*Bioinformatics*, 35(14), 2019, 2427–2433
doi: 10.1093/bioinformatics/bty985
Advance Access Publication Date: 30 November 2018
Original Paper





Structural bioinformatics

# Validation of electron microscopy initial models via small angle X-ray scattering curves

**Amaya Jiménez[1], Slavica Jonic[2], Tomas Majtner[1], Joaquín Otón[1], Jose Luis Vilas[1], David Maluenda[1], Javier Mota[1], Erney Ramírez-Aportela[1], Marta Martínez[1], Yaiza Rancel[1], Joan Segura[1], Ruben Sánchez-García[1], Roberto Melero[1], Laura del Cano[1], Pablo Conesa[1], Lars Skjaerven[4], Roberto Marabini[1,3], Jose M. Carazo[1] and Carlos Oscar S. Sorzano[1,5,*]**

[1]Biocomputing Unit, Centro Nac. Biotecnología (CSIC), Cantoblanco, Madrid 28049, Spain, [2]UMR CNRS 7590, Muséum National d 'Histoire Naturelle, IRD, Institut de Minéralogie, de Physique des Matériaux et de Cosmochimie, IMPMC, Sorbonne Université, Paris 75005, France, [3]Department of Computer Science, University Autónoma de Madrid, Cantoblanco, Madrid 28049, Spain, [4]Department of Biomedicine, University of Bergen, 5020 Bergen, Norway and [5]Department of Engineering of Electronic and Telecommunication System, University San Pablo–CEU, Campus Urb. Montepríncipe, Boadilla del Monte, Madrid 28668, Spain

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on May 9, 2018; revised on October 29, 2018; editorial decision on November 26, 2018; accepted on November 29, 2018

## Abstract

**Motivation:** Cryo electron microscopy (EM) is currently one of the main tools to reveal the structural information of biological macromolecules. The re-construction of three-dimensional (3D) maps is typically carried out following an iterative process that requires an initial estimation of the 3D map to be refined in subsequent steps. Therefore, its determination is key in the quality of the final results, and there are cases in which it is still an open issue in single particle analysis (SPA). Small angle X-ray scattering (SAXS) is a well-known technique applied to structural biology. It is useful from small nanostructures up to macromolecular ensembles for its ability to obtain low resolution information of the biological sample measuring its X-ray scattering curve. These curves, together with further analysis, are able to yield information on the sizes, shapes and structures of the analyzed particles.

**Results:** In this paper, we show how the low resolution structural information revealed by SAXS is very useful for the validation of EM initial 3D models in SPA, helping the following refinement process to obtain more accurate 3D structures. For this purpose, we approximate the initial map by pseudo-atoms and predict the SAXS curve expected for this pseudo-atomic structure. The match between the predicted and experimental SAXS curves is considered as a good sign of the correctness of the EM initial map.

**Availability and implementation:** The algorithm is freely available as part of the Scipion 1.2 software at `http://scipion.i2pc.es/`.

**Contact:** coss@cnb.csic.es

## 1 Introduction

The knowledge of the three-dimensional (3D) structures of macromolecules is key to understand the molecular interactions and the functional operation of biological ensembles. Single particle analysis (SPA) by Cryo Electron Microscopy (EM) has established as one of the key players to provide the knowledge of their 3D maps at near-atomic resolution (Nogales, 2016). The transmission electron microscope acquires thousands of experimental images, coming from projections of the original structure under analysis, which are then used to estimate a 3D map representing the structure. Usually, an iterative refinement process is applied to obtain this 3D representation. Starting from an initial estimation of the 3D structure, the experimental images are used to improve the accuracy of the estimated 3D model.

However, refinement algorithms easily end up in a local minima of the solution space (Sorzano *et al.*, 2006). Therefore, the initial 3D map, which represents the starting point in the solution space, is key in the accuracy of the final estimation, i.e. the behavior of the refinement process greatly depends on the initial 3D map. This is known as the 'initial model problem', and its determination is, in some cases, one of the major issues in the field (Voss *et al.*, 2010).

