IDENTIFICATION OF REPRODUCIBLE 3-D STRUCTURES IN HETEROGENOUS CRYO-EM DATA SETS

Zhong Huang, Pawel Penczek

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

UTHealth
The University of Texas
Health Science Center at Houston

McGovern
Medical School

Heterogeneity and single particle analysis

- Assumption of single particle analysis: Specimen has structural 'integrity', so all particles can be treated as copies of (in principle) the same structure (JF, 1996)
- In practice, all EM datasets have a degree of structural heterogeneity.
- Causes/types of protein heterogeneity:
  - mixture of different types of proteins
  - substoichiometric ligand binding
  - multiple conformations/ functional scale (large scale - open/close states, subunit rearrangement)
  - flexibility of protein
  - fluctuations of the structure around the ground state

General principle of clustering

Cluster analysis is a poorly defined 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

ß

- Assign *n* object to *K* classes such that an overall mathematical criterion is optimized:

  SSE - Sum of Squared Errors/Sum of Within Class Variances

SSE *K*-means

how to arrive at optimum partition?

Compute SSE for all possible partitions and select partition with the smallest SSE.

There are approximately $K^n / K!$ possibilities in which *n* objects can be partitioned among *K* classes.

for *K*=5; *n*=10,000 it is ~10^{6974}

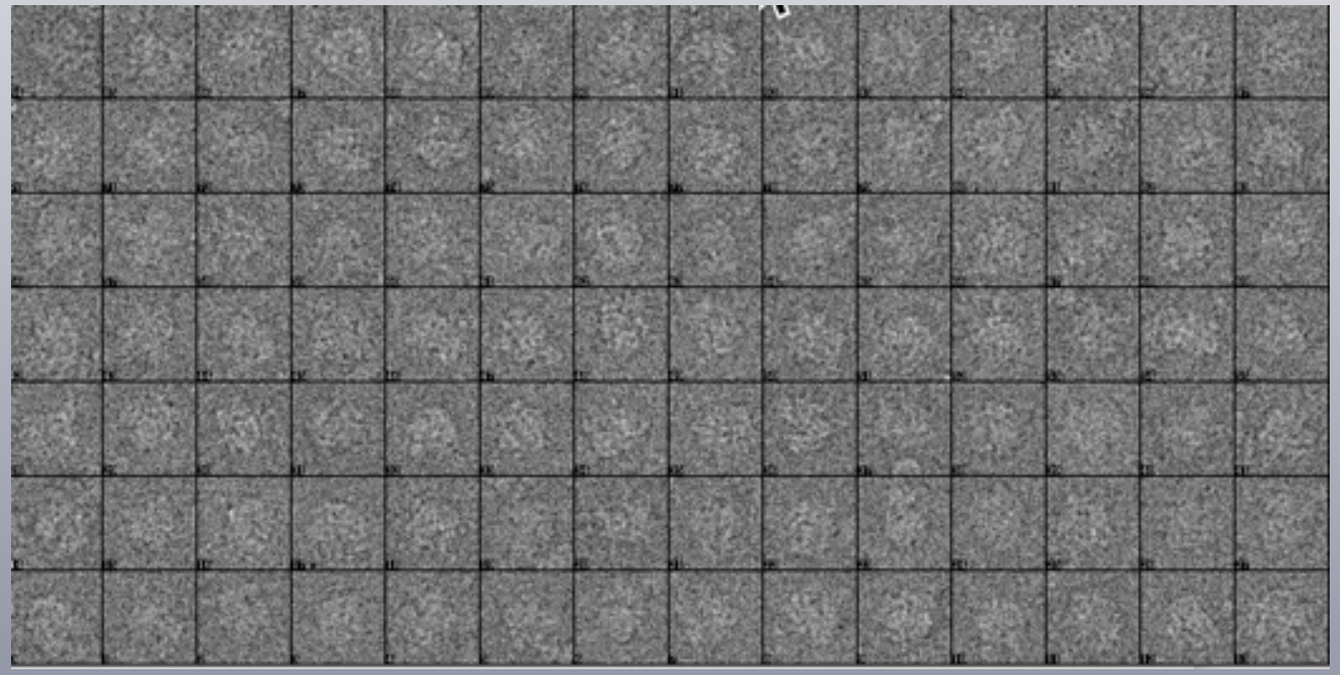
Current approaches: multi-reference refinement (MRR)

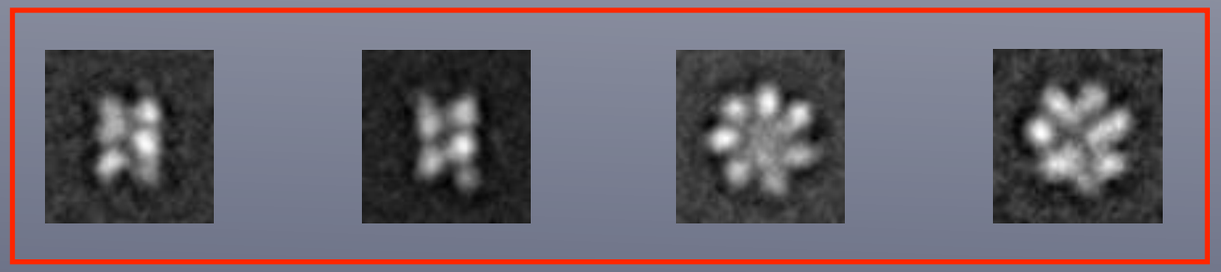
MRR is equivalent to *K*-means clustering, with the distance between images defined as a maximum similarity over the permissible range of image rotations and translations.

K-means results also depend on another nontrivial problem: the 3-D alignment of 2-D images.

Because the 2-D images are all aligned to one reference volume, the adjustment of 3-D parameters might subtly affect the 3-D structures of subsets

n images



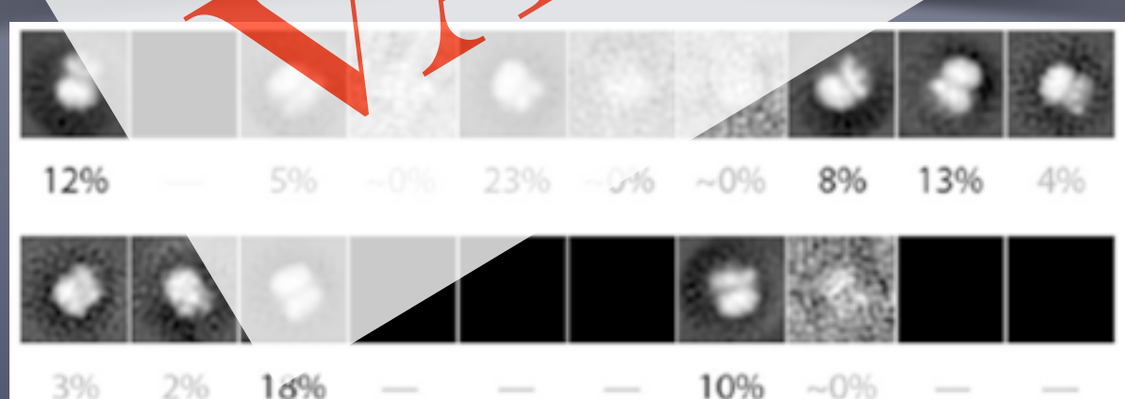


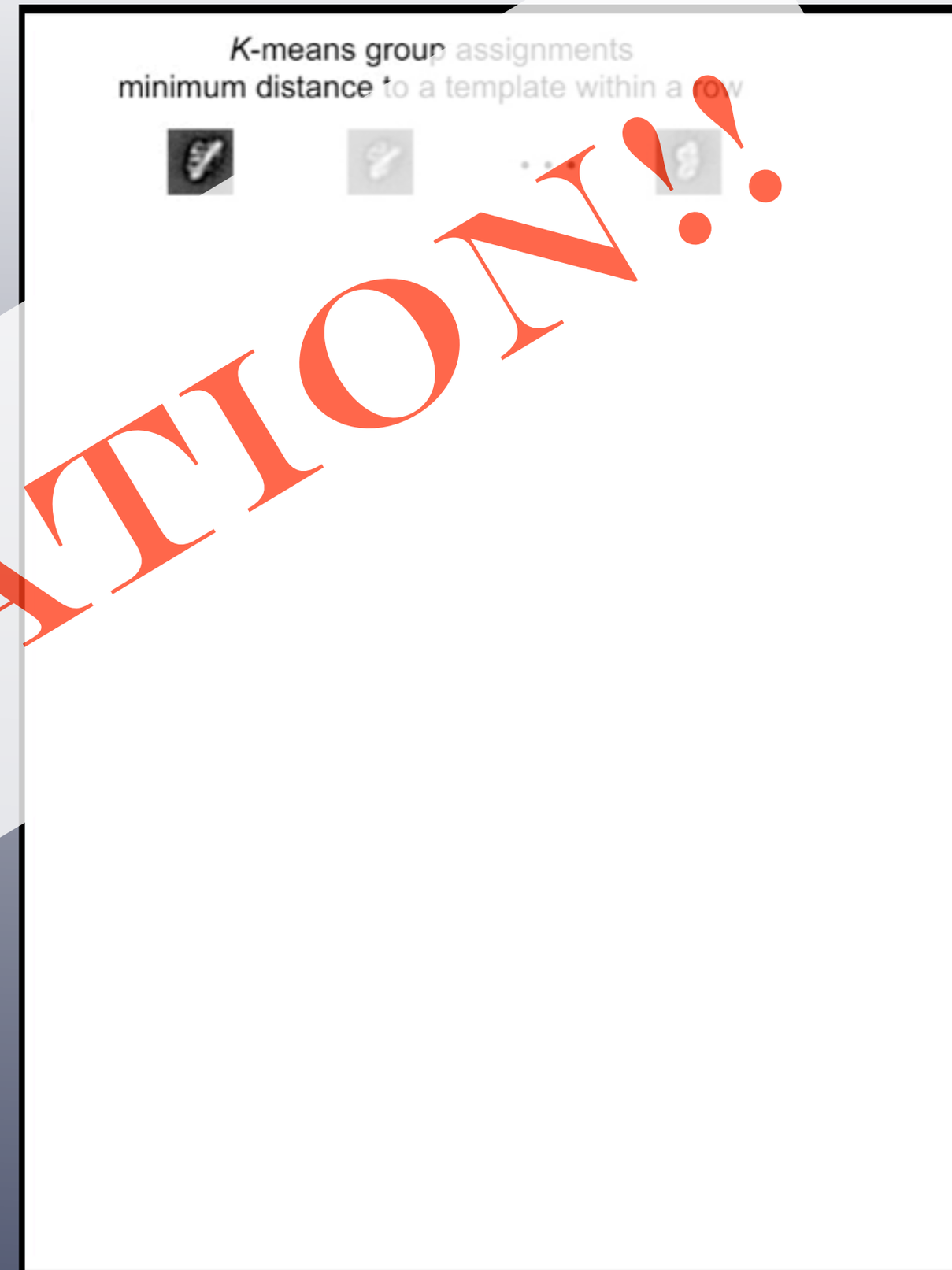
K averages (clusters)

K-means clustering

KNOWN PROPERTIES:

- Very fast convergence guaranteed in a finite number of steps
- Converges only to a local minimum
- Unclear how to determine the appropriate number of classes (*K*)
- The solution (final structure) depends on the initial set of particle images, and will change if clustering is repeated using different initializations.





