The Inaugural Flatiron Institute Cryo-EM Conformational Heterogeneity Challenge

Miro A. Astore[1,2†], Geoffrey Woollard[1,2,3†], David Silva-Sánchez[2,4†], Wenda Zhou[2], Mykhailo Kopylov[5], Khanh Dao Duc[6], Roy R. Lederman[7], Yilai Li[8], Yi Zhou[8], Jing Yuan[8], Fei Ye[8], Quanquan Gu[8], Rémi Vuillemot[9,10], Slavica Jonic[10], Lan Dang[11], Steven J. Ludtke[11], Hannah Bridges[12], Serena Liu[12], Michael McLean[12], Valentin Peretroukhin[12], Johannes Schwab[13], Eduardo R. Cruz-Chú[14], Peter Schwander[14], Marc Aurèle Gilles[15], Amit Singer[15,16], David Herreros[17], Jose-Maria Carazo[17], Carlos O.S. Sorzano[17], J. Ryan Feathers[18,19], Ellen D. Zhong[19], Nikolaus Grigorieff[20], Pilar Cossio[1,2*], Sonya M. Hanson[1,2*]

[1]Center for Computational Biology, Flatiron Institute, 162 5th Ave., New York, 10010, NY, USA.
[2]Center for Computational Mathematics, Flatiron Institute, 162 5th Ave., New York, 10010, NY, USA.
[3]Department of Computer Science, University of British Columbia, ICICS/CS Building 201-2366 Main Mall, Vancouver, V6T 1Z4, BC, Canada.
[4]Department of Applied Mathematics, Yale University, 12 Hillhouse Ave., New Haven, 06511, CT, USA.
[5]City College of New York, New York Structural Biology Center, 89 Convent Ave., New York, 10027, NY, USA.
[6]Department of Mathematics, University of British Columbia, 1984 Mathematics Road, Vancouver, V6T 1Z2, BC, Canada.
[7]Department of Statistics & Data Science, Yale University, 260 Whitney Ave., New Haven, 06511, CT, USA.
[8]ByteDance Seed, Los Angeles / San Jose / Shanghai, CA, USA / China.
[9]CNRS, Grenoble INP, LJK, Univ. Grenoble Alpes, Grenoble, 38000, France.

[10]IMPMC - CNRS UMR 7590, Sorbonne Université, 4 Pl. Jussieu, Paris, 75005, France.
[11]Department of Biochemistry and Molecular Pharmacology, Baylor College of Medicine, 1 Baylor Plaza, Houston, 77030, TX, USA.
[12]Structura Biotechnology Inc., 400 University Ave., Toronto, M5G 1S5, ON, Canada.
[13]MRC Laboratory of Molecular Biology, Cambridge Biomedical Campus, Francis Crick Ave., Cambridge, CB2 0QH, Cambridge, United Kingdom.
[14]Physics Department, University of Wisconsin–Milwaukee, 3135 North Maryland Ave., Milwaukee, 53211, WI, USA.
[15]Department of Mathematics, Princeton University, Fine Hall, Washington Road, Princeton, 08544, NJ, USA.
[16]Program in Applied and Computational Mathematics, Princeton University, Fine Hall, Washington Road, Princeton, 08544, NJ, USA.
[17]The National Center for Biotechnology, CSIC, Spanish National Research Council, C. Darwin, 3, Fuencarral-El Pardo, Madrid, 28049, Spain.
[18]Department of Molecular Biology, Princeton University, 119 Lewis Thomas Laboratory Washington Rd., Princeton, 08544, NJ, USA.
[19]Department of Computer Science, Princeton University, 35 Olden St., Princeton, 08540, NJ, USA.
[20]Howard Hughes Medical Institute, University of Massachusetts Chan Medical School, 55N Lake Ave., Worchester, 01655, MA, USA.

*Corresponding author(s). E-mail(s): pcossio@flatironinstitute.org; shanson@flatironinstitute.org;
†These authors contributed equally to this work.

Abstract

Despite the rise of single particle cryo-electron microscopy (cryo-EM) as a premier method for resolving macromolecular structures at atomic resolution, methods to address molecular heterogeneity in vitrified samples have yet to reach maturity. With an increasing number of new methods to analyze the multitude of heterogeneous states captured in single particle images, a systematic approach to validation in this field is needed. With this motivation, we issued a challenge to the community to analyze two cryo-EM particle image sets of thyroglobulin that exhibit continuous conformational heterogeneity. The first dataset was experimental and the second was generated with a simulator, allowing control over the distribution of molecular structures and enabled direct comparison between participants' submissions and the ground truth molecular structures and

distributions. Participants were asked to submit 80 volumes representing the heterogeneous ensemble and estimate their respective populations in the image sets provided. Participation of the research community in the challenge was strong, with submissions from nearly all developers of heterogeneity methods, resulting in 41 submissions across both datasets. Submissions qualitatively exceeded expectations, with the molecular motions identified by methods resembling both each other and the ground truth motion. However, quantitatively assessing these similarities was a challenge in and of itself. In the process of assessing the submissions, we developed several validation metrics, most of which require reference to the underlying ground truth volumes. However, we have also explored the use of metrics that do not necessarily reference ground truth. This is particularly apt for experimental datasets where ground truth is inaccessible. These approaches allowed us to assess the similarity and accuracy in volume quality, molecular motions, and conformational distribution of different submissions. These metrics and the efforts of all participants help chart a path forward for the improvements of heterogeneity methods for cryo-EM and for future challenges to validate these new methods as they continue to be developed by the community.

Over the last decade, cryo-electron microscopy (cryo-EM) has become mainstream for the determination of high resolution structures of biological macromolecules. Especially for difficult targets, it has become a preferred method over X-ray crystallography and nuclear magnetic resonance [1, 2](Fig. 1A). Several key advantages are that crystals are not a prerequisite, biomolecules can be large, samples can be prepared at near-native environments, and that single molecules are imaged directly. In a typical cryo-EM dataset, images of tens of thousands to millions of biomolecules are captured within a thin layer of vitreous ice. While standard cryo-EM reconstruction methods [3, 4] generally classify these images into a few states so that tens of thousands of images can then be averaged into a single high resolution structure, information about a broader distribution of states can also be captured from cryo-EM datasets due to its single-molecule nature. Indeed, an increasing number of methods are being developed and used to analyze heterogeneity in cryo-EM particles (Fig. 1A), ranging from linear methods to machine-learning algorithms that map the images onto a low-dimensional space representing the main conformational motions, to methods coupled to physical simulators of the conformational motions. We refer readers elsewhere for an overview of approaches [5–7].

The capacity for cryo-EM to capture conformational heterogeneity of biomolecules has potential impacts from basic biology to drug design. Understanding the conformational change, for example, of the SARS-CoV2 spike protein [8, 9] or the capacity for cryo-EM to capture multiple states of drug targets like GPCRs [10] has lead to important therapeutic advancements [11, 12]. Generally, it is acknowledged that static structures of proteins do not tell the whole story of their function. Some key examples of this in cryo-EM heterogeneity analysis include gaining insights into allosteric modulation [13, 14], entropy-driven processes [15], or the transition state of enzymes [16, 17]. However, analyzing the conformational heterogeneity present in cryo-EM data is challenging, because applying different methods to the same dataset

can yield varying results [18], and there are limited metrics and tools for quantitatively comparing and validating these outputs [19, 20]. This stems from a fundamental difference in objectives: while traditional cryo-EM reconstruction prioritizes high-resolution structure determination, supported by well-established validation metrics, heterogeneity analysis focuses on resolving conformational motions and estimating population distributions; metrics for validating these aspects remain underdeveloped.

Community challenges in cryo-EM have been launched to compare algorithmic performance on a variety of cryo-EM data processing tasks: CTF estimation [21], volume estimation [22], atomic model building [23], and particle picking [24]. However until now, a community challenge for conformational heterogeneity in cryo-EM has yet to be conducted. One key barrier to running a community challenge for conformational heterogeneity is the lack of metrics for assessing the quality of results. For example, gold-standard metrics that avoid overfitting like the half-set Fourier shell correlation (FSC) for reconstruction or the Q-score for model building do not exist for analyzing output series of volumes from cryo-EM variability methods.

Despite this, the age of machine learning has made it clear that benchmarks and metrics can rapidly accelerate algorithmic advances [25], especially obvious in the field of structural biology with the role of CASP in facilitating the success of AlphaFold [26]. Motivated both by these successes, and our own questions about how much to trust the results produced by conformational heterogeneity methods in cryo-EM, we launched the Flatiron Institute Cryo-EM Conformational Heterogeneity Challenge to the community in the summer of 2023, which included two datasets: one simulated with a ground truth, for which metrics would be more straightforward, and one experimental dataset. We had strong participation from the cryo-EM method's developer community across the globe, with 41 submissions for both datasets. Here, we first describe the construction of the challenge, and the participant's submissions. Then, we present the metrics we developed to assess and compare the methods' outputs for several specific tasks, assessing volumes, motions, and distribution quality across submissions. Overall, we do not find a single method that consistently outperforms the others; rather, different methods highlight different trends.

