

Ensemble refinement of mismodeled cryo-EM RNA structures using all-atom simulations

Received: 14 October 2024

Accepted: 2 May 2025

Published online: 16 May 2025

Elisa Posani¹, Pavel Janoš², Daniel Haack³, Navtej Toor³,
Massimiliano Bonomi⁴, Alessandra Magistrato^{2,5}✉ & Giovanni Bussi^{1,5}✉

The advent of single-particle cryogenic electron microscopy (cryo-EM) has enabled near-atomic resolution imaging of large macromolecules, enhancing functional insights. However, current cryo-EM refinement tools condense all single-particle images into a single structure, which can misrepresent highly flexible molecules like RNAs. Here, we combine molecular dynamics simulations with cryo-EM density maps to better account for the structural dynamics of a complex and biologically relevant RNA macromolecule. Namely, using metainference, a Bayesian method, we reconstruct an ensemble of structures of the group II intron ribozyme, which better matches experimental data, and we reveal inaccuracies of single-structure approaches in modeling flexible regions. An analysis of all RNA-containing structures deposited in the Protein Data Bank reveals that this issue affects most cryo-EM structures in the 2.5–4 Å range. Thus, RNA structures determined by cryo-EM require careful handling, and our method may be broadly applicable to other RNA systems.

RNA molecules play a vital role in cells by acting as both carriers of genetic information and catalysts (ribozymes) to regulate gene expression, proteome diversification, and to perform protein synthesis^{1–3}. RNA function is often dictated by its 3D structure, and this is particularly true when considering catalytic RNA molecules or proteins/RNA and small-molecules/RNA interactions⁴. With the advent of single-particle cryogenic electron microscopy (cryo-EM) techniques, it is now possible to resolve large macromolecular complexes, including RNA-containing systems, with a resolution comparable with that of traditional X-ray diffraction experiments^{5–9}. However, the number of solved RNA structures remains limited in comparison with protein structures. The standard approach for the analysis of cryo-EM data consists in classifying millions of acquired 2D single-particle images into structurally homogeneous 2D class averages, and then reconstructing one or more 3D density maps. The following step, namely generating a 3D model that fits the obtained density map, usually assumes that a single 3D structure can reproduce the density associated with a very large number of 2D images. For high-resolution

maps (<3 Å), structures can be built by using the same software employed for X-ray crystallography^{10,11}. However, for cryo-EM maps with resolutions >4 Å, de novo model building is also adopted^{12–16}.

All the methods mentioned, as single structure approaches, aim at reproducing a cryo-EM potential density map by building a single model. However, applying these approaches to heterogeneous and dynamic biomolecules, where the observed density maps might result from a mixture of structures (conformational states), is highly challenging. This is particularly true for disordered or multi-domain proteins and for RNA systems^{17,18}. Fitting such a potential density map with a single structure might indeed lead to non-biologically-relevant models or structural artifacts if the employed density map originates from a mixture of heterogeneous conformations. In these approaches, structures are normally refined by performing energy minimizations. Accurate MD simulations can also be used for cryo-EM structure refinement^{19,20}. Standard approaches are based on empirical potentials enforcing the atoms positions of the simulated structure to match the measured density maps^{21–27}. Although accounting for some conformational variability in

¹Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy. ²CNR-IOM at SISSA, Trieste, Italy. ³Department of Chemistry and Biochemistry, University of California, San Diego, La Jolla, CA, USA. ⁴Institut Pasteur, Université Paris Cité, CNRS UMR 3528, Computational Structural Biology Unit, Paris, France. ⁵These authors contributed equally: Alessandra Magistrato and Giovanni Bussi. ✉ e-mail: alessandra.magistrato@sisssa.it; bussi@sisssa.it

the space confined to the EM map potential, the MD simulation is here used to search for a global minimum. Importantly, the conformational landscape may become very rugged and feature multiple proximal local minima, possibly causing the fitted structure to become “trapped” within local minima and resulting in structurally poor or functionally irrelevant models. This problem can be tackled using a replica-exchange approach²⁸. More advanced methods allow an automatic inclusion of conformational averaging^{29,30}. Among these methods, metainference-based MD simulations³¹ can be fruitfully used to reconstruct structural ensembles. Metainference is an integrative method for modeling structural ensembles of biological systems using noisy, ensemble-averaged data based on a Bayesian probabilistic framework. In this approach, an ensemble is generated by a multi-replica MD simulation guided by a hybrid energy function. This function combines physico-chemical information with additional spatial restraints that enforce the agreement of the modeled ensemble with the experimental data. Metainference enables: (i) automatically determining the accuracy of the input data; (ii) optimally weighting multiple sources of information based on their relative accuracy; (iii) ultimately modeling structural ensembles by improving the prior description of the system with experimental information. Specifically, metainference enforces the agreement between the experimental density map and an average computed simulating a number of replicas of the system. The replica averaging procedure results in a conformational ensemble that is as close as possible to the underlying MD ensembles^{32,33} and, at the same time, minimizes model discrepancy with respect to experiment³¹. Additionally, the dynamical nature of the refinement process allows to include soft constraints that can enforce independently available structural information. The application of cryo-EM-guided metainference to refine structural ensembles of macromolecular complexes has so far been limited to proteins^{30,34–38}. Importantly, this method has never been tested on large RNA molecules.

Here we apply cryo-EM guided metainference-based MD simulations to the group II intron ribozyme from *Thermosynechococcus elongatus*, a ~800-nucleotide long RNA macromolecule that can function as a catalytic RNA by performing self-splicing reactions and by acting as a retroelement through its insertion into double-stranded DNA³⁹. Besides its biological significance as evolutionary ancestor of the spliceosome and its potential application as a gene editing tool, this structure was selected because it is one of the few cryo-EM RNA-structures deposited in the PDB data bank, and was obtained using a single-structure refinement procedure⁴⁰. We reveal that, for such a highly plastic system, a single structure cannot be at the same time compatible with the experimental data and with the expected structure of base-paired RNA helices, resulting in mis-modeling of some regions. For this system, the remodeling mostly affects flexible regions, which are located in the exterior, solvent-exposed stem loops, and are not phylogenetically conserved. In contrast, functional domains involved in catalysis are well-ordered, exhibit high-quality cryo-EM density, and are phylogenetically conserved. As a result, these core catalytic regions of the ribozyme are relatively rigid and did not require remodeling. Interestingly, a detailed analysis of all the deposited RNA-containing structures revealed that the modeling problem with flexible helical regions broadly applies to RNA-containing cryo-EM derived structures in the 2.5–4 Å resolution range.

Results

Test system and preparation

First, we visually inspected the deposited structure of the group II intron ribozyme (PDB code: 6ME0, cryo-EM map: EMD 9105, resolution 3.6 Å)⁴⁰. The secondary and tertiary structures are reported in Fig. 1a, c. A 38-nucleotide long gap was present, which we modeled using DeepFoldRNA⁴¹ (see Methods). This modeled part was included in all the simulations but not explicitly analyzed, since the reported density map was truncated in this region. In addition, a number of supposedly

canonical RNA helices could be visually seen to be not properly paired (see Fig. 1). A subsequent more systematic analysis, based on a combination of annotation⁴² and secondary structure prediction⁴³ (see Methods), confirmed the presence of six helices that should be present based on the predicted secondary structure, but are not properly folded in the solved structure (Fig. 1 and Supplementary Table 1). These helices are also present in the reference secondary structure of this system⁴⁰, which validates the secondary structure prediction. We then run 2.5 ns-long molecular dynamics (MD) simulations in explicit solvent (see Methods) to remodel correctly these helices by restraining them to the structure of canonical template RNA duplexes with the same sequence (Fig. 1b). To this aim we used the ERMSD metric⁴⁴, which accounts for the presence of properly paired strands and has been used to reversibly fold RNA motifs⁴⁵ or to model structures starting from a coarse-grained prediction⁴⁶. See Supplementary Fig. 1 for a schematic view of the performed simulations.

