cryoSPARC: Algorithms for Rapid Unsupervised cryo-EM Structure Determination

Group Meeting

連昱翔 Yu-Hsiang Lien

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]

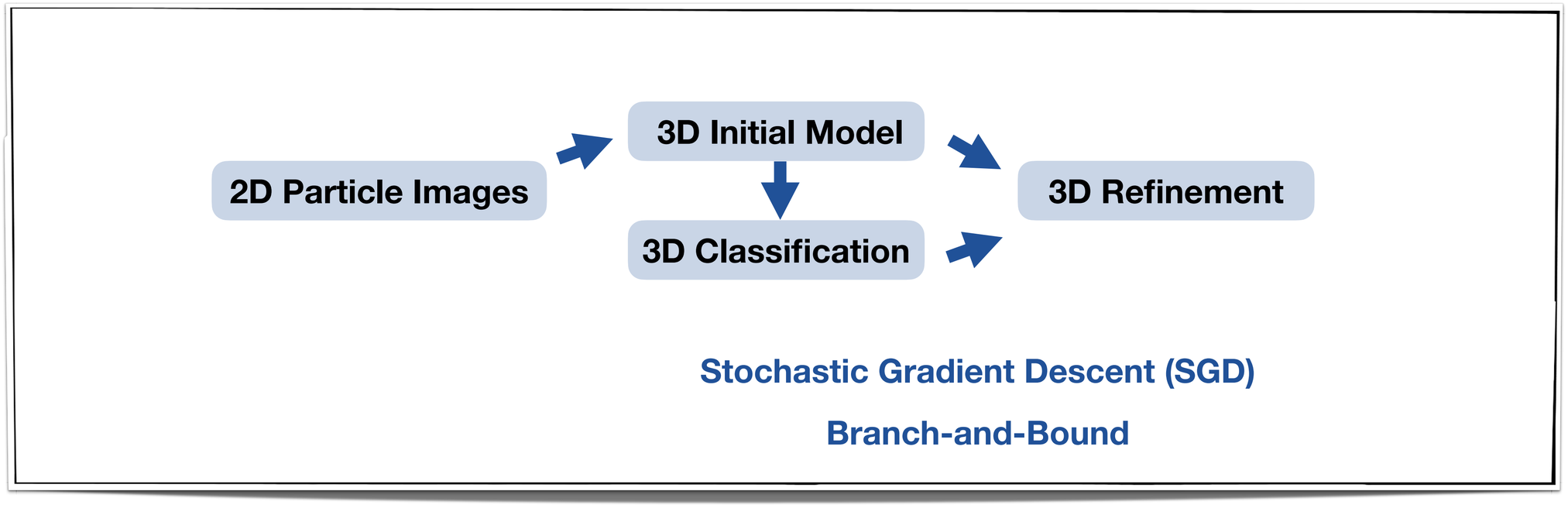
[1]Department of Computer Science, The University of Toronto, Toronto, Ontario, Canada. [2]Molecular Structure and Function Program, The Hospital for Sick Children Research Institute, Toronto, Ontario, Canada. [3]Department of Biochemistry, The University of Toronto, Toronto, Ontario, Canada. [4]Department 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

Single Particle 3D Reconstruction

cryoSPARC: cryo-EM Single-Particle *Ab initio* Reconstruction and Classification

- Two algorithms included in cryoSPARC:
  - Stochastic Gradient Descent (SGD)
  - Branch-and-Bound maximum likelihood optimization



Optimization of Objective Function

- The aim of the optimization is to find the 3D structures $V_1 \cdots V_K$ that best explain the observed images $X_1 \cdots X_N$ by marginalizing over class assignment j and the unknown pose variable ϕ_i (3D rotation + 2D translation) for each single particle image.
- The optimization is formulated as the maximization of an **objective function**:

$$\begin{aligned}
\arg\max_{V_{1\cdots K}} \log p(V_1, \cdots, V_K \,|\, X_1, \cdots, X_N) &= \arg\max_{V_{1\cdots K}} \log p(X_1, \cdots, X_N \,|\, V_1, \cdots, V_K) + \log p(V_1, \cdots, V_K) \\
&= \arg\max_{V_{1\cdots K}} \sum_{i=1}^{N} \log p(X_i \,|\, \mathbf{V}) + \sum_{j=1}^{K} \log p(V_j) \equiv f(\mathbf{V}) \\
&= \arg\max_{\mathbf{V}} f(\mathbf{V})
\end{aligned}$$

$$p(X_i \,|\, \mathbf{V}) = \sum_{j=1}^{K} \pi_j p(X_i \,|\, V_j) \equiv U_i \qquad p(X_i \,|\, V_j) = \int p(X_i \,|\, \phi, V_j) p(\phi) \mathrm{d}\phi$$

Objective function

$$f(\mathbf{V}) = \sum_{i=1}^{N} \log \left(\sum_{j=1}^{K} \pi_j \int p(X_i \,|\, \phi, V_j) p(\phi) \mathrm{d}\phi \right) + \sum_{j=1}^{K} \log p(V_j)$$

