Supplemental Information

Separating distinct structures of multiple macromolecular assemblies from cryo-EM projections

Eric J. Verbeke[1-3], Yi Zhou[1-3], Andrew P. Horton[1-3], Anna L. Mallam[1-3], David W. Taylor[1–4*], Edward M. Marcotte[1–3*]

Affiliations:

[1]Department of Molecular Biosciences

[2]Center for Systems and Synthetic Biology

[3]Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, TX 78712, USA

[4]LIVESTRONG Cancer Institutes, Dell Medical School, Austin, TX 78712, USA

*Correspondence: marcotte@icmb.utexas.edu (E.M.M), dtaylor@utexas.edu (D.W.T)

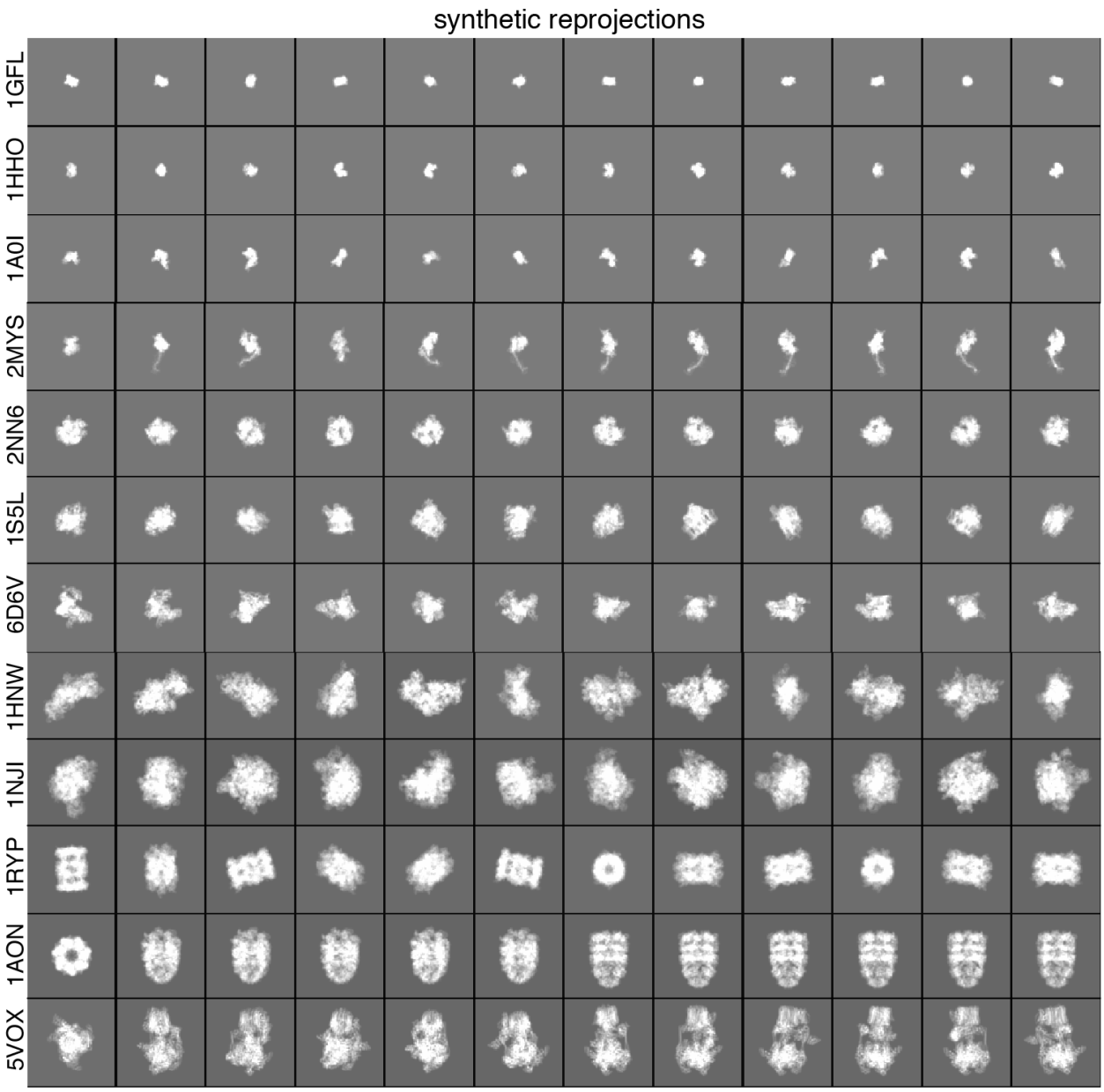


Figure S1. 2D reprojections from synthetic dataset

(A) Subset of 2D reprojections from 12 of the 35 structures in our synthetic dataset. Box size corresponds to 300 Å.