Ensemble refinement

After having rebuilt a complete structure with properly folded helices, starting from the final structure of the restrained MD simulation, we performed cryo-EM guided metainference simulations³¹, to maximize the accordance with the experimental density map while accounting for the ribozyme plasticity. According to a well-established protocol devised for proteins³⁶, we initially performed a single-replica refinement in which we enforced one single model to match the experimental data. As expected, when this single-replica refinement was run without any restraint on the six remodeled helices, these quickly unfolded. This confirms that the properly folded helices are not directly compatible with the experimental density within the single-structure assumption. Conversely, when restraining their helicity, within this single-replica refinement approach, we observed the unfolding of three additional helices (helices B, D, and I in Fig. 1) possibly due to other structural constraints or to the incompatibility with the experimental map within the single-structure approximation. Indeed, as detailed below, all these helices are recovered and identified as stable in our ensemble refinement simulations. The only way to preserve the folding of all helices listed in Fig. 1 during the single-structure refinement was to explicitly restrain all of them. Next, we performed ensemble refinement metainference simulations by using an increasing number of replicas (8, 16, 32, and 64) initialized taking equally spaced snapshots from the single-structure refinement simulation. We then collected a 10 ns long trajectory for each replica. Interestingly, a minimum number of 8 replicas was required to satisfy the experimental restraints in the metainference MD simulation. Specifically, all the simulations attempted with a lower number of replicas crashed reporting missing convergence in enforcing bond constraints, which indicates that the experimental and helical restraints were mutually incompatible. This observation per se is suggestive of the substantial dynamics of this ribozyme. Importantly, helical restraints were released after the first 5 ns, and only the remaining part of the MD simulations trajectories was used for the analysis. As a result, helices not compatible with the cryo-EM map were allowed to unfold. As a control, the simulation with 16 replicas was repeated using a different force field for the RNA molecule⁴⁷. Another control simulation was performed replacing K⁺ ions with Na⁺.

The most representative structures extracted from the 32-replicas simulations (Fig. 2b) exemplify the inherent plasticity of this macromolecular complex. The most flexible region is the ab initio modeled loop, in agreement with the fact that its structure was eluding experimental determination in the single structure refinement procedure. However, also other parts of the ribozyme display significant dynamics, as was expected from the high values of B-factors, shown in Fig. 2a. The flexibility of the different regions of the ribozyme can be also visualized in Supplementary Fig. 2, where the root-mean-square fluctuations is reported.

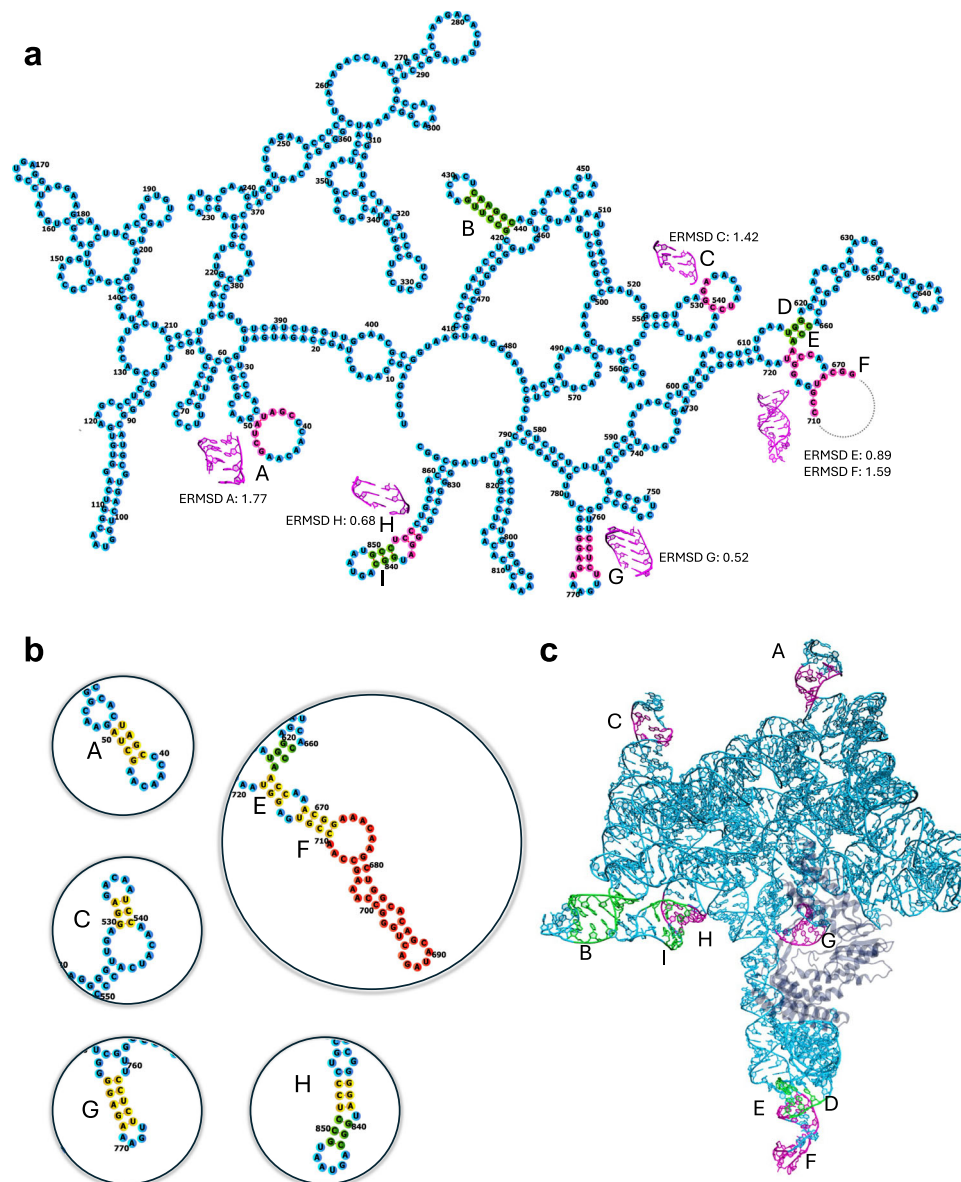


Fig. 1 | Secondary structure and identification of unfolded helices. a Secondary structure of the group II intron from *Thermosynechococcus elongatus*, as obtained by annotating the deposited PDB structure (6ME0). Helices object of investigation are labeled as (A-I). Regions predicted as base paired by secondary structure prediction tools, but that are not properly paired in the PDB are highlighted in magenta. The 3D structure of these regions as reported in the deposited PDB is also shown in magenta. Deviations of the deposited helices from an ideal helix are measured using the ERMSD metric⁴⁴ and reported. Base pairs that were correctly

formed in the PDB structure, but that unfolded during the all-atom simulations are shown in green. The dashed line indicates the missing part in the PDB structure, close to helix F. **b** The regions properly paired after remodeling are enclosed in circles, with the corresponding base pairs highlighted in yellow. The missing domain, modeled with DeepFoldRNA⁴¹, is shown in red. **c** 3D structure as deposited in the PDB (PDB code 6ME0) with the helices highlighted using the same colors and numbering of panel a. The protein part is depicted in transparent blue.

Back-calculation of density maps

The trajectories obtained with metainference simulations were then analyzed by back-calculating the corresponding averaged density map and comparing it to the experimental one (Fig. 2c). The cross-correlation coefficient (CC_{mask}), which accounts for agreement between the calculated and the experimental map inside the mask calculated around the macromolecule⁴⁸, clearly increases with the number of replicas, irrespective of the exact definition of the region used to do the comparison. This is expected since the agreement with the experiment is enforced, and the flexibility of the model increases with the number of replicas. The standard error of the CC_{mask} between the predicted and experimental density markedly decreased with an increasing number of replicas, consistently with the corresponding increase in statistics. Indeed, as discussed below, most of the

conformational heterogeneity originates from the differences between replicas. As a result, 32 replicas appeared to be the best compromise between agreement with experiment and computational cost.

Analysis of helical regions

Next, we monitored the folding of the nine helices that were restrained in the initial part of the metainference simulation. The CC_{mask} computed only on those helical regions was significantly larger than the global CC_{mask} (Supplementary Fig. 3), suggesting that, when properly modeling helical flexibility, the experimental density map could be correctly reproduced.

