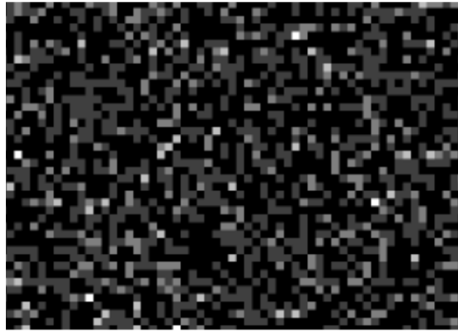
C.O.S. Sorzano

Biocomputing Unit, CNB-CSIC

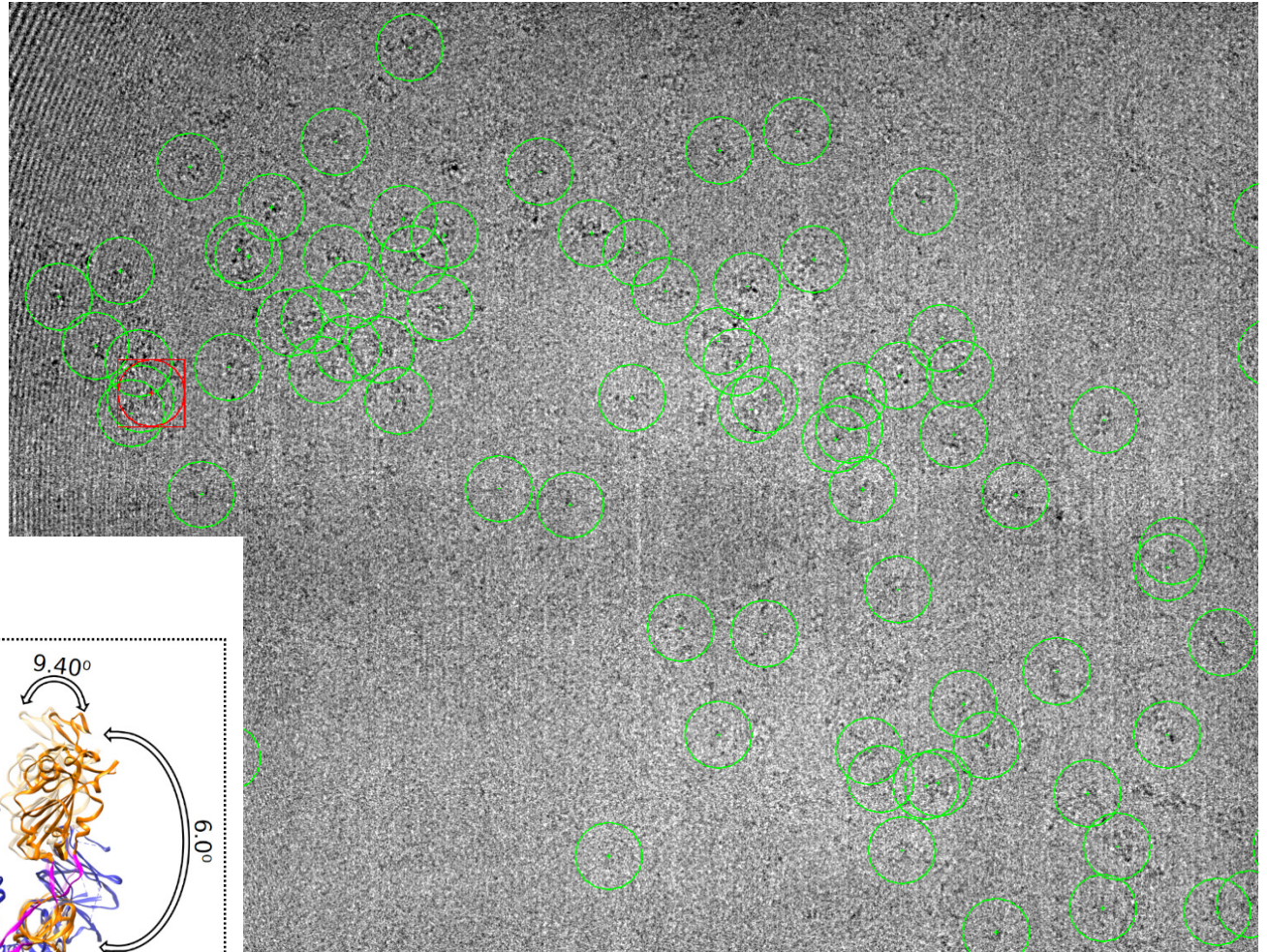
Instruct Image Processing Center



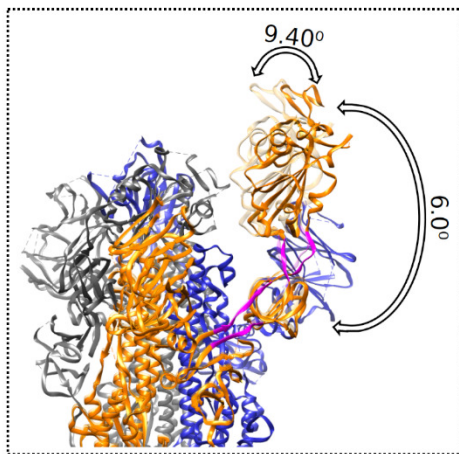
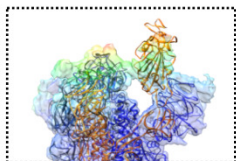
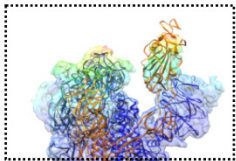
Raw data



SNR \approx 0.001



SNR \approx 0.01



For more information

CHEMICAL REVIEWS

pubs.acs.org/CR



Review

Emerging Themes in CryoEM—Single Particle Analysis Image Processing

Jose Luis Vilas, Jose Maria Carazo,* and Carlos Oscar S. Sorzano*



Cite This: <https://doi.org/10.1021/acs.chemrev.1c00850>



Read Online

ACCESS |

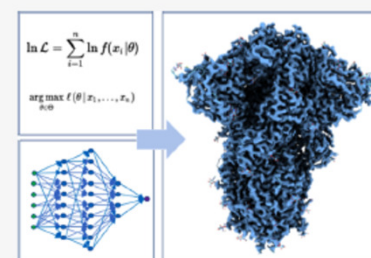


Metrics & More



Article Recommendations

ABSTRACT: Cryo-electron microscopy (CryoEM) has become a vital technique in structural biology. It is an interdisciplinary field that takes advantage of advances in biochemistry, physics, and image processing, among other disciplines. Innovations in these three basic pillars have contributed to the boosting of CryoEM in the past decade. This work reviews the main contributions in image processing to the current reconstruction workflow of single particle analysis (SPA) by CryoEM. Our review emphasizes the time evolution of the algorithms across the different steps of the workflow differentiating between two groups of approaches: analytical methods and deep learning algorithms. We present an analysis of the current state of the art. Finally, we discuss the emerging problems and challenges still to be addressed in the evolution of CryoEM image processing methods in SPA.



For more information

Journal of Structural Biology 214 (2022) 107861



Contents lists available at ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi



Cryo-Electron Microscopy: The field of 1,000⁺ methods

C.O.S. Sorzano^{*}, J.M. Carazo

Natl. Center of Biotechnology, CSIC, c/Darwin, 3, Campus Univ. Autónoma de Madrid, 28049 Madrid, Spain

ARTICLE INFO

Keywords:
cryoEM
Method development
Image processing

ABSTRACT

Cryo-Electron Microscopy (CryoEM) is currently a well-established method to elucidate a biological macromolecule's three-dimensional (3D) structure. Its success is due to technological and methodological advances in several fronts: sample preparation, electron optics and detection, image acquisition, image processing, and map interpretation. The first methods started in the late 1960s and, since then, new methods on all fronts have continuously been published, maturing the field as we know it now.

In terms of publications, we can distinguish several periods, witnessing a substantial acceleration of methodological publications in recent years, pointing out to an increased interest in the domain. On the other hand, this accelerated increase of methods development may confuse practitioners about which method they should be using (and how) and highlight the importance of paying attention to establishing best practices for methods reporting and usage.

In this paper, we analyze the trends identified in over 1,000 methodological papers. Our focus is primarily on computational image processing methods. However, our list also covers some aspects of sample preparation and image acquisition.

Several interesting ideas stem out from this study: (1) Single Particle Analysis (SPA) has largely accelerated in the last decade and sample preparation methods in the last five years; (2) Electron Tomography is not yet in a rapidly growing phase, but it is foreseeable that it will soon be; (3) the work horses of SPA are 3D classification, 3D reconstruction, and 3D alignment, and there have been many papers on these topics, which are not considered to be solved yet, but ever improving; and (4) since the resolution revolution, atomic modelling has also caught on as a hot topic.



For more information

research papers



STRUCTURAL
BIOLOGY

ISSN 2059-7983

Received 9 August 2021
Accepted 18 February 2022

Edited by T. Burnley, Rutherford Appleton
Laboratory, United Kingdom

Keywords: single-particle analysis; cryo-electron
microscopy; parameter estimation; image
processing; bias; variance; overfitting; gold
standard.

Supporting information: this article has
supporting information at journals.iucr.org/d

On bias, variance, overfitting, gold standard and consensus in single-particle analysis by cryo-electron microscopy

C. O. S. Sorzano,^{a,*} A. Jiménez-Moreno,^a D. Maluenda,^a M. Martínez,^a E. Ramírez-
Aportela,^a J. Krieger,^a R. Melero,^a A. Cuervo,^a J. Conesa,^a J. Filipovic,^b P. Conesa,^a
L. del Caño,^a Y. C. Fonseca,^a J. Jiménez-de la Morena,^a P. Losana,^a R. Sánchez-
García,^a D. Strelak,^{a,b} E. Fernández-Giménez,^a F. P. de Isidro-Gómez,^a
D. Herreros,^a J. L. Vilas,^c R. Marabini^d and J. M. Carazo^{a,*}

^aBiocomputing Unit, Centro Nacional de Biotecnología (CNB-CSIC), Calle Darwin 3, 28049 Cantoblanco, Madrid, Spain,
^bMasaryk University, Brno, Czech Republic, ^cSchool of Engineering and Applied Science, Yale University, New Haven,
CT 06520-829, USA, and ^dEscuela Politécnica Superior, Universidad Autónoma de Madrid, 28049 Cantoblanco, Madrid,
Spain. *Correspondence e-mail: cos@cnb.csic.es, carazo@cnb.csic.es

Cryo-electron microscopy (cryoEM) has become a well established technique to elucidate the 3D structures of biological macromolecules. Projection images from thousands of macromolecules that are assumed to be structurally identical are combined into a single 3D map representing the Coulomb potential of the macromolecule under study. This article discusses possible caveats along the image-processing path and how to avoid them to obtain a reliable 3D structure.



For more information

Faraday Discussions









Cite this: DOI: 10.1039/d2fd00059h



PAPER

[View Article Online](#)
[View Journal](#)

Image processing tools for the validation of CryoEM maps†

C. O. S. Sorzano, ^{*a} J. L. Vilas,^a E. Ramírez-Aportela,^a J. Krieger, ^a
D. del Hoyo,^a D. Herreros,^a E. Fernandez-Giménez,^a D. Marchán, ^a
J. R. Macías, ^a I. Sánchez, ^a L. del Caño,^a Y. Fonseca-Reyna,^a
P. Conesa,^a A. García-Mena,^a J. Burguet, ^b J. García Condado, ^c
J. Méndez García,^d M. Martínez, ^a A. Muñoz-Barrutia,^e R. Marabini,^f
J. Vargas^b and J. M. Carazo^a

Received 7th March 2022, Accepted 4th April 2022

DOI: 10.1039/d2fd00059h

The number of maps deposited in public databases (Electron Microscopy Data Bank, EMDB) determined by cryo-electron microscopy has quickly grown in recent years. With this rapid growth, it is critical to guarantee their quality. So far, map validation has primarily focused on the agreement between maps and models. From the image



Machine learning

$$\arg \max_{\hat{Y}_0} f_{Y_0|Y}(\hat{Y}_0|Y) = \arg \max_{\hat{Y}_0} \frac{f_{Y|Y_0}(Y|\hat{Y}_0)f_{Y_0}(\hat{Y}_0)}{\int f_{Y|Y_0}(Y|Y'_0)f_{Y_0}(Y'_0)dY'_0}$$

Y are the observations
 Y_0 is the model

Gaussian noise
 Statistical prior on parameters

$$\arg \min_{\Theta} \sum_{j=1}^P \left\| Y_j - f_{\Theta}(X_j) \right\|^2 + \lambda \Phi(\Theta)$$

- The role of X and Y are played by different elements.
- Deep learning is simply a “sophisticated” f .

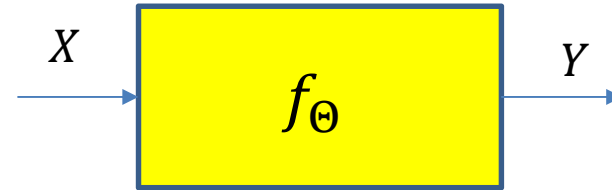


J.L. Vilas, J.M. Carazo, C.O.S. Sorzano. Emerging themes in cryoEM-SPA Image processing. Chemical reviews (in press)



Machine learning

- Movie alignment:
 - Y : micrograph
 - X : frames
 - Θ : alignment parameters
- CTF determination:
 - Y : Power Spectrum Density
 - X : White noise
 - Θ : microscope parameters
- Particle picking:
 - Y : a coordinate in a micrograph is the center of a particle (T or F)
 - X : micrograph
 - Θ : picking model



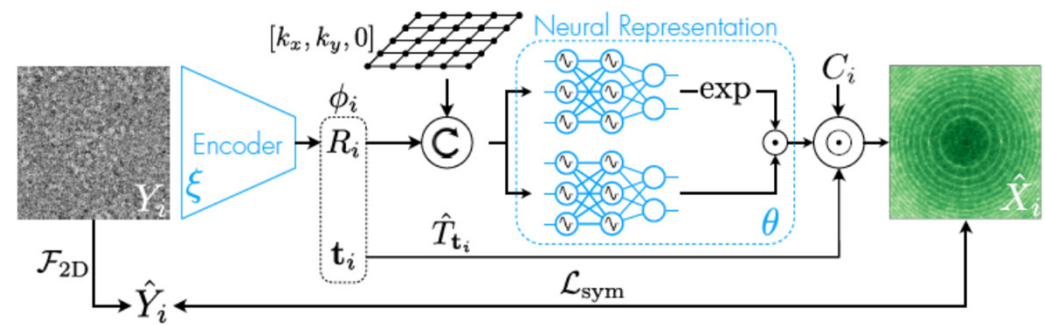
J.L. Vilas, J.M. Carazo, C.O.S. Sorzano. Emerging themes in cryoEM-SPA Image processing. Chemical reviews (in press)



Classical Image Processing or Deep Learning

$$\begin{aligned}
 I_{\tilde{A},i} &= \int_{-\infty}^{\infty} V(\tilde{A}^{-1} \tilde{H}^T \tilde{s}_i) dt \\
 &= \int_{-\infty}^{\infty} \left(\sum_j x_j b(\tilde{A}^{-1} \tilde{H}^T \tilde{s}_i - \tilde{r}_j) \right) dt \\
 &= \sum_j x_j \left(\int_{-\infty}^{\infty} b(\tilde{A}^{-1} \tilde{H}^T \tilde{s}_i - \tilde{r}_j) dt \right) = \sum_j x_j a_{\tilde{A},ij},
 \end{aligned}$$

C.O.S. Sorzano, J. Vargas, J. Oton, J.L. Vilas, M. Kazemi, R. Melero, L. del Caño, J. Cuenca, P. Conesa, J. Gomez-Blanco, R. Marabini, J.M. Carazo. A survey of the use of iterative reconstruction algorithms in Electron Microscopy. BioMed Research Intl. 6482567 (2017)



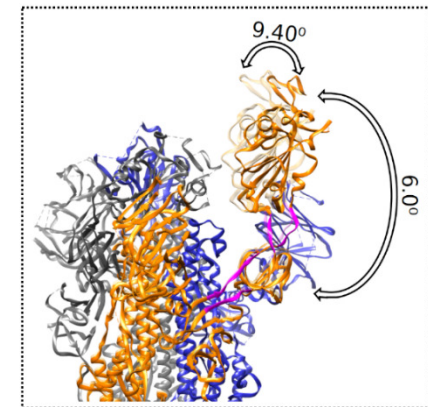
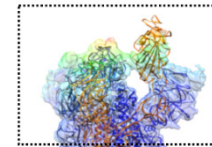
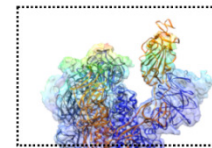
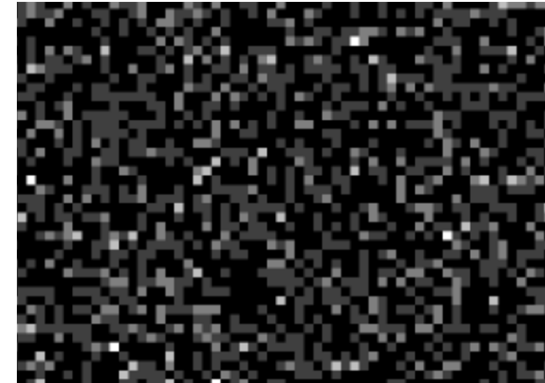
Levy, A., Poitevin, F., Martel, J., Nashed, Y., Peck, A., Miolane, N., ... & Wetzstein, G. (2022). Cryoai: Amortized inference of poses for ab initio reconstruction of 3d molecular volumes from real cryo-em images. arXiv preprint arXiv:2203.08138.



