Ab initio cryo-EM structure determination as a validation problem

Pawel A. Penczek

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

ACKNOWLEDGMENTS

Francisco J. Asturias
La Jolla, CA

Christian M.T. Spahn
Charité, Berlin

NIH

CONCLUSIONS

1. Validation should be an integral part of the structure determination process.

2. Any method should be permitted to fail under controlled circumstances as the failure can be as informative as success.

3. EM projection images are of very poor quality. Therefore, they should not be evaluated individually but as members of statistical assemblies.

4. Implementation in SPARX http://sparx-em.org/sparxwiki/ with new additions of tools for the analysis of local variability (please see the poster).

Statistical cross-validation
for
detecting and preventing overfitting

Problem of model selection

EM DATA AND PARAMETER ERROR ESTIMATION

- A typical EM experiment generates a single dataset and it is not possible to derive an analytical expression to determine (alignment) parameter errors
- The challenge is then to estimate parameter errors in the absence of independent sample sets
- Statistical Resampling offers the best option for accurate estimation of parameter errors independent of assumptions about their statistical properties

EM DATA AND PARAMETER ERROR ESTIMATION

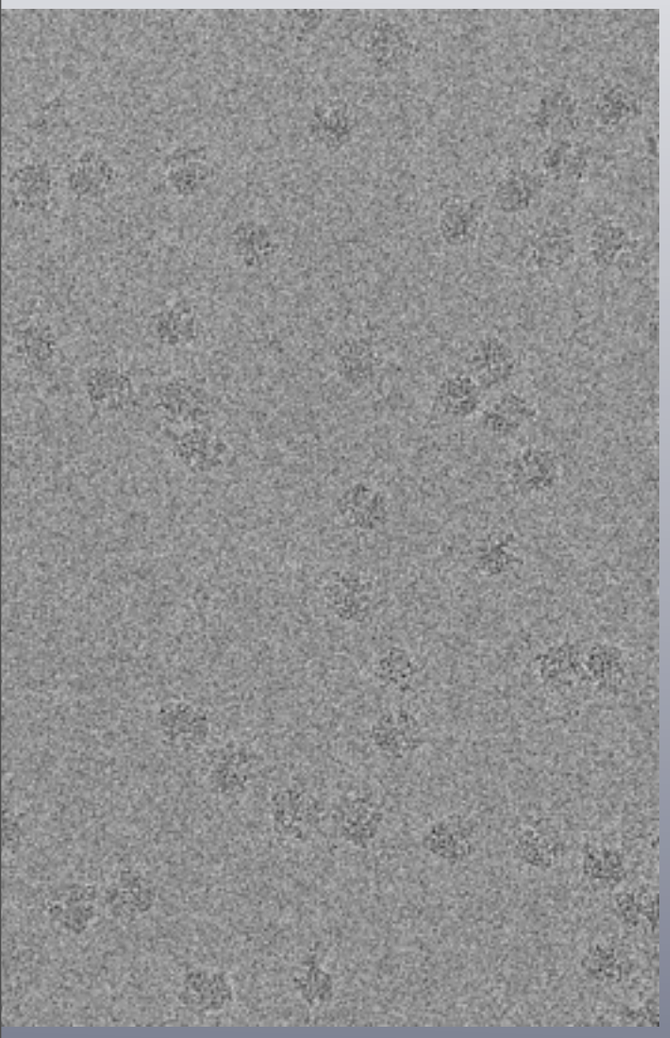
- A typical EM experiment generates a single dataset and it is not possible to derive an analytical expression to determine (alignment) parameter errors

- The challenge is then to estimate parameter errors in the absence of independent sample sets

- Statistical Resampling offers the best option for accurate estimation of parameter errors independent of assumptions about their statistical properties

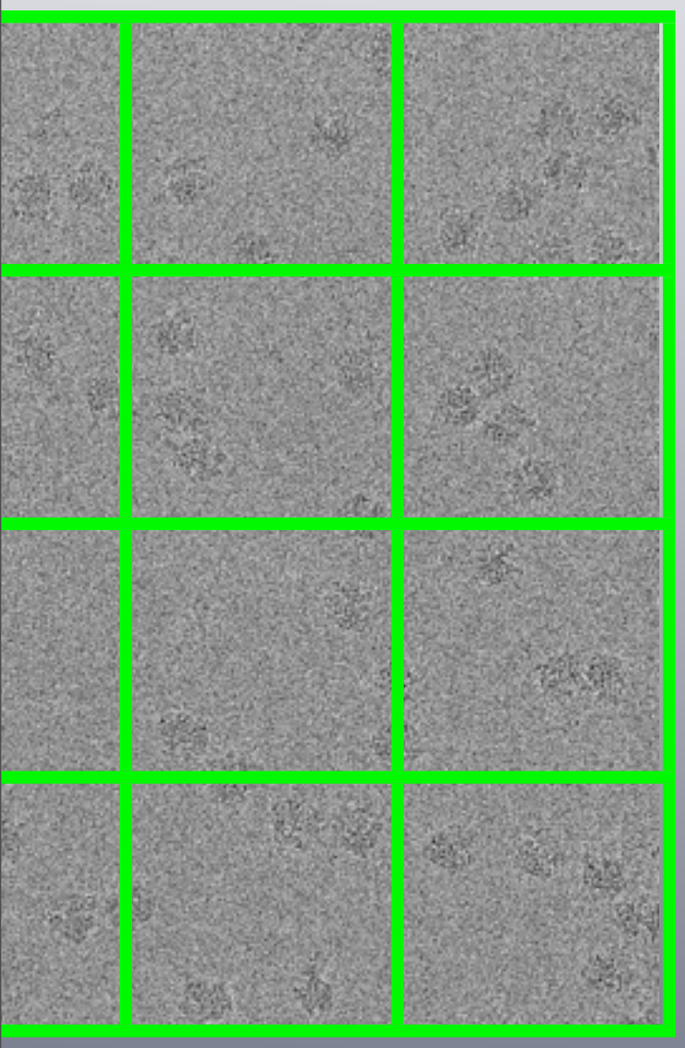
If we treat the observed sample (EM dataset) as though it exactly represented the entire population,
evaluating artificial variability generated through resampling allows us to accurately estimate variability of a sample statistic

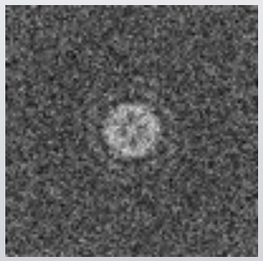
CTF parameter estimation and error assessment through bootstrap resampling (CTER)

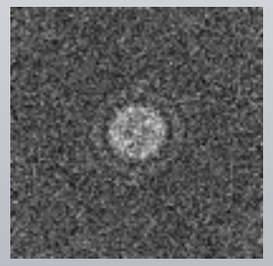


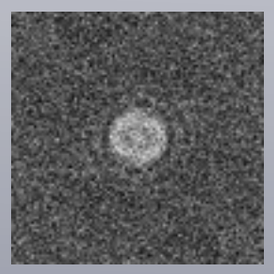
Penczek, P. A., Fang, J., X. Li, X., Cheng, Y., Loerke, J., Spahn, Ch.M.T.: CTER-Rapid estimation of CTF parameters with error assessment. *Ultramicroscopy*, **140**:9-19, 2014.

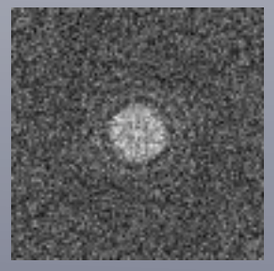
CTF parameter estimation and error assessment through bootstrap resampling (CTER)

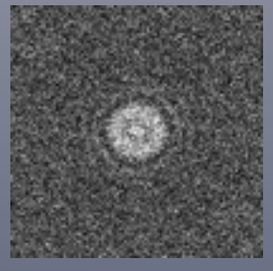


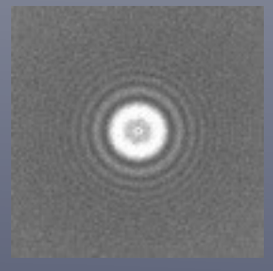








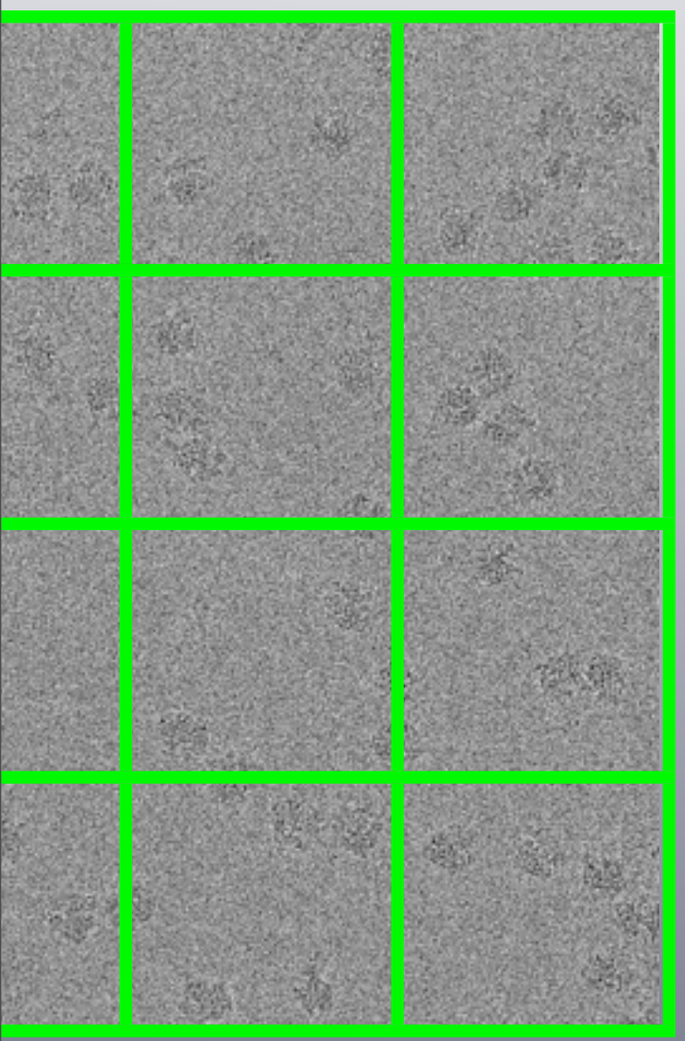


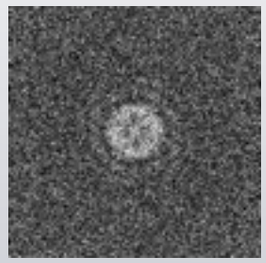


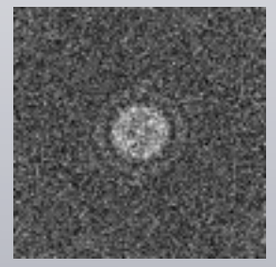
Average power spectrum and its variance

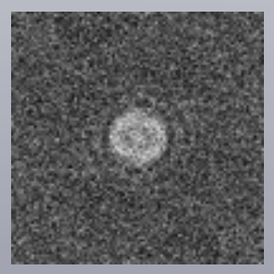
Penczek, P. A., Fang, J., X. Li, X., Cheng, Y., Loerke, J., Spahn, Ch.M.T.: CTER-Rapid estimation of CTF parameters with error assessment. *Ultramicroscopy*, **140**:9-19, 2014.

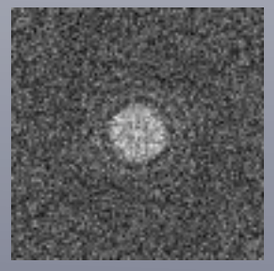
CTF parameter estimation and error assessment through bootstrap resampling (CTER)

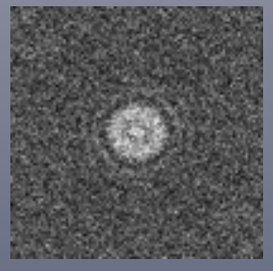


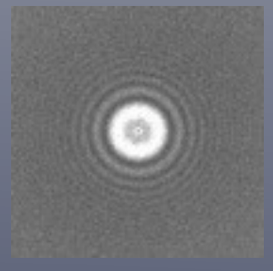


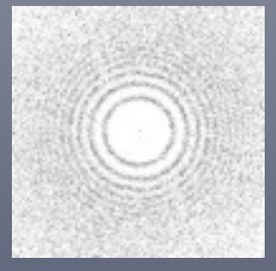








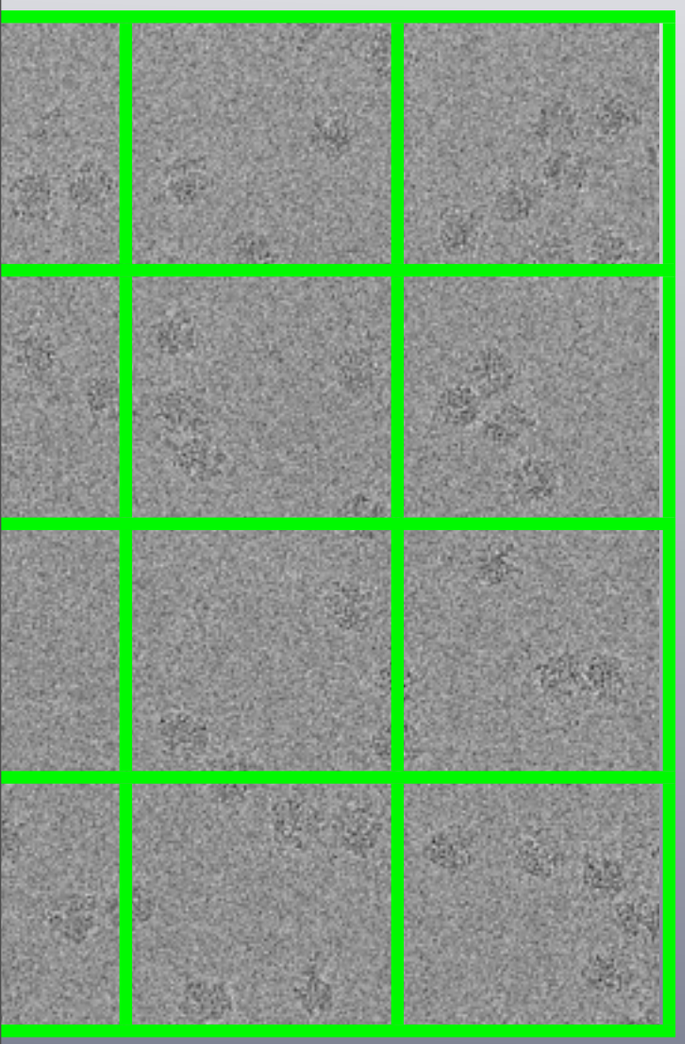


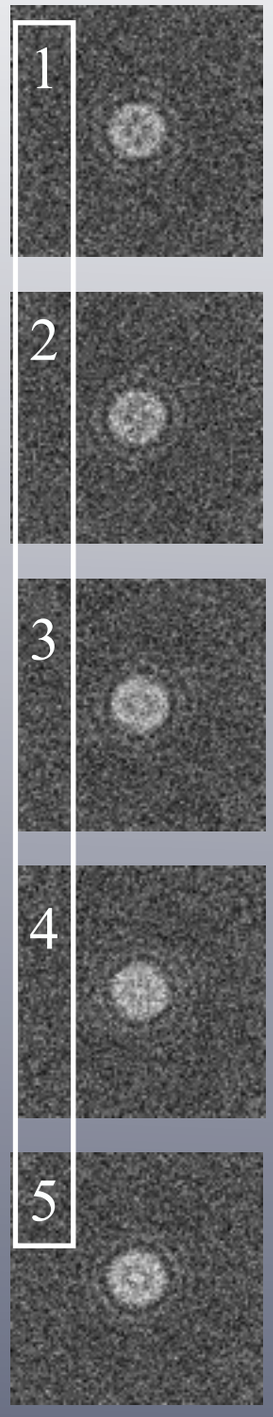


Average power spectrum and its variance

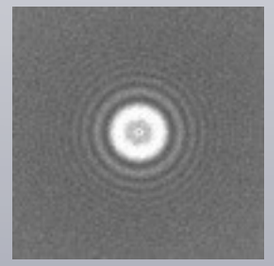
Penczek, P. A., Fang, J., X. Li, X., Cheng, Y., Loerke, J., Spahn, Ch.M.T.: CTER-Rapid estimation of CTF parameters with error assessment. *Ultramicroscopy*, **140**:9-19, 2014.

CTF parameter estimation and error assessment through bootstrap resampling (CTER)





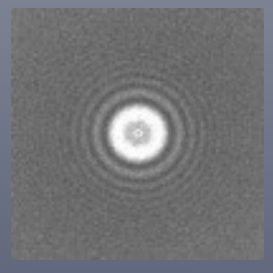
Average of selected power spectra

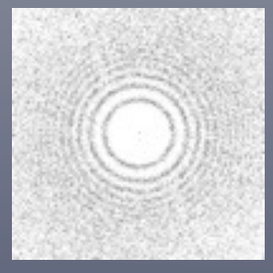


Determine:
1. defocus
2. astigmatism amplitude
3. astigmatism angle

Repeat *B* times

BOOTSTRAP RESAMPLING OF TILED POWER SPECTRA

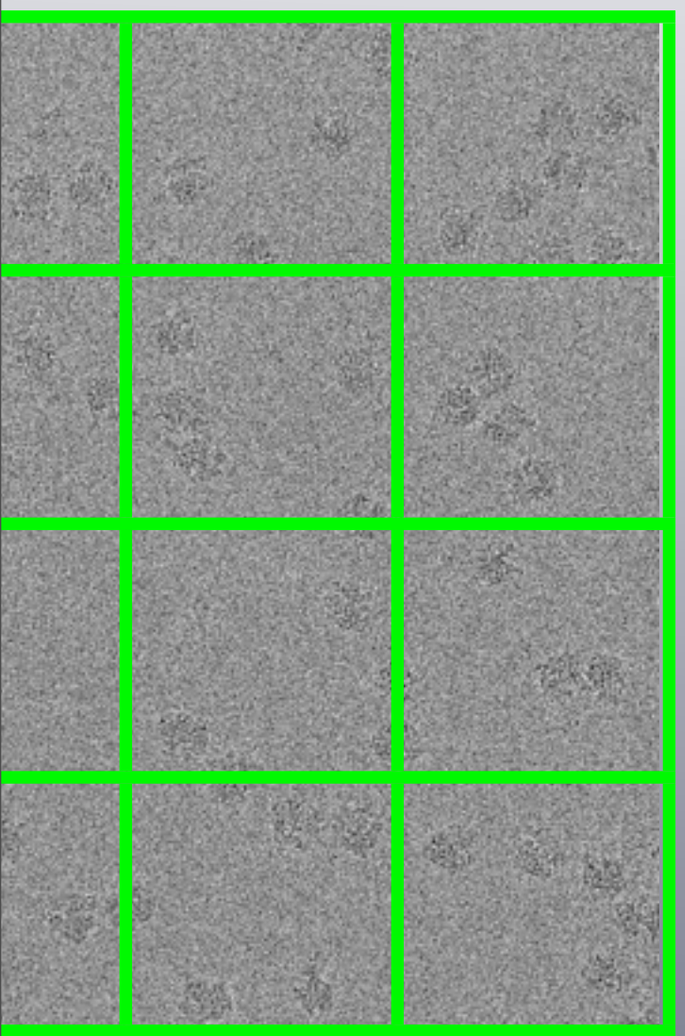


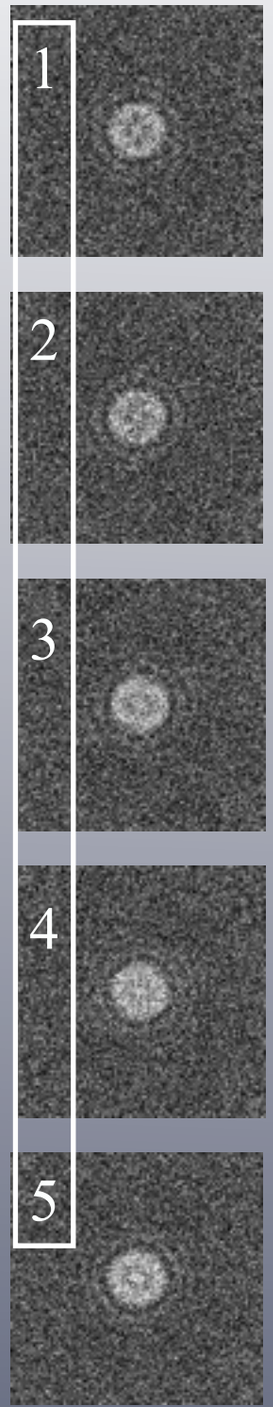


Average power spectrum and its variance

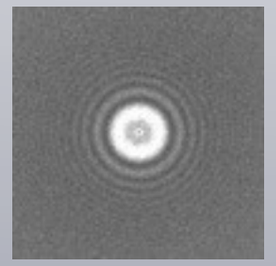
Penczek, P. A., Fang, J., X. Li, X., Cheng, Y., Loerke, J., Spahn, Ch.M.T.: CTER-Rapid estimation of CTF parameters with error assessment. *Ultramicroscopy*, **140**:9-19, 2014.

