Published online 3 July 2023

Nucleic Acids Research, 2023, Vol. 51, No. 15 ***8255–8269***
https://doi.org/10.1093/nar/gkad553

DoubleHelix: nucleic acid sequence identification, assignment and validation tool for cryo-EM and crystal structure models

Grzegorz Chojnowski [*]

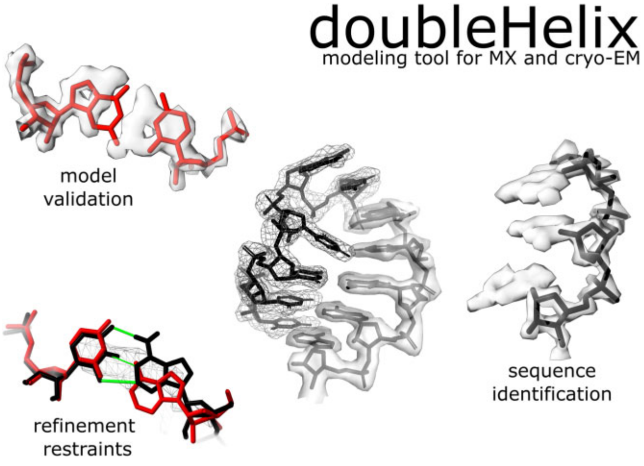
European Molecular Biology Laboratory, Hamburg Unit, Notkestraße 85, 22607 Hamburg, Germany

Received February 17, 2023; Revised May 29, 2023; Editorial Decision June 12, 2023; Accepted June 16, 2023

ABSTRACT

Sequence assignment is a key step of the model building process in both cryogenic electron microscopy (cryo-EM) and macromolecular crystallography (MX). If the assignment fails, it can result in difficult to identify errors affecting the interpretation of a model. There are many model validation strategies that help experimentalists in this step of protein model building, but they are virtually non-existent for nucleic acids. Here, I present doubleHelix—a comprehensive method for assignment, identification, and validation of nucleic acid sequences in structures determined using cryo-EM and MX. The method combines a neural network classifier of nucleobase identities and a sequence-independent secondary structure assignment approach. I show that the presented method can successfully assist sequence-assignment step in nucleic-acid model building at lower resolutions, where visual map interpretation is very difficult. Moreover, I present examples of sequence assignment errors detected using doubleHelix in cryo-EM and MX structures of ribosomes deposited in the Protein Data Bank, which escaped the scrutiny of available model-validation approaches. The doubleHelix program source code is available under BSD-3 license at https://gitlab.com/gchojnowski/doublehelix.

GRAPHICAL ABSTRACT



INTRODUCTION

Nucleic acids are key players in many cellular processes ranging from gene expression to the catalysis of chemical reactions. For many nucleic acid molecules tertiary structure determines function, much like for proteins. Nevertheless, our understanding of the structure-function relationship in nucleic acids clearly lags behind proteins, which is reflected by the disproportion of structures deposited in the Protein Data Bank (PDB) (1). As of January 2023, out of 200 708 available structures only 15 374 (7%) contained a nucleic acid component. The resolution revolution in cryogenic electron microscopy (cryo-EM) seems to be slowly changing this picture as more and more challenging nucleic-acid complexes are being determined using this technique. In 2022, out of 1454 structures with nucleic-acid components deposited in the PDB as many as 804 (55%) were determined using cryo-EM. Many of these cryo-EM structures would be very difficult to solve using other techniques owing to their size and structural heterogeneity (2).

The release of the Artificial Intelligence (AI) based structure prediction programs AlphaFold2 (3) and RoseTTAFold (4) provided means for accurate and widely

[*] To whom correspondence should be addressed. Tel: +49 40 89 902 110; Fax: +49 40 89 902 149; Email: gchojnowski@embl-hamburg.de

accessible structure prediction of protein structures. Although they did not solve the problem of protein structure determination the accurate predictions proved useful for solving the phase problem in macromolecular X-ray crystallography and the interpretation of cryo-EM maps (5). There is no AlphaFold2 or RoseTTAFold equivalent for nucleic acids and most likely won't be soon as building AI 3D structure prediction tools requires huge and diverse training sets of reliable structural models that are currently not available for nucleic acids. The PDB-deposited models are strongly biased towards ribosomal RNA that is usually highly conserved across all kingdoms of live. Moreover, as will be shown later in this work, available experimental RNA structures contain difficult to identify errors that may reduce generalization properties of structure prediction methods. Although attempts to build such tools are already taken (e.g. RoseTTAFoldNA (6) and ARES (7)), experimental techniques will continue to be the method of choice for detailed structural studies of nucleic acid complexes, with all their limitations and bottlenecks.

MX and cryo-EM remain the most frequently used experimental approaches for the structure determination of large biomolecules. The main result of both these methods is an atomic model traced into a map - an interpretation of experimental observations given *a priori* knowledge of biomolecular structure. Although, the main effort of the method developer community is clearly focused on proteins, several techniques facilitating experimental determination of nucleic acid structures in cryo-EM and MX have been developed, e.g. NAUTILUS (8), ARP/wARP (9), PHENIX (10), RCrane (11), COOT (12), ISOLDE (13), DeepTracer (14) and ModelAngelo (15). As with proteins, nucleic acid model building in these methods usually starts with tracing into a density map a ribose-phosphate backbone that makes up two-thirds of a polynucleotide mass. The backbone model is subsequently assigned to a target sequence and complemented with base moieties.

Sequence assignment is a crucial step in macromolecular model building. It is required for the identification and completion of missing fragments in initial models. It is also a fundamental prerequisite of a model interpretation. Failure may lead to register-shift errors, where residues are systematically assigned an identity of a residue a few positions before or ahead in sequence. Although register-shifts may bias model interpretation, they remain one of the most difficult problems to identify and correct in macromolecular models (16). In protein models, register-shifts often result in backbone-geometry outliers when several sidechains are forced into too small density volumes, which can be detected using geometry validation approaches like CaBLAM (17). Moreover, backbone tracing issues (e.g. deletion or insertion) that caused register shift can be occasionally detected as a geometry outlier (18). Nevertheless, it has been shown that regardless of the effort made to validate protein models, register-shift errors are relatively common in PDB (19,20). Particularly affected are very large structures, for example ribosomes, where detailed inspection of a model using interactive tools (e.g. COOT or ISOLDE) is rarely feasible.

Sequence assignment errors in nucleic acids are even more difficult to identify than in proteins. They rarely result in severe geometry issues during model refinement as the ribose-phosphate backbone dominates scattering and its geometry is weakly affected by the presence of misassigned nucleobases. Moreover, small differences between different types of purines and pyrimidines makes visual validation of a sequence assignment very challenging unless high-resolution maps are available.

The most prominent issue related to a sequence assignment error in nucleic acid structures are steric clashes arising from the presence of base-paired nucleobases that don't fit their secondary structure context - non-isostericity (21). For example, a Watson–Crick pair in *cis* orientation that erroneously involves two guanines is too large to fit into a double-helical region (22). At lower resolutions, however, this will be promptly masked by a refinement software and the bases that are weakly restrained by a map shifted to a non-clashing conformation. Nevertheless, these relatively rare issues can be in principle be detected using standard model-validation software, e.g. Molprobity (17).

Another, rarely recognised, issue related to the sequence assignment in model building are unknown target sequences. Until recently, structural studies of macromolecules of unknown identity, e.g. extracted from natural sources, were attempted predominantly using MX (23). Recent developments in cryo-EM, however, forged a completely new path to the studies of uncharacterised macromolecules. It has been shown that cryo-EM reconstructions of protein nucleic-acid complexes, at resolutions high enough for *de-novo* model building, can be determined directly in a cell using subtomogram averaging (24). High-resolution structural information can be also retrieved from a systematic cryo-EM analysis of cell lysate fractions (25,26). In a recent study Skalidis and co-workers (27) presented a complete workflow for identifying biomolecules directly from native cell extracts combining cryo-EM with AI structure prediction methods. Although there are in principle no technical limitations to the identification of nucleic-acid sequences directly from cryo-EM reconstruction, to the best of my knowledge there is no computational tool available that could be used for this purpose.

In this work, I present *doubleHelix*; a computer program for comprehensive nucleic-acid sequence identification, assignment, and validation in MX and cryo-EM models. Similarly to a previously developed program *findMySequence* (28) for protein-sequence identification, *doubleHelix* uses neural network classifiers for estimating residue-type probabilities given a backbone model and a density map. What makes the *doubleHelix* program unique is the way it addresses the inherent nucleobase-type ambiguity that makes it impossible to distinguish adenine from guanine and cytosine from uracil or thymine unless a very high-resolution experimental data is available. The program estimates only the probabilities of purines and pyrimidines in a model. This information is complemented with base-pairing restraints obtained using a new approach that relies on a backbone conformation ignoring nucleobase identities that are not known before the sequence assignment. The base-pair identification approach is based on alignment of recurrent nucleic-acid structural motifs of known secondary structure (e.g. A- or B-form double helices) to the target model. I show that despite its simplicity this approach is both highly specific and accurate. Moreover, the secondary structure information it provides readily improves sequence assignment and identification performance at lower resolutions where base-type classification reliability is reduced. I also show an example of a previously unidentified RNA-sequence assignment errors in mammalian and bacterial ribosome structures deposited in the PDB that could have been avoided if *doubleHelix* had been used for model building and validation.

AN OVERVIEW OF THE DOUBLEHELIX METHOD

The *doubleHelix* program requires on input a model in PDB or mmCIF format. For the sequence identification and assignment, it also requires a corresponding map, which can be provided in CCP4/MRC format for cryo-EM models or as a MTZ file with structure factor amplitudes and phases for crystal structures. The *doubleHelix* program provides four basic functionalities:

- Secondary structure restraints generator for nucleic-acid models

Given RNA or DNA model on input the program generates base-pair and -stacking restraints in formats accepted by COOT and popular refinement programs REFMAC5 (29), PHENIX (30) and ISOLDE (13). Additionally, it generates a PYMOL (31) script that can be used for visualising the restraints (Figure 1C). The restraints are also generated for interacting model fragments modelled as separate chains (e.g. DNA duplexes).

- Identification of unknown sequences of nucleic-acid models

For a nucleic acid model and a corresponding map (CCP4/MRC or MTZ formats for cryo-EM and MX, respectively), the program identifies the most plausible sequence from a sequence-database in FASTA format given estimated nucleobase-type probabilities and input model secondary structure (Figure 1D). By default, the program identifies a sequence that best matches all nucleic-acid chains fragments in the input model.

- Assignment of nucleic acid models to known target sequences

For a nucleic acid model and a corresponding map (CCP4/MRC or MTZ formats for cryo-EM and MX, respectively), the program assigns continuous polynucleotide chain fragments to the target sequence and rebuilds the bases accordingly. Apart from the estimated nucleobase-type probabilities, base-pairs identified within the fragments are used as additional restraints (Figure 1E).