The actual structure of the nine enforced helices was quantitatively analyzed by computing their ERMSD⁴⁴ with respect to their corresponding template A-form helices. An analysis of the cumulative

distribution of the ERMSD value sampled in the 32-replica-simulation (Supplementary Figs. 4–7) revealed that some of them retained a highly dynamic behavior even in the metainference MD simulation. We note that in metainference simulations each replica contributes equally to the ensemble average, so that the measured populations reflect the fraction of replicas in which each helix is formed. Sample distributions are shown in the main text for helix B, which adopted both a folded and unfolded state (Fig. 2d) during the simulation, and for helix F, which never folded. Fig. 2e displays the fraction of structures with a properly folded helix, for every helix and in each of the simulation setups. Most helices retained the ideal folded state in more than 50% of the ensemble, with helices e and f showing the least percentage of folded structures. However, these helices are placed nearby the gap in the structure that we modeled *ab initio* (Fig. 1b, c). Therefore

they lie in the region of the map with low signal, which corresponds to a high dynamics of the structure. Equivalent analyses performed on the control simulation using (i) the DESRES force field, or (ii) Na⁺ ions replacing the K⁺, are reported in Supplementary Figs. 8 and 9. In both cases the results were not affected by the force field or monovalent ion type adopted in the simulations. These helices stability has also been verified in the plain MD simulation, which was prolonged to 2 microseconds (see Supplementary Fig. 10). Most helices are stable on this timescale, except for helix A, which displays a transition to a trans-Hoogsteen-Watson-Crick pair, and helix E, where the register is shifted and non-canonical interactions appear. These transitions happen after 800 and 1500 ns, respectively, indicating that these helices are meta-stable on this time scale.

Convergence of the simulations

In order to assess how much the percentage of ideally folded helices was depending on the length of the trajectory we also prolonged the 32 replica metainference simulations to 20 ns per replica, achieving very similar results (Supplementary Fig. 11). This finding might be surprising, considering that the time scale of the metainference simulation (10 ns per replica) was much shorter than both the time scale actually required to observe these fluctuations in RNA molecules¹⁷ and the typical time scale spanned by MD simulations⁴⁹. Indeed, the additional 10 ns are mostly equilibrating local fluctuations, and the structural heterogeneity originates from the presence of multiple replicas, as discussed below. The choice of using relatively short simulation timescales was mostly dictated by the high computational cost associated with the on-the-fly back-calculation of the cryo-EM map needed for enforcing agreement with experiment and the necessity to run a large number of replicas in parallel. To verify that a 10 ns long MD simulation within the metainference scheme was indeed sufficient to sample the structural dynamics of the investigated system, we made a systematic comparison of the generated structures, both intra replica and inter replica. Specifically, we performed a full annotation⁴² of the structures produced during (a) a reference 1 μ s MD simulation and (b) during the metainference simulation, and quantitatively estimated their heterogeneity (see Methods). Results for the 32-replica simulation and for the plain MD simulation are shown in Fig. 3, other

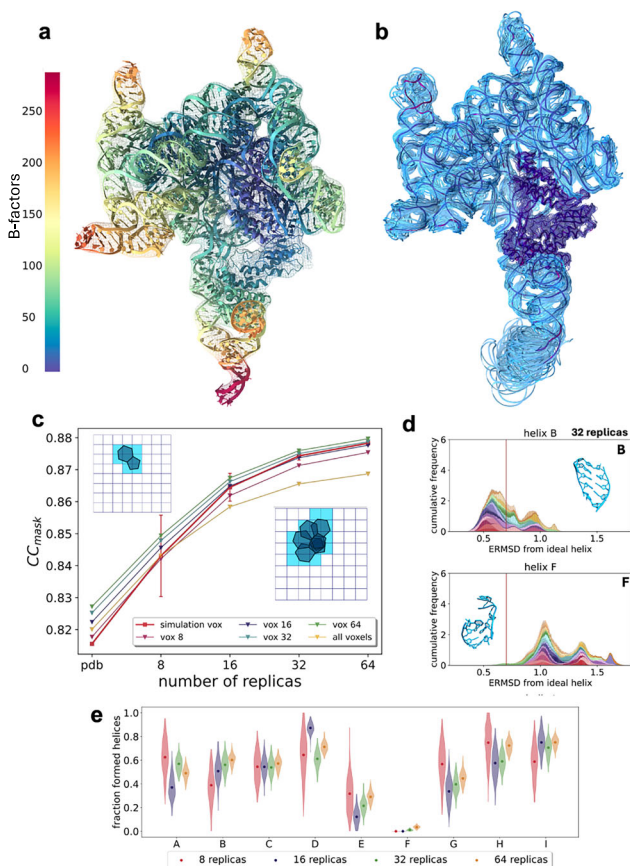


Fig. 2 | Ensemble refinement results. **a** PDB structure colored by B-factor values (increasing value from blue to red) to highlight the most flexible regions of the structure. **b** centroids of the 32 replica simulations (blue), together with the original PDB structure (purple), that display dynamics in the higher B-factor external helices and in the modeled gap part. **c** Accordance with the experimental potential density map, tested with different sets of voxels. The voxels were selected around the initial structure (simulation vox) using all the voxels explored during the replica simulations with 8, 16, 32, and 64 replicas, or including all the voxels. The accordance was monitored for the PDB structure and the multiple-replicas simulations (8, 16, 32, and 64). The error bars have been evaluated bootstrapping over the independent replicas for 200 iterations. Source data are provided as a Source Data file. **d** Examples of the ERMSD distributions in the 32-replica simulation for helix B, which displays both folded and unfolded conformations, and helix F, which is always unfolded. The red line indicates the standard threshold for ERMSD value to discern the two states⁴². Source data are provided as a Source Data file. **e** Fraction of ensembles from the replica simulations with formed helices during the second part of the simulation, where helix restraints were removed. The violin plots show 200 data points obtained by bootstrapping over the independent replicas. The center is the value obtained by using all the different replicas. Source data are provided as a Source Data file.

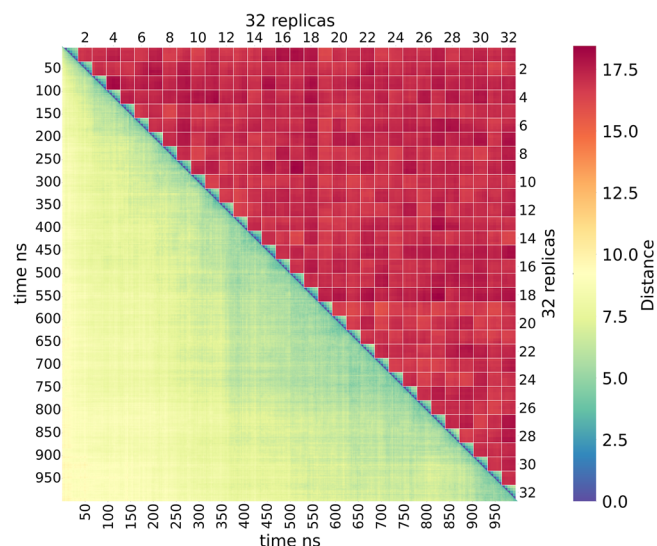


Fig. 3 | Heterogeneity of the generated ensembles. Annotation distances among the 32-replica MD simulations (each 5 ns-long), which quantifies how heterogeneous different blocks or replicas are. Intra- and inter-replica distances are reported along the diagonal and in the upper triangle, respectively. Distances of parts of a 1 μ s long MD simulation used as a reference are shown in lower triangle. Source data are provided as a Source Data file.

