ENRICH: a fast method to improve the quality of flexible macromolecular reconstructions

M. Kazemi^{1*}, C. O. S. Sorzano², A. des Georges^{3,4,5}, V. Abrishami⁶, J. M. Carazo², J. Vargas^{7*}

¹*Dep. of Biochemistry and Molecular Biology and Bio21 Molecular Science and Biotechnology Institute, The
University of Melbourne, Parkville, VIC 3010, Australia*

²*Biocomputing Unit, Centro Nacional de Biotecnología-CSIC,
C/ Darwin 3, 28049, Cantoblanco (Madrid), Spain*

³*Structural Biology Initiative, CUNY Advanced Science Research Center,
New York, NY 10031, USA.*

⁴*Dept. of Chemistry & Biochemistry, City College of New York,
New York, NY 10031, USA.*

⁵*Ph.D. Program in Biochemistry, The Graduate Center of the City University of New
York, New York, NY 10016, USA.*

⁶*Laboratory of Structural Biology, Helsinki Institute of Life Science HiLIFE*

⁷*Departamento de Optica, Universidad Complutense de Madrid, Avda. Computense s/n, Ciudad Universitaria
28040 Madrid, Spain, and Department of Anatomy and Cell Biology, McGill University
3640 Rue University, Montréal, QC H3A 0C7, Canada*

*Correspondence to jvargas@fis.ucm.es , mohsen.kazemi@unimelb.edu.au

27 **Summary**

28 Cryo-electron microscopy using single particle analysis requires the computational averaging of
29 thousands of projection images captured from identical macromolecules. However, macromolecules
30 usually present some degree of flexibility showing different conformations. Computational
31 approaches are then required to classify heterogeneous single particle images into homogeneous sets
32 corresponding to different structural states. Nonetheless, sometimes the attainable resolution of
33 reconstructions obtained from these smaller homogeneous sets is compromised because of reduced
34 number of particles or lack of images at certain macromolecular orientations. In these situations, the
35 current solution to improve map resolution is returning to the electron microscope and collect more
36 data. In this work, we present a fast approach to partially overcome this limitation for heterogeneous
37 data sets. Our method is based on deforming and then moving particles between different
38 conformations using an optical flow approach. Particles are then merged into a unique conformation
39 obtaining reconstructions with improved resolution, contrast and signal-to-noise ratio. We present
40 experimental results that show clear improvements in the quality of obtained 3D maps, however,
41 there are also limits to this approach, i.e., the method is restricted to small deformations and cannot
42 determine local patterns of flexibility of small elements, such as secondary structures, which we
43 discuss in the manuscript.

44 **Keywords**

45 Cryo-electron microscopy; single-particle analysis; image processing; optical flow; heterogeneous
46 particles; flexible macromolecular reconstructions

47 **1. Introduction**

48 Cryogenic electron microscopy (cryo-EM) using single particle analysis is a well-known structural
49 technique to achieve three-dimensional reconstructions of macromolecular complexes in their native
50 state. This imaging modality has recently transitioned between being considered a promising
51 structural technique to become a well-established approach, capable of providing high-resolution

52 structural maps routinely. One of the main advantages of this technique is its ability to visualize
53 macromolecule complexes that would be too large or flexible to be tackled by nuclear magnetic
54 resonance or X-ray crystallography. Biological macromolecular complexes are not rigid entities and
55 usually exist in a range of conformations that play a key role in their function. Macromolecular
56 structures at sufficient resolution and in different physiological conformations provide essential
57 insights into understanding their function and are key to efficiently design new drugs [Boland2017].
58 However, obtaining high-resolution reconstructions from highly flexible samples remains a
59 challenge. The reconstruction workflow is an inverse ill-posed problem. In cryo-EM of isolated
60 macromolecules, normally referred to as Single Particle Analysis (SPA), macromolecular projection
61 images are obtained without knowledge of the particle orientation. This lack of orientation
62 information added to the low signal-to-noise ratio that characterizes cryo-EM images impose
63 computationally demanding reconstruction procedures as marginalized maximum-likelihood based
64 approaches [Scheres2007; Scheres2012a; Lyumkis2013] or statistically robust techniques
65 [Sorzano2018]. High-resolution structure determination requires obtaining homogeneous sets of
66 particles in the same state. Flexible regions of complexes will display lower resolutions correlated
67 with the amplitude of movement. In case of macromolecular complexes exhibiting important
68 flexibility or multiple conformations, an added level of complication is a drop in the accuracy of the
69 particle orientation assignment. Thus, particles from a structurally heterogeneous data set must be
70 divided into structurally homogeneous sets or classes. This processing step is called three-
71 dimensional (3D) classification and is usually performed in combination with 3D reconstruction
72 [Scheres2016]. Computational developments in 3D classification methods have made possible the
73 reconstruction of different conformations at high resolution from heterogeneous cryo-EM datasets
74 [Scheres2007; Scheres2012a; Lyumkis2013; Punjani2017]. The most successful classification
75 method so far is based on an iterative regularized empirical Bayesian strategy [Scheres2012a]. This
76 approach is based on determining the parameters of a model that shows the highest probability of
77 being the true one taking into the account the observed data (images) and statistical prior knowledge.
78 Most classification methods divide the data set into discrete groups (classes) with typically different

