CESPED: A benchmark for supervised particle pose estimation in cryo-EM

Ruben Sanchez-Garcia,[1,2,*] Michael Saur,[2,†] Javier Vargas,[3,‡] Carl Poelking,[2,§] and Charlotte M. Deane[1,‖]
[1]*Department of Statistics, University of Oxford, Oxford OX1 3LB, United Kingdom*
[2]*Astex Pharmaceuticals, Cambridge CB4 0QA, United Kingdom*
[3]*Departamento de Optica, Universidad Complutense de Madrid, Madrid 28040, Spain*

(Received 29 January 2024; accepted 3 May 2024; published 4 June 2024)

Cryo-EM is a powerful tool for understanding macromolecular structures, yet current methods for structure reconstruction are slow and computationally demanding. To accelerate research on pose estimation, we present CESPED, a data set specifically designed for supervised pose estimation in cryo-EM. Alongside CESPED, we provide a PyTorch package to simplify cryo-EM data handling and model evaluation. We evaluate the performance of a baseline model, Image2Sphere, on CESPED, which shows promising results but also highlights the need for further improvements. Additionally, we illustrate the potential of deep learning-based pose estimators to generalize across different samples, suggesting a promising path toward more efficient processing strategies.

DOI: 10.1103/PhysRevResearch.6.023245

I. INTRODUCTION

A. Cryo-EM single-particle analysis

Determining the structure of macromolecules is crucial to deciphering the intricacies of biological processes and the underlying mechanisms of diseases. With the advent of the resolution revolution, cryogenic electron microscopy (cryo-EM) has emerged as a leading technique for elucidating structures [1,2]. This revolution, driven by significant advances in direct electron detectors and image-processing algorithms, has made cryo-EM a routine, often unrivaled, method for many complex samples [3]. Its advantages include, among others, the relative ease of sample preparation compared to other techniques (e.g., x-ray crystallography), the capability to analyze protein complexes previously considered out of reach, and the ability to recover different conformations, offering a dynamic view of molecules in action [4]. The pivotal role of cryo-EM in structural biology was globally recognized in 2017 when the technique was awarded the Nobel Prize in chemistry.

The primary aim of cryo-EM single-particle analysis (SPA) is to reconstruct the three-dimensional (3D) structure of a given macromolecule at near-atomic resolution, ideally better than 3 Å. This process uses electron beams to capture thousands of two-dimensional (2D) images of the macromolecules, which are flash frozen in vitreous ice to preserve their native state without the distortions typical of crystalline ice or other fixation methods [5]. Each image, called a micrograph, can display several hundred snapshots of the macromolecule (referred to as particle images or just particles) in unknown random orientations. If the orientations of these images were known, the reconstruction task would closely resemble the algorithms used in tomography, which reconstruct 3D volumes from 2D projections taken at predetermined angles [6]. However, the unknown orientations of the particles in SPA present a unique challenge not encountered in tomography [7]. Compounding this challenge is the inherently low contrast and extremely poor signal-to-noise ratio (SNR) of the images, a consequence of the delicate biological nature of the samples. Given these challenges, a highly sophisticated image-processing pipeline is essential to accurately resolve the 3D structure of the macromolecule [8,9].

The fundamental principle of image processing in SPA is grounded in the intuitive strategy of employing averaging to mitigate noise. Since images are characterized by a low SNR, averaging multiple images of the same particle, assumed to be identical, can significantly enhance the underlying signal [4]. However, before averaging can be effectively carried out, each particle projection must be aligned to a common orientation. This ensures that the differences observed across the images are solely due to noise, allowing its effective cancellation during the averaging process.

The standard cryo-EM image processing pipeline encompasses several key steps, beginning with various preprocessing operations to correct errors, such as beam-induced movement blur, followed by particle picking, which extracts the individual particle images from the micrographs [8,9]. Subsequent stages include, among others, clustering (commonly referred to as 2D classification in the context of cryo-EM) and particle alignment against references, leading to a cleaner subset of the data and an initial low-resolution 3D volume of the protein. This preparatory work sets the stage for the refinement step, a critical phase where the poses of the particles are precisely estimated, a requirement to achieve the high-resolution volumes needed to reveal atomic-level details.

*ruben.sanchez-garcia@stats.ox.ac.uk
†michael.saur@astx.com
‡jvargas@fis.ucm.es
§carl.poelking@astx.com
‖deane@stats.ox.ac.uk

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

Traditional refinement algorithms perform pose estimation by exhaustive comparison of experimental particle images and simulated projections of 3D volumes that are iteratively improved [10–14]. When sample homogeneity can be assumed, the simplest approach to the pose estimation problem is the projection-matching algorithm [15], which consists of T iterations of two steps: alignment and reconstruction. First, in the alignment phase, the pose $(R, s)_i \in \mathrm{SO}(3) \times \mathbb{R}^2$ of each experimental particle image x_i is set to be the same as the one of the most similar 2D projection of the reference volume V^t at iteration t,

$$(R, s)_i = \arg \min_{(R,s)\in \mathrm{SO}(3)\times\mathbb{R}^2} \|x_i - f_i * P_{(R,s)} V^t\|^2, \qquad (1)$$

where $P_{(R,s)}$ is the projector operator, f_i is the point spread function of the microscope for the ith particle, and $*$ the convolution operator. Then, in the reconstruction phase, a new volume (in reciprocal space) is computed from the estimated poses as

$$\hat{V}^{(t+1)} = \frac{\sum_{i=1}^{N} P^{-1}_{(R,s)_i} \hat{f}_i \hat{x}_i}{\sum_{i=1}^{N} P^{-1}_{(R,s)_i} \hat{f}_i^2 + C_i}, \qquad (2)$$

with $\hat{V}$ being the Fourier transform of the volume V, C_i a constant depending on the SNR, $\hat{f}_i$ the Fourier transform of the point spread function (CTF, contrast transfer function), and N the number of particles. This iterative process continues until convergence. State-of-the-art methods build on this approach, for example, RELION [13] employs a Bayesian probabilistic model with a prior for the map, making it much more robust. CRYOSPARC [12] accelerates Bayesian methods through branch-and-bound search and gradient descent optimization. See Ref. [7] for a review.

Despite the innovations aimed at enhancing efficiency, the refinement process still poses significant computational challenges. The primary factor contributing to these challenges is the large number of image comparisons required for each experimental image. Furthermore, the iterative refinement of the volumes, beginning with an initial low-resolution model and progressively improving it, further increases the computational cost, making the refinement stage the most computationally intensive step in the cryo-EM workflow.

B. Deep learning for pose estimation in real-world objects

Similar to refinement algorithms in cryo-EM, traditional pose estimation techniques for real-world images primarily focus on matching 2D images with 3D objects. The significantly higher SNRs characteristic of real-world images enable the use of more sophisticated and efficient methods beyond simple template matching. Among these, landmark-based registration methods are particularly prevalent. Such methods involve extracting distinctive landmarks through various feature extraction techniques [16,17], followed by a registration process to identify the relative orientation of the landmarks in the image with respect to the reference landmarks [18].

POSENET [19] was a groundbreaking development in this field, leveraging a convolutional neural network (CNN) to directly regress the absolute pose of an object using quaternions and *xyz* shifts. This direct approach contrasts with earlier techniques that relied heavily on feature extraction and landmark identification, allowing for end-to-end pose estimation. Subsequent innovations have built on the foundation laid by POSENET. Improvements in network architectures [20], the introduction of more sophisticated loss functions [21], and the incorporation of multitask learning [22] have contributed to significant improvements in pose estimation performance.

Addressing the inherent challenges of symmetry and occlusion in pose estimation has also seen considerable progress through deep learning. Strategies have evolved from breaking symmetry during the data labeling process [23] to implementing loss functions specifically designed to accommodate known symmetries [22]. Probabilistic models offer alternative approaches that either classify poses within a discretized space or explicitly learn the parameters of probability distributions [24–28]. Due to their probabilistic nature, these models are better suited for challenging data sets with high levels of ambiguity or noise.

C. Deep-learning methods for cryo-EM structure determination or pose estimation

While traditional cryo-EM refinement algorithms tend to be relatively robust and accurate, they are computationally intensive and slow. In an attempt to overcome this, deep learning (DL) alternatives have begun to emerge.

Unsupervised DL methods aim to determine the 3D structure of macromolecules from experimental images alone. Some of them tackle the problem using a distance learning approach in which the angular distance between pairs of images is estimated as a preprocessing step to retrieve their relative poses [29]. Other unsupervised DL methods mirror traditional techniques by maintaining a 3D volume representation to compute 2D projections in a differentiable manner [30]. Unlike traditional refinement methods, which compare each experimental particle against all images in an SO(3) projection gallery with up to millions of members, these methods try to limit the number of comparisons between experimental images and projections. For instance, in CRYOGAN, a 3D volume, randomly initialized, serves as the generator in the generative adversarial network (GAN) framework [31]. This generator produces a set of projections from random orientations that are then fed to a discriminator network along with real experimental images. The objective of the training process is to refine the generator until the discriminator can no longer distinguish between the generated projections and the actual experimental images, effectively capturing the underlying 3D structure present in the experimental data. In some other approaches [32,33], particle images are first processed by an encoder designed to predict particle orientations. Following this prediction, a projection of the representation of the volume corresponding to the inferred orientation is rendered. This projection is then directly compared to the original experimental particle image. A loss function is utilized to concurrently refine both the encoder's parameters and the representation of the volume, improving the accuracy of orientation predictions and the fidelity of the reconstructed volume.