CTF parameter estimation and error assessment through bootstrap resampling (CTER)





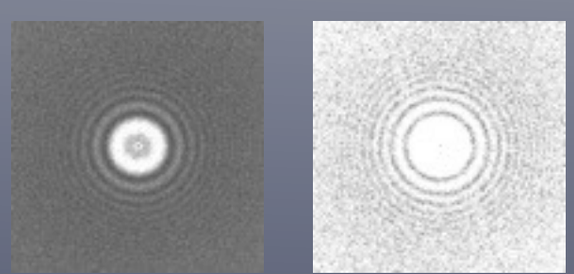
Average of selected power spectra



Determine:
1. defocus
2. astigmatism amplitude
3. astigmatism angle

Repeat ***B*** times

BOOTSTRAP RESAMPLING OF TILED POWER SPECTRA



Average power spectrum and its variance

RESULT

Based on *B* estimates compute average value and error (std. dev.) of
<defocus>
<astigmatism amplitude>
<astigmatism angle>

Penczek, P. A., Fang, J., X. Li, X., Cheng, Y., Loerke, J., Spahn, Ch.M.T.: CTER-Rapid estimation of CTF parameters with error assessment. *Ultramicroscopy*, **140**:9-19, 2014.

ISAC: VALIDATION OF 2D MULTI-REFERENCE ALIGNMENT THROUGH STABILITY TESTING

1. If a set of images is homogeneous, the result from reference-free alignment is stable even for very low SNR data.

2. The converse is true, i.e., if a set of images is stable, it must be homogeneous.

2D alignment is stable if perturbation of initial alignment parameters does not produce dramatically different results.

ISAC: VALIDATION OF 2D MULTI-REFERENCE ALIGNMENT THROUGH STABILITY TESTING

1. If a set of images is homogeneous, the result from reference-free alignment is stable even for very low SNR data.

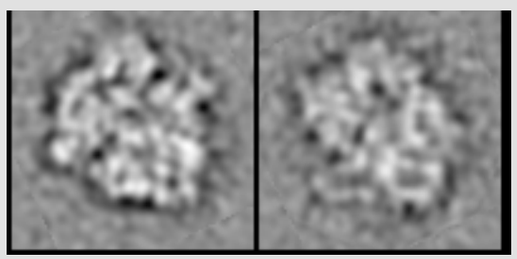
2. The converse is true, i.e., if a set of images is stable, it must be homogeneous.

2D alignment is stable if perturbation of initial alignment parameters does not produce dramatically different results.

Assuming 1 and 2 are correct:
If we can find homogeneous subsets of images,
we can solve the multi-reference alignment problem.

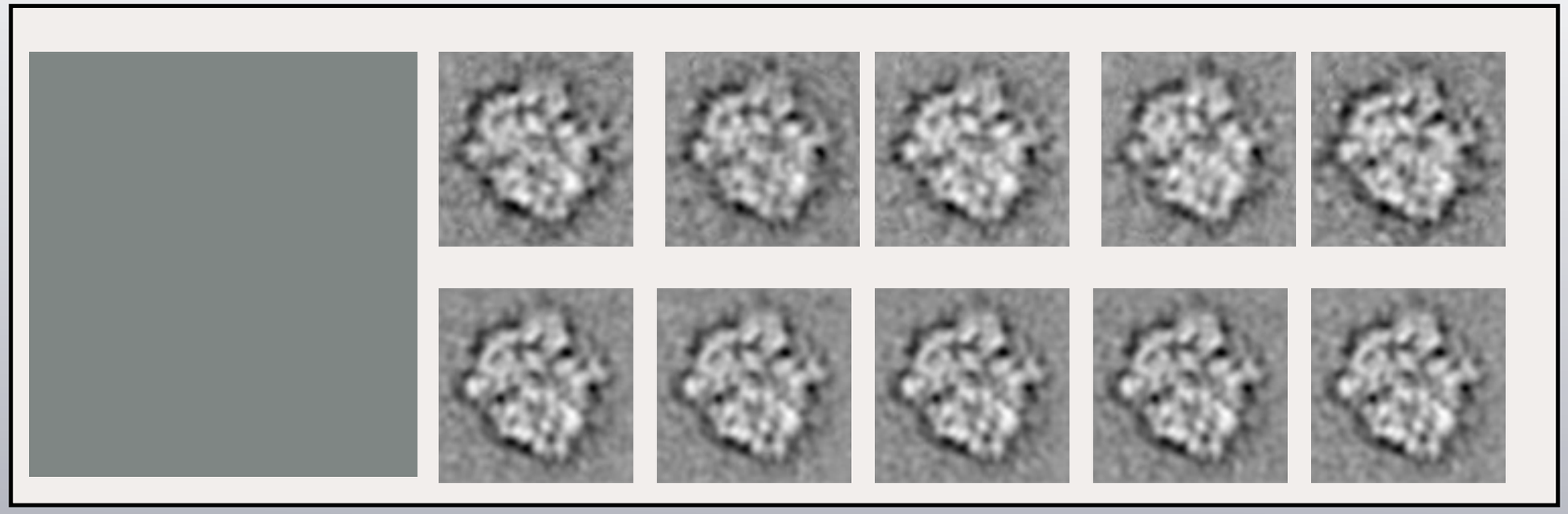
STABLE VS. UNSTABLE CLASSES: A TEST CASE

- Two groups were mixed 50-50, their respective averages are:

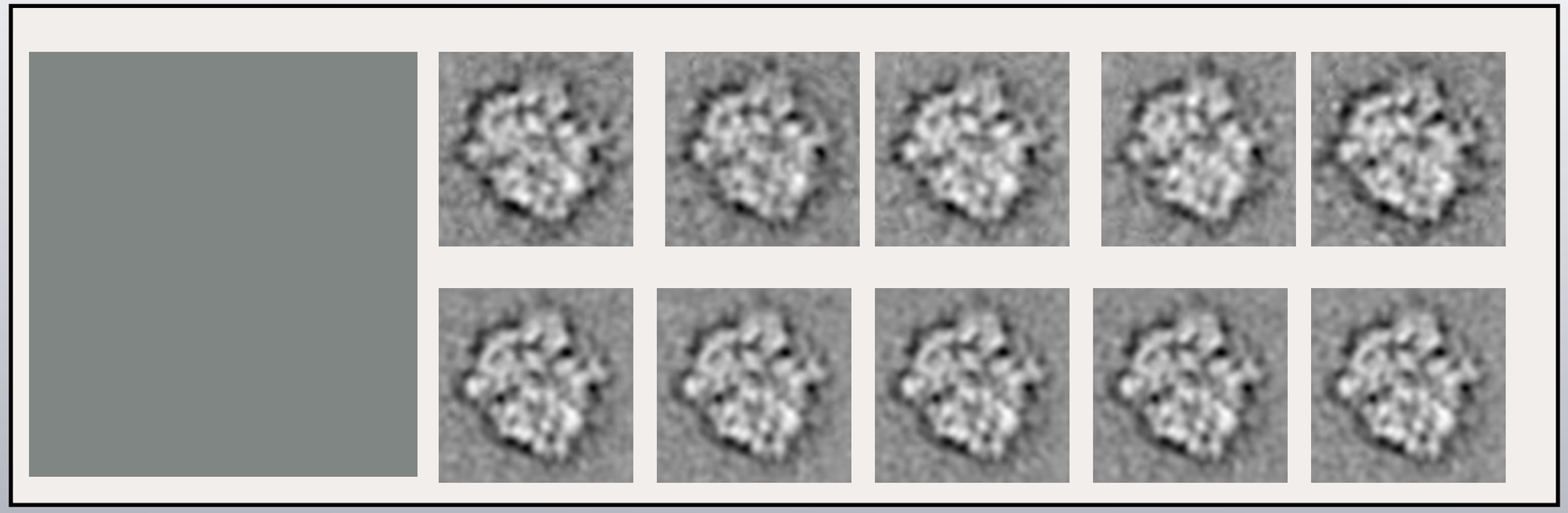


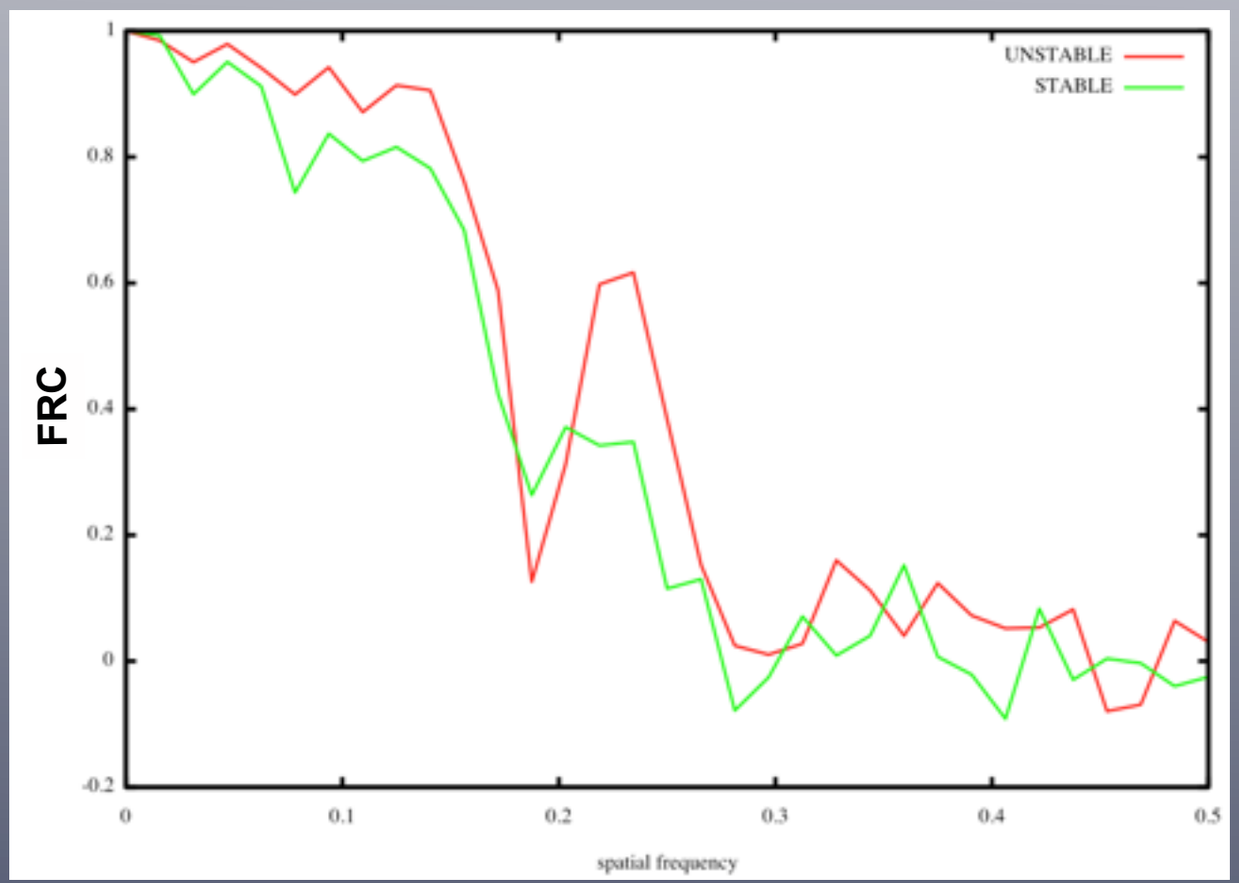
- Sum of these two averages:

STABLE VS. UNSTABLE CLASSES: TEST RESULTS

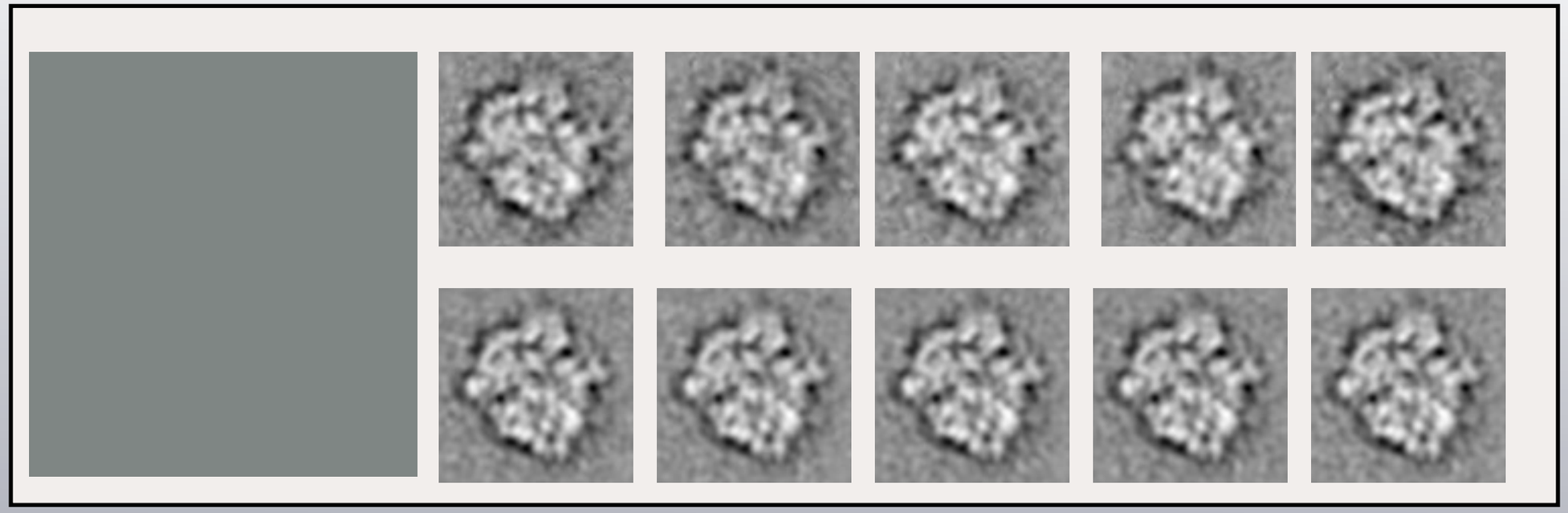


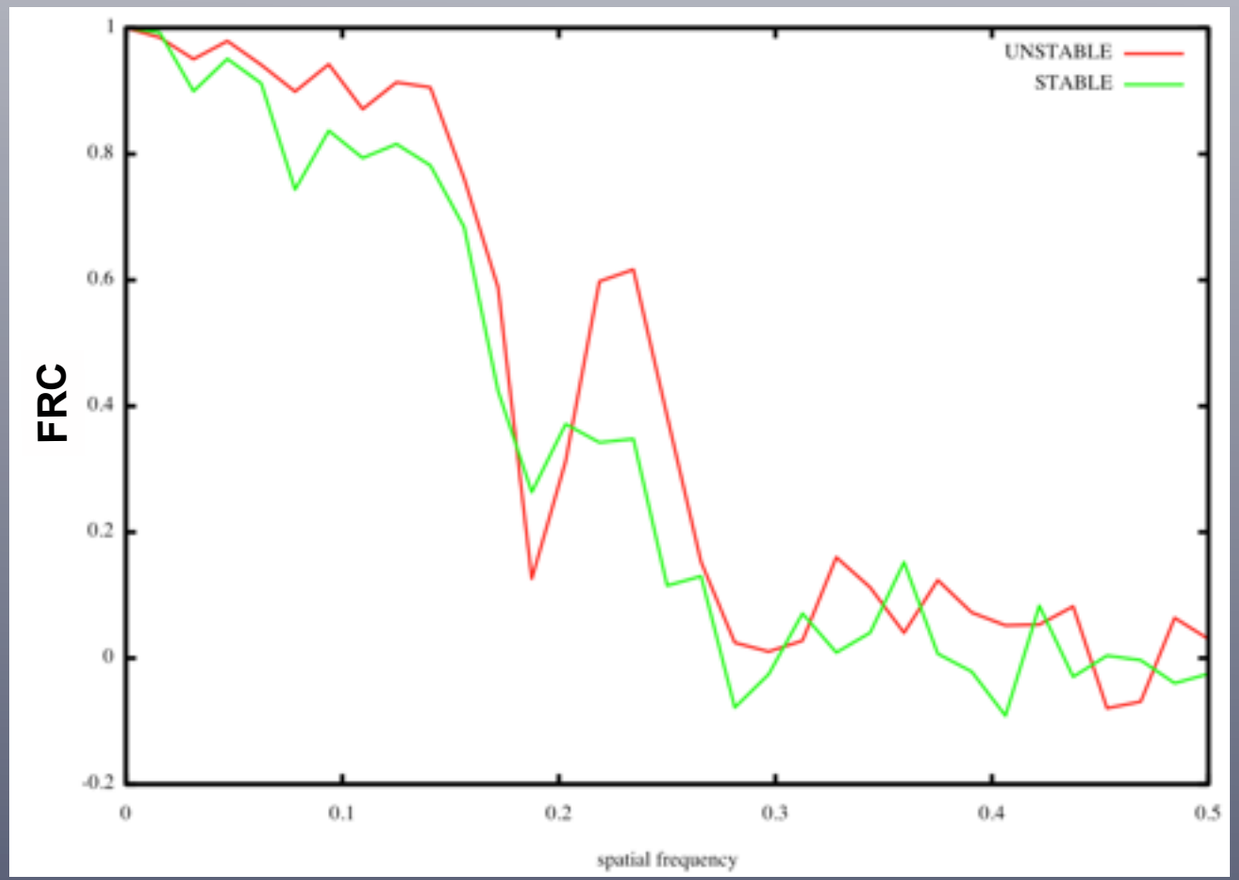
STABLE VS. UNSTABLE CLASSES: TEST RESULTS

