Likelihood-based docking of models into cryo-EM maps

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

[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 Exploiting analogies to molecular replacement, a strategy for docking into cryo-EM maps is informed by calculations of expected log-likelihood-gain scores.

Abstract Optimized docking of models into cryo-EM maps requires exploiting expected signal in the data to minimize the calculation time while maintaining sufficient signal. The likelihood-based rotation function used in crystallography can be employed to establish plausible orientations in a docking search. A phased likelihood translation function yields scores for the placement and rigid-body refinement of oriented models. Optimised strategies for choices of the resolution of data and the size of search volumes are based on expected log-likelihood-gain scores, computed in advance of the search calculation. Tests demonstrate that the new procedure is fast, robust and effective at placing models into even challenging cryo-EM maps.

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

1. Introduction

Advances in cryo-EM hardware and software are improving the resolution and quality of cryo-EM maps, often yielding maps that allow model-building from scratch. Nevertheless, for various sample-specific and technical reasons, a substantial proportion of cryo-EM maps from single-particle reconstructions and a larger proportion of maps from sub-tomogram averaging lack the necessary resolution and quality for *ab initio* model building. In this situation, the density may be interpreted by docking one or more pre-existing experimental or predicted atomic models. We explore here the development of new likelihood-based docking tools to fill this need.

A large number of tools have been developed to carry out manual or automated docking. The automated tools include DockEM (Roseman, 2000; Titarenko & Roseman, 2021), Situs (Kovacs & Wriggers, 2002; Wriggers, 2012), Powerfit (Zundert *et al.*, 2015), OffGridFit (Hoffmann *et al.*, 2017), phenix.dock_in_map (Liebschner *et al.*, 2019), and MrBUMP (Simpkin *et al.*, 2021). DockEM, Powerfit and phenix.dock_in_map all carry out an exhaustive exploration of orientations. Situs and OffGridFit both use 6-dimensional FFT-based algorithms for an exhaustive 6D search. Among these, MrBUMP is unique in carrying out the translation search first with the spherically-averaged phased translation function, followed by an orientation search centered on the point found in the translation search.

We have not attempted to carry out head-to-head comparisons of our software with existing tools for two reasons. First, the half-maps needed for our approach are not generally available for the published test cases for existing tools. Second, we are not experts in the use of the other tools and would therefore not be able to show them to their best advantage.

2. Expected LLG-based search strategy

Docking problems can differ dramatically in their difficulty, from trivial cases where distinctive features of the search model could be spotted by eye in an excellent map, to extremely challenging cases where there is barely enough signal to recognise that a docked model agrees with a very noisy map. Great gains can be made in the efficiency and effectiveness of docking calculations by adopting a case-dependent strategy that is informed by considering the expected LLG (eLLG).

In molecular replacement (MR), we have found that searches yielding an LLG value of 60 or greater after a combined rotation/translation search are almost always correct (McCoy *et al.*, 2017; Oeffner *et al.*, 2018). In cryo-EM, after correcting for oversampling, our experiences suggest that a similar threshold applies. Given uncertainties about the sizes of coordinate errors prior to structure solution, trials of different choices in a database of MR problems showed that it is more efficient, overall, to choose strategy parameters expected to give a higher LLG score than 60, with 225 being a choice that works well to balance an increased initial search cost with a lower chance of having to rerun an unsuccessful search with modified parameters.

The pivotal decisions in the docking search strategy are determined by the rotation search, because it gives the lowest signal-to-noise; if this search is expected to succeed (or at least to give sufficient signal that a chosen subset of orientations is likely to include the correct orientation), then the subsequent translation search will be almost certain to succeed.

Given uncertainties in the eLLG calculations, particularly in estimating the quality of the search model in advance, searches that aim for the minimum required LLG have a significant chance of failure, and it is safer to be somewhat more conservative so that fewer searches need to be repeated. For the rotation search, an LLG score of 30 or more is expected to correspond to a correct solution, as this is equivalent to the P1 search score required for confidence in a crystallographic MR search (McCoy *et al.*, 2017). As a more conservative estimate, the initial target $eLLG_{rot}$ is set to 60.

2.1. Searching over the whole map with one rotation search

A major decision in the search strategy is whether a rotation search over the whole map is likely to succeed. For good maps and good models that comprise a sufficient fraction of the total structure, a strong signal will be expected in the rotation search. Searches can then be carried out over the whole map, but the efficiency can be optimised by reducing the resolution to what is required to achieve $eLLG_{rot} = 60$. In principle, an even lower resolution limit for Fourier terms could be used in the translation search, but in practice the translation search is computationally very efficient, and only a few translation searches will be required if there is good signal in the rotation search. Even if $eLLG_{rot} = 7.5$, the correct orientation is likely to be found in an orientation list of modest size, so carrying out a rotation plus translation search over the entire map is abandoned only if $eLLG_{rot} < 7.5$.