79 sizes, although continuous classification algorithms exist [Sorzano2019]. Homogeneous datasets
80 show a linear relationship between the inverse of the obtained resolution (spatial frequency) and the
81 logarithm of the number of particles employed in the 3D reconstruction [Stagg2014; Stagg2008].
82 Then, a common problem after 3D classification is that the resolution obtained from the computed
83 homogeneous classes may be limited by the number of particles belonging to these sets. Moreover,
84 there are technical reconstruction issues that become more severe when the number of particles is
85 small, such as unevenness of the angular distribution [Sorzano2001] and even artefacts of the
86 reconstruction method. Indeed, it is common that after 3D classification only one or maybe a few
87 abundant classes (corresponding to the most stable states) can be resolved at high-resolution after
88 refinement. In the worst possible scenario, only modest resolution reconstructions (between 9 and 20
89 Å) will be attainable for all obtained classes. This situation is typically found when the number of
90 particles was low, or the number of selected classes was high. In all these cases, and even assuming
91 that the 3D classification worked correctly, the number of images in these low populated classes
92 represent a technical problem when aiming for high resolution. Modest resolutions between 9 and 20
93 Å are not sufficient to observe secondary element structures; thus, it is not possible to fully visualize
94 the conformational changes performed by the macromolecule, complicating the molecular
95 understanding of its function. Currently, there is no computational solution to assist users with such
96 issues. The only option users have is returning to the electron microscope again with the hope of
97 collecting enough additional data to resolve the different conformations at sufficient resolution. If
98 resources are available, of course, this is the best approach, but in many practical cases, it may not be
99 feasible. Recently, it has been proposed solutions to process flexible samples [Nakane2018;
100 Schilbach2017]. In [Nakane2018] is presented an approach that performs separated focused
101 refinements with iteratively signal subtractions of rigid bodies defined by the user. In
102 [Schilbach2017] authors showed the capabilities of a new software called WarpCraft, which uses
103 normal mode analysis to model the motions between different regions of a cryo-EM map.

104 In this work, we propose a fast and efficient computational approach that can increase the
105 resolution of modestly populated classes without requiring collecting more data, so it is aimed at

106 “making the best” from relatively limited data sets. As an example, our approach required only 2
107 hours and 50 minutes to process 45059 particles from the 80S ribosome presented below using 10
108 CPUs. The idea has been inspired by classical work in the field of two-dimensional (2D) electron
109 crystallography, where a less-than-perfect 2D crystal is modified (distorted) so that an artificially
110 improved crystal is created [Henderson1986; Gil2006]. The computational method proposed shares
111 some conceptual similarities with the above described 2D crystallography approach. Our method
112 focuses on artificially increasing the number of particles in a class by distorting particles from other
113 similar classes and “moving” them to this class. In this way, as the approach artificially increases the
114 number of images in a small class, then, many technical issues associated with the calculation of a
115 map can be better tolerated, leading to increased resolution reconstructions. However, the big
116 question here is how particles can be “moved” between different conformations, so that, and within
117 specified limits, we can increase the quality of these maps in an objective and quantitative manner.

118 In this work, we use a regularized optical flow (OF) approach [Bouguet2001] to move
119 (deform) particles between different classes/conformations. We have applied this OF method in the
120 past in the context of frame alignment in cryo-EM [Abrishami2015], where we showed that this
121 algorithm is useful to accurately capture local movements between movie frames. In this case, we use
122 OF with a different goal, which is moving (distorting) particles between two (or more) different
123 conformations (input and reference). In more concrete terms, for a given input particle with
124 orientation Θ and belonging to the input conformation, our approach first computes the 2D OF
125 motion field between the two conformations (input and reference) at orientation Θ . To this end, the
126 map conformations are low-pass filtered and projected at the given orientation Θ , then, the 2D
127 motion field is determined by optical flow approach that has been elaborated in the supplemental
128 information, STAR Methods section. Map projections are not highly affected by noise, decreasing
129 the difficulties for a robust estimation of the 2D motion field. The obtained motion field is then
130 applied to the input particle which is, therefore, moved from the input conformation to the reference
131 conformation. This process is repeated for all particles belonging to the input conformation. Here, we

refer to the reference conformation as the one to be improved and to the input —or inputs— conformations as the ones from which particles are to be moved to the reference one. We envision two clear applications of the proposed approach, both aimed at mitigating map reconstruction problems in classes having a limited number of particles. One is the case when 3D classification results in one abundant and one or several additional scarce classes. Then, our aim is to push the resolution of the scarce classes moving particles from the abundant conformation. The other application is based on increasing the number of particles in conformations where all the obtained classes are not highly populated and then the resultant reconstructions are limited by the number of particles. In this case, particles from all classes are moved to one reference conformation.

