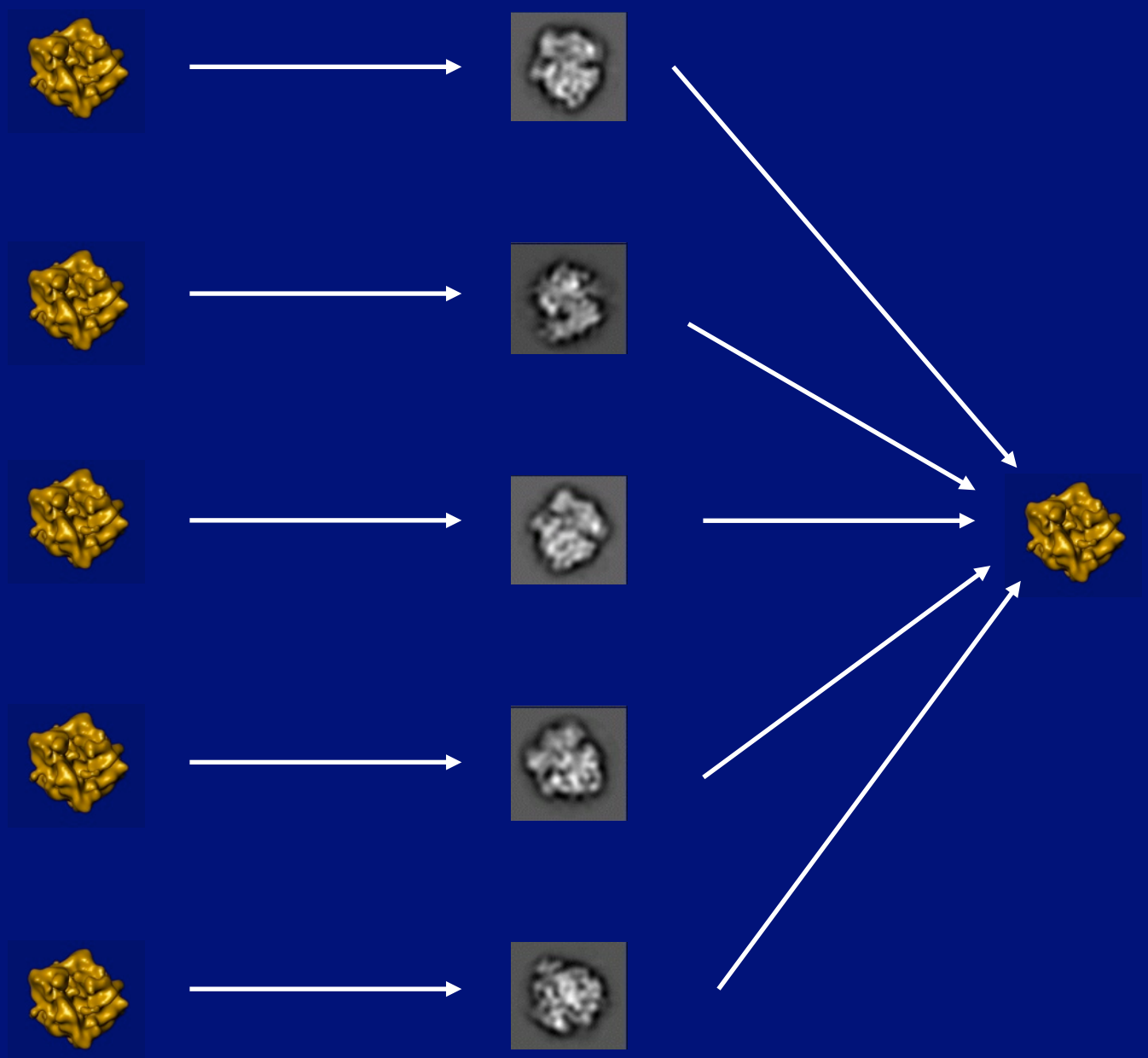
Identification of conformational states by codimensional PCA

Pawel A. Penczek

**The University of Texas – Houston Medical School,
Department of Biochemistry.**

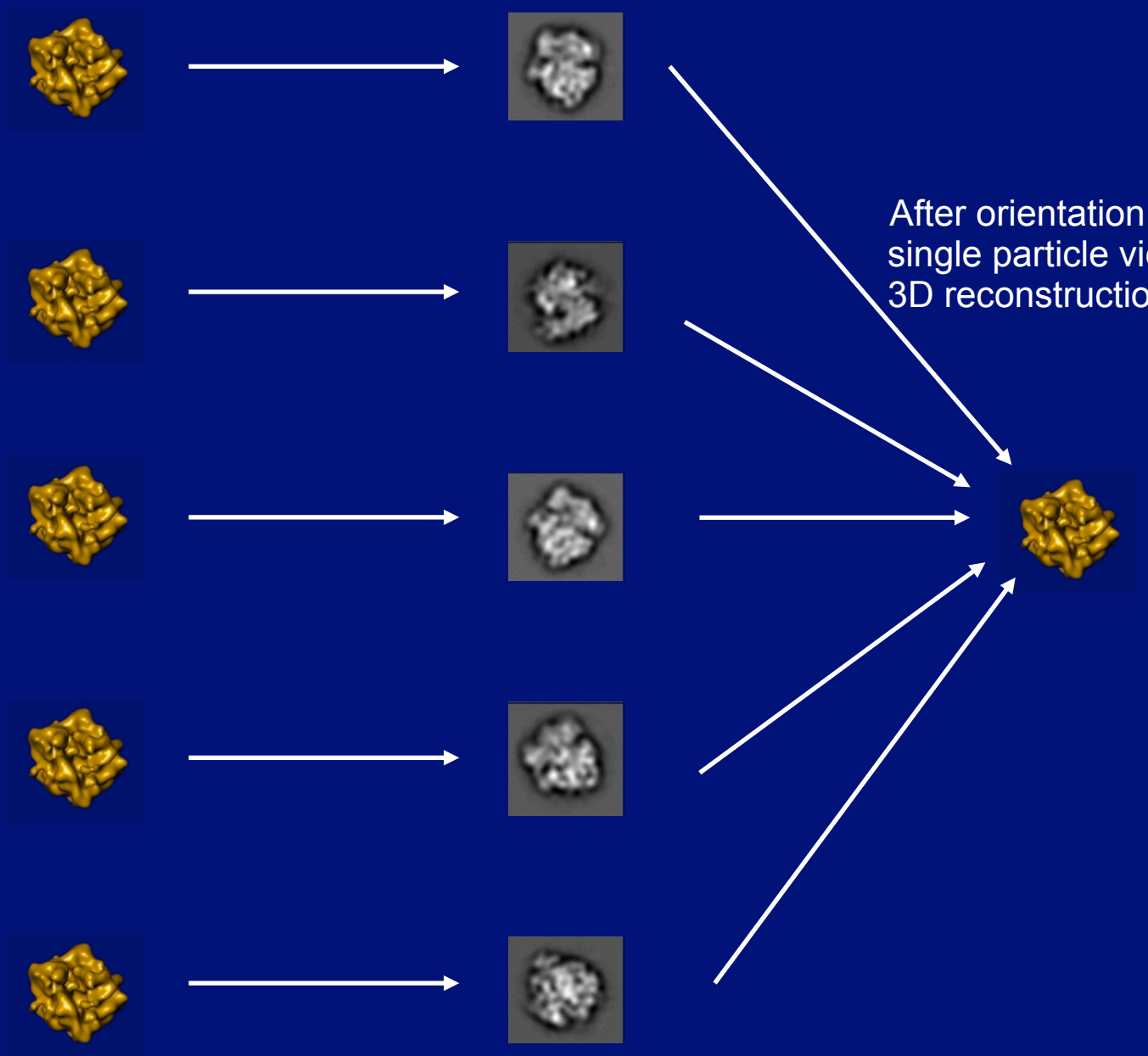
In single particle cryo-EM, projections originate from different macromolecules that in principle have the same structure.



Different copies of the same macromolecule (3D).

Electron microscope forms 2D projections macromolecules.

After orientation parameters of single particle views are found, 3D reconstruction is calculated.

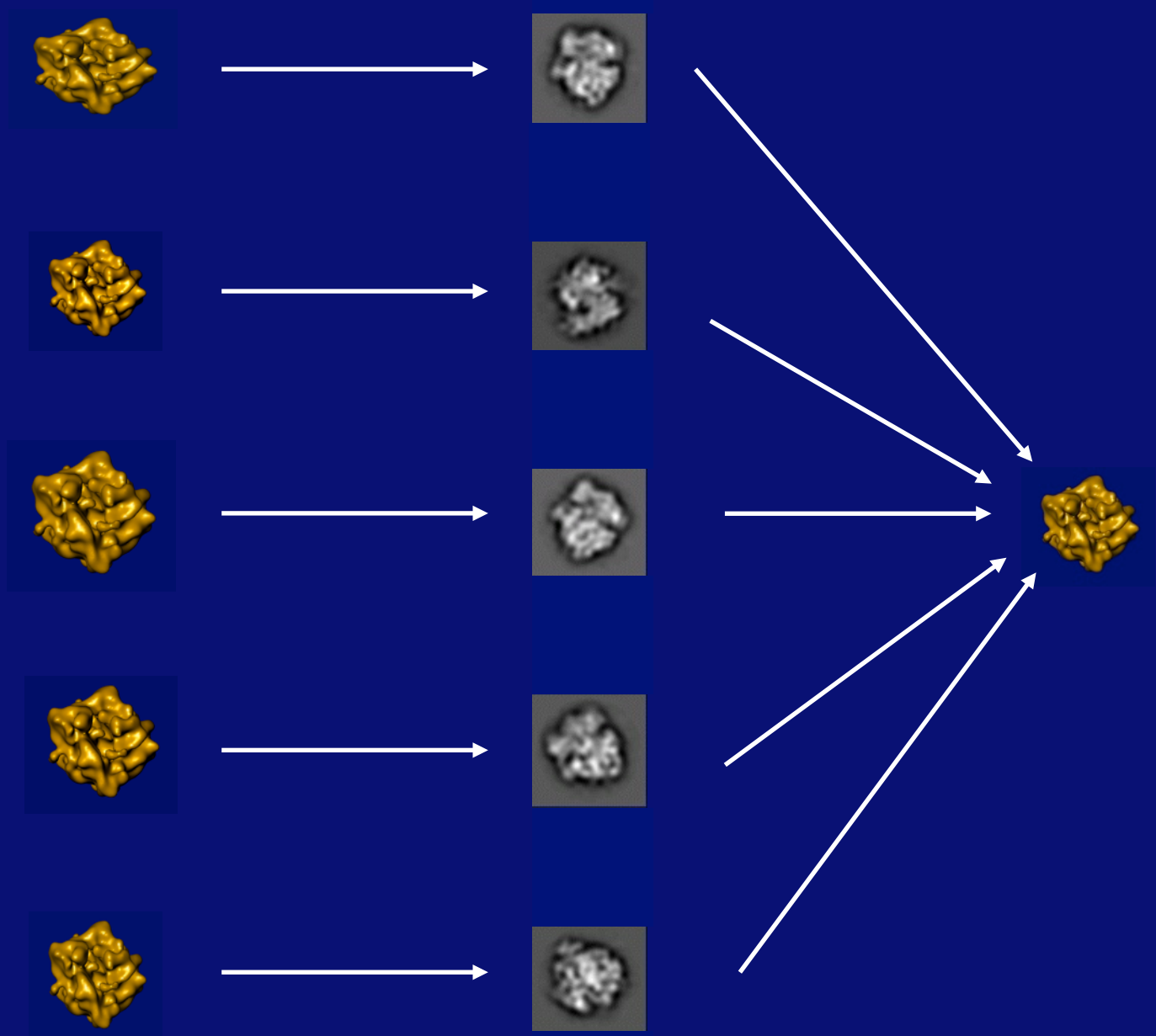


There is mounting evidence that macromolecules occur naturally in a mixture of conformational states:

- ribosome
- RNA polymerase
- human transcription factor
- pyruvate dehydrogenase complex (breathing core)

In addition to the expected conformational heterogeneity of the assemblies that is due to fluctuations of the structure around the ground state, one can expect to capture molecules in different functional states, especially if the binding of a ligand induces a conformational change in the macromolecular assembly. Therefore, *data set of images from an EM experiment must be interpreted as a mixture of projections from similar but not identical structures.*