Supervised DL models, on the other hand, are trained using experimental images and some form of (possibly noisy)

labels, such as the poses of prealigned sets of particles. The simplest alternative consists of only an encoder module that predicts the orientation of the particle directly from its image [34,35]. Although supervised approaches offer remarkable efficiency and speed, they require labeled data for training, thus limiting their applicability in *de novo* situations. However, there are use cases where supervised DL methods could offer an advantage. For instance, it should be possible to apply them to on-the-fly pipelines in which a first batch of particles is prealigned before the end of the data stream. This initial alignment could be used to train a supervised model to be applied to subsequent batches of data, inferring their poses in real time. Even more interestingly, a pretrained supervised model could be used to infer poses in different projects, providing the new samples are similar to the training data. This second use case relies on the fact that pose estimation in classical methods is mainly driven by low- to mid-resolution frequencies [36]. As similar proteins have similar low- to mid-resolution frequencies, trained models are expected to generalize to these new samples. In addition, because ligand binding does not generally modify the overall shape of proteins, supervised approaches can be especially valuable in drug discovery, where prealigned data for target proteins is often available.

In the context of cryo-EM, only two supervised methods have been proposed to perform direct pose estimation given prealigned particles. DeepAlign [34], a set of CNNs that perform binary classification over a discretization of S^2, and the approach of Lian *et al.* [35], who implemented a CNN to perform direct regression of quaternions. However, due to its limitations, especially for symmetric data, Lian *et al.* finally adopted a hybrid model with a projector as in some cryo-EM unsupervised estimators. Neither of the two methods has been used in practical scenarios.

Much slower, classical refinement methods still outperform DL pose estimation models in terms of performance and reliability. This gap can be partly attributed to the unique characteristics of cryo-EM data, which differ from the natural images DL architectures were designed for and, importantly, to the lack of a standardized benchmark that would allow for a direct comparison of methods to stimulate progress, much as IMAGENET [37] did for image classification. In this paper, we introduce CESPED (Cryo-EM Supervised Pose Estimation Dataset), a benchmark specifically designed to evaluate supervised pose estimation methods. As the first benchmark dedicated to pose estimation in cryo-EM, CESPED addresses a crucial gap in the array of available data sets, which have, until now, primarily focused on other cryo-EM challenges, such as model building [38] and particle picking [39,40]. CESPED aims to foster advancements in DL methods for particle processing by promoting improvements in supervised pose estimation models, which, due to shared architectural building blocks and data challenges, are likely to benefit methods for related tasks as well.

D. Main contributions

In this paper, we provide an accessible entry point for a wider scientific audience to engage with the challenges of SPA in cryo-EM. Toward this goal:

(1) We compile CESPED, an easy-to-use benchmark specifically designed for supervised pose estimation in cryo-EM.

(2) We implement a PYTORCH-based [41] package to handle cryo-EM particle data and to easily compute cryo-EM quality metrics.

(3) We train and evaluate the Image2Sphere model [42], originally developed for real-world pose estimation, on our benchmark, illustrating the utility of our benchmark and shedding light on the transferability of real-world pose estimation models to the cryo-EM domain.

(4) We present a use case demonstrating that deep learning-based supervised pose estimators have the potential to generalize across related but different samples.

II. METHODS

A. Benchmark compilation and preprocessing

In our effort to build a comprehensive benchmark, our primary goal was to identify a diverse set of EMPIAR entries containing at least 200 000 particles, a number deemed sufficient for effective model training. Due to the limitations of EMPIAR's search functionality and the inconsistencies in data-set annotations, we conducted a manual search for entries exceeding this particle count and containing standard RELION files (.star and .mrcs). Subsequently, we verified the consistency and accuracy of the metadata by running `relion_reconstruct` [13] and visually assessing the resulting volumes. This step was crucial for eliminating a significant number of entries due to metadata issues that either crashed the reconstruction process or led to incorrect volumes. To ensure consistent estimation of particle poses, the data was reprocessed using the RELION version 4 autorefine program [13,43] (see Appendix A for details). Only entries for which the reconstructed volume exhibited resolution values close to those reported in the literature were selected for inclusion in the benchmark. Finally, for consistency, all images were downsampled to 1.5 Å/pixel, with different image dimensions in each entry, as macromolecules vary in size. See Appendix A for a list of the entries and their properties.

The images fed to the deep-learning model were preprocessed on the fly. We performed per-image normalization following the standard cryo-EM procedure, which involves rescaling the intensity so the background (noise) has a mean of 0 and a standard deviation of 1. We also corrected the contrast inversion caused by the defocus via phase flipping [44]. Finally, since the macromolecule typically represents only between 25% to 50% of the whole particle image, the images were cropped so neighboring particles are not included. It is important to note that our benchmark package allows users the flexibility to choose whether or not to apply any of these normalization steps.

The data labels are represented as rotation matrices and then converted into grid indices by finding the closest rotation matrix in the $\mathrm{SO}(3)_{\mathrm{grid}}$. For the cases in which the macromolecule exhibits point symmetry, the labels are expanded as $L_i = \{g_j R_i | g_j \in G\}$, where G is the set of rotation matrices given a point symmetry group (e.g., $C1$), and R_i

the ground-truth rotation matrix. As a result, the labels consist of vectors with $|G|$ nonzero values and $|\mathrm{SO}(3)_{\mathrm{grid}}| - |G|$ zeros.

B. Baseline model

We adapted the state-of-the-art Image2Sphere model [42]. Image2Sphere is a hybrid architecture that uses a ResNet to produce a 2D feature map of the input image, which is then orthographically projected onto a 3D hemisphere and expanded in spherical harmonics. Then, equivariant group convolutions are applied, first with global support on the S^2 sphere, and finally as a refinement step, on SO(3). The output of the model is a probability distribution over a discretized grid of rotation matrices. Other supervised cryo-EM methods for pose estimation were not considered for this paper due to the lack of publicly available code [35] or their GUI requirements [34].

C. Evaluation metrics

The most widely used metric in pose estimation is the mean angular error (MAnE), averaged across all poses:

$$\mathrm{MAnE} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{angError}_i. \tag{3}$$

The angular error (angError) measures the geodesic distance between the predicted and ground truth poses, typically expressed in degrees or radians. This distance can be directly calculated from the rotation matrix of the ground-truth pose trueR_i and the predicted rotation matrix predR_i as

$$\mathrm{angError}_i = \arccos\left(\frac{\mathrm{trace}\left(\mathrm{trueR}_i \cdot \mathrm{predR}_i^T\right) - 1}{2}\right). \tag{4}$$

When evaluating predicted orientations of macromolecules exhibiting point symmetry, it is necessary to adjust the angular error, as several rotation matrices become equivalent. In this context, the angular error is defined as the minimum geodesic distance between the predicted orientation and any orientation equivalent to the ground truth under the molecule's symmetry group,

$$\mathrm{angError}_i = \min_{g_j \in G} \arccos\left(\frac{\mathrm{trace}\left(g_j \cdot \mathrm{trueR}_i \cdot \mathrm{predR}_i^T\right) - 1}{2}\right), \tag{5}$$

with G being the set of rotation matrices given a point symmetry group.

However, due to the uncertainty in the estimated poses [34], we propose additional metrics. The first one is the confidence-weighted mean-angular-error,

$$\mathrm{wMAnE} = \frac{\sum_{i=1}^{N} \mathrm{conf}_i \cdot \mathrm{angError}_i}{\sum_{i=1}^{N} \mathrm{conf}_i}, \tag{6}$$

which weights the $\mathrm{angError_i}$ by $\mathrm{conf_i}$, the confidence in the ground-truth pose, measured as RELION's rlnMaxValueProbDistribution. This confidence estimation is a number between 0 and 1 that measures the probability of the particle having the reported ground-truth orientation according to the RELION model. While wMAnE is still sensitive to ground truth and confidence estimation errors, due to its simplicity, we used it as the criterion for hyperparameter tuning.

The quality of volumes reconstructed from the predicted poses is assessed by comparing them with the ground-truth volumes generated from the original poses (see Appendix H). For this comparison, we employ the real space Pearson's correlation coefficient (PCC) and the Fouriershell correlation (FSC) Resolution as metrics.

The Pearson's correlation coefficient is a value between -1 and 1, where values closer to 1 indicate a higher similarity. It measures the linear correlation between the pixels of the two volumes as follows:

$$\mathrm{PCC}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}, \tag{7}$$

with X_i and Y_i being the pixel i of the two volumes, n the number of voxels, and $\bar{X}$ and $\bar{Y}$ the average value of the volumes.

The FSC quantifies the correlation between two signals at different spatial frequencies. For each frequency k, a value between -1 and 1 (with higher values indicating greater similarity) is computed by comparing the concentric shells in the Fourier transforms of the two volumes corresponding to k:

$$\mathrm{FSC}_k(X,Y) = \frac{\sum_{\mathbf{r}\in\mathrm{shell}(k)} \hat{X}(\mathbf{r}) \cdot \hat{Y}^*(\mathbf{r})}{\sqrt{\left(\sum_{\mathbf{r}\in\mathrm{shell}(k)} |\hat{X}(\mathbf{r})|^2\right)\cdot\left(\sum_{\mathbf{r}\in\mathrm{shell}(k)} F\,|\hat{Y}(\mathbf{r})|^2\right)}}, \tag{8}$$

where $\hat{X}(\mathbf{r})$ and $\hat{Y}(\mathbf{r})$ represent the Fourier transforms of the two volumes at frequency $\mathbf{r}$, $\mathrm{shell}(k)$ is the shell of frequency k, and $\hat{Y}^*(\mathbf{r})$ is the complex conjugate of $\hat{Y}(\mathbf{r})$.