2. Results

The goal of this method is to improve the quality of a reference 3D class by moving particles from an input conformation (or conformations) to that reference one; we refer to this approach as enRICH (Resolution Improvement in Conformational Heterogeneity maps). The inputs of the method are therefore the maps representing the different conformations and the particles to move. The particles have to be aligned previously, thus, the method requires the following steps: 1) 3D classification to obtain the different map conformations and the particles associated with them; 2) 3D refinement to obtain accurate particle angular assignments; 3) enRICH to move particles from one (or several) input conformations to a reference one; 4) 3D refinement of merged particles belonging to the reference conformation. This pipeline is summarized in Figure 1. In this figure, as an example, our goal is to improve the quality of Class 3 using data from Class 2. The input of enRICH is then the map and particles from Class 2 and the map of Class 3. Therefore, Class 3 is the reference conformation and Class 2 is the input conformation. Steps 1) and 2) from the list above ends with the refinement of particles of each Class. enRICH then starts in Step 3), where we take as input maps from Class 2 and 3 and particles from Class 2. EnRICH produces as output a new set of Class 2 optical flow-deformed particles that we refer to in Figure 1 as “Particles Class2 to Class3”. The workflow continues in Step 4) by merging and refining together corrected particles from Class 2 and

158 particles from Class 3 against Class 3 map, then, obtaining an improved quality map (within certain
159 limits). Observe that enRICH can be used between any classes of Figure 1, but that in this example
160 we have illustrated its use with Class 2 and Class 3 only. Our workflow uses 3D classification and
161 refinement methods from Relion 2.0 electron microscopy package [Kimanius2016], however, any
162 method for 3D classification and refinement can be used as well. In supplementary material we
163 provide detailed information about how particles are moved between different conformations by
164 optical-flow approach, strategies followed to avoid model bias and discussion about limitations of the
165 method.

Figure 1 Diagram of the workflow used to improve reconstructions of heterogeneous samples with enRICH.

We have used the proposed method with data coming from the *Plasmodium falciparum* 80S ribosome, EMPIAR ID 10028 [Wong2014], and from the type-1 ryanodine receptor (RyR1) [DesGeorges2016].

171 2.1 Case with moderate flexibility: 80S ribosome data set

172 Ribosomal particles were obtained with a FEI Polara 300 microscope equipped with a Falcon II
173 camera. The number of projection images deposited (EMPIAR-10028) is 105,145. We first classified
174 the data set in three classes using Relion through Scipion framework [DelaRosa2016]. Class 1
175 (25,323 particles) was composed by particles without 40S subunit, so these particles were not taken
176 into consideration in subsequent analyses. The rest of particles (79,822 particles), distributed in Class
177 2 (45,059 particles) and Class 3 (34,763 particles), were further refined independently obtaining
178 reconstructions of 4.4 Å and 4.6 Å, which were improved to 3.4 Å and 3.6 Å respectively after
179 postprocessing (masking and B-factor correction). We then used enRICH to improve first Class 2 and
180 then Class 3 reconstructions. The processing times were 2 hours and 50 minutes and 6 hours to
181 deform the sets of 34,763 and 45,059 particles using 10 and 5 CPUs respectively. After refinement of
182 merged particles, we obtained resolutions of 4 Å and 4 Å (without postprocessing), and of 3.2 Å and
183 3.2 Å (with postprocessing) for Class 2 and 3 respectively. The local resolution analysis of these
184 results obtained by MonoRes [Vilas2018] are shown in Figures S5 (Class 2) and S6 (Class 3). These
185 figures clearly show that enRICH improves the local resolution for Class 2 and Class 3 when
186 compared with the reconstructions obtained without any correction. The corresponding FSCs of these
187 results are shown in Figure 2(a) and (b). As can be seen from these figures, there is a substantial
188 improvement in the resolution for both Class 2 and Class 3 after running enRICH. In all cases, the
189 obtained FSCs using enRICH surpass the ones attained without using enRICH. In Figure 2(c-f) we
190 show map slides of reconstructions obtained for Class 2 with (c) and without using enRICH (d) and
191 for Class 3 with (e) and without using enRICH (f). Figures 2(c-f) are shown without any
192 postprocessing. Figure 2(g) shows 3D maps of the “head” of the smaller ribosomal subunit obtained
193 with enRICH (left) and with the consensus reconstruction using the complete set of 79,822 particles
194 but without any correction (right). This region corresponds to the more flexible part of the structure
195 and is marked with a red rectangle in Figure 2(c). These maps were masked by the same soft-mask
196 and are shown at the same density threshold value. As can be seen from Figure 2(g) the map obtained
197 with enRICH shows better structural details and the densities are not broken as happen in the other

case. We have also compared the FSC curves obtained from reconstructions computed using all particles (79,822 particles) without applying any 3D classification or correction (consensus reconstruction) with the reconstruction computed from all particles after correction with enRICH, where we used Class 2 as the reference map. These results are shown in Figure S7, which shows an improvement in the FSC obtained after correction with enRICH. These results show that enRICH is able, in this case, to correct the movements of the small ribosomal subunit improving, therefore, the resolution of the 3D reconstruction. Moreover, in Figure S8, we show zoomed density structures obtained from the reconstruction computed using all particles after correction with enRICH (left) and from the consensus reconstruction computed using all particles without applying any 3D classification or correction (right). In all cases, the corresponding atomic model fitted to the density map is also shown. In Figure S8 (A-B), we show a loop and a β -sheet located in the small and flexible 40S ribosomal subunit, while in C) an α -helix and a loop located in the large 60S subunit are rendered. Black arrows and rectangles are used to show the regions that present improvements after correction with enRICH. Note that in Figure S8 (C) we had to set the threshold to a low value where the noise started to corrupt the consensus map (right) to completely render the α -helix in the case of the consensus reconstruction (right). In general, the consensus reconstruction is noisier than the enRICHed one.

