




OPEN

Bayesian reweighting of biomolecular structural ensembles using heterogeneous cryo-EM maps with the cryoENsemble method

Tomasz Włodarski^{1,2}, Julian O. Streit¹, Alkistis Mitropoulou¹, Lisa D. Cabrita¹, Michele Vendruscolo⁴ & John Christodoulou^{1,3}

Cryogenic electron microscopy (cryo-EM) has emerged as a powerful method for the determination of structures of complex biological molecules. The accurate characterisation of the dynamics of such systems, however, remains a challenge. To address this problem, we introduce cryoENsemble, a method that applies Bayesian reweighting to conformational ensembles derived from molecular dynamics simulations to improve their agreement with cryo-EM data, thus enabling the extraction of dynamics information. We illustrate the use of cryoENsemble to determine the dynamics of the ribosome-bound state of the co-translational chaperone trigger factor (TF). We also show that cryoENsemble can assist with the interpretation of low-resolution, noisy or unaccounted regions of cryo-EM maps. Notably, we are able to link an unaccounted part of the cryo-EM map to the presence of another protein (methionine aminopeptidase, or MetAP), rather than to the dynamics of TF, and model its TF-bound state. Based on these results, we anticipate that cryoENsemble will find use for challenging heterogeneous cryo-EM maps for biomolecular systems encompassing dynamic components.

Keywords Statistical inference, Molecular dynamics simulations, Cryo-EM, Trigger factor, Structural biology

Describing the dynamics of complex macromolecular systems presents significant challenges^{1–6}. The main techniques to achieve this goal have been nuclear magnetic resonance (NMR) spectroscopy and single-molecule fluorescence methods^{2–5,7,8}. More recently, technological and methodological advancements in single-particle cryogenic electron microscopy (cryo-EM), including improvements in electron detectors, image processing software and motion correction algorithms, have offered a new means to investigate protein dynamics^{9–12}. By recording large numbers of two-dimensional (2D) projection images of biomolecules captured by flash freezing in various compositional or conformational states, cryo-EM offers a glimpse into the diverse conformational landscape of dynamic macromolecular complexes.

A variety of computational methods for fitting and refining atomic models against single-particle cryo-EM maps have been developed^{13–19}. These methods include rigid body fitting of available X-ray structures into low-resolution cryo-EM maps²⁰, incorporation of protein flexibility through normal mode analysis²¹ and flexible fitting^{22–24}, and density-based molecular dynamics (MD) simulations^{25–28}. Despite these advances, however, characterising the conformational heterogeneity underlying the dynamics of the systems under observation in cryo-EM samples remains a significant challenge^{19,29,30}. Structural regions that display conformational heterogeneity can be incorrectly aligned and then erroneously averaged with other images, causing these regions to become blurred, or even invisible, in the reconstruction, leading to lower final resolution and less detailed or incomplete maps. Separating these regions into homogeneous subsets during post-processing can be achieved, for example,

¹Institute of Structural and Molecular Biology, University College London, Gower Street, London WC1E 6BT, UK. ²Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Pawinskiego 5a, 02-106 Warsaw, Poland. ³Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK. ⁴Centre for Misfolding Diseases, Yusuf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK. ✉email: t.wlodarski@ucl.ac.uk; twlodarski@ibb.edu.pl

by using heterogeneous refinement with maximum likelihood classification methods³¹. This approach, however, tends to work better for discrete heterogeneity when the system can be characterised by a finite number of states. For continuous conformational heterogeneity, other methods have been developed that can be applied to either 2D particle images or 3D maps³², including focus refinement, where a mask is applied to different regions of the structure³³, multi-body refinement³⁴, manifold embedding³⁵ or deep neural networks^{36,37}.

Typically, to generate dynamical descriptions, structural models can be fine-tuned with experimental data^{38–43}. This approach, however, presents significant challenges as the experimental data are affected by a combination of the experimental errors and approximations included in post-processing into molecular simulations. As a result, different conformations can lead to a similar agreement with experimental observables, particularly when the data are incomplete and noisy, or when the forward model is dependent on many approximations, such as being based only on distances or angles between atoms to back-calculate experimental properties. Solutions that combine structural information from various experimental techniques (e.g. NMR spectroscopy, cryo-EM, small-angle X-ray scattering (SAXS)) with computational methods (e.g. molecular dynamics) and Bayesian inference have been proposed to produce structural ensembles^{38–41,44}. This integrative structural biology approach has been applied to many biological systems^{42,43}. Bayesian inference can be applied during MD simulations by adding a bias energy term to constrain simulations to sample conformations in agreement with experimental data⁴³, or it can be applied a posteriori when the experimental data are used to reweight the MD ensemble. The utility of these methods depends on the nature of the system under study, as well as the available experimental data⁴⁵.

Bayesian methods incorporating cryo-EM 2D particle images or 3D maps have been used within the integrative modelling framework^{40,46,47} and during MD simulations to enforce the agreement with experimental data^{39,48}. A notable example is EMMI³⁹ (Electron Microscopy MetaInference), which generates a new structural ensemble during MD simulations by incorporating 3D maps as restraints and modelling the errors present in the map. Here, we describe cryoENsemble, a computational approach that instead combines molecular dynamics simulations with Bayesian reweighting utilising cryo-EM maps (Fig. 1). We present it both in a standard mode and in an iterative mode, in which we repeatedly apply reweighting to a more refined structural ensemble. These methods allow the interpretation of discrete, continuous and compositional heterogeneity from cryo-EM maps to accurately describe the underlying structural ensembles. To accomplish this, we adapted and extended the Bayesian Inference Of ENsembles (BioEn) method⁴⁹, which uses various experimental data (e.g. NMR, SAXS, DEER) to refine structural ensembles from MD simulations. We first tested the cryoENsemble method and its iterative extension with synthetic cryo-EM maps from two well-characterised systems, namely adenylate kinase (ADK) and ribosomal nascent chain complex (RNC) (Supplementary Fig. 1). We could effectively reweight the structural ensembles in both cases, capture important structural and dynamic features and, at the same time, account for variations in resolution and noise levels present in the maps that correspond to discrete and continuous cryo-EM heterogeneity.

Next, we applied iterative cryoENsemble to the cryo-EM map of the ribosome-bound state of trigger factor (TF) stabilised by the presence of peptide deformylase (PDF) and methionine aminopeptidase (MetAP) (Supplementary Fig. 1). TF is the only ribosome-associated chaperone in bacteria, whereas PDF and MetAP are essential enzymes involved in the co-translational removal of formylated methionine in nascent protein chains, and both bind in the proximity of the ribosomal exit tunnel⁵⁰. Despite intensive research^{51–56}, the detailed role of TF in the co-translational folding process remains incompletely understood due to the experimental challenges presented by its dynamic nature, even in its ribosome-bound state, and only low-resolution or incomplete cryo-EM maps, which often encompass merely the ribosome-binding domain (RBD)^{57–59} are available.

By using all-atom MD simulations combined with cryoENsemble, we provide insights into the dynamics of TF, as captured within this cryo-EM map and explain the additional density present around TF. Our findings indicate that an ensemble of TF structures obtained with MD can better explain cryo-EM maps compared to a single model. Furthermore, using cryoENsemble, we confirmed that the additional density localised close to TF is not due to the dynamics of TF, as was initially hypothesised⁵⁷. Instead, by fitting MetAP to this unaccounted density, we found a compelling overlap, further confirming that this density stems from a novel binding site of the MetAP, as suggested recently⁵⁰. Based on this observation, we set up another MD simulation with MetAP bound to TF. The obtained structural ensemble of TF + MetAP and the previously generated TF ensemble were used for the final cryoENsemble reweighting. By combining these two ensembles, we aimed to improve our structural ensemble further and analyse the compositional heterogeneity of the cryo-EM map. We obtained a reweighted minimal ensemble that combined TF and MetAP, and found that the MetAP population in this map is ~40%.

Overall, we demonstrate that cryoENsemble can extract otherwise elusive information about populations of the macromolecular states and their dynamics from heterogeneous cryo-EM data. It also proves valuable in modelling biomolecular complexes when it is challenging to assign the regions of density due to their dynamics or structural changes upon binding, making it a much-needed addition to the structural biology toolbox.

Results

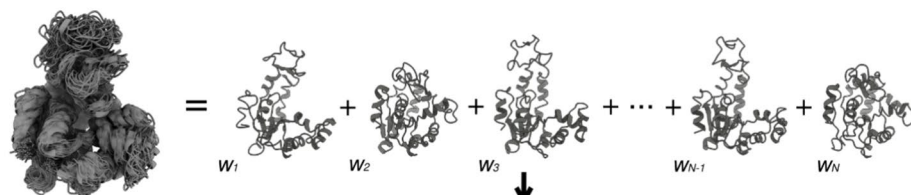
The cryoENsemble method for Bayesian reweighting with cryo-EM maps

To derive an ensemble of structures, each with a corresponding set of weights that adequately represent the experimental data, we based cryoENsemble on the BioEn method^{49,60,61} and incorporated a single-particle cryo-EM data framework. The BioEn algorithm uses Bayes' theorem to define the posterior probability as a function of the statistical weights of each member of the structural ensemble (w_i), where i is the index of the member, given the experimental data (D) and the prior knowledge about the system (I)

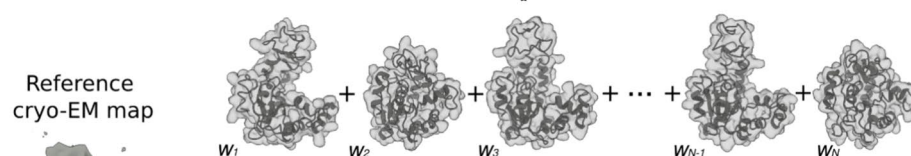
$$P(w|D, I) \approx P(D|w, I)P(w, I) \quad (1)$$

A) Standard reweighting

Prior structural ensemble

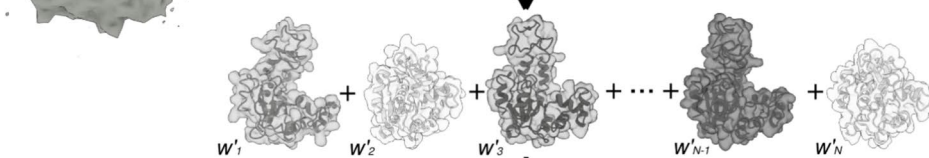


Map generation



Reference cryo-EM map

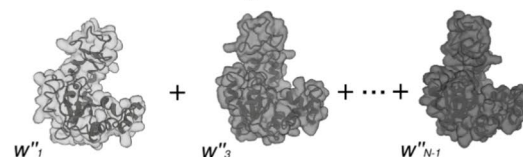
Bayesian reweighting



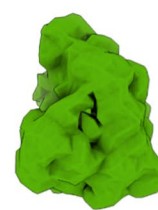
B) Iterative reweighting

Selection (based on weights)

Bayesian reweighting

If $\chi^2_{i+1} < \chi^2_i$ 

Posterior structural ensemble



Posterior cryo-EM map

Figure 1. Schematic illustration of the standard cryoENsemble method and its iterative extension. The input includes a structural ensemble (depicted in grey), typically obtained from molecular dynamics simulations, and a cryo-EM map of the biological system under investigation. Each model from the prior structural ensemble is fitted into the reference cryo-EM density before the cryoENsemble calculations. Subsequently, density is generated for each structure and starting weights (w_i) are assigned. In the standard mode, Bayesian reweighting generates new weights (w'_i), indicated in the figure by varying shades of grey, for each structure in the ensemble (posterior structural ensemble) as well as posterior density for the system (see Methods). In the iterative mode, following the standard reweighting, a sub-ensemble of structures that meets weight criteria is selected for another round of Bayesian reweighting, which generates new weights (w''_i) for the sub-ensemble. This process is repeated until the agreement with the experimental data decreases (reflected by an increase in χ^2). The iterative mode returns the minimal set of structures that maintains good agreement with the experimental cryo-EM data and posterior density.