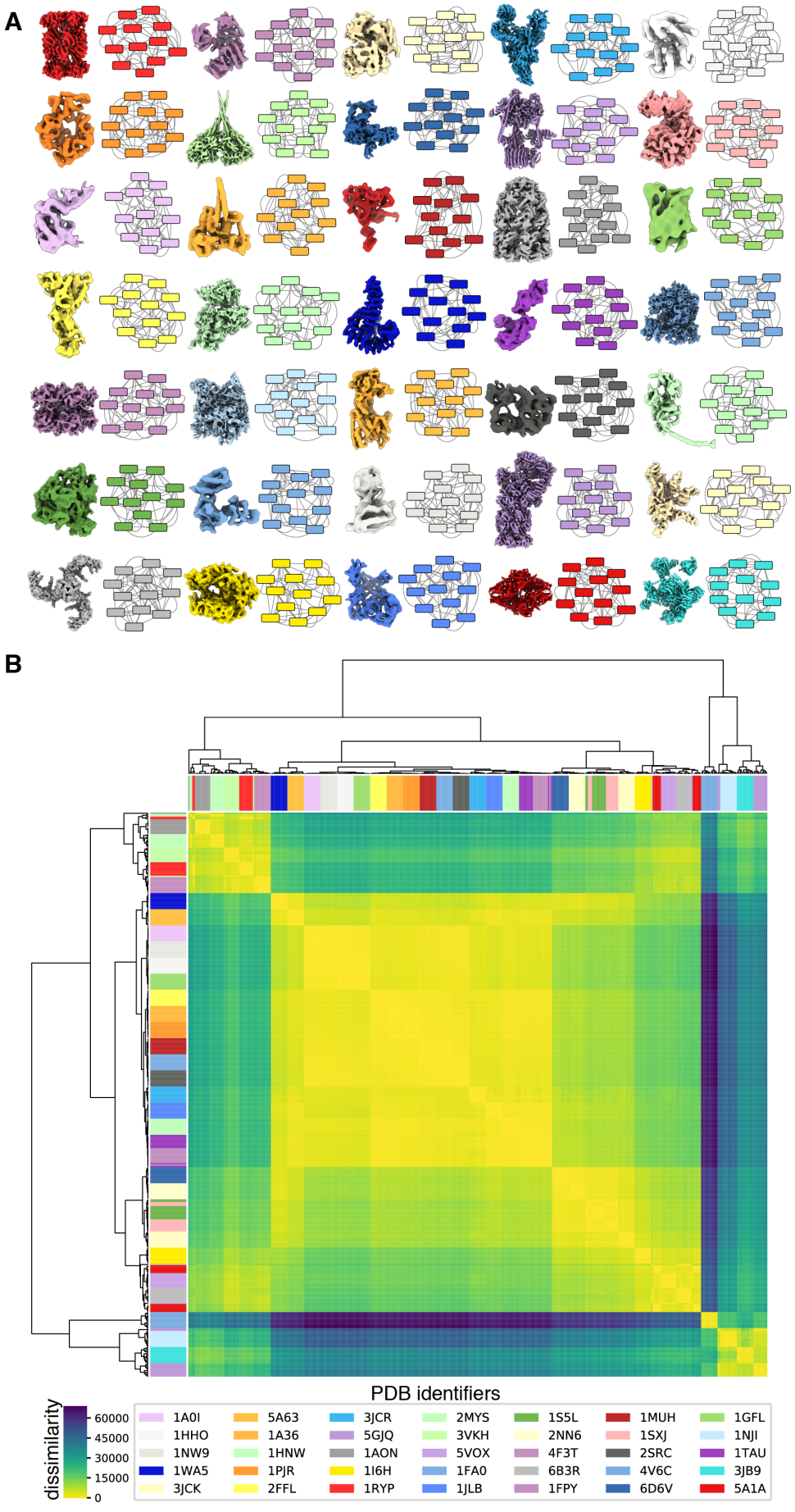

A
B
PDB identifiers
dissimilarity
60000
45000
30000
15000
0
1A0I
1HHO
1NW9
1WA5
3JCK
5A63
1A36
1HNW
1PJR
2FFL
3JCR
5GJQ
1AON
1I6H
1RYP
2MYS
3VKH
5VOX
1FA0
1JLB
1S5L
2NN6
4F3T
6B3R
1FPY
1MUH
1SXJ
2SRC
4V6C
6D6V
1GFL
1NJI
1TAU
3JB9
5A1A

Figure S2. Synthetic dataset network and clustergram

(A) Network displaying communities of 2D reprojection images determined using SLICEM. Each node represents a 2D reprojection with 5 connecting edges to the most similar reprojections as scored using Euclidean distance. The color of each node matches the structure from which it was reprojected (shown as a surface). (B) Clustergram showing the block structure of similarity between synthetic 2D reprojections based on the scoring of their 1D line projections by Euclidean distance. The distance metric used for hierarchical clustering was Euclidean and the linkage method used was “average”. Row and column colors correspond to PDB structure identity and individual pixels reflect the dissimilarity between them (i.e. the lower the dissimilarity, the better the match between 1D line projections).