215

216 **Figure 2** (a-b) obtained FSCs for Class 2 (a) and Class 3 (b) after refinement for particles belonging to these
 217 classes (blue and green lines) and for the complete data set after correction by enRICH (red and cyan). Blue
 218 and green curves show results without applying any postprocessing, while red and cyan curves present results

219 after masking and performing B-factor correction to reconstructions. (c-f) respective map slides of
 220 reconstructions obtained for Class 2 with (c) and without using enRICH (d) and for Class 3 with (e) and
 221 without using enRICH (f). The red mark corresponds to the “head” region of the smaller subunit. In (g) 3D
 222 reconstructions of the “head” region are shown after correction with enRICH (left) and without any correction
 223 by enRICH and using the complete data set (consensus reconstruction). In both cases the number of particles is
 224 79,822.
 225

226

227 **Figure 3** FSCs of reconstructions obtained after refinement and particle correction with enRICH following
 228 and not following a gold-standard approach for Class 2 and Class 3.
 229

230 We could not find differences between the maps obtained with enRICH using and without
 231 using the gold-standard approach. In this context, enRICH gold-standard means that every particle is
 232 deformed using the 2D motion field calculated from input and reference maps having the same
 233 half_id value than this particle, while non using gold-standard in enRICH just means that all particles
 234 were processed together using only one map as reference and another one as input. In other words,
 235 we used the same concept of the gold-standard in calculating the density maps resolution (see
 236 Avoiding model bias in enRICH in supplementary information). Therefore, particles with half_id=1
 237 (or 2) are deformed by the OF 2D motion field using map projections from maps with half_id=1 (or
 238 2) exclusively. This result shows that enRICH is not affected by overfitting. In Figure 3 we compare
 239 FSCs of reconstructions obtained with enRICH using and not using a gold-standard approach for
 240 Classes 2 and 3. Figure 3 shows no essential differences between reconstructions in both cases. As
 241 can be seen from Figure 2(a-b), the FSCs obtained after Relion auto-refine without postprocessing

(blue lines) drops off to ~0.9-0.8 in the range between 10-20 Å. These small differences between maps can be propagated to the slightly differences in the FSCs shown in Figure 3. The different 0.143-FSC resolution results are summarized in Table S1.

We compared also the obtained maps with the corresponding PDBs (PDB codes: 3j79/3j7a), since it was critical to show that at this relatively high-resolution reconstruction, the improved maps still conserved coherent structural information. PDB 3j79 corresponds to the bigger 60S subunit, while PDB 3j7a refers to the smaller 40S one. Figure 4 shows the obtained results for Class 2 and 3 using (solid line) and not using (dash line) enRICH approach. The obtained FSC resolutions are presented in Table S2. As can be seen from Figure 4, enRICH improves the different FSC curves for both classes at almost all frequencies.

252

Figure 4 FSC curves obtained confronting corresponding PDBs with maps obtained when only particles belonging to Class 2 and 3 (dashed curve) were used, and when we employed all the particles after correction by enRICH (solid line).

256

Moreover, we compared Class 2 and 3 maps with the corresponding PDBs, using EMRinger, before and after applying enRICH, to show the improvement in enRICHed maps. EMRinger is a tool that assesses the precise fitting of an atomic model into the map and is useful for monitoring progress in resolving and modeling high-resolution features in cryo-EM [Barad2015]. EMRinger scores can quantify improvements in the resolvability of atomic features owing to improvements obtained by different ways such as motion correction algorithms, and new data collection procedures that balance dose and radiation damages. The higher EMRinger score reports the better-refined structures and higher resolution maps. For the maps in the 3 to 4 Å range, this value is normally above 1.0. To calculate the EMRinger score for our maps, we used the implemented protocol in Scipion in the name of Phenix-EMRinger [DelaRosa2016]. In the first step, we separated two subunit of each map. Thus, for each map, we have a bigger 60S subunit corresponding and aligned to PDB model 3j79, and a smaller 40S subunit aligned with PDB model 3j7a. Table 1 shows the results of applying EMRinger for each sub-map and corresponding PDB model.

Table 1 EMRinger scores obtained from maps reconstructed using Class 2 and Class 3 particles (with and without using enRICH), and all particles with and without correction with enRICH.