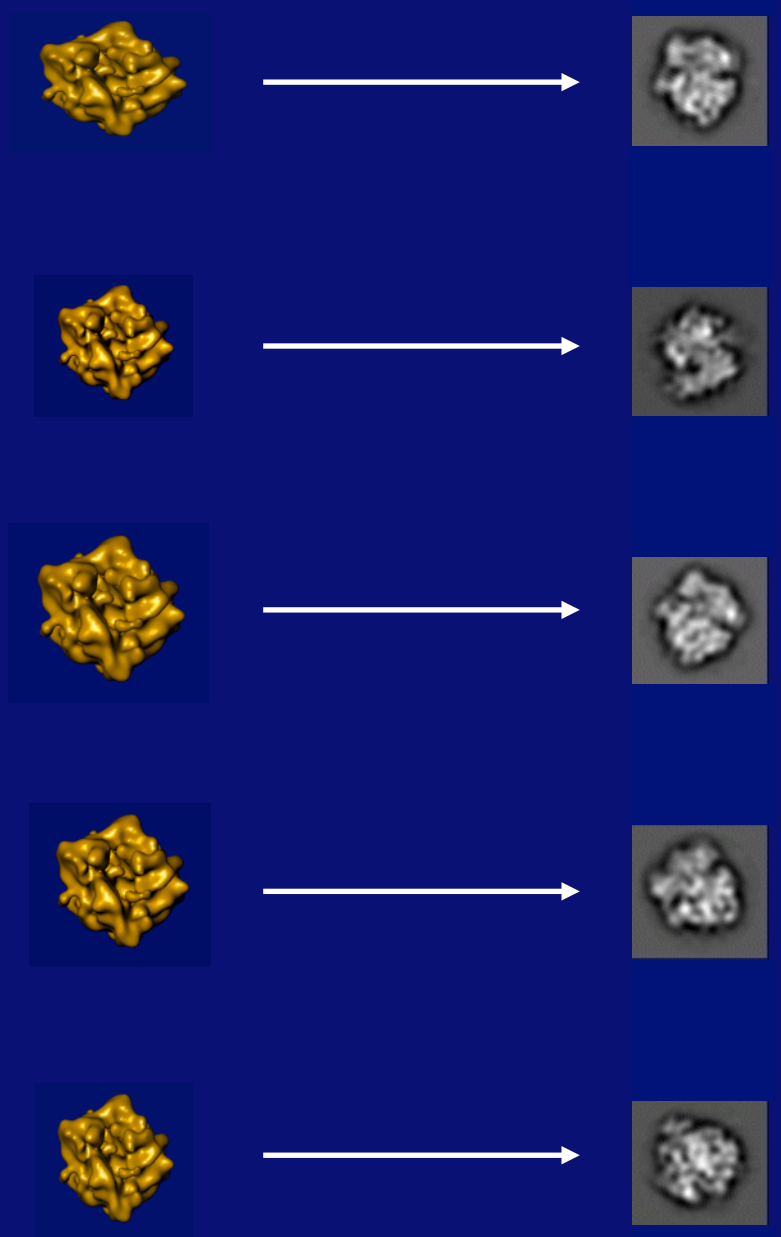
In single particle analysis (cryo-EM), projections may originate from different 3D structures.



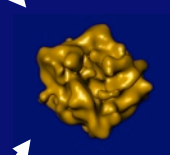
In single particle analysis (cryo-EM), projections may originate from different 3D structures.

Different states of the same macromolecule (3D).

Electron microscope forms 2D projections of macromolecules.



After orientation parameters of single particle views are found, 3D reconstruction is calculated.

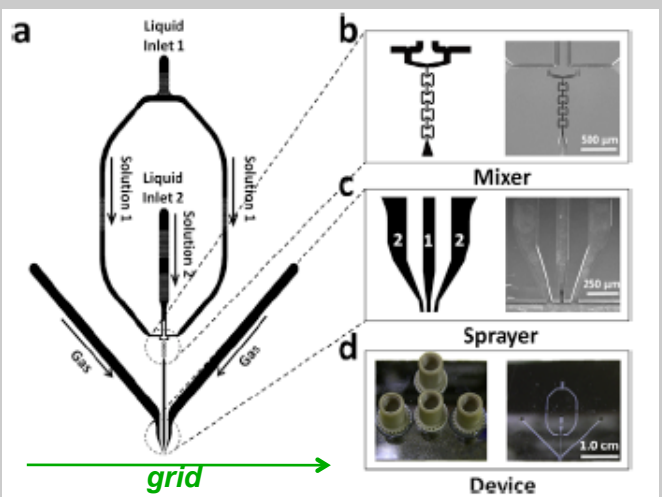


Experimental time-resolved cryo-EM

Structures of various conformers are determined using cryo-EM data that are taken at successive times from a system that is known to be developing in time.

Heymann, J. B., Conway, J. F., Steven, A. C., 2004. Molecular dynamics of protein complexes from four-dimensional cryo-electron microscopy. JSB **147**, 291-301.

The components are mixed and allowed to react, then are sprayed onto an EM grid as it is being plunged into cryogen.

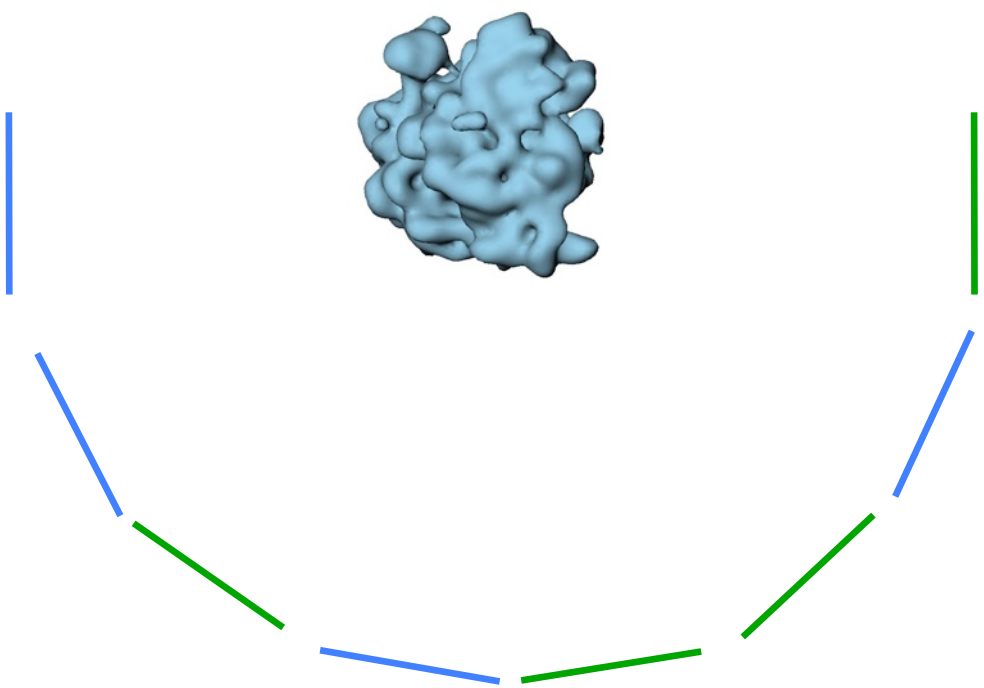


Lu, Z., Shaikh, T. R., Barnard, D., Meng, X., Mohamed, H., Yassin, A., Mannella, C. A., Agrawal, R. K., Lu, T. M., and Wagenknecht, T., 2009. Monolithic microfluidic mixing-spraying devices for time-resolved cryo-electron microscopy, J Struct Biol **168**, 388-395.

Computational time-resolved cryo-EM

Class A

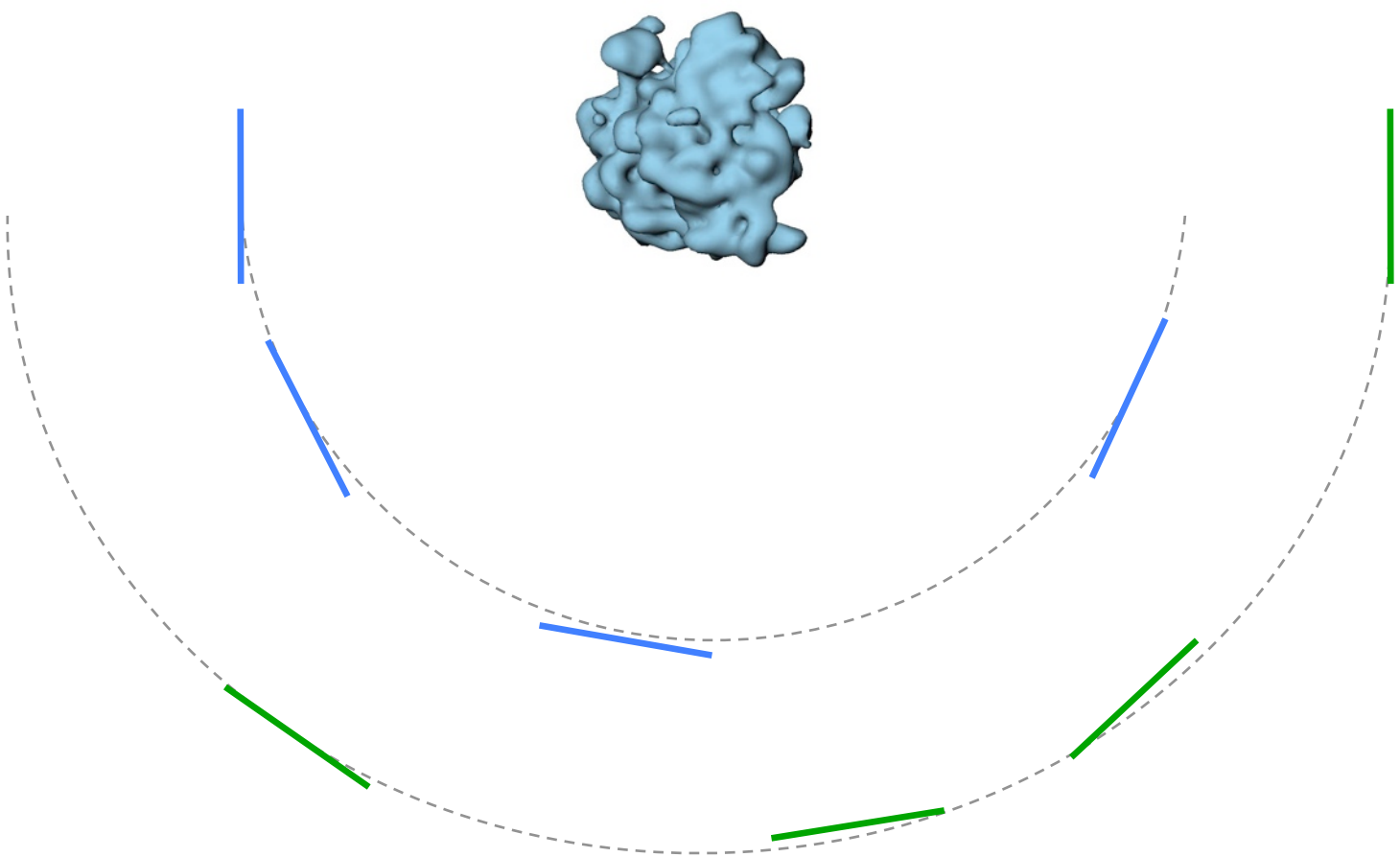
Class B



Computational time-resolved cryo-EM

Class A

Class B



Computational time-resolved cryo-EM

- From 2D to 3D: using tilted data (RCT)
  Clustering in 2D can yield structures of 3D conformers through experimentally established angular relations.
  Requires large number of dominating views.
- From 2D to 3D: common lines approaches, angular continuity, sorting of 2D projections, focused classification (last three mainly for two classes).
  Common lines approaches can establish angular relations and conformers simultaneously
  Effectiveness easy to demonstrate using simulated data, but very weak performance on EM data.
  Based on faulty theoretical premises (pair-wise discrepancies instead of simultaneous agreement).
- Sorting in 3D: multi-reference (or competitive) refinement (including ML)
  Based on K-means clustering principle with discrepancies computed between 2D projection data and reprojections of template structures.
  Requires good guess of initial structures and their number, computationally extremely demanding (particularly ML).
- codimensional PCA
  based on PCA of the covariance matrix of resampled volumes.
  Computationally very efficient, allows easy exploration of a range of possible solution.
  Applicable only if conformers are reasonably similar, limited resolution thus major effects only.

Spahn, C. M., and Penczek, P. A. (2009) Exploring conformational modes of macromolecular assemblies by multiparticle cryo-EM, Curr Opin Struct Biol **19**, 623-631.

Real-space variance in single particle analysis

Images from an EM experiment must be interpreted as a mixture of projections from similar but not identical structures

- Detection of different functional states (caused by binding of a ligand)
- Significance of small details in 3-D reconstructions
- Conformational heterogeneity of the assemblies due to fluctuations of the structure around the ground state
- Significance of details in difference maps
- Fitting (docking) of known structural domains into EM density maps

Calculation of a real space variance in 3D reconstruction from projections is a difficult problem.

- The data is available in form of projections, i.e., information is partial.
- In single particle analysis (cryo-EM), the projections originate from different 3D structures.
- The main difficulty is that *there is only one data set*. In addition, even if we know that some macromolecules on the grid are identical, *we do not know which particle view corresponds to which macromolecule*.
- Exact inversion of the projection process is impossible. Thus, the step of 3D reconstruction itself is a source of noise.

3D reconstruction – weighted sum of the input projections with the weights dependent on the number and distribution of projections.