2.2. Searching over sub-volumes

If it is not judged possible to search successfully for rotations using the full map, a decision is made whether it will be possible, instead, to find a solution by searching over sub-volumes. The target sub-volume is set according to the inverse relationship between the size of the sub-volume and the $eLLG_{rot}$ that would be achieved by searching in that sub-volume (if it contained the object being sought). As a simple example, if a value of 3.75 were found for $eLLG_{rot}$ when computed over the whole map, the $eLLG_{rot}$ for a map containing one-half of the total volume would be 7.5. This calculation depends on the assumption that one of the sub-volumes will contain the entire object being sought, so there is a lower limit to the smallest relevant sub-volume. It is also implicitly assumed that the map quality in local regions is not much worse than the overall average map quality. This can lead to failures when the component being sought corresponds to a poor part of the map. Note that there is a practical limit to how small a sub-volume can be; the number of sub-volumes required to ensure that at least one of them contains the full volume of the model grows dramatically once the search volume is less than about 1.15 times the volume of the sphere enclosing the model.

When a suitable size has been defined for the sub-volumes, target sub-volumes for docking searches are constructed as follows. First, a hexagonal close-packed grid is defined, such that spheres with the target volume that are centered on the grid points will overlap sufficiently that at least one of the spheres is guaranteed to cover the volume containing the target object. Second, any spheres that lack sufficient ordered volume (defined as regions of the map with high local variance) to contain the search object are discarded. Following this, the spheres of density are analysed to evaluate signal and noise (to calibrate the likelihood targets), and then rotation and translation searches are carried out, followed by rigid-body refinement. To avoid Fourier artefacts from sharp boundaries in the map, the target sphere is cut out inside a cube large enough to allow a smooth masking of the density to the edges.

2.3. Brute-force six-dimensional search

If the rotation search cannot be carried out with sufficient signal even with sub-volumes, then the final fall-back in the search algorithm is to carry out a brute-force six-dimensional search. To make this search as efficient as possible, data are used only to the resolution required to obtain a value of $eLLG_{tra}$ sufficient to yield a clear solution for the correct rotation and translation. Based on experiences with crystallographic MR, searches given an LLG of 60 should almost always be correct, but to be safe the target for $eLLG_{tra}$ is set to 225, a value that has been adopted in *Phaser* to give a good compromise between efficiency and the danger of missing the solution. Using the lowest resolution possible improves efficiency by allowing orientations and translations to be sampled more coarsely, and by reducing the number of Fourier terms over which the likelihood scores must be calculated. Even so, it is not uncommon for such a brute-force search to take hours to run.

2.4. Focused docking

The final step in the docking strategy is to evaluate all potential docking solutions in a common framework. For each potential solution, the size of the sphere of density required to accommodate the entire search model is evaluated, a sphere of density of that size is cut out, the analysis of signal and noise is carried out, and then a rigid-body refinement is carried out to obtain an LLG score, a final model placement and a map correlation with the processed density sphere.

Two types of map coefficients (equations 18 and 19 from the accompanying paper) for the processed density sphere have been evaluated in the set of tests described. The first type ($\mathbf{F}_{map} = D_{obs}\mathbf{E}_{mean}$) should give a map that minimises the error from the true sharpened map. The second type ($\mathbf{F}_{map} = \frac{2}{1-D_{obs}^2\sigma_A^2} D_{obs}\sigma_A\mathbf{E}_{mean}$) includes an additional weighting term from the likelihood target and therefore gives a map for which the sum of densities at atomic positions should be roughly proportional to the likelihood score for that model. To compute the second map, a choice has to be made for the value of σ_A. The current default is to assign a value of 0.9, which would correspond to a model that accounts for about 80% of the scattering in the volume under consideration but has no other errors. The choice for σ_A could potentially be improved by considering deficiencies in the ability of atomic models to account for the bulk solvent region. The second choice for map coefficients yielded higher map correlations than the first in the test calculations reported below. Qualitatively, the blurring that comes from giving higher weight to well-determined (typically lower-resolution) Fourier terms seems to give more readily interpretable maps. The second choice, therefore, is the default and was used for the calculations reported below.

3. Methods

3.1. Target selection

A set of single particle cryo-EM structures was chosen that would convey a representative sample of experimental reconstructions covering a range of resolutions (1.7-8.5 Å), overall quality and symmetry conditions (1-24 symmetric copies). The test cases were restricted to EMDataBank (EMDB) (Lawson *et al.*, 2016), entries for which half-maps had been deposited. Table 1 shows a summary of the selected test cases.

