cryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination

Ali Punjani[1], John L. Rubinstein[2,3,4], David J. Fleet[5], and Marcus A. Brubaker[6]

1. Department of Computer Science, The University of Toronto, Toronto, Ontario, Canada
2. Molecular Structure and Function Program, The Hospital for Sick Children Research Institute, Toronto, Ontario, Canada
3. Department of Biochemistry, The University of Toronto, Toronto, Ontario, Canada
4. Department of Medical Biophysics, The University of Toronto, Toronto, Ontario, Canada
5. Department of Computer Science, The University of Toronto, Toronto, Ontario, Canada
6. Department of Electrical Engineering and Computer Science, York University, Toronto, Ontario, Canada

Correspondence to:
Ali Punjani (alipunjani@cs.toronto.edu) and Marcus Brubaker (mab@eecs.yorku.ca)

Abstract

Single particle electron cryomicroscopy (cryo-EM) is a powerful method for determining the structures of biological macromolecules. With automated microscopes, cryo-EM data can often be obtained in a few days. However, processing cryo-EM image data to reveal heterogeneity in the protein structure and to refine 3-D maps to high resolution frequently becomes a severe bottleneck, requiring expert intervention, prior structural knowledge, and weeks of calculations on expensive computer clusters. Here we show that stochastic gradient descent (SGD) and branch and bound maximum likelihood optimization algorithms permit the major steps in cryo-EM structure determination to be performed in hours or minutes on an inexpensive desktop computer. Furthermore, SGD with Bayesian marginalization allows *ab initio* 3-D classification, enabling automated analysis and discovery of unexpected structures without bias from a reference map. These algorithms are combined in a user-friendly computer program named cryoSPARC.

Introduction

Scientific approaches can be transformed by innovations that decrease the cost and improve non-expert usability of technology, as seen with DNA sequencing and synthesis, microarray technology, and even the use of computers themselves. These changes can occur both quantitatively, by allowing more experiments to be done in a shorter time by experts and non-specialists, and qualitatively by changing the type and scope of experiments that are feasible. Recent advances in single particle cryo-EM[1,2] have enabled near-atomic resolution structure determination of biomedically important protein complexes[3–5], bringing the technique to the attention of the general biological research community and pharmaceutical companies. The throughput and automation of cryo-EM becomes increasingly important as the technique is used for structure-based drug design[6] and time-critical structural studies of pathogens[7]. Automated electron microscopes can collect datasets for atomic resolution structure determination in as little as 24 or 48 hours given appropriately prepared specimens, and centralized cryo-EM facilities are now providing instrument access to non-specialist investigators. Calculation of 3-D maps from cryo-EM images, however, can require weeks of computational analysis by an expert user. With routine collection of cryo-EM datasets that contain millions of single particle images corresponding to different 3-D conformations of the sample[8], the cost of image analysis can exceed 500,000 CPU hours on large, expensive computer clusters[9]. Further, without significant user expertise, there are a variety of ways in which incorrect and misleading 3-D maps can be generated at various stages in the image analysis pipeline[10,11]. The computational cost and the requirement for user input are bottlenecks both for automation and widespread use of cryo-EM.

To address these issues, we developed two new algorithms. The first of these algorithms, for the first time, makes it possible to perform unsupervised *ab initio* 3-D classification, whereby multiple 3-D states of a protein can be discovered from a single sample without user input of prior structural knowledge, and without the assumption that all 3-D states resemble each other. In contrast, existing techniques for 3-D refinement of cryo-EM maps require an initial structure that is close to the correct target structure[12,13]. The second algorithmic development radically speeds up high-resolution refinement of cryo-EM maps by exploiting characteristics of image alignment to achieve massive computational savings by removing redundant computation. These two abilities are combined in a standalone Graphics Processing Unit (GPU) accelerated software package that we have named cryoSPARC (cryo-EM single particle *ab initio* reconstruction and classification). CryoSPARC can refine multiple high-resolution 3D structures directly from single particle images, with no user input or expertise required. These combined steps are done in a matter of hours on a single consumer-grade desktop computer. GPU hardware has been used previously to accelerate cryo-EM contrast transfer function estimation[14] and identification of particles within images[15]. Related work has shown that exploiting GPU hardware in the popular program RELION can significantly speed up existing algorithms for reference-based 3-D classification and refinement[9]. The algorithms presented here provide a further order-of-magnitude reduction in computational cost compared to GPU acceleration, which would require at least an additional ~7 years if driven by hardware advances alone[16]. Based on the combination of algorithms, inexpensive hardware, and an easy-to-use graphical user interface, cryoSPARC can allow new non-specialist cryo-EM users to process data rapidly without needing to purchase or set up their own computer clusters and with minimal user input and expertise.

Results

Formally, structure determination by cryo-EM is an optimization problem and may be described in a Bayesian likelihood framework[12,17]:

$$\arg\max_{V_{1\ldots K}} \log p(V_{1\ldots K}|X_{1\ldots N}) = \arg\max_{V_{1\ldots K}} \sum_{i=1}^{N} \log \sum_{j=1}^{K} \frac{1}{K} \int p(X_i,\phi_i|V_j) d\phi_i + \log p(V_{1\ldots K}) \quad (1)$$

The aim of the optimization is to find the 3-D structures (V_1 to V_K) that best explain the observed images (X_1 to X_N), by marginalizing over class assignment (j) and the unknown pose variable (ϕ_i), which describes a 3-D rotation and a 2-D translation for each particle image.

Numerical optimization problems have been studied extensively in computer science[18]. Traditionally, optimization is formulated as the maximization of a single, monolithic objective function. With this approach, the variables of a function are iteratively altered until the 'best' values, which give an optimum value to the function, are identified. Sophisticated algorithms for iterative optimization have been developed[19] and are central to a myriad of problems in data modeling and engineering. In the case of cryo-EM map calculation, the objective function (Equation 1) quantifies how well cryo-EM maps explain the collected experimental images, and the variables in the function include the 3-D maps represented as density voxels on a 3-D grid.

We use a stochastic gradient descent (SGD) optimization scheme to quickly identify one or several low-resolution 3-D structures that are consistent with a set of observed images. This algorithm allows for *ab initio* heterogeneous structure determination with no prior model of the molecule's structure. Once approximate structures are determined, a branch and bound algorithm for image alignment helps rapidly refine structures to high resolution. The speed and robustness of these approaches allow structure determination in a matter of minutes or hours on a single inexpensive desktop workstation.