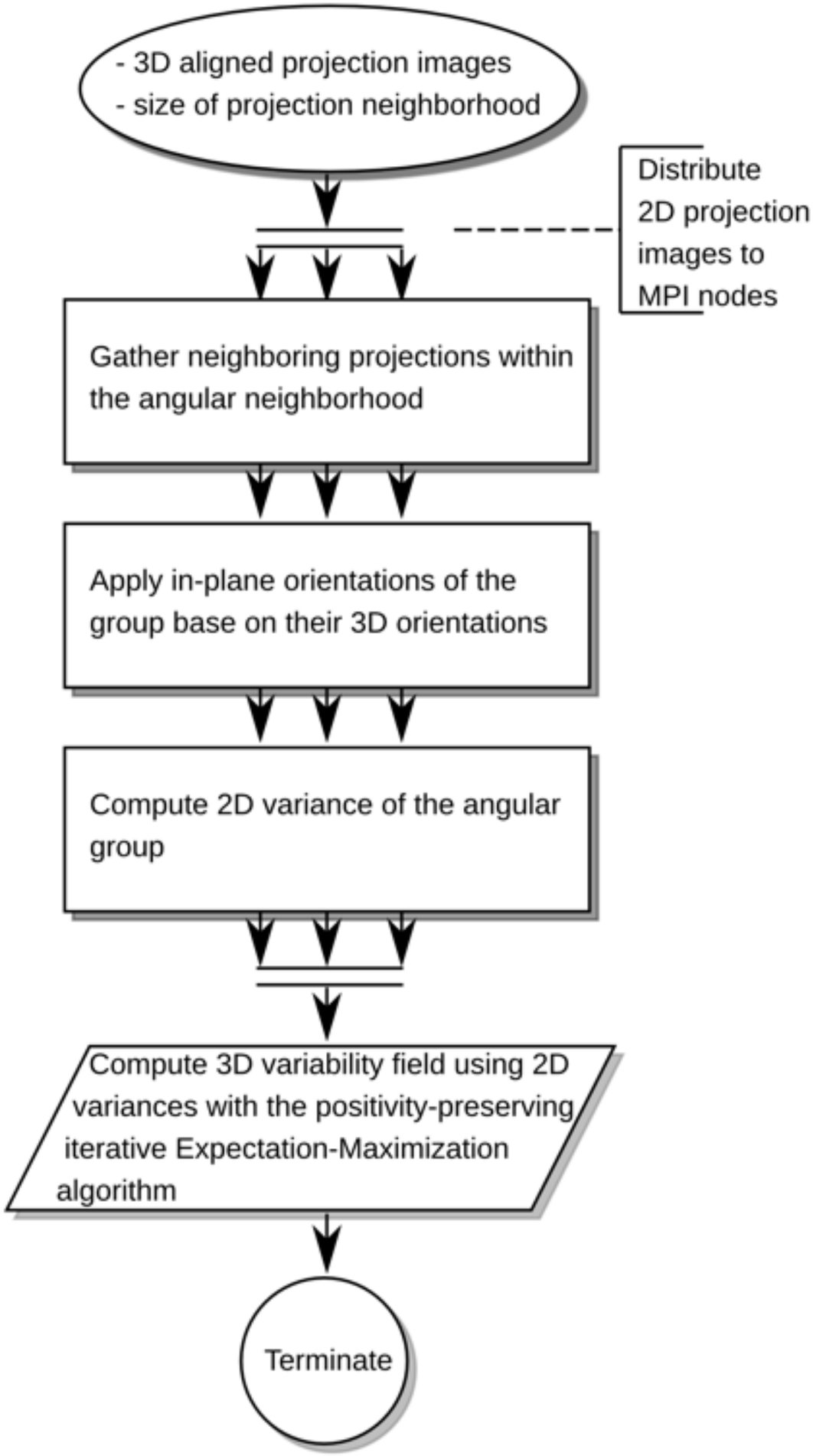
VALID ATTENTION!!

3D sorting is guided/validated by the analysis of 3D variability

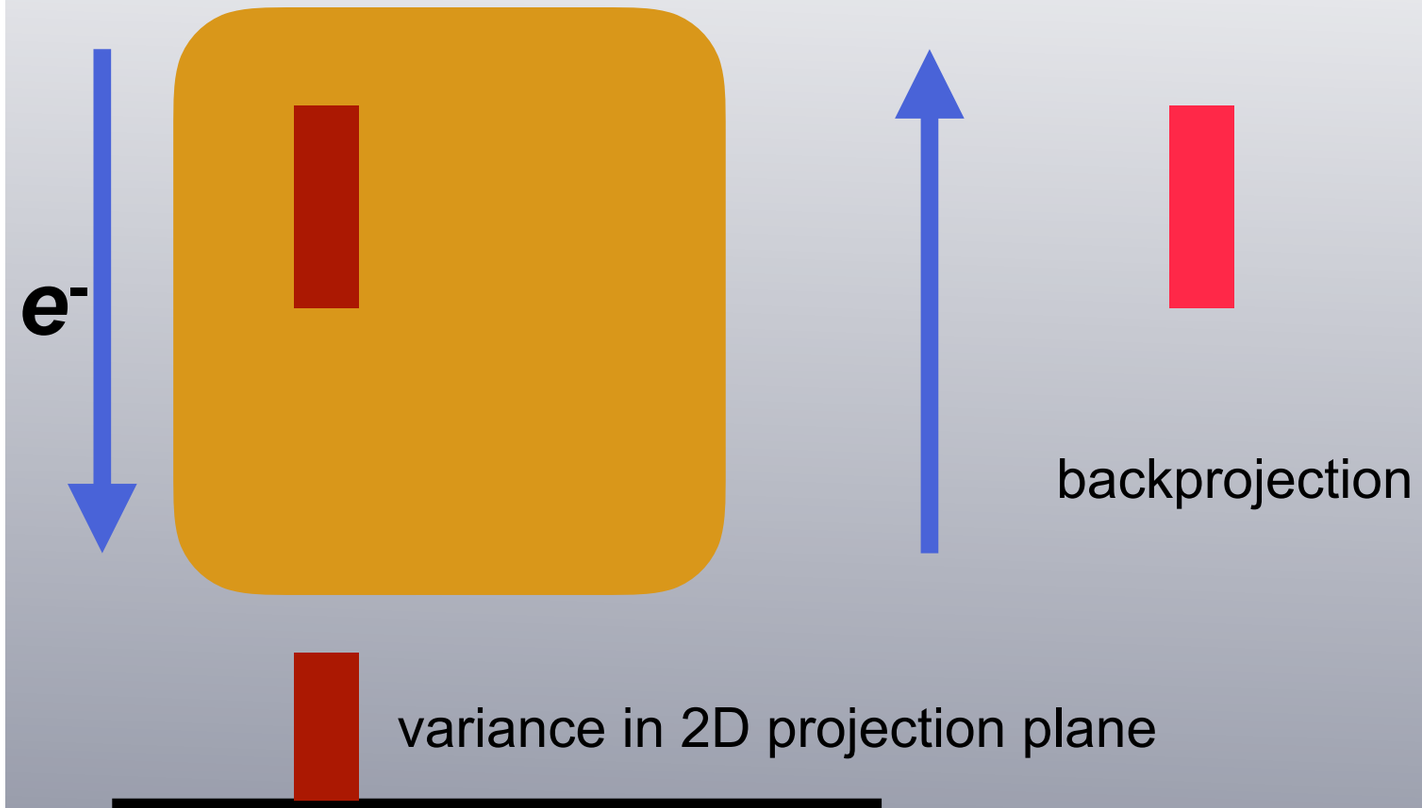
Two dedicated methods implemented in SPARX:

- 3D variability
  http://sparx-em.org/sparxwiki/sx3dvariability

- 3D local resolution
  http://sparx-em.org/sparxwiki/sxlocres



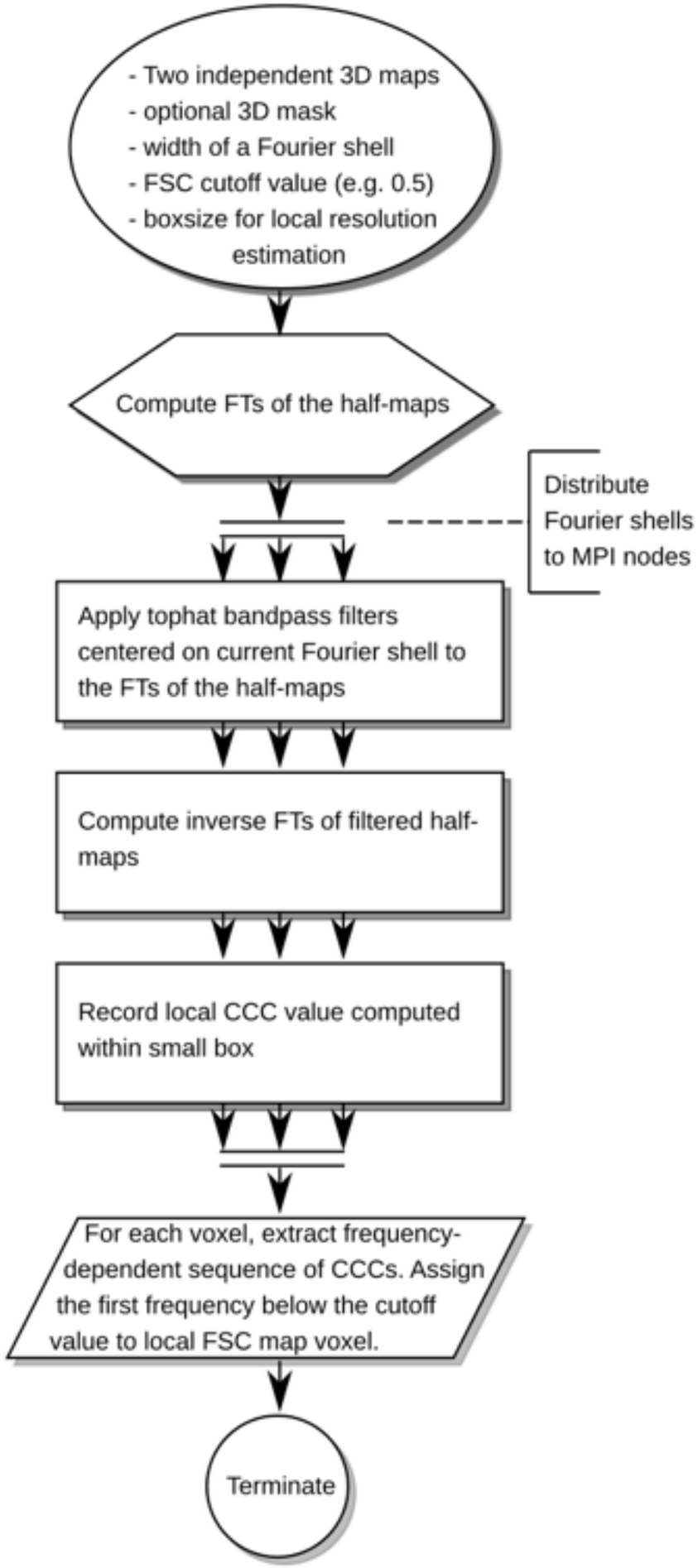
3D variability



$$\sigma_{2D}^2 = \sum_k \sigma_k^2 + \sum_k \sum_{l \neq k} cov_{kl}$$

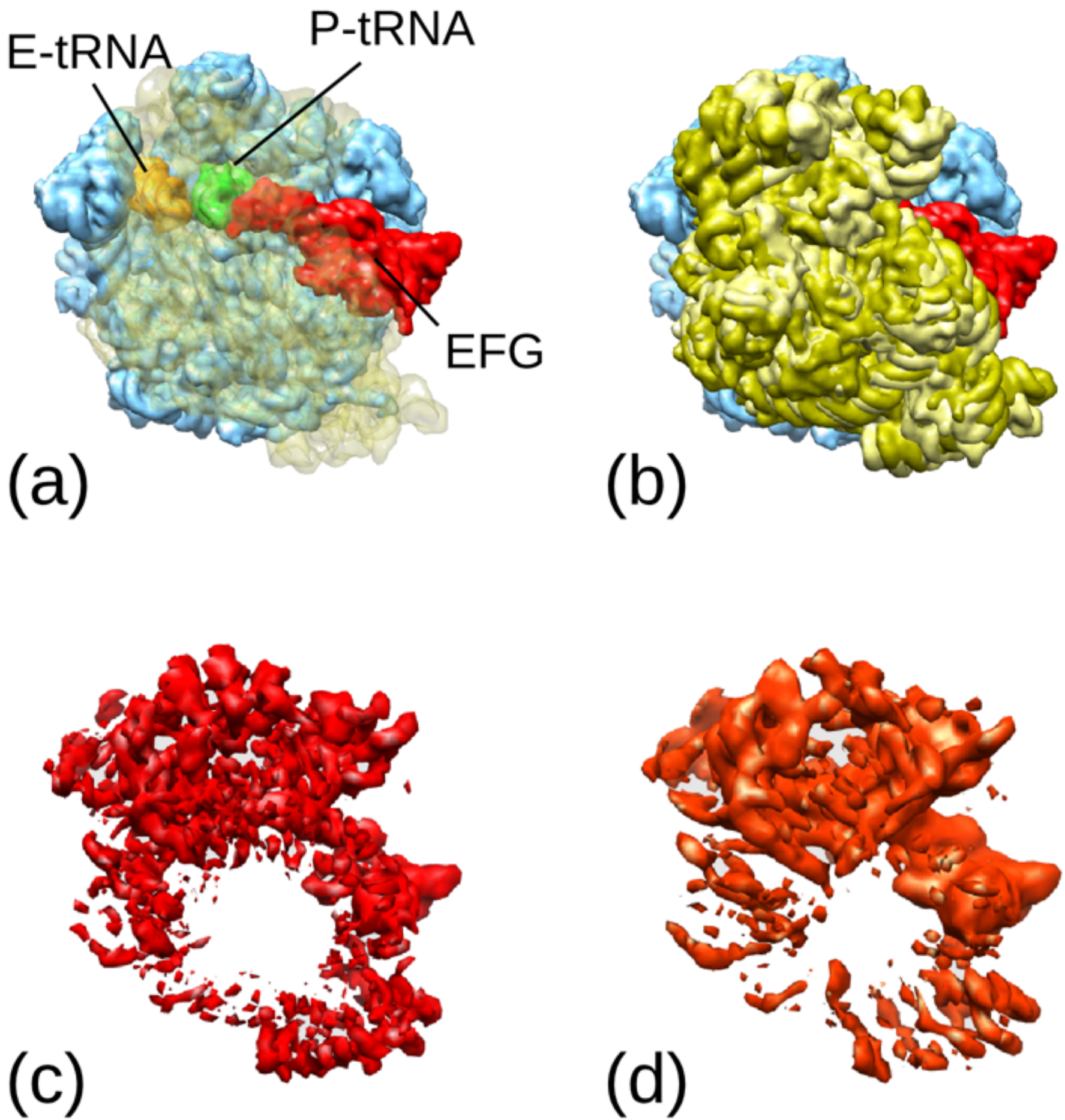
- 2D variance can be zero even if 3D variance is non-zero
- For substoichiometric binding of large ligands 2D variance is increased by positive covariances

3D local resolution computed as proper FSC



Reciprocal space	Real space
FSC	*ccc*
localized in frequency	spatially localized

Simulated data example



Five states

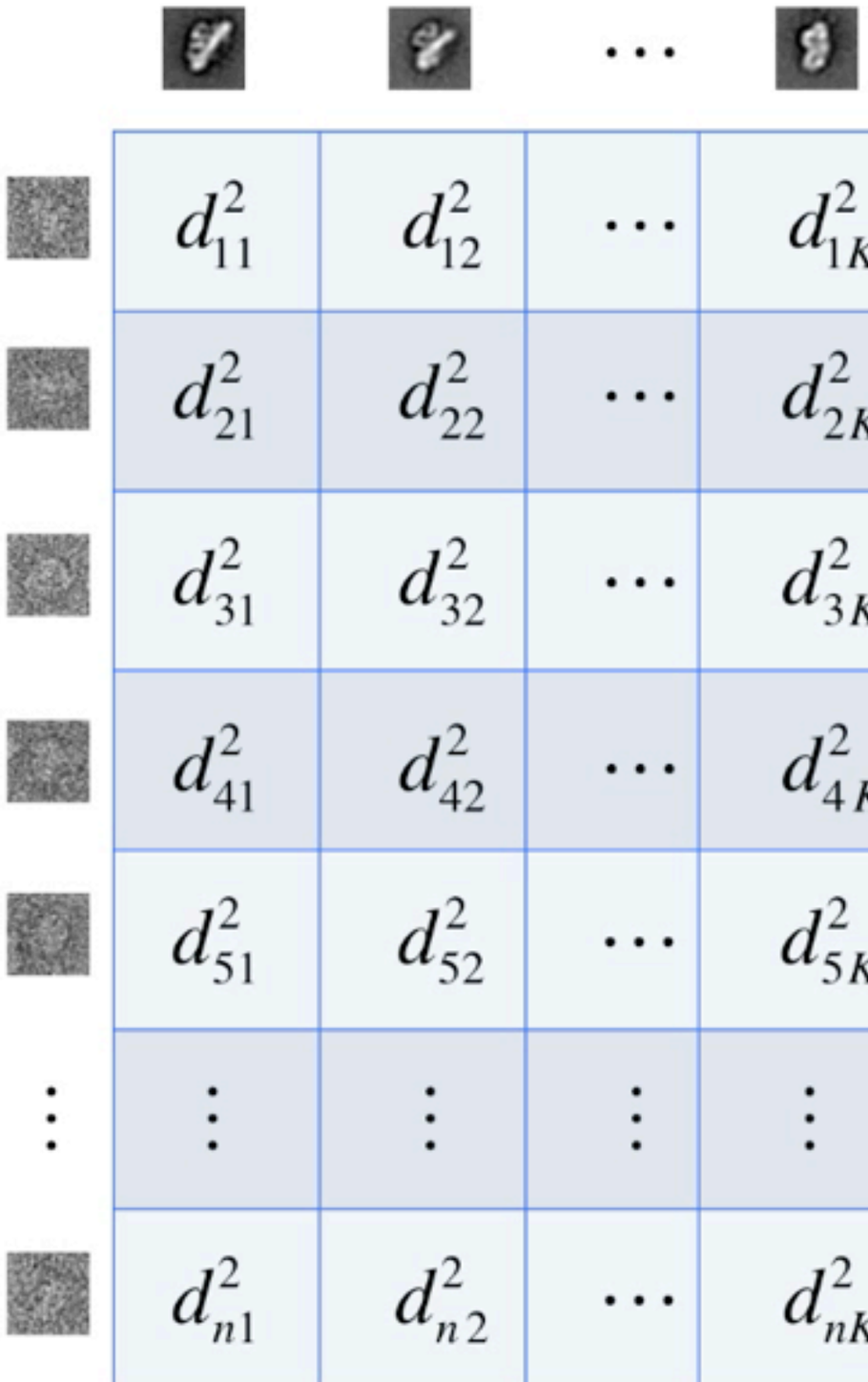
EQK(EQUAL GROUP SIZE)- means clustering-converge-under-control