We have a problem to solve



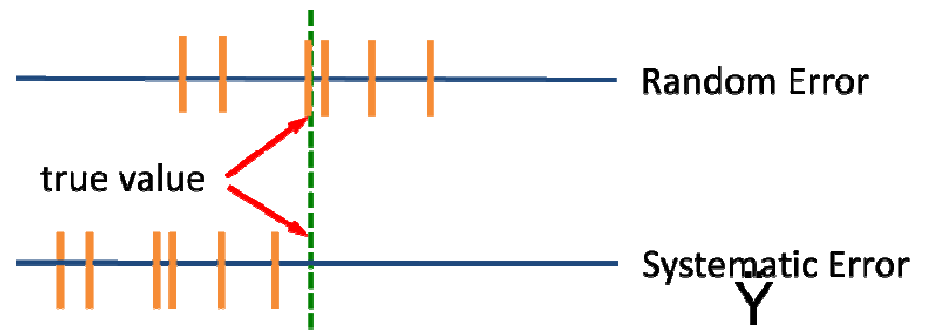
No matter if it is a white cat or a black cat;
as long as it can catch mice, it is a good cat



Bias and variance

$$V_{reconstructed} = V_{ideal} + \Delta V$$

Random Error vs. Systematic Error



$E(\Delta V)$ **Bias:** systematic errors

$\text{Cov}(\Delta V)$ **Variance:** random errors with zero mean

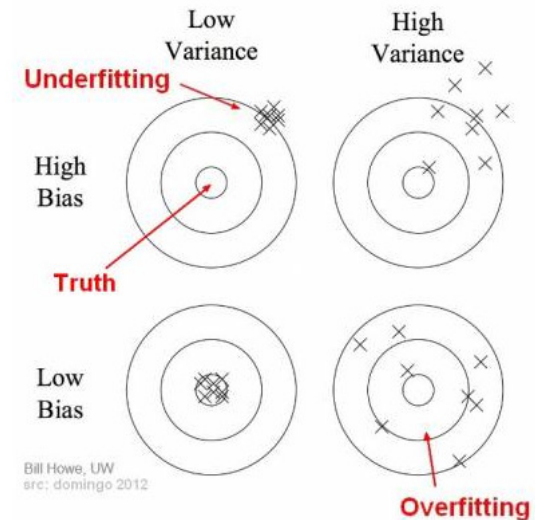
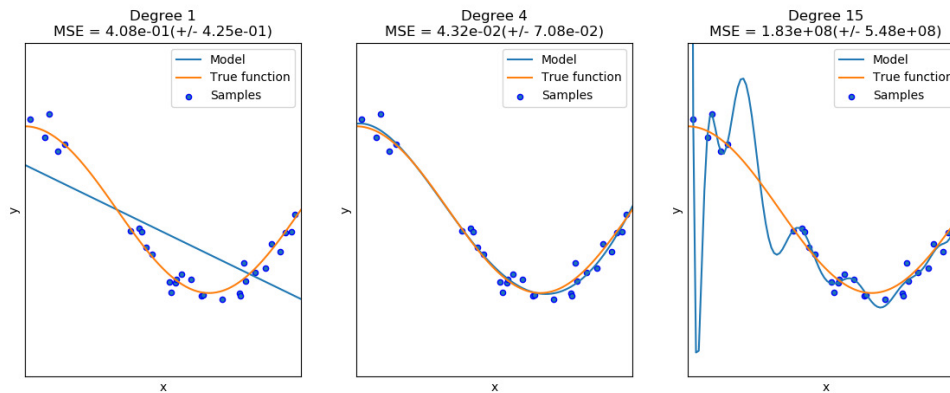
$$V_{reconstructed} = pV_{correct} + (1 - p)V_{incorrect}$$



C.O.S. Sorzano, ... On bias, variance, overfitting, gold standard and consensus in Single Particle Analysis by Cryo-electron microscopy. Acta Crystallographica Section D, D78: 410-423 (2022)

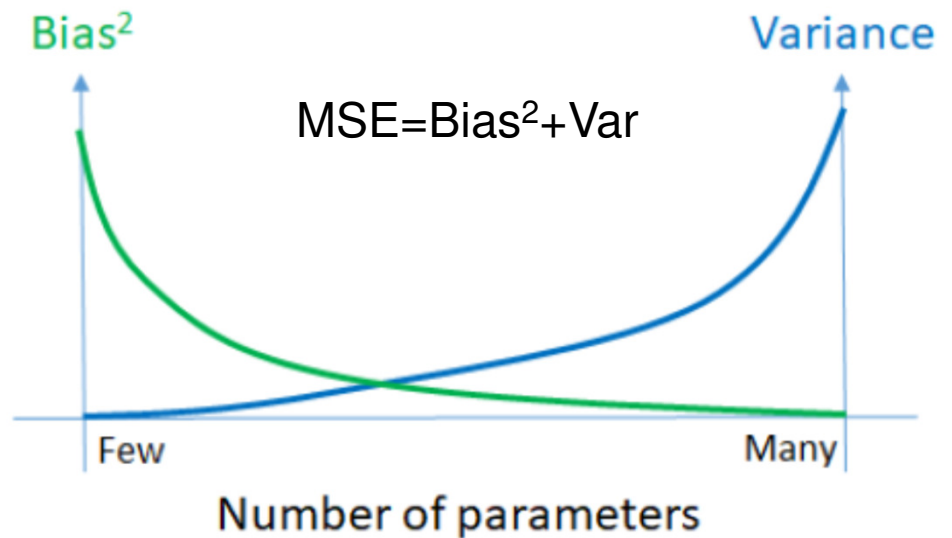


Reconstruction is all about parameter estimation

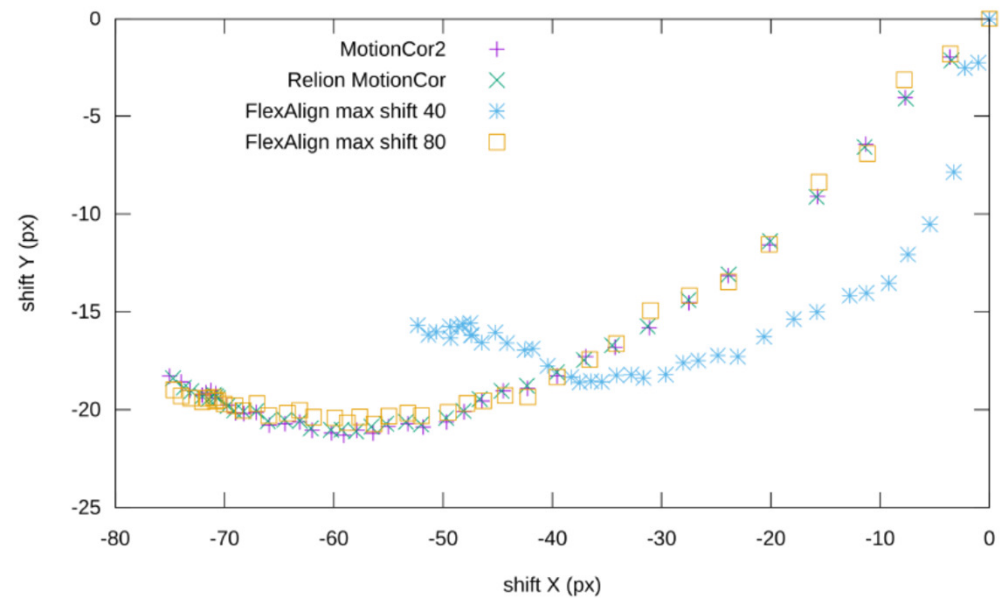
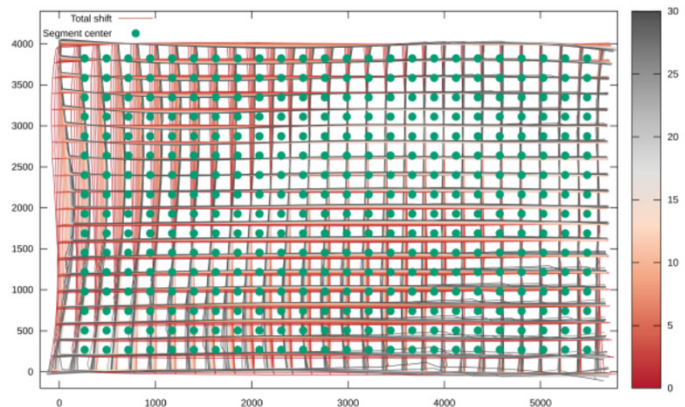
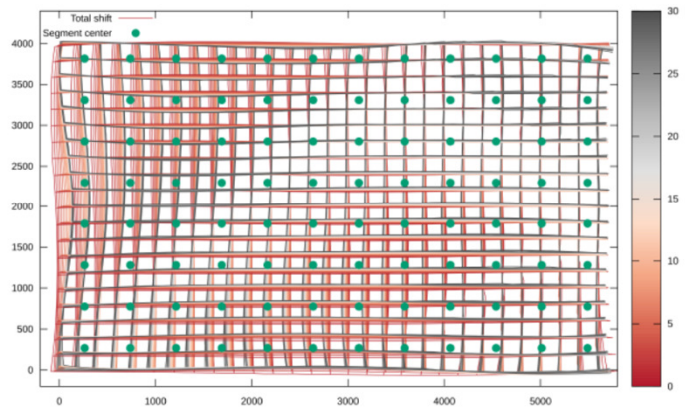
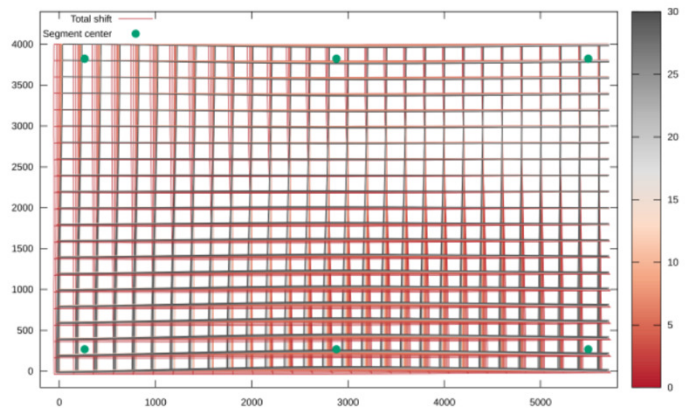
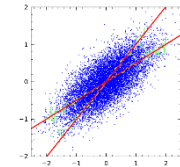


$$BIC = 2 \log P\{y|\hat{\theta}\} - k \log N$$

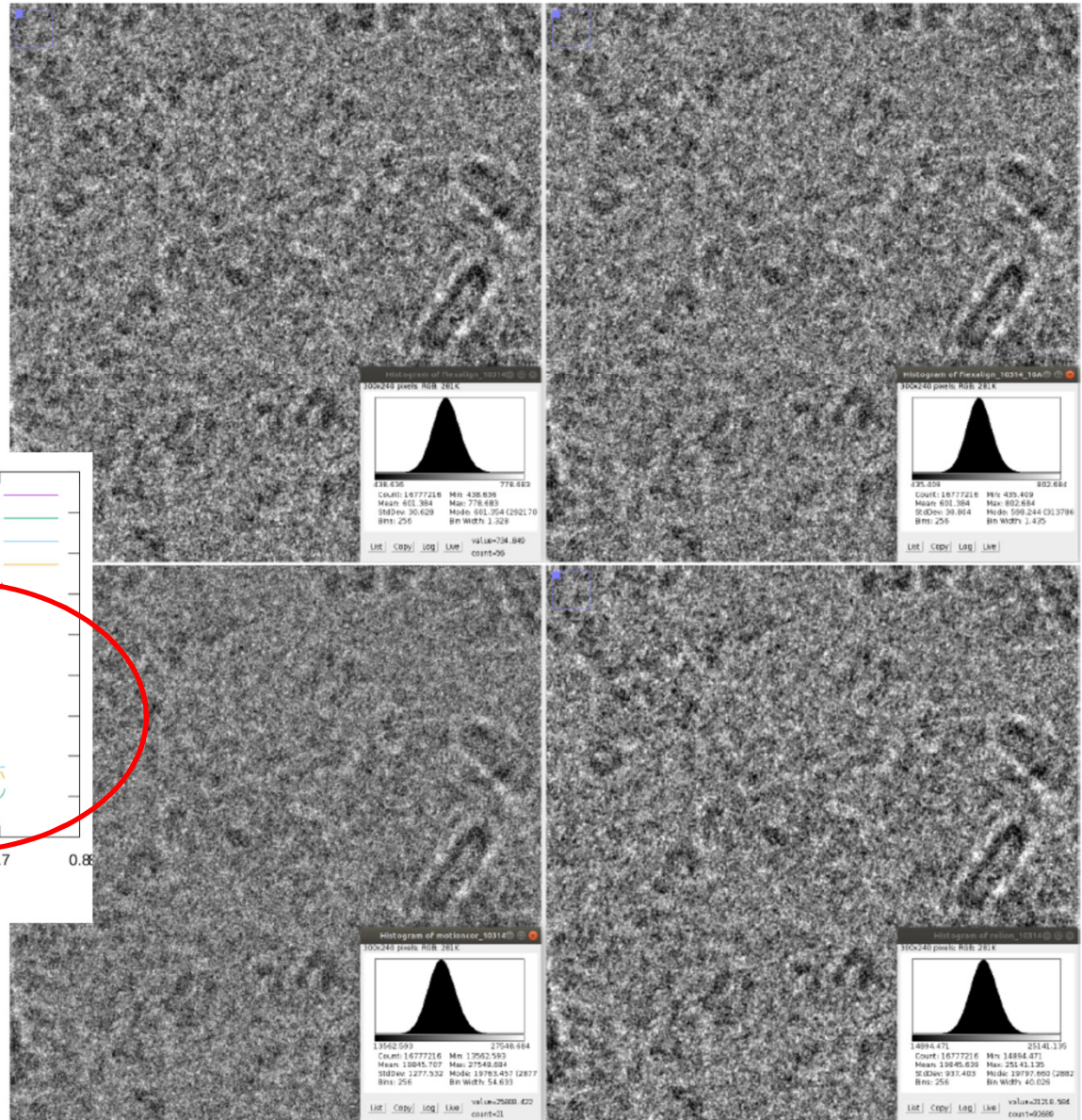
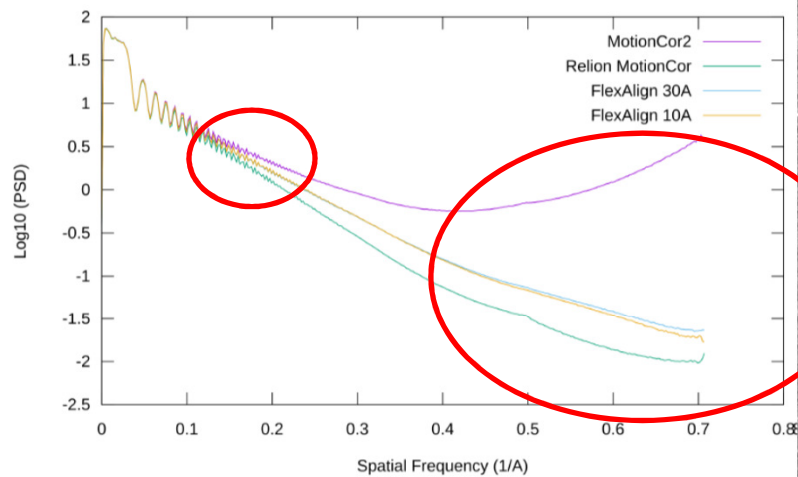
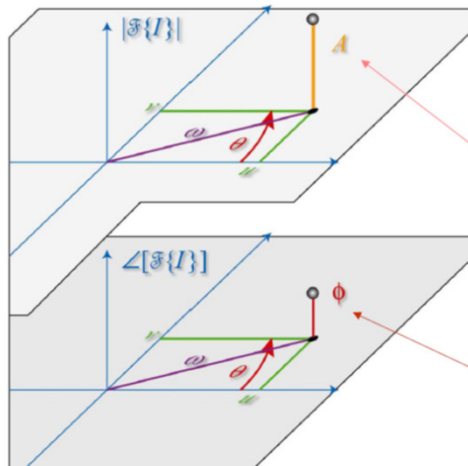
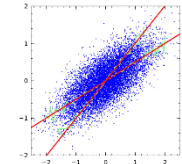
6,666 measurements/parameter