Stochastic Gradient Descent: Discovery of protein structure from random initialization

Cryo-EM map calculation is a non-convex optimization problem. These problems are among the most computationally challenging optimization problems known and are characterized by the presence of multiple locally-optimal settings of variables, each of which forms an attractor where typical iterative optimization algorithms can become stuck if poorly initialized[19] (Figure 1A). Sensitivity to local optima is seen in most optimization algorithms, including those used in cryo-EM[12,13] and as a result, refinement programs require a reasonably accurate initial model for the structure that initializes the search near the global optimum. However, recent methods have been discovered that perform well on non-convex problems. One such method is stochastic gradient descent (SGD)[20] (Figure 1). SGD was popularized as a key tool in Deep Learning for the optimization of non-convex functions, resulting in near-human level performance in tasks like image and speech recognition[21,22].

In Equation 1, each term of the outer sum represents the contribution of a single particle image to the overall likelihood of the 3-D map. SGD repeatedly approximates this objective function by selecting a different random subset of terms (i.e., single particle images) at each iteration, and

computes the sum of those terms (Figure 1B). In a single iteration, the optimization variables (i.e., the 3-D map) are updated based on the gradient of this approximate objective (Supplementary Note 1). Each iteration requires analyzing only a small subset of single particle images. As a consequence, a single iteration is inexpensive and hundreds or thousands of iterative changes can be made during each pass through the full dataset. It is commonly believed that it is because of these many noisy changes that SGD is insensitive to local optima and often finds effective solutions to non-convex problems (Figure 1C).

We implemented an SGD method for *ab initio* structure determination and 3-D classification. Applied to several different datasets, the use of SGD enables convergence to correct low-resolution structures from arbitrary random initialization, allowing both *ab initio* structure determination and *ab initio* 3-D classification (Figure 2). With 35,645 TRPV1 particle images[3] SGD optimization resulted in a low-resolution 3-D map in 75 minutes from random initialization (Figure 2A) using a single inexpensive desktop workstation with an Intel i7-5820K Processor and a single NVIDIA Tesla K40 GPU. When applied to a dataset of conformationally heterogeneous *Thermus thermophilus* V/A-ATPase particle images[23], the algorithm was able to discern three different conformational states for the enzyme, again from random initializations (Figure 2B). These three states correspond to the three different rotational positions of the central rotor of the enzyme[24]. This finding is particularly notable as previous analysis with reference-based classification[12] and the same dataset of images was only able to detect two of the three states[23]. The newly identified third rotational state is the conformation of the enzyme that differs the most from the other two. This observation illustrates the importance of reference-free *ab initio* classification for unbiased identification of states that differ from the expected structures present in the dataset.

Branch and bound: rapid refinement of maps to high resolution

The primary computational burden in map refinement is the search for orientation parameters that best align each 2-D single particle image to a 3-D density map. The branch and bound algorithm design paradigm[25] can accelerate this search by quickly and inexpensively ruling out large regions of the search space that cannot contain the optimum of the objective function (Figure 3A).

In cryo-EM map refinement, the optimal pose for a particle image minimizes the error between the observed image and a projection of the 3-D map. To find this optimal pose using the branch and bound approach (Figure 3B), an inexpensive lower bound on the error is first computed across the entire space of poses. At the pose that minimizes this lower bound, the computationally expensive true error function is evaluated. All regions of the search space where the lower bound exceeds this computed value of the true error function cannot contain the optimal pose and can be excluded from further search. A new lower bound is then calculated that fits more tightly to the true error function but is more expensive to calculate. The process of evaluating the error function at the optimum of the lower bound, discarding regions of search space where the true error is above the lower bound, and recalculating a tighter-fitting lower bound, is repeated until only the optimal pose remains.

Although conceptually straightforward, application of the branch and bound strategy requires an informative and inexpensive lower bound for the objective function. Suitable lower bounds are well known for other problems[26,27] but use of the method for determining the orientations of single particle cryo-EM images required derivation of an appropriate bound (Supplementary Note 2). The derivation we describe was based on the signal-to-noise ratio of single particle images over a range of resolutions. It is worth emphasizing that the branch and bound approach is a global pose search that requires no prior estimate of an optimal pose. In contrast, strategies to accelerate orientation determination based solely on local search risk selection of a pose that is not the global optimum[12,13]. In practice, an approximation to this branch and bound search is used (Supplementary Note 2) that was found to be equally effective but even more efficient.

We implemented the branch and bound approach and applied it to high-resolution structure determination from several published datasets: the 20S proteasome from *Thermoplasma acidophilum*[28], the 80S ribosome from *Plasmodium falciparum*[29], amphipol-solubilized rat TRPV1[3], as well as the *T. thermophilus* V/A-ATPase[23]. Computations were carried out with the same desktop workstation and single NVIDIA Tesla K40 GPU used for *ab initio* SGD calculations. Applied to 35,645 TRPV1 particle images, branch and bound orientation determination produced a 3.3 Å resolution map in 66 minutes with C4 symmetry enforced using a gold-standard refinement procedure[30], the FSC=0.143 resolution criterion[31], and correction for effects of masking on the FSC by high-resolution noise substitution[32] (Figure 2C). This resolution slightly exceeds the previously published resolution of 3.4 Å from the same dataset[3]. With *T. thermophilus* V/A-ATPase particle images sorted into three classes by SGD, the branch and bound search produced maps of all three states in a total of 2.4 hours (Figure 2D). The resolutions estimated for the states were 6.4 Å, 7.6 Å, and 7.9 Å, compared to 6.4 Å and 9.5 Å for the two states identified in the previously published analysis[23].

Following SGD *ab initio* structure determination, the application of the branch and bound method allowed high-resolution refinement of the 80S ribosome to 3.2 Å resolution, equivalent to the published resolution[29], in 2.2 hours (Figure 4A), demonstrating the capability of the method to deal with large and asymmetric protein complexes. Notably, on the same computer hardware (desktop computer with one GPU), this dataset of particle images would take approximately 20 hours for refinement using the GPU accelerated program RELION[9]. Similarly, the 20S proteasome structure was refined to 2.8 Å with D7 symmetry enforced, matching the published sub-3 Å resolution from the dataset[28] (Figure 4B) but in only 70 minutes. These refined maps show clear high-resolution detail and side-chain density, illustrating the performance of the method at near-atomic resolution.