1 Results

1.1 Dataset design

In the design of this heterogeneity challenge, we sought a system that, though simple, would present a non-trivial task for the participants. We were also motivated to include both experimental and simulated cryo-EM images (Fig. 1B, left) that exhibited continuous conformational heterogeneity. The experimental dataset was a 33,742 particle stack of thyroglobulin that reconstructed to a resolution of 3.1 Å (see Methods for details). We subsequently used this experimental dataset to design the simulated dataset. The simulated dataset was constructed using the Euler angles, defocus values, and particle number of the experimental dataset, and was generated using the *cis*TEM (Computational Imaging System for Transmission Electron Microscopy) simulator [27, 28]. This simulator is more complex than simulators that only make the typical weak phase and projection approximations and employ Gaussian white noise.

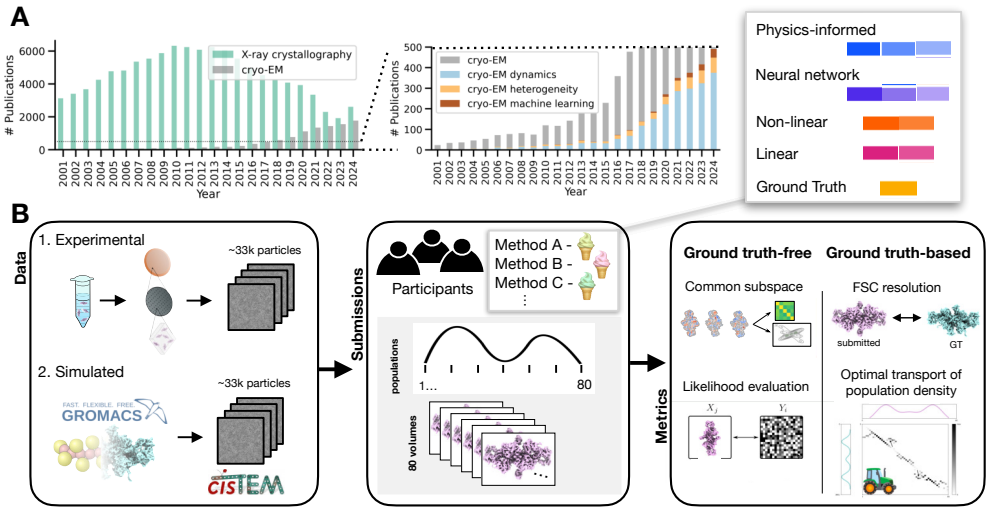

Fig. 1: **A.** Number of publications per year for 'X-ray crystallography' vs. 'cryo-EM' search terms (left) or 'cryo-EM' with additional search terms 'dynamics', 'heterogeneity', or 'machine learning' (right). Note zoomed-in y-axis for cryo-EM subterms. Data gathered *via* PubMed Trends. **B.** Schematic of the challenge: dataset generation, submissions, and comparison metrics. Upper right panel indicates the legend for the method type, subsequently used throughout the rest of the manuscript.

We chose this more complex simulator because we wanted to model experimental cryo-EM data more closely with *cis*TEM's multislice integration, explicit solvent modeling, and inclusion of radiation damage. Notably, not all participants were able to distinguish between the experimental and simulated datasets during the initial phase of the challenge because the simulated dataset exhibits a higher level of noise compared to the experimental dataset..

A main motivation for including a simulated dataset in this challenge was to have ground truth conformations and a distribution to which we could compare participants' submissions. We ran coarse-grained molecular dynamics simulations of thyroglobulin whose conformational space we then redistributed along the first principal component of its motion in molecular space (MD-PC1), which we defined as the "ground truth heterogeneity coordinate" (see section 3.1). We re-weighted the conformational distribution according to a nonuniform mixture of three Gaussians that included a less-populated state in the middle of this space, which we believed would represent an interesting challenge for methods to identify (Fig. 4). A total of 3,861 conformations were sampled from this distribution, backmapped to atomic coordinates and used to generate the 33,742 image simulated particle stack. While re-weighting the distribution along MD-PC1 and quenching motions MD-PC2 (see Methods for details) was done to emphasize one degree of freedom, the simulated volumes still contained other degrees of freedom and exhibit considerable side chain motion.

1.2 Submission summary

For both of the provided datasets, we asked for participants to submit 80 volumes along the dominant degree of freedom and the relative populations of each volume (Fig. 1B, middle). This format was influenced by our knowledge that most cryo-EM heterogeneity methods produce volumes as output, rather than molecular ensembles for example, and the request for 80 volumes was to increase the chance of picking up low-probability states. To facilitate comparison, we asked that participants align their submitted volumes to a provided reference and match its voxel size (2.146 Å). Additionally, we designed two mock submissions for the simulated dataset built from the Ground Truth (GT) which we refer to as Averaged GT and Sampled GT. To generate these, first we sorted the volumes by MD-PC1, and split up the volumes into 80 equispaced groups with an approximately equal number of volumes. Averaged GT consists of 80 volumes generated by Euclidean averaging each equispaced set. Sampled GT consists of 80 volumes selected by choosing the most commonly sampled volume in each equispaced set.

All submissions underwent the same pre-processing of being aligned to the consensus reconstruction for experimental dataset, or to the average of all the ground truth volumes for the simulated data, before being passed through the separate analysis pipelines (see Methods for further details). This was done to avoid favoring one volume pair, or having to perform multiple alignments between all pairs. We labeled the submissions by ice cream flavor to preserve the anonymity of the participants. We have grouped methods, as shown in the top right of Fig. 1, as physics-informed (blue), neural network without physics (purple), non-linear (orange), linear (magenta), and ground truth for the simulated datasets (yellow). We note that some ice cream flavors have additional number labels (e.g. Ice Cream Flavor 1, Ice Cream Flavor 2, etc). These count the number of rounds the same method made a submission. Note that only submissions from "round 1" are blind; after that participants had seen a presentation illustrating the ground truth distribution for the simulated dataset (although not the ground truth latent values of the simulated images, and not the ground truth volumes).

1.3 High level overview of submissions

As an initial analysis of the submissions, the simplest procedure is to visually inspect the submitted volumes and the conformational motions (Fig. 2A). From a qualitative perspective, the submissions were diverse in terms of resolution and, in general, the motions represented were consistent (Supplementary Movie 1). Despite asking for submissions to be aligned to a reference volume at a specific pixel and box size, substantial pre-processing of the submitted volumes was still required. For the simulated dataset, since the ground truth is available, we could also assess the resolution of the submissions by comparing the volumes to the ground truth structures using the Fourier shell correlation (FSC) with a 0.5 threshold (Fig. 2E). We computed the FSC for all 3861 GT volumes against all volumes in every submission, and matched each of the 80 volumes with the closest GT volume using the resolution at FSC 0.5 threshold. Some submissions had relatively tight resolution distributions, either higher (Cookie Dough

