Likelihood-based signal and noise analysis for docking of models into cryo-EM maps

Randy J. Read[a*], Claudia Millán[a,b], Airlie J. McCoy[a], Thomas C. Terwilliger[c]

[a] Department of Haematology, Cambridge Institute for Medical Research, University of Cambridge, Hills Road, Cambridge CB2 0XY, United Kingdom

[b] Present address: SciBite Limited, BioData Innovation Centre, Wellcome Genome Campus, Hinxton, Cambridge CB10 1DR, United Kingdom

[c] New Mexico Consortium, Los Alamos National Laboratory, 100 Entrada Drive, Los Alamos, NM 87544, USA

* Correspondence email: rjr27@cam.ac.uk

Synopsis Likelihood-based rotation, translation and refinement targets have been derived for docking models into cryo-EM reconstructions.

Abstract Fast, reliable docking of models into cryo-EM maps requires understanding of the errors in the maps and the models. Likelihood-based approaches to errors have proven to be powerful and adaptable in experimental structural biology, finding applications in both crystallography and cryo-EM. Indeed, previous crystallographic work on the errors in structural models is directly applicable to likelihood targets in cryo-EM. Likelihood targets in Fourier space are derived here to characterise, based on the comparison of half-maps, the direction- and resolution-dependent variation in the strength of both signal and noise in the data. Because the signal depends on local features, the signal and noise are analysed in local regions of the cryo-EM reconstruction. The likelihood analysis extends to prediction of the signal that will be achieved in any docking calculation for a model of specified quality and completeness. A related calculation generalises a previous measure of the information gained by making the cryo-EM reconstruction.

Keywords: Likelihood, cryo-EM, docking, information gain

1. Introduction

The problem of docking models into cryo-EM maps is similar to the molecular replacement (MR) problem in crystallography. The key difference is that cryo-EM data are enriched by the phase information that is lost in crystallography, and the resulting increase in signal-to-noise greatly simplifies the task of translating an oriented model with FFT-based correlation functions. However, this phase information cannot be used directly in assessing different

model orientations prior to the translation search, so many existing docking algorithms rely on a systematic six-dimensional search over possible orientations and positions.

In crystallography, MR algorithms have been made significantly more sensitive by using likelihood scores for rotation, translation and rigid-body refinement tasks (McCoy *et al.*, 2007). In addition, an understanding of the relationship connecting data and model quality with the likelihood scores that can be expected in a particular calculation has opened up new possibilities for tailoring the MR calculations to the problem at hand (McCoy *et al.*, 2017; Oeffner *et al.*, 2018). These concepts can be applied to the related problems in cryo-EM.

Most existing docking methods for cryo-EM are scored by variants of cross-correlation functions. Comprehensive reviews of these score functions have been compiled by others: (Zundert *et al.*, 2015; Cragnolini *et al.*, 2021). Some examples include cross-correlation of the experimental cryo-EM map and a map computed from coordinates (Stewart *et al.*, 1993), local cross-correlation (Roseman, 2000), Laplacian filtered cross-correlation (Wriggers, 2012) and core-weighted cross-correlation (Wu *et al.*, 2003).

When comparing a variety of scores of fit to model, including cross-correlations, Joseph *et al.* (2017) found that mutual information was a better discriminator for low to medium resolution maps. Like the likelihood score proposed here, mutual information is a probabilistic measure, but it works with real-space voxel values, not Fourier terms. In addition, mutual information does not explicitly account for errors in the reconstruction itself.

As noted below, our docking target is based on similar ideas to the likelihood-based refinement target for models against cryo-EM maps used in Refmac (Murshudov, 2016), but differs importantly in using a more sophisticated error model for experimental data that takes account of the directional dependence of both the signal and the noise in Fourier space.

2. Probabilities and likelihood targets

2.1. Error model for single-particle cryo-EM data

For a cryo-EM reconstruction, the aim is that each individual molecule or molecular assembly in a particle is essentially a rigid object, either by nature or as a result of particle selection.

Errors in cryo-EM reconstructions come from a combination of suboptimal relative orientations of individual images, structural differences and imaging limitations and artefacts among the collection of particles used in the reconstruction, reviewed for instance by

Ramlaul *et al.* (2019). In the individual 2D particle images derived from a series collected over the total exposure, the images are smeared by any uncorrected sample motion, degraded by effects of any radiation damage and limited in resolution by the detector pixel size. Additional random shot noise comes from counting statistics and the presence of irreproducible features in the vitrified solvent around them.

For reconstruction, the information contained in the 2D image is converted into its Fourier transform, which comprises a 2D slice through the Fourier transform of the molecule; errors in the Fourier terms can arise, for instance, from errors in the contrast transfer function correction. If the correction terms have been optimized, their values and errors will differ in different images, so we can expect the remaining errors in data from these individual images to be largely uncorrelated with particle orientation or with the images themselves. Nonetheless, if systematic errors remained it would be difficult to distinguish them from signal.

Each particle imaged in a data set will be in a different orientation and (to a greater or lesser extent) a different conformation. 3D classification will allow significantly different conformations to be grouped together, but variation will remain within the groups, corresponding in real space to blurring of the atoms over their range of possible relative positions when constructing a 3D image. Further blurring will come from uncertainties in the orientation and position of the particle in each image, when averaging the Fourier terms from different images to obtain a 3D data set.

In our error model, we consider that the signal in an individual Fourier term in the reconstruction comes from the Fourier transform of the image of atoms at rest, blurred by the effects of global and local variations in orientation and position. These blurring effects are similar to what is modeled locally in crystallography by anisotropic displacements or, on a larger scale, by translation-libration-screw (TLS) models (Schomaker & Trueblood, 1968).