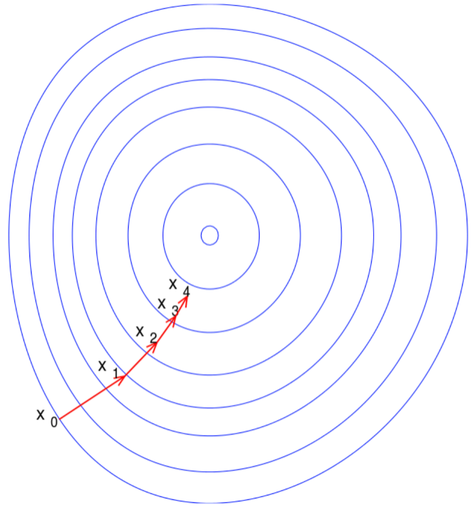
Gradient Descent

- **Gradient descent** is a first-order iterative optimization algorithm for finding a local minimum/maximum of a differentiable function.

$$\mathbf{V}^{(n+1)} = \mathbf{V}^{(n)} \pm \eta \nabla f\left(\mathbf{V}^{(n)}\right)$$

Step size

  - If the step size is too small, the movement in the search space will be small and the search will take a long time. If the step size is too large, the search may bounce around the search space and skip over the optima.



Ref: https://en.wikipedia.org/wiki/Gradient_descent

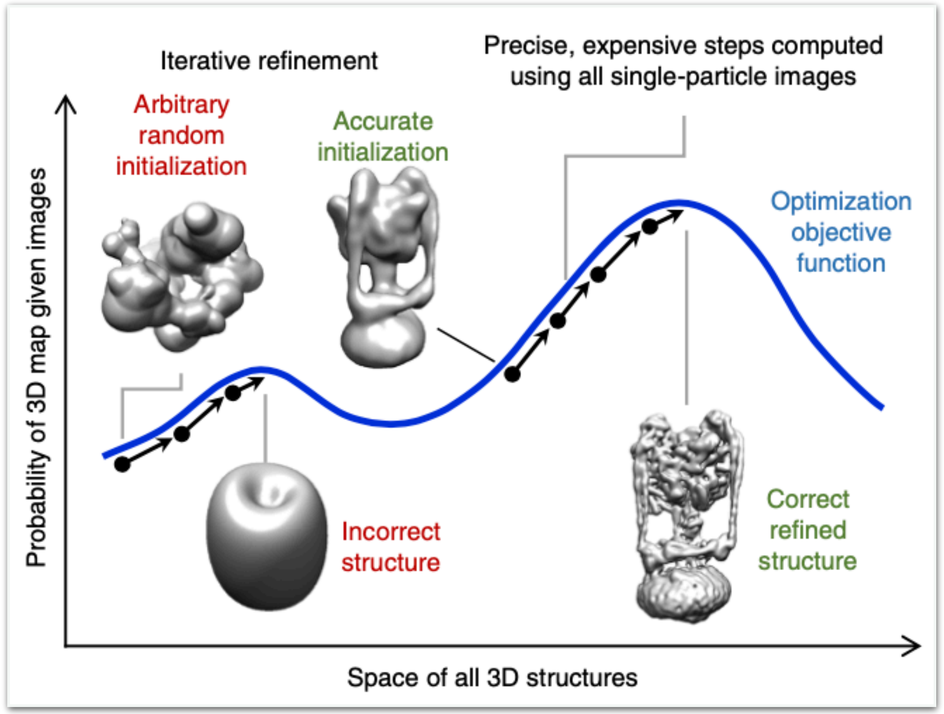
- **Gradient Descent with momentum**: an extension to the gradient descent optimization algorithm, which is designed to accelerate the optimization process.

$$\mathbf{V}^{(n+1)} = \mathbf{V}^{(n)} \pm \left\{ \mu \, \eta \nabla f\left(\mathbf{V}^{(n-1)}\right) + (1-\mu)\, \eta \nabla f\left(\mathbf{V}^{(n)}\right) \right\}$$

