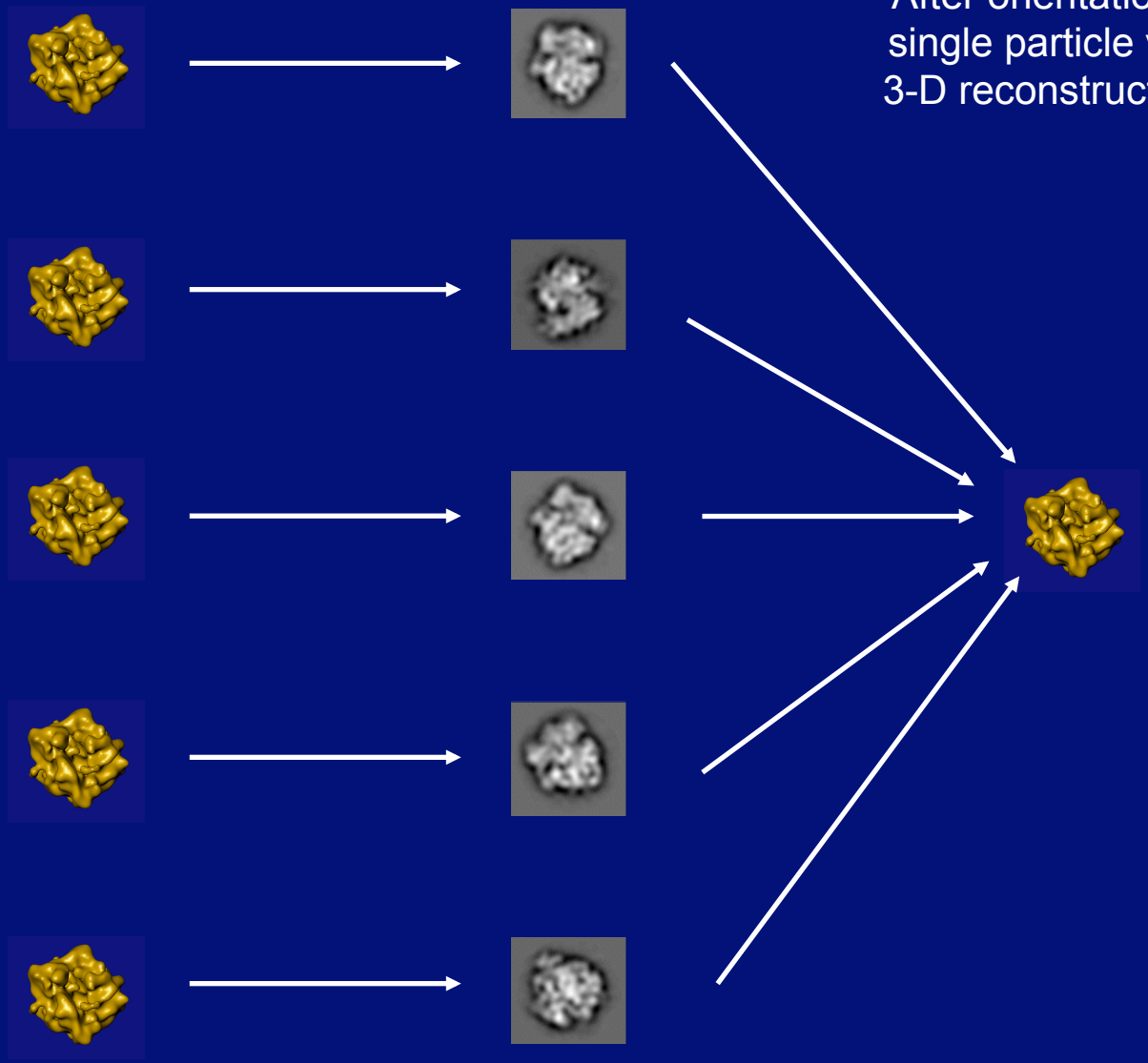
Analysis of conformational variability of macromolecules in cryo-electron microscopy

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.



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

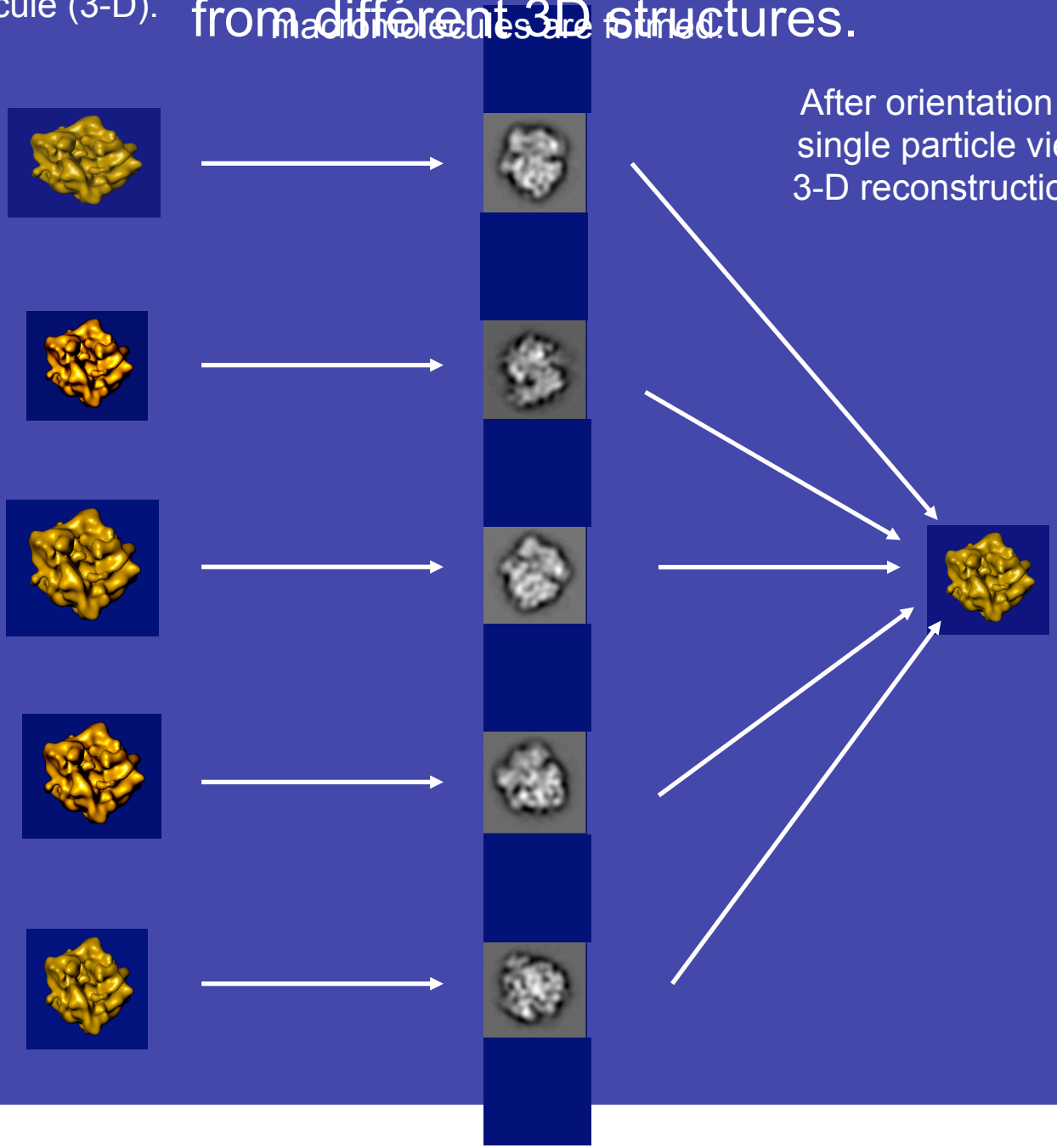
- ribosome
- RNA polymerase
- human transcription factor
- pyruvate dehydrogenase complex (breathing core)

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.

Different states of the same macromolecule (3-D).

In electron microscope, 2-D projections of observed macromolecules are formed.



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

Computational time-resolved cryo-EM

- Multi-reference alignment
- Focused classification
- Multiple particle analysis

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. J. Struct. Biol. 147, 291-301.

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 3-D 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.

3-D 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

Bootstrap technique

Resampling with replacements

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

1 2 3 4 5 6 7 8 9

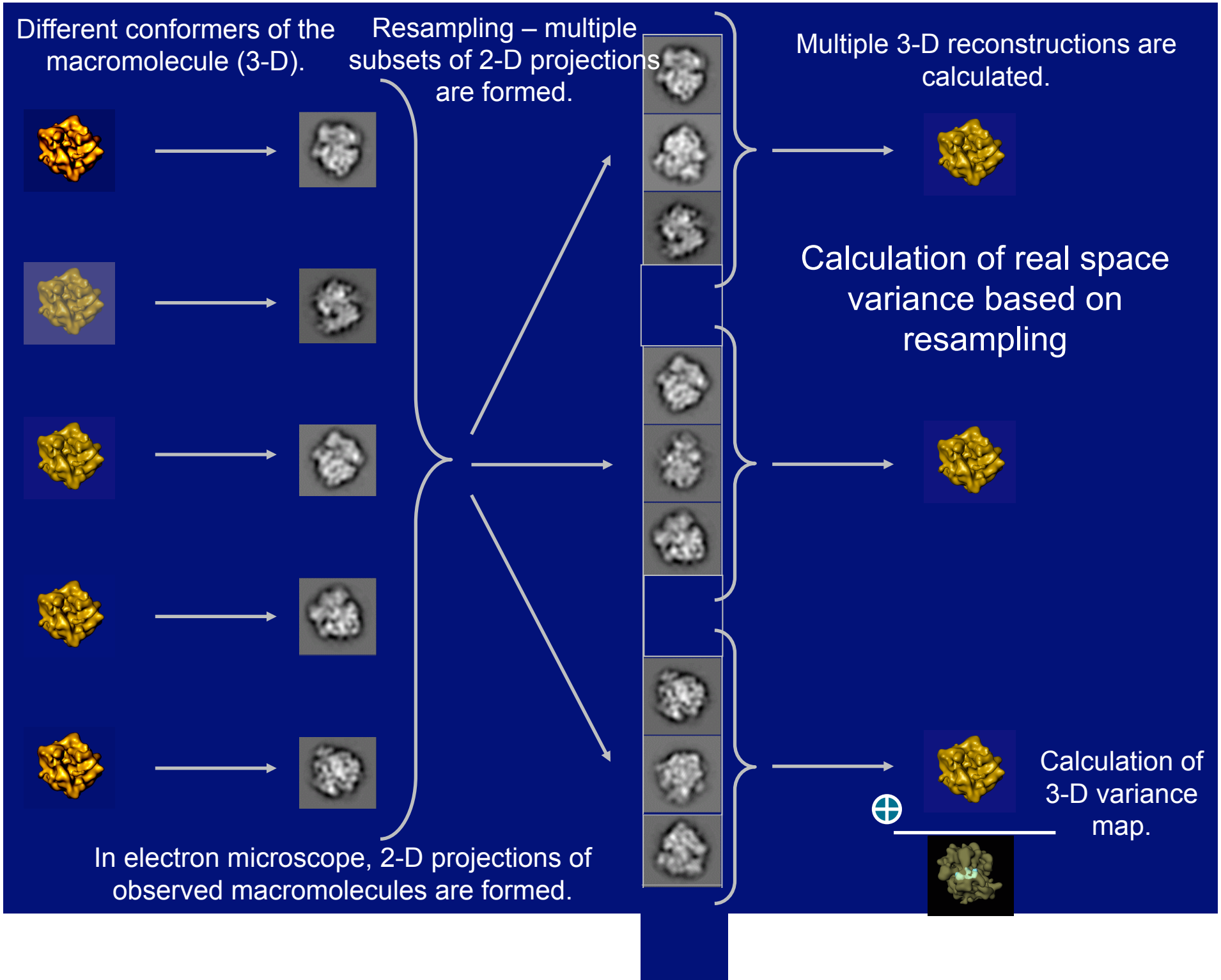
Resampled data sets of 2-D projections, each contains nine projections.