3.2. Model selection

Models were selected to cover a variety of scenarios. Some models correspond to what could be called “reference” models, in the sense that they are the deposited models associated with the EMDB entry; these provide a reference docking with nearly zero rotation or translation. Others correspond to crystal structures of the same protein. Finally, we have tested some predicted models produced by AlphaFold (Jumper *et al.*, 2021) (AF); such models will be used frequently, so understanding how they should be treated and how they will perform in our algorithm is essential. In all cases, we processed the predicted models with the process_predicted_model tool (Oeffner *et al.*, 2022), which replaces the predicted values for the local distance difference test (Mariani *et al.*, 2013) in the B-factor column of the coordinate file with appropriate B-factors to down-weight the less-confident parts of the model, as well as trimming off residues with a predicted local difference distance test less than 70 (on a scale of 0-100).

To determine the effect of model completeness, as well as local map quality, we also tested the effect of using smaller pieces of the structural model (individual chains, domains or sub-domains). The models are also summarised in Table 1.

Table 1 Cryo-EM structures and models used for docking tests

Target	**code**[*]	**d_{min}**	**copies**[†]	**model**[‡]	**fraction**[‖]	**model type**
GABA receptor	7a5v_11657	1.7	5	4cofA (291-447) (membrane domain)	0.055	crystal structure
				AF model of megabody	0.051	prediction
Beta-galactosidase	5a1a_2984	2.2	4	1jz7A	0.25	crystal structure
				5a1aA	0.25	reference
				5a1aA beta-barrel domain (626-726)	0.025	reference
Apoferritin	5xb1_6714	3.0	24	5xb1A	0.042	reference
				2cei (5-159)	0.042	crystal structure
Respiratory complex	7nyu_12654	3.8	1	3rkoAJKLMN (membrane domain)	0.42	crystal structure
				3rkoN	0.11	crystal structure
				3rkoM	0.11	crystal structure
				3rkoL (1-546)	0.11	crystal structure
				7nyuM	0.11	reference
MutS	7ai6_11792	6.9	(2)	6i5f (566-799)	0.13	crystal structure
CFTRΔF508 mutant	8ej1_28172	6.9	1	AF (4- 263, 282-379, 844-871, 933-1170) (membrane domain)	0.52	AF, no template
				AF (264-281, 1204-1429)	0.19	AF, no template
				AF (391-400, 440-633)	0.17	AF, no template
Get3, closed	EMD_25375	8.46	(2)	7spz[¶]	0.5	crystal structure

[*] Code for PDB plus EMDB pair, with the PDB identifier followed by the EMDB deposition number.

[†] Number of symmetry-related copies (or pseudo-symmetric copies if in parentheses)

[‡] Models from PDB entries are defined in terms of the PDB identifier, optionally followed by a chain identifier and/or a range of residue numbers. AF indicates AlphaFold model.

[‖] Fraction of the entire reconstruction explained by one copy of the model

[¶] Structure of Get3 in the open conformation

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).

The framework for the docking search has been implemented in the Python program *em_placement*, which is part of the Voyager structural biology framework built on *phasertng* (McCoy *et al.*, 2021). Associated tools required to evaluate the map eLLG, map information gain, fast phased translation search and cryo-EM likelihood target have been added to *phasertng*, which already contained tools to compute the rotation function eLLG (McCoy *et al.*, 2017), fast searches and LLG rescoring for rotations (Storoni *et al.*, 2004), and phased rigid-body refinement (Millán *et al.*, 2021).

Note that the symmetry of the reconstruction is not yet used to aid model placement in the current version of the program.

The *em_placement* program is controlled using a set of keywords in the phil syntax used in Phenix (Liebschner *et al.*, 2019). An example keyword script is given in Appendix A. Most keywords (map files, model file, composition of the reconstruction defined in terms of sequences of the components) will not usually be altered. The nominal resolution of the map is required, and the author-defined value in the EMDB entry was used in all cases reported here. An appropriate choice is the FSC derived resolution. Since the nominal resolution is used as the high-resolution limit for all the calculations, if the map actually contains valid higher resolution features than the user-entered value, some signal will be lost. If there is really no information in the highest resolution data used, CPU time is wasted but the search results should not be degraded unless the nominal resolution is very over-optimistic. The only parameter that might be varied by the user is the equivalent RMS error that defines the expected model quality. For the test cases, a value of 0.8 Å was used for models obtained from experimental structures of the same protein, 1.0 Å for models predicted by AlphaFold, and 1.2 Å for the model of apoferritin derived from a structure that contains the helix deleted in the target structure.

Data used for test calculations are all available through the EMDataBank (Lawson *et al.*, 2016). Cryo-EM and crystallographic models are available from the worldwide Protein Data Bank (Berman *et al.*, 2007), except for the AF models, which were computed using the community ColabFold version (Mirdita *et al.*, 2022) of AlphaFold (Jumper *et al.*, 2021).

