Supplemental Information

Mathematical background

We discretize the volume containing a macromolecule as having $N \times N \times N$ voxels and model it as a N^3-dimensional random column vector X. A projection image from the data is modeled as a noisy approximation to the line integrals across the volume in a given direction, which we write

$$Y = RX, \tag{S1}$$

where R contains the orientation-dependent coefficients in the line integrals. We also assume all projection orientations to be known. Since X is random, so is Y, and their respective covariance matrices $cov(.)$ are related by

$$cov(Y) = R\, cov(X)\, R^T, \tag{S2}$$

where R^T denotes the transpose of R. The equation above can be re-written as a system of linear equations:

$$C_Y = W C_X, \tag{S3}$$

where C_Y is the covariance of the line integrals (to be referred to as "2D covariance") and C_X the unknown three-dimensional covariance ("3D covariance"), and the elements of W are products of elements of R. To see how Equation S2 reduces to Equation S3, note that the covariance between two linear combinations of random variables equals a linear combination of covariances, each of which is between a random variable from one combination and another random variable from the other combination. The coefficients in the newly formed linear combination are simply the products of the corresponding coefficients. Linearity between C_X and C_Y has already been derived in (Katsevich, et al., 2014).

In practice, many projections exist, and one would therefore concatenate all these equations and solve the entire system. However, to reduce computational burden, the projections are grouped based on the similarity of their orientations; and one equation like Equation S3 is created for each of, say, J groups (hence one W is given for each group):

$$\begin{cases} C_Y^1 = W_1 C_X \\ \dots \\ C_Y^J = W_J C_X \end{cases} \tag{S4}$$

The aim now is to estimate the 3D covariance from the set of measured 2D covariances.

To correct for noise, which is substantial in cryo-EM, we subtract the 2D covariance of a pure-noise projection (estimated by shifting the projection images by one-half of their size in both vertical and

horizontal direction) from the corresponding measured 2D covariance. This is allowed under the assumption that the measurement noise and the structural heterogeneity of the macromolecules are statistically independent. Due to uneven ice thickness and uneven illumination, data are normalized to compensate for such data imperfections. Projection data are assumed to be correctly aligned and corrected for the CTF (Frank). Binning the orientations creates an unwanted variability, which nevertheless is considerably reduced by subtracting the reprojection of a volume reconstructed from the normalized projection data. Supplementary Figure S1 explains how we preprocess the data and compute the 2D covariance for each group (see also Section "Variability due to grouping of orientations" below).

To solve the system of equations, we use an iterative algorithm known as block-Algebraic Reconstruction Technique (Herman, 1970; Censor & Zenios, 1997), with relaxation parameter β=0.005 (note that this value depends on the entries in W). To speed up the convergence, we imposed at each iteration the condition for the solution to have the properties of a variance/covariance matrix: the variance must be non-negative, and the squared covariance must be no greater than the product of the corresponding variances. We found that usually twenty or fewer iterations are adequate to obtain a stable solution. Specifically, starting with an initial 3D covariance $C_X^{(0)}$, at the k-th iteration of our algorithm we compute

$$C_X^{(k+1)} = C_X^{(k)} + \beta \left(C_Y^{\tilde{k}} - W_{\tilde{k}} C_X^{(k)} \right), \tag{S5}$$

where $\tilde{k} = k \bmod(J) + 1$, and impose the constraints

$$\left(C_X^{(k+1)}\right)_{j,j} = 0, \text{ if } \left(C_X^{(k+1)}\right)_{j,j} < 0, \ 1 \le j \le J, \tag{S6}$$

followed by these constraints

$$\left(C_X^{(k+1)}\right)_{i,j} = \operatorname{sgn}\left[\left(C_X^{(k+1)}\right)_{i,j}\right] \sqrt{\left(C_X^{(k+1)}\right)_{i,i} \left(C_X^{(k+1)}\right)_{j,j}}, \text{if } \left|\left(C_X^{(k+1)}\right)_{i,j}\right| > \sqrt{\left(C_X^{(k+1)}\right)_{i,i} \left(C_X^{(k+1)}\right)_{j,j}}, \tag{S7}$$

for $1 \le i \ne j \le J$, where sgn() is the sign function, and $(C_X)_{i,j}$ denotes the $i + j * J$ element of the vector C_X, for $1 \le i, j \le J$.

