bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Automated model building and protein identification in cryo-EM maps

Kiarash Jamali [1*], Lukas Käll [2], Rui Zhang [3],
Alan Brown [4], Dari Kimanius [1*] & Sjors H.W. Scheres [1*]

[1] MRC Laboratory of Molecular Biology,
Cambridge, UK
[2] Science for Life Laboratory, KTH Royal Institute of Technology,
Stockholm, Sweden
[3] Washington University in St. Louis,
St. Louis, MO, USA
[4] Blavatnik Institute, Harvard Medical School,
Boston, MA, USA
[*] Correspondence to {kjamali,dari,scheres}@mrc-lmb.cam.ac.uk

Abstract

Interpreting electron cryo-microscopy (cryo-EM) maps with atomic models requires high levels of expertise and labour-intensive manual intervention. We present ModelAngelo, a machine-learning approach for automated atomic model building in cryo-EM maps. By combining information from the cryo-EM map with information from protein sequence and structure in a single graph neural network, ModelAngelo builds atomic models for proteins that are of similar quality as those generated by human experts. For nucleotides, ModelAngelo builds backbones with similar accuracy as humans. By using its predicted amino acid probabilities for each residue in hidden Markov model sequence searches, ModelAngelo outperforms human experts in the identification of proteins with unknown sequences. ModelAngelo will thus remove bottlenecks

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and increase objectivity in cryo-EM structure determination.

Introduction

Knowledge of the three-dimensional atomic structures of proteins and nucleic acids is pivotal for our understanding of the molecular processes of life. In recent years, considerable advances have been made in the determination of structures of biological macromolecules by electron cryo-microscopy (cryo-EM), culminating in cryo-EM maps of proteins with sufficient resolution to resolve individual atoms (*1, 2*). Accordingly, the number of new cryo-EM structures in the electron microscopy database (EMDB) (*3*) is growing exponentially. If this trend continues, approximately 100,000 cryo-EM structures will be determined in the next 5 years (*4*).

Over two-thirds of last year's (2022) structures had resolutions better than 4 Å. Although individual atoms are not resolved at resolutions between 2-4 Å, reliable atomic models can be built by exploiting prior knowledge of the chemical structures of the proteins and nucleic acids in the sample, including their amino acid and nucleic acid sequences. Typically, atomic model building in cryo-EM maps is performed using manual procedures in three-dimensional computer graphics programs (*5, 6*). Atomic model building is often time-consuming and requires substantial levels of expertise to produce accurate models. At resolutions better than 3 Å, experts can build atomic models with few errors, whereas at resolutions below 4 Å, avoiding mistakes is challenging. It is therefore not uncommon for atomic models of biological complexes to contain errors (*7*), with potentially grievous consequences (*8*).

Structure determination by cryo-EM is also an increasingly important tool for the discovery of new subunits in biological complexes. Because of its relaxed requirements for sample quantity and purity compared to other structural biology techniques, cryo-EM is capable of determining structures of complexes purified from endogenous sources. Many such complexes contain subunits of unknown identities. Without prior knowledge of the amino acid sequence,

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

identifying the chemical identity of individual amino acids in cryo-EM maps is difficult, and requires relatively high resolutions. Yet, provided one can build stretches of several consecutive amino acids, database searches with the sequence fragments can lead to the identification of the corresponding protein. Recent examples include the identification of TMEM106B in amyloid filaments from human brains (*9–11*) and the detection of subunits of axonemal complexes (*12, 13*).

Here, we introduce a machine-learning approach, called ModelAngelo, for the automated building of atomic models and the identification of proteins in cryo-EM maps. Machine learning approaches often require large amounts of training data. For example, recent protein language models were trained on tens of millions of sequences (*14*) and AlphaFold2 was trained on more than 200,000 structures (*15*). In contrast, fewer than 13,000 cryo-EM structures with resolutions better than 4 Å have been determined to date and many of these are redundant. The limited amount of available training data prompted us to design a multi-modal machine-learning approach that combines local information from the cryo-EM map surrounding each protein or nucleic acid residue with additional information from the protein sequences in the sample and the local geometry of the structure. Similar sources of information are exploited by human experts when manually building atomic models in cryo-EM maps.

The sudden availability of atomic models for millions of proteins from protein structure prediction by AlphaFold2 (*15, 16*) has helped to guide and accelerate model building (*17*). However, previous attempts to fully automate atomic modelling (*18–24*) or the identification of unknown proteins (*25–27*) have so far failed to become mainstream, although DeepTracer (*21, 24*) and findMySequence (*25*) have gained some traction. Still, atomic modelling remains a time-consuming and expert-dependent process in many structure determination projects. With the ongoing exponential growth in cryo-EM structures and the continuing influx of newcomers to the cryo-EM field, automation will be key in removing bottlenecks and replacing the depen-

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

dence on human experts with objective methods that are accessible to all. In what follows, we demonstrate that ModelAngelo can meet this need. ModelAngelo outperforms human experts in identifying unknown proteins and produces near-complete atomic models of comparable quality to those obtained by human experts.

Approach

Automated model building of proteins and nucleic acids in ModelAngelo comprises three steps (Figure 1a). Details about the network architectures that underlie these steps and how they are trained have been described in (*28*).

Graph initialisation Positions for the backbone Cα atom of amino acids and the phosphor atom of nucleic acids are predicted using a convolutional neural network (CNN). This CNN is a modified feature-pyramid network (*30*) that predicts whether each voxel in the cryo-EM map contains the Cα atom of an amino acid, the phosphor atom of a nucleic acid residue, or neither. A graph is then constructed, where each residue is a node, and edges are formed between each residue and its twenty nearest neighbours.

Graph optimisation A graph neural network (GNN) is used to optimise the positions and orientations of the residues, to predict their amino or nucleic acid identity, and to predict torsion angles for their side chains or bases. The GNN consists of three modules: a cryo-EM module, a sequence module, and an invariant point attention (IPA) module (Figure 1b). Each node of the graph is associated with a residue feature vector. Each module takes the residue feature vector as input, combines it with new information, and outputs an updated residue feature vector that is passed to the next module. The sequential application of the three modules in eight layers (Figure 1b) allows the gradual extraction of more information from the different inputs.