To summarize the FSC curves into a single number, the FSC resolution is computed by selecting a threshold t and identifying the highest frequency k such that $\mathrm{FSC}_k(X,Y) < t$. As thresholds, we employ the commonly used values of 0.5 ($\mathrm{FSCR}_{0.5}$) and 0.143 ($\mathrm{FSCR}_{0.143}$), which correspond to the highest frequency at which the two maps agree with an SNR of 1 and 0.5, respectively [45].

To decouple the different quality levels of the different benchmark entries, we report the differences of the metrics with respect to the ground-truth levels, estimated from the half maps of the ground truth,

$$\Delta\mathrm{PCC} = \mathrm{PCC}(GT_0, GT_1) - \mathrm{PCC}(GT, V) \tag{9}$$

and

$$\Delta\mathrm{FSC} = \mathrm{FSC}(GT_0, GT_1) - \mathrm{FSC}(GT, V), \tag{10}$$

where GT is the ground-truth map, GT_i is the ground truth map reconstructed with the ith half of the data, and V is the 3D volume reconstructed with the predicted poses (see Fig. 1).

D. Training

Each benchmark entry was trained independently with the same hyperparameters (see Appendix B). Due to the uncertainty in the estimated orientations, we employed a weighted cross-entropy loss using the pose reliability estimate of each

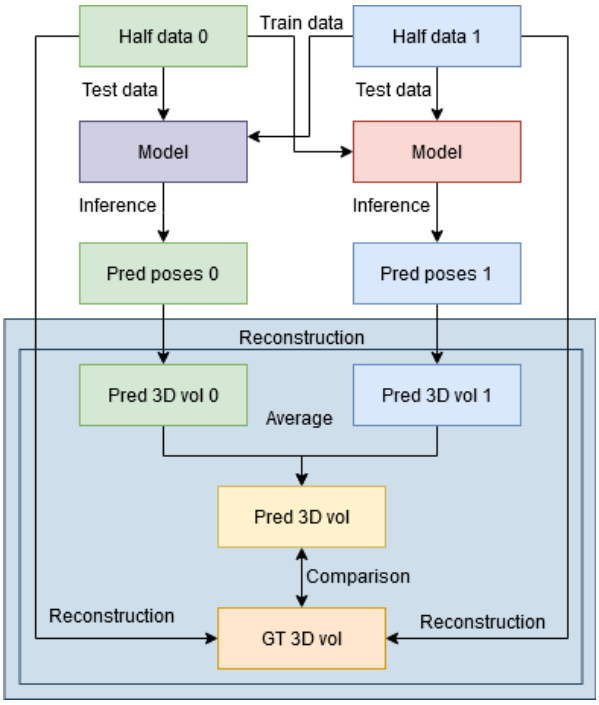

FIG. 1. Evaluation protocol inspired by the cryo-EM gold standard. For each entry, the data set is randomly split into two subsets (half-data 0 and half-data 1) that are processed independently. Then, each half of the data is used to train a different model that will be used to infer the poses of the other half of the data. From the inferred poses, two reconstructed volumes can be obtained, one for each half of the data. The two reconstructed volumes can be combined and compared to the ground-truth volume, which is obtained from the ground-truth poses. The grey box represents the automatic evaluation tool that takes as input the predictions for both data half sets and internally computes the required reconstructed volumes to perform the comparisons.

particle as the per-image weight

$$L = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} -\text{conf}_i \cdot P(R_{c,i})\ln(Q(R_{c,i})), \qquad (11)$$

where $Q(R_{c,i})$ is the predicted probability for the rotation matrix with grid index c and $P(R_{c,i})$ is $1/|G|$ when any of the ground truth matrices is c and zero otherwise.

E. Evaluation protocol

Due to the uncertainty in the ground-truth labels and the fact that what matters to cryo-EM practitioners is the quality of the reconstructed volume, we devised an evaluation protocol inspired by the cryo-EM gold standard [46], which is a per-entry twofold cross-validation strategy in which the poses of each half of the data are independently estimated and used to reconstruct two volumes (half-maps). For benchmarking supervised methods, it involves training an independent model for each half of the data set to infer the poses of the other half of the data set. After that, the final 3D volume is computed by reconstructing the two half maps and averaging them. The final averaged map can then be compared with the ground-truth map obtained from the original poses (see Fig. 1). It is important to note that the FSC resolution values derived from this comparison are analogous to map-to-model FSC resolution estimations and not equivalent to the gold standard half-to-half resolution.

Since Image2Sphere predicts only rotation matrices but not image shifts, when reconstructing the volumes, we employed the ground-truth translations. This could result in an overoptimistic estimation of performance, however, since the effect of the translations is tightly coupled with the accuracy of the angular estimation, this overestimation should be small. We leave for future work the full inference of both the rotations and translations. Finally, to avoid overfitting to the validation set, we performed hyperparameter tuning on only one half of the data using wMAnE as a metric.

III. RESULTS AND DISCUSSION

A. Benchmark, ParticlesDataset class, and evaluation tool

Our benchmark consists of a diverse set of eight macromolecules, with an average number of 300 000 particles including soluble and membrane macromolecules, symmetric and asymmetric complexes, and resolutions ranging from 5 Å to 3.2 Å (see Appendix A). For each particle in the data set, we provide its image and estimated pose together with an estimate of the reliability of the poses. The benchmark can be automatically downloaded from Zenodo [47] using our CESPD Python package.

The package includes a `ParticlesDataset` class, which implements the PyTorch `Dataset` API for seamless integration. It also offers optional yet recommended preprocessing steps commonly adopted in cryo-EM (e.g., image normalization, phase flipping), and specialized data augmentation techniques, like affine transformations that adjust both the image and its corresponding pose (see Appendix B). While the CESPD package was designed with PyTorch in mind, the benchmark is accessible to a broader audience, as the data is stored in standard formats and accompanied by utility programs to assist users of other frameworks in adopting the CESPED benchmark.

Additionally, the package offers an automatic evaluation pipeline that only requires as inputs the predicted poses (grey box in Fig. 1). For ease of use, a Singularity [48] image definition file is included, eliminating the need to install cryo-EM-specific software like RELION. This design enables those without cryo-EM experience to utilize the CESPD benchmark and package as effortlessly as they would with standard data sets such as MNIST. Usage examples can be found in Appendix C.

B. Performance of the baseline model on the benchmark

Table I summarizes the results of the Image2Sphere [42] model on our benchmark, with per-entry results in Appendix D. While the wMAnE is $\sim 24^\circ$, for the best cases, the error is as small as 9°. The ΔPCC for the worse cases is >0.1, highlighting that, for some entries, the reconstructed volumes are far from the ground-truth solution. For a few cases, the results are much better, with $\Delta\text{PCC} < 0.03$. In terms of prediction vs ground truth $\text{FSCR}_{0.5}$, most maps are in the 8–6 Å range, with $\Delta\text{FSCR}_{0.5}$ of 3.6 Å. However, the $\text{FSCR}_{0.143}$ values between

TABLE I. Image2Sphere results on CESPED. MAnE and wMAnE measure angular errors in predicted poses. ΔPCC and ΔFSCR measure the reduction in quality of the predicted volumes compared to the ground truth. The mean and standard deviation (std) of the metrics are computed over the seven benchmark entries.

	MAnE (°)	wMAnE (°)	ΔPCC	$\Delta\text{FSCR}_{0.5}$ (Å)	$\Delta\text{FSCR}_{0.143}$ (Å)
Mean (std)	28.7 (12.7)	23.8 (12.2)	0.059 (0.033)	3.4 (0.6)	1.3 (0.7)

4–5 Å, indicate better correlation at lower signal levels. This visually translates into a relatively well-resolved central part of the map that becomes blurrier away from the center (see Fig. 2 and Appendix E). For the top-performing cases, a simple and fast local refinement of the predicted poses is sufficient to obtain high-resolution reconstructions comparable to ground-truth volumes, at a computational cost threefold less than global refinement (Appendix F). Since the Image2Sphere model inference takes only minutes, far less than the hours needed for traditional refinement, further improvements could reduce computing times by at least one order of magnitude if local refinement is no longer needed (see Appendix G for running times).

Given the inherent difficulties of cryo-EM data, the fact that a generic pose estimation model can produce meaningful results in some examples without major modifications suggests that equivariant architectures can be useful for the cryo-EM data domain.

C. Example of model generalizability across samples

One of the main potential applications of supervised pose estimation models is to infer poses on similar, yet different projects. In this section, we illustrate this use case by using an Image2Sphere model trained on the EMPIAR-10280 data set to predict poses of the same protein under different experimental conditions (EMPIAR-10278 data set).

Figure 3 showcases three reconstructed volumes: (1) EMPIAR-10278 using ground-truth poses (grey); (2) EMPIAR-10278 with poses predicted by the model trained on EMPIAR-10280 (yellow), illustrating the model's generalizability; and (3) EMPIAR-10280 using poses inferred by the model trained on its own data set, serving as a control for model performance. As expected, the EMPIAR-10278 map reconstructed with original poses shows superior quality compared to the others. Similarly, the EMPIAR-10280 map generated from the model trained on EMPIAR-10280 exhibits better quality than the EMPIAR-10278 map inferred using the EMPIAR-10280 model, reflecting the differences between the two data sets despite containing the same protein. Independently of these quality differences, the model's capacity for generalization across data sets is evident through visual inspection of the EMPIAR-10278 inferred map (yellow), as the overall shape of the protein and several key secondary structure elements are clearly recognizable. This suggests that further improvements in the model could lead to the desired goal of training the model once and then inferring the poses of similar data sets at much faster speeds.