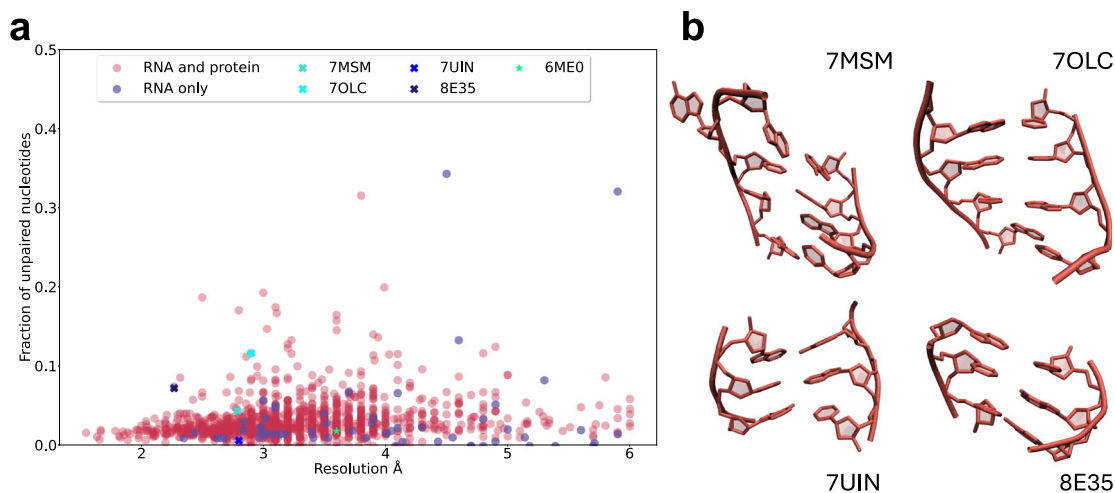


Fig. 4 | PDB wide analysis on mismodeled helices. **a** Structural analysis performed on 1395 deposited cryo-EM structures from the Protein Data Bank, with resolution <6 Å. The structures have been analysed to predict base pairings that are not present in the deposited PDB, but that could exist based on the sequence. The total number of unpaired nucleotides was then divided by the length of the sequence (the values range from 0 when all the base pairings are formed, to 1 if no base pairs

exist). The blue data points represent RNA-only structures, while the pink ones are RNA-protein complexes. The group II intron ribozyme investigated here (6MEO) is highlighted in green and examples of PDBs where many unformed base pairs are found (7MSM, 7OLC, 7UIN, 8E35) are marked with Xs. Source data are provided as a Source Data file. **b** Examples of mis-modeled base pairings, in more pathological cases (7MSM) and less serious ones (8E35), taken from the aforementioned PDBs.

simulations are analyzed in Supplementary Fig. 12. The structural pairwise distance between pairs of snapshots reports on the difference in their pairing and stacking pattern. As a result, the typical distance between structures produced in different replicas is significantly larger than the typical distance between any pair of structures produced in the long MD simulation. In fact, the typical distance between structures produced in the 1 μ s long MD simulation is comparable to the distance observed between structures produced in the same replica of the 10 ns long metainference MD simulation. Therefore the structures extracted from different replicas appear to be more heterogeneous than those sampled from the much longer plain MD simulation. This analysis shows that the use of replicas has a major benefit on sampling and confirms the effectiveness of the multi-replica metainference approach in sampling the structural ensembles of flexible biomacromolecules. To further support this finding, we evaluated the CC_{mask} of the 32 replicas using a single representative frame for every trajectory (i.e., the centroid of each MD simulation replica, Fig. 2b). As a result, we obtained a CC_{mask} value of 0.87, which is compatible with the value obtained by analyzing the whole trajectories (see Supplementary Fig. 13). We also tested different lengths of the simulations, obtaining the same result for 1, 5 and 15 ns. Conversely, the CC_{mask} value of the 1 μ s unbiased MD simulation (see Methods), was only 0.33.

PDB-wide analysis of incorrectly folded RNA helices

To assess the general impact of the canonical single-structure assumption in refining RNA structures, we performed a systematic analysis of all RNA-containing structures determined by cryo-EM currently deposited in the Protein Data Bank (PDB). Specifically, we inspected all the cryo-EM structures with resolution <6 Å that contained more than 100 nucleotides-long RNA strands. Structures were extracted and annotated as discussed in Methods. Some of the structures present in the PDB database were discarded due to issues in the annotation or incomplete nomenclature in the PDB database. Hence, this analysis covers 1395 structures (full list in the associated GitHub repository).

For each structure, we performed a secondary structure analysis similar to that performed on the structure 6MEO. Namely, a secondary structure prediction was done using as constraints the annotated base pairs, and the number of nucleotides that were predicted to be paired, but that were not paired in the annotation, was calculated, and then normalized over the sequence length. At variance with the manual

analysis done for 6MEO, the predicted secondary structures were not validated experimentally. Figure 4a shows a scatter plot comparing the reported resolution of each of the analyzed structures with the fraction of nucleotides in which the secondary structure prediction and the actual annotation of base pairing were not corresponding. The result is expected to be dependent on the specific annotation tool employed, and might falsely report unpaired nucleotides in cases where the secondary structure prediction tools fail in identifying the most stable structure. However, we observed that even structures in the 2.5–4 Å high-resolution range have a very significant fraction of nucleotides that are annotated as unpaired. Examples of improperly folded helices belonging to these high-resolution structures are shown in Fig. 4b, where it can be seen that Watson-Crick hydrogen bonds are not correctly formed.

Discussion

In recent years, remarkable advances in single-particle cryo-EM have enabled the structure determination of an increasing number of biomolecular assemblies with unprecedented detail and resolution. Nevertheless, although cryo-EM maps can frequently reach the atomic-level resolution, the information that they provide often covers a wide range of resolutions across the entire density map, with conformationally variable regions having a substantially decreased resolution. These low-resolution regions often correspond to flexible domains. In these cases, the ensemble of coexisting conformations contributing to the measured cryo-EM images poses challenges to structural biologists. Indeed, the common approach of single structure fitting to cryo-EM potential density maps may result in artifacts due to flexible domains.

Here, we have evaluated the difficulties of using cryo-EM maps of inherently dynamic RNA macromolecules for structural determination. As a prototypical example we chose the group II intron ribozyme, considering an experimental structure that was previously derived using single-structure refinement⁴⁰. This specific structure is representative as it covers a range of relevant cases. The core of the molecule is very well structured and solved at high resolution, while a highly flexible peripheral region could not be determined from the experimental map. At the same time, several helices, that are experimentally known to be part of the secondary structure, were in regions considered rigid enough to be accurately determined by single-structure

refinement but, after careful analysis, resulted to be incorrectly paired in the deposited PDB structure. After having modeled the missing region and enforced the correct pairing in the helical regions, we proceeded with the systematic application of cryo-EM based metainference simulations with an increasing number of replicas (8, 16, 32, and 64), which showed better agreement with the experimental data. At the initial stage of our simulations, we applied helical restraints to enforce a putative secondary structure. These restraints were then removed during the production phase and should therefore be considered only as a means to accelerate system equilibration. If any of the enforced helices were incompatible with the cryo-EM density map, they should unfold during the subsequent simulation. As a control, we repeated the simulation with 16 replicas using a different force field, obtaining compatible results. Similarly, we tested the effect of replacing K^+ with Na^+ in the simulation buffer. Our results clearly indicate how the structural heterogeneity underlying the plasticity of this large ribozyme can not be accounted for by the single PDB structure, which we showed to have a lower CC_{mask} with experiment when compared to the generated ensemble. Indeed, an ensemble refinement based on 32 replica simulation represents for this system a good trade-off between agreement with the experimental data, correct helix modeling, and computational cost of the simulation. The 32 representative structures are also available for download in the associated GitHub repository. We notice that the idea of combining molecular dynamics simulations and experiments to reconstruct the structure of dynamical RNAs is not new. However, this has been so far done on much smaller RNA molecules, mostly using nuclear-magnetic-resonance or scattering data (see ref. 50 for a recent review). Cryo-EM data has been traditionally combined with MD within the single-structure approximation. We note that backcalculation of the density map was performed ignoring the contribution of water and ions. We think these might give important contributions, that could be very relevant in high resolution maps⁵¹, and will be the subject of another work.

In spite of its relatively large computational cost, mostly related to the need to simulate the system in explicit solvent and to back-calculate the cryo-EM map on-the-fly during such simulation, the computational approach used here is ultimately cheap when compared with state-of-the-art unbiased MD simulations, which typically covers the multi-microsecond timescale, yet being less efficient in accounting for the conformational variability of this RNA macromolecule. This is due to the conformational heterogeneity guaranteed by the presence of multiple replicas and of the restraint on their average, which allows us to obtain converged results in relatively short simulation timescales. Hence, even relatively short simulations can be used to generate heterogeneous ensembles provided that enough replicas are simulated, making the approach highly parallel and suitable for high-performance computing setups. Importantly, all the simulations were performed using open source software and all the relevant input scripts and files are made available, so as to facilitate the application of our protocol to other RNA systems.