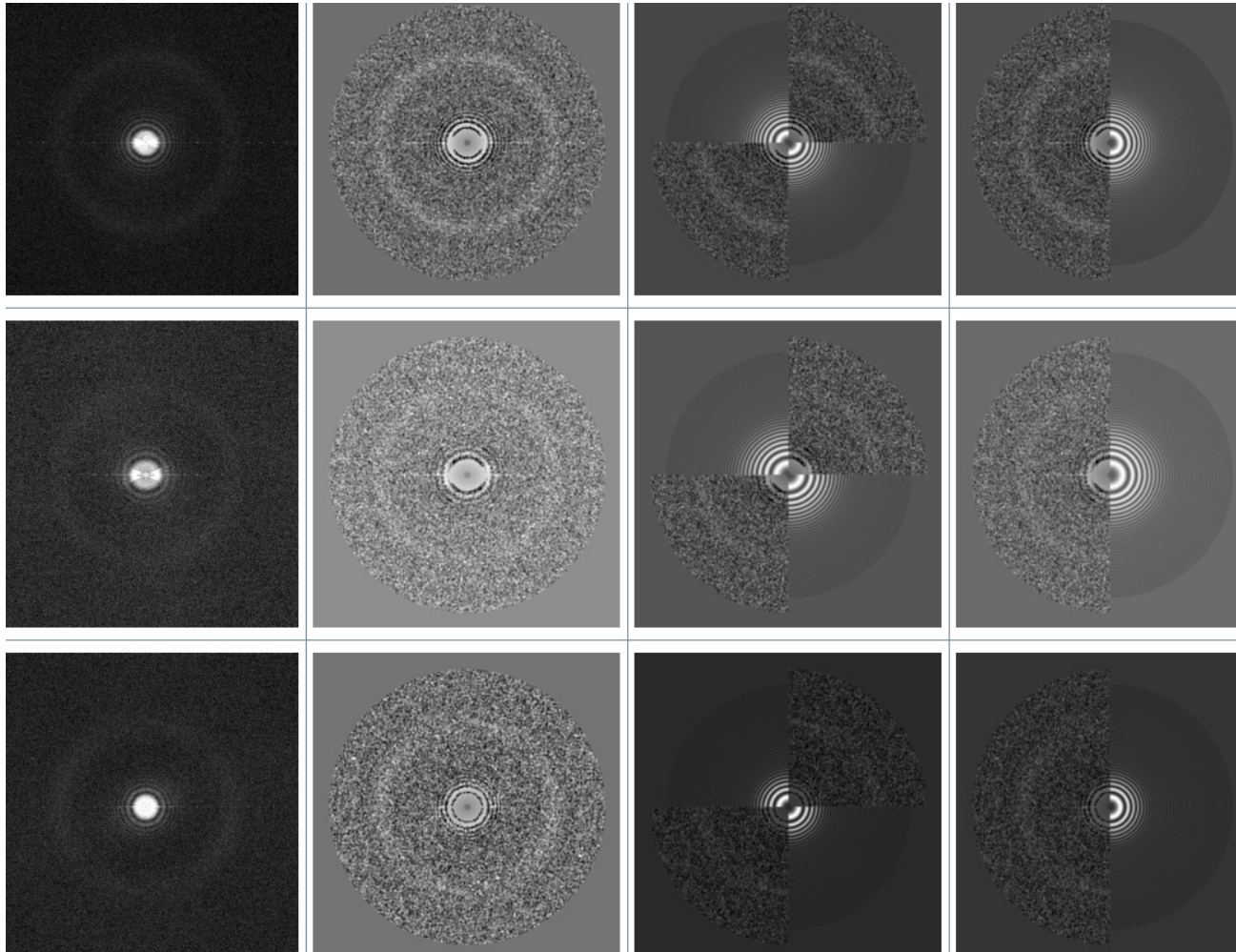
Perfect frame alignment



Perfect frame alignment



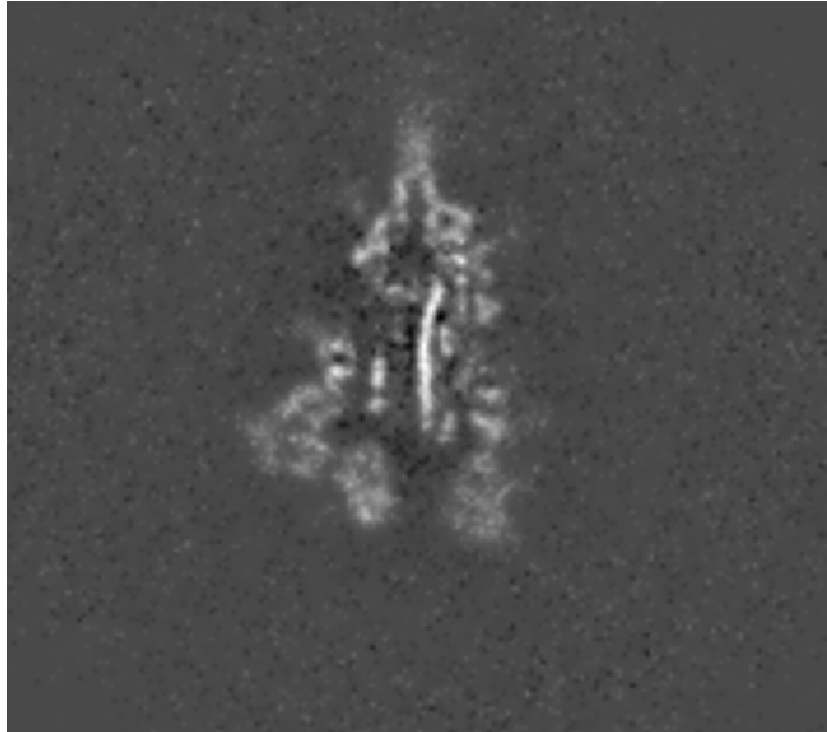
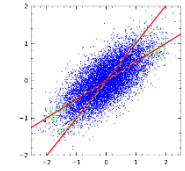
Accurate defocus



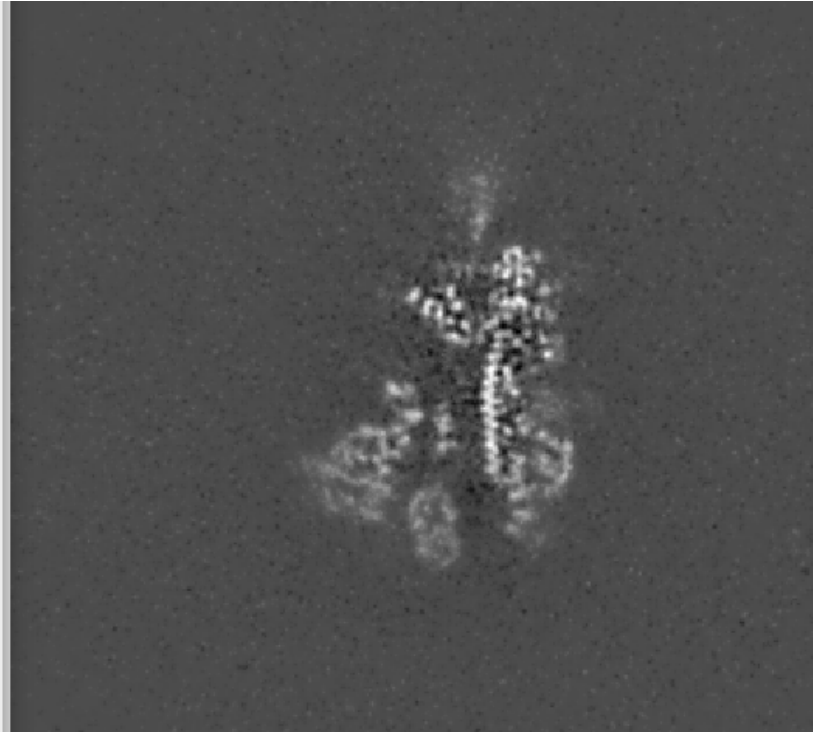
C.O.S. Sorzano, ... On bias, variance, overfitting, gold standard and consensus in Single Particle Analysis by Cryo-electron microscopy. Acta Crystallographica Section D, D78: 410-423 (2022)



Accurate defocus



CTF standard



CTF agrees up to 2.1Å

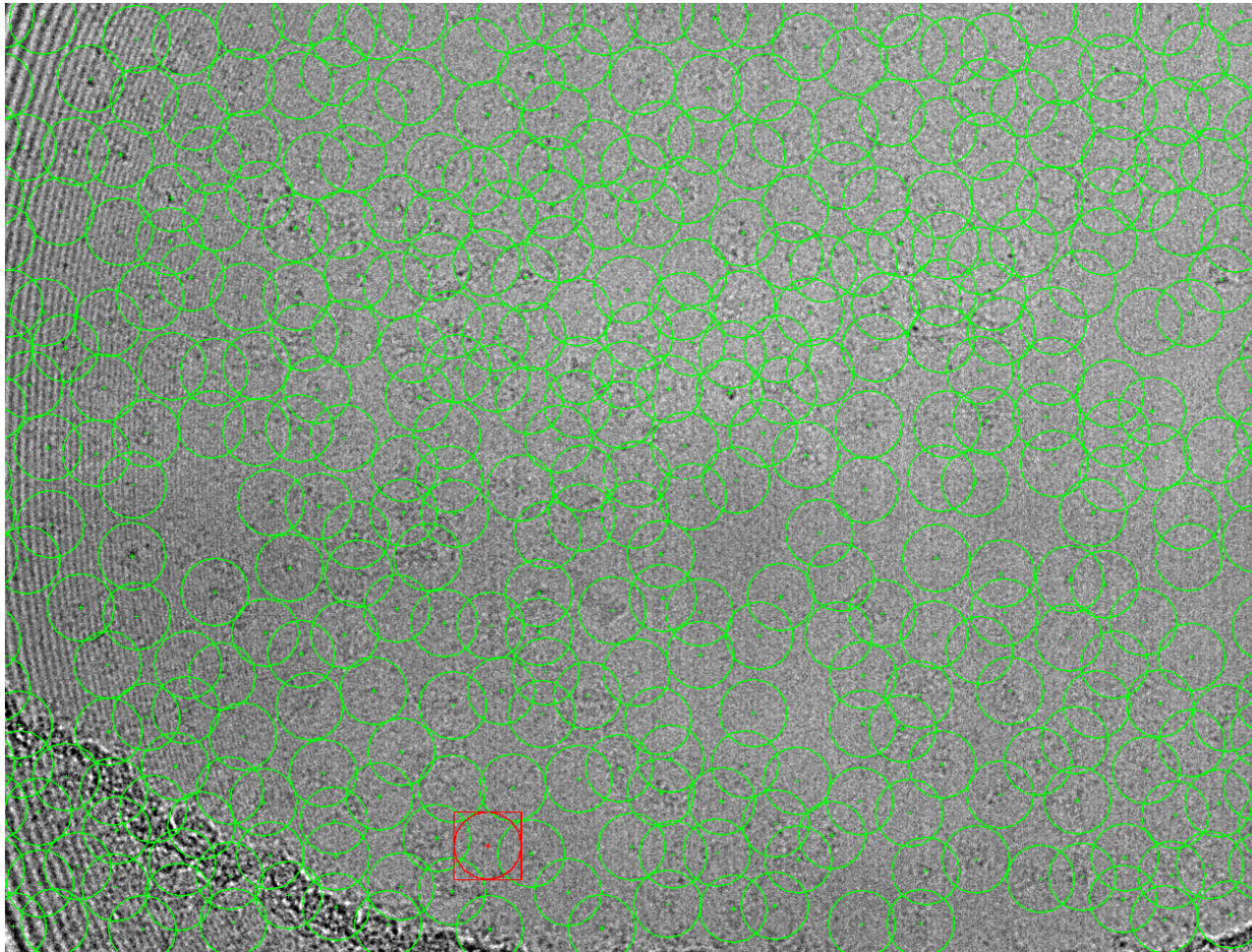
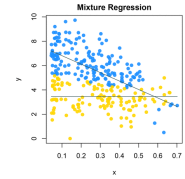
GCTF+CTFFind+Consensus 2.1Å
Xmipp CTF for the envelope



C.O.S. Sorzano, ... On bias, variance, overfitting, gold standard and consensus in Single Particle Analysis by Cryo-electron microscopy. Acta Crystallographica Section D, D78: 410-423 (2022)



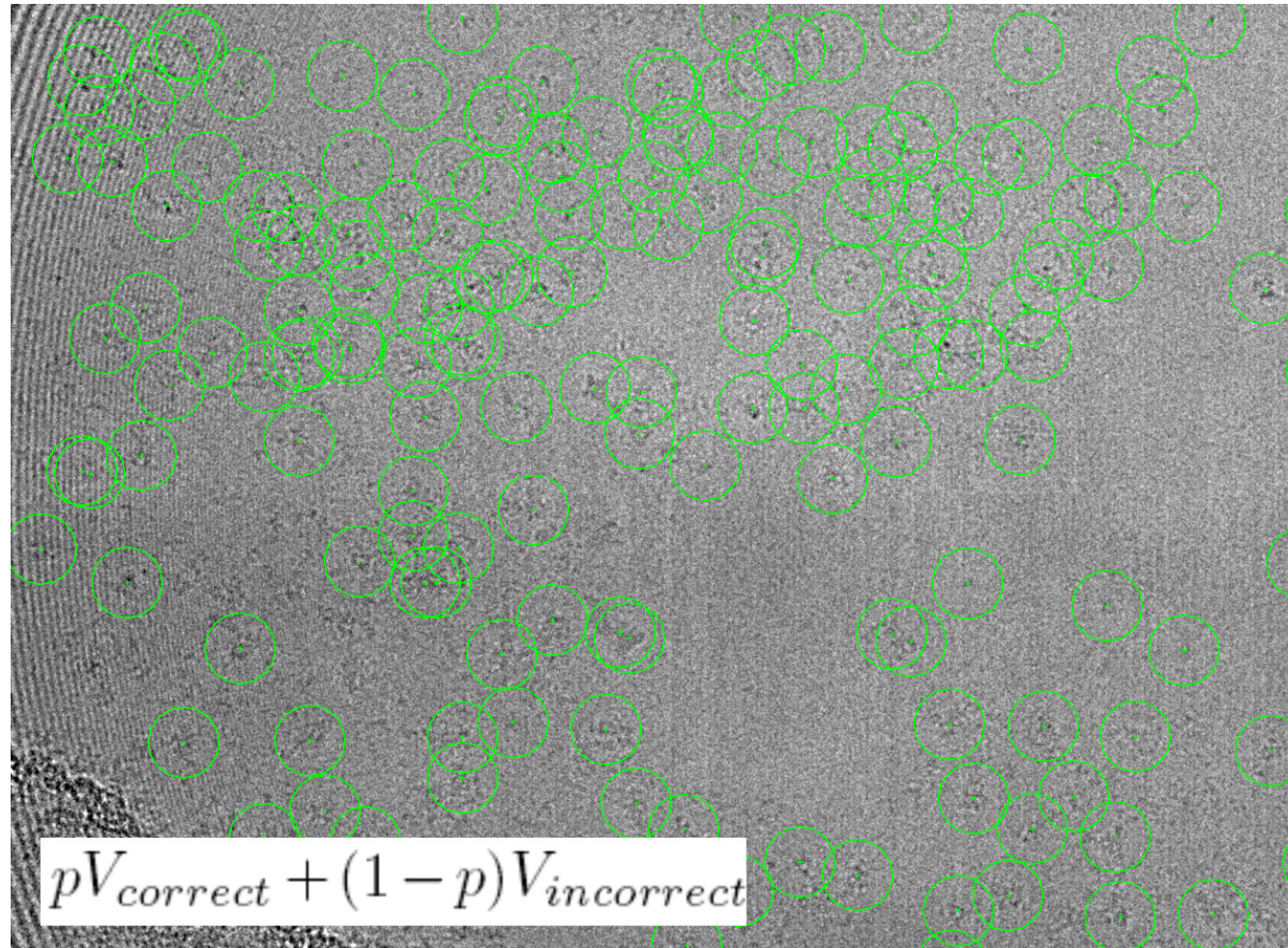
Picking (1.2 M)



C.O.S. Sorzano, ... On bias, variance, overfitting, gold standard and consensus in Single Particle Analysis by Cryo-electron microscopy. Acta Crystallographica Section D, D78: 410-423 (2022)



Picking (Multiple pickers + Deep Consensus + Mic. Cleaner 569k)