5. Results

5.1. Docking results

The results of the docking trials are summarized in Table 2. The majority of the searches succeeded, and many of these required only a single search over the entire reconstruction. The time required for the searches ranged from half a minute to about 31 minutes, averaging about 12 minutes over the set of test cases. When multiple spherical sub-volumes were searched, the number varied from 4 to 43. None of the test cases triggered the fall-back of carrying out a brute-force six-dimensional search.

Table 2 Results of docking trials

Target	model	docking spheres[*]	copies placed[†]	LLG score[‡]	mapCC[‡]	run time (seconds)[‖]
GABA receptor	4cofA (291-447) (membrane domain)	1	5/5	5660-5899	0.759-0.762	474
	AF model of megabody	6	1/5	348	0.369	1884
Beta-galactosidase	1jz7A	1	4/4	23532-24117	0.814-0.815	589
	5a1aA	1	4/4	23420-24089	0.817-0.820	606
	5a1aA beta-barrel domain (626-726)	14	2/4	1615, 1616	0.778, 0.778	1111
Apoferritin	5xb1A	1	24/24	2672-2675	0.843-0.843	923
	2cei (5-159)	1	24/24	1834-1839	0.732-0.733	851
Respiratory complex	3rkoAJKLMN (membrane domain)	1	1/1	560	0.336	666
	3rkoN	4	1/1	620	0.463	392
	3rkoM	4	1/1 (1)	213 (189)	0.472 (0.268)	777
	3rkoL (1-546)	4	0/1 (2)	(66, 141)	(0.249, 0.263)	1056
	7nyuM	4	1/1 (1)	176 (150)	0.368 (0.268)	714
MutS	6i5f (566-799)	43	2/2	167, 174	0.605, 0.620	306
CFTRΔF508 mutant	AF (4- 263, 282-379, 844-871, 933-1170)	1	1/1	495	0.637	120

	(membrane domain)					
	AF (264-281, 1204-1429)	14	1/1	66	0.654	466
	AF (391-400, 440-633)	32	0/1 (5)	(8-14)	(0.142-0.264)	1565
Get3, closed	7spz	1	2/2	216, 319	0.632, 0.688	28

[*] Number of sub-volume spheres used for docking search, 1 for a single sphere covering the entire reconstruction.

[†] Number of copies placed correctly (or incorrectly in parentheses)

[‡] Scores for incorrectly-placed copies are in parentheses

[‖] Linux workstation with 3.8GHz Intel Core i7-9800X CPU with 16 cores, but running primarily on a single thread

5.1.1. GABA receptor

The highest-resolution (1.7 Å) cryo-EM structure in our test set is that of the human γ-aminobutyric acid receptor bound to a megabody, PDB entry 7a5v, EMDB entry 11657 (Nakane *et al.*, 2020).

To provide a reasonable challenge at such high resolution, only small models were tested, each comprising about 1/20 of the full pentamer or 1/4 of a single copy. The membrane domain is well-ordered and is easy to place when using the membrane component of a single subunit of a crystal structure, PDB entry 4cof (Miller & Aricescu, 2014), as a model. However, an AlphaFold model of the bound megabody is more difficult to place, as the associated density is the least well-ordered in the map. Only 1 of the 5 copies was placed successfully, in spite of the 5-fold symmetry of the reconstruction. The sensitivity of the search to the boundaries of the sub-volumes is an indicator that this is a marginal model for searching in this map. In principle, the missing copies could be generated by application of the 5-fold symmetry.

5.1.2. Beta-galactosidase

Beta-galactosidase is commonly used as a test object for cryo-EM methodology, as it is well-behaved and possesses D2 tetrameric symmetry. We chose a medium-resolution (2.2 Å) representative: PDB entry 5a1a, EMDB entry 2984 (Bartesaghi *et al.*, 2015).

Docking a full chain, either from the associated PDB entry or from a crystal structure, PDB entry 1jz7 (Juers *et al.*, 2001), is straightforward to achieve by searching over the full map. On the other hand, docking just the beta-barrel domain of one subunit is substantially more challenging, and the map is divided into 14 sub-volumes. Because of the marginal nature of this case, only 2 of 4 copies were found successfully. Again, the other two copies could have been recovered by exploiting the symmetry of the map.

5.1.3. Apoferritin

Because of its stability and high octahedral (432) symmetry, apoferritin is another very common test object for cryo-EM. We chose a relatively low-resolution (3.0 Å) representative: PDB entry 5xb1, EMDB entry 6714, a deletion mutant of the E-helix (Ahn *et al.*, 2018).