We suggest the resulting ensemble is a more faithful representation of the dynamics of the group II intron compared to the deposited single-structure model. Interestingly, not all the helices that we initially enforced were formed in the final ensemble, suggesting that some of them might have been incorrectly predicted by the thermodynamic model used here or that their population could be too low to be observable. The secondary structure for the group II intron considered in this work⁴⁰ has been constructed based on homology with a previously published structure⁵², which is consistent with published chemical probing data⁵³. However, this previously published structure was truncated in domain IV, which comprises the helices labeled here as D, E, and F. Hence, it is possible that the stability of helices E and F is limited, as suggested by the populations predicted by our simulations. More generally, if a cryo-EM experiment is conducted under slightly denaturing conditions that disrupt the least stable helices, our

approach can reliably capture this effect through the reconstructed dynamics. Hence, our protocol only relies on the putative secondary structure to facilitate the initial equilibration but can detect if the actual cryo-EM measurement corresponds to a different structure. Overall, the combination of system equilibration using helical restraints and cryo-EM based metainference simulations provides an ensemble that optimally combines physical modeling, previously known information, and experimental data. Other factors could contribute to the fuzzy density observed in the map, such as misalignments of the single particle images. However, based on the high quality of the map, we hypothesized that the main explanation is the conformational dynamics. Our results confirm that conformational dynamics can explain to a large extent the observed density.

It is worth mentioning that the group II intron deposited in ref. 40 had two distinct structures (6ME0 and 6MEC). These structures correspond to separate conformational states from the same cryo-EM dataset. Hence, the analysis performed in ref. 40 was already able to correctly split the two conformations. The remodeling performed here is complementary to this process and works at a finer scale: once the major conformations have been classified, the local dynamics pertaining to each of them can be resolved with molecular dynamics simulations. We also notice that this classification results in a large fraction of the initial particles to be discarded. The number of particles used in the reconstruction of the 3D map would affect its resolution. This would provide less information to the MD simulation. Ideally, at lower resolution, the metainference simulation could be able to sample more heterogeneous ensembles, but observing large conformational changes might be very difficult and computationally expensive. It is also relevant to mention that RNA has a tendency to bind to the air-water interface, which results in denaturation causing disordered density in 3D classes. Our analysis of other low resolution 3D classes led us to the conclusion that they were damaged particles. We decided to focus on a single high-resolution map (6ME0) for this work. A similar analysis could be performed on the alternative conformation (6MEC).

Most importantly, a systematic analysis of the protein data bank, revealed that most of the RNA-containing or RNA-only structures deposited therein are affected by modeling artifacts. Indeed, we observed many long mismodeled helices at an intermediate resolution range. Namely, 65% (11%) of the RNA-containing structures present at least 2% (5%) mismodeled nucleotides. This is most likely due to fitting single-structure models into cryo-EM density maps collected from highly dynamic biomolecules. Furthermore, the use of simplified force fields might also contribute to mismodeling artifacts. These structures must be therefore handled with care, and the computational approach presented here offers an effective and viable way to significantly improve them. In order to process automatically a large number of systems, this analysis relies on secondary structures predicted by RNAfold⁴³. A potential source of error is the use of standard nearest neighbor parameters, measured in NaCl canonical conditions, which might affect the stability of the helices. Even more important is the complete absence of pseudoknots and tertiary contacts in the secondary-structure model. We consider these inaccuracies to be largely alleviated by the constraints applied to the annotated base-pairs, which ensure that a large fraction of the base pairings are consistent with experiment. In addition, we remark that our analysis is focusing on base-pairs that are predicted by the secondary structure model, but not present in the deposited structure. Hence, pseudoknots that are not predicted have no impact on the results. Furthermore, the precise results of this analysis might depend on the annotation method used. Here we used RNAview⁵⁴, which has been optimized to analyze X-ray structures and might be slightly too restrictive in identifying correctly folded helices. However, we made our analysis very conservative by only including nucleobases that are not involved in other interactions on their Watson-Crick edges, which corroborates the idea that a missing canonical pair is the result of an incorrect refinement. For

instance, applying the automatized analysis to 6MEO would result in 16 nucleobases detected as mispaired, to be compared to 22 nucleobases identified by using the experimentally validated secondary structure. An analysis performed with less stringent criteria is available in a previous version of this work⁵⁵. More importantly, the visual inspection of randomly chosen structures clearly shows that modeling RNA helices constrained to cryo-EM maps is affected by the single-structure assumption. Our results suggest that the correct base pairing can be easily recovered in MD simulations by suitably designed restraints, at the price of decreasing the agreement with the experimental data.

We notice that Auto-DRRAFTER¹⁶, a recently developed tool for modeling RNA molecules based on the observed cryo-EM density map, is designed to take into account the secondary structure of the modeled system and, in this sense, should be able to achieve a good performance in modeling helical regions. However, it is still designed to fit a single model to the cryoEM density map. One could therefore envision using this model as starting point for ensemble refinement with our proposed approach. The metainference-based approach share some similarity with variability analysis methods such as cryoDRGN⁵⁶. However, these methods model dynamics by constructing multiple 3D maps corresponding to different conformations from the individual single-particle images. The approach employed here, instead, is capable of accurately describing both the structure and population of substates that have been averaged out in a single 3D reconstruction. In this regard, combining these different approaches could be an interesting direction for future research.

We also remark that, given the growing number of cryo-EM structures in the structural databases and the fact that these models are routinely used to train artificial-intelligence methods to predict new structures from sequence, artifacts in the deposited models could easily propagate to new predictions. We tested the recently released AlphaFold3³⁷ on the group II intron ribozyme, but unfortunately the predicted structure was completely different from the deposited one, making the comparison very difficult. We stress that the performance of AlphaFold3 on systems where RNA-RNA interactions are predominant is currently under investigation⁵⁸. Additional explorations in this direction, for example using cryo-EM data to refine initial models generated by artificial intelligence approaches, are certainly a new avenue that we believe should be further explored in the future. Furthermore, we envision the possibility to use reweighting methods based on the direct analysis of cryo-EM images to further refine the ensembles generated with our approach^{59,60}.

In summary, we have illustrated the application of metainference, an ensemble refinement technique based on all-atom MD simulations, to a large RNA-containing macromolecular complex. The accurate physico-chemical models of RNA and the surrounding environment result in a higher computational cost with respect to the fast model prediction of previously reported and more recent deep-learning based algorithms. However, our approach can provide an accurate structural characterization of RNA systems at atomic level, while accounting for its dynamics, a key requirement for broadening our understanding of RNA functions. Furthermore, in the future one could envision generating several structural ensembles of RNA molecules and complexes using our approach and then train a deep-learning model to predict ensembles from cryo-EM maps in a more computational efficient way, thus enabling large-scale determination of accurate RNA ensembles.

Methods

Model building

We built our model of the group II intron ribozyme starting from the model deposited in the PDB (id: 6MEO). Since this structure presents a gap of 38 nucleotides, we modeled the missing part with the online webserver of DeepFoldRNA⁴¹, using default settings and the fasta sequence of the missing part (residues 673 to 710) with 17 additional nucleotides for a total of 55 nucleotides (complete sequence from 663 to

717: ACCAAACGGAAACAAGCUGGCACAGCAUAGACUGGGCCAAAGC CAACCGUGAGGU). A portion larger than the missing gap was modeled in order to facilitate the alignment and attachment of the modeled part to the rest of the structure. We also estimated the secondary structure of this sequence with RNAfold and checked that the output of DeepFoldRNA was consistent with the RNAfold prediction⁴³. We then aligned the nucleotides in common with the PDB structure and merged the two structures. We also added the 2'-5' lariat bond between residues U1 and A860, using Amber tleap⁶¹. Finally, we ran a minimization in vacuo with GROMACS-2021.5⁶².