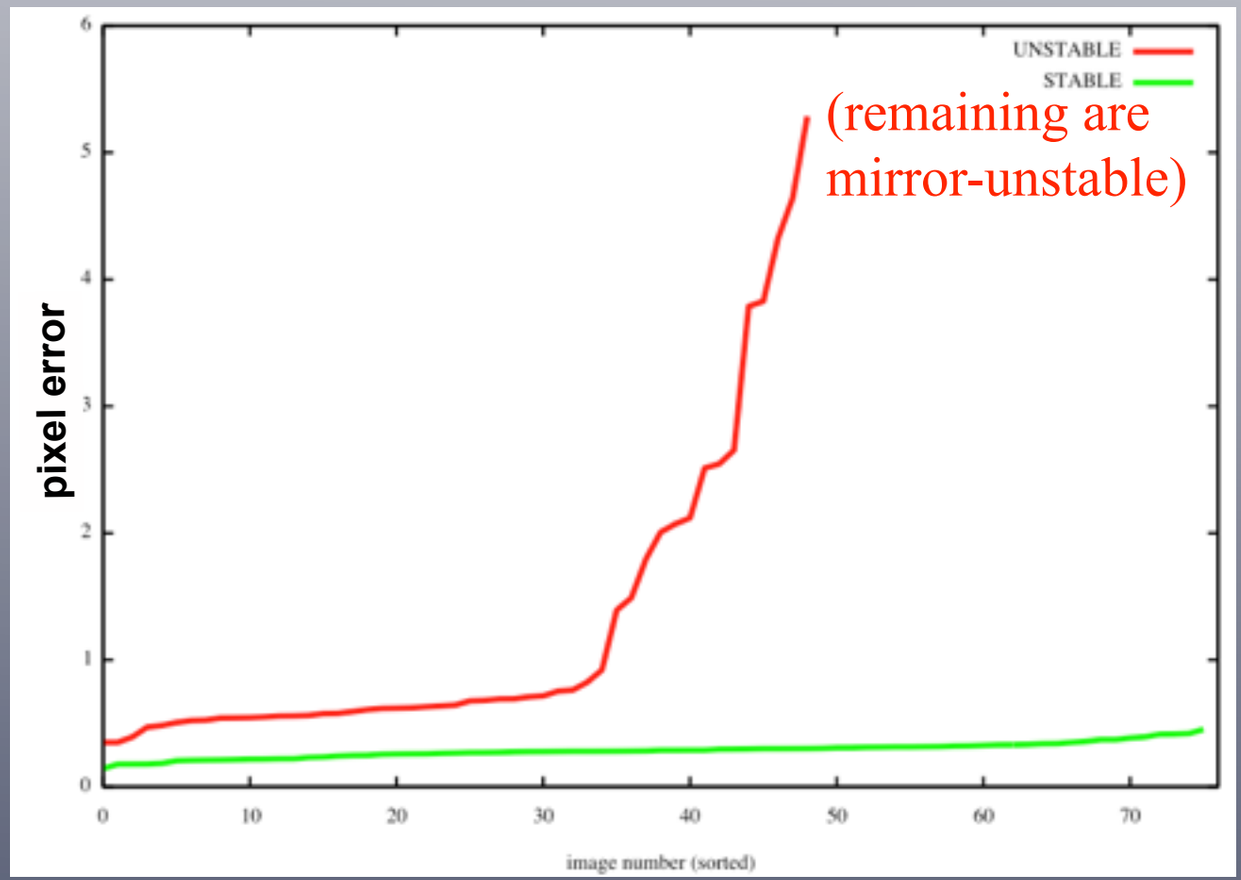




STABLE VS. UNSTABLE CLASSES: TEST RESULTS







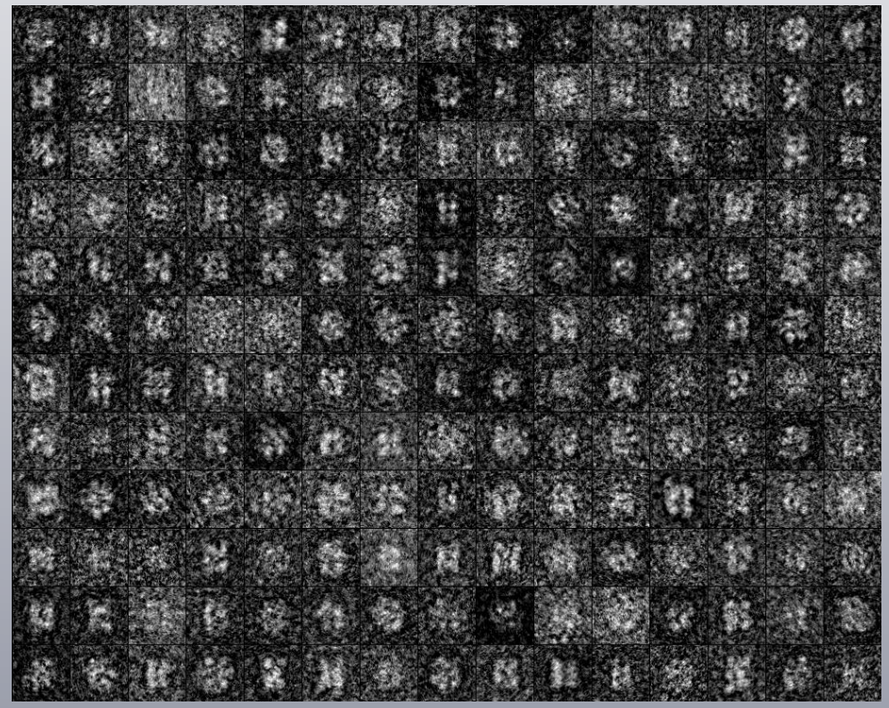
2D MULTI-REFERENCE ALIGNMENT (MRA)

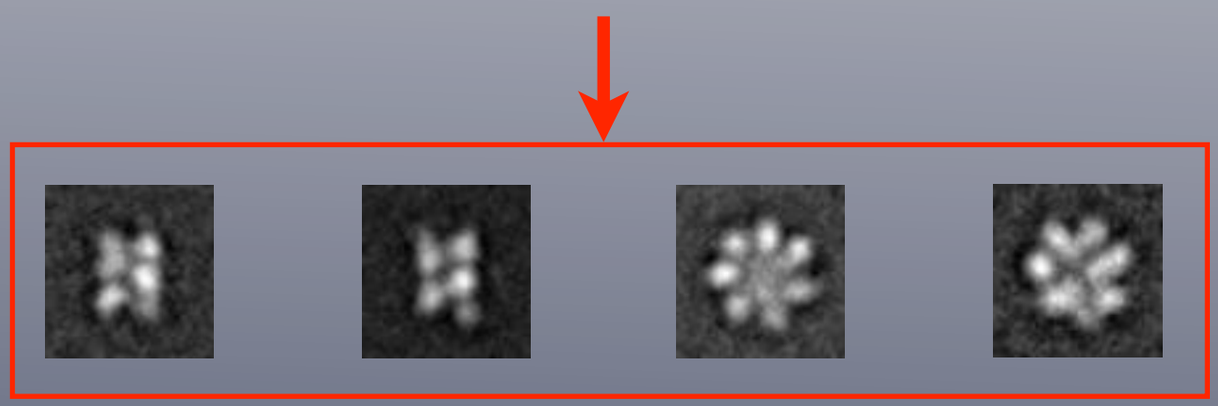
MRA is equivalent to *K*-means clustering, with the distance between images defined as a maximum similarity over the permissible range of image rotations and translations.

K-means results depend on the solution to another nontrivial problem: the alignment of a set of 2D images.

Because neither of these two problems can be easily solved, the difficulty is compounded.

n images



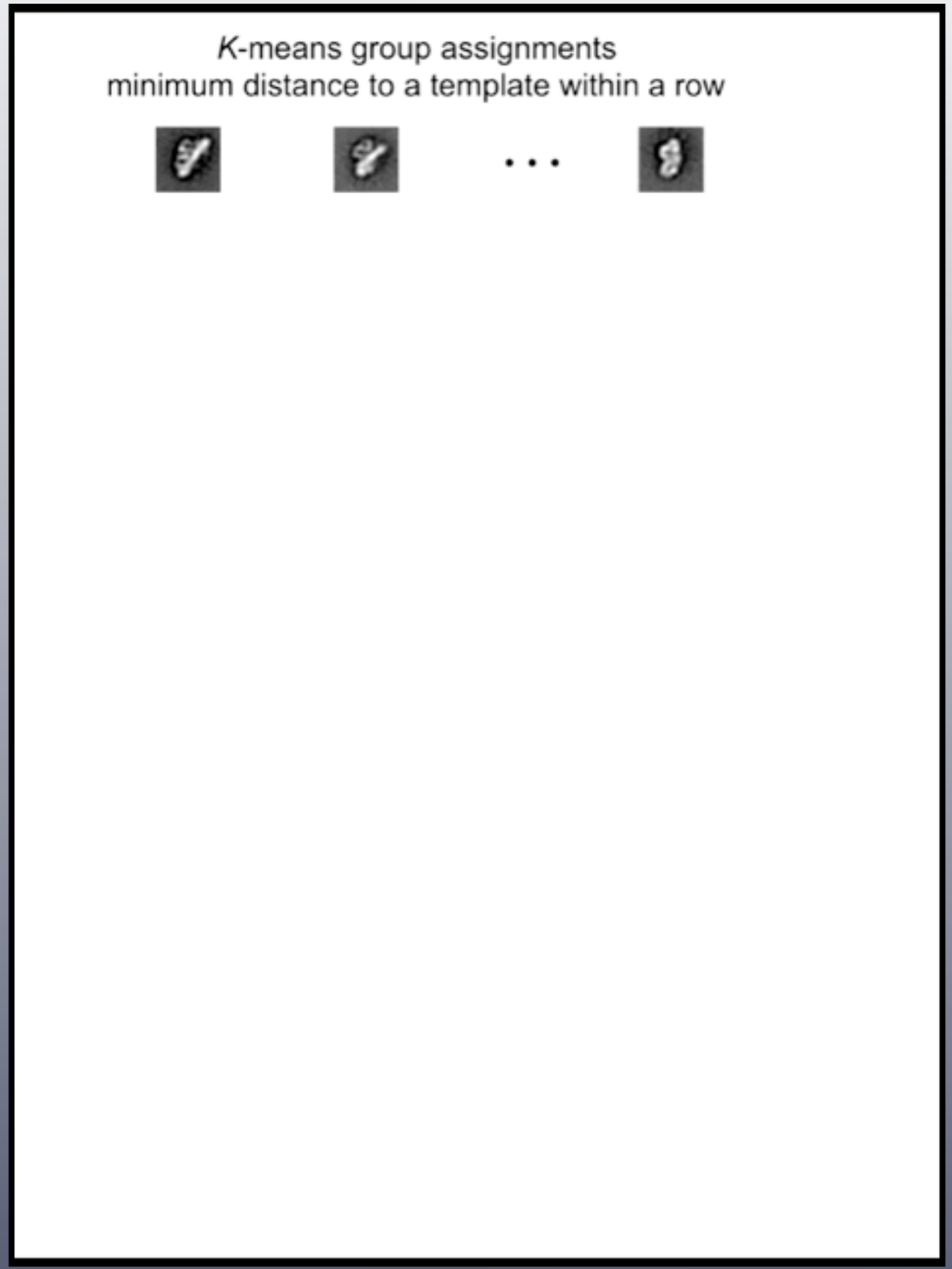


K averages (clusters)

K-MEANS CLUSTERING

KNOWN PROPERTIES:

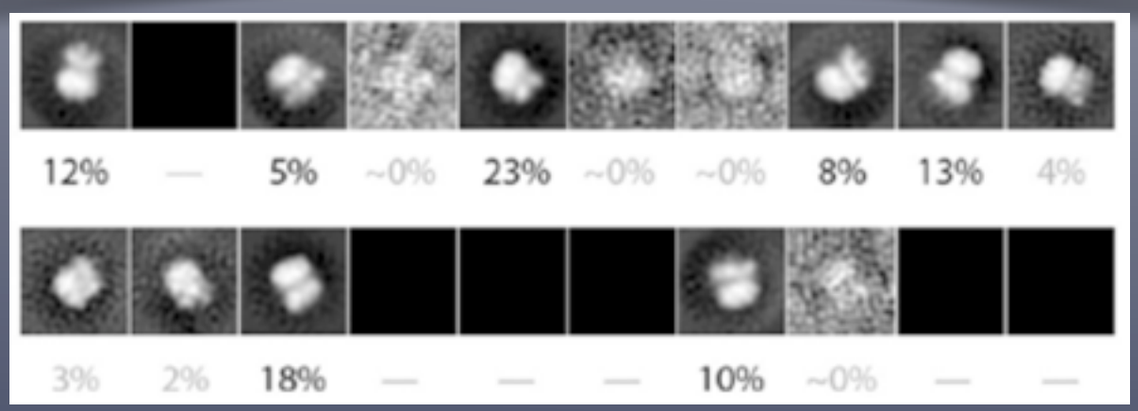
- Very fast convergence guaranteed in a finite number of steps
- Converges only to a local minimum
- Unclear how to determine the appropriate number of classes (*K*)
- All images must be assigned to an average
- The solution (final averages) depends on the initial set of averages, and will change if clustering is repeated using different initial averages
- In EM, when alignment is added, classes tend to collapse

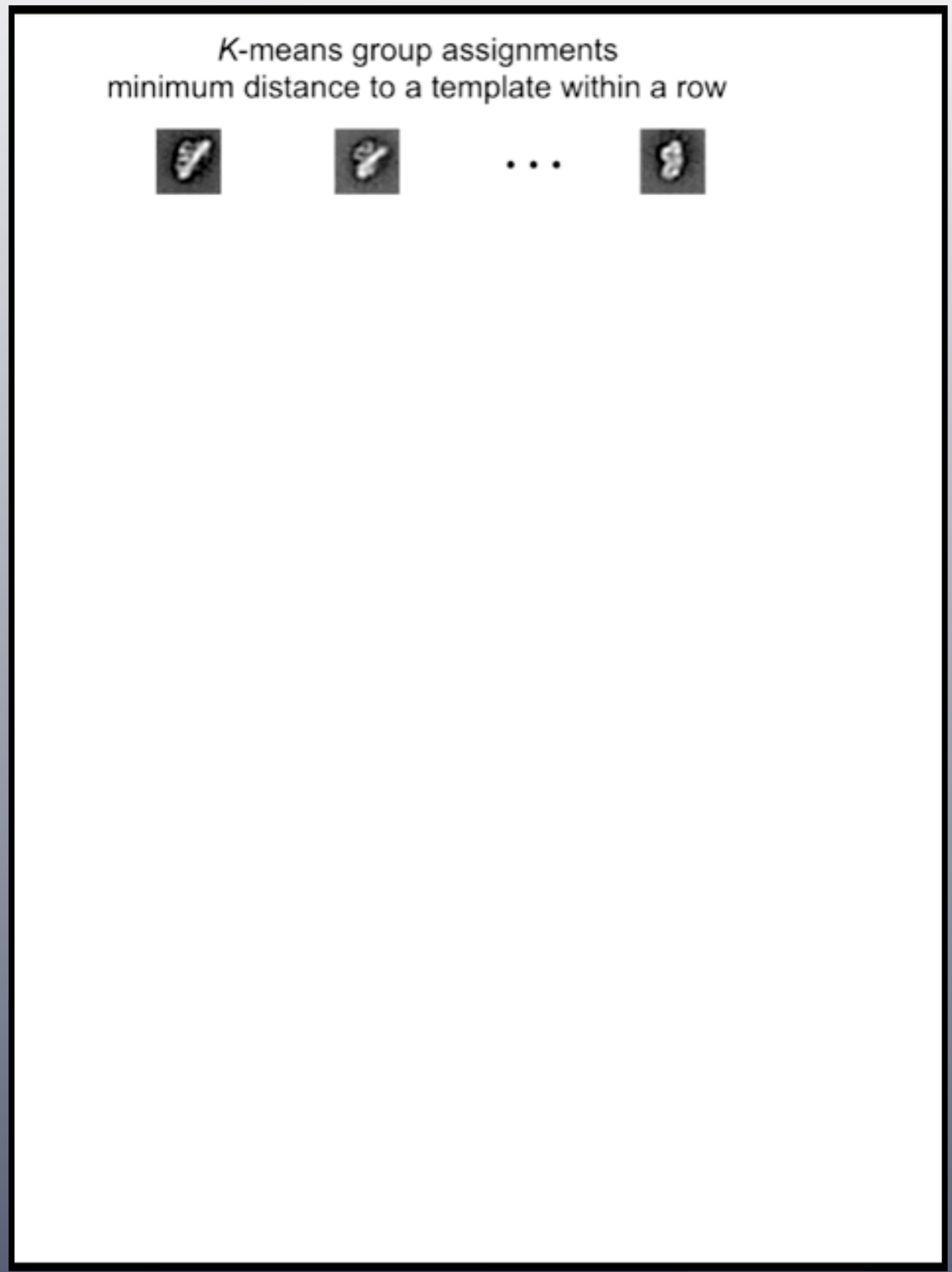


K-MEANS CLUSTERING

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*)
- All images must be assigned to an average
- The solution (final averages) depends on the initial set of averages, and will change if clustering is repeated using different initial averages
- In EM, when alignment is added, classes tend to collapse