In the context of cryo-EM data, the experimental data points are defined as a set of voxels with a density exceeding a predetermined threshold value, the latter established based on the noise levels present in the data (see Methods). To take into account the resolution anisotropy that can be present in the map, we can filter the map based on local resolution prior to reweighting. The likelihood function, $P(D|w, I)$ assesses the probability of observing a given set of experimental data (D), considering the actual ensemble of structures and their corresponding statistical weights (w). The prior probability term, $P(w|I)$, encapsulates the knowledge about the structural ensemble and weights (w). This knowledge is typically derived from the molecular dynamics ensemble prior to the incorporation of the experimental data. The prior can be constructed in several ways, but keeping in line with the BioEn methodology, we utilise Kullback–Leibler (KL) divergence (S_{KL})

$$P(w, I) \approx \exp(-\theta S_{KL}) \quad (2)$$

where S_{KL} is defined as $S_{KL} = \sum_{i=1}^M w_i \ln \frac{w_i}{w_i^0}$ and both reference (w_i^0) and refined weights (w_i) are normalised and positive, and M is the size of the ensemble. Generally, the reference weights of the prior structural ensemble (w_i^0) are selected from the uniform distribution, though they can also be set according to populations derived from biased MD simulations (e.g. from metadynamics⁶²). An additional hyperparameter, θ , describes our confidence in the initial structural ensemble. A high value of θ indicates high confidence in the MD simulations and generated ensemble, causing the refined weights (w_i) to stay close to the initial ones (w_i^0). Conversely, a low value of θ , suggests that the initial ensemble may be far from optimal, allowing the weights (w_i) to deviate significantly from the starting one (w_i^0). θ is automatically selected during optimisation based on the developed automatic L-curve analysis.

The likelihood function is modelled via a Gaussian distribution⁶³

$$P(D|w, I) \propto \exp\left(-\sum_{n=1}^N \frac{[\rho_n^0 - \alpha \sum_{i=1}^M w_i \rho_n^i(\sigma)]^2}{2\sigma_L^2}\right) \quad (3)$$

where ρ_n^0 represents the experimental/reference density from the n -th voxel of our map, whereas $\rho_n^i(\sigma)$ is simulated density from the same voxel generated from the i -th model of the structural ensemble with the use of Gaussian functions with the width equal to σ , which is a nuisance parameter (see Methods). This likelihood function contains two additional parameters: the variance of the Gaussian likelihood (σ_L^2), which is equivalent to the experimental error, and the scaling factor (α). We approximate σ_L^2 using the variance of noise distribution outside of the molecular system density, while α and σ are estimated during the reweighting. To determine the optimal value of θ , we perform calculations over a range of θ values and use an automatic L-curve analysis with the Kneedle algorithm⁶⁴. This allows the selection of a θ value that yields good agreement with experimental data (low χ^2) and also prevents overfitting (maintaining a small difference from the distribution of the initial weights).

Having defined both the likelihood and prior functions, we can express the negative log-posterior function, which we will minimise to find the optimal weights, along with nuisance parameters, using the log-weights optimisation as encoded in BioEn

$$-\log P(w|D, I) = \theta \sum_{i=1}^M w_i \ln \frac{w_i}{w_i^0} + \sum_{n=1}^N \frac{[\rho_n^0 - \alpha \sum_{i=1}^M w_i \rho_n^i(\sigma)]^2}{2\sigma_L^2} \quad (4)$$

The execution of standard cryoENsemble calculations yields optimal (non-zero) weights for every structure in our structural ensemble, along with the values of θ , σ and α . A schematic of our methodology is shown in Fig. 1A. Additionally, we extended the cryoENsemble standard run into the iterative mode, where we select a sub-ensemble of structures from the initial run and perform subsequent rounds of cryoEnsemble. In this mode, we select $N_{\text{eff}} \cdot M$ structures with the highest weight during each round, where N_{eff} is an effective sample size ($N_{\text{eff}} = \exp(S_{KL})$). This iterative process continues until the agreement with the experimental data decreases (reflected by an increase in χ^2) (Fig. 1B).

Both approaches have been tested using two extensive synthetic cryo-EM datasets from the adenylate kinase and ribosomal nascent chain complex (Supplementary Fig. 1) and showed that they can accurately reproduce the structural properties of the underlying conformational ensembles from the heterogeneous and noisy cryo-EM data.

CryoENsemble reweighting of a structural ensemble of adenylate kinase

Adenylate kinase (ADK) is an enzyme that catalyses the transfer of a phosphoryl group from ATP to AMP. This enzyme comprises three domains (CORE, NMP and LID) and undergoes a significant conformational change from open (apo) to closed (holo) conformation upon ligand binding, with RMSD = 7.16 Å (Supplementary Fig. 1A). Both states have been structurally characterised by X-ray crystallography (PDB IDs: 1AKE for the closed⁶⁵ and 4AKE for open⁶⁶ conformation). To test our method, we generated synthetic density maps based on these structures using different populations of each state, resolution and noise levels, culminating in a total of 66 maps for reweighting (Supplementary Fig. 2). The prior structural ensemble consisted of structures obtained from two structure-based all-atom MD simulations (see Methods).

For each ADK dataset, consisting of a structural ensemble and a selected set of voxels from a combination of reference map and simulated map (Supplementary Fig. 3), we ran the cryoENsemble reweighting method both in a standard and iterative approach. Initially, we assessed the effectiveness of our methodology in reproducing the reference population of the open state used to generate the reference maps (Fig. 2A). Our findings from the standard cryoENsemble run indicate consistency across all density maps, which decreases with both a reduction in resolution (10 Å) and an increase in the noise level (10%) (Fig. 2A). The consistency was lower for reference

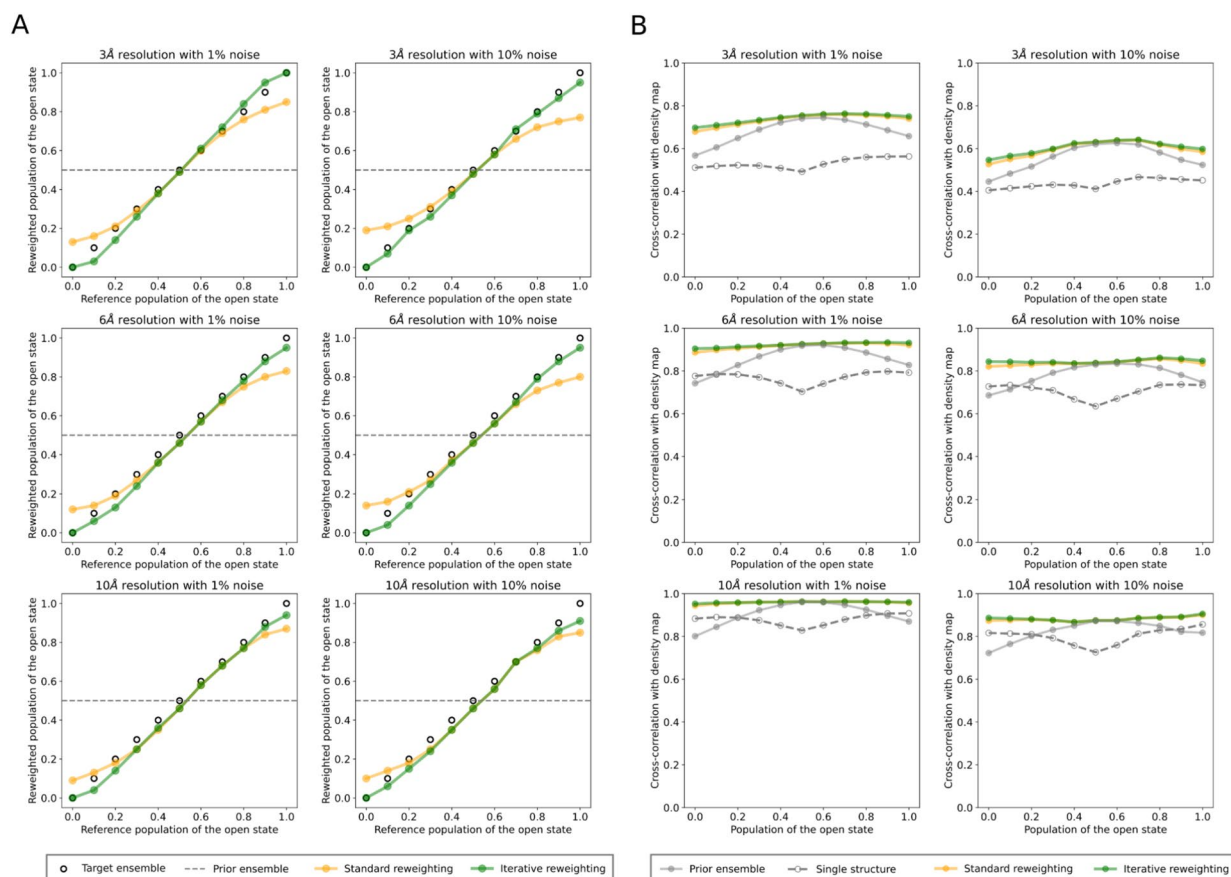


Figure 2. CryoENSEMBLE reweighting of ADK. **(A)** Open state populations calculated from cryoENSEMBLE reweighting of the ADK dataset. The open state populations obtained after structural ensemble reweighting for each ADK dataset are shown in orange for the standard mode and green for the iterative mode, along with the target values (black circle). **(B)** Correlations between reference maps and posterior maps upon cryoENSEMBLE reweighting of the ADK dataset in standard (orange) and iterative mode (green), including values for the prior ensemble and the best single structure. The datasets vary in resolution, noise level, and reference populations of the open state.

maps with a very small (≤ 0.2) or very large (≥ 0.8) population of the open state. This disparity is likely due to our prior structural ensemble equally representing the open and closed states, a significant deviation from these reference maps. This discrepancy was resolved in the iterative mode, where by selecting progressively smaller subsets (Supplementary Fig. 4), we managed to achieve the correct population across all analysed open states (Fig. 2A).