The 'initial model problem' has been addressed multiple times in the literature. One approach was the Random Conical Tilt and other orthogonal tilt methods (Sorzano *et al.*, 2015a), where the sample images were taken using several known tilt angles. There were also a great variety of approaches based on Common Lines (Elmlund and Elmlund, 2012; Shkolnisky and Singer, 2012). Other methods were based on computer-generated shapes (Bilbao-Castro *et al.*, 2004; Ludtke *et al.*, 2004). Several algorithms were developed using the experimental images, from re-constructions based on one image of a particle assuming certain symmetry (Cantele *et al.*, 2003) to the assignment of random orientations to the image class averages (Harauz and Van Heel, 1985). Other algorithms thoroughly exploit robust statistical techniques to estimate it (Sorzano *et al.*, 2015b; Vargas *et al.*, 2014). A method was devised in Sorzano *et al.* (2018) to obtain a consensus among the initial maps obtained with different algorithms. However, in spite of the great effort of the research community to solve this problem, it is still common that many of these techniques suggest an ensemble of possible solutions, mixing correct and incorrect ones. Therefore, it is still necessary to discover new ways to identify the initial 3D maps that could result in an accurate final 3D model of a biological structure.

Small angle scattering of X-rays (SAXS) is an alternative technique to study biological structures that provides low resolution information of macromolecules in solution (Koch *et al.*, 2003). Specifically, the scattered intensity is the Fourier transform of the correlation function of the electronic density, which means that X-ray scattering experiments are able to reveal low resolution features of the spatial correlations in the sample. Small angle scattering experiments measure the scattered intensity at very small scattering distances, typically from few micro-radians to a 10 of radians.

Moreover, SAXS is a complementary tool to higher resolution methods, e.g. EM and can be used for low resolution structure validation. SAXS experiments give a one-dimensional profile, called scattering curve (SAXS curve), where the scattered intensity recorded by a detector is represented as a function of the scattering angle. It must be highlighted that there is a large variety of tools available to work with SAXS data. DARA (Kikhney *et al.*, 2016) is a web server with hundreds of thousands of scattering profiles pre-computed from models deposited in the Protein Data Bank to compare with any given SAXS curve finding the most similar ones. This information can give a valuable insight on the structure of the particle. Other example is ATSAS (Franke *et al.*, 2017) which is a program suite covering the pipeline for SAXS data processing.

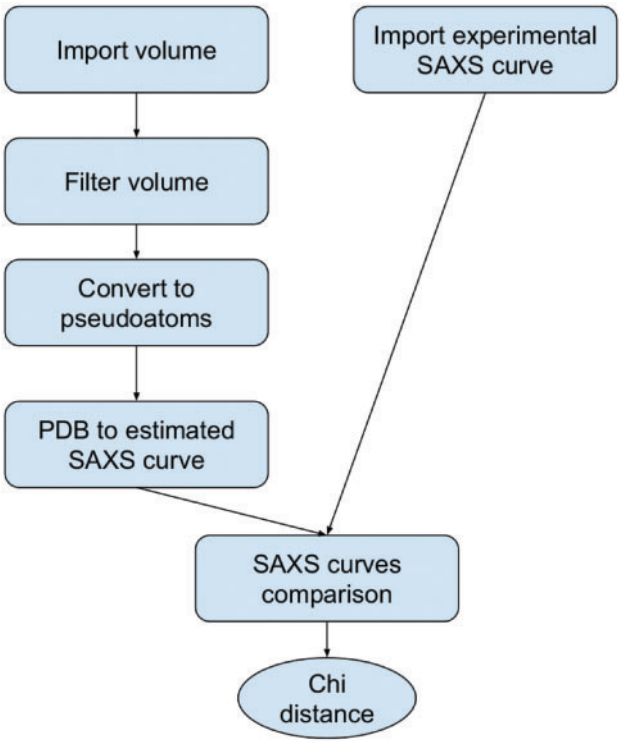


**Fig. 1.** Workflow for comparing a predicted SAXS curve to a set of initial 3D maps

In practice, a non-negligible number of users has available SAXS and EM information from the same biological sample. Our proposal relies on the combination of both pieces of structural information, SAXS and EM, to validate the initial 3D map estimation in EM. It must be highlighted that from an estimated initial map at low resolution there is no atomic structure from which the SAXS curve can be easily calculated. As we demonstrate in this work, from a pseudo-atomic representation (Jonić and Sorzano, 2016), we are able to simulate a SAXS curve that, afterwards, can be compared with the experimental one. In this way, the ensemble of solutions suggested by most of current Cryo EM image processing methods can be pruned using SAXS data, revealing the most suitable initial maps for the subsequent refinement steps.