Discussion

Ab initio reconstruction of 3-D maps from cryo-EM images has long been known as a significant problem. While random initialization can be successful for highly-symmetric particles[33], this has not been the case for asymmetric or low-degree of symmetry structures where incorrect structures have been published[34]. Previous approaches for determining low-resolution initial maps often involve collecting image tilt pairs[35,36]. In that method, the need to switch to a different experimental procedure to generate an initial map is unwieldy and presents a barrier to automated structure determination. Other investigators have proposed algorithms to generate initial maps from images obtained under standard conditions. The approaches have included evolutionary algorithms[37], a statistical weighted least squares approach[38], complex annealing procedures[39], matching of common lines[40] and statistical weighting[41]. However, all of these algorithms rely on analyzing all images in batch, making them intrinsically slower than our approach, particularly as the number of particle images in datasets grow. In contrast, SGD processes random subsets of data at each iteration, making it efficient, even in the face of large datasets. We previously showed that SGD could produce a reasonable low-resolution map *ab initio* for a homogenous dataset[42]. Here we have demonstrated that SGD, unlike other approaches, is sufficiently robust to enable reconstruction of multiple 3-D classes from independent arbitrary initializations. To our knowledge, all existing techniques for 3-D classification use a single initial reference from which analysis of heterogeneity proceeds. Removal of the assumption that all 3-D classes are similar to the single input reference is particularly advantageous for discovering 3-D states that are unexpected and different from the consensus structure. It is important to note that, like other algorithms, SGD will fail when the particle images do not contain a sufficient series of views to define the 3-D structure of the molecule. It can also fail if there are sufficient views, but strongly preferred orientations for particles. Other pathological situations may include analysis of datasets with little contrast at low-resolution. This situation can occur when insufficient defocus is used with a cryo-EM microscope that does not posses a phase plate or when imaging low molecular weight complexes.

Combination of the SGD approach and branch and bound refinement provides a complete framework for rapid *ab initio* calculation of multiple high-resolution maps from a heterogeneous dataset on inexpensive computer hardware. The bound derived and used in this work is based on, and provides a mathematical basis for, the common intuition that high-resolution features in an image contribute less to alignment than low-resolution features. This intuition has previously been used in heuristic methods that perform alignment and reconstruction at iteratively increasing resolution levels[12] or decompose the space of particle images into basis vectors that contain low-resolution features[43]. A number of heuristic methods have also been employed to accelerate the alignment of particle images to a structure at a fixed resolution. Most commonly, locally restricted high-resolution searches are used in later iterations of refinement, after exhaustive search at early iterations provides a guess for the optimal pose of each image[12,13]. These approaches can still be computationally expensive, require extra tunable parameters for when to start and how much to restrict local search, and run the risk of missing the optimal alignment. Branch and bound optimization provides a risk-free, parameter-free approach to accelerating computationally expensive search problems, is significantly faster than heuristic methods, and will likely find other applications in cryo-EM image analysis.

With the recent push to re-implement existing algorithms on new hardware (e.g., GPUs), attempts have also been made to simplify the task of accessing and using computer clusters through cloud computing service providers, notably Amazon EC2[44]. However, even with computer clusters available for rent, existing software methods do not scale well, providing diminishing returns with larger clusters. As the pace of cryo-EM data collection grows, and studies aim to distinguish increasingly subtle structural differences between 3-D classes[8,45], improved computational efficiency through algorithm development will be a critical enabler for both academic and industrial researchers using cryo-EM.

The new cryoSPARC software is available as a standalone program that can run on either commodity desktop workstations or rackmount servers. CryoSPARC is also available as a web-service, for new users to try prior to installing locally. Once particle images are selected and corrected for anisotropic beam-induced movement[46] and the effects of radiation damage[46,47] they may be processed through the program's web-browser graphical user interface (GUI). At minimum, a single consumer or professional grade NVIDIA GPU is required. The easy-to-use GUI (Supplementary Video 1) provides the same interface through both the web-service and in local installations. This GUI allows for multiple users within a laboratory to have separate accounts, access the program remotely, upload and share datasets, manage experimental results, launch computational tasks, and view results streaming in real-time as they are computed. A protocol detailing use of the software package has been prepared[49] (Supplementary Protocol).

Software availability

The software package, including source code, is available for non-commercial use as a download and as a web service at www.cryosparc.com. Results reported in this work were computed using cryoSPARC version 0.2.36.

Accession Codes

Data Availability Statement

The cryo-EM images used to experimentally demonstrate the effectiveness of algorithms were taken from previously published studies. Several datasets were downloaded directly from EMPIAR[48], including the 80S Ribosome (EMPIAR-10028), 20S proteasome (EMPIAR-10025), TRVP1 channel (EMPIAR-10005). Images of the *T. thermophilus* V/A-ATPase are available from the authors upon request. In all cases, the single particle images that were used in the original studies were input directly into cryoSPARC, with no further preprocessing.

Acknowledgements

We thank Suhail Dawood for construction of the GUI front end and members of the Rubinstein laboratory for testing cryoSPARC. AP was supported by a scholarship from the Natural Sciences and Engineering Research Council (NSERC), JLR was supported by the Canada Research Chairs program, and DJF was supported in part by the Learning in Machines and Brains program of the Canadian Institute for Advanced Research. This research was also supported by NSERC Discovery Grants RGPIN 2015-05630 (DJF) and 401724-12 (JLR), and an NVIDIA Academic Hardware Grant (MAB, AP). Part of this work was performed while MAB was a postdoctoral fellow at the University of Toronto.

Author Contributions

AP and MAB designed algorithms and implemented software. AP, MAB and JLR performed experimental work. JLR, DJF and MAB contributed expertise and supervision. All authors contributed to manuscript preparation.

Competing Financial Interests

All authors are engaged in a venture to commercially support cryoSPARC for industrial use.