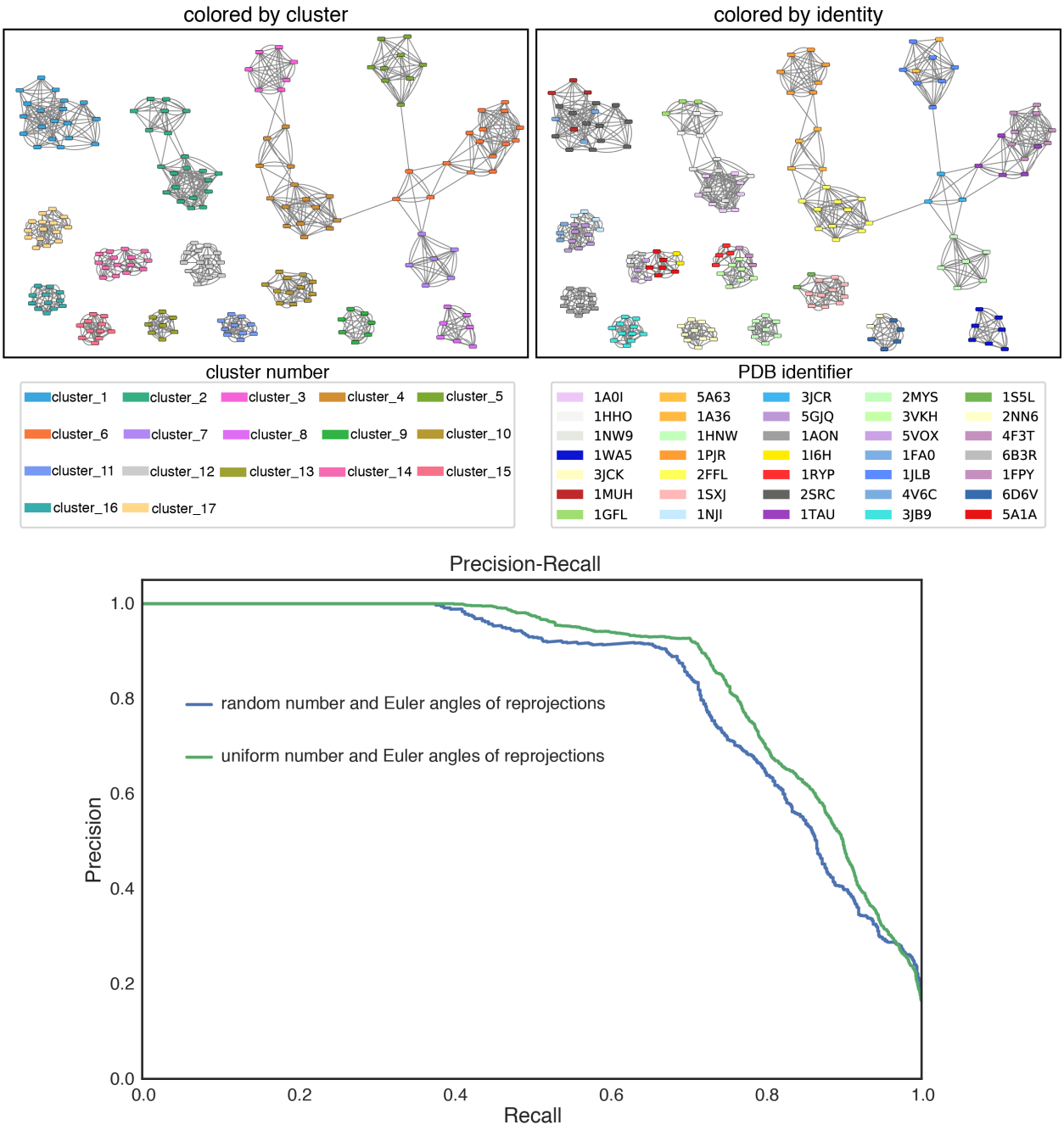


Figure S3. Effect of non-uniform projection angles and number of projections

The 35 PDB structures from Figure 2A were randomly reprojected between 2-12 times and at random Euler angles each between 0 degrees – 360 degrees. These reprojections were then subjected to our SLICEM algorithm using Euclidean distance to score 1D line projections and walktrap for clustering. The side-by-side networks show the nodes labeled after clustering and

the nodes labeled by PDB identity. Projections belonging to the same PDB structure are well grouped but not always clustered correctly. Precision and recall was used to compare the non-uniform dataset to the uniform dataset and shows only a slight decrease in performance for the more realistic case of non-uniform projections.