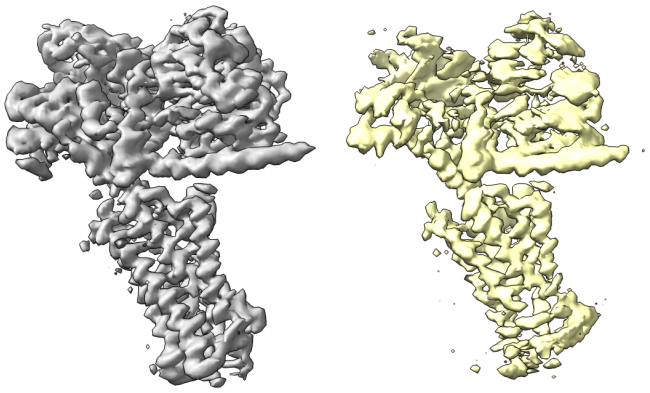

FIG. 2. Ground-truth reconstruction for EMPIAR-10786 (grey) and reconstruction using the angles predicted with the Image2Sphere model (yellow).

D. Challenges and future directions

Cryo-EM particle images are fundamentally different from the kinds of images encountered in other fields. One of the most critical challenges is their poor SNR, which can be as low as 0.01 [7]. While some methods have tried to mitigate this issue by applying filtering techniques [33] or using CNNs with larger kernel sizes [34,49,50], these solutions are not entirely effective.

Symmetry presents another complex facet of cryo-EM data. Exploiting symmetry can drastically reduce the computational requirements for pose estimation, but it can also prevent simple models from learning. The unique combination of rotationally equivariant convolutions with the probabilistic estimation of poses makes the Image2Sphere model an ideal candidate to exploit this feature. However, the hybrid $S^2/\text{SO}(3)$ formalism means that the separation of rotational degrees of freedom from translational in-plane shifts is not easily achieved within this framework. A significant area for future work lies in leveraging rotational equivariance and translational equivariance for the joint estimation of the rotational and translational components of the poses [e.g., SE(3) equivariance].

In this paper, we have considered only the case of homogeneous refinement, which assumes that all particles are projections from a single macromolecule in a unique conformation. However, this is not always the case and our benchmark could potentially be extended to deal with such examples. Models would then need to perform conformation classification alongside pose estimation.

IV. CONCLUSIONS

Pose estimation is one of the most critical steps of the cryo-EM processing pipeline, and while current algorithms are relatively robust and reliable, they are also computationally slow. Deep learning holds the promise of overcoming

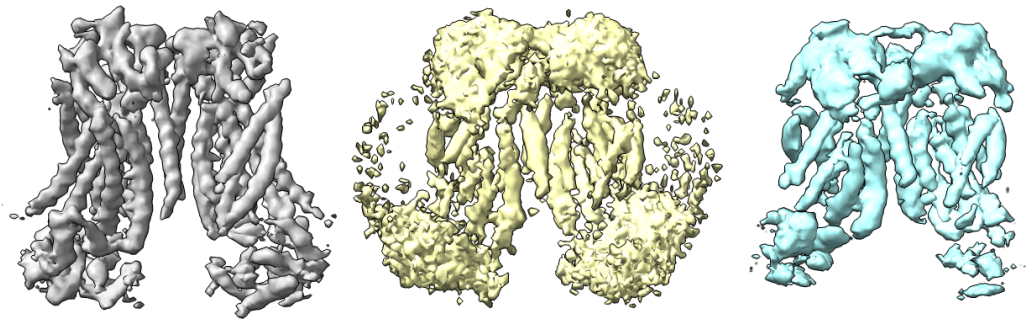

FIG. 3. Generalizability use case. Left and center: Reconstructions for the EMPIAR-10278 data set from the ground truth poses (grey) and the predicted poses (yellow) using a model trained on the EMPIAR-10280 data. Right: Reconstruction for the EMPIAR-10280 data set using a model trained on the EMPIAR-10280 data used as control (cyan). The model used for the yellow and cyan volumes is the same, but the particles fed to the model come from different data sets of the same protein in different conditions. Although of lower quality than the cyan volume, the yellow volume demonstrates similarities to the ground truth map (grey).

these challenges, but achieving this potential hinges on improvements in accuracy and reliability, for which systematic benchmarking is required. In this paper, we introduced a benchmark specifically designed for supervised pose inference of cryo-EM particles, along with a suite of code utilities to assist machine-learning practitioners unfamiliar with cryo-EM. We also present a real-world image pose prediction model applied to our benchmark, demonstrating promising preliminary results on a subset of the data. This preliminary success suggests that addressing cryo-EM-specific challenges, such as high noise levels and label inaccuracies, could lead to even better performance. The improvements in models for this benchmark will not only pave the way for more effective supervised pose prediction models but are also likely to give rise to innovative approaches to closely related challenges like unsupervised pose estimation and heterogeneity analysis. Ultimately, those advancements could serve as a catalyst for even further developments, leading to another paradigm in cryo-EM image processing.

CESPED repository is available at [51]. The repository contains the code and the documentation, as well as the downloading scripts to automatically download the data used in this work. Relion singularity image [48] is also accessible at the repository. The masks used in this work are available at [52].

ACKNOWLEDGMENTS

R.S.-G. is funded by an Astex Pharmaceuticals Sustaining Innovation postdoctoral award. J.V. is financially supported by the Spanish Ministerio de Ciencia e Innovación, Grant No. PID2022-137548OB-I00 funded by MCIN/AEI/10.13039/501100011033/.

APPENDIX A: BENCHMARK COMPOSITION

Table II shows the composition of the CESPED benchmark. Particle poses were estimated using the RELION version 4 autorefine program [13,43]. As a starting model, we used the map obtained with

```
relion_reconstruct --pad 2.0 --ctf --i original_poses.star
--sym $SYMMETRY  --o reconstructd_map.mrc.
```

The mask was created using

```
relion_mask_create --i reconstructed_map.mrc --o mask.mrc --lowpass 15.0
-extend_inimask 3 --width_soft_edge 6 --ini_threshold $THRESHOLD,
```

with `$THRESHOLD` manually selected for each entry.

The autorefine command used was

```
mpirun -np 5 relion_refine_mpi --i original_poses.star --particle_diameter
$DIAMETER --ctf --zero_mask --firstiter_cc --ini_high 40.0
--sym $SYMMETRY --ref reconstructd_map.mrc --norm --scale
--solvent_mask mask.mrc --o outputdir/run --oversampling 1 --flatten_solvent
--solvent_correct_fsc --pad 2 --auto_local_healpix_order 4 --healpix_order 2
--offset_range 5.0 --offset_step 2.0 --auto_refine --split_random_halves
 --low_resol_join_halves 40 --dont_combine_weights_via_disc.
```

The simulated data set was generated with the following command:

```
 relion_project --i original_poses.star --ang original_poses.star
 --ang_simulate original_poses.star --o simulated_dir/simulated
 --simulate --adjust_simulation_SNR 2.0 --ctf.
```

The consensus data set was generated using the compare angles protocol from Scipion Xmipp [34,53], incorporating both our original RELION refinement output and a refinement performed with cisTEM [11]. An angular distance threshold of 5° was employed.

TABLE II. CESPED benchmark entries.

EMPIAR ID	Composition	Symmetry	Image pixels	**FSCR$_{0.143}$** (Å)	Masked**FSCR**$_{0.143}$(Å)	Number of particles
10166	Human 26S proteasome bound to the chemotherapeutic Oprozomib	C1	284	5.0	3.9	238 631
10786	Substance P-neurokinin receptor G protein complexes (SP-NK1R-miniGs399)	C1	184	3.3	3.0[a]	288 659
10280	Calcium-bound TMEM16F in nanodisc with supplement of PIP2	C2	182	3.6	3.0[a]	459 504
11120	M22 bound TSHR Gs 7TM G protein	C1	232	3.4	3.0[a]	244 973
10409	Replicating SARS-CoV-2 polymerase (map 1)	C1	240	3.3	3.0[a]	406 001
10374	Human ABCG2 transporter with inhibitor MZ29 and 5D3-Fab	C2	216	3.7	3.0	323 681
10399	Arabinofuranosyltransferase AftD from mycobacteria	C1	184	3.2	3.1	490 616
10648	PKM2 in complex with compound 5	D2	222	3.7	3.3	234 956
Simulated 10648	Same PKM2 data set as in 10648, but with simulated images	D2	222	3.5	3.4	234 956
Consensus 10648	Same PKM2 data set as in 10648, but with consensus angles	D2	222	3.8	3.4	138 848

[a]Nyquist Frequency at 1.5 Å/pixel; resolution is estimated at the usual threshold 0.143. Reported FSCR$_{0.143}$ values were obtained directly from the relion_refine logs while Masked FSCR$_{0.143}$ values were collected from the relion_postprocess logs.

APPENDIX B: IMAGE2SPHERE AND TRAINING HYPERPARAMETERS

Our Image2Sphere model follows the implementation of Klee *et al.* [42] with the following configuration:

(1) Feature extractor: ResNet152 [54] with default parameters as implemented in torchvision using imageNet weights. The input images are resized to 256 pixels before being fed, giving a feature map of shape 2048 x 8 x 8. Since the input images only contain one channel, but the ResNet expects three channels, two additional channels were added by applying a Gaussian filter with sigma 1 and 2 to the input image.