Searching with a single chain from the reference structure finds all 24 copies with strong signal in a search over the full volume. As a more challenging test, we based a search model on a single chain from PDB entry 2cei, the crystal structure of a full-length version of apoferritin, removing the E-helix from the search model. Again, all 24 copies were found with strong (though slightly lower) signal. Note that much of the computing time in these two tests is expended on evaluating the map correlations for the 24 solutions.

5.1.4. *E. coli* respiratory complex I

The largest series of trials was carried out with the reconstruction for conformation 2 of the *E. coli* respiratory complex I: PDB entry 7nyu, EMDB entry 12654 (Kolata & Efremov, 2021). This reconstruction presents a variety of challenges, as the overall resolution (3.8 Å) is already relatively low but also varies substantially over the different subunits. Parts of the membrane domain are particularly poorly resolved; the local resolution of chain L is estimated by the authors as being in the range of 9-11 Å. An additional challenge comes from the fact that three of the membrane domain components (chains L, M and N) have related sequences and structures, with pairwise sequence identities of 25-26%.

Models were taken either from the reference structure or from the crystallographic structure of the membrane domain, PDB entry 3rko (Efremov & Sazanov, 2011). Searching for the entire membrane domain gives a clear solution using the full reconstruction. In searches for individual chains, such as the three related membrane domain components, the reconstruction is divided into sub-volumes. For the best-ordered of the three related subunits, chain N, an unambiguous solution is found. Chain M is more poorly-ordered, and two potential solutions are found. The solution with higher scores is correctly placed, while the second solution superimposes the chain M model on the better-ordered density of chain N. Chain L is the least well-ordered, and the search places the model on either the density for chain N or chain M, but not on the correct density that corresponds to chain L. Fig. 1a illustrates the most difficult successful result, showing the docked model of chain M.

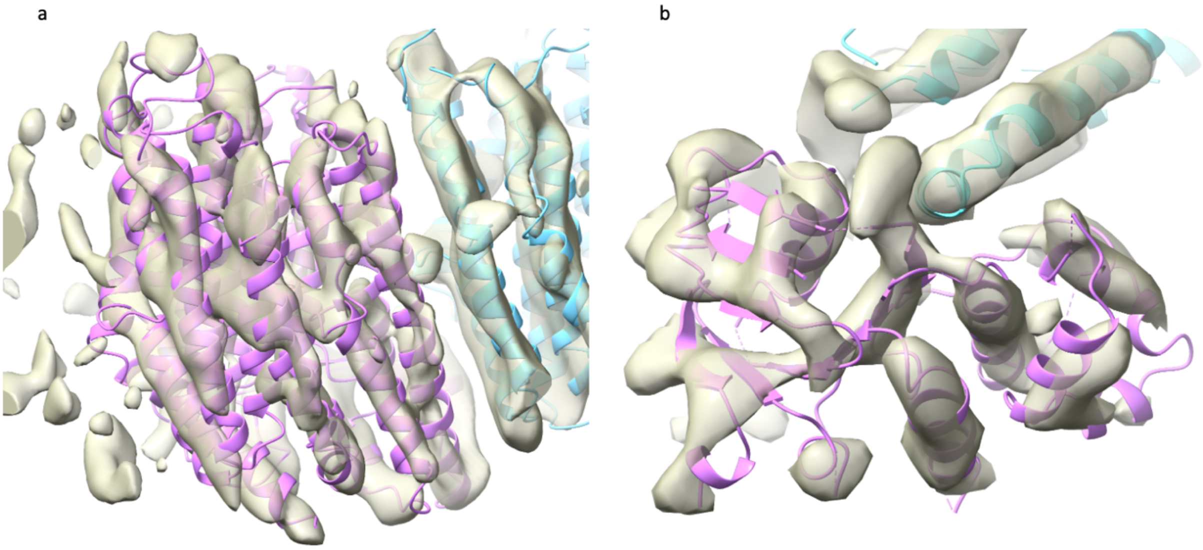

Figure 1 Docked models in maps for challenging cases. Both maps are computed using the Fourier coefficients $D_{obs}\mathbf{E}_{mean}$ arising from the analysis of the local map volumes, and the images were made with ChimeraX (Goddard *et al.*, 2018). a) Chain M (magenta) of PDB entry 3rko, docked into the region of the map corresponding to chain M of PDB entry 7nyu (associated with EMDB entry 12654). Chain N is shown in light blue. B) The AlphaFold model of the smallest domain of the ΔF508 mutant of CFTR (magenta), docked into the corresponding region of the map derived from EMDB entry 28172. The membrane domain is shown in light blue.

5.1.5. DNA mismatch repair protein, MutS

For one representative of a low-resolution (6.9 Å) reconstruction, we chose the *E. coli* DNA mismatch repair protein, MutS, in its mismatch-bound state: PDB entry 7ai6, EMDB entry 11792. In this bound state, the protein is a pseudosymmetric dimer, so there are two independent copies to find.