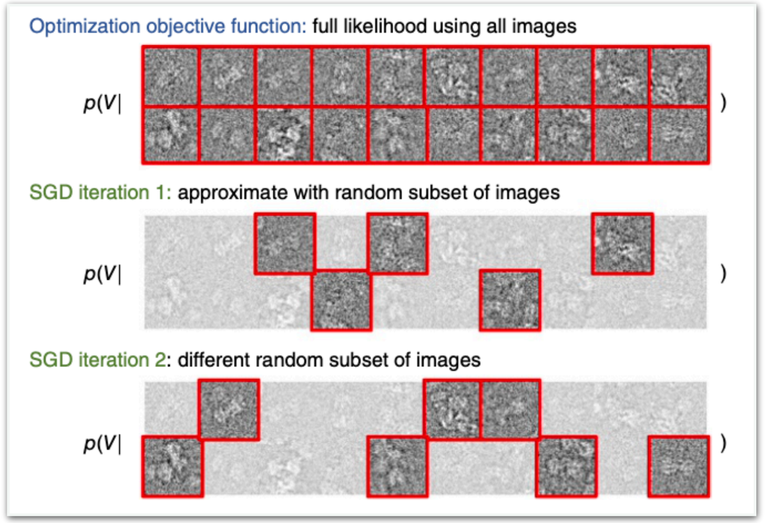
Momentum

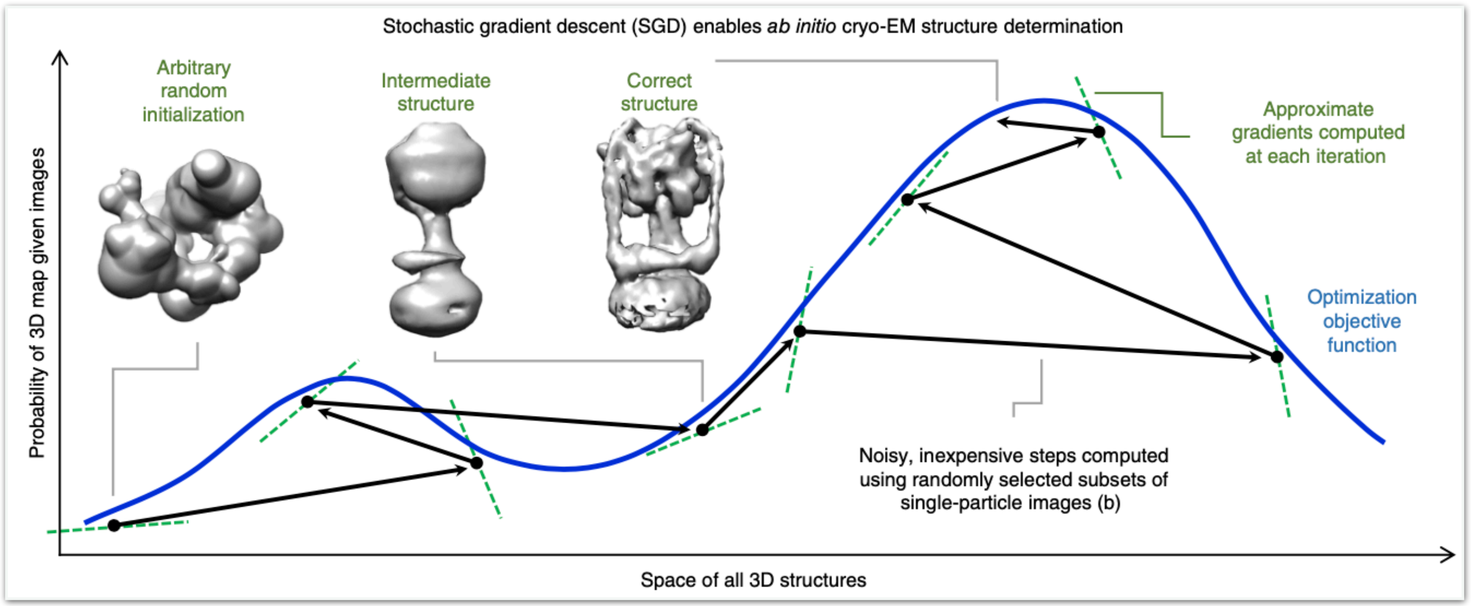
  - The momentum algorithm accumulates an exponentially decaying moving average of past gradients and continues to move in their direction.

Stochastic Gradient Descent (SGD)



SGD





- The two key parts of SGD:
  - Computation of the approximate gradients.
  - Determination of an appropriate parameter update step based on the gradient.

Gradient and Approximate Gradient

- SGD optimize the objective function $f(\mathbf{V})$ by iteratively updating the parameters $V_1 \cdots V_K$.
- The **gradient of** $f(\mathbf{V})$ w.r.t each structure V_k is computed in order to **take steps**:

Gradient

$$\frac{\partial f(\mathbf{V})}{\partial V_k} = \sum_{i=1}^{N} \frac{1}{U_i} \pi_k \int \frac{\partial}{\partial V_k} p(X_i \mid \phi, V_k) p(\phi) \mathrm{d}\phi + \frac{\partial}{\partial V_k} \log p(V_k)$$

- In SGD, the sum is approximated using **subsampling** (at each iteration, SGD selects a random subset of images from the dataset, called **minibatch**):

Approximate Gradient

$$\frac{\partial f(\mathbf{V})}{\partial V_k} \approx \frac{N}{M} \sum_{i \in \mathbf{M}} \frac{1}{U_i} \pi_k \int \frac{\partial}{\partial V_k} p(X_i \mid \phi, V_k) p(\phi) \mathrm{d}\phi + \frac{\partial}{\partial V_k} \log p(V_k) \equiv G_k$$

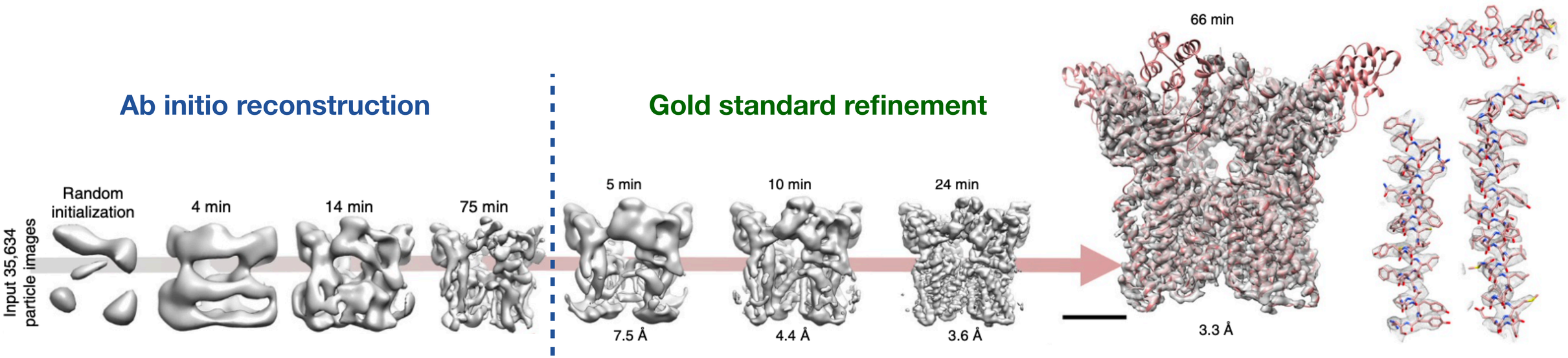
- SGD update rule with **momentum**:

$$V_k^{(n+1)} = V_k^{(n)} + \Delta V_k^{(n)}$$

$$\Delta V_k^{(n)} = (\mu) \Delta V_k^{(n-1)} + (1 - \mu)(\eta_k) G_k^{(n)}$$

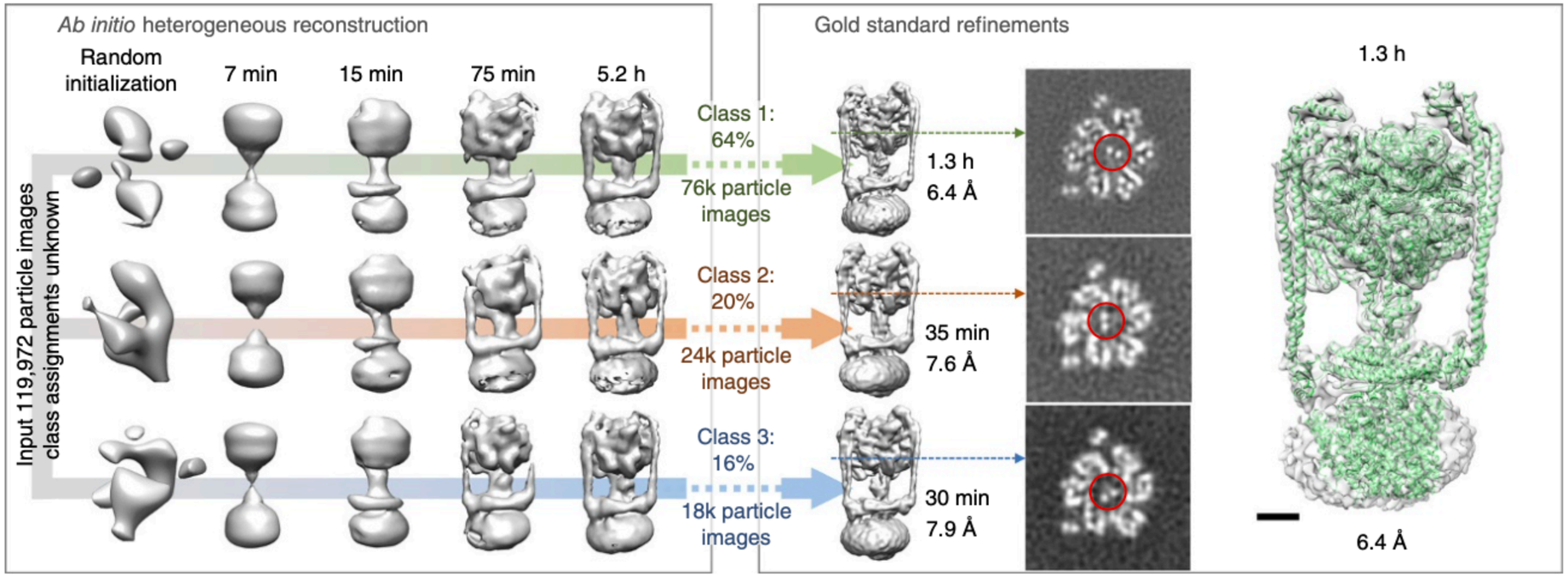
Momentum — Step size

SGD for 3D Reconstruction



- SGD is able to compute ab initio structures to medium resolution.
- Once SGD has converged, the **Expectation-Maximization algorithm** (a.k.a. iterative refinement) is used to refine the resulting structures to high resolution.
- The computationally expensive part of iterative refinement is the expectation step, in which each 2D image is aligned over 3D orientations and 2D translations to the current estimate of each 3D structure ➟ Solved efficiently using a **branch-and-bound search** technique in cryoSPARC.

SGD for 3D Classification



- When applied to a data set of conformationally heterogeneous Thermus thermophilus V/A-ATPase particle images, the SGD algorithm discerned three different conformational states for the enzyme from random initializations.
- This finding is particularly notable, as previous analysis with reference-based classification and the same data set of images only detected two of the three states.