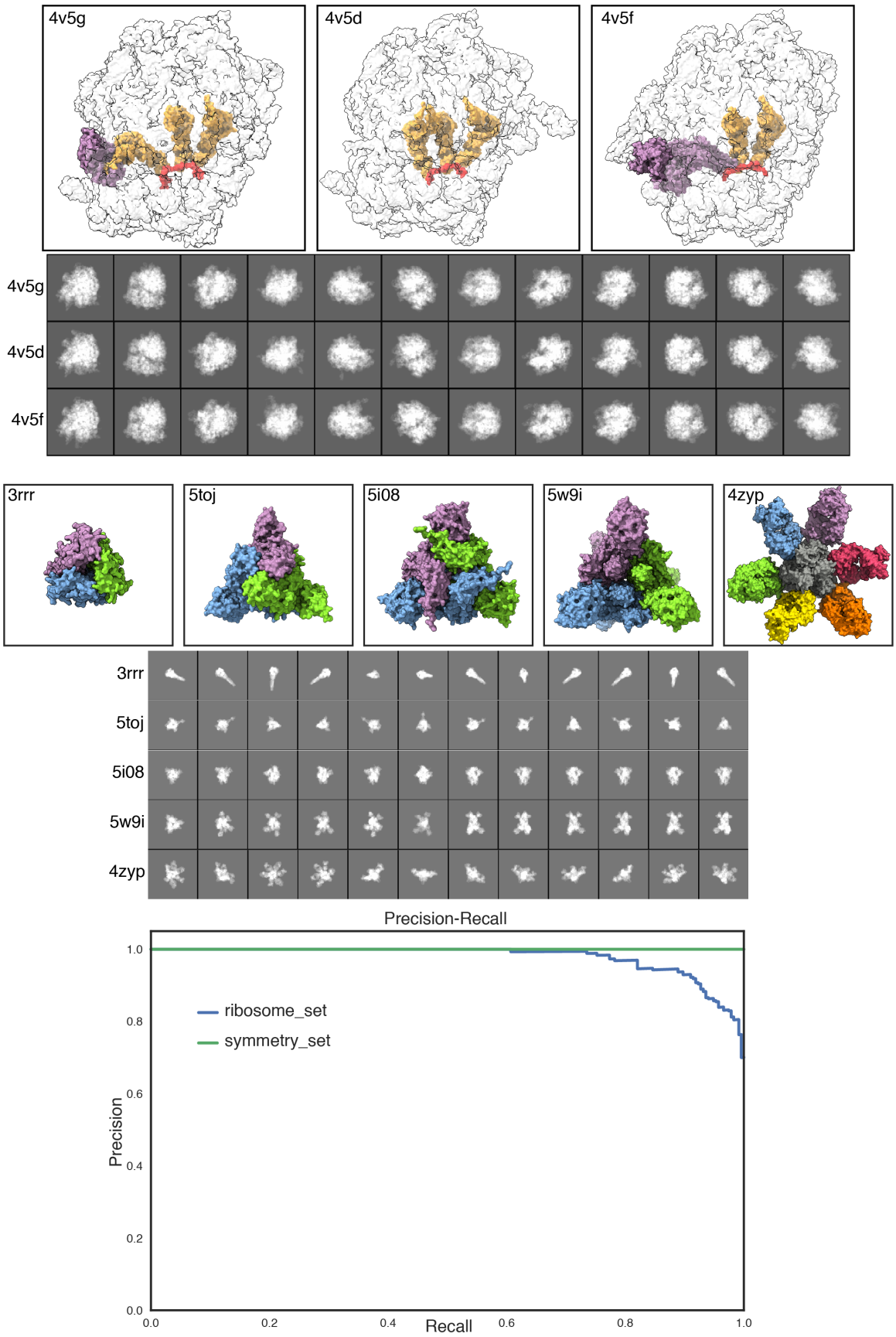
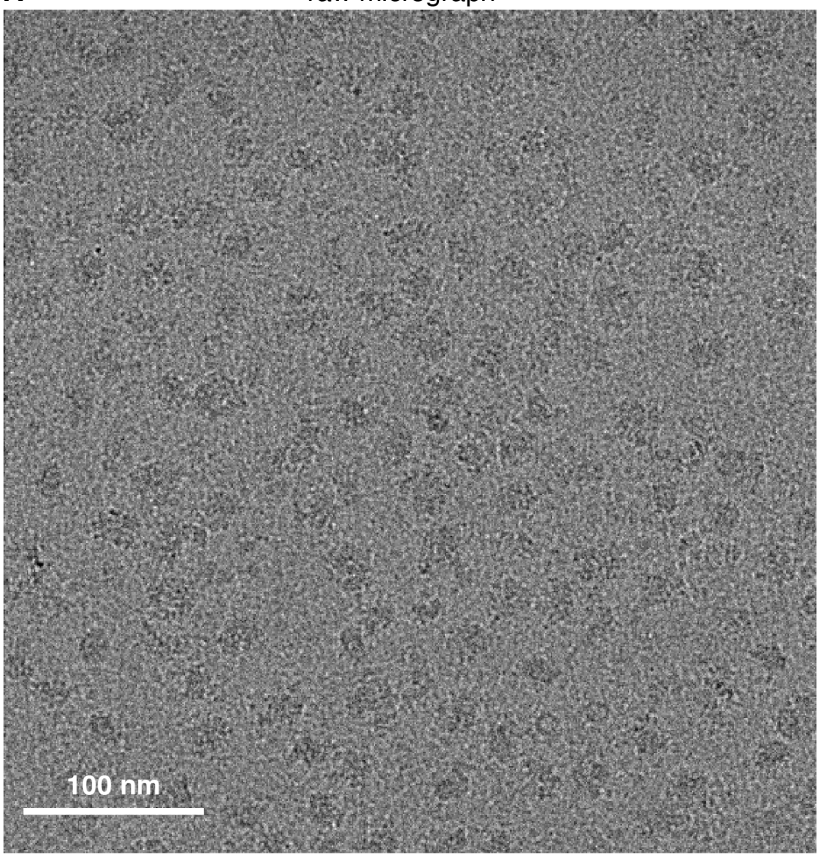