Random noise in contributions from individual particle images will be reduced when corresponding Fourier terms from different images are averaged. However, the existence of preferred orientations will mean that the magnitude of the random noise terms, after averaging over different numbers of observations, will vary with direction in Fourier space. In principle, this could be modeled by keeping track of redundancy during the reconstruction process, but in the current implementation we are starting from conventional half-maps rather than individual particle images. We approximate the directional and resolution dependence as

a smoothly-varying function in Fourier space. Since variations in conformation need not be correlated with orientation preference in the sample, the two sources of variation in signal and noise are evaluated independently.

Because the estimation of noise requires the comparison of independent measurements, all of our signal and error evaluation is carried out using the Fourier terms computed from the half-maps. The signal power is deduced from correlations between the half-map terms and the error power from their differences.

The signal in matching pairs of Fourier terms derived from the half maps can be expressed in terms of the underlying Fourier transform of atoms at rest (represented as $\mathbf{T}_{hkl}$ and drawn from a complex normal distribution with variance Σ_T representing the scattering power), multiplied by a scale factor combined with a term that varies with resolution and direction in Fourier space (represented as A_{hkl}). The noise term, $\boldsymbol{\varepsilon}_{hkl}$, is drawn independently for each half map from a complex normal distribution with variance Σ_E. Note that both Σ_T and Σ_E will vary with resolution; as noted above, Σ_E will also vary with direction in Fourier space.

$$\mathbf{F}_{hkl} = A_{hkl}\mathbf{T}_{hkl} + \boldsymbol{\varepsilon}_{hkl} \tag{1}$$

When describing the individual half-map terms, the subscripts hkl will be implicit for simplicity of notation:

$$\mathbf{F}_1 = A\mathbf{T} + \boldsymbol{\varepsilon}_1 \tag{1a}$$

$$\mathbf{F}_2 = A\mathbf{T} + \boldsymbol{\varepsilon}_2 \tag{1b}$$

Because $\mathbf{T}$ and $\boldsymbol{\varepsilon}$ are both drawn from complex normal distributions, the joint distribution of $\mathbf{F}_1$ and $\mathbf{F}_2$ can be defined in terms of a bivariate complex normal distribution. The covariance matrix for this distribution is given by:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \langle \mathbf{F}_1\mathbf{F}_1^* \rangle & \langle \mathbf{F}_1\mathbf{F}_2^* \rangle \\ \langle \mathbf{F}_1^*\mathbf{F}_2 \rangle & \langle \mathbf{F}_2\mathbf{F}_2^* \rangle \end{bmatrix} \tag{2}$$

The terms in the covariance matrix can be simplified in terms of the variances of the distributions for $\mathbf{T}$ and $\boldsymbol{\varepsilon}$, noting that $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ are independent so that their covariance is zero.

$$\langle \mathbf{F}_1\mathbf{F}_1^* \rangle = \langle (A\mathbf{T} + \boldsymbol{\varepsilon}_1)(A\mathbf{T} + \boldsymbol{\varepsilon}_1)^* \rangle = \Sigma_S + \Sigma_E \tag{3a}$$

$$\langle \mathbf{F}_2\mathbf{F}_2^* \rangle = \langle (A\mathbf{T} + \boldsymbol{\varepsilon}_2)(A\mathbf{T} + \boldsymbol{\varepsilon}_2)^* \rangle = \Sigma_S + \Sigma_E \tag{3b}$$

$$\langle \mathbf{F}_1\mathbf{F}_2^* \rangle = \langle (A\mathbf{T} + \boldsymbol{\varepsilon}_1)(A\mathbf{T} + \boldsymbol{\varepsilon}_2)^* \rangle = \Sigma_S = \langle \mathbf{F}_1^*\mathbf{F}_2 \rangle, \text{ where} \tag{3c}$$

$$\Sigma_S = A^2 \Sigma_T \tag{3d}$$

The parameters characterising the bivariate complex normal distribution can be estimated by maximising the likelihood of measuring the data derived from the two half-maps.

The determinant and the inverse of the covariance matrix are needed to compute the likelihood target:

$$\det = \det\left(\begin{bmatrix} \Sigma_S + \Sigma_E & \Sigma_S \\ \Sigma_S & \Sigma_S + \Sigma_E \end{bmatrix}\right) = 2\Sigma_S\Sigma_E + \Sigma_E^2 \tag{4a}$$

$$\mathbf{\Sigma}^{-1} = \frac{1}{2\Sigma_S\Sigma_E + \Sigma_E^2}\begin{bmatrix} \Sigma_S + \Sigma_E & -\Sigma_S \\ -\Sigma_S & \Sigma_S + \Sigma_E \end{bmatrix} \tag{4b}$$

Considering that, before we know the values of $\mathbf{F}_1$ and $\mathbf{F}_2$ their expected values are zero, the joint probability distribution is given by

$$p(\mathbf{F}_1, \mathbf{F}_2) = \frac{1}{|\pi\mathbf{\Sigma}|}\exp[-(\mathbf{F}_1, \mathbf{F}_2)^H \mathbf{\Sigma}^{-1}(\mathbf{F}_1, \mathbf{F}_2)] =$$

$$\frac{1}{\pi^2(2\Sigma_S\Sigma_E + \Sigma_E^2)}\exp\left(\frac{-(F_1^2 + F_2^2)(\Sigma_S + \Sigma_E) + 2\Sigma_S F_1 F_2 \cos(\varphi_1 - \varphi_2)}{2\Sigma_S\Sigma_E + \Sigma_E^2}\right) \tag{5}$$

where superscript H indicates the Hermitian transpose and the Fourier terms are represented in terms of their amplitudes and phases. The contribution of a single Fourier term to the log-likelihood function is therefore given by

$$\ln(L) = -\ln\left(\pi^2(2\Sigma_S\Sigma_E + \Sigma_E^2)\right) - \frac{(F_1^2 + F_2^2)(\Sigma_S + \Sigma_E)}{2\Sigma_S\Sigma_E + \Sigma_E^2} + \frac{2\Sigma_S F_1 F_2 \cos(\varphi_1 - \varphi_2)}{2\Sigma_S\Sigma_E + \Sigma_E^2} \tag{6}$$