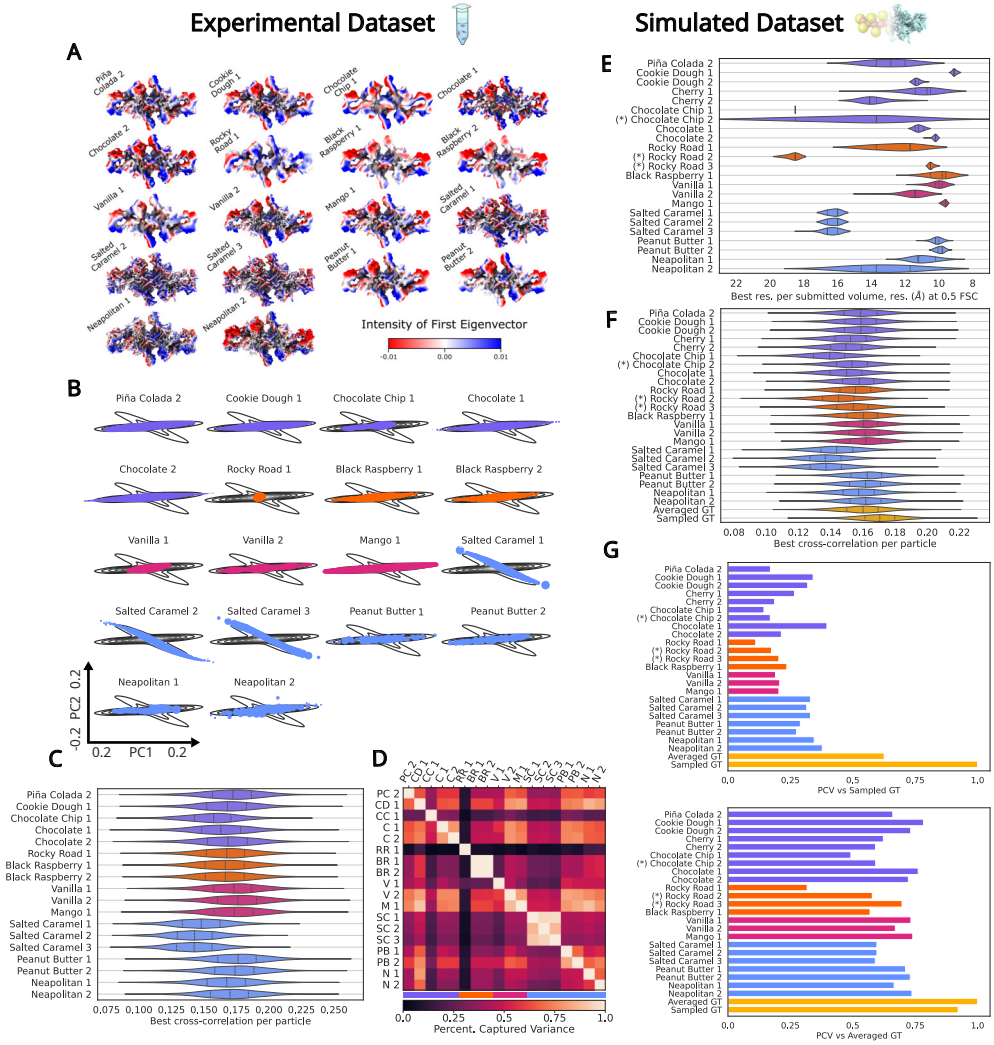

Fig. 2: **A.** The 40th volume from each submission for the experimental dataset, colored by the first PCA eigenvector (blue is positive, red is negative, white is zero). **B.** Projection of submissions onto PC1 and PC2 of the common subspace for the experimental dataset. Each dot represents a submitted volume, with the size defined by the submitted population. The contour plot is a weighted kernel density estimate computed with the projection of all submissions and their populations. In **C.** and **F.** the best volume-to-image cross-correlation for each particle is shown for the experimental and simulated dataset, respectively, for a subset of particles. **D.** The percentage of captured variance (PCV) between the submissions in the experimental dataset, specifically, the amount of variance captured between each pair of submissions. Similarly, **G.** shows the PCV between the submissions and the ground truth. **E.** Best resolution for each submitted volume obtained by computing the Fourier shell correlation (FSC) against the GT volumes. The resolution is defined using the 0.5 FSC threshold. (*) Indicates submissions that made use of an expanded simulated dataset with 674,840 particles.

1 around 9 Å; Mango 1 around 9.5 Å) or lower (Chocolate Chip 1 and Rocky Road 2 around 18.5 Å). Other submissions had a large range of values, for example Chocolate Chip 2, Salted Caramel 3. As an alternate validation metric not needing access to ground truth volumes, we directly assessed each submitted volume against a subset of the individual images by exhaustively searching the pose space and identifying the optimal cross-correlation for each image (see Methods for details). We observed that the distributions of image-to-volume cross-correlations are broad for both the experimental and simulated datasets (Fig. 2C and F, respectively). However, some methods produce volumes that align more closely with the images than others. Notably, the patterns observed in image-to-volume comparisons do not always align with those seen in volume-to-volume evaluations.

1.4 Comparing conformational motions through PCA

For both experimental and simulated datasets, a subspace can be computed for each submission. These subspaces can then be quantitatively compared, and a common subspace can be derived to visualize the similarity between the inferred conformational motions. This is accomplished through Principal Component Analysis (PCA) on each submission, which provides an orthogonal basis defined by the eigenvectors $\mathcal{V}$ and the eigenvalues S^2 of each submission's volume covariance matrix. These bases can be used to compute a common subspace, as detailed in Section 6.1.4. We compute the similarity between the different bases using the Percentage of Captured Variance (PCV) introduced in ref. [29].

Fig. 2A-B shows the behavior of each submission in both the individual and common subspaces for the experimental dataset. Fig. 2A illustrates the regions of variance in the first principal component eigenvector for each submission's embedding. A change in color indicates regions of variance between the submitted volumes, providing insights into the motions predicted by each submission. Conversely, Fig. 2B shows the projection onto PC1 and PC2 of the common subspace obtained for the simulated dataset submissions. Most methods cover a similar range in PC1, demonstrating that most methods share at least one high variance conformational motion. The dominance of Salted Caramel in the second principal component can be attributed to it having more submissions and possibly presenting higher frequency variance. Supplementary Fig. S1B shows the projection of the motions for the simulated dataset, and the conformational motion with the highest variance aligns well with that of the ground truth and most of the submissions.

To quantify the similarity between different submissions, we measure the PCV between each pair of submissions, which we denote by $PCV(i, j)$, where i and j refer to the i-th and j-th submission, respectively. Submission i serves as the reference, and we compute how much of its variance is captured by the subspace of submission j (see Methods for details). The asymmetry in this metric allows us to quantify, when comparing two submissions, which submission is more effective at capturing variance. For instance, when comparing the two ground truth mock submissions through this metric in Fig.2G, we observe that the Sampled GT effectively captures the Averaged GT, while the opposite is not as true. This finding aligns with our expectations, as averaging tends to result in the loss of high-frequency features. Fig. 2D illustrates the

variance captured among the submissions in the experimental dataset. In this figure, the rows correspond to the reference submissions (i.e., submission i). We notice that submissions produced using the same method, as well as those belonging to the same category, tend to cluster together. This is evident from the higher PCV values that are situated close to the diagonal.

Fig. 2G shows the PCV for the simulated dataset using the Ground Truth mock submissions–Averaged GT and Sampled GT–as references, top and bottom, respectively. The results indicate that methods with higher resolution outputs (e.g., Salted Caramel, Chocolate, and Neapolitan) are better at capturing the high frequency features of the conformational heterogeneity from the ground truth volumes. Once these features are averaged out, lower resolution submissions (e.g., Rocky Road, Vanilla, Mango) perform much better. There are also methods that have a good balance of both cases, such as Chocolate and Cookie Dough. Supplementary Fig. S1C shows the percentage of captured variance between all methods as a distance matrix. As for the simulated dataset (Fig. 2D), we observe that methods assigned to the same categories have a tendency to form clusters, which can be seen from bright spots along the diagonal. Additionally, we notice that Piña Colada, Peanut Butter and Chocolate, methods that presented a balanced behavior in the simulated dataset, are easily captured by other methods (bright rows); while Neapolitan and Peanut Butter are good at capturing other methods (bright columns).

1.5 Comparing conformational distributions

A key objective of the challenge was to assess the capacity for heterogeneity methods to predict the population distribution along the 1D conformational coordinate. Since the true populations are unknown for the experimental dataset, our analyses will first focus on the simulated dataset, where the ground truth distribution is available.

1.5.1 *Via* linear subspace embedding

One way to compare the conformational distributions is by using the linear subspaces introduced in Section 1.4. Specifically, we projected the submissions onto the subspace defined by the Sampled GT and visualized the distributions with respect to the first principal component along with the population of submitted volumes. This is illustrated in Supplementary Fig. S2, where these distributions are compared to the distribution of the Averaged GT. Most submissions report a distribution with two modes, either missing the middle mode or showing shifted modes that possibly accumulate the middle mode into the lateral modes. Other submissions simply report a single mode in the middle. Only three submissions (Cherry 2, Chocolate 2, and Vanilla 2) infer the three modes present in the ground truth, although corresponding to the second (i.e. non-blind) round of submissions. On the other hand, most submissions obtain a good coverage of the first principal component, showing that most methods are good at estimating the conformational variability coverage.