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

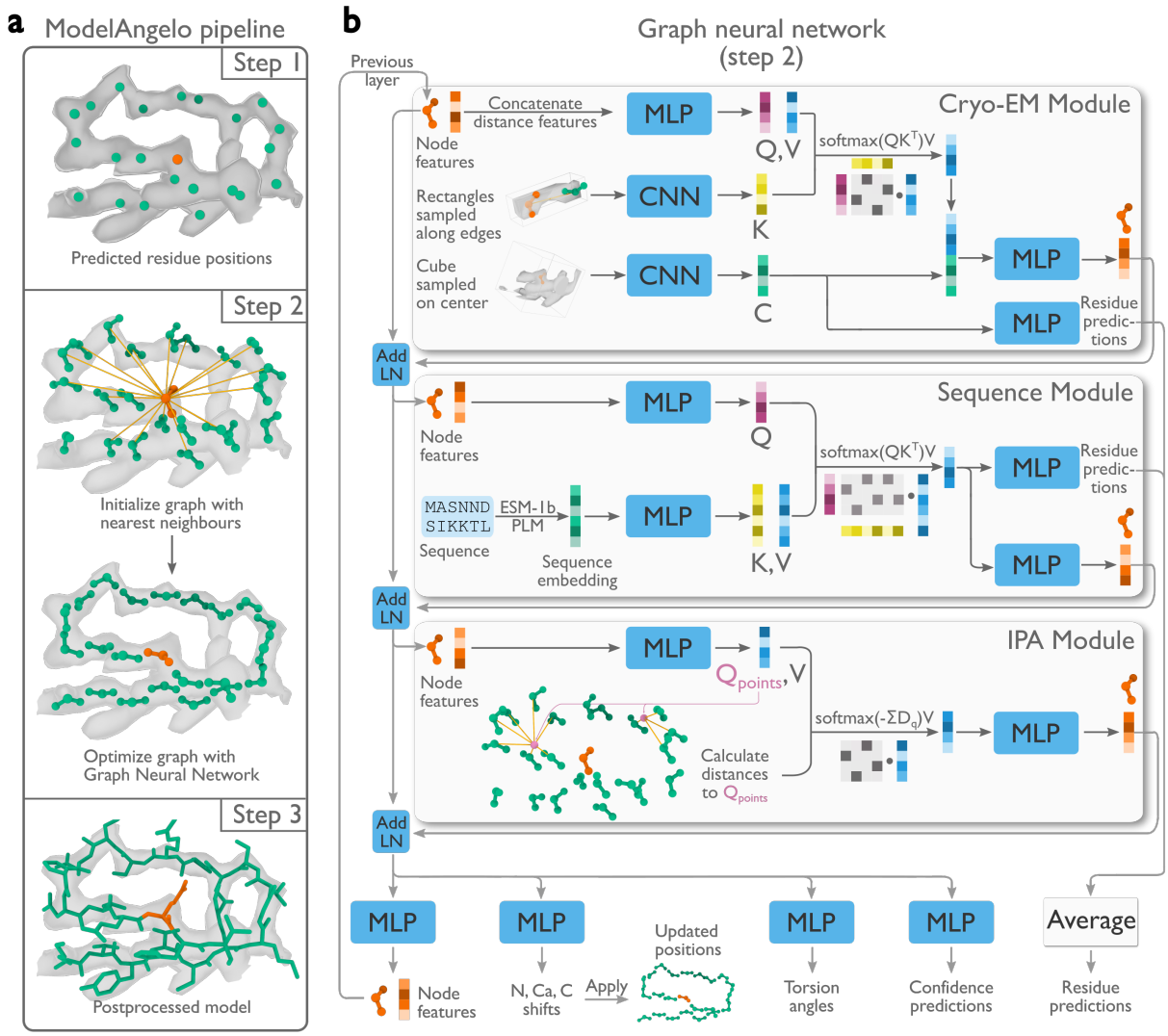


Figure 1: Atomic modelling in ModelAngelo. **a**, Three steps build atomic models: 1) A convolutional neural network (CNN) predicts protein and nucleic acid residue positions; 2) A graph neural network (GNN) optimises these positions and orientations (panel b); 3) A post-processed, optimised graph forms a complete atomic model. **b**, The GNN, arranged in eight layers with three modules, uses a feature vector per residue passed through multi-layer perceptrons (MLP), integrated with additional data via attention mechanisms that have query (Q), key (K), and value (V) vectors. Stable gradient propagation is ensured by residual connections with layer norms (Add LN) (*29*). Residue feature vectors update residue positions, orientations, predict torsion angles, confidence scores, and identities at the end of each layer.

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The cryo-EM module incorporates information from the cryo-EM map and comprises two parts. Firstly, the input feature vector is passed through a multi-layer perceptron (MLP) network to generate query and value vectors. These vectors are used for cross-attention (*31*) with key vectors that are calculated from a CNN on rectangular boxes that are extracted from the cryo-EM density map that point from the current residue to its twenty nearest neighbours. Intuitively, the cross-attention mechanism allows mixing information from each residue with that of its twenty nearest neighbours, depending on whether the cryo-EM density between them looks connected. Secondly, a cubic box is extracted from the cryo-EM map around the position of the current residue and passed through another CNN. The resulting vector is used in two ways: to generate amino and nucleic acid identity predictions through an MLP; and after concatenation with the vector from the cross-attention, it is passed through another MLP to generate the output residue feature vector of the cryo-EM module.

The sequence module performs cross-attention for each residue with the user-provided amino acid sequences, which are embedded using the pre-trained protein language model, ESM-1b (*32*). This incorporates information that is learned by the language model from many amino acid sequences, including multiple homologues. The information in protein language models has been shown to be sufficient for protein structure prediction (*14*). The vector from the cross-attention is used in two ways: a first MLP is used to generate amino and nucleic acid identity predictions; a second MLP generates the output residue feature vector of the sequence module. For nucleic acid residues, the sequence module is not used.

The IPA module incorporates information from the geometry of the nodes in the graph and was inspired by the module with the same name in AlphaFold2 (*15*). An MLP calculates four query points per residue and the Euclidean distance between the query points and the location of the neighbouring nodes is used to replace the cosine similarity of the attention algorithm between the query and key vectors. Intuitively, this allows the model to learn information about

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the topology of neighbouring residues, for example about secondary structure. In fact, disabling this module in an ablation study led to atomic models with incorrect secondary structure geometry (*28*).

Post-processing Finally, the residue feature vectors are post-processed to generate an atomic model. The residue feature vectors are used as inputs into two separate MLPs to predict new positions and orientations for each residue, as well as torsion angles for amino acid side chains and nucleic acid bases. In addition, the predictions for the amino or nucleic acid identities from the cryo-EM and sequence modules are averaged to generate probabilities for each possible identity for all residues. These vectors are converted into a hidden Markov model (HMM) profile (see below) that is used for a search against the input sequences using HMMER (*33*). Matched residues, as defined in (*34*), are mutated to the corresponding amino or nucleic acid in the input sequences, and separate chains are connected based on their assigned sequences and proximity. Finally, chains shorter than four residues are pruned from the model, and a full atomic model is generated from the predicted positions and orientations of each residue and their corresponding amino acid or nucleic base torsion angle predictions, using idealised geometries.

Generating HMM profiles A profile HMM is a probabilistic model representing the multiple sequence alignment (MSA) of a set of related sequences. The parameters of a profile HMM are normally estimated from the MSA it strives to model, however, here, they are instead estimated from ModelAngelo predictions. There are three types of states in the profile HMM. For each position of the MSA's consensus sequence, there is a match (M), a delete (D), and an insert (I) state with respect to the query sequences. There are two types of probabilities in a profile HMM: transition and emission. The transition probabilities reflect the probability of a sequence going between the M, I, and D states of one position of the profile to the next. ModelAngelo

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

uses the confidence metric, $c^{(i)}$, that it predicts for each residue i to construct the transition probabilities as follows:

$$P^{(i)}_{M\to M} = \max(c^{(i)} - d, 0.5) \qquad P^{(i)}_{D\to M} = 1 - d \qquad P^{(i)}_{I\to M} = 1 - d$$
$$p^{(i)}_{M\to D} = \frac{1 - p^{(i)}_{M\to M}}{2} \qquad P^{(i)}_{D\to D} = d \qquad P^{(i)}_{I\to D} = 0$$
$$p^{(i)}_{M\to I} = \frac{1 - p^{(i)}_{M\to M}}{2} \qquad P^{(i)}_{D\to I} = 0 \qquad P^{(i)}_{I\to I} = d$$

The strategy to set $P^{(i)}_{M\to I} = P^{(i)}_{M\to D}$; the constant $d = 0.5$; and the minimum value of $P^{(i)}_{M\to M} = 0.5$ were chosen arbitrarily and are never optimised. The emission probabilities represent the probability of each amino acid being produced in a M or I state. For these, ModelAngelo uses its predicted probability distribution of the amino acids for each residue. The resulting HMM profiles are compatible with HMMER3 (*35*) and HHblits (*36*).

Recycling Inspired by AlphaFold (*15*), we recycle the post-processed model from one round of the GNN as the starting point of a subsequent round of graph optimisation. For this purpose, ModelAngelo was trained with a random number of 1-3 recycling steps. During inference, we perform three rounds of recycling, as the performance plateaus after three rounds.

Training ModelAngelo was trained on maps deposited in the EMDB (*3*) before April 1st, 2022 with resolutions better than 4 Å and paired with models in the PDB (*37*) that cover the entire map correctly, as detailed in (*28*). PDB files that included insertion codes, i.e. additional residues relative to the reference sequence, were removed. This resulted in 3715 map-model pairs that were used during training. All cryo-EM maps were resampled to a common pixel size of 1 Å. As a point of comparison, findMySequence only uses 117 pairs while DeepTracer uses approximately 1400 (*21, 25*).