Variability due to grouping of orientations

In this section we analyze the effect of the grouping/binning of orientations, which generates an additional unwanted variability in the covariance map. We show that the effect is lowered if we subtract the reprojection of the average structure from the projection data.

We consider one row of Equation S1 in the noiseless case; i.e.,

$$y_1 = \sum_i r_i x_i \tag{S8}$$

That is, y_1 is a measurement of image y, which is a projection of a 3D structure x. Because of the structural heterogeneity, we write $x = \bar{x} + \Delta x$, where $\bar{x} = E(x)$ is the expected value of x. Because of the

grouping of orientations, several projections of different but similar orientations are assigned to the same orientation. To reflect this fact, we write the coefficients r_i as $\overline{r_i} + \Delta r_i$, where $\overline{r_i}$ corresponds to the assigned orientation of the group. Thus, with consideration of variability in the orientation, a measurement becomes

$$y_1 = \sum_i r_i(\overline{x_i} + \Delta x_i) = \sum_i [r_i \overline{x_i} + (\overline{r_i} + \Delta r_i)\Delta x_i] = \sum_i [r_i \overline{x_i} + \overline{r_i}\Delta x_i + \Delta r_i \Delta x_i]. \quad \text{(S9)}$$

That is, the variability in y_1 is created by the variability in each of the three terms inside the bracket of the right-most formula in (S9). Since the variability we want to see is the one coming from Δx_i and not Δr_i, the first and the last terms need to be removed -- note that the first term is in fact $r_i \overline{x_i} = (\overline{r_i} + \Delta r_i)\overline{x_i}$. The first term, however, is a reprojection of the average 3D structure, so it can be easily reproduced and subtracted. The second term is a reprojection (in the direction of the assigned orientation) of the average-subtracted 3D structure, which is in line with our model assumption. Finally, the third term is a "residual" projection of the average-subtracted 3D structure.

By removing the first term, we eliminate an important amount of variability due to orientation uncertainty, but not completely because of the third term. A precondition for our approach to work is, thus, that the bins must not be too large.

Supplemental Figure S1

Data preprocessing

After normalizing the data by setting the background to zero mean and unit variance, we subtract from each projection the corresponding reprojection of a volume reconstructed from the normalized data. In the section "Variability due to grouping of orientations" above, we show that the removal of the reprojection helps remove the unwanted variability due to the slightly different orientations in a binned group. The projection binning/grouping comes in the next step, which is applied to both the projections after the reprojection removal and their shifted (by one-half of their size in both vertical and horizontal direction) version. The shifted version contains only noise. We estimate the 2D covariance for both versions and their difference is the final estimated 2D covariance.