	EMRinger Score
Class2 40S subunit	0.83
Class2 40S subunit (enRICHed)	1.24
Class2 60S subunit	1.5
Class2 60S subunit (enRICHed)	1.55
Class3 40S subunit	0.93
Class3 40S subunit (enRICHed)	1.23
Class3 60S subunit	1.44
Class2+Class3 60S subunit (enRICHed)	1.59
Class2+Class3 40S subunit	1.11
Class2+Class3 40S subunit (enRICHed)	1.48
Class2+Class3 60S subunit	1.47
Class2+Class3 60S subunit (enRICHed)	1.53

As Table 1 shows, all maps obtained with applying our method (enRICH) have the higher EMRinger score that shows improvements in term of consistency between maps and model. Specifically, smaller 40S subunit that corresponds to the more flexible part of the structure shows better improvement in both enRICHed class 2 and 3.

For the sake of clarity and completeness, we have repeated this experiment comparing the result obtained from enRICH for Class 2 with the results computed using all particles (79,822 particles) without applying any 3D classification or correction (consensus reconstructions) when atomic models are used as references. These results are shown in Figures S9 (bigger 60S subunit) and S10 (smaller 40S subunit). As can be seen from these figures, enRICH improves the FSCs, especially for the small moving 40S subunit when compared with the results provided by the consensus reconstruction. In Table 1 we show also the EMRinger scores provided by reconstructions obtained from all particles when compared with both the 40S and 60S atomic models. Table 1 shows higher values from enRICHed maps in comparison with results computed from the consensus reconstruction without any 3D classification or correction, which approves the application of our developed method.

Finally, to show the capacity of the approach for classes with very low number of particles, we further classified the dataset composed by merged particles (without any correction) from Classes 2 and 3, into 4 classes. The population of each class was 25,087, 23,702, 5,744 and 24,289 respectively. These classes refined to resolutions of 5.3, 5.7, 15.1 and 5.8 Å without any applied mask or B-factor correction. We then applied enRICH to move particles from Class1, 2 and 4 to Class 3, which is the least populated class. Figure 5 shows the gold-standard FSCs obtained after 3D refinement by Relion auto-refine, without masking or performing B-factor correction for Classes 1, 2, 3 and 4 and after merging all particles to Class 3 using enRICH (blue curve). As can be seen from Figure 5, the resolution is substantially improved after using enRICH.

Figure 5 FSCs obtained for Class 1, 2, 3 and 4 reconstructions (green, red, cyan and pink) and after merging all particles to Class 3 using enRICH (blue line).

In Figure S11, we have compared the FSC curves of the consensus reconstruction using all particles without applying any 3D classification or correction, with the reconstruction computed from all particles after correcting and merging particles from the four 3D classes with enRICH, where we used Class 3 as reference map. Figure S11 clearly shows that using enRICH to turn the particles to the class with lowest resolution, although improves the resolution of that class even when compared with the consensus reconstruction using all particles, it produces a reconstruction with lower resolution than the enRICHed one shown in Figure S7. The main reason is that the reference map in this case has fewer number of particles (~5k). Thus, this reference map is of lower quality causing that the deformation maps computed from enRICH are not accurate enough to provide a better enRICHed reconstruction.

2.2 Case with large flexibility: RyR1 data set:

This data set was composed by 277,044 single particles of RyR1 bound to Ca²⁺, ATP and caffeine. We first classified this data set into two sets composed by 129,873 (Class 1) and 147,171 (Class 2) particles respectively. Class 1 and Class 2 refined until 4.8 Å and 4.9 Å (0.143-FSC gold standard) using Relion. In this case, we found that the large conformational changes shown by the RyR1

317 prevented our method to improve these results. Indeed, after moving particles with enRICH from
318 Class 2 to Class 1 and refine the merged set, the resolution achieved was 4.9 Å (0.143-FSC gold
319 standard). As mentioned in supplemental information (section limitations of the method), complexes
320 with large conformation movements, as the RyR1 specimen, prevent our method to obtain high-
321 resolution reconstructions. In these situations, our approach should be used to improve maps at
322 intermediate resolutions only. Therefore, in this case it is not realistic the use of enRICH for
323 improving map resolutions better than 5 Å. However, although we could not improve these results,
324 we employed this data set to show a different use of our approach, where enRICH was able to
325 improve intermediate resolution reconstructions coming from low populated classes.

326 In the second experiment, we decided to further classify Class 2 (147,171 particles) into four
327 classes. The particle distribution between classes and the obtained resolution after refinement were
328 Class 2.1: 38,543 particles, resolution 7.2 Å; Class 2.2: 39,239 particles, resolution 7.2 Å; Class 2.3:
329 35,174 particles, resolution 7.3 Å; Class 2.4: 34,215 particles, resolution 7.3 Å. We then moved
330 particles from Class 2.2 to Class 2.1, particles from Class 2.2 and from Class 2.3 to Class 2.1, and
331 particles from Class 2.2, Class 2.3 and Class 2.4 to Class 2.1. The processing time to deform ~40,000
332 particles of 400x400 pixels using 44 CPUs was 1 hour and 35 minutes. In each case, we refined the
333 merged sets obtaining respective FSCs from Relion that are shown in Figure 6(a). As can be seen
334 from Figure 6(a), the resolution of Class 2.1 was improved from 7.2 (38,543 particles) to 5.3 Å when
335 the dataset was composed by 147,171 particles. Additionally, in Figure 6(b-c) we show respective
336 map slides (without any postprocessing) obtained after merging all particles by enRICH and refining
337 with Relion (Figure 6(b)), and after refining the data set of 38,543 particles alone (Figure 6(c)). As
338 can be seen from these figures, the map slide obtained after merging all particles with enRICH has
339 better resolution.