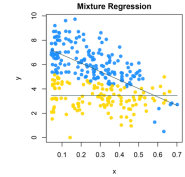
instruct
Integrating
Biology

C.O.S. Sorzano, ... On bias, variance, overfitting, gold standard and consensus in Single Particle Analysis by Cryo-electron microscopy. Acta Crystallographica Section D, D78: 410-423 (2022)



instruct
image
Processing
Center

2D Analysis, Round 1



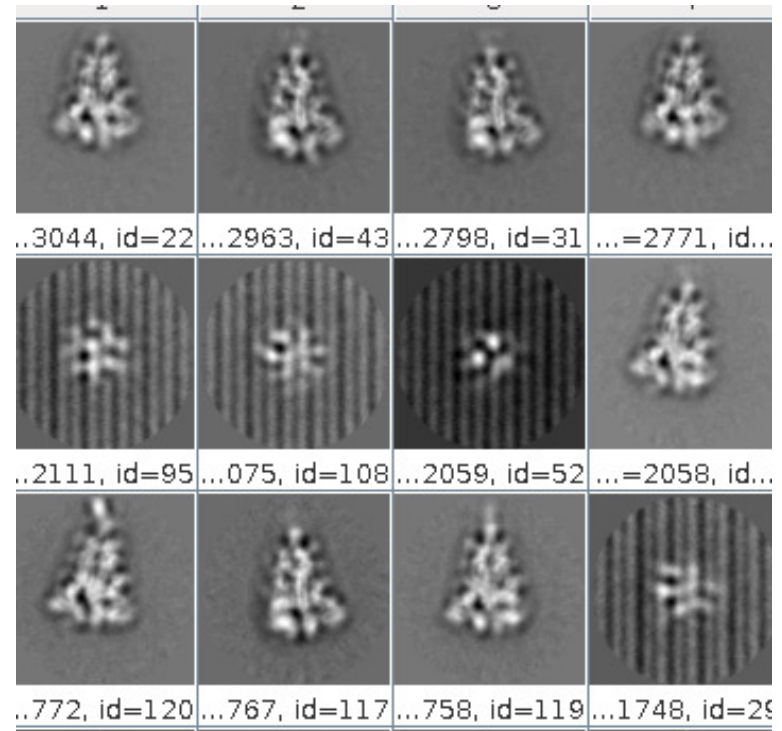
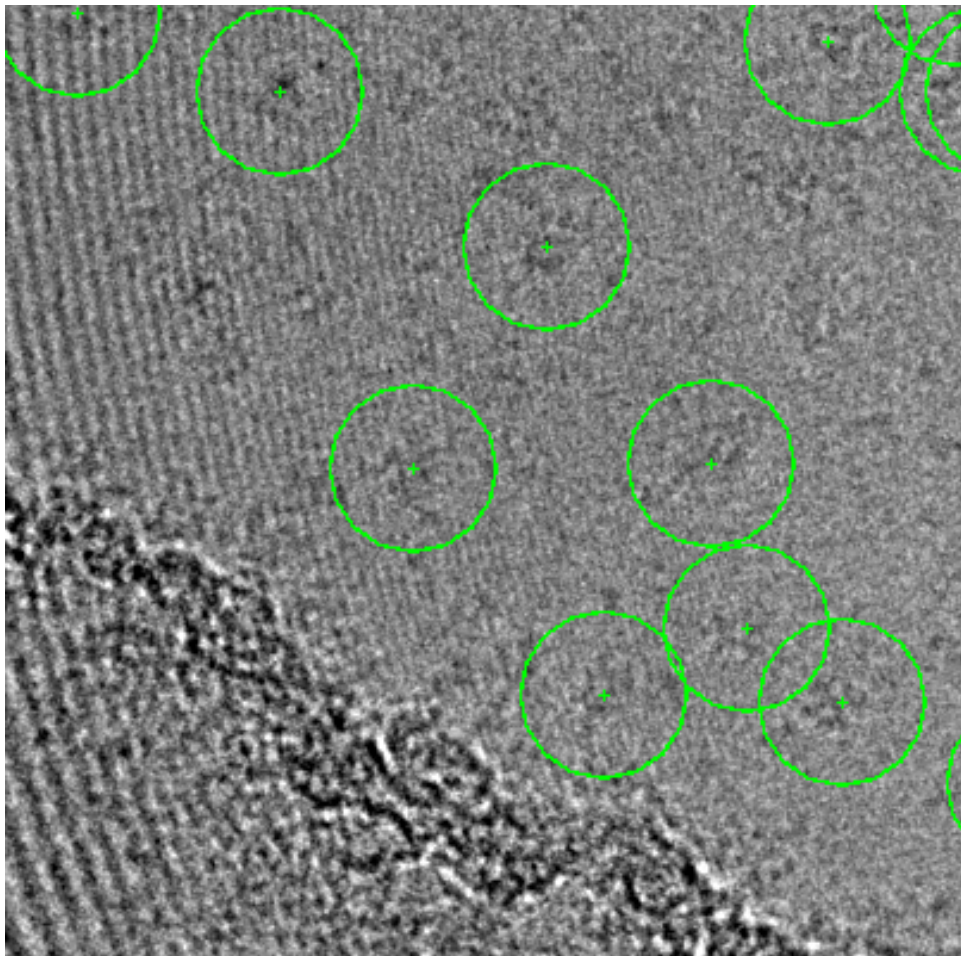
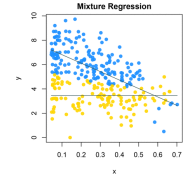
$$pV_{correct} + (1 - p)V_{incorrect}$$



C.O.S. Sorzano, ... On bias, variance, overfitting, gold standard and consensus in Single Particle Analysis by Cryo-electron microscopy. Acta Crystallographica Section D, D78: 410-423 (2022)



Avoid aperture?



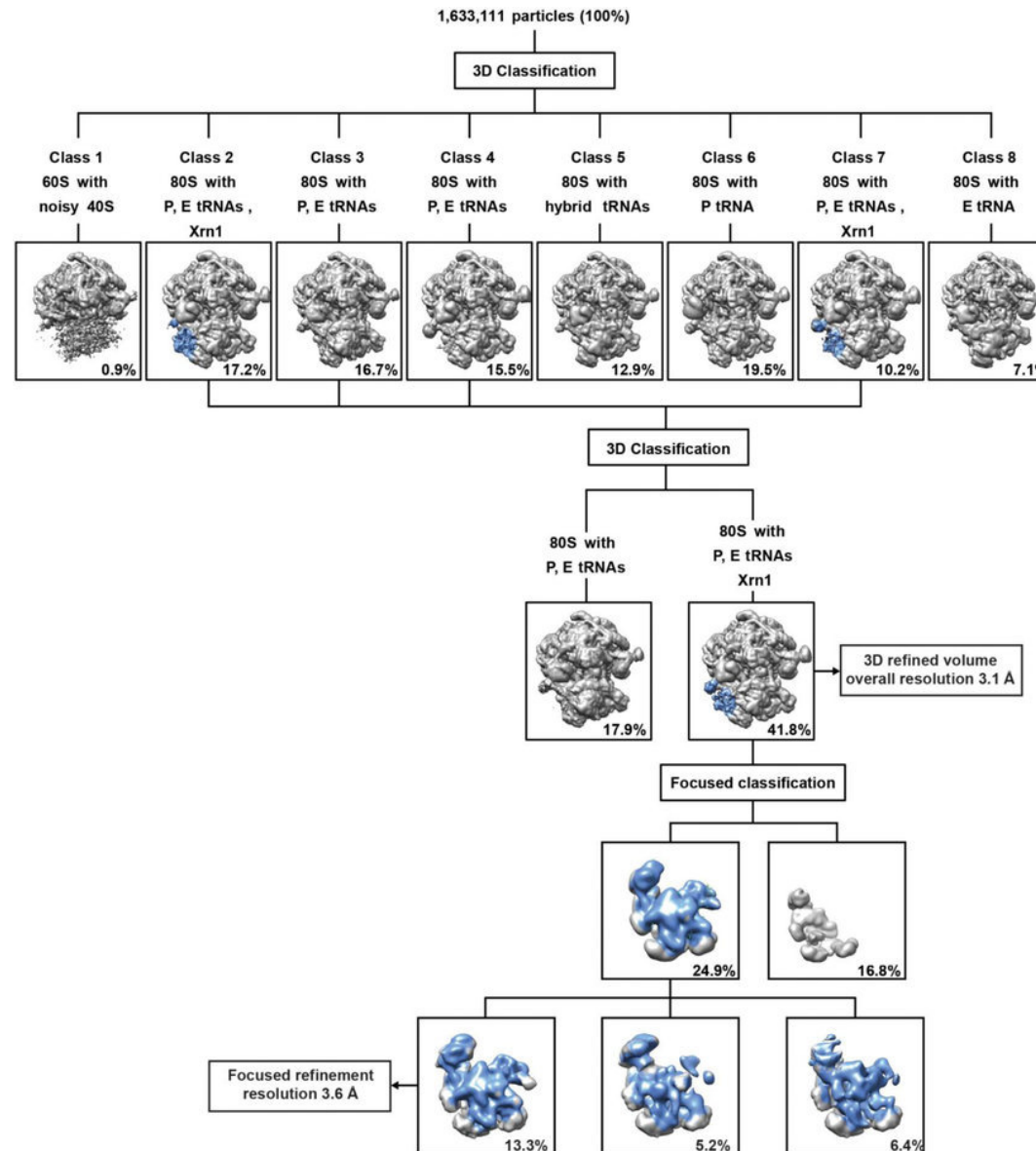
$$pV_{correct} + (1 - p)V_{incorrect}$$



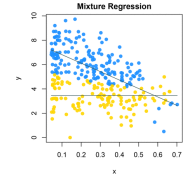
C.O.S. Sorzano, ... On bias, variance, overfitting, gold standard and consensus in Single Particle Analysis by Cryo-electron microscopy. Acta Crystallographica Section D, D78: 410-423 (2022)



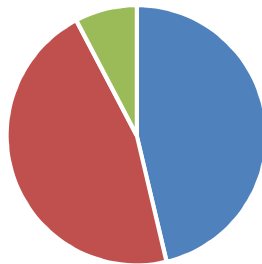
3D Analysis



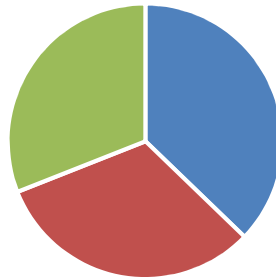
3D Analysis



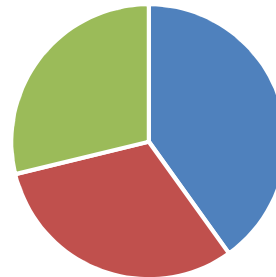
Classification 1



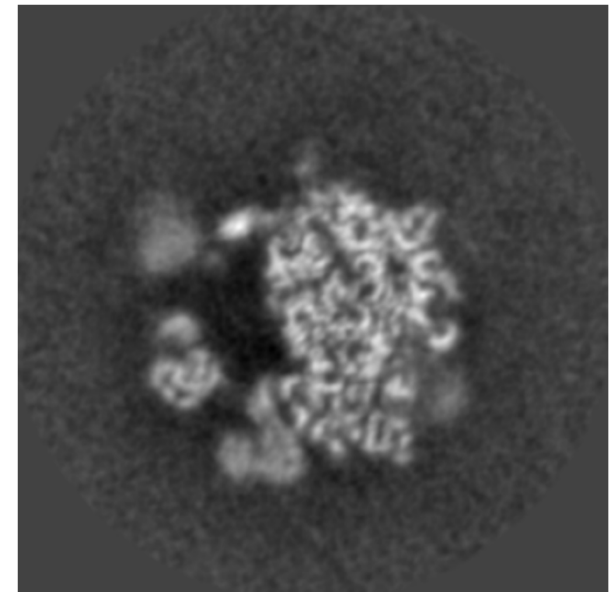
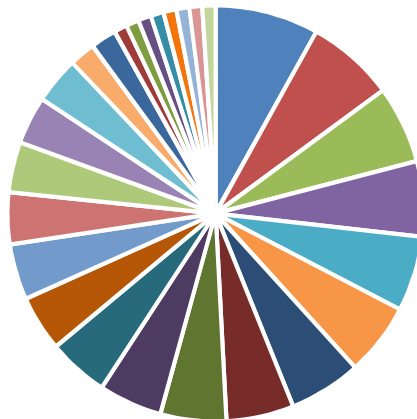
Classification 2



Classification 3



Same class



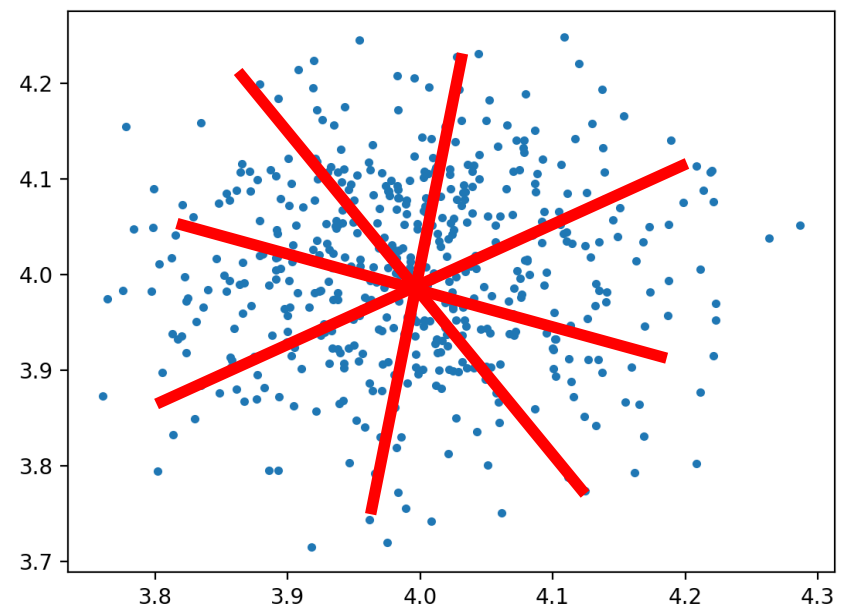
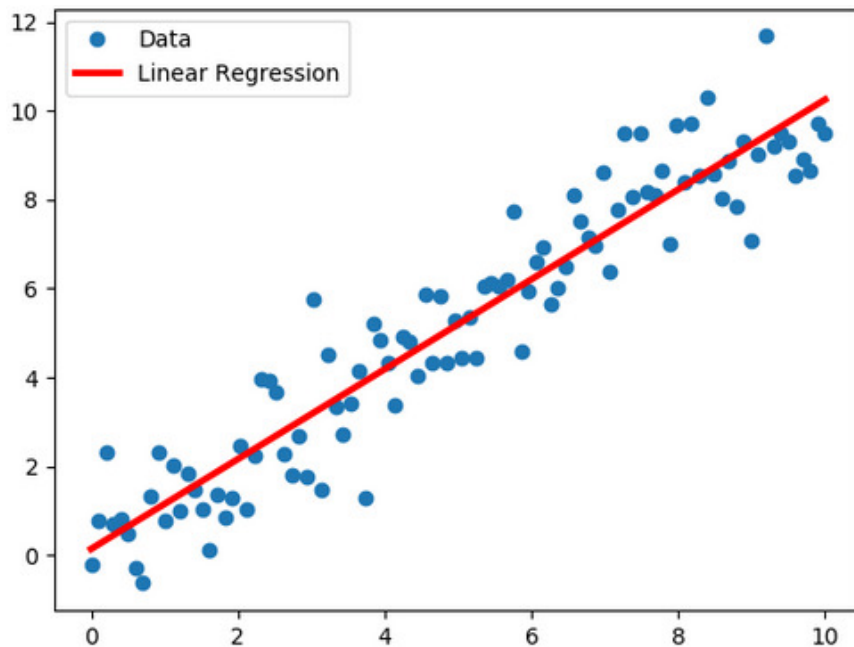
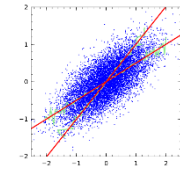
$$pV_{correct} + (1 - p)V_{incorrect}$$



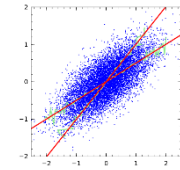
C.O.S. Sorzano, ... On bias, variance, overfitting, gold standard and consensus in Single Particle Analysis by Cryo-electron microscopy. Acta Crystallographica Section D, D78: 410-423 (2022)