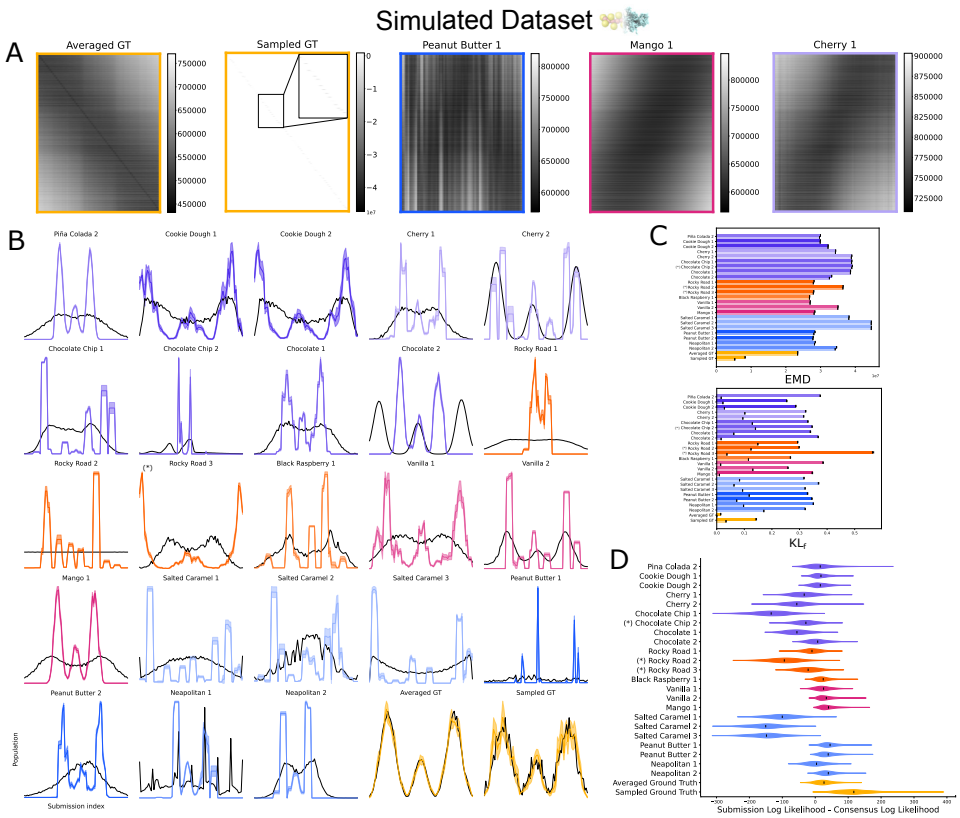

Fig. 3: **A.** Volume-to-volume BioEM3D distance, Eq. (17), for select submissions. The distance matrices for all submissions can be found in Supplementary Fig. S3. **B.** Optimal population weight (in color, Eq. (33)) *via* the earth movers distance (EMD), and original submitted population weight (black). Replicates average volume-to-volume distances over 50 nearest neighbors by ground truth ordering (MD-PC1). **C.** EMD and KL_f comparing submitted volumes and population weights to the ground truth. The error bars are from averaging using different neighborhood sizes of 40,45 and 50. **D.** Ensemble-to-image likelihood (Eq. 39) for a subset of 1224 images corresponding to Sampled GT, relative to the volume-to-image likelihood of the consensus GT volume.

1.5.2 *Via* Earth Mover's distance and Kullback–Leibler divergence

In order to quantitatively compare the information in the volumes and their population weights, we used a distance over distributions known as the earth mover's distance (EMD) [30]. The EMD compares two probability distributions over a metric space and is also known as the "Wasserstein distance" or "Kantorovich–Rubinstein metric" in the optimal transport literature. In discrete settings it involves finding a soft correspondence matrix between two weighted spaces, referred to as an optimal transport plan, which is a joint probability matrix between these spaces. In other words, the

optimal transport plan is the joint distribution that minimizes the distance, d, defined between these spaces. Here we compare two discrete probability distributions (of size 3861 for the ground truth, and 80 for the submission) where we use the EMD, or alternatively the Kullback-Leibler (KL) divergence, with a volume-to-volume distance, d (see Section 6.2 for details). The EMD has already been used to compare distributions of atomic models of biomolecules [31], but in our case, we compared voxelized volumes, which do not generally include correspondence information between labeled coordinates.

We experimented with various "volume-to-volume" distance functions suitable for comparing these discretized scalar densities (broadly BioEM3D, FSC, Zernike3D, transport distances — see Methods for details), balancing computational feasibility with robustness to pose and high frequency conformational heterogeneity. Furthermore, since inferring the relative populations of their submitted volumes was a new task for many challenge participants, we also optimized their relative populations under the EMD and KL objectives. In other words we used (i) ground truth volumes and (ii) ground truth relative populations, and (iii) submitted volumes, but optimized the submitted relative populations.

As reference, we present here the results using one specific volume-to-volume distance, the BioEM3D distance, which marginalizes out the overall offset, magnitude and noise level of the submitted volumes, under a Gaussian white noise model in 3D voxel space. Results for other volume-to-volume distances are shown in supplementary Section 6.4. In Fig. 3A, we show the volume-to-volume distance matrix for several representative submissions. Using these, we compare the optimal relative population (colored) to the submitted one (black) in the submission's volume space (Supplementary Fig. 3B). As illustrated by the Averaged GT, the optimized populations match the reference with high accuracy. Also note that with the BioEM3D distance in an EMD volume distribution objective yielded a smooth optimal population for Sampled GT, which was also seen for the KL conformational distribution objective, but not seen in the other volume-to-volume distances (data not shown).

We averaged the volume-to-volume distance matrix along neighborhoods of ground truth ordering (MD-PC1). However, we observed that when large neighborhoods are used and the number of (averaged) ground truth states subsequently decreases, the sparsity of the solution increases. This is expected with the unregularized EMD and KL objectives. We therefore averaged over random replicates, where each replicate drops 90% of ground truth volumes. While each replicate yields sparse optimized relative populations, the average is more smooth. In order to promote smoothness without assuming a ground truth ordering we used entropic regularization, but it did not give the same trend (Supplementary Fig. S4).

Supplementary Fig. S5D compares the conformational distribution with the EMD using a BioEM3D volume-to-volume distance, and different trends are seen even among submissions of the same ice cream flavor: for instance Cherry, Chocolate, Rocky Road, Vanilla, Peanut Butter, Neapolitan, and finally Sampled/Averaged GT. Note that while almost identical results are seen for the weighted populations optimized by a KL vs. EMD objective (data not shown), the magnitude of change in the KL objective is more pronounced (Fig. S5C). Small changes in the EMD objective are expected

since the volume-to-volume distances have an intrinsic scale: for example, Averaged GT ranges from 0.4 - 0.75 (Fig. S5A).

In order to put submissions on the same scale, we rank normalized the volume-to-volume distances. Supplementary Fig. S6 shows that rank normalization had a minimal effect on the optimized relative population (although Sampled GT did dramatically change), but changed the trends of the objective values. Note that the BioEM3D volume-to-volume distance for Sampled GT atypically shows low values on the diagonal, and is high elsewhere, regardless of the proximity of the submitted volume along the ground truth ordering (MD-PC1), which could explain why its behavior was different to other submissions. This effect is also seen in the BioEM3D volume-to-volume distance matrix of Averaged GT compared with itself in Supplementary Fig, S4B (lower). In both cases there is an exact match between some volume pairs being compared, which have a BioEM3D value several order of magnitude lower than non-exact matches.

For roughly half of the submissions, Fig. 3B illustrates that the optimized probabilities are a trimodal mixture, with two symmetric large modes close to the two large modes from the submitted population and another mode ranging from around half to a third its size in the middle, which resembles the ground truth distribution, as expected. In a related way, various bimodal submissions are optimized to be trimodal, indicating the difficulty of inferring low probability surrounding a "hidden" middle mode. This is not surprising given the high level of noise in the images, small scale of the motion, and low number of images that represent the trough in the ground truth relative population distribution.

1.6 Ensemble-to-images comparison metric

The previous metrics for assessing distributions relied on ground truth volumes and populations, making them infeasible in the evaluation of submissions for the experimental dataset. Here, we introduce a comparison metric that bypasses this requirement by directly evaluating how well an ensemble (i.e., volumes and populations) explains a given set of cryo-EM images. This metric is independent of identifying a common subspace, and does not require the use of ground truth volumes, because it directly compares volumes with populations to the images using the ensemble-to-image likelihood (Eq. 39) [32]. Each volume is compared to each image by finely searching the pose and particle center, assuming a forward model of image formation with Gaussian white noise [33]. As this pose-search is computationally expensive, only a subset of particles from each dataset were chosen to evaluate this metric (see Section 6.6 for details). The higher the log-likelihood the more likely the particle images are to be generated by the volumes at the given relative population. With this metric, we ranked the performance of submissions for 2000 of the experimental images (Supplementary Fig. S7), where we had no ground truth for the poses or volumes. While for the simulated dataset, we specifically chose to compare to a subset of 1224 images which corresponded to the 80 volumes in the Sampled GT, allowing us access to a ground truth comparison for this metric (Fig. 3D). These results show that the log-likelihoods from the ensemble-image comparison agree to some extent with the EMD/KL results, especially for the submissions that are far from the data (e.g. Salted Caramel, Rocky Road 2, Chocolate Chip

1). Unlike the EMD/KL, many submissions show better agreement with the images than Averaged GT. However, the volumes corresponding to the exact images (Sampled GT) have the best agreement. In this case, we found we could use Markov chain Monte Carlo (MCMC) and ensemble reweighting [32] to recover the correct relative populations between the volumes in the Sampled GT submission directly using the image-to-structure likelihoods (Supplementary Fig. S8). Overall, this suggests that the ensemble-to-image likelihood comparison is able to discern if the ensemble is a poor match, but is sensitive to an exact match between the images and volumes, and care must be taken if volumes are smoothed or averaged in some manner.