- Sequence assignment validation in nucleic-acid models

Given a nucleic-acid model, a corresponding map (CCP4/MRC or MTZ formats for EM and MX, respectively) and the set of all model sequences, the program evaluates the plausibility of the model's sequence assignment. This feature is implemented as an extension of the *checkMySequence* program and uses an algorithm described previously (19). Users interested in the validation of nucleic acid model sequence assignment should refer to instructions available on the *checkMySequence* project page.

MATERIALS AND METHODS

Recurrent structural motifs in nucleic acids

The *doubleHelix* program identifies base pairs in RNA and DNA models from a local similarity of backbone coordinates with small 'search-fragments' of known secondary structure. Model, double helical A-RNA and B-DNA search-fragments were generated using the X3DNA suite (32). Non-helical search fragments were selected using RNA Bricks (33) database. Selected sets of recurrent RNA fragments classified as 'loops' with at least 500, 100, 25 and 10 occurrences in the database correspond to 83, 532, 1430 and 2664 search-fragments respectively (as of 28 March 2020).

Ribosome crystal structures for secondary structure assignment benchmarks

As a reference for the secondary structure assignment benchmarks, I used crystal structures of ribosomes available in PDB as of 28 March 2020. From all structures determined at a resolution better than 3.0 Å, I selected ones with crystallographic R-free factor below 0.3. To reduce the set redundancy, from each group of similar structures (e.g. originating from the same publication) I selected models with the lowest R-work/R-free difference. The resulting set contained eight structures originating from *Haloarcula marismortui*, *Thermus thermophilus*, and *Deinococcus radiodurans* (PDB entries 1s72, 4ybb, 7rqa, 1hnx, 1fjg, 1k73, 4y4o, 6oxi). For each of the models, secondary structure was determined using the ClaRNA program (34) and used as a ground-truth.

Reference set of ribosome cryo-EM structures for sequence identification benchmarks

From the PDB, I selected cryo-EM structures of ribosomes determined at a resolution better than 3.5 Å. Among 102 such structures available as of 4 February 2020, I selected 17 with half-maps available for download in the Electron Microscopy Data Bank (EMDB). For each of the half-map pairs local resolution maps were calculated using Resmap version 1.1.4 (35) with default parameters.

The selected models (PDB entries 3j79, 3j7a, 3j7q, 5iqr, 5mdv, 5mdw, 5mdy, 5ngm, 5umd, 5wdt, 5we4, 5wfs, 6okk, 6p5i, 6p5j, 6p5k, 6p5n) originated from five different organisms: *Plasmodium falciparum, Escherichia coli, Staphylococcus aureus, Sus scrofa* and *Oryctolagus cuniculus.* For each of them, nucleotide sequences corresponding to RNA features annotated based on the genome sequence were downloaded from NCBI (36) and used as references for the sequence identification benchmarks. The reference sets contain tRNA, rRNA and ncRNA sequences, except for those corresponding to eukaryotic organisms that additionally contain mRNA sequences. To ensure that exact matches are available in the reference sets I added target rRNA sequences to each of them.

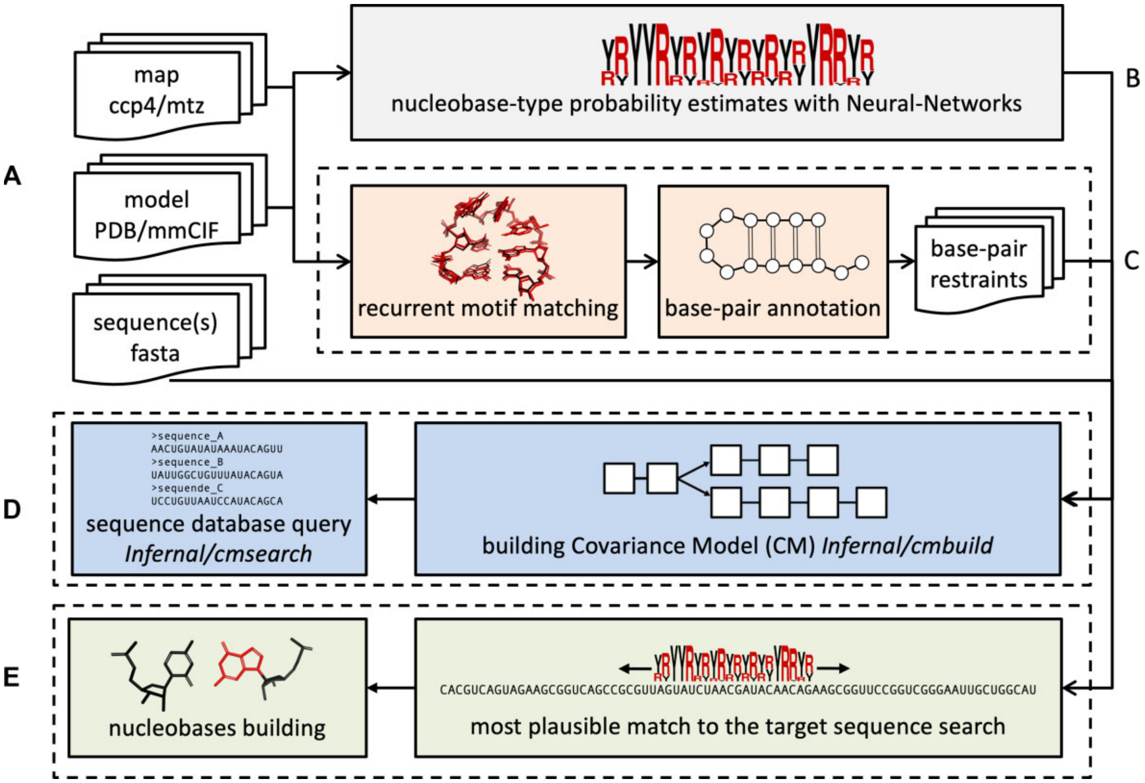


Figure 1. Schematic representation of the *doubleHelix* workflow. Key steps are color-coded and grouped in dashed rectangles; (**A**) input map (cryo-EM or MX), nucleic acid model, and target sequences (**B**) nucleobase-type probability estimation, (**C**) base-pair and refinement restraints assignment based on matched recurrent structural motifs, (**D**) sequence identification and (**E**) assignment based on estimated nucleobase-type probabilities and secondary structure. All steps are integrated in the software and performed automatically.

Structures used for training neural network classifiers

As map features observed in cryo-EM and MX maps differ in fine detail, two separate neural networks were trained for each of these experimental methods.

For training the cryo-EM nucleobase-type classifier from the cryo-EM structures of ribosomes initially selected for sequence identification benchmarks, for which half-maps were not available in EMDB, I randomly selected 10 (PDB entries 5afi, 5mmi, 5u9f, 5wdt, 6eri, 6h4n, 6ogi, 6om0, 6q8y, 6sgc). Additionally, 142 PDB-deposited cryo-EM structures containing a DNA, but not an RNA component determined at resolution 3.5 Å or better with map-to-model correlation coefficient above 0.8 as estimated for complete models (including non-NA components) using phenix.map_model_cc (37) were added to the training set.

For training a crystal structure nucleobase-type classifier, I selected eight structures and corresponding maps that were also used for benchmarking secondary structure assignment procedure. Moreover, 100 crystal structures randomly selected from PDB that contain a DNA, but not an RNA component, determined at resolution 3.5 Å or better with R-free–R-work below 0.3 were added to the training set.

Ribosome structures for the base-type classifier benchmarks

For benchmarking the residue type neural network classifiers, I selected five the highest resolution cryo-EM and MX structures available in PDB as of 24 April 2023 and released after training the base-type classifiers. The resulting set contained crystal structurers refined between 2.3 and 2.5 Å resolution (PDB entries 8cvl, 6xhv, 8cvj, 7rqe and 7rqa). The resolution of selected cryo-EM structures varied between 1.5 and 1.9 Å (PDB entries 8b0x, 8glp, 8a3d, 8aye and 7k00). All the models contain modified nucleic acids residues as modelled by their authors. Each modified residue in the set was classified as a purine or pyrimidine based on the presence of imidazole rings. Specifically, the classification was based on the presence of both N1 and N9 atoms within a base, which are located approximately 4.1 Å apart from each other.

Ribosome crystal structures for sequence identification and assignment benchmarks

For RNA sequence identification benchmarks, I arbitrarily selected two crystal structure models of *Thermus thermophilus* 30S ribosomal subunit determined at a resolution 2.8Å (PDB entry 2uub) and 3.3Å (PDB entry 6mpi). For both targets, re-refined structures were downloaded from the PDB_REDO server (38). Initially, a randomly selected 90% of ribosomal RNA nucleobases were mutated in both models (purines to pyrimidines and vice versa) keeping canonical base-pairing interactions identified using *doubleHelix* (the procedure is implemented in *doubleHelix* program and can be enabled with an option "--*randomize = 0.9*"). The model coordinates were subsequently randomised with 0.2Å RMSD ignoring any geometry restraints and automatically refined using the PDB_REDO web server. For both models the randomisation procedure clearly affected the R-work/R-free factor values that increased from 0.19/0.23 to 0.25/0.29 and from 0.20/0.25 to 0.23/0.27 for better and worse resolution structures respectively. The automatically refined randomised models with corresponding maximum likelihood, combined 2mFo-DFc maps were used for sequence identification and assignment benchmarks.

For the sequence identification benchmarks, nucleotide sequences corresponding to RNA features annotated based on the *Thermus thermophilus* genome were downloaded from NCBI and used for making queries.

Neural network base-type classifier

To estimate the probability that a given nucleotide fitted into a map corresponds to a purine or pyrimidine two independent neural-network classifiers were prepared. The classifiers have identical architecture but are trained on distinct training sets derived from crystal structures or cryo-EM models and their respective maps.

Nucleotides are described with a vector of map values sampled around a putative base moiety (a residue descriptor). The map is sampled on a regular grid with 1.0 Å spacing. The grid is centred at the N1 or N9 atom for purine and pyrimidine respectively and spanned by orthonormal vectors defined by glycosidic bond (e_x), the normal vector of the ribose best-fitting plane (e_y), and their cross product ($e_z = e_x \times e_y$). For a given nucleotide the input to the classifier contains a cloud of 403 grid points that are within 1.0 Å distance from any atom of a nucleobase mutated to Guanine in any rotation around the glycosidic bond. In practice a precomputed cloud is aligned to each nucleotide using C2', C1' and O4' ribose or deoxyribose atoms.