References for main text

1. Kühlbrandt, W. *et al.* Biochemistry. The resolution revolution. *Science* **343,** 1443–4 (2014).
2. Smith, M. T. J. & Rubinstein, J. L. Beyond blob-ology. *Science (80-.).* **345,** 617–619 (2014).
3. Liao, M., Cao, E., Julius, D. & Cheng, Y. Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* **504,** 107–12 (2013).
4. Bai, X. C., Fernandez, I. S., McMullan, G. & Scheres, S. H. W. Ribosome structures to near-atomic resolution from thirty thousand cryo-EM particles. *Elife* **2013,** 2–13 (2013).
5. Yan, C. *et al.* Structure of a yeast spliceosome at 3.6-angstrom resolution. *Science* **349,** 1182–1191 (2015).
6. Banerjee, S. *et al.* 2.3 Å resolution cryo-EM structure of human p97 and mechanism of allosteric inhibition. *Science* **351,** 871–5 (2016).
7. Sirohi, D. *et al.* The 3.8 Å resolution cryo-EM structure of Zika virus. *Science* **352,** 467–70 (2016).
8. Abeyrathne, P. D., Koh, C. S., Grant, T., Grigorieff, N. & Korostelev, A. A. Ensemble cryo-EM uncovers inchworm-like translocation of a viral IRES through the ribosome. *Elife* **5,** e14874 (2016).
9. Kimanius, D., Forsberg, B. O., Scheres, S. & Lindahl, E. *Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2*. *bioRxiv* (Cold Spring Harbor Labs Journals, 2016). doi:10.1101/059717
10. Henderson, R. Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 18037–41 (2013).
11. Henderson, R. *et al.* Outcome of the first electron microscopy validation task force meeting. *Structure* **20,** 205–214 (2012).
12. Scheres, S. H. W. A bayesian view on cryo-EM structure determination. *J. Mol. Biol.* **415,** 406–418 (2012).
13. Grigorieff, N. FREALIGN: High-resolution refinement of single particle structures. *J. Struct. Biol.* **157,** 117–125 (2007).
14. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193,** 1–12 (2016).
15. Hoang, T. V, Cavin, X., Schultz, P. & Ritchie, D. W. gEMpicker: a highly parallel GPU-accelerated particle picking tool for cryo-electron microscopy. *BMC Struct. Biol.* **13,** 25 (2013).
16. Moore, G. E. Progress in digital integrated electronics. *1975 Int. Electron Devices Meet.* **21,** 11–13 (1975).
17. Sigworth, F. J. A maximum-likelihood approach to single-particle image refinement. *J. Struct. Biol.* **122,** 328–39 (1998).
18. Nocedal, J. & Wright, S. J. *Numerical Optimization*. (Springer New York, 2000). doi:10.1007/BF01068601
19. Calafiore, G. C. & El Ghaoui, L. *Optimization Models*. (Cambridge University Press, 2014).
20. Bottou, L. Large-Scale Machine Learning with Stochastic Gradient Descent. *Proc. COMPSTAT'2010* 177–186 (2010). doi:10.1007/978-3-7908-2604-3_16
21. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* 1–9 (2012).

doi:http://dx.doi.org/10.1016/j.protcy.2014.09.007
22. Taigman, Y., Yang, M., Ranzato, M. & Wolf, L. DeepFace: Closing the gap to human-level performance in face verification. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* 1701–1708 (2014). doi:10.1109/CVPR.2014.220
23. Schep, D. G., Zhao, J. & Rubinstein, J. L. Models for the a subunits of the Thermus thermophilus V/A-ATPase and Saccharomyces cerevisiae V-ATPase enzymes by cryo-EM and evolutionary covariance. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 3245–3250 (2016).
24. Zhao, J., Benlekbir, S. & Rubinstein, J. Electron cryomicroscopy observation of rotational states in a eukaryotic V-ATPase. *Nature* **521,** 241–245 (2015).
25. Kearfott, R. B. *Rigorous global search: Continuous problems*. *Igarss 2014* (Springer US, 2014). doi:10.1007/s13398-014-0173-7.2
26. Little, J. D. C., Karel, C., Murty, K. G. & Sweeney, D. W. An algorithm for the traveling salesman problem. *Oper. Res.* **11,** 972–989 (1963).
27. Yang, J., Li, H. & Jia, Y. Go-ICP: Solving 3D Registration Efficiently and Globally Optimally. *2013 IEEE Int. Conf. Comput. Vis.* 1457–1464 (2013). doi:10.1109/ICCV.2013.184
28. Campbell, M. G., Veesler, D., Cheng, A., Potter, C. S. & Carragher, B. 2.8 ?? Resolution Reconstruction of the Thermoplasma Acidophilum 20 S Proteasome Using Cryo-Electron Microscopy. *Elife* **2015,** 1–22 (2015).
29. Wong, W. *et al.* Cryo-EM structure of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. *Elife* **2014,** 1–20 (2014).
30. Scheres, S. H. W. & Chen, S. Prevention of overfitting in cryo-EM structure determination. *Nat. Methods* **9,** 853–854 (2012).
31. Rosenthal, P. B. & Henderson, R. Optimal determination of particle orientation, absolute hand, and contrast loss in single-particle electron cryomicroscopy. *J. Mol. Biol.* **333,** 721–745 (2003).
32. Chen, S. *et al.* High-resolution noise substitution to measure overfitting and validate resolution in 3D structure determination by single particle electron cryomicroscopy. *Ultramicroscopy* **135,** 24–35 (2013).
33. Yan, X., Cardone, G., Zhang, X., Zhou, Z. H. & Baker, T. S. Single particle analysis integrated with microscopy: A high-throughput approach for reconstructing icosahedral particles. *J. Struct. Biol.* **186,** 8–18 (2014).
34. Murray, S. C. *et al.* Validation of cryo-EM structure of IP3R1 channel. *Structure* **21,** 900–909 (2013).
35. Radermacher, M., Wagenknecht, T., Verschoor, A. & Frank, J. Three-dimensional reconstruction from a single-exposure, random conical tilt series applied to the 50S ribosomal subunit of Escherichia coli. *J. Microsc.* **146,** 113–136 (1987).
36. Leschziner, A. E. & Nogales, E. The orthogonal tilt reconstruction method: An approach to generating single-class volumes with no missing cone for ab initio reconstruction of asymmetric particles. *J. Struct. Biol.* **153,** 284–299 (2006).
37. Penczek, P. A. & Asturias, F. J. Ab initio cryo-EM structure determination as a validation problem. in *2014 IEEE International Conference on Image Processing (ICIP)* 2090–2094 (2014). doi:10.1109/ICIP.2014.7025419
38. Sorzano, C. O. S. *et al.* A statistical approach to the initial volume problem in Single particle analysis by electron microscopy. *J. Struct. Biol.* **189,** 213–219 (2015).
39. Jaitly, N., Brubaker, M. A., Rubinstein, J. L. & Lilien, R. H. A Bayesian method for 3D