We have tested two approaches to determining values for Σ_S and Σ_E. One is to assume that their values are close to constant in a small region of Fourier space such as a sphere around a particular Fourier term (or at least that their variation over that sphere is such that their mean value is representative of the Fourier term at the centre). This approach makes no assumption about the functional form for their dependence on resolution or direction in Fourier space. The second approach is to assume that the variation can be captured by some combination of a resolution term (such as a constant for each spherical shell) and an anisotropic tensor. Considering that the distribution of favoured orientations can have a number of modes that do not necessarily obey any symmetry, only the first approach would be strongly justified for the Σ_E error terms. This was borne out by preliminary work using the anisotropic tensor approach, not reported here in further detail. On the other hand, the assumption that the local molecular structure undergoes displacements that can be captured by an anisotropic tensor seems easier to justify, based on bonding constraints. For this reason, our tests have

concentrated on an error model in which local variation in Fourier space is used for the error terms, and both approaches are used for the signal terms.

For the local variation approach, there is an analytical solution for the Σ_S and Σ_E terms that maximise the log-likelihood for a local region in Fourier space. This is obtained by taking the derivatives of the sum of the log-likelihood, over a set of n Fourier terms, with respect to Σ_S and Σ_E, then solving the simultaneous equations to find the values where the two derivatives are equal to zero. The results for the maximum likelihood estimators of these parameters are intuitively reasonable, from inspection of equations (3) defining Σ_S and Σ_E above:

$$\widehat{\Sigma_S} = \frac{1}{n}\sum F_1 F_2 \cos(\varphi_1 - \varphi_2) \tag{7a}$$

$$\widehat{\Sigma_E} = \left(\frac{1}{2n}\sum F_1^2 + F_2^2\right) - \widehat{\Sigma_S} \tag{7b}$$

For the anisotropic tensor approach, there does not appear to be an analytical solution, so an iterative refinement is required. The refinable parameters in the log-likelihood function are the parameters determining the values of A and Σ_T, while Σ_E can be estimated using the result above.

$$A = A_0 \exp(-\mathbf{h}^T \beta_A \mathbf{h}), \tag{8}$$

where A_0 is an overall scale and β_A is an anisotropic tensor that captures the overall anisotropic displacements of the object in the map.

Σ_T is a function of resolution because the spectral variation of the Fourier transform reflects both the width of atomic features and favoured interatomic distances within the imaged object. If the signal-to-noise were reasonably high for all resolution ranges, Σ_T could be estimated reliably in resolution bins, but this is not usually a safe assumption towards the resolution limit. For a similar problem in normalising crystallographic data (Read & McCoy, 2016), we have found that a Bayesian framework using a prior probability distribution is useful: we assume that the overall spectral variation of data should be similar to the average seen in a large variety of structures, as captured by the BEST curve tabulated by Popov and Bourenkov (2003). This curve will not be exact for any particular data set, so some variation must be allowed; this is accomplished by refining bin-wise resolution parameters that are set initially to one, and weakly restraining the logarithm of their values to zero. Weaker restraints are used at low resolution than at high resolution, because the low-resolution Fourier terms depend more on molecular shape than favoured interatomic distances, and thus vary more from structure to structure. The restraints have very little effect on the refined parameters for

strong data but dramatically improve the behaviour of refinement for weak data. Note that the BEST curve was derived using a large set of X-ray diffraction data to high resolution. A related curve is not yet available for cryo-EM or electron diffraction data, although a similar use of power spectra has been suggested in the past (Scheres, 2012), but we note that the spectral variation of the Fourier terms from cryo-EM reconstructions at high resolution show similar behaviour to those from X-ray diffraction, because of the predominant effect of favoured distances. Using the BEST curve, the refinable parameters for Σ_T are the resolution bin parameters in the following equation:

$$\Sigma_T = \Sigma_{T,bin}\Sigma_{T,best} \tag{9}$$

Note that the assumption that the spectral variation will follow the BEST curve will be violated when the half-maps have been manipulated, for instance by applying band-pass filters. For this reason (among others), our method requires the availability of unfiltered, unmasked half-maps.