The neural-network model input is a vector of length 403 (the residue descriptor described above). The model contains two, fully connected hidden layers. The first layer has a ReLU (Rectified Linear Unit) activation function, which sets all negative neuron inputs to zero, and 403 output features. The second layer has 2 output features and uses the log-softmax normalisation function enabling estimation of output classification probabilities. To avoid overfitting, an additional dropout layer was inserted between the two hidden layers. The dropout layer at each training step disables neuron connections selected at random with probability P. The models were trained for 1000 epochs with $P = 0.5$, a batch size of 20 residue descriptors in each parameters update cycle, and a 10% validation set. The models were trained using the ADAM optimization algorithm (39) with a learning rate of 1e–5 that resulted in the best test-set accuracies.

For training the crystal structure classifier I used 84 887 and 9431 nucleobase descriptors for training and test-set respectively. The accuracies of a resulting model were 0.98 and 0.96 for the training and test sets, respectively. Similarly, for training the EM classifier I used 85 092 and 9454 residue descriptors in training and test-sets, respectively. The resulting model estimated accuracies were 0.95 and 0.92 for training and test set, respectively.

Secondary structure assignment procedure

Unlike DNA, which occurs in nature predominantly in a double-helical form, RNA molecules are often single-stranded and fold into complex structures stabilised by stacking and base-pairing interactions (40). Folded RNA molecules have a modular architecture in which the double-helical regions are intertwined with different types of loops that define the topology of the structure and stabilise it through long-range interactions. Many of these loops are recurrent and can be found in a similar structural context in many, possibly evolutionary unrelated, RNA molecules (33). Most importantly, it is the overall module geometry and base-pairing pattern rather than the nucleotide sequence that is conserved across different occurrences of the same module (41). This feature of RNA molecules is used in the *doubleHelix* program for the inference of base-pairing interactions from the local geometry of sugar-phosphate backbone. This approach ignores both identities and mutual orientation of bases, which is particularly important in the analysis of preliminary, not fully refined models, where base identities are not yet known and their coordinates, unlike relatively heavy backbone, may be inaccurate.

The program superposes small RNA or DNA search-fragments of known secondary structure onto the input model using an algorithm described previously as a part of a model-building program *Brickworx* (2). First, all possible triplets of phosphate group P-atoms in a search-fragment are structurally aligned with similar P-atom triplets from the input structure. Resulting rigid body transformation is applied to the complete fragment to identify matching nucleotides in the search fragment and input structure. Finally, the match is refined using all sugar-phosphate backbone atoms. If the resulting root-mean-square deviation (RMSD) is below 1.0 Å (the threshold defined in the Results section), search-fragment base pairs are assigned to the corresponding residues in the input model. If multiple, overlapping matches of search-fragments are identified, the one with the lowest RMSD is selected.

For the sake of computational efficiency, the input model processing is divided into two steps. Firstly, only a double-helical fragment is matched to identify Watson-Crick base-pairs. In this step, A-RNA or B-DNA search fragments are used depending on the target nucleic acid type. Next, all nucleotides within stacked Watson-Crick base-paired regions (except flanking residues) are removed from the input model. In the second step, used only in case of RNA targets, a predefined set of recurrent RNA motifs is matched to the remaining nucleotides in the input model. In case of base-pair assignment conflicts, the ones detected in the second step are given preference.

Sequence identification procedure

For the identification of the most plausible sequence in a database, given input model residue-type probabilities and secondary structure, *doubleHelix* uses sequence-comparison tools from the INFERNAL suite (42) (Figure 1D). Initially, predicted residue-type probabilities are converted into a multiple sequence alignment (MSA), where fractions of residue types in each column correspond to predicted probabilities. The MSA, combined with base-pairing pattern is encoded in a Stockholm format (STO), which is an input to the *cmbuild* program. The resulting Covariance Model (CM) is further calibrated with *cmalibrate* and used to query sequence databases using the *cmsearch* program with default parameters. The Hidden Markov Model (profile-HMM) queries are enabled by adding the --*hmmonly* keyword to *cmsearch* and skipping the calibration step. Sequence hits with the lowest E-values are returned to the user (3 by default).

Sequence assignment procedure

Analogously to the sequence identification procedure, *doubleHelix* uses the estimated base-type probabilities to assign RNA or DNA models to a target sequence. For a continuous polynucleotide fragment in the input model a neural network classifier is used to estimate base-type (purine or pyrimidine) probabilities for each residue. The resulting scoring matrix is aligned to the target sequence and the probability of each tentative alignment is approximated by a product of the probability estimates for each residue in the fragment, assuming their independence. If any residue pair in the fragment forms a Watson–Crick base-pair (detected using a procedure described above) an additional, low-probability correction factor (0.1) is used if for a tentative alignment the two bases are either both purines or pyrimidines, which is very unlikely for a Watson–Crick interaction. Otherwise, the correction term is 1.

Although the neural-network classifier has been calibrated and the predicted residue-type probabilities generally reflect expected frequencies, the accuracy of predictions may vary depending on local map resolution and quality of the models (9). Therefore, for each tentative assignment of a fragment to a target sequence the method estimates a p-value, or probability that a given alignment has been observed by chance. A tentative alignment probability is compared to a background distribution of the fragment alignment probabilities for a long, random sequence. To additionally account for the varying target-sequence lengths an additional extreme-value distribution theorem correction is applied as described before (19).

Implementation

The *doubleHelix* program was implemented using Python3 with an extensive use of numpy (43), scipy (44), CCTBX (45) and CCP4 (46) libraries and utility programs. The neural network classifier used in this work was trained using Pytorch (47) and deployed to a C code using keras2c (https://github.com/f0uriest/keras2c). For making the rRNA sequence database queries, the program uses INFERNAL suite version 1.1.4.

RESULTS AND DISCUSSION

Secondary structure assignment

The base-pair assignment in the *doubleHelix* program relies on structural alignment of recurrent motifs with known secondary structure (search-fragments) to the input model. The method uses five different sets of search-fragments. First, A- or B-form double helices, for RNA and DNA target models, respectively, are tried. Additionally, for RNA models, the method uses four sets of the most frequent recurrent RNA motifs from the RNA Bricks database (see Materials and Methods for details). Although the double-helical search fragments can be used for the identification of the canonical Watson–Crick base-pairs only, the other search fragments allow for the identification of any base-pairing interaction type.

The base-pair assignment procedure parameters providing maximum performance are 2-base-pair double-helical search fragments with 1.0 Å RMSD threshold (Figure 2A). This can be explained with a relatively high structural heterogeneity observed in double-helical structures (48), which cannot be represented using longer, idealised search models. Interestingly, an additional search step with recurrent RNA motifs with at least 500 occurrences in the RNA Bricks database further improves precision of the Watson-Crick base-pairs assignment. The resulting Watson-Crick base-pair assignment *f*1-score 0.94 is comparable to a value of 0.95 reported for a recently described program CSSR (49), which also relies on backbone conformations only. With *doubleHelix* this translates to the recall of 0.91 and precision 0.98 (no corresponding results were reported for CSSR). Unlike doubleHelix, however, CSSR focuses exclusively on pairs of nucleotides compatible with canonical base pairing (A/U, G/C or G/U) that makes the two methods not directly comparable. Another advantage of the *doubleHelix* approach over CSSR is its ability to identify stackings and non-canonical base-pairs with recall/precision of 0.63/0.89 and 0.47/0.94, respectively (Figure 2B). This, however, requires the use of a larger set of recurrent RNA motifs with at least 100 instances in the RNA Bricks database that results in an increased computational cost. For example, for a tRNA model (76 nucleotides, PDB entry 1ehz) processing time on a standard laptop increases from 4 s in default configuration to 13 s. For a complete porcine 28S rRNA (3938 nucleotides, PDB entry 3j7q) these times increase from 1 to almost 8 h. This, however, is needed only for the accurate assignment of non-canonical base-pairs in RNA models (e.g. as refinement restraints). By default, for the identification of Watson–Crick base-pairs the *doubleHelix* uses the smallest set of RNA recurrent motifs providing maximum performance at a reasonable computational cost.

Base-type classifier benchmarks

The neural network residue-type classifiers used in this work were trained ignoring any base modifications in the structures. However, these modifications can effectively alter the scattering properties of a base and introduce bias in the classification results. To examine the impact of the base modifications on the classification performance, I selected ten high-resolution models of ribosomes containing modified ribonucleotides, as described in Materials and Methods section.

The crystal structure models of ribosomes contained in total 45 784 standard ribonucleotides out of which 99% have been correctly classified. Among 358 modified bases the most frequent were pseudouridines (81) and all of them were correctly classified as pyrimidines. For other ribonu-

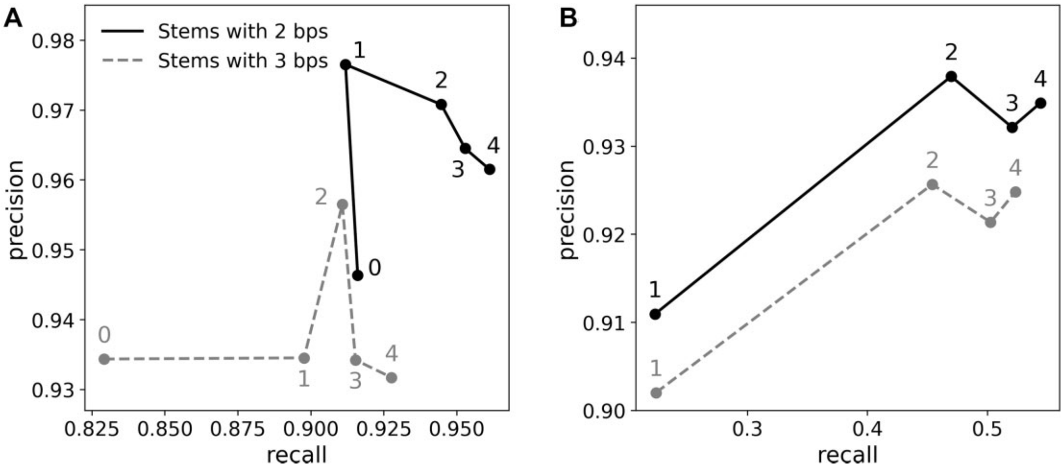


Figure 2. Performance of RNA secondary structure assignment based on structural alignment of recurrent motifs for (**A**) canonical (Watson–Crick) and (**B**) non-canonical base-pairs. Each data point represents performance for a given stem length (2 or 3 bp) and main-chain atoms RMSD maximising classification *f*1-score (harmonic mean of precision and recall). Data-point labels correspond to the number of recurrent RNA motifs used for model interpretation; RNA stem only (0), motifs with at least 500 (1), 100 (2), 25 (3) and 10 (4) occurrences in RNA Bricks database. Precision and recall were estimated using ClaRNA (34) base-pair classification as ground truth.

cleotides with modified bases, 222 out of 234 (95%) were correctly classified as purines or pyrimidines. For the remaining modifications that do not affect the base (e.g. O2'-methyluridine), 39 out of 42 (93%) were classified correctly.