Figure 6 (a) Gold-standard FSCs obtained from Class 2.1 particles (blue curve), Class 2.1 + Class 2.2 (corrected) (green curve), Class 2.1 + Class 2.2 (corrected) + Class 2.3 (corrected) (red curve), Class 2.1 + Class 2.2 (corrected) + Class 2.3 (corrected) + Class 2.4 (corrected) (cyan curve). (b-c) respective slides of maps obtained after merging all particles by enRICH and then refining with Relion (b), and after refining Class 2.1 data set (38,543 particles) alone (c).

For in this experiment, we have also compared the FSC curves of reconstructions obtained from all particles with and without using enRICH in Figure S12. In this case, using enRICH to turn the particles to the class with lowest resolution (Class2.2), although improves the resolution of that class substantially so it possibilities to visualize that conformation better, also provides a reconstruction showing lower global resolution than the consensus reconstruction using all particles without any correction and 3D classification. Note that this result still is interesting as allow us to visualize the different conformations after 3D classification with improved resolution after correction with enRICH.

356 **3. Discussion**

357 Macromolecules are not static entities and typically exhibit different conformations when performing
358 their function. Cryo-EM is a high-resolution structure determination technique that can capture these
359 different states providing essential information about the macromolecule's function. Hence, 3D
360 classification is a routine and crucial task when processing cryo-EM data. However, classifying data
361 requires dividing the collected data into sets of homogeneous particles, a process that usually limits
362 the attainable resolution for the different conformations. This situation is more dramatic in complexes
363 showing a large (or infinite) number of different possible conformations [Sorzano2016]. In this work,
364 we propose a method to merge particles from different conformations. This approach can be used in
365 cases where the resulting number of particles after 3D classification limits the attainable resolution.
366 Then, this method can be used to improve (within limits) the quality of reconstructions coming from
367 scarce classes, for example. The approach is based on obtaining the 2D motion or deformation field
368 by a regularized optical flow approach. As shown from our results, the approach is fast and
369 computationally efficient. The OF is computed between projections of different conformations at
370 certain orientations given by the experimental particles to process. These motion fields are then
371 applied to corresponding particle images to move them between conformations. The conceptual idea
372 of the proposed method is close to crystal unbending implemented in 2D electron crystallography
373 [Henderson1985, Gil2006] that enabled a dramatic improvement in resolution. The main goal of
374 crystal unbending is to correct the crystal distortions, that is, its deviation from regular lattice
375 repetition. This technique first obtains the 2D motion field of unit cells with respect to their perfect
376 regular positions and then correct these shifts by interpolation obtaining a corrected crystal image. In
377 both 2D crystallography and single particle analysis underlying physical limitations are crystal local
378 deformations in the first case and large-scale low frequencies motions in the second.

379 We have taken special attention in avoiding problems related with model-bias when using our
380 approach. First, we use a robust optical flow approach which imposes strong regularization in the

381 obtained 2D motion fields. Typical integration kernels and imposed resolutions to map projections
382 used in the estimation of the motion fields by OF are of 25-50 Å and 15-20 Å, respectively. All the
383 results shown in this work have been obtained using an OF integration kernel of 50 Å and the maps
384 have been filtered to 20 Å resolution. Additionally, enRICH can be used in a gold-standard fashion.
385 In this case the method uses half maps of each conformation (input and reference) to determine
386 motion fields. Particles are corrected according to their corresponding halves.

387   It is important to mention that our method imposes a deformation model. This model assumes
388 that the 3D motion field between conformations is small, smooth, continuous and then obtainable
389 from map projection images. This assumption is only fulfilled when the 3D movements between
390 conformations is small and, better still, when there is an intrinsic element of continuous flexibility in
391 the macromolecule under study. That is the difference between enRICH and Relion multibody
392 refinement, as if the domains or regions are small, due to masking in Relion multibody approach, the
393 accurate alignment is challenging and most of the times is not possible. However, our focus is on a
394 different problem which is small continuous movement without defining any mask. Also, our
395 approach is addressed to capture and correct movements of large regions as subunits, but not to
396 determine small local changes of smaller elements such as secondary structures. An important
397 advance of enRICH is that it can deal with samples affected by flexibility. In these cases, the data set
398 can be 3D classified into a large number of classes, each affected by low resolution because the
399 limited number of particles per class. Then enrich can merge all particles from these 3D classes into
400 one corrected class with improved resolution.

401   We have tested our approach on two different classes of ryanodine receptor (EMD-2751, and
402 EMD-2752), and ribosome (EMD-6647, and EMD-6653) by simulating the particle images as has
403 been discussed in supplemental information (section Simulations).