(2) Image projector to S2: Default orthographic projector with HEALPix [55] grid order 3 ($\sim$7.5°), where only 50% of the grid points are sampled. The feature map is projected from 2048 channels to 512 using a 1 x 1 Conv2d and then converted to spherical harmonics with $l_{\max} = 8$.

(3) S2 convolution: 512 filters with global support on a HEALPix grid of order 3.

(4) SO(3) convolution: 16 filters with local support (max_beta= $\pi/8$, max_gamma=2π, n_alpha=8, n_beta=3).

(5) Probability distribution discretization: HEALPix grid of order 4 ($\sim$3.7°).

Training was conducted using RAdam [56] as the optimizer with an initial learning rate of 10^{-3}. A weight decay of 10^{-5} was employed. The learning rate was halved each time the validation loss stagnated during 10 epochs. The training was stopped when the number of epochs reached 400 or the validation loss did not improve for 12 epochs.

Data augmentation was conducted with the following composed transformations:

(1) Random shift from −5% to 5% with probability 0.5.

(2) Random rotation from −20° to 20° with probability 0.5.

(3) Random 90° rotation with probability 1.

(4) Uniform noise addition with a random scale from 0 to 2 with probability 0.2.

(5) Gaussian noise addition with a random standard deviation from 0 to 0.5 with probability 0.2.

(6) Random zoom-in of size 0% to 5% with probability 0.2.

(7) Random erasing of patches of size 0% to 2% with probability 0.1.

Notice that rotation transformations require adjustments in the ground-truth labels.

APPENDIX C: CESPED PACKAGE USAGE EXAMPLE

Data-set instantiation only requires providing the name of the target (a string like 10280) and the half-set number (0 or 1) (Listing 1). `ParticlesDatasets` can be directly used as data sets in PyTorch `DataLoader`(s).

```python
import torch
from cesped.particlesDataset import ParticlesDataset

listOfEntries = ParticlesDataset.getCESPEDEntries()
targetName, halfset = listOfEntries[0] #We will work with the first example
dataset = ParticlesDataset(targetName, halfset)
dl = DataLoader(dataset, batch_size=32, num_workers=4)
for batch in dl:
    iid, img, (rotMat, xyShiftAngs, confidence), metadata = batch

    #iid is the id of the particle (a string)
    #img is a batch of Bx1xNxN images
    #rotMat is a batch of rotation matrices Bx3x3
    #xyShiftAngs is a batch of image shifts in Angstroms Bx2
    #confidence is a batch of numbers between 0 and 1, Bx1
    #metata is a dict of names:values with particle information

    predRot = model(img)
    loss = loss_function(predRot, rotMat)
    loss.backward()
    optimizer.step()
    optimizer.zero_grad()
```

Listing 1: Example of how to load and use a CESPED benchmark entry in a training loop.

`ParticlesDataset` objects can also be used to update the metadata with the predicted poses and to save the results in RELION star format, commonly used in cryo-EM software (Listing 2).

```python
for iid, pred_rotmats, maxprob in predictions:
    #iid is the list of ids of the particles (string)
    #pred_rotmats is a batch of predicted rotation matrices Bx3x3
    #maxprob is a batch of numbers, between 0 and 1, Bx1
    #that indicates the confidence in the prediction (e.g., softmax values)
    n_preds = pred_rotmats.shape[0]
    dataset.updateMd(ids=iid, angles=pred_rotmats,
                     shifts=torch.zeros(n_preds, 2),
                     confidence=maxprob,
                     angles_format="rotmat")
dataset.saveMd("predictions.star") #Save dataset as an starfile
```

Listing 2: Example of how to save predictions for usage in CRYO-EM packages and evaluation.

Once the predictions are computed for the two halves of the benchmark entry, evaluation can be automatically computed by providing the starfiles of both predictions via a command line tool (Listing 3) or a function. While you can use your local installation of RELION, we also provide a singularity definition file so you do not need to manually install it. See instructions at [57],

```
python -m cesped.evaluateEntry  --predictionType SO3 --targetName 10280
--half0PredsFname particles_preds_0.star
--half1PredsFname particles_preds_1.star
--n_cpus 12 --outdir evaluation/
```

Listing 3: Evaluation script execution.

```
from cesped.evaluateEntry import evaluate
evaluation_metrics = evaluate(targetName="10280",
            half0PredsFname="particles_preds_0.star",
            half1PredsFname="particles_preds_1.star",
            predictionType="SO3", #Literal["S2", "SO3", "SO3xR2"],
            usePredConfidence=True,
            n_cpus=4,
            outdir="output/directory")
```

Listing 4: Evaluation function example.

APPENDIX D: IMAGE2SPHERE PER-ENTRY RESULTS

This section contains per-entry statistics for the Image2Sphere model predictions using the evaluation protocol proposed in the main text (Table III). The last two rows correspond to different versions of the 10648 entry and have not been included in Table I. In addition to angular error measurements, the other metrics compare the ground-truth map (GT) against the map reconstructed from the predicted poses (V), namely, PCC(GT, V) and $\mathrm{FSCR_t}$(GT, V), where t denotes the threshold 0.5 or 0.143, where reported. GT is obtained by employing `relion_reconstruct` on the ground truth poses (that were estimated with `relion_refine --auto_refine`). The reconstructed map V is generated with `relion_reconstruct` from the predicted poses.

In addition, we also report half-to-half map metrics, which are commonly employed in traditional cryo-EM algorithms and in unsupervised DL methods and can be used to compare them to Supervised DL methods. In particular, we compute PCC($\mathrm{GT_0}$, $\mathrm{GT_1}$), PCC(V_0, V_1), $\mathrm{FSCR_t}$($\mathrm{GT_0}$, $\mathrm{GT_1}$), and $\mathrm{FSCR_t}$(V_0, V_1), where 0 and 1 denote the data-set half. Thus, V_0 is the map reconstructed from the predicted poses of the half data-set 0 using a model trained on data-set 1. $\mathrm{GT_0}$ is obtained as GT, but using only the ground truth poses of the half data-set 0.

APPENDIX E: RECONSTRUCTED VOLUMES

This Appendix shows the volumes reconstructed for some of the best performing examples of the Image2Sphere model on our benchmark (Figs. 4 and 5). As is shown in all the cases, the quality of the central region of the protein is quite close to the one of the ground truth. However, the density for the regions that are at the edges of the macromolecule is much worse. This is in line to what could be expected if there were some degree of inaccuracy in the angular estimation, as the magnitude of the errors in the volume is proportional to both the angular error and the radius of the macromolecule.

APPENDIX F: LOCALLY REFINED SOLUTION

In this Appendix, we illustrate the usefulness of our approach by showing the effect of classical local refinement on the Image2Sphere results for benchmark entry 10374. In this case, the Image2Sphere model predicted poses with a wMAnE of 24.8° that lead to a reconstructed map with $\mathrm{FSCR_{0.143}}$(V_0, V_1) of 3.7 Å. When the predicted poses are used as priors for a local refinement in RELION with `--sigma_angle 2.0`, the refined map achieved a PCC(GT, V) of 0.997 compared to the original 0.969, showing that the refined map is much more similar to the ground

TABLE III. Per-entry CESPED benchmark results using an Image2Sphere model. MAnE: mean angular error; wMAnE: weighted mean angular error; PCC($\mathrm{V_0}$, $\mathrm{V_1}$): reconstructed half-to-half Pearsons's correlation coefficient; PCC(GT, V): reconstructed to ground truth Pearson's correlation Ccoefficient; $\mathrm{FSCR_{0.143}}$($\mathrm{V_0}$, $\mathrm{V_1}$): reconstructed half-to-half FSC resolution at threshold 0.143 and $\mathrm{FSCR_{0.5}}$($\mathrm{V_0}$, $\mathrm{V_1}$) at threshold 0.5; $\mathrm{FSCR_{0.143}}$(GT, V): reconstructed to ground truth resolution at threshold 0.143, and $\mathrm{FSCR_{0.5}}$(GT, V) at threshold 0.5; $\mathrm{FSCR_{0.143}}$($\mathrm{GT_0}$, $\mathrm{GT_1}$): ground-truth half-to-half FSC resolution at threshold 0.143 and $\mathrm{FSCR_{0.5}}$($\mathrm{GT_0}$, $\mathrm{GT_1}$) at threshold 0.5. PCC($\mathrm{GT_0}$, $\mathrm{GT_1}$): Ground-truth half-to-half Pearsons's correlation coefficient. All reported resolutions were obtained using manually computed masks that are available at Ref. [52].