Assign *n* images to *K* classes

such that each class contains

$\frac{n}{K}$ images

EQK-means group assignments
minimum distance to all templates, maximum number per group=3

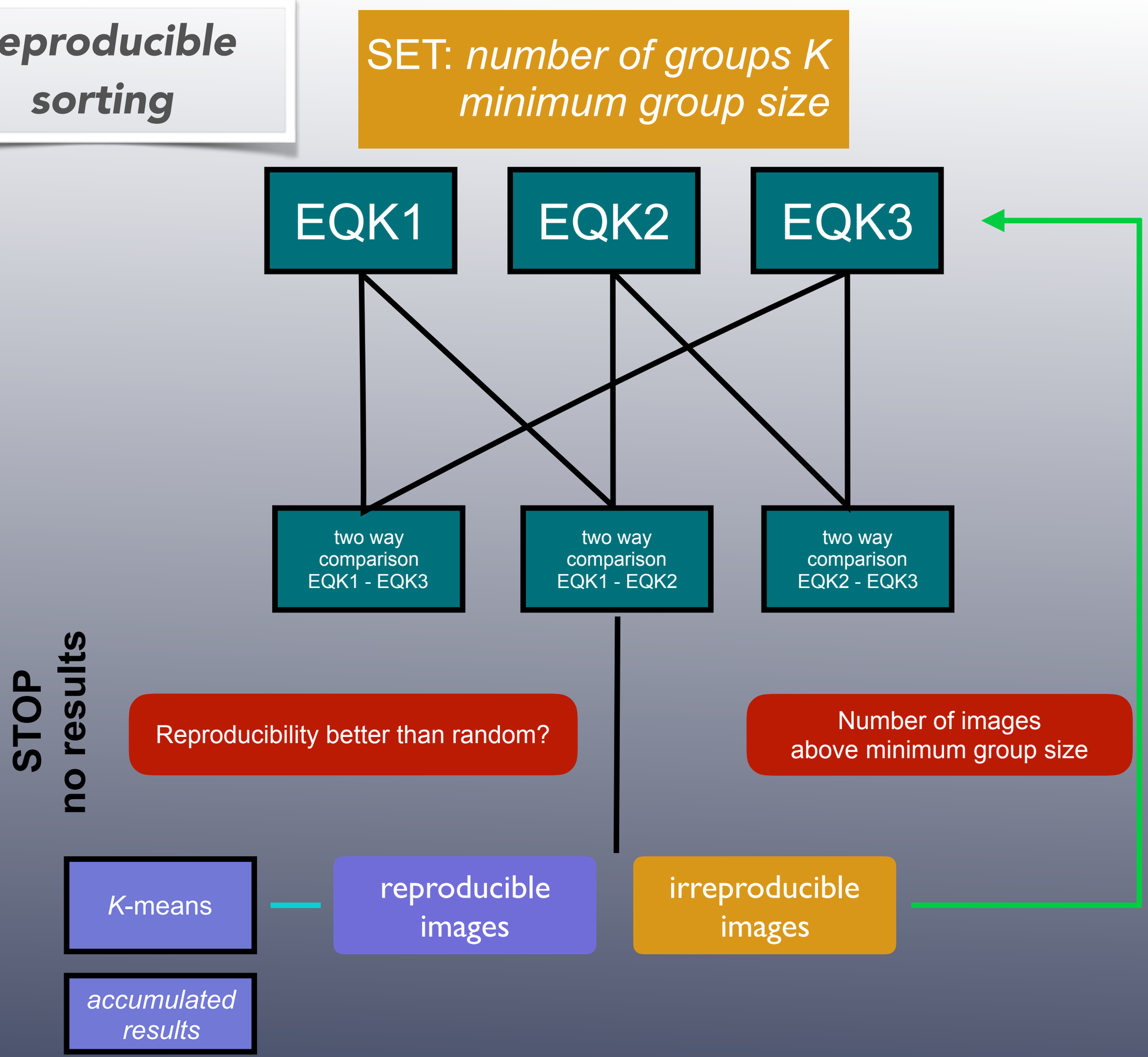


Multiple assignment problem

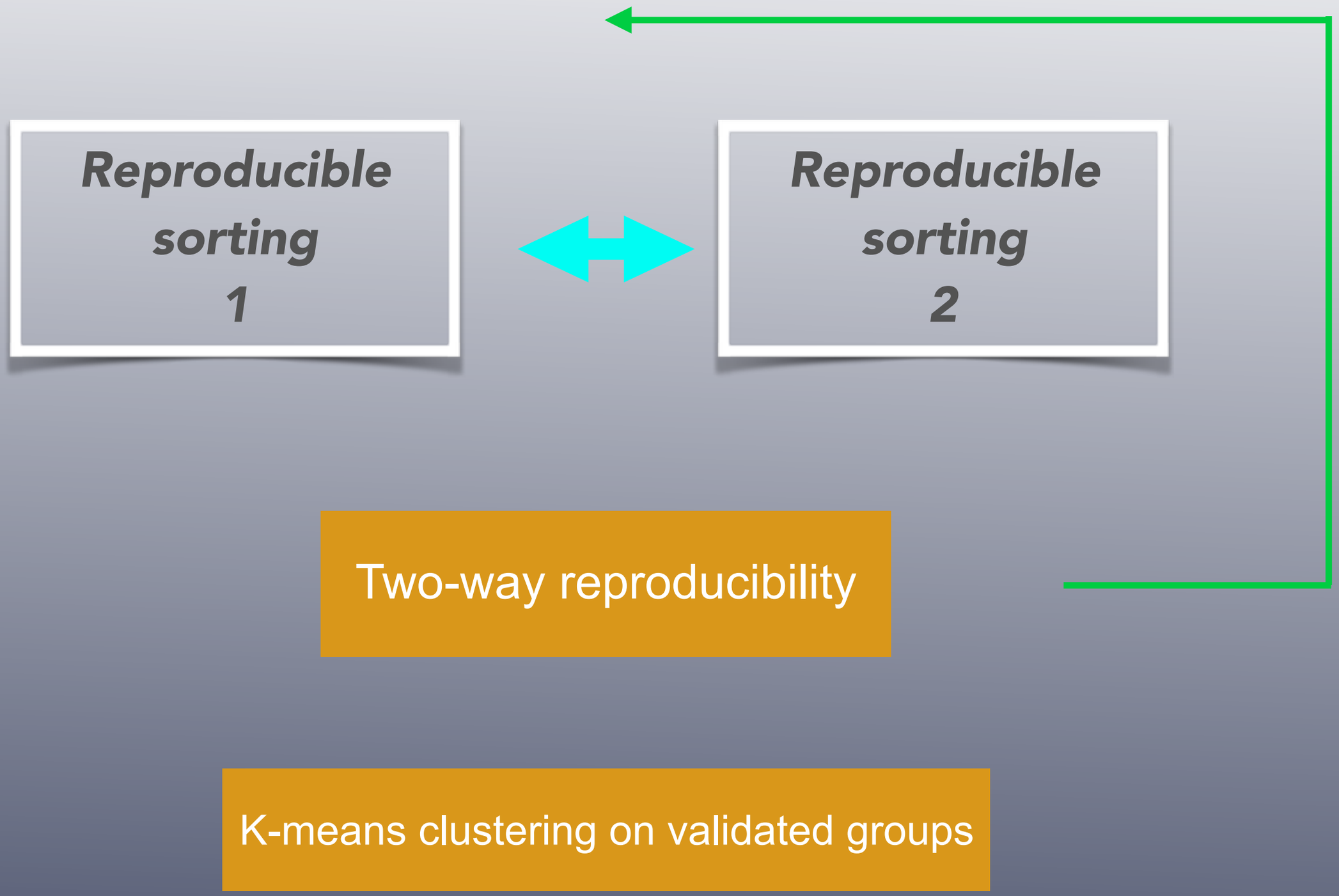
Second clustering \ First clustering	{1,3,7}	{6,8,9}	{2,4,5}
{1,4,5}	1	0	2
{3,7,8}	2	1	0
{2,6,9}	0	2	1

- For matching two sets of assignments, the solution is given in polynomial time by the Hungarian algorithm