The SAXS curves for the pseudo-atomic representation of an initial map can be obtained using Crysol (Svergun *et al.*, 1995), as part of the ATSAS suite (Franke *et al.*, 2017).

## 2 Materials and methods

Our aim is to compare the predicted SAXS curves of a set of initial 3D maps with the experimental SAXS curves obtained from the real structure. In this way, the power of SAXS and Cryo EM can then be combined in the selection of that candidate map whose predicted SAXS curve is most similar to the experimentally observed one.

The workflow proposed to generate the predicted SAXS curves is as follows. Given any candidate initial Cryo EM map: (i) the map is low-pass filtered to discard any high resolution information (a typical cut-off resolution may be 15 Å), optionally the model may also be masked to remove artifacts clearly not belonging to the

structure; (ii) the map is converted to a pseudo-atoms model with a pseudo-atom Gaussian sigma of 2 voxels (Jonić and Sorzano, 2016); (iii) from the previous model, the predicted SAXS curves are calculated (Svergun *et al.*, 1995) and compared to the experimental one using the $\chi^2$ distance. This workflow is summarized in Figure 1.

## 3 Results

To validate the workflow outlined above, we have used three different structures. The data of two of them were taken from the Electron Microscopy Data Bank (EMDB) and Scattering Biological Data Bank (SASBDB); specifically, these structures are *Gallus gallus* lysozyme (EMDB entry 8217 and SASBDB entry SASDAG2) and *Bos taurus* catalase (EMDB 6314 and SASBDB SASDA92). The third one, Human tyrosine hydroxylase isoform 1, was experimentally obtained by some of the authors of the present work and is described in Bezem *et al.* (2016) (currently also deposited in SASBDB entry SASDBZ4). The molecular weight of lysozyme is 14 kDa, of catalase is 240 kDa and for tyrosine hydroxylase is 284 kDa, so our test are carried out with a variety of molecular weights. We simulated the construction of the initial map in EM by generating projections from low pass filtered versions of the known 3D structures of the three macromolecules, and giving these projections as inputs to the initial map estimation by two different algorithms, EMAN2 and RANSAC [as implemented in Scipion v1.1 (de la Rosa-Trevín *et al.*, 2016)]. These algorithms returned a total of 10 candidate structures, representing an ensemble of maps mixing suitable and unsuitable starting solutions for a re-construction process. From these maps, the pseudo-atoms models were generated following the method in Jonić and Sorzano (2016). The pseudo-atomic representation can be seen as an approximation problem between an input density map $V$ and an approximated one $V_{\text{pseudo}}$ with 3D Gaussian as radial basis functions (RBFs). Considering each pseudo-atom as a RBF, $V_{\text{pseudo}}$ can be mathematically described as:

$$V_{\text{pseudo}}(\bar{r}) = \sum_{i=1}^{M} c_i b_\sigma(||\bar{r} - \bar{r}_i||), \tag{1}$$

where $\bar{r} \in \mathbb{R}^3$, $M$ is the number of RBFs, with $b$ being one of them with amplitude $c_i$, width $\sigma$ and $\bar{r}_i$ its center position. With this definition, the goal is to find the set of parameters for $V_{\text{pseudo}}$ that generate a representation error that satisfies:

$$e \propto \frac{1}{V}\sum_{j=1}^{V} |V(\bar{r}_j) - V_{\text{pseudo}}(\bar{r}_j)| < \epsilon, \tag{2}$$

with $V$ the number of voxels and $\epsilon$ the target volume approximation error. In our experiments, the approximation error was 5%.

From the pseudo-atomic maps, an estimation of the SAXS curves is obtained with Crysol (Svergun *et al.*, 1995). Our aim is to measure the differences between the experimental SAXS curve and the estimated one from a correct initial map and an incorrect one. The fitting between experimental and estimated curves is measured using $\chi^2$, defined as:

$$\chi^2 = \int_0^{R_{\max}} \frac{|\log(I_{\exp}(R)) - \log(I_{\text{est}}(R))|^2}{\log(I_{\exp}(R))} dR, \tag{3}$$

being $I_{\exp}$ and $I_{\text{est}}$ the experimental and estimated curves in the Fourier domain, respectively, and taking into account the frequency range of interest, $R_{\max}$. $\chi^2$ gives us a quantitative value to objectively measure the different behavior of our proposal with correct and incorrect initial volumes. Note that the area under the intensity curves is related to the molecular mass of the macromolecule. Crysol automatically scales the estimated SAXS curve to minimize the weighted distance between the estimated and experimental curves, as described in Svergun *et al.* (1995). In this way, the macromolecular mass associated to both curves are as similar as possible.