In the set of cryo-EM models, 98% out of 28 791 standard ribonucleotides, 297 out of 301 (99%) modifications not affecting bases, and all 294 pseudouridines were classified correctly. Among ribonucleotides with base modifications 89 out of 91 (98%) were classified correctly.

Overall, the performance of base-type classifiers for both cryo-EM and MX agrees with the estimates from the training procedure. This includes modified bases, even though given high-resolution of the maps, the modifications are usually resolved in the density. It can be expected that at lower resolutions the effect of base-modifications will be also negligible as modifications are not resolved in the maps anyway.

Sequence identification in cryo-EM

For the identification of a nucleic acid model the *doubleHelix* program finds the most plausible sequence in a database given nucleobase-type probabilities (purine or pyrimidine) estimated based on a backbone model and corresponding cryo-EM map. Secondary structure restraints, derived directly from a backbone model, are used as an additional source of information. Both base-type probabilities and base-pairing information are used to query sequence databases using the INFERNAL suite as described in the Materials and Methods section. I observed that this approach allows for a sequence identification up to 4.5Å local resolution for fragments of 50 amino acid residues (Figure 3A) when Covariance Models (CMs) and secondary structure information is used. The use of Hidden Markov Models (HMMs), which neglects base-pairing information, clearly reduces the method performance (Figure 3A). The use of longer fragments of 100 residues, further increases the resolution limit of the method applicability up to 5.5 Å when the base-pairing information and CMs are used (Figure 3A). Overall, the *E*-value provided by the INFERNAL suite is a reliable measure of the sequence identification reliability. There are, however, a few model fragments for which the identified sequence has a relatively low sequence identity to the target, despite a reliable *E*-value score (Figure 3B). This issue can be attributed to the reduction of the sequence identification problem to purines and pyrimidines only that may occasionally make different sequence fragments practically indistinguishable.

Sequence assignment in cryo-EM

It has been shown that a neural network-based assignment of protein model sequence can provide reliable results up to local map resolutions where *de novo* model building would be very challenging (50). Moreover, a carefully derived sequence assignment reliability score, the p-value, accurately separates correct from wrong alignments independently of local map resolution. The assignment of polynucleotide sequences, however, presents a particular challenge compared to proteins. First of all, nucleic acid models are often built *de novo* into lower resolution map regions, as double-helical fragments are excellent and universal models for map interpretation. Moreover, even at moderate resolutions the target sequence is effectively reduced to only two types of nucleobases (purines and pyrimidines) that greatly increases sequence assignment ambiguity.

I observed that similarly to proteins the neural-network classifier implemented in *doubleHelix* provides a reliable means of assigning polynucleotide sequences to model fragments at local resolutions as low as 5Å, even though the required fragment lengths are clearly longer (Figure 4A). For a total number of 18 655 RNA test-fragments of 20 residues (see Materials and Methods for details), the assigned sequence matched the corresponding model in 83% (15 461) of cases. For longer fragments of 40 residues the number of correctly assigned sequences increases to 95% (17 383).

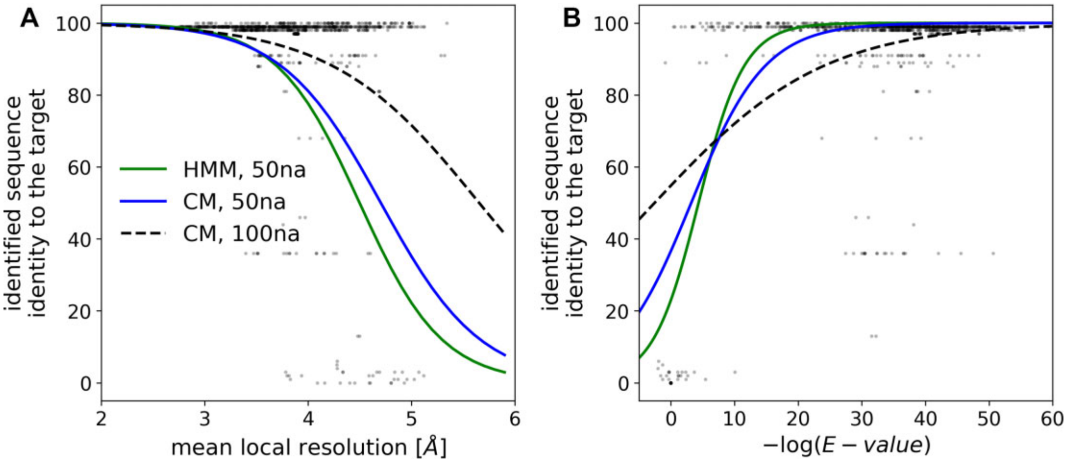


Figure 3. Sequence-identification benchmarks with continuous fragments of 50 and 100 nucleic acid residues selected at random from cryo-EM ribosomal RNA models. Comparison of the method performances for an identification of model sequence with (CM) and without (HMM) the use of base-pairing information. The sequence identification performance is shown as a function of (**A**) local resolution of EM maps and (**B**) *E*-value of the sequence assignment estimated by the INFERNAL suite. The (B) plot horizontal axis shows -log(E-value); higher values correspond to lower *E*-values and more reliable sequence identification results. The continuous and dashed curves are logistic regression estimates of a probability that an identified sequence will have at least 90% sequence identity to the target sequence.

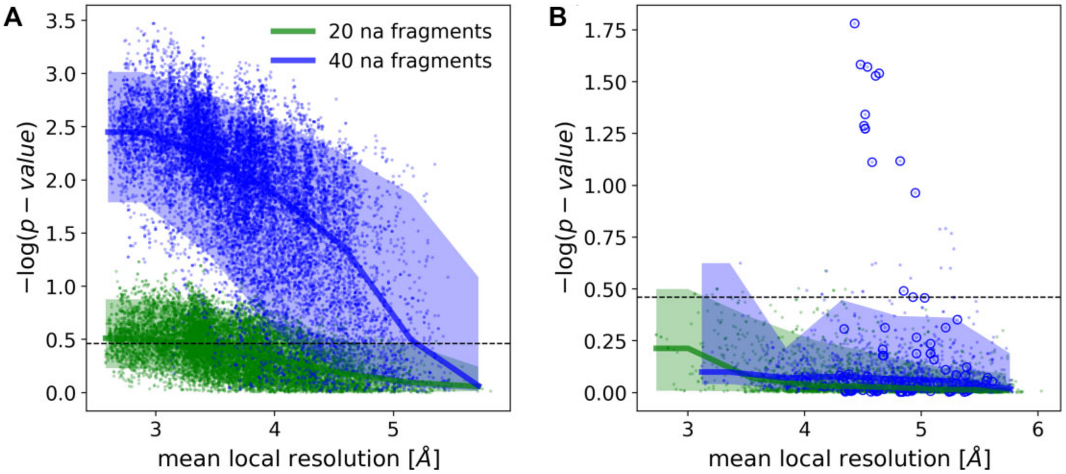


Figure 4. Medians and 90% confidence intervals for sequence assignment p-value as a function of local resolution for RNA chain test-fragments of 20 and 40 nucleic acid residues. Panel (**A**) shows fragments with sequences matching reference. Fragments for which assigned and reference model sequences differ are presented in panel (**B**). The dashed line corresponds to a 99.5% one-sided confidence interval estimated for fragments of 20 nucleic acids with an assigned sequence different from the input-model sequence. Blue circles depict an outlier presented in the text (porcine 28S rRNA, PDB entry 3j7q). The plots' ordinate axes show –log(p-value) for the test-fragments; higher values correspond to lower p-values and more reliable sequence assignments.

The p-value, or the probability of observing a given sequence assignment by chance, provides a reliable estimate of the alignment accuracy that doesn't depend on local map resolution (Figure 4B). Indeed, a one-sided 99.5% confidence interval for fragments with sequence assignment that doesn't match the reference (dashed line on Figure 4) corresponds to 62% and 17% of cases with matching sequences for fragments of 20 and 40 residues respectively. Although sequence mismatches with a p-value outside this region are expected to be very rare, I observed several such test-fragments in the benchmark set (blue circles on Figure 4B), all of them correspond to a single model of porcine 28S rRNA (PDB entry 3j7q). I will discuss this outlier in more detail in the next section.

Sequence assignment outlier: mammalian 28S rRNA

In the cryo-EM benchmark set, I identified several clear outliers, where RNA test-fragments were assigned reliable (low p-value) sequences different from the reference model (Figure 4B). All the fragments originate from an expansion segment (ES7a) of a cryo-EM structure of porcine 28S ribosomal subunit determined at 3.4Å resolution (PDB entry 3j7q). Closer inspection of the model revealed several nucleobases poorly fitting the EM reconstruction, which is understandable given limited resolution, but no clearly visible sequence assignment issues. As there is no higher resolution structure for the porcine ribosome available, which could be used to validate the sequence register, I decided to use as a reference the closest homologue from rabbit (98% of sequence identity), for which a structure determined at 3Å resolution has been recently deposited (PDB entry 6r5q). For a detailed comparison I selected a fragment of the ES7a that has a strictly conserved sequence in both organisms according to an alignment generated using R-coffee (51). Structural alignment of the corresponding model fragments, however, revealed multiple differences between corresponding nucleobase identities, several of them resulting in base-pairing violations (Figure 5C). I also observed that several differences in sequence preserving secondary structure are visible as clear density-fit outliers (Figure 5A). The problem is easily solved by shifting the sequences of the two chain fragments by one residue, as suggested by the *doubleHelix* program, which results in a perfect fit between the porcine and rabbit models (Figure 5B).

RNA sequence identification and assignment in MX

A crystallographic diffraction experiment provides only amplitudes of complex structure factors required for calculating electron-density maps. Missing phases need to be derived from other sources, for example a tentative model of the unknown crystal structure from Molecular Replacement procedure. The use of model-derived phases for calculating electron-density maps inevitably results in so-called 'model bias' - the presence in a map of model features that are absent in a crystal structure. The same issue may be expected when a tentative crystal structure model polynucleotide sequence doesn't match an unknown crystal structure. Although the model bias is reduced in maximum likelihood maps (54), routinely used for model building and refinement in MX, the problem is not completely eliminated. To investigate how the model-bias issue affects the sequence identification and assignment procedures I benchmarked them using ribosome crystal structures with randomised rRNA sequences. I used two crystal structure models of *Thermus thermophilus* 30S ribosomal subunit determined at resolutions 2.8Å (PDB entry 2uub) and 3.3Å (PDB entry 6mpi). Additionally, to remove any effect related to the presence of protein chain models that were refined in the presence of the correct rRNA sequences all atomic coordinates were randomised with 0.2Å RMSD. The resulting, randomized 30S models were refined using REFMAC5 and the PDB_REDO webserver. Interestingly, observed model bias in both randomised structures is moderate and clear sequence mismatches can be noticed for few (but not all) nucleobases (Figure 6B and D).