2.2. Likelihood target for evaluating models in cryo-EM reconstructions

To derive a likelihood target for evaluating the fit of models to data, we need to account for errors in the model (in addition to estimates of measurement error discussed above). Both structure refinement and docking can be carried out using a likelihood target that evaluates the likelihood of the map given the model. We start by considering the errors between the Fourier terms corresponding to the (unknown) true map (**T**) and the average map coefficients obtained from the two half-maps. This can be evaluating by considering the definition of the half-map Fourier coefficients in terms of the true coefficients:

$$\mathbf{F}_{mean} = \frac{(\mathbf{F}_1+\mathbf{F}_2)}{2} = \frac{(A\mathbf{T}+\boldsymbol{\varepsilon}_1+A\mathbf{T}+\boldsymbol{\varepsilon}_2)}{2} = A\mathbf{T} + \boldsymbol{\varepsilon}_{\text{mean}}, \text{ where} \tag{10a}$$

$$\boldsymbol{\varepsilon}_{\text{mean}} = \frac{(\boldsymbol{\varepsilon}_1+\boldsymbol{\varepsilon}_2)}{2} \tag{10b}$$

When $\boldsymbol{\varepsilon}_1$ and $\boldsymbol{\varepsilon}_2$ can be considered independent with equal variance (as we assume for half-maps), the variance of their mean is reduced by a factor of two. This allows us to work out the terms of the covariance matrix relating **T** and $\mathbf{F}_{mean}$.

$$\langle \mathbf{T}\,\mathbf{T}^* \rangle = \Sigma_T \tag{11a}$$

$$\langle \mathbf{F}_{mean}\mathbf{F}^*_{mean} \rangle = A^2\Sigma_T + \Sigma_E/2 \tag{11b}$$

$$\langle \mathbf{T}\,\mathbf{F}^*_{mean} \rangle = A\Sigma_T = \langle \mathbf{T}^*\mathbf{F}_{mean} \rangle \tag{11c}$$

The likelihood target has a simpler form if it is expressed in terms of normalised map coefficients (E-values), in which the mean-square value of E is expected to be one.

$$\mathbf{E}_T = \mathbf{T}/\sqrt{\Sigma_T} \tag{12a}$$

$$\mathbf{E}_{mean} = \mathbf{F}_{mean}/\sqrt{A^2\Sigma_T + \Sigma_E/2} \tag{12b}$$

$$\langle \mathbf{E}_T \mathbf{E}^*_{mean} \rangle = \frac{A\Sigma_T}{\sqrt{\Sigma_T(A^2\Sigma_T+\Sigma_E/2)}} = \sqrt{\frac{A^2\Sigma_T}{A^2\Sigma_T+\Sigma_E/2}} = \sqrt{\frac{\Sigma_S}{\Sigma_S+\Sigma_E/2}} = D_{obs} \tag{12c}$$

D_{obs} is the complex correlation relating $\mathbf{E}_{mean}$ and the true value, $\mathbf{E}_T$. It plays the same role, for a single Fourier term, as FSC_{ref} (Fourier shell correlation to the true map) does for a whole resolution shell in Fourier space. Note that, if Σ_E is zero, D_{obs} is one, but it becomes smaller as the ratio between Σ_E and Σ_S increases, reaching zero when Σ_S is zero.

The other source of error in the likelihood target is model error. For docking, it is generally safe to assume that the errors in the map and the errors in the model are independent, prior to any refinement against the map, so there are no concerns about overfitting. The relationship between the Fourier coefficients computed from a model and those that would be obtained from the true map is the same as that between calculated and true structure factors in crystallography: the Central Limit Theorem allows us to conclude that the errors in the calculated Fourier coefficients that arise from the sum of many small errors from the individual atoms in the model can be described in terms of a complex normal distribution, like the errors between the true map and the experimental reconstruction. In crystallography, this is described by the complex correlation, termed σ_A, between the normalised structure factors (Srinivasan & Ramachandran, 1965; Read, 1990). The σ_A term combines the effects of completeness of the model (the fraction *f* of the scattering accounted for by the model) and the accuracy of the model; if we make the simplifying assumption that the errors in the coordinates of all the atoms are all drawn from the same 3D Gaussian distribution, σ_A can be calculated with the following formula:

$$\sigma_A(s) = \sqrt{f}\exp\left(-\frac{2\pi^2}{3}s^2\Delta^2\right) \tag{13}$$

where s is the inverse resolution and Δ is the rms radial coordinate error (Read, 1990). Violation of this assumption can change the resolution dependence of the σ_A curve, but a compromise effective overall rms error is determined by refinement after placing the model.

Because the errors between the true map and either the calculated map or the observed map are independent, the complex correlation between the observed map and the model is simply the product of the two individual complex correlations, D_{obs} and σ_A. Therefore, the joint distribution of $\mathbf{E}_{mean}$ and the normalised calculated Fourier coefficient, $\mathbf{E}_C$, is a bivariate complex normal distribution with expected values of zero (prior to any knowledge of either) and the following covariance matrix:

$$\Sigma = \begin{bmatrix} 1 & D_{obs}\sigma_A \\ D_{obs}\sigma_A & 1 \end{bmatrix} \tag{14}$$

The likelihood function for judging the fit of a model in a map (whether derived by docking or structure refinement) is the conditional probability distribution for the observed normalised Fourier coefficient given the corresponding term computed from the model. This conditional distribution is obtained by a simple manipulation of the joint distribution, yielding a complex normal distribution with a variance of $1 - D_{obs}^2\sigma_A^2$ and an expected value of $D_{obs}\sigma_A\mathbf{E}_C$:

$$p(\mathbf{E}_{mean}; \mathbf{E}_C) = \frac{1}{\pi(1-D_{obs}^2\sigma_A^2)} \exp\left(-\frac{|\mathbf{E}_{mean} - D_{obs}\sigma_A\mathbf{E}_C|^2}{1-D_{obs}^2\sigma_A^2}\right) \tag{15}$$