To test a workflow in which individual domains are docked, in order to approximate a conformational change, we used the C-terminal domain of one chain of MutS in the DNA-free conformation, from PDB entry 6i5f (Bhairosing-Kok *et al.*, 2019). For such a small fraction of the full structure at such low resolution, the signal in the rotation search would be extremely low, so the sub-volume determination algorithm chose to carry out the search with 43 spherical sub-volumes of the maps. Although this is a large number, each calculation is fast with low-resolution data, and an unambiguous docking of both copies was achieved in about 5 minutes.

5.1.6. Cystic fibrosis transmembrane regulator, ΔF508 mutant

The ΔF508 mutant of the cystic fibrosis transmembrane regulator (CFTR), with bound folding modulators, was chosen as a second low-resolution (6.9 Å) reconstruction: PDB entry 8ej1, EMDB entry 28172 (Fiedorczuk & Chen, 2022).

Rather than testing other experimental structures of the same protein, we chose to make AlphaFold (Jumper *et al.*, 2021) models in the ColabFold environment (Mirdita *et al.*, 2022). Although structures of the CFTR would have been present in the training data for AlphaFold, their influence was reduced by turning off the option to include explicit templates of related structure in the structure prediction process. As for the MutS case, the difficulty of the docking calculations was increased by extracting models of individual domains from the full predicted structure. As expected, it was more difficult to place smaller models. The membrane domain, the largest with 585 residues in the processed model, was placed easily (LLG = 495) in a search over the entire reconstruction. A mid-sized domain, comprising 214 residues, gave a clear solution with an LLG of 66 but searching over 14 sub-volumes and taking nearly four times as long. The smallest domain, comprising 187 residues, failed to yield a convincing solution, as judged from the fact that there were five potential solutions that had low LLG values in the range of 8-14, much lower than the value of 60 that should be achieved for an unambiguous solution. On examination, all five potential solutions were indeed incorrect.

5.1.7. Get3, closed conformation

The lowest-resolution (8.46 Å) map in the test set is of the closed conformation of the ER targeting factor Get3: EMDB entry 25375 (Fry *et al.*, 2022). The authors did not deposit coordinates in the PDB for this reconstruction, presumably because it had the lowest resolution of a series of maps. Therefore, it makes a good example for the circumstance in which a structural biologist would like to examine a published map in the context of a docked model from another structure.

We chose the crystal structure the same authors determined for the open conformation, from PDB entry 7spz (Fry *et al.*, 2022), as the model. The reconstruction is pseudosymmetric, so there are two independent copies to find. Both of them can be found in a straightforward search over the full reconstruction that takes only about half a minute.

5.2. Checking the $eLLG_{rot}$-guided sub-volume criterion

In the global search, the $eLLG_{rot}$ criterion suggested that a single search sphere covering the entire ordered volume of the reconstruction would give sufficient rotation function signal for 8 of the 17 test cases. Validating this, all 8 of these searches succeeded (Table 2). However, if the $eLLG_{rot}$ criterion were too pessimistic about the ability to find the model in the whole map, the 7 searches that found at least one copy when searching over sub-volumes might have succeeded with a global search. To test this, we used a manual override in the *em_placement* program to force a search over the single sphere covering the entire ordered volume. Because the two models for chain M of the *E. coli* respiratory complex I are very similar, we only tested chain M from the crystal structure in PDB entry 3rko. The results are given in Table 3.

Table 3 Results of trials searching globally over a single sphere

Target	**model**	**original docking spheres**[*]	**copies placed**[†]	**LLG score**[‡]	**mapCC**[‡]	**run time (seconds)**[‖]
GABA receptor	AF model of megabody	6	0/5	(14-28)	(0.026-0.035)	1045
Beta-galactosidase	5a1aA beta-barrel domain (626-726)	14	0/4	(18-57)	(0.100-0.215)	680
Respiratory complex	3rkoN	4	0/1	(11-23)	(0.057-0.099)	670
	3rkoM	4	0/1	(12-23)	(0.049-0.083)	777
MutS	6i5f (566-799)	43	1/2	174	0.618	28
CFTRΔF508 mutant	AF (264-281, 1204-1429)	14	1/1	67	0.655	132

[*] Number of sub-volume spheres used for original automated docking search
[†] Number of copies placed correctly
[‡] Scores for incorrectly-placed copies are in parentheses
[‖] Linux workstation with 3.8GHz Intel Core i7-9800X CPU with 16 cores, but running primarily on a single thread

The results support the $eLLG_{rot}$ criterion as an effective guide to search strategy. No correct solution is found for four of the six test cases, and only one of two solutions is found when searching for the C-terminal domain of MutS. The only case where the criterion was clearly too pessimistic about the ability to find the model in the whole map is the search for the mid-sized domain of the AlphaFold model for the ΔF508 mutant of CFTR. Here the correct solution is found in a little over two minutes in the whole map, whereas the global sub-volume search took about 26 minutes (Table 2).