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

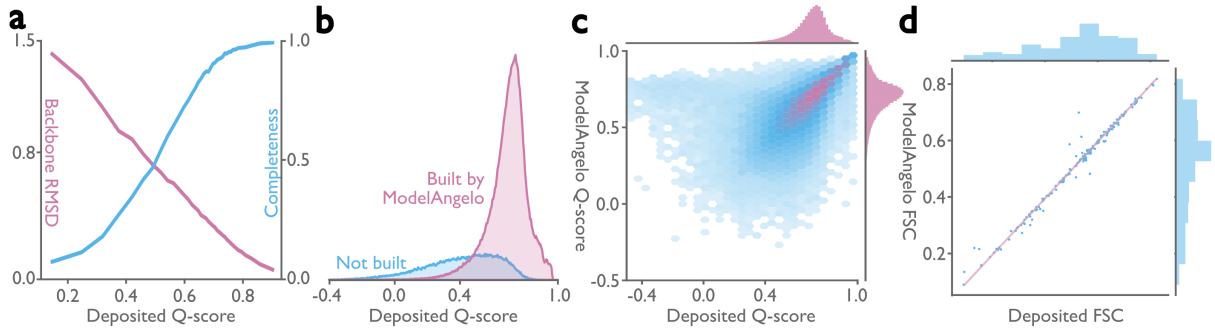


Figure 2: Performance of ModelAngelo for atomic modelling of proteins. **a**, Backbone root-mean-square deviation (RMSD) and model completeness plotted as a function of the target model Q-scores. **b**, Q-score distribution of residues in the deposited models, comparing those built by ModelAngelo with those not built. **c**, Q-score comparison between ModelAngelo predicted models and the deposited models. **d**, Model-to-map Fourier shell correlation (FSC), as calculated by Servalcat (*38*), after refining both models and using only residues present in both ModelAngelo and deposited models.

Protein identification To allow model building for structures with unknown sequences, we also trained a version of ModelAngelo without its sequence module. Still, for each protein residue, ModelAngelo predicts probabilities for all twenty amino acids. Within ModelAngelo, these probabilities are converted into HMM profiles and used for searches in HMMER3 (*35*) as described above, but using a larger proteome, rather than only the sequences known to be present in the structure.

Results

ModelAngelo builds protein models of comparable quality to those built by humans To test ModelAngelo, we considered all cryo-EM structures determined to at least 4 Å resolution and released from the EMDB between the cutoff date for training, April 1st, 2022, and February 9th, 2023. To reduce computational costs, we excluded structures with more than 30,000 protein residues. We also removed viruses with icosahedral symmetry, for which typically only the asymmetric unit was built. To ensure none of the sequences were seen before during training,

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

we removed structures that had protein chains with more than 10% sequence identity to any of the proteins in the training set. Finally, we removed structures with insertion codes and other irregularities. This resulted in a test set of 177 structures, on which we ran ModelAngelo. Using a single A100 GPU, the smallest structure (PDB ID: 8DWI, with a molecular weight of 54.7 kDa) took 2 minutes; the largest structure (PDB ID: 7UMS, with a molecular weight of 1.85 MDa) took 53 minutes. The output coordinates from ModelAngelo were refined against the cryo-EM map using Servalcat (*38*), and the refined models were compared to the deposited ones.

To assess the quality of the models generated by ModelAngelo, we analyzed the Q-scores (*39*) of all structures in the test set. The Q-score measures the resolvability of individual atoms in cryo-EM maps, thus reflecting the quality of the built model. Provided the model is built well, Q-scores also correlate with the local resolution, which can vary in cryo-EM maps: Q-scores of 0.4 are typical for cryo-EM maps at 4 Å resolution, values of 0.6 for maps at 3 Å, and values better than 0.7 for maps beyond 2 Å resolution (*39*). We implemented Q-score calculation in ModelAngelo and calculated average Q-scores for all atoms in each residue of both the deposited models and those built by ModelAngelo. Next, we calculated backbone root mean squared deviations (RMSDs) between the protein models built by ModelAngelo and those deposited, and plotted these against the Q-scores of the deposited residues (Figure 2a; pink line). As expected, ModelAngelo builds models with lower RMSDs for residues with higher (better) Q-scores. Even for residues with Q-scores as low as 0.4, ModelAngelo builds models with backbone RMSDs lower than 1.0 Å. We also measured the completeness of the models built by ModelAngelo. We define completeness as the fraction of residues that are built with their Cα atom within 3 Å of the deposited model and with the correct amino acid assignment. As with backbone RMSD, completeness improves for residues with higher Q-scores (Figure 2a; blue line). Overall, ModelAngelo built 77% of all 410,585 residues in the test set. Analysis of the

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

deposited Q-scores shows that those residues not built by ModelAngelo have lower Q-scores than those that are built (Figure 2b). In the deposited models, many of the residues with the lowest Q-scores were probably obtained by rigid-body docking of protein domains into poorly resolved regions of the cryo-EM maps. Excluding the 51,446 residues with Q-scores below 0.4, ModelAngelo built 85% of the residues in the test set. A comparison of Q-scores calculated for the models built by ModelAngelo with those calculated for the deposited models shows that models from ModelAngelo are as good as the deposited ones (Figure 2c). The same is also true for overall Fourier shell correlation values between the cryo-EM maps and those parts of the models that were both built by ModelAngelo and present in the deposited models (Figure 2d).

ModelAngelo builds good nucleic acid backbones The test set described above contained only 103 nucleic acid chains, many with just a few nucleotides. Therefore, instead of conducting a systematic analysis as done for the proteins, we present a few test cases to illustrate the quality of nucleotide building (Figure 3). We applied ModelAngelo to 11 different ribosome structures that were determined to resolutions ranging from 1.98 to 3.80 Å (Figures 3a,b), as well as a CRISPR-associated transpososome from *Scytonema hofmanni* (Figures 3c,d) (*41*). Although ribosome structures were included in ModelAngelo's training set, the nucleotide sequences were not. When plotting backbone RMSDs and backbone completeness against the Q-scores of the deposited nucleotide coordinates (Figure 3e), we observed similar trends as for the protein chains. Backbone RMSDs range from 2 Å in the worst regions of the map to values better than 0.5 Å in the best regions. Likewise, near-complete backbones are built in the best regions, while backbone completeness drops to below 80 % for the worst regions. However, ModelAngelo struggles to distinguish between the two purines or the two pyrimidines, echoing the difficulty humans face in building nucleotide sequences based solely on the cryo-EM density, if the resolution does not extend beyond 2.5 Å. Consequently, when considering only