Image Alignment Problem Setup

- The probability of observing an image from a particular pose is given in the Fourier domain as following:

$$p(X \mid \phi, V) = p(X \mid r, t, V) = \frac{1}{Z} \exp\left(\sum_l \frac{-1}{2\sigma_l^2} \mid C_l Y_l(r) - S_l(t) X_l \mid^2 \right)$$

- r: 3D orientation
- t: 2D translation
- l: Two-component index of Fourier coefficient (wavevector)
- $S_l(t)$: 2D translation of t pixels
- $Y_l(r)$: Projection of V according to pose r ($Y_l(r) = \Theta_l(r) V$)
- C_l: CTF
- σ_l: Gaussian noise (allowing white/colored noise)

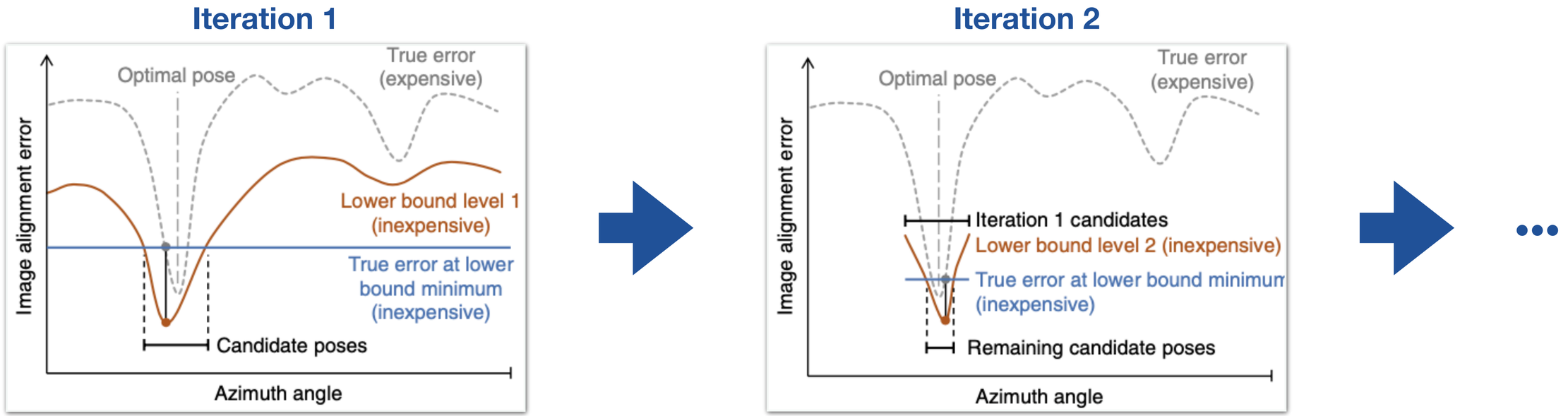
- Taking the **negative log of probability** gives the **image alignment error** (negative log likelihood):

$$E(r, t) = \sum_l \frac{1}{2} \mid C_l Y_l(r) - S_l(t) X_l \mid^2$$

Assume a white noise model $\sigma_l = \sigma = 1$ just for notational clarity.

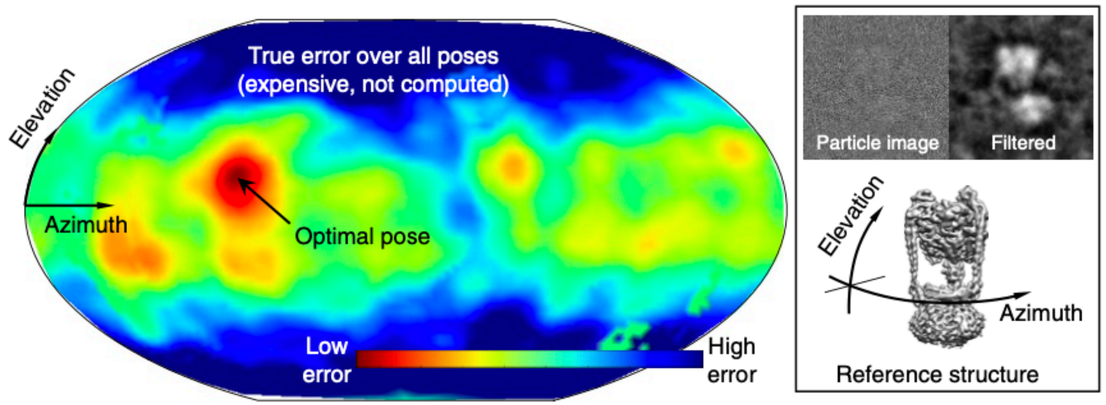
- The aim of image alignment is to **find r and t that minimize this function for the given image X and model V.**

