cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination

Ali Punjani[1], John L Rubinstein[2–4], David J Fleet[5] & Marcus A Brubaker[6]

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 3D 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* 3D 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 (http://www.cryosparc.com).

Scientific approaches can be transformed by innovations that decrease the cost and improve nonexpert 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 both experts and nonspecialists and qualitatively by changing the type and scope of feasible experiments. 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 become increasingly important as the technique is used for structure-based drug design[6] and time-critical structural studies of pathogens[7]. Given appropriately prepared specimens, automated electron microscopes can collect data sets for atomic-resolution structure determination in as little as 24 or 48 h; and centralized cryo-EM facilities are now providing instrument access to nonspecialist investigators. Calculation of 3D maps from cryo-EM images, however, can require weeks of computational analysis by an expert user. With routine collection of cryo-EM data sets that contain millions of single-particle images corresponding to different 3D conformations of the sample[8], the cost of image analysis can exceed 500,000 CPU hours on large, expensive computer clusters[9]. Furthermore, without significant user expertise, there are a variety of ways in which incorrect and misleading 3D 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 for both automation and widespread use of cryo-EM.

To address these issues, we developed two new algorithms. The first of these algorithms makes it possible to perform unsupervised *ab initio* 3D classification, whereby multiple 3D states of a protein can be discovered from a single sample without user input of prior structural knowledge and without the assumption that all 3D states resemble each other. In contrast, existing techniques for 3D 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 techniques are combined in a standalone graphics processing unit (GPU) accelerated software package that we have named cryo-EM single-particle *ab initio* reconstruction and classification (cryoSPARC). 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 previously been used 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 3D classification and refinement[9]. The algorithms presented here provide a further order-of-magnitude reduction in computational cost compared with that of GPU acceleration, which would require at least an additional ~7 years

[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 should be addressed to A.P. (alipunjani@cs.toronto.edu) or M.A.B. (mab@eecs.yorku.ca).

RECEIVED 18 AUGUST 2016; ACCEPTED 27 DECEMBER 2016; PUBLISHED ONLINE 6 FEBRUARY 2017; DOI:10.1038/NMETH.4169

if driven by hardware advances alone[16]. Based on the combination of algorithms, inexpensive hardware, and an easy-to-use graphical user interface, cryoSPARC will enable nonspecialist 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]:

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

The aim of the optimization is to find the 3D 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 3D rotation and a 2D translation for each single-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 3D maps represented as density voxels on a 3D grid.

We use an SGD optimization scheme to quickly identify one or several low-resolution 3D 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