EMPIAR ID	MAnE (°)	wMAnE (°)	PCC ($\mathrm{V_0}$, $\mathrm{V_1}$)	PCC (GT, V)	$\mathrm{FSCR_{0.143}}$ ($\mathrm{V_0}$, $\mathrm{V_1}$) (Å)	$\mathrm{FSCR_{0.5}}$ ($\mathrm{V_0}$, $\mathrm{V_1}$) (Å)	$\mathrm{FSCR_{0.143}}$ (GT, V) (Å)	$\mathrm{FSCR_{0.5}}$ (GT, V) (Å)	$\mathrm{FSCR_{0.143}}$ ($\mathrm{GT_0}$, $\mathrm{GT_1}$) (Å)	$\mathrm{FSCR_{0.5}}$ ($\mathrm{GT_0}$, $\mathrm{GT_1}$) (Å)	PCC ($\mathrm{GT_0}$, $\mathrm{GT_1}$)
10166	15.7	9.1	0.986	0.974	5.1	6.8	6.2	8.1	4.4	4.8	0.992
10786	32.6	29.5	0.957	0.925	3.8	4.3	3.4	7.6	3.1	3.5	0.974
10280	17.8	14.9	0.981	0.957	3.9	4.4	4.3	7.0	3.4	3.8	0.991
11120	44.7	41.1	0.989	0.863	4.1	4.6	6.0	8.3	3.2	3.7	0.965
10409	45.3	39.2	0.960	0.884	3.5	4.0	4.0	8.3	3.0	3.3	0.988
10374	35.0	24.8	0.991	0.969	3.7	4.1	4.1	6.5	3.0	3.5	0.996
10399	25.5	21.6	0.992	0.917	3.7	4.1	4.0	6.1	3.1	3.4	0.996
10648	13.3	10.6	0.982	0.934	3.8	4.1	4.3	6.5	3.4	3.6	0.994
Simulated 10648	6.0	NA	0.996	0.935	4.5	4.6	4.6	4.8	3.5	4.6	0.998
Consensus 10648	8.3	8.1	0.971	0.893	3.8	4.1	4.2	6.8	3.4	3.6	0.986

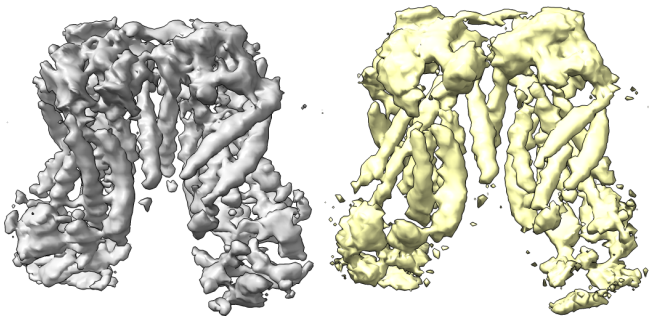

FIG. 4. Ground-truth reconstruction for EMPIAR-10280 (grey) and reconstruction using the angles predicted with the Image2Sphere model (yellow).

truth map. Indeed, as can be appreciated in Fig. 6, after the local refinement, not only the quality of the core of the protein is comparable to the quality of the ground truth, but also the quality of the distant parts of the maps is much better, almost as good as in the ground truth. Equally important, since we are limiting the angular search to the neighborhood around the predicted poses ($\pm$ 6°), the number of image comparisons carried out by RELION is much smaller, resulting in a threefold speedup in computational time, even when including the time required for pose inference using the Image2Sphere model.

APPENDIX G: RUNNING TIMES

Table IV collects the running times of RELION autorefine and the Image2Sphere inference using the same hardware configuration (4 Nvidia A100 cards and 32 CPU cores).

APPENDIX H: IMPACT OF PARTICLE MISALIGNMENT IN MAP QUALITY ESTIMATION

To study the sensitivity to angular inaccuracy of the map quality estimations used in this paper, namely, the FSC resolution and PCC, we ran two experiments. In the first experiment, we added uniform random noise within the ranges of $\pm$ 1°, 3°, and 5° to the Euler angles of each particle (Table V and Figs. 7 and 8, right columns). In the second experiment, we

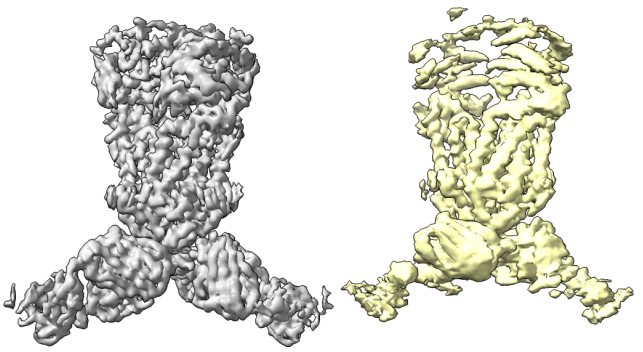

FIG. 5. Ground-truth reconstruction for EMPIAR-10374 (grey) and reconstruction using the angles predicted with the Image2Sphere model (yellow).

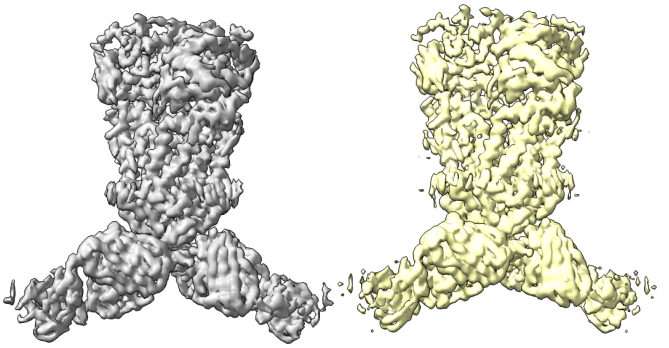

FIG. 6. Ground truth reconstruction for EMPIAR-10374 (grey) and reconstruction using the angles predicted with the Image2Sphere model and locally refined using RELION with priors (yellow).

randomized the Euler angles of 10%, 20%, and 30% of the particles for each entry in the benchmark (Table VI and Figs. 7 and 8, left column). In the absence of symmetry, the expected angular error (geodesic distance) for randomized angles is approximately 126.9°, whereas for the uniform random noise, the expected angular error is of 1.0°, 2.9°, and 4.8°, respectively (as estimated through simulation).

Table V and the right column of Figs. 7 and 8 illustrate a clear trend in which increasing angular errors lead to a reduction in the FSC resolution and PCC. Since in this experiment we corrupted the alignment of all particles, this underscores that map global quality measurements are effective proxies for estimating overall mean angular accuracy.

Table VI and the left column of Figs. 7 and 8 show that as the fraction of misaligned particles increases, both the resolution and the correlation of the maps decreases as well. While it is true that the effect of this type of corruption is smaller than when the angles of all particles are perturbed, it remains noticeable. In most cases, the FSC resolution at threshold 0.5 is clearly different, even when as little as 10% of the particles are perturbed. Given that the number of misaligned particles in refined maps using methods such as RELION is quite large, with some cases reporting up to 60% misalignment levels,[2] the sensitivity of the FSC resolution should be enough to compare the accuracy of different methods. A similar trend is observed in the PCC values, which steadily decline as the fraction of misaligned particles increases.

These two experiments confirm that it is possible to distinguish between different levels of alignment corruption using map quality measurements; hence, they serve as sensible proxies for assessing angular alignment accuracy. However,

TABLE IV. Running time for RELION Autorefine and the Image2Sphere model on CESPED benchmark entries.

EMPIAR ID	RELION (min)	Image2Sphere (min)
10166	521	22
10786	227	12
10280	192	8
11120	102	3
10648	91	5
10409	190	5
10374	133	4

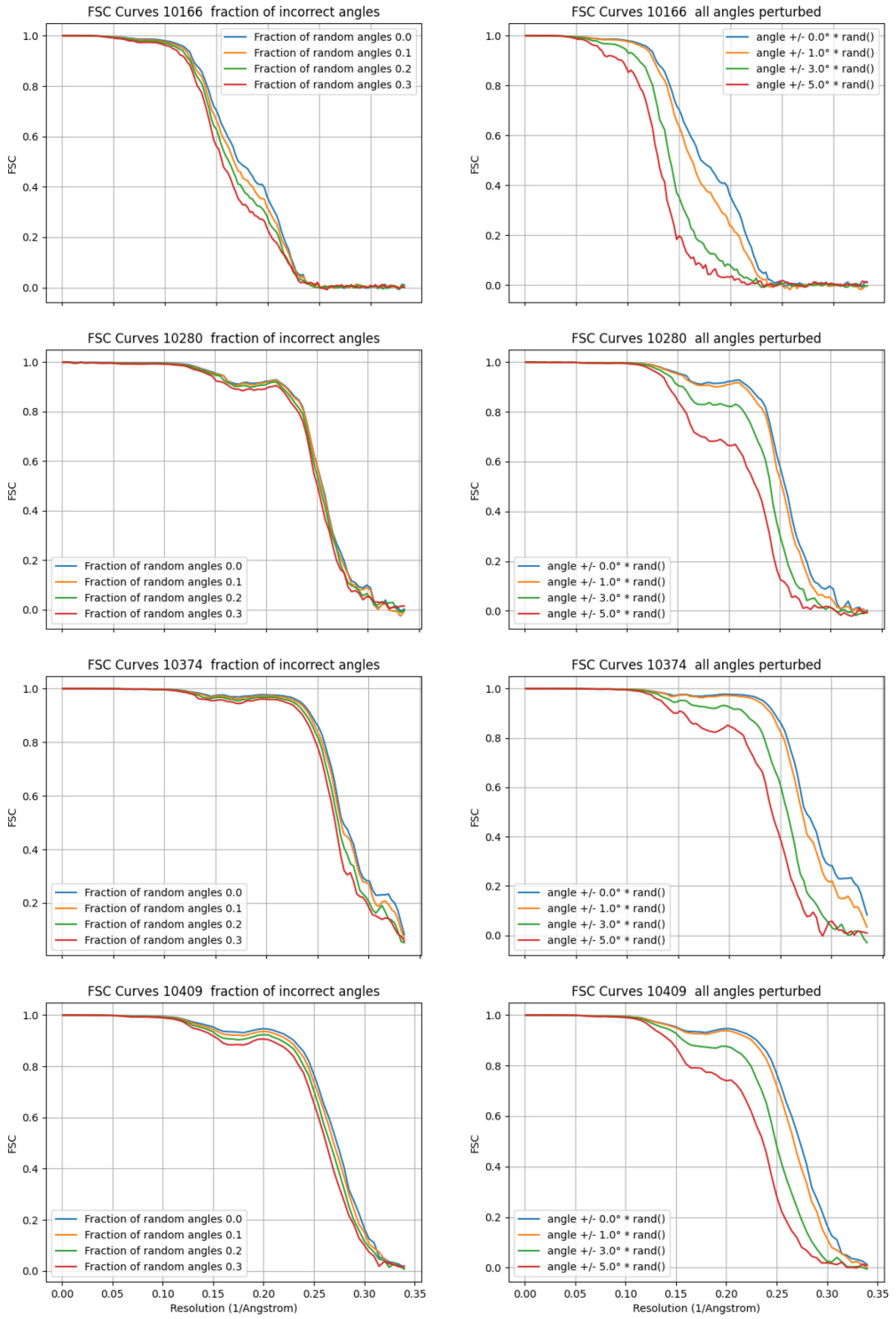

FIG. 7. Effect of particle misalignment in FSC resolution for entries 10166, 10280, 10374, and 10409. Left: FSC curves at different amounts of incorrectly aligned particles: 0% (blue), 10% (orange), 20% (green), and 30% (red) of the particles were assigned random angles. Right: FSC curves at different levels of induced angular inaccuracy. Each particle alignment was perturbed using uniform random noise of $\pm 0^\circ$ (blue), 1° (orange), 3° (green), and 5° (red).