Backprojection

(in real space)

Voxel = algebraic (weighted) sum of projection pixels

Weighting

(in Fourier space)

Compensation for uneven distribution of projections in Fourier space

Resampling without replacements

Original data set of nine 2-D projections (n=9)

1 2 3 4 5 6 7 8 9

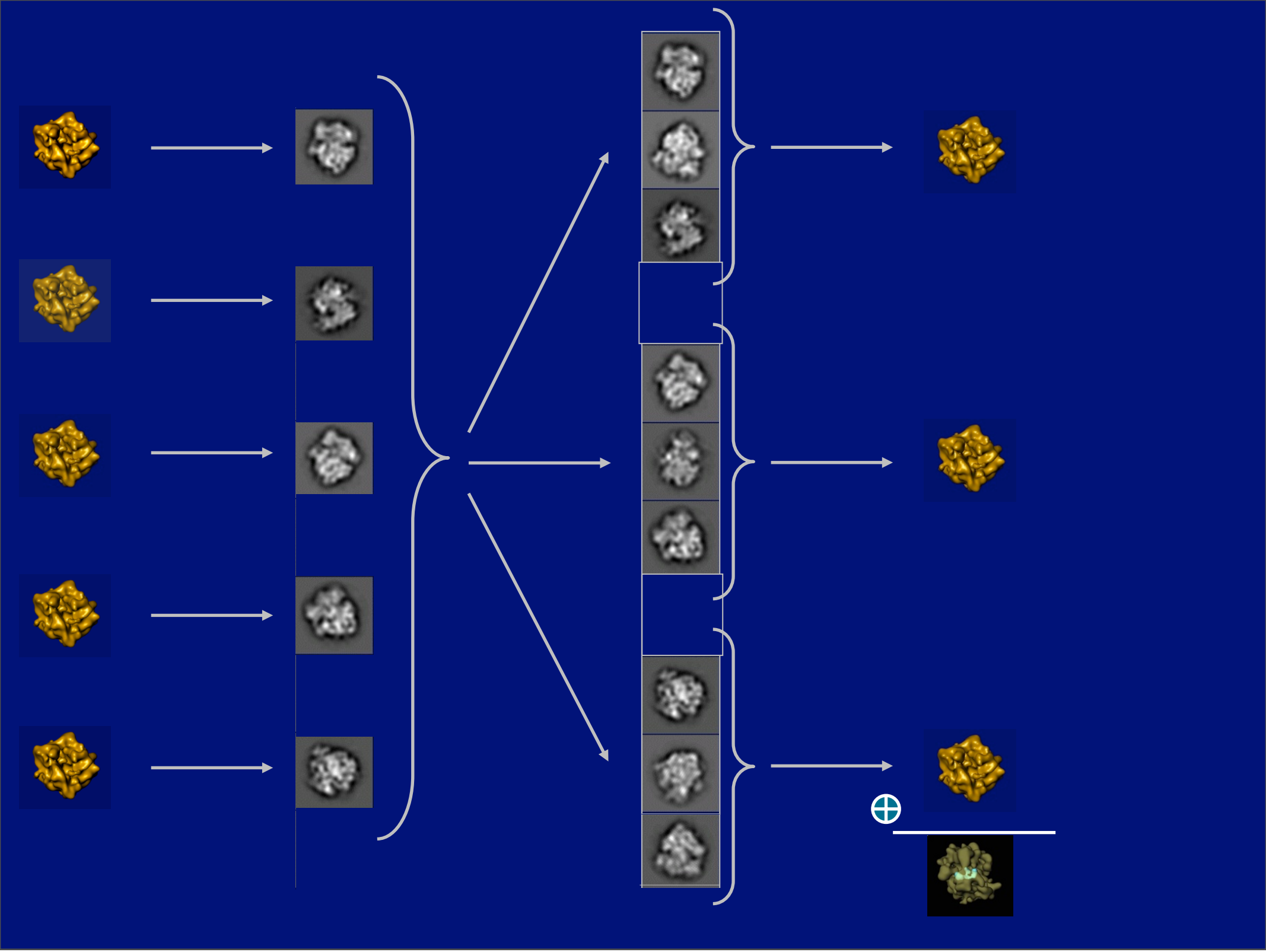
Resampled data sets of 2-D projections, each contains six projections.

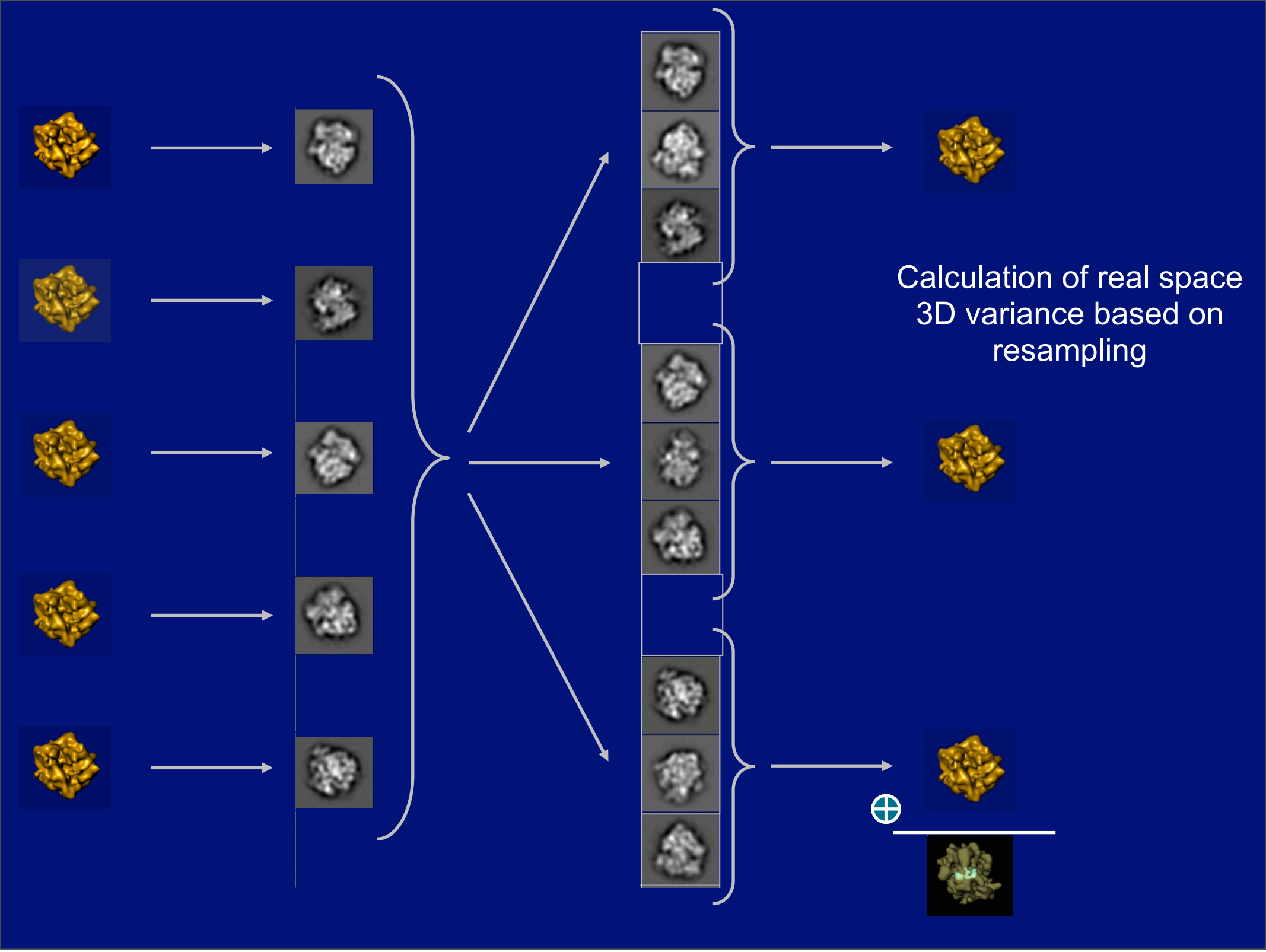
1 3 4 5 8 9

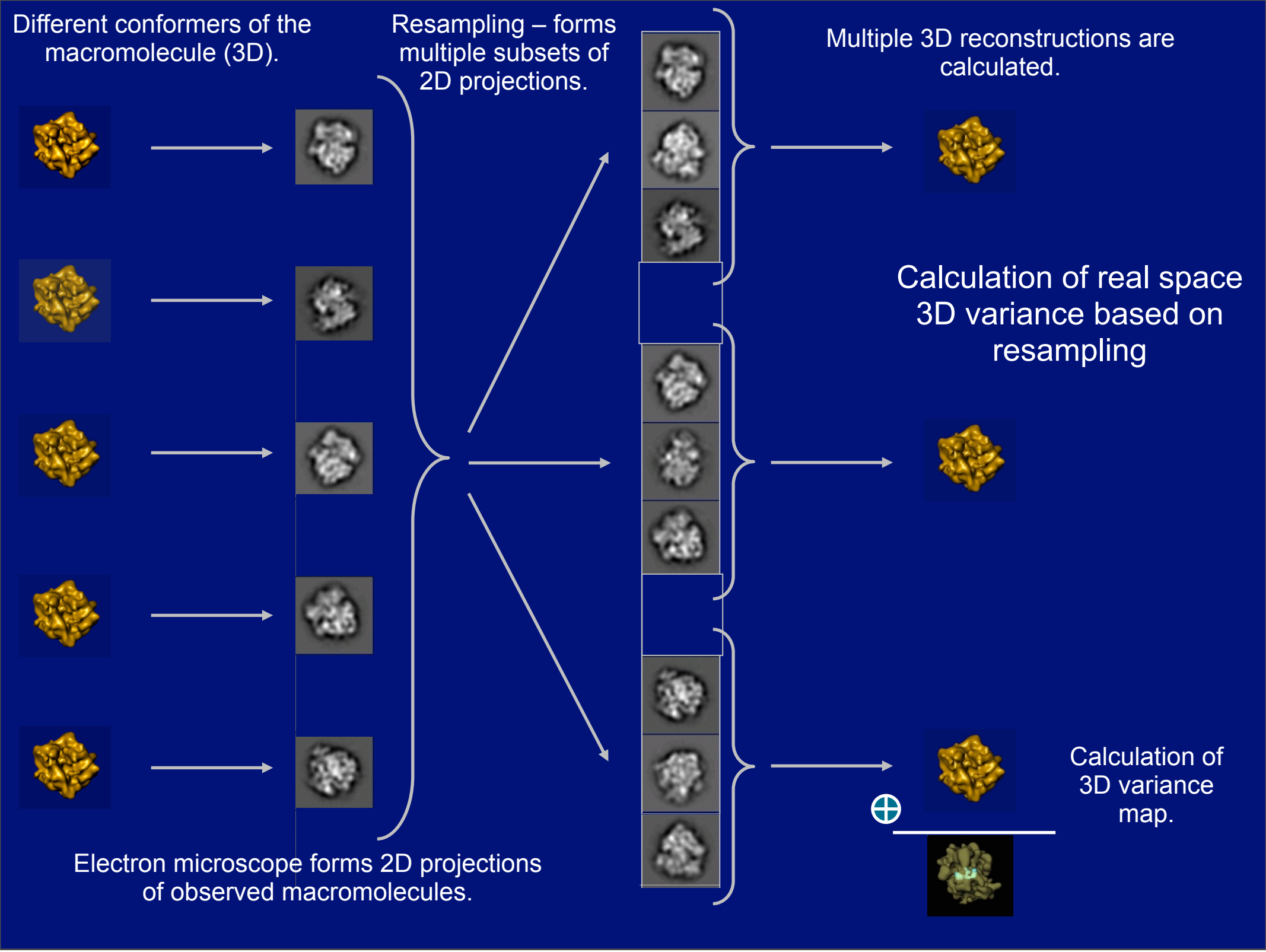
2 4 5 7 8 9

3D reconstruction
Large number of “different” volumes

Variance/covariance!



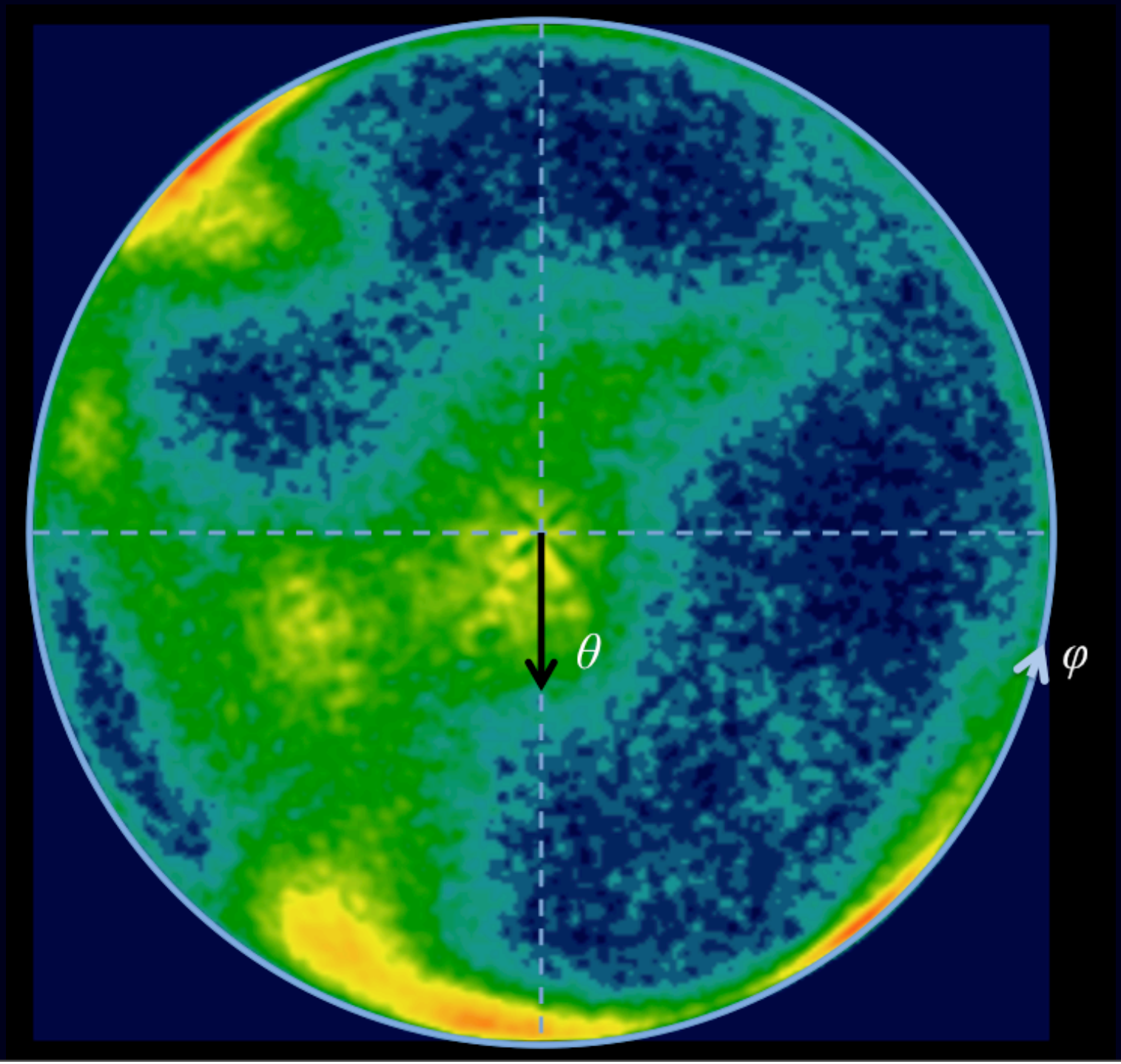

Calculation of real space
3D variance based on
resampling


Different conformers of the macromolecule (3D).
Resampling – forms multiple subsets of 2D projections.
Multiple 3D reconstructions are calculated.
Calculation of real space 3D variance based on resampling
Calculation of 3D variance map.
Electron microscope forms 2D projections of observed macromolecules.