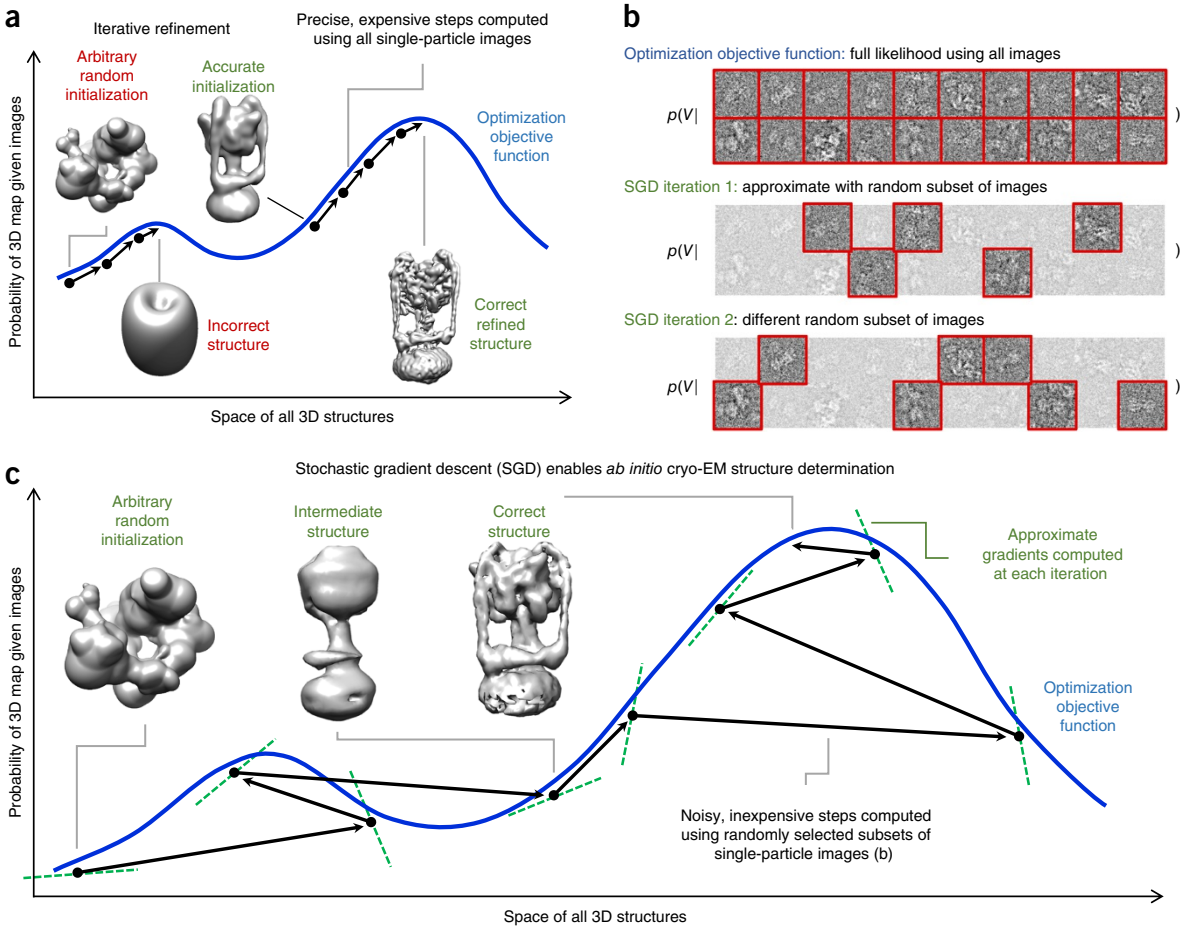

Figure 1 | Stochastic gradient descent for cryo-EM map calculation. (**a**) Iterative refinement methods are sensitive to initialization. An arbitrary initialization far from the correct 3D map will be refined into an incorrect structure that attains a locally optimal probability within the space of all 3D maps. An accurate initialization will be refined to the correct structure. Iterative refinement uses all single-particle images in a data set 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 3D maps to finally arrive near the correct structure.

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 nonconvex optimization problem. This type of problem is among the most computationally challenging optimization problems known and is 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] (**Fig. 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. Recently, however, methods have been discovered that perform well on nonconvex problems. One such method is SGD[20] (**Fig. 1**). SGD was popularized as a key tool in deep learning for the optimization of nonconvex functions, and it results 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 3D map. SGD repeatedly approximates this objective function by selecting a different random subset of terms (i.e., single-particle images) at each iteration, and it computes the sum of those terms (**Fig. 1b**). In a single iteration, the optimization variables (i.e., the 3D 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. Consequently, a single iteration is inexpensive, and hundreds or thousands of iterative changes can be made during each pass through the full data set. 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 nonconvex problems (**Fig. 1c**).

We implemented an SGD method for *ab initio* structure determination and 3D classification. Applied to several different data sets, the use of SGD enabled convergence to correct low-resolution structures from arbitrary random initialization, allowing both *ab initio* structure determination and *ab initio* 3D classification (**Fig. 2**). With 35,645 TRPV1 particle images[3], SGD optimization resulted in a low-resolution 3D map in 75 min from random initialization (**Fig. 2a**) using a single inexpensive desktop workstation with an Intel i7-5820K Processor and a single NVIDIA Tesla K40 GPU. When applied to a data set of conformationally heterogeneous *Thermus thermophilus*