To check the pseudo-atomic maps obtained, in parts (a) and (c) of the Figures 2–4, it can be seen the initial map (left side of the figure) along with the pseudo-atomic representation showed by spheres surrounded by a semi-transparent low resolution version of the map (right side of the figure). In parts (b) and (d) of the Figures 2–4, it can be seen the experimental SAXS curve, in continuous line and the obtained SAXS curve for the pseudo-atomic representation, in dashed line, up to the cutoff frequency of 15 Å used to low-pass filtering the maps. In all the examples, the first row of the figures [parts (a) and (b)] corresponds to a correct estimated initial map and second row [parts (c) and (d)] corresponds to an incorrect one. Just a look to the experimental and estimated SAXS curves gives an intuitive idea that the most similar initial map to the structure under analysis corresponds to that with more similar SAXS curves. The quantitative results of the fitting measured by means of the $\chi^2$ value are presented in Table 1.

As can be seen, the correlation between the predicted and the experimental SAXS curves was always lower for the incorrect initial volumes (higher $\chi^2$ values). Therefore, we only need to know the $\chi^2$ between the curves to select the best input volumes for going to the following refinement steps.

We also wanted to compare our proposal with the DARA server (Kikhney *et al.*, 2016). DARA searches for the most similar SAXS curves from a pre-calculated dataset from a large number of atomic structures (PDB), giving as result a list with the most similar atomic structures in SAXS terms. Consequently, comparison with DARA as a performance test is a very good way to analyze if the results of our newly proposed method and DARA are able to reveal similar structures for those test cases in which the true solution is already known as an structural model. To carry out the comparison, we use the most similar PDB structure given by DARA server and the best map selected with our workflow for the tyrosine hydroxylase. The results shown in Figure 5 reveal that the original density map (that was calculated converting the PDB) and the one obtained with DARA are quite far from being similar; however, the highest similarity map selected by comparison of SAXS curves maintains the general shape quite similar to that of the original. It must be also highlighted that, while we select a reasonable good initial map for the lysozyme and the catalase, DARA server in the first case was not able to find the correct PDB and in the second one it was found as the fourth most similar structure. This test with DARA shows the difficulty of trying to find similar structures to the one that produced the input SAXS curve. For this reason, it is useful to start from images, reconstruct an initial map and to validate it, as we proposed.

All these results confirm the value of having alternative tools, as the one presented in this paper, to select among candidate initial maps generated by current methods in EM. Moreover, our proposal based on SAXS curves matching has shown improvements in obtaining similar structures to the one being analyzed when compared to DARA, a method with similar basis.

The results presented in this work show several advantages thanks to the use of SAXS curves to complement the information obtained by EM. One of the main difficulties to generate 3D initial map estimations arises from poor angular assignments of 2D images, generating an inaccurate 3D map estimation that our proposal is able to discriminate. Currently, there are plenty of methods