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

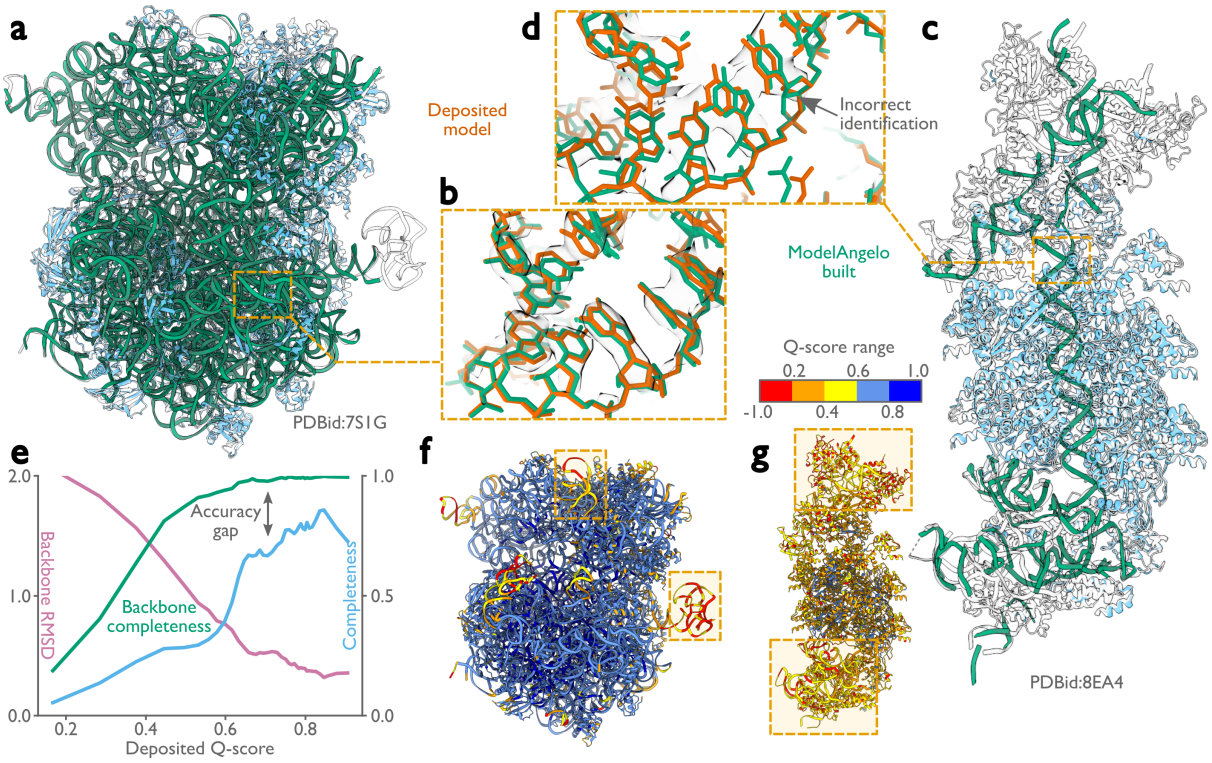


Figure 3: Performance of ModelAngelo for atomic modelling of nucleic acids. **a**, *Escherichia coli* ribosome built by ModelAngelo (with ribosomal RNA in green and proteins in blue) compared with the deposited model (PDB ID: 7S1G, black outline) (*40*). **b**, Zoomed-in view with nucleotide bases showing high accuracy compared to the deposited model (orange). **c**, ModelAngelo model of the V-K CAST transpososome from *Scytonema hofmanni* compared with the deposited model (PDB ID: 8EA4) (*41*). Sections not built by ModelAngelo (black outline) are in regions of low Q-score (see panel g). **d**, Zoomed-in view comparing the nucleotide bases of both models showing a sequence incorrectly identified by ModelAngelo. **e**, Backbone RMSD, backbone completeness, and sequence completeness plotted against the deposited Q-score for six ribosome structures. **f, g,** Deposited models for the structures in a and c, coloured by Q-score, with low Q-score regions boxed.

correctly built sequences, the completeness of the models built by ModelAngelo drops to 80 % for the best parts of the map, and to as low as 20 % for the worst parts (Fig 3E). Users should therefore carefully validate the nucleotide chains of models built by ModelAngelo, for example by using nucleotide secondary structure predictors (*42–46*). Nonetheless, ModelAngelo considerably accelerates the process of building the nucleotide backbone. Subsequent nucleotide

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

base changes can be made with minimal manual intervention.

ModelAngelo identifies protein chains that were not built by human experts To illustrate the performance of ModelAngelo in identifying protein chains in cryo-EM maps, we applied ModelAngelo to two examples of large cryo-EM structures that were recently determined from endogenous sources. The first example is a structure of the supercomplex of the phycobilisome (PBS), photosystem I and II (PSI and PSII), and the transmembrane light-harvesting complexes (LHCs) that was imaged *in situ* in the red alga *Porphyridium purpureum* (*47*). The second example is a structure of the ciliary central apparatus and radial spokes of the green alga *Chlamydomonas reinhardtii* that was obtained by single-particle analysis following purification from cilia (*12, 13*).

At 16.7 MDa, the PBS-PSII-PSI-LHC supercomplex is one of the largest complexes determined by single-particle cryo-EM. The deposited model (PDB ID: 7Y5E) consists of 158,730 residues in 81 unique protein chains, including six chains for which the authors were unable to identify the corresponding protein. The unidentified chains were termed $L_{PP}1$ (linker of PBS–PSII 1); CNT (for "connector"); PsbW and Psb34 (two of the core subunits of PSII); LRH (a linker protein); and $L_{PS}1$ (photosystem linker protein 1). To identify these chains, we ran ModelAngelo without using its sequence module (using the `build_no_seq` option) to calculate an initial atomic model with HMM profiles for all chains, and we searched these profiles against the proteome constructed in (*48*) (using the `hmm_search` option). Because of local pseudo-symmetry, all six unidentified proteins occur more than once in the cryo-EM map. This allows us to bootstrap weaker individual hits by cross-referencing their matches to the other instances. Specifically, the same six protein chains were identified for all instances, with E-values in the range of 5.8e-66 to 6.4e-2. For comparison, using the backbone traces in the deposited model as inputs to findMySequence (*25*) identified only two of the unassigned

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

proteins (Psb34 and PsbW). We then constructed an input sequence file that included all chains in the deposited model plus the six newly identified chains and ran ModelAngelo again. This calculation took 23 hours on an A100 GPU. The resulting model, containing 110,742 residues, is shown in Figure 4a. For most sections of the unidentified chains, ModelAngelo built better models than those in the deposited structure, most notably for LRH and CNT. ModelAngelo did not build models for parts of the unidentified proteins that were in regions of poor cryo-EM density. Besides excellent agreement between side chain densities in the cryo-EM map and the predicted sequences (Extended Data Figure 1), the structures built by ModelAngelo were also highly similar to AlphaFold2 predictions for the unidentified chains (*15, 49*) (Extended Data Figure 2).

Like the PBS-PSII-PSI-LHC supercomplex, the central apparatus (CA) and radial spoke complexes isolated from *C. reinhardtii* ciliary axonemes are large complexes with poorly characterized subunit compositions. Although recent cryo-EM structures had identified 23 different radial spoke proteins (RSPs) and 48 different CA proteins (*12, 13*), the deposited maps (EMD-22475, EMD-24481, and EMD-25381) contained densities that were left unassigned despite considerable manual effort. To identify these proteins, we applied ModelAngelo without using its sequence module to the deposited maps and searched the resulting HMM profiles against the latest version of the *C. reinhardtii* predicted proteome (*50*) (Figure 4b and Supplementary Information). This approach identified four additional radial spoke proteins: FAP109, Cre05.g240450, Cre08.g800895, Cre17.g802036), which we rename to RSP24, RSP25, RSP26 and RSP27, respectively, and two additional CA proteins (FAP92 and FAP374) (Extended Data Table 1). Neither RSP24 (Cre08.g800895) nor RSP26 (Cre17.g802036) were annotated in earlier versions of the *C. reinhardtii* genome, explaining their absence from proteomic studies, and demonstrating the importance of high-quality genome annotations for *de novo* identification of proteins by cryo-EM. RSP27 (Cre05.g240450) was identified from a fragment of just 33