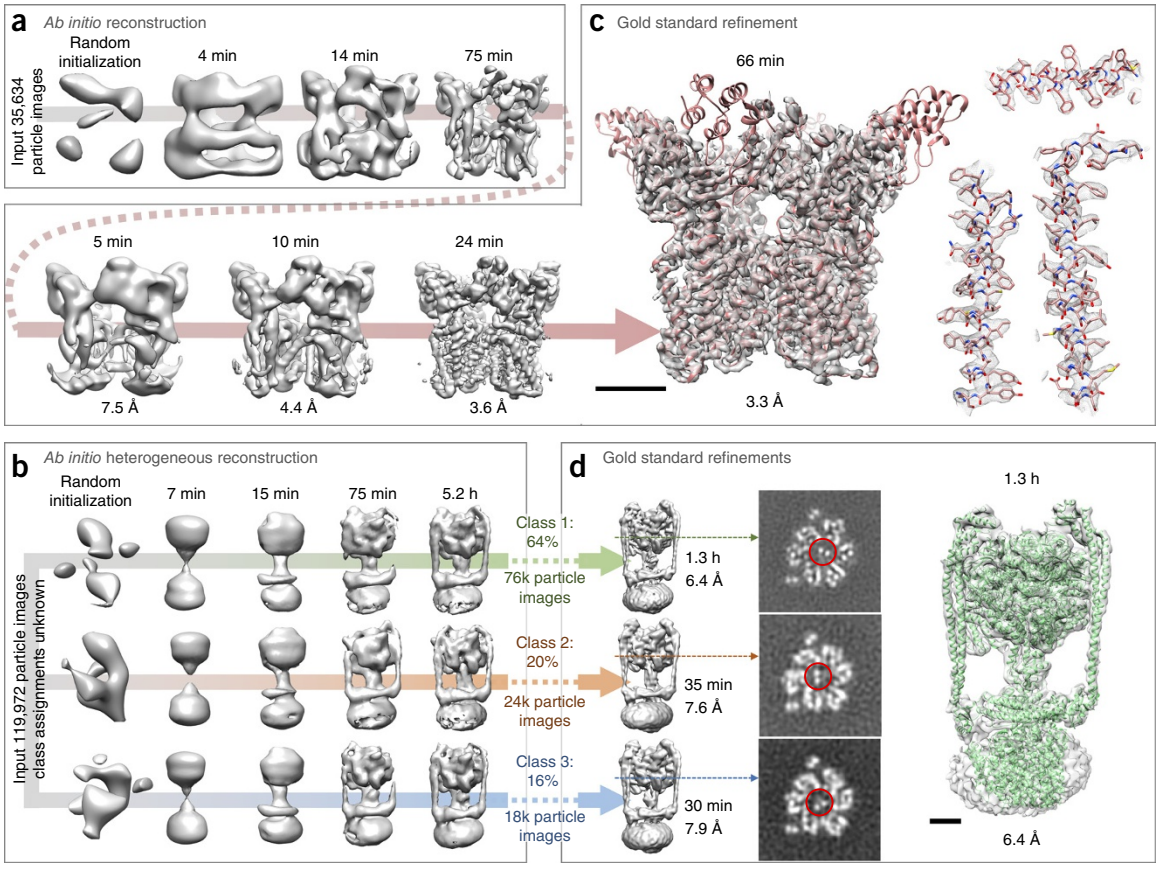

Figure 2 | Evolution of 3D cryo-EM maps as computation progresses using the SGD algorithm and branch-and-bound refinement. (**a**) Low-resolution map of the TRPV1 channel calculated in 75 min from random initialization. (**b**) Multiple conformations of the *Thermus thermophilus* V/A-ATPase calculated simultaneously from separate random initializations. (**c**) Refinement of TRPV1 to 3.3-Å resolution on a single GPU desktop workstation in 66 min with C4 symmetry enforced. (**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 3D maps. All computations were performed on a single desktop computer with a single NVIDIA Tesla K40 GPU. Scale bars, 25 Å.