When comparing the log likelihood results for submissions to both the experimental and simulated datasets, we find that no submissions perform significantly better or worse on either dataset. Similar trends between the relative ranks of submissions are observed in Fig. 3D and Supplementary Fig. S7. This can also be seen in the cross-correlation scores (Fig. 2C, F). However, because of higher noise levels in the simulated dataset, the difference between the log likelihoods compared to the consensus structure in the submissions to the simulated dataset are usually less than those for the experimental dataset.

Discussion and conclusions

The capacity to accurately discern the structures and populations of conformational ensembles of biological macromolecules has vast implications for the understanding of biological processes. Cryo-EM offers an unprecedented window into this conformational heterogeneity. However, cryo-EM methodologies and validation metrics have primarily focused on achieving high-resolution reconstructions, while quantitative frameworks for assessing the accuracy of approaches that recover conformational motions and population distributions are still emerging. Here, we presented both experimental and simulated datasets analyzed by different methods in the hands of the developers themselves and subsequently compared to one another and to ground truth when available. The submissions overall were able to capture similar conformational motions and many methods even captured populations comparable to the ground truth distribution in the simulated dataset.

We launched this challenge knowing that the metrics with which the submissions would be assessed had yet to be developed. Therefore, in the process of analyzing the challenge submissions, we have developed and tested new quantitative cryo-EM heterogeneity metrics, with the aim of providing computationally tractable, reliable indicators of motions and population prediction. We developed three distinct classes of metrics to evaluate the quality of submissions across different tasks: *i*) a PCA-based method that compares conformational subspaces, *ii*) an optimal transport method that allows for the comparison of populations under volume-to-volume similarity measures, and *iii*) an image-to-ensemble likelihood that allows for direct comparisons of the volumes and populations to the provided images. One common advantage of all these metrics is that they do not require analysis of the latent spaces produced by the heterogeneity analysis methods themselves. One issue we have had with all metrics, however, is their sensitivity to the resolution of the submitted volumes and other

aspects of the preprocessing of the volumes. When comparing volume series through PCA, it is important that the volumes in all series are close in resolution, as we found that otherwise methods with high resolution outputs would be unfairly penalized. We solved this issue by normalizing the power spectrum of submissions and ground truth volumes, allowing us to find that submissions like Neapolitan, Salted Caramel, and Peanut Butter excel at matching the ground truth volumes. For the volume-to-volume metrics, one can downsample volumes or smooth out the distance to high frequency information (e.g. truncate the FSC at a lower resolution). For the volume-to-images comparison, one can likewise simply downsample images and volumes.

Overall, while some methods show similar patterns, no single method consistently outperforms the others; instead, different approaches demonstrate strengths in different areas. We also did not observe strong evidence that particular group types systematically outperform others across specific tasks. For instance, some ML-based methods—such as Cookie Dough—achieved strong EMD scores but performed average in terms of image cross-correlation. Likewise, performance among physics-based methods varied significantly depending on whether the metric related to resolution. These findings suggest that the specific design and implementation details of each method play a critical role in determining its effectiveness for a given task.

Throughout the analysis of the challenge submissions, some of the limitations of the dataset choice and submission format have become clearer, though ultimately we believe this task has been a suitable one for a first attempt at a conformational heterogeneity challenge in cryo-EM. For example, while thyroglobulin is a biomolecule of typical size (660 kDa) and conformational heterogeneity can be detected with a modest particle count (33,742 images), it is a symmetric molecule, hence conformational heterogeneity can be washed out due to degeneracies in pose estimation (Supplementary Fig. S9). For the simulated dataset, we aimed to restrict heterogeneity to a single dimension by selecting conformations along PC1 and limiting variance to a narrow band in PC2. However, because higher-order modes were not filtered, residual variability from these modes in addition to high-frequency motions like side chain dynamics remain present in the data (Supplementary Movie 2). In the future, while multi-dimensional state spaces are realistic in molecular systems, a more careful limitation of the motion in question to a single degree of freedom may be warranted. Additionally, these analyses provided greater insight into the choice of the number of volumes. For example, 80 volumes were initially selected under the assumption that this would increase the likelihood of participants identifying a low-lying state of interest. However, subsequent studies evaluating the number of ground-truth average volumes required to achieve robust distributions under the EMD metric revealed that 80, 40, 20, 16, 10, and even 8 ground-truth–derived volumes all yielded similarly suitable matches (Supplementary Fig. S10). Finally, we note that requesting submissions in the form of volume series rather than atomic models introduced certain challenges, as this format requires standardization steps such as filtering, sharpening, and alignment—procedures that are non-trivial using volumes but essential to ensure fair comparisons. Nevertheless, volumes remain the most widely accessible format for most heterogeneity-focused methods and offer valuable insights into conformational variability, even when only low-resolution data is available.

In summary, with 41 submissions across both datasets and the development of metrics to assess these submissions, this first conformational heterogeneity challenge in cryo-EM has been informative on many fronts. First, while benchmark datasets based on simulations are immensely valuable [19, 34], this challenge represents the first coordinated effort by the community to systematically compare cryo-EM heterogeneity analysis methods on both experimental and simulated datasets. Participants were given the opportunity to optimize their approaches as they saw fit, enabling a realistic assessment of each method's capabilities. Second, we established a foundation for cryo-EM heterogeneity metrics tailored to specific tasks—capturing motions and distributions—both in the presence and absence of ground truth data. We have brought to the forefront the importance of accurately estimating populations [35], crucial for understanding the thermodynamics of biomolecules. Third, we hope this challenge serves as only the first of many that can drive the field forward towards more confident estimates of conformational ensembles from cryo-EM data. Finally, we are grateful to a community of scientists whose contributions made this project possible and who will remain essential to advancing the field for the many practitioners seeking to confidently use computational methods to extract the information inherent in their cryo-EM datasets.

Data Availability

All submitted volumes and relative populations, anonymized and pre-processed, are available at https://osf.io/8h6fz/. All analysis code is available at https://github.com/flatironinstitute/Cryo-EM-Heterogeneity-Challenge-1/. The two challenge datasets remain available at https://www.simonsfoundation.org/flatiron/center-for-computational-biology/structural-and-molecular-biophysics-collaboration/heterogeneity-in-cryo-electron-microscopy/.

Author contributions

YL, YZ, JY, FY, QG, RV, SJ, LD, SJL, HB, SR, MM, VP, JS, ERCC, PS, MAG, AS, DH, JMC, COSS, JRF, and EDZ submitted to the challenge as participants. MAA, DSS, WZ, NG, PC, and SMH devised the challenge, generated the simulated data, and launched the challenge. MK collected and analyzed the electron microscopy data. MAA, GW, DSS, and DH implemented metrics, plotted figures, and developed automated workflow for data processing. MAA, GW, DSS, PC, and SMH devised methodology, coordinated correspondence with co-authors, interpreted results, and jointly wrote the manuscript with feedback from all authors.

Acknowledgments

The first authors and corresponding authors would like to thank the general support of the Structural and Molecular Biophysics team and the Center for Computational Biology throughout this process. In general we thank our admin team at the Flatiron Institute, including but not limited to Rebecca Sesny who helped with website

creation and Christina Julien who helped in compiling co-author information. We are grateful to Reza Paraan and Will Conway for their involvement in discussions at early stages of the project. GW acknowledges various colleagues for helpful discussions: Frederik Kunstner on multiplicative gradient descent for a KL objective; Bruce Shepherd on the linear EMD objective as a min flow problems; Daniel Lee on using Frank Wolfe to solve the Gromov-Wasserstein distance; Minhuan Li on the self-continuity; and Luke Evans on connections to his work on ensemble-reweighting. KDD and GW acknowledge the support of the Government of Canada's New Frontiers in Research Fund / fonds Nouvelles frontières en recherche du gouvernement du Canada (NFRF), NFRFE-2019-00486. JS acknowledges support from the Medical Research Council (MRC), as part of United Kingdom Research and Innovation (UKRI) (MC UP A025 1013 to Sjors Scheres). AS and MAG are supported in part by the Simons Foundation Math+X Investigator Award and NIH/NIGMS R01GM136780-01. PS acknowledges support from the U.S. National Science Foundation under Award No. STC1231306. The Flatiron Institute is a division of the Simons Foundation. Technical support was provided by the Scientific Computing Core of the Flatiron Institute.