It is more convenient to work with the log-likelihood gain or LLG, *i.e.* the gain in the log-likelihood score compared to an uninformative model (for which σ_A is zero). The contribution of a single Fourier term to the total LLG is:

$$LLG(\mathbf{E}_{mean}; \mathbf{E}_C) = \frac{2D_{obs}\sigma_A E_{mean} E_C \cos(\Delta\phi) - D_{obs}^2\sigma_A^2(E_{mean}^2 + E_C^2)}{1-D_{obs}^2\sigma_A^2} - \ln(1 - D_{obs}^2\sigma_A^2) \tag{16}$$

where $\Delta\phi$ is the difference between the phases of $\mathbf{E}_{mean}$ and $\mathbf{E}_C$, and E_{mean} and E_C are amplitudes. Note that this can alternatively be expressed in terms of a correlation function between the weighted averaged map and the model, a scale factor and an offset:

$$LLG(\mathbf{E}_{mean}; \mathbf{E}_C) = \frac{2}{1-D_{obs}^2\sigma_A^2} D_{obs}\sigma_A E_{mean} E_C \cos(\Delta\phi) - \frac{D_{obs}^2\sigma_A^2(E_{mean}^2 + E_C^2)}{1-D_{obs}^2\sigma_A^2} - \ln(1 - D_{obs}^2\sigma_A^2) \tag{17}$$

The total LLG score is the sum over all Fourier terms. However, it is important to note that cryo-EM differs from crystallography in that the Fourier transform of the reconstruction is typically highly over-sampled: proteins in crystals pack in a lattice where molecules are in contact with each other, whereas the cryo-EM reconstruction is computed in a much larger box than required to contain the particle. Oversampling leads to strong correlations among neighbouring Fourier terms. This can be accounted for simply by applying a correction factor equal to the ratio of the volume required to contain the particle and the volume of the box in

which the reconstruction was carried out, as proposed by van Heel and Schatz (van Heel & Schatz, 2020) for computing information content. The same correction factor must be applied to all fast approximations, expected values and information gain discussed below.

The likelihood target described here has the same basic functional form as the likelihood target used to refine cryo-EM models in Refmac (Murshudov, 2016), differing importantly in taking account of the dependence of the signal and error terms on direction in reciprocal space, instead of depending only on resolution.

2.3. Map coefficients

Two sets of map coefficients can be generated for evaluating the fit of a docked model. The first type uses the Fourier coefficients

$$\mathbf{F}_{map} = D_{obs}\mathbf{E}_{mean} \tag{18}$$

Since this is the expected value of the true sharpened map coefficient (the centroid of the probability distribution), this should give a map that minimises the error from the true sharpened map.

The second type uses Fourier coefficients that include the other weighting terms from the correlation function in the log-likelihood target (17),

$$\mathbf{F}_{map} = \frac{2}{1-D_{obs}^2\sigma_A^2} D_{obs}\sigma_A\mathbf{E}_{mean}. \tag{19}$$

The correlation of this map (19) with the sharpened map computed from a docked model should be proportional to the likelihood target. To compute such a map, a choice has to be made for the value of σ_A that is used, which primarily depends on the scattering in the volume under consideration but also coordinate error and on the ability of atomic models to account for the bulk solvent region.

2.4. Fast rotation target for scoring orientations of models

In crystallographic MR, the six-dimensional problem of finding the orientation and position of a model to fit the diffraction data is typically divided into a sequence of two three-dimensional problems: an orientation search (rotation function) followed by translation searches with models in a number of plausible orientations determined from the orientation search (translation function). The crystallographic rotation functions can be directly adapted to the docking problem in cryo-EM, because the phase information in the complex Fourier terms cannot be used without some knowledge of the position of the search model. The

rotation search thus depends solely on the amplitudes of the Fourier terms, so the crystallographic likelihood-based rotation function (Storoni *et al.*, 2004) can be used without alteration: the cryo-EM D_{obs} plays the same role as the crystallographic D_{obs} parameter in the log-likelihood-gain on intensities (LLGI) target (Read & McCoy, 2016; Jamshidiha *et al.*, 2019). As in crystallography, an approximation of the likelihood-based rotation function can be computed rapidly by FFT methods before being scored by the exact likelihood function.

It should be noted that phase information can be used indirectly in the rotation search. If there is a hypothesis for the location of a particular component in a full reconstruction, the rotation search can use the Fourier terms computed from a portion extracted from the full reconstruction. The use of such a procedure is essential to the sub-volume searches mentioned below and discussed in detail in the accompanying paper.

2.5. Fast translation target for scoring positions of oriented models

In crystallography, where there is typically no prior phase information in an MR search, only an approximation to the likelihood target can be computed by FFT methods. However, the LLG score for the fit of a model to cryo-EM data (17) takes the form of a correlation function, which can therefore be calculated exactly as a function of translation using a single FFT, as long as scaling parameters and contributions that do not change with translation are accounted for.

2.6. Rigid-body refinement

Refinement of a docked model involves optimizing parameters of (17) to maximise the LLG. The orientation and translation parameters affect the calculated Fourier terms, while the estimated RMSD of the model changes the σ_A term. As in the related MR case, a careful choice of parameterization can improve the refinement behaviour. For instance, correlations between rotation and translation parameters can be minimised by defining the rotation in terms of a rotation about the center of mass of the component. In addition, defining the rotation in terms of a perturbation applied to the current orientation by rotating sequentially about orthogonal x, y and z axes (rather than, for instance, Euler angles) makes the rotation parameters locally close to orthogonal.

Although improvements in hardware and cryo-EM protocols have generally reduced the uncertainty about voxel size (or magnification factor) in modern cryo-EM reconstructions, we

have implemented a cell scale factor parameter, which affects the calculated Fourier terms and therefore can be refined to compensate for any error in voxel size.