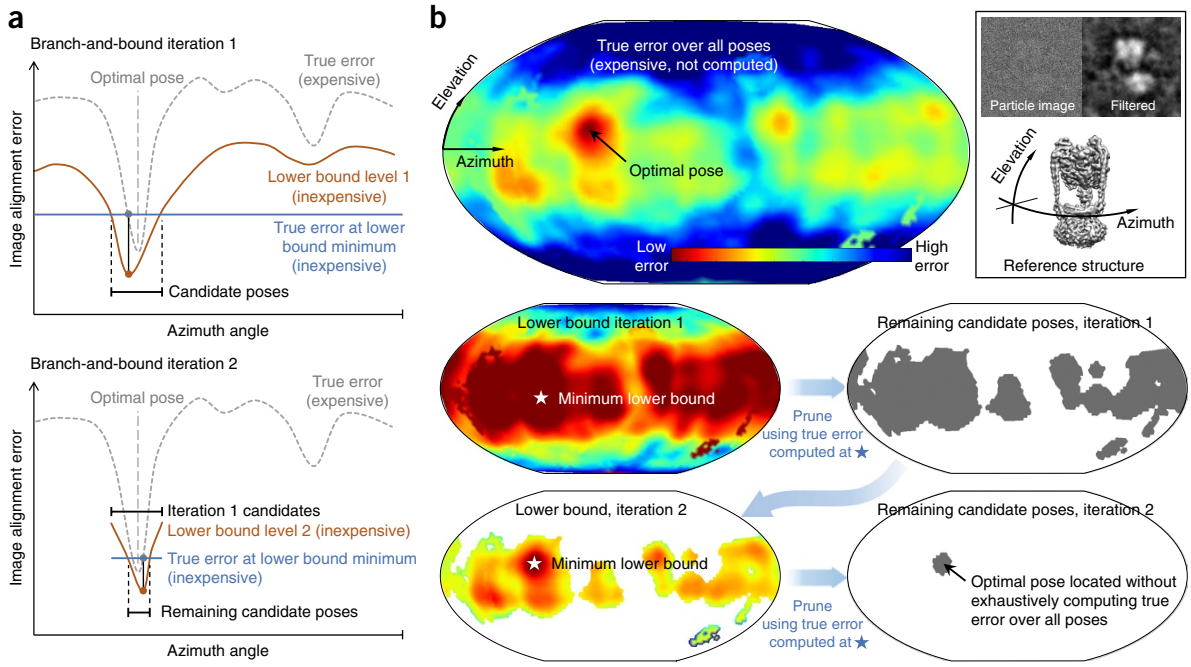

Figure 3 | The branch-and-bound approach to high-resolution cryo-EM map refinement. (**a**) Two iterations of a simplified 1D representation of the branch-and-bound approach. Candidate poses are iteratively eliminated by evaluation of an inexpensive lower bound over all poses, and the true error function 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).

V/A-ATPase particle images[23], the algorithm discerned three different conformational states for the enzyme, again from random initializations (**Fig. 2b**). These three states corresponded 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 data set of images only detected 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 structures expected to be present in the data set.

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 2D single-particle image to a 3D 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 (**Fig. 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 3D map. To find this optimal pose using the branch-and-bound approach (**Fig. 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 is 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 that was found to be equally effective but even more efficient is used (**Supplementary Note 2**).

We implemented the branch-and-bound approach and applied it to high-resolution structure determination from several published data sets: the 20S proteasome from *Thermoplasma acidophilum*[28], the 80S ribosome from *Plasmodium falciparum*[29], amphipol-solubilized rat TRPV1 (ref. 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

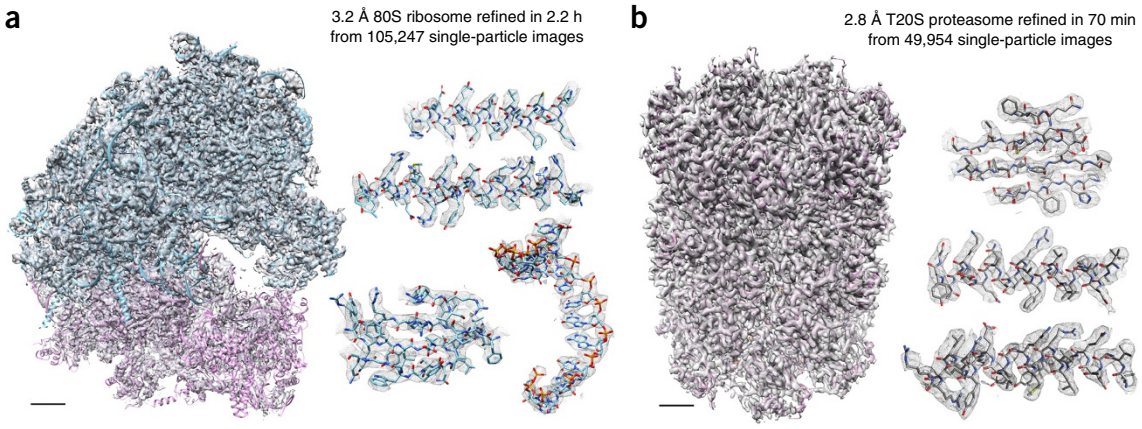

Figure 4 | High-resolution structures from branch-and-bound refinement. (**a**) 80S ribosome structure refined to 3.2 Å resolution in 2.2 h with 105,247 particle images. Amino acid side chain and RNA base densities are clearly visible in α-helices, β-strands, and rRNA (inset). (**b**) A 20S proteasome 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.