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

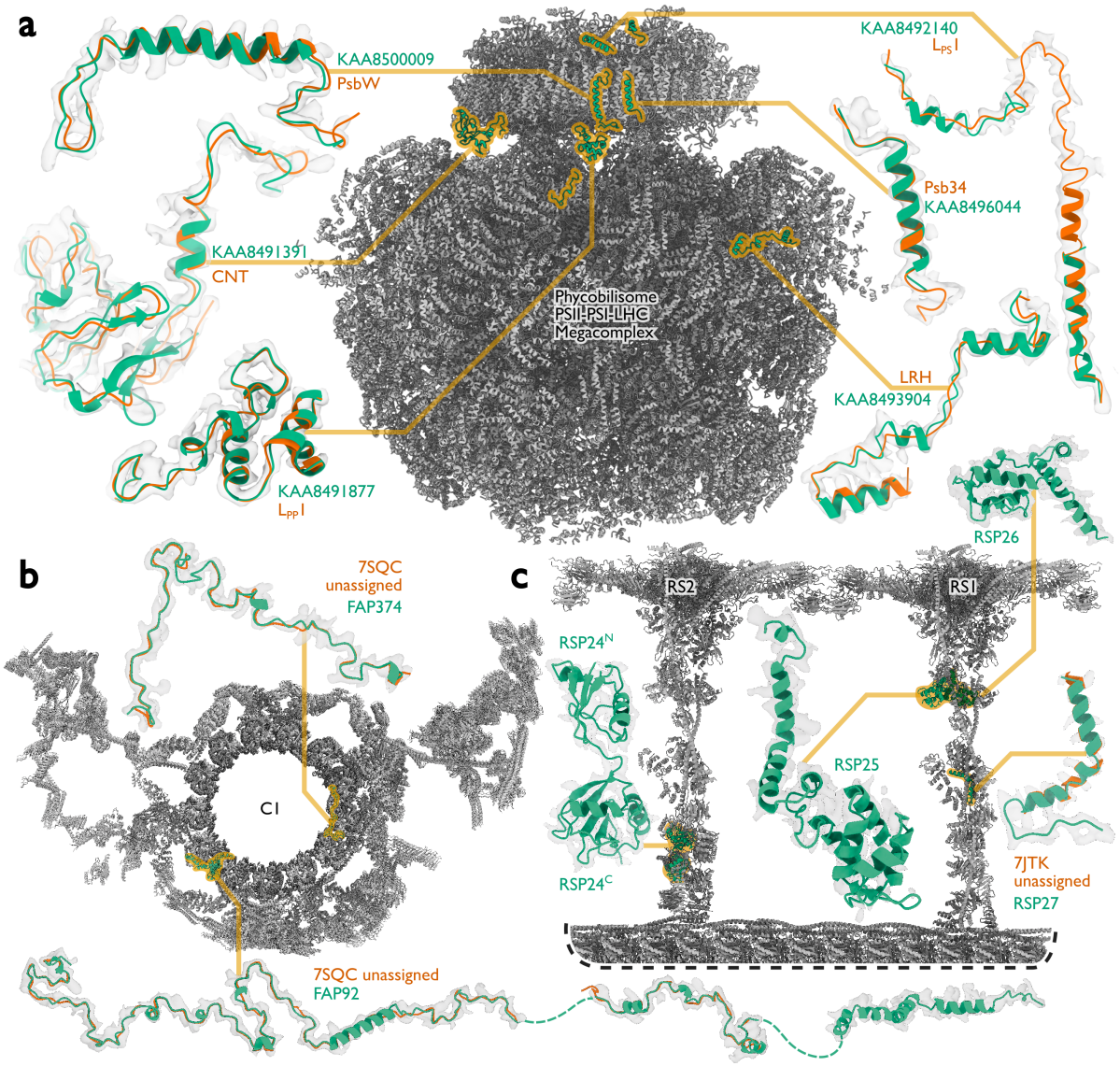


Figure 4: Examples of protein identification by ModelAngelo. a, The ModelAngelo model of the single-PBS-PSII-PSI-LHC's supercomplex (gray) showing the positions, models, and map densities of six newly identified proteins (green). Backbone traces in the deposited model (PDB ID: 7Y5E) are shown in orange. **b**, Atomic model of the central apparatus microtubule C1 showing the positions, models, and map densities of two newly identified proteins: FAP92 and FAP374. Orange cartoons represent poly(UNK) chains deposited in the original model (PDB ID: 7SQC). **c**, An atomic model of radial spokes 1 and 2 (RS1 and RS2) bound to a doublet microtubule (gray) showing the positions, models, and map densities of four proteins (RSP24-27, green) newly identified by ModelAngelo. Only RSP27 had a backbone trace in the deposited model (orange).

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

residues, demonstrating the power of ModelAngelo to identify proteins from small sections of well-resolved density. Both CA proteins (FAP92 and FAP374) bind directly to the microtubule surface and have tertiary structures poorly predicted by AlphaFold2 (Extended Data Figure 3); sidechain density was, therefore, essential for their successful identification (Extended Data Figure 4). The identification of these proteins will allow their functional relevance to the regulation of ciliary motility to be investigated through targeted genetic manipulation.

Discussion

ModelAngelo automates atomic modelling in cryo-EM maps, building protein models of comparable quality to those built by human experts and nucleic acid models with near-complete and accurate backbones. ModelAngelo outperforms existing approaches like DeepTracer for the automated modelling of both proteins (*28*) and nucleotides (Extended Data Figure 5). Furthermore, ModelAngelo builds these models within hours on a modern GPU, thereby removing an important bottleneck in cryo-EM structure determination. Future incorporation of ModelAngelo into automated cryo-EM image processing pipelines (*51–57*), will enable users to go from data acquisition to atomic models in a single automated procedure.

By introducing objectivity in the model-building process, ModelAngelo also informs which parts of the map can be confidently interpreted with an atomic model and which should be left uninterpreted. In this way, ModelAngelo will not only reduce the number of errors in atomic models but also play a role in making cryo-EM structure determination more accessible to the large numbers of newcomers that the field has experienced in recent years. Still, some degree of human supervision and intervention will remain necessary, particularly for (parts of) cryo-EM maps with resolutions worse than 3.5-4.0 Å. For example, rigid body fitting of a known domain in lower-resolution map regions to obtain a more complete model falls outside the scope of ModelAngelo.

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Besides accelerating cryo-EM structure determination and providing objectivity in atomic modelling, ModelAngelo also identifies protein chains in cryo-EM maps better than human experts. The reason why ModelAngelo outperforms the human expert in this task likely lies in the implementation of its sequence searches. While human experts typically base their identifications on discrete assignments of individual amino acids to various residues in unknown chains, ModelAngelo exploits predicted probabilities for all twenty amino acids for every protein residue and combines this information with its predicted confidence in each residue in a full HMM search. This not only allows better identification of unknown chains but also helps ModelAngelo during the building of atomic models with known sequences, where it may potentially outperform human experts in placing protein chains for which ambiguity exists, for example when multiple homologous chains coexist in a single structure. The ability to identify proteins in cryo-EM maps will increase in importance as ongoing advances in sample preparation, microscopy, and image processing allow ever more structures to be determined for samples purified from native sources or visualised in situ by electron tomography of frozen cells or thin tissue sections.

Acknowledgements

We thank George Ghanim, Joe Greener, Katerina Naydenova, Johannes Schwab, Zala Sekne, Sofia Lövestam, and Keitaro Yamashita for helpful discussions; Miao Gui for contributions to atomic modelling of the ciliary axonemes; and Jake Grimmett, Toby Darling, and Ivan Clayson for help with high-performance computing. This work was supported by the Medical Research Council as part of the United Kingdom Research and Innovation (MC_UP_A025_1013 to SHWS); the European Union's Horizon 2020 research and innovation programme (under grant agreement no. 895412 to DK); the National Institutes of Health (R01-GM141109 to AB and R01-GM138854 to RZ); and the Knut and Alice Wallenberg Foundation (2022.0032 to LK).

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

For the purpose of open access, the MRC Laboratory of Molecular Biology has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising

Author contributions

KJ designed and implemented ModelAngelo. LK designed and implemented the HMM search algorithm. RZ and AB analyzed ciliary axoneme data. DK and SHWS jointly supervised the project. All authors contributed to the writing of the manuscript.

Competing interests

The authors declare no competing interests.

Data availability

ModelAngelo is freely available under the open-source MIT license and can be downloaded from `https://github.com/3dem/model-angelo`.

References

1. Nakane, T. *et al.* Single-particle cryo-em at atomic resolution. *Nature* **587**, 152–156 (2020).

2. Yip, K. M., Fischer, N., Paknia, E., Chari, A. & Stark, H. Atomic-resolution protein structure determination by cryo-em. *Nature* **587**, 157–161 (2020).

3. Lawson, C. L. *et al.* Emdatabank unified data resource for 3dem. *Nucleic acids research* **44**, D396–D403 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

4. Russo, C. J., Dickerson, J. L. & Naydenova, K. Cryomicroscopy in situ: what is the smallest molecule that can be directly identified without labels in a cell? *Faraday Discussions* **240**, 277–302 (2022).

5. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of coot. *Acta Crystallographica Section D - Biological Crystallography* **66**, 486–501 (2010).

6. Croll, T. I. Isolde: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallographica Section D: Structural Biology* **74**, 519–530 (2018).

7. Gao, Y., Thorn, V. & Thorn, A. Errors in structural biology are not the exception. *Acta Crystallographica Section D: Structural Biology* **79** (2023).