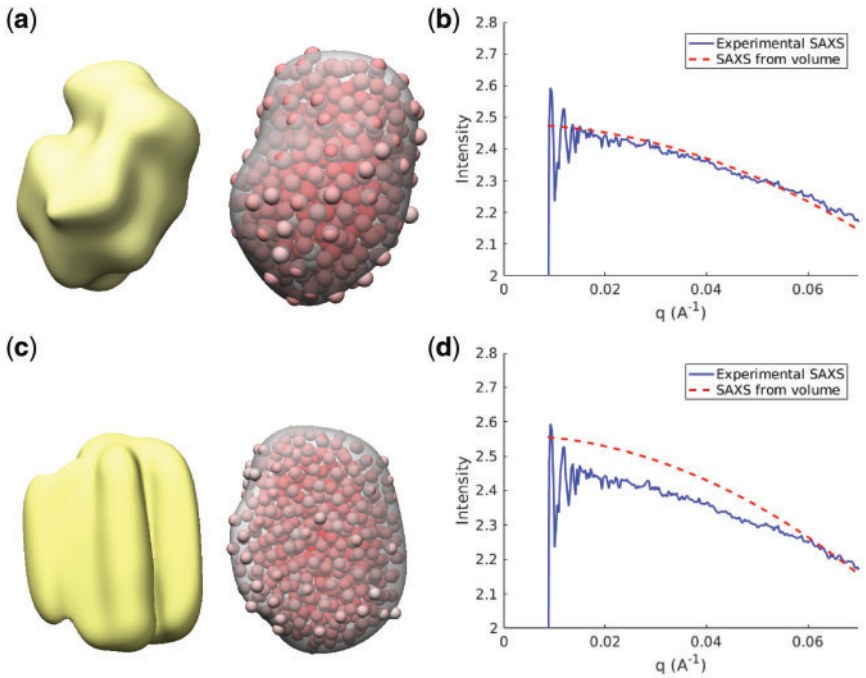


**Fig. 2.** Lysozyme results. First row for a high similarity initial volume. Second row for a low similarity initial volume. (**a**) and (**c**) The initial volume to the left, along with its pseudo-atomic low pass filtered version to the right. (**b**) and (**d**) Comparison between experimental and predicted SAXS curves in semi-log scale in the low frequency region up to 15 Å

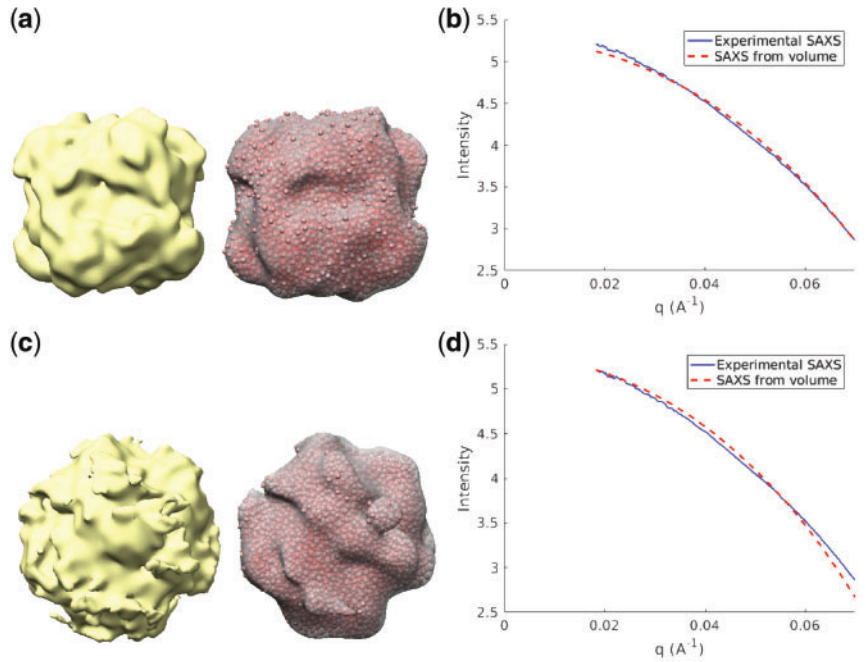


**Fig. 3.** Catalase results. First row for a high similarity initial volume. Second row for a low similarity initial volume. (**a**) and (**c**) The initial volume to the left, along with its pseudo-atomic low pass filtered version to the right. (**b**) and (**d**) Comparison between experimental and predicted SAXS curves in semi-log scale in the low frequency region up to 15 Å

for initial map estimation able to generate a variety of initial volumes, so the possibility of obtaining a ranking based on our measures ($\chi^2$) could be very useful to select the best ones for the following refinement steps, or even to discard all of them if they are very far from the experimental SAXS curve.

### 3.1 Evaluating pseudo-atomic approach

We also evaluated the kind of information yielded by the pseudo-atomic representation. In our simulations the effective diameter of the pseudo-atoms is between 8 Å and 12 Å, which means to model the molecule under study as an entity formed by very simple base