1 1 3 4 4 4 4 8 8

2 3 4 5 6 8 8 9 9

3-D reconstruction
Large number of “different” volumes

Variance/covariance!


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

Sources of variance in 3-D reconstructions

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

Calculation of the variance of structures

$$\sigma^2_{Struct} = K\left(\sigma^2_B - \bar{\sigma}^2_{Back}\right)$$

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.

Test in the presence of additive noise $N(0,30)$, SNR = 2.3 in the projection data. B = 500 bootstrap volumes

(a) Average of low-passed model structures.

(b) The variance calculated using 1,253 simulated low-passed model structures.

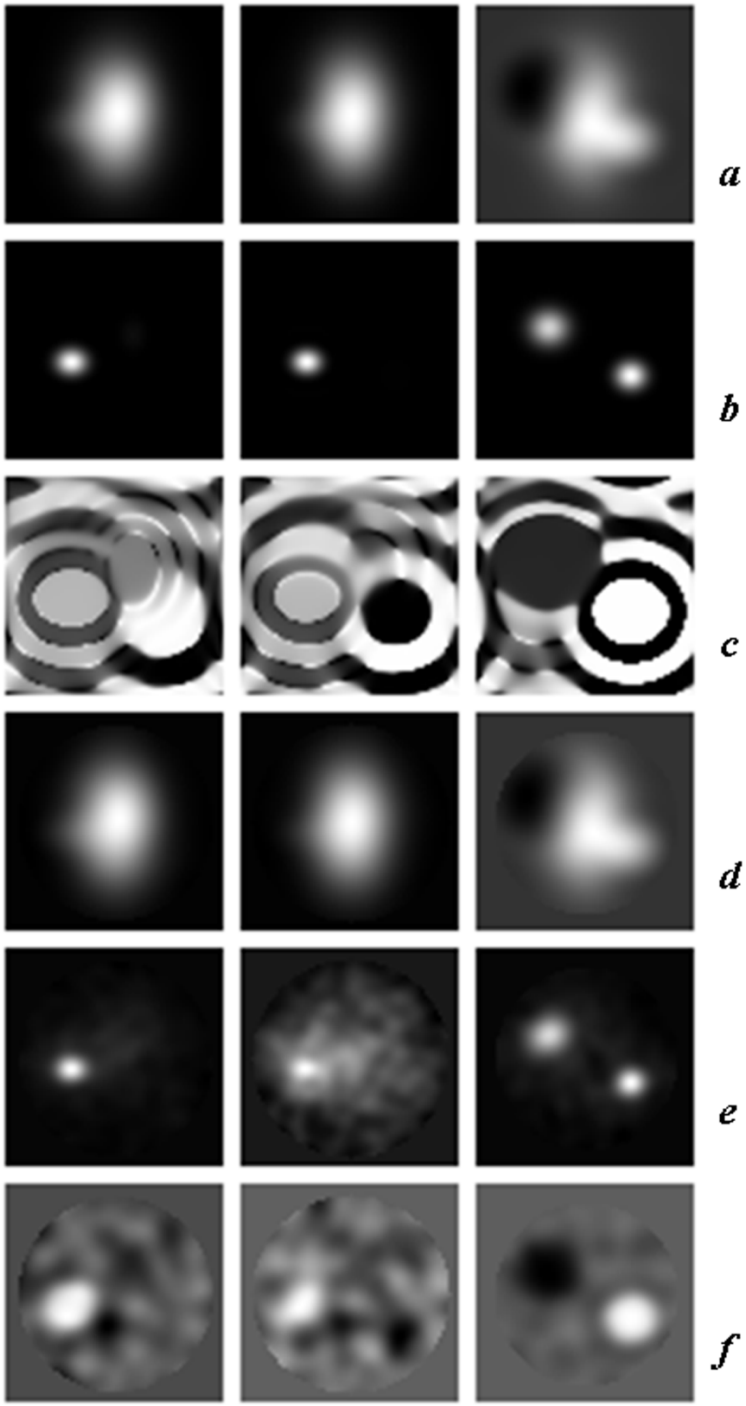
(c) Correlation map between the center of the feature *A* and the remaining voxels calculated for simulated low-passed volumes. The unusual pattern is due to correlations introduced into the volumes by the process of low-pass filtration.

(d) The average of low-passed bootstrap structures.

(e) Structure variance calculated using the bootstrap method and estimated from low-passed sample volumes.

(f) Correlation map between the center of the feature *A* and the remaining voxels calculated using low-passed bootstrap volumes.

Contrast within each slice adjusted independently, so the intensities do not reflect absolute values in respective slices.

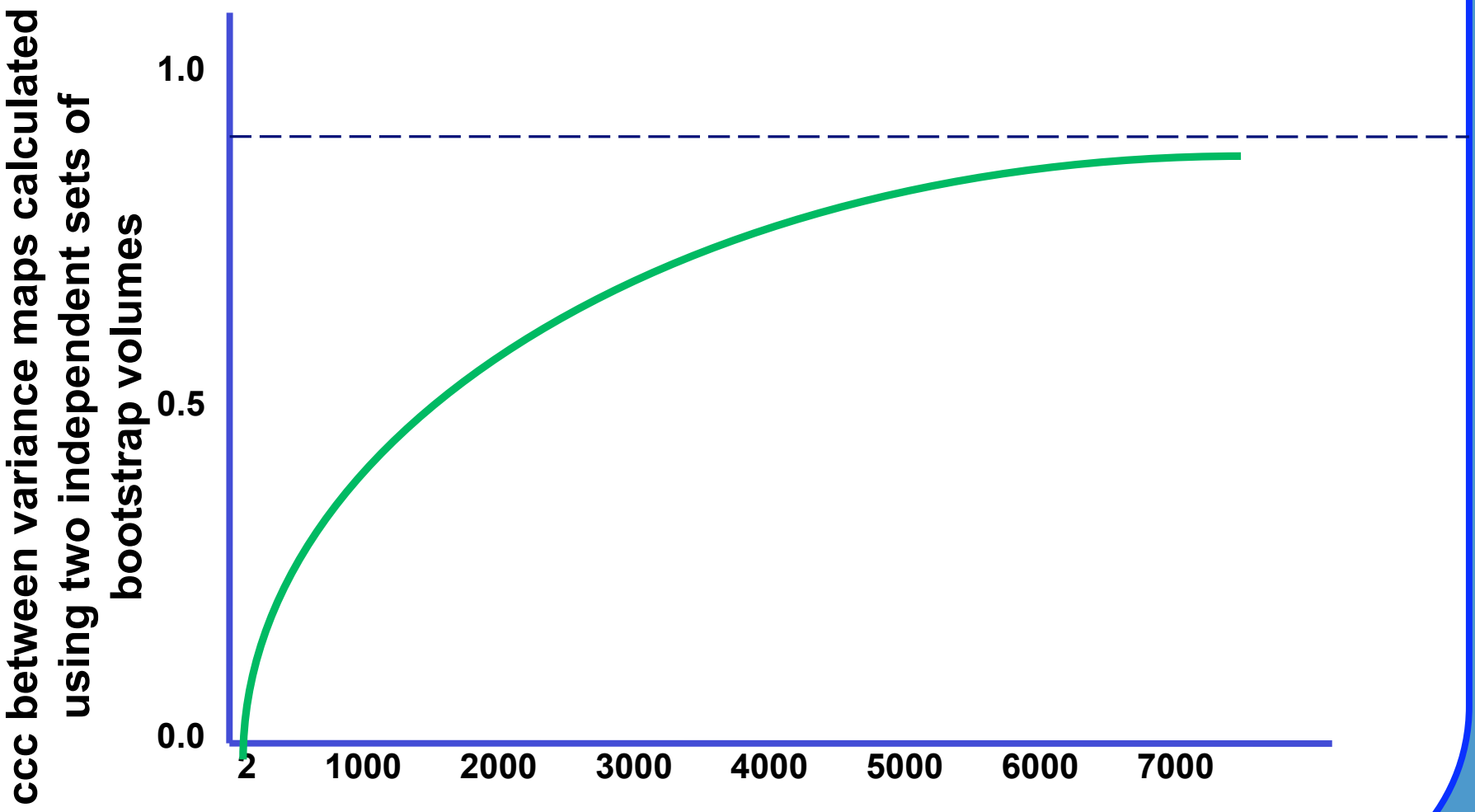


Test of the estimation of the structure variance using the bootstrap method in the presence of additive independent Gaussian in projections.

	A	*B*	*C*	*D*	*Center*
σ_f^2	1.17	1.24	0.65	1.35	$4\text{x}10^{-2}$
$K\sigma_{SVar}^2$	1.51	1.59	1.14	1.83	0.77
σ_{Struct}^2	1.19	1.28	0.82	1.52	0.46
r_f	1.00	0.00	-0.70	0.44	0.19
r_{SVar}	0.87	-0.02	-0.44	0.34	$-7\text{x}10^{-3}$

$$\sigma_{Struct}^2 = K\left(\sigma_{SVar}^2 - \bar{\sigma}_{Back}^2\right)$$

Determination of the number of bootstrap volumes



How well bootstrap technique approximates the variance of the structure?

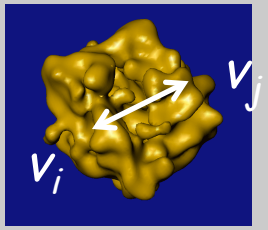
Expectation value of the correlation coefficient (ρ_γ) of the variance of original distribution ($S^2(\sigma^2_{Struct})$) and the bootstrap variance.
N - sample size; *B* – number of bootstrap samples.

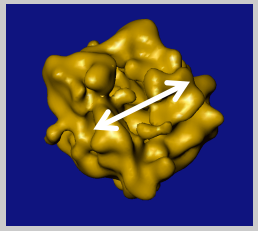
$$E(\rho_\gamma) \cong \sqrt{\frac{S^2(\sigma^2_{Struct})}{S^2(\sigma^2_{Struct}) + \frac{\bar{m}_4 - \bar{m}_2^2}{N} + \frac{2\bar{m}_2^2}{B}}}.$$

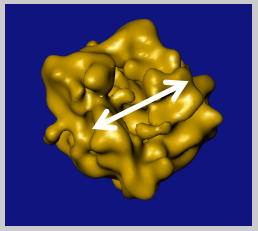
For small *N*, ρ_γ is always less than 1, no matter how large *B* is.
The only way to obtain more accurate variance map is to increase the data set.

Zhang, W., Kimmel, M., Spahn, C.M., Penczek, P.A.: Heterogeneity of large macromolecular complexes revealed by 3D cryo-EM variance analysis. Structure 16:1770-1776, 2008.

B bootstrap 3-D 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 of bootstrap volumes: eigenvectors (eigenvolumes) provide information about variability of the structure, i.e., conformational modes of the structure.