Following these calculations, we generated a reweighted and averaged cryo-EM map for each dataset and compared it with the reference density map to evaluate the impact of the reweighting (Fig. 2B). For the comparison, we applied two metrics recommended by the 2019 Cryo-EM Model Challenge⁶⁷: the correlation coefficient (CC) and Segment Based Manders' Overlap Coefficient (SMOC)⁶⁸ as they capture global and local aspects of similarity between maps.

The correlation coefficient calculated between the reference maps and posterior densities revealed that the effect of reweighting is evident across all open state populations in each ADK dataset (Fig. 2B). The correlation coefficient for medium and low-resolution maps reaches values of up to 0.9 or 0.8 for low (1%) and high (10%) noise levels, respectively. Higher resolution maps (3 Å), which contain more detail, present a more significant challenge in reweighting the structural ensemble to achieve high correlation coefficients, in particular when high noise levels (10%) are present in the data (Fig. 2B). However, reweighting consistently yields higher correlation coefficients than those obtained with the prior weights (for 3 Å with high noise levels (10%) on average $CC = 0.597$ and $CC = 0.557$ for posterior and prior weights, respectively). We also compared our reweighting results with the correlation coefficients derived from maps generated based on the best single structure fit. In the majority of the cases, the entire ensemble obtained after reweighting provides a more accurate representation of the map than any individual structure (Fig. 2B). Interestingly, when lowering the map resolution, e.g. from 6 to 10 Å, the quality of a single structure fit increases and in the cases of either entirely open or closed state it is higher than the CC of the prior ensemble, and can even equal the value of the reweighted ensemble. The single structure CC deteriorates when maps are close to an equal mixture of open and closed states. Although using iterative mode does not significantly improve the correlation coefficient (Fig. 2B), this method selects smaller sub-ensembles with at least as good agreement with the data as the whole reweighted ensemble (Supplementary Fig. 4). These

observations show that cryoENsemble not only can provide a reweighted structural ensemble but also inform on when a single structure may be sufficient to describe a cryoEM map satisfactorily.

For the second score, we used SMOC, which captures the local similarity between the reference map and the fitted model. We calculated the average SMOC score across all residues and models from the MD ensemble and observed that it improved after reweighting, in particular for the entirely open and closed states (Supplementary Fig. 5). With iterative mode, we observed small further improvement, especially again for maps capturing completely open or close state. Altogether, for both global and local metrics, we see a clear effect of the reweighting on the quality of the ADK structural ensemble, with the iterative mode having the highest impact on the estimation of the correct populations. We subsequently analysed how the weights of each model from the structural ensemble were updated during the reweighting. We found that cryoENsemble shifted the weights from the uniform prior distribution to correctly capture the reference map open/closed state population in standard (Supplementary Figs. 6–11) and iterative (Supplementary Figs. 12–17) reweighting. Finally, a visual comparison of the prior and posterior average density maps alongside the reference maps shows the impact of the reweighting, especially pronounced for the open state maps where posterior maps combine only the open state conformations (Supplementary Figs. 18–20).

Overall, we have demonstrated the effectiveness of cryoENsemble in characterising discrete heterogeneity in cryo-EM maps. We derived weights that can generate a map in good agreement with the experimental data (Supplementary Figs. 18–20) and are able to describe the correct populations of each state (Fig. 2A). Our reweighted structural ensemble better explains the experimental data, using both global (Fig. 2B) and local metrics (Supplementary Fig. 5), than the starting ensemble. Furthermore, our method can also suggest when a single structure fitted into the reference map is insufficient, highlighting the necessity of using a structural ensemble for heterogeneous cryo-EM map fitting.

Reweighting the structural ensemble of the ribosome-bound nascent chain of FLN5-6

The second system we used for the test is the FLN5-6 ribosome nascent chain complex encompassing the immunoglobulin-like domain (FLN5 protein), captured during its biosynthesis on the bacterial 70S ribosome (Supplementary Fig. 1B). The FLN5 is the fifth filamin domain (residues 646–750) of the *Dictyostelium discoideum* gelation factor, and its co-translational folding has been extensively studied through a combination of experimental and computational techniques^{69–72}. The FLN5-6 nascent chain sequence also consists of the 31 amino-acid linker comprising the fragment of the subsequent filamin domain (FLN6) and the SecM stalling sequence⁷³. For our study, we obtained a starting structural ensemble of FLN5-6 based on the previous all-atom structure-based MD simulations⁷⁴ encompassing a diverse set of 100 structures of the FLN5-6 nascent chain (Supplementary Fig. 21A,C). Additionally, we created reference density maps based on random selections of 10 structures from this structural ensemble (Supplementary Fig. 21B) (see Methods section).

We applied the cryoENsemble protocol to the prior structural ensemble using the combined data from the reference map and maps generated from the structural ensemble. We used the entire structural ensemble (100 models) for the reweighting, including the ten conformations used to generate the reference density maps. Average density maps were generated based on the prior and posterior weights, and their correlation coefficients with the reference density maps were calculated (Fig. 3A). The prior ensemble, despite its significant structural heterogeneity, already displayed a good agreement with the reference density maps with average CC varying between 0.8 (3 Å maps) and 0.95 (10 Å maps) for 1% noise and from 0.47 (3 Å maps) to 0.84 (10 Å maps) for 10% noise (Fig. 3A). After standard reweighting, the correlation increased in all cases, reaching a value close to 1.0 for low noise levels (1%) and 0.9 for high noise levels (10%), with the only exception of the 3 Å maps, which, as in the ADK case, present a more significant challenge in reweighting the structural ensemble, in particular at the higher noise levels (10%) where CC reached 0.54 versus 0.47 with the prior weights. This difficulty is further apparent upon examining the extent of density in this highly noisy system (Supplementary Fig. 21B). Iterative reweighting improved the CC further, especially for the highest resolution maps. A comparison with maps generated based on a single structure shows that, in contrast to the ADK system, a single structure cannot represent the dynamic heterogeneity present in the nascent chain cryo-EM maps for any of the systems we tested (Fig. 3A). We also evaluated the reweighted ensemble using SMOC metrics (Supplementary Figs. 22), finding that the reweighting improved the agreement with experimental data both globally and locally in all cases. Additionally, in order to assess the structural similarity between the obtained reweighted ensemble and the ten structures used to generate the reference map, we used the Jensen-Shannon (JS) divergence. We found significantly closer matching values upon reweighting, especially upon iterative reweighting (Supplementary Fig. 23). The high initial correlation between the prior ensemble and reference map can result in relatively minor changes to the correlation coefficients after reweighting (Fig. 3A). However, we observed significant shifts from the uniform distribution of the weights of the prior structural ensembles due to reweighting (Fig. 3B) especially visible after iterative reweighting, which significantly narrowed down the structural ensembles, especially for the high-resolution maps (Supplementary Figs. 24, 25A). The weights of the ten models used for reference map generation (circled in Fig. 3B, Supplementary Fig. 24) are substantially higher than those of the remaining structures (e.g. 5.6% vs. 0.5% on average for 3 Å maps with 1% noise), a trend not significantly affected by the resolution of the density map or its noise level. These observations highlight the sensitivity of our method, which became particularly apparent when we analysed the entire dataset to determine how many of the ten models used to generate the reference map received the highest weight after the standard reweighting (Supplementary Fig. 25B). For high- and medium-resolution maps (3 and 6 Å), our method assigned the highest weights to the correct models in all datasets. While lower-resolution maps (10 Å) posed greater challenges, over half of the reference models were correctly identified to receive the top ten highest weights.

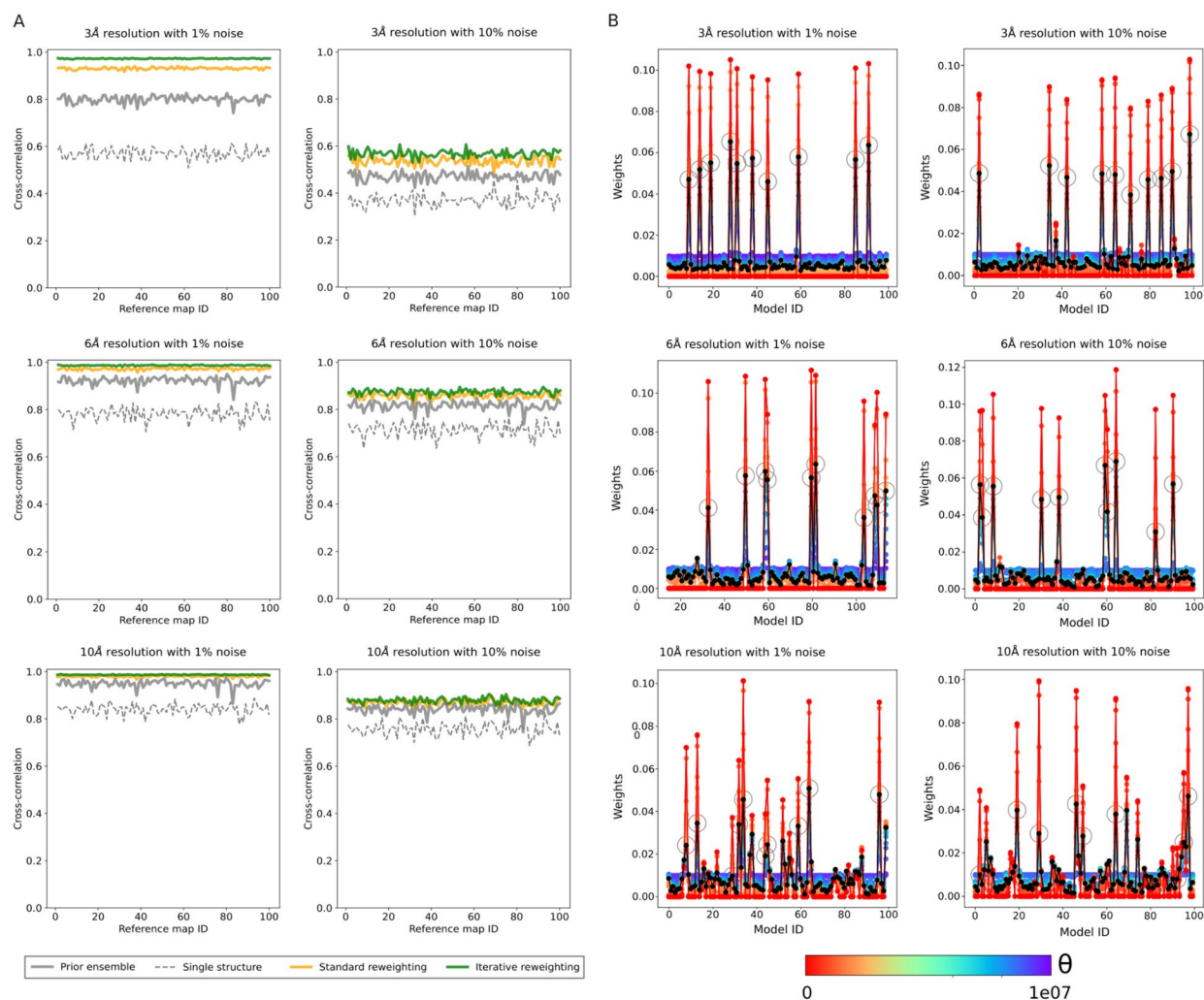


Figure 3. CryoENSEMBLE reweighting of the FLN5-6 nascent chain dataset. **(A)** Correlations between the reference maps and posterior maps upon standard and iterative cryoENSEMBLE reweighting of the FLN5-6 nascent chain dataset. Correlation coefficients calculated between the FLN5-6 nascent chain reference density maps and maps obtained before and after the reweighting, as well as the maps derived from the best single structure fitted into the reference density map. The 100 reference density maps (at resolutions of 3, 6, and 10 Å, and with noise levels of 1% and 10%) were generated based on ten randomly selected structures from the MD ensemble. **(B)** Weights obtained upon standard cryoENSEMBLE reweighting of the FLN5-6 nascent chain dataset. Examples of the reweighting process for FLN5-6 nascent chain based on the reference map (at resolutions of 3, 6, and 10 Å, and with noise levels of 1% and 10%). Weights are calculated with different theta (θ) values ranging from 0 to 10^7 , and with black lines, we depict optimal weights selected based on the L-curve analysis. Additionally, weights corresponding to the ten models used to generate the reference map are circled.