References

[1] Cheng, Y.: Single-particle cryo-em—how did it get here and where will it go. Science **361**(6405), 876–880 (2018) https://doi.org/10.1126/science.aat4346 https://www.science.org/doi/pdf/10.1126/science.aat4346

[2] Egelman, E.H.: The current revolution in cryo-em. Biophysical Journal **10**(5), 1008–1012 (2016) https://doi.org/10.1016/j.bpj.2016.02.001

[3] Scheres, S.H.W.: Relion: Implementation of a bayesian approach to cryo-em structure determination. Journal of Structural Biology **180**(3), 519–530 (2012) https://doi.org/10.1016/j.jsb.2012.09.006

[4] Punjani, A., Rubinstein, J.L., Fleet, D.J., Brubaker, M.A.: cryosparc: algorithms for rapid unsupervised cryo-em structure determination. Nature Methods **14**(3), 290–295 (2017) https://doi.org/10.1038/nmeth.4169

[5] Donnat, C., Levy, A., Poitevin, F., Zhong, E.D., Miolane, N.: Deep generative modeling for volume reconstruction in cryo-electron microscopy. Journal of Structural Biology **214**(4), 107920 (2022) https://doi.org/10.1016/j.jsb.2022.107920

[6] Tang, W.S., Zhong, E.D., Hanson, S.M., Thiede, E.H., Cossio, P.: Conformational heterogeneity and probability distributions from single-particle cryo-electron microscopy. Current Opinion in Structural Biology **81**, 102626 (2023) https://doi.org/10.1016/j.sbi.2023.102626

[7] Kimanius, D., Schwab, J.: Confronting heterogeneity in cryogenic electron microscopy data: Innovative strategies and future perspectives with data-driven

methods. Current Opinion in Structural Biology **86**, 102815 (2024) https://doi.org/10.1016/j.sbi.2024.102815

[8] Sztain, T., Ahn, S.-H., Bogetti, A.T., Casalino, L., Goldsmith, J.A., Seitz, E., McCool, R.S., Kearns, F.L., Acosta-Reyes, F., Maji, S., Mashayekhi, G., McCammon, J.A., Ourmazd, A., Frank, J., McLellan, J.S., Chong, L.T., Amaro, R.E.: A glycan gate controls opening of the sars-cov-2 spike protein. Nature Chemistry **13**(10), 963–968 (2021) https://doi.org/10.1038/s41557-021-00758-3

[9] Rapp, M., Shapiro, L., Frank, J.: Contributions of single-particle cryoelectron microscopy toward fighting covid-19. Trends in Biochemical Sciences **47**(2), 117–123 (2022) https://doi.org/10.1016/j.tibs.2021.10.005 . Special Issue: Pushing boundaries of cryo-EM

[10] Papasergi-Scott, M.M., Pérez-Hernández, G., Batebi, H., Gao, Y., Eskici, G., Seven, A.B., Panova, O., Hilger, D., Casiraghi, M., He, F., Maul, L., Gmeiner, P., Kobilka, B.K., Hildebrand, P.W., Skiniotis, G.: Time-resolved cryo-em of g-protein activation by a gpcr. Nature **629**(8014), 1182–1191 (2024) https://doi.org/10.1038/s41586-024-07153-1

[11] Renaud, J.-P., Chari, A., Ciferri, C., Liu, W.-t., Rémigy, H.-W., Stark, H., Wiesmann, C.: Cryo-em in drug discovery: achievements, limitations and prospects. Nature Reviews Drug Discovery **17**(7), 471–492 (2018) https://doi.org/10.1038/nrd.2018.77

[12] Robertson, M.J., Meyerowitz, J.G., Skiniotis, G.: Drug discovery in the era of cryo-electron microscopy. Trends in Biochemical Sciences **47**(2), 124–135 (2022) https://doi.org/10.1016/j.tibs.2021.06.008

[13] Lü, W., Du, J., Goehring, A., Gouaux, E.: Cryo-em structures of the triheteromeric nmda receptor and its allosteric modulation. Science **355**(6331), 3729 (2017) https://doi.org/10.1126/science.aal3729 https://www.science.org/doi/pdf/10.1126/science.aal3729

[14] Zhang, L., Mobbs, J.I., May, L.T., Glukhova, A., Thal, D.M.: The impact of cryo-em on determining allosteric modulator-bound structures of g protein-coupled receptors. Current Opinion in Structural Biology **79**, 102560 (2023) https://doi.org/10.1016/j.sbi.2023.102560

[15] Jin, M., Seed, R.I., Cai, G., Shing, T., Wang, L., Ito, S., Cormier, A., Wankowicz, S.A., Jespersen, J.M., Baron, J.L., Carey, N.D., Campbell, M.G., Yu, Z., Tang, P.K., Cossio, P., Wen, W., Lou, J., Marks, J., Nishimura, S.L., Cheng, Y.: Dynamic allostery drives autocrine and paracrine tgf-β signaling. Cell **187**(22), 6200–621923 (2024) https://doi.org/10.1016/j.cell.2024.08.036

[16] Li, Y., Zhang, R., Wang, C., Forouhar, F., Clarke, O.B., Vorobiev, S., Singh, S., Montelione, G.T., Szyperski, T., Xu, Y., Hunt, J.F.: Oligomeric

interactions maintain active-site structure in a noncooperative enzyme family. The EMBO Journal **41**(17), 108368 (2022) https://doi.org/10.15252/embj.2021108368 https://www.embopress.org/doi/pdf/10.15252/embj.2021108368

[17] Tsai, M.-D., Wu, W.-J., Ho, M.-C.: Enzymology and dynamics by cryogenic electron microscopy. Annual Review of Biophysics **51**(Volume 51, 2022), 19–38 (2022) https://doi.org/10.1146/annurev-biophys-100121-075228

[18] Astore, M.A., Blackwell, R., Silva-Sánchez, D., Cossio, P., Hanson, S.M.: Heterogeneity in ligand-bound trpv1: A comparison of methods in cryo-em and molecular dynamics simulation. bioRxiv (2024) https://doi.org/10.1101/2024.10.07.617120 https://www.biorxiv.org/content/early/2024/10/11/2024.10.07.617120.full.pdf

[19] Jeon, M., Raghu, R., Astore, M., Woollard, G., Feathers, R., Kaz, A., Hanson, S.M., Cossio, P., Zhong, E.D.: CryoBench: Diverse and challenging datasets for the heterogeneity problem in cryo-EM. Advances in Neural Information Processing Systems **37**, 89468–89512 (2024) https://doi.org/10.48550/arXiv.2408.05526 arXiv:2408.05526

[20] Shi, B., Zhang, K., Fleet, D.J., McLeod, R.A., Dwayne Miller, R.J., Howe, J.Y.: Deep generative priors for biomolecular 3D heterogeneous reconstruction from cryo-EM projections. Journal of Structural Biology **216**(2), 108073 (2024) https://doi.org/10.1016/j.jsb.2024.108073

[21] Marabini, R., Carragher, B., Chen, S., Chen, J., Cheng, A., Downing, K.H., Frank, J., Grassucci, R.A., Bernard Heymann, J., Jiang, W., Jonic, S., Liao, H.Y., Ludtke, S.J., Patwari, S., Piotrowski, A.L., Quintana, A., Sorzano, C.O.S., Stahlberg, H., Vargas, J., Voss, N.R., Chiu, W., Carazo, J.M.: CTF Challenge: Result summary. Journal of Structural Biology **190**(3), 348–359 (2015) https://doi.org/10.1016/j.jsb.2015.04.003

[22] Heymann, J.B., Marabini, R., Kazemi, M., Sorzano, C.O.S., Holmdahl, M., Mendez, J.H., Stagg, S.M., Jonic, S., Palovcak, E., Armache, J.-P., Zhao, J., Cheng, Y., Pintilie, G., Chiu, W., Patwardhan, A., Carazo, J.-M.: The first single particle analysis Map Challenge: A summary of the assessments. Journal of Structural Biology **204**(2), 291–300 (2018) https://doi.org/10.1016/j.jsb.2018.08.010