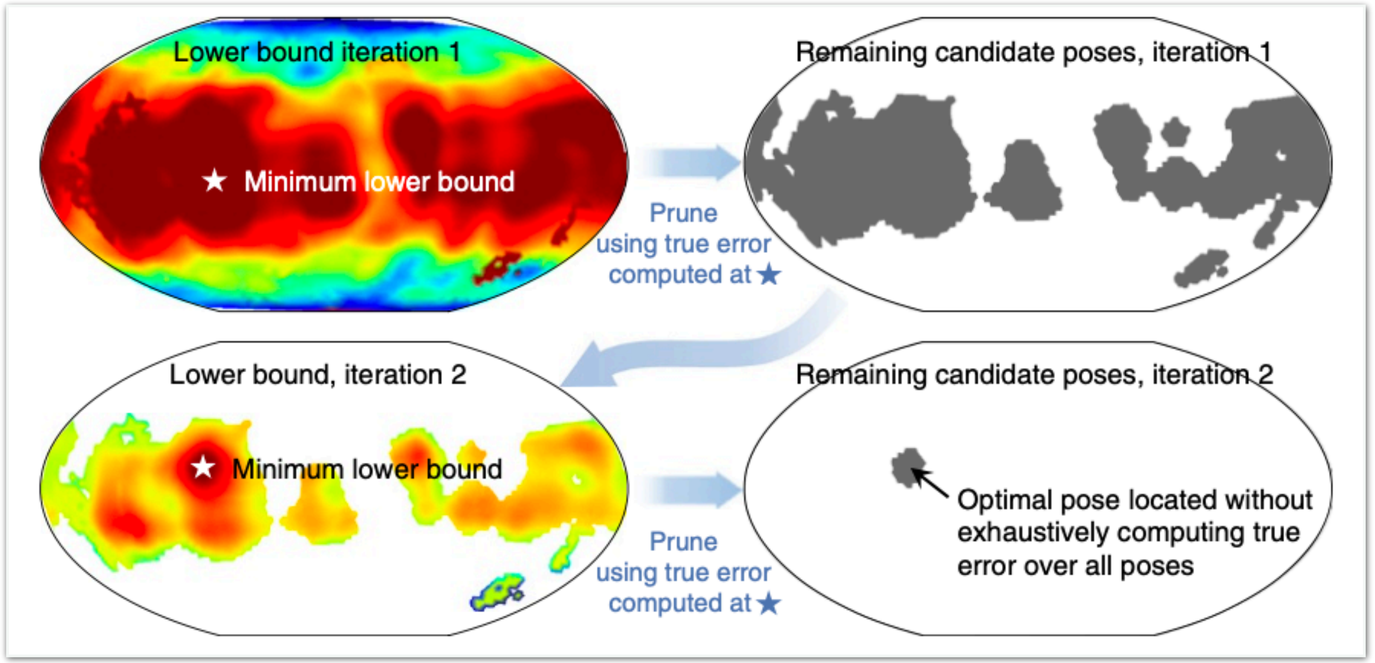
Branch and Bound: Rapid Refinement



- 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, 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.

Branch-and-Bound Approach to 3D Refinement





- The branch-and-bound approach is a global pose search that requires no prior estimate of an optimal pose.
- For cryo-EM images, the true error function over all poses for an individual particle is never evaluated.

Intuition behind a Lower Bound

- The core challenge in employing a branch-and-bound method for cryo-EM is to derive a lower bound that is inexpensive to evaluate but informative about the image alignment error function $E(r, t)$.

- Intuition: *If an image aligns poorly to a structure at low resolution, it will not align well at high resolution*.

- If the low-frequency coefficients already have a given error at a particular pose, the high-frequency coefficients cannot make this error much better or worse.

Derivation of a Lower Bound — I

Image alignment error

$$E(r,t) = \sum_l \frac{1}{2} |C_l Y_l(r) - S_l(t) X_l|^2$$

Split E into two parts: A and B by radius L

L: a certain radius in Fourier space

$$E(r,t) = \underbrace{\sum_{||l||\le L} \frac{1}{2} |C_l Y_l(r) - S_l(t) X_l|^2}_{\equiv A(r,t)} + \underbrace{\sum_{||l||>L} \frac{1}{2} |C_l Y_l(r) - S_l(t) X_l|^2}_{\equiv B(r,t)}$$

Compute A directly (inexpensive when L is small)

$|S_l(t)| = 1$ and the CTF is real-valued

$$B(r,t) = \underbrace{\sum_{||l||>L} \frac{1}{2} |X_l|^2}_{\equiv B_1} + \underbrace{\sum_{||l||>L} \frac{1}{2} C_l^2 |Y_l(r)|^2}_{\equiv B_2} - \underbrace{\sum_{||l||>L} C_l \Re(Y_l(r) * S_l(t) X_l)}_{\equiv B_3}$$

$\Re(z)$: real part of complex value z

$*$: complex conjugation

$\equiv B_1$: The total power of the image at high frequencies.

$\equiv B_2$: The total power of a slice of the model from pose r at high frequencies.

$\equiv B_3$: The correlation between the shifted image X and the slice of the 3D model in the Fourier domain.

Derivation of a Lower Bound — II

$$B(r,t) = \underbrace{\sum_{||l||>L} \frac{1}{2}|X_l|^2}_{\equiv B_1} + \underbrace{\sum_{||l||>L} \frac{1}{2}C_l^2|Y_l(r)|^2}_{\equiv B_2} - \underbrace{\sum_{||l||>L} C_l\Re(Y_l(r) * S_l(t)X_l)}_{\equiv B_3}$$

B_3 *An upper bound on B_3 contributes to a lower bound on $B(r,t)$.*

The cryo-EM image formation model: $X_l = C_l\tilde{X}_l + \epsilon_l$, $\epsilon_l \sim \mathcal{CN}\left(0, \frac{1}{2}\right)$

(True signal: $\tilde{X}_l$; Gaussian noise: ϵ_l)

$$B_3 = \sum_{||l||>L} C_l^2\Re(Y_l(r) * S_l(t)\tilde{X}_l) + \underbrace{\sum_{||l||>L} C_l\Re(Y_l(r) * S_l(t)\epsilon_l)}_{\equiv H}$$

$$\leq \sum_{||l||>L} C_l^2|Y_l(r)||\tilde{X}_l| + H$$

H is normally distributed with variance $\sigma_H^2 = \sum_{||l||>L} \frac{1}{2}C_l^2|Y_l(r)|^2$