macromolecular structure inference using class average images from single particle electron microscopy. *Bioinformatics* **26,** 2406–2415 (2010).
40. Elmlund, D. & Elmlund, H. SIMPLE: Software for ab initio reconstruction of heterogeneous single-particles. *J. Struct. Biol.* **180,** 420–427 (2012).
41. Elmlund, H., Elmlund, D. & Bengio, S. PRIME: Probabilistic initial 3D model generation for single-particle cryo-electron microscopy. *Structure* **21,** 1299–1306 (2013).
42. Brubaker, M. A., Punjani, A. & Fleet, D. J. Building proteins in a day: Efficient 3D molecular reconstruction. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **07–12–June,** (2015).
43. Dvornek, N. C., Sigworth, F. J. & Tagare, H. D. SubspaceEM: A fast maximum-a-posteriori algorithm for cryo-EM single particle reconstruction. *J. Struct. Biol.* **190,** 200–214 (2015).
44. Cianfrocco, M. A. & Leschziner, A. E. Low cost, high performance processing of single particle cryo-electron microscopy data in the cloud. *Elife* **4,** e06664 (2015).
45. Bai, X.-C., Rajendra, E., Yang, G., Shi, Y. & Scheres, S. H. Sampling the conformational space of the catalytic subunit of human γ-secretase. *Elife* **4,** e11182 (2015).
46. Rubinstein, J. L. & Brubaker, M. A. Alignment of cryo-EM movies of individual particles by optimization of image translations. *J. Struct. Biol.* **192,** 188–195 (2015).
47. Grant, T. & Grigorieff, N. Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. *Elife* **4,** e06980 (2015).
48. Iudin, A., Korir, P. K., Salavert-Torres, J., Kleywegt, G. J. & Patwardhan, A. EMPIAR: a public archive for raw electron microscopy image data. *Nat. Methods* **13,** 387–388 (2016).
49. Punjani, A., Rubinstein, J., Fleet, D. and Brubaker, M. "Protocol for rapid unsupervised cryo-EM structure determination using cryoSPARC software" *Protocol Exchange* (2016) DOI: [to-be-provided].

Figure Legends for main text

Figure 1. Stochastic gradient descent for cryo-EM map calculation. A, Iterative refinement methods are sensitive to initialization. An arbitrary initialization far from the correct 3-D map will be refined into an incorrect structure that attains a locally optimal probability within the space of all 3-D maps. An accurate initialization will be refined to the correct structure. Iterative refinement uses all single particle images in a dataset to compute each step. **B,** Random selection of particle images in the SGD algorithm. At each iteration, a different small random selection of images is used to approximate the true optimization objective. Each iteration may use a different number of images. **C,** Stochastic Gradient Descent (SGD) algorithm enables *ab initio* structure determination through insensitivity to initialization. An arbitrary computer generated random initialization is incrementally improved by many noisy steps. Each step is based on the gradient of the approximated objective function obtained by random selection in (B). These approximate gradients do not exactly match the overall optimization objective. The success of SGD is commonly explained by the noisy sampling approximation allowing the algorithm to widely explore the space of all 3-D maps to arrive finally near the correct structure.

Figure 2. Evolution of 3-D cryo-EM maps as computation progresses. A. Low-resolution map of the TRPV1 channel calculated in 75 minutes from random initialization. **B.** Multiple conformations of the *Thermus thermophilus* V/A-ATPase calculated simultaneously from separate random initializations. **C.** Refinement of TRPV1 to 3.3 Å resolution on a single GPU desktop workstation in 66 minutes with C4 symmetry enforced. Density is apparent that corresponds to amino acid side chains. **D.** Refinement of each of three V/A-ATPase rotational states. The rotational state of the central rotor (indicated by red circles) is seen in cross sections through the 3-D maps. All computations were performed on a single desktop computer with a single NVIDIA Tesla K40 GPU. Scale bars, 25 Å.

Figure 3. The branch and bound approach to high-resolution cryo-EM map refinement. A, Two iterations of a simplified 1-D representation of the branch and bound approach. Candidate poses are iteratively eliminated by evaluation of an inexpensive lower bound over all poses, and the true error function at the minimum of the lower bound. **B,** For cryo-EM images, the true error function over all poses (top left) for an individual particle (top right) is never evaluated. Instead, the entire lower bound is computed (middle left), the true error is calculated at the minimum of the bound, and all poses where the lower bound exceeds this calculated error are eliminated (middle right). A tighter lower bound is evaluated and the process repeated until the optimum pose is identified (bottom left and right).

Figure 4. High-resolution structures from branch and bound refinement. A, 80S ribosome structure refined to 3.2 Å resolution in 2.2 h with 105,247 particle images. Amino acids side chain and RNA base densities are clearly visible in α -helices, β-strands, and rRNA (inset). **B,** A 20S proteasome map refined to 2.8 Å in 70 min with 49,954 particle images and D7 symmetry enforced. Well-resolved densities are apparent for small and large residues (inset). Branch and bound refinement of both structures was initialized with *ab initio* maps from SGD. Scale bars, 25 Å.

Tables: None.

Online Methods

Statistics

In all 3-D map refinement experiments, the Fourier shell correlation (FSC) between two independently refined half-maps (the "gold standard") was used to assess resolution[30], along with the FSC=0.143 resolution criterion[31] and correction of the FSC for effects of masking by high-resolution noise substitution[32].

Computational Hardware

All experiments were carried out on a single desktop workstation, equipped with an Intel i7-5820K 6-core CPU, NVIDIA Tesla K40 GPU, 64GB of CPU RAM, and a 512GB SSD for file storage. Tests were also run and equivalent running times were achieved using the consumer-grade NVIDIA Titan Z GPU. It should be noted that at the time of writing, the Tesla K40 GPU is over two years old, and more recent GPU cards will perform significantly faster.

Implementation

CryoSPARC is a software package written in a mixture of Python, CUDA C, and Javascript. Algorithms are implemented in Python and the GPU computation routines are written in CUDA C. Computations are parallelized over images, pixels, and search parameters. Two CPU threads are used for the GPU to improve utilization, and images are loaded from SSD and prepared by the CPU simultaneously with GPU processing of a different batch of images.