Sources of variance in 3D reconstruction from projections

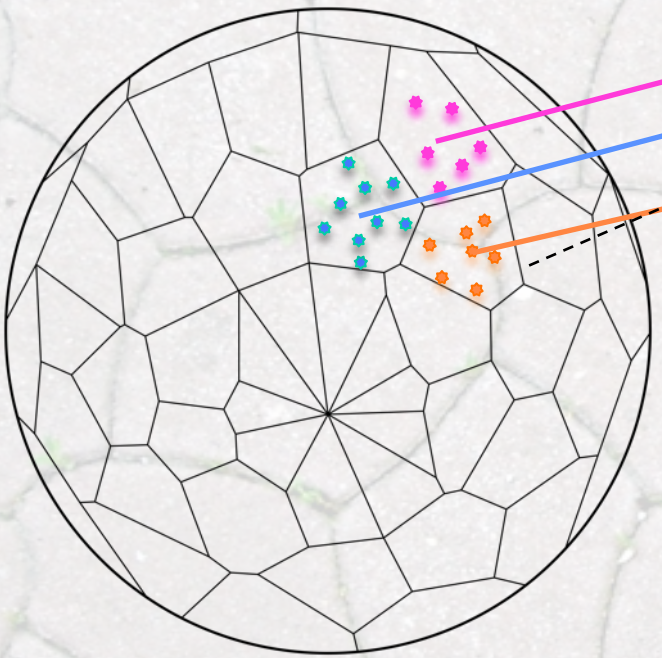
- Variability of the structure
- Noise in projection data
- Uneven distribution of projections
- Normalization errors in projections
- Numerical accuracy of the reconstruction algorithm

Uneven angular distribution of projection images



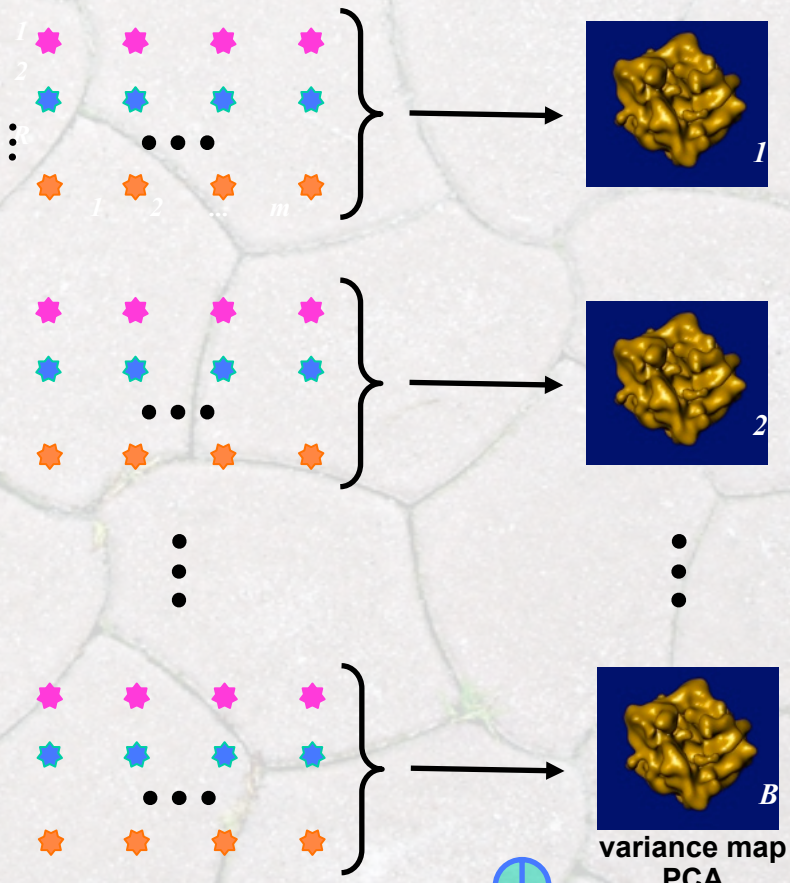
HYPERgeometric **STRATI**fied Re**S**ampling (HYPERSTRATIS) designed to compensate for uneven distribution of projection directions

Cells on the sphere are sampling strata and from each the same number of projections is selected without replacements.



Tessellation of a sphere yields areas containing projection directions.

3D reconstruction using resampled set of 2D projections



The expected value of the HYPERSTRATIS variance S^2 is related to the distribution variance σ^2 by:

$$E\left[S^2\right] = \sigma^2 \frac{1 - \frac{m}{R}\sum_{r=1}^{R} n_r^{-1}}{mR}$$

where:

- n - total number of projection images $\sum_{r=1}^{R} n_r = n$
- R - number approximately equal sized "angular areas" (stratums)
- n_r – number of projection images within r'th area
- m - the number of images to be retained in each area such that

$$0 < m < \min_r n_r$$

Assuming all angular areas are equally populated:

$$\sigma^2 = \frac{q}{1-q} nE\left[S^2\right]$$

$$0 < q < 1$$

Small fraction selected from each area (q~0) – quickly converges, large uncertainties.

Large fraction selected from each area (q~1) – slowly converges, higher reliability.

Calculation of the variance of structures

$$\sigma^2_{Struct} = Q\left(E\left[S^2\right] - \overline{\sigma}^2_{Back}\right)$$

Where *Q* is a scaling factor for HYPERSTRATIS/

We disregard the variance arising from alignment errors, as there is no method to estimate it independently.

Penczek, P.A., Chao, Y., Frank, J., Spahn, Ch.M.T.: Estimation of variance in single particle reconstruction using the bootstrap technique. *J. Struct. Biol.*, 154:168-183, 2006.

Penczek, P.A., Frank, J., Spahn, Ch.M.T.: A method of focused classification, based on the bootstrap 3-D variance analysis, and its application to EF-G-dependent translocation. *J. Struct. Biol.*, 154: 184-194, 2006.

GTPase activation of elongation factor EF-Tu by the ribosome during decoding

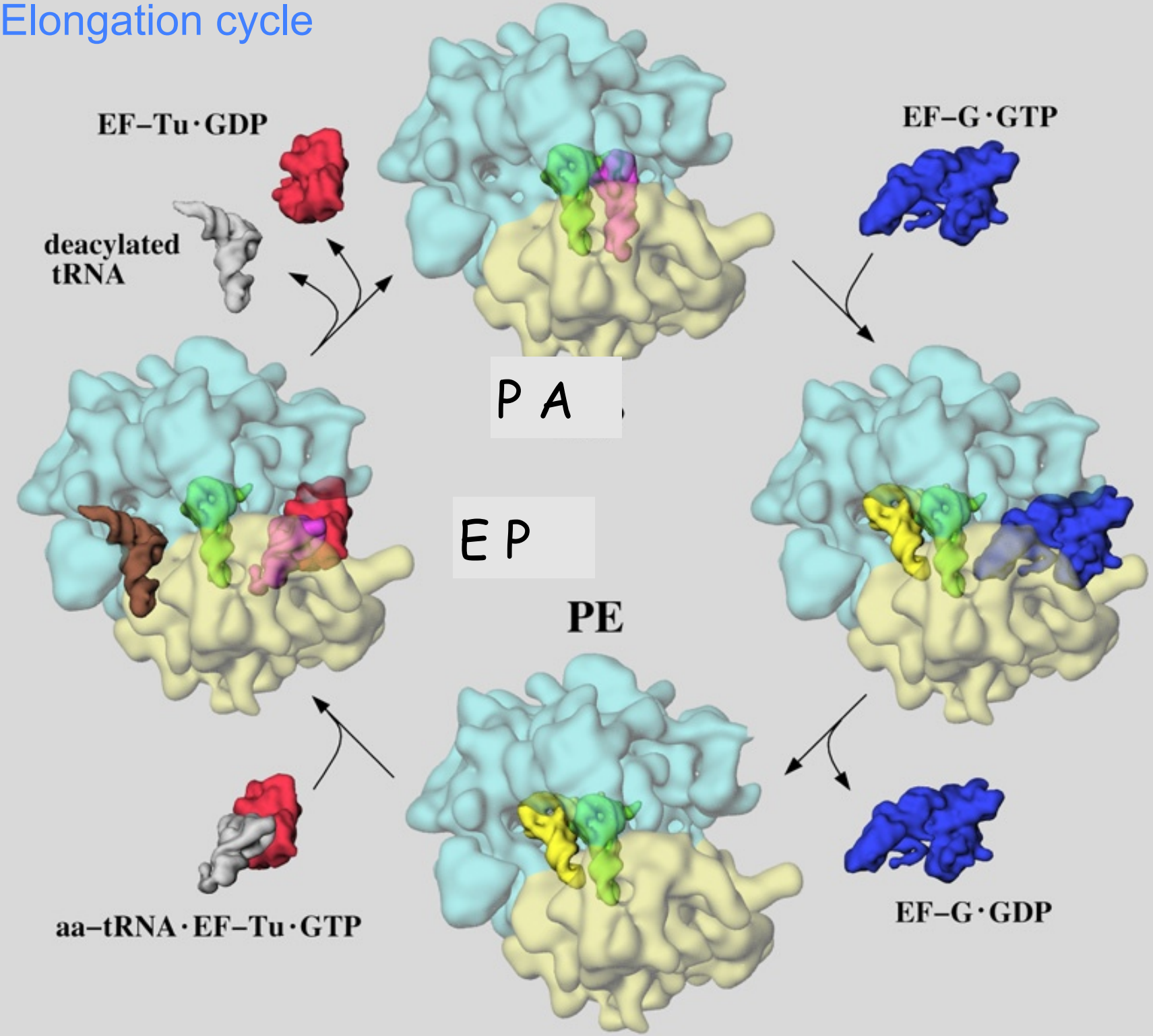
323,688 cryo-EM projection images of *Thermus thermophilus* 70S ribosome in which the ternary complex of elongation factor Tu (EF-Tu), tRNA and guanine nucleotide has been trapped on the ribosome using the antibiotic kirromycin.

Resolution: 6.5 Å.

Total set reprocessed here contained 586,329 images.

Schuette, J.C., Murphy, F.Vt., Kelley, A.C., Weir, J.R., Giesebrecht, J., Connell, S.R., Loerke, J., Mielke, T., Zhang, W., Penczek, P.A., Ramakrishnan, V., Spahn, Ch.M.T.: GTPase activation of elongation factor EF-Tu by the ribosome during decoding. EMBO J 2009, **28**:755-765.