As the maps were generated based on a selection of an ensemble of 10 structures (target ensemble), we can use this synthetic data to test cryoENSEMBLE's capacity to retrieve not only structural ensembles that correspond to the cryoEM map but also the embedded dynamics. To analyse if local flexibility can be retrieved upon reweighting, we calculated the root-mean-square fluctuations (RMSF) for prior and reweighted ensembles and compared them to the target RMSF using root-mean-square deviation (Fig. 4). We observed that, in all cases apart from the very challenging map with 3 Å resolution and 10% noise, the description of the RMSF improved, especially with the use of iterative mode. Furthermore, we performed a principal component analysis (PCA) to study global dynamics for each ensemble and compared the three main principle components (PC1, PC2, PC3) with the target values. We found significant improvement for PC1 and PC2 (Supplementary Figs. 26 and 27), whereas PC3 proved more challenging to improve for some maps (Supplementary Fig. 28).

Overall, we have shown that cryoENSEMBLE can retrieve structural ensembles that globally and locally capture structural (CC, SMOC, JS divergence) and dynamical (RMSF, PCA) aspects ingrained in the cryo-EM map that represents continuous heterogeneity.

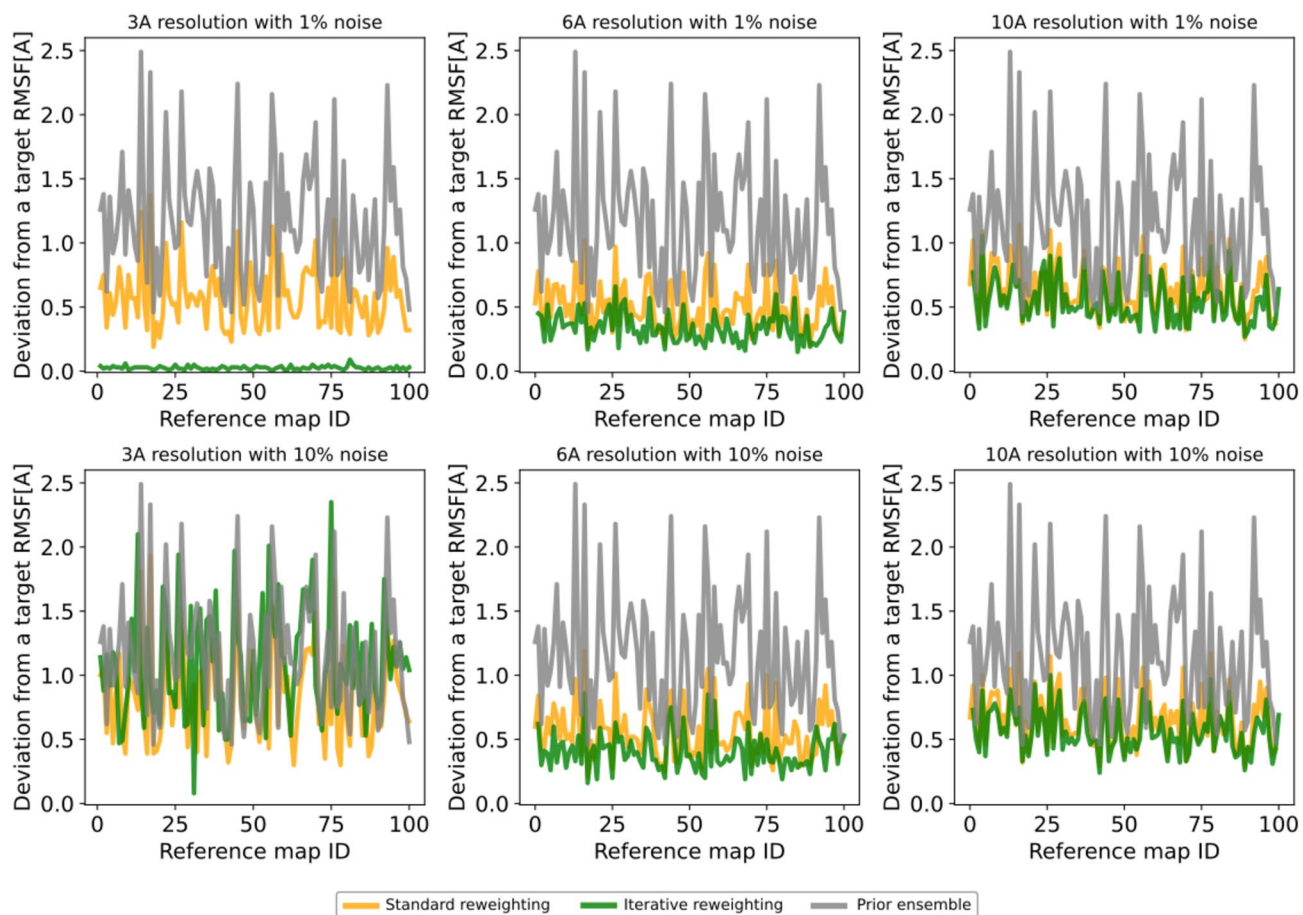


Figure 4. CryoENsemble reweighting can retrieve local dynamics from cryo-EM maps. The root-mean-square deviations calculated between the RMSF of the target ensemble and the RMSFs from the prior ensemble (in grey), standard reweighting (in orange), and iterative reweighting (in green).

Dynamics of the ribosome-bound TF in complex with PDF

Upon binding to the ribosome, TF remains highly dynamic, making it a challenging system for structural studies. We applied the iterative cryoENsemble methodology to the cryo-EM map, which represents the *E. coli* 70S ribosome in complex with PDF, TF and MetAP⁵⁰. PDF and MetAP also bind around the ribosomal exit tunnel and compete for the s uL22--uL32 protein region⁵⁹. MetAP additionally has a secondary binding site, which overlaps with the TF one^{50,59}. When TF is bound to the ribosome in the presence of PDF or MetAP, it exhibits reduced dynamics and is, therefore, easier to characterise via cryo-EM.

We exploited this and ran a long all-atom structure-based⁷⁵ MD simulation with TF bound to the surface of the 70S ribosome complexed with PDF (Methods). Despite the fact that the dynamics of TF is restricted in the MD simulations by the presence of the bound PDF, it remains mobile, in particular within the peptidyl-prolyl *cis-trans* isomerase (PPI-ase) domain region (Supplementary Fig. 29). We next reweighted the MD trajectory using iterative cryoENsemble and the available cryo-EM map (EMDB: 30611⁵⁰) (Supplementary Fig. 1C). Our initial ensemble was already in good agreement with the cryo-EM data, with a correlation coefficient (CC) of 0.68 and after four rounds of iterative reweighting, the agreement improved to CC=0.73, respectively (Supplementary Fig. 30). Moreover, as the reweighting process increased the weights of selected members of the ensemble (Fig. 5A, Supplementary Fig. 30), we identified the best-fit model within the density map, with CC=0.65 (Fig. 5B). The improvement of the agreement with the experimental data for the MD ensemble, upon the reweighting, underscores the significance of utilising a structural ensemble instead of a single model when analysing heterogeneous cryo-EM maps. The most significant change in the ensemble composition was evident in our clustering analysis (see Methods). We found that, of the five main clusters encompassing 77% of the total trajectory, only one had an increased population following the reweighting (cluster_4 from 7.7 to 12%). Cluster_2 remained at a similar level (12%), and the remaining three experienced decreased populations, particularly apparent for cluster_1 (Fig. 5A), comprising 40% of the MD trajectory, which dropped to 21.5% in the reweighted ensemble.

Interestingly, cluster_14 had the highest absolute increase in the population upon reweighting (from 0.8 to 12%). The three main clusters obtained from reweighting combine to 45.5% of the population (Fig. 5C). Altogether, these findings show that reweighting using cryoENsemble can significantly improve the quality of the MD ensemble and its agreement with the cryo-EM data. Importantly, the reweighting process is not a simple