404   In addition, we have applied our approach to two different heterogeneous samples: 80S
405 ribosome and RyR1. In the case of the 80S ribosome, we clearly improved resolution results obtained
406 using pure classes after 3D classification. This result shows that the deformation pattern followed by
407 these ribosomal particles aligns well with our model. On the other hand, we also showed a case

where enRICH could not improve resolutions of homogeneous sets after classification. In this situation our deformation model may not describe well the large movements exhibited by the RyR1 macromolecules at high-resolution, limiting the attainable resolution to ~ 5 Å. However, we employed this dataset to show other practical uses of our approach. Interestingly, our defined control checks (please see ‘Theoretical Analysis of enRICH’ in supplementary material) to evaluate the capacity of enRICH to process heterogeneous data were 0.26 and 1.17 for the ribosome and RyR1 respectively. These scores agree with the different results obtained.

Acknowledgments

We are grateful to Rouslan Efremov for helpful discussions and critical reading of the manuscript and to Josue Gomez-Blanco to help us with Chimera and in the evaluation of the maps. We are also thankful for HPC resources provided by Bio21 Institute to perform part of this analysis. This work was supported by McGill start-up funds,), the Spanish Ministry of Science and Innovation through the call 2019 Proyectos de I+D+i - RTI Tipo A (PID2019-108850RA-I00), Comunidad de Madrid (S2017/BMD-3817) and the MINECO – Spain (BIO2016-76400-R (AEI/FEDER, UE)). J.V. acknowledges economical support from the Ramón y Cajal 2018 program (RYC2018-024087-I). M.K. acknowledges Isabelle Rouiller for her support and providing some computational resources.

Author Contributions

J.V. had the main idea and wrote the manuscript, J.V and M.K developed the method, A.G provided the RyR1 data. All authors reviewed the manuscript, supervised the experiments and discussed the results.

Declaration of Interests

The authors declare no competing interests.

432 **References**

- 433 [Abrishami2015] Abrishami, V., Vargas, J., Li, X., Cheng, Y., Marabini, R., Sorzano, COS, Carazo,
434 JM., “Alignment of direct detection device micrographs using a robust Optical Flow approach” J.
435 Struct. Biol. 189(3) 163–176 (2015).
- 436 [Barad2015] Barad, BA., Echols, N., Wang, R., Cheng, Y., DiMaio, F., Adams, PD., Fraser, JS.,
437 “EMRinger: side chain-directed model and map validation for #D cryo-electron microscopy” Nat
438 Methods, 12(10), 943-948 (2015).
- 439 [Boland2017] Boland, A., Chang, L., Barford, D., “The potential of cryo-electron microscopy for
440 structure-based drug design” Essays Biochem., 61(5) 543-560 (2017).
- 441 [Bouguet2001] Bouguet, JY., “Pyramidal Implementation of the Lucas Kanade Feature Tracker
442 Description of the algorithm” Intel Corporation 2 (2001).
- 443 [Cressey2017] Cressey, D., Callaway, E., “Cryo-electron microscopy wins chemistry Nobel” Nature,
444 550 167 (2017).
- 445 [DelaRosa2016] Dela Rosa, JM., Quintana, A., Del Cano, L., Zaldivar, A., Foche, I., Gutierrez, J.,
446 Gomez Blanco, J., Burguet, J., Cuenca Alba, J., Abrishami, V., Vargas, J., Oton, J., Sharov, G.,
447 Vilas, JL., Navas, J., Conesa, P., Kazemi, M., Marabini, R., Sorzano, COS., Carazo, JM., “Scipion: a
448 software framework toward integration, reproducibility and validation in 3D electron microscopy” J.
449 Struct. Biol. 195(1) 93–99 (2016).
- 450 [DesGeorges2016] Des Georges, A., Clarke, OB., Zalk, R., Yuan, Q., Condon, KJ., Grassucci RA.,
451 Hendrickson WA., Marks, AQ., Frank, J., “Structural Basis for Gating and Activation of RyR1” Cell
452 167(22) 145-157 (2016).
- 453 [Gil2006] Gil, D., Carazo, JM., Marabini, R., “On the nature of 2D crystal unbending” 156(3) 546-
454 555 (2006).
- 455 [Kimanius2016] Kimanius, D., Forsberg, BO., Scheres, SHW., Lindahl, E., “Accelerated cryo-EM
456 structure determination with parallelisation using GPUs in RELION-2” ELife 15(5) e18722 (2015).

457 [Henderson1986] Henderson, R., Baldwin JM., Downing, KH, Lepault, J., Zemlin, F., “Structure of
 458 purple membrane from Halobacterium Halobium: recording, measurement and evaluation of electron
 459 micrographs at 3.5 Å resolution” Ultramicroscopy 19 147-178 (1986).

460 [Lyumkis2013] Lyumkis, D., Brilot, AF., Theobald, DL., Grigorieff, N., “Likelihood-based
 461 classification of cryo-EM images using FREALIGN” J. Struct. Biol. 183(3) 377–388 (2013).