these measurements are not directly comparable across different samples; thus, comparisons are only valid when examining different alignment results for the same sample, as we do in this benchmark.

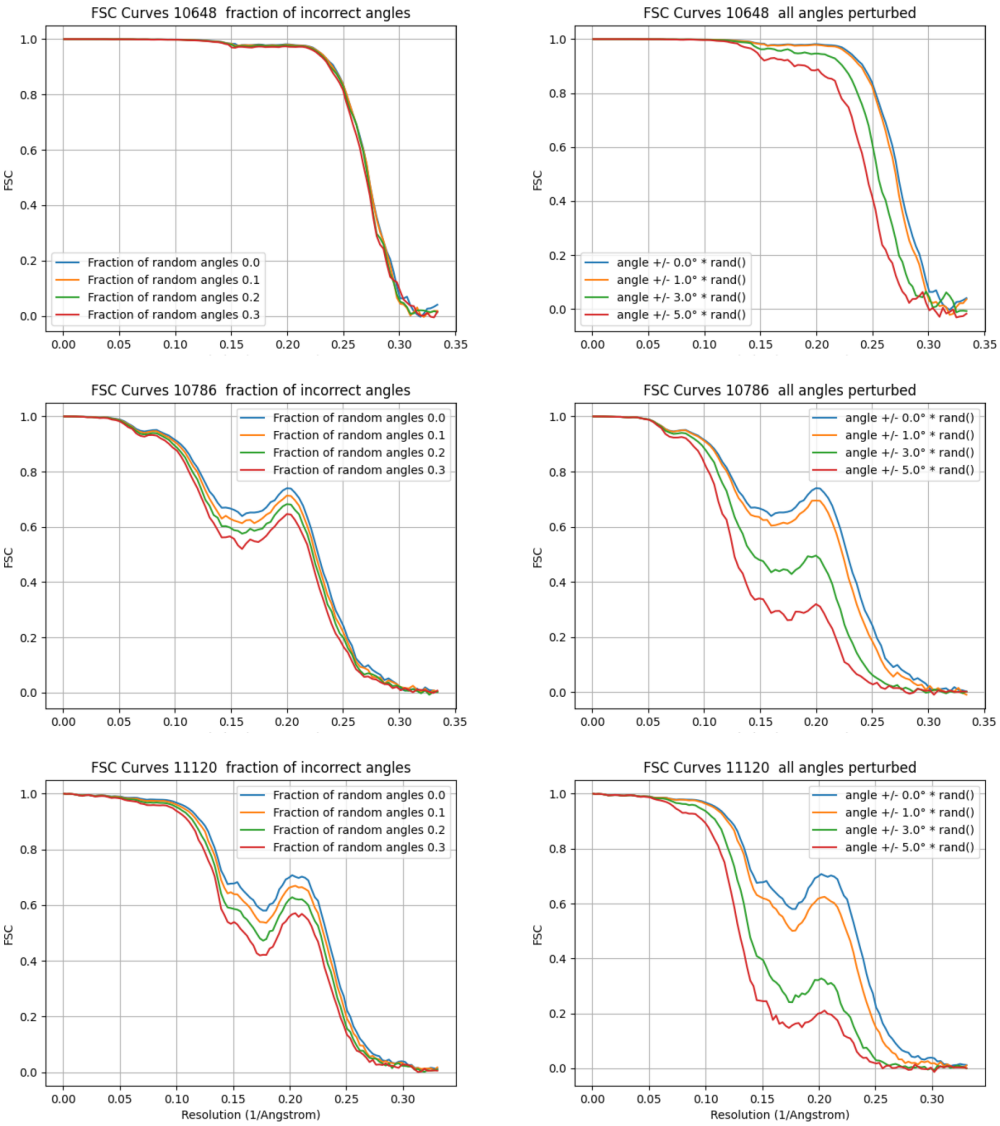

FIG. 8. Effect of particle misalignment in FSC resolution for entries 10648, 10786, and 11120. Left: FSC curves at different amounts of incorrectly aligned particles: 0% (blue), 10% (orange), 20% (green), and 30% (red) of the particles were assigned random angles. Right: FSC curves at different levels of induced angular inaccuracy. Each particle alignment was perturbed using uniform random noise of $\pm 0^\circ$ (blue), 1° (orange), 3° (green), and 5° (red).

TABLE V. Volume Pearson's correlation coefficients (PCCs) at different levels of induced angular inaccuracy. For each particle in each entry, uniform random noise of $\pm 0^\circ$, 1°, 3°, and 5° is added to all the components of the Euler angles. PCC values are reported without using a mask, whereas Masked PCC values are computed using the mask described in the text.

	PCC				Masked PCC			
Entry	0°	1°	3°	5°	0°	1°	3°	5°
10166	0.981	0.950	0.921	0.894	0.995	0.989	0.968	0.941
10280	0.978	0.946	0.911	0.887	0.995	0.991	0.972	0.949
10374	0.980	0.950	0.925	0.908	0.998	0.996	0.980	0.962
10409	0.960	0.914	0.833	0.781	0.976	0.951	0.886	0.835
10648	0.966	0.916	0.868	0.821	0.997	0.995	0.969	0.917
10786	0.931	0.840	0.749	0.700	0.938	0.858	0.772	0.722
11120	0.852	0.571	0.433	0.391	0.905	0.736	0.606	0.555

TABLE VI. Volume Pearson's correlation coefficients (PCCs) at different amounts of incorrectly aligned particles. For each entry, 0%, 10%, 20%, and 30% of the particles were assigned random angles. PCC values are reported without using a mask, whereas masked PCC values are computed using the mask described in the text.

	PCC				Masked PCC			
Entry	0%	10%	20%	30%	0%	10%	20%	30%
10166	0.981	0.971	0.956	0.935	0.995	0.992	0.985	0.975
10280	0.978	0.964	0.946	0.926	0.995	0.988	0.975	0.958
10374	0.980	0.971	0.959	0.945	0.998	0.995	0.987	0.976
10409	0.960	0.944	0.924	0.899	0.976	0.965	0.949	0.928
10648	0.966	0.941	0.909	0.875	0.997	0.988	0.970	0.949
10786	0.931	0.904	0.874	0.838	0.938	0.914	0.886	0.853
11120	0.852	0.805	0.753	0.699	0.905	0.872	0.834	0.792

[1] E. Nogales, The development of cryo-EM into a mainstream structural biology technique, Nat. Methods **13**, 24 (2015).

[2] E. Callaway, Revolutionary cryo-EM is taking over structural biology, Nature (London) **578**, 201 (2020).

[3] E. H. Egelman, The current revolution in Cryo-EM, Biophys. J. **110**, 1008 (2016).

[4] E. Nogales and S. H. W. Scheres, Cryo-EM: A unique tool for the visualization of macromolecular complexity, Mol. Cell **58**, 677 (2015).

[5] L. A. Passmore and C. J. Russo, Specimen preparation for high-resolution cryo-EM, Methods Enzymol. **579**, 51 (2016).

[6] G. Harauz and M. van Heel, Exact filters for general geometry three dimensional reconstruction, Optik 73 (1986).

[7] T. Bendory, A. Bartesaghi, and A. Singer, Single-particle cryo-electron microscopy: Mathematical theory, computational challenges, and opportunities, IEEE Signal Process. Mag. **37**, 58 (2020).

[8] D. Maluenda, T. Majtner, P. Horvath, J. L. L. Vilas, A. Jiménez-Moreno, J. Mota, E. Ramírez-Aportela, R. Sánchez-García, P. Conesa, L. Del Caño, Y. Rancel, Y. Fonseca, M. Martínez, G. Sharov, C. A. A. García, D. Strelak, R. Melero, R. Marabini, J. M. M. Carazo, and C. O. S. O. S. Sorzano, Flexible workflows for on-the-fly electron-microscopy single-particle image processing using Scipion, Acta Crystallogr. Sect. D **75**, 882 (2019).

[9] D. Tegunov and P. Cramer, Real-time cryo-electron microscopy data preprocessing with Warp, Nat. Methods **16**, 1146 (2019).