EQK(EQUAL GROUP SIZE)-MEANS CLUSTERING

Assign n images to K classes

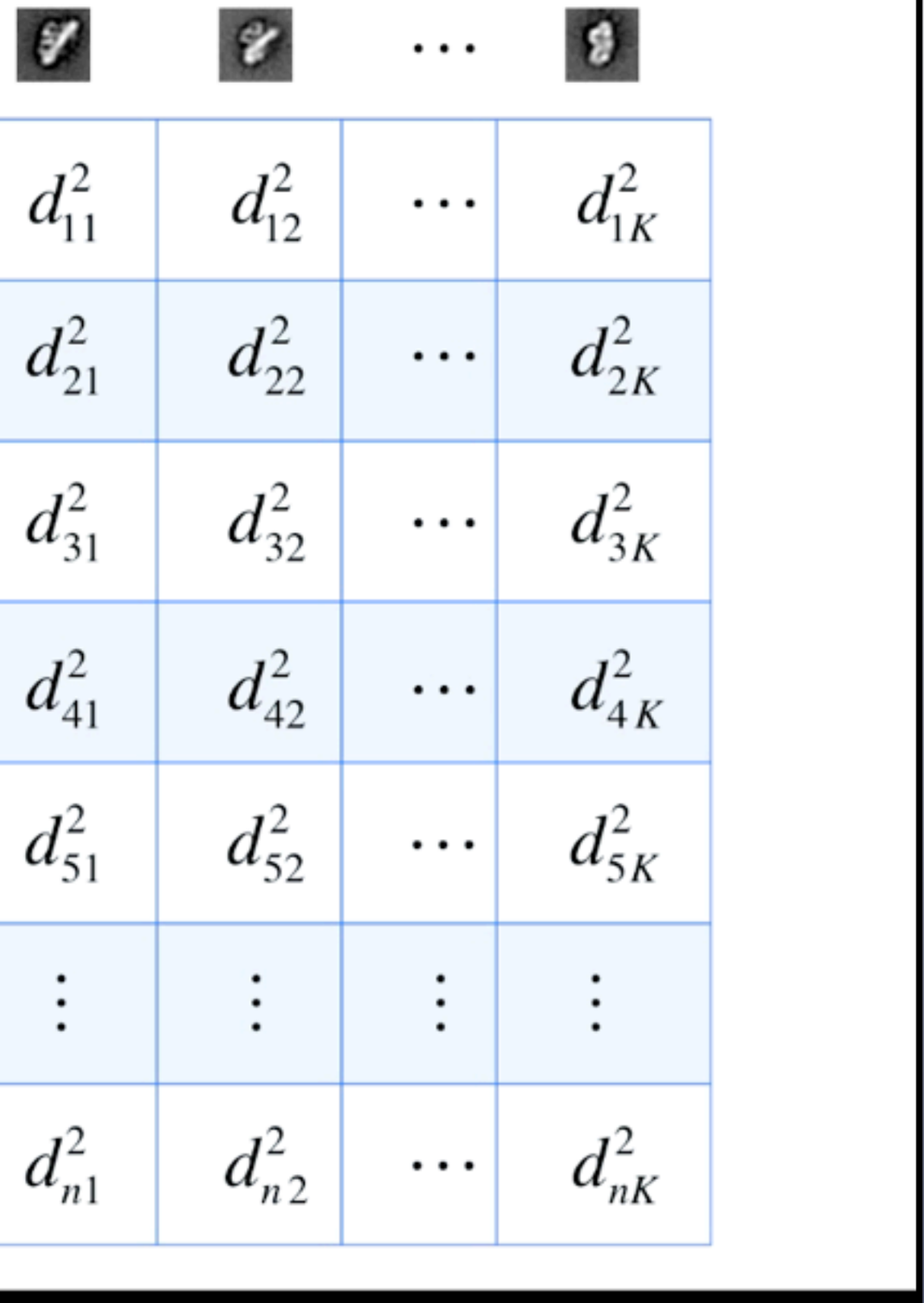
such that each class contains

$\frac{n}{K}$ images

EQK(EQUAL GROUP SIZE)-MEANS CLUSTERING

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



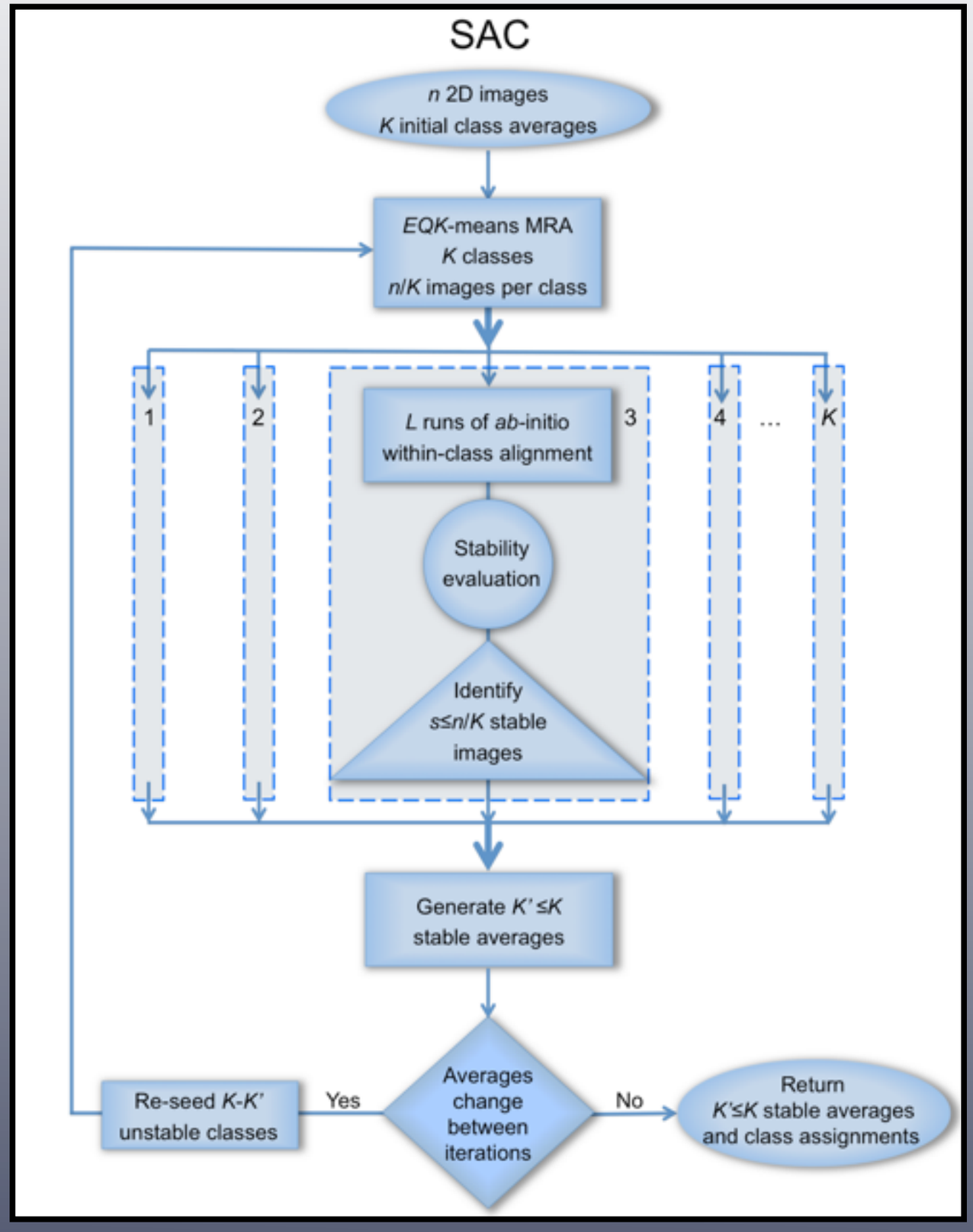
A PROTOCOL FOR TESTING ALIGNMENT STABILITY

1. Run reference-free alignment *L*-times, using randomized initial orientation parameters

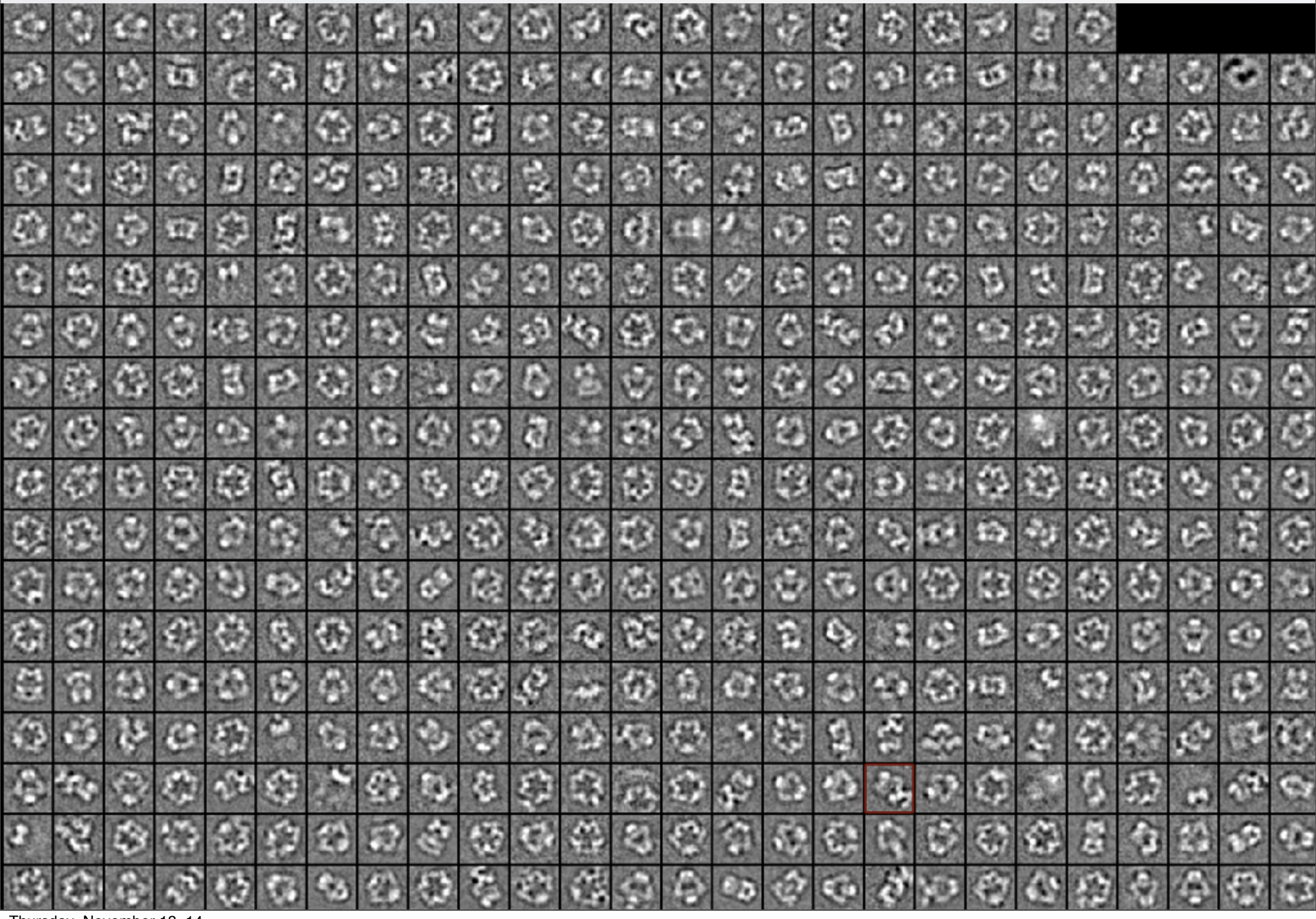
2. Bring all L sets of solutions into register by simultaneous minimization of the variance of orientation parameters (similar but not equivalent to alignment of resulting averages)

3. Compute pixel error for each image using orientation parameters for *L* positions it adopted

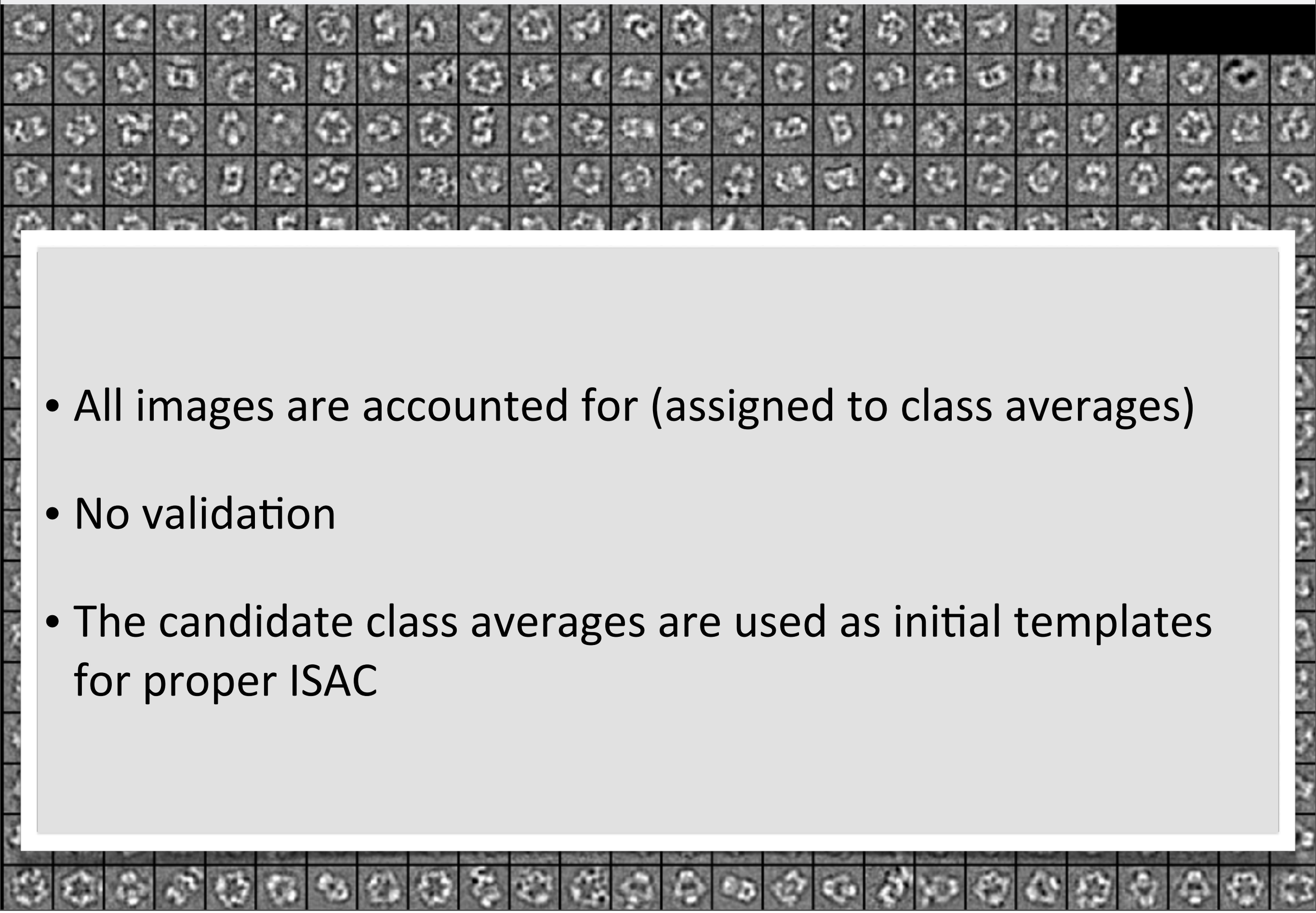
4. The set is called stable if the average of pixel errors for all images in *L* alignments is less than a predefined threshold (usually one pixel).



CANDIDATE CLASS AVERAGES



CANDIDATE CLASS AVERAGES



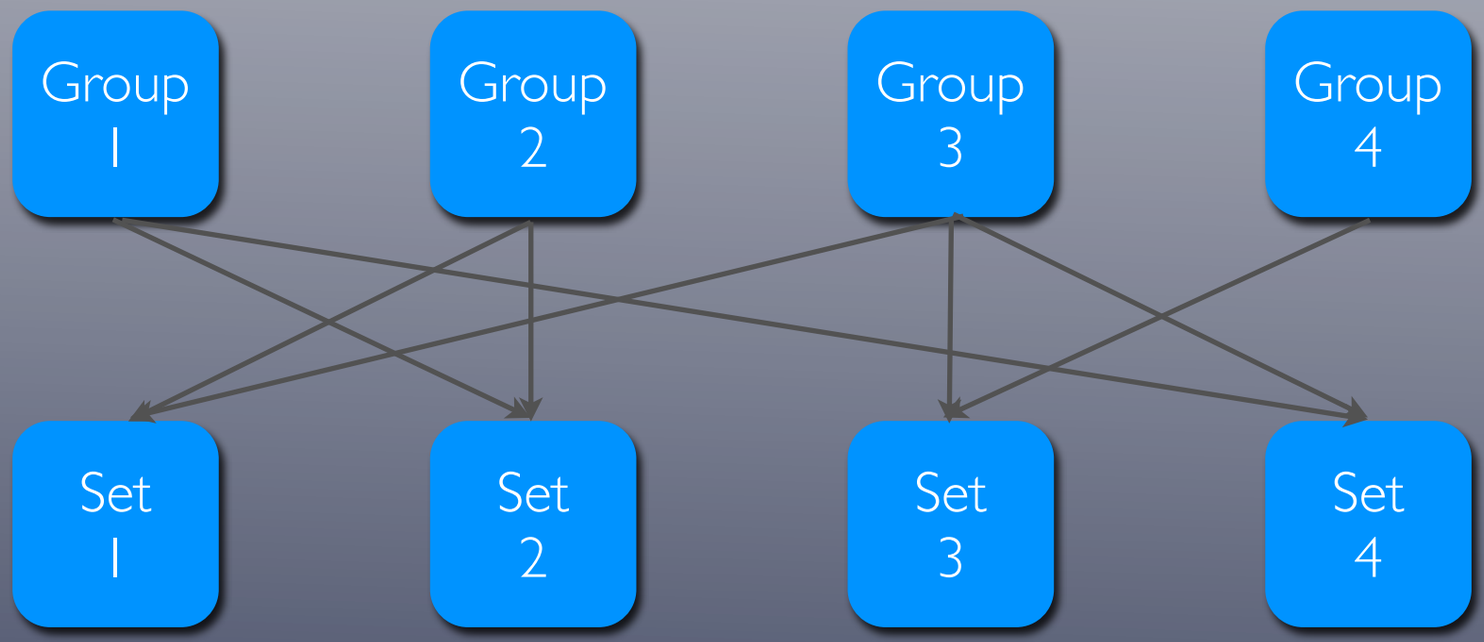
- All images are accounted for (assigned to class averages)
- No validation
- The candidate class averages are used as initial templates for proper ISAC

REPRODUCIBILITY

- Since *EQK*-means, even if combined with an alignment stability test, does not guarantee an optimum solution (global minimum) and stable groups can be fake, we require the solution to be reproducible over a number of quasi-independent runs.