The sequence identification procedure was very effective for the randomised 2.8Å resolution model. Among 1000 continuous, randomly selected rRNA fragments of 100 residues, *doubleHelix* identified a correct target sequence in 99% of cases. For shorter fragments of 50 residues this fraction reduces to 76%. At lower resolution (randomised 6mpi model at 3.3Å resolution) the performance clearly reduces. The program provided a correct hit in 86% and 62% and of cases for test-fragments of 100 and 50 residues respectively. Interestingly, I also observed that the use of secondary structure information for the sequence identification clearly improves the method performance. Without base-pairing restraints the fraction of correctly identified sequences for fragments of 100 residues reduces to 90% and 72% for better and worse resolution structures. In all cases the incorrect hits can be easily filtered based on E-value returned by INFERNAL suite where values smaller than 0.1 guarantee a correct solution (Figure 7A).

For sequence assignment the method correctly identified sequences of 46% and 65% of 1000 continuous rRNA fragments of 20 nucleic acid residues selected at random from randomised models at worse and better resolution respectively. These numbers increase to 82% and 94% for longer fragments of 40 residues. The correct sequence assignments are also clearly separated from incorrect ones by p-value with the 99.5% one-sided confidence interval for wrongly assigned fragments matching a value estimated for EM models (Figures 4B and 7B).

Case study: sequence register errors in crystal structure of the large ribosomal subunit of *S. Aureus*

The *doubleHelix* program has been integrated into a previously developed tool called *checkMySequence,* which is used for validating sequence assignment in protein models. This integration enables a fully automated validation of sequence assignment for complete protein-nucleic acid complexes. Benchmarks of the updated *checkMySequence* method revealed a particularly interesting structure of the large ribosomal subunit from *S. aureus*. The program identified in this model plausible register shift errors in two ribosomal proteins (L18 and L4) and 5S rRNA. The structure was determined at 3.5Å resolution (PDB entry 4wce, (55)) and refined to R-work/R-free factor values of 0.202/0.246 and *clashscore* 11.

A detailed discussion of the register shift error correction in ribosomal proteins is out of scope of this work and will be presented only briefly. The protein chains were replaced with corresponding predictions from AlphaFold2 database (release 4 for UniProt entries Q2FW07 and Q2FW22 for L4 and L18 respectively) subsequently refined using COOT in real space with self-restrains generated at 5 Å cut-off. Both predicted models were assigned a very high confidence score (pLDDT > 90), which was observed to usually correspond to minor differences in loop and side-chain conformations compared to reliable experimental models (56). Comparison of the deposited and corrected protein chains of both proteins revealed multiple plausible tracing issues and confirmed register-shifts suggested by *checkMySequence*.

The target sequence of the 5S rRNA chain consists of 114 residues that were all traced in the map (chain Y in the deposited model). The *checkMySequence* scan revealed an alternative sequence assignment with p-value of 0.01 (very reliable according to Figure 7B) for a chain fragment following residue 81. The alternative sequence corresponds to a register shift by one base which suggests that a residue could have been omitted in the deposited model (deletion). Although closer inspection of the crystal structure revealed several clear density outliers for bases (Figure 8A), there were no signs of tracing issues. To confirm the chain sequence correctness, I performed sequence identification using *doubleHelix* with the deposited 5S chain coordinates, corresponding maximum likelihood 2mFo-DFc map, and a set of RNA sequences for *S. aureus* strain NCTC8325 downloaded from NCBI. At the time of writing (25.11.2022) there were two assemblies

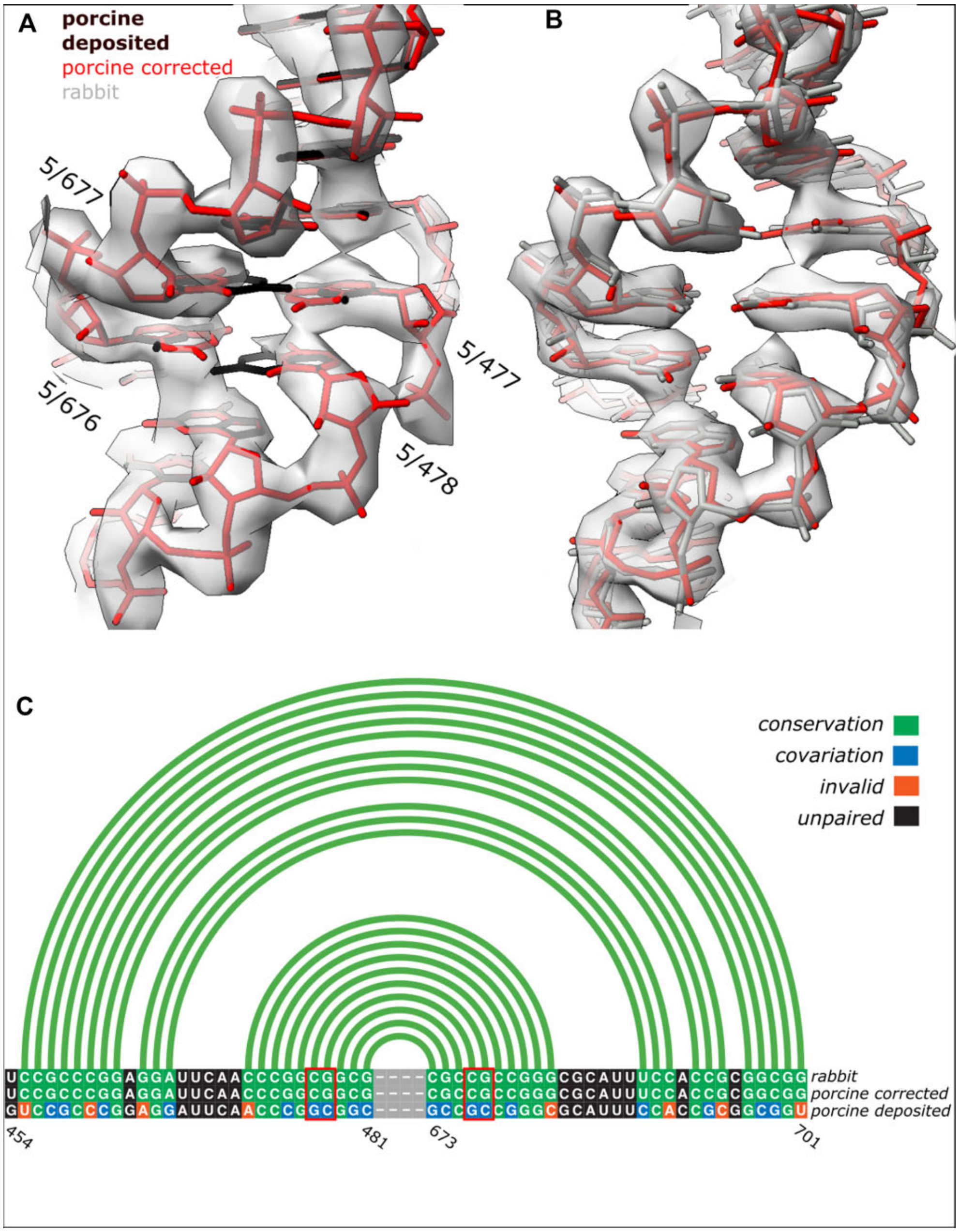

A
porcine
deposited
porcine corrected
rabbit
5/677
5/676
5/477
5/478
B
C
conservation 🟩
covariation 🟦
invalid 🟧
unpaired ⬛
UCCGCCCGGAGGAUUCAACCCGGCGGCG----CGCCGCCGGGCGCAUUUCCACCGCGGCGG rabbit
UCCGCCCGGAGGAUUCAACCCGGCGGCG----CGCCGCCGGGCGCAUUUCCACCGCGGCGG porcine corrected
GUCCGCCCGGAGGAUUCAACCCGGCGGC----GCCGCCGGGCGCAUUUCCACCGCGGCGGU porcine deposited
454
481
673
701

Figure 5. Fragments with strictly conserved sequence of porcine (**A**) and rabbit (**B**) expansion segments (ES7a) in 28S rRNA with corresponding cryo-EM reconstructions at 3.4Å and 3.0Å resolution, respectively. Black model on panel (A) and grey on panel (B) represent deposited coordinates whereas porcine structure with sequence re-assigned using *doubleHelix* is depicted in red. Aligned sequences and secondary structures of the rabbit and porcine models are presented on panel (**C**). Although the register shift in deposited porcine structure preserves most of the base-pairs (green and blue boxes), several of them are visible as clear density-fit outliers (bases in red rectangles and labelled on panel A). There are also multiple secondary structure violations (shown in orange). Secondary structure presented on panel (C) was identified from the corrected model using ClaRNA. The figure was prepared using ChimeraX (52) and R-chie webserver (53).

available (GCF_900475245 and GCF_900475245) each containing roughly 100 tRNA, rRNA, and ncRNA sequences.

In the RNA sequence sets *doubleHelix* identified two, different 5S genes (96% sequence identity), both very reliable hits with *E*-value below 1e-20 (see Figure 7A). One of the sequences, however, scored visibly better that the other (7e-25 versus 4e-21 for NC_007795.1_rrna_7 and NC_007795.1_rrna_6 respectively), which has been previously shown to be a good evidence for a better fit to the data for protein models (28). The deposited 5S model was assigned the sequence variant with worse E-value, which resulted in the map-model fit outliers mentioned above (Figure 8A). The model, after re-assigning the 5S sequence variant identified with better E-value and subsequent restrained refinement with REFMAC5 and PDB_REDO shows much better fit to the map (Figure 8B). The new base identities are also more favourable for the formation of canonical base-pairs in the model (Figure 8C). The fi-