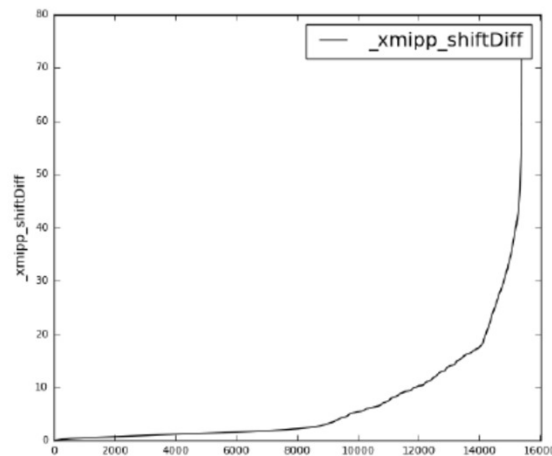
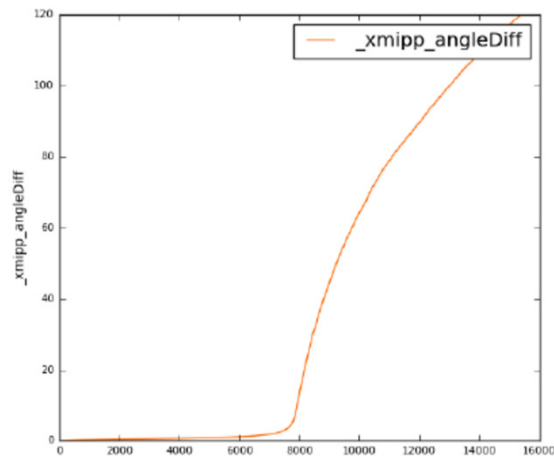
Angular assignment



Angular assignment



$$pV_{correct} + (1 - p)V_{incorrect}$$



→ Slightly incorrect
→ **Totally incorrect**

Xmipp Highres vs Relion autorefine: 10-50%
Relion autorefine vs Relion autorefine: 12-38%

Detection:

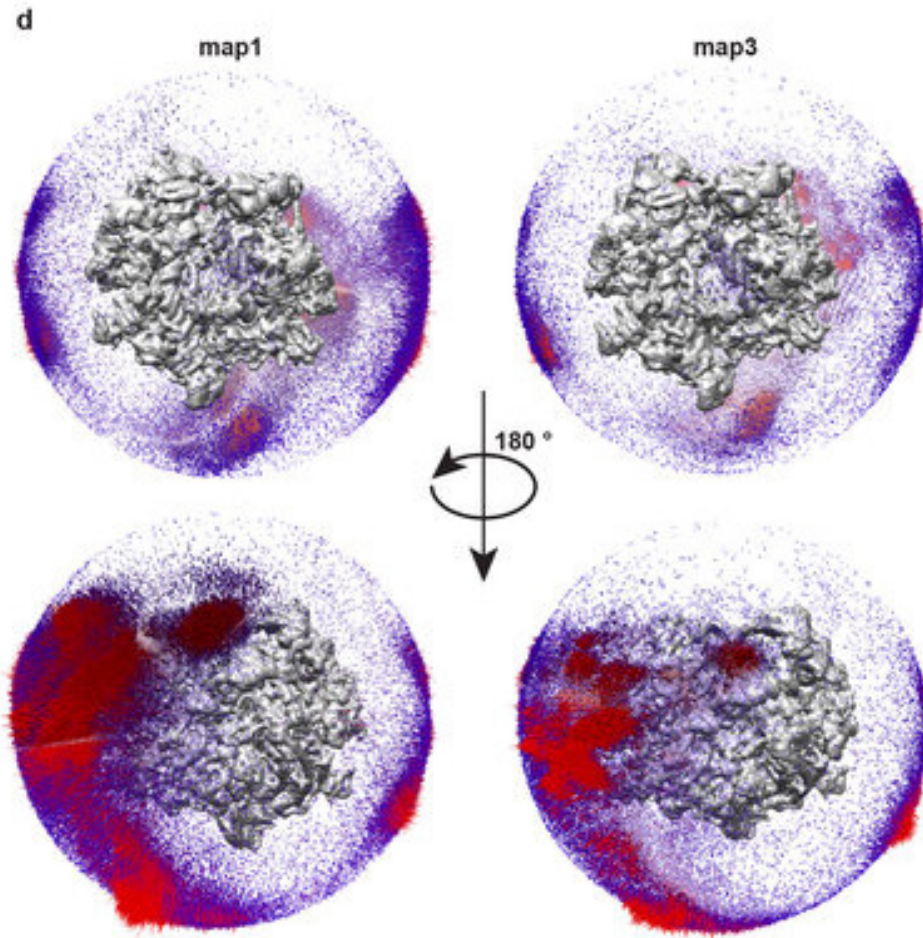
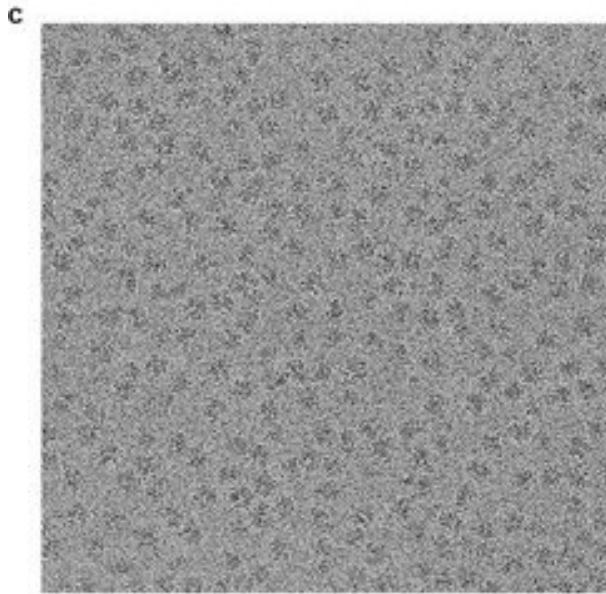
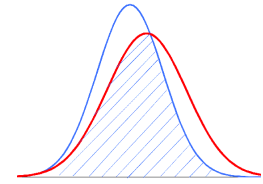
- **Consensus:** angular comparison (scipion)



C.O.S. Sorzano, ... On bias, variance, overfitting, gold standard and consensus in Single Particle Analysis by Cryo-electron microscopy. Acta Crystallographica Section D, D78: 410-423 (2022)



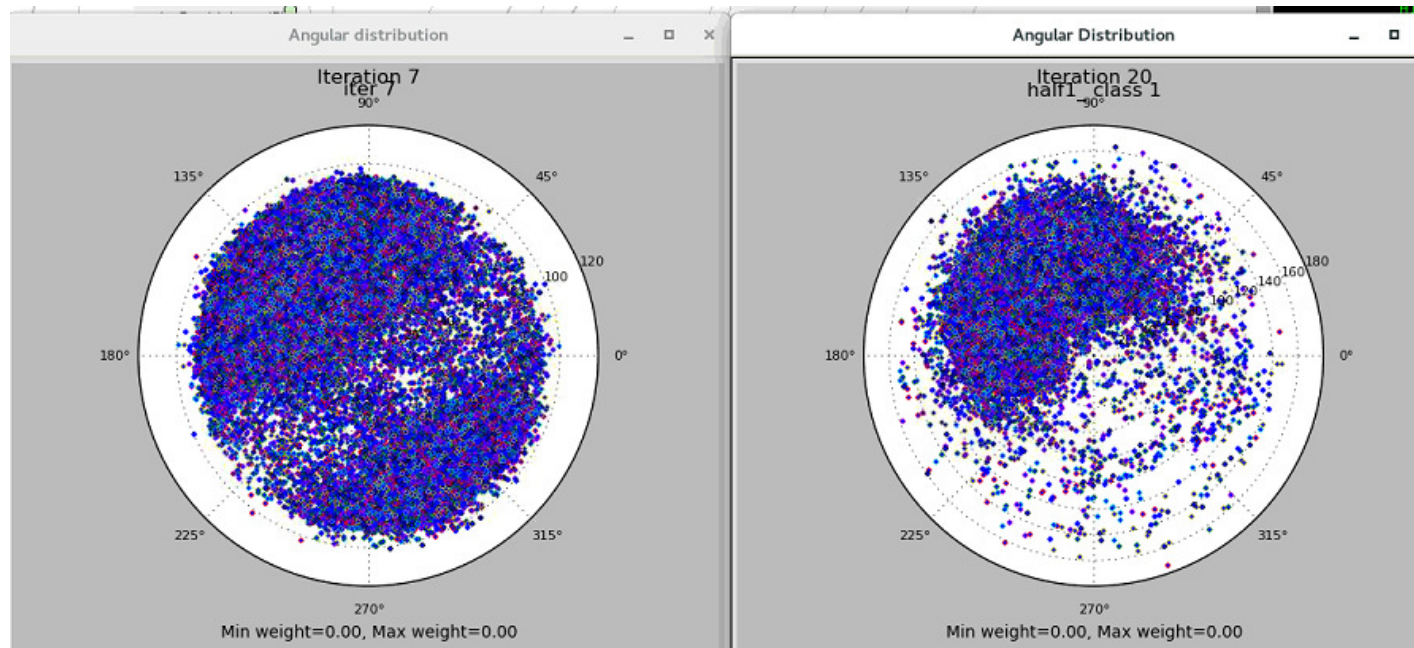
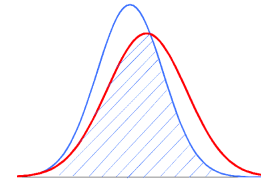
Attraction problem



C.O.S. Sorzano, D. Semchonok, S.C. Lin, J.L. Vilas, A. Jimenez-Moreno, M. Gragera, D. Maluenda, M. Martinez, E. Ramirez-Aportela, R. Melero, A. Cuervo, P. Conesa, L. del Caño, Y.C. Fonseca, R. Sánchez-García, D. Strelak, E. Fernández-Giménez, F. de Isidro, P. Kastitis, R. Marabini, B. Bruce, J.M. Carazo. [Algorithmic robustness to preferred orientations in Single Particle Analysis by CryoEM](#). J. Structural Biology 213: 107695 (2021)



Attraction problem



$$pV_{correct} + (1 - p)V_{incorrect}$$

C.O.S. Sorzano, D. Semchonok, S.C. Lin, J.L. Vilas, A. Jimenez-Moreno, M. Gragera, D. Maluenda, M. Martinez, E. Ramirez-Aportela, R. Melero, A. Cuervo, P. Conesa, L. del Caño, Y.C. Fonseca, R. Sánchez-García, D. Strelak, E. Fernández-Giménez, F. de Isidro, P. Kastitis, R. Marabini, B. Bruce, J.M. Carazo. [Algorithmic robustness to preferred orientations in Single Particle Analysis by CryoEM](#). J. Structural Biology 213: 107695 (2021)



Conclusions 1

$$pV_{correct} + (1 - p)V_{incorrect}$$

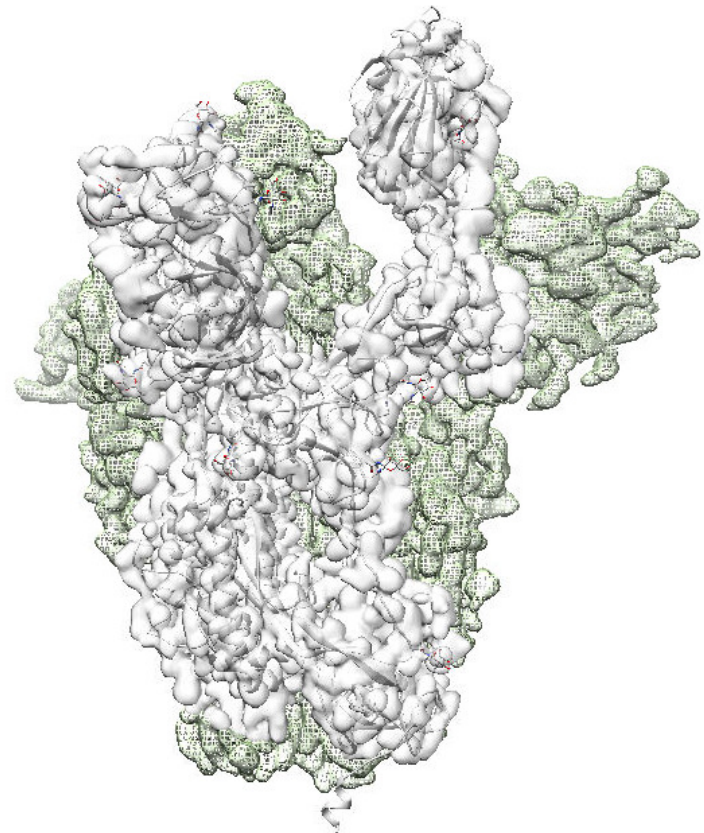
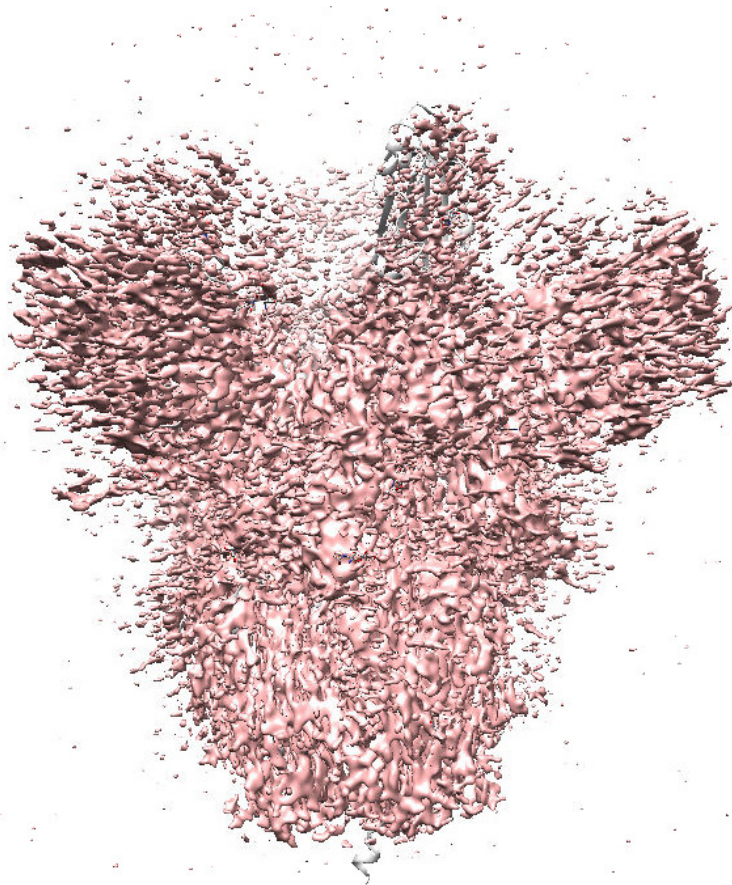
- 3D Reconstruction is all about parameter estimation.
- How do you assure that you got them right?
 - Consistency of alternatives estimates
- “Putting everything in and the algorithm will know” is suboptimal

Full talk at <https://www.youtube.com/watch?v=B30hA2DzpPg>



Current situation

EMDB 22301



Server

<https://biocomp.cnb.csic.es/EMValidationService/>

 VRS Actions ▾ Help About us ▾

UUID

Job ID

 Check

(cryo-EM) Validation Report Service

The number of maps solved by Cryo-Electron Microscopy is quickly growing in recent years. With this rapid growth, it is key to guarantee their quality.

VRS provides a map validation grading system that allows to assess consistency with the different provided elements: 2D classes, particles, angles, coordinates, defoci, and micrographs, among others. The level of validation of the map will be defined as the highest consecutive number up to which there is information available: $0 \Rightarrow 5$ [A,W,O]

Don't worry, we will guide you through the different steps. You can provide as much or less data as you have available, or wish.

 Submit a job

 Check job results



C.O.S. Sorzano, ... Image processing tools for the validation of CryoEM maps. Faraday Discussions (in press)



Validation scheme

- 1.0 Level 0: Map
- 1.1 Level 1: ... +Half maps
- 1.2 Level 2: ... +2D classes
- 1.3 Level 3: ... +Particles
- 1.4 Level 4: ... +Angular assignment
- 1.5 Level 5: ... +Micrographs and Coordinates
- 1.6 Level A: ... +Atomic model
- 1.7 Qualifier W: ... +Workflow
- 1.8 Qualifier O: ... +Other techniques

Validation report of Level(s)
0, 1, 2, 3, 4, 5, A, W, O

I²PC Validation server

February 25, 2022
4:07pm



C.O.S. Sorzano, ... Image processing tools for the validation of CryoEM maps. Faraday Discussions (in press)



Level 0: Map

0.a Center analysis. Centering of the mass and extra space available to correct for the Contrast Transfer Function (CTF). There should be at least 30-40 Å on each side for a proper correction.

0.b Mask analysis. At the threshold value specified by the user, most of the mass should be collected in a single connected component.

0.c Background analysis. If we analyze the gray values outside the mask, they should not have too negative values (e.g., values below five times the standard deviation of the background noise).

0.d Bfactor analysis. The B-factor line¹⁰ fitted between 15 Å, and the resolution reported should have a slope that is between 0 and 300 Å².

0.e DeepRes¹¹: This method is based on a deep learning algorithm that assesses the similarity of the texture features present in the map to the texture features observed in atomic structures.

0.f LocBfactor¹²: This method estimates a local resolution B-factor by decomposing the input map into a local magnitude and phase term using the spiral transform.