4v5g
4v5d
4v5f
4v5g
4v5d
4v5f
3rrr
5toj
5i08
5w9i
4zyp
3rrr
5toj
5i08
5w9i
4zyp
Precision-Recall
ribosome_set
symmetry_set
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

Figure S4. Mixtures with molecular symmetries or conformational and compositional heterogeneity

Uniform reprojections of a ribosome, a ribosome with EF-Tu, and a ribosome with EF-G (PDB 4V5D, PDB 4V5G, PDB 4V5F) low-pass filtered to 9 Å were created to test the effect of conformational and compositional heterogeneity on our algorithm. Similarly, uniform reprojections of five C-3 symmetric structures (PDB 3RRR, PDB 5TOJ, PDB 5I08, PDB 5W9I, PDB 4ZYP) were created to test the effect of molecular symmetry on our algorithm. The box size of all reprojections is 350 Å. The similarity of 1D line projections between 2D reprojections in each set were calculated using Euclidean distance and performance was measured using precision and recall. Here, the high scoring precision-recall curves indicate that common line scores from projections of the same structure outperform scores to similar, but slightly different structures.

A raw micrograph



B reference-free 2D class averages

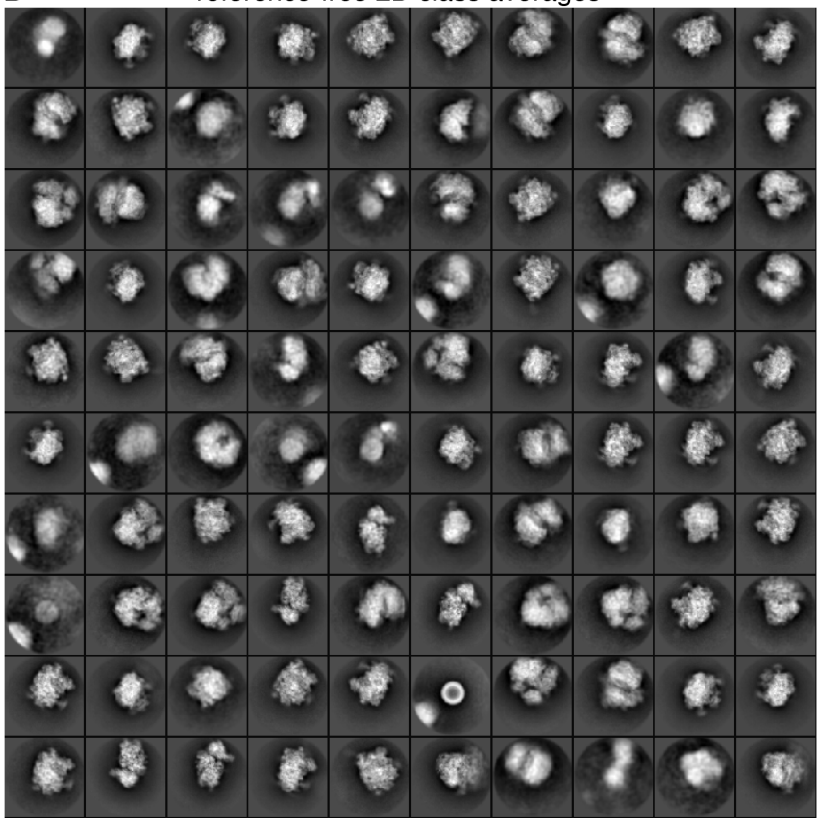


Figure S5. 2D classification of particles using RELION

(A) Representative raw micrograph of a mixture containing 40S, 60S and 80S ribosomes, apoferritin and β-galactosidase. (B) Reference-free 2D class averages generated using RELION of ~203,000 template-picked particle images. Box size corresponds to 422 Å.