System setup and molecular dynamics (MD) simulations

To model the ribozyme at physiological conditions we replaced the Na⁺ ion in the PDB structure with a K⁺ ion. The replacement was done because group II introns are known to be functionally active in the presence of K⁺. We inserted the model in a rhombic dodecahedron periodic box and solvated with 398121 water molecules, 1563 K⁺ ions and 788 Cl⁻ ions to reach charge neutrality at physiological salt concentration of 150 mM. We then run a new minimization with the solvent. The system topology was built with the OL3 force field (FF) for RNA⁶³, OL15 FF for DNA⁶⁴, both with the bsc0 corrections⁶⁵, and ff14SB FF for the protein⁶⁶. Joung-Cheatham parameters were used for the monovalent ions⁶⁷, while Li-Merz parameters were used for the Mg²⁺⁶⁸. Water molecules were described using the TIP3P model⁶⁹. Only deposited Mg²⁺ ions were included in the simulation. Sampling correctly additional divalent cations that might directly bind RNA is not possible in the simulated timescales. In addition, it has been shown that the force field capabilities to reproduce precisely the effects of salts on RNA structural dynamics are limited⁷⁰. All the simulations were performed using GROMACS-2021.5, at constant temperature T = 300 K⁷¹. Plain MD simulations were performed at constant pressure p = 1 bar⁷², whereas metainference simulations were performed at constant volume. We used LINCS bond constraints⁷³ and treated electrostatics with particle-mesh Ewald⁷⁴.

Identification of new base pairings

In order to identify possible base pairings that are not present in the deposited PDB, we started from the original structure 6MEO. We annotated the Watson-Crick base pairings present in the original structure using Barnaba⁴² in dot-bracket notation. Then, we run a prediction of possible additional base-pairings not present in the original structure with RNAfold⁴³. Here we enforced in the prediction all the base-pairings already found ((· ·)). Since this tool does not handle pseudo-knots, we prevented any bond with all the bases involved in this kind of bonds from the annotation ([· ·], { · · }, < · · >). From this prediction, we were able to find new base pairings that were missing in the original structure, and correspond to the helices shown in Fig. 1 and Supplementary Table 1.

Helix restraints

After having identified the mis modeled helices lacking the predicted base-pairings, we generated ideal helix models by using fd_helix (https://casegroup.rutgers.edu/fd_helix.c). Starting from the energy-minimized model described in the previous section, we performed an MD simulation with restraints on all the mis modeled helices (Fig. 1). To this end we used the ERMSD⁴⁴ metric as implemented in PLUMED⁷⁵, which accounts for both the distance between nucleobases and the orientation, and we forced the ERMSD value of the simulated helix to assume a value of 0, corresponding to that of the ideal helix of the same sequence, with harmonic constant 500 kJ/mol. We ran a 2.5 ns long restrained simulation to attain a ERMSD < 0.7 for all targeted helices. The final value of ERMSD from the target helices was < 0.3. These simulations were performed with no additional restraint and in the constant pressure ensemble. Two structures are supposedly equivalent when the ERMSD value is below the threshold of 0.7⁴².

Plain molecular dynamics

Starting from the structure with all the targeted helices in their ideal conformation, we performed a reference 2 μ s long MD simulation, at constant pressure and without any restraint.

Metainference simulations

To perform the metainference simulations, we followed the procedure described in ref. 36 and at <https://github.com/COSBILab/EMMIVox>. This method requires a modified version of PLUMED 2.9 and will be available in the coming version 2.10. Namely, we started from the structure with all the targeted helices in their ideal conformation and we selected the experimental density map within 3.5 Å from the starting model. We decided not to use the entire map since the computational cost of the simulations would have been much larger without a concrete improvement of the results. Using this cropped map, we initially ran a single-structure refinement for 10 ns, enforcing one single conformation to match the experimental data. In this single-structure refinement step, we sampled the B-factors using a Monte Carlo approach to maximize the accordance with the experimental data. During this step, we also kept the restraints on the helices. The single-structure refinement resulted in the unfolding of 3 additional helices (Fig. 1, in green). Therefore we performed again the refinement protocol by adding restraints also to these 3 helices. The refined model that retained all properly folded helices was the starting point of the metainference multi-replica simulations. Multiple metainference simulations were then performed using a different number of replicas: 8, 16, 32 and 64. The initial frames of each replica were selected from the single-structure refinement trajectory as equally spaced. In the first 5 ns of each simulation, we kept the restraints on all the helices in Fig. 1 to prevent them from unfolding. In the second part, we removed the restraints to allow the system to explore the conformational space without additional restraints beside the cryo-EM density map. In all the cryo-EM guided simulations, agreement with density maps was enforced every 4 steps to decrease the computational cost⁷⁶.

Accordance with the experimental data

In order to assess the accordance of the simulated structures with the experimental data, we followed the protocol described in ref. 36. We generated a map for each frame of the trajectory and evaluated the average map along the trajectory. We then performed a linear regression among the voxels of this map and the corresponding ones in the experimental potential density map and we calculated the Pearson correlation coefficient (here called CC_{mask}). In order to assess the robustness of the results, we used different sets of voxels for the calculation of the CC_{mask} . Namely, we selected (i) the voxels around the initial structure, which are those where we enforced the experimental data during the simulations; (ii) all the voxels explored during the simulations of the replicas (8, 16, 32 and 64, separately); (iii) all the voxels in the map. To test a single structure, we estimated the map that would be generated from the structure, limited to the selected voxels.

RMSF

During the single structure refinement, we estimated new B-factors using a Monte Carlo approach, needed to maximize the accordance with the experimental density map. Then, for each ensemble-refinement simulation, we compared the average RMSF per nucleotide, among the metainference replicas, with the RMSF obtained from the B-factors determined during single-structure refinement using:

$$BF = \frac{8}{3} \pi^2 rmsf^2 \quad (1)$$

From Supplementary Fig. 2, it can be seen that the RMSF values do not depend on the number of replicas used during the simulation. In addition, there is a high correlation between the RMSF values obtained

from the single-structure refinement B-factors and those calculated from the metainference ensembles. As shown visually in Supplementary Fig. 2, the targeted helices display peaks in the RMSF, denoting a greater dynamics. This is particularly true for the modeled part in the gap. Interestingly, this confirms the hypothesis that the corrupted density in that region is due to a higher mobility of the molecule.

Distances

To assess the distances between the replicas, we started by dividing the second part of the trajectory of each replica (after removing the restraints) into 5 parts, each 1 ns long, and annotated each frame using Barnaba. For every part of the trajectory, we counted how many times each base pairing⁷⁷ and stacking (using the Barnaba definition) occurred, thus obtaining a histogram of interactions along the trajectory. We evaluated the distances between parts of the trajectories by calculating the Euclidean distance between the corresponding histograms divided by number of frames. The same procedure was applied to the unbiased MD simulation trajectory, that we instead divided into 1000 parts of 1 ns each.

Base pairing analysis of the Protein Data Bank

For the systematic analysis of the base pairs in the experimental structures deposited in the PDB database, we developed an automated protocol. Given the difficulty of doing an explicit validation for the large number of analyzed structures, the approach used here to identify incorrectly paired nucleobases was more conservative than the one that we used to analyze 6ME0, which is discussed above. Specifically, we performed the following steps: (i) download of the structures (PDB or cif format); (ii) annotation of the Watson-Crick base pairings present in the structure using the RCSB tool RNAView⁵⁴, using the default parameters. (iii) From the annotated interactions, we extracted the Watson-Crick pairings, with standard base pairings (AU and GC), GU pairs and non-canonical Watson-Crick pairs. We rewrote the annotation in dot-bracket notation, inserting all the aforementioned pairs. All the chemically modified nucleotides have been excluded from the following analysis, hence we called them as a generic X in the sequence; (iv) rewriting the sequences of the RNA molecules in FASTA format; (v) prediction of possible additional base-pairings not present in the original structure with RNAfold. As explained previously, we enforced in the prediction all the base-pairings found in the PDB and prevented any bond with all the bases involved in pseudoknots. Also the nucleotides that are already involved in an interaction that involves the Watson-Crick edge⁷⁷, have been restrained out of the prediction. In addition, if there were multiple sequences in the structure, we added a UUUU sequence to connect them, and prevented RNAfold to pair these 4 nucleotides with the rest of the structure. From the prediction, we were able to propose new base pairings that were missing in the original structure; (vi) These were finally filtered based on the distances between the two nucleotides in the pair (N1/N9, N1 for pyrimidines and N9 for purines). The correct distance in a proper Watson-Crick bond should be around 9 Å⁷⁸. As we are aimed to find non properly paired nucleotides, this threshold was softened to 12 Å. This value corresponds to the largest of the distances between the pairs considered for 6ME0 (helices a to h in Fig. 1 and Supplementary Table 1).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All the input files and centroids corresponding to the 32 replicas are available at on GitHub <https://github.com/ElisPo/Cryo-EM-refinement/>. Metainference input files are available on the PLUMED NEST (plumID:24.016)⁷⁹. Full trajectories for the metainference simulations,

excluding solvent, are available on Zenodo at <https://doi.org/10.5281/zenodo.12761120>. Source data are provided with this paper.