0.g LocOccupancy¹²: This method estimates the occupancy of a voxel by the macromolecule.

0.h DeepHand¹³: This method determines for maps whose resolution is higher than 5 Å whether the map has the right hand or, on the contrary, they are the mirrored versions of the correct map.

0.a Mass analysis	Sec. 2.1	OK
0.b Mask analysis	Sec. 2.2	OK
0.c Background analysis	Sec. 2.3	2 warnings
0.d B-factor analysis	Sec. 2.4	OK
0.e DeepRes	Sec. 2.5	1 warnings
0.f LocBfactor	Sec. 2.6	OK
0.g LocOccupancy	Sec. 2.7	OK
0.h DeepHand	Sec. 2.8	OK

Section 2.3 (0.c Background analysis)

1. The null hypothesis that the background mean is 0 has been rejected because the p-value of the comparison is smaller than 0.001
2. There is a significant proportion of outlier values in the background (cdf5 ratio=2031.06)

Section 2.5 (0.e DeepRes)

1. The reported resolution, 2.60 Å, is particularly with respect to the local resolution distribution. It occupies the 0.00 percentile



C.O.S. Sorzano, ... Image processing tools for the validation of CryoEM maps. Faraday Discussions (in press)



Level 1: ...+ Half maps

1.a Global resolution¹⁴: The Fourier Shell Correlation (FSC) between the two half maps is the most standard method to determine the global resolution of a map. However, other measures exist, such as the Spectral Signal-to-Noise Ratio and the Differential Phase Residual. There is a long debate about the correct thresholds for these measures. Probably, the clearest threshold is the one of the SSNR (SSNR=1). For the DPR, we have chosen 103.9°¹⁴ and for the FSC, the standard 0.143.

1.b Permutation test FSC¹⁵: This method calculates a global resolution by formulating a hypothesis test in which the distribution of the FSC of noise is calculated from the two maps.

1.c BlocRes¹⁶: This method computes a local Fourier Shell Correlation (FSC) between the two half maps.

1.d Resmap¹⁷: This method is based on a test hypothesis testing the superiority of signal over noise at different frequencies.

1.e MonoRes¹⁸: This method evaluates the local energy of a point to the distribution of energy in the noise. This comparison is performed at multiple frequencies, and for each one, the monogenic transformation separates the amplitude and phase of the input map.

1.f MonoDir¹⁹: This method extends the concept of local resolution to local and directional resolution by changing the shape of the filter applied to the input map. The directional analysis can reveal image alignment problems.

1.g FSO: This method calculates the anisotropy of the energy distribution in Fourier shells. It is an indirect measure of anisotropy of the angular distribution or the presence of heterogeneity.

1.h FSC Directional²⁰: This method analyzes the FSC in different directions and evaluates its homogeneity through the sphericity of FSC surface.

1.a Global resolution	Sec. 4.1	OK
1.b FSC permutation	Sec. 4.2	OK
1.c Blocres	Sec. 4.3	OK
1.d Resmap	Sec. 4.4	1 warnings
1.e MonoRes	Sec. 4.5	OK
1.f MonoDir	Sec. 4.6	1 warnings
1.g FSO	Sec. 4.7	OK
1.h FSC3D	Sec. 6.1	OK



C.O.S. Sorzano, ... Image processing tools for the validation of CryoEM maps. Faraday Discussions (in press)

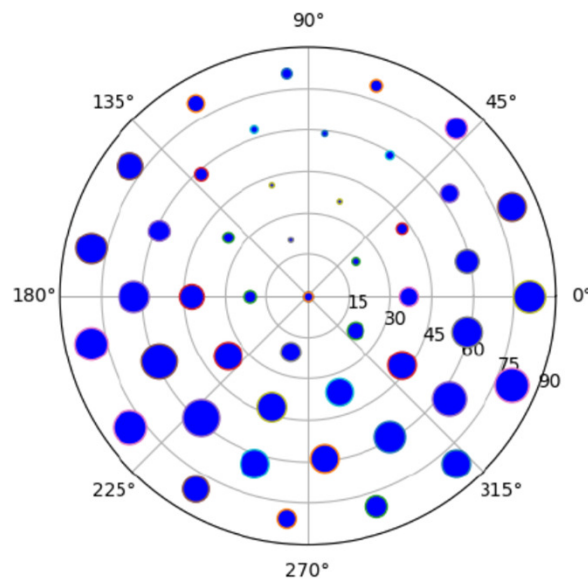
Level 1: ...+ Half maps

Section 4.4 (1.d Resmap)

1. The reported resolution, 2.60 Å, is particularly with respect to the local resolution distribution. It occupies the 0.00 percentile

Section 4.6 (1.f MonoDir)

1. The distribution of best resolution is not uniform in all directions. The associated p-value is 0.000000.



C.O.S. Sorzano, ... Image processing tools for the validation of CryoEM maps. Faraday Discussions (in press)

4.4 Level 1.d Local resolution with Resmap

Explanation:

This method [Kucukelbir et al., 2014] is based on a test hypothesis testing of the superiority of signal over noise at different frequencies.

Results:

Fig. 26 shows the histogram of the local resolution according to Resmap. Some representative percentiles are:

Percentile	Resolution(Å)
2.5%	3.13
25%	3.45
50%	3.52
75%	3.55
97.5%	3.58

The reported resolution, 2.60 Å, is at the percentile 0. Fig. 27 shows some representative views of the local resolution.

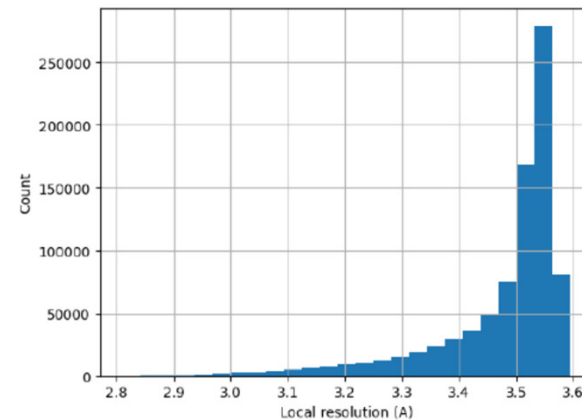


Figure 26: Histogram of the local resolution according to Resmap.

Level 2: ...+2D classes

2.a Reprojection consistency: The 2D classes can be aligned against the reconstructed map, then the correlation between reprojections of the map and the 2D classes can be analyzed. Also, analyzing the residuals (2D class minus the corresponding reprojection) can reveal systematic differences.

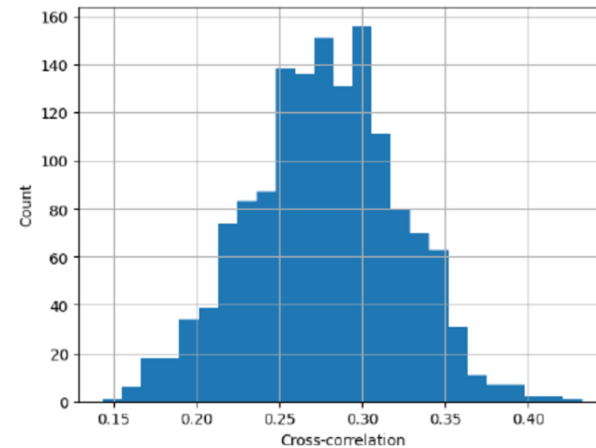
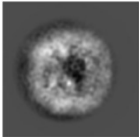
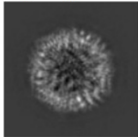
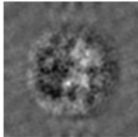
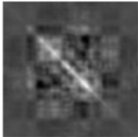
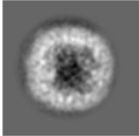
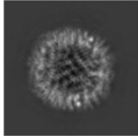
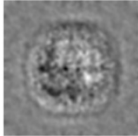
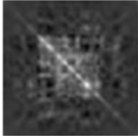
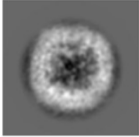
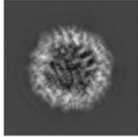
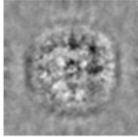
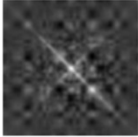


Figure 39: Histogram of the correlation coefficient between the 2D classes provided by the user and the corresponding reprojections.

2.a Reprojection consistency

Sec. 6.1

OK

2D Class	Reprojection	Residual	Covariance	Correlation
				0.72
				0.78
				0.79



Level 3: ...+Particles

- 3.a Outlier detection: The set of particles is classified into the input set of 2D classes of Level 2. The number of particles considered to be outliers in those classes is reported. A particle is an outlier if its Mahalanobis distance to the centroid of the class is larger than 3^{21} .
- 3.b 2D Classification internal consistency: The input particles are classified in 2D clusters. The quality of the 2D clusters is assessed through Fourier Ring Correlation.
- 3.c 2D Classification external consistency: We measure the overlap between the subspace spanned by the classes in Level 2 and the classes of Level 3.

3.a Outlier detection	Sec. 9.1	OK
3.b 2D Classification internal consistency	Sec. 8.2	Cannot be automated
3.c 2D Classification external consistency	Sec. 8.3	OK

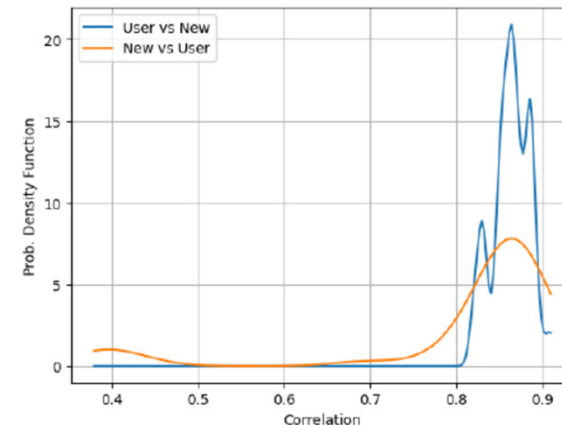
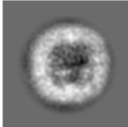
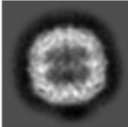
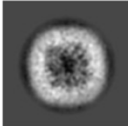
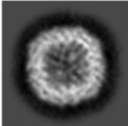


Figure 46: Probability density function of the correlation of the user classes compared to the newly computed classes and vice versa.

The following table shows for each class in the User set which is the best match in the New set and its correlation coefficient.

User class	New class	Correlation
		0.856
		0.887



Level 4: ...+ Angular assignment

4.a Analysis of the distribution of the similarity between the input particles and the reprojection from the same angular orientation by different scores.

4.b Alignability smoothness²²: This algorithm analyzes the smoothness of the correlation function over the projection sphere and the stability of its maximum.

4.c Alignability precision and accuracy: The precision²³ analyzes the orientation distribution of the best matching reprojections from the reference volume. If the high values are clustered around the same orientation, the precision is close to 1. Otherwise, it is closer to -1. Below 0.5, the best directions tend to be scattered. The alignability accuracy²⁴ compares the final angular assignment with the result of a new angular assignment. The similarity between both is again encoded between -1 and 1.

4.d Angular error distribution between the provided angles and an independent angular assignment performed with state-of-the-art algorithms.

4.e 3D Classification of the input particles without angular refinement.

4.f Detection of overfitting²⁵: This method compares the resolution achieved by subsets of images of increasing size and by subsets of noise images of the same size.

4.g Angular distribution efficiency²⁶: This method evaluates the ability of the angular distribution to fill the Fourier space.

4.h Sampling compensation factor²⁷: This method is another way of measuring the ability of the angular distribution to fill the Fourier space.

4.i Analysis of the stability of the defocus parameters. For this purpose, defocus, B-factor, astigmatism, and phase shift can be estimated from the given particles, and these refined parameters' deviations are reported. Ideally, the differences in defoci cannot be larger than the ice thickness. The same can be done with local magnification offsets (which should be around 0) and the B-factor.

4.a Similarity criteria	Sec. 9.1	Cannot be automated
4.b Alignability smoothness	Sec. 9.2	1 warnings
4.c Alignability precision and accuracy	Sec. 9.3	OK
4.d1 Relion alignment	Sec. 9.4	OK
4.d2 CryoSparc alignment	Sec. 9.5	1 warnings
4.d3 Relion/CryoSparc alignments	Sec. 9.6	1 warnings
4.e Classification without alignment	Sec. 9.8	OK
4.f Overfitting detection	Sec. 9.8	OK
4.g Angular distribution efficiency	Sec. 9.9	OK
4.h SCF	Sec. 9.10	OK
4.i CTF stability	Sec. 9.11	1 warnings

Section 9.2 (4.b Alignability smoothness)

1. The percentage of images whose angular assignment is significantly away from the smoothed maximum is too high, 50.2%

Section 9.5 (4.d2 CryoSparc alignment)

1. The percentage of images with uncertain shift is larger than 20%

Section 9.6 (4.d3 Relion/CryoSparc alignments)

1. The percentage of images with uncertain shift is larger than 20%

Section 9.11 (4.i CTF stability)

1. The 95% confidence interval of scale factor is not centered.



Level 4: ...+ Angular assignment

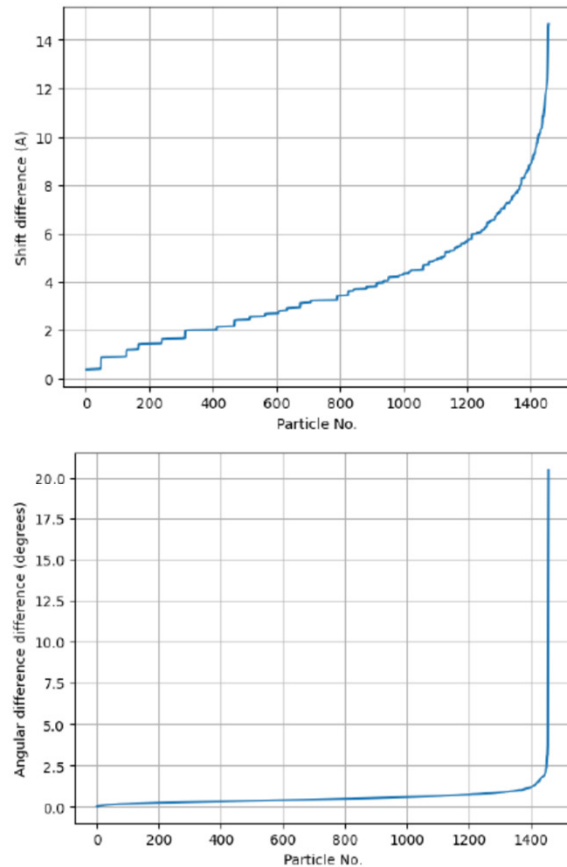


Figure 53: Top: Shift difference between the alignment given by the user and the one calculated by CryoSparc. Bottom: Angular difference. The X-axis represents all particles sorted by their difference.

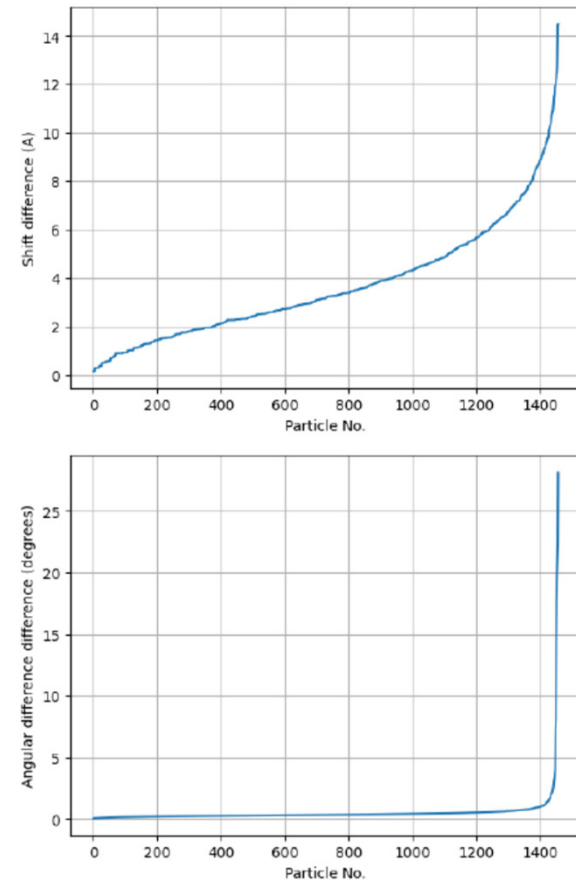


Figure 54: Top: Shift difference between the alignment given by Relion and the one calculated by CryoSparc. Bottom: Angular difference. The X-axis represents all particles sorted by their difference.

Level 5: ...+ Coordinates

5.a Micrograph cleaner²⁸: This method assigns a score between 0 and 1, reflecting the probability that the coordinate is outside a region with aggregations, ice crystals, carbon edges, etc.