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

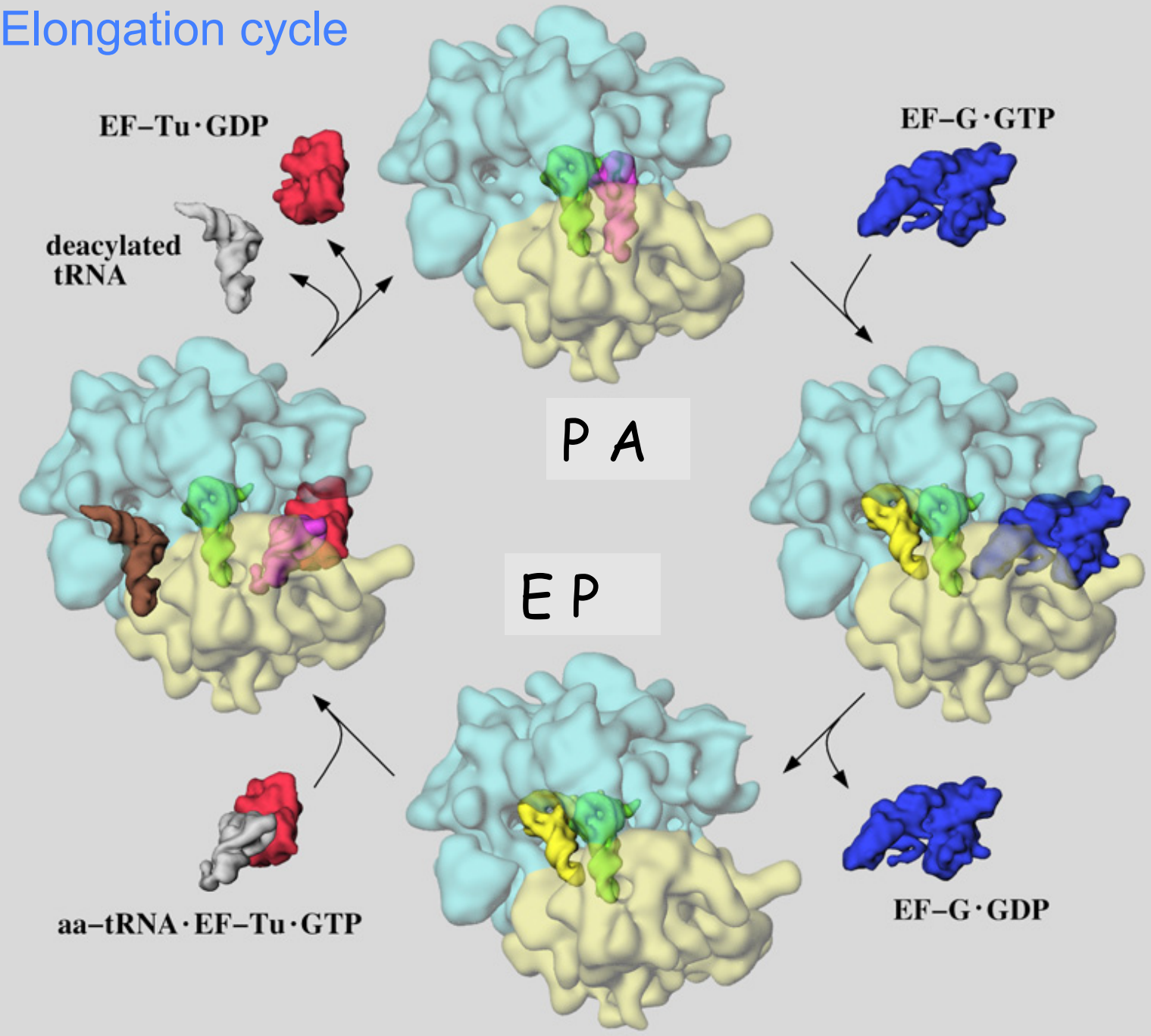
GTPase activation of elongation factor EF-Tu by the ribosome during decoding

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.

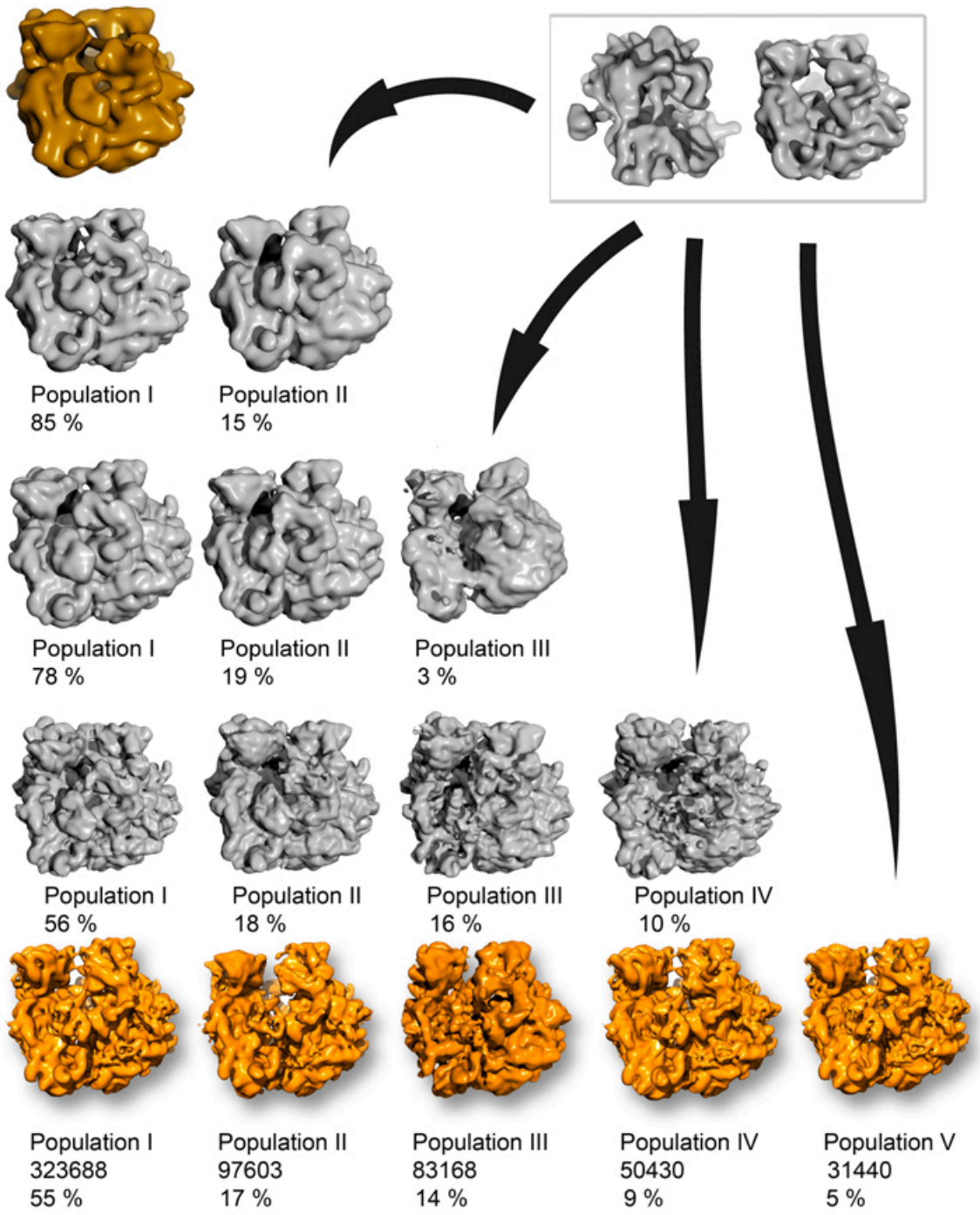
Resolution: 6.5 Å.