- We have m=4 *EQK*-means runs analyzing the data in parallel. Once all runs produce their respective averages, we compare assignments of images to class averages and select as reproducible subsets shared among quasi-independent runs.

m= 2

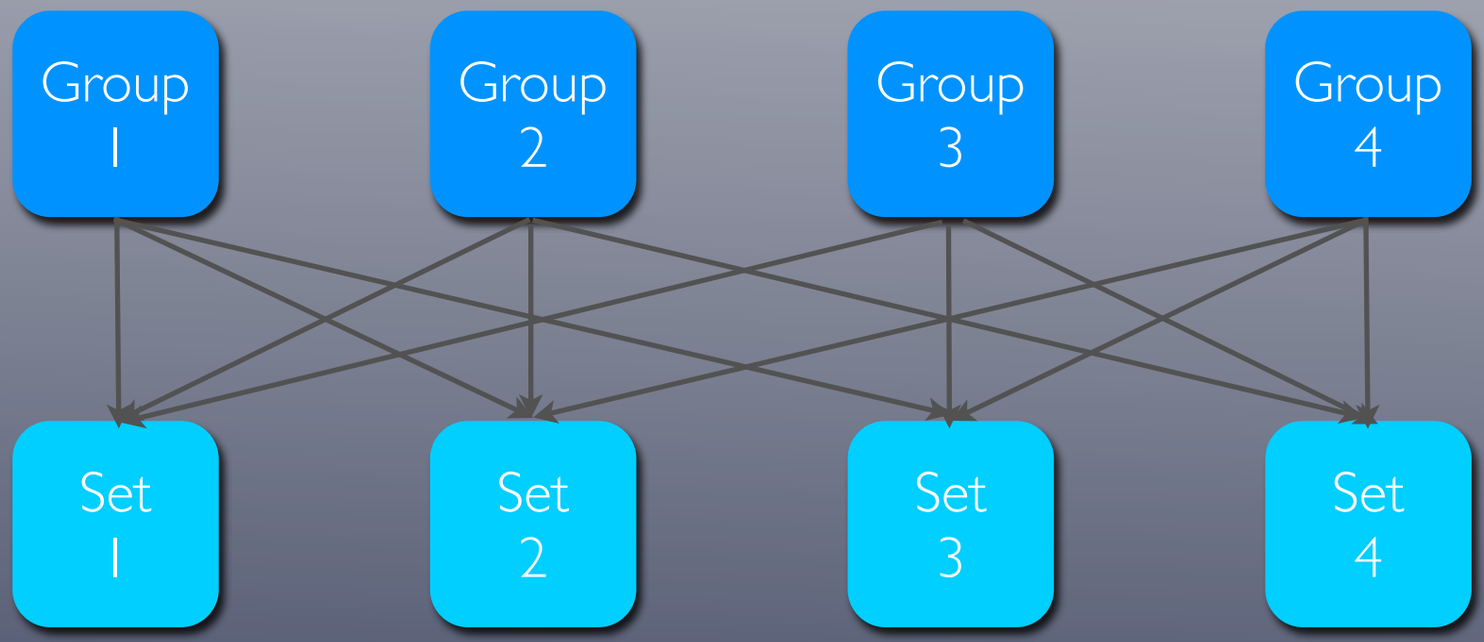


REPRODUCIBILITY

- Since *EQK*-means, even if combined with an alignment stability test, does not guarantee an optimum solution (global minimum) and stable groups can be fake, we require the solution to be reproducible over a number of quasi-independent runs.

- We have m=4 *EQK*-means runs analyzing the data in parallel. Once all runs produce their respective averages, we compare assignments of images to class averages and select as reproducible subsets shared among quasi-independent runs.

m= 3

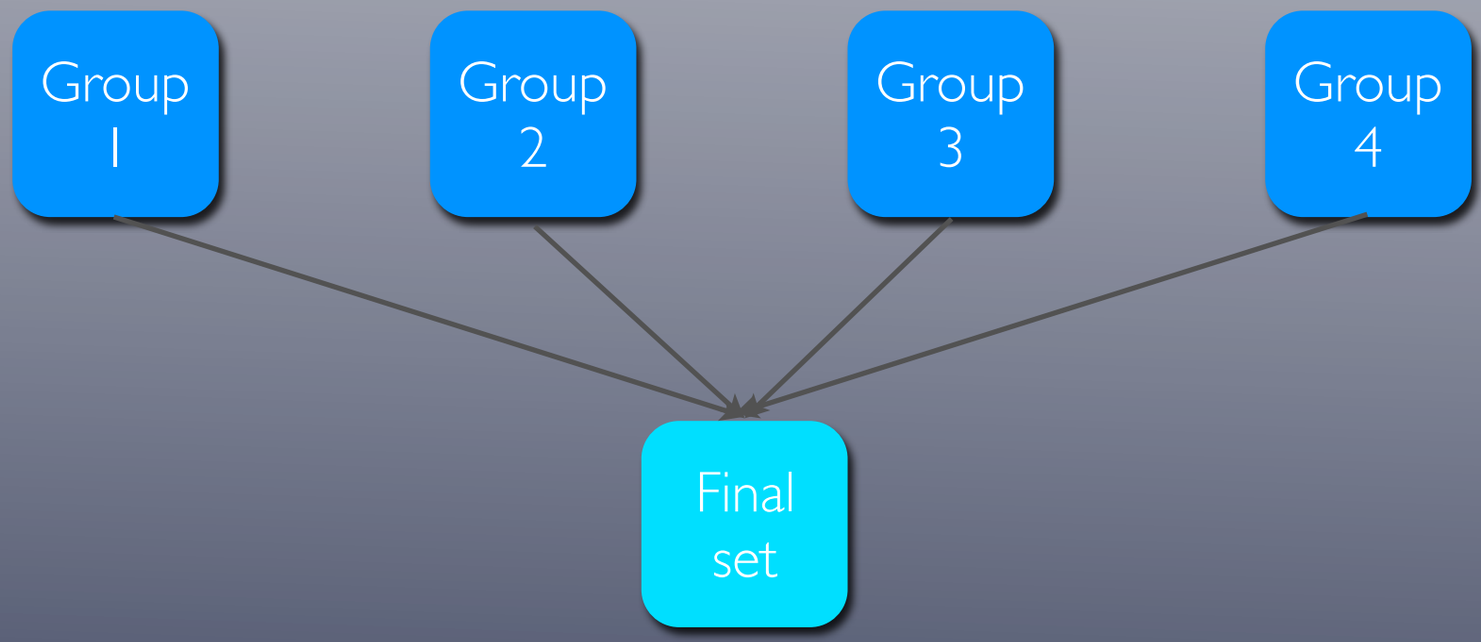


REPRODUCIBILITY

- Since *EQK*-means, even if combined with an alignment stability test, does not guarantee an optimum solution (global minimum) and stable groups can be fake, we require the solution to be reproducible over a number of quasi-independent runs.

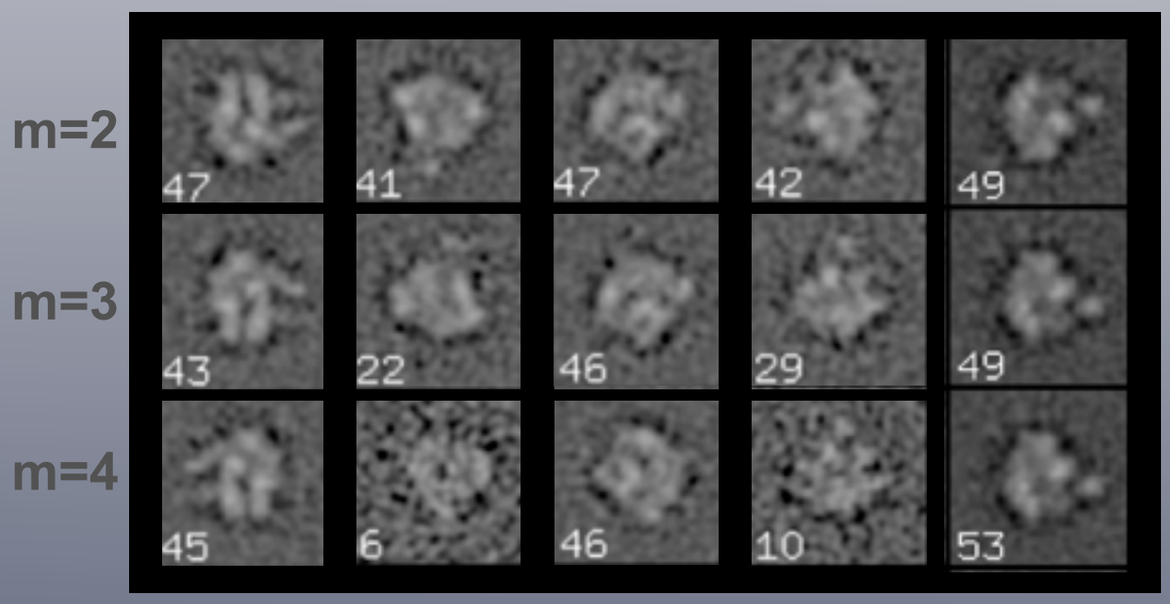
- We have m=4 *EQK*-means runs analyzing the data in parallel. Once all runs produce their respective averages, we compare assignments of images to class averages and select as reproducible subsets shared among quasi-independent runs.

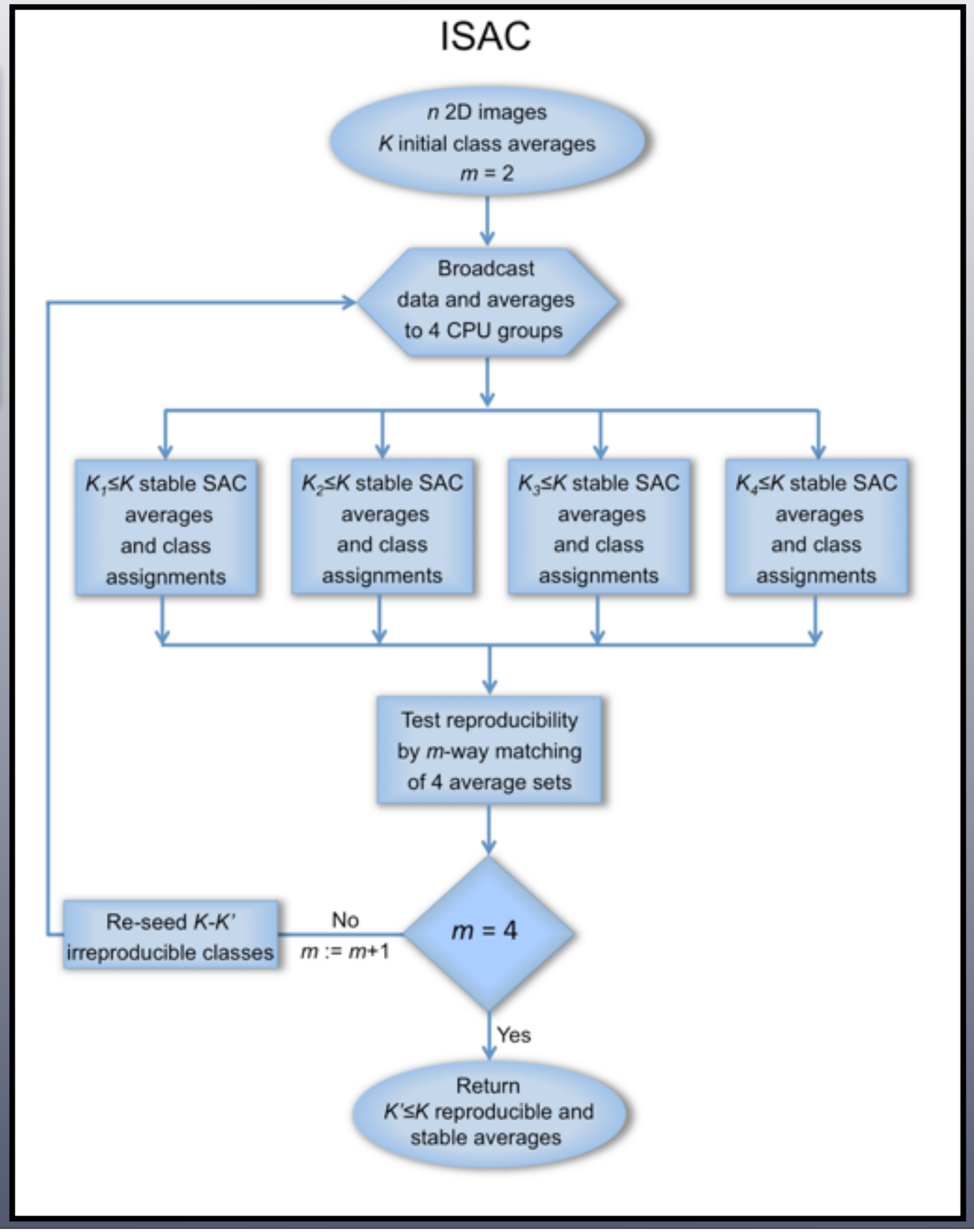
m= 4



ISAC: ITERATIVE STABLE ALIGNMENT AND CLUSTERING

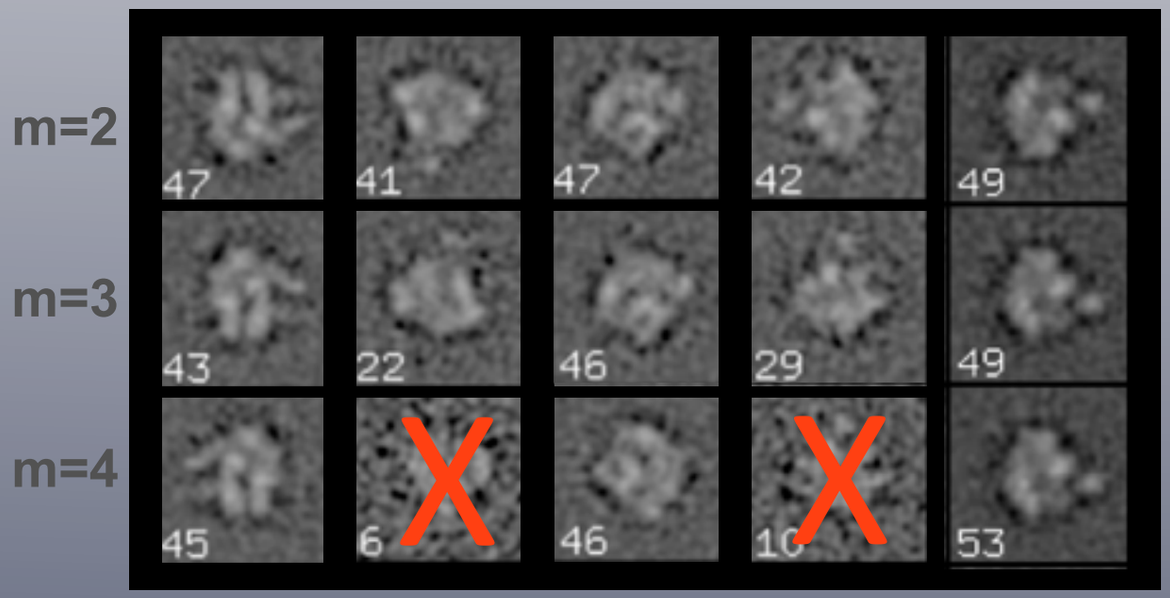
- We use 4 CPU groups to analyze the data set simultaneously
- Irreproducible averages are eliminated