Elongation cycle



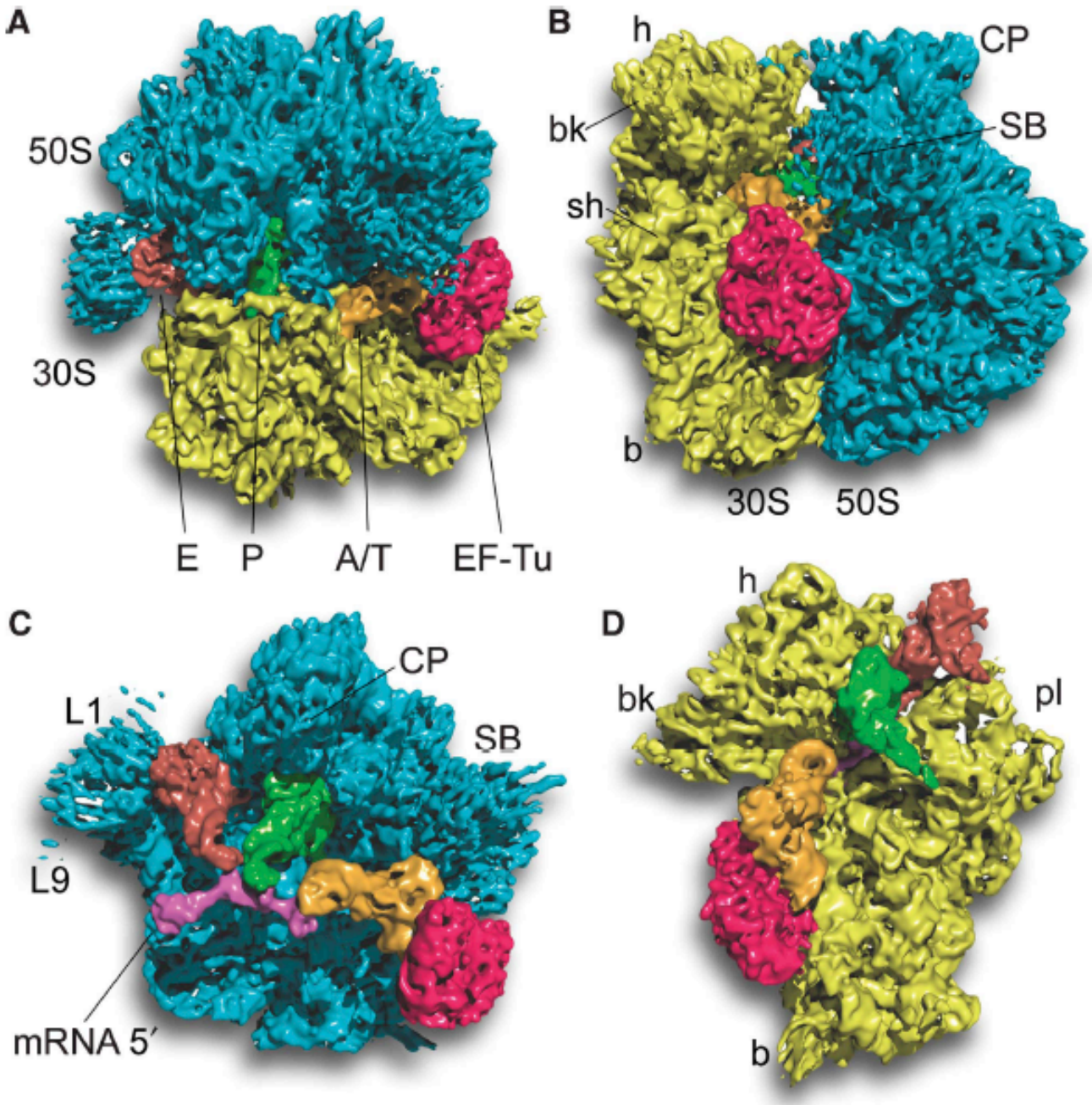
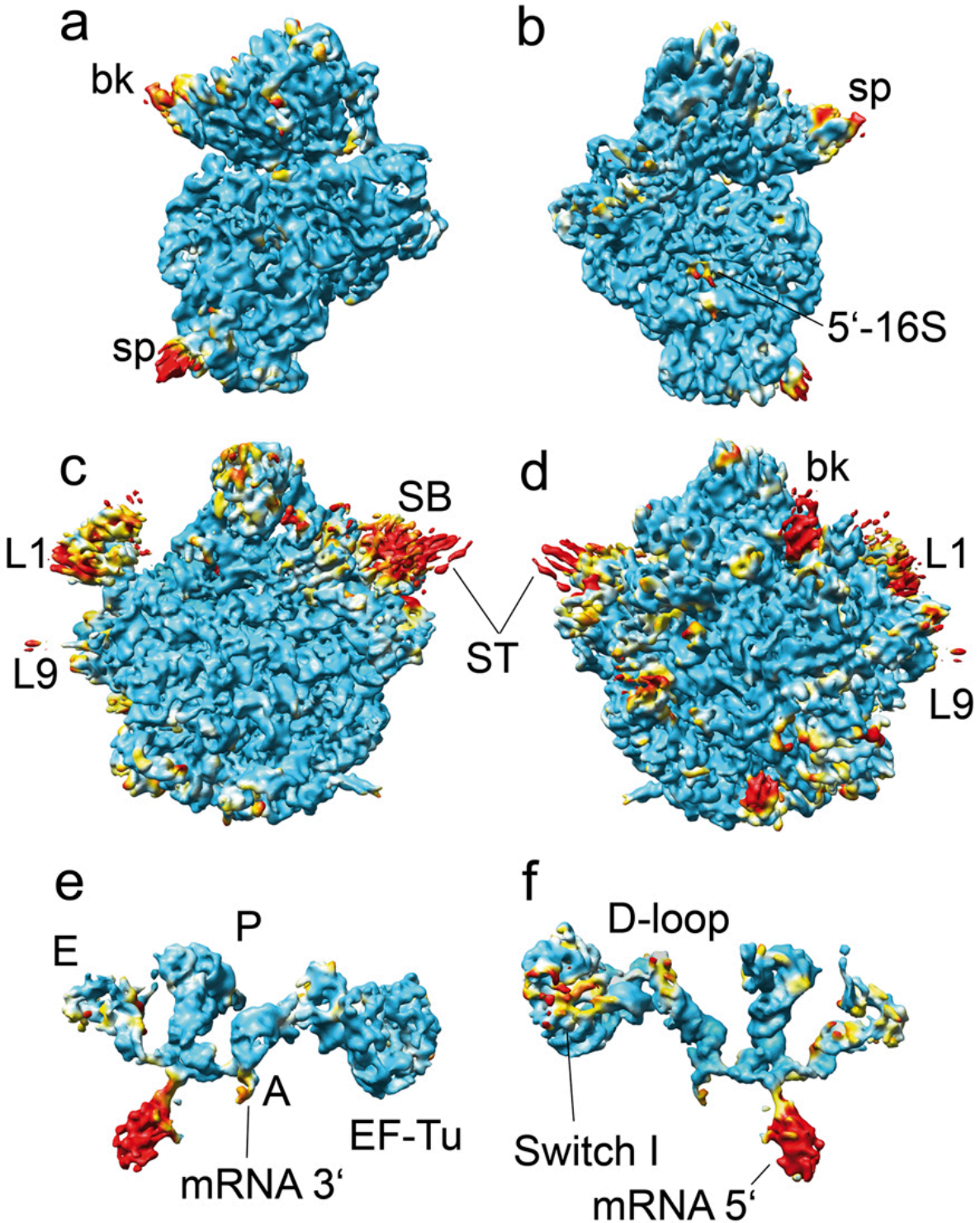
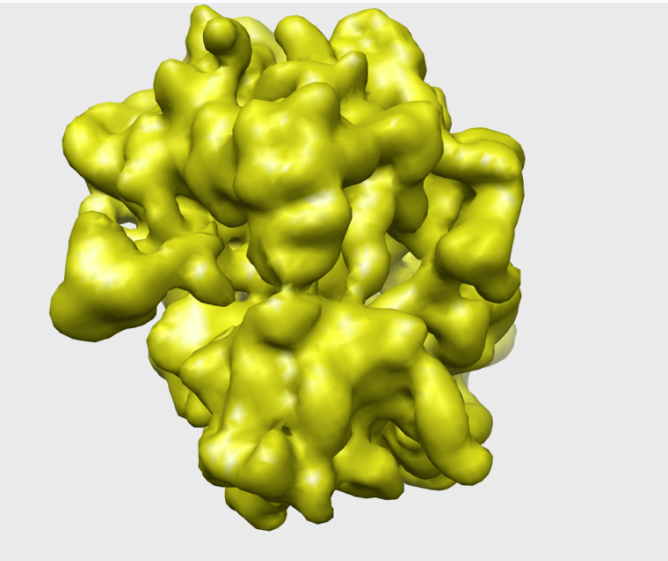


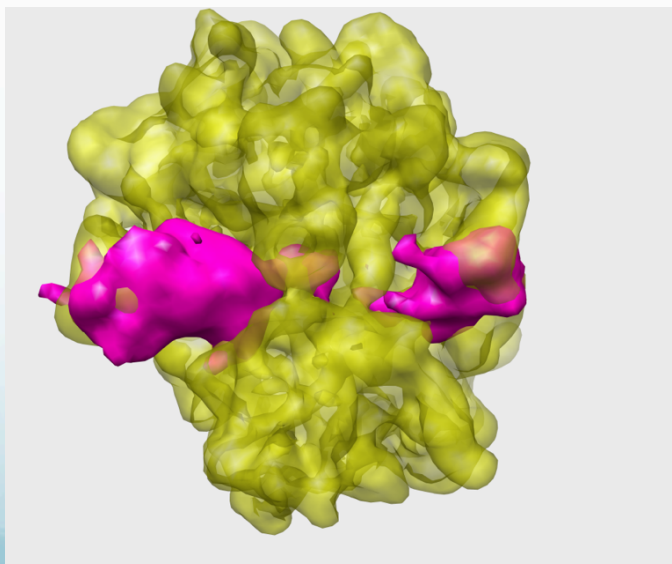
Figure 1 Overview of the 70S●EF-Tu●Phe-tRNA●GDP●kirromycin complex. A surface representation of the cryo-EM map is shown (**A**) from the top; (**B**) from the L7/L12 side; (**C**) from the 30S side, with 30S removed and (**D**) from the 50S side, with 50S removed. The components are coloured distinctly (30S subunit, yellow; 50S subunit, blue; EF-Tu, red; A/T-tRNA, orange; P-tRNA, green; E-tRNA, brown; mRNA, pink).

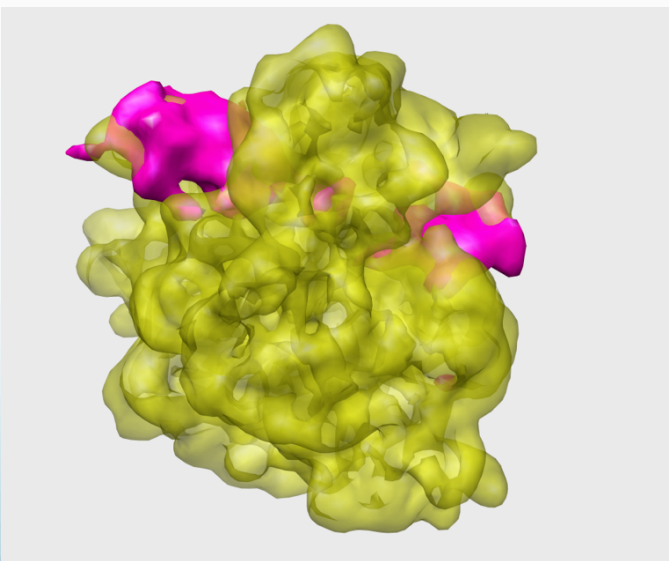
Disordered regions of the EF-Tu ribsomal complex



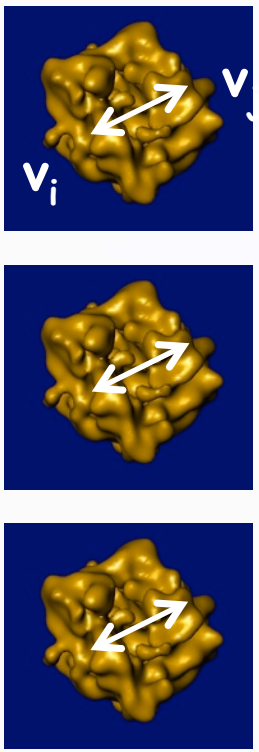
The distribution of variance in the 70S EF-Tu ribosome complex.





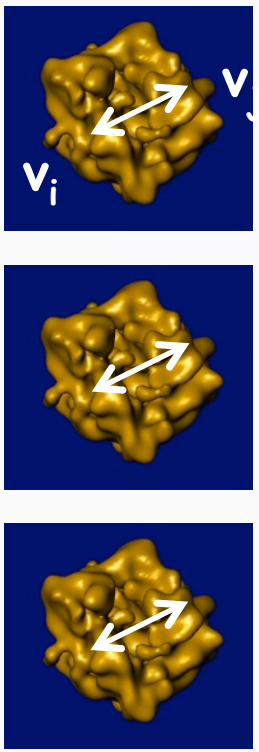


B resampled 3D reconstructions, pair-wise correlations



$$c_{ij} = \sum_{l=1}^{B} \left(v_i^l - \overline{v}_i \right) \left(v_j^l - \overline{v}_j \right)$$

B resampled 3D reconstructions, pair-wise correlations

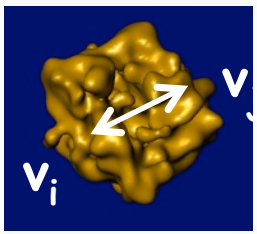


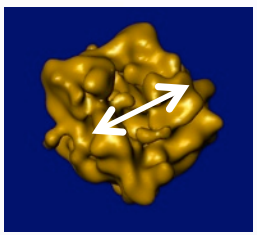
For a volume size n^3,
there are ~n^6 pair-wise correlations (~10^{12})!

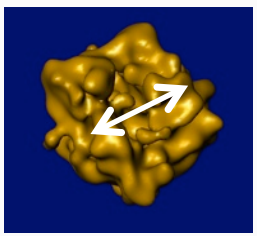
Impossible to visualize/analyze.

$$c_{ij} = \sum_{l=1}^{B} \left(v_i^l - \overline{v}_i \right) \left(v_j^l - \overline{v}_j \right)$$

B resampled 3D reconstructions, pair-wise correlations







For a volume size n^3,
there are ~n^6 pair-wise correlations (~10^{12})!

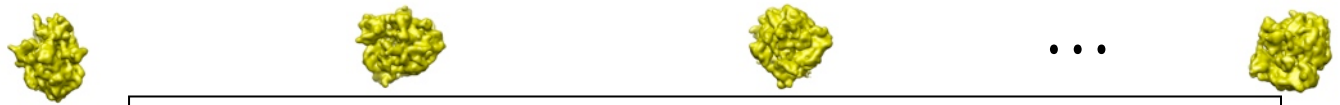
Impossible to visualize/analyze.