Reproducible sorting

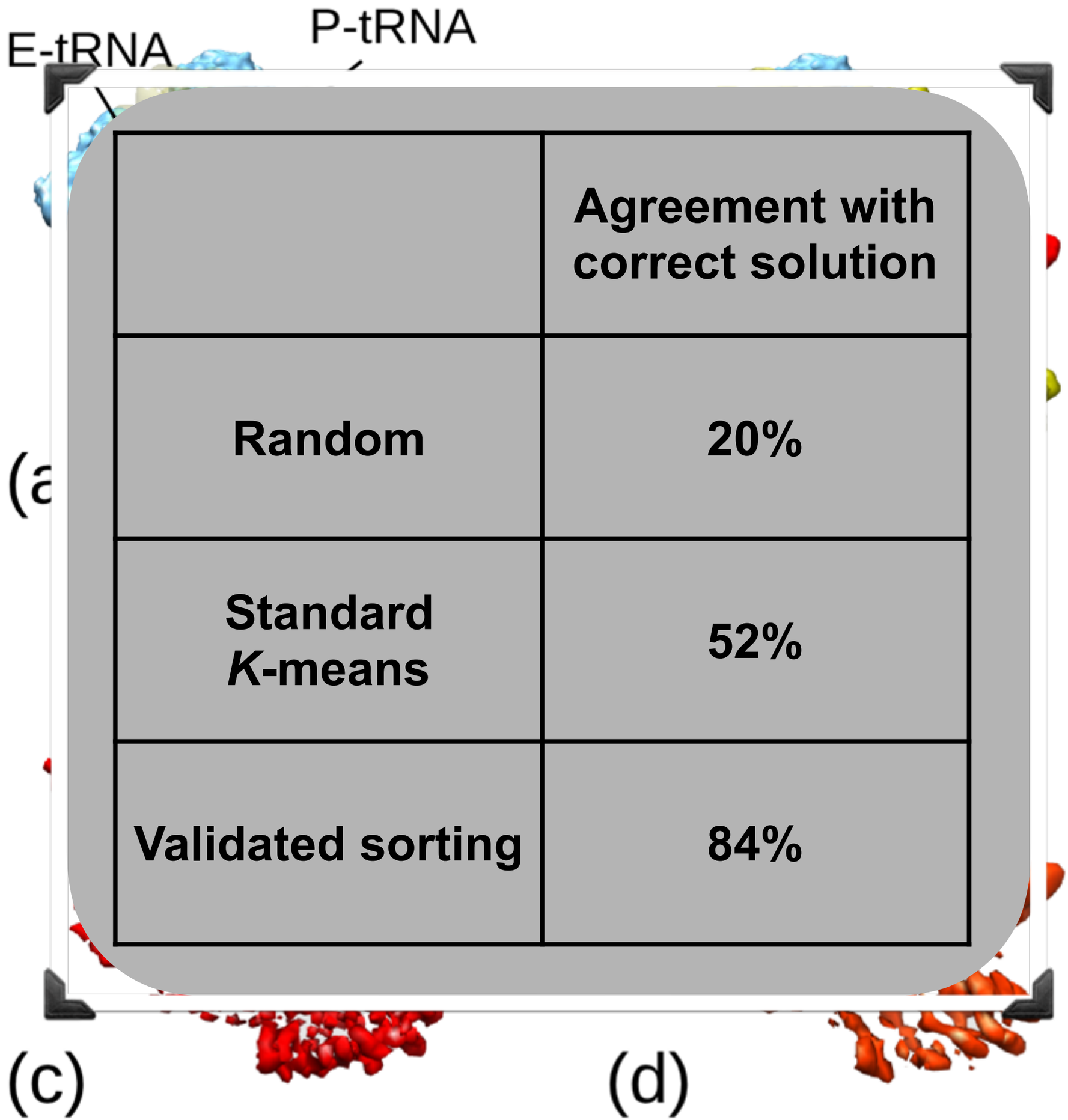


Sorting, phase two: validation



Simulated data example

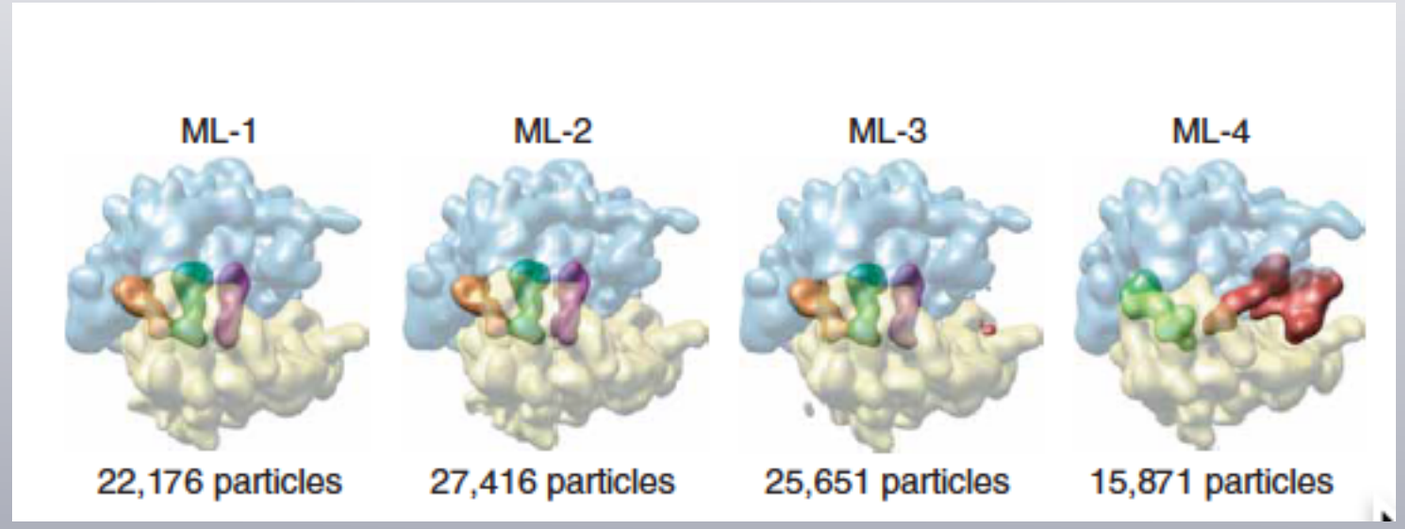
Five states



	Agreement with correct solution
Random	20%
Standard *K*-means	52%
Validated sorting	84%

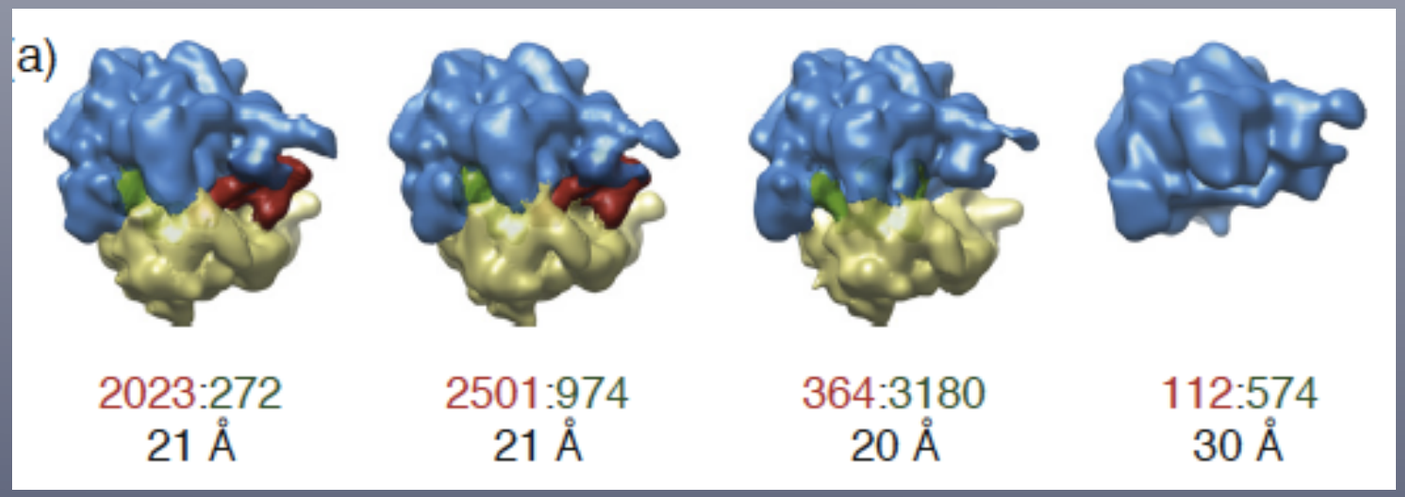
Analysis of an experimental cryoEM ribosome data set.

Scheres, S.H., Gao, H., Valle, M., Herman, G.T., Eggermont, P.P., Frank, J., Carazo, J.M., **2007**.
Disentangling conformational states of macromolecules in 3D-EM through likelihood optimization. Nature Methods 4, 27-29.



Subset reprocessed.

Scheres SH, **2012**. A Bayesian view on cryo-EM structure determination. J Mol Biol 415, p.406.