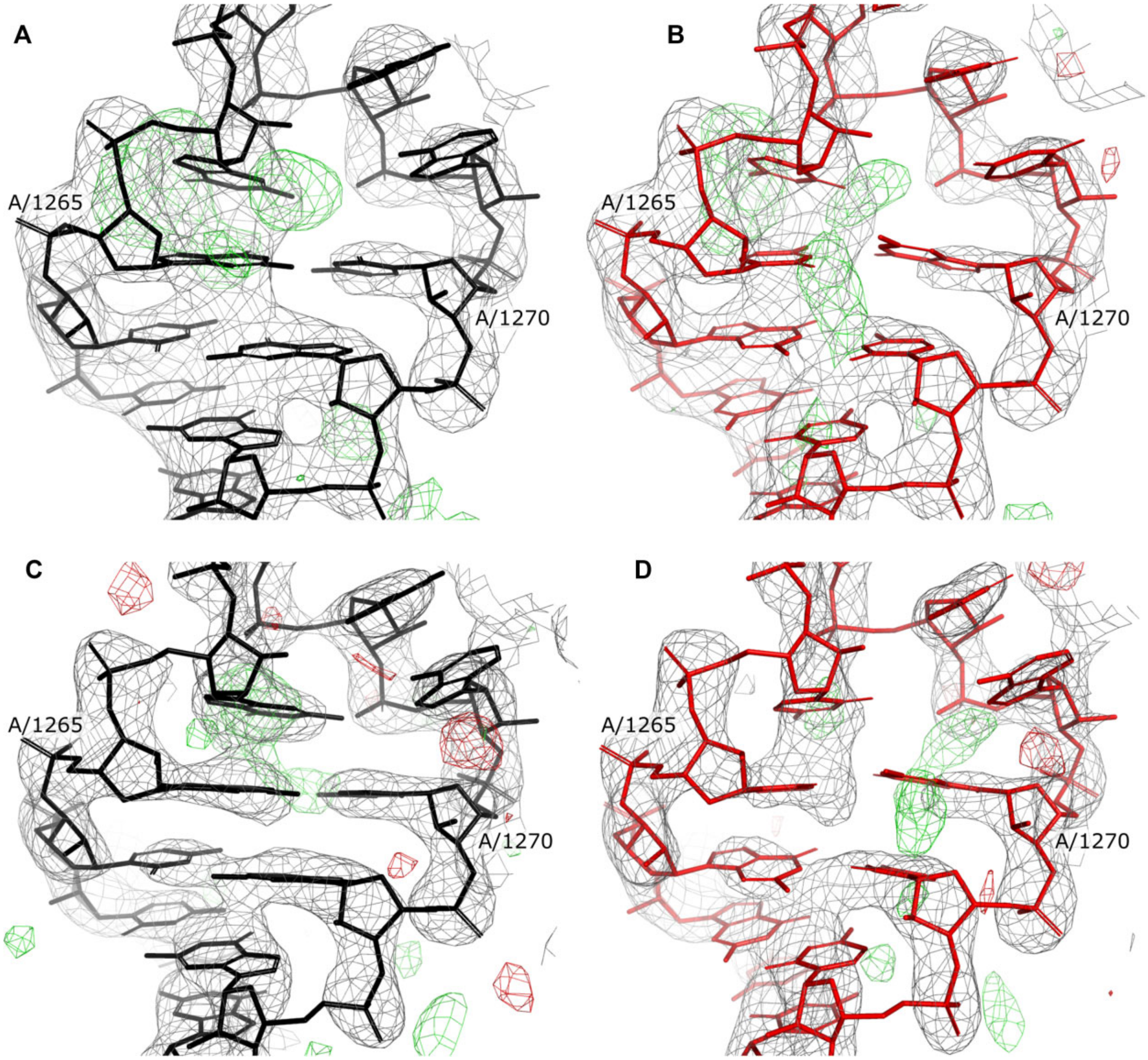


Figure 6. Corresponding fragments of *Thermus thermophilus* 28S ribosomal subunit crystal structures used for the sequence identification and assignment benchmarks. Two models with randomised rRNA sequence were generated based on crystal structures determined at a resolution of 2.8 Å (PDB entry 2uub, panels **A** and **B**) and 3.3Å (PDB entry 6mpi, panels **C** and **D**). The panels depict residue range 1262–1273 with corresponding combined 2mFo-DFc (grey) and difference mFo-DFc (red/green) maximum likelihood maps contoured at 2σ and 3σ levels respectively; as deposited (A, C) and after randomising atomic coordinates and mutating 90% of nucleobases (B, D, see Materials and Methods for details). Both deposited and randomised structures were automatically refined using the PDB_REDO web server.

nal model of complete ribosome, after correcting tracing issues in the two protein chains (L4 and L18) and the 5S chain sequence refines with clearly better scores; R-work/R-free reduces to 0.195/0.237 (from 0.202/0.247) and clashscore to 5 (from 11). The additional base-pairs, visually better map-to-model fit and reduced global model-quality scores together provide good evidence of improved agreement of the corrected model with the experimental data.

CONCLUSIONS

Sequence assignment is a key step of macromolecular model building that may lead to difficult to identify errors affecting structure interpretation. Nevertheless, it has been shown that protein models deposited in the PDB, despite expensive model validation efforts, contain many register-shift errors (16,20,57). Validation and assignment of nucleic acid sequences presents a particular challenge compared to proteins; the models are usually poorly resolved in cryo-EM and MX maps and available sequence-information is in practice limited to two nucleobase-types. Moreover, validation of nucleic acid models addresses predominantly backbone conformations that are rarely affected by nucleobase-sequence assignment issues. In consequence, the reliability of available, experimentally determined nucleic acid models is very difficult to assess. This may result in unintended error propagation, where a newly deposited model contains an error inherited from an earlier one used for model building. Errors can also detrimentally affect efforts of bioinformaticians working on large-scale structural analyses or structure prediction methods.

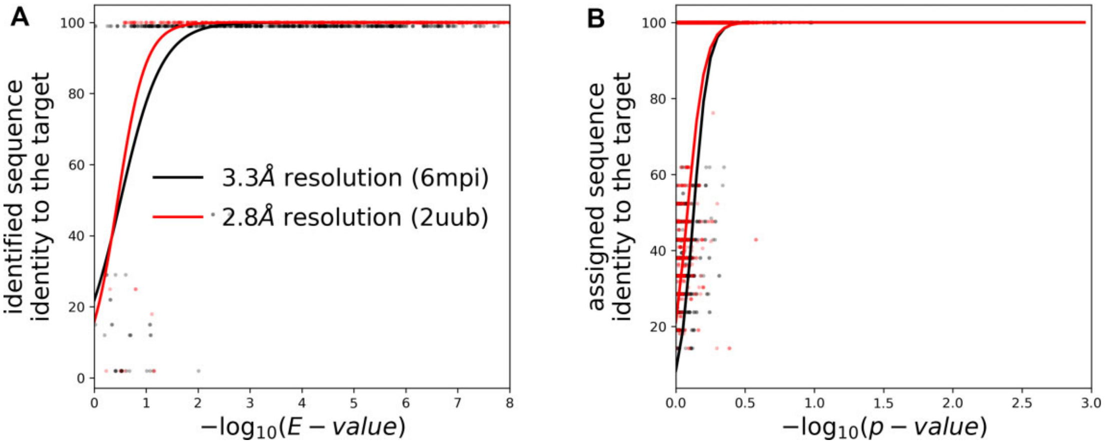


Figure 7. Performance of (**A**) sequence-identification and (**B**) sequence-assignment for fragments selected at random from two MX ribosomal RNA models determined at a resolution of 2.8Å (PDB entry 2uub) and 3.3Å (PDB entry 6mpi). The random fragments used for sequence identification and assignment had 50 and 20 nucleic acids, respectively. The continuous curves are logistic regression estimates of a probability that an identified sequence will have at least 90% sequence identity to the target sequence.

Here, I presented *doubleHelix*; a new program for a comprehensive assignment, identification, and validation of nucleic acid sequences in cryo-EM and MX. I show that the approach, which relies on a neural-network base-type classifier, can successfully identify and assign sequences of cryo-EM model fragments at local resolution as low as 5 Å. I also show that base-pairing information, derived from backbone-geometry, clearly improves the program's performance at lower resolutions but is not essential. This is particularly important for large nucleic acid structures with very low content of paired bases; for example the trypanosomal mitochondrial ribosome (58).

Nucleic acid model building in cryo-EM and MX usually begins with tracing a ribose-phosphate backbone, which is then assigned to a target sequence. Although the former step can be addressed by several fully automated and interactive tools, the sequence assignment remains an open issue. Popular crystal structure building programs and modern, AI-based EM model building tools (e.g. DeepTracer or ModelAngelo) usually produce intermediate models that need to be completed and partially assigned to the target sequence using interactive software (14,15). The *doubleHelix* program should prove useful in this model building step. It can be used for assigning nucleic acid model fragments to the target sequence after each round of model rebuilding (and refinement in case of MX). It can help a user identify model connectivity and make decisions about consecutive modelling steps. Particularly important for this purpose is the sequence-assignment score provided by *doubleHelix* (p-value) that can help assessing local correctness of intermediate models.

The *doubleHelix* software can be also used for generating base-pairing restraints ready-to-use with the most popular cryo-EM and MX model-building and refinement tools (REFMAC5, PHENIX, COOT, ISOLDE). I have shown that the approach can successfully identify 91% of canonical base-pairs with 98% precision without relying on base conformation or identity. This may be particularly useful in early stages of the model building process as available base-pairing restraints generation approaches are strictly dependent on base-identities and the detection of hydrogen-bonding patterns that requires accurate relative positioning of paired-bases (LibG (59), Phenix suite (60)). An exception here is a pipeline implemented in PDB_REDO (61) relying on base-pair assignment by DSSR suite, which is by design less sensitive to the structure distortions (62).

The presented *doubleHelix* benchmarks revealed a plausible sequence-register error in an expansion segment ES7a of a mammalian ribosome model deposited in the PDB (PDB entry 3j7q/EMDB entry EMD-2650). Expansion segments (ES) are present only in eukaryotic ribosomes and exhibit a surprising level of variability between different organisms. Nevertheless, the function of ES remains poorly understood, which makes them an important research target (63). Ribosomes are usually highly conserved across all kingdoms of live and ribosome models already available in the PDB can be often used to greatly simplify model building and refinement process of newly determined structures. The high variability of the ES at a structural and sequence level makes them one of the few rRNA regions that require in-depth modelling. Not surprisingly, this results in sporadic errors, which may hinder efforts aimed at understanding ES function. The problem is even more important for newly determined nucleic-acid complex structures for which at least partial models that could be used for model building are not available in the PDB. This makes the *doubleHelix* program particularly useful in structure determination using cryo-EM and MX as a reliable tool for nucleic-acid model sequence assignment and validation.