Stochastic Gradient Descent

SGD is initialized from a computer generated random initialization for each 3D class (Supplementary Note 1). The number of images used in each iteration of SGD is automatically determined based on the current resolution. A model of the noise level in single particle images is initialized with an over-estimate relative to measured noise levels. Approximate gradients of Equation 1 are computed along with second-order curvature information to enable estimation of an optimal step size for descent at each iteration. Step directions are averaged over iterations using a classical momentum method[50]. Resulting iterative steps are applied to the 3-D maps and a projection operation is used to enforce non-negativity of 3-D map density. The noise model is refined based on errors between the images and projections of the 3-D map at each iteration, converging to the optimal noise model over several iterations. The descent step size is decreased monotonically over iterations to improve convergence once an approximately correct structure is found. Further details can be found in Supplementary Note 1.

Branch and Bound Image Alignment

The branch and bound method is applied to each image individually at each iteration of high-resolution map refinement. A space partitioning tree-structure is used to segment the space of orientation parameters, which are represented using axis-angle coordinates. A coarse initial sampling of the orientation space forms the first level of the tree, and each stage of branch and

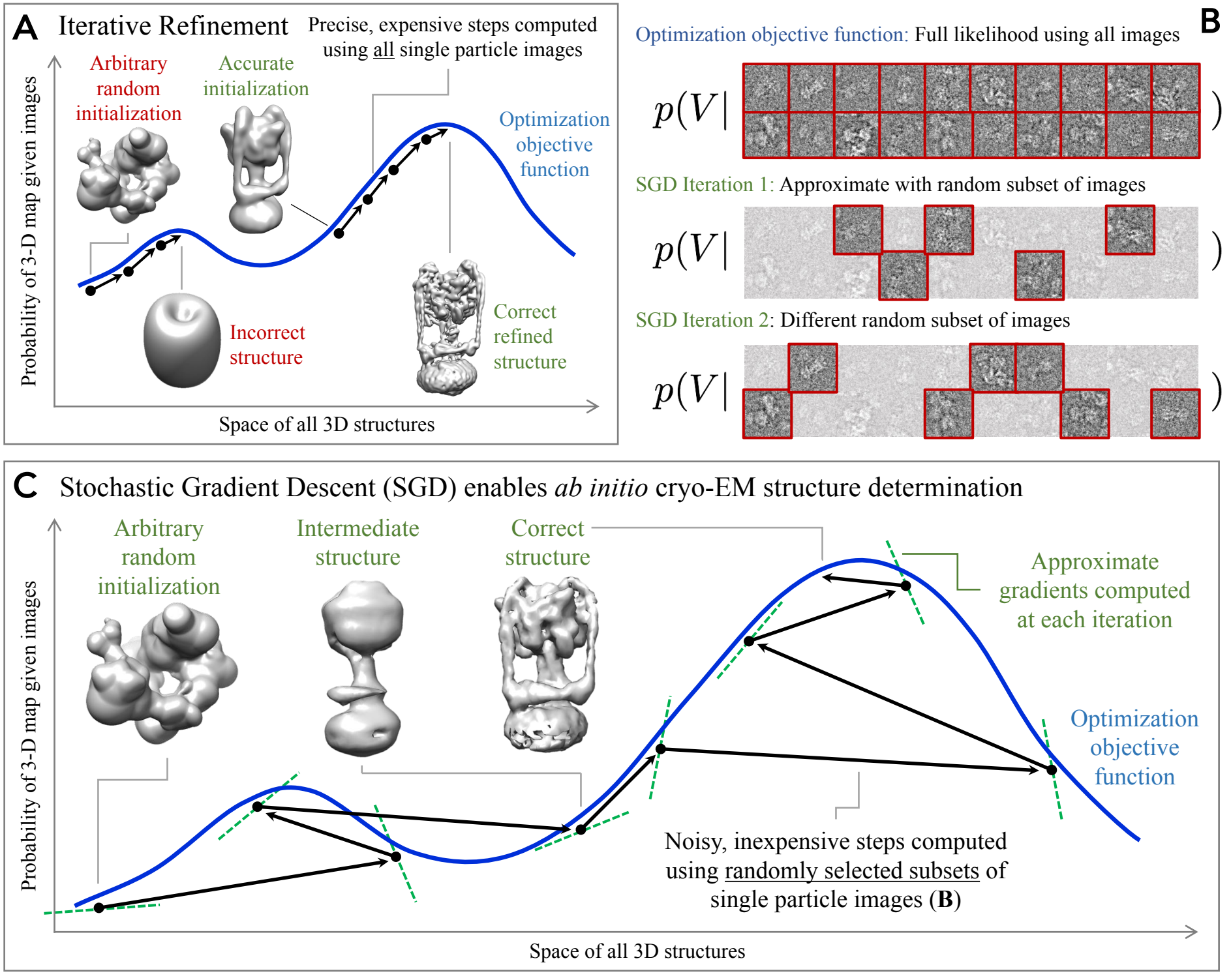
bound subdivides and prunes branches in the tree until only the optimal pose remains to within a specified angular precision of 0.18°. A similar tree structure is used to segment and subdivide the 2-D space of in-plane shifts for each image, resulting in a specified translational precision of 0.04 pixels. Further details including the derived lower bound and approximations can be found in Supplementary Note 2.