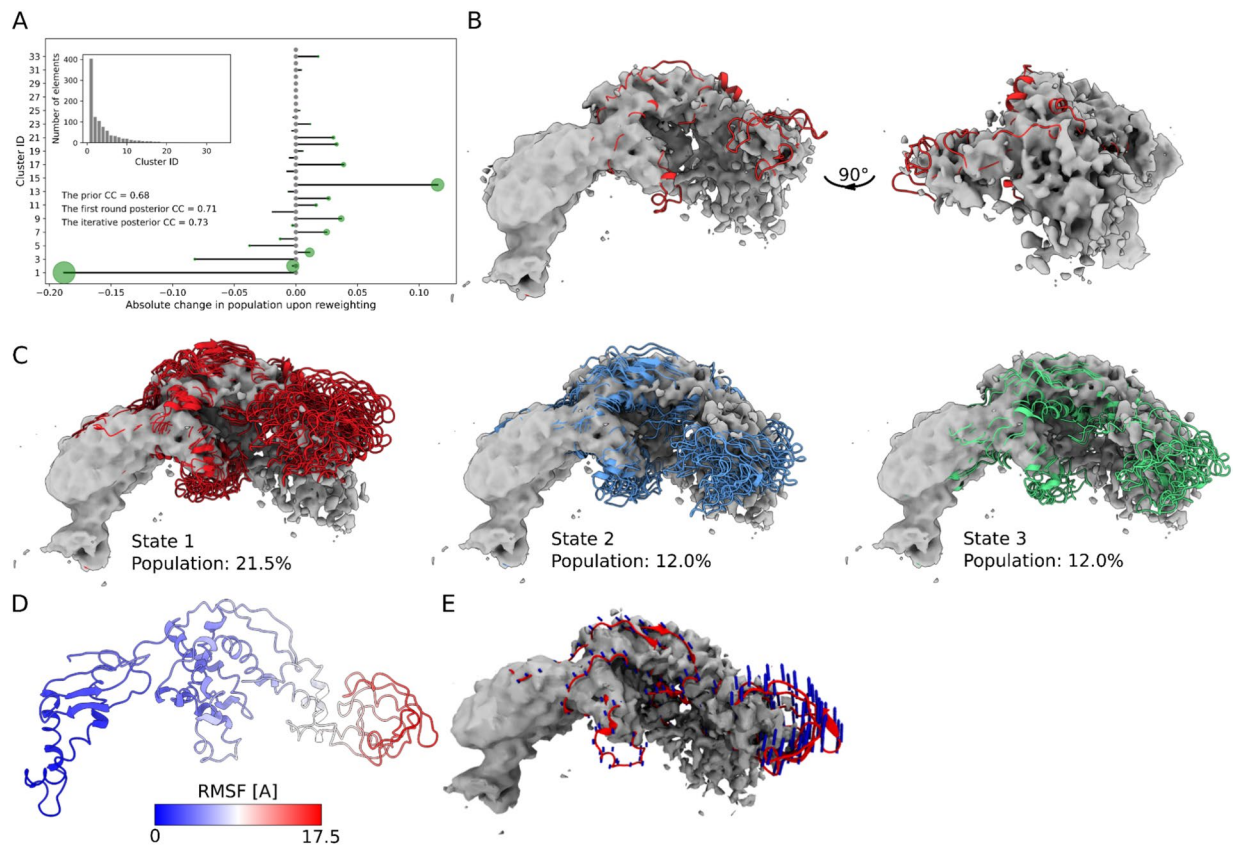


Figure 5. CryoENSEMBLE iterative reweighting of the TF dataset. **(A)** Analysis of the effect of reweighting on the weights of each cluster obtained from the MD simulations. The green circle size is proportional to the population of the cluster upon reweighting. **(B)** The structural model with the highest weight selected by cryoENSEMBLE (Supplementary Fig. 30) is visualised in two different orientations inside the cryo-EM map. **(C)** The three main states obtained upon reweighting corresponding to the cluster_1, cluster_2 and cluster_14, along with their populations. **(D)** RMSF calculated on the reweighted ensemble and mapped on the structure of TF. **(E)** Visualisation of the principal mode PC1 that captures the most dominant motion (indicated by blue arrows) within the reweighted ensemble.

increase of weights for all structures with high CC, as we found no correlation between new weights and corresponding CC scores (Supplementary Fig. 31); however, the structures that have the highest weights also tend to have high CC scores – supporting our observation that an ensemble explains heterogeneous cryo-EM data better than a single structure.

Additionally, when we compared the experimental cryo-EM map with the map obtained from the single best-fit model and the map representing the entire reweighted MD ensemble, the reweighted map is more similar to the cryo-EM when visualised at different density thresholds (Supplementary Fig. 32). This emphasises not only the necessity to use an ensemble of structures instead of a single structure to capture information from highly heterogeneous cryo-EM maps but also underscores the importance of reweighting.

Finally, we analysed the TF dynamics retrieved from the cryo-EM map using RMSF and the first principal component (PC1) from the PCA analysis of the reweighted ensemble. We found that the most dynamic region corresponds to the PPI domain (Fig. 5D), which is also visible in the three main structural states (Fig. 5C). By conducting PCA on the reweighted ensemble, we characterised the movement of the PPI domain embedded in the cryo-EM map further and found that it moves towards and away from the ribosome surface (Fig. 5E). These observations align with previous work, which depicted, through coarse-grained MD simulations, that this domain fluctuates between bound and unbound state⁵⁷. Altogether, these results show that cryoENSEMBLE captures both the structural ensemble and its possible dynamics from the corresponding cryo-EM map.

The unaccounted cryo-EM density corresponds to a TF-bound methionine aminopeptidase, not TF dynamics

The initial study of TF, MetAP and PDF assembly on the ribosome provided several low-resolution cryo-EM maps of the 70S ribosome in various configurations⁵⁹. Notably, in the cryo-EM map of MetAP-PDF-TF (12.2 Å, EMDB:9778), the MetAP density was unannotated, and an additional density near TF was attributed to the possible dynamics of TF. A subsequent study obtained a higher resolution cryo-EM map (4.1 Å) of the 70S ribosome with MetAP, PDF and TF with again additional density near TF, but now annotated as a tertiary binding site for MetAP⁵⁰.

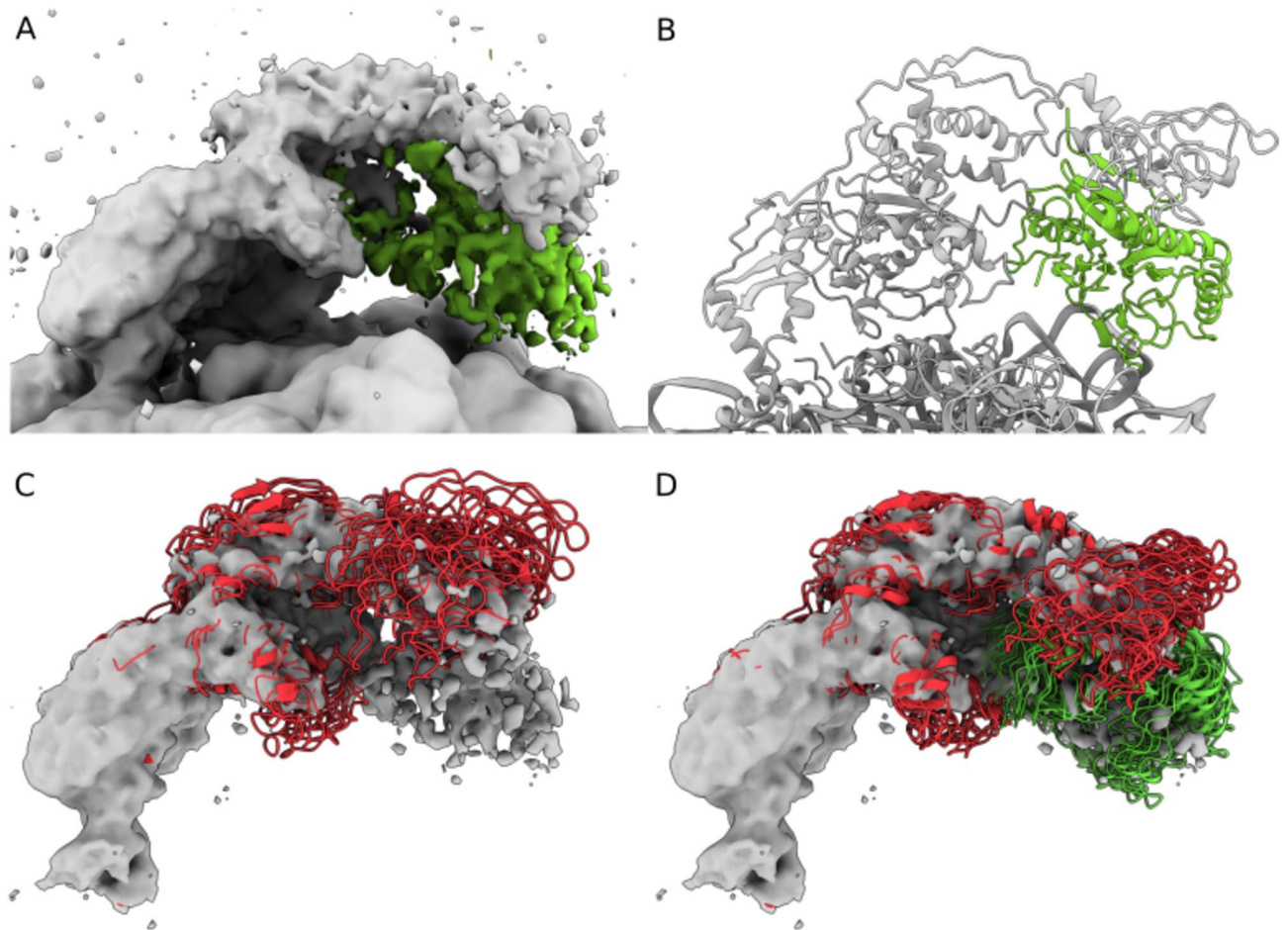


Figure 6. CryoENsemble iterative reweighting captures conformational and compositional variability in cryoEM maps. (A) The cryo-EM map (EMDB: 30611⁵⁰) with unaccounted density coloured in green. (B) The outcome of fitting the MetAP structure (PDB ID: 1MAT⁷⁷) into the unaccounted density (from (A)), presented along the 70S-Trigger factor structure (PDB ID: 7D80⁵⁰). (C) The main cluster of TF structures, with a population of 10%, obtained after the reweighting of combined structural ensembles (TF and TF + MetAP). (D) The main cluster of TF + MetAP structures, with a population of 12%, obtained after the reweighting of combined structural ensembles (TF and TF + MetAP).

Seeking to clarify the nature of this additional density, we took advantage of the unique combination of the MD simulations and cryoENsemble. After accounting for the TF structural ensemble obtained upon reweighting, we observed that there is still an unaccounted density present (Fig. 6A), which confirms the suggestion of a tertiary binding site for MetAP⁵⁰. To further validate this observation, we fitted the MetAP structure using a rigid-body procedure, starting with the orientation where the positively charged loops faced the ribosome surface, as indicated by biochemical studies to be a probable ribosome-binding mode⁷⁶, and found a compelling overlap (Fig. 6B, Supplementary Fig. 33).

Next, we ran another structure-based MD simulation using this TF + MetAP structure. The resulting structural ensemble was combined with the previous TF structural ensemble, and we conducted iterative reweighting with cryoENsemble. After five iterations (Supplementary Fig. 34), we obtained a reweighted ensemble with 79 structures and a correlation coefficient of 0.76, which is higher than when using only the TF ensemble. We identified two main states for the TF and TF + MetAP ensemble, finding that the population of the MetAP in the cryo-EM map is ~ 40% (Fig. 6C,D and Supplementary Fig. 35).

These findings demonstrate how MD simulations, in combination with cryoENsemble reweighting, can help explain unmodelled and unaccounted for parts of cryo-EM density maps corresponding to dynamic regions of biomolecular complexes with potential compositional heterogeneity.

Discussion

Characterising complex biological processes through cryo-EM presents many unique challenges, especially for systems that are dynamic, exist in multiple conformational or compositional states. Among such systems is the ribosome-bound trigger factor, which possible dynamics we have elucidated using cryoENsemble, a method that takes advantage of both the molecular dynamics simulation and Bayesian methodology to yield accurate structural and dynamic representations of complex biomolecular systems. Notably, unlike most of the existing

methods, it adjusts the weights of the structural ensembles to improve their agreement with 3D cryo-EM maps, rather than fit or refine a single structure. Our approach uses Bayesian inference to reweight pre-calculated structural ensembles, which differs from EMMI³⁹, where inference is applied during the MD simulations to refine the ensemble with data from the cryo-EM map.