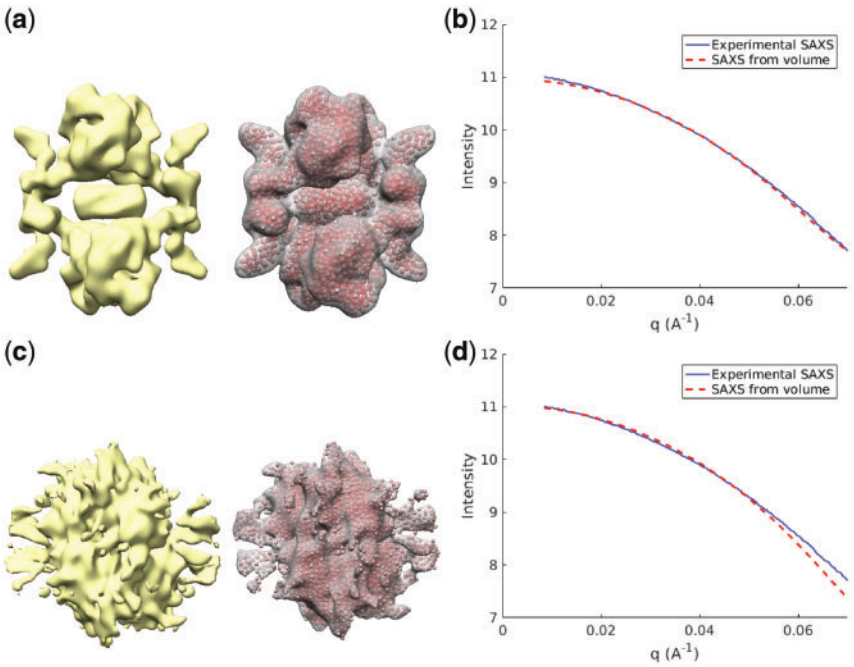


**Fig. 4.** Tyrosine hydroxylase results. First row for a high similarity initial volume. Second row for a low similarity initial volume. (**a**) and (**c**) The initial volume to the left, along with its pseudo-atomic low pass filtered version to the right. (**b**) and (**d**) Comparison between experimental and predicted SAXS curves in semi-log scale in the low frequency region up to 15 Å

functions. In this way, we are including some *a priori* information in our maps, with which we are able to simulate, at some extent, the physical reality.

Figure 6a and c show the radial average of the Fourier transform for the non-filtered, filtered and pseudo-atomic versions of a correct initial map and an incorrect one, respectively, for the lysozyme, as a representative example. As can be seen from these curves, despite the fact that the EM initial map is filtered to a relatively low frequency (15 Å in our examples), filling this structure with pseudo-atoms provides us to, at some extent, high-resolution information that can be subsequently exploited by the SAXS curve simulator.

In parts (b) and (d) of Figure 6, the SAXS curve obtained from the pseudo-atomic map and the experimental one can be seen in the whole range. Measuring the fitting between both curves using the $\chi^2$ value, we obtain 0.79 and 4.84 for the correct initial map and the incorrect one, respectively, taking into account the frequencies up to 6 Å, as above this frequency the experimental noise dominates that curve. Thanks to the high frequency information yielded by the pseudo-atoms, the match between the predicted and the experimental SAXS curves can be measured beyond the low frequency cutoff applied to the candidate initial map.

We must highlight that same trend is obtained with the remaining analyzed molecules: $\chi^2$ values 16.27 and 2677.24 for a correct initial map and an incorrect one for catalase and 1.58 and 1.86 for tyrosine hydroxylase, all of them calculated in the region of the curves without experimental noise. Moreover, these results confirm the previous ones calculated in the regions below the cutoff frequency of the low pass filter.

**Table 1.** $\chi^2$ results for the three evaluated structures

| | Correct volume | Incorrect volume |
|---|---|---|
| Lysozyme | 0.58 | 1.07 |
| Catalase | 0.07 | 0.13 |
| Tyrosine hydroxylase | 0.03 | 0.23 |

## 4 Conclusions

In this work, we have proposed an approach to automatically select the best initial 3D maps, helping the following refinement steps to build a higher accuracy 3D model of a biological structure in SPA. Our proposal relies on the SAXS curves to recognize, among all the available 3D initial maps, the best ones. When the two curves overlap (the $\chi^2$ is low) the corresponding map is more likely to be a good representative of the original structure; otherwise, the map will more likely represent a local minimum of the 3D reconstruction landscape. We carried out several experiments supporting the previous statement, with three different biological structures and using several methods to generate the initial 3D maps. The method proposed here can be also useful with more challenging structures, as flexible proteins or multi-subunit complexes. In the case of flexibility, we are using low resolution initial maps to calculate the estimated SAXS curves; so, our method must be able to give valuable results also in this situation. With multi-subunits complexes, whenever the SAXS curves to compare come from SAXS and Cryo EM experiments with the same subunits, the comparison will be fair and useful. Moreover, we are providing a sort of the initial maps, so if any of them presents any problem, i.e. if it is very far from the experimental SAXS curve, we can easily discard it for the following refinement steps and select just the best ones. Moreover, we have compared our method with DARA server obtaining more accurate results, which shows the value of our proposal for combining SAXS and EM information to validate the initial map estimations coming from 2D images.