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



3D clustering with predetermined seed



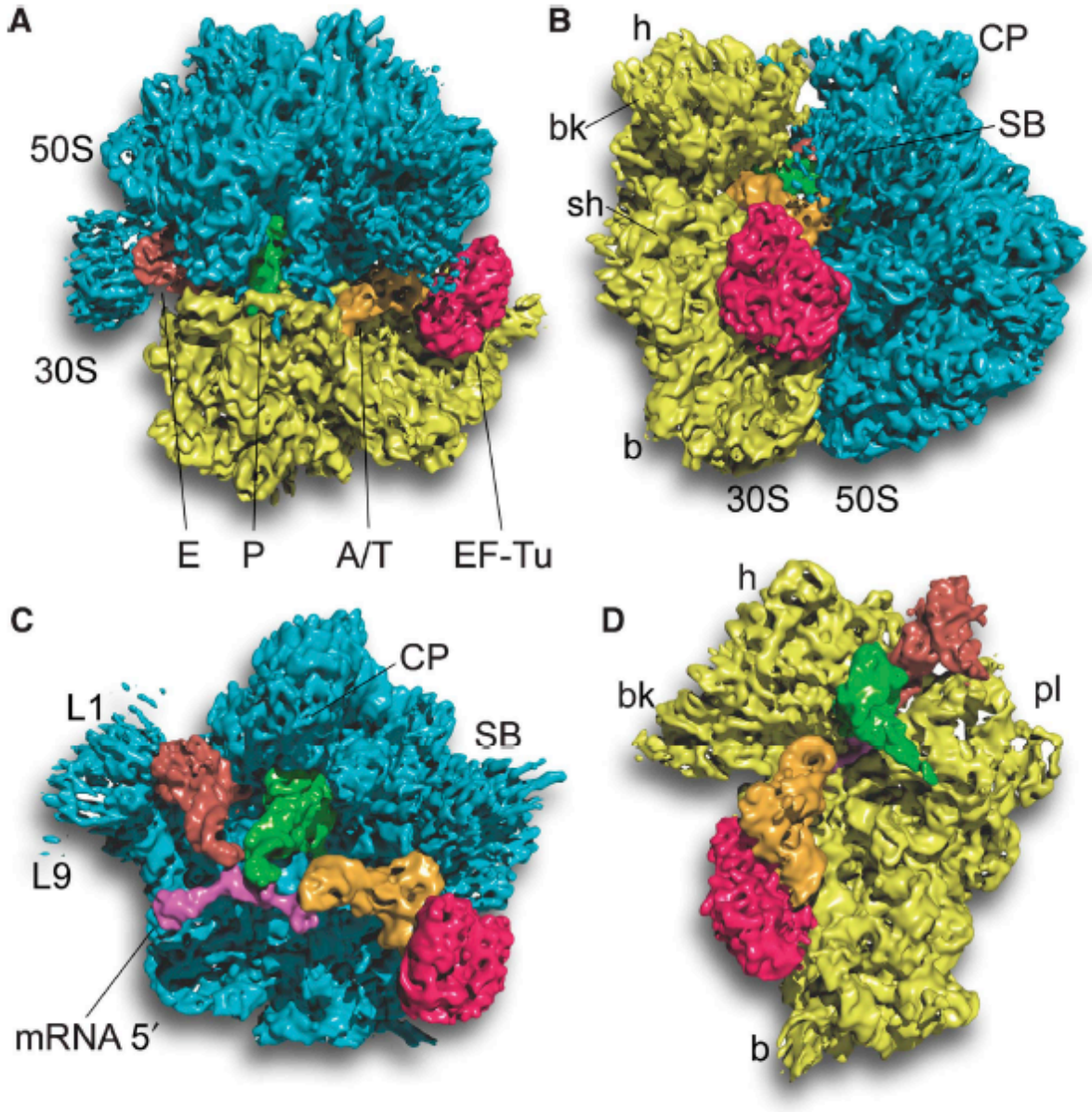
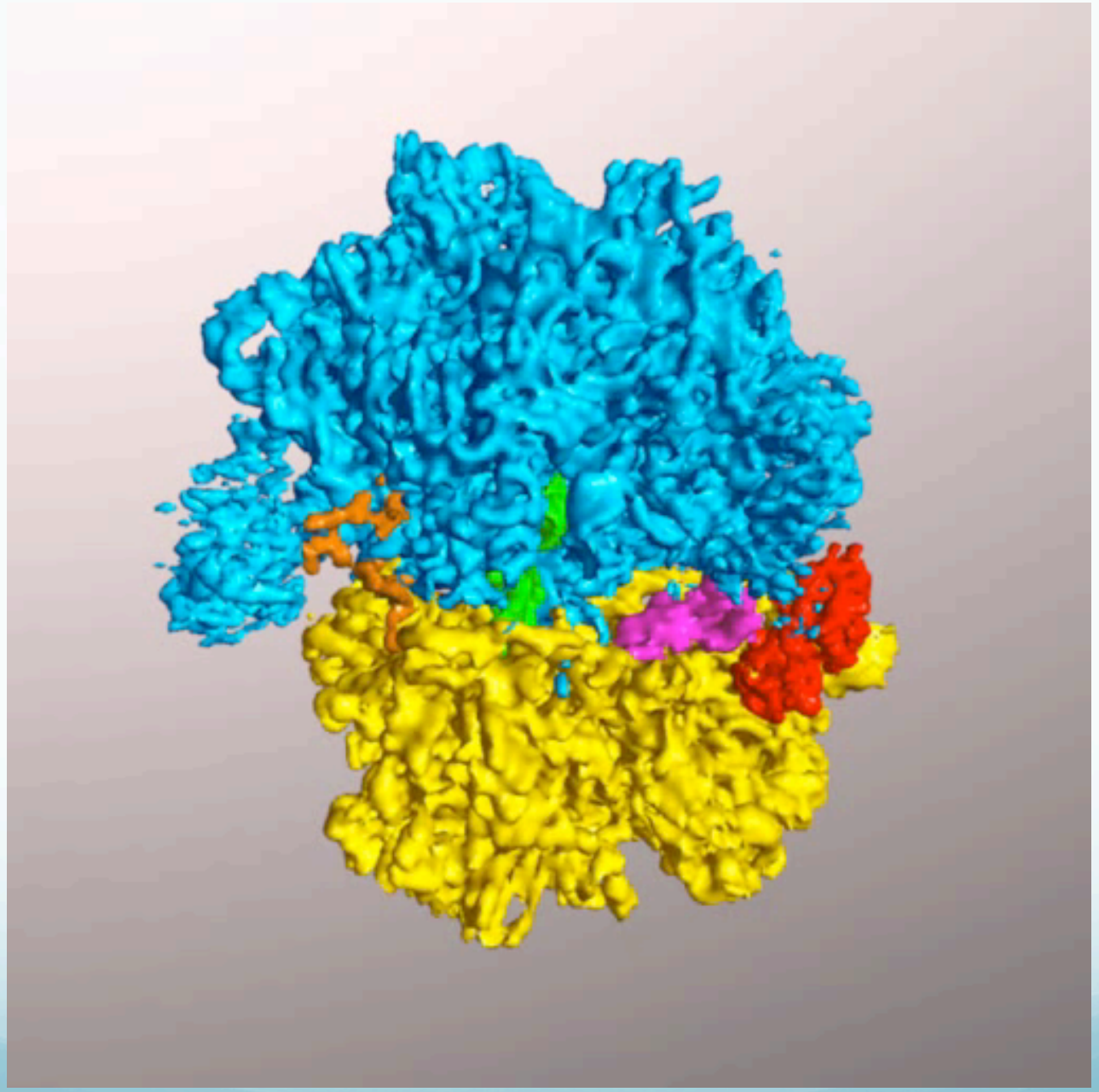
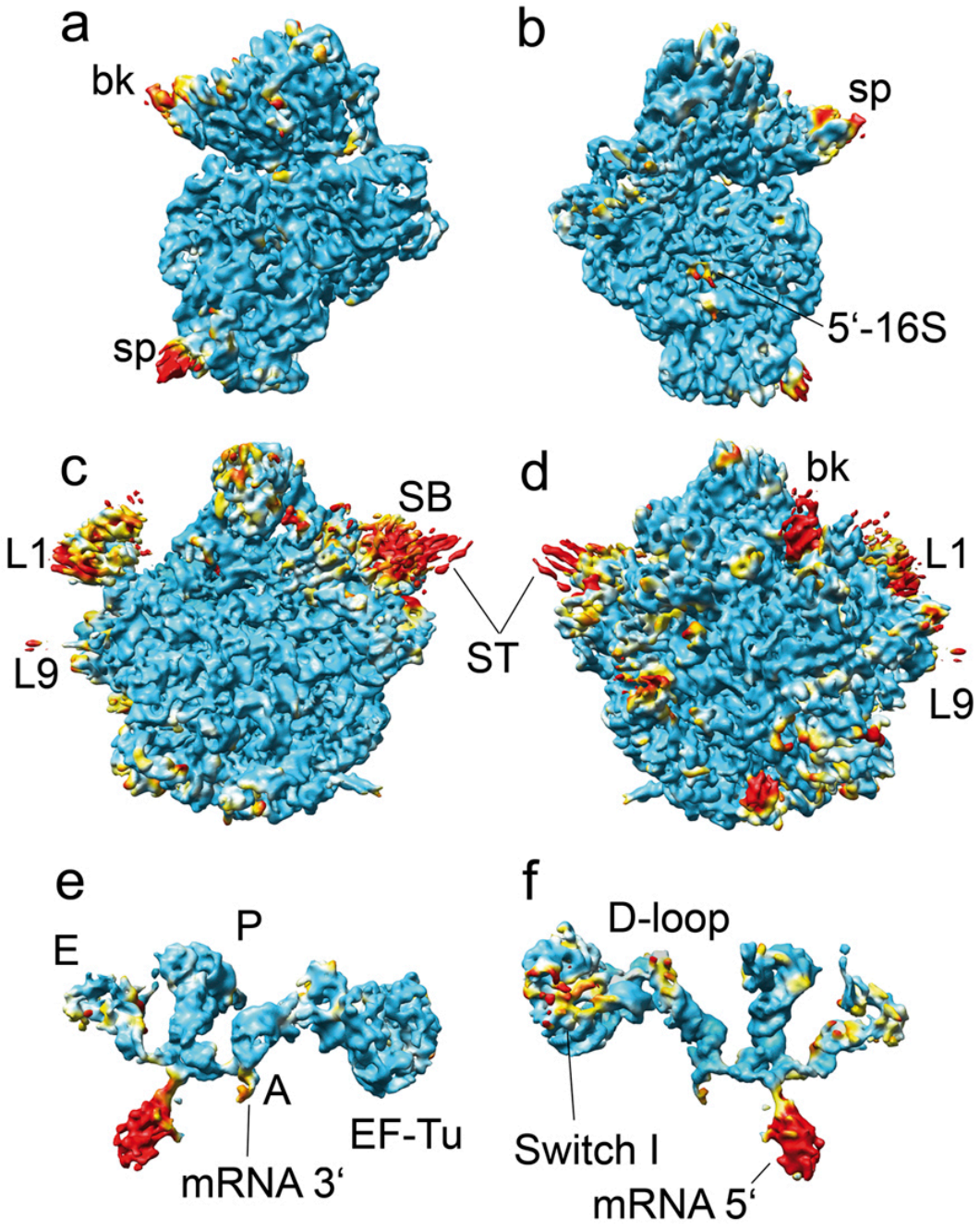


Figure 1 Overview of the 70S●EF-Tu●Phe-tRNA●GDP●kirromycin complex. A surface representation of the cryo-EM map is shown (**A**) from the top; (**B**) from the L7/L12 side; (**C**) from the 30S side, with 30S removed and (**D**) from the 50S side, with 50S removed. The components are coloured distinctly (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 robisomal complex



Analysis of conformational variability of EF-Tu ribosomal complex

① Data set of 323,688 cryo-EM projection images

② 140,000 bootstrap volumes

③ Voxel-by-voxel 3D variance

④ 9 eigenvolumes

⑤ Factorial coordinates

⑥ Two clusters

⑦ 3D multireference refinement.

3-D classification of projections using bootstrap technique

1. Calculation of the large set of resampled volumes using bootstrap technique.

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

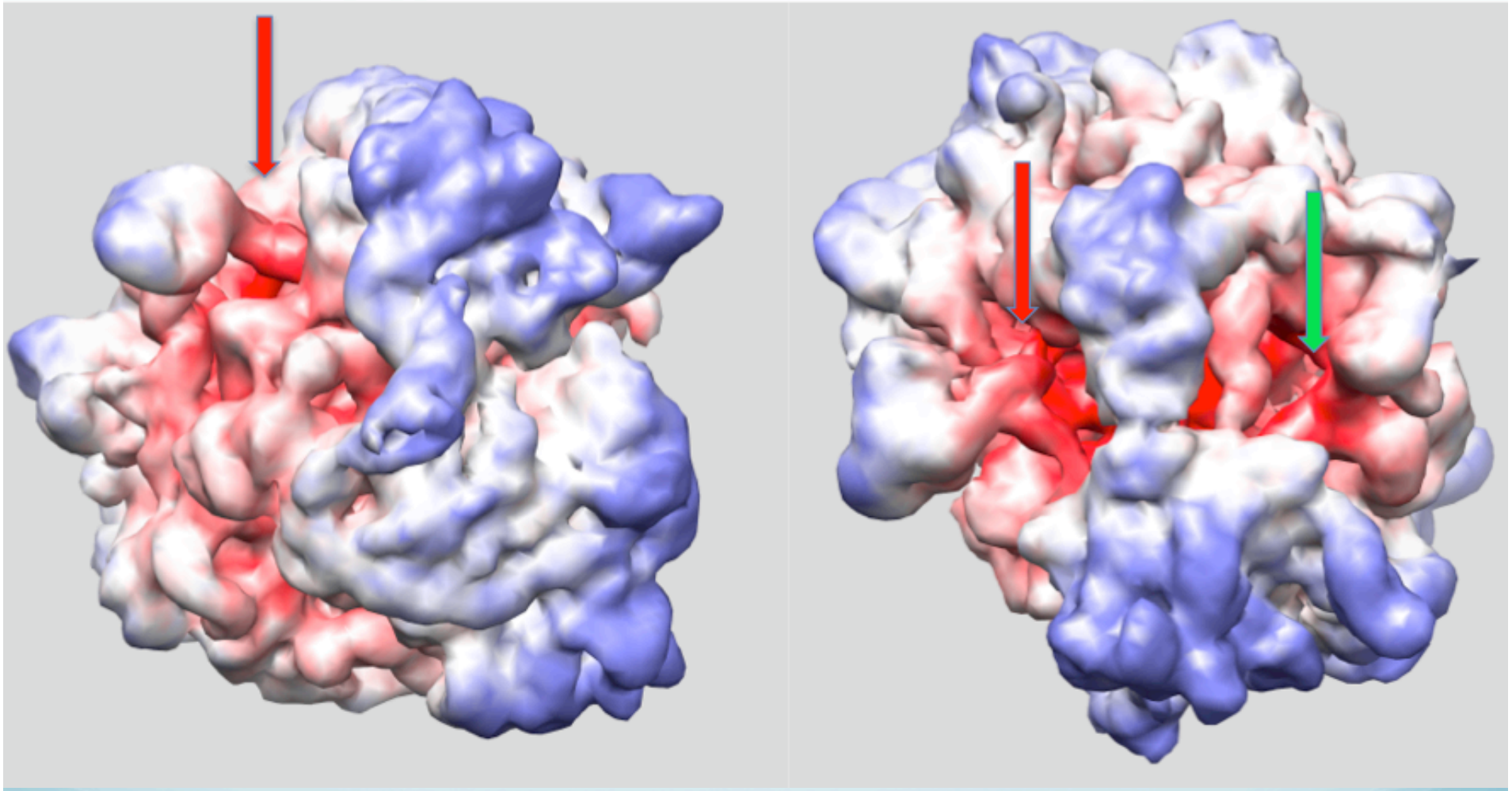
4. Cluster analysis of particle projections using factorial coordinates yields assignments of projections to *K* groups.