Perform eigenanalysis (PCA) of resampled volumes: eigenvectors (eigenvolumes) provide information about variability of the structure, i.e.,

conformational modes of the structure.

$$c_{ij} = \sum_{l=1}^{B} \left(v_i^l - \overline{v}_i \right) \left(v_j^l - \overline{v}_j \right)$$

n 3D structures in random orientations in cryo preparation



Electron microscope records multiple single 2D projections of each 3D structure

Computational structure determination establishes orientation parameters of all 2D projections and yields an "average" 3D map

HYPERSTRATIS: *B* resampled 3D maps

3D PCA of resampled maps yields eigenvolumes

Factorial coordinates **f** computed using 2D projections of eigenvolumes and 2D EM projection data

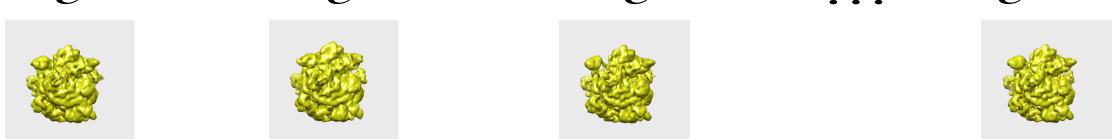
$\mathbf{f}_1$ $\mathbf{f}_2$ $\mathbf{f}_3$ ... $\mathbf{f}_n$

K-means clustering

c_1^3 c_2^1 c_3^K ... c_n^1

3D reconstructions using subsets of projections determined by clustering

c^1 c^2 c^3 ... c^K



Putative conformers

Codimensional - PCA

3D classification of projection images using codimensional PCA

1. Calculation of the large set of resampled volumes using HYPERSTRATIS. **3-D**

2. Eigenanalysis (PCA) of the resampled volumes yields eigenvolumes. **3-D**

3. Calculation of factorial coordinates using of particle projections using a small subset of dominating eigenvolumes. **2-D**

4. Cluster analysis of particle projections using factorial coordinates yields assignments of projections to *K* groups. **factorial**

5. Calculation of *K* 3D structures. **3-D**

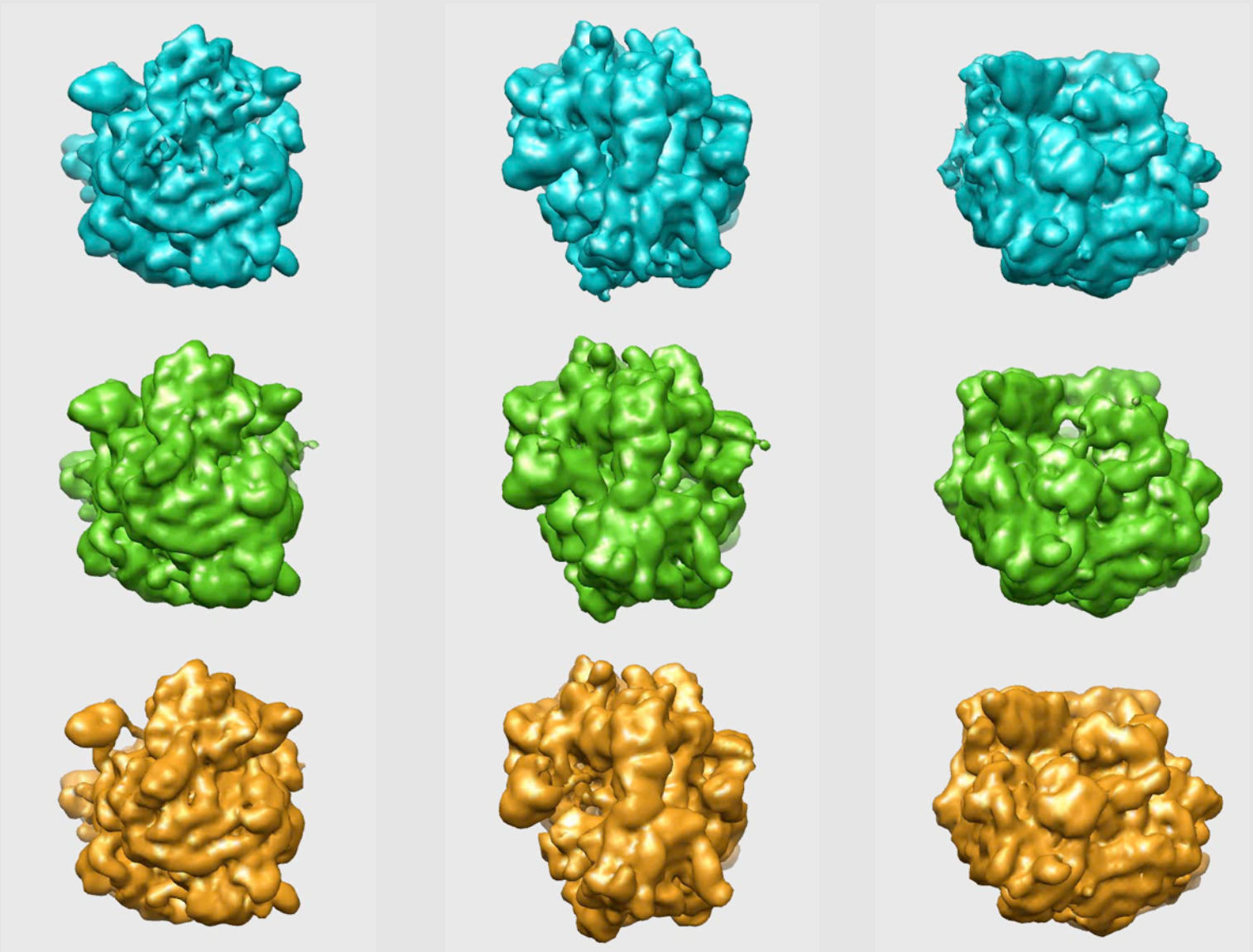
GTPase activation of elongation factor EF-Tu by the ribosome during decoding

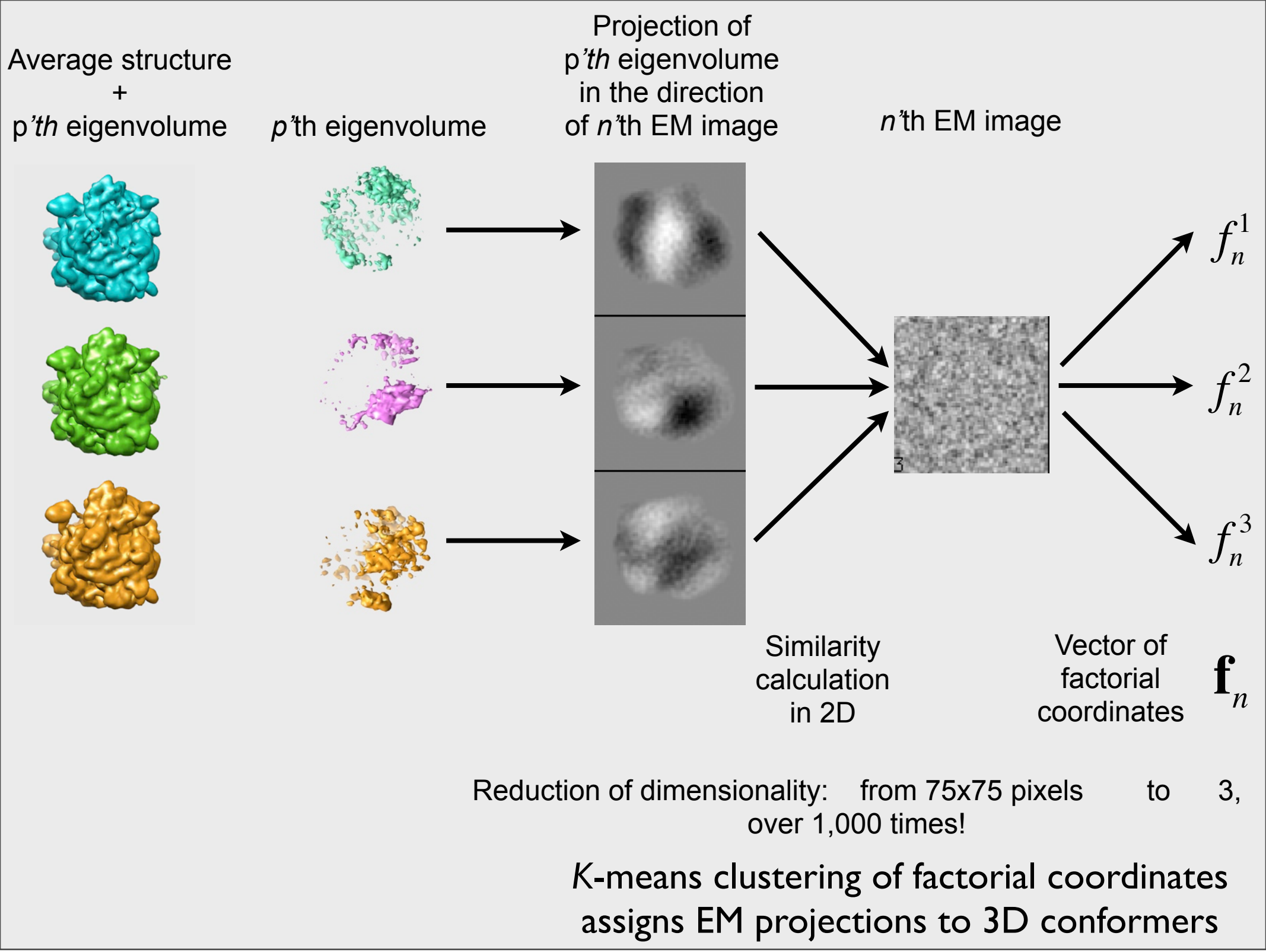
Analysis of the full set of 586,329 cryo-EM projection images of *Thermus thermophilus* 70S ribosome in which the ternary complex of elongation factor Tu (EF-Tu), tRNA and guanine nucleotide has been trapped on the ribosome using the antibiotic kirromycin.

Eigenvolumes

movie = average +/− w × eigenvolume

−1.0<w<1.0





Reduction of dimensionality: from 75x75 pixels to 3, over 1,000 times!

K-means clustering of factorial coordinates assigns EM projections to 3D conformers

Six initial structures obtained from clustering of factorial coordinates

(1) Data set of 586,329 cryo-EM projection images

(2) 100,000 resampled volumes computed using 137,605 projection images each

(3) Voxel-by-voxel 3D variance

(4) Three eigenvolumes

(5) Factorial coordinates

(6) Six clusters

(7) 3D multireference refinement

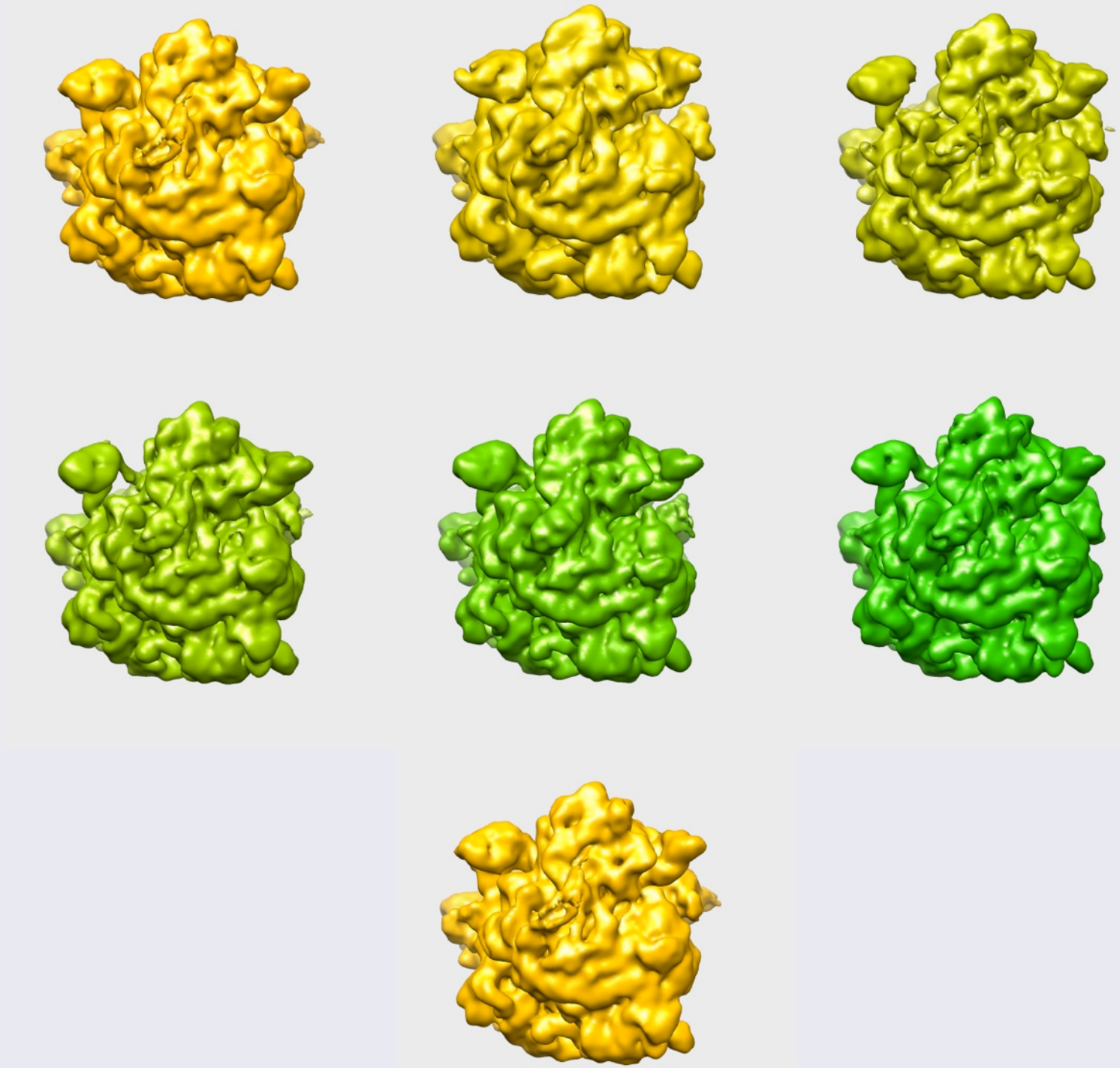
(8) Two clusters collapsed

(9) 3D multireference refinement of four structures (138,900; 87,641; 133,757; 113,743)