The effectiveness of our method highly depends on the quality of the prior structural ensemble since our reweighting strategy by design does not produce new conformations. The advantage of this approach is that it is computationally less demanding, easier to set up and can use structural ensembles obtained from any type of molecular dynamics simulations (all-atom or coarse-grained force fields). Additionally, we can combine multiple different ensembles to study compositional heterogeneity. However, if the prior ensemble is too different from the cryo-EM map (e.g. low correlation coefficient upon reweighting), it may be necessary to restart the simulation in a different force field, from a different starting structure or use enhanced sampling methods to sample conformational space more efficiently. Finally, as new, more refined maps emerge, we can rerun the cryoENsemble without the need to restart the MD simulations, allowing us to use the method alongside ongoing cryo-EM map processing.

In this study, we used all-atom structure-based models⁷⁵ to sample available conformational space efficiently. While structure-based potentials have been previously used to fit models in the cryo-EM maps via MDfit⁷⁸, we have instead employed them to generate a prior structural ensemble. Despite approximations, structure-based models allow exploration of the dynamics of large biological complexes that are otherwise inaccessible to more detailed computational approaches and can accurately describe their functional dynamics^{79,80}. Coarse-grained simulations, once converted to the all-atom resolution, could be used in a similar manner to generate prior structural ensembles for cryoENsemble, thereby further expanding the accessible system size and complexity.

For more detailed systems, the use of more advanced force fields, such as CHARMM36m⁸¹ or DES-AMBER⁸², may be necessary to generate more apt initial structural ensembles – potentially even guided by density-driven MD simulations²⁶, where lower resolution map or one of the half-maps can be used to restrict sampled conformational space. Moreover, the structural ensembles derived from MD simulations with enhanced sampling methods can also be used to further increase the capacity to extensively sample the conformational landscape. In this scenario, weights obtained from the reweighted MD simulations would serve to define our initial ensemble.

CryoENsemble can also be used to improve or test molecular dynamics simulations. By setting up multiple simulations with different force fields, cryoENsemble can assess which of the force fields most accurately captures the cryo-EM data. Furthermore, this method could be integrated into various force field parameterisation schemes, thereby enabling the utilisation of cryo-EM data^{83–85}.

In standard cryoENsemble reweighting, the weights can't be equal to 0 (see method description). By introducing the iterative mode, where in each step we select a sub-ensemble of structures (Fig. 1), we define the minimal set of structures that achieves a good agreement with the cryo-EM map. Each new sub-ensemble is treated as a new prior ensemble with uniform weights and undergoes a standard cryoENsemble run. The iterative mode improved the results in both of our test systems (ADK and FLN5-6 + 31 RNC).

While our method is computationally efficient, the reweighting time and memory usage can depend on the size of both the cryo-EM map (number of voxels) and the structural ensemble. This can be mitigated by clustering the MD structural ensemble before the reweighting to eliminate highly similar structures, as each requires calculating and storing a density map. The largest structural ensemble we tested comprised 2000 structures, which should suffice to capture the heterogeneity present in the cryo-EM reconstructions for most of the cases. We used a maximum of ~ 120,000 voxels from the cryo-EM map; however, one can apply initial down-sampling of the map to reduce the number of voxels for particularly large datasets. This method can also be helpful when working with large structural datasets. A stepwise reweighting with a downsampled map can be applied to obtain a minimal set of structures, which can be subsequently reweighted again using the high-resolution map. In a similar fashion, cryo-EM maps can be split into sections with varying resolutions and noise levels or be utilised through separate half-maps. Reweighting the structural ensemble first to the much more refined maps and then subsequently reweighting it with a less resolved map region could therefore help mitigate some of the challenges with highly heterogeneous maps.

Through the use of two synthetic datasets and one experimentally derived cryo-EM map, we have demonstrated that cryoENsemble can generate structural ensembles with averaged density maps closely mirroring the experimental maps and accurately reproducing the structural and dynamic properties of the underlying conformational ensembles. We have also presented that a fitted structural ensemble captures experimental data better than a single structure in these cases.

Each 3D cryo-EM map is obtained through extensive processing involving particle-sorting steps like 2D and 3D classification and thus represents only a fragment of the conformational space sampled by the system under study, which limits our complete understanding of its dynamics. Depending on the number of iterative 3D classifications and sorting rounds driven by the system dynamics and the scientific question, the maps can combine different levels of heterogeneity and dynamics. Developing methods such as cryoENsemble, which can capture dynamics in well-defined maps and model the various types of heterogeneity in low-resolution maps, will enable the user to choose maps in earlier processing stages, reducing the time from data collection to model.

The cryoENsemble approach is particularly suited for complex biological systems featuring convoluted dynamics. However, highly dynamic systems may be challenging if they do not present detectable cryo-EM density. Systems accessible to the cryoENsemble approach often yield cryo-EM maps with high-resolution regions associated with more static components and lower-resolution and ambiguous cryo-EM density describing dynamic elements. This includes nascent chain polypeptides or ribosome auxiliary factors bound to the ribosome. In these cases, the rigid and well-resolved structure of the ribosome contrasts with the low-resolution cryo-EM density of the NC or auxiliary factor. The dynamic character of these components implies the search for a solution in the form of a true structural ensemble rather than a selection of structures, which individually

fit into the density or just a single structure⁴³. We have also shown how cryoENsemble can be applied to analyse unaccounted densities and compositional heterogeneity, which allowed us to model TF-bound methionine aminopeptidase. By enabling the analysis of cryo-EM maps for regions that are more dynamic and therefore have less well-defined density, our method opens up new avenues for structural studies. Finally, cryoENsemble can be extended to utilise other experimental data within the BioEn or similar framework, making it a potent tool for integrative structural biology.

Methods

We tested cryoENsemble using two synthetic datasets. The first one is adenylate kinase (ADK) in both the open and closed conformations, capturing the discrete heterogeneity present in cryo-EM maps. The second system is a ribosome-bound nascent chain of the immunoglobulin-like domain (FLN5-6 + 31), exemplifying continuous heterogeneity. Characterising ribosome-bound nascent chains using cryo-EM is especially challenging, given that they combine flexible and predominantly unstructured linkers in the exit tunnel with folded or partially folded domains outside of the exit tunnel⁶⁹; the latter are only transiently interacting with the ribosome⁷².

Generating the synthetic reference density maps

In our Bayesian framework, under typical circumstances, the reference density map would correspond to the experimentally derived cryo-EM map ($\rho^{ref} = \rho^{exp}$). However, to test our methodology, we utilised synthetic reference density maps. They were either generated based on the crystal structures of the open or closed ADK state ($\rho^{ref} = \rho^{X-ray}$) or on randomly selected models from the all-atom MD ensemble of the FLN5-6 + 31 ribosomal nascent chain ($\rho^{ref} = \sum_{i=1}^N \frac{\rho_i}{N}$, where ρ_i is a density map of the i -th model). These synthetic reference density maps were generated using a protocol that mimics *molmap* command from ChimeraX with the bandwidth of the blur kernel σ set at $0.225 \times$ resolution⁸⁶. Maps were produced at three differing resolutions (3, 6, and 10 Å) to explore the influence of resolution on the reweighting process. To further investigate the effect of noise on reweighting, we added different levels of Gaussian noise to the map. The noise had a mean of 0 and a standard deviation based on either 1% or 10% of the map's maximum density.

Generating the density maps for the prior structural ensemble

In addition to the synthetic reference density maps (ρ^{ref}), we generated density maps for every structure from the MD ensemble (ρ^i). If not initially aligned, each structure was aligned to the reference density map using Situs²⁰. Following this alignment and using an approach similar to one from the modified *gmconvert* script⁴⁰, density maps were generated with the same voxel size, number of voxels and origin as the reference density map. The process involved positioning a spherical 3D-Gaussian function at each atom position with parameters for the corresponding atom obtained by fitting the electron atomic scattering factors specific to each atom type^{40,87}.

Synthetic density map processing

The generated density maps, both the reference and those from the structural ensemble, were further processed using *mrcfile* python library⁸⁸. From our reweighting dataset, we excluded voxels with negative values and rescaled the remaining ones to a molecular density value of 1, making the different maps easier to compare. Our reweighting methodology operates only on the selected voxels, both from the reference density map and the density maps generated based on the MD ensemble, that have density above the corresponding thresholds. The reference density map threshold is set up to be equal to $3 * \sigma_{noise}$, where σ_{noise} was either 1% or 10% of the maximum density, whereas the threshold for maps generated based on MD was equal to $3 * \sigma_{map}$, where σ_{map} is the standard deviation of the synthetic map (Supplementary Fig. 3).

Generation of adenylate kinase synthetic cryo-EM densities

We generated synthetic density maps based on available X-ray structures (1AKE for closed and 4AKE for open conformation), and for the final reference map, we averaged the different populations of open and closed states maps, starting from fully open state conformation and changing the population progressively using 10% intervals until the fully closed conformation was arrived at. In our test protocol, we operated under the assumption that during the cryo-EM image processing, these states could not be separated into individual 3D reconstructions but were averaged into a single one. We produced 11 averaged reference maps at three different resolutions (3, 6 or 10 Å) and with varying levels of Gaussian noise (with a mean of zero and a standard deviation corresponding to 1% or 10% of the maximum ADK density) (Supplementary Fig. 2). In total, we generated 66 synthetic maps for analysis.

Generation of the prior structural ensemble for adenylate kinase

Two short ($1.5 * 10^7$ steps) molecular dynamics simulations, carried out in GROMACS 4.5.7⁸⁹, were used to obtain the prior structural ensemble. We used an all-atom structure-based model generated in SMOG 2.0⁷⁵ with native contacts defined with the Shadow map algorithm⁹⁰ based on the X-ray structures of either the open or closed state. These two ensembles encapsulate the local dynamics around the native state of the apo or holo form. The combined structural ensemble consists of a total of 100 ADK conformations, with 50 randomly selected from each simulation.

Generation of the prior structural ensemble and synthetic cryo-EM densities for the FLN5-6 + 31 RNCs

For our study, we generated a starting ensemble by randomly selecting 100 conformations of the NC from the FLN5-6 structural ensemble obtained from the previous all-atom structure-based MD simulation⁷⁴ (Supplementary Figs. 1B and 21A). This ensemble exhibits significant structural heterogeneity, with RMSD values up to 28 Å (Supplementary Fig. 21C), reflecting the dynamic nature of the RNCs. To obtain the reference density maps, we randomly selected ten structures from this starting ensemble, generated an average density map at one of the three different resolutions (3, 6 or 10 Å), and repeated this procedure 100 times with Gaussian noise, corresponding to either 1% or 10% of the main density added (Supplementary Fig. 21B). This system enables us to evaluate our methodology in the case of continuous heterogeneity present in the cryo-EM maps.