factorial

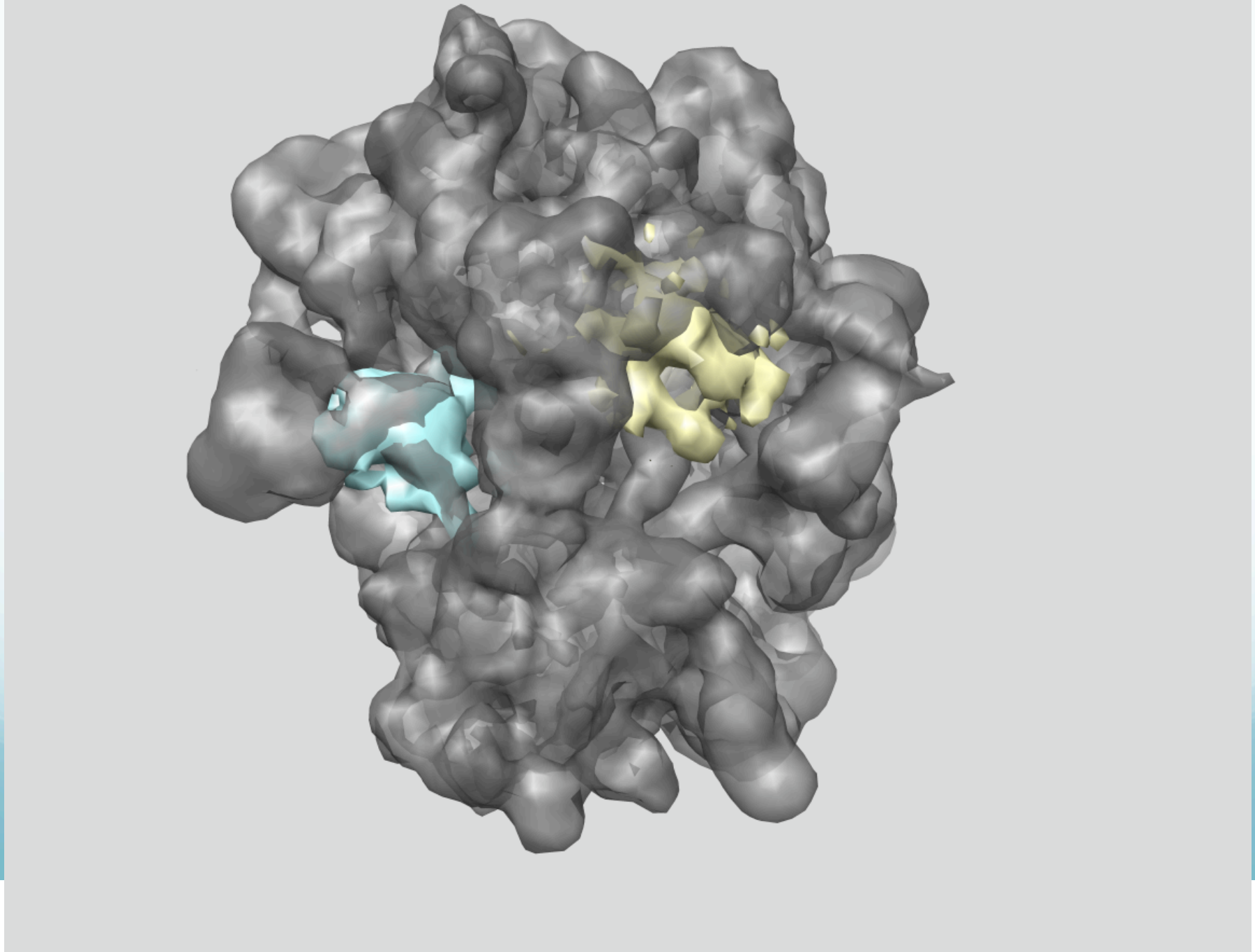
5. Calculation of *K* 3-D structures.

3-D

The color-coded distribution of variance on the surface of 70S ribosome complex.

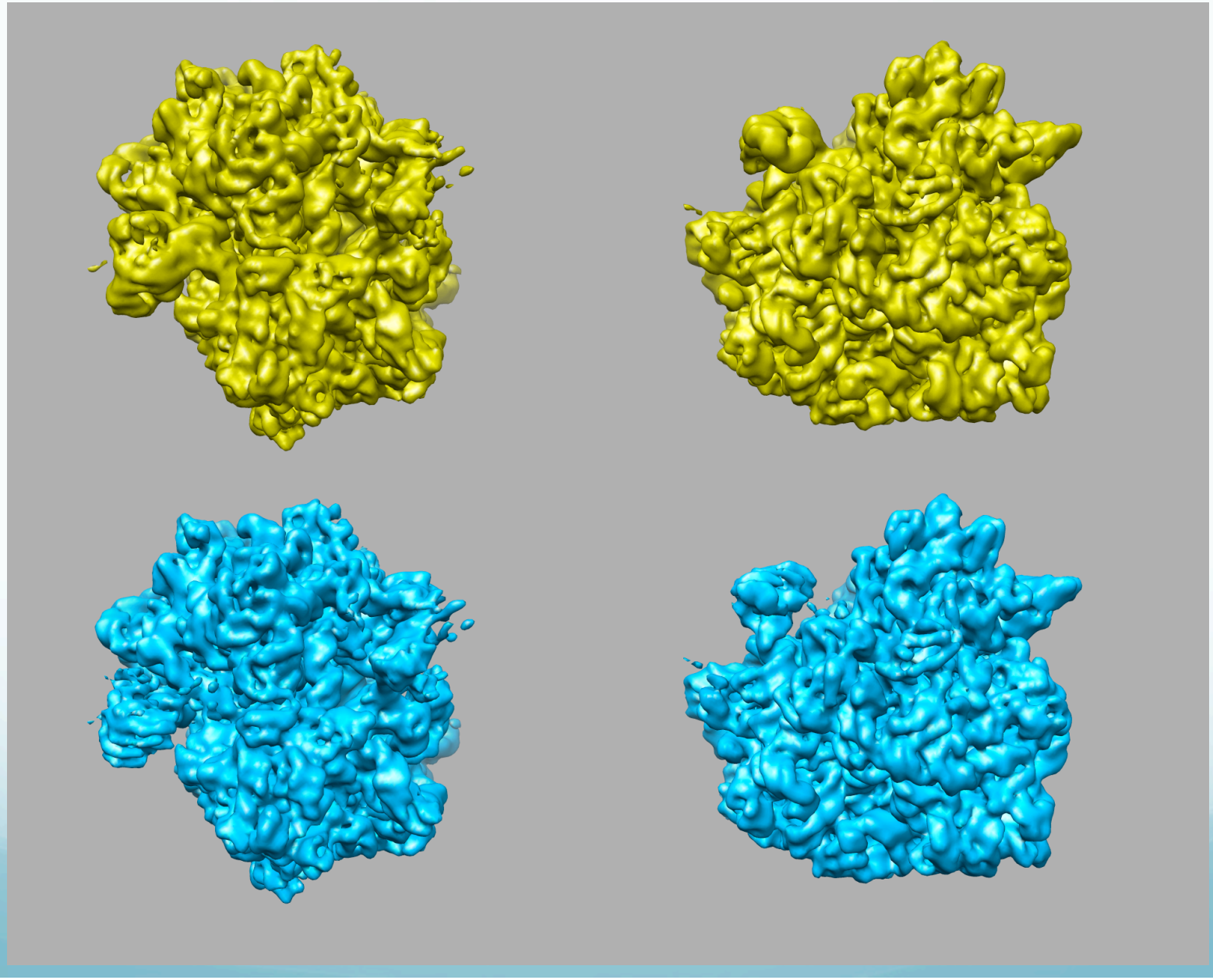


First eigenvolume
(positive: yellow; negative: blue) after Varimax transformation



3D multi-reference alignment

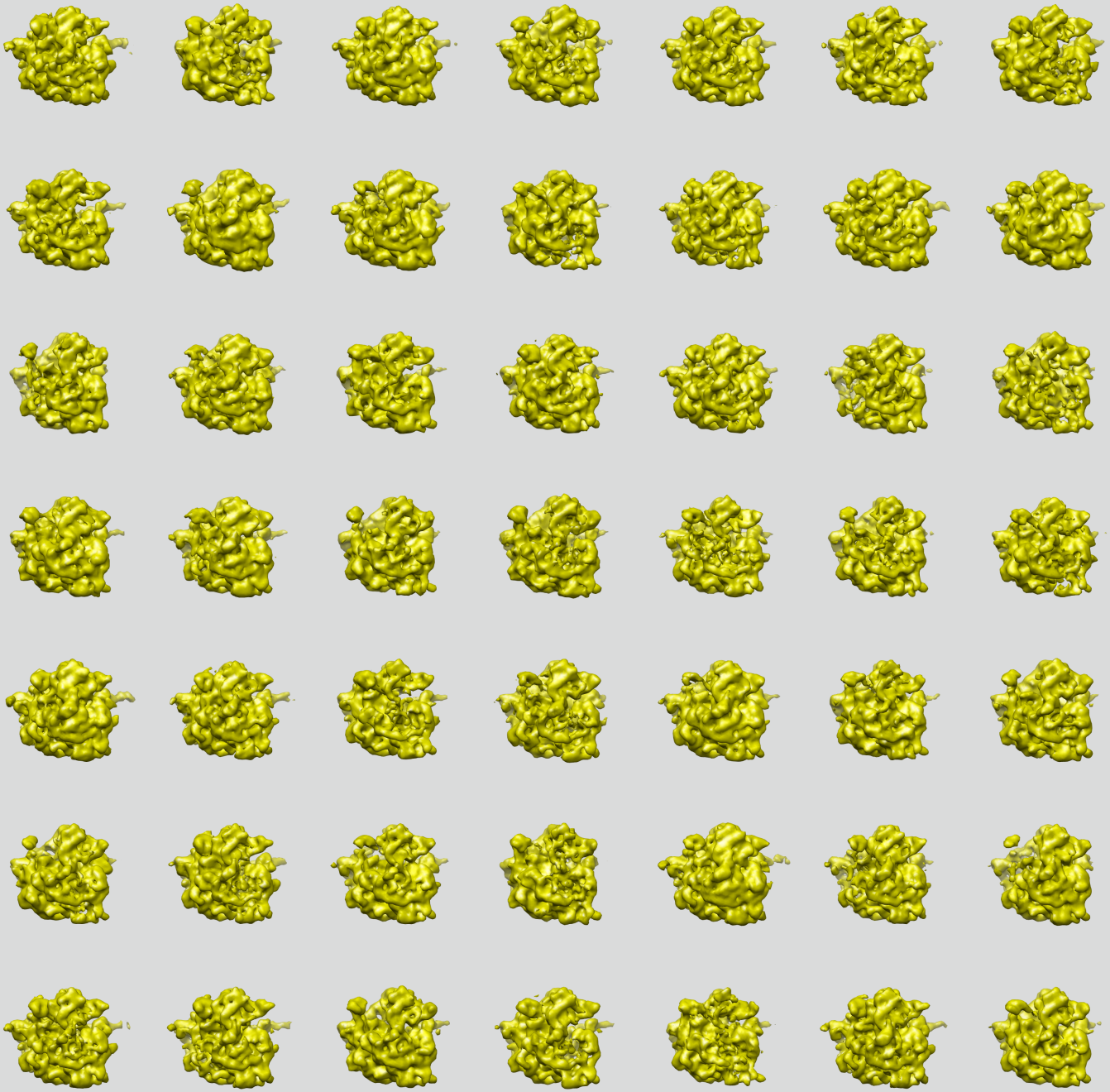
152,104



171,584

Bootstrap combined with PCA of projections works very well….

too well!