5.a Micrograph cleaner	Sec. 11.1	OK
------------------------	-----------	----

11 Level 5 analysis

11.1 Level 5.a Micrograph cleaner

Explanation:

This method assigns a score between 0 (bad coordinate) and 1 (good coordinate) reflecting the probability that the coordinate is outside a region with aggregations, ice crystals, carbon edges, etc. [Sanchez-Garcia et al., 2020]

Results:

0 coordinates out of 1457 (0.0 %) were scored below 0.9 by Micrograph-Cleaner.

Automatic criteria: The validation is OK if less than 20% of the coordinates are suspected to lie in aggregations, contaminations, ice crystals, etc.

STATUS: OK



Level A: ...+ Atomic model

A.a Map-Q²⁹: This method computes the local correlation between the map and each one of its atoms assumed to have a Gaussian shape.

A.b FSC-Q³⁰: This method compares the local FSC between the map and the atomic model to the local FSC of the two half maps.

A.c Model ambiguity by molecular dynamics³¹: This method estimates the ambiguity of the atomic model in each region of the CryoEM map due to the different local resolutions or local heterogeneity.

A.d Guinier plot of model and map³²: This method compares the falloff in Fourier space between the map and atomic model.

A.e Phenix CryoEM validation tools³³: Phenix provides several tools to assess the agreement between the experimental map and its atomic model. Two large clusters of these measurements are: 1) different ways of measuring the cross-correlation between the map and model, and 2) different ways of measuring the resolution between the map and model.

A.f EMRinger³⁴: This algorithm compares the side chains of the atomic model to the CryoEM map.

A.g DAQ³⁵: This algorithm uses deep learning that can estimate the residue-wise local quality for protein models from cryo-Electron Microscopy (EM) maps. The method calculates the likelihood that a given density feature corresponds to an aminoacid, atom, and secondary structure. These likelihoods are combined into a score that ranges from -1 (bad quality) to 1 (good quality).

A.a MapQ	Sec. 13.1	OK
A.b FSC-Q	Sec. 13.2	OK
A.c Multimodel	Sec. 13.3	OK
A.d Map-Model Guinier	Sec. 13.4	OK
A.e Phenix validation	Sec. 13.5	1 warnings
A.f EMRinger	Sec. 13.6	1 warnings
A.g DAQ	Sec. 13.7	1 warnings

Section 13.5 (A.e Phenix validation)

1. The resolution reported by the user, 2.6 Å, is significantly smaller than the resolution estimated between map and model (FSC=0.5), 4.4 Å

Section 13.6 (A.f EMRinger)

1. The EMRinger score is smaller than 1, it is 0.892.

Section 13.7 (A.g DAQ)

1. The average DAQ is smaller than 0.5.



Level A: ...+ Atomic model

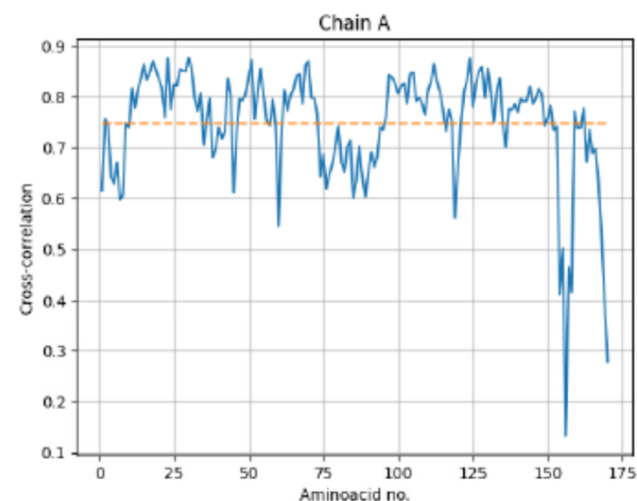
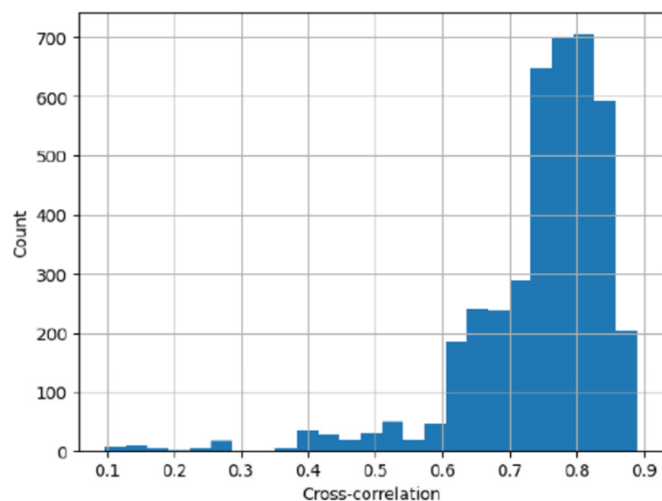


Figure 73: Histogram of the cross-correlation between the map and model evaluated for all residues.

Resolutions estimated from the model:

Resolution (Å)	Masked	Unmasked
d99	1.7	1.6
d_model	3.8	3.8
d_model (B-factor=0)	4.3	4.3
FSC_model=0	3.0	2.9
FSC_model=0.143	3.4	3.4
FSC_model=0.5	4.4	4.5

Overall isotropic B factor:

B factor	Masked	Unmasked
Overall B-iso	85.0	85.0

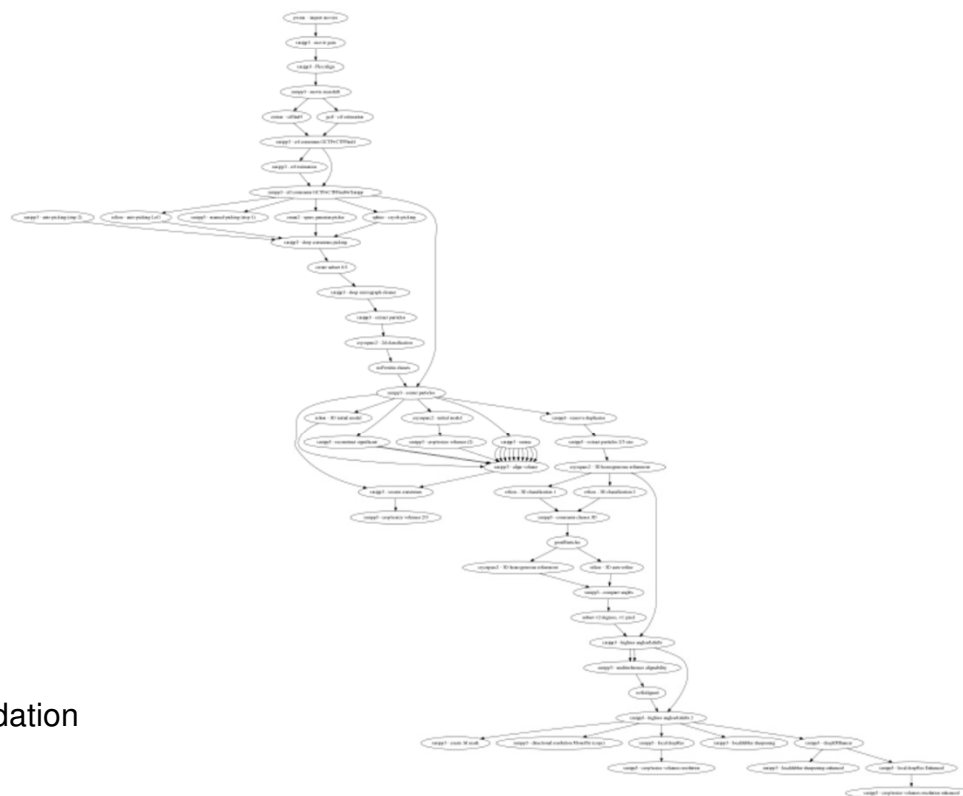


Level W: ...+ Workflow

14 Workflow

Workflow file: <http://nolan.cnb.csic.es/cryoemworkflowviewer/workflow/637ca2bbcd57e45e88f6fabb7f6b1095a3ca0de6>
SHA256 hash: 5d8c5ff8948f4ac986f5d43f819515e25668bfdcd954b8fb8c41d15cdf00fda2

Fig. 79 shows the image processing workflow followed in Scipion to achieve these results.



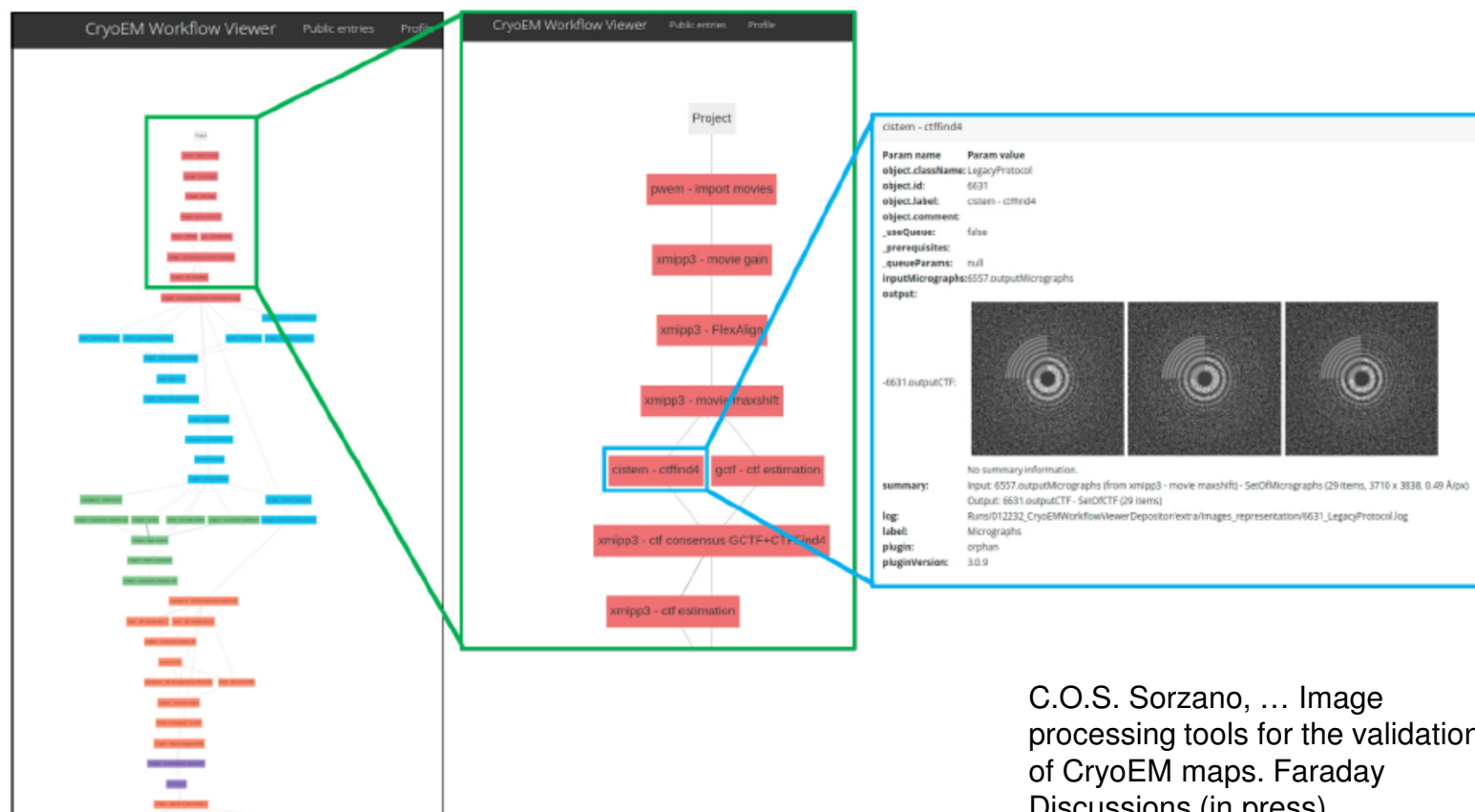
C.O.S. Sorzano, ... Image processing tools for the validation of CryoEM maps. Faraday Discussions (in press)

Level W: ...+ Workflow

14 Workflow

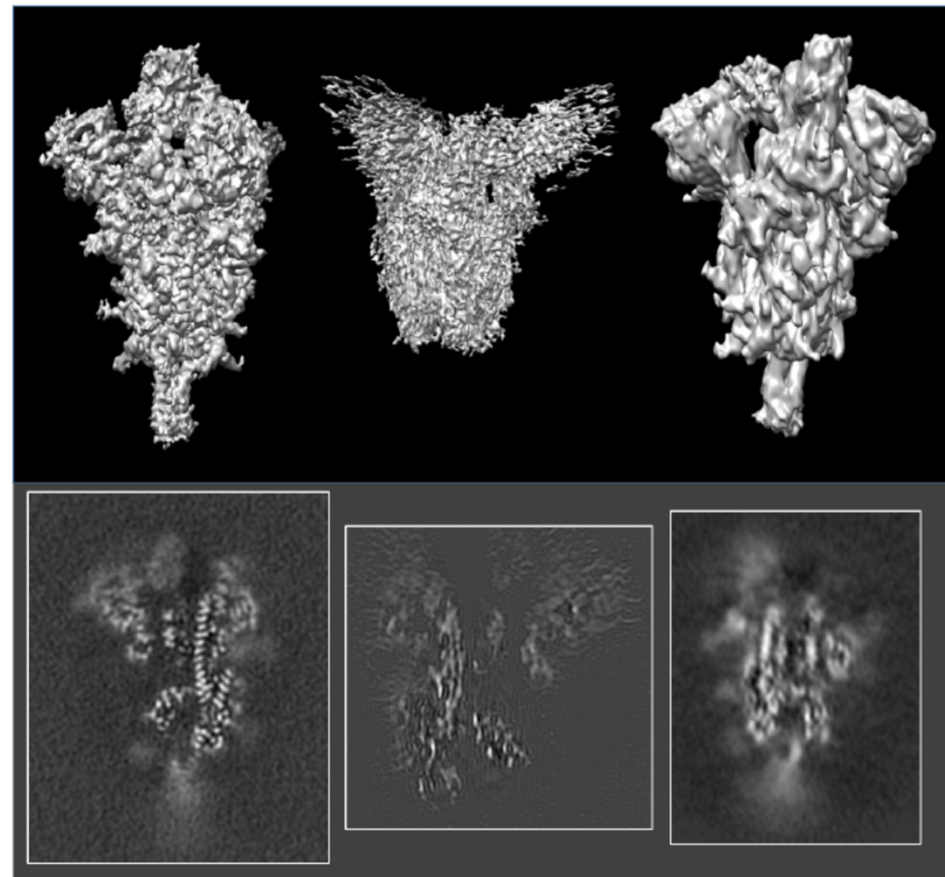
Workflow file: <http://nolan.cnb.csic.es/cryoemworkflowviewer/workflow/637ca2bbcd57e45e88f6fabb7f6b1095a3ca0de6>
SHA256 hash: 5d8c5ff8948f4ac986f5d43f819515e25668bfdcd954b8fb8c41d15cdf00fda2

Fig. 79 shows the image processing workflow followed in Scipion to achieve these results.



C.O.S. Sorzano, ... Image processing tools for the validation of CryoEM maps. Faraday Discussions (in press)

Example



C.O.S. Sorzano, ... Image processing tools for the validation of CryoEM maps. Faraday Discussions (in press)

Fig. 6 Isosurface and central slice of EMDB 11337 (left), EMDB 22301 (middle), and EMDB 22838 (right), all of them are SARS-CoV2 spikes.