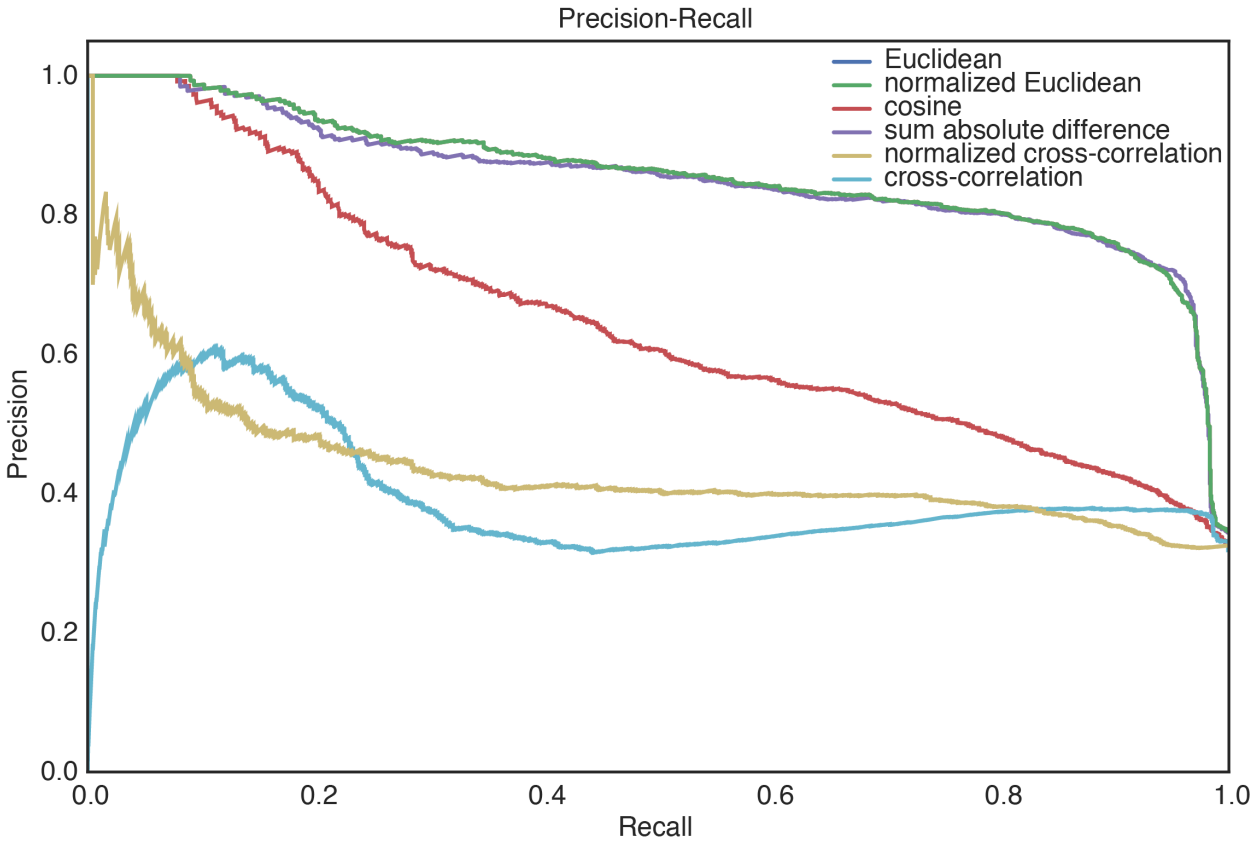


Figure S6. Precision-recall curves for experimental cryo-EM data

Precision-recall plot displaying 6 different metrics for scoring the similarity between 1D line projections from the entire set of 2D class averages. Euclidean and normalized Euclidean metrics scored identically and are overlaid.

