Why is a single execution of a single algorithm not enough in CryoEM?

C.O.S. Sorzano[1], A. Jiménez-Moreno[1], E. Ramírez-Aportela[1], J. Krieger[1], J. Filipovic[1], P. Conesa[1], Y.C. Fonseca[1], J. Jiménez-de la Morena[1], D. Strelak[12], E. Fernández-Giménez, F. de Isidro[1], D. Herreros[1], D. Marchán[1], J.L. Vilas[1], R. Marabini[3], J.M. Carazo[1]

[1.] Centro Nacional de Biotecnología (CNB-CSIC), Madrid, Spain
[2.] Masaryk University, Brno, Czech Republic
[3.] Escuela Politécnica Superior, Universidad Autónoma de Madrid, Madrid, Spain

Single-particle analysis by cryo-electron microscopy has become a significant player in the field of Structural Biology. The number of structures solved by this technique is steadily increasing, and many X-ray crystallography and Nuclear Magnetic Resonance Structural Biology groups have adopted it as another complementary technique. Additionally, many sophisticated but easy-to-use software programs allow many practitioners to analyze the data collected by the microscope in a relatively straightforward way. In this abstract, we argue that it should be standard practice to confirm all the steps in the image processing pipeline with multiple algorithms and parameter estimates. Unfortunately, this procedure is rarely observed in published structures.

In a simplified view, the image processing requires alignment of the frames of the movies collected by the microscope, determination of the microscope aberrations for each micrograph, identification of the particles of interest in the field of view, grouping them into structurally homogeneous subpopulations, estimating orientation of those particles and finding a 3D structure that is compatible with the collected data. All these steps require estimating parameters (for a complete description, see [1]). However, these parameters must be determined in an environment with extremely low Signal-to-Noise Ratios (SNRs). As a rough guide, the SNR at the level of the micrograph is between 0.1 and 0.01, while at the level of the frame it is 0.001, meaning that we have 10, 100, and 1,000 times more noise than signal. That makes this parameter detection an extremely challenging problem. To illustrate this idea, we show in Fig. 1 a one-dimensional equivalent of locating a particle in noise.

In this context, it is well understood that the parameters estimated for each particle may also be very noisy; see Fig. 2 for an example in angular assignment. It is clear from the figure that only a small fraction of the particles gets a relatively consistent angular assignment (angular difference less than 10 degrees and shift difference less than 5 pixels, the particle size is 100x100).

There are two kinds of errors when estimating any parameter: small and large. Small errors can be easily reduced by averaging multiple estimates of the same parameter. If there is no bias, the averaged parameter will be closer to the underlying ground truth. However, for large errors, even in the case of no bias, the average of few parameter estimates would still be expected to be quite far from the underlying ground truth. Consequently, this brings two interesting recommendations for image processing in Single Particles by CryoEM:

1. With a single run of a single algorithm to estimate any given parameter, it is impossible to know whether the estimate can be relied on or not. We need at least two estimates. If they agree, there is no evidence that they are incorrect. But if they disagree, at least one of the two (and we do not know which one) has to be wrong. We could try to identify the actual value by many more estimates and clustering their values, and if we do not perform more estimates, we would need to

drop both.

2. For those estimates that approximately agree, a better estimate of the underlying parameter would be given by their average, as the noise variance is reduced by the number of estimates averaged.

This principle of multiple estimates applies to parameters at all levels (frame alignment, microscope aberrations, particle identification and extraction from the micrograph, membership to a given structural class, angular assignment, volume reconstruction, etc.). Unfortunately, it is rarely seen in published structures that these parameters are estimated more than once. In that regard, it should be expected that many of the maps we produce are a mixture of two structures: one in which all the parameters have been estimated with small errors, and another one in which at least one of the parameters has been estimated with a large error.

References:

[1] C.O.S. Sorzano, et al. Structural Proteomics, 3rd ed. (2021), p. 257-289

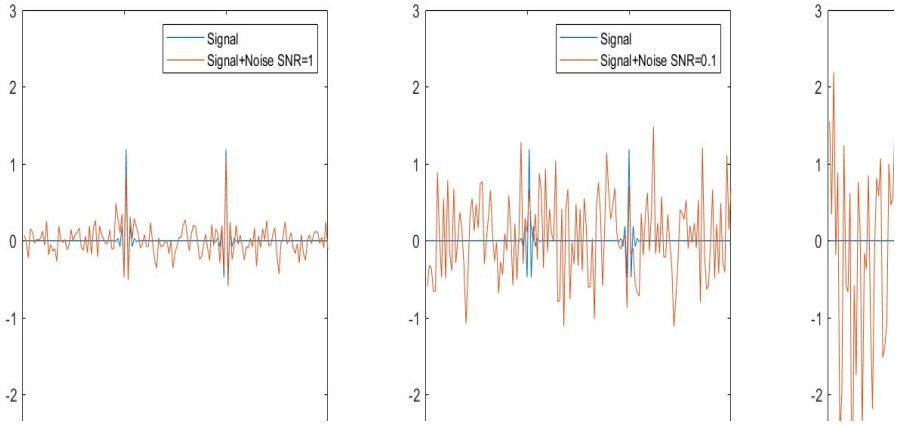

Figure 1. Example of a parameter detection in noise whose SNRs are 1, 0.1, and 0.01, respectively. The parameter here would be locating the center point between the two peaks (the equivalent of particle picking).

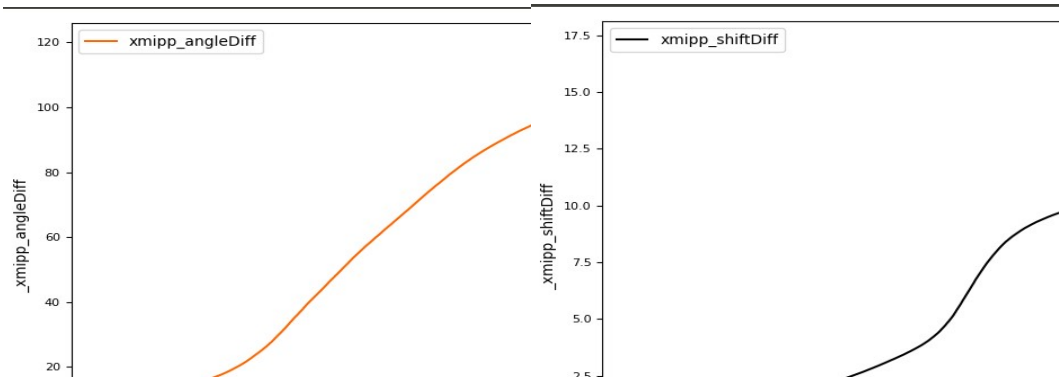

Figure 2. Angular and shift difference of two executions of CryoSparc non-homogeneous refinement with exactly the same inputs (121k images) and execution parameters.