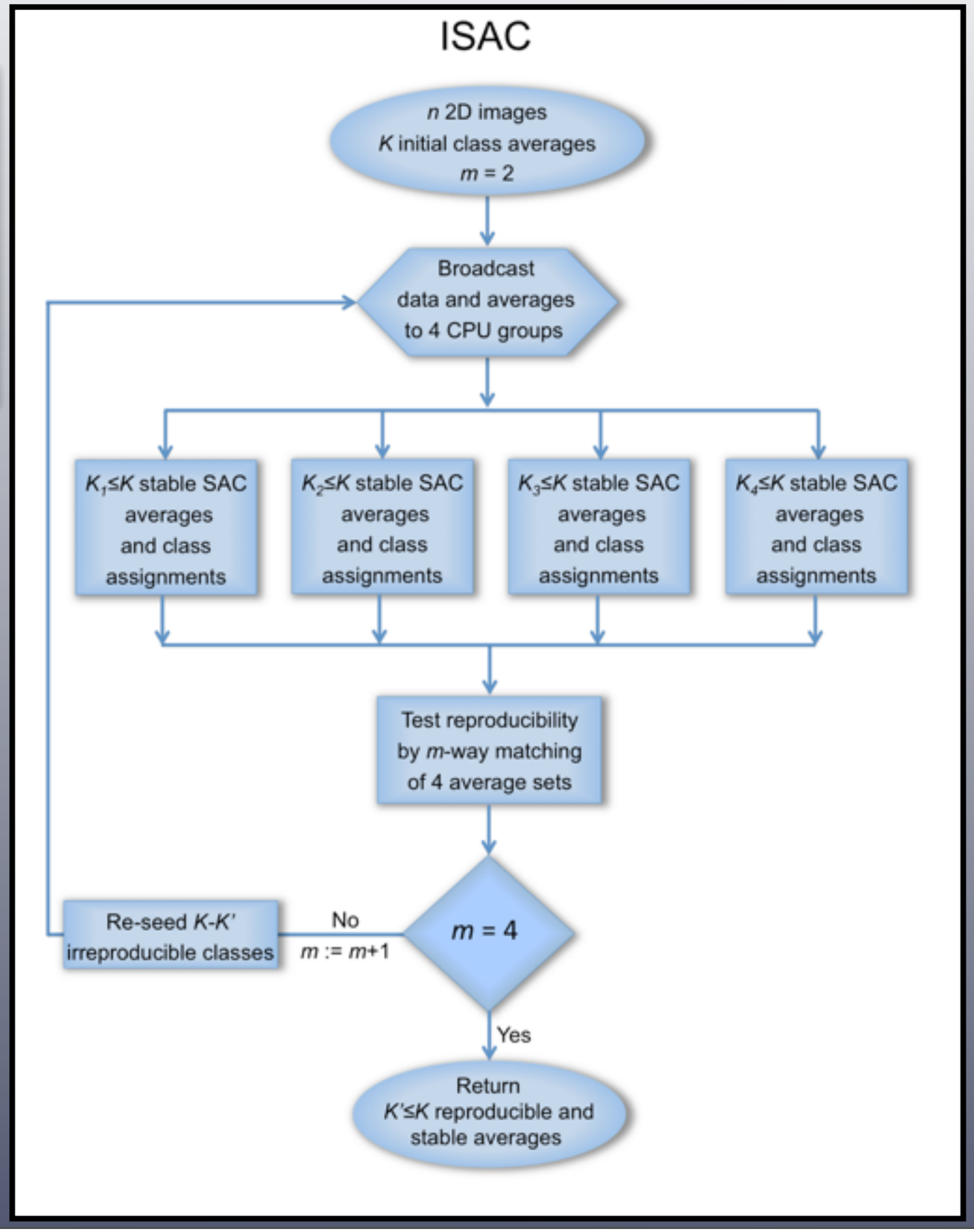


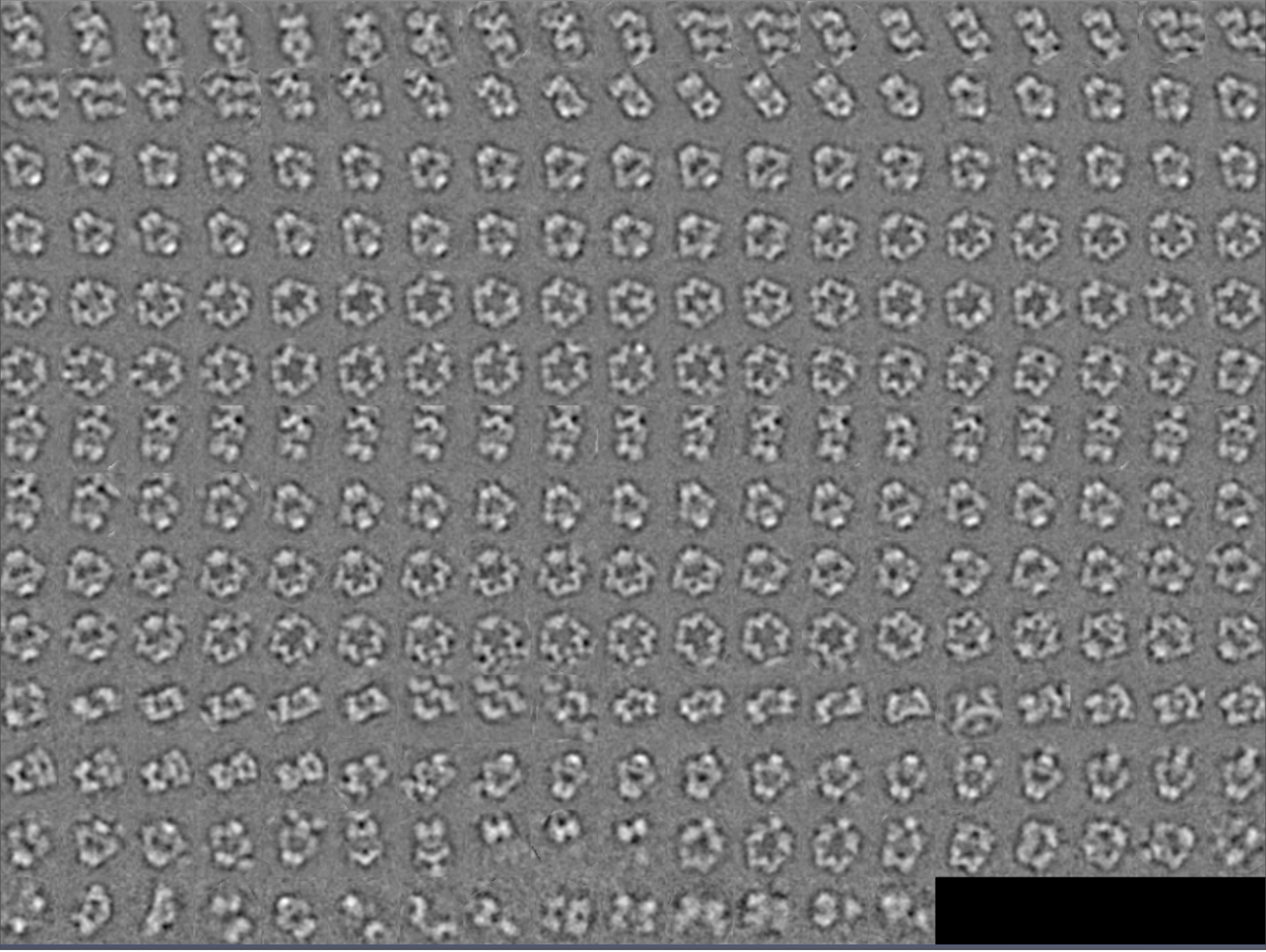


ISAC: ITERATIVE STABLE ALIGNMENT AND CLUSTERING

- We use 4 CPU groups to analyze the data set simultaneously
- Irreproducible averages are eliminated

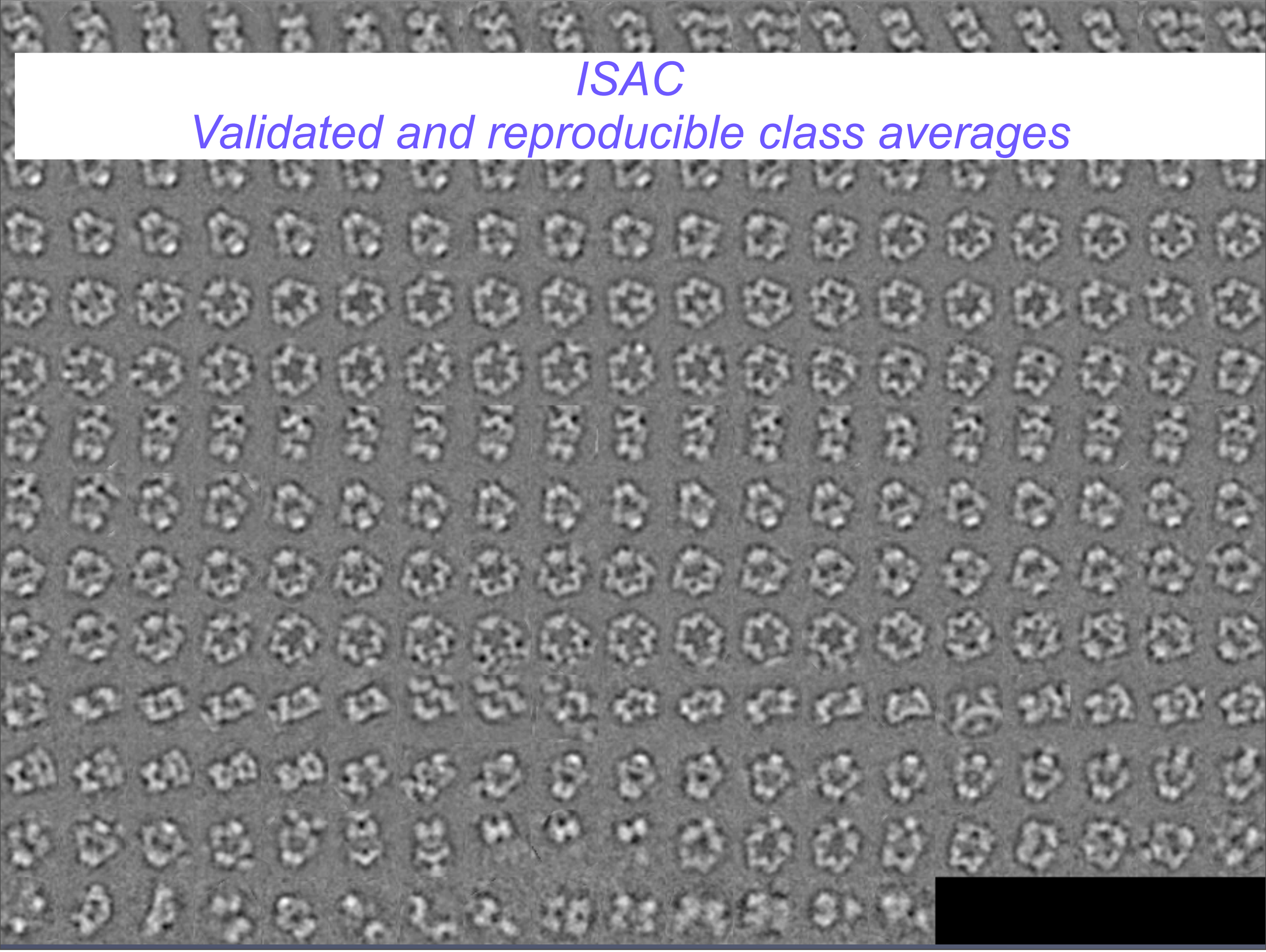






ISAC

Validated and reproducible class averages

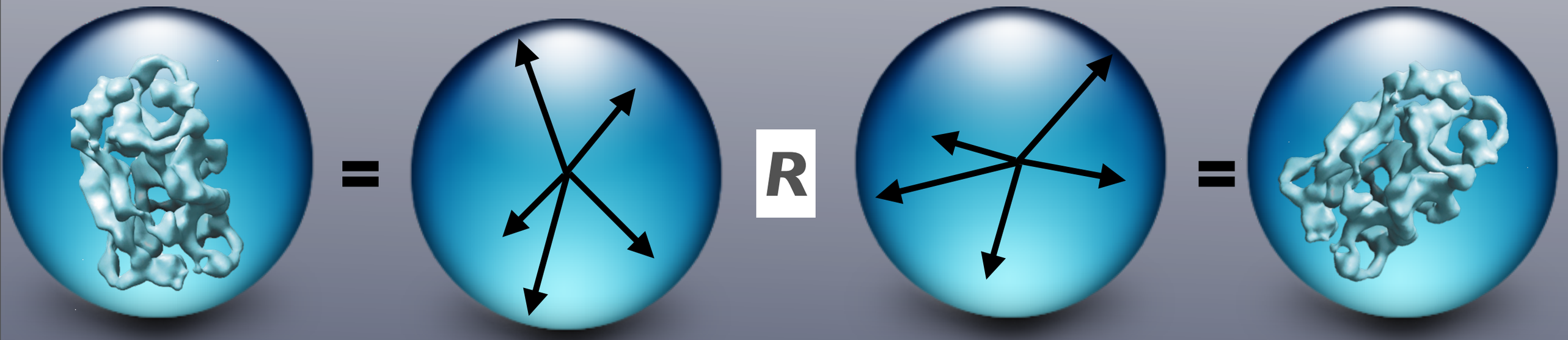


Constructive validation:
from *ab initio* EM map determination to map refinement

3D structure = 2D projection data + Orientation parameters

$(\phi, \theta), \psi, s_x, s_y$

$\boldsymbol{\tau}, \psi, s_x, s_y$

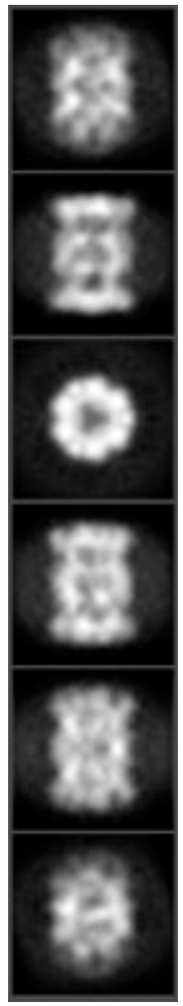


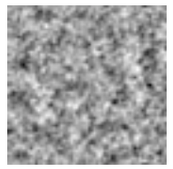
$$\max_{\mathbf{R}} \sum_{n=1}^{N} \boldsymbol{\tau}_n^{1\prime} \mathbf{R} \boldsymbol{\tau}_n^2$$

STEP 1: GENERATING A MAP

template structure

systematically generated reprojections $(\varphi,\theta)_k$





2D ccf

(ψ, s_x, s_y)

ccf_1

ccf_2

ccf_3

ccf_4

ccf_5

ccf_6

projection matching

best ➡ Orientation parameters.

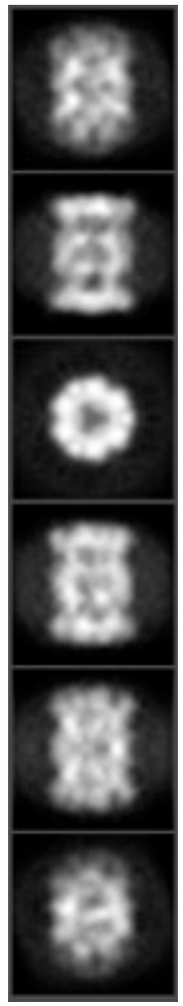
3D reconstruction from projections

low-pass filtration masking?

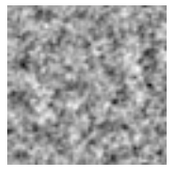
STEP 1: GENERATING A MAP

template structure

systematically generated reprojections $(\varphi,\theta)_k$



randomize order



2D ccf

(ψ, s_x, s_y)

ccf_1

ccf_2

ccf_3

.

.

.

SHC projection matching

ccf_n >previous best ➡ Orientation parameters. New best.

3D reconstruction from projections

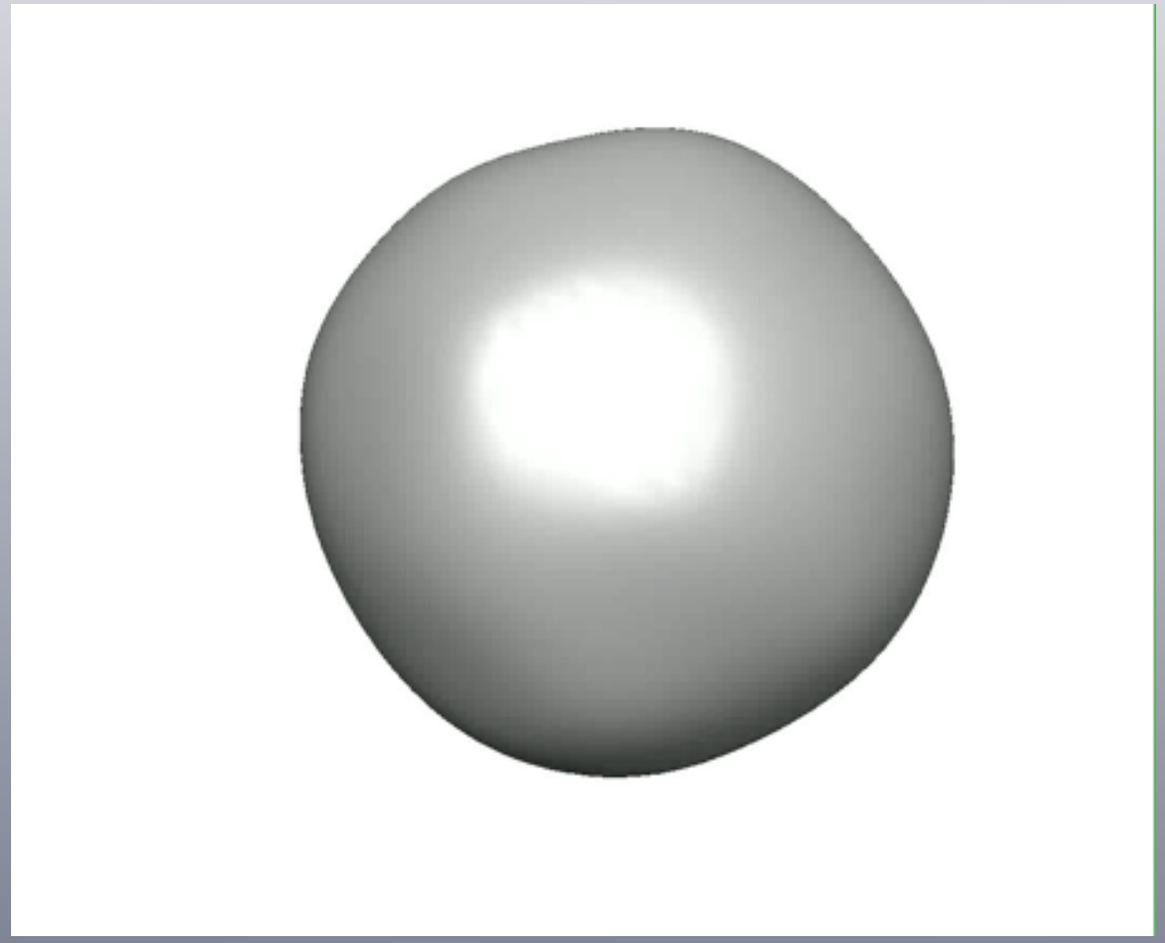
low-pass filtration masking?

H. Elmlund, D. Elmlund, S. Bengio, PRIME: probabilistic initial 3D model generation for single-particle cryo-electron microscopy, Structure, 21 (2013) 1299-1306.

SHC - CONVERGENCE

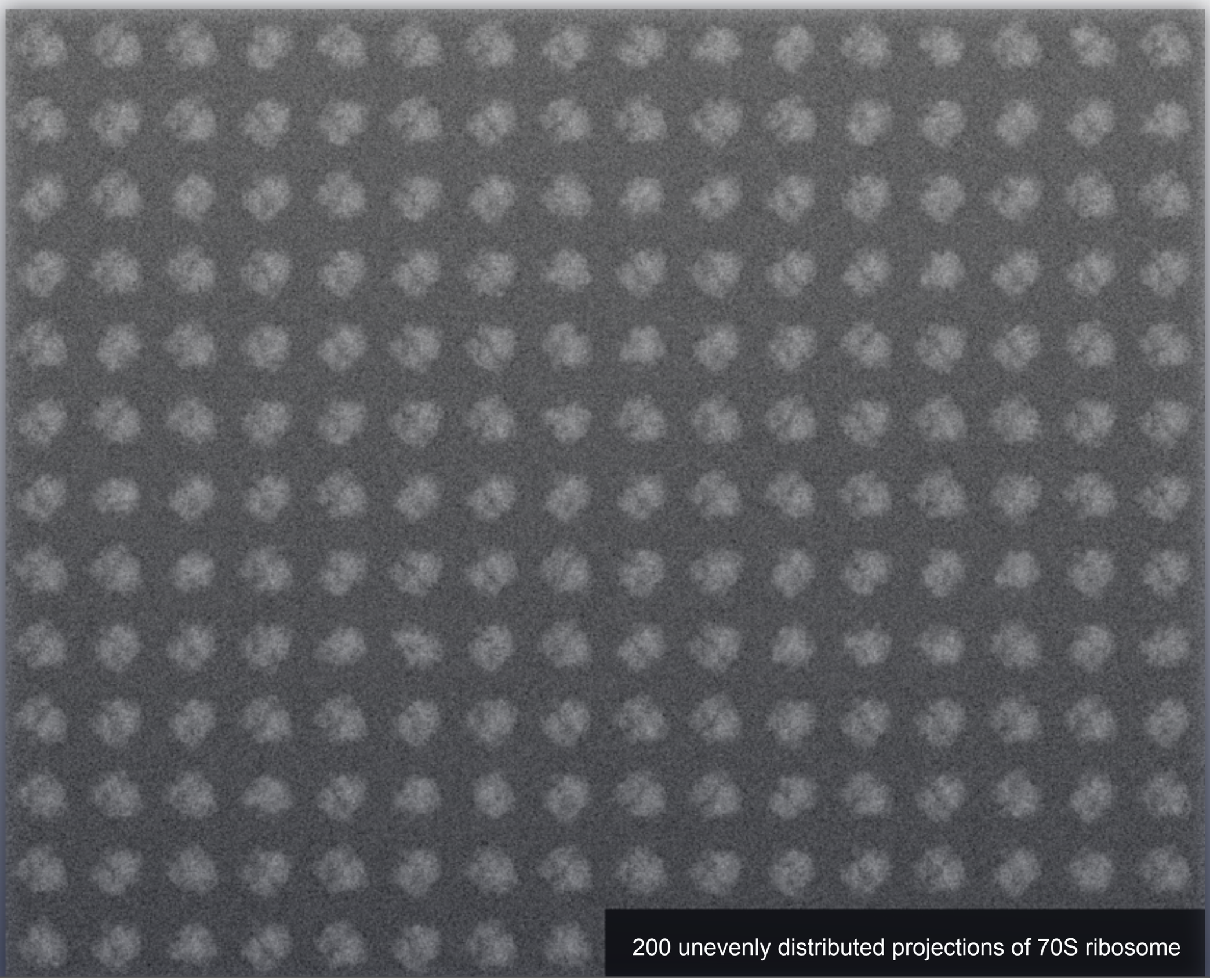
H. Elmlund, D. Elmlund, S. Bengio, PRIME: probabilistic initial 3D model generation for single-particle cryo-electron microscopy, Structure, 21 (2013) 1299-1306.