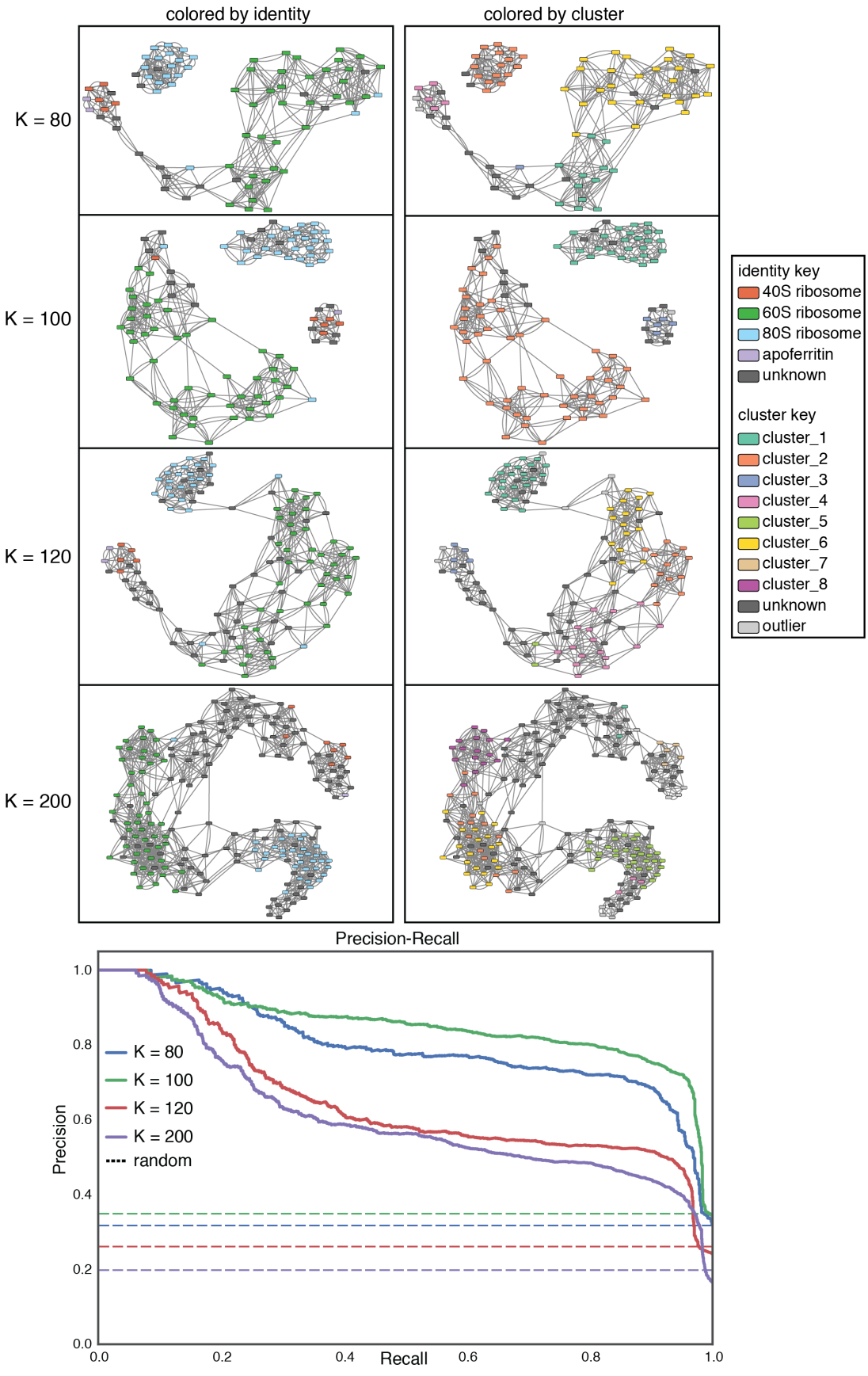

colored by identity
colored by cluster
K = 80
K = 100
K = 120
K = 200
identity key
40S ribosome
60S ribosome
80S ribosome
apoferritin
unknown
cluster key
cluster_1
cluster_2
cluster_3
cluster_4
cluster_5
cluster_6
cluster_7
cluster_8
unknown
outlier
Precision-Recall
Precision
Recall
K = 80
K = 100
K = 120
K = 200
random
0.0
0.2
0.4
0.6
0.8
1.0

Figure S7. Effect of varying the number of 2D class averages

Using RELION, 2D class averages with K = 80, 100, 120 and 200 were created from the final stack of ~203,000 particle images. The 2D class averages were then subjected to our SLICEM algorithm using the sum of absolute difference to score 1D line projections and edge betweenness for clustering. The side-by-side networks show the nodes labeled by manually annotated identity and by cluster. Precision and recall was used to compare the performance at various numbers of 2D classes. All precision-recall curves show substantial improvement over random assignment.

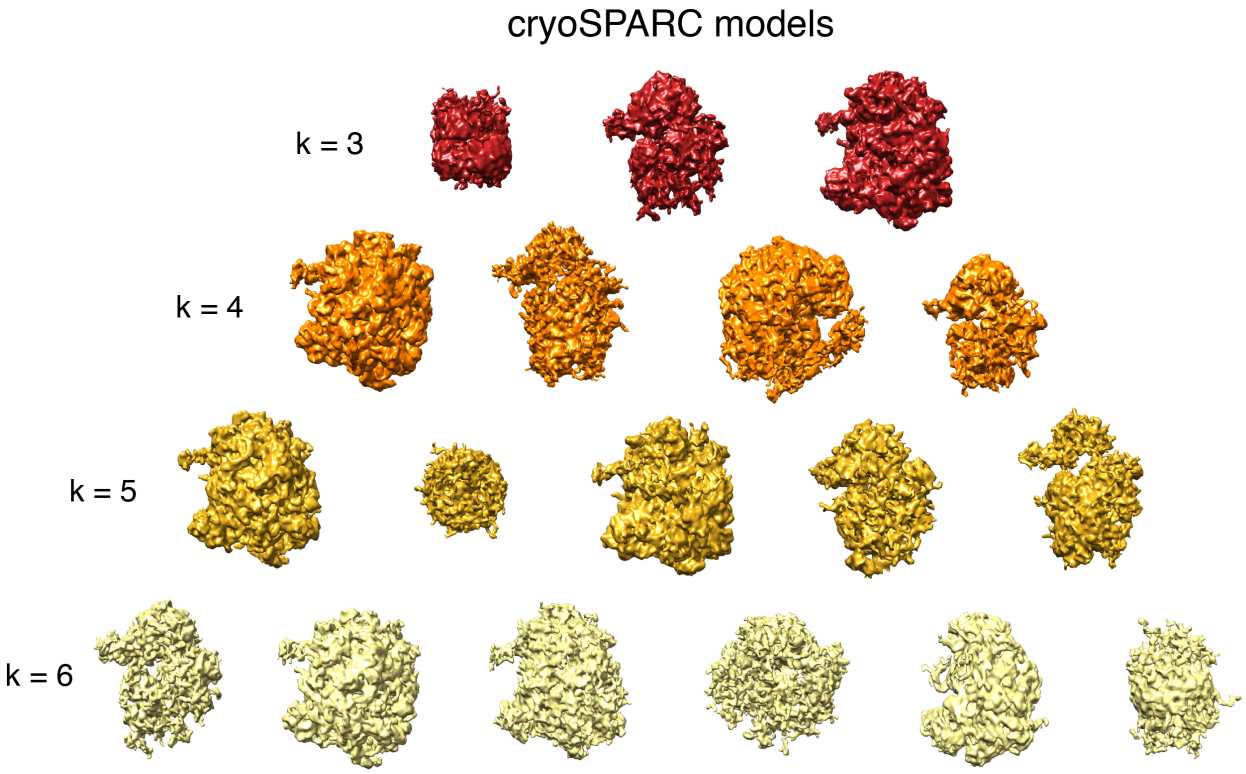


Figure S8. *Ab initio* reconstructions in cryoSPARC with varying class number

3D reconstructions using *ab initio* reconstruction in cryoSPARC from the entire data set with K = 3, 4, 5 and 6 classes, respectively.