[23] Lawson, C.L., Kryshtafovych, A., Adams, P.D., Afonine, P.V., Baker, M.L., Barad, B.A., Bond, P., Burnley, T., Cao, R., Cheng, J., Chojnowski, G., Cowtan, K., Dill, K.A., DiMaio, F., Farrell, D.P., Fraser, J.S., Herzik, M.A., Hoh, S.W., Hou, J., Hung, L.-W., Igaev, M., Joseph, A.P., Kihara, D., Kumar, D., Mittal, S., Monastyrskyy, B., Olek, M., Palmer, C.M., Patwardhan, A., Perez, A., Pfab, J., Pintilie, G.D., Richardson, J.S., Rosenthal, P.B., Sarkar, D., Schäfer, L.U., Schmid, M.F., Schröder, G.F., Shekhar, M., Si, D., Singharoy, A., Terashi, G., Terwilliger, T.C., Vaiana, A., Wang, L., Wang, Z., Wankowicz, S.A., Williams, C.J., Winn, M., Wu, T., Yu, X., Zhang, K., Berman, H.M.,

Chiu, W.: Cryo-EM model validation recommendations based on outcomes of the 2019 EMDataResource challenge. Nature Methods **18**(2), 156–164 (2021) https://doi.org/10.1038/s41592-020-01051-w

[24] Dhakal, A., Gyawali, R., Wang, L., Cheng, J.: A large expert-curated cryo-EM image dataset for machine learning protein particle picking. Scientific Data **10**(1), 392 (2023) https://doi.org/10.1038/s41597-023-02280-2

[25] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **28**(521), 436–44 (2015) https://doi.org/10.1038/nature14539

[26] AlQuraishi, M.: Alphafold at casp13. Bioinformatics **35**(22), 4862–4865 (2019) https://doi.org/10.1093/bioinformatics/btz422 https://academic.oup.com/bioinformatics/article-pdf/35/22/4862/48979279/bioinformatics_35_22_4862.pdf

[27] Himes, B., Grigorieff, N.: Cryo-TEM simulations of amorphous radiation-sensitive samples using multislice wave propagation **8**(6), 943–953 https://doi.org/10.1107/S2052252521008538 . Publisher: International Union of Crystallography. Accessed 2024-12-23

[28] Grant, T., Rohou, A., Grigorieff, N.: *cis*tem, user-friendly software for single-particle image processing. eLife **7**, 35383 (2018) https://doi.org/10.7554/eLife.35383

[29] Gilles, M.A., Singer, A.: Cryo-em heterogeneity analysis using regularized covariance estimation and kernel regression. Proceedings of the National Academy of Sciences **122**(9), 2419140122 (2025)

[30] Peyré, G., Cuturi, M.: Computational optimal transport. Foundations and Trends in Machine Learning **11**(5-6), 1–257 (2019) https://doi.org/10.1561/2200000073 arXiv:1803.00567

[31] Prins, V., Diepeveen, W., Bekkers, E.J., Oktem, O.: Physics-informed geometric regularization of heterogeneous reconstructions in cryo-EM. GEM workshop, ICLR 2024, 1–14 (2024)

[32] Tang, W.S., Silva-Sánchez, D., Giraldo-Barreto, J., Carpenter, B., Hanson, S.M., Barnett, A.H., Thiede, E.H., Cossio, P.: Ensemble reweighting using cryo-em particle images. The Journal of Physical Chemistry B **127**(24), 5410–5421 (2023)

[33] Tang, W.S., Soules, J., Rangan, A., Cossio, P.: Cryolike: A python package for cryo-electron microscopy image-to-structure likelihood calculations. bioRxiv, 2024–10 (2024)

[34] Joosten, M., Greer, J., Parkhurst, J., Burnley, T., Jakobi, A.J.: *Roodmus*: a toolkit for benchmarking heterogeneous electron cryo-microscopy reconstructions. IUCrJ

11(6), 951–965 (2024) https://doi.org/10.1107/S2052252524009321

[35] Evans, L., Dingeldein, L., Covino, R., Gilles, M.A., Thiede, E., Cossio, P.: Counting particles could give wrong probabilities in cryo-electron microscopy. bioRxiv, 2025–03 (2025)

[36] Kim, K., Kopylov, M., Bobe, D., Kelley, K., Eng, E.T., Arvan, P., Clarke, O.B.: The structure of natively iodinated bovine thyroglobulin. Acta Crystallographica Section D: Structural Biology **77**(11), 1451–1459 (2021) https://doi.org/10.1107/S2059798321010056

[37] Zheng, S.Q., Palovcak, E., Armache, J.-P., Verba, K.A., Cheng, Y., Agard, D.A.: MotionCor2: Anisotropic correction of beam-induced motion for improved cryo-electron microscopy **14**(4), 331–332 https://doi.org/10.1038/nmeth.4193 . Accessed 2025-06-02

[38] Webb, B., Sali, A.: Comparative Protein Structure Modeling Using MODELLER **54**, 5–615637 https://doi.org/10.1002/cpbi.3 27322406

[39] Souza, P.C.T., Alessandri, R., Barnoud, J., Thallmair, S., Faustino, I., Grünewald, F., Patmanidis, I., Abdizadeh, H., Bruininks, B.M.H., Wassenaar, T.A., Kroon, P.C., Melcr, J., Nieto, V., Corradi, V., Khan, H.M., Domański, J., Javanainen, M., Martinez-Seara, H., Reuter, N., Best, R.B., Vattulainen, I., Monticelli, L., Periole, X., Tieleman, D.P.T., Marrink, A.H.d.V., J., S.: Martini 3: A general purpose force field for coarse-grained molecular dynamics **18**(4), 382–388 https://doi.org/10.1038/s41592-021-01098-3 . Accessed 2025-06-02

[40] Kabsch, W., Sander, C.: Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers **22**(12), 2577–2637 (1983) https://doi.org/10.1002/bip.360221211 https://onlinelibrary.wiley.com/doi/pdf/10.1002/bip.360221211

[41] Abraham, M.J., Murtola, T., Schulz, R., Páll, S., Smith, J.C., Hess, B., Lindahl, E.: Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX **1-2**, 19–25 (2015) https://doi.org/10.1016/j.softx.2015.06.001

[42] Wassenaar, T.A., Pluhackova, K., Böckmann, R.A., Marrink, S.J., Tieleman, D.P.: Going backward: A flexible geometric approach to reverse transformation from coarse grained to atomistic models **10**(2), 676–690 https://doi.org/10.1021/ct400617g . Publisher: American Chemical Society. Accessed 2024-12-23

[43] Phillips, J.C., Hardy, D.J., Maia, J.D.C., Stone, J.E., Ribeiro, J.V., Bernardi, R.C., Buch, R., Fiorin, G., Hénin, J., Jiang, W., McGreevy, R., Melo, M.C.R., Radak, B.K., Skeel, R.D., Singharoy, A., Wang, Y., Roux, B., Aksimentiev, A., Luthey-Schulten, Z., Kalé, L.V., Schulten, K., Chipot, C., Tajkhorshid, E.: Scalable molecular dynamics on CPU and GPU architectures with NAMD **153**(4),

044130 https://doi.org/10.1063/5.0014475 . Accessed 2024-12-23

[44] David, C.C., Jacobs, D.J.: Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins. In: Methods in Molecular Biology vol. 1084, pp. 193–226 (2014). https://doi.org/10.1007/978-1-62703-658-0_11 . https://link.springer.com/10.1007/978-1-62703-658-0_11

[45] Palma, J., Pierdominici-Sottile, G.: On the Uses of PCA to Characterise Molecular Dynamics Simulations of Biological Macromolecules: Basics and Tips for an Effective Use. Chemphyschem: A European Journal of Chemical Physics and Physical Chemistry **24**(2), 202200491 (2023) https://doi.org/10.1002/cphc.202200491

[46] Bock, L.V., Grubmüller, H.: Effects of cryo-em cooling on structural ensembles. Nature Communications **13**(11), 1709 (2022) https://doi.org/10.1038/s41467-022-29332-2

[47] Schwab, J., Kimanius, D., Burt, A., Dendooven, T., Scheres, S.H.W.: DynaMight: estimating molecular motions with improved reconstruction from cryo-EM images. Nature Methods **21**(October) (2024) https://doi.org/10.1038/s41592-024-02377-5

[48] Chen, M., Ludtke, S.J.: Deep learning-based mixed-dimensional Gaussian mixture model for characterizing variability in cryo-EM. Nature Methods **18**(8), 930–936 (2021) https://doi.org/10.1038/s41592-021-01220-5 arXiv:2101.10356

[49] Punjani, A., Fleet, D.J.: 3D variability analysis: Resolving continuous flexibility and discrete heterogeneity from single particle cryo-EM. Journal of Structural Biology **213**(2), 107702 (2021) https://doi.org/10.1016/j.jsb.2021.107702 . Accessed 2023-08-25