Code availability

GROMACS 2021.5 is available at <https://www.gromacs.org>. PLUMED 2.10 is available at <https://www.plumed.org>. Both codes are available under an open source license. Scripts used to prepare input files and to run structure analysis are available at <https://github.com/ElisPo/Cryo-EM-refinement>, released as v1.0.0, on <https://doi.org/10.5281/zenodo.15039869>.

References

- Gilbert, W. Origin of life: the RNA world. *Nature* **319**, 618 (1986).
- Doudna, J. A. & Cech, T. R. The chemical repertoire of natural ribozymes. *Nature* **418**, 222–228 (2002).
- Morris, K. V. & Mattick, J. S. The rise of regulatory RNA. *Nat. Rev. Genet.* **15**, 423–437 (2014).
- Vicens, Q. & Kieft, J. S. Thoughts on how to think (and talk) about RNA structure. *Proc. Natl. Acad. Sci. USA* **119**, e2112677119 (2022).
- Kühlbrandt, W. The resolution revolution. *Science* **343**, 1443–1444 (2014).
- Galej, W. P., Toor, N., Newman, A. J. & Nagai, K. Molecular mechanism and evolution of nuclear pre-mrna and group ii intron splicing: insights from cryo-electron microscopy structures. *Chem. Rev.* **118**, 4156–4176 (2018).
- Chari, A. & Stark, H. Prospects and limitations of high-resolution single-particle cryo-electron microscopy. *Annu. Rev. Biophys.* **52**, 391–411 (2023).
- Ma, H., Jia, X., Zhang, K. & Su, Z. Cryo-em advances in RNA structure determination. *Signal Transduct. Target. Ther.* **7**, 58 (2022).
- Bonilla, S. L. & Kieft, J. S. The promise of cryo-em to explore RNA structural dynamics. *J. Mol. Biol.* **434**, 167802 (2022).
- Terwilliger, T. C. et al. Iterative model building, structure refinement and density modification with the phenix autobuild wizard. *Acta Crystallogr. D Biol. Crystallogr.* **64**, 61–69 (2008).
- Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
- Kappel, K. et al. De novo computational RNA modeling into cryo-em maps of large ribonucleoprotein complexes. *Nat. Methods* **15**, 947–954 (2018).
- Terashi, G. & Kihara, D. De novo main-chain modeling for em maps using mainmast. *Nat. Commun.* **9**, 1618 (2018).
- Pfab, J., Phan, N. M. & Si, D. Deeptimizer for fast de novo cryo-em protein structure modeling and special studies on cov-related complexes. *Proc. Natl. Acad. Sci. USA* **118**, e2017525118 (2021).
- He, J. & Huang, S.-Y. Full-length de novo protein structure determination from cryo-em maps using deep learning. *Bioinformatics* **37**, 3480–3490 (2021).
- Ma, H. et al. Auto-drafter: automated RNA modeling based on cryo-em density. *Methods Mol. Biol.* **2568**, 193–211 (2023).
- Al-Hashimi, H. M. & Walter, N. G. Rna dynamics: it is about time. *Curr. Opin. Struct. Biol.* **18**, 321–329 (2008).
- Bussi, G. et al. Rna dynamics from experimental and computational approaches. *Structure* **32**, 1281–1287 (2024).
- Fraser, J. S., Lindorff-Larsen, K. & Bonomi, M. What will computational modeling approaches have to say in the era of atomistic cryo-em data? *J. Chem. Inf. Model.* **60**, 2410–2412 (2020).
- Nierzwicki, Ł. & Palermo, G. Molecular dynamics to predict cryo-em: capturing transitions and short-lived conformational states of biomolecules. *Front. Mol. Biosci.* **8**, 641208 (2021).
- Trabuco, L. G., Villa, E., Schreiner, E., Harrison, C. B. & Schulten, K. Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and x-ray crystallography. *Methods* **49**, 174–180 (2009).
- Kaushal, P. S. et al. Cryo-em structure of the small subunit of the mammalian mitochondrial ribosome. *Proc. Natl. Acad. Sci. USA* **111**, 7284–7289 (2014).
- Kim, D. N. et al. Cryo_fit: Democratization of flexible fitting for cryo-em. *J. Struct. Biol.* **208**, 1–6 (2019).
- Bonilla, S. L., Sherlock, M. E., MacFadden, A. & Kieft, J. S. A viral RNA hijacks host machinery using dynamic conformational changes of a tRNA-like structure. *Science* **374**, 955–960 (2021).
- Mori, T., Terashi, G., Matsuoka, D., Kihara, D. & Sugita, Y. Efficient flexible fitting refinement with automatic error fixing for de novo structure modeling from cryo-em density maps. *J. Chem. Inf. Model.* **61**, 3516–3528 (2021).
- Bonilla, S. L., Vicens, Q. & Kieft, J. S. Cryo-em reveals an entangled kinetic trap in the folding of a catalytic RNA. *Sci. Adv.* **8**, eabq4144 (2022).
- Blau, C., Yvonesdotter, L. & Lindahl, E. Gentle and fast all-atom model refinement to cryo-em densities via a maximum likelihood approach. *PLoS Comput. Biol.* **19**, e1011255 (2023).
- Dodd, T., Yan, C. & Ivanov, I. Simulation-based methods for model building and refinement in cryoelectron microscopy. *J. Chem. Inf. Model.* **60**, 2470–2483 (2020).
- Bonomi, M. et al. Bayesian weighing of electron cryo-microscopy data for integrative structural modeling. *Structure* **27**, 175–188 (2019).
- Garib Singh, R.-A. A. et al. Rational design of asct2 inhibitors using an integrated experimental-computational approach. *Proc. Natl. Acad. Sci. USA* **118**, e2104093118 (2021).
- Bonomi, M., Cavalli, A. & Vendruscolo, M. Metainference: a bayesian inference method for heterogeneous systems. *Sci. Adv.* **2**, e1501177 (2016).
- Cavalli, A., Camilloni, C. & Vendruscolo, M. Molecular dynamics simulations with replica-averaged structural restraints generate structural ensembles according to the maximum entropy principle. *J. Chem. Phys.* **138**, 094112 (2013).
- Roux, B. & Weare, J. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *J. Chem. Phys.* **138**, 084107 (2013).
- Brotzakis, Z. F. et al. A structural ensemble of a tau-microtubule complex reveals regulatory tau phosphorylation and acetylation mechanisms. *ACS Cent. Sci.* **7**, 1986–1995 (2021).
- Brotzakis, Z. F., Löhr, T. & Vendruscolo, M. Determination of intermediate state structures in the opening pathway of sars-cov-2 spike using cryo-electron microscopy. *Chem. Sci.* **12**, 9168–9175 (2021).
- Hoff, S. E., Thomasen, F. E., Lindorff-Larsen, K. & Bonomi, M. Accurate model and ensemble refinement using cryo-electron microscopy maps and bayesian inference. *PLoS Comput. Biol.* **20**, e1012180 (2024).
- Faidon Brotzakis, Z. et al. Determination of the structure and dynamics of the fuzzy coat of an amyloid fibril of iapp using cryo-electron microscopy. *Biochemistry* **62**, 2407–2416 (2023).
- Kim, S., Mollaei, P., Farimani, A. B. & Robinson, A. S. Characterization of phosphorylated tau-microtubule complex with molecular dynamics (md) simulation. Preprint at arXiv: 2312.11700 (2023).
- Lambowitz, A. M. & Zimmerly, S. Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb. Perspect. Biol.* **3**, a003616 (2011).
- Haack, D. B. et al. Cryo-em structures of a group ii intron reverse splicing into DNA. *Cell* **178**, 612–623 (2019).
- Pearce, R., Omenn, G. S. & Zhang, Y. De novo RNA tertiary structure prediction at atomic resolution using geometric potentials from deep learning. Preprint at <https://doi.org/10.1101/2022.05.15.491755> (2022).
- Bottaro, S. et al. Barnaba: software for analysis of nucleic acid structures and trajectories. *RNA* **25**, 219–231 (2019).
- Lorenz, R. et al. Viennarna package 2.0. *Algorithms Mol. Biol.* **6**, 1–14 (2011).