Determination of conformational heterogeneity as a clustering problem

- *Clustering* is the process of identifying natural groupings in the data
- *Clustering* is the assignment of a set of objects into subsets so that objects in the same cluster are similar in some sense

Unsupervised learning technique

- No predefined class labels

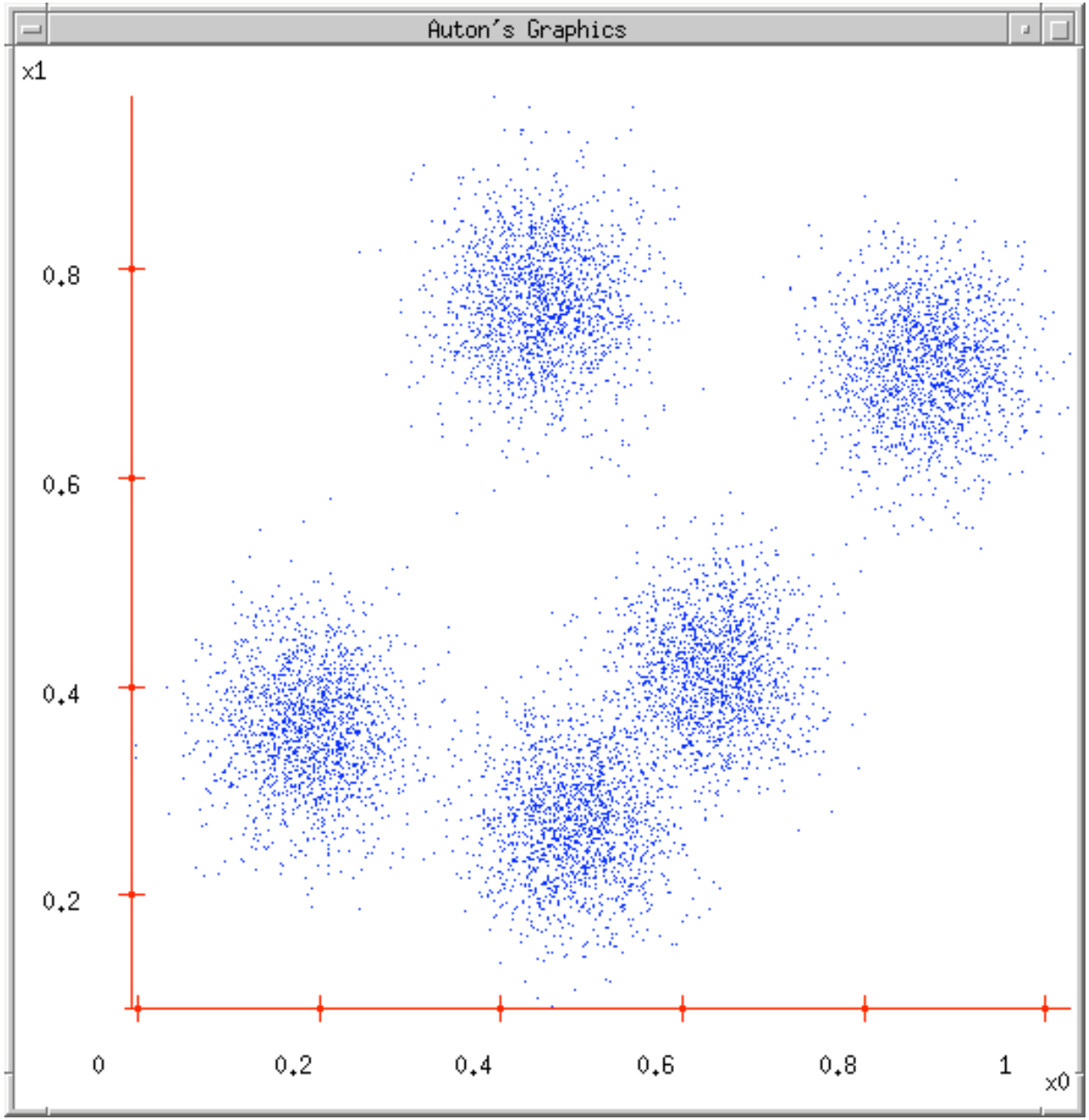
The *K*-means method is by far the most popular clustering algorithm used in scientific and industrial application.

K-means is both very simple and very fast, which makes it appealing in practice.

K-means begins with an arbitrary clustering based on *K* centers, and then repeatedly makes local improvements until the clustering stabilizes.