3. Expected likelihood scores and information gain

In crystallographic MR, it has been possible to optimise the choice of search strategy by, first, knowing what absolute LLG score is required for correct solutions to be recognised and, second, being able to predict the LLG score that can be achieved in a particular search, given the quality of the model, the quality of the data and the resolution limit applied to the data (McCoy *et al.*, 2017; Oeffner *et al.*, 2018). The same considerations of expected LLG (or eLLG) can be applied to docking in cryo-EM, as discussed in detail in the accompanying manuscript.

3.1. Rotation eLLG

In a rotation search for a cryo-EM reconstruction lacking symmetry, the LLG score for an orientation is the same as the crystallographic LLG score for a model of a crystal in space group P1. Therefore, the rotation eLLG, $eLLG_{rot}$, can be computed with the same formula as the crystallographic eLLG for P1 (McCoy *et al.*, 2017):

$$eLLG_{rot} = \sum_{hkl} \frac{D_{obs}^4 \sigma_A^4}{2} \tag{20}$$

It is instructive to consider the effect of increasing the volume of a sphere extracted from the total reconstruction. If a sphere containing the correct volume for the component under investigation were doubled in volume, the number of Fourier terms would double. At the same time, the fraction of the map accounted for by the model would decrease by the same factor of two. Because σ_A is proportional to the square root of the model completeness, each term in the sum would be reduced by a factor of 4, so that the total $eLLG_{rot}$ would be reduced by a factor of two. More generally, all else being equal, $eLLG_{rot}$ is inversely proportional to the volume of the part of the map being used for the search.

3.2. Translation eLLG

The expected value of the LLG for an individual Fourier term is given by the probability-weighted average of the LLG over all possible values of the calculated Fourier term, where the weighting is the conditional probability of that calculated term given the observed Fourier term. Because the joint probability distribution of the calculated and observed Fourier terms

is symmetric, the required conditional probability has the same functional form as the likelihood of the data given the model.

$$eLLG_{tra} = \int p(\mathbf{E}_C; \mathbf{E}_{mean}) LLG(\mathbf{E}_{mean}; \mathbf{E}_C) d\mathbf{E}_C, \text{ where} \tag{21a}$$

$$p(\mathbf{E}_C; \mathbf{E}_{mean}) = \frac{1}{\pi(1-D_{obs}^2\sigma_A^2)} \exp\left(-\frac{|\mathbf{E}_C - D_{obs}\sigma_A \mathbf{E}_{mean}|^2}{1-D_{obs}^2\sigma_A^2}\right) \tag{21b}$$

The integral has a simple analytical solution:

$$eLLG_{tra} = (E_{mean}^2 - 1)D_{obs}^2\sigma_A^2 - \ln(1 - D_{obs}^2\sigma_A^2) \tag{22}$$

Considering that the expected value of E_{mean}^2 is one, if we assume that there is no correlation between E_{mean}^2 and $D_{obs}^2\sigma_A^2$, the expected value of the first term is zero, so that

$$eLLG_{tra} \approx -\ln(1 - D_{obs}^2\sigma_A^2) \tag{23}$$

For all values of $D_{obs}\sigma_A$, $eLLG_{tra}$ is greater than $eLLG_{rot}$, especially for the poorest combinations of map and model quality; when $D_{obs}\sigma_A$ is 0.01, for instance, the ratio is about 20,000. This is an indication of the extent to which phase information enhances the likelihood scores. The implication is that the trade-off between the size of the sub-volume and the sensitivity to the correct solution is very different for the rotation and translation parts of the search.

In contrast to $eLLG_{rot}$, $eLLG_{tra}$ is relatively insensitive to the size of sub-volume, especially when either the map or the model is poor ($D_{obs}\sigma_A \ll 1$), in which case the change in number of Fourier terms counterbalances the change in the logarithmic term.

3.3. Information gained by cryo-EM reconstruction

Information theory and likelihood are closely connected, and the information gained by measuring the data in a cryo-EM reconstruction, computed using the Kullback-Leibler divergence (Kullback & Leibler, 1951) can be derived with methods related to those used for the eLLG. Essentially, the Kullback-Leibler divergence (if measured with the natural logarithm in units of natural units of information, nats, rather than the conventional bits obtained with the logarithm base 2) is equivalent to the eLLG that would be expected for a perfect model.

The Kullback-Leibler divergence for the information gained about the true map, given the reconstruction, can be computed for one Fourier term with the following integral over all possible values of the true Fourier term, **E**:

$$D_{KL} = \int p(\mathbf{E}; \mathbf{E}_{mean}) \ln\left(\frac{p(\mathbf{E};\mathbf{E}_{mean})}{p(\mathbf{E})}\right) d\mathbf{E} \tag{24}$$

Applying Bayes' theorem, we can substitute

$$\frac{p(\mathbf{E};\mathbf{E}_{mean})}{p(\mathbf{E})} = \frac{p(\mathbf{E}_{mean};\mathbf{E})}{p(\mathbf{E}_{mean})} \tag{25}$$

so we see that the logarithm in the integral is the LLG that would be achieved with a perfect model, and the expression for D_{KL} is equivalent to eLLG assuming a perfect model. Because a perfect model would have $\sigma_A = 1$ for all Fourier terms,

$$D_{KL} = (E_{mean}^2 - 1)D_{obs}^2 - \ln(1 - D_{obs}^2) \tag{26}$$

Noting as before that the mean value of E_{mean}^2 should be one, if E_{mean}^2 and D_{obs}^2 are uncorrelated we have

$$D_{KL} = -\ln(1 - D_{obs}^2) \tag{27}$$

Information in units of bits instead of nats can be obtained by using the logarithm base 2, which differs by a factor of ln(2).