44. Bottaro, S., Di Palma, F. & Bussi, G. The role of nucleobase interactions in RNA structure and dynamics. *Nucleic Acids Res.* **42**, 13306–13314 (2014).
45. Bottaro, S., Banáš, P., Sponer, J. & Bussi, G. Free energy landscape of gaga and uucg RNA tetraloops. *J. Phys. Chem. Lett.* **7**, 4032–4038 (2016).
46. Poblete, S., Bottaro, S. & Bussi, G. A nucleobase-centered coarse-grained representation for structure prediction of RNA motifs. *Nucleic Acids Res.* **46**, 1674–1683 (2018).
47. Tan, D., Piana, S., Dirks, R. M. & Shaw, D. E. RNA force field with accuracy comparable to state-of-the-art protein force fields. *Proc. Natl. Acad. Sci. USA* **115**, E1346–E1355 (2018).
48. Jiang, J.-S. & Brünger, A. T. Protein hydration observed by x-ray diffraction: solvation properties of penicillopepsin and neuraminidase crystal structures. *J. Mol. Biol.* **243**, 100–115 (1994).
49. Sponer, J. et al. RNA structural dynamics as captured by molecular simulations: a comprehensive overview. *Chem. Rev.* **118**, 4177–4338 (2018).
50. Bernetti, M. & Bussi, G. Integrating experimental data with molecular simulations to investigate RNA structural dynamics. *Curri. Open. Struct. Biol.* **78**, 102503 (2023).
51. Kretsch, R. C. et al. Complex water networks visualized by cryogenic electron microscopy of RNA. *Nature*. <https://doi.org/10.1038/s41586-025-08855-w> (2025).
52. Robart, A. R., Chan, R. T., Peters, J. K., Rajashankar, K. R. & Toor, N. Crystal structure of a eukaryotic group ii intron lariat. *Nature* **514**, 193–197 (2014).
53. Wiryaman, T. & Toor, N. Shape profiling to probe group ii intron conformational dynamics during splicing. in *Ribozymes: Methods and Protocols*, 171–182 (Springer, 2020).
54. Yang, H. et al. Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Res.* **31**, 3450–3460 (2003).
55. Posani, E. et al. All atom molecular dynamics simulations enable ensemble refinement of flexible and mismodelled cryo-em derived rna structures. Preprint at: <https://www.biorxiv.org/content/10.1101/2024.07.24.604258v2> 2024–07 (2024).
56. Zhong, E. D., Bepler, T., Berger, B. & Davis, J. H. Cryodrgn: reconstruction of heterogeneous cryo-em structures using neural networks. *Nat. Methods* **18**, 176–185 (2021).
57. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature* **630**, 493–500 (2024).
58. Bernard, C., Postic, G., Ghannay, S. & Tah, F. Has AlphaFold3 achieved success for RNA? *Acta. Crystallogr. D Struct. Biol.* **81**, 49–62 (2025).
59. Cossio, P. & Hummer, G. Likelihood-based structural analysis of electron microscopy images. *Curr. Opin. Struct. Biol.* **49**, 162–168 (2018).
60. Tang, W. S. et al. Ensemble reweighting using cryo-em particle images. *J. Phys. Chem. B* **127**, 5410–5421 (2023).
61. Case, D. A. et al. Ambertools. *J. Chem. Inf. Model.* **63**, 6183–6191 (2023).
62. Abraham, M. J. et al. Gromacs: High-performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1**, 19–25 (2015).
63. Zgarbová, M. et al. Refinement of the cornell et al. nucleic acids force field based on reference quantum chemical calculations of glycosidic torsion profiles. *J. Chem. Theory Comput.* **7**, 2886–2902 (2011).
64. Zgarbová, M. et al. Refinement of the sugar–phosphate backbone torsion beta for amber force fields improves the description of z-and b-dna. *J. Chem. Theory Comput.* **11**, 5723–5736 (2015).
65. Pérez, A. et al. Refinement of the amber force field for nucleic acids: improving the description of α/γ conformers. *Biophys. J.* **92**, 3817–3829 (2007).
66. Maier, J. A. et al. ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb. *J. Chem. Theory Comput.* **11**, 3696–3713 (2015).
67. Joung, S. & Cheatham, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J. Phys. Chem. B* **112**, 9020–9041 (2008).
68. Li, P., Roberts, B. P., Chakravorty, D. K. & Merz, K. M. J. Rational design of particle mesh Ewald compatible Lennard-Jones parameters for +2 metal cations in explicit solvent. *J. Chem. Theory Comput.* **9**, 2733–2748 (2013).
69. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).
70. Kuhrova, P., Mlynsky, V., Otyepka, M., Sponer, J. & Banas, P. Sensitivity of the RNA structure to ion conditions as probed by molecular dynamics simulations of common canonical RNA duplexes. *J. Chem. Inf. Model.* **63**, 2133–2146 (2023).
71. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
72. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).
73. Hess, B., Bekker, H., Berendsen, H. J. & Fraaije, J. G. Lincs: a linear constraint solver for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).
74. Essmann, U. et al. A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577–8593 (1995).
75. Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. Plumed 2: new feathers for an old bird. *Comput. Phys. Commun.* **185**, 604–613 (2014).
76. Ferrarotti, M. J., Bottaro, S., Pérez-Villa, A. & Bussi, G. Accurate multiple time step in biased molecular simulations. *J. Chem. Theory Comput.* **11**, 139–146 (2015).
77. Leontis, N. B. & Westhof, E. Geometric nomenclature and classification of rna base pairs. *RNA* **7**, 499–512 (2001).
78. Pietal, M. J., Szostak, N., Rother, K. M. & Bujnicki, J. M. Rnmap2d—calculation, visualization and analysis of contact and distance maps for rna and protein-rna complex structures. *BMC Bioinform.* **13**, 1–11 (2012).
79. The PLUMED consortium. Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods* **16**, 670–673 (2019).

Acknowledgements

Samuel Hoff (Institut Pasteur) is acknowledged for useful discussions. This work has been funded by the Next Generation EU project PRIN 2022 (2022Z4FZE9), EMBO Scientific Exchange Grant (10033) and EURO-HPC (2023R03-136). A.M. and P.J. thank PNRR: National Center for Gene Therapy and Drugs based on RNA Technology CUPB83C22002860006 CN00000004. G.B. thanks the Italian National Centre for HPC, Big Data, and Quantum Computing (grant No. CN000000013). N.T. is supported by a grant from the National Institutes of Health grant R35GM141706. The authors of the open review posted at <https://zenodo.org/records/14004843> are acknowledged for their constructive comments.

Author contributions

A.M. and G.B. conceptualized the study. E.P., P.J., M.B., A.M. and G.B. developed the methodology and conducted the investigation and data curation. E.P., A.M. and G.B. wrote the original draft. E.P., P.J., D.H., N.T., M.B., A.M. and G.B. contributed to review and editing. A.M. and G.B. supervised the project.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-59769-0>.

Correspondence and requests for materials should be addressed to Alessandra Magistrato or Giovanni Bussi.

Peer review information *Nature Communications* thanks Giulia Palermo, who co-reviewed with Mohd Ahsan, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025