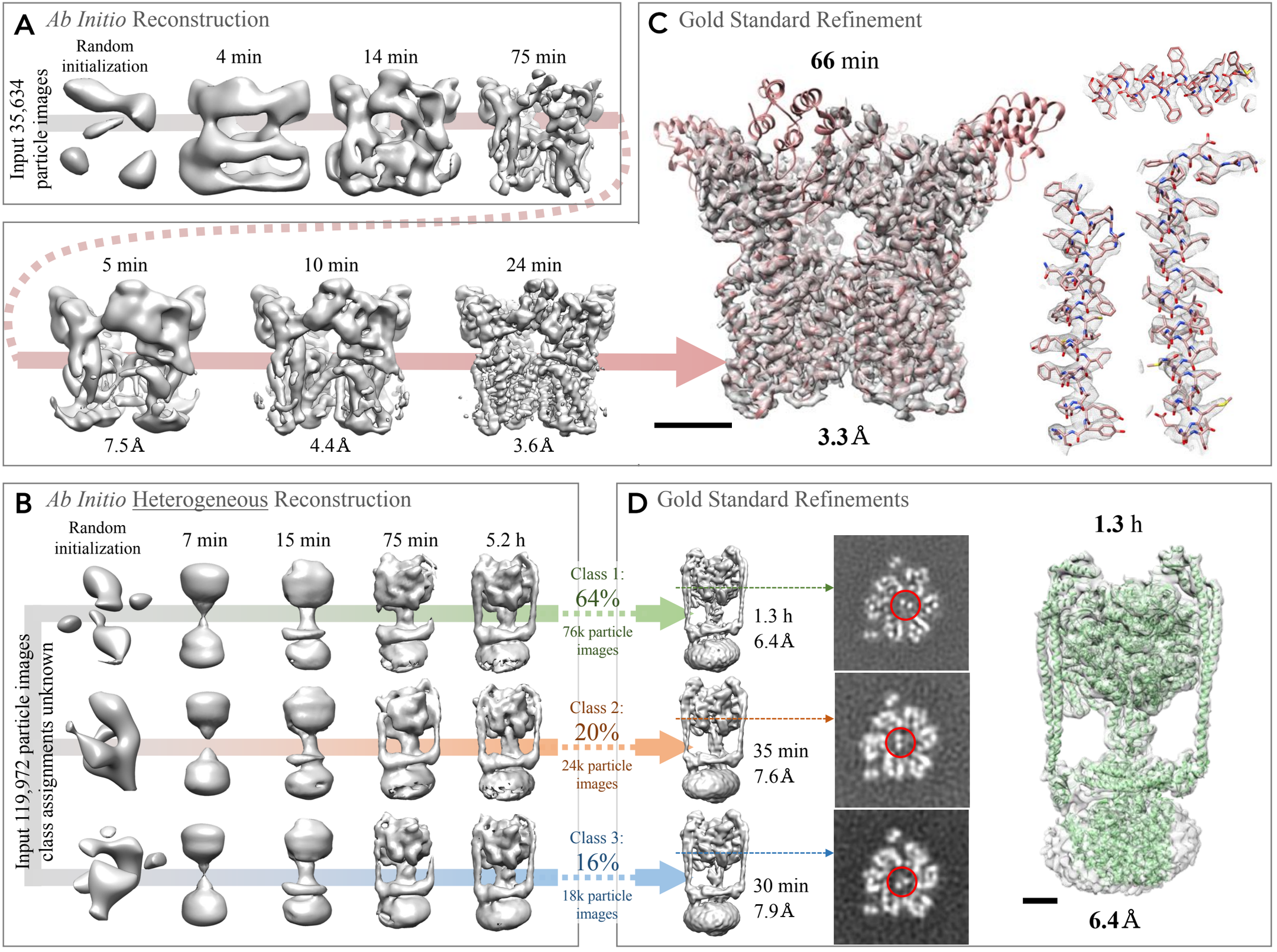
Program Settings

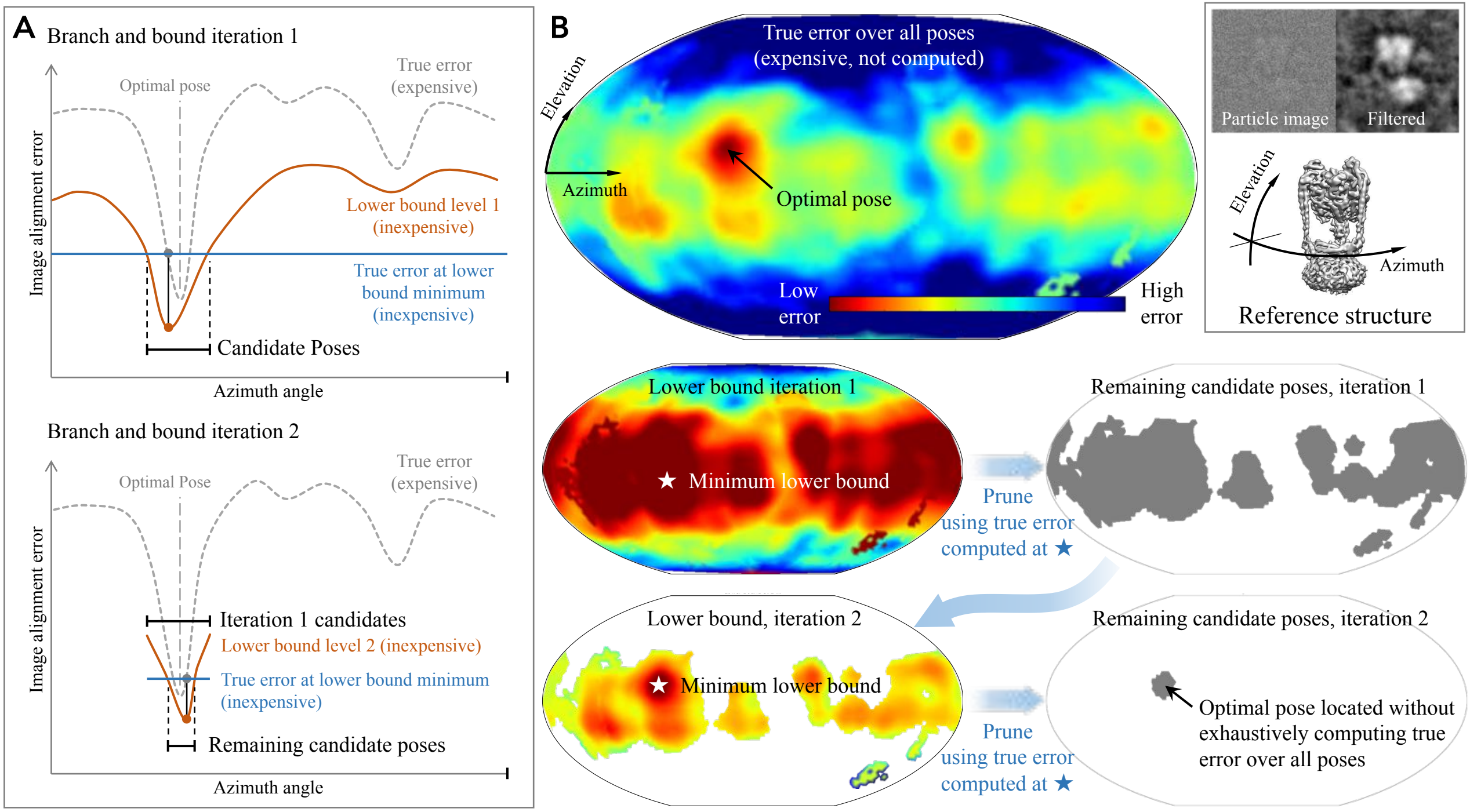
Default cryoSPARC settings were used in all refinement experiments, and the number of classes was set in each *ab initio* reconstruction experiment. Symmetry was enforced in refinement experiments where noted, but not in *ab initio* reconstruction.