images, branch-and-bound orientation determination produced a 3.3-Å-resolution map (**Supplementary Data 1**) in 66 min with C4 symmetry enforced using a gold-standard refinement procedure[30], the Fourier shell correlation (FSC) = 0.143 resolution criterion[31], and correction for effects of masking on the FSC by high-resolution noise substitution[32] (**Fig. 2c**). This resolution slightly exceeds the previously published resolution of 3.4 Å from the same data set[3] (EMDB, 5778; PDB, 3J5P). 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 h (**Fig. 2d** and **Supplementary Data 2**). The resolutions estimated for the states were 6.4 Å, 7.6 Å, and 7.9 Å compared with 6.4 Å and 9.5 Å for the two states identified in the previously published analysis[23] (EMDB, 8016; EMDB, 8017; PDB, 5GAR; PDB, 5GAS).

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 (**Supplementary Data 3**), equivalent to the published resolution[29] (EMDB, 2660; PDB, 3J79; PDB, 3J7A), in 2.2 h (**Fig. 4a**), demonstrating the method's capability to deal with large and asymmetric protein complexes. Notably, on the same computer hardware (desktop computer with one GPU), this data set of particle images would take approximately 20 h for refinement using the GPU accelerated program RELION[9]. Similarly, the 20S proteasome structure was refined to 2.8 Å with D7 symmetry enforced (**Supplementary Data 4**), matching the published sub-3-Å resolution (EMDB, 6287; PDB, 1YAR) from the data set[28] but in only 70 min (**Fig. 4b**). 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 3D maps from cryo-EM images has long been known to be 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 such a 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, and they are thus intrinsically slower than our approach, particularly as the number of particle images in data sets grow. In contrast, SGD processes random subsets of data at each iteration, making it efficient even when applied to large data sets.

We previously showed that SGD could produce a reasonable low-resolution map *ab initio* for a homogenous data set[42]. Here we have demonstrated that SGD is sufficiently robust to enable reconstruction of multiple 3D classes from independent arbitrary initializations. To our knowledge, all existing techniques for 3D classification use a single initial reference from which analysis of heterogeneity proceeds. Removal of the assumption that all 3D classes are similar to the single input reference is particularly advantageous for discovering 3D 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 3D 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 data sets with little contrast at low resolution. This situation can occur when insufficient defocus is used with a cryo-EM microscope that does not possess 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 data set 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 reimplement 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 (ref. 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 3D classes[8,45], improved computational efficiency through algorithm development will be a critical enabler for both academic and industrial researchers using cryo-EM.

The cryoSPARC software is available as a standalone program that can run on either commodity desktop workstations or rack-mount servers. 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 for both local and remote usage. This GUI allows multiple users within a laboratory to have separate accounts, access the program remotely, upload and share data sets, manage experimental results, launch computational tasks, and view results streaming in real time as they are computed.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank S. Dawood for construction of the GUI front end and members of the Rubinstein laboratory for testing cryoSPARC. A.P. was supported by a scholarship from the Natural Sciences and Engineering Research Council (NSERC), J.L.R. was supported by the Canada Research Chairs program, and D.J.F. 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 (D.J.F.) and 401724-12 (J.L.R.)) and an NVIDIA Academic Hardware Grant (M.A.B. and A.P.). Part of this work was performed while M.A.B. was a postdoctoral fellow at the University of Toronto.