Preparation of the trigger factor cryo-EM map for reweighting

For the final system, we used an experimentally derived cryo-EM map capturing the dynamics of the ribosome-associated chaperone (trigger factor) bound to the ribosome in the presence of the peptide deformylase and excess of methionine aminopeptidase (Supplementary Fig. 1C)⁵⁰. The obtained cryo-EM map had a clear density for the 70S ribosome, TF and PDF, which enabled authors to fit and refine models. However, the presence of the incomplete MetAP density suggested a novel tertiary binding site but did not allow for modelling the bound state. To create a reference map for the reweighting process, we used ChimeraX to select only the density that corresponds to either the Trigger Factor (TF) or the unmodeled MetAP (Supplementary Fig. 1C), subsequently saving it as a smaller, cropped map. This map was then normalised using the same methodology that we previously outlined for the synthetic reference map. We used the map threshold suggested by the authors (0.005) to select significant voxels for the reweighting.

Generation of the prior structural ensemble for the TF system

Using the available structure of the 70S ribosome from *E. coli* with bound TF and PDF (from PDB ID: 7D80⁵⁰), we prepared a starting structure for the MD simulation that encompassed the surface of the 70S ribosome around the ribosomal exit tunnel and bound both TF and PDF (Supplementary Fig. 36). We used an all-atom structure-based model generated with SMOG 2.4.4^{75,91} with bond lengths and angles based on the AMBER03 force field^{92,93}. Native contacts that are used in structure-based potential were defined based on TF cryo-EM structure with the use of the Shadow Map⁹⁰. For the structure-based MD simulations setup in SMOG, reduced units were applied with length, time, mass and energy scale all set to 1, except for the Boltzmann constant, which is $k_B = 0.00831451$ (kJmol⁻¹ K⁻¹, default in GROMACS). Simulations were performed for 5×10^8 steps in GROMACS 2021.2⁹⁴ in NVT ensemble at a reduced temperature of 0.5 (60 in GROMACS units), which is slightly below the temperature for this model to capture physiological conditions (0.582 reduced unit⁹²). The constant temperature was maintained via the Langevin Dynamics protocol. Taking advantage of a recent comparison of diffusion coefficients in the SMOG model and an all-atom explicit-solvent model⁹⁵, we estimated the effective simulated time to be in a range of hundreds of microseconds. During simulations, we kept the atoms of the ribosome surface frozen. We sampled the trajectory every 5×10^5 steps generating 1000 structures, and clustered them based on the RMSD using the *gmx cluster* method from GROMACS (Fig. 5A). Obtained structural ensemble, we used as a prior during the reweighting process carried out in cryoENsemble.

Fitting of the MetAP and generation of the prior structural ensemble for the TF + MetAP system

To isolate the MetAP cryo-EM density, we utilised the ChimeraX⁸⁶ command ‘volume subtract’ to create a difference map between the original (EMDB: 30611⁵⁰) and the posterior map derived from cryoENsemble reweighting of the TF MD trajectory. The *E. coli* methionine aminopeptidase structure (PDB ID: 1MAT⁷⁷) was fitted into the obtained density using ChimeraX, orienting positively charged loops towards the ribosome, in accordance with previous studies⁷⁶. For subsequent rigid-body fitting, we utilised the ‘Fit in Map’ command, setting the simulated map resolution to 8 Å. Generated TF + MetAP structure was used as a starting structure in the all-atom structure-based MD simulations with the same setup as the TF ensemble. The trajectory was sampled every 5×10^5 steps, generating 1000 structures, which were clustered based on the RMSD using the *gmx cluster* method from GROMACS. The obtained structural ensemble was combined with the TF ensemble and used as a prior during the reweighting process carried out in cryoENsemble.

Code availability

The source code of cryoENsemble, accompanied by a basic tutorial, is freely available on GitHub at: <https://github.com/dydyμος/cryoENsemble>.

Received: 21 February 2024; Accepted: 24 July 2024

Published online: 05 August 2024

References

1. Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. The energy landscapes and motions of proteins. *Science* **254**, 1598–1603 (1991).
2. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* **450**, 964–972 (2007).
3. Alderson, T. R. & Kay, L. E. NMR spectroscopy captures the essential role of dynamics in regulating biomolecular function. *Cell* **184**, 577–595 (2021).
4. Lerner, E. *et al.* Toward dynamic structural biology: Two decades of single-molecule Förster resonance energy transfer. *Science* **359**, eaan1133 (2018).

5. Michalet, X., Weiss, S. & Jäger, M. Single-molecule fluorescence studies of protein folding and conformational dynamics. *Chem. Rev.* **106**, 1785–1813 (2006).
6. Vendruscolo, M. & Dobson, C. M. Structural biology. Dynamic visions of enzymatic reactions. *Science* **313**, 1586–1587 (2006).
7. Baldwin, A. J. & Kay, L. E. NMR spectroscopy brings invisible protein states into focus. *Nat. Chem. Biol.* **5**, 808–814 (2009).
8. Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* **5**, 789–796 (2009).
9. Bai, X., McMullan, G. & Scheres, S. H. W. How cryo-EM is revolutionizing structural biology. *Trends Biochem. Sci.* **40**, 49–57 (2015).
10. Cheng, Y. Single-particle cryo-EM at crystallographic resolution. *Cell* **161**, 450–457 (2015).
11. Nakane, T. *et al.* Single-particle cryo-EM at atomic resolution. *Nature* **587**, 152–156 (2020).
12. Fernandez-Leiro, R. & Scheres, S. H. W. Unravelling biological macromolecules with cryo-electron microscopy. *Nature* **537**, 339–346 (2016).
13. Vant, J. W. *et al.* Exploring cryo-electron microscopy with molecular dynamics. *Biochem. Soc. Trans.* **50**, 569–581 (2022).
14. Kirmizialtin, S., Loerke, J., Behrmann, E., Spahn, C. M. T. & Sanbonmatsu, K. Y. Using molecular simulation to model high-resolution cryo-EM reconstructions. *Meth. Enzymol.* **558**, 497–514 (2015).
15. Nierzwicki, I. & Palermo, G. Molecular dynamics to predict cryo-EM: Capturing transitions and short-lived conformational states of biomolecules. *Front. Mol. Biosci.* **8**, 641208 (2021).
16. Fraser, J. S., Lindorff-Larsen, K. & Bonomi, M. What will computational modeling approaches have to say in the era of atomistic cryo-EM data?. *J. Chem. Inf. Model.* <https://doi.org/10.1021/acs.jcim.0c00123> (2020).
17. Malhotra, S., Träger, S., Dal Peraro, M. & Topf, M. Modelling structures in cryo-EM maps. *Curr. Opin. Struct. Biol.* **58**, 105–114 (2019).
18. Giri, N., Roy, R. S. & Cheng, J. Deep learning for reconstructing protein structures from cryo-EM density maps: Recent advances and future directions. *Curr. Opin. Struct. Biol.* **79**, 102536 (2023).
19. Bonomi, M. & Vendruscolo, M. Determination of protein structural ensembles using cryo-electron microscopy. *Curr. Opin. Struct. Biol.* **56**, 37–45 (2019).
20. Wriggers, W., Milligan, R. A. & McCammon, J. A. Situs: A package for docking crystal structures into low-resolution maps from electron microscopy. *J. Struct. Biol.* **125**, 185–195 (1999).
21. Tama, F., Miyashita, O. & Brooks, C. L. Flexible multi-scale fitting of atomic structures into low-resolution electron density maps with elastic network normal mode analysis. *J. Mol. Biol.* **337**, 985–999 (2004).
22. Topf, M. *et al.* Protein structure fitting and refinement guided by cryo-EM density. *Structure* **16**, 295–307 (2008).
23. Joseph, A. P. *et al.* Refinement of atomic models in high resolution EM reconstructions using Flex-EM and local assessment. *Methods* **100**, 42–49 (2016).
24. Croll, T. I. ISOLDE: A physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr. D Struct. Biol.* **74**, 519–530 (2018).
25. Trabuco, L. G., Villa, E., Mitra, K., Frank, J. & Schulten, K. Flexible fitting of atomic structures into electron microscopy maps using molecular dynamics. *Structure* **16**, 673–683 (2008).
26. Igaev, M., Kutzner, C., Bock, L. V., Vaiana, A. C. & Grubmüller, H. Automated cryo-EM structure refinement using correlation-driven molecular dynamics. *eLife* **8**, e43542 (2019).
27. Singharoy, A. *et al.* Molecular dynamics-based refinement and validation for sub-5 Å cryo-electron microscopy maps. *eLife* **5**, e16105 (2016).
28. Blau, C., Yvonesdotter, L. & Lindahl, E. Gentle and fast all-atom model refinement to cryo-EM densities via a maximum likelihood approach. *PLoS Comput. Biol.* **19**, e1011255 (2023).
29. Serna, M. Hands on methods for high resolution cryo-electron microscopy structures of heterogeneous macromolecular complexes. *Front. Mol. Biosci.* **6**, 33 (2019).
30. Lyumkis, D. Challenges and opportunities in cryo-EM single-particle analysis. *J. Biol. Chem.* **294**, 5181–5197 (2019).
31. Scheres, S. H. W. *et al.* Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. *Nat. Methods* **4**, 27–29 (2007).
32. 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. *Curr. Opin. Struct. Biol.* **81**, 102626 (2023).
33. Wong, W. *et al.* Cryo-EM structure of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. *eLife* **3**, e03080 (2014).
34. Nakane, T., Kimanius, D., Lindahl, E. & Scheres, S. H. Characterisation of molecular motions in cryo-EM single-particle data by multi-body refinement in RELION. *eLife* **7**, e36861 (2018).
35. Frank, J. & Ourmazd, A. Continuous changes in structure mapped by manifold embedding of single-particle data in cryo-EM. *Methods* **100**, 61–67 (2016).
36. Zhong, E. D., Bepler, T., Berger, B. & Davis, J. H. CryoDRGN: Reconstruction of heterogeneous cryo-EM structures using neural networks. *Nat. Methods* **18**, 176–185 (2021).
37. Zhong, E. D., Lerer, A., Davis, J. H. & Berger, B. Cryodr2: Ab initio neural reconstruction of dynamic protein complexes. *MAM* **28**, 1216–1217 (2022).
38. Koukos, P. I. & Bonvin, A. M. J. Integrative modelling of biomolecular complexes. *J. Mol. Biol.* **432**, 2861–2881 (2020).
39. Bonomi, M., Pellarin, R. & Vendruscolo, M. Simultaneous determination of protein structure and dynamics using Cryo-electron microscopy. *Biophys. J.* **114**, 1604–1613 (2018).
40. Bonomi, M. *et al.* Bayesian weighing of electron cryo-microscopy data for integrative structural modeling. *Structure* **27**, 175–188 (2019).
41. Rieping, W., Habeck, M. & Nilges, M. Inferential structure determination. *Science* **309**, 303–306 (2005).
42. Rout, M. P. & Sali, A. Principles for integrative structural biology studies. *Cell* **177**, 1384–1403 (2019).
43. Bonomi, M., Heller, G. T., Camilloni, C. & Vendruscolo, M. Principles of protein structural ensemble determination. *Curr. Opin. Struct. Biol.* **42**, 106–116 (2017).
44. Bonomi, M., Camilloni, C., Cavalli, A. & Vendruscolo, M. MetaInference: A Bayesian inference method for heterogeneous systems. *Sci. Adv.* **2**, e1501177 (2016).
45. Rangan, R. *et al.* Determination of structural ensembles of proteins: restraining vs reweighting. *J. Chem. Theory Comput.* **14**, 6632–6641 (2018).
46. Tang, W. S. *et al.* Ensemble reweighting using cryo-EM particle images. *J. Phys. Chem. B* **127**, 5410–5421 (2023).
47. Giraldo-Barreto, J. *et al.* A Bayesian approach to extracting free-energy profiles from cryo-electron microscopy experiments. *Sci. Rep.* **11**, 13657 (2021).
48. Hoff, S. E., Thomasen, F. E., Lindorff-Larsen, K. & Bonomi, M. Accurate model and ensemble refinement using cryo-electron microscopy maps and Bayesian inference. *BioRxiv* <https://doi.org/10.1101/2023.10.18.562710> (2023).
49. Hummer, G. & Köfinger, J. Bayesian ensemble refinement by replica simulations and reweighting. *J. Chem. Phys.* **143**, 243150 (2015).
50. Akbar, S., Bhakta, S. & Sengupta, J. Structural insights into the interplay of protein biogenesis factors with the 70S ribosome. *Structure* **29**, 755–767 (2021).