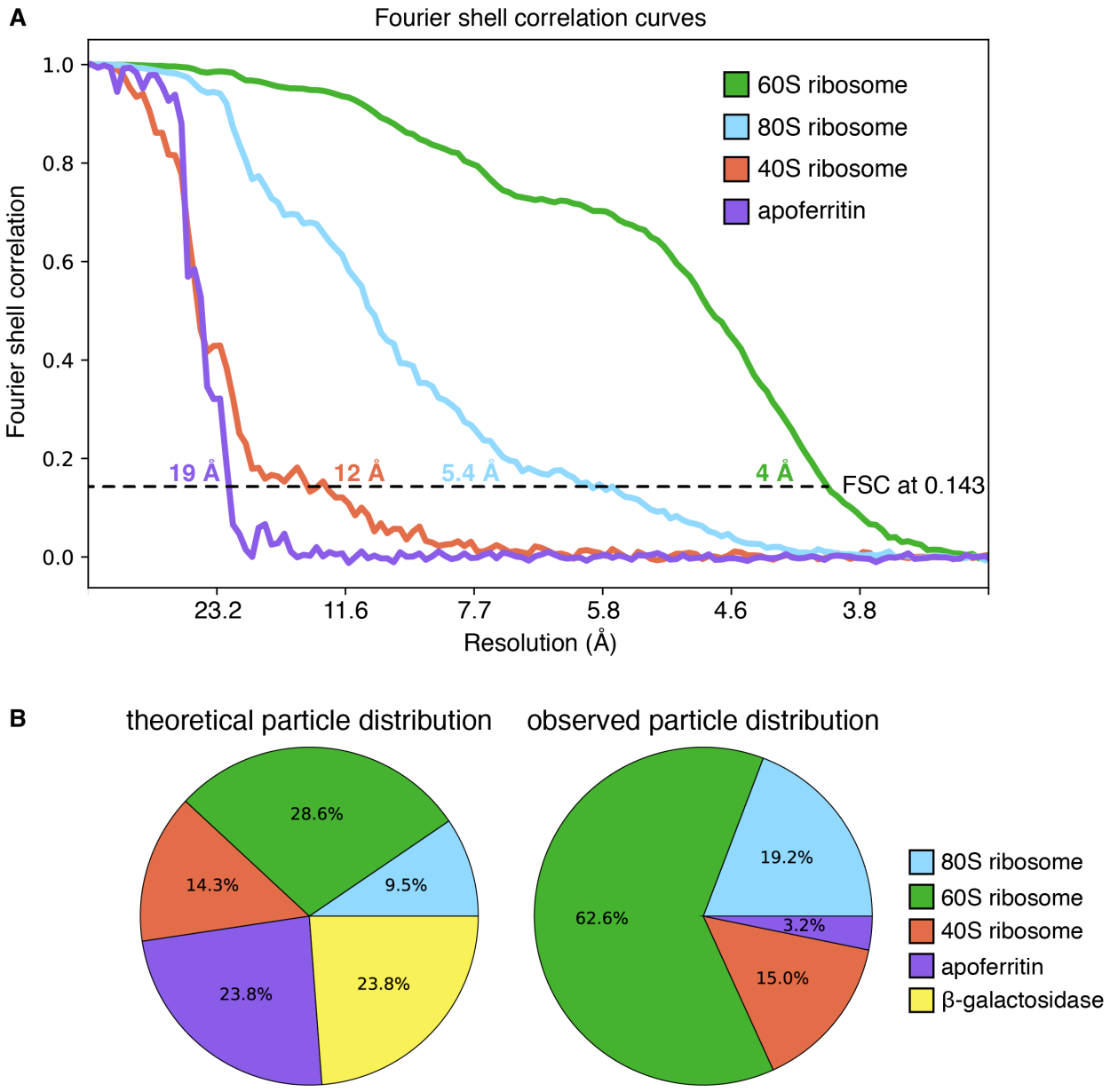


Figure S9. Fourier shell correlations curves

(A) FSC curves for our clustered 80S ribosome (blue), 60S ribosome (green), 40S ribosome (red) and apoferritin (purple) shown in Figure 5. Nominal resolutions were estimated to be 5.4, 4, 12 and 19 Å, respectively, using the 0.143 gold-standard FSC criterion. β-galactosidase was not observed in our dataset and therefore no reconstruction was calculated. (B) Theoretical ratio of particles from the input concentrations compared to observed ratio of particles after clustering of the data. The input concentrations for the 80S, 60S, 40S, apoferritin and β-galactosidase were

50, 150, 75, 125 and 125 nM and the observed particles for each were 38957, 126776, 30353, 6525 and 0, respectively.