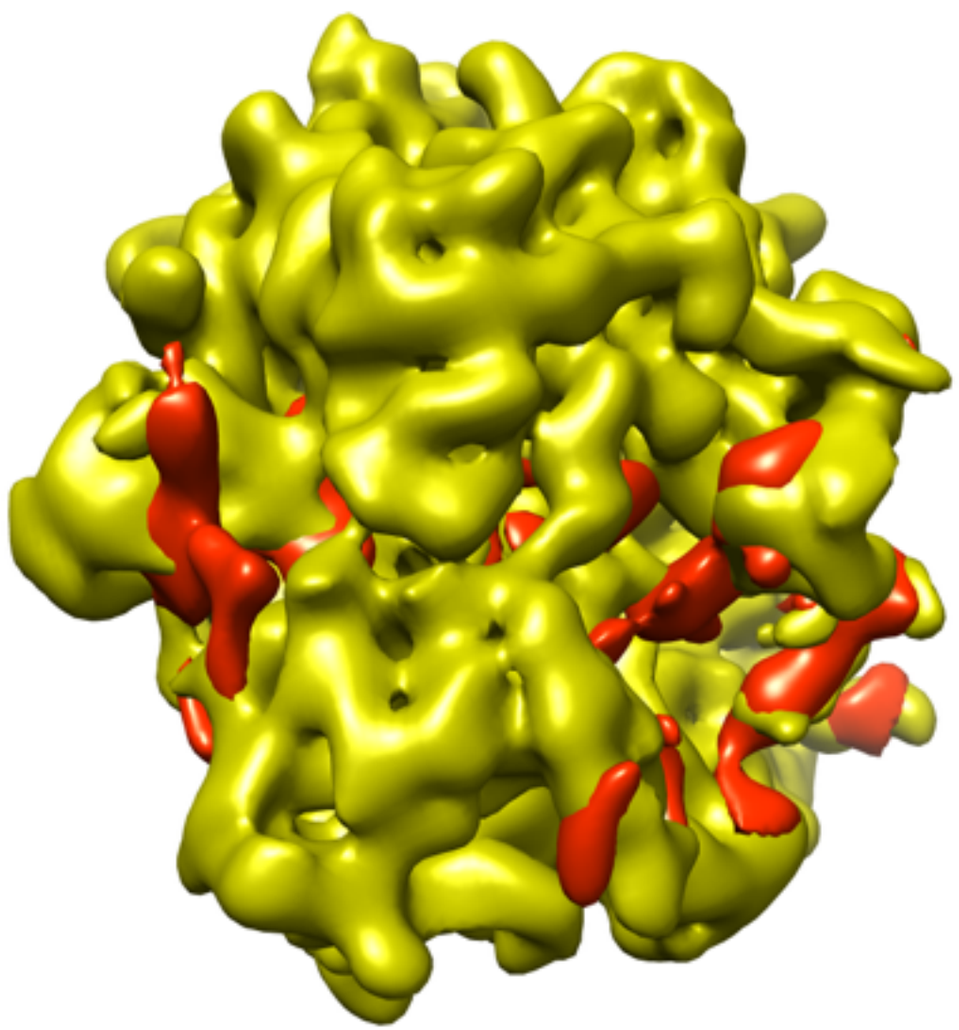


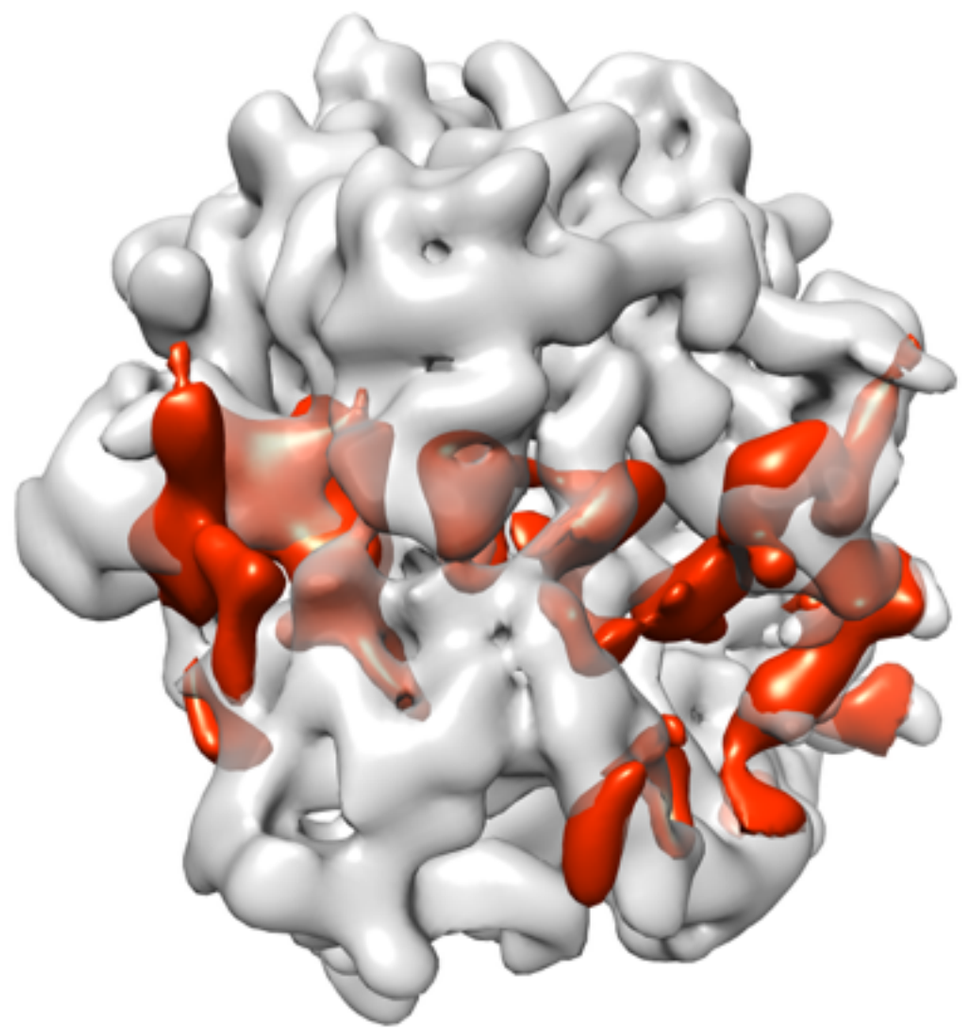
racheted
EF-G
E-site tRNA.

unratcheted
no EF-G
three tRNAs.

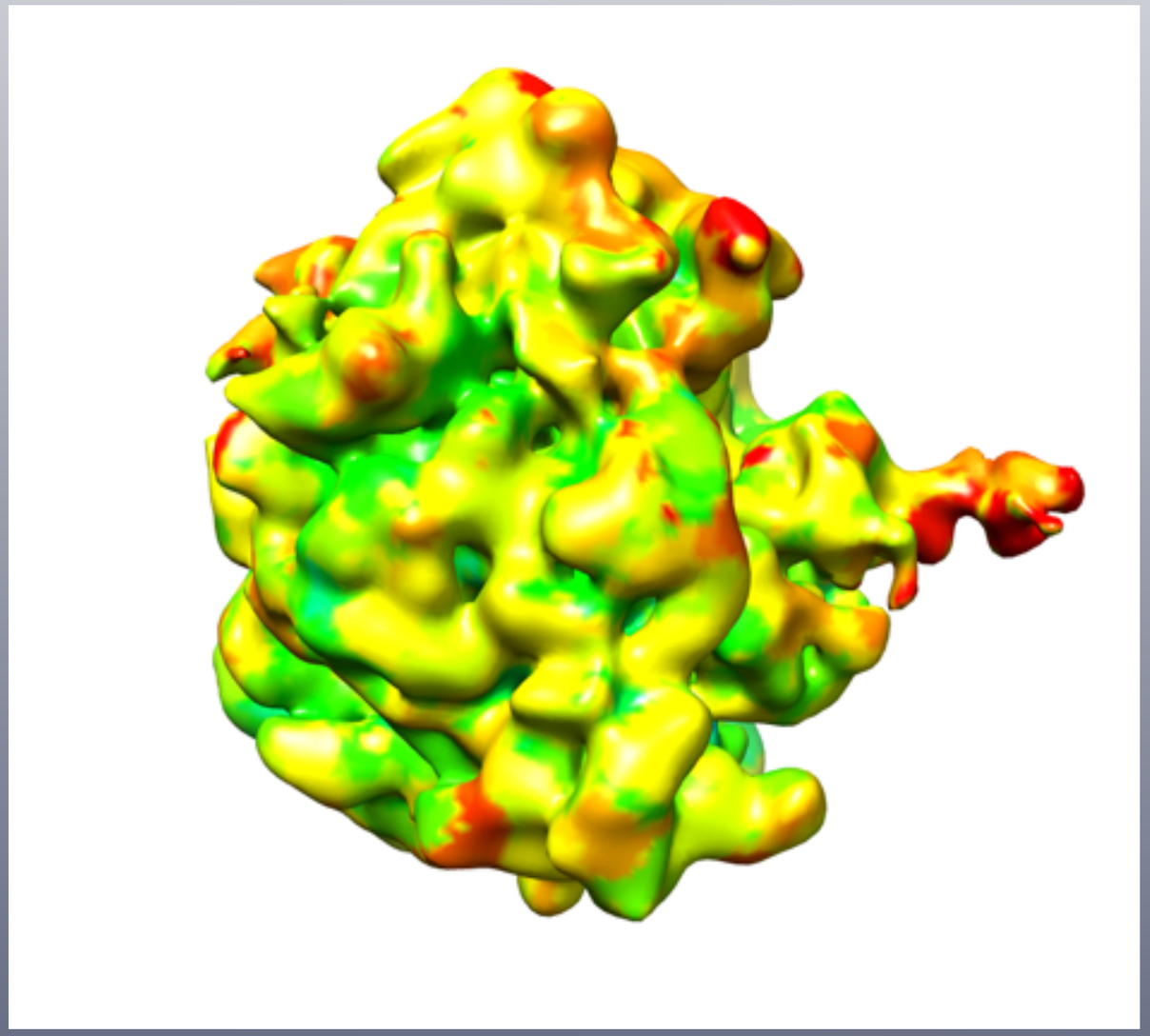
Processing in SPARX with validation

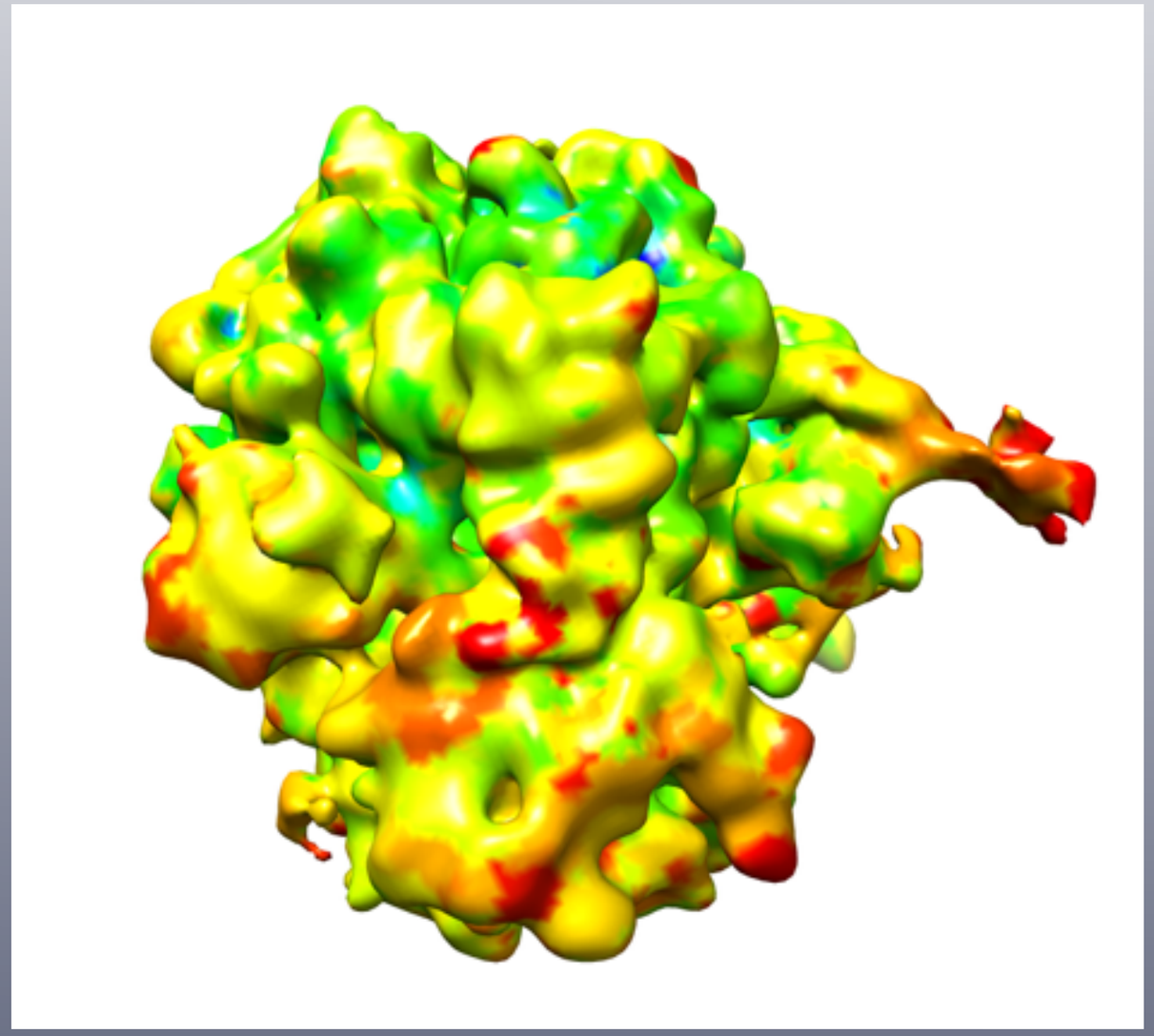
- 10,000 particles
- image size 128x128
- pixel size 2.8 Angstrom/pixel
- Refined to 9.6Å @0.5-FSC.