The search for a beta-barrel domain in beta-galactosidase gave a surprising result. The top three solutions have better than random LLG scores of 57. On examination, these correspond to superpositions of the model (comprising residues 626-726) on three of four copies of another beta-barrel domain comprising residues 220-320, detecting the low structural similarity between the domains. These solutions were either not found or were rejected in the default search, because the correct placement yields dramatically higher scores (Table 2).

5.3. Tests of brute-force six-dimensional searches

The two cases where the global search failed, as well as the MutS case in which 43 sub-volumes were explored, provided tests of the brute-force 6D fall-back algorithm. These were carried out to examine whether the global 6D search could succeed for cases where rotation searches for the smallest practical sub-volume would have insufficient signal, and also how it compares in efficiency to searching over a large number of sub-volumes.

5.3.1. Chain L of the *E. coli* respiratory complex I

The brute-force 6D search fails to find the correct position of chain L, but does reproduce the results of the automated search using multiple sub-volumes as the model for chain L is superimposed on the map regions for chains M and N. The run time is dramatically longer at approximately 10 hours, compared to about 18 minutes for the automated search with multiple sub-volumes.

5.3.2. Smallest domain from AlphaFold model of the ΔF508 mutant of CFTR

The brute-force 6D search, taking 102 minutes (compared to about 26 minutes for the automated search), succeeds in placing this domain correctly, with LLG=37 and a map correlation of 0.641 (Fig. 1b). In addition, it finds a solution superimposing this domain on the mid-sized domain, with LLG=29 and a map correlation of 0.553. The two domains are in fact related to each other, with a sequence identity of 27% over 168 matched residues (of 187 in the smaller of the two domains). However, the marginal quality of signal-to-noise in this case is indicated by the presence of a third, incorrect, placement, with LLG=21 and a map correlation of 0.229.

5.3.3. C-terminal domain of MutS

Both copies of the C-terminal domain of PDB entry 6i5f are found in the brute-force 6D search, with the same scores. However, the search using 43 sub-volume spheres is considerably more efficient, taking about 5 minutes compared to 98 minutes for the brute-force 6D search.

6. Discussion and conclusions

The strength of likelihood as a criterion is supported by the success of our new likelihood-based approach to docking models in a series of progressively more challenging cryo-EM maps. Since the successful application of likelihood to a problem requires a good model of the sources of error and their propagation, these results also support our approach to characterising the signal power and noise power as independent smoothly-varying functions in Fourier space.

The outcomes of different search strategies can be predicted by an analysis of the expected log-likelihood-gain (eLLG) score for both the rotation and translation search components of the docking algorithm. The rotation function eLLG can be used to predict how large a volume of the map can be explored in one rotation search, allowing automated decisions about the subdivision of the full map into spherical sub-volumes. The choices made by this criterion have been validated by comparing the success of searches over the full map with those carried out over the suggested sub-volumes.

Docking models into the most poorly-ordered part of a map is difficult, partly because of the reduced signal to noise but also because the assessment of global map quality can mislead the algorithm determining search strategy into choices that provide insufficient signal in the worst regions of the map. This could potentially be mitigated by adapting the strategic choices to local levels of signal to noise in the reconstruction.

Plans for future enhancements include accounting for symmetry in the search space, which will be significantly more efficient in the case of high-symmetry reconstructions such as the ones for apoferritin. Searches for multiple components will be implemented, which requires accounting for the contribution of previously-placed components in the fit to the experimental data, as well as avoiding clashes between components.

Appendix A. Example script for *em_placement*

The following script defines the search parameters for the *em_placement* script used to run the first test case, docking the membrane component of a model of the GABA receptor derived from PDB entry 4cof into the cryo-EM reconstruction deposited as EMDB entry 11657.

```
voyager
{
  remove_phasertng_folder = True

  map_model
  {
     half_map = emd_11657_half_map_1.map
     half_map = emd_11657_half_map_2.map
     best_resolution = 1.7
     point_group_symmetry = C5
     sequence_composition = 7a5v.fa
  }

  biological_unit {
    molecule
    {
      molecule_name = 4cofA_membrane
      map_or_model_file = 4cofA_membrane.pdb
      starting_model_vrms = 0.8
    }
  }

}
```

Using *Phenix* version dev-4820 or newer, this script can be run with the command: phenix.voyager.em_placement docking_script.phil.