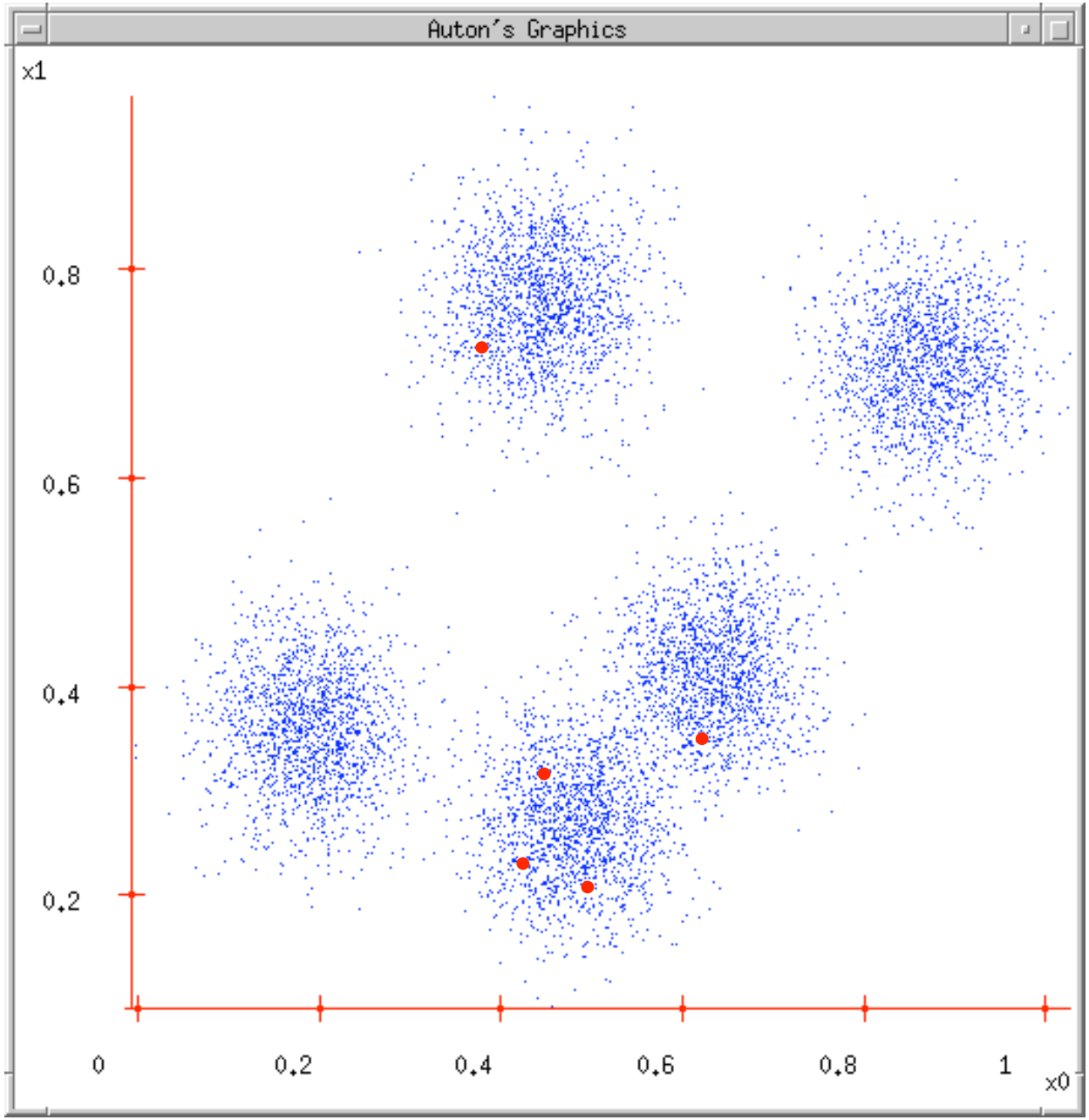
K-means

1. Ask user how many clusters they'd like. *(e.g. K=5)*



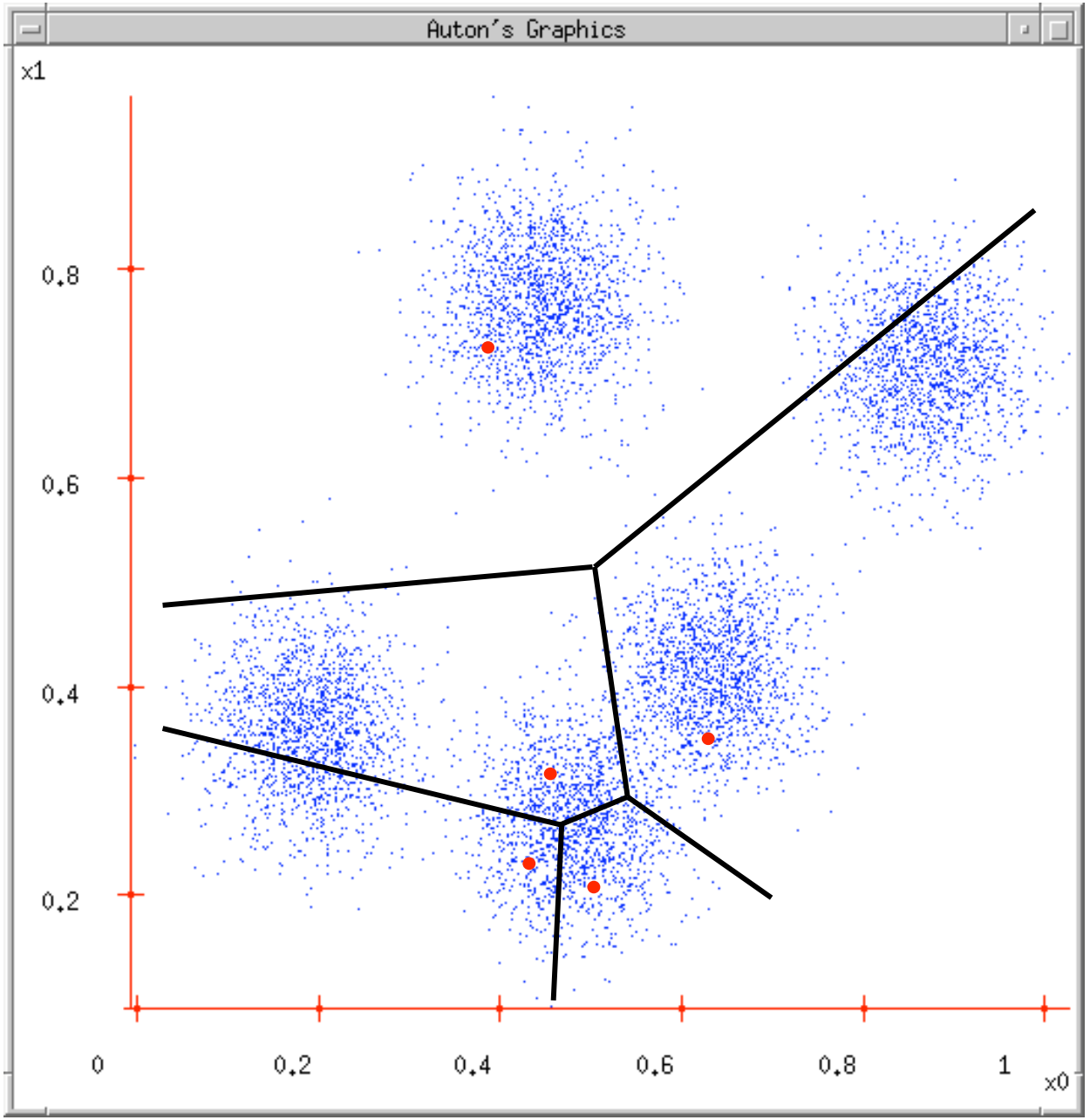
K-means

1. Ask user how many clusters they'd like. (e.g. *K*=5)
2. Randomly guess *K* cluster Center locations



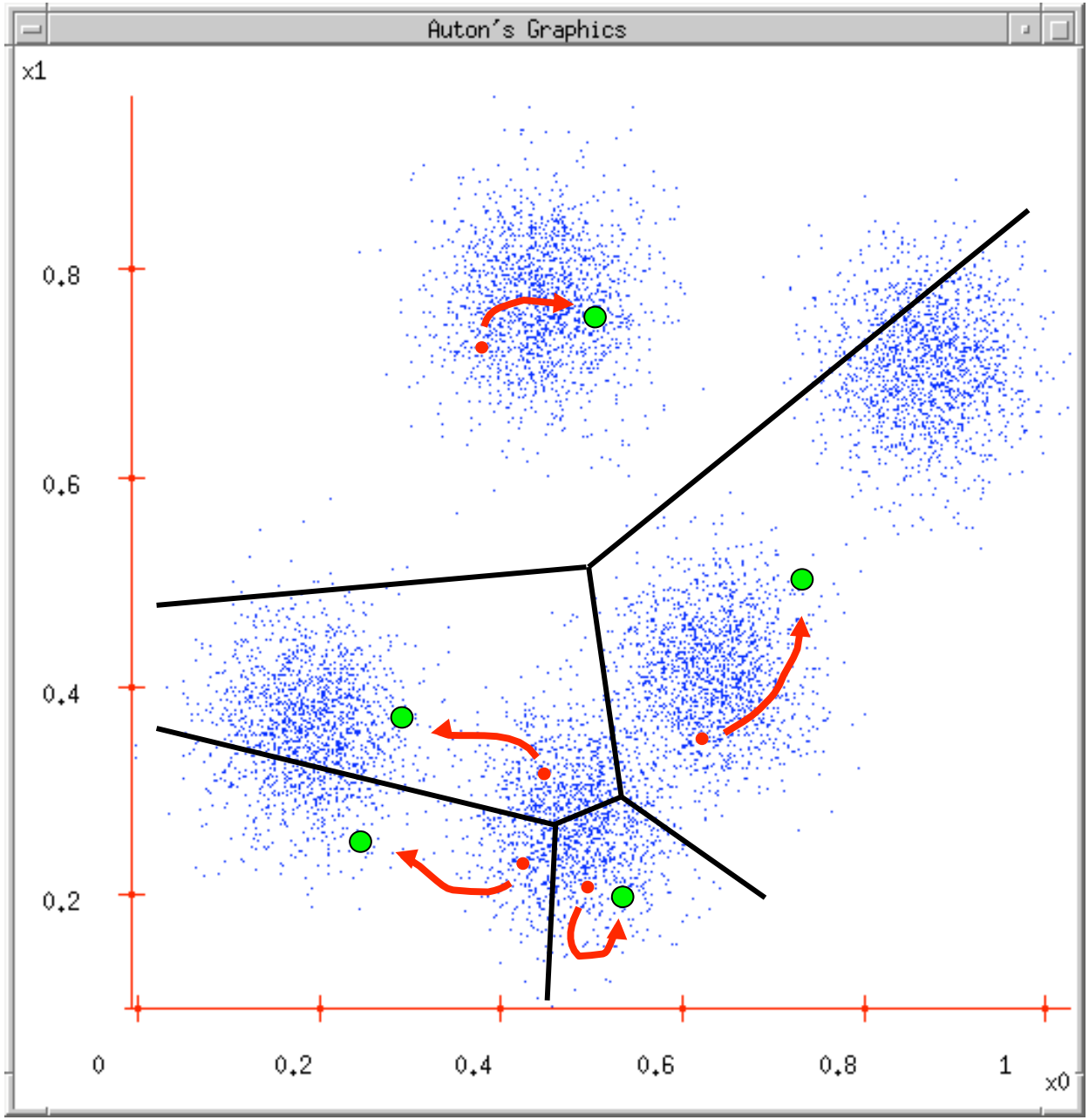
K-means

1. Ask user how many clusters they'd like. ***(e.g. k=5)***
2. Randomly guess *K* cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)



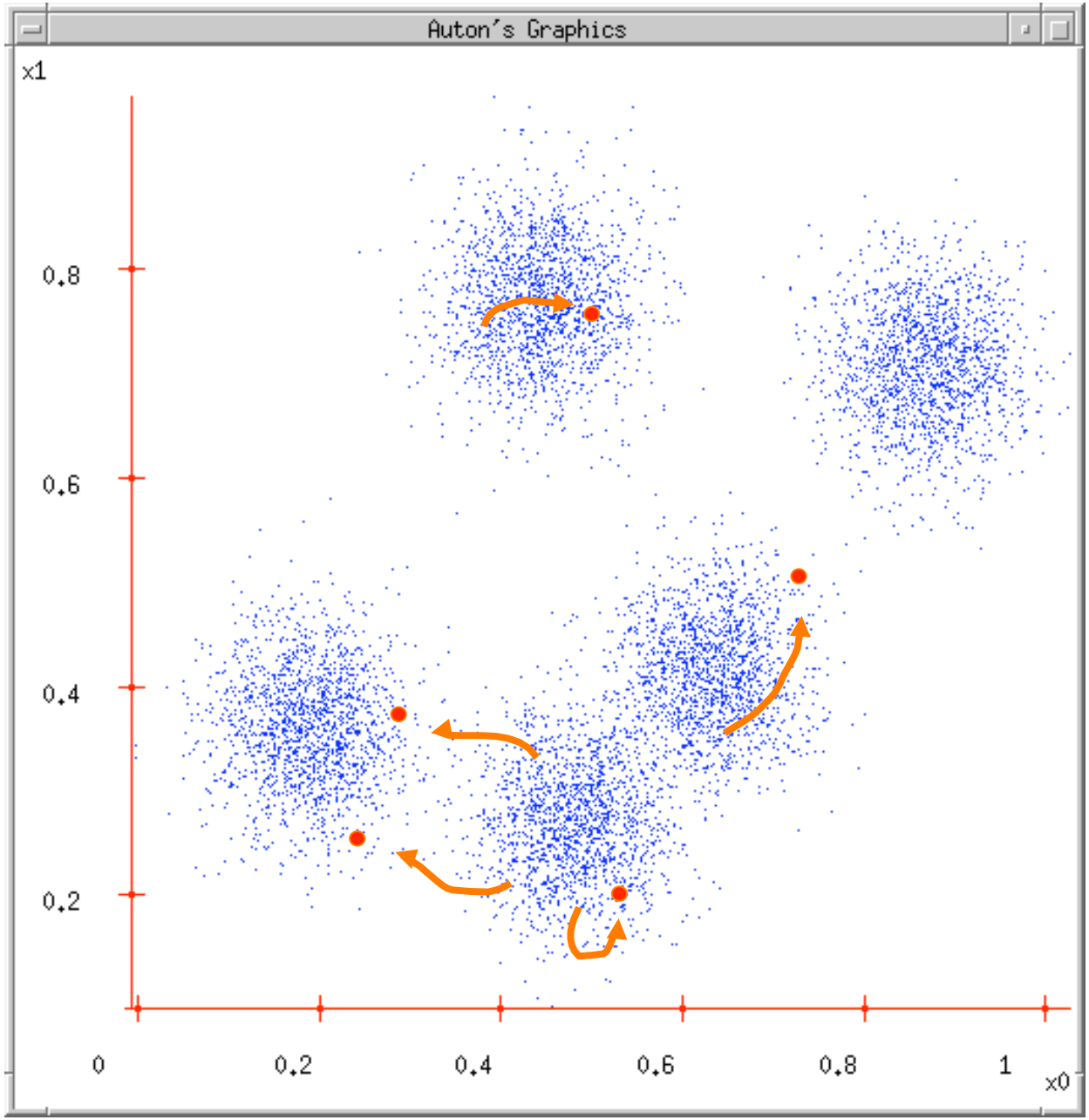
K-means

1. Ask user how many clusters they'd like. *(e.g. K=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns



K-means

1. Ask user how many clusters they'd like. *(e.g. K=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns…
5. …and jumps there
6. …Repeat until terminated!



Single particle methods that use *K*-means principle

(explicitly or not)

1. Generating of 2D class averages.
2. 2D alignment by classification (IMAGIC)
3. 2D multireference alignment
4. 3D multireference (multiparticle) alignment
5. Maximum-likelihood multireference alignment (2D & 3D)
6. Bootstrap with 3D PCA (Penczek)
7. Cross-correlation of common lines (Nogales)
8. …..

K-means properties

+ Very simple algorithm
+ Works very well if groups are well separated and number of groups *K* was guessed correctly
+ $O(KNt)$ time complexity
+ Guaranteed to converge in a finite number of steps
+ In the SSE version, optimizes well-defined and intuitive notion of "natural grouping" (i.e., within-group variance)

− Circular cluster shape only
− Not guaranteed to converge to a global minimum
− Finding global minimum not feasible in practice
− Outliers can have very negative impact
− If *K* not guessed correctly and/or groups are not well separated (i.e., almost always), the result dramatically depends on initialization.