To facilitate the use of *doubleHelix* for model validation, it has been incorporated into a previously released, open-source sequence-assignment validation tool *checkMySequence* that can now process complete protein-nucleic acid complexes. The method enables a conceptually simple and fully automated detection of the most common sequence assignment issues in models of proteins, RNA, DNA, and their complexes that include register-shifts, sequence mismatches (single-residue differences between model and

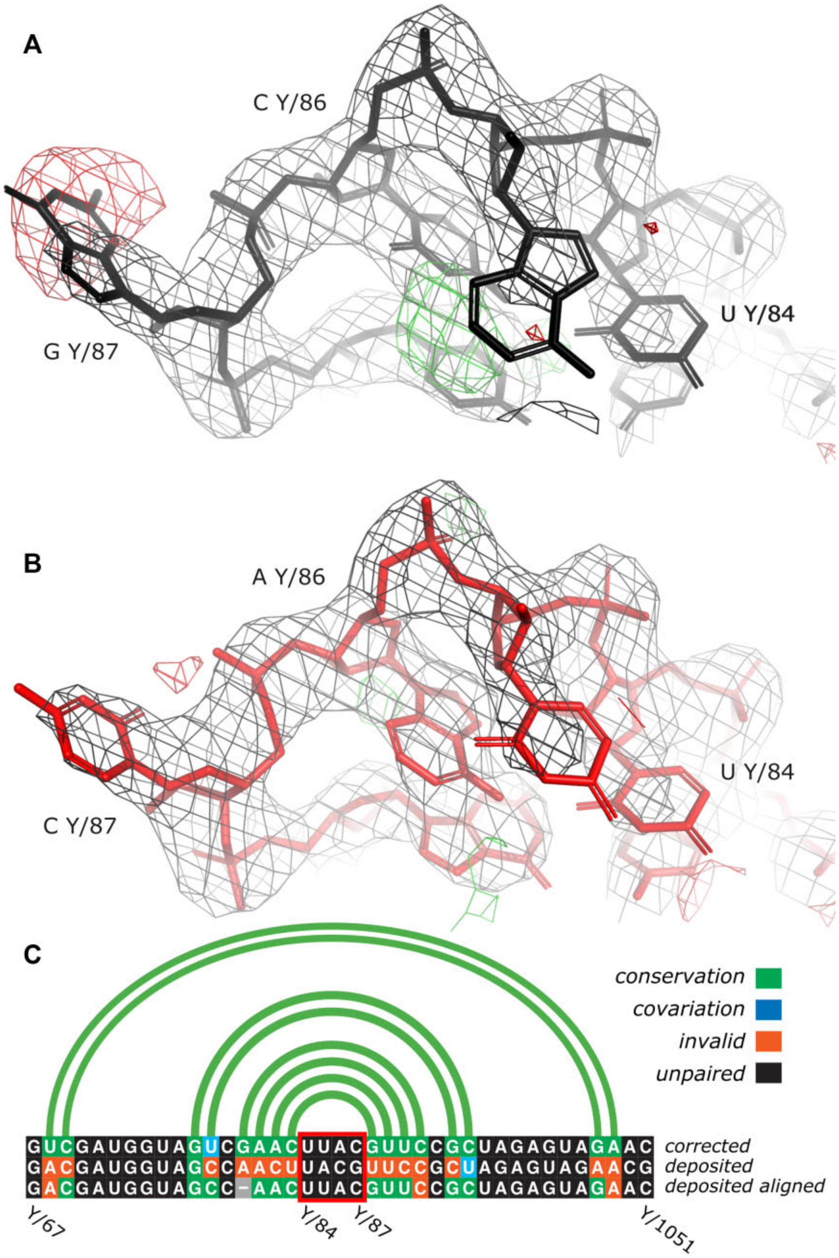

A
C Y/86
G Y/87
U Y/84
B
A Y/86
C Y/87
U Y/84
C
conservation 🟩
covariation 🟦
invalid 🟧
unpaired ⬛
GUCGAUGGUAGUCGAACUUACGUUCCGCUAGAGUAGAAC corrected
GACGAUGGUAGCCAACUUACGUUCCGCUAGAGUAGAACG deposited
GACGAUGGUAGCC-AACUUACGUUCCGCUAGAGUAGAAC deposited aligned
Y/67
Y/84
Y/87
Y/1051

Figure 8. Fragment of 5S rRNA model from a crystal structure of large ribosomal subunit of *S. aureus* refined at 3.5 Å resolution. High negative (red) and positive (green) difference density map values near Guanine Y/87 and Cytosine Y/86 correspond to excess and missing atoms in the model (**A**). After re-assigning the model to a sequence of different 5S gene variant and subsequent restrained refinement in REFMAC5 and PDB_REDO map-model fit clearly improves (**B**). The new sequence also clearly improves base-pairing pattern of the model (**C**). For clarity, only a sequence fragment including the stem loop depicted in panels (A, B) is shown (highlighted with a red box). Most sequence mismatches between the two 5S sequence variants can be corrected by introducing a gap to the alignment (shown as a grey box) that corresponds to a single base register-shift identified by *checkMySequence*. Maximum likelihood combined 2mFo-DFc and difference mFo-DFc maps on panels (A, B) are contoured at 2σ and 3σ level, respectively. Secondary structure presented on panel (C) was identified in the corrected model using ClaRNA. The figure was prepared using Pymol (31) and R-chie webserver (53).

target sequence), problems with residue numbering (e.g. continuous residue numbering ignoring parts of a model that were not traced). With an example of bacterial ribosome crystal structure, I show that the *checkMySequence* can successfully identify errors in both protein and nucleic acid components of complex structures, resulting from model tracing issues and errors in reference sequences that would be otherwise very difficult to identify. Owing to its performance and full automation, *checkMySequence* is applicable to the analysis of very large models. For example, validation of a complete cryo-EM structures of 80S ribosome at 3.4 Å resolution discussed in this work, with 48 chains and over 11 000 protein and nucleic acid residues takes less than six minutes on a laptop.

DATA AVAILABILITY

The doubleHelix program source code and installation instructions are available at https://gitlab.com/gchojnowski/doublehelix. The doubleHelix sequence assignment algorithm has been added to the *checkMySequence* program to enable comprehensive validation of sequence assignment in the models of protein-nucleic acid complexes. The updated method is available at https://gitlab.com/gchojnowski/checkmysequence. Corrected models of mammalian and bacterial ribosome structures presented in this work are available at https://doi.org/10.5281/zenodo.7650444.

ACKNOWLEDGEMENTS

I would like to thank all the microscopists who decided to deposit half-maps to EMDB which was of great help while developing *doubleHelix*. I would also like to thank Sean Eddy for his help with the configuration of INFERNAL suite. I am very grateful to Daniel J. Rigden and Katherine S. H. Beckham for critical reading of the manuscript and very helpful comments.

FUNDING

Funding for open access charge: European Molecular Biology Laboratory.

Conflict of interest statement. None declared.

REFERENCES

1. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
2. Chen,J., Wang,Q., Malone,B., Llewellyn,E., Pechersky,Y., Maruthi,K., Eng,E.T., Perry,J.K., Campbell,E.A., Shaw,D.E. *et al.* (2022) Ensemble cryo-EM reveals conformational states of the nsp13 helicase in the SARS-CoV-2 helicase replication-transcription complex. *Nat. Struct. Mol. Biol.*, **29**, 250–260.
3. Jumper,J. and Hassabis,D. (2022) Protein structure predictions to atomic accuracy with AlphaFold. *Nat. Methods*, **19**, 11–12.
4. Baek,M., DiMaio,F., Anishchenko,I., Dauparas,J., Ovchinnikov,S., Lee,G.R., Wang,J., Cong,Q., Kinch,L.N., Schaeffer,R.D. *et al.* (2021) Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, **373**, 871–876.
5. Mosalaganti,S., Obarska-Kosinska,A., Siggel,M., Taniguchi,R., Turonova,B., Zimmerli,C.E., Buczak,K., Schmidt,F.H., Margiotta,E., Mackmull,M.T. *et al.* (2022) AI-based structure prediction empowers integrative structural analysis of human nuclear pores. *Science*, **376**, eabm9506.
6. Baek,M., McHugh,R., Anishchenko,I., Baker,D. and DiMaio,F. (2022) Accurate prediction of nucleic acid and protein-nucleic acid complexes using RoseTTAFoldNA. bioRxiv doi: https://doi.org/10.1101/2022.09.09.507333, 10 September 2022, preprint: not peer reviewed.
7. Townshend,R.J.L., Eismann,S., Watkins,A.M., Rangan,R., Karelina,M., Das,R. and Dror,R.O. (2021) Geometric deep learning of RNA structure. *Science*, **373**, 1047–1051.
8. Hoh,S.W., Burnley,T. and Cowtan,K. (2020) Current approaches for automated model building into cryo-EM maps using Buccaneer with CCP-EM. *Acta Crystallogr. D: Struct. Biol.*, **76**, 531–541.
9. Chojnowski,G., Sobolev,E., Heuser,P. and Lamzin,V.S. (2021) The accuracy of protein models automatically built into cryo-EM maps with ARP/wARP. *Acta Crystallogr. D Struct. Biol.*, **77**, 142–150.
10. Terwilliger,T.C., Adams,P.D., Afonine,P.V. and Sobolev,O.V. (2018) A fully automatic method yielding initial models from high-resolution cryo-electron microscopy maps. *Nat. Methods*, **15**, 905–908.
11. Keating,K.S. and Pyle,A.M. (2012) RCrane: semi-automated RNA model building. *Acta Crystallogr. D: Biol. Crystallogr.*, **68**, 985–995.
12. Casañal,A., Lohkamp,B. and Emsley,P. (2020) Current developments in Coot for macromolecular model building of electron cryo-microscopy and crystallographic data. *Protein Science*, **29**, 1055–1064.
13. Croll,T.I. (2018) ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. *Acta Crystallogr. D: Struct. Biol.*, **74**, 519–530.
14. Nakamura,A., Meng,H., Zhao,M., Wang,F., Hou,J., Cao,R. and Si,D. (2023) Fast and automated protein-DNA/RNA macromolecular complex modeling from cryo-EM maps. *Brief. Bioinf.*, **24**, bbac632.
15. Jamali,K., Kall,L., Zhang,R., Brown,A., Kimanius,D. and Scheres,S. (2023) Automated model building and protein identification in cryo-EM maps. bioRxiv doi: https://doi.org/10.1101/2023.05.16.541002, 16 May 2023, preprint: not peer reviewed.
16. Wlodawer,A., Dauter,Z., Porebski,P.J., Minor,W., Stanfield,R., Jaskolski,M., Pozharski,E., Weichenberger,C.X. and Rupp,B. (2018) Detect, correct, retract: how to manage incorrect structural models. *FEBS J.*, **285**, 444–466.
17. Prisant,M.G., Williams,C.J., Chen,V.B., Richardson,J.S. and Richardson,D.C. (2020) New tools in MolProbity validation: caBLAM for CryoEM backbone, UnDowser to rethink "waters," and NGL Viewer to recapture online 3D graphics. *Protein Sci.*, **29**, 315–329.
18. Lawson,C.L., Kryshtafovych,A., Adams,P.D., Afonine,P.V., Baker,M.L., Barad,B.A., Bond,P., Burnley,T., Cao,R. and Cheng,J. (2021) Cryo-EM model validation recommendations based on outcomes of the 2019 EMDataResource challenge. *Nat. Methods*, **18**, 156–164.
19. Chojnowski,G. (2022) Sequence-assignment validation in cryo-EM models with checkMySequence. *Acta Crystallogr. D*, **78**, 806–816.
20. Sánchez Rodríguez,F., Chojnowski,G., Keegan,R.M. and Rigden,D.J. (2022) Using deep-learning predictions of inter-residue distances for model validation. *Acta Crystallogr. D*, **78**, 1412–1427.
21. Leontis,N.B., Stombaugh,J. and Westhof,E. (2002) The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.*, **30**, 3497–3531.
22. Stombaugh,J., Zirbel,C.L., Westhof,E. and Leontis,N.B. (2009) Frequency and isostericity of RNA base pairs. *Nucleic Acids Res.*, **37**, 2294–2312.
23. Niedzialkowska,E., Gasiorowska,O., Handing,K.B., Majorek,K.A., Porebski,P.J., Shabalin,I.G., Zasadzinska,E., Cymborowski,M. and Minor,W. (2016) Protein purification and crystallization artifacts: the tale usually not told. *Protein Sci.*, **25**, 720–733.
24. Tegunov,D., Xue,L., Dienemann,C., Cramer,P. and Mahamid,J. (2021) Multi-particle cryo-EM refinement with M visualizes ribosome-antibiotic complex at 3.5 A in cells. *Nat. Methods*, **18**, 186–193.
25. Ho,C.M., Li,X., Lai,M., Terwilliger,T.C., Beck,J.R., Wohlschlegel,J., Goldberg,D.E., Fitzpatrick,A.W.P. and Zhou,Z.H. (2020) Bottom-up structural proteomics: cryoEM of protein complexes enriched from the cellular milieu. *Nat. Methods*, **17**, 79–85.
26. Su,C.C., Lyu,M., Morgan,C.E., Bolla,J.R., Robinson,C.V. and Yu,E.W. (2021) A 'build and retrieve' methodology to simultaneously solve cryo-EM structures of membrane proteins. *Nat. Methods*, **18**, 69–75.
27. Skalidis,I., Kyrilis,F.L., Tu,C., Hamdi,F., Chojnowski,G. and Kastritis,P.L. (2022) Cryo-EM and artificial intelligence visualize endogenous protein community members. *Structure*, **30**, 575–589.
28. Chojnowski,G., Simpkin,A.J., Leonardo,D.A., Seifert-Davila,W., Vivas-Ruiz,D.E., Keegan,R.M. and Rigden,D.J. (2022) findMySequence: a neural-network-based approach for identification of unknown proteins in X-ray crystallography and cryo-EM. *IUCrJ*, **9**, 86–97.
29. Kovalevskiy,O., Nicholls,R.A. and Murshudov,G.N. (2016) Automated refinement of macromolecular structures at low resolution using prior information. *Acta Crystallogr. D Struct. Biol.*, **72**, 1149–1161.
30. Afonine,P.V., Poon,B.K., Read,R.J., Sobolev,O.V., Terwilliger,T.C., Urzhumtsev,A. and Adams,P.D. (2018) Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. D: Struct. Biol.*, **74**, 531–544.