8. Croll, T. I. *et al.* Making the invisible enemy visible. *Nature structural & molecular biology* **28**, 404–408 (2021).

9. Schweighauser, M. *et al.* Age-dependent formation of tmem106b amyloid filaments in human brains. *Nature* **605**, 310–314 (2022).

10. Jiang, Y. X. *et al.* Amyloid fibrils in ftld-tdp are composed of tmem106b and not tdp-43. *Nature* **605**, 304–309 (2022).

11. Chang, A. *et al.* Homotypic fibrillization of tmem106b across diverse neurodegenerative diseases. *Cell* **185**, 1346–1355 (2022).

12. Gui, M. *et al.* Structures of radial spokes and associated complexes important for ciliary motility. *Nature structural & molecular biology* **28**, 29–37 (2021).

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

13. Gui, M., Wang, X., Dutcher, S. K., Brown, A. & Zhang, R. Ciliary central apparatus structure reveals mechanisms of microtubule patterning. *Nature Structural & Molecular Biology* **29**, 483–492 (2022).

14. Lin, Z. *et al.* Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv* (2022). URL `https://www.biorxiv.org/content/early/2022/07/21/2022.07.20.500902`. `https://www.biorxiv.org/content/early/2022/07/21/2022.07.20.500902.full.pdf`.

15. Jumper, J. *et al.* Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).

16. Varadi, M. *et al.* Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research* **50**, D439–D444 (2022).

17. Oeffner, R. D. *et al.* Putting alphafold models to work with phenix. process_predicted_model and isolde. *Acta Crystallographica Section D: Structural Biology* **78** (2022).

18. Terashi, G. & Kihara, D. De novo main-chain modeling for em maps using mainmast. *Nature communications* **9**, 1618 (2018).

19. Liebschner, D. *et al.* Macromolecular structure determination using x-rays, neutrons and electrons: recent developments in phenix. *Acta Crystallographica Section D: Structural Biology* **75**, 861–877 (2019).

20. Hoh, S. W., Burnley, T. & Cowtan, K. Current approaches for automated model building into cryo-em maps using buccaneer with ccp-em. *Acta Crystallographica Section D: Structural Biology* **76**, 531–541 (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

21. Pfab, J., Phan, N. M. & Si, D. Deeptracer for fast de novo cryo-em protein structure modeling and special studies on cov-related complexes. *Proceedings of the National Academy of Sciences of the United States of America* **118** (2021).

22. Zhang, X., Zhang, B., Freddolino, P. L. & Zhang, Y. Cr-i-tasser: assemble protein structures from cryo-em density maps using deep convolutional neural networks. *Nature methods* **19**, 195–204 (2022).

23. He, J., Lin, P., Chen, J., Cao, H. & Huang, S.-Y. Model building of protein complexes from intermediate-resolution cryo-em maps with deep learning-guided automatic assembly. *Nature Communications* **13**, 4066 (2022).

24. Nakamura, A. *et al.* Fast and automated protein-dna/rna macromolecular complex modeling from cryo-em maps. *Briefings in Bioinformatics* **24**, bbac632 (2023).

25. Chojnowski, G. *et al.* findmysequence: a neural-network-based approach for identification of unknown proteins in x-ray crystallography and cryo-em. *IUCrJ* **9** (2022).

26. Chang, L. *et al.* Deeptracer-id: De novo protein identification from cryo-em maps. *Biophysical Journal* **121**, 2840–2848 (2022).

27. Terwilliger, T. C. *et al.* Protein identification from electron cryomicroscopy maps by automated model building and side-chain matching. *Acta Crystallographica Section D: Structural Biology* **77**, 457–462 (2021).

28. Jamali, K., Kimanius, D. & Scheres, S. H. A graph neural network approach to automated model building in cryo-EM maps. In *The Eleventh International Conference on Learning Representations* (2023). URL `https://openreview.net/forum?id=65XDF_nwI61`.

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

29. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).

30. Lin, T.-Y. *et al.* Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125 (2017).

31. Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).

32. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118**, e2016239118 (2021).

33. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: Hmmer3 and convergent evolution of coiled-coil regions. *Nucleic acids research* **41**, e121–e121 (2013).

34. Krogh, A., Brown, M., Mian, I. S., Sjölander, K. & Haussler, D. Hidden markov models in computational biology: Applications to protein modeling. *Journal of molecular biology* **235**, 1501–1531 (1994).

35. Eddy, S. R. Accelerated profile hmm searches. *PLoS computational biology* **7**, e1002195 (2011).

36. Remmert, M., Biegert, A., Hauser, A. & Söding, J. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature methods* **9**, 173–175 (2012).

37. Burley, S. K. *et al.* Rcsb protein data bank: powerful new tools for exploring 3d structures of biological macromolecules for basic and applied research and education in fundamental

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic acids research* **49**, D437–D451 (2021).

38. Yamashita, K., Palmer, C. M., Burnley, T. & Murshudov, G. N. Cryo-em single-particle structure refinement and map calculation using servalcat. *Acta Crystallographica Section D: Structural Biology* **77** (2021).

39. Pintilie, G. *et al.* Measurement of atom resolvability in cryo-em maps with q-scores. *Nature methods* **17**, 328–334 (2020).

40. Tsai, K. *et al.* Structural basis for context-specific inhibition of translation by oxazolidinone antibiotics. *Nature structural & molecular biology* **29**, 162–171 (2022).

41. Park, J.-U. *et al.* Structures of the holo crispr rna-guided transposon integration complex. *Nature* **613**, 775–782 (2023).

42. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic acids research* **31**, 3406–3415 (2003).

43. Markham, N. R. & Zuker, M. Unafold: software for nucleic acid folding and hybridization. *Bioinformatics: structure, function and applications* 3–31 (2008).

44. Hofacker, I. L. Vienna rna secondary structure server. *Nucleic acids research* **31**, 3429–3431 (2003).

45. Lorenz, R. *et al.* Viennarna package 2.0. *Algorithms for molecular biology* **6**, 1–14 (2011).

46. Reuter, J. S. & Mathews, D. H. Rnastructure: software for rna secondary structure prediction and analysis. *BMC bioinformatics* **11**, 1–9 (2010).

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

47. You, X. *et al.* In situ structure of the red algal phycobilisome–psii–psi–lhc megacomplex. *Nature* 1–8 (2023).

48. Lee, J., Kim, D., Bhattacharya, D. & Yoon, H. S. Expansion of phycobilisome linker gene families in mesophilic red algae. *Nature communications* **10**, 4823 (2019).

49. Mirdita, M. *et al.* Colabfold: making protein folding accessible to all. *Nature methods* **19**, 679–682 (2022).

50. Craig, R. J. *et al.* The chlamydomonas genome project, version 6: Reference assemblies for mating-type plus and minus strains reveal extensive structural mutation in the laboratory. *The Plant Cell* **35**, 644–672 (2023).

51. Kimanius, D., Dong, L., Sharov, G., Nakane, T. & Scheres, S. H. New tools for automated cryo-em single-particle analysis in relion-4.0. *Biochemical Journal* **478**, 4169–4185 (2021).

52. Biyani, N. *et al.* Focus: The interface between data collection and data processing in cryo-em. *Journal of structural biology* **198**, 124–133 (2017).

53. Gómez-Blanco, J. *et al.* Using scipion for stream image processing at cryo-em facilities. *Journal of structural biology* **204**, 457–463 (2018).

54. Tegunov, D. & Cramer, P. Real-time cryo-electron microscopy data preprocessing with warp. *Nature methods* **16**, 1146–1152 (2019).

55. Stabrin, M. *et al.* Transphire: automated and feedback-optimized on-the-fly processing for cryo-em. nat commun 11, 5716 (2020).