Local resolution (FSC-based)

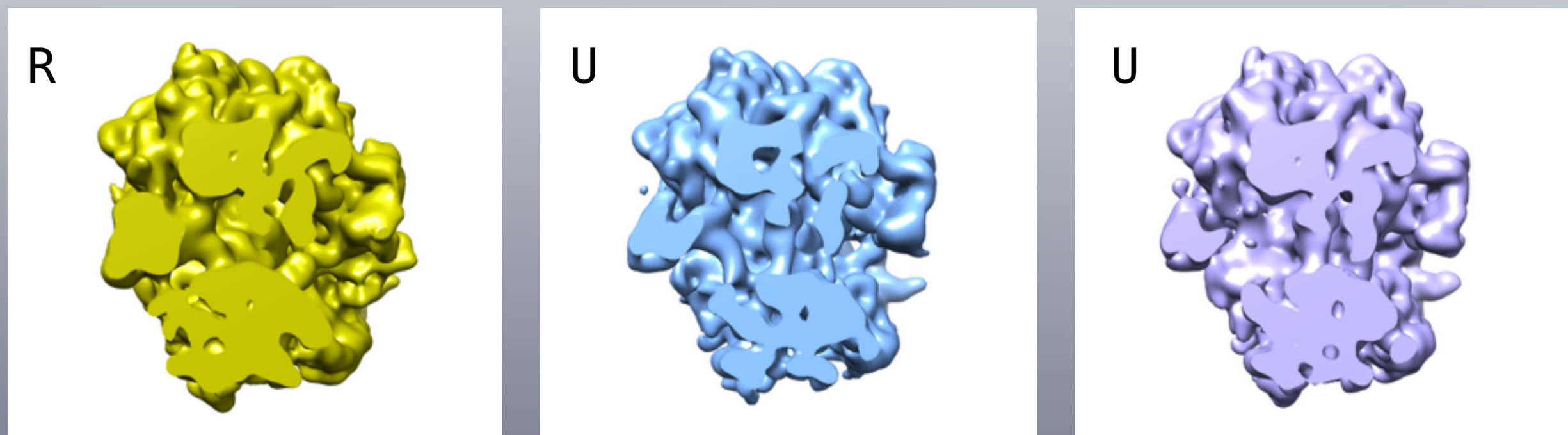




Sorting results

number of images: 10001; window size: 64
requested number of images per group: 2500
minimum group size: 100

Unaccounted images:	**7%**
Reproducibility:	**93%**



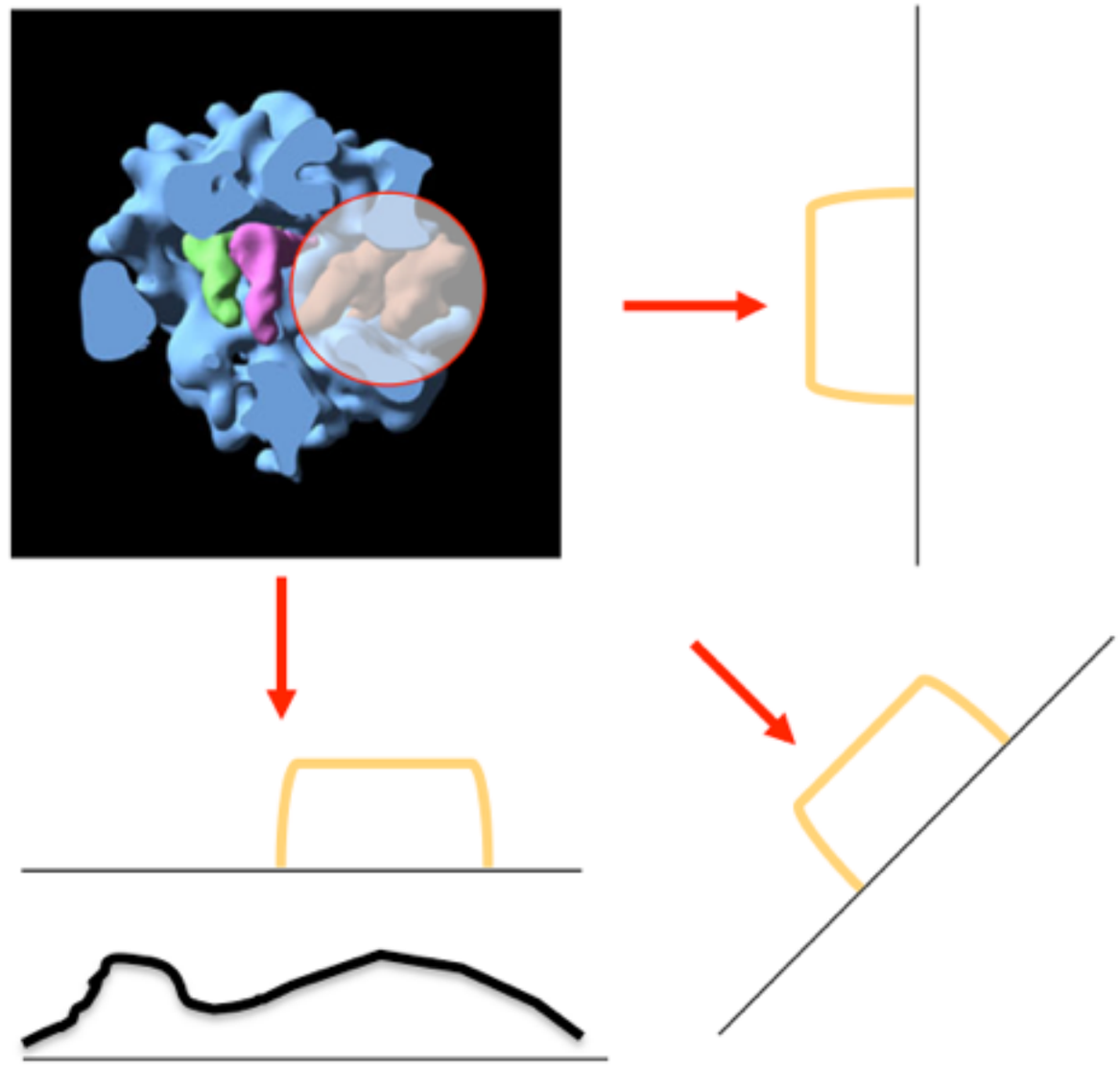
Ratcheted	Unratcheted		50S
E, EFG	E, P	P, A	not shown
43% (3 groups)	12%	25%	6%

Principle of focused classification

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**.

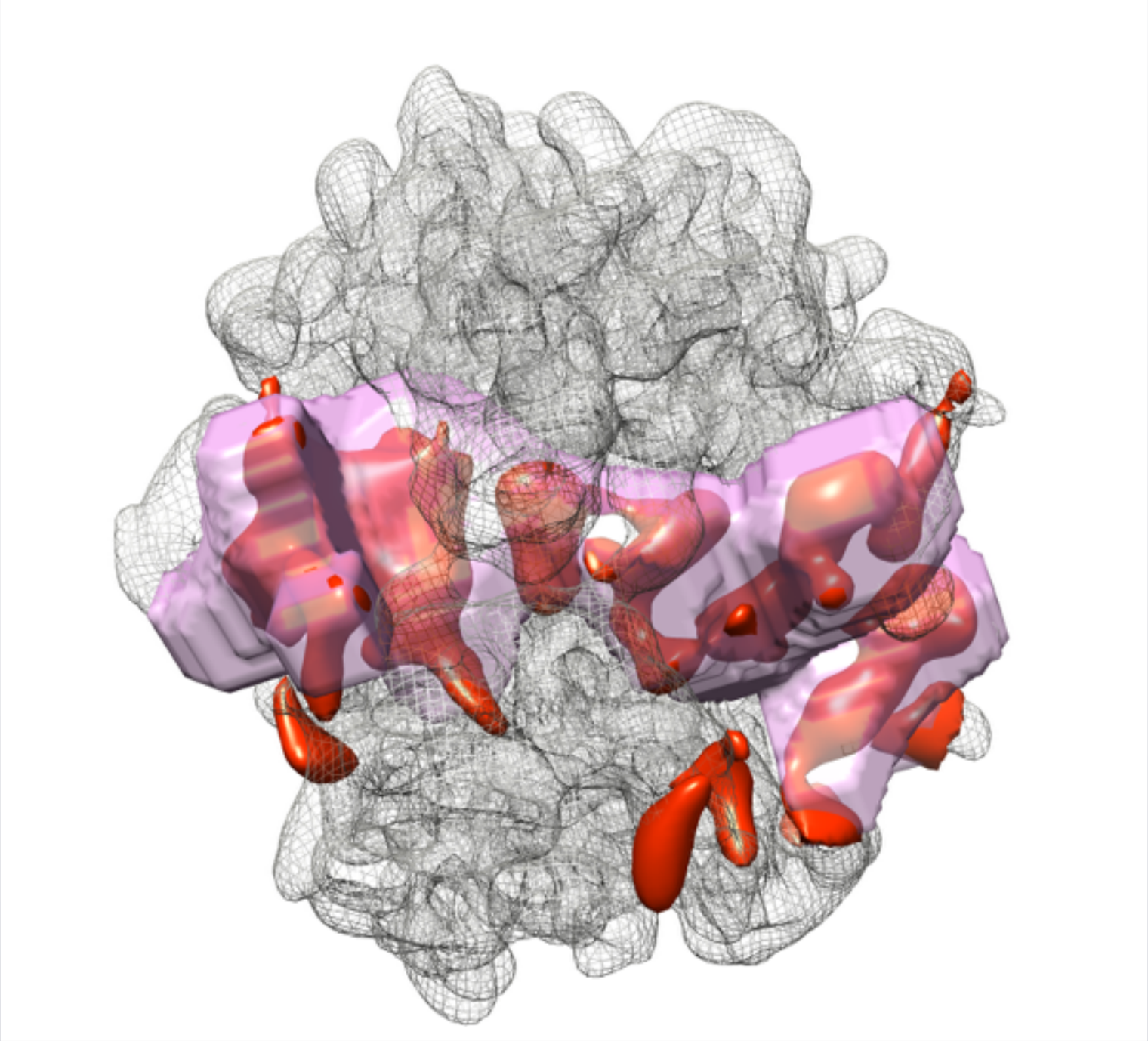
1. Create a 3D mask around locations of high 3D variability