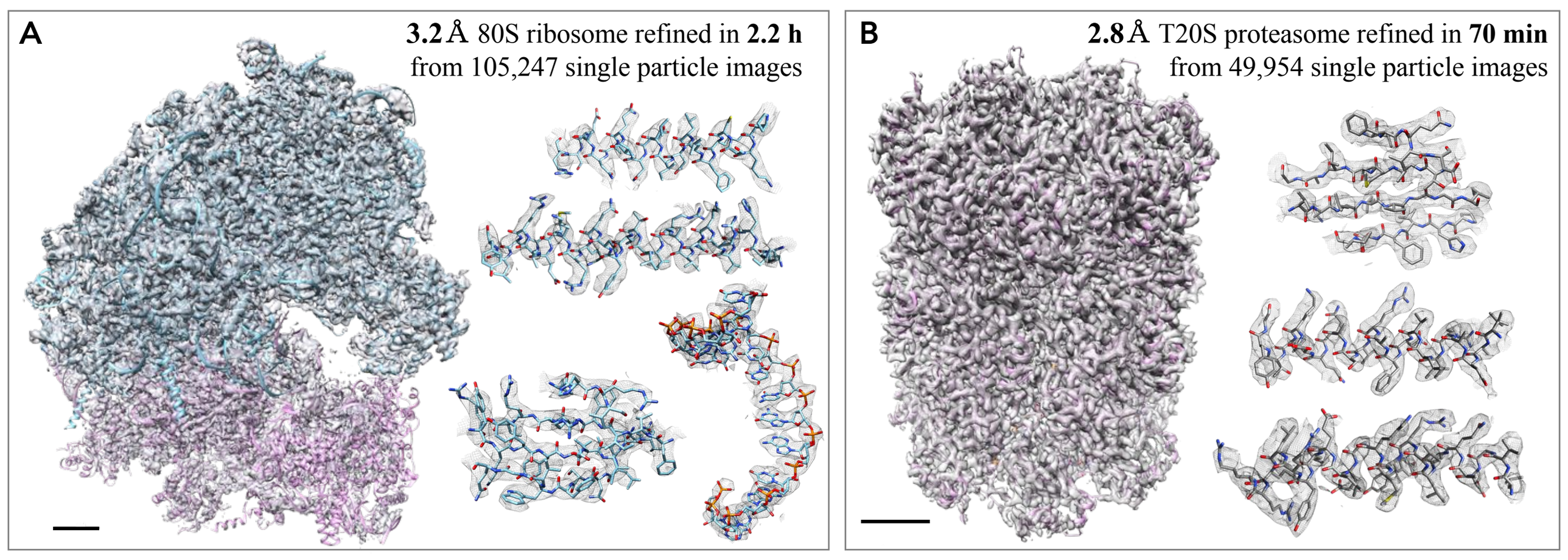
Methods-only References

50. Sutskever, I., Martens, J., Dahl, G. E. & Hinton, G. E. On the importance of initialization and momentum in deep learning. *ICML* **28,** 1139–1147 (2013).


A
Iterative Refinement
Precise, expensive steps computed using all single particle images
Arbitrary random initialization
Accurate initialization
Optimization objective function
Incorrect structure
Correct refined structure
Probability of 3-D map given images
Space of all 3D structures
B
Optimization objective function: Full likelihood using all images
$p(V|$)
SGD Iteration 1: Approximate with random subset of images
$p(V|$)
SGD Iteration 2: Different random subset of images
$p(V|$)
C
Stochastic Gradient Descent (SGD) enables *ab initio* cryo-EM structure determination
Arbitrary random initialization
Intermediate structure
Correct structure
Approximate gradients computed at each iteration
Optimization objective function
Noisy, inexpensive steps computed using randomly selected subsets of single particle images (**B**)
Probability of 3-D map given images
Space of all 3D structures


A
Ab Initio Reconstruction
Random initialization
4 min
14 min
75 min
Input 35,634 particle images
5 min
10 min
24 min
7.5Å
4.4Å
3.6Å
C
Gold Standard Refinement
66 min
3.3Å
B
Ab Initio Heterogeneous Reconstruction
Random initialization
7 min
15 min
75 min
5.2 h
Input 119,972 particle images
class assignments unknown
Class 1:
64%
76k particle images
Class 2:
20%
24k particle images
Class 3:
16%
18k particle images
D
Gold Standard Refinements
1.3 h
6.4Å
35 min
7.6Å
30 min
7.9Å
1.3 h
6.4Å


A
Branch and bound iteration 1
Optimal pose
True error
(expensive)
Image alignment error
Lower bound level 1
(inexpensive)
True error at lower
bound minimum
(inexpensive)
Candidate Poses
Azimuth angle
Branch and bound iteration 2
Optimal Pose
True error
(expensive)
Image alignment error
Iteration 1 candidates
Lower bound level 2 (inexpensive)
True error at lower bound minimum
(inexpensive)
Remaining candidate poses
Azimuth angle
B
True error over all poses
(expensive, not computed)
Elevation
Azimuth
Optimal pose
Low
error
High
error
Particle image
Filtered
Elevation
Azimuth
Reference structure
Lower bound iteration 1
★ Minimum lower bound
Prune
using true error
computed at ★
Remaining candidate poses, iteration 1
Lower bound, iteration 2
★ Minimum lower bound
Prune
using true error
computed at ★
Remaining candidate poses, iteration 2
Optimal pose located without
exhaustively computing true
error over all poses


A
3.2 Å 80S ribosome refined in 2.2 h
from 105,247 single particle images
B
2.8 Å T20S proteasome refined in 70 min
from 49,954 single particle images