14/20
3.3A

7/13
3.7A

6/13
3.8A



Report

SM1_report_ScipionTutorial.pdf - Adobe Acrobat Reader DC (64-bit)

Archivo Edición Ver Firmar Ventana Ayuda

Inicio Herramientas SM1_report_Scipion... x

1 (1 de 113)

Marcadores

- Input data
- > Level 0 analysis
- Half maps
- > Level 1 analysis
- 2D Classes
- > Level 2 analysis
- Particles
- > Level 3 analysis
- > Level 4 analysis
- Micrographs
- > Level 5 analysis
- Atomic model
- > Level A analysis
- Workflow
- > Other experimental techniques

Validation report of Level(s)
0, 1, 2, 3, 4, 5, A, W, O

I²PC Validation server
February 25, 2022
4:07pm

1



Trends

← → ↻ ⚠ Not secure | 3demmethods.i2pc.es/index.php/Main_Page

main page discussion view source history

3DEM Methods

navigation

- Main Page
- Recent changes
- All articles

search

Search 3DEM-Methc

Go Search

tools

- What links here
- Related changes
- Special pages
- Printable version
- Permanent link
- Page information

Main Page

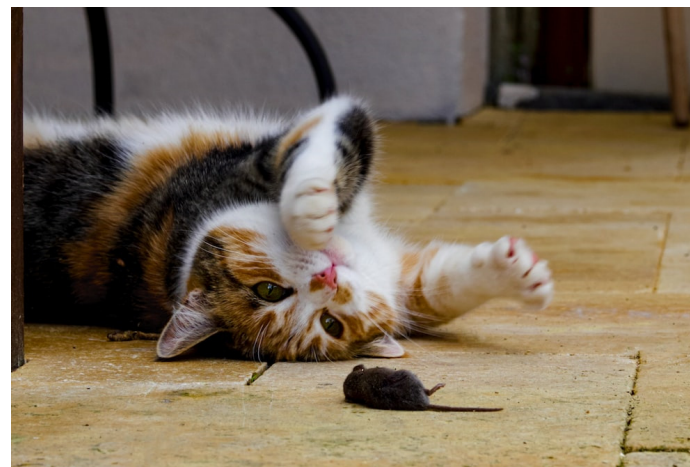
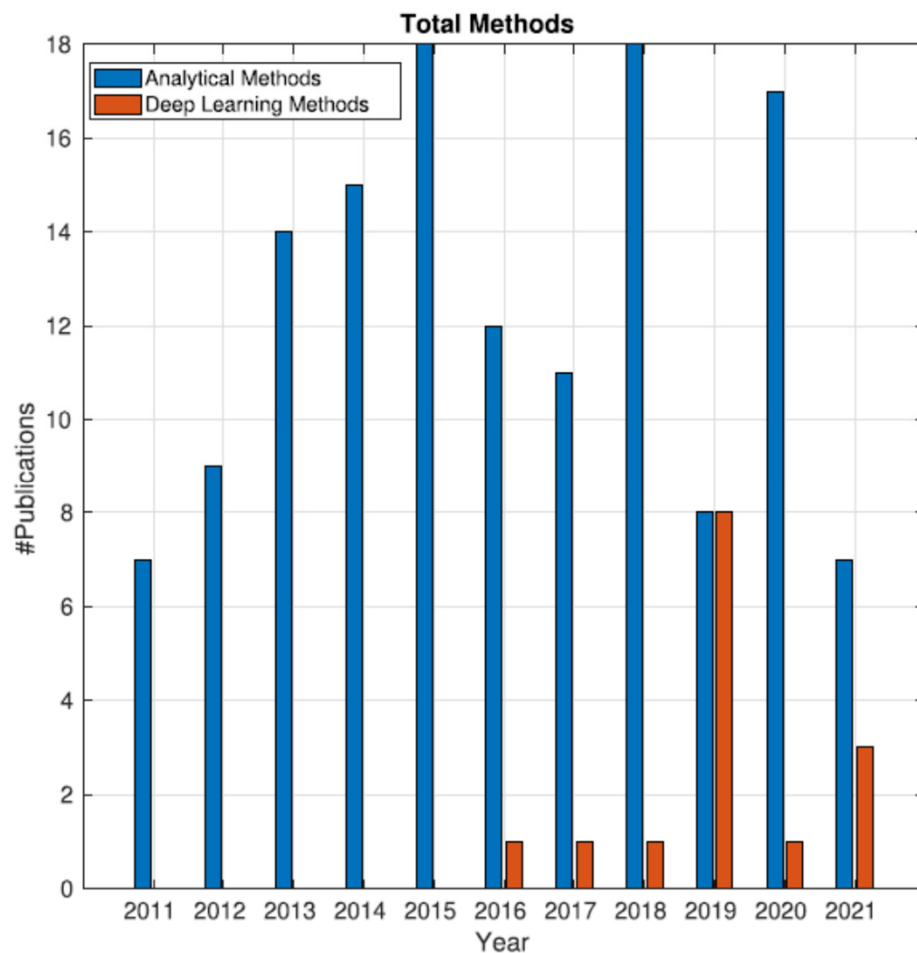
Contents [hide]

- 1 Presentation
- 2 How to subscribe to this page
- 3 Electron microscopy images
 - 3.1 Online courses and Learning material
 - 3.2 Image formation
 - 3.3 Collection geometry
 - 3.4 Sample preparation
 - 3.5 Automated data collection
- 4 Single particles
 - 4.1 Automatic particle picking
 - 4.2 2D Preprocessing
 - 4.3 2D Alignment
 - 4.4 2D Classification and clustering
 - 4.5 3D Alignment
 - 4.6 3D Reconstruction
 - 4.7 3D Heterogeneity
 - 4.8 Validation

<http://3demmethods.i2pc.es/>



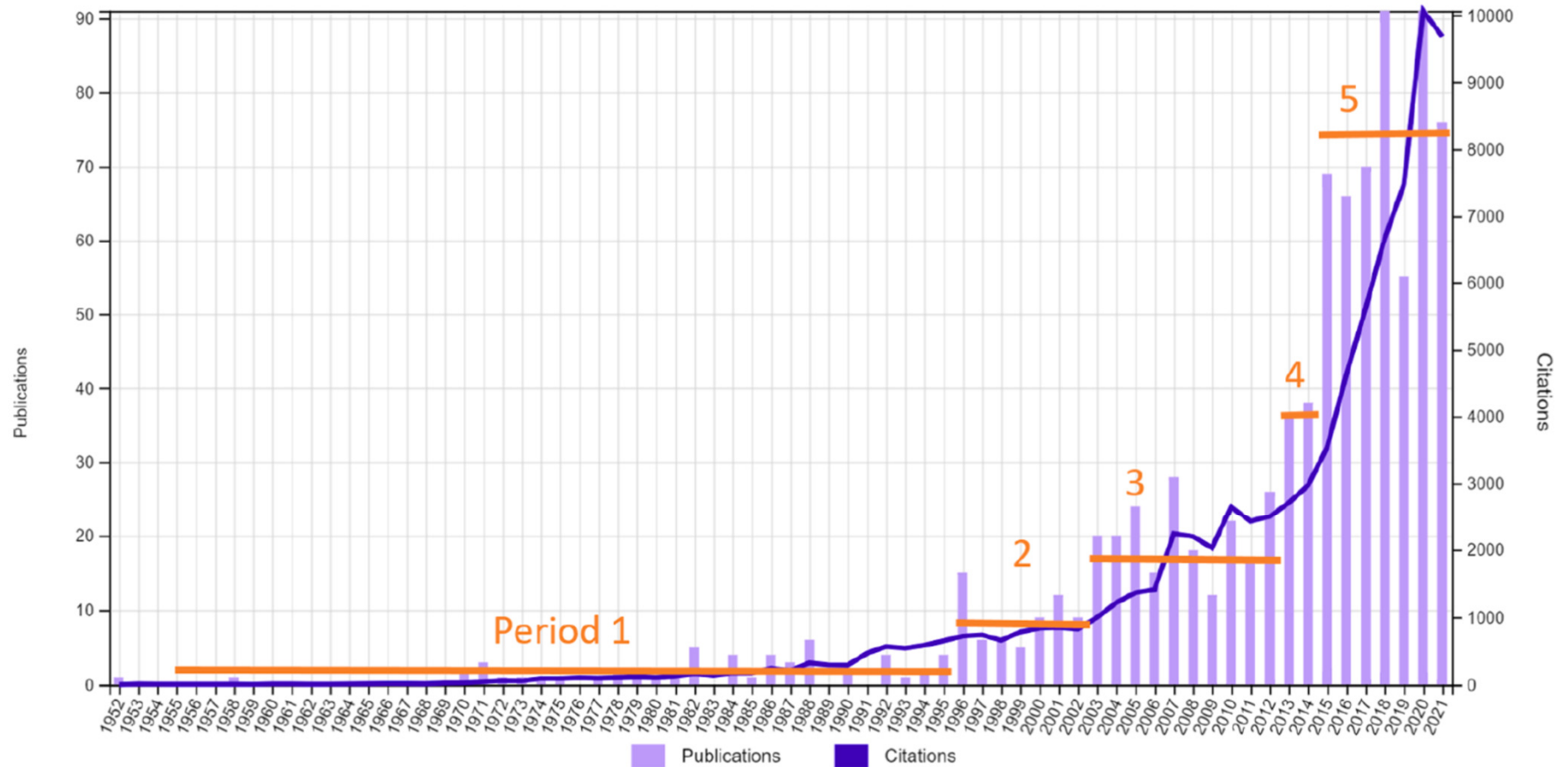
Trends



J.L. Vilas, J.M. Carazo, C.O.S. Sorzano. Emerging themes in cryoEM-SPA Image processing. Chemical reviews (in press)



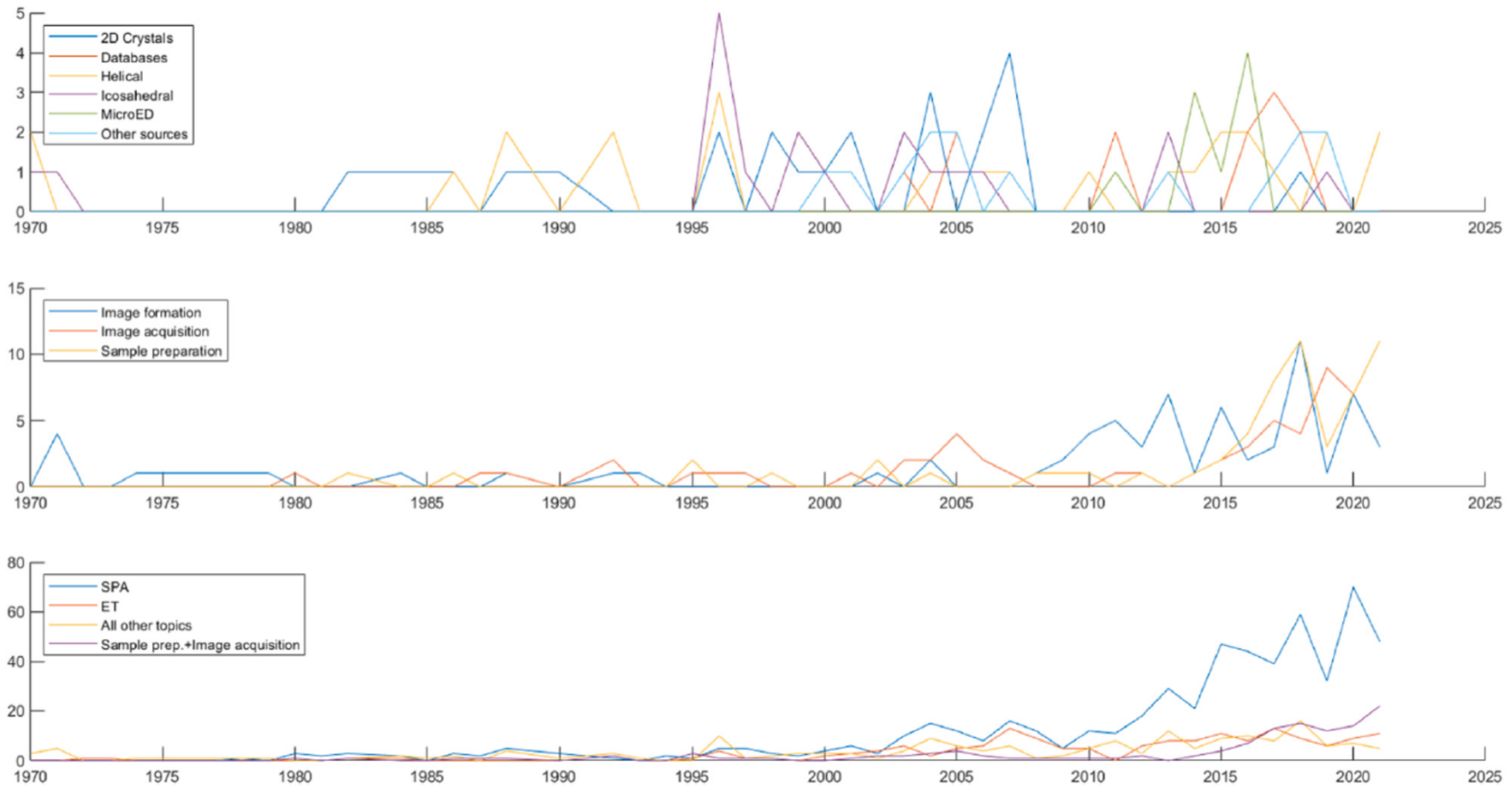
Trends



C.O.S. Sorzano, J.M. Carazo. Cryo-Electron Microscopy: the field of 1,000+ methods. J. Structural Biology, 214: 107861 (2022)



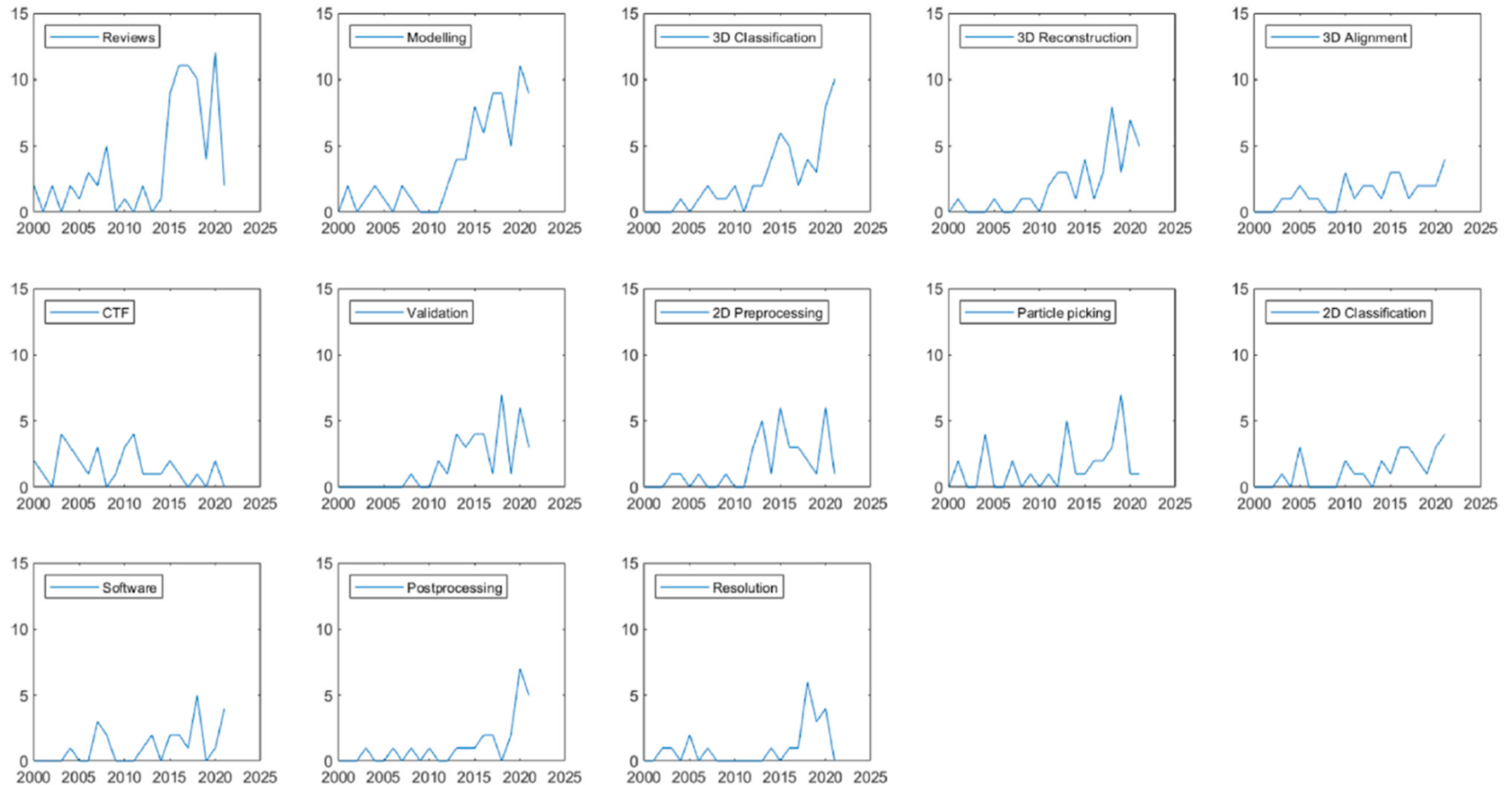
Trends



C.O.S. Sorzano, J.M. Carazo. Cryo-Electron Microscopy: the field of 1,000+ methods. J. Structural Biology, 214: 107861 (2022)



Trends



C.O.S. Sorzano, J.M. Carazo. Cryo-Electron Microscopy: the field of 1,000+ methods. J. Structural Biology, 214: 107861 (2022)



Conclusions

- SPA has to operate in a **very noisy** environment, which imply small errors (**noise**) and big errors (**bias**).
- The only way to detect bias is by **comparing** the estimates of several algorithms
- $V_{reconstructed} = pV_{correct} + (1 - p)V_{incorrect}$
- Then, **we should validate** our result with the many tools available.
- Part of the information needed for **others to validate** is the **raw data** and the disclosure of our image processing **pipeline/decisions**



Thanks



+Collaborators