[10] J. M. De la Rosa-Trevín, J. Otón, R. Marabini, A. Zaldívar, J. Vargas, J. M. Carazo, and C. O. S. Sorzano, Xmipp 3.0: An improved software suite for image processing in electron microscopy, J. Struct. Biol. **184**, 321 (2013).

[11] T. Grant, A. Rohou, and N. Grigorieff, *cis*TEM, user-friendly software for single-particle image processing, eLife **7**, e35383 (2018).

[12] A. Punjani, J. L. Rubinstein, D. J. Fleet, and M. A. Brubaker, cryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination, Nat. Methods **14**, 290 (2017).

[13] S. H. W. Scheres, RELION: Implementation of a Bayesian approach to cryo-EM structure determination, J. Struct. Biol. **180**, 519 (2012).

[14] G. Tang, L. Peng, P. R. Baldwin, D. S. Mann, W. Jiang, I. Rees, and S. J. Ludtke, EMAN2: An extensible image processing suite for electron microscopy, J. Struct. Biol. **157**, 38 (2007).

[15] P. A. Penczek, R. A. Grassucci, and J. Frank, The ribosome at improved resolution: New techniques for merging and orientation refinement in 3D cryo-electron microscopy of biological particles, Ultramicroscopy **53**, 251 (1994).

[16] D. G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision **60**, 91 (2004).

[17] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, Speeded-up robust features (SURF), Comput. Vis. Image Underst. **110**, 346 (2008).

[18] P. J. Besl and N. D. McKay, *Method for registration of 3-D shapes*, in *Sensor Fusion IV: Control Paradigms and Data Structures* (SPIE, Boston, MA, United States, 1992), Vol. 1611, pp. 586–606.

[19] A. Kendall, M. Grimes, and R. Cipolla, PoseNet: A convolutional network for real-time 6-DOF camera relocalization, in *2015 IEEE International Conference on Computer Vision (ICCV)* (Santiago, Chile, 2015), pp. 2938–2946.

[20] I. Melekhov, J. Ylioinas, J. Kannala, and E. Rahtu, Image-based localization using hourglass networks, in *2017 IEEE International Conference on Computer Vision Workshop (ICCVW)* (Venice, Italy, 2017), pp. 870–877.

[21] A. Kendall and R. Cipolla, Geometric loss functions for camera pose regression with deep learning, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI, USA, 2017), pp. 6555–6564.

[22] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes, Robotics: Science and Systems (2018), doi: 10.15607/RSS.2018.XIV.019.

[23] Y. Xiang, R. Mottaghi, and S. Savarese, Beyond PASCAL: A benchmark for 3D object detection in the wild, in *IEEE Winter Conference on Applications of Computer Vision, Steamboat Springs* (CO, USA, 2014), pp. 75–82.

[24] H. Deng, M. Bui, N. Navab, L. Guibas, S. Ilic, and T. Birdal, Deep Bingham Networks: Dealing with uncertainty and ambiguity in pose estimation, Int. J. Comput. Vis. **130**, 1627 (2022).

[25] S. Mahendran, H. Ali, and R. Vidal, A mixed classification-regression framework for 3D pose estimation from 2D images, in *British Machine Vision Conference 2018, BMVC 2018* (2019).

[26] D. Mohlin, G. B. T. Danderyd, and J. Sullivan, Probabilistic orientation estimation with matrix fisher distributions, Adv. Neural Inf. Process. Syst. **33**, 4884 (2020).

[27] K. Murphy, C. Esteves, V. Jampani, S. Ramalingam, and A. Makadia, Implicit-PDF: Non-parametric representation of probability distributions on the rotation manifold, in *Proceedings of the 38th International Conference on Machine Learning*, Proceedings of Machine Learning Research Vol. 139 (PMLR, 2021), pp. 7882–7893.

[28] S. Prokudin, P. Gehler, and S. Nowozin, Pose estimation with uncertainty quantification, in *Proceedings of the European Conference on Computer Vision*, (Springer, Verlag, Munich, Germany, 2018), pp. 534–551.

[29] J. Banjac, L. Donati, and M. Defferrard, Learning to recover orientations from projections in single-particle cryo-EM, arXiv:2104.06237.

[30] C. Donnat, A. Levy, F. Poitevin, E. D. Zhong, and N. Miolane, Deep generative modeling for volume reconstruction in cryo-electron microscopy, J. Struct. Biol. **214**, 107920 (2022).

[31] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2016).

[32] A. Levy, G. Wetzstein, J. Martel, F. Poitevin, and E. D. Zhong, Amortized inference for heterogeneous reconstruction in Cryo-EM, Adv. Neural Inf. Process. Syst. **35**, 13038 (2022).

[33] A. Levy, F. Poitevin, J. Martel, Y. Nashed, A. Peck, N. Miolane, D. Ratner, M. Dunne, and G. Wetzstein, CryoAI: Amortized inference of poses for *ab initio* reconstruction of 3D molecular volumes from real Cryo-EM Images, Comput Vis ECCV **13681**, 540 (2022).

[34] A. Jiménez-Moreno, D. Střelák, J. Filipovič, J. M. Carazo, and C. O. S. Sorzano, DeepAlign, a 3D alignment method based on regionalized deep learning for Cryo-EM, J. Struct. Biol. **213**, 107712 (2021).

[35] R. Lian, B. Huang, L. Wang, Q. Liu, Y. Lin, and H. Ling, End-to-end orientation estimation from 2D cryo-EM images, Acta Crystallogr. Sect. D **78**, 174 (2022).

[36] S. H. W. Scheres and S. Chen, Prevention of overfitting in cryo-EM structure determination, Nat. Methods **9**, 853 (2012).

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, Miami, FL, USA, 2009), pp. 248–255.

[38] N. Giri, L. Wang, and J. Cheng, Cryo2StructData: A large labeled cryo-em density map dataset for AI-based modeling of protein structures, Sci Data **11**, 458 (2024).

[39] R. Gyawali, A. Dhakal, L. Wang, and J. Cheng, CryoVirusDB: A labeled cryo-EM image dataset for AI-driven virus particle picking, bioRxiv (2023), doi: 10.1101/2023.12.25.573312.

[40] A. Dhakal, R. Gyawali, L. Wang, and J. Cheng, A large expert-curated cryo-EM image dataset for machine learning protein particle picking, Sci Data **10**, 392 (2023).

[41] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, PyTorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. **32** (2019).

[42] D. M. Klee, O. Biza, R. Platt, and R. Walters, Image to sphere: Learning equivariant features for efficient pose prediction, *International Conference on Learning Representations* (2023).

[43] D. Kimanius, L. Dong, G. Sharov, T. Nakane, and S. H. W. Scheres, New tools for automated cryo-EM single-particle analysis in RELION-4.0, Biochem. J. **478**, 4169 (2021).

[44] P. A. Penczek, Image restoration in cryo-electron microscopy, Methods Enzymol. **482**, 35 (2010).

[45] P. B. Rosenthal and R. Henderson, Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy, J. Mol. Biol. **333**, 721 (2003).

[46] R. Henderson, A. Sali, M. L. Baker, B. Carragher, B. Devkota, K. H. Downing, E. H. Egelman, Z. Feng, J. Frank, N. Grigorieff, W. Jiang, S. J. Ludtke, O. Medalia, P. A. Penczek, P. B. Rosenthal, M. G. Rossmann, M. F. Schmid, G. F. Schröder, A. C. Steven, D. L. Stokes, J. D. Westbrook, W. Wriggers, H. Yang, J. Young, H. M. Berman, W. Chiu, G. J. Kleywegt, and C. L. Lawson, Outcome of the first electron microscopy validation task force meeting, Structure **20**, 205 (2012).

[47] Zenodo, Zenodo: Research, shared (2013), https://zenodo.org/.

[48] https://zenodo.org/records/4667718.

[49] T. Bepler, A. Morin, M. Rapp, J. Brasch, L. Shapiro, A. J. Noble, and B. Berger, Positive-unlabeled convolutional neural networks for particle picking in cryo-electron micrographs, Nat. Methods **16**, 1153 (2019).

[50] R. Sanchez-Garcia, J. Segura, D. Maluenda, Jose M. Carazo, and C. O. Sorzano Sorzano, Deep Consensus, a deep learning-based approach for particle pruning in cryo-electron microscopy Ruben, IUCrJ **5**, 854 (2018).

[51] https://github.com/rsanchezgarc/cesped.

[52] https://zenodo.org/record/8392782.

[53] C. O. S. Sorzano, A. Jimenez-Moreno, D. Maluenda, M. Martinez, E. Ramirez-Aportela, J. Krieger, R. Melero, A. Cuervo, J. Conesa, J. Filipovic, P. Conesa, L. Del Cano, Y. C. Fonseca, J. Jiménez-De La Morena, P. Losana, R. Sanchez-Garcia, D. Strelak, E. Fernandez-Gimenez, F. P. De Isidro-Gómez, D. Herreros, J. L. Vilas, R. Marabini, and J. M. Carazo, On bias, variance, overfitting, gold standard and consensus in single-particle analysis by cryo-electron microscopy, Acta Crystallogr. Sect. D **78**, 410 (2022).

[54] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV, USA, 2016), pp. 770–778.

[55] K. M. Gorski, E. Hivon, A. J. Banday, B. D. Wandelt, F. K. Hansen, M. Reinecke, and M. Bartelman, HEALPix-A framework for high resolution discretization, and fast analysis of data distributed on the sphere, Astrophys. J. **622**, 759 (2005).

[56] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, On the variance of the adaptive learning rate and beyond, *8th International Conference on Learning Representations, ICLR 2020* (2019).

[57] https://github.com/oxpig/cesped.