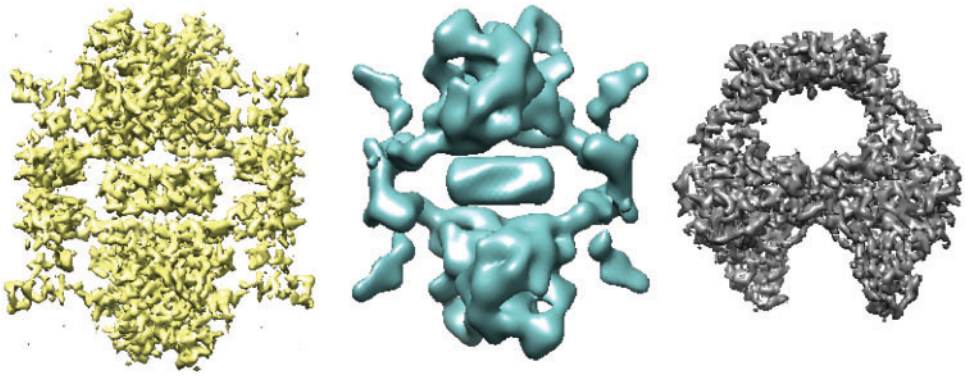

**Fig. 5.** Comparison between the original density map volume obtained from the PDB (left part of the figure), highest similarity volume build by the workflow proposed in this paper (central part of the figure) and highest similarity volume given by DARA server (right part of the figure) for the tyrosine hydroxylase

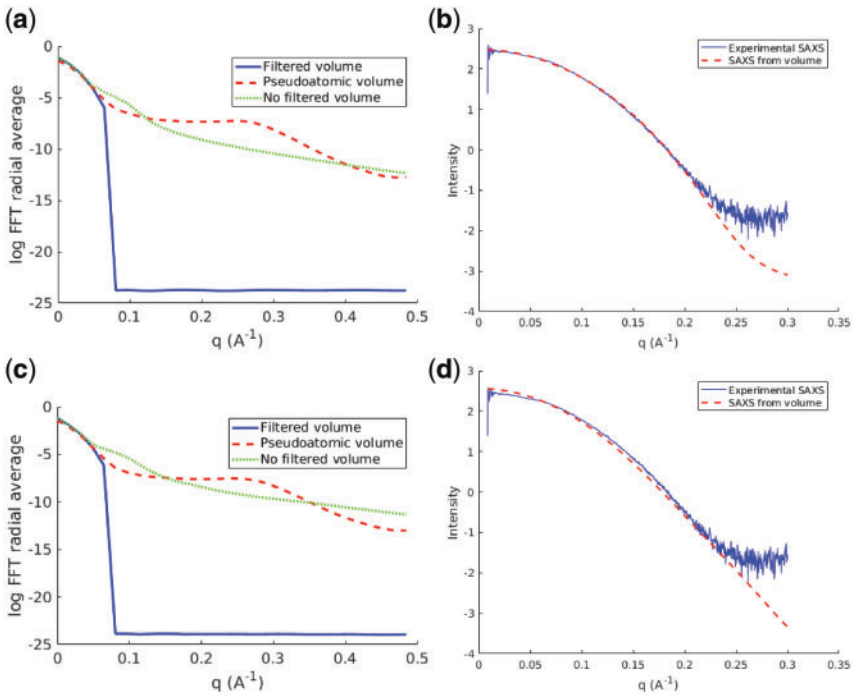


**Fig. 6.** Lysozyme pseudo-atomic evaluation. First row for a high similarity initial volume. Second row for a low similarity initial volume. (**a**) and (**c**) Radial average of the Fourier transform for the non-filtered, filtered and the pseudo-atomic maps. (**b**) and (**d**) Comparison between experimental and predicted SAXS curves in semi-log scale up to the high frequency region

## Acknowledgements

The authors would like to thank A. Martínez for useful discussions about the tyrosine hydroxylase.