AUTHOR CONTRIBUTIONS

A.P. and M.A.B. designed algorithms and implemented software. A.P., M.A.B. and J.L.R. performed experimental work. J.L.R., D.J.F., and M.A.B. contributed expertise and supervision. All authors contributed to manuscript preparation.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Kühlbrandt, W. Biochemistry. The resolution revolution. *Science* **343**, 1443–1444 (2014).
2. Smith, M.T.J. & Rubinstein, J.L. Structural biology. Beyond blob-ology. *Science* **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**, 107–112 (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* **2**, e00461 (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**, 871–875 (2016).
7. Sirohi, D. *et al.* The 3.8 Å resolution cryo-EM structure of Zika virus. *Science* **352**, 467–470 (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.H. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5**, e18722 (2016).
10. Henderson, R. Avoiding the pitfalls of single particle cryo-electron microscopy: Einstein from noise. *Proc. Natl. Acad. Sci. USA* **110**, 18037–18041 (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. In *Proc. Int. Elect. Devices Meet* 11–13 (IEEE, 1975).
17. Sigworth, F.J. A maximum-likelihood approach to single-particle image refinement. *J. Struct. Biol.* **122**, 328–339 (1998).
18. Nocedal, J. & Wright, S.J. *Numerical Optimization* (Springer, 2000).
19. Calafiore, G.C. & El Ghaoui, L. *Optimization Models* (Cambridge University Press, 2014).
20. Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proc. COMPSTAT'2010* (eds. Lechevallier, Y. & Saporta, G.) 177–186 (2010).
21. Krizhevsky, A., Sutskever, I. & Hinton, G.E. In *Adv. Neural Inf. Process. Syst.* (eds. Pereira, F., Burges, C.J.C. *et al.*) 1–9 (NIPS, 2012).
22. Taigman, Y., Yang, M., Ranzato, M. & Wolf, L. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* (eds. Dickinson, S. *et al.*) 1701–1708 (IEEE Computer Society, 2014).
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. USA* **113**, 3245–3250 (2016).
24. Zhao, J., Benlekbir, S. & Rubinstein, J.L. 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* (Springer, 2014).
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. In *Proc. IEEE Int. Conf. Comput. Vis.* (eds. Davis, L. & Hartley, R.) 1457–1464 (IEEE, 2013).

28. Campbell, M.G., Veesler, D., Cheng, A., Potter, C.S. & Carragher, B. 2.8 Å resolution reconstruction of the thermoplasma *Acidophilum* 20S proteasome using cryo-electron microscopy. *eLife* **4**, e06380 (2015).
29. Wong, W. *et al.* Cryo-EM structure of the Plasmodium falciparum 80S ribosome bound to the anti-protozoan drug emetine. *eLife* **3**, 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 IP_3R1 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 *Proc. IEEE Int. Conf. on Image Process.* (eds. Pesquet-Popescu, B. & Fowler, J.) 2090–2094 (IEEE, 2014).
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 *Proc. IEEE Comp. Soc. Conf. on Comput. Vis. Pattern Rec.* (eds. Bischof, H. *et al.*) (IEEE, 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).

ONLINE METHODS

A protocol detailing use of the software package has been prepared[48] (**Supplementary Protocol**).

Experimental results and test data sets. All cryo-EM images used to experimentally demonstrate the effectiveness of algorithms were taken from previously published studies. Several data sets were downloaded directly from EMPIAR[49]. In all cases, the single-particle images, their CTF parameters, and the microscope parameters that were used in the original studies were input directly into cryoSPARC with no further preprocessing. Result figures were generated by rigidly docking models provided in previously published studies into the new 3D maps generated by cryoSPARC.

Statistics. In all 3D 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, 64 GB of CPU RAM, and a 512 GB 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 2 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 overestimate 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 3D maps, and a projection operation is used to enforce non-negativity of 3D map density. The noise model is refined based on errors between the images and projections of the 3D 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 2D 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.

Software availability. The software package, including source code, is available for noncommercial use as a download at http://www.cryosparc.com. The software will also be available as a web service for new users to try before installing locally. Results reported in this work were computed using cryoSPARC version 0.2.36.

Data availability statement. The cryo-EM images used to experimentally demonstrate the effectiveness of algorithms were taken from previously published studies. Several data sets were downloaded directly from EMPIAR[49], including the 80S Ribosome (EMPIAR-10028), 20S proteasome (EMPIAR-10025), and 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. Refined maps are available in **Supplementary Data 1–4**.

48. Punjani, A., Rubinstein, J., Fleet, D. & Brubaker, M. Protocol for rapid unsupervised cryo-EM structure determination using cryoSPARC software. *Protocol Exchange* http://dx.doi.org/10.1038/protex.2017.009 (2016).
49. 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).
50. Sutskever, I., Martens, J., Dahl, G.E. & Hinton, G.E. On the importance of initialization and momentum in deep learning. *J. Mach. Learn. Res.* **28**, 1139–1147 (2013).