SHC - CONVERGENCE



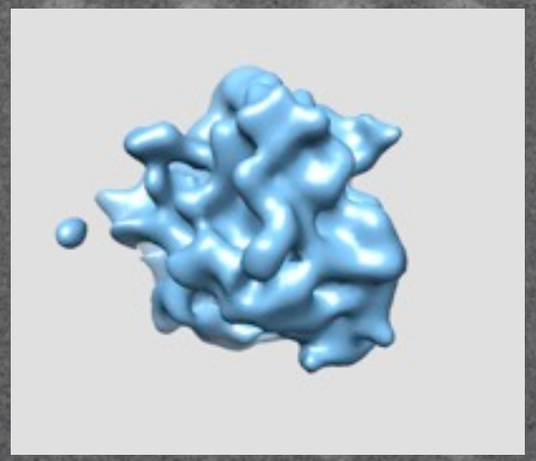
H. Elmlund, D. Elmlund, S. Bengio, PRIME: probabilistic initial 3D model generation for single-particle cryo-electron microscopy, Structure, 21 (2013) 1299-1306.

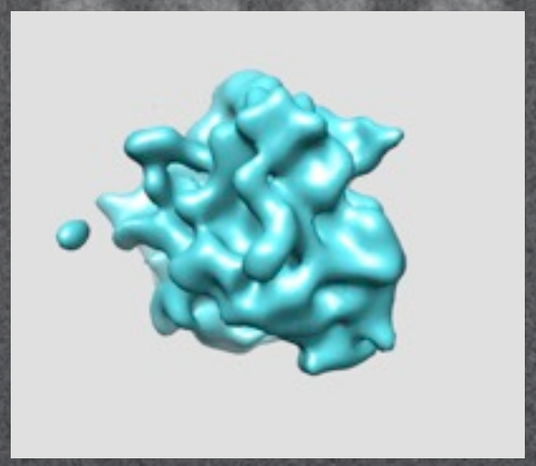
OVERCOMING SHC CONVERGENCE LIMITATIONS BY MONITORING PARAMETER REPRODUCIBILITY

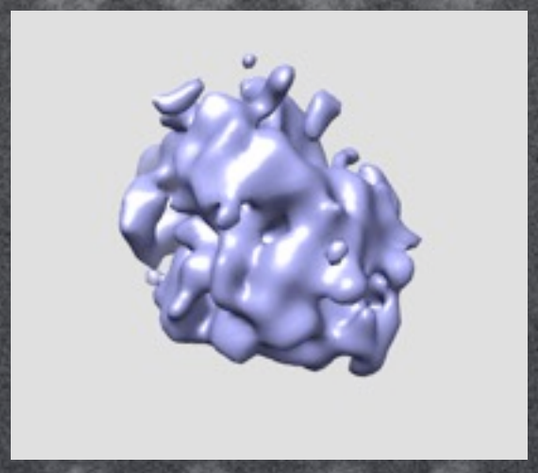


200 unevenly distributed projections of 70S ribosome

OVERCOMING SHC CONVERGENCE LIMITATIONS BY MONITORING PARAMETER REPRODUCIBILITY







GOOD:

- No bias towards the initial structure, in normal use always randomized start
- Often converges to a plausible solution
- Very good for structure refinement

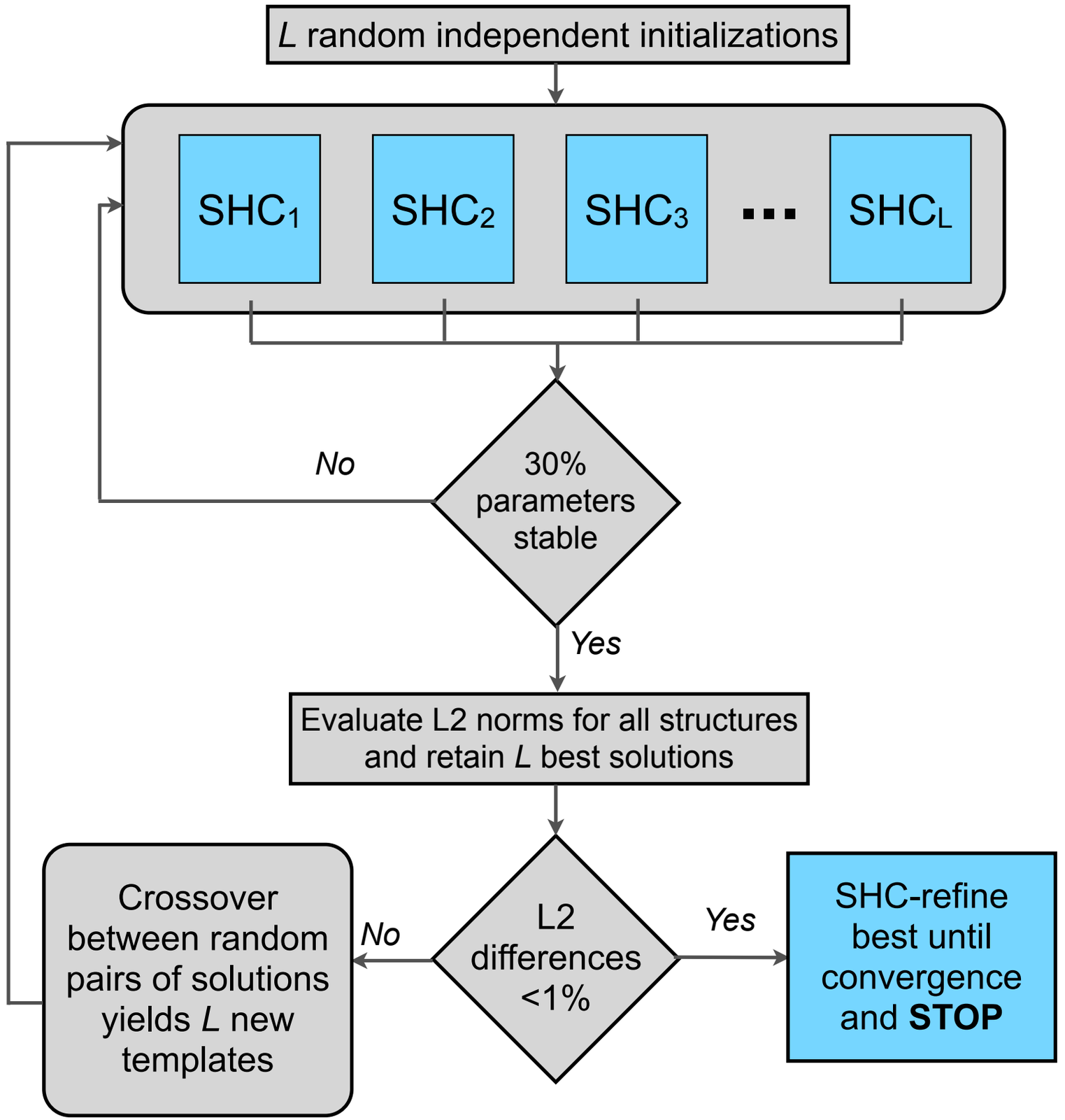
NOT SO GOOD:

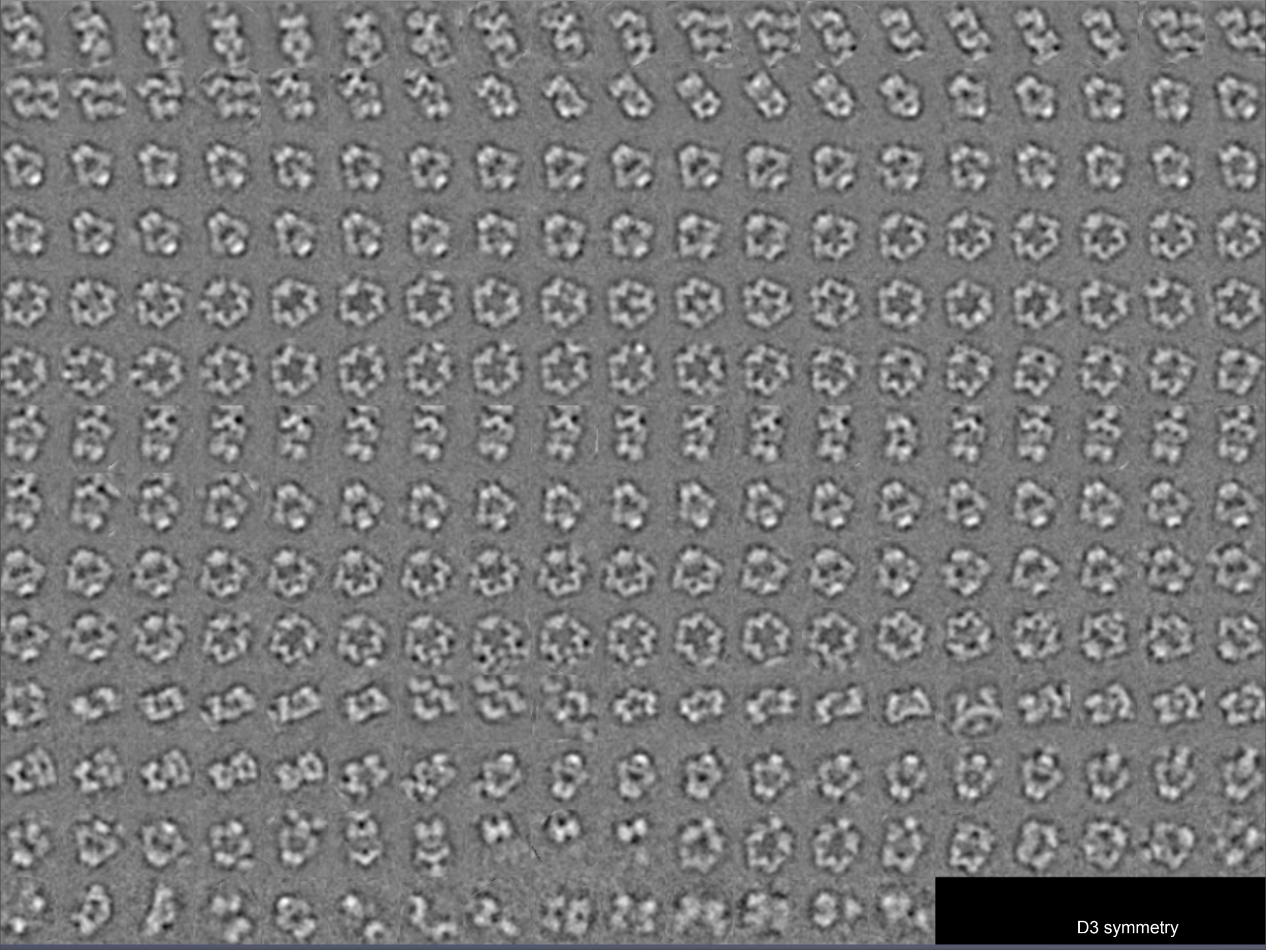
- Convergence properties poorly characterized/ understood, unclear how often it converges and what does it depend on
- Sometimes gets stuck in a completely wrong solution
- Plausible solutions somewhat different

200 unevenly distributed projections of 70S ribosome

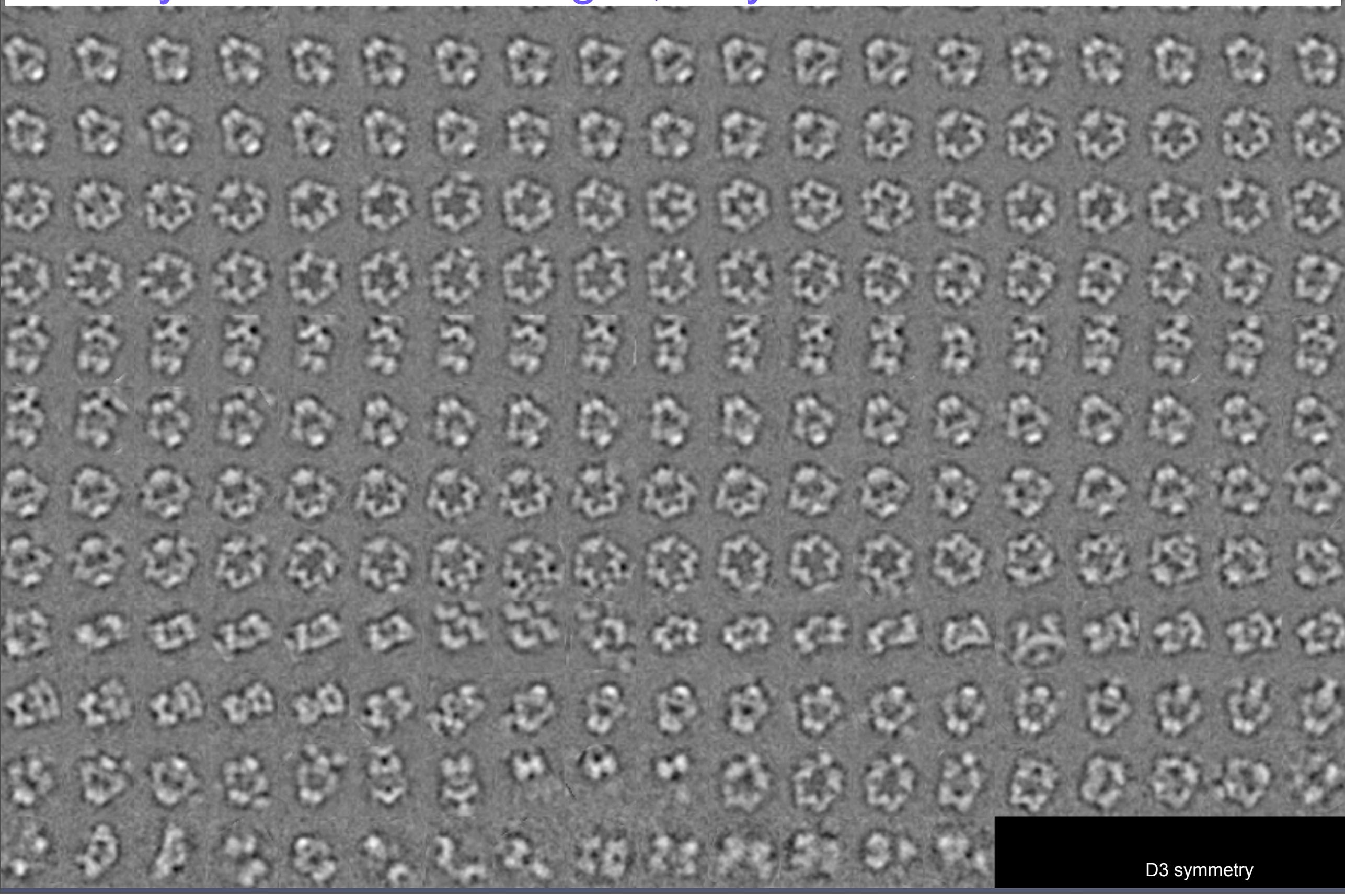
STEP 2: VIPER

(Validation of Individual Parameter Reproducibility)



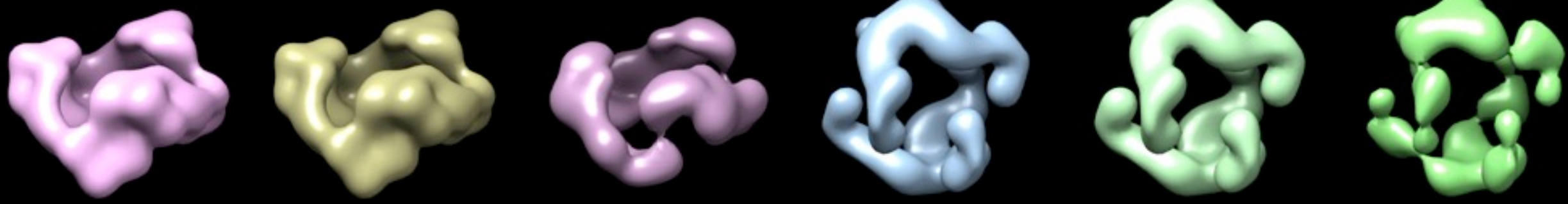

D3 symmetry

Ab initio structure determination with ISAC/VIPER: only the correct averages, only the correct structure



Ab initio structure determination with ISAC/VIPER: only the correct averages, only the correct structure