[50] Punjani, A., Fleet, D.J.: 3DFlex: determining structure and motion of flexible proteins from cryo-EM. Nature Methods **20**(6), 860–870 (2023) https://doi.org/10.1038/s41592-023-01853-8 . Number: 6 Publisher: Nature Publishing Group. Accessed 2023-08-25

[51] Dashti, A., Schwander, P., Langlois, R., Fung, R., Li, W., Hosseinizadeh, A., Liao, H.Y., Pallesen, J., Sharma, G., Stupina, V.A., Simon, A.E., Dinman, J.D., Frank, J., Ourmazd, A.: Trajectories of the ribosome as a Brownian nanomachine. Proceedings of the National Academy of Sciences **111**(49), 17492–17497 (2014) https://doi.org/10.1073/pnas.1419276111 . Publisher: Proceedings of the National Academy of Sciences. Accessed 2023-08-25

[52] Dashti, A., Mashayekhi, G., Shekhar, M., Ben Hail, D., Salah, S., Schwander, P., Georges, A., Singharoy, A., Frank, J., Ourmazd, A.: Retrieving functional pathways of biomolecules from single-particle snapshots. Nature Communications **11**(1), 4734 (2020) https://doi.org/10.1038/s41467-020-18403-x . Number:

1 Publisher: Nature Publishing Group. Accessed 2023-08-25

[53] Herreros, D., Lederman, R.R., Krieger, J.M., Jiménez-Moreno, A., Martínez, M., Myška, D., Strelak, D., Filipovic, J., Sorzano, C.O.S., Carazo, J.M.: Estimating conformational landscapes from Cryo-EM particles by 3D Zernike polynomials. Nature Communications **14**(1) (2023) https://doi.org/10.1038/s41467-023-35791-y

[54] Herreros, D., Kiska, J., Ramírez-Aportela, E., Filipovic, J., Carazo, J.M., Sorzano, C.O.S.: Zart: A novel multiresolution reconstruction algorithm with motion-blur correction for single particle analysis. Journal of Molecular Biology **435**(9), 168088 (2023) https://doi.org/10.1016/j.jmb.2023.168088 . New Frontier of Cryo-Electron Microscopy Technology

[55] McInnes, L., Healy, J., Saul, N., Großberger, L.: Umap: Uniform manifold approximation and projection. Journal of Open Source Software **3**(29), 861 (2018) https://doi.org/10.21105/joss.00861

[56] Herreros, D., Mata, C.P., Noddings, C., Irene, D., Krieger, J., Agard, D.A., Tsai, M.-D., Sorzano, C.O.S., Carazo, J.M.: Real-space heterogeneous reconstruction, refinement, and disentanglement of CryoEM conformational states with HetSIREN. Nature Communications **16**(1) (2025) https://doi.org/10.1038/s41467-025-59135-0

[57] Zhong, E.D., Bepler, T., Berger, B., Davis, J.H.: Cryodrgn: reconstruction of heterogeneous cryo-em structures using neural networks. Nature methods **18**(2), 176–185 (2021)

[58] Vuillemot, R., Mirzaei, A., Harastani, M., Hamitouche, I., Fréchin, L., Klaholz, B.P., Miyashita, O., Tama, F., Rouiller, I., Jonic, S.: Mdspace: Extracting continuous conformational landscapes from cryo-em single particle datasets using 3d-to-2d flexible fitting based on molecular dynamics simulation. Journal of Molecular Biology **435**(9), 167951 (2023) https://doi.org/10.1016/j.jmb.2023.167951 . New Frontier of Cryo-Electron Microscopy Technology

[59] Vuillemot, R., Harastani, M., Hamitouche, I., Jonic, S.: Mdspace and mdtomo software for extracting continuous conformational landscapes from datasets of single particle images and subtomograms based on molecular dynamics simulations: Latest developments in continuousflex software package. International Journal of Molecular Sciences **25**(1) (2024) https://doi.org/10.3390/ijms25010020

[60] Li, Y., Zhou, Y., Yuan, J., Ye, F., Gu, Q.: CryoSTAR: leveraging structural priors and constraints for cryo-EM heterogeneous reconstruction. Nature Methods **21**(12), 2318–2326 (2024) https://doi.org/10.1038/s41592-024-02486-1

[61] Singer, A., Yang, R.: Alignment of density maps in wasserstein distance. Biological Imaging **4**, 5 (2024)

[62] Langfield, C., Carmichael, J., Wright, G., Andén, J., Singer, A.: Aspire: Cryo-electron microscopy image processing in python

[63] Martinsson, P.-G., Tropp, J.A.: Randomized numerical linear algebra: Foundations and algorithms. Acta Numerica **29**, 403–572 (2020)

[64] Paszke, A.: Pytorch: An imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703 (2019)

[65] Woollard, G., Zhou, W., Thiede, E.H., Lin, C., Grigorieff, N., Cossio, P., Dao Duc, K., Hanson, S.M.: InstaMap : instant-NGP for cryo-EM density maps. Acta Crystallographica Section D Structural Biology **81**(4), 147–169 (2025) https://doi.org/10.1107/S2059798325002025

[66] Cossio, P., Hummer, G.: Bayesian analysis of individual electron microscopy images: Towards structures of dynamic and heterogeneous biomolecular assemblies. Journal of structural biology **184**(3), 427–437 (2013)

[67] Flamary, R., Courty, N., Gramfort, A., Alaya, M.Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N.T.H., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D.J., Tavenard, R., Tong, A., Vayer, T.: Pot: Python optimal transport. Journal of Machine Learning Research **22**(78), 1–8 (2021)

[68] Peyré, G., Cuturi, M.: Computational optimal transport. Foundations and Trends in Machine Learning **11**(5-6), 1–257 (2019) https://doi.org/10.1561/2200000073 arXiv:1803.00567

[69] Santambrogio, F.: Optimal Transport for Applied Mathematicians. Progress in Nonlinear Differential Equations and Their Applications, vol. 87. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-20828-2 . https://link.springer.com/10.1007/978-3-319-20828-2

[70] Grave, E., Joulin, A., Berthet, Q.: Unsupervised alignment of embeddings with Wasserstein Procrustes. AISTATS 2019 - 22nd International Conference on Artificial Intelligence and Statistics **89** (2020) arXiv:1805.11222

[71] Adamo, D., Corneli, M., Vuillien, M., Vila, E., Adamo, D., Corneli, M., Vuillien, M., Vila, E., Adamo, D., Corneli, M., Vuillien, M., Vila, E.: An in depth look at the Procrustes-Wasserstein distance: properties and barycenters. HAL **hal-049362** (2025)

[72] Ryner, M., Kronqvist, J., Karlsson, J.: Globally solving the Gromov-Wasserstein problem for point clouds in low dimensional Euclidean spaces. Advances in Neural Information Processing Systems **36**(NeurIPS 2023) (2023) arXiv:2307.09057

[73] Parthasarathy, K.R.: Extreme points of the convex set of joint probability

distributions with fixed marginals. Proceedings of the Indian Academy of Sciences: Mathematical Sciences **117**(4), 505–515 (2007) https://doi.org/10.1007/s12044-007-0042-9

[74] Herreros, D., Lederman, R.R., Krieger, J., Jimenez-Moreno, A., Martinez, M., Myska, D., Strelak, D., Filipovic, J., Bahar, I., Carazo, J.M., Sanchez, C.O.S.: Approximating deformation fields for the analysis of continuous heterogeneity of biological macromolecules by 3D Zernike polynomials. IUCrJ **8**, 992–1005 (2021) https://doi.org/10.1107/S2052252521008903

[75] Herreros, D., Lederman, R.R., Krieger, J.M., Jiménez-Moreno, A., Martínez, M., Myška, D., Strelak, D., Filipovic, J., Sorzano, C.O.S., Carazo, J.M.: Estimating conformational landscapes from Cryo-EM particles by 3D Zernike polynomials. Nature Communications **14**(1) (2023) https://doi.org/10.1038/s41467-023-35791-y

[76] Herreros, D.: xmipp-metadata: Xmipp Metadata Handler (version 1.1.4). https://pypi.org/project/xmipp-metadata/. Xmipp Metadata handling functionality with image binary accession in Python (2024)

[77] Zhao, R.: Convergence rate analysis of the multiplicative gradient method for PET-type problems. Operations Research Letters **51**(1), 26–32 (2023) https://doi.org/10.1016/j.orl.2022.11.010 arXiv:2109.05601

[78] Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2009)

[79] Diamond, S., Boyd, S.: CVXPY: A Python-embedded modeling language for convex optimization. Journal of Machine Learning Research **17**(83), 1–5 (2016)

[80] Agrawal, A., Verschueren, R., Diamond, S., Boyd, S.: A rewriting system for convex optimization problems. Journal of Control and Decision **5**(1), 42–60 (2018)