51. Saio, T., Guan, X., Rossi, P., Economou, A. & Kalodimos, C. G. Structural basis for protein antiaggregation activity of the trigger factor chaperone. *Science* **344**, 1250494 (2014).
52. Deuerling, E., Schulze-Specking, A., Tomoyasu, T., Mogk, A. & Bukau, B. Trigger factor and DnaK cooperate in folding of newly synthesized proteins. *Nature* **400**, 693–696 (1999).
53. Mashaghi, A. *et al.* Reshaping of the conformational search of a protein by the chaperone trigger factor. *Nature* **500**, 98–101 (2013).
54. Hoffmann, A. *et al.* Concerted action of the ribosome and the associated chaperone trigger factor confines nascent polypeptide folding. *Mol. Cell* **48**, 63–74 (2012).
55. Wu, K., Minshull, T. C., Radford, S. E., Calabrese, A. N. & Bardwell, J. C. A. Trigger factor both holds and folds its client proteins. *Nat. Commun.* **13**, 4126 (2022).
56. Arhar, T., Shkedi, A., Nadel, C. M. & Gestwicki, J. E. The interactions of molecular chaperones with client proteins: Why are they so weak?. *J. Biol. Chem.* **297**, 101282 (2021).
57. Deeng, J. *et al.* Dynamic behavior of trigger factor on the ribosome. *J. Mol. Biol.* **428**, 3588–3602 (2016).
58. Merz, F. *et al.* Molecular mechanism and structure of Trigger Factor bound to the translating ribosome. *EMBO J.* **27**, 1622–1632 (2008).
59. Bhakta, S., Akbar, S. & Sengupta, J. Cryo-EM structures reveal relocalization of MetAP in the presence of other protein biogenesis factors at the ribosomal tunnel exit. *J. Mol. Biol.* **431**, 1426–1439 (2019).
60. Köfinger, J. *et al.* Efficient ensemble refinement by reweighting. *J. Chem. Theory Comput.* **15**, 3390–3401 (2019).
61. Köfinger, J. & Hummer, G. Encoding prior knowledge in ensemble refinement. *J. Chem. Phys.* <https://doi.org/10.26434/chemrxiv-2023-71rr6> (2023).
62. Bussi, G. & Laio, A. Using metadynamics to explore complex free-energy landscapes. *Nat. Rev. Phys.* <https://doi.org/10.1038/s42254-020-0153-0> (2020).
63. Habeck, M. Bayesian modeling of biomolecular assemblies with Cryo-EM maps. *Front. Mol. Biosci.* **4**, 15 (2017).
64. Satopaa, V., Albrecht, J., Irwin, D. & Raghavan, B. Finding a “kneede” in a haystack: Detecting knee points in system behavior. in *2011 31st International Conference on Distributed Computing Systems Workshops* 166–171 (IEEE, 2011). <https://doi.org/10.1109/ICDCSW.2011.20>.
65. Müller, C. W. & Schulz, G. E. Structure of the complex between adenylate kinase from *Escherichia coli* and the inhibitor Ap5A refined at 1.9 Å resolution. A model for a catalytic transition state. *J. Mol. Biol.* **224**, 159–177 (1992).
66. Müller, C. W., Schlauderer, G. J., Reinstein, J. & Schulz, G. E. Adenylate kinase motions during catalysis: An energetic counterweight balancing substrate binding. *Structure* **4**, 147–156 (1996).
67. Lawson, C. L. *et al.* Cryo-EM model validation recommendations based on outcomes of the 2019 EMDataResource challenge. *Nat. Methods* **18**, 156–164 (2021).
68. Joseph, A. P., Lagerstedt, I., Patwardhan, A., Topf, M. & Winn, M. Improved metrics for comparing structures of macromolecular assemblies determined by 3D electron-microscopy. *J. Struct. Biol.* **199**, 12–26 (2017).
69. Cassaignau, A. M. E., Cabrita, L. D. & Christodoulou, J. How does the ribosome fold the proteome?. *Annu. Rev. Biochem.* **89**, 389–415 (2020).
70. Waudby, C. A., Burrige, C., Cabrita, L. D. & Christodoulou, J. Thermodynamics of co-translational folding and ribosome-nascent chain interactions. *Curr. Opin. Struct. Biol.* **74**, 102357 (2022).
71. Waudby, C. A., Dobson, C. M. & Christodoulou, J. Nature and regulation of protein folding on the ribosome. *Trends Biochem. Sci.* **44**, 914–926 (2019).
72. Chan, S. H. S. *et al.* The ribosome stabilizes partially folded intermediates of a nascent multi-domain protein. *Nat. Chem.* **14**, 1165–1173 (2022).
73. Nakatogawa, H. & Ito, K. Secretion monitor, SecM, undergoes self-translation arrest in the cytosol. *Mol. Cell* **7**, 185–192 (2001).
74. Ahn, M. *et al.* Modulating co-translational protein folding by rational design and ribosome engineering. *Nat. Commun.* **13**, 4243 (2022).
75. Noel, J. K. *et al.* SMOG 2: A versatile software package for generating structure-based models. *PLoS Comput. Biol.* **12**, e1004794 (2016).
76. Sandikci, A. *et al.* Dynamic enzyme docking to the ribosome coordinates N-terminal processing with polypeptide folding. *Nat. Struct. Mol. Biol.* **20**, 843–850 (2013).
77. Roderick, S. L. & Matthews, B. W. Structure of the cobalt-dependent methionine aminopeptidase from *Escherichia coli*: A new type of proteolytic enzyme. *Biochemistry* **32**, 3907–3912 (1993).
78. Whitford, P. C. *et al.* Excited states of ribosome translocation revealed through integrative molecular modeling. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 18943–18948 (2011).
79. de Oliveira, A. B. *et al.* SMOG 2 and OpenSMOG: Extending the limits of structure-based models. *Protein Sci.* **31**, 158–172 (2022).
80. Jackson, J., Nguyen, K. & Whitford, P. C. Exploring the balance between folding and functional dynamics in proteins and RNA. *Int. J. Mol. Sci.* **16**, 6868–6889 (2015).
81. Huang, J. *et al.* CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2017).
82. Tucker, M. R., Piana, S., Tan, D., LeVine, M. V. & Shaw, D. E. Development of force field parameters for the simulation of single- and double-stranded DNA molecules and DNA-protein complexes. *J. Phys. Chem. B* **126**, 4442–4457 (2022).
83. Tesei, G., Schulze, T. K., Crehuet, R. & Lindorff-Larsen, K. Accurate model of liquid-liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2111696118 (2021).
84. Polêto, M. D. & Lemkul, J. A. Integration of experimental data and use of automated fitting methods in developing protein force fields. *Commun. Chem.* **5**, 38 (2022).
85. Köfinger, J. & Hummer, G. Empirical optimization of molecular simulation force fields by Bayesian inference. *Eur. Phys. J. B* **94**, 245 (2021).
86. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci.* **30**, 70–82 (2021).
87. *International Tables for Crystallography: Mathematical, Physical and Chemical Tables*. (International Union of Crystallography, 2006). <https://doi.org/10.1107/97809553602060000103>.
88. Burnley, T., Palmer, C. M. & Winn, M. Recent developments in the CCP-EM software suite. *Acta Crystallogr. D Struct. Biol.* **73**, 469–477 (2017).
89. Pronk, S. *et al.* GROMACS 4.5: A high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845–854 (2013).
90. Noel, J. K., Whitford, P. C. & Onuchic, J. N. The shadow map: A general contact definition for capturing the dynamics of biomolecular folding and function. *J. Phys. Chem. B* **116**, 8692–8702 (2012).
91. Whitford, P. C. *et al.* Nonlocal helix formation is key to understanding S-adenosylmethionine-1 riboswitch function. *Biophys. J.* **96**, L7–L9 (2009).
92. Whitford, P. C., Jiang, W. & Serwer, P. Simulations of phage T7 capsid expansion reveal the role of molecular sterics on dynamics. *Viruses* **12**, 1273 (2020).
93. Duan, Y. *et al.* A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.* **24**, 1999–2012 (2003).

94. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
95. Yang, H. *et al.* Diffusion of tRNA inside the ribosome is position-dependent. *J. Chem. Phys.* **151**, 085102 (2019).

Acknowledgements

The study was supported by a Wellcome Trust Investigator Award (to J.C. 206409/Z/17/ Z). This project made use of time on HPC granted via the UK High-End Computing Consortium for Biomolecular Simulation, HECBio-Sim (<http://hecbiosim.ac.uk>), supported by EPSRC (Grants Nos. EP/R029407/1 and EP/X035603/1). We also acknowledge the EuroHPC Joint Undertaking for awarding this project access to the EuroHPC supercomputer LUMI, hosted by CSC (Finland) and the LUMI consortium through a EuroHPC Regular Access call and the Baskerville Tier 2 HPC service (<https://www.baskerville.ac.uk/>). Baskerville was funded by the EPSRC and UKRI through the World Class Labs scheme (EP/T022221/1) and the Digital Research Infrastructure programme (EP/W032244/1) and is operated by Advanced Research Computing at the University of Birmingham. We acknowledge the ISMB EM facility (Birkbeck College, University of London).

Author contributions

T.W., J.O.S., A.M., L.D.C., M.V. and J.C. designed the project. T.W. performed the research. T.W., J.O.S., A.M., L.D.C., M.V. and J.C. prepared the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-68468-7>.

Correspondence and requests for materials should be addressed to T.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024