The total information gain in an entire data set or in a resolution shell will be the sum from the individual Fourier terms, but corrected for the correlations arising from oversampling in Fourier space. As above, following similar reasoning to that invoked by van Heel and Schatz (2020), the correction for oversampling can be made by comparing the volume of the map to the volume occupied by the ordered part from which the signal is obtained.

Although it is not immediately obvious, the D_{KL} measure proposed here is closely related to the information content measure proposed by van Heel and Schatz (2020), in which they followed a different line of reasoning. They proposed a Fourier Shell Information (*FSI*) measure for the information, in bits, gained by a shell of data in Fourier space, expressed in terms of the *FSC* between two half-maps for that resolution shell:

$$FSI(s) = K \log_2\left(\frac{1+FSC(s)}{1-FSC(s)}\right), \tag{28}$$

where K is the effective number of independent Fourier terms in the shell under consideration. As noted above, D_{obs} plays the same role for a single Fourier term as FSC_{ref} does for a resolution shell. If we assume that all Fourier terms in a resolution shell have the same value of D_{obs}, we can express the *FSI* equation in terms of D_{obs} using the relationship between FSC and FSC_{ref} derived by Rosenthal and Henderson (2003).

$$D_{obs} = FSC_{ref} = \sqrt{\frac{2FSC}{1+FSC}} \tag{29}$$

Solving for *FSC* yields

$$FSC = \frac{D_{obs}^2}{2-D_{obs}^2} \tag{30}$$

Substituting this for FSC in (28) and simplifying yields the following:

$$FSI(s) = -K \log_2(1 - D_{obs}^2) \tag{31}$$

Interpreting *K* as the number of independent Fourier terms in the shell, this is equivalent to the expression given above for the Kullback-Leibler divergence measured in bits. The expression given here is more general, because it allows for differences in accuracy of different Fourier terms around a shell, arising from anisotropy and the effects of favoured orientations, which will lead to variation among the values of D_{obs} for different terms.

In our docking calculations, the information gain calculation is used to save computing time by omitting Fourier terms that will have almost no effect on the likelihood calculation. As done in the related molecular replacement calculation (Jamshidiha *et al.*, 2019), Fourier terms with an information gain of less than 0.01 bit are ignored after the error analysis step.

4. Implementation of algorithms

The algorithms have been implemented as a combination of Python scripts and C++ code, both making substantial use of the Computational Crystallography Toolbox, cctbx (Grosse-Kunstleve *et al.*, 2002).

Tools to analyze the maps, determine the parameters characterizing the signal and noise, and compute modified Fourier coefficients for the docking calculation have been implemented in the Python program *prepare_map_for_docking*.

The *prepare_map_for_docking* tool is available as a Python script within the *maptbx* section of the open-source Computational Crystallography Toolbox, *cctbx* (Grosse-Kunstleve *et al.*, 2002). This is available standalone and also as part of the *Phenix* (Liebschner *et al.*, 2019) and *CCP4* (Winn *et al.*, 2011) software suites.

5. Results

5.1. Behaviour of signal and error analysis

As noted by Palmer and Aylett (2022), errors are similar throughout a cryo-EM reconstruction, but signal-to-noise ratios can vary dramatically within the reconstruction because of variations in the strength of the signal. This can be demonstrated by looking at the local behaviour of the signal power (Σ_S) and noise power (Σ_E) in reciprocal space, after the analysis using the *prepare_map_for_docking* tool. One informative example is the map for conformation 2 of the *E. coli* respiratory complex I (EMDB entry 12654, PDB entry 7nyu), for which the local reconstruction quality varies widely (Kolata & Efremov, 2021). An analysis of one of the best and one of the worst regions of the map is given in Fig. 1, illustrating that the noise power is similar in the two regions whereas the signal power, and its variation in Fourier space, differs substantially.

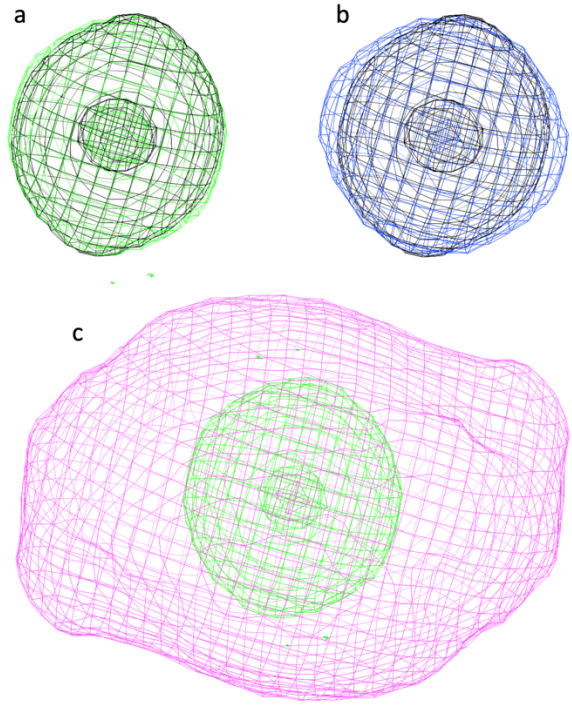

Figure 1 Variation of signal power (Σ_S) and noise power (Σ_E) in Fourier space for well-ordered (center of cytoplasmic domain) and poorly-ordered (chain L) regions of the *E. coli* respiratory complex I. The local mean signal and noise variances are all contoured at a level where the two are approximately equal for values computed from a sphere with a radius of 30 Å from the poor region of the map. a) The distribution of Σ_E computed for the poorly-ordered region around chain L (black) is seen to be similar to the distribution of Σ_E computed for the well-ordered region (green). b) At the chosen contour level the distribution of Σ_S for the poorly-ordered region (blue) is similar to that for the distribution of Σ_E, with the contoured surface corresponding roughly to a resolution of 10 Å. c) The signal power for the well-ordered region (magenta) extends to much higher resolution than the noise power for the same region of the map (green).