(10) Codimensional focused analysis of group #4 to reveal flexibility of stalk

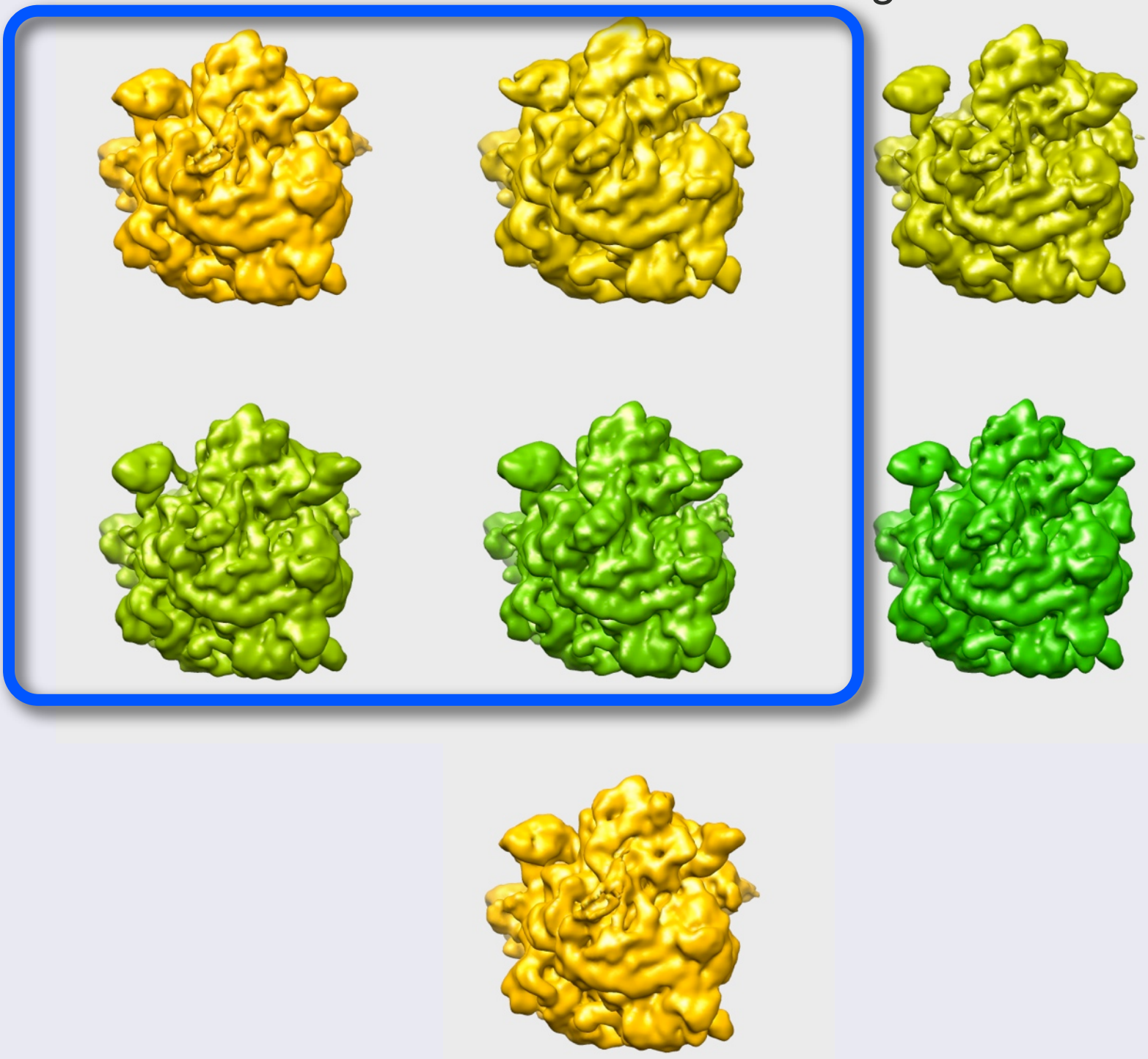
Group #4

Six initial structures obtained from clustering of factorial coordinates

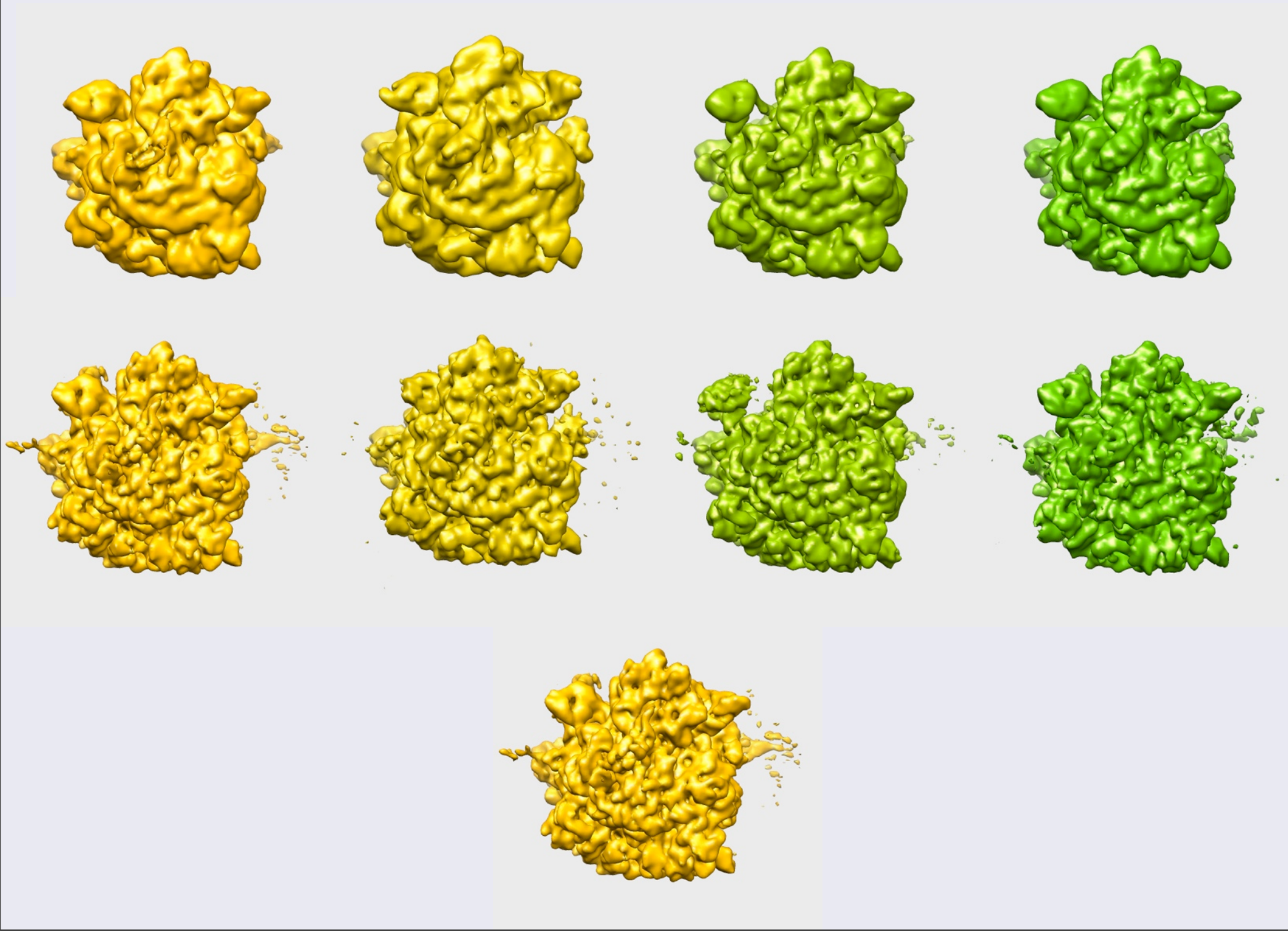


Group #4

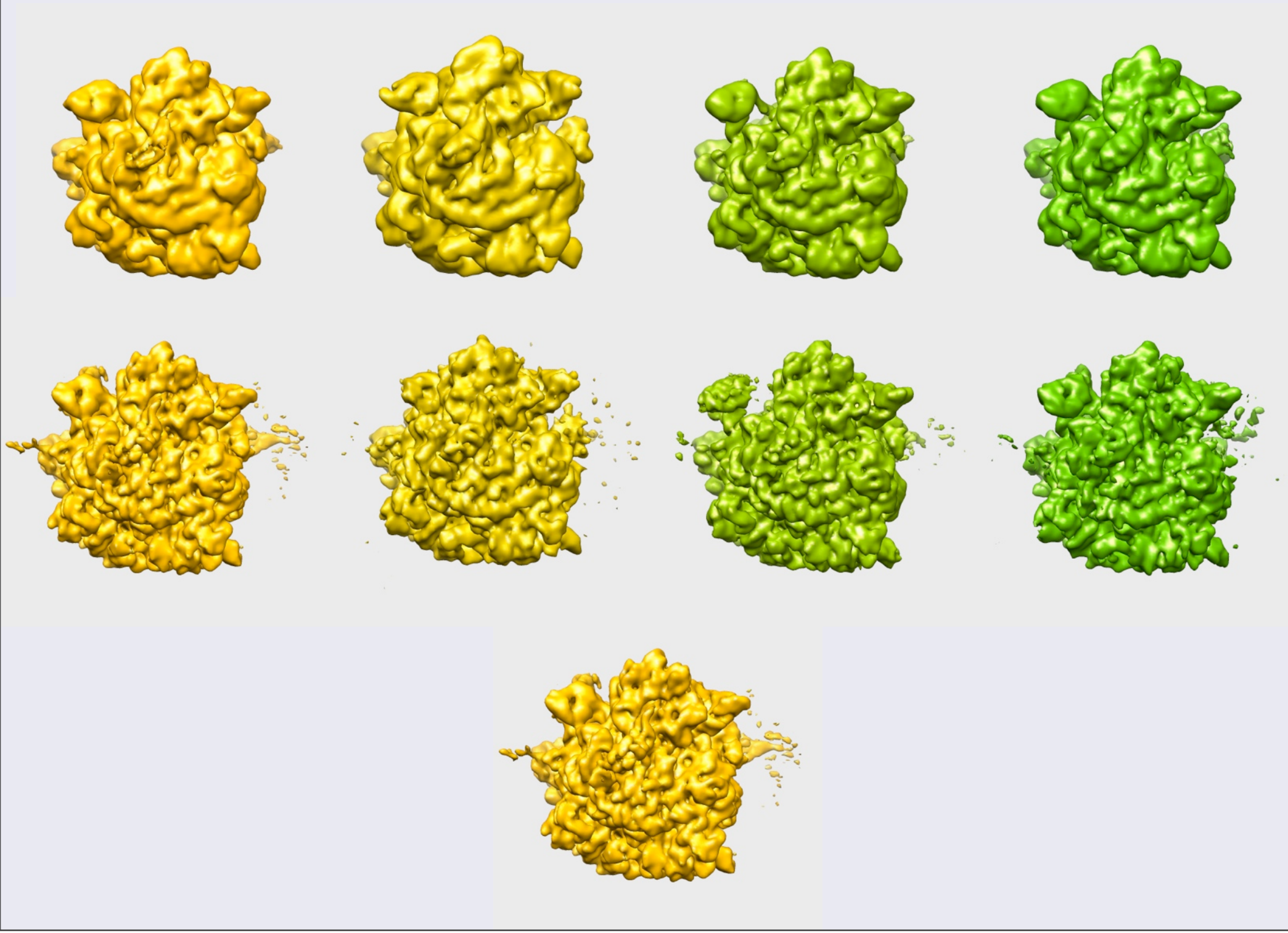
Six initial structures obtained from clustering of factorial coordinates