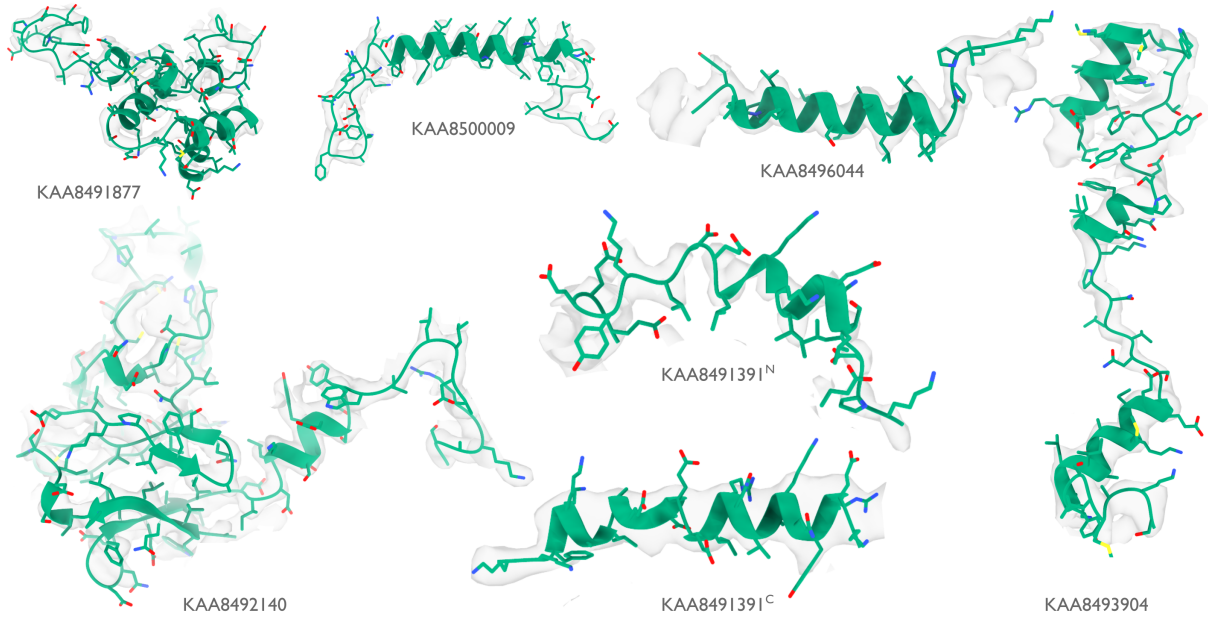
56. Xie, R., Chen, Y.-X., Cai, J.-M., Yang, Y. & Shen, H.-B. Spread: a fully automated toolkit for single-particle cryogenic electron microscopy data 3d reconstruction with image-

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

network-aided orientation assignment. *Journal of chemical information and modeling* **60**, 2614–2625 (2020).

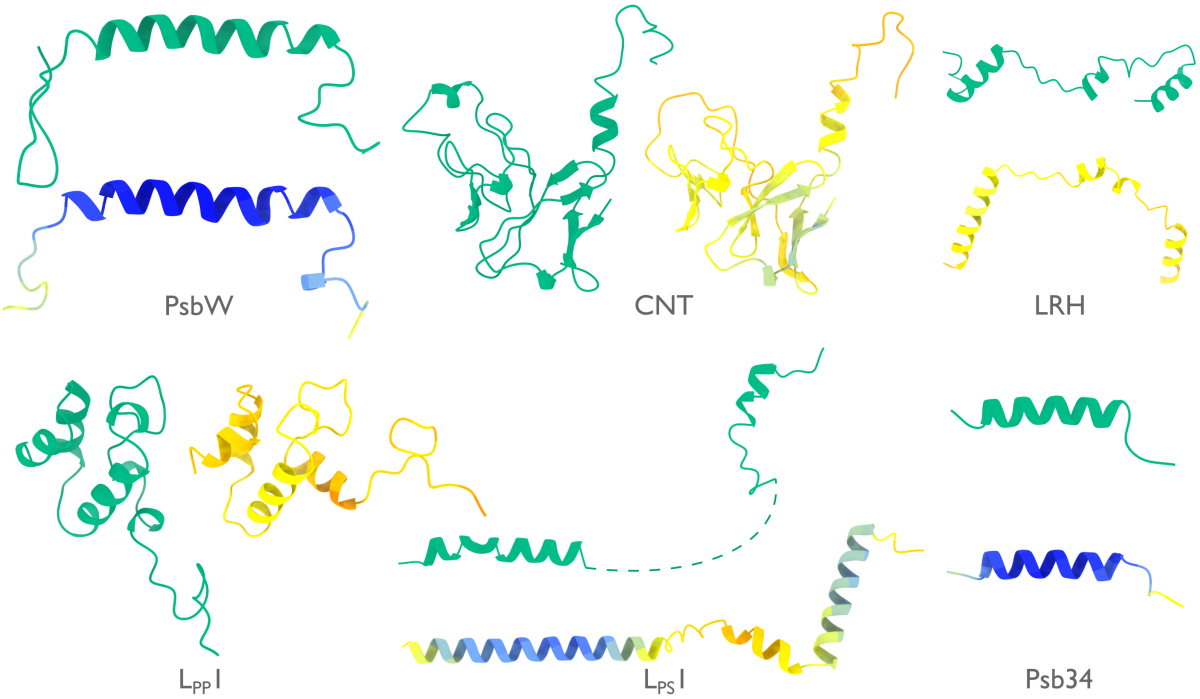
57. Schenk, A. D., Cavadini, S., Thoma, N. H. & Genoud, C. Live analysis and reconstruction of single-particle cryo-electron microscopy data with cryoflare. *Journal of chemical information and modeling* **60**, 2561–2569 (2020).

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



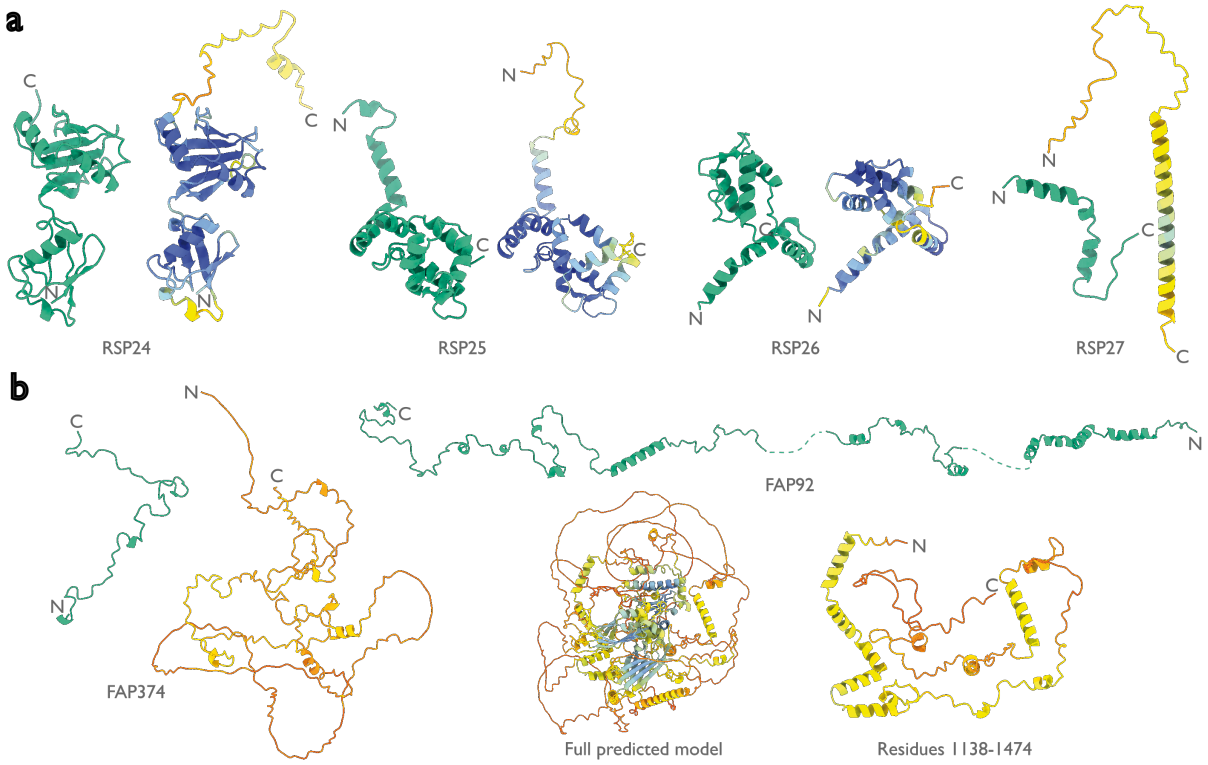
Extended Data Figure 1: Identified proteins in the phycobilisome Atomic models built by ModelAngelo (green) for the six proteins that were identified by ModelAngelo. Side chain densities in the cryo-EM map (transparent grey) are in agreement with those of the atomic models.