Most parameters specified in the script have been named in a way intended to convey the purpose of that parameter. The remove_phasertng_folder parameter is activated to clean up the graph database produced by *phasertng*, which could be used in other circumstances as part of a larger automation framework or for debugging. The point_group_symmetry feature is only used at the moment to optionally generate a full assembly from a single copy. The sequence_composition parameter specifies the name of a file containing the sequences of all the components in the reconstruction.

7. References

Ahn, B., Lee, S.-G., Yoon, H. R., Lee, J. M., Oh, H. J., Kim, H. M. & Jung, Y. (2018). *Angew Chem Int Ed Engl*. **57**, 2909–2913.

Bartesaghi, A., Merk, A., Banerjee, S., Matthies, D., Wu, X., Milne, J. L. S. & Subramaniam, S. (2015). *Science*. **348**, 1147–1151.

Berman, H., Henrick, K., Nakamura, H. & Markley, J. L. (2007). *Nucleic Acids Research*.

Bhairosing-Kok, D., Groothuizen, F. S., Fish, A., Dharadhar, S., Winterwerp, H. H. K. & Sixma, T. K. (2019). *Nucleic Acids Res*. **47**, 8888–8898.

Efremov, R. G. & Sazanov, L. A. (2011). *Nature*. **476**, 414–420.

Fiedorczuk, K. & Chen, J. (2022). *Science*. **378**, 284–290.

Fry, M. Y., Najdrová, V., Maggiolo, A. O., Saladi, S. M., Doležal, P. & Clemons, W. M. (2022). *Nat Struct Mol Biol*. **29**, 820–830.

Goddard, T. D., Huang, C. C., Meng, E. C., Pettersen, E. F., Couch, G. S., Morris, J. H. & Ferrin, T. E. (2018). *Protein Science*. **27**, 14–25.

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

Hoffmann, A., Perrier, V. & Grudinin, S. (2017). *J Appl Cryst*. **50**, 1036–1047.

Juers, D. H., Heightman, T. D., Vasella, A., McCarter, J. D., Mackenzie, L., Withers, S. G. & Matthews, B. W. (2001). *Biochemistry*. **40**, 14781–14794.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P. & Hassabis, D. (2021). *Nature*. **596**, 583–589.

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

Kovacs, J. A. & Wriggers, W. (2002). *Acta Cryst D*. **58**, 1282–1286.

Lawson, C. L., Patwardhan, A., Baker, M. L., Hryc, C., Garcia, E. S., Hudson, B. P., Lagerstedt, I., Ludtke, S. J., Pintilie, G., Sala, R., Westbrook, J. D., Berman, H. M., Kleywegt, G. J. & Chiu, W. (2016). *Nucleic Acids Research*. **44**, D396–D403.

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.

Mariani, V., Biasini, M., Barbato, A. & Schwede, T. (2013). *Bioinformatics*. **29**, 2722–2728.

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.

McCoy, A. J., Stockwell, D. H., Sammito, M. D., Oeffner, R. D., Hatti, K. S., Croll, T. I. & Read, R. J. (2021). *Acta Crystallographica Section D Structural Biology*. **77**, 1–10.

Millán, C., Keegan, R. M., Pereira, J., Sammito, M. D., Simpkin, A. J., McCoy, A. J., Lupas, A. N., Hartmann, M. D., Rigden, D. J. & Read, R. J. (2021). *Proteins: Structure, Function, and Bioinformatics*. prot.26214-prot.26214.

Miller, P. S. & Aricescu, A. R. (2014). *Nature*. **512**, 270–275.

Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S. & Steinegger, M. (2022). *Nat Methods*. **19**, 679–682.

Nakane, T., Kotecha, A., Sente, A., McMullan, G., Masiulis, S., Brown, P. M. G. E., Grigoras, I. T., Malinauskaite, L., Malinauskas, T., Miehling, J., Uchański, T., Yu, L., Karia, D., Pechnikova,

E. V., de Jong, E., Keizer, J., Bischoff, M., McCormack, J., Tiemeijer, P., Hardwick, S. W., Chirgadze, D. Y., Murshudov, G., Aricescu, A. R. & Scheres, S. H. W. (2020). *Nature*. **587**, 152–156.

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.

Oeffner, R. D., Croll, T. I., Millán, C., Poon, B. K., Schlicksup, C. J., Read, R. J. & Terwilliger, T. C. (2022). *Acta Crystallogr D Struct Biol*. **78**, 1303–1314.

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

Simpkin, A. J., Winn, M. D., Rigden, D. J. & Keegan, R. M. (2021). *Acta Cryst D*. **77**, 1378–1385.

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

Titarenko, V. & Roseman, A. M. (2021). *Acta Cryst D*. **77**, 447–456.

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

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.