H is a random variable which captures the expectation of the effect of noise on E:

$$\begin{aligned}
H &= \sum_{||l||>L} C_l\Re(Y_l(r) * S_l(t)\epsilon_l) \\
&= \sum_{||l||>L} C_l\Re(Y_l(r) * \epsilon_l) \\
&= \sum_{||l||>L} C_l\Re\left(Y_l(r) * \mathcal{CN}\left(0, \frac{1}{2}\right)\right) \\
&= \sum_{||l||>L} C_l\Re\left(\mathcal{CN}\left(0, \frac{1}{2}|Y_l(r)|^2\right)\right) \\
&= \sum_{||l||>L} \mathcal{N}\left(0, \frac{1}{2}C_l^2|Y_l(r)|^2\right) \\
&= \mathcal{N}\left(0, \sum_{||l||>L} \frac{1}{2}C_l^2|Y_l(r)|^2\right)
\end{aligned}$$

Derivation of a Lower Bound — III

$$B(r,t) = \underbrace{\sum_{||l||>L} \frac{1}{2}|X_l|^2}_{\equiv B_1} + \underbrace{\sum_{||l||>L} \frac{1}{2}C_l^2|Y_l(r)|^2}_{\equiv B_2} - \underbrace{\sum_{||l||>L} C_l \Re(Y_l(r) * S_l(t)X_l)}_{\equiv B_3}$$

$B_2 - B_3$

$$B_2 - B_3 \geq \sum_{||l||>L} \frac{1}{2}C_l^2|Y_l(r)|^2 - \sum_{||l||>L} C_l^2|Y_l(r)||\tilde{X}_l| - H$$

$$= \underbrace{\sum_{||l||>L} \frac{1}{2}\left(C_l^2|Y_l(r)|^2 - 2C_l^2|Y_l(r)||\tilde{X}_l|\right)}_{\equiv Q} - H$$

The maximum of lower bound of Q is attained at $r = \hat{r}$, $\hat{Y}_l \equiv Y_l(\hat{r})$

$\hat{Y}$ is **the slice of model V that has the maximum CTF-modulated total power**.

$$Q \geq -\sum_{||l||>L} \frac{1}{2}C_l^2|\hat{Y}_l|^2 \equiv \hat{Q}$$

Once $\hat{Y}$ is identified, the bound $\hat{Q}$ on Q is fixed.

Each term of Q is a positive-definite quadratic function of $Y_l(r)$.
Therefore Q can be bounded from below:

$$Q \geq \min_{Y_l(r)} Q$$

$$= \min_{Y_l(r)} \sum_{||l||>L} \frac{1}{2}\left(C_l^2|Y_l(r)|^2 - 2C_l^2|Y_l(r)||\tilde{X}_l|\right)$$

$$= \sum_{||l||>L} -\frac{1}{2}C_l^2|\tilde{X}_l|^2$$

$$= \sum_{||l||>L} -\frac{1}{2}C_l^2|Y_l(r^*)|^2$$

$$\geq \min_{r} \sum_{||l||>L} -\frac{1}{2}C_l^2|Y_l(r)|^2$$

$$= -\max_{r} \sum_{||l||>L} \frac{1}{2}C_l^2|Y_l(r)|^2$$

$\tilde{X}_l = Y_l(r^*)$ ↓ Unknown true pose

Derivation of a Lower Bound — IV

Lower bound on $B(r,t)$

$$B(r,t) \geq \sum_{||l||>L} \frac{1}{2}|X_l|^2 - \sum_{||l||>L} \frac{1}{2}C_l^2|\hat{Y}_l|^2 - H$$

Lower bound on $E(r,t)$

$$E(r,t) \geq \sum_{||l||\leq L} \frac{1}{2}|C_l Y_l(r) - S_l(t)X_l|^2 + \sum_{||l||>L} \frac{1}{2}|X_l|^2 - \sum_{||l||>L} \frac{1}{2}C_l^2|\hat{Y}_l|^2 - H$$

Due to the presence of H, the above expression is a **probabilistic** bound on $E(r,t)$, giving the probability of $E(r,t)$ being greater than the value of the expression.

In practice, consider a probability of 0.999936 of H ($4\sigma_H$)

$$H \leq 4\sigma_H \leq 4\sqrt{\max_r \sum_{||l||>L} \frac{1}{2}C_l^2|Y_l(r)|^2}$$

Complete lower bound on $E(r,t)$

$$E(r,t) \geq \sum_{||l||\leq L} \frac{1}{2}|C_l Y_l(r) - S_l(t)X_l|^2 + \sum_{||l||>L} \frac{1}{2}|X_l|^2 - \sum_{||l||>L} \frac{1}{2}C_l^2|\hat{Y}_l|^2 - 4\sqrt{\sum_{||l||>L} \frac{1}{2}C_l^2|\hat{Y}_l|^2} \equiv \beta_L(r,t)$$

The bound is inexpensive to compute for a particular r, t (since only $A(r,t)$ depends on these).