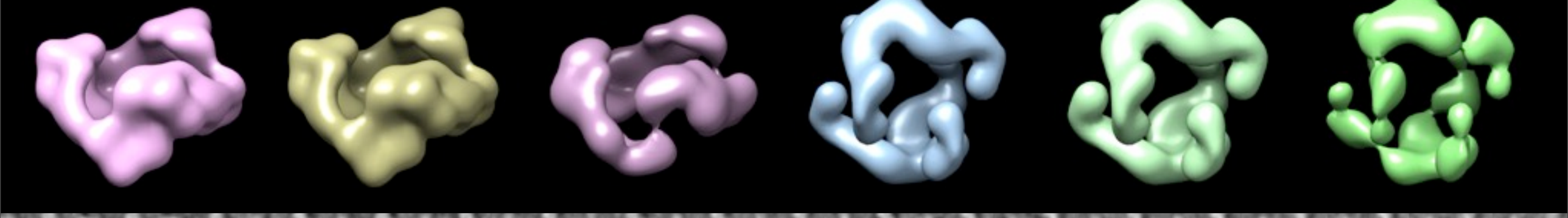
GA - generation I



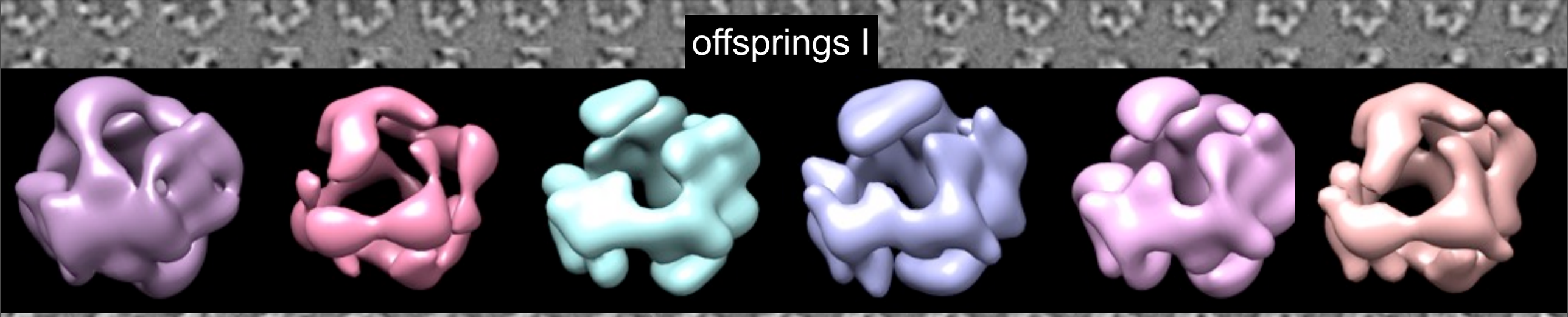
D3 symmetry

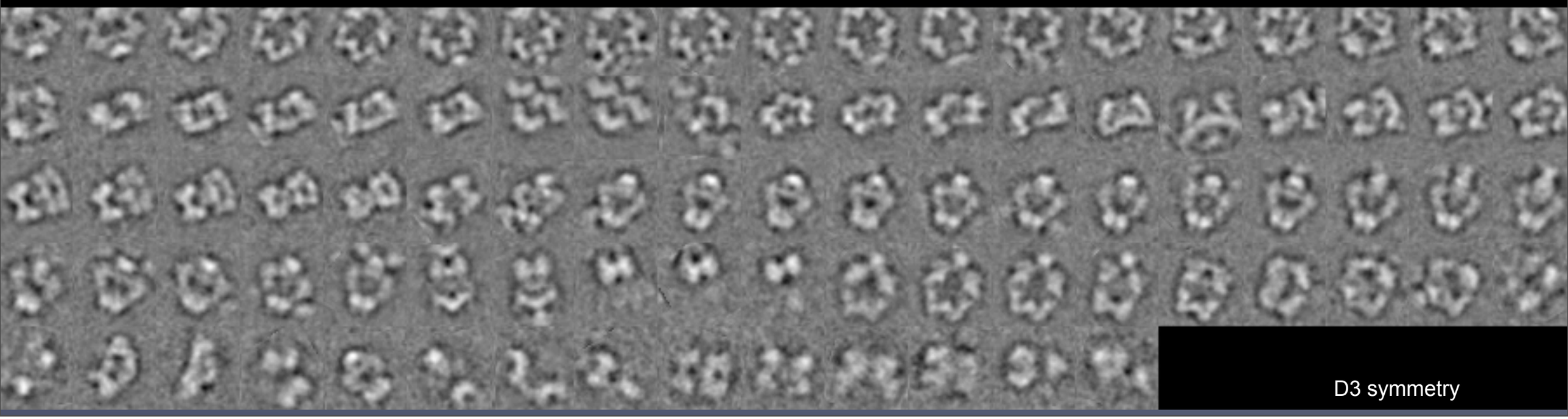
Ab initio structure determination with ISAC/VIPER: only the correct averages, only the correct structure

GA - generation I



offsprings I

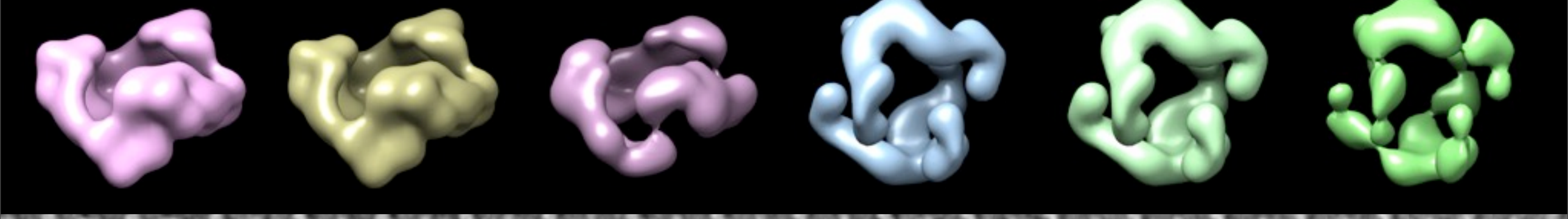




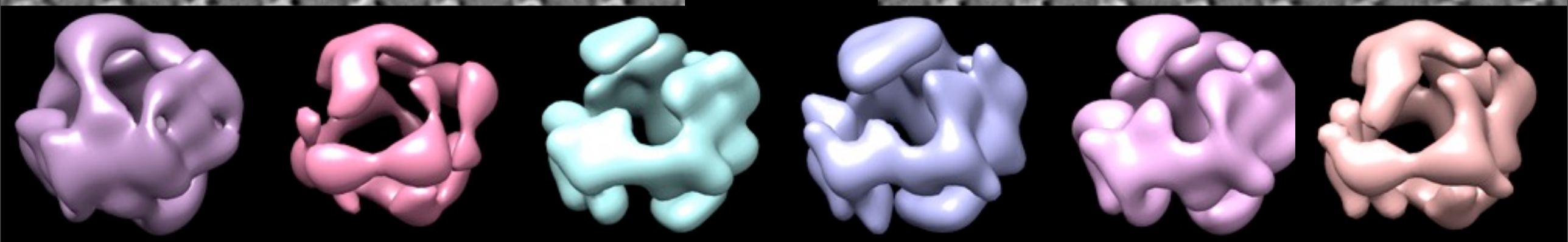
D3 symmetry

Ab initio structure determination with ISAC/VIPER: only the correct averages, only the correct structure

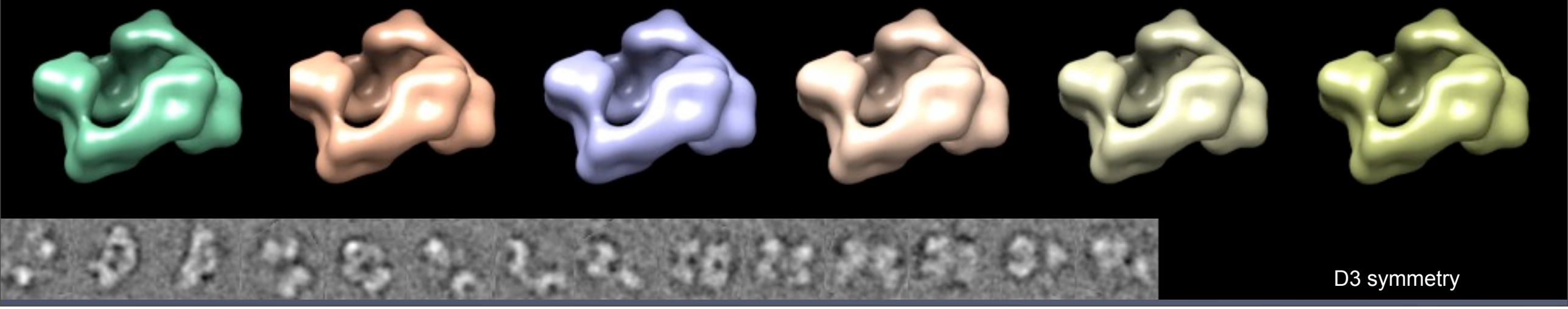
GA - generation I



offsprings I

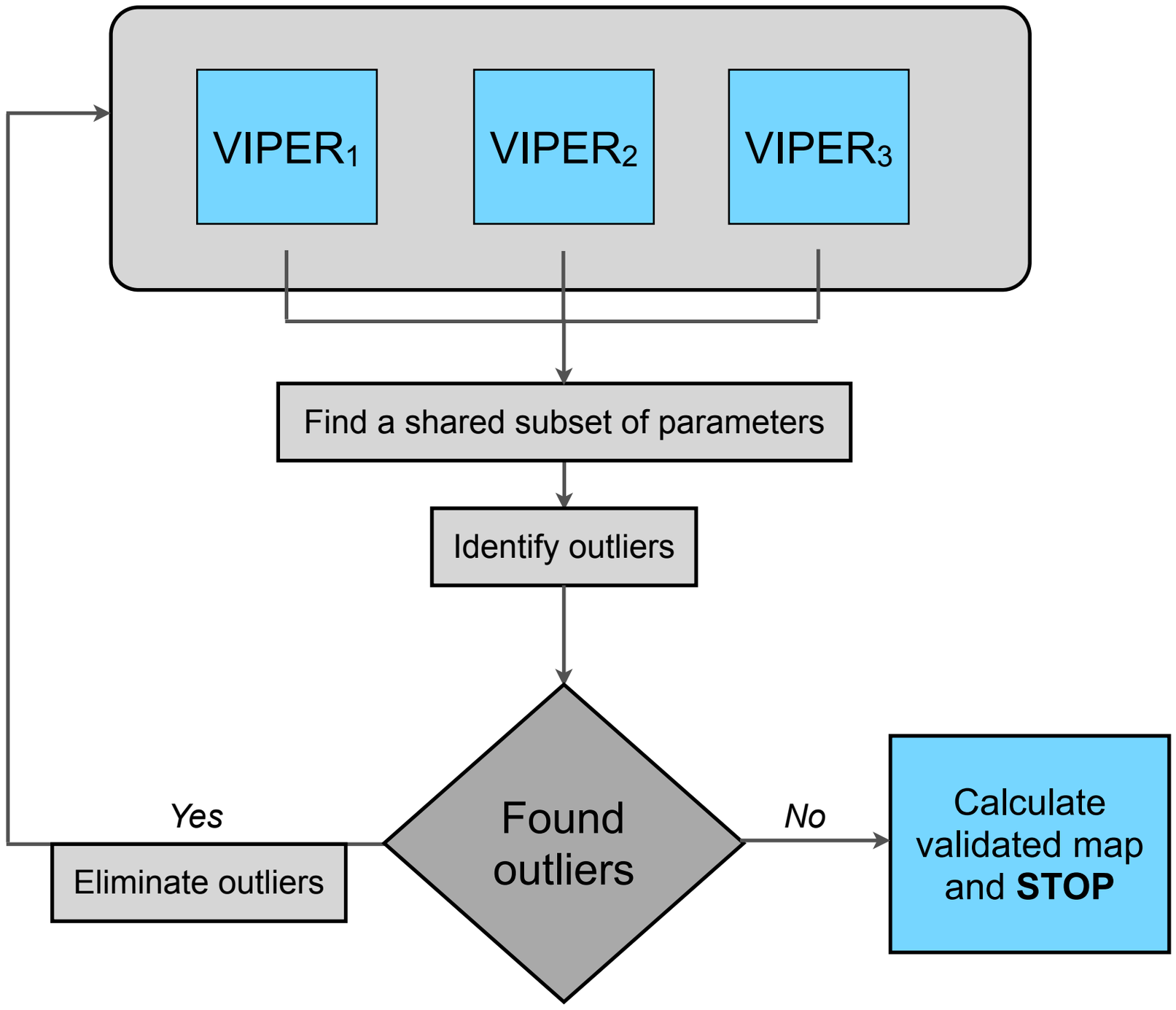


GA - generation II

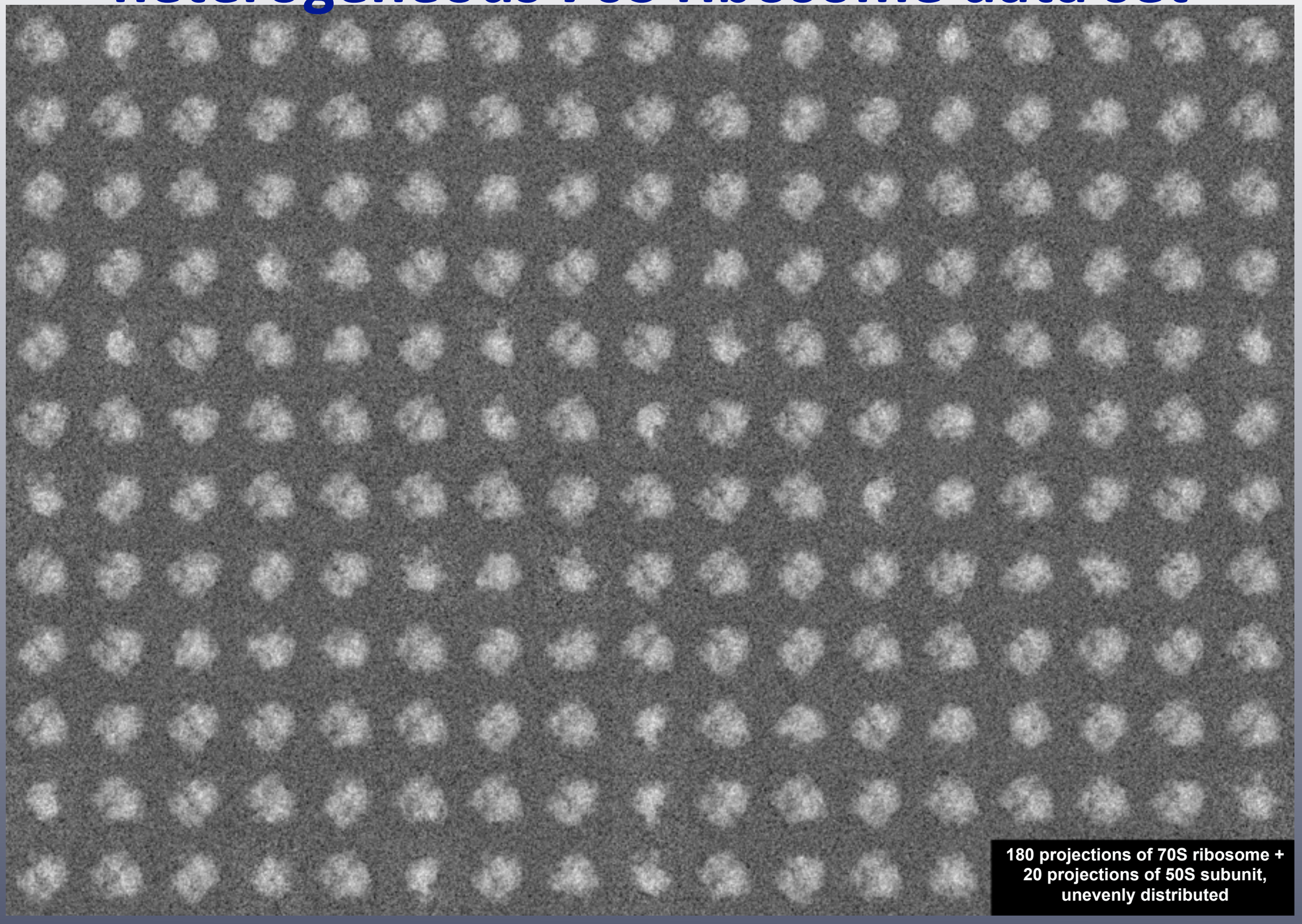


D3 symmetry

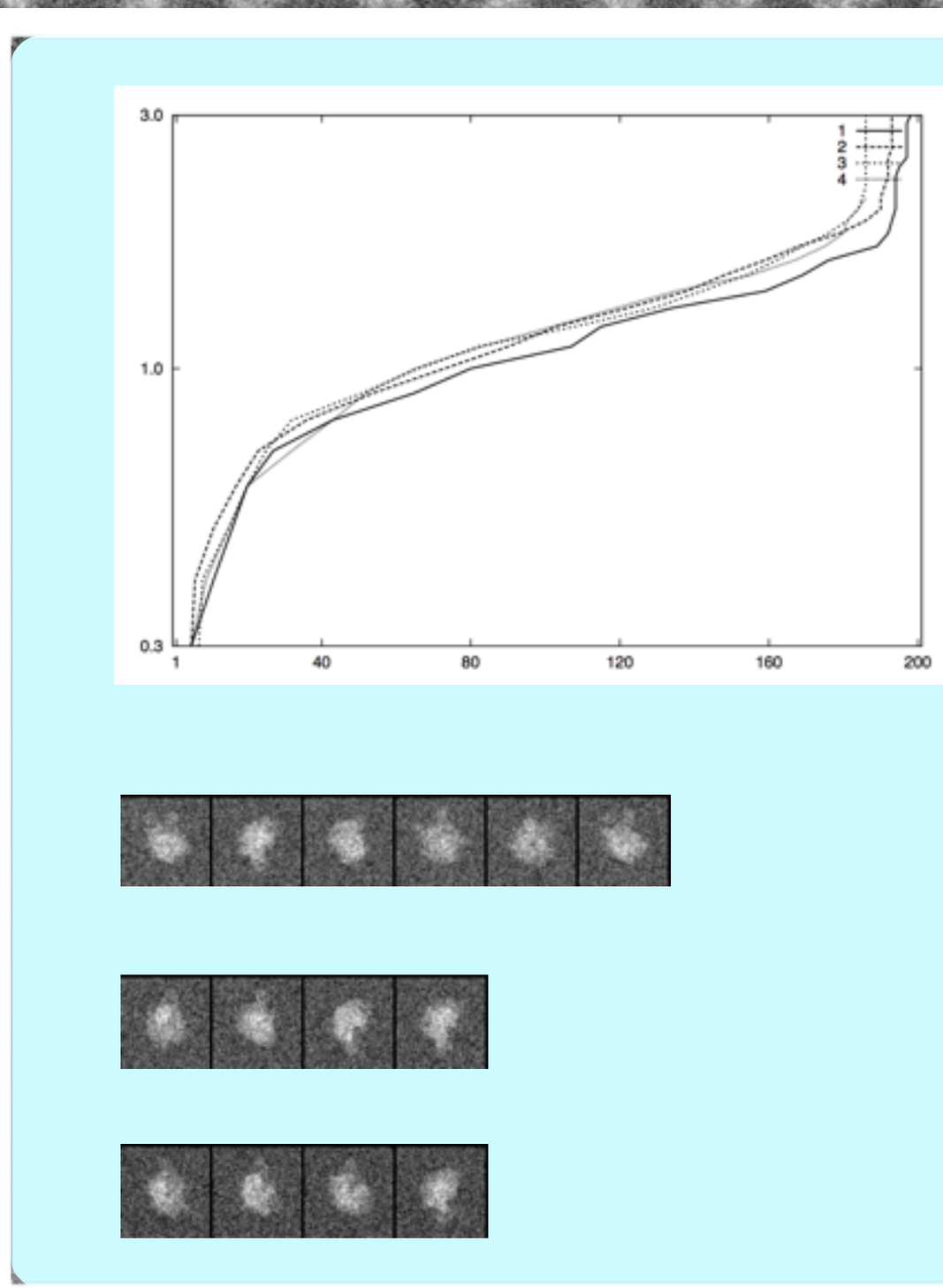
STEP 3: VALIDATION WITH REPRODUCIBLE VIPER
R-VIPER YIELDS A VALIDATED *ab initio* MAP



Application of VIPER to a simulated heterogeneous 70S ribosome data set



Application of VIPER to a simulated heterogeneous 70S ribosome data set



180 projections of 70S ribosome + 20 projections of 50S subunit, unevenly distributed