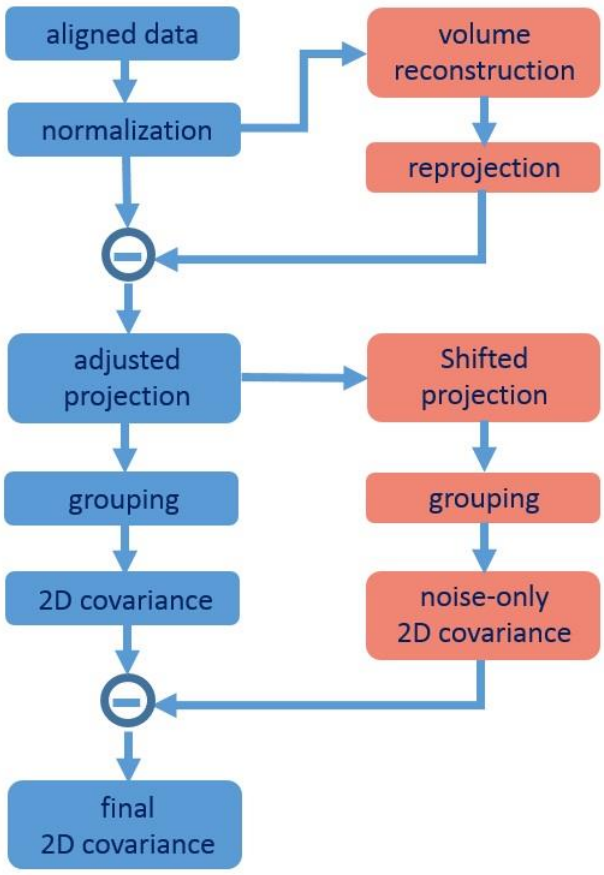

Figure S1. Estimation of the 2D covariance; related to Figure 1

Supplemental Figure S2

43S ribosomal pre-initiation complex with DHX29 bound

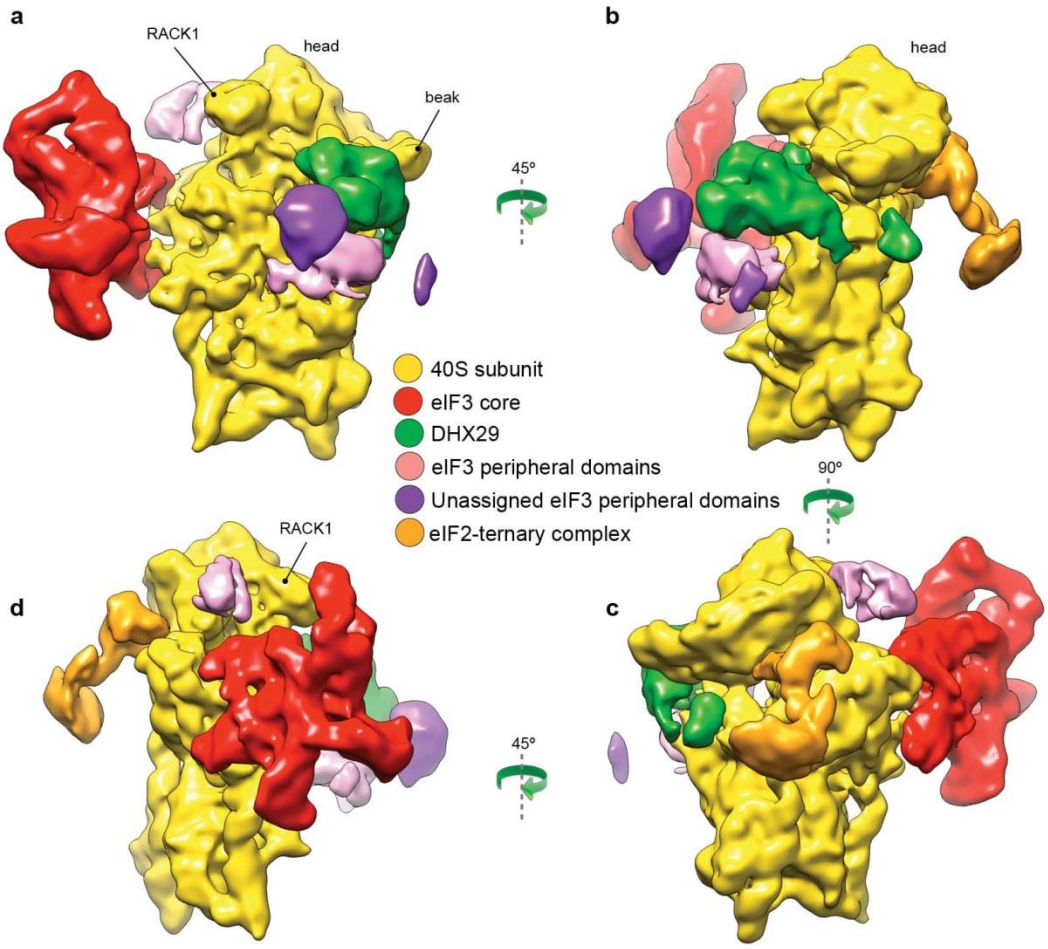

Figure S2. 43S ribosomal pre-initiation complex with DHX29 bound; related to Experimental Procedures