2. Project 3D mask in the directions of the particle views



3. Calculate distances between reprojections of a 3D map and 2D EM data **only** within 2D regions outlined by respective projections of 3D mask.

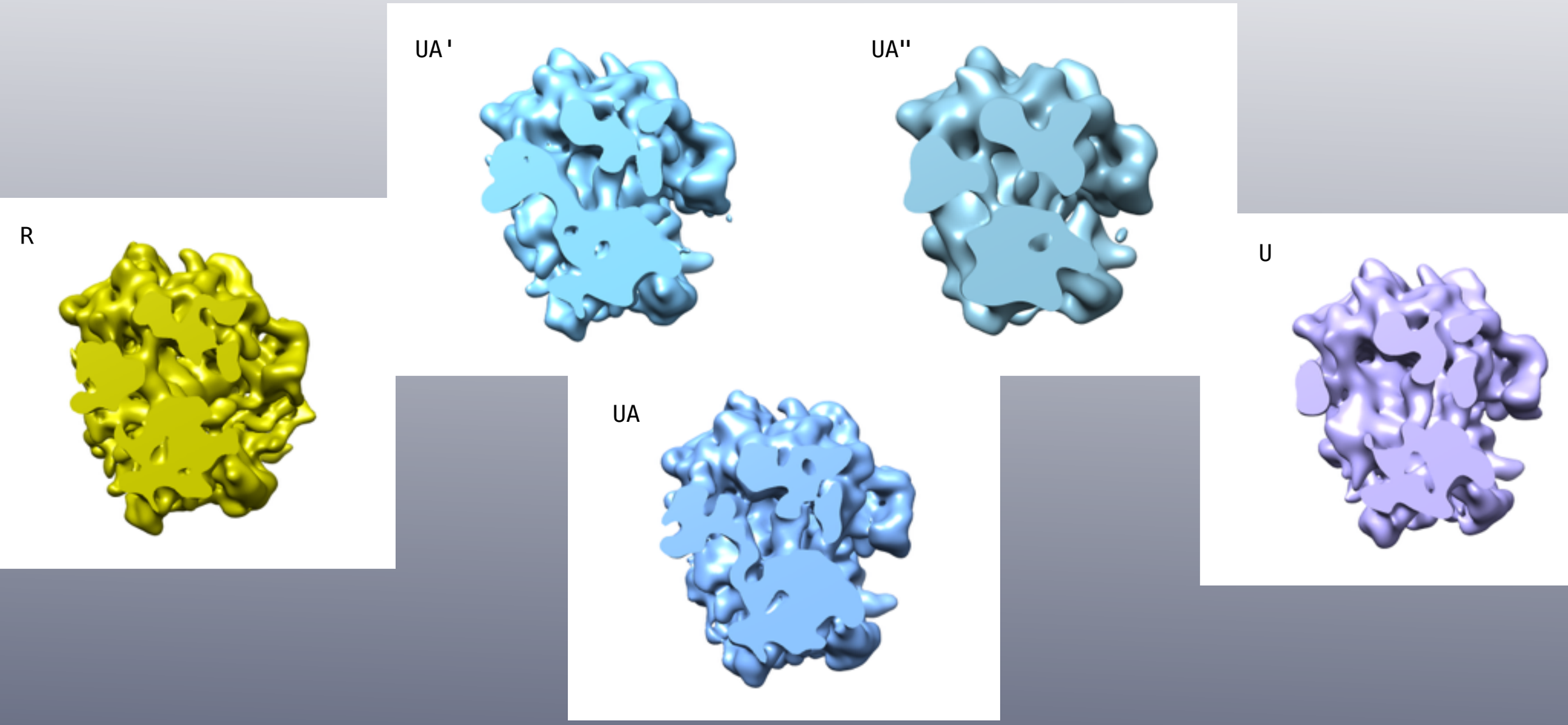
Focused mask



Focused sorting results

number of images: 10001; window size: 64
requested number of images per group: 2000
minimum group size: 100

Unaccounted images: 17%
Reproducibility: 83%



Ratcheted	Unratcheted				50S
E, EFG	E, P,(A,	A',	A'')	P, A	not shown
39% (3 groups)	16%	9%	2%	7%	6%

Known problems

1. Ribosome has "binary" groups.
2. Distance problem
3. Sensitivity to the number of groups

Conclusions

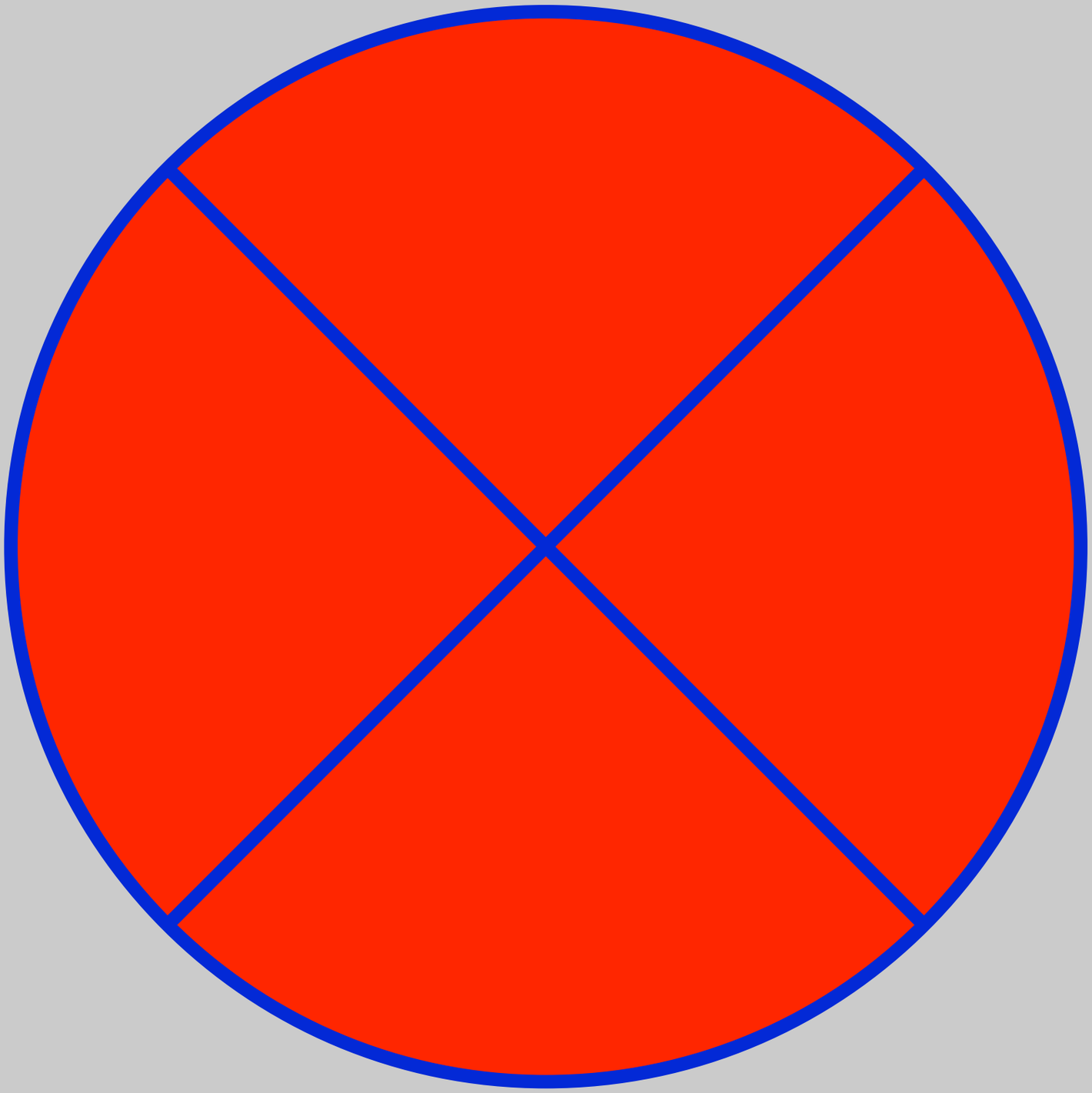
1. A simple and intuitive approach with outcome validation based on reproducibility concept.
2. sort3d requires a minimal number of parameters:
    1. Desirable number of images per group
    2. Minimum group size
3. Reproducibility and optimization result in a relatively long computation time.
4. Extensive statistical diagnostics.

ACKNOWLEDGMENTS

Zhong Huang

Justus Loerke
Christian M.T. Spahn

Francisco Asturias

Steve Ludtke

NIH