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



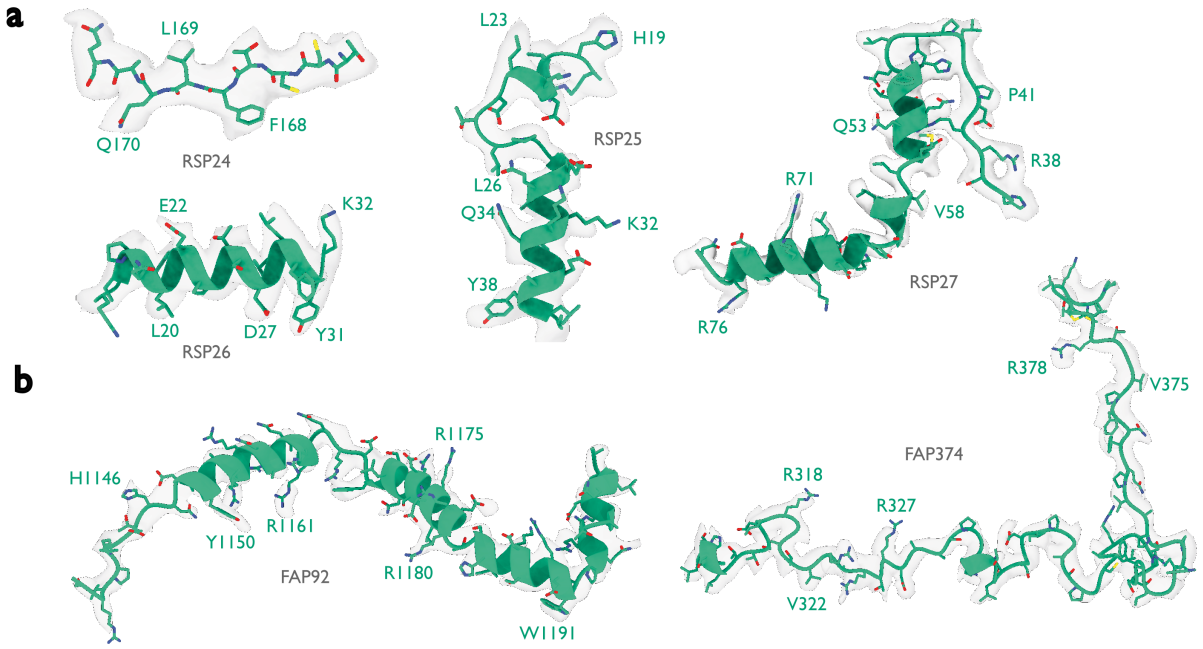
Extended Data Figure 2: Models by ModelAngelo and AlphaFold for identified proteins in the phycobilisome. Models built by ModelAngelo (green) are shown next to predictions of the corresponding sequences by AlphaFold (*15*) (coloured by AlphaFold's confidence from high in blue, to low in red).

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



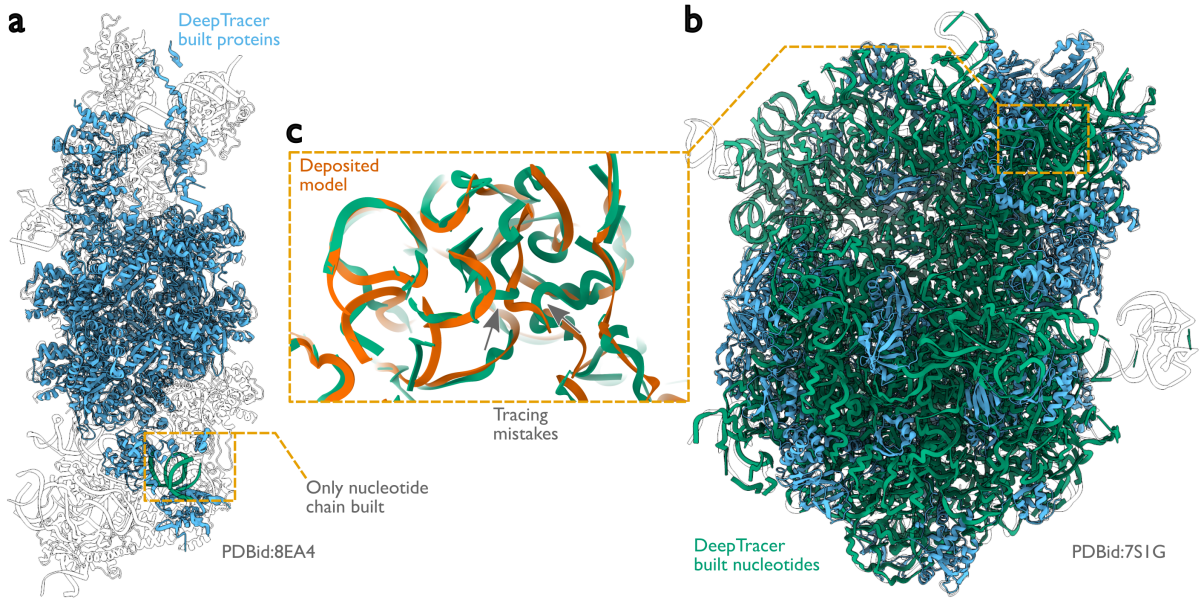
Extended Data Figure 3: Models by ModelAngelo and AlphaFold for identified proteins in the ciliary axoneme. Models built by ModelAngelo (green) are shown next to predictions of the corresponding sequences by AlphaFold (*15*) (coloured by AlphaFold's confidence from high in blue, to low in red). These are split between **a**, the radial spoke proteins, and **b**, the central apparatus microtubule proteins.

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Extended Data Figure 4: Identified proteins in the ciliary axoneme Atomic models built by ModelAngelo (green) for the six proteins that were identified by ModelAngelo. Side chain densities in the cryo-EM map (transparent grey) are in agreement with those of the atomic models. These are split between **a**, the radial spoke proteins, and **b**, the central apparatus microtubule proteins.

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Extended Data Figure 5: Performance of DeepTracer for building nucleotides. a, Protein (blue) and nucleotide (green) chains built by DeepTracer for the V-K CAST transpososome from *S. hofmanni* compared with the deposited model (in black outlines, PDB ID: 8EA4). Only one nucleotide chain is built, in an incorrect position. **b**, DeepTracer model of the *E. coli* ribosome (PDB ID: 7S1G), depicted as in panel a. **c**, For the ribosome, DeepTracer only positioned phosphor atoms for the nucleotides. Errors in the nucleotide phosphor atoms led to incorrect nucleotide chains, as evidenced by comparison with the deposited model (orange).

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.16.541002; this version posted May 16, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Protein	Phytozome ID	Nr. of residues	Built residues	EMDB entry	Map res. (Å)	Location
RSP24	Cre08.g800895	226	1-187	22481	3.4	RS2 stalk
RSP25	Cre01.g034550	176	18-176	22475	3.2	RS1 neck*
RSP26	Cre17.g802036	128	2-125	22475	3.2	RS1 neck*
RSP27	Cre05.g240450	91	36-78	22475	3.2	RS1 stalk
FAP92	Cre13.g562250	1471	1138-1471	25381	3.8	C1 microtubule/ Protofilaments 3-4
FAP374	Cre03.g176600	400	308-386	25381	3.8	C1 microtubule/ Protofilaments 7-9

Extended Data Table 1: Proteins identified in the *C. reinhardtii* axoneme using ModelAngelo.

*RSP25 and RSP26 are also expected to occur in the neck of RS2, which is thought to be identical to the neck of RS1 (*12*).