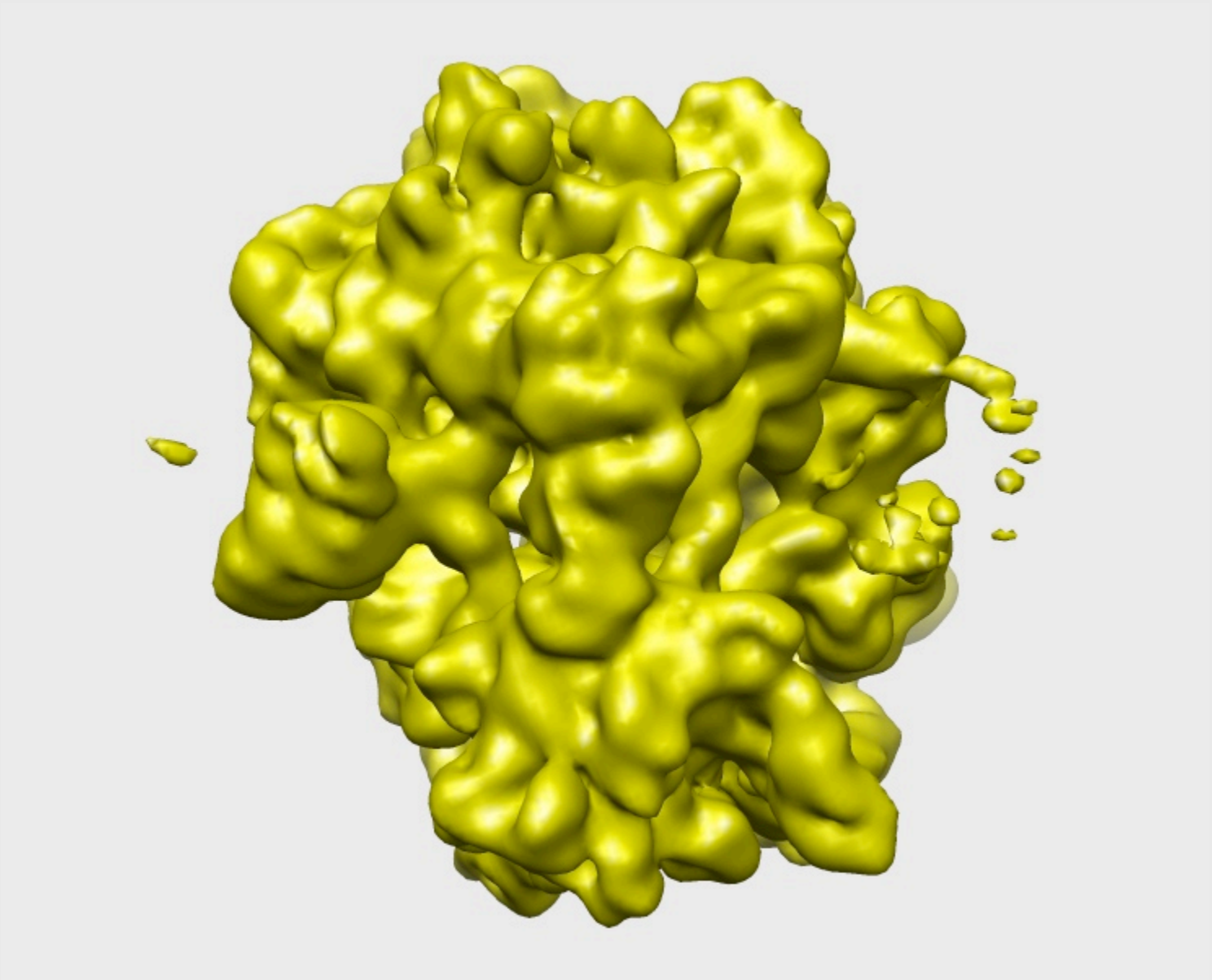
Four structures obtained after multireference refinement



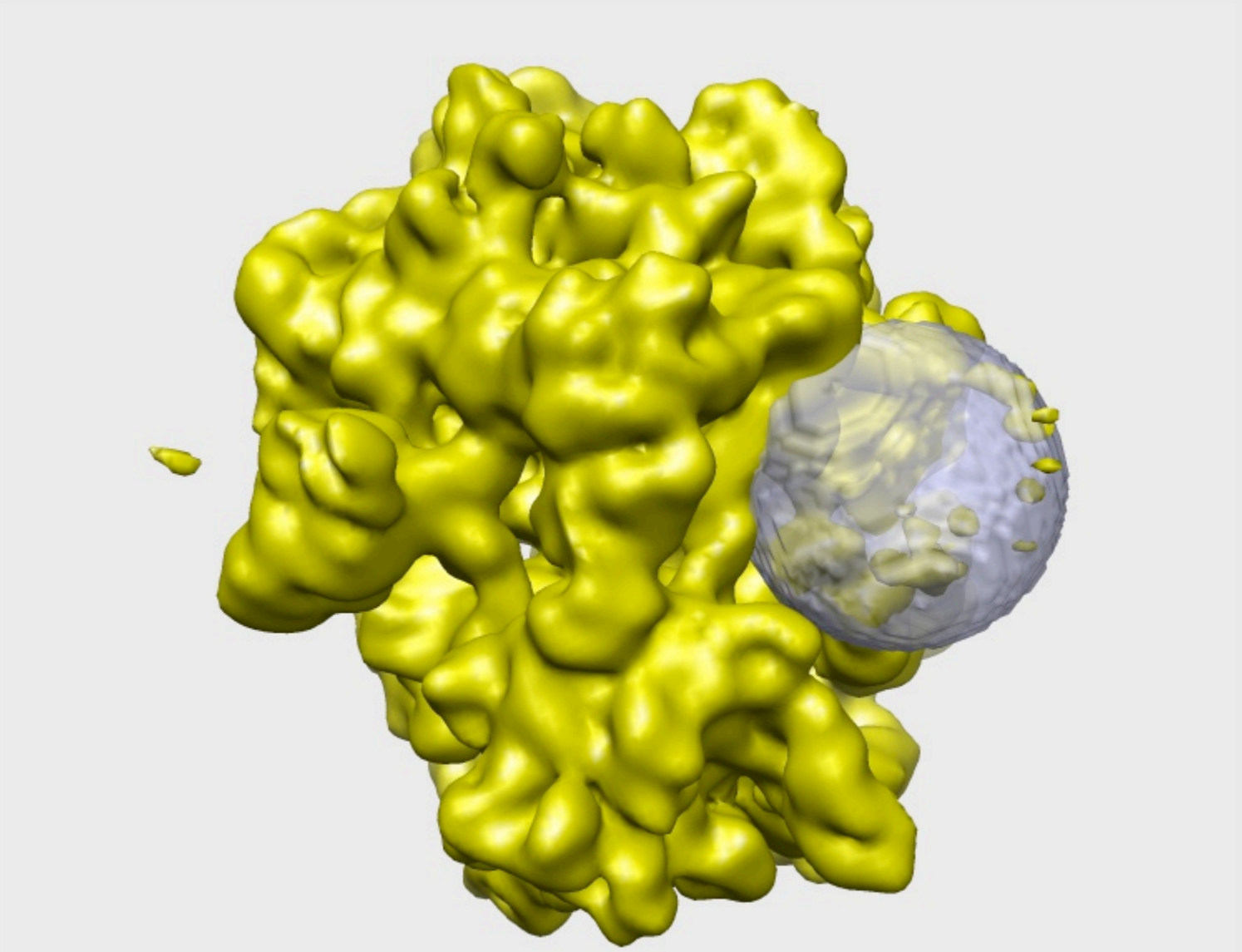
Four structures obtained after multireference refinement

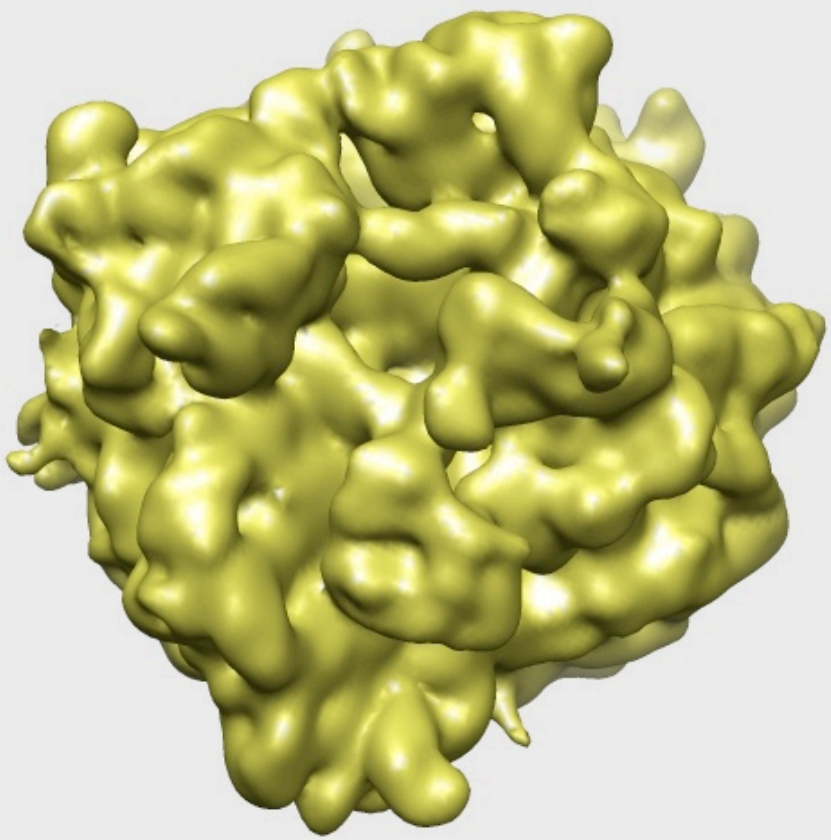


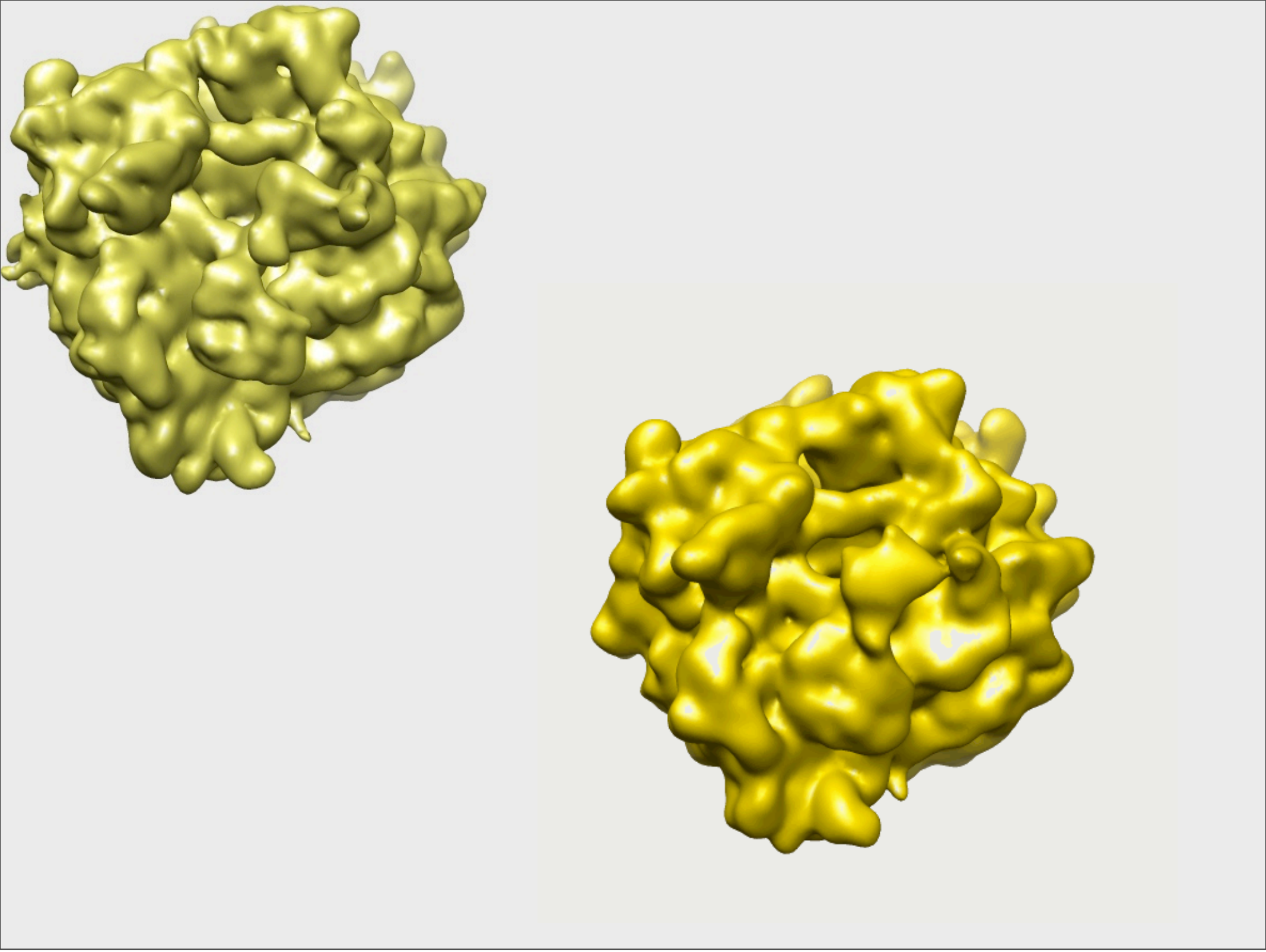
"Focused" analysis of stalk region in group #4



"Focused" analysis of stalk region in group #4







Acknowledgments

Christian M.T. Spahn,
Charité, Berlin

Marek Kimmel,
Rice University, Houston

SPARX

Please download file with the test data for the afternoon sessions (3pm).

File: codim.tar.gz

http://sparx-em.org/sparxwiki/codim