## Funding

This work was supported by the Spanish Ministry of Economy and Competitiveness through Grants BIO2016-76400-R(AEI/FEDER, UE), Comunidad Autonoma de Madrid through Grant: S2017/BMD-3817, Instituto de Salud Carlos III, PT13/0001/0009, PT17/0009/0010 and European Union (EU) and Horizon 2020 through Grants: Elixir—EXCELERATE (INFRADEV-3-2015, Proposal: 676559), iNEXT (INFRAIA-1-2014-2015, Proposal: 653706) and INSTRUCT—ULTRA (INFRADEV-03-2016-2017, Proposal: 731005).

*Conflict of Interest*: none declared.

## References

Bezem,M.T. *et al.* (2016) Stable preparations of tyrosine hydroxylase provide the solution structure of the full-length enzyme. *Sci. Rep.*, **6**, 30390.

Bilbao-Castro,J.R. *et al.* (2004) Phan3D: design of biological phantoms in 3D electron microscopy. *Bioinformatics*, **20**, 3286–3288.

Cantele,F. *et al.* (2003) The variance of icosahedral virus models is a key indicator in the structure determination: a model-free reconstruction of viruses, suitable for refractory particles. *J. Struct. Biol.*, **141**, 84–92.

de la Rosa-Trevín,J.M. *et al.* (2016) Scipion: a software framework toward integration, reproducibility and validation in 3d electron microscopy. *J. Struct. Biol.*, **195**, 93–99.

Elmlund,D. and Elmlund,H. (2012) Simple: software for *ab initio* reconstruction of heterogeneous single-particles. *J. Struct. Biol.*, **180**, 420–427.

Franke,D. *et al.* (2017) *ATSAS 2.8*: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions. *J. Appl. Crystal.*, **50**, 1212–1225.

Harauz,G. and Van Heel,M. (1985) *Direct 3D Reconstruction from Projections with Initially Unknown Angles*. Springer, Berlin, Heidelberg, pp. 649–653.

Jonić,S. and Sorzano,C.O.S. (2016) Coarse-graining of volumes for modeling of structure and dynamics in electron microscopy: algorithm to automatically control accuracy of approximation. *IEEE J. Sel. Topics Signal Process.*, **10**, 161–173.

Kikhney,A.G. *et al.* (2016) DARA: a web server for rapid search of structural neighbours using solution small angle X-ray scattering data. *Bioinformatics*, **32**, 616–618.

Koch,M.H.J. *et al.* (2003) Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. *Q. Rev. Biophys.*, **36**, 147–227.

Ludtke,S.J. *et al.* (2004) Seeing GroEL at 6Å resolution by single particle electron cryomicroscopy. *Structure*, **12**, 1129–1136.

Nogales,E. (2016) The development of cryo-EM into a mainstream structural biology technique. *Nat. Methods*, **13**, 24–27.

Shkolnisky,Y. and Singer,A. (2012) Viewing direction estimation in cryo-EM using synchronization. *SIAM J. Imaging Sci.*, **5**, 1088–1110.

Sorzano,C. *et al.* (2018) Swarm optimization as a consensus technique for Electron Microscopy Initial Volume. *Appl. Anal. Optim.*, **2**, 299–313.

Sorzano,C.O.S. *et al.* (2006) Optimization problems in electron microscopy of single particles. *Ann. Oper. Res.*, **148**, 133–165.

Sorzano,C.O.S. *et al.* (2015a) Cryo-EM and the elucidation of new macromolecular structures: random conical tilt revisited. *Sci. Rep.*, **5**, 14290.

Sorzano,C.O.S. *et al.* (2015b) A statistical approach to the initial volume problem in single particle analysis by electron microscopy. *J. Struct. Biol.*, **189**, 213–219.

Svergun,D. *et al.* (1995) *CRYSOL*—a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *J. Appl. Crystal.*, **28**, 768–773.

Vargas,J. *et al.* (2014) Efficient initial volume determination from electron microscopy images of single particles. *Bioinformatics*, **30**, 2891–2898.

Voss,N.R. *et al.* (2010) A toolbox for ab initio 3-D reconstructions in single-particle electron microscopy. *J. Struct. Biol.*, **169**, 389–398.