6. Discussion and conclusions

The problem of docking an atomic model into a cryo-EM reconstruction is reminiscent of the molecular replacement problem in crystallography. The similarity is more than superficial, as both problems can be addressed using likelihood functions that start from joint distributions of complex Fourier terms. In both cases, the model is represented by its Fourier transform (either of its electrostatic potential or its electron density), but cryo-EM differs in the important fact that the data retain the phase information lost in the crystallographic diffraction experiment.

Applying likelihood requires characterising all sources of error, which differ between the methods. In cryo-EM, the typical presence of favoured particle orientations leads to large

differences in the reliability of the Fourier terms. The variation of noise contributions to the Fourier terms is expected to vary smoothly over Fourier space, and a method to assess this variation has been developed.

The likelihood framework allows the implementation of tools that have been found useful in molecular replacement. In particular the expected log-likelihood-gain (eLLG) score can be calculated in advance of any molecular replacement search, as well as the information gained by making the cryo-EM reconstruction.

The accompanying paper describes the implementation of these ideas in software tools for docking, and the success of those tools demonstrates the validity of the approach described here, including the use of eLLG to choose optimal strategies.

7. References

Cragnolini, T., Sahota, H., Joseph, A. P., Sweeney, A., Malhotra, S., Vasishtan, D. & Topf, M. (2021). *Acta Cryst D*. **77**, 41–47.

Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). *J Appl Cryst*. **35**, 126–136.

van Heel, M. & Schatz, M. (2020). Information: to Harvest, to Have and to Hold arXiv.

Jamshidiha, M., Pérez-Dorado, I., Murray, J. W., Tate, E. W., Cota, E. & Read, R. J. (2019). *Acta Crystallographica Section D: Structural Biology*. **75**,.

Joseph, A. P., Lagerstedt, I., Patwardhan, A., Topf, M. & Winn, M. (2017). *Journal of Structural Biology*. **199**, 12–26.

Kolata, P. & Efremov, R. G. (2021). *ELife*. **10**, e68710.

Kullback, S. & Leibler, R. A. (1951). *Ann. Math. Stat.* **22**, 79–86.

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkoczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L.-W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., Stockwell, D. H., Terwilliger, T. C., Urzhumtsev, A. G., Videau, L. L., Williams, C. J. & Adams, P. D. (2019). *Acta Crystallogr D Struct Biol*. **75**, 861–877.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *Journal of Applied Crystallography*. **40**, 658–674.

McCoy, A. J., Oeffner, R. D., Wrobel, A. G., Ojala, J. R. M., Tryggvason, K., Lohkamp, B. & Read, R. J. (2017). *Proceedings of the National Academy of Sciences of the United States of America*. **114**, 3637–3641.

Murshudov, G. N. (2016). *Methods Enzymol*. **579**, 277–305.

Oeffner, R. D., Afonine, P. V., Millán, C., Sammito, M., Usón, I., Read, R. J. & McCoy, A. J. (2018). *Acta Crystallographica Section D: Structural Biology*. **74**, 245–255.

Palmer, C. M. & Aylett, C. H. S. (2022). *Acta Crystallogr D Struct Biol*. **78**, 136–143.

Popov, A. N. & Bourenkov, G. P. (2003). *Acta Cryst D*. **59**, 1145–1153.

Ramlaul, K., Palmer, C. M. & Aylett, C. H. S. (2019). *Journal of Structural Biology*. **205**, 30–40.

Read, R. J. (1990). *Acta Crystallographica Section A*. **46**, 900–912.

Read, R. J. & McCoy, A. J. (2016). *Acta Crystallographica. Section D, Structural Biology*. **72**, 375–387.

Roseman, A. M. (2000). *Acta Crystallogr D Biol Crystallogr*. **56**, 1332–1340.

Rosenthal, P. B. & Henderson, R. (2003). *Journal of Molecular Biology*. **333**, 721–745.

Scheres, S. H. W. (2012). *Journal of Molecular Biology*. **415**, 406–418.

Schomaker, V. & Trueblood, K. N. (1968). *Acta Crystallogr B Struct Crystallogr Cryst Chem*. **24**, 63–76.

Srinivasan, R. & Ramachandran, G. (1965). *Acta Crystallographica*. **19**, 1008–1014.

Stewart, P. L., Fuller, S. D. & Burnett, R. M. (1993). *EMBO J*. **12**, 2589–2599.

Storoni, L. C., McCoy, A. J. & Read, R. J. (2004). *Acta Crystallogr D Biol Crystallogr*. **60**, 432–438.

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A. J., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A. & Wilson, K. S. (2011). *Acta Crystallogr D Biol Crystallogr*. **67**, 235–242.

Wriggers, W. (2012). *Acta Cryst D*. **68**, 344–351.

Wu, X., Milne, J. L. S., Borgnia, M. J., Rostapshov, A. V., Subramaniam, S. & Brooks, B. R. (2003). *Journal of Structural Biology*. **141**, 63–76.

Zundert, G. C. P. van, Bonvin, A. M. J. J., Zundert, G. C. P. van & Bonvin, A. M. J. J. (2015). *AIMSBPOA*. **2**, 73–87.