Subdivision Scheme

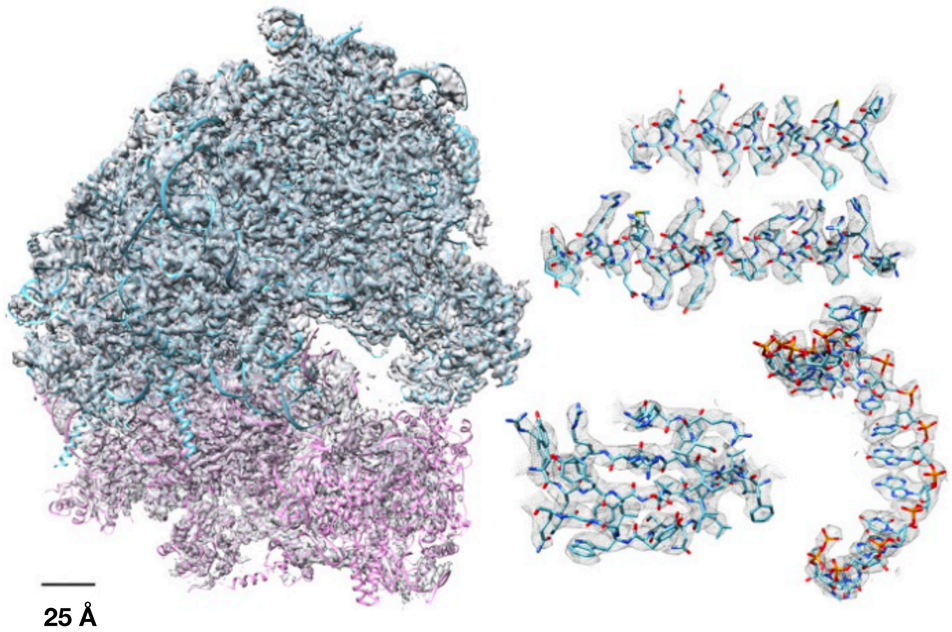
- The branch and bound algorithm relies on a method for representing regions in the space of 3D poses and 2D shifts so that when the bound above is computed, it can be used to discard regions and the remaining candidate poses are recorded in an organized fashion.
- This work uses a cartesian grid in the axis-angle representation of 3D pose, and a second cartesian grid in the 2D space of pixel shifts.
- Each subsequent iteration uses double the radius in Fourier space, up to a maximum radius that is determined from the current resolution of the 3D map.

Approximations

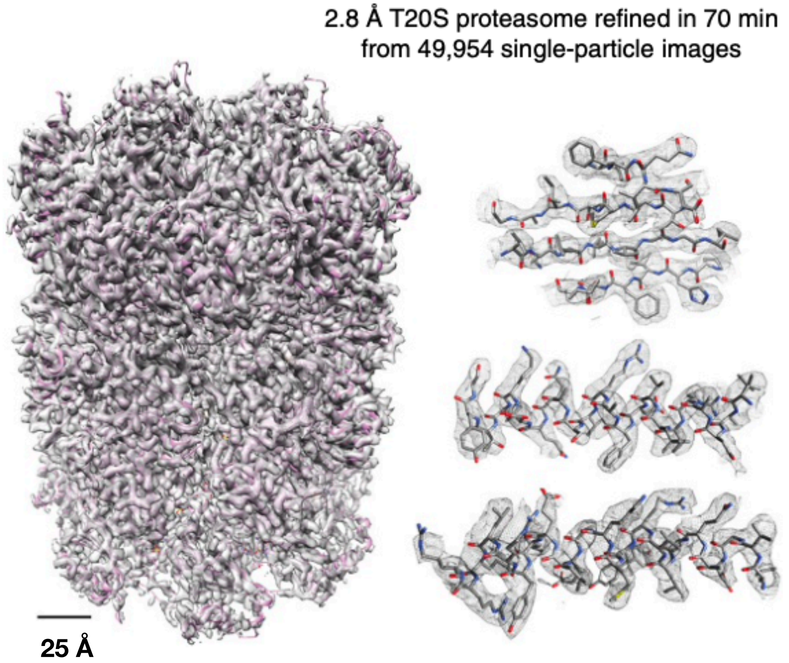
- Consider the case of an image that is an outlier (non-particle or noise), the assumptions of the lower bound are violated as the image does not come from the standard cryo-EM image formation model ➟ The bound will not be able to reject large portions of the pose space.
  - An upper limit is set on the number of candidate poses (on the current discrete grid) that can remain after each iteration of branch and bound search. (In this work, the limit is set to 12.5% of 3D orientations, and 25% of 2D image shifts)
- The bound $\beta_L(r,t)$ depends on the CTF of the image ➟ The components of the bound must be recomputed for each micrograph.
  - In this work approximate the magnitude of the oscillating CTF at high resolutions using the root mean squared value of the CTF, which is a constant $1/\sqrt{2}$. ➟ removes the dependence on the CTF so that the last three terms of the bound only needs to be computed once for all images, given the structure V.
- It is assumed that the lower bound is sufficiently smooth that it does not need to be sampled at full resolution, which would require a prohibitively large number of poses to be evaluated even with small values of L.

High-Resolution Structure from Branch-and-Bound Refinement

3.2 Å 80S ribosome refined in 2.2h
from 105,247 single-particle images



2.8 Å T20S proteasome refined in 70min
from 49,954 single-particle images



- The application of the branch-and-bound method allowed high-resolution refinement of the 80S ribosome to **3.2 Å** resolution in **2.2 h**. (**Take ~20h using RELION**)
- The 20S proteasome structure was refined to **2.8 Å** with D7 symmetry enforced, matching the published resolution but only in **70 min**.