31. DeLano,W.L. (2002) Pymol: an open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr.*, **40**, 82–92.
32. Colasanti,A.V., Lu,X.J. and Olson,W.K. (2013) Analyzing and building nucleic acid structures with 3DNA. *J. Vis. Exp.*, e4401.
33. Chojnowski,G., Walen,T. and Bujnicki,J.M. (2014) RNA bricks–a database of RNA 3D motifs and their interactions. *Nucleic Acids Res.*, **42**, D123–D131.
34. Walen,T., Chojnowski,G., Gierski,P. and Bujnicki,J.M. (2014) ClaRNA: a classifier of contacts in RNA 3D structures based on a comparative analysis of various classification schemes. *Nucleic Acids Res.*, **42**, e151.
35. Kucukelbir,A., Sigworth,F.J. and Tagare,H.D. (2014) Quantifying the local resolution of cryo-EM density maps. *Nat. Methods*, **11**, 63–65.
36. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
37. Afonine,P.V., Klaholz,B.P., Moriarty,N.W., Poon,B.K., Sobolev,O.V., Terwilliger,T.C., Adams,P.D. and Urzhumtsev,A. (2018) New tools for the analysis and validation of cryo-EM maps and atomic models. *Acta Crystallogr. D Struct. Biol.*, **74**, 814–840.
38. Joosten,R.P., Salzemann,J., Bloch,V., Stockinger,H., Berglund,A.C., Blanchet,C., Bongcam-Rudloff,E., Combet,C., Da Costa,A.L., Deleage,G. *et al.* (2009) PDB_REDO: automated re-refinement of X-ray structure models in the PDB. *J. Appl. Crystallogr.*, **42**, 376–384.
39. Kingma,D.P. and Ba,J. (2014) Adam: A Method for Stochastic Optimization. arXiv doi: https://arxiv.org/abs/1412.6980, 22 December 2014, preprint: not peer reviewed.
40. Butcher,S.E. and Pyle,A.M. (2011) The molecular interactions that stabilize RNA tertiary structure: RNA motifs, patterns, and networks. *Acc. Chem. Res.*, **44**, 1302–1311.
41. Zirbel,C.L., Roll,J., Sweeney,B.A., Petrov,A.I., Pirrung,M. and Leontis,N.B. (2015) Identifying novel sequence variants of RNA 3D motifs. *Nucleic Acids Res.*, **43**, 7504–7520.
42. Nawrocki,E.P. and Eddy,S.R. (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, **29**, 2933–2935.
43. Oliphant,T.E. (2006) In: *A guide to NumPy*. Trelgol Publishing, USA.
44. Virtanen,P., Gommers,R., Oliphant,T.E., Haberland,M., Reddy,T., Cournapeau,D., Burovski,E., Peterson,P., Weckesser,W. and Bright,J. (2020) SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, **17**, 261–272.
45. Grosse-Kunstleve,R.W., Sauter,N.K., Moriarty,N.W. and Adams,P.D. (2002) The Computational Crystallography Toolbox: crystallographic algorithms in a reusable software framework. *J. Appl. Crystallogr.*, **35**, 126–136.
46. Winn,M.D., Ballard,C.C., Cowtan,K.D., Dodson,E.J., Emsley,P., Evans,P.R., Keegan,R.M., Krissinel,E.B., Leslie,A.G.W. and McCoy,A. (2011) Overview of the CCP4 suite and current developments. *Acta Crystallogr. D: Biol. Crystallogr.*, **67**, 235–242.
47. Paszke,A., Gross,S., Massa,F., Lerer,A., Bradbury,J., Chanan,G., Killeen,T., Lin,Z., Gimelshein,N. and Antiga,L. (2019) Pytorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.*, **32**, 1–12.
48. Yesselman,J.D., Denny,S.K., Bisaria,N., Herschlag,D., Greenleaf,W.J. and Das,R. (2019) Sequence-dependent RNA helix conformational preferences predictably impact tertiary structure formation. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 16847–16855.
49. Zhang,C. and Pyle,A.M. (2022) CSSR: assignment of secondary structure to coarse-grained RNA tertiary structures. *Acta Crystallogr. D Struct. Biol.*, **78**, 466–471.
50. Beckham,K.S.H., Ritter,C., Chojnowski,G., Ziemianowicz,D.S., Mullapudi,E., Rettel,M., Savitski,M.M., Mortensen,S.A., Kosinski,J. and Wilmanns,M. (2021) Structure of the mycobacterial ESX-5 type VII secretion system pore complex. *Sci. Adv.*, **7**, eabg9923.
51. Wilm,A., Higgins,D.G. and Notredame,C. (2008) R-coffee: a method for multiple alignment of non-coding RNA. *Nucl. Acids. Res.*, **36**, e52.
52. Pettersen,E.F., Goddard,T.D., Huang,C.C., Meng,E.C., Couch,G.S., Croll,T.I., Morris,J.H. and Ferrin,T.E. (2021) UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Sci.*, **30**, 70–82.
53. Lai,D., Proctor,J.R., Zhu,J.Y. and Meyer,I.M. (2012) R-CHIE: a web server and R package for visualizing RNA secondary structures. *Nucleic Acids Res.*, **40**, e95.
54. Read,R.J. (1986) Improved fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr. A: Found. Crystallogr.*, **42**, 140–149
55. Eyal,Z., Matzov,D., Krupkin,M., Wekselman,I., Paukner,S., Zimmerman,E., Rozenberg,H., Bashan,A. and Yonath,A. (2015) Structural insights into species-specific features of the ribosome from the pathogen Staphylococcus aureus. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E5805–E5814.
56. Terwilliger,T.C., Leibschner,D.L., Croll,T., Williams,C.J., McCoy,A.J., Poon,B.K., Afonine,P., Oeffner,R.D., Richardson,J.S., Read,R.J. and Adams,P.D. (2023) AlphaFold predictionsare valuable hypotheses, and accelerate but do not replace experimental structure determination. bioRxiv doi: https://doi.org/10.1101/2022.11.21.517405, 19 May 2023, preprint: not peer reviewed.
57. Croll,T.I., Williams,C.J., Chen,V.B., Richardson,D.C. and Richardson,J.S. (2021) Improving SARS-CoV-2 structures: peer review by early coordinate release. *Biophys. J.*, **120**, 1085–1096.
58. Ramrath,D.J.F., Niemann,M., Leibundgut,M., Bieri,P., Prange,C., Horn,E.K., Leitner,A., Boehringer,D., Schneider,A. and Ban,N. (2018) Evolutionary shift toward protein-based architecture in trypanosomal mitochondrial ribosomes. *Science*, **362**, eaau7735.
59. Brown,A., Long,F., Nicholls,R.A., Toots,J., Emsley,P. and Murshudov,G. (2015) Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta. Crystallogr. D Biol. Crystallogr.*, **71**, 136–153.
60. Liebschner,D., Afonine,P.V., Baker,M.L., Bunkóczi,G., Chen,V.B., Croll,T.I., Hintze,B., Hung,L.W., Jain,S. and McCoy,A.J. (2019) Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta Crystallogr. D: Struct. Biol.*, **75**, 861–877.
61. de Vries,I., Kwakman,T., Lu,X.J., Hekkelman,M.L., Deshpande,M., Velankar,S., Perrakis,A. and Joosten,R.P. (2021) New restraints and validation approaches for nucleic acid structures in PDB-REDO. *Acta Crystallogr. D: Struct. Biol.*, **77**, 1127–1141.
62. Lu,X.-J., Bussemaker,H.J. and Olson,W.K. (2015) DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res.*, **43**, e142.
63. Fujii,K., Susanto,T.T., Saurabh,S. and Barna,M. (2018) Decoding the function of expansion segments in ribosomes. *Mol. Cell*, **72**, 1013–1020.