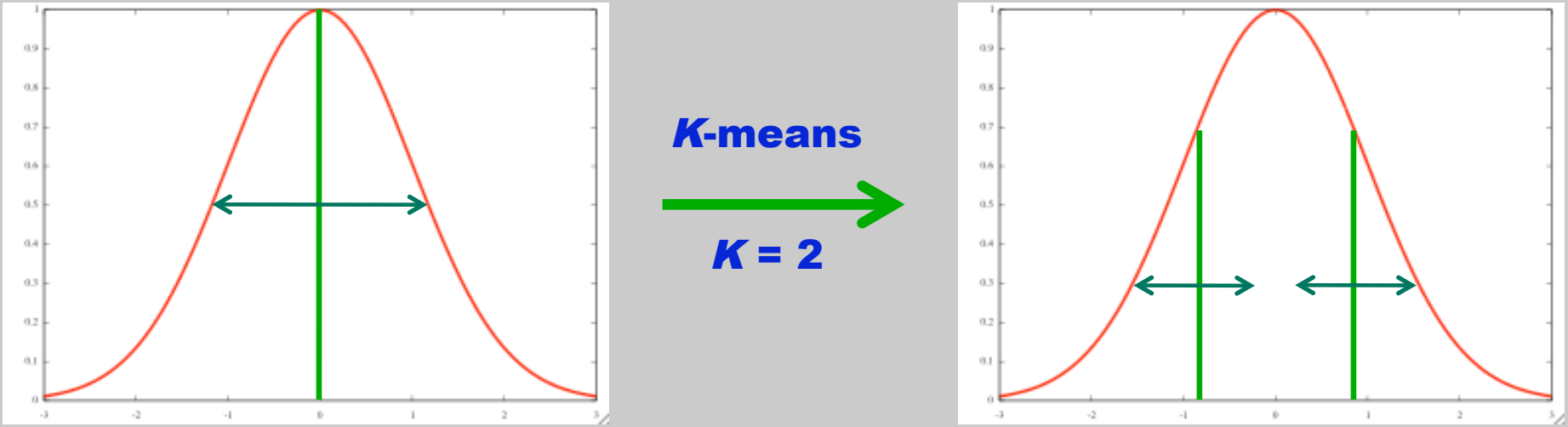
Consequences of *K*-means properties for cryo-EM

- Virtually all methods will work well if the conformers are *sufficiently different* and their number is known.
- None of the methods will work well if the conformational changes are *small*, number of conformers is not known, and/or instead of discrete states we face continuous conformational changes.
- *Small* subgroups are difficult if not impossible to detect.
- If the analysis is initialized using guessed conformers and a guess of their number, the results are *more than likely* to reproduce/confirm this guess.
- In the absence of a priori information about the number of conformers, it is *all but impossible* to decide whether obtained groups are “pure”.

Consequences of *K*-means properties for cryo-EM

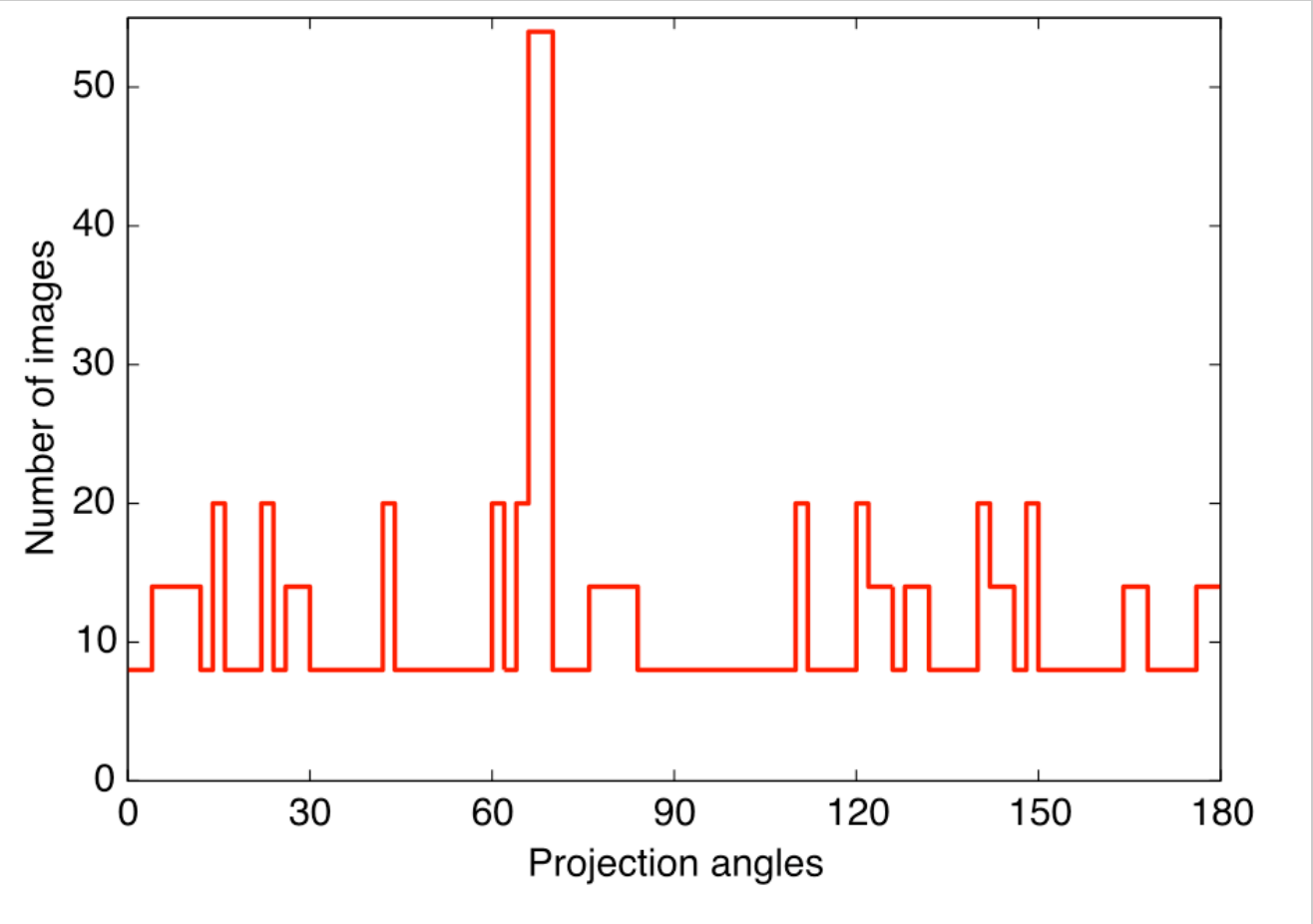
K-means will impose a partition on the data
(*if so determined, one is bound to find conformational changes*):

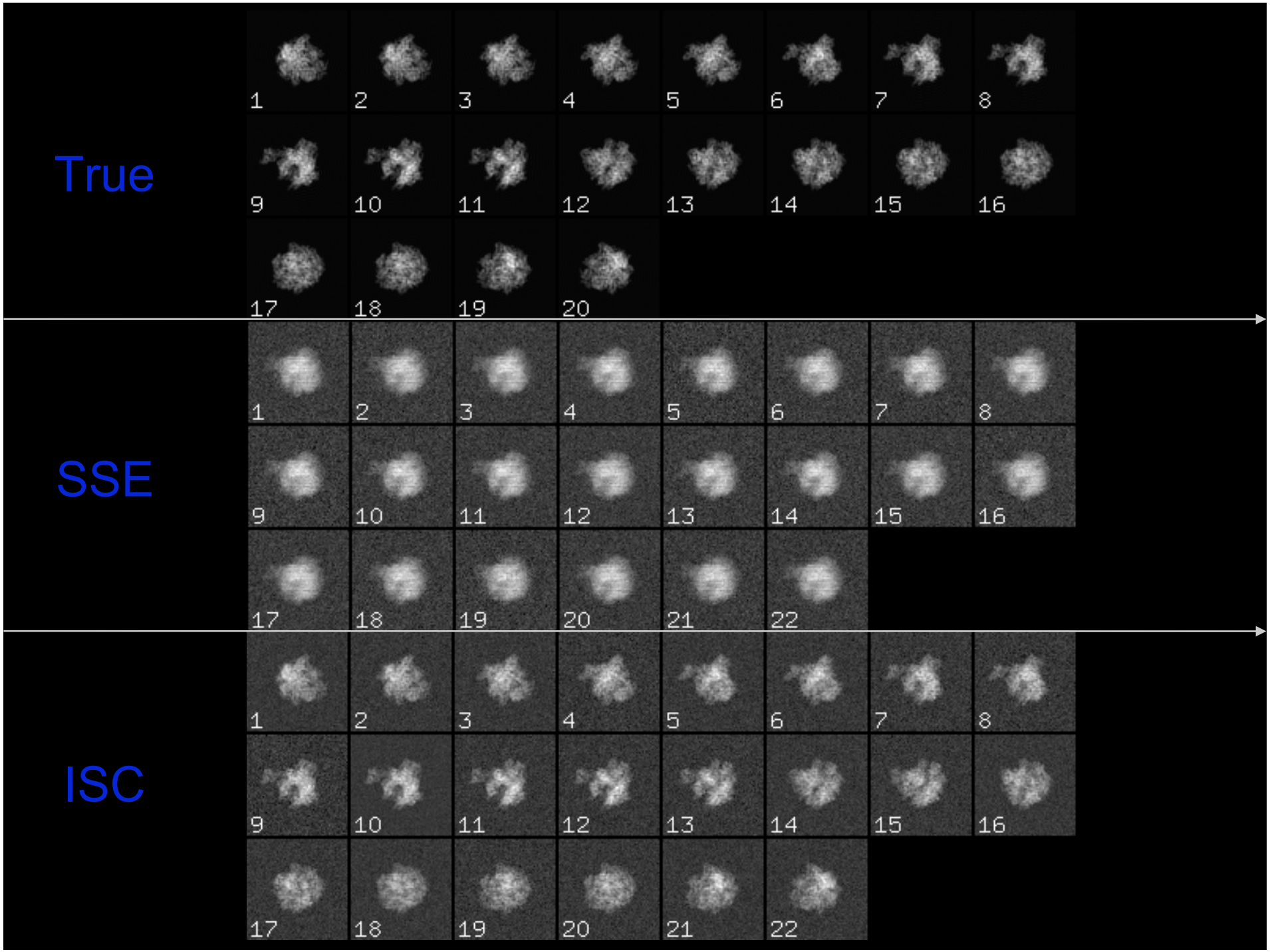
- ✓ *K*-means is not a statistical method, so it is impossible to say how reliable the results are
- ✓ if there are no clusters in the data, *K*-means will still return a solution and the resulting means (conformers) will appear to be significantly different.



We propose a new approach we call
Iterated Stable Clustering

Crystal structure of RNAP II (PDB entry 1NT9).
Test projections generated using single-axis tilt geometry:
The set of computationally generated projections comprised three subsets:
(1) a basic evenly distributed set,
(2) 19 subsets generated around randomly selected directions,
(3) one dominating group.
In the data set of 1,040 2D projection images no two were identical and one group dominated the remaining ones.




True
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
SSE
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
ISC
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

SPARX

Open source code developed in python with C++ libraries shared with EMAN2.
Novel algorithms for single particle reconstruction.
Parallelized for MPI.
Some code ported on GPUs.

Available at:
http://sparx-em.org/sparxwiki

EMAN2 (Steve Ludtke)
Single particle reconstruction, GUI interfaces, structure modeling:

http://ncmi.bcm.edu/ncmi

Acknowledgments

Wei Zhang,
U of Texas, Houston

Christian M.T. Spahn,
Charité, Berlin

Marek Kimmel,
Rice University, Houston

Francisco Asturias,
Scripps Institute, La Jolla

NIH