462 [Nakane2018] Nakane, T., Kimanius, D., Lindahl, E., Scheres, SHW, “Characterisation of molecular
 463 motions in cryo-EM single article data by multibody refinement in Relion” eLife, 7 (2018)

464 [Punjani2017] Punjani, A., Rubinstein, JL., Fleet, DJ., Brubaker, MA., “cryoSPARC: algorithms for
 465 rapid unsupervised cryo-EM structure determination” Nat. Methods, 14(3) 290-296 (2017).

466 [Scheres2007] Scheres, SHW, Gao, H., Valle, M., Herman, GT., Eggermont, PPB., Frank, J., Carazo,
 467 JM., “Disentangling conformational states of macromolecules in 3D-EM through likelihood
 468 optimization” Nat. Methods, 4 27-29 (2007).

469 [Scheres2012a] Scheres, SHW., “RELION: Implementation of a Bayesian approach to cryo-EM
 470 structure determination” J. Struct. Biol. 180 519–530 (2012).

471 [Scheres2012b] Scheres, SHW., Chen, S., “Prevention of overfitting in cryo-EM structure
 472 determination” Nat. Methods, 9(9) 853-854 (2012).

473 [Scheres2016] Scheres, SHW., “Chapter Six - Processing of Structurally Heterogeneous Cryo-EM
 474 Data in RELION” Methods in Enzymology, 579 125-157 (2016).

475 [Schilbach2017] Schilbach, S., Hantsche, M., Tegunov, D., Wigge, C., Urlaub, H., Cramer, P.,
 476 “Structures of transcription pre-initiation complex with TFIIF and Mediator” Nature 551 204-209
 477 (2017).

478 [Sorzano2001] C.O.S.Sorzano, R.Marabini, N.Boisset, E.Rietzel, R.Schröder, G.T.Herman,
 479 J.M.Carazo. The effect of overabundant projection directions on 3D reconstruction algorithms.
 480 Journal of Structural Biology, 133 (2,3), 108-118 (2001)

481 [Sorzano2018] C.O.S. Sorzano, J. Vargas, J.M. de la Rosa-Trevín, A. Jiménez-Moreno, R. Melero,
 482 M. Martínez, E. Ramírez-Aportela, P. Conesa, J.L. Vilas, R. Marabini, J.M. Carazo. A new algorithm

for high-resolution reconstruction of Single Particles by Electron Microscopy. *J. Structural Biology*,
204: 329-337 (2018)

[Sorzano2019] C.O.S. Sorzano, A. Jiménez, J. Mota, J.L. Vilas, D. Maluenda, M. Martínez, E.
Ramírez-Aportela, T. Majtner, J. Segura, R. Sánchez-García, Y. Rancel, L. del Caño, P. Conesa, S.
Jonic, J. Vargas, F. Cazals, Z. Freyberg, J. Krieger, I. Bahar, R. Marabini, J.M. Carazo. Survey of the
analysis of continuous conformational variability of biological macromolecules by electron
microscopy. *Acta Crystallographica Section F* 75: 19-32 (2019)

[Stagg2008] Stagg, SM., Lander, GC., Quispe, J., Voss, NR., Cheng, A., Bradlow, H., Bradlow, S.,
Carragher, B., Potter, CS., “A test-bed for optimizing high-resolution single particle
reconstructions” *J Struct Biol.* 163(1) 29–39 (2008).

[Stagg2014] Stagg, SM., Noble, AJ, Spilman, M., Chapman MS., “ResLog plots as an empirical
metric of the quality of cryo-EM reconstructions” *J Struct Biol.* 185(3) 418–426 (2014).

[Sorzano2016] Sorzano, COS., Alvarez-Cabrera, AL., Kazemi, M., Carazo, JM., Jonic, S.,
“StructMap: Elastic distance analysis of electron microscopy maps for studying conformational
changes” *Biophysical J.* 110 1753-1765 (2016).

[Vilas2018] Vilas, JL., Gómez-Blanco J., Conesa P., Melero R., Miguel de la Rosa-Trevín J., Otón
J., Cuenca J., Marabini R., Carazo JM., Vargas J., Sorzano COS., “MonoRes: Automatic and
accurate estimation of local resolution for electron microscopy maps” *Structure* 26(2) 337-344
(2018).

[Wong2014] Wong, W., Bai, XC., Brown, A., Fernandez IS., Hanssen, E., Condrón, M., Tan YH.,
Baum, J., Scheres, SHW., “Cryo-EM structure of the Plasmodium falciparum 80S ribosome bound to
the anti-protozoan drug emetine” *eLife* 3 e03080 (2014).

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Kazemi, M;Sorzano, COS;Carazo, JM;des Georges, A;Abrishami, V;Vargas, J

Title:

ENRICH: A fast method to improve the quality of flexible macromolecular reconstructions.

Date:

2021

Citation:

Kazemi, M., Sorzano, C. O. S., Carazo, J. M., des Georges, A., Abrishami, V. & Vargas, J. (2021). ENRICH: A fast method to improve the quality of flexible macromolecular reconstructions.. Progress in Biophysics and Molecular Biology, 164, pp.92-100. <https://doi.org/10.1016/j.pbiomolbio.2021.01.001>.

Persistent Link:

<http://hdl.handle.net/11343/263711>