Supplemental Figure S3

Solution domain of finer grid (mesh)

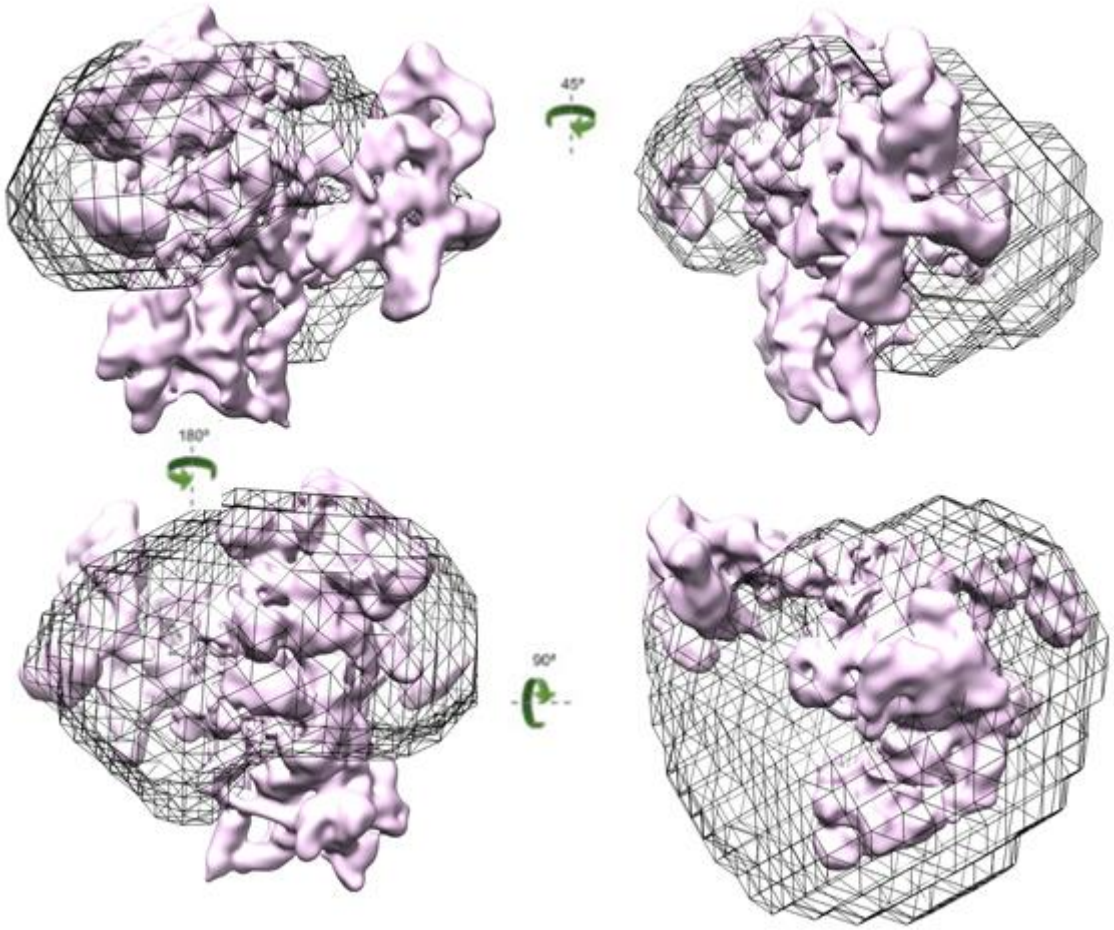

Figure S3. Solution domain (mesh) for higher resolution (32^3-voxel tessellation) estimation of the covariance. First, a low resolution (16^3-voxel tessellation) estimation was calculated in the whole sphere inscribed in a cube of 16^3 voxels. Next, regions of relatively high variance were determined and smoothed, yielding a new region that is used as a solution domain for